Predicting Influential Users in Online Social Network Groups

ANDREA DE SALVE, Institute of Applied Sciences and Intelligent Systems—CNR
PAOLO MORI, Institute of Informatics and Telematics—CNR
BARBARA GUIDI and LAURA RICCI, Department of Computer Science, University of Pisa
ROBERTO DI PIETRO, ICT Division, College of Science and Engineering, Hamad Bin Khalifa University and Institute of Informatics and Telematics—CNR

The widespread adoption of Online Social Networks (OSNs), the ever-increasing amount of information produced by their users, and the corresponding capacity to influence markets, politics, and society, have led both industrial and academic researchers to focus on how such systems could be *influenced*. While previous work has mainly focused on measuring current influential users, contents, or pages on the overall OSNs, the problem of predicting influencers in OSNs has remained relatively unexplored from a research perspective. Indeed, one of the main characteristics of OSNs is the ability of users to create different groups types, as well as to join groups defined by other users, in order to share information and opinions.

In this article, we formulate the Influencers Prediction problem in the context of groups created in OSNs, and we define a general framework and an effective methodology to predict which users will be able to influence the behavior of the other ones in a future time period, based on historical interactions that occurred within the group. Our contribution, while rooted in solid rationale and established analytical tools, is also supported by an extensive experimental campaign. We investigate the accuracy of the predictions collecting data concerning the interactions among about 800,000 users from 18 Facebook groups belonging to different categories (i.e., News, Education, Sport, Entertainment, and Work). The achieved results show the quality and viability of our approach. For instance, we are able to predict, on average, for each group, around a third of what an ex-post analysis will show being the 10 most influential members of that group. While our contribution is interesting on its own and—to the best of our knowledge—unique, it is worth noticing that it also paves the way for further research in this field.

CCS Concepts: • **Information systems** → *Social networking sites*; *Social recommendation*; *Social networks*; • **Human-centered computing** → *Social networks*; *Social recommendation*; *Social network analysis*; • **Theory of computation** → *Models of learning*;

Additional Key Words and Phrases: Influencer prediction, online social network, centrality measures, behavior analysis

This publication was partially supported by award NPRP11S-0109-180242, from the QNRF-Qatar National Research Fund, a member of Qatar Foundation.

Authors' addresses: A. De Salve, Institute of Applied Sciences and Intelligent Systems—CNR, Dhitech scarl Campus Universitario via Monteroni, Lecce 73100, Italy; email: andrea.desalve@isasi.cnr.it; P. Mori, Institute of Informatics and Telematics—CNR, Via Giuseppe Moruzzi, 1, Pisa 56124, Italy; email: paolo.mori@iit.cnr.it; B. Guidi and L. Ricci, Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo, 3, Pisa 56127, Italy; emails: {guidi, laura.ricci}@di.unipi.it; R. Di Pietro, ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Education City, Doha, Qatar, and Institute of Informatics and Telematics—CNR, Via Giuseppe Moruzzi, 1, Pisa 56124, Italy; email: rdipietro@hbku.edu.qa.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2021 Copyright held by the owner/author(s).
1556-4681/2021/04-ART35
https://doi.org/10.1145/3441447

ACM Reference format:
Andrea De salve, Paolo Mori, Barbara Guidi, Laura Ricci, and Roberto di Pietro. 2021. Predicting Influential Users in Online Social Network Groups. *ACM Trans. Knowl. Discov. Data* 15, 3, Article 35 (April 2021), 50 pages.
https://doi.org/10.1145/3441447

1 INTRODUCTION

Over the past few years, Online Social Networks (OSNs) have become one of the most important communication channels among people. They are analyzed in several research fields and for several purposes [26, 61]. Usually, they are used for spreading information, opinions, ideas, and even for advertising events, products, and services. Indeed, OSNs' users post their contents and receive feedback (e.g., comments, reactions) from the other users who approve or criticize the posted material. Most existing OSNs enable their users to create social groups [54], which are communities of users that, typically, are characterized by a common interest and meant to facilitate the sharing of contents among such users.

Some OSN users are able to attract the attention of many other users: Their posts receive a large number of comments and reactions, or they are reposted several times, thus being visible to a very large number of other users. The above introduced users are called *influencers*, and they are believed to be capable of influencing other users' behavior (e.g., purchasing decisions, voting) because of their real (or perceived) authority/trust. Two very large field experiments conducted on Facebook users confirmed the importance of social influence measuring the response of users to ads in terms of clicks [6]. In the context of social groups, the role of influencers is even more clear, because they are able to influence a community with respect to the common interests that characterize it.

The influencers' phenomenon has been studied by researchers from academia and industry [10, 43, 46]. Indeed, the task of identifying current influencers or predicting future ones is of paramount importance, because it brings a wide range of opportunities in many fields, such as science and economy. Quantifying influence allows us to capture real-world properties of users, which are very important for both understanding the evolution of the system [54] and providing practical solutions to relevant real-world tasks, such as finding audience and artists who influence musical tastes of users [41], identifying promoters for a brand [30, 39], measuring the importance of users in micro-blogs or games [7, 42, 57], identifying travel bloggers who affect tourism destinations [38], driving political opinions [56], and affecting health and medical aspects of our life [28, 62]. In most of the previously cited fields, the ability of influencers in conditioning the opinions and the decisions of a large number of other users has also a relevant economic value. For instance, several companies rely on one or more OSNs to maximize the spread of their brand, services, and products. As a matter of fact, the influencer marketing [13] is a new concept defined over the last few years, which is becoming increasingly popular because it is very effective. The advertisement of a product would be even more effective if the influencers promoting it are chosen from the social groups of the OSNs whose topics are related to the product to be advertised. Typically, this market strategy is carried out by identifying a set of *micro-influencers* [5], i.e., users who are relevant for a community and who are able to influence the largest number of community members with respect to the advertisement [18, 47]. The very same reward of an influencer for conducting an advertising campaign on a given OSN typically depends on the popularity of the influencer in that OSN. For the above reasons, being able to know in advance which will be the influencers in a near future in a given OSN with respect to a given topic would be a competitive advantage for a company, because it would allow such a company to enter in a long-term advertising contract with such future influencers at a considerably lower price with respect to already well-known ones.

1.1 Motivation

Despite the advancement in identifying the influential members of social groups in advance, an in-depth analysis of the available scientific literature (see Section 9) revealed that such a problem has not received sufficient attention so far. As a matter of fact, we found out that most of existing studies are based on Twitter, where the concept of social groups does not exist since tweets are addressed to everybody or to all the followers of the posting user. Instead, social groups play a fundamental role in Facebook, which has recently announced the *Facebook Community Leadership Program*[1]: a global initiative that funds with tens of millions of dollars the administrators who will successfully create and lead groups.[2] However, to the best of our knowledge, the studies that examined *influencers* in Facebook were not focused on social groups, but they took into account different scenarios, i.e., identifying influential contents published by users (such as posts [50] or photos [22]), influential Facebook pages of companies or brands [23, 39], or influential users of Facebook's applications [4, 42]. Other than being not focused on social groups, the great majority of the approaches concerning OSNs' influential users do not even deal with the prediction of future influential users, but they are only meant to identify the current ones. Moreover, the few works performing a prediction [7, 23, 52, 60] are not meant to predict the most influential members within a social group, but they are actually focused on different aspects of influence with respect to our proposal (e.g., [60] predicts active neighbors of the users who have an influence on retweet behaviors; see Section 9 for further details).

Finally, most of the research that has been conducted in the area of social influence has been mainly focused on proposing new measures to compute influence taking into account distinct aspects of the OSNs. Consequently, more than 70 different features are already available in the literature to assess the influence of users [46]. Most cited features measure global influencers of the OSN, and assume that it is easy to obtain a large sample of data from the OSN to compute such features. This is the case for Twitter, which has been widely investigated as a platform of choice for measuring influence of users [7, 14, 17, 48, 51, 52, 57–59] because of its (relative) open platform: the followers of users and the number of retweets on tweets are public. Instead, as shown in Section 9, a study of a general, comprehensive and parametric methodology describing in detail all the phases of the Influencers Prediction process in the context of social groups has not been performed yet. And this is the gap we intend to fill with this work, as described in the next subsection.

1.2 Contributions

The main contribution of this article is the definition of a framework, instantiated over a complete methodology, to predict which users will be able to influence the behavior of other users in a future time period in the context of OSNs' groups. In particular, our framework defines a workflow to deal with the peculiar aspects of the Influencers Prediction process, such as the collection of information about users' activities from OSNs, the modeling of the collected information, the data transformation, the prediction of future influential members, and the evaluation of the accuracy of the prediction. A relevant feature of our framework is that it has been designed to be modular and general: It is not based on a specific data analysis method, influence measure, or prediction model. Consequently, a considerable advantage w.r.t. existing proposals is that our framework allows us to instantiate several distinct versions of the Influencers Prediction process by implementing the phases of such a process leveraging distinct tools and technologies (e.g., by adopting different prediction models for the prediction phase). The framework itself allows us to determine which of such versions provides the best prediction through the prediction accuracy

[1] https://communities.fb.com/ [last accessed on September 23, 2020].
[2] https://newsroom.fb.com/news/2018/02/investment-in-community-leaders/.

evaluation phase. Furthermore, the modular and general proposed approach also shows great applicability and portability to other scenarios as well. To the best of our knowledge, there is no similar work in the literature taking into account the prediction of influential members in social groups. As a matter of fact, our novelty stems from the modeling of the dynamical properties of the social group arising from time-ordered interactions. Moreover, instead of defining yet another influence measure, we define a strategy to properly combine existing Time-Aware Centrality measures to perform the Influencers Prediction process.

Another relevant contribution is the development of a methodology prototype within the proposed framework that allowed us to conduct an extensive experimental validation of the accuracy of the prediction, i.e., the number of predicted influencers that become real ones in the future time period the prediction refers to. In particular, we collected the behavior of the users of 18 Facebook groups belonging to five different categories (Education, News, Politics, Entertainment, Sport), which account for about 800,000 users at the time of writing, and we performed in-depth investigation of the Influencers Prediction accuracy varying the methodology parameters. For instance, we experimentally observed how the adoption of distinct prediction algorithms, the length of the time period we monitored the activities of the group members to train the predictor, and the values of the other parameters of the methodology (see Table 8, given later) affected the accuracy of the prediction. The overall results we obtained from our experiments are striking. Indeed, by properly configuring the methodology parameters, the proposed framework, on average, has achieved a remarkable prediction accuracy: It has been able to correctly predict at least three of the top-10 influential members for half of the examined group (see Figure 16(a)). The best prediction accuracy has been obtained on a group of the Entertainment category for which about six of the top-10 influential members have been correctly predicted by our framework.

1.3 Outline of the Article

In Section 2, we present a comprehensive review of the scientific literature on modeling and predicting influence in OSNs. In Section 3, we formulate the Influencers Prediction problem, and we provide a general overview of the framework we propose in order to solve it. We implemented a methodology within the proposed framework, and we discuss in Section 4 the approach used by the methodology to collect information from Facebook's groups, as well as the description of the collected data. In Section 5, we provide a detailed model of the collected information that allows us to capture different degrees of importance of the group's members. In Section 6, we select some centrality metrics used to compute importance, and we discuss how such metrics are prepared for the training and the prediction of the most influential group's members. Then, we present in Section 7 a set of prediction algorithms that can be used in our methodology to predict the most influential members of the groups and to quantify the strength of such influence. Section 8 presents an assessment of the prediction accuracy of our framework performed over a set of real Facebook's groups, while Section 9 provides a discussion of the main novelty of the proposed approach over the existing ones. Finally, in Section 10, we report conclusions and discuss possible improvements and future works.

2 RELATED WORK

The task of measuring the users' influence has attracted the attention of several research communities because of its potential applications in different domains such as product/item recommendation [41, 42], marketing [30], diffusion of ideas [38] and opinions [56], health [28, 62], and so forth. In this section, we review the relevant literature concerning the measurement and prediction of the users' influence in OSNs.

The authors in [48] show that most Twitter's users act with a passive behavior (i.e., do not forward content). For this reason, they propose an algorithm that measures the influence and passivity of the users by using the acceptance rate of their followers and the amount of interactions that the user rejected from a follower (rejection rate), respectively. Furthermore, they found evidence of weak correlation between the popularity and influence of users.

The authors in [7] focus on the diffusion of tweets that contain a specific tracking URL provided by the *bit.ly* service. They analyze the chains of users who posted the URLs and measure the influence of the initiator by counting the number of distinct users of each chain. Experimental results indicate that the most influential current users have also been influential in the past and that they have a large number of followers. Furthermore, they show that the type of content shared by users (such as Media, News, or Blog) does not affect the length of the chain created by such tweets.

In [57], the authors design a social structure based on the retweet information and use it to infer the level of social interactions. They propose to measure the influence of a user by combining several standard centrality metrics with weighted parameters. In particular, the proposed measure combines (through F-measure) the strength of ties of retweets and the topological importance of the node expressed with the following centrality measures: Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, and Page Rank.

In [59], the authors propose a score (named TURank) for measuring the importance of Twitter's users by using ObjectRank, i.e., an extension of Page Rank that takes into account the types of edges and nodes of the user's tweet graph. In order to evaluate TURank, the authors examine different examinees that compare the TURank score against the number of followers, the number of retweets, Page Rank, and HITS.

The work proposed in [17] presents an empirical analysis of the influence by using a large dataset collected from Twitter. The influence of a Twitter's user is measured by using the number of followers of a user, the number of retweets, and the number of mentions across different topics and time windows. They revealed that the number of followers is not enough to measure the influence of a user. In addition, the most influential users need to put some effort in order to gain and maintain influence.

The authors in [14] propose a method to evaluate the influence of people by using the user connectivity. The method is a variant of the k-shell decomposition algorithm, and takes into account the effectiveness of the k-core decomposition to identify influential spreaders in complex networks. The authors make two changes to the basic algorithm in order to use the method for the Twitter network: the first change produces the k-shell logarithmic value and the second one concerns how the network is interpreted. User influence is measured by the k-shell value produced by the modified algorithm.

In [58], the authors propose Twitter Rank, an extension of the PageRank algorithm. TwitterRank measures the influence by exploiting both the topical similarity between users and the link structure. Further details about the Twitter influence measures that exist in the literature can be found in [46].

In [60], the authors propose the notion of social influence locality, which quantifies how likely a user will also retweet a content, given that a specific set of the user's friends already retweeted the same content. The measure is computed by considering the sum of the random walk probabilities of all active neighbors and the collection of clusters formed by the active neighbors. To predict the retweet behavior, the framework uses a logistic regression learner trained by using the influence locality function.

The authors in [52] propose a supervised rank aggregation to compute the influence of users. They combine ideas from Sociometry and Social Choice Theory and show the effectiveness of the

approach by using a collection of 40 million Twitter users. In detail, the authors use the sociometrics to predict which users will have a viral outburst of retweets in the following week.

An analysis of FriendFeed is described in [27]. Therein, the authors propose a recursive measure similar to Page Rank (named *influence rank*) for identifying influential users in FriendFeed. The measure computes the influence of a user by considering the number of the user's followers, the average number of retweets given by followers to the user's tweets, and the influence rank of the user's followers. In addition, the value of influence is adjusted by considering the root mean square of all the tweets published by users.

The work proposed in [41] investigates the role of time in social influence and proposes a method for influence prediction in *Last.fm*, an OSN where the profiles of users concern their favorite artists and songs. The artists recommended to a user u at time t depends on both the strength of the influence between the u and u's friends and the time elapsed since the friend of u discovered the artists. The strength of the influence between users a and b depends on the period of time between consecutive listens of a and b on the same artists.

The authors in [31] propose a method to compute influential bloggers in the BlogCatalog blogging community. They propose a method to aggregate different network measurements into an influence score by demonstrating that the variation in one metric does not affect the aggregate value. Furthermore, they discover that influential bloggers form a densely connected core, while non-influential bloggers remain at the periphery of the network.

In [42], the authors focus on a Facebook quiz application (named Dek.D) and apply association rule mining to find the relationship between the degree of influence of a user and the characteristics of users (i.e., age, gender, number of shares and responders in each game category).

The authors in [4] derive the concept of influential and susceptible users from [55] in order to investigate in more detail if the influence process is mainly driven by the former, the latter, or both. The experiment was conducted on a Facebook dataset containing the interactions about movies exchanged between users of a commercial application. The influence and susceptibility of a user are estimated by modeling the time required to influence a user, as a function of the user's attributes (such as age, gender, and relationship status). Results indicate that susceptibility decreases with age, while women are apparently less influential than men. In addition, they show that the most influential users are less susceptible to influence, and that they cluster in the network.

A similar approach for measuring influential and susceptible users based on interactions performed in OSNs is proposed in [44]. In particular, the authors define the influence of a user u as a weighed combination of the out-degree of u, the out-clustering coefficient, and the strength of u's links. The same formula is also used to compute the susceptible users (i.e., the users who are likely to be influenced), but in such a case only the incoming interactions of users are considered (i.e., the in-degree and the clustering coefficient).

The authors in [39] focus on Facebook brand pages, and they propose a method to identify the set of users having a higher association value with the pages. The association value is computed by combining the similarity between the user and the page, and the frequency of the interactions (namely, posts, comment, and like) with different weights.

The study proposed in [50] investigates the problem of predicting the best time for a user to publish a content in order to maximize the probability of receiving reactions from the audience. They compare different predictors that consider the volume of reactions received by a post/tweet in a specific period of time after its publication. Experimental results on Facebook and Twitter reveal that the *when-to-post* problem can be efficiently solved by considering (i) the number of reactions generated by the friends of a user after the publication time of the post; and (ii) the tendency of a friend to react to the user's posts.

In [23], the authors propose a methodology based on association rules to discover relationships between users. Furthermore, they use association rules to predict if a user will participate in a post discussion in the future based on the users' activity.

Multiple platforms have been analyzed in [45, 63]. In [45], a score is assigned to each user (on a daily basis) through a supervised model that assigns weights to 3,550 features related to the network structure (i.e., graph properties), to the profile of a user (e.g., education), and to the interactions among users (e.g., comment or post).

The authors in [63] propose a framework to efficiently compute the influence of users by considering heterogeneous networks that model information coming from different sources (i.e., the social relationships, the activities among objects, and the relationships between users and objects). In particular, the proposed approach uses the previous model to cluster the users in K different clusters based on similarities between two members. Each cluster represents a category of activities that people have engaged in, and the similarities are computed by considering both the influence derived from social connection and the influence of the social activity.

In [9], the authors propose a method to identify active micro-bloggers by using the social interactions between them. Indeed, InfRank is a PageRank-like algorithm that measures users' influence by using the user–retweets graph, where nodes are users and edges are retweets. In detail, the algorithm measures the ability to spread information in the network.

An alternative approach is proposed in [64]. Therein, the authors introduce SIRank to measure the spread influence of users in a general micro-blog. The method uses several features: user interactions, retweet intervals, location of users in information cascades, comments, and so forth.

The authors in [51] propose the problem of influence maximization in the context of information flows. They provide an algorithm called InFlowMine to discover information flow patterns using content propagation.

Other research works on this topic focus mainly on proposing efficient algorithms for solving specific influence problems, such as the *topic-based social influence* problem [53] or the *influence maximization* problem [34].

In contrast to most research works, which are mainly oriented toward influencing prediction in Twitter, little work has been conducted on Facebook. Indeed, most of the works on Facebook focus on investigating aspects of the influence process with respect to a single content: such as a photo [22], a single Facebook page [39], or within a Facebook application [4, 42].

3 A GENERAL FRAMEWORK FOR IDENTIFYING INFLUENTIAL MEMBERS WITHIN GROUPS

Inspired by the increasing importance of OSN groups as discussion venues, and given the limited attention they have received from researchers, in this article, we focus on a specific yet relevant aspect of OSNs: influencers. In particular, we designed a general framework where we have instantiated our methodology for identifying the current most influential members of a group, and to predict the future ones. In this section, we start by defining the concept of influence on social groups, and we describe how it can be quantitatively measured starting from the interactions that occurred among group members. Then, we provide an overview of the proposed framework that will be described in detail in next sections. Table 1 summarizes the general notation used in the rest of this article.

3.1 Social Influence

Influence is a very broad concept and a certain ambiguity exists in both industry and academia about its meaning. Some people equate the influence of users to their popularity, while others see it

Table 1. Notations and Terms Used to Represent Different Aspects of the Framework

Variable	Description
$\mathcal{G}$	a generic group of members of the OSN
u	a generic member of the group
$\mathcal{T}$	a specified time period
$\mathcal{E}_{\mathcal{T}}$	the set of interactions that occurred in $\mathcal{T}$ among the members of the group
$I_{\mathcal{E}_{\mathcal{T}}} : \mathcal{G} \rightarrow IR$	influence measure that assigns to $u \in \mathcal{G}$ a numerical value in IR
$\overrightarrow{G}(V, E)$	a directed multigraph (*Group Interaction Graph*)
V	the set of users in the group
E	the set of directed edges representing the events between users of V
$\overrightarrow{e}(v, w)$	a directed edge of E where $v, w \in V$
Δ	duration of a slice allocated on the time period $\mathcal{T}$
$\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$	a *Group Interaction Graph* for a generic slice $[t_i, t_i + \Delta)$
V_{t_i}	the set of the group's members who interacted with each other in the slice $[t_i, t_i + \Delta)$
E_{t_i}	the edges that occurred between the members of a group in the slice $[t_i, t_i + \Delta)$
$\{\overrightarrow{G_{t_1}}, \overrightarrow{G_{t_2}}, \ldots, \overrightarrow{G_{t_s}}\}$	a *temporal network* of a group where $t_1 < t_2 < t_3 < \cdots < t_s$
$M_u^{t_i} = [m_1, \ldots, m_N]$	the vector of metrics computed for a member $u \in V_{t_i}$
$C_u^{t_i} = \{c_1, c_2, \ldots, c_{N'}\}$	the vector of metrics of size N' (named principal components), where $N' < N$

as an indicator of their activity on the OSN. However, there are important differences that need to be understood in order to design a new system for identifying the influential members of a group.

In the context of social groups defined in OSNs, the influence of members in a group can be quantified through a measure that leverages the knowledge of this group, such as the activities performed by the members or the information in their profiles [43, 46]. In this manuscript, we define the influence of a member u as the ability of u to attract the attention of other members as a result of the activity she performed. We restrict the scope of our study to the activities performed within the group by its members, while the information on users' profiles is not considered. This choice is dictated by the fact that, typically, profile information is private and cannot be accessed [39]. Furthermore, the activities performed by the members of a group can be categorized as active or passive. Passive actions cannot be observed by the other members of the group because they correspond to viewing posts, scrolling comments and replies, or clicking a link of the group. For this reason, such actions are outside the scope of this work. Instead, we will focus on active actions that can be observed by all the members of the group (such as publication of contents, comments, reply, and reactions) [45].

Let $\mathcal{E}_{\mathcal{T}}$ be the set of interactions that occurred in a specified time period $\mathcal{T}$ among the members of the group $\mathcal{G}$. We define the influence measure $I_{\mathcal{E}_{\mathcal{T}}}(u)$ of a group member u as a function $I_{\mathcal{E}_{\mathcal{T}}} : \mathcal{G} \rightarrow \mathbb{R}$, which assigns to $u \in \mathcal{G}$ a numerical value. Such a value determines to which extent a member u influences the activities of the other members of the group in the period $\mathcal{T}$, attracting a high fraction of the interactions in $\mathcal{E}_{\mathcal{T}}$. Once the influence measure has been calculated for each member of the group, the most influential members can be selected by comparing their influence scores.

Due to the complexity of the OSN scenario, the influence measure is typically obtained by combining different metrics capturing important aspects of the group (see Section 9). For instance, a metric for influence can take into consideration the local properties of the members (such as the type and the volume of the interactions received by other members) and/or global properties related to the whole structure of the group (such as the importance of a member with respect the other members of the group and/or the length of the information diffusion triggered by a member).

Based on the above concept, we can formulate the problem of predicting the most influential members of a group defined in OSNs as follows:

Problem setting. Given the set $\mathcal{E}_{\mathcal{T}}$ of the interactions that occurred among the members of the group $\mathcal{G}$ over a specified time period $\mathcal{T}$, the Influencers Prediction is the problem of finding the future influencers of the group $\mathcal{G}$, i.e., the members who are more likely to influence the other members of the group in a future time interval $\mathcal{T}'$ on the basis of the score obtained by the influence measure $I_{\mathcal{E}_{\mathcal{T}}}()$. Since in the following we compute the influence measure of each group separately from the others, we simplify the notation by using $I_{\mathcal{T}}()$ instead of $I_{\mathcal{E}_{\mathcal{T}}}()$.

Since our formulation of the problem is based on social groups within OSNs, it is necessary to consider the requirements and constraints of the target scenario. Indeed, social groups provided by most current OSNs are based on a group communication model that provides content delivery from one member (sender) to all the other members of the group (receivers). Furthermore, there are several aspects that must be considered when predicting influential members of a group.

Group type. Typically, in social groups, there are no defined relationships between members. However, some OSNs provide the capability to assign specific roles to one or more members of the group (such as administrator or moderator). Since such special members could exhibit a high number of interactions with the other members of the group because of administrative reasons, our framework discards them from the analysis.

Group dynamism. In general, any user of the OSN can request to join a group to its administrator. Furthermore, some groups can be freely joined without the need to request a permission to the group administrator. In addition to the join operation, the members of a group can decide to leave it, for instance, because they are no longer interested in the group topic. Moreover, a member can also be evicted (*banned*) from a group, for instance, because of an inappropriate behavior. As a result, social groups in OSNs are highly dynamic. In order to capture such dynamism, the membership information of the group is updated periodically by our framework.

Temporal factor. The members of a group interact over time and the influence measure should consider the age of such interactions. In particular, to reflect a current phenomenon we all experience in life, the importance associated with each interaction should decrease over time. As such, in order to capture the variation in influence of members, the contribution of very old interactions should be mitigated, while the importance of the recent interactions that occurred in the group should be increased by considering proper weights.

3.2 Framework Overview and Influencers Prediction Workflow

This section defines the general architecture of the proposed framework and describes the workflow of the Influencers Prediction process. The workflow consists of a sequence of phases performing different operations on data, namely: *Data collection, Data modeling, Data transformation, Training, and Prediction*. A further phase, called *Evaluation*, can be optionally performed at the end of the time interval the prediction refers to, i.e., when the real influencers for that interval can be computed, in order to evaluate the accuracy of the prediction, i.e., how many of the predicted influential members turned out to be real influencers. The input data and the results produced by each phase are shown in Figure 1. Moreover, each of these phases consists of one or more tasks to be executed to produce the results, as shown in Figure 2. The proposed framework has been designed to be general, i.e., it is not based on a specific data analysis method. As a result, several data analysis methods and tools can be selected and applied in our framework to implement each task, depending on the algorithmic approach used to find the most influential members.

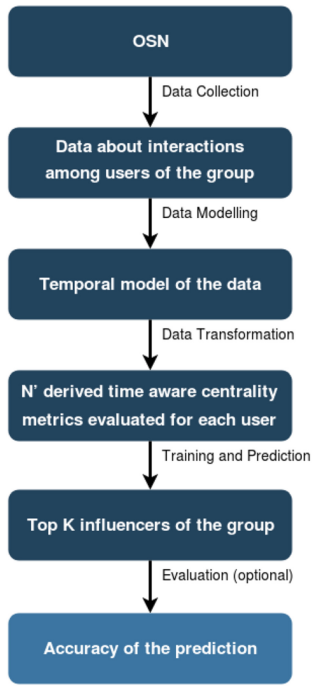


Fig. 1. Workflow of the Influencers Prediction process focusing on data transformation.

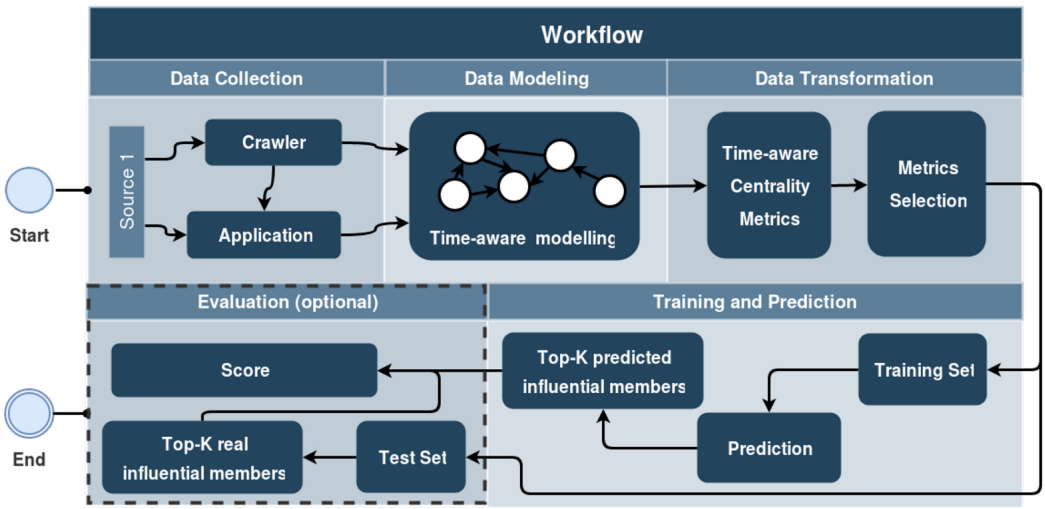


Fig. 2. Workflow of the Influencers Prediction process focusing on the tasks performed in each phase.

Data collection. The first phase of the workflow deals with data collection, and it ensures that updated information related to the activities performed by the members in a selected social group concerning a given time interval $\mathcal{T}$ is gathered from the OSN. The data collection time interval could cover the entire OSN life, or could include only recent activities, e.g., the last n months. The social group to be considered for Influencers Prediction is selected based either on target marketing

strategies or on the domain of interest. To collect from the OSN the relevant data concerning the interactions that occurred among the members of the group (e.g., likes, comments, replies, and creation time) in the time period $\mathcal{T}$, two main approaches are typically used. The first approach consists of implementing a program, named crawler [16, 25], which periodically explores the social group on behalf of a given user using the HTTP protocol, thus behaving like a web browser, and extracts information about the item published by the group members. The second approach, instead, uses the APIs provided by the OSNs' infrastructures to create an application that collects data from groups. In this case, OSNs' users have to authorize such applications to collect their data by typically issuing an *access token* that is valid for a limited period of time only [20]. Regardless of the method used to collect social group information, it is important to be able to collect updated information in order to both maximize the accuracy of the prediction and avoid the timeliness problem. In fact, the Influencers Prediction could lead to imprecise results if the *Data collection* phase provides out-of-date information about the social group. Due to the dynamic nature of these systems, changes occur in social group very frequently. As a matter of fact, new interactions are ceaselessly generated by the members through the publication of new posts, comments, replies, and so forth. Furthermore, the administrator(s) of a social group can add new users to the group or remove existing members at any time. For this reason, the *Data collection* phase must be iterated periodically, defining a proper sampling frequency. The result of the *Data collection* phase is the list of all the interactions (posts, comments, replies) that occurred among the members of the group in the chosen time period $\mathcal{T}$.

Data modeling. The choice of a suitable model to represent the data gathered in the *Data collection* phase is crucial to facilitate data understanding and to allow the execution of the Influencers Prediction process. Very often such data are represented using the graph formalism, where the members of the group are represented by vertexes and their activities are represented by edges connecting vertexes. For example, if user A commented on a post published by user B, an edge from the vertex representing A to the vertex representing B will be added to the graph. Moreover, in order to represent the time when an interaction among two users occurred, the time dimension must be included in the graph model. The resulting formalism captures the structural–temporal aspects of the social group arising from time-ordered interactions, and it is known under various names in scientific literature, such as *temporal network/graph* [33, 35], *time-varying graphs* [15, 40], *contact sequence graph* [8, 29], or *evolving graph* [24]. In these models, the system is defined as a graph $\mathcal{G} = \langle V, E, T \rangle$, where V is the set of vertices representing the group members, E is the set of edges between two vertices (a, b) $\forall a, b \in V$, and T is a function that given an edge $(a, b) \in E$ returns a time reference for each of the interactions that occurred between a and b in the period $\mathcal{T}$. In particular, the previously presented models differ in the representation of such time references. In fact, the model used by [8, 24, 29] does not take into account the duration of interactions among users, and, consequently, the function T applied to an edge $(a, b) \in E$ simply returns the set of time instants $t'_1, \ldots, t'_s$ when the interactions between user a and user b occurred. The purpose of this model is to enable the analysis of network systems where the duration of the interactions between members is not important, because the main focus is on the amount of interactions happening at the same discrete time. Instead, the model proposed by [15, 33, 35, 40] specifies that each interaction starts at a specific point in time and has a duration. In this case, the time function T applied to an edge (a, b) returns a set of time intervals $\{(t'_1, \delta t'_1), \ldots, (t'_s, \delta t'_s)\}$, where t'_i (for $i = 1, \ldots, s$) is the starting time of the ith interaction between a and b and $\delta t'_i$ specifies the duration of such interactions. Also, in this case, we say that the edge e is defined (or present) for some extent of time and the model is very suitable in all those cases where the duration of the events plays a key role in understanding the system.

Another important aspect to consider is the direction of the interactions that occurred among group's members. In our case, directed graph $\overrightarrow{G}$ is used, where the position of vertices in each edge represents the direction of the interaction.

Finally, since typical social groups provided by current OSNs (such as Facebook and Linkedin) do not exhibit the relationships between the group's members (e.g., friend relationships), the most general model is to consider V as the set of members of a group and to use a member-to-member edge e for representing each interaction between members. In particular, the edge e specifies the type of the related interaction (e.g., comment or reply), and it can also contain implicit information about the interaction (such as the number of reactions received and the type of contents published in the interaction). Another possibility is to consider a more complete data model where vertexes $V = V_1 \cup V_2$ represent both the members of the group, V_1, and the set of contents published by the members (e.g., posts, comments, and replies), V_2. In this case, an edge e can represent different types of relationships between users and contents, such as, user "publishes a" post, user "comments a" post, or user "replies to a" comment. The result of the *Data modeling* phase is a model that properly represents the time-ordered activities performed by the group's members.

Data transformation. The *Data transformation* phase uses the data model built in the previous phase to compute a set of relevant centrality metrics capturing distinct aspects of the group's activities that quantify the importance of the group's members. In the context of information diffusion, centrality metrics represent the most popular measures used to derive the importance of users in the form of a numerical value. According to [46], centrality metrics can be classified into three different categories: activity metrics, popularity metrics, and influence metrics. Activity metrics aim at quantifying the participation of a member over time (such as the weekly number of posts of a member), regardless of the number of other users who noticed the participation of such members (e.g., by posting a reaction). Popularity metrics, instead, focus on the identification of the most popular members, i.e., those users who have been involved in the interactions among a large number of other members of the group (e.g., they have been tagged in a large number of posts by other users), independently of their active participation in the group activities. Finally, the purpose of influence metrics is to identify members whose interactions have both attracted the attention of other users and affected their interactions (e.g., the number of comments received on posts or the number of replies received by comments). Once a relevant set of N centrality metrics has been properly selected and instantiated, the `Time-Aware Centrality Measures` task computes the value of each of these metrics for each member of the group. The resulting values are then processed by the `Metrics Selection` task, which reduces the number of metrics from N to N' (where $N' < N$). In particular, the `Metrics Selection` task applies data dimension reduction techniques and factor analysis in order to preserve as much as possible the structure of the original centrality metrics while reducing the number of dimensions. In addition, the `Metrics Selection` task reduces the number of dimension taking into account the correlation among the centrality metrics, because it can affect the accuracy of the predictions and lead to multicollinearity problems. The result of the *Data transformation* phase is a set of N' derived time-aware centrality metrics computed in the data collection time interval $\mathcal{T}$ for each member of the group.

Training and Prediction. Once the derived centrality metrics have been chosen and the related results have been computed, the *Training and Prediction* phase applies predictive tools to extract the most relevant members of the group. The first task of this phase consists of creating a training set that contains the data used for the `Prediction` task. Normally, all the data received from the *Data transformation* phase are included in the training set. However, if the *Evaluation* phase were to be performed after the prediction, then the training set should include only a part of the data

received from the *Data transformation* phase, since a subset of them consisting of the most recent should be reserved as a test set for the *Evaluation* phase. Hence, supposing that the *Data collection* phase gathered data concerning the time interval $[t_s, t_e]$, the data concerning the interval $[t_s, t_t)$ would belong to the training set, while the data related to the interval $[t_t, t_e]$ would belong to the test set. We notice that, since the `Prediction` task has been performed using the data in the interval $[t_s, t_t)$, the prediction result should be valid for a future period with respect to t_t, e.g., for the period $[t_t, t_e]$.

The `Prediction` task represents a building block of our framework, and it can be instantiated by using the most suitable predictor. As a matter of fact, finding the prediction method that provides the optimal result for our goal is a very complex task because it requires trying out different prediction methods. In order to alleviate such an effort, we designed our framework to accommodate any prediction algorithm. Consequently, a set of possible candidates can be considered, and the `Prediction` task can be instantiated by using each possible prediction method (such as linear regression, support vector machine, or neural networks). The result of the `Prediction` task is a list of members of the group, each paired with a numeric score, which enables us to compare their influences.

The `Top-k Predicted Influential Members` task, given the list of members and the scores resulting from the `Prediction` task, is responsible for selecting a subset of k of the most influential members. A straightforward and fast method to select such k members is to consider the score of each member and to take the k top-ranked ones. Hence, the result of the *Training and Prediction* phase is the list of the k most influential members of the group.

Evaluation. The workflow can be optionally extended by executing the *Evaluation* phase, which is in charge of evaluating the accuracy of the results generated by the *Training and Prediction* phase. As a matter of fact, the accuracy of the prediction depends on the configuration chosen when the framework is implemented. This phase is executed before using the framework for real predictions in order to fine-tune the parameters of the framework, i.e., to find the configuration that results in the best accuracy of the prediction.

The *Evaluation* phase executes the `Top-k Real Influential Members` task on the data belonging to the test set to determine the top-k real most influential members for the time interval $[t_t, t_e]$. Then, the `Score` task compares the list of the top-k real most influential members for the time interval $[t_t, t_e]$ with the list of the top-k predicted influential members produced in the previous phase, in order to evaluate the performance achieved by the predictor. The result of this phase is the accuracy of the prediction, i.e., how many real influencers have been correctly predicted.

Finally, it is worth noting that the whole Influencers Prediction process can be executed several times using different tools for implementing their tasks (e.g., data representation models, centrality metrics, prediction methods) and different values for the other configuration parameters (e.g., data collection time interval, reduced number of centrality metrics). The *Evaluation* phase allows us to select the combination of tools and parameters that obtains the best accuracy in detecting influencers.

4 DATA COLLECTION

In this section, we focus on the *Data collection* phase of the Influencers Prediction workflow by describing in more detail the methodology we use to collect a dataset containing the required information. Nowadays, several OSN services exist. While most of them are for general purposes (such as Facebook and Twitter), other ones are more service oriented and targeted to a particular type of content (e.g., YouTube and Flickr) or usage (e.g., such as LinkedIn). We focus on general-purpose OSNs because they have general diffusion and they are very popular among the

Table 2. Types of Group that Users Can Create on Facebook

Type of group	Join request	Read	Publish
Public	Anyone	Anyone	Current members
Closed	Anyone	Current members	Current members

users. Since we are interested in measuring the influence of users in social groups, we restrict our analysis only to those OSNs that allow their users to define groups. In particular, we selected Facebook as a subject of our examination because of the following reasons:

— Facebook is the most used OSN in the world, which (at the time of writing) attracts about 3 billion active users.[3]
— Since there is no group concept in Twitter, it is difficult to find communities of users in Twitter because it requires analyzing the tags used by the OSN in order to collect a subset of users related to a specific topic.
— With respect to Facebook, Twitter has been widely studied in the context of users' influence and information diffusion because its users publicly declare the people they follow. Instead, the role of influence in Facebook's groups still remains unexplored in current scientific literature (see Sections 2 and 9).

Facebook has widely extended its features over time, by giving to its users the ability to define groups having different characteristics. In particular, according to Facebook Help Center,[4] users can create two different types of groups whose characteristics are summarized in Table 2. Both *public* and *closed groups* require the approval of a new member by either the administrator or another member, depending on the membership settings of the group. In *public groups,* anyone can see the group's members and read posts published in such groups, but only the current group members can publish a post. Anyone can ask to join a public group. *Closed groups* consist of posts that can be published and accessed only by the current members of such groups. Anyone can request to join a closed group. In order to obtain a large collection of heterogeneous groups having different characteristics, we implemented a Facebook crawler application that allows us to retrieve the interactions that occurred on a set of selected Facebook groups. In particular, we focus only on *public* or *closed groups* because they can be searched, respectively, by any users or by becoming a member. Instead, accessing secret groups requires receiving an invitation from an existing member.

4.1 Data Collection Methodology

The approach we used to retrieve the information about Facebook groups uses an HTTP-crawler that directly interacts with the Facebook groups' pages using the web browser. Our crawler based on Selenium[5] automates browsers' library, which, given a Facebook group G, periodically retrieves the following sets of information:

— *Members.* We retrieve the member list of the group, which contains the users participating in the group.
— *Interactions.* We collect interactions that occurred between members of the groups, such as posts, comments, replies, likes to post, likes to comment, and reactions.

[3] https://newsroom.fb.com/company-info/ [accessed on May 2020].
[4] https://www.facebook.com/help/220336891328465?helpref=about_content [last accessed on May 2020].
[5] https://www.seleniumhq.org/ [last accessed on May 2020].

Table 3. Information Collected about Groups by Our Crawler

Posts		Comments/Replies		Members	
Name	**Description**	**Name**	**Description**	**Name**	**Description**
postId	identifier of the post	*commentId*	the identifier of the comment	*memberId*	the identifier of a member
authorId	identifier of the author	*authorId*	the identifier of the author	*memberType*	the type of the members (0 = administrator, 1 = members with things in common, 2 = recently joined)
time	creation time	*time*	creation time		
isShared	true if the post contains an original object shared on the group, false otherwise	*commentType*	indicates if it is a comment to a post (C) or it is a reply to a comment (R)		
contentType	the type of content (Video, Link, Album, Post, Photo, Event, Page, Gif, Group, Offer, Memory, Text)	*targetId*	it indicates the identifier of the post (if commentType = C) or the identifier of the comment (if commentType = R)		
share	the number of times a content has been shared	*reactions*	the number of reactions (Like, Love, Haha, Wow, Sad, Angry)		
reactions	the number of reactions (Like, Love, Haha, Wow, Sad, Angry)				

The collected information is parsed and preprocessed at a later stage, when the crawler has finished running.

4.2 Description of the Collected Data

Table 3 describes in more detail the information the crawler collects from each Facebook's group. The information related to the members of a group is essential to uniquely identify such members (*memberId*) and to understand the organization of the groups (*memberType*). In particular, the latter indicates whether a member is (i) an administrator of the group, (ii) a member who has been added recently, and (iii) a member who has something in common with the other members of the group. Since members of the groups can restrict the access to their profiles, we collected the previous information only for those members who have a visible profile.

For each group, we were able to retrieve the activity occurring on the posts published by the group's members. The posts published by members of the groups consist of the identifier of the post (*postId*), the identifier of the author (*authorId*), the time the post was initially published (*time*), a flag that indicates whether a post contains a shared content or not (*isShared*), the type of content published in this post (*contentType*), the share count of this post (*share*), and the number of reactions to this post (*reactions*).

Other important sources of information are comments and replies related to posts. In particular, these interactions consist of the identifier of the comment/reply (*commentId*), the identifier of the member who made the comment/reply (*authorId*), the time the comment/reply was made (*time*), a flag that indicates whether the interaction is either a comment on a post or a reply to a comment (*commentType*), the target identifier of the interaction (*targetId*), and the number of reactions that occurred on it (*reactions*).

4.3 Analysis of the Social Groups

We identified five group's categories that gather a number of Facebook's groups having similar subjects (i.e., News, Education, Sport, Entertainment, and Work), and we selected 18 Facebook groups that belong to these categories. Table 4 shows the selected Facebook groups, the topics discussed by the groups, the number of members, and the category to which they belong. The collected groups have the advantage of representing a heterogeneous set of real-world groups having

Table 4. General Description of the Selected Facebook's Groups

GroupId	Group Topics	Members	Category
N1	Gossip, News, Rumors, TV	49,761	News
N2	Politics, News, Current affair	37,253	News
N3	Online newspapers, Magazines, News	5,083	News
E1	Students, Programming, Code, Nerd	10,771	Education
E2	Students, Education, University	46,040	Education
E3	Research, Innovation, Science	21,286	Education
S1	Tennis, Players, Tournaments	10,451	Sport
S2	Swimming, Swimmers, Competitions	3,583	Sport
S3	Gym, Fitness, Exercise	108,078	Sport
S4	Well-being, Lifestyles	36,558	Sport
T1	Thriller, Series, Movie	6,941	Entertainment
T2	Horror, Fiction, Film	15,028	Entertainment
T3	Motion pictures, Photography, Music	39,079	Entertainment
T4	Guitar, Acoustics, Musicians	9,392	Entertainment
W1	Startup, Business, Market	26,496	Work
W2	Administration, Systems, Engineers	4,592	Work
W3	Temporary work, Employment, Workers	25,391	Work
W4	Jobs, Positions, Occupation	11,842	Work

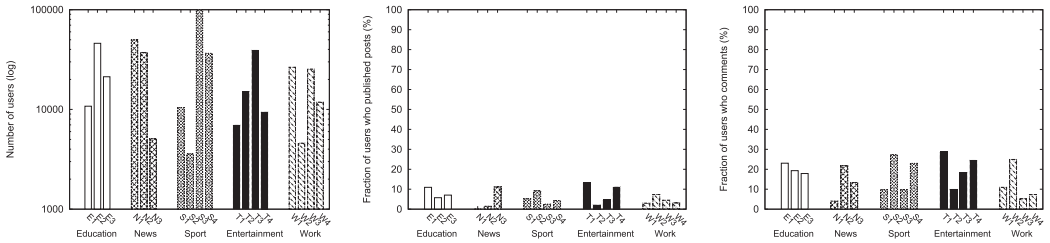
different backgrounds and purposes. We use our crawler application to retrieve the information concerning the activities performed by members of the selected groups, and we choose as the sampling frequency of the crawler the value of 1 day. Hence, the information about groups is retrieved every day of the monitored time period and is stored in an XHTML format. Due to technical reasons, we collected the information related to the interactions that occurred within the selected groups in the last 16 months. We will see in the following sections that this choice will not impact the prediction performance of the proposed approach. Table 5 summarizes the main characteristics of each group by showing the date of both the first post (*min Date*) and the last post (*max Date*) retrieved by our crawler, the number of consecutive days on which our crawler collected the interactions (*Days*), the total number of posts retrieved from the group in the monitored period (*Posts*), the size of the set of members who are authors of at least a post (*Authors*), and the average number of posts published each day by groups' members (*Post X day*). Results indicate that we were able to collect the activities performed by members on 365 days (i.e., 1 year) only for about half of the selected groups. Indeed, groups having a higher activity rate (i.e., more than 15 posts per day) can overload the crawler because of the huge amount of resource needed to load all the interactions collected over the last years. In general, the number of posts collected from each group depends mainly on the social activity of the members, and it ranges between 155 (for N1 of the News category) and 5,271 (for E2 of the Education category) posts.

The collected groups exhibit a high variation in the number of distinct authors, and this value is positively correlated with the total number of posts (Pearson correlation $R = 0.64$) and with the average number of posts per day (Pearson correlation $R = 0.12$). However, it is important to note that group N1 of the News category has only two users with an account for the whole set of posts published in the group.

The *Data transformation* phase will be configured to filter out the shortcomings resulting from the *Data collection* phase by removing group S3 of the Sport category and group N1 of the News category because they are considered abnormal cases (outliers).

Table 5. General Statistics about the Groups

Category	Group	min Date	max Date	Days	Posts	Authors	Posts x day
News	N1	07/10/2017	26/01/2018	111	155	2	1.396
	N2	08/11/2017	07/02/2018	91	3,397	543	37.330
	N3	02/01/2017	12/02/2018	406	1,133	565	2.791
Education	E1	01/01/2017	24/01/2018	388	3,555	1,180	9.162
	E2	06/04/2017	18/02/2018	317	5,271	2,643	16.628
	E3	25/01/2017	22/02/2018	393	5,060	1,499	12.875
Sport	S1	21/05/2017	27/01/2018	251	5,100	544	20.319
	S2	04/02/2017	09/02/2018	370	708	331	1.914
	S3	13/02/2018	14/03/2018	28	6,353	2,754	226.893
	S4	27/08/2017	03/05/2018	249	5,588	1,562	22.441
Entertainment	T1	30/09/2017	08/02/2018	130	5,009	915	38.531
	T2	22/10/2017	23/02/2018	123	3,777	319	30.707
	T3	02/01/2018	03/05/2018	120	4,904	1,836	40.867
	T4	09/09/2017	06/03/2018	178	3,543	1,041	19.905
Work	W1	02/01/2017	12/02/2018	406	1,444	800	3.557
	W2	04/01/2017	26/02/2018	418	945	337	2.261
	W3	13/06/2017	27/04/2018	318	4,809	1,140	15.122
	W4	03/01/2017	04/05/2018	485	2,651	374	5.466

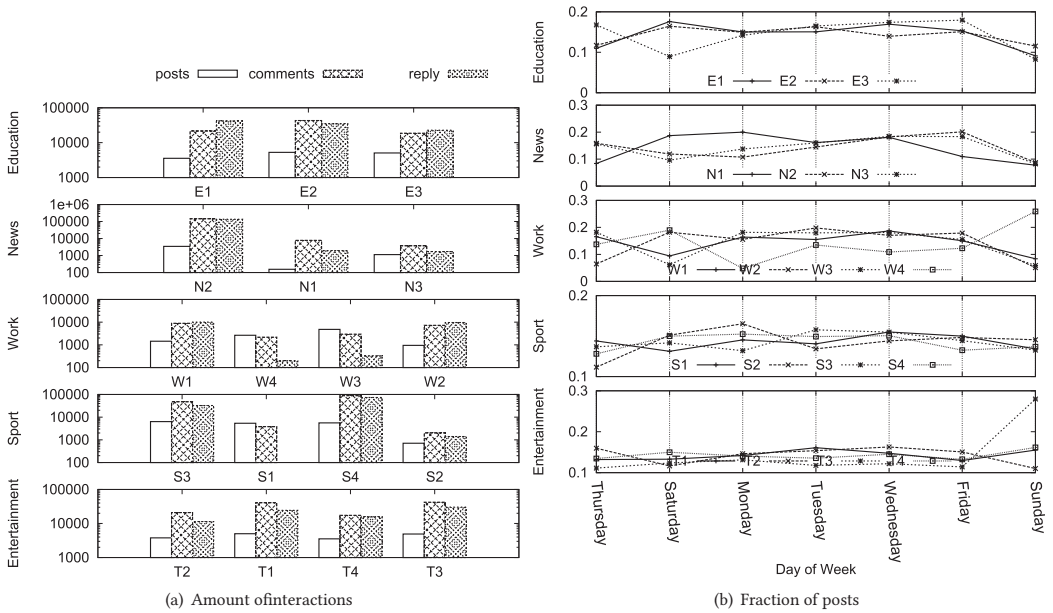


(a) Number of members (b) Fraction of posts writers (c) Fraction of comments writers

Fig. 3. Analysis on the members of the groups.

Group members. We investigated in more detail how our groups are internally organized and the activities performed by groups' members. Figure 3(a) shows the total number of users allocated to each group, while Figure 3(b) shows the percentage of members who interacted on the group by publishing at least one post. The users are distributed among different groups in a heterogeneous way, and the total number of members who joined a group does not depend on the group category, but it depends on the interests of the users. However, the plots clearly indicate that most group members (about 98%) primarily use the service for passive interactions, such as browsing the posts or comments from the groups, while only a small fraction of the members (less than 15%) are involved in active interaction, i.e., writing a post or a comment on the group. For instance, groups E2 and E3 of Figure 3 have more than 20,000 members, but at most 2,700 of them (i.e., less than 13%) have been engaged in writing at least one post within the groups.

Figure 3(c) shows the percentage of members who commented a post or replied to a comment on the posts of a group. As expected, the fraction of members who interact over a post is quite higher, but it does not exceed 30% of the members for all the selected groups. Finally, we can observe that,



(a) Amount ofinteractions (b) Fraction of posts

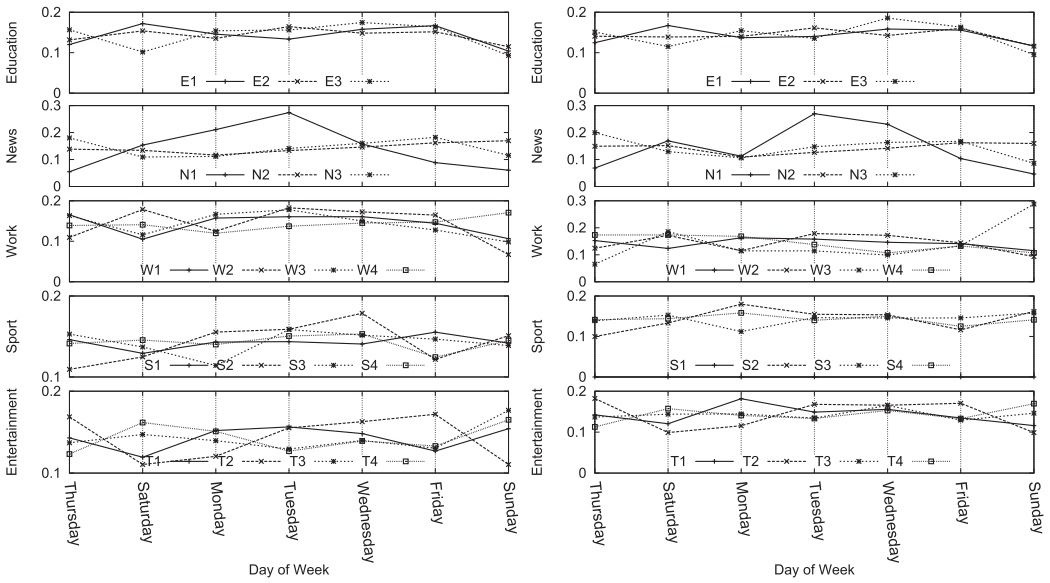
Fig. 4. Amount of interactions and fraction of posts published by members of the groups each day of the week.

during the monitored period, all the social activities on Facebook's groups involve only a small fraction of users (i.e., at most 40% of the group members).

Interactions. The next step in our study involved a detailed analysis of the type and amount of interactions performed by members of a group during the monitored period. Figure 4(a) shows the total number of posts, comments, and replies published by members of groups for each category. About 33% of the posts created by users embed contents linked to other sources (e.g., the website of an online newspaper), such as album (0.01%), event (0.5%), gif (0.008%), group (0.08%), link (16%), memory (0.04%), offer (0.008%), page (0.19%), photo (1.2%), post (13%), and video (1.9%). Instead, most posts (about 66%) contain contents that have been uploaded by the user.

It is important to notice that comments and replies are the most frequent iterations performed by members of groups. However, there may be groups that exhibit different characteristics, such as group S1 of the Sport category, which is the only one that has no replies to the comments published on the posts.

We investigated in more detail how interactions are distributed across the days of the week. For this reason, we plot in Figure 4(b) the fraction of posts published by users each day of the week (from Sunday to Saturday). We can notice that members of the group prefer to publish posts during the working days of the week, while the number of posts published during the weekends has been significantly reduced. Indeed, about 15% of the posts created by users are published during a day of the week, while there is a decrease of about 5% in the number of posts during the weekends (i.e., Sunday and Saturday). Starting from Sunday, the number of posts published on groups each day increases as long as the middle of the week is reached. Indeed, the majority of posts (about 17%) are published on Wednesday, and, after that, the number of posts decreases for the second half of the week. We performed the same analysis on comments and replies, and we show in Figure 5(a) the fraction of comments published by the members. We can observe that the number of comments published by groups' members exhibit a different pattern from those found in posts. The majority



(a) Fraction of comments (b) Fraction of replies

Fig. 5. Fraction of comments and replies published by members of the groups each day of the week.

of users publish their comments on Tuesday (about 16.3% of the comments), while the number of comments published in the second half of the week is distributed uniformly across the period (i.e., about 15% of the comments are published on Wednesday, Thursday, and Friday). Figure 5(a) indicates that members significantly reduce their activity on posts over the weekend (only 11% of comments published on Sunday and Saturday), while the activity on groups starts to increase from Monday (about 14%).

As shown in Figure 5(b), the fraction of replies on comments follows a similar trend, i.e., the users published most replies (about 15%) in the middle of the week (i.e., on Wednesday), while they lose interest in publishing replies on the weekend. Indeed, the number of replies on the weekend decreases by about 5% as we get close to Sunday and Saturday.

5 DATA MODELING

This section focuses on the *Data modeling* phase by formally defining the interaction graphs we used to model the activities performed by users on Facebook groups. Thereafter, we refined the interaction graph models to capture the dynamic nature of the network, the evolution of interactions, and their dynamical properties.

5.1 Definition of Interaction Graphs

A *Group Interaction Graph* $\overrightarrow{G}(V, E)$ is a directed multigraph where the set of vertexes (or nodes) V represents the set of users in the group, while E is the set of directed edges representing the events that occurred between the nodes in V, e.g., $\overrightarrow{e}(v, w)$, where $v, w \in V$. Such events correspond to comments to posts or replies to comments. Reactions are not considered events themselves, but they represented as attributes of the events they refer to. The function $l : E \rightarrow \Sigma_E$ defines the label of the interaction, and $\Sigma_E = \{Comment, Reply\}$ is the set of possible labels. In our scenario, the graph $\overrightarrow{G}$ is directed and the order of nodes in each edge $\overrightarrow{e}(v, w)$ determines the source and the target of the interaction. In particular, each edge $\overrightarrow{e}(v, w)$ consists of the source $s(e) = v$ and

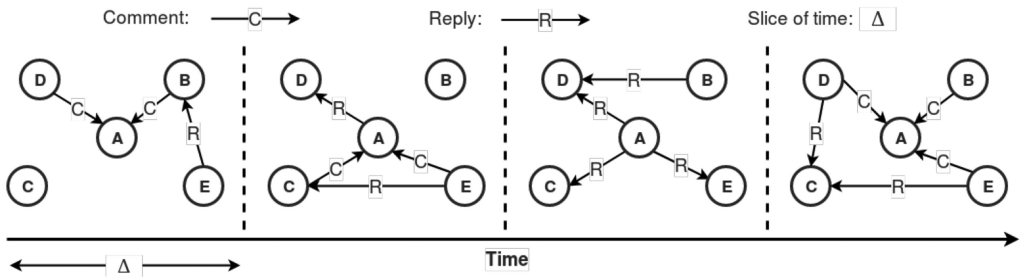


Fig. 6. Contact sequences of a temporal network.

target vertex $d(e) = w$, respectively. The graph $\overrightarrow{G}$ can have multiple arcs between two vertexes, i.e., there may be different edges $\overrightarrow{e_1}(v, w), \overrightarrow{e_2}(v, w), \ldots, \overrightarrow{e_n}(v, w)$ having the same source and target in the graph $\overrightarrow{G}$. In addition, each edge includes information on the corresponding interaction, such as the number of reactions, the type of the interaction (such as comment or reply), or the creation time.

5.2 Modeling Temporal Networks

In this section, we refined the interaction graph model to capture the time-variant nature of the network. In particular, we used the temporal network formalism in order to model the evolution of interactions and their dynamical properties.

Temporal networks: terminology and notation. Static graphs, such as the *Group Interaction Graph*, are widely used to analyze the structural properties of systems, but this model cannot capture the structural–temporal aspects of the systems that arise from the time ordering of events. Since our dataset is a collection of interactions that occurred at some points in time, we used the graph formalisms introduced in Section 3.2 to take into account the temporal dimension of the network and to capture basic schemes of communication between users.

Even though there may be a substantial distinction between the different terminologies proposed in the scientific literature, we will use the formalism introduced in [8, 24, 29] as it is one of the most intuitive and easy to understand. Such a graph formalism is widely adopted in several scenarios to measure the evolution of the centrality, influence, or popularity of the users/contents over time.

In particular, a temporal network is represented by dividing the time period to be monitored into different slices of equal duration Δ, and by creating a *Group Interaction Graph* $\overrightarrow{G_t} = \langle V_t, E_t \rangle$ for each slice $[t, t + \Delta)$. The set E_t represents the interactions $\overrightarrow{e}(v, w)$, which occurred between the members of a group in the slice $[t, t + \Delta)$, while the set V_t contains the endpoints of such interactions. Indeed, each interaction $\overrightarrow{e} \in E_t$ is paired with a time stamp $t^{\overrightarrow{e}}$, which indicates the time when the event occurred. It holds that $t^{\overrightarrow{e}} \in [t, t + \Delta)$. Hence, we indicate with $\overrightarrow{G_t} = \langle V_t, E_t \rangle$ the *Group Interaction Graph* that considers only the interactions that occurred within the time interval $[t, t + \Delta)$, while $\overrightarrow{G} = \langle V, E \rangle$ represents the *Group Interaction Graph* for the entire duration of the monitored period. Figure 6 shows a simple *Group Interaction Graph* obtained for a group of five members when the duration of the slice is equal to Δ.

However, if we consider a general *Group Interaction Graph* $\overrightarrow{G_t} = \langle V_t, E_t \rangle$ at a time t, it is possible that a reply to a comment made in the current slice $[t, t + \Delta)$ refers to a post p previously published at a time $t^p < t$. In such a case, we include in the temporal network $\overrightarrow{G_t} = \langle V_t, E_t \rangle$ the comment

Table 6. Metrics Used to Compute Centrality

Metrics	Geometric	Path	Spectral	Value	Variable
Reactions	x			$[0, \infty)$	*reactions*
Reshare	x			$[0, \infty)$	*reshare*
Edge Degree Centrality	x			$[0, \infty)$	d_v^+, d_v^-
Node Degree Centrality	x			$[0, \infty)$	k_v^+, k_v^-
Interaction Rate	x			$[0, \infty)$	r_v^+, r_v^-
Activity Rate	x			$[0, 1]$	a_v
H-Index	x			$[0, \infty)$	h_v
Closeness Centrality		x		$[0, \infty]$	c_v
Betweenness Centrality		x		$[0, \infty]$	bc_v
Page Rank			x	$[0, 1]$	pr_v
Eigenvector Centrality			x	$[0, 1]$	en_v

interaction between the author of the comment and the author of the post in order to lead the importance of the interaction back to the person who initially created the post. In our scenario, the duration of the interactions between users is not important because comments and replies occur suddenly. For this reason, the edges of $\overrightarrow{G_t} = \langle V_t, E_t \rangle$ are defined for the entire period of the slice $[t, t + \Delta)$. The result of the *Data modeling* phase is a *temporal network* that consists of an ordered sequence of contiguous directed *Group Interaction Graph*s:

$$\overrightarrow{G_{t_1}}, \overrightarrow{G_{t_2}}, \overrightarrow{G_{t_3}}, \ldots, \overrightarrow{G_{t_s}} \qquad t_1 < t_2 < t_3 < \cdots < t_s, \tag{1}$$

where each *Group Interaction Graph* G_{t_i} refers to the slice $[t_i, t_i + \Delta)$ and $t_{i+1} - t_i = \Delta$.

6 DATA TRANSFORMATION

The data model defined in the previous section is used by the *Data transformation* phase to derive a number of metrics of importance that might be relevant for measuring the influence of members within the group. In particular, the metrics we consider use the formalism of the temporal network defined above to derive the importance of the members at different time intervals $t_1, t_2, \ldots, t_s$. Hereafter, we identify a set of metrics that can be considered to implement the framework in order to predict the influence of the members, and we show the process we defined to select the ones that will be used in the *Training and Prediction* phase.

6.1 Time-Aware Centrality Metrics

Given a temporal network $\{\overrightarrow{G_{t_1}}, \overrightarrow{G_{t_2}}, \overrightarrow{G_{t_3}}, \ldots, \overrightarrow{G_{t_s}}\}$ of a group (where $t_1 < t_2 < t_3 < \cdots < t_s$) and a general *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network (i.e., $t_i \in \{t_1, \ldots, t_s\}$) representing the group in the time interval $[t_i, t_i + \Delta)$, the `Time-Aware Centrality Metrics` task computes for each member $u \in V_{t_i}$ the vector of metrics $M_u^{t_i} = [m_1, m_2, \ldots, m_N]$, where N is the number of metrics and the element m_j of the vector (for j=1, 2, ..., N) represents the value of the metric j for user u in the time interval $[t_i, t_i + \Delta)$ under consideration.

As shown in Table 6, the collected data permit us to select 11 metrics of importance, which, according to [11], belong to three categories: geometric, path, or spectral. The geometric-based metrics focus on a distance function between users, and they consider the number of users existing at a specified distance. Among the most popular geometric-based centrality metrics, we mention the *Edge/Node Degree Centrality* and the *H-Index*. In addition to those classical centrality

metrics, we also take into account metrics that consider the relationship between the outgoing and incoming interactions (*activity rate*), the amount of interactions intended for each user (*interaction rate*), the number of *reactions*, and the number of *reshares* obtained by interactions published by a user. Such metrics have the advantage of being computed by using only the information of the node of $\overrightarrow{G_{t_i}}$ they refer to, or the information from the direct neighbors of such a node (such as the number of reactions, comments, reshares obtained by a friend). Instead, path-based metrics compute the shortest paths between nodes, like *Betweenness Centrality* and *Closeness Centrality*. Finally, we also consider spectral-based centrality, such as the *Page Rank* and the *Eigenvector Centrality*, which take into account spectral properties of the graph matrices. Hereinafter, we focus on describing the calculation of the metrics listed in Table 6, omitting to discuss the first two, number of reactions and reshares, because they are intuitive.

Edge Degree Centrality. It measures the number of links incident upon a node. Since we consider a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network, we have two different versions of the Edge Degree Centrality depending on the direction of the edges. The Edge Out-Degree Centrality d_v^+ of a node v measures the number of interactions that occurred from node v to any node w in the graph. Similarly, the In-Degree Centrality d_v^- of a node v measures the number of distinct edges in the graph connecting any vertex w to the vertex v. As a result, the Edge In-Degree and the Edge Out-Degree of a node v is computed by measuring the size of the following sets:

$$d_v^+ = \{\overrightarrow{e}(v, w) \in E | w \in V_{t_i}\}, \tag{2}$$

$$d_v^- = \{\overrightarrow{e}(w, v) \in E | w \in V_{t_i}\}. \tag{3}$$

Node Degree Centrality. It measures the number of users who interacted with a specific user $v \in V_{t_i}$. As for the case of the Edge Degree Centrality, we consider two different versions of the Node Degree Centrality depending on the direction of the edges. The Node Out-Degree Centrality k_v^+ measures the number of distinct users who received one or more interactions from node v in the graph. More formally, it considers the collection of nodes that are connected through ingoing edges incident on v. Similarly, the Node In-Degree Centrality k_v^- of v measures the number of distinct users who initiated one or more interactions with vertex v. Given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network, the Node In-Degree and the Node Out-Degree of a node v are given by the sizes of the following sets:

$$k_v^+ = \{w \in V | \overrightarrow{e}(v, w) \in E_{t_i}\}, \tag{4}$$

$$k_v^- = \{w \in V | \overrightarrow{e}(w, v) \in E_{t_i}\}. \tag{5}$$

Interaction rate. It measures the average number of interactions performed by user v. We consider two different versions of the interaction rate depending on the direction of the edges. The Output Interaction Rate r_v^+ is the ratio between the number of distinct interactions initiated by v (i.e., d_v^+) and the number of distinct users who received these interactions from node v (i.e., k_v^+). Similarly, the Input Interaction Rate r_v^- of node v is the ratio between the number of distinct interactions received by v (i.e., d_v^-) and the number of distinct users who initiated these interactions (i.e., k_v^-). Given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network, the Output/Input Interaction Rate of a node v is computed by using the following equations:

$$r_v^+ = \begin{cases} \frac{|d_v^+|}{|k_v^+|} & \text{if } k_v^+ \neq \emptyset \\ 0 & \text{if } k_v^+ = \emptyset \end{cases}, \qquad r_v^- = \begin{cases} \frac{|d_v^-|}{|k_v^-|} & \text{if } k_v^- \neq \emptyset \\ 0 & \text{if } k_v^- = \emptyset \end{cases}. \tag{6}$$

It is important to note how users with no ingoing interactions have an Input Interaction Rate equal to 0, while users who receive only one ingoing interaction for each ingoing neighbor in k_v^- have an Input Interaction Rate equal to 1.

Activity rate. It measures the activity level of the user by using both ingoing and outgoing interactions. The Activity Rate a_v is the fraction of the ingoing interactions of node v. Given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network, the Activity Rate of a node v is computed by using the following equations:

$$a_v = \begin{cases} 0 & \text{if } |d_v^+| + |d_v^-| = 0 \\ \frac{|d_v^-|}{|d_v^+| + |d_v^-|} & \text{otherwise} \end{cases}. \tag{7}$$

It is important to note how the Activity Rate of users is within the range of $[0, 1]$. In particular, the Activity Rate of a user v is 0 if the user v has not performed any ingoing interactions in the graph (i.e., $|d_v^-| = 0$), while it is equal to 1 if v has not performed any outgoing interactions (i.e., $|d_v^+| = 0$). When the Activity Rate of v is greater than 0.5, we say that v is a *consumer* because v has more ingoing interactions than the outgoing ones. Instead, we say that v is a *producer* if v has more outgoing interactions than the ingoing ones, i.e., the Activity Rate of v is less than 0.5.

H-Index. It is typically used to measure both the productivity and citation impact of the publications of a scientist or scholar. A user with an index of h has published h papers each of which has been cited in other papers at least h times. We extended the H-Index measure to the case of a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network, by using the definition proposed in [37]. In particular, the H-Index $H(X)$ of a finite number of reals $X = (x_1, x_2, \ldots, x_i)$ returns the maximum integer h such that there exist at least h elements in $(x_1, x_2, \ldots, x_i)$, each of which is greater than h. Given the collection k_v^- of nodes who have interacted with v in $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$, the H-Index h_v of a node $v \in V$ computes, for each neighbor $w \in k_v^-$, the number of directed edges from w to u, i.e., $|\overrightarrow{e}(w, v)|$. Finally, the finite sequence of the interactions between the neighbours and the user v is given as an input parameter to the function H. Formally, we have

$$h_v = H(|\overrightarrow{e}(u_1, v)|, |\overrightarrow{e}(u_2, v)|, \ldots, |\overrightarrow{e}(u_i, v)|), \qquad u_1, u_2, \ldots, u_i \in k_v^-. \tag{8}$$

Closeness Centrality. The Closeness Centrality measures the global importance of a node in the graph by calculating the sum of the lengths of the directed shortest paths between the node and all other nodes in the graph. Given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ and a node $v \in V$, the Closeness Centrality c_v is computed by using the following equations:

$$c_v = \sum_{w \in V_{t_i}} \frac{1}{d(w, u)}, \tag{9}$$

where d is a function that measures the distance from node w to node u. The Dijkstra shortest path algorithm is used to compute distances in a specified graph based on the minimum cost path. The classical definition of Closeness Centrality assumes that all edge weights in the graph are equal to 1, and it does not consider multiple edges between nodes. We refined this definition by applying a transformation function that computes the weight between two nodes as the reciprocal of the number of interactions. A node v who frequently interacts with several members of the group will have a high Closeness Centrality c_v because v is close to all the members of the group.

Betweenness Centrality. The Betweenness Centrality measures the importance of a node in terms of aiding information diffusion toward disconnected groups of members. The Betweenness Centrality of a node is computed by calculating the fraction of directed shortest paths traversing that

specific node. Given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ and a node $v \in V_{t_i}$, the Betweenness Centrality bc_v is computed by using Equation (10), where σ_{st} is the total number of the shortest paths between users s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t that cross node v:

$$bc_v = \sum_{s \neq v \neq w} \frac{\sigma_{st}(v)}{\sigma_{st}}. \tag{10}$$

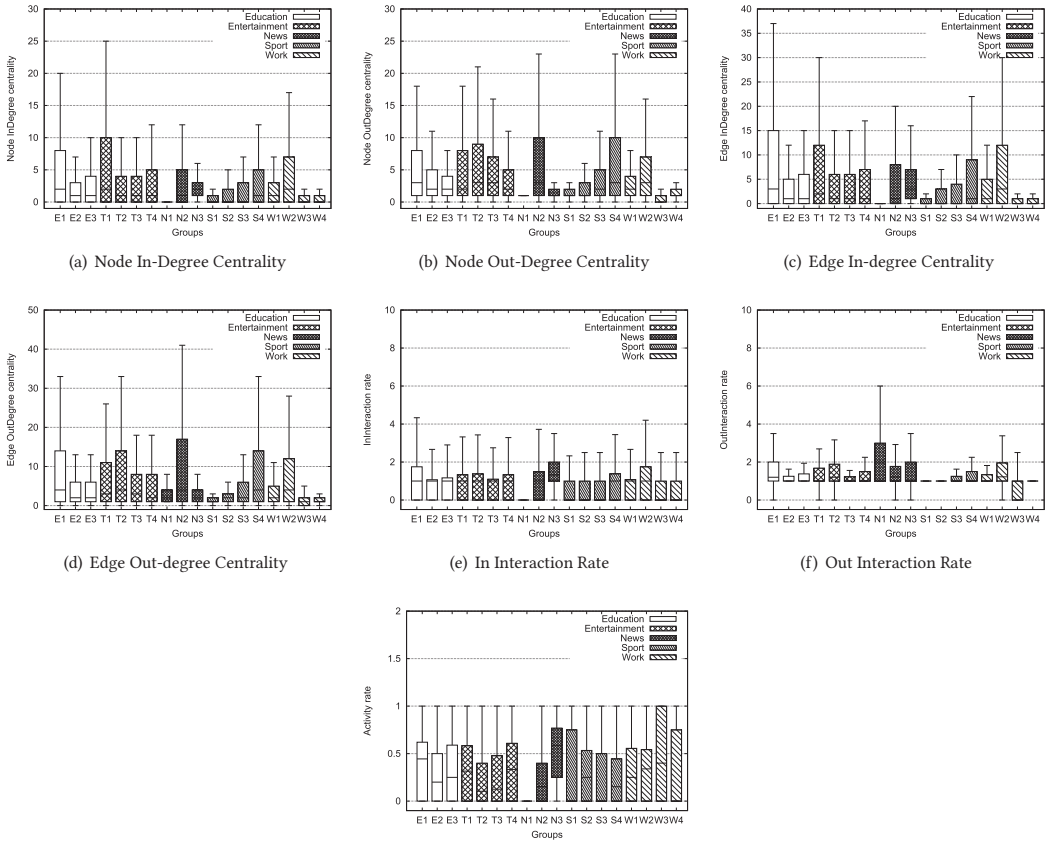
A user v with high Betweenness Centrality bc_v acts as a bridge between members of the group who do not interact with each other. We selected the number of shortest paths between two nodes by using the Dijkstra shortest path algorithm, which is used to calculate distances in a specified graph based on the minimum cost path.

Page Rank. It measures the importance of web pages by search engine, and the underlying assumption is that more important websites are likely to receive more links from other websites. Given the likelihood of choosing a random link from the current page and the likelihood of jumping to a page chosen at random from the entire Web, the Page Rank indicates the probability that a user will reach a specific page. According to the PageRank algorithm, each vertex represents a web page, and the Page Rank score assigned to it can be considered as the fraction of time spent by a user visiting that vertex in a random walk (following outgoing edges from each vertex). The Page Rank pr_v of a user v modifies this random walk by adding to the model a probability, identified as *alpha*, of jumping to any vertex of the given *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$. If *alpha* is 0, this is equivalent to the Eigenvector Centrality algorithm; if *alpha* is 1, all vertices will receive the same score ($1/|V_{t_i}|$). We set the parameter *alpha* equal to 0.15 because the typical values for *alpha* are in the range [0.1, 0.2], but it may be any value between 0 and 1 inclusive.

Eigenvector Centrality. It is a measure of the centrality of a node, which considers the neighbors of that node. In particular, the Eigenvector Centrality en_v of a user v depends on both the number and the importance of neighbors. Typically, it is computed by considering the sum of the Eigenvector centralities of direct neighbors, but, in some cases, it could be approximated by using iterative algorithms based on random walk.

6.1.1 Evaluation of the Centrality Measures. In this section, we investigate in more detail how the values of the centrality measures are distributed among groups of different types. For this reason, we built the *Group Interaction Graph* $\overrightarrow{G} = \langle V, E \rangle$ of each group, which contains all the interactions that occurred in the group during the monitored period, and we computed the influence of each user in $\overrightarrow{G}$ by using the centrality measures described above. The graphs in Figure 7 show the box plot of each centrality measure for each of the groups we selected by indicating the minimum, lower quartile (corresponding to the 25th percentile), median, upper quartile (corresponding to the 75th percentile), and maximum values. It is worthwhile to point out that, in Figure 7, for all those boxes where the median line is not shown, the median value is the same as the lower quartile. For instance, for group S1, the median value is equal to the lower quartile for all the metrics. Instead, for group S2, the median value is equal to the lower quartile for all the metrics but Edge Out-Degree Centrality and Activity Rate.

The Node In-Degree (Figure 7(a)) and the Node Out-Degree (Figure 7(b)) metrics have a right-skewed distribution, and the median values are very close to the minimum ones. From Figure 7(a), we see that the highest maximum value among the groups for the Node In-Degree Centrality is 25 (group T1), while the average of the maximum values of all the groups is equal to 9.08. In order to understand whether there are influential members in the groups, we compare the previous values with the maximum and the average upper quartile values of all the groups. The highest upper



(a) Node In-Degree Centrality

(b) Node Out-Degree Centrality

(c) Edge In-degree Centrality

(d) Edge Out-degree Centrality

(e) In Interaction Rate

(f) Out Interaction Rate

(g) Activity Rate

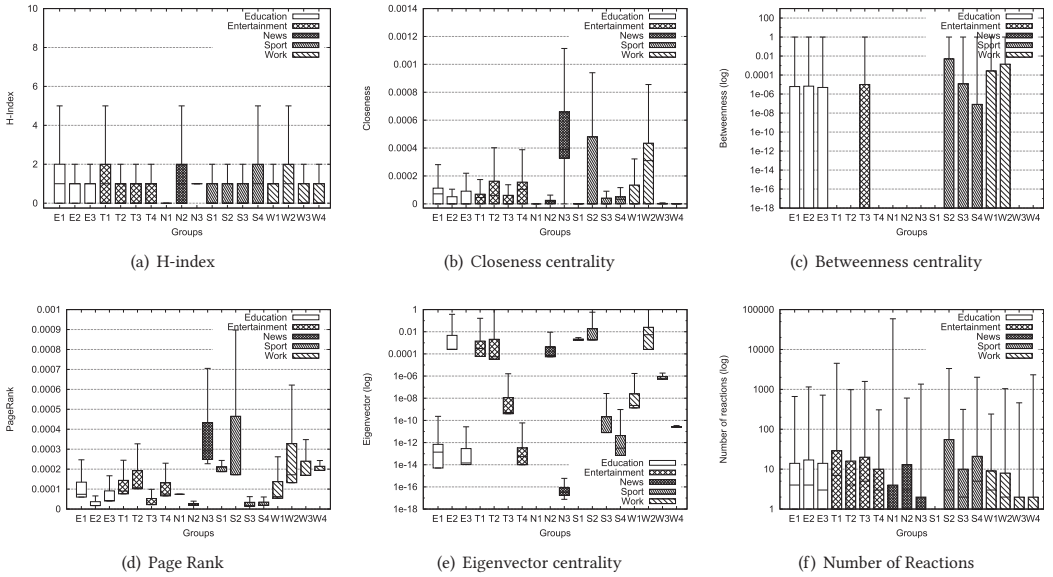
Fig. 7. Evaluation of the Node Degree Centrality, Edge Degree Centrality, and Interaction Rate of the groups' members.

quartile value for In-Degree Centrality among all the groups is 10, and it is obtained for group T1. Hence, for group T1, the maximum value (25) is 250% of the highest upper quartile (10). Instead, the average of the upper quartile values of all the groups is 3.83. Consequently, the average of the maximum values (9.08) is 237% of the average of the upper quartiles.

Figure 7(b) shows that the highest maximum value and the upper quartile of the Node Out-Degree Centrality among all the groups are 24 and 10, respectively (groups N2 and S4). Instead, the average of the maximum values of all the groups is 11.22, while the average upper quartile value of all the groups is 5.2. Again, the highest/average maximum value is more than twice the highest/average upper quartiles.

We notice that the Node In-Degree Centrality for most groups exhibits slightly smaller values than the Node Out-Degree Centrality. (The average upper quartile of all the groups for the In-Degree and the Out-Degree Centrality is 3.83 and 5.2, respectively.). A possible interpretation of this difference could be that the comments and replies published by most of the members of the groups are all directed to the posts and comments published by a smaller set of other members.

The Edge In-Degree (Figure 7(c)) and the Edge Out-Degree (Figure 7(d)) metrics exhibit a right-skewed distribution as well. In particular, most members perform and receive few interactions, while a small fraction of members exhibit a higher activity level. As a matter of fact, for the Edge



(a) H-index (b) Closeness centrality (c) Betweenness centrality

(d) Page Rank (e) Eigenvector centrality (f) Number of Reactions

Fig. 8. Evaluation of the H-Index Centrality, Closeness Centrality, and Betweenness Centrality of the *Group Interaction Graph*.

In-Degree Centrality, the highest maximum value is 37 (group E1), and the average maximum value among all the groups is 14, while the highest upper quartile is 15 (group E1), and the average upper quartile value among all the groups is 6. Instead, for the Edge Out-Degree Centrality, the highest and average maximum values among all the groups are, respectively, 41 (group N2) and 17, while the highest and average upper quartile values among all the groups are, respectively, 17 (group N2) and 6.

The interaction rate of the groups depends on the direction of the links, as shown in Figure 7(e) and (e). For the In Interaction Rate Centrality metric, group E1 exhibits the highest maximum value (4.3) and the highest upper quartile value (1.8) among all the groups. Moreover, the average of the maximum values of all the groups is 3.3, while the average of the upper quartile values of all groups is 1.21. Concerning the Out Interaction Rate metric, group N1 exhibits both the highest maximum value (6) and the highest upper quartile value (3) of all the groups. Instead, the average maximum value and the average upper quartile value among all the groups are, respectively, 2.6 and 1.54.

Figure 7(g) shows the Activity Rate distributions of the groups. Most members of all the groups except for N3 are producers because the median Activity Rate is below 0.5, i.e., the majority of the group's members posted more comments/replies than those they received from the other members of the group. Group W3 is more balanced with respect to the other groups, and it has the highest upper quartile value (1) of Activity Rate among all the groups. Most groups have upper quartile values greater than 0.5, indicating that about 25% of the group's members consist of consumers, i.e., they attracted many interactions from the other members of the group.

Another important metric that produces very interesting results is the H-Index of members, which is shown in Figure 8(a). The majority of the groups (E2, E3, T2, T3, T4, S1, S2, S3, W1, W3, and W4) have very right-skewed distribution and their maximum H-Index is equal to 2. The highest maximum value and the highest upper quartile value of the H-Index Centrality for all the groups are 5 and 2 (groups E1, T1, N2, S4, and W2), respectively. Figure 8(b) shows the distribution

of the normalized Closeness Centrality of each group. Most groups exhibit a very low Closeness Centrality value and right-skewed distribution (such as groups of the Education and Entertainment categories). Group N3 of the News category exhibits the highest maximum value (0.00112) and the highest upper quartile value (0.00064) of Closeness Centrality among all the groups. The average of the maximum values of all the groups is equal to 0.0003, while the average of the upper quartile values of all the groups is 0.00014.

The normalized Betweenness Centrality of the groups' members is shown in Figure 8(c). The graph clearly indicates the presence of two different trends. About half of the groups include a significant number of members with very low Betweenness Centrality, while the other groups have a large fraction of members having Betweenness Centrality equal to zero. The highest maximum value of Betweenness Centrality is 1 (groups E1, E2, E3, T3, S2, S3, S4, W1, and W2), while the highest upper quartile value is 0.004 (group S2).

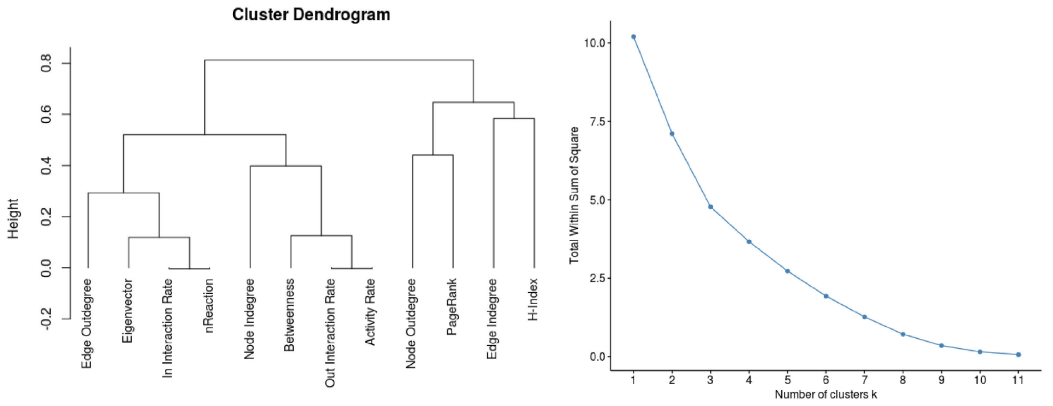
As shown in Figure 8(d), the Page Rank distributions of the groups' members do not exceed the value of 0.001, and the median Page Rank of the members is very close to the minimum value. The highest maximum value of Betweenness Centrality (0.0009) and the highest upper quartile value (0.00058) are exhibited by members of group S2. Furthermore, the length of the upper segments indicates that the Page Rank distribution exhibits a long tail. Indeed, the average of the maximum Page Rank values for all the groups is 0.00028, while the average upper quartile value for all the groups is 0.00017.

The distributions of the normalized Eigenvector Centrality are shown in Figure 8(e), and they are very different in terms of median Eigenvector Centrality. Indeed, the median Eigenvector Centrality of the groups is uniformly distributed over the entire domain, while the majority of the distributions have positive skewness. The highest maximum values of Eigenvector Centrality (1) are exhibited by members of groups T2 and W2, while the highest upper quartile value (0.03) is obtained only by group W2. The average upper quartile value of the Eigenvector Centrality for all the groups is 0.003, indicating the presence of a small fraction of members (about 25%) who are very central for the groups.

Finally, as shown in Figure 8(f), the median number of reactions received by groups' members for their interactions is fewer than 10 for all groups. However, the plot reveals the presence of a small fraction of members who received a relevant number of reactions (more than 100) for their interactions, while most interactions (about 74%) exhibit fewer than 10 reactions. The highest maximum number of reactions is obtained by members of group N1 (500,000), while the highest upper quartile value of the number of reactions is equal to 55 (group S2). The average upper quartile value of the number of reactions is equal to 14 for all the groups, while the average maximum number of reactions for all the groups is equal to 4,000.

6.2 Metrics Selection

Although the metrics defined in the previous section allow us to capture interesting aspects related to the importance of the members within their groups, some of them may not be very useful for identifying the most influential ones. Moreover, some metrics could exhibit close dependencies among them, thus being not useful in applying them simultaneously in the Influencers Prediction process. Hence, we analyze the metrics to discover or confirm dependencies between them. To this aim, we compute the correlation matrix C_G of each group G by obtaining the Pearson correlation coefficient (r) between each pair of centrality measures. The value of r ranges between +1 and −1, where +1 indicates a positive linear correlation, 0 indicates no linear correlation, and −1 indicates a negative linear correlation. The correlation among metrics is computed by considering the real observations $M_u^{t_1}, M_u^{t_2}, \ldots, M_u^{t_s}$ collected for each member u of the group at each time interval t_i of the temporal networks (for $i = 1, \ldots, s$). The results are represented by the *Dendrogram* [36] shown



(a) Cluster Dendrogram (b) Optimal Number of clusters

Fig. 9. Evaluation of the optimal number of clusters.

in Figure 9(a), where we use cumulative hierarchical clustering to group the metrics according to how closely correlated they are.

Figure 9(b), instead, shows the *Total Within Sum of Square* of the clusters obtained by using a hierarchical clustering algorithm. The *Total Within Sum of Square* is computed as the sum, over all observations, of the squared differences of each observation from its cluster's centroid.

The plot of Figure 9(b) can help determine a good number of clusters, while the *Dendrogram* in Figure 9(a) shows how to build such clusters of metrics, and the height is the euclidean distance metric between clusters. For instance, to build three clusters (to compute the value of the *Total Within Sum of Square* for *number of clusters* $k = 3$), we see from the *Dendrogram* that we should choose *height* $= 0.6$, thus enabling H-Index and the Edge In-Degree form one cluster, Page Rank and Node Out-Degree metrics form another cluster, and the remaining metrics constitute the third cluster.

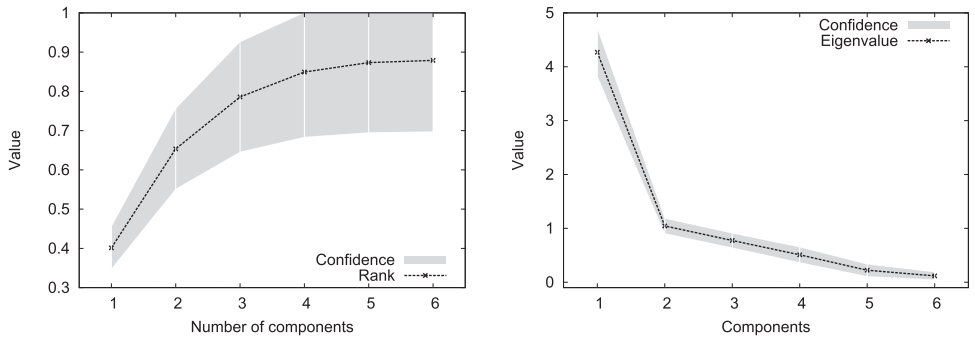
Summarizing, from Figure 9(a) and (b), we observe that the metrics we considered exhibit non-trivial relationships among them, which must be taken into account for the purpose of the Influencers Prediction.

The `Metric Selection` task performs a first reduction of the number of metrics used for the prediction process by filtering out the ones that are not relevant for identifying influential members. Since we are interested in identifying those users who are more likely to attract a high number of interactions, we consider only the metrics that evaluate the incoming activity of users, namely, Reactions, Reshare, Node In-Degree, Edge In-Degree, In Interaction Rate, Activity Rate, H-Index, Closeness, Betweenness, Page Rank, and Eigenvector metrics. In addition to this filter, the `Metric Selection` task removes the metrics whose variance is below a specified threshold. Indeed, the metrics with very low variance are likely to affect the performance of certain prediction algorithms, and they are filtered out from the dataset. The Low Variance Filter suggests removing the Eigenvector, Reshare, and Betweenness metrics because they carry little information (such as the Reshare metric, which is equal to 0 for almost all groups). The remaining metrics, which will be used in the next task, as well as the groups on which they will be applied, are summarized in Table 7.

6.2.1 Principal Component Analysis. In order to further reduce the number of metrics and to mitigate the multicollinearity issues among them, we performed a Principal Components Analysis (PCA) of the selected metrics. Indeed, PCA allows us to transform the original set of N observed

Table 7. Metrics and Groups Considered by the `Metric Selection` Task

Metrics calculated for all groups	Category	Groups
Reactions, Node In-Degree, Edge In-Degree, In Interaction Rate, Activity Rate, H-Index, Closeness, Page Rank	News	N2, N3
	Education	E1, E2, E3
	Sport	S1, S2, S4
	Entertainment	T1, T2, T3, T4
	Work	W1, W2, W3, W4



(a) Variance explained by principal components

(b) Eigenvalues of each component

Fig. 10. Results of the PCA.

metrics $M_u^{t_i} = \{m_1, m_2, \ldots, m_N\}$ of user u into another set of metrics $C_u^{t_i} = \{c_1, c_2, \ldots, c_{N'}\}$ of size N' (named principal components) having a smaller dimension, i.e., $N' < N$. In particular, given a generic *Group Interaction Graph* $\overrightarrow{G_{t_i}} = \langle V_{t_i}, E_{t_i} \rangle$ of the temporal network representing the interaction graph of the group in the time interval $[t_i, t_i + \Delta)$, the vector of metrics $M_u^{t_i} = [m_1, m_2, \ldots, m_N]$ computed by the `Time-Aware Centrality Metrics` task for each member $u \in V_{t_i}$ can be transformed (or projected) into another vector $C_u^{t_i} = [c_1, c_2, \ldots, c_{N'}]$ of components such that the dimension of $C_u^{t_i}$ is less than the dimension of $M_u^{t_i}$.

Besides being orthogonal, the resulting components $C_u^{t_i}$ have most of the information of the original metrics, i.e., they preserve a large fraction of the original variance of $M_u^{t_i}$. In order to compute PCA, we create a dataset that consists of the observations $\{M_u^{t_1}, M_u^{t_2}, \ldots, M_u^{t_s}\}$ obtained from the temporal network on each user u of the group, and we use *z-score standardization* to avoid dependencies on the scale of the metrics. The principal components of the group's members are obtained from the covariance matrix of the standardized metrics, by computing the normalized eigenvectors for the corresponding eigenvalues. Figure 10 summarizes the results of the PCA by showing the eigenvalue of each component (Figure 10(b)) and the cumulative fraction of explained variance (Figure 10(a)). As shown in Figure 10(b), the order of the components is given by the eigenvalues where the vector corresponding to the largest eigenvalue is the first component, while the vector of the smallest eigenvalue is the last component. Indeed, Figure 10(a) indicates that the first principal component preserves a large fraction of the original variance (about 40%), whereas adopting a two-dimensional transformation of the original metrics, the variance preserved by the first two principal components is about 60%. In order to maximize the advantage of PCA for the prediction of the influential members, the number of components to be considered for data transformation is taken as an input parameter by our framework. Hence, the Influencers Prediction process can be

performed on the same data several times varying the number of components in order to find the optimal one, i.e., the one that results in the best prediction accuracy.

As a result, the *Data transformation* phase will return a set of metrics $\{c_1, \ldots, c_{N'}\}$ for each member of the group and for each *Group Interaction Graph* of the temporal network, where N' is the number of principal components chosen as an input parameter.

7 TRAINING AND PREDICTION

The *Training and Prediction* phase uses the transformed metrics obtained from the *Data transformation* phase to forecast the most influential members of the groups and to quantify the amount of such influence. As usual, in order to build the model for the prediction, a training set is created.

7.1 Training Set

The training set consists of a sequence of S transformed observations $\{C_u^{t_1}, C_u^{t_2}, \ldots, C_u^{t_s}\}$ related to a user u, and they are ordered according to the time in which they occurred. In particular, each transformed observation $C_u^{t_i}$ corresponds to a real observation $M_u^{t_i}$ and s is an input parameter that represents the number of observations of the user u on which the `Prediction` task is executed. In our framework, each observation is performed at discrete time (such as hourly, daily, or weekly), depending on the duration of the slice period of the temporal network. For instance, if the granularity Δ of the temporal network is equal to 1 day, one possible strategy is to provide a training set that consists of the observations made in the last 7, 14, or 21 days for the user u. Alternatively, when the granularity of the temporal network is 1 week, a possible training set may consist of the observations made in the last 4, 8, or 12 weeks. To investigate whether the size of the training set and the time resolutions of the temporal network affect the prediction, the Influencers Prediction process can be performed varying the size of the training set and considering different values of granularity Δ for the temporal network (such as daily or weekly) in order to compare the resulting prediction accuracy.

Each transformed observation $C_u^{t_i} = [c_1, c_2, \ldots, c_{N'}]$ of the training set is related to a specific member u at a time t_i, and it consists of N' values representing the transformed metrics returned from the *Data selection* phase (described in Section 6). In order to obtain individual influence scores of u from the transformed observation $C_u^{t_i} = [c_1, c_2, \ldots, c_{N'}]$, we combine the contribution of each component by using the following approach. If the number of components N' of the observation is equal to 1, then the influence score of the observation is equal to the value of the individual component. Instead, if the number of components N' of the observation is greater than 1 (i.e., $N' > 1$), the influence score must be computed by combining the values of components. In such a case, the corresponding influence score is computed by simply summing the values of individual components. Formally, the influence score $I_{t_i}(u)$ of the group's member u at a time t_i is equal to $I_{t_i}(u) = \sum_{j=1}^{N'} C_u^{t_i}[j]$.

7.2 Prediction

As observed in Section 3, several algorithms can be used for the prediction of the most influential members, each one having its own specific strengths and limitations. Hence, the adoption of distinct prediction algorithms in the Influencers Prediction process could result in different accuracies of the results. For this purpose, we used WEKA,[6] a very popular open source Machine Learning (ML) software package that provides stable implementations of the following prediction algorithms commonly used for `Prediction` tasks:

[6] https://www.cs.waikato.ac.nz/~ml/weka/ [last accessed on May 2020].

—Linear predictor (LP) algorithm is a simple and intuitive linear prediction model that has been successfully used in the field of OSN [3, 21]. Formally, given the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ computed on the member u until the time t_s, a linear predictor is defined as a linear combination $f(I_{t_{s+1}})$ constructed from a set of S terms $I_{t_1}(u), \ldots, I_{t_s}(u)$ by multiplying each term I_{t_i} with the corresponding weight β_i:

$$f(I_{t_{s+1}}(u)) = \beta_1 \cdot I_{t_1}(u) + \beta_2 \cdot I_{t_2} + \cdots + \beta_s \cdot I_{t_s}(u). \tag{11}$$

In particular, weights β_i for $i = 1, \ldots, s$ are named coefficients, and they are used to fine-tune the impact of each independent variable $I_{t_i}(u)$ in predicting the outcome of a dependent variable $f(I_{t_{s+1}}(u))$. In the context of Influencers Prediction problem, the linear predictors can be used to calculate the most influential members of a group in a given future period of time, by taking into account the influence score of members in the last s different time instants. The coefficient vector can be used in the fitting process to indicate how much each day contributes to the prediction of the availability status. To be able to compare probability measures, we performed a normalization by dividing it by the sum of the weights β, obtaining a normalized value between 0 and 1.

—Linear regression (LR) allows us to find the straight line that best fits a set of data points. Formally, given the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ computed on the member u until the time t_s, a linear regression estimates the line $\widehat{I}$ that best represents the influence score at different times $t_1, t_2, \ldots, t_s$:

$$\widehat{I}(t) = b + w \cdot t, \tag{12}$$

where t is the time predictor variable, $\widehat{I}$ is the estimate of the influence score, while b and w are regression coefficients corresponding to the Y-intercept and slope of the line, respectively. These coefficients are obtained from actual data according to the least-squares method, which minimizes the error between them and the estimate of the line. The resulting function $\widehat{I}$ can be used to predict the influence score of a user u after the time t_s.

—K-nearest neighbours (KNN) is a very popular prediction method for both classification and regression problems, which is widely used in the area of pattern recognition [1]. Given the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ of the member u until the time t_s, the approach used to predict the influence of u at time t_{s+1} consists of assigning the average influence score of its KNN. For this reason, a distance metric (such as Euclidean or Manhattan distance) is used to identify the KNN at time t_{s+1}. The number of neighbors k used to predict the influence score is an input parameter that impacts the accuracy of the model.

—Classification And Regression Trees (CART) [12] are decision tree induction algorithms adapted to predict continuous (ordered) values. Given the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ of the member u until the time t_s, the first step consists of using a greedy approach to divide the time space at different points, according to a cost function (i.e., the information gain). For each split, the algorithm creates two child nodes, assigns to them the corresponding points resulting from the split, and recursively repeats the procedure on the child nodes, until there are enough instances to split. Each leaf of the regression tree consists of a set of influence scores, and the average value of this set represents the influence score predicted by the regression tree. The last step prunes the tree by using a reduced-error pruning technique that cuts off the less important nodes of the tree to further refine the performance.

—Support Vector Regression (SVR) is a popular regression approach that relies on kernel functions, and it has demonstrated excellent performance in several real-world applications [49]. Given the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ of the member u until the

Table 8. Input and Configuration Parameters of the Influencers Prediction Process

Workflow phase	Variable	Description	Value
Data selection	SF	sampling frequency of the crawler	1 day
Data modeling	Δ	duration of the period of the slice for temporal networks	1 day, 1 week
Data transformation	N	number of centrality metrics to compute	1, 2, 3, 4...
	N'	number of components for PCA	1, 2, 3, 4...
Training and Prediction	s	number of observations in the training set	1, 2, 3, 4...
	Top-k	number of the most influential members to predict	5, 10, 15, 20
	P	prediction algorithm	LP, KNN, SVR, LR, MLP, CART
Evaluation	z	number of observations in the test set (validation check points)	{1 day, 1 week}

time t_s, the goal of the SVR is to estimate a function

$$\widehat{I}(t) = w \cdot t + b \tag{13}$$

that is as close as possible to the influence score $I(t_i)$ for each t_i. The term w is a weight vector that is used in the optimization problem to minimize the size of the margin, while b is a scalar number named bias. In some cases, the input data cannot be approximated by a linear hyperplane and they must be transformed into a higher dimensional space using kernel functions and then searching for a linear hyperplane in the new space [19]. The most used types of kernels are radial basis function kernels and polynomial kernels.

— Multi-Layer Perceptron (MLP) is a neural network that uses a back-propagation algorithm for learning input data. In general, an MLP consists of several layers each of which has one or more units. The input layer takes the set of influence scores $\{I_{t_1}(u), I_{t_2}(u), \ldots, I_{t_s}(u)\}$ of the member u until the time t_s, multiplies them by a set of weights, and sends the values to the next level, known as a hidden layer. The number of hidden layers is an input parameter of the method, and the result of the last hidden layer is a set of weighted outputs. The last layer of the MLP is the output layer, which returns the result of the prediction. The weights used in each level are updated in order to minimize the mean squared error between the predicted influence score and the actual influence score.

8 EVALUATION

The *Evaluation* phase is meant to estimate the accuracy of the prediction generated by the Influencers Prediction process by verifying how many of the k-predicted influential members turned out to be really the most influential members of the group in a given future time period. The result of the prediction (and, consequently, its accuracy) depends on the values of the parameters chosen for the execution of the operations building up the Influencers Prediction process (described in the previous sections and summarized in Table 8). For instance, Figure 12 shows that the choice of the duration of the slice used to build the temporal network in the *Data modeling* phase really affects the accuracy of the prediction. Hence, in this section, we execute the Influencers Prediction process several times varying the values of such parameters, and, by computing the accuracy of each of these predictions through the *Evaluation* phase, we determine the parameter values that ensure the best prediction accuracy. Once the optimal parameters have been determined, the *Evaluation* phase is not executed anymore, and the next executions of the Influencers Prediction process will leverage such values in order to obtain the best prediction accuracy.

In order to execute the *Evaluation* phase, we need to know the real influencers in the future time period the prediction refers to. To this aim, in the following experiments, the training set created in the *Training and Prediction* phase does not include all the data produced by the *Data transformation* phase, but it considers only the older of them. As a matter of fact, the most recent observations are used as a test set in the *Evaluation* phase to compute the real influencers. In particular, let $\{C_u^{t_0}, C_u^{t_1}, \ldots, C_u^{t_s}\}$ be the set of observations included in the training set, for each member u of the group. Consequently, t_s is the period of time of the last observation in the training set, and the corresponding test set consists of the observations $\{C_u^{t_{s+1}}, C_u^{t_{s+2}}, \ldots, C_u^{t_{s+z}}\}$, which are subsequent to t_s. Each observation $C_u^{t_{s+j}}$ related to a member u (for $j = 1, \ldots, z$) refers to a period of time $[t_{s+j}, t_{s+j} + \Delta)$ having duration equals to Δ (i.e., the granularity of the temporal network). The size of the test set z determines the amplitude of the prediction, i.e., the future time interval to which the prediction refers. As done in the `Prediction` task, the influence score $I_{t_{s+j}}(u)$ is computed for each member u in the test set and for each time interval t_{s+j} with $j = 1, \ldots, z$. The influence scores $\{I_{t_{s+1}}(u), I_{t_{s+2}}(u), \ldots, I_{t_{s+z}}(u)\}$ resulting from the observations are used to identify the top-k real influential members of the test set. Finally, the `Score` Task of the *Evaluation* phase assesses the accuracy of the proposed framework by comparing the top-k predicted influential members of the training set and the top-k real influential members of the test set in order to evaluate the number of members correctly identified by the framework. Table 8 summarizes the configuration and input parameters defined in each phase of the framework, as well as the values of the parameters used during the experiments.

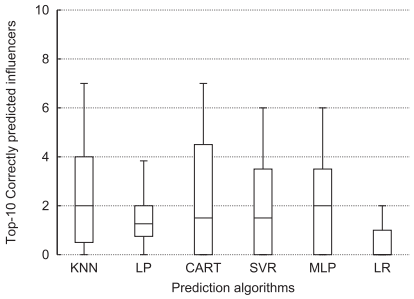
8.1 Comparison of Prediction Algorithms

The first set of experiments we performed are meant to identify the best prediction algorithm for the `Prediction` task. To this aim, we performed a model selection analysis by evaluating the accuracy obtained by using each of the algorithms described in Section 7.2, namely, LP, LR, KNN, CART, SVR, and MLP.

In particular, we run our framework with these prediction algorithms, and for each of them, we compared the first Top-10 predicted influential users with the real ones. Each prediction algorithm learns from the same training data, which consist of the 80% the data of each group, while the remaining 20% are used for testing. Furthermore, in this experiment, we fixed the granularity of the temporal network (i.e., the duration Δ of the slide) to 1 week and the number of principal components elected by the *Data transformation* phase, N', to 1, while in the next section we investigate in more detail the impact of the duration of the slice and of the number of principal components.

However, some prediction algorithms have their own specific input parameters. For instance, the KNN takes as input parameters the size of the neighborhood and the function used to measure the distance. Another important parameter is the kernel function used by the SVR algorithm, which determines how input data are transformed into a higher dimensional space. Hence, for each of the adopted algorithms, we found the input parameter combination resulting in the best accuracy using a grid search method over the training data. The boxplot at the top of Figure 11 shows the distribution of accuracy resulting from the execution of each prediction algorithm to predict the top-10 influencers from all the groups, while the table at the bottom summarizes the results by showing the lower quartile (Q1), the median, the upper quartile (Q3), and the standard deviation.

We can clearly see that the best accuracy has been obtained by using the KNN algorithm (with at most four similar neighbors and by using the Manhattan distance). As a matter of fact, the KNN algorithm allowed us to find at least two real influential members for 50% of the groups, and at least four influential members for 25% of the groups. Moreover, for some groups of the Entertainment category, we found seven influential members. The MLP algorithm as well obtained similar results by allowing us to find at least two real influential members for 50% of the groups, but the first



Prediction Algorithms	Top-10 Correctly Predicted Influencers			
	Q1	Median	Q3	Std.Dev
KNN	0.75	2	4	2.08
LP	1	1.25	2	1.22
CART	0	1.5	4.25	2.45
SVR	0	1.5	3.25	1.8
MLP	0	2	3.25	1.98
LR	0	0	1	0.70

Fig. 11. Analysis of the prediction algorithms: the figure at the top shows the distribution of the accuracy resulting from the execution of each prediction algorithm, while the table at the bottom shows the lower quartile (Q1), the median, the upper quartile (Q3), and the standard deviation.

and third quartiles are slightly lower w.r.t. the KNN algorithm. The network topology that has the best performance consists of only one hidden layer, one input layer, and one output layer. The performances achieved by CART and SVR are very similar, and both algorithms allowed us to find 1.5 real influential members out of 10 for half of the groups. The CART algorithm exhibits a high variation in the number of correctly predicted influencers, and the third quartile is greater than 4. Instead, the SVR algorithm achieves excellent performance with the radial basis function kernel, and the distribution of the results exhibits lower variability with respect CART. The median number of influencers correctly predicted by the LP algorithm is equal to 1.26 and the distribution of the results has a low deviation. In addition, the outliers in the plot indicate that, for some groups, the LP achieves the same results as obtained by the previous algorithms. The LR algorithm has the worst performance because the number of top-10 correctly predicted influencers ranges between 0 and 2.

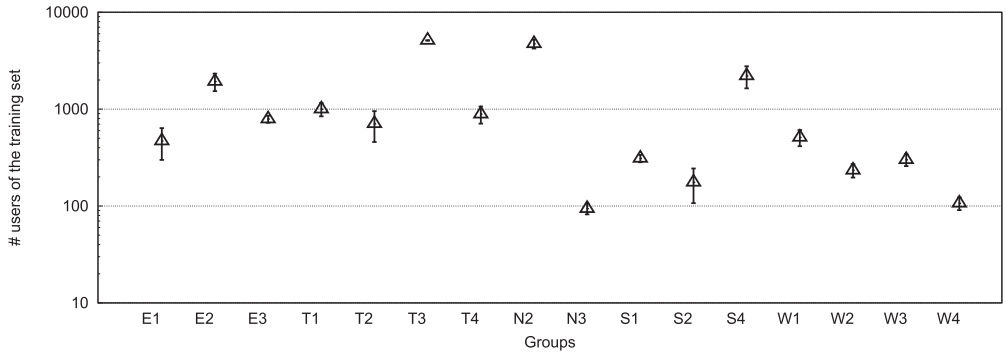
Hence, the KNN results to be one of the most accurate algorithms for Influencers Prediction among all the most popular prediction algorithms used in this experiment. Consequently, the experiments presented in the following were executed using the KNN algorithm.

8.2 Granularity of the Temporal Network

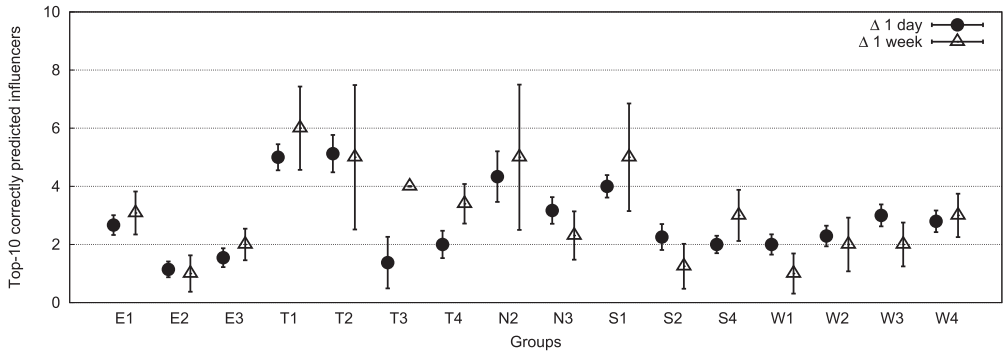
In this section, we investigate the effect of the duration of the slice Δ of the temporal network on the accuracy of the predictor. For this purpose, we compared the predictions obtained with $\Delta = 1$ day and $\Delta = 1$ week.

For the training of the predictor, both experiments use a training set that, for each group's member u, considers the interactions of u occurred in the month preceding the period to predict (i.e., 4 weeks or 28 days). In particular, in the case of $\Delta = 1$ day, such a training set consists of the observations related to the 28 interaction graphs of the previous month, one for each slice Δ. Instead, in the case of $\Delta = 1$ week, the training set contains the observations obtained from the four interaction graphs, one for each week of the previous month. From this point onward, for each group, we perform several experiments choosing distinct blocks of data within the dataset we collected to be used as a training set, and we compute the average number of correctly predicted influencers in such experiments. In order to reduce the number of dimensions of the data, the metrics obtained from the observations are transformed and projected on the first principal component. As a result, for each group's member, the transformed training set consists of one principal component for each observation.

At first, we evaluate for each group the number of members belonging to the training set, i.e., the members who interacted at least once in the reference period, since the most influential users



(a) Average number of potential influencers per group



(b) Top-10 influential members

Fig. 12. Accuracy of the prediction: number of the most influential members correctly predicted.

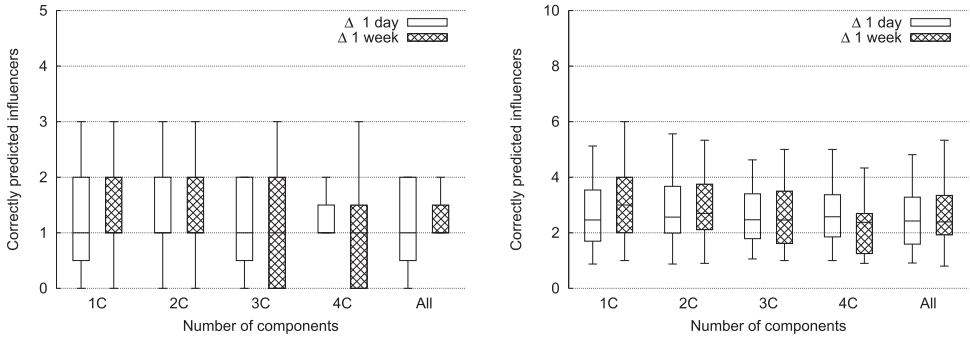
of a group will be chosen among them. Figure 12(a) shows for each group the average number of members in the training set.

Instead, Figure 12(b) depicts, for $\Delta = 1$ day and $\Delta = 1$ week, the 10 most influential members, which have been correctly predicted on average. The test sets adopted for checking the correctness of the predictions include one observation over a period of duration Δ. As shown by Figure 12(a), most groups (about 70%) have a number of possible influential members between 100 and 1000, while the other groups exhibit a larger number of potential influential members (between 1,000 and 10,000).

As expected, the duration of the slide Δ of the temporal network affects the accuracy of the prediction (see Figure 12(b)). Most groups (about 60%) exhibit better results when the slide Δ of the temporal network used for data modeling is equal to 1 week. In this case, the framework successfully identifies, on average, about 3 of the most 10 influential members of a group in a future time interval. As shown by Figure 12(b), some groups of the Entertainment (T1, T2), News (N2), and Sport (S1) categories exhibit high accuracy because the average number of influential members correctly predicted by the framework w.r.t. the top-10 real influencers for T1 is 6, while for S1, N2, and T2, it is 5.

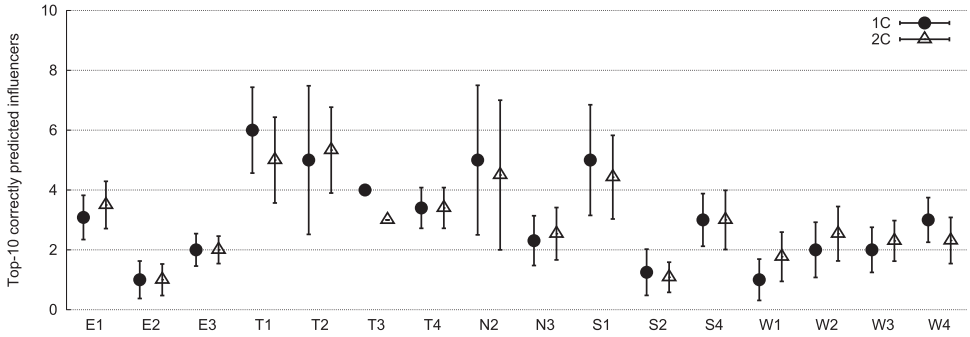
8.3 Number of Principal Components

In the previous experiments, we fixed the number of principal components to 1 for each observation. In this section, we investigate how the number of principal components used for data



(a) Top-5 influential members

(b) Top-10 influential members



(c) Top-10 predicted influential members by using temporal network with Δ =1 week

Fig. 13. Impact of the number of principal components on the accuracy of the prediction.

transformation affects the accuracy of the prediction. For this purpose, we analyze temporal networks having duration of the slide Δ equal to 1 week and 1 day. For the training of the predictor, the framework considers a training set that consists of the interactions occurred in the last month (i.e., 4 weeks or 28 days). The test sets adopted for checking the correctness of the predictions include one observation over a period of duration Δ. We executed several experiments where the number N' of principal components selected for the *Data transformation* phase varies among $\{1, 2, 3, 4, All\}$.

Figure 13 depicts the results of such experiments by showing the distribution of the average number of members correctly predicted w.r.t. the top-5 (Figure 13(a)) and the top-10 (Figure 13(b)) real influential members, for different numbers of components. From the numerical point of view, we observe that the best accuracy has been obtained using one component and a slide duration of $\Delta = 1$ week. However, in general, we could say that the accuracy achieved by using one or two components is quite similar, and it is in general better than the performance obtained using more components.

In particular, Figure 13(a) indicates that the number of influential members correctly predicted w.r.t. the top-5 real ones improves when the number of selected components ranges between 1 and 2. Indeed, for $\Delta = 1$ day, the maximum and the median numbers of correctly predicted influential members are, respectively, equal to 3 and 1 for both one and two components. For $\Delta = 1$ week, the best accuracy is obtained when the *Data transformation* phase selects either one component or two components. In particular, in this case, the maximum and the median numbers of correctly predicted influential members are equal, respectively, to 3 and 1 when we use one, two, three, and four components. However, the distribution of the results in the case of one or two components also

exhibits a high first quartile value, i.e., equal to 1, while on choosing three and four components, the value of the first quartile is 0.

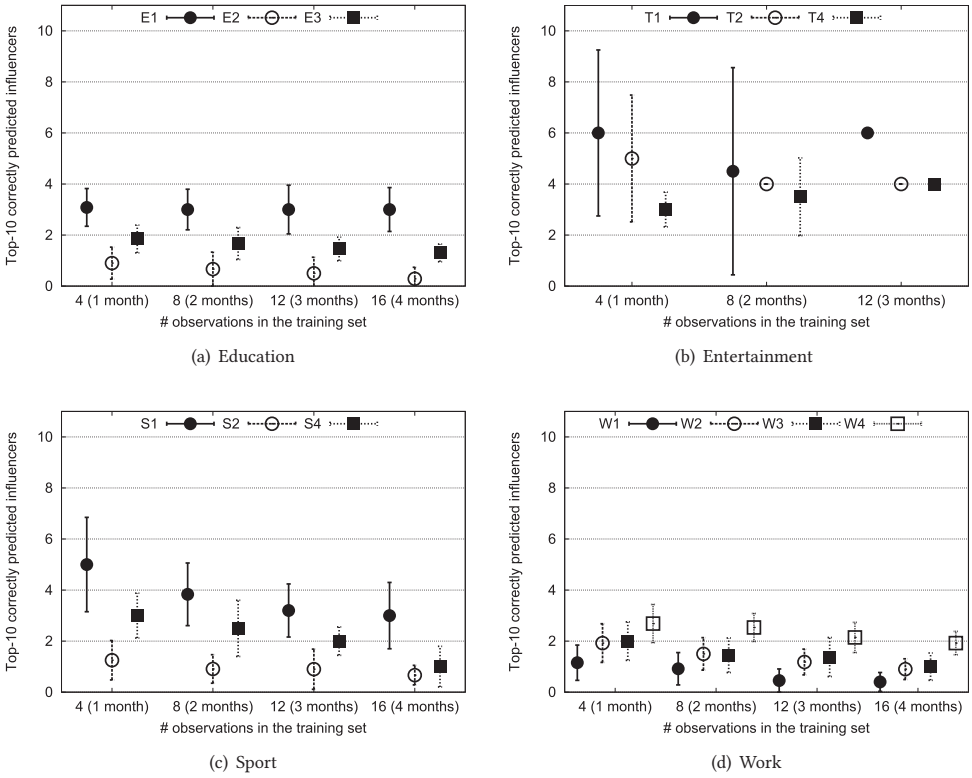
Figure 13(b) shows the accuracy of the framework in recognizing the top-10 real influential members for different numbers of components. In such a case, the median prediction results achieved for $\Delta = 1$ day do not change much as the number of components increases. Indeed, the median number of correctly predicted influencers ranges between 2.46 (for one component) and 2.57 (for two and four components). However, we could say that the best prediction results are achieved when selecting the first two components, as the highest value for the maximum number of correctly predicted influencers, equal to 5.56, is obtained with two components. Moreover, the highest values of the first and third quartiles are obtained with two components as well. The selection of one component improves the accuracy of the framework for $\Delta = 1$ week because the maximum and the median numbers of correctly predicted influencers are equal to 6 and 3, respectively. Instead, the maximum and median numbers of correctly predicted influencers obtained by using $\Delta = 1$ week and two components is equal to, respectively, 5.33 and 2.70, and these values become smaller using more components. We analyze in more detail the accuracy of the predictions on each group by showing in Figure 13(c) the average number of influential members successfully recognized among the top-10 real ones, when the slide Δ of the temporal networks used for data modeling is equal to 1 week and the number of selected components is equal to 1 and 2. For four groups (E2, E3, T4, S4), using one or two components led to the same accuracy in the results, i.e., the same number of predicted influencers turned out to be real ones. For half of the remaining 12 groups, the number of predicted influencers who turned out to be real ones was higher using one component, and for the other half, it was higher using two components. However, if we take into account the total number of real influencers who have been correctly predicted among all the groups, we see that using one component we were able to correctly predict 47 influencers, while using two components, the number of predicted influencers who turned out to be real ones was 46.

Summarizing, the framework achieves the best prediction accuracy when the first component is used and the slide Δ of the temporal networks used for data modeling is equal to 1 week. Increasing the number of components does not produce significant improvements in the number of predicted influencers, while making the training of the predictors more complex.

8.4 Size of the Training Set

As a further step, we analyze the impact of the size of training set on the performance of the framework. In particular, we vary the number of past observations used for the training of the `Prediction` task, and we evaluate the ability of the framework to identify the most influential members of the groups.

Figure 14 shows the average number of influencers (along with the 95% confidence interval (CI)) correctly predicted by the framework w.r.t. the top-10 real ones varying the number l of months covered by the training set in $\{1, 2, 3, 4\}$. The number of observations in l months depends on the duration of the slide Δ of the temporal network. Our experiments have been conducted setting Δ= 1 week, and the related numbers of observations are shown on the x-axis. Moreover, the *Data transformation* phase of the framework has been configured with N'=1 component for the prediction, while the test set consists of the real influential members in the week following the training set. The results show that the accuracy of the framework, in most of the cases, slightly decreases when the number of observations considered in the training set increases. Indeed, for the majority of the groups (i.e., groups E2, E3, T2, S1, S2, S4, W1, W2. W3, and W4), the average number of correctly predicted influencers is higher when the observations in the training set span over a



(a) Education (b) Entertainment

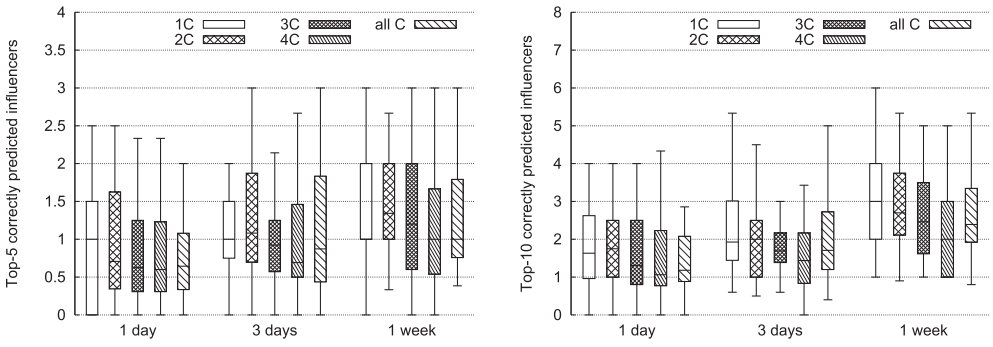
(c) Sport (d) Work

Fig. 14. Impact of the size of the training set on the accuracy of the prediction (Δ = 1 week).

month period. Furthermore, each group leads to a specific performance value of the predictor, depending on the category to which it belongs. In particular, for the Education category, using a training set of 1 month, the average number of correctly predicted influential members is equal to 0.9 for group E2 and to 1.9 for group E3. For group T2, belonging to the Entertainment category, the average number of correctly predicted influential members using a training set of 1 month is equal to 5. The groups of the Sport category exhibit an average number of correctly predicted influential members equal to 5 for S1, 1.25 for S2, and 3 for S4. Finally, for the groups of the Work category, the average number of influential members correctly predicted using a training set of 1 month is equal to 1.15 for W1, 1.92 for W2, 2 for W3, and 2.70 for W4.

Instead, the average number of correctly predicted influential members for group T1 is equal to 6. This latter result is achieved with two training sets: 1 month and 3 months, respectively. Finally, group E1 (Education category) and group T4 (Entertainment category) exhibit a different trend because the average number of correctly predicted influential members for group E1 is about 3, independently of the size of the training set. Instead, the results obtained from group T4 show that the accuracy of the framework improves when the number of observations considered in the training set increases.

Summarizing, for most groups, increasing the number of observations considered for the training of the predictor introduces some degradation on the accuracy of the prediction. This suggests that, given our settings, the influential members of the groups are more likely to be identified by our framework when the observations of the training set are related to a period of 1 month.



(a) Top-5 influential members (b) Top-10 influential members

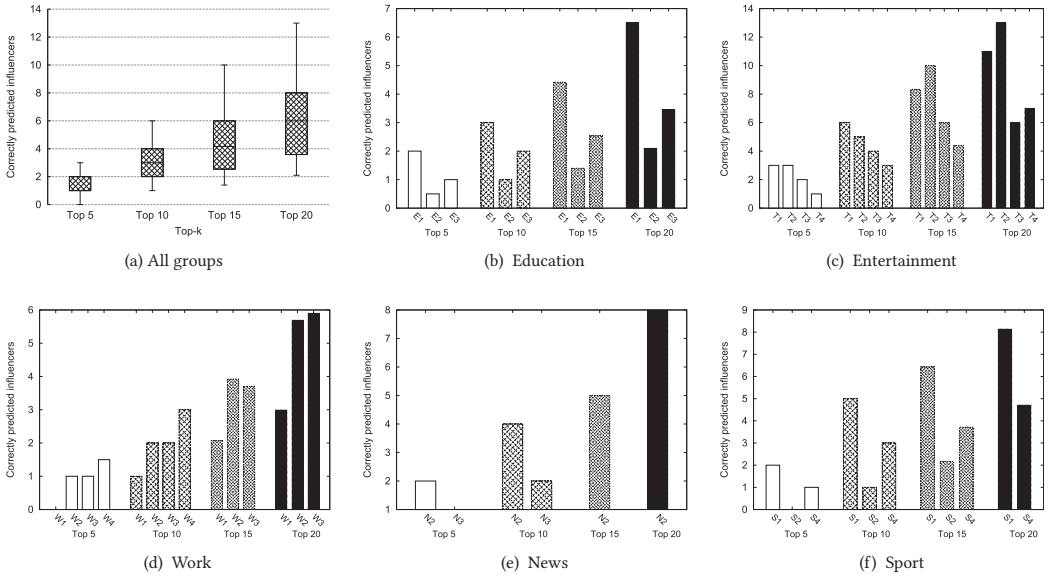
Fig. 15. Accuracy of the prediction by varying the future time interval (Δ = 1 week).

8.5 Amplitude of the Prediction

Another important aspect for the prediction of influential members is the accuracy achieved by the framework varying the size of the future time interval to predict. Hence, we use the framework to compute the top-k influential members of the groups, and we evaluate the accuracy of this prediction by using three different test sets consisting of the observations over future time intervals of, respectively, 1 day, 3 days, and 1 week. All the experiments use the same training set, which consists of the observations made in the last month, and the same granularity Δ = 1 week of the temporal network. Figure 15 shows the distribution of the average number of influential members of the groups correctly predicted by the framework by using different values for N', the number of components selected by the *Data transformation* phase, and different lengths of the test set. The graphs confirm the global trend, showing that predictions over a future period of 1 week are more effective, regardless of the number k of the top-k members selected by the framework. In fact, on one hand, Figure 15(b) shows that the number of correctly predicted influential members over the top-10 increases with the size of the test set used to evaluate the predictions in most of the cases. The highest median accuracy is achieved in the case of predictions for the next week and N'= 1 component. In this case, for half of the groups, the framework was able to identify at least three influencers, and for group T1, the framework correctly predicted six influencers. Instead, the lowest number of correctly predicted influential members is obtained by using all the components and a test set equal to 1 day. On the other hand, from Figure 15(a), we notice that the highest value for the median number of influential members correctly predicted over the top-5 is equal to 1.34, and it is obtained taking into account a future period of 1 week and selecting two components for the *Data transformation* phase, while by selecting only one component, the median value is equal to 1. Both in the case of N'= 1 component and in the case of N'= 2 components, the values of the first and third quartiles are the same, i.e., respectively, 1 and 2, while the maximum value is higher in the case of one component.

8.6 Number of Predicted Influencers

In general, we could say that the number k of the influential group's members to be predicted depends on the influencer marketing campaign strategy and budget. For instance, a company could decide to base its advertisement campaign on five potential influencers. Hence, in this section, we discuss the accuracy of the predictions of our framework having the number k of predicted influential members taking the values {5, 10, 15, 20}. To perform a fair evaluation of the prediction accuracy, it is important to notice that the number k of predicted influencers that will be compared



(a) All groups (b) Education (c) Entertainment

(d) Work (e) News (f) Sport

Fig. 16. Top-k influential members.

with real ones should be small enough with respect to the total number V of members of the group, (i.e., $k \ll V$). Otherwise, when k is close to V, the significance of the prediction could be overestimated. For this reason, for each value of k in $\{5, 10, 15, 20\}$, we determined for which group it is meaningful to compute the prediction accuracy by calculating the ratio between k and the number of potential influential members V of the group, and selecting only the groups for which that value is greater than a given threshold (set to 0.1). By using this methodology, we observe that group N3 and group W4 have about 100 members (see Figure 12(a)) and that the ratio between k and 100 is less than or equal to 0.1 when the value of k does not exceed 10. Hence, for N3 and W4, we will predict up to 10 influencers. Similarly, group S2 consists of about 180 potential influential members, and the maximum number k of predicted influential members we take into account for computing prediction accuracy is equal to 15. Instead, the other groups have a number of potential influential members between 234 (group W2) and 5,120 (group T3). Consequently, even for k equal to 20, the ratio between k and the number of potential influential members V of each group is below 0.1. Hence, we will predict up to 20 influencers. The experimental results are represented in Figure 16(a). The predictions have been conducted by using the input parameter values that resulted in the best prediction accuracy in our previous analysis: the granularity Δ of the temporal network is equal to 1 week (see Section 8.2), the number of principal components, N', is equal to 1 (see Section 8.3), the size of the training set is equal to 1 month (see Section 8.4), and the size of the future time interval to predict is equal to 1 week (see Section 8.5). As shown by the plot, the prediction of the top-k most influential users of the groups is more challenging when the number of members k to be recognized is small. In fact, we can see that our framework, for half of the groups we examined, was able to predict at least one influencer among the top-5, three influencers among the top-10, about four influencers among the top-15, and about six influencers among the top-20. We investigate in more detail the prediction accuracy by showing in Figure 16 the top-k most influential users of the groups for each category. The number of predicted users among the top-5 most influential members of the groups ranges between 0 (for a group of the Work category, a group of Sport, and a group of the News category) and 3 (for two groups of the

Entertainment category). Instead, for $k = 10$, the lowest number of correctly predicted influencers is 1 (for groups belonging to the categories Education, Work, and Sport), while the highest value is 6, and it was obtained for a group in the Entertainment category. In the case of top-15 most influential members, the methodology obtained the lowest number of correctly predicted users, 1.4, on a group of the Education category, while the highest value is 10 and it was obtained on a group of the Entertainment category. Finally, the number of correctly predicted members over the top-20 most influential members ranges between 2 for a group of the Education category and 13 for a group of the Entertainment category.

Finally, we computed the average number of correctly predicted influential members among all the groups, which resulted in 1.37 for the top-5, 2.94 for the top-10, 4.58 for the top-15, and 6.5 for the top-20.

9 COMPARISON WITH THE LITERATURE

In this section, we compare the approaches for Influencers Identification or Influencers Prediction described in Section 2 with the one we propose in this article. Table 9 summarizes the characteristics of existing approaches indicating the OSN considered by the approach (column *OSNs*); the type of measures used by the approach to compute influence (column *Measures*); the features considered by such measures (column *Features*); the ability of the approaches to predict future influential users (column *Prediction*); the type of dataset used to evaluate performance (column *Dataset*); and the type of approach used to divide the proposals into three categories: the Users Network Structure (UN) consists of the approaches taking into account users and their activities, the Content Flow (CF) category takes into account how the information flows in the network, and the Other Methods (OM) category takes into account other models, such as Association Rules and Features Analysis (column *Category*).

According to [2], the measures proposed in order to identify influential users in OSNs can be classified into six different classes. The first class includes works based on Local Measures (LM), which refers to local metrics calculated through the direct neighborhood of a given user. Examples of LM are in-degree and out-degree centralities. The second class is based on Short Path-Based Measures (SPM), which consider all the shortest paths that go through a user in a network to calculate the user's influence. Examples are the Betweenness Centraility and the Closeness Centrality. The third class considers Iterative Calculation-Based Measures (IC) that compute the score by considering not only the contribution of a network user, but also his direct connections, which means that each user contributes his ranking value to his connected users—similar to what happens with page ranking algorithms. Another class includes the Coreness-Based Measures (CM), which consider the location of users more important than their direct connection. Moreover, the fifth class includes ML approaches focused on features (such as the number of reactions or the length of the text) used by ML algorithms (e.g., Support Vector Machine and Decision Tree) to predict influential users. Finally, the categorization proposes a last class to classify the Hybrid Methods (HM), which are based on more than one technique.

In the following, we compare our approach with the ones listed in Table 9, taking into account three aspects: the elements taken into account for influencers identification and influencers prediction, the methodology, and the prediction.

9.1 Taxonomy Elements for OSN Influencers Identification and Influencers Prediction

As shown in Table 9, most of the current proposals are focused on Twitter [7, 14, 17, 48, 51, 52, 57–59]. Even if the main goal of all of these approaches is to find influencers, they differ from each other for the information they analyze to identify them. For instance, in [48], the authors focused on the tweets containing the *http* string, and analyze their propagation over time on Twitter. Their

Table 9. Summary of the Approaches Used to Identify or Predict Influential Users in OSNs

Reference	OSNs	Measures	Features	Prediction	Dataset	Category
[48]	Twitter	Influence, Passivity (LM)	retweet of tweets, including URLs	×	Real	UN
[7]	Twitter	Influence (LM)	tweet with bit.ly URLs	✓	Real	CF
[57]	Twitter	Influence (HM)	retweet from blogs associated with health subject	×	Real	UN
[59]	Twitter	Importance (IC)	Japanese tweet and retweet	×	Real	UN
[17]	Twitter	Influence (LM)	follower, tweet, retweet, mentions	×	Real	UN
[51]	Twitter	Influence (HM)	tweet filtered by tags, retweet, mention	×	Real	CF
[14]	Twitter	Influence (CM)	follower	×	Real	UN
[60]	Weibo	Influence (ML)	follower, micro-blog	✓	Real	CF
[27]	FriendFeed	Influence (IC)	follower's comments, likes, and tweet	×	Real	CF
[41]	Last.fm	Influence (ML)	user listening, listening time	×	Real	CF
[23]	Facebook (Pages)	Influence (HM)	posts, likes, comments	✓	Real	OM
[42]	Facebook (Game App)	Influence (ML)	age, gender, number of shares and responders in each game category	×	Real	OM
[4]	Facebook (Movie App)	Influence, Susceptibility (HM)	notifications' messages, age, gender, relationship status	×	Real	CF
[44]	All	Influence, Susceptibility (HM)	interaction	×	Synthetic	UN
[39]	Facebook (Pages)	Influence (HM)	similarity, interaction, proximity	×	Real	UN
[45]	Multiplatform	Influence (LM)	3,550 features related to the platform, user, interaction	×	Real	OM
[63]	Multiplatform	Influence (ML)	users' relationships, activities, relationships between users and objects	×	Real	OM
[9]	Micro-blog	Influence. (IC)	followerships, retweets, mentions, from micro-blogs filtered by topics	×	Real	UN
[58]	Twitter	Influence. (IC)	tweets, followers, and friends, from top-1000 Singapore-based twitters from twitterholic.com	×	Real	UN
[64]	Micro-blog	Influence (HM)	user interactions, retweet intervals comments	×	Real	CF
[31]	BlogCatalog	Influence (HM)	relationships	×	Real	UN
[52]	Twitter	Influence (ML)	retweets and mentions around a Pepsi discussion	✓	Real	UN

influence measure takes into account both structural properties of the network and the behavior of users in retweeting such links. The authors in [7] focus on tweets, including bit.ly URLs. In particular, they compute the influence of such posts tracking their diffusion and predict influential users by taking into account users' attributes and the influence of their past posts. In [59], the principal analyzed information is the tweet and the proposed approach considers the user–tweet graph, where nodes correspond to user accounts and tweets, and edges correspond to follow and retweet relationships.

Other proposals based on relationships are [14, 17, 58], which differ from our scenario because relationships do not exist in OSNs' groups. Indeed, the Twitter dataset used in [17] consists of following links among users. Instead, [14] measures the user influence by analyzing the user connectivity network through a customized version of the *k*-shell decomposition algorithm. It analyzes two Twitter datasets. The first one includes user profiles and social relations, and the second one a sampling of tweets representing more than 7 million users. In [58], a set of Twitter data about Singapore-based twitterers is analyzed, starting by a set of top-1000 Singapore-based twitterers from twitterholic.com. The influence measure they define takes into account the reciprocity of the following relation. In [57], the authors crawled with about 100 users in accordance with the link "how to follow" and later "browse interests", and then they searched for the topic "health" during March 2011 by extracting 200 retweets per user.

Works based on hashtags or filtered information are [51, 52, 58]. In [51], the authors extracted the tweets filtering them by following a specific set of hashtags. From this dataset, the authors constructed the network structure using the mention or retweet activities between users using the past 10 months of twitter data. Examining the tweets having hashtags related to the same topic can be considered somehow similar to examining the posts of a social group, although this way needs to find a relevant set of hashtags for the topic to be analyzed; instead, a group collects a relevant number of interactions concerning the topic it is focused on. In [52], the dataset is based on a discussion around Pepsi, in particular retweets, and it was extracted in two different periods, from November 11, 2009 to November 26, 2009. Instead, the authors in [58] use an ML technique for grouping tweets concerning the same topic.

A smaller number of proposals are focused on Facebook [4, 23, 39, 42], but none of them takes into account Facebook groups. Indeed, the information collected by these proposals is taken from Facebook applications and Facebook pages. For instance, in [23], the dataset is composed of a sample of 195 Facebook pages containing posts with all likes and comments, while the authors in [39] worked with a brand page containing information about a particular company, or brand, or product, taking into account the activities on the page carried out by the administrator and/or subscribers of the page. Instead, authors in [42] focused on the Facebook application, and they considered information that was related to the specific app. A similar approach is proposed in [4], where the authors study the effects of the users' attributes on the product adoption decisions by using influence-mediating messages sent from a commercial Facebook application.

Finally, there are other works based on different Social Media, such as Weibo, FriendFeed, and Last.fm. Weibo is principally used in China, and it is a mix between Twitter and Facebook. For this reason, it has been used in [60] to study following relationships and retweets. FriendFeed was closed in 2015, and Last.fm is active and useful in evaluating recommendation systems, as shown in [41], where a set of scrobbles are used to recommend musical similarity.

In [31], the authors use BlogCatalog, which is one of the largest blog directories available on the Internet. BlogCatalog allows you to search blogs, connect with bloggers, learn more about blogging and tips to promote blogs for oneself, and it is considered a social network of bloggers.

In summary, the literature presents several approaches based on various social media, and is tailored to analyze specific pieces of information of social media. In the following, we highlight the differences between such approaches and our proposal. First of all, we notice that our proposal is based on the function of a social groups. Indeed, social groups are used to share information and contents concerning specific interests. Information spread inside a group is visible to each member of the group, and it can be considered useful for the members of the group, but probably not so much useful outside the group. Hence, most of the works taking into account Twitter differ from our proposal because they consider influencers that are not related to a specific group (that is, a community of people interested in a given topic), but tweets are addressed to everybody or to all the followers of the posting user. Furthermore, taking into account only tweets, including specific hashtags [51, 52], differs from taking into account posts on social groups because hashtags are typically related to very specific topics, and are dynamic and short-lived. To collect a relevant number of tweets related to a given topic, the related hashtags should be derived earlier, using a similarity measure able to recognize hashtags belonging to that topic—for instance, the approach of [58] uses an ML technique for identifying tweets concerning the same topic. Only few approaches are based on Facebook, but they consider a different scenario, mainly Facebook applications [4, 42] (data were collected when the old Facebook privacy policies still allowed applications to access information concerning users) or Facebook pages of companies or brands [23, 39], and no one is focused, unlike our proposal, on Facebook groups.

9.2 Methodology

Concerning the methodology, we highlighted in Table 9 (column *Category*) that the approaches presented in the literature can be categorized into three main categories: approaches based on UN, approaches based on CF, and OM. Our approach falls into the first category, because it takes into account the activity of members inside groups, and it does not consider the content flow due to the possible lack of interest in sharing the information shared within a group with outside the group itself. As a matter of fact, a specific content is created in a group because it is typically related to the topic of the group and hence relevant for the group members only. Only a subset of the approaches presented in Table 9 ([9, 14, 17, 31, 39, 44, 48, 52, 57–59]) falls under the same category as we propose in this article. In [9], the authors propose a multigraph approach to evaluate three different categories of users: Influencers, Leaders, and Discussers. They propose three different link analysis algorithms, one for each specific role to identify. In [59], the TuRank algorithm is proposed in order to evaluate influential node. The algorithm considers both a Twitter social graph and how tweets actually flow among users. The approach presented in [57] is based on combining a set of standard metrics, i.e., Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, and Page Rank, with adjustable weighted parameters, considering not only the topological importance of a node, but also the strength of ties of retweets. Instead, [14] proposes to measure the user influence by analyzing the user connectivity network through a customized version of the k-shell decomposition algorithm. In this approach, the most influential users are the ones at the core of the network built by the k-shell decomposition algorithm. In [17], the proposed model is designed to consider only three specific influence measures: in-degree, retweet, and mention influence. The authors in [48] defined an influence measure that takes into account both structural properties of the network and the link diffusion behavior of users. In particular, they define the user passivity to measure the percentage of tweets that are retweeted by each user, and, to measure the influence of a user, they take into account both the number of retweets of tweets and the passivity of the users who retweeted them. In [52], the authors produce the Retweet and the Mention Graph, and the users who will have viral outbursts of retweets in the following week are computed by taking into account the in-degree, out-degree, and Page Rank (with a damping factor of 0.85). In [44], the

authors studied influence as a function of link strength and incoming and outgoing clustering value defined for each node in the network. The methodology measures the link strength according to the volume of interactions among users, while the clustering value is measured by the closeness of a node to highly interconnected communities. In [39], the methodology considers an association value composed of three metrics: proximity, similarity, and interaction. In [58], the authors provide a methodology where the initial dataset is distilled in order to identify topics in the text. Then, a topic-specific algorithm is applied. A proposal that shows some limited similarity to our work is presented in [31], where authors aggregate different centrality metrics to compute an influence score as the average of the positions of a node in decreasing order of centrality scores over various centrality metrics. They selected six metrics to evaluate the methodology. However, the cited work does not provide details about the possibility to use other combinations of measures, and it seems to be correct only when the measures are correlated between them. Furthermore, it uses averaged values computed on the selected metrics.

With respect to the previously cited approaches, our proposal is novel, defining a complete and general framework that can be applied to particular scenarios in order to instantiate a personalized influencer prediction process by choosing a customizable set of centrality measures and a prediction model. Furthermore, the practical methodology that we have proposed defines all the phases required for the Influencers Prediction process, starting from the data collection from OSNs. The methodology also includes, as an optional step, an evaluation phase, to be performed at the end of the Influencers Prediction process, which allows its user to evaluate the accuracy of the prediction obtained with the specific configuration they chose, in order to fine-tune this configuration (e.g., changing the set of centrality measures, the number of components selected by the PCA, the prediction algorithm, or other parameters of the methodology) to improve the prediction accuracy in their specific domain.

9.3 Influencers Prediction

The prediction of future influencers has been considered only by the approaches described in [7, 23, 52, 60], as shown in Table 9. In [23], association rules are applied in order to assess user participation in a post, based on previous interactions. Consequently, the approach relates to how users perform actions (e.g., comments or likes) on posts in Facebook pages, being able to predict if a specific user will or will not participate on a post discussion. This is different from our work, in which we predict the influential members of a group by using the interactions as well, but without focusing on predicting the participation of specific users to post discussions, which in itself is a different problem.

In [60], the authors propose a novel notion of social influence for modeling the retweet behaviors in the context of the ego network of a user. In particular, they predict for a specific user v whether his/her active neighbors have an influence on retweet behaviors of v by analyzing the influence locality-based features, personal attributes, topic propensity, and instantaneity. The six personal attributes taken into account are the number of followees, the number of followers, the number of bi-followers (i.e., the reciprocal of "following" relationships), the longevity (age of the account), gender, and verification status. The principal difference with our work is that since the prediction takes into account the ego network of a specific user, such a prediction is valid only for that specific user. Instead, in our approach, a prediction concerns all the users of the group taken into account.

In [7], the authors consider the user's ability to seed the content containing URLs that generate large cascades of retweets. To compute the influence score for a given post, the approach tracks the diffusion of the URL from its original post until the diffusion stops. The prediction uses a regression tree model taking into account both users' features and the past influence of their posts, also using folded cross-validation to terminate partitioning to prevent over-fitting. The main difference

from the approach we propose, besides the scenario, is that this approach defines a customized prediction process tailored for the specific context, while our approach allows the choice of any prediction algorithm, also allowing the evaluation of the resulting accuracy.

The authors in [52] compare different measures of influence on Twitter, such as Degree Centrality and Page Rank, on their ability to accurately predict which tweet will be virally retweeted in the near future, and also combine them for obtaining better predictions. To this aim, they introduce a supervised version of Kemeny Ranking [32] in order to aggregate individual rankings obtained evaluating such measures for the task of predicting influencers. In this case too, the main difference from our approach, besides the scenario, is that this approach is focused on defining a customized prediction process maximizing the prediction accuracy in the specific scenario of interest, while our approach allows the choice of any prediction algorithm.

9.4 Discussion

In order to highlight the main differences between our work and the state-of-the-art solutions, we analyzed three distinct aspects: the taxonomy elements for OSN Influencers Prediction, the methodology, and the `Prediction` task. To summarize, there is no other previous work that takes into account the prediction of influencers in Facebook groups. As a matter of fact, the majority of the existing approaches are based on Twitter, where the concept of group does not exist, and the only one that takes into account groups of tweets related to the same topic (selected through an ML technique) does not perform Influencers Prediction. The few approaches based on Facebook consider information gathered by Facebook applications (collected in the past, when the old Facebook privacy policies allowed applications to access a wealth of information concerning users) or from Facebook pages, which are related to companies or products. Moreover, most of the existing approaches are meant to identify current influencers, and only four of them also deal with the prediction of future ones. Only one of these four is related with Facebook pages, but it performs a very different kind of prediction with respect to our proposal. In fact, it predicts whether a given user will or will not participate in the discussion on a given post.

Finally, our contribution is very different from the existing ones because it defines a general and complete framework focused on Influencers Prediction in Facebook groups, which covers all the phases of the Influencers Prediction process, starting from the collection of data from the OSN and ending with the evaluation of the accuracy of the prediction. For each phase of the process, we describe the input data and the results, and a set of algorithms and tools that could be adopted to perform the specific elaboration on such input data to obtain the results. Hence, our framework is not fixed to a specific set of measures or to a specific prediction algorithm, but it allows its users to choose a customized set of measures and a prediction algorithm. The framework also identifies a number of other parameters of each phase of the Influencers Prediction process, such as the number of component selected by the PCA when combining the chosen measures (see Table 8), which can be properly configured to obtain the best prediction accuracy in each specific scenario. As a matter of fact, our framework also comprises an evaluation phase aimed at evaluating the prediction accuracy obtained with a specific parameter configuration, in order to fine-tune these parameters to improve the prediction accuracy. The framework has been instantiated with a practical and viable methodology, thoroughly discussed. The results achieved by the proposed methodology, reported and discussed in previous sections, show the quality and viability of our proposal.

9.5 Experimental Comparison

The previous qualitative comparison revealed that our proposal is very different from the state-of-the-art solutions, which do not take into account the problem of Influencers Prediction in Facebook

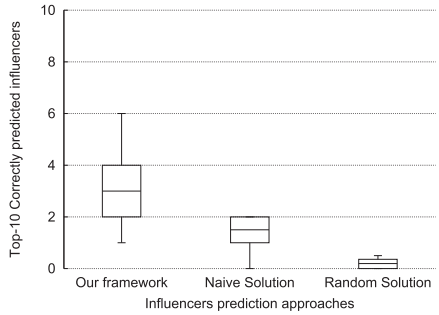


Fig. 17. Experimental comparison with other Influencers Prediction approaches.

groups. For this reason, a direct experimental comparison between our approach and the those listened in Table 9 would not be meaningful in this scenario. In order to show that the framework implementing our methodology works effectively, we selected and implemented two reference solutions to the Influencers Prediction problem as baselines for comparison: the random solution and the naive solution. The naive solution computes, for each group, the number of comments or replies received by each member in the training set and predicts as most influential members the members who attracted the highest number of interactions. Instead, the random solution generates for each member of the group a random number with equal probability in a given interval and takes as most influential members the ones who obtained the highest random numbers, regardless of the interactions that occurred in the groups. For the naive and random solutions, the number of interactions received by each member of the group in the test set is used to identify the top-10 real influential members, and the accuracy of each solution is assessed by comparing the latter with the predicted ones. We execute the above experiment for all the groups in our dataset and summarize in Figure 17 the distribution of results we obtained for the three solutions. In particular, each solution uses the same training data, which consist of the 80% the data of each group, while the remaining 20% are used for testing. Furthermore, in this experiment, the values of the configuration parameters used to execute our framework are the ones we determined in Section 8 (see Figure 16(a)), i.e., we fixed the granularity Δ of the temporal network to 1 week, the number of principal components to 1, and the size of the neighborhood used by the KNN to 4. Compared to these alternatives solutions, the proposed approach achieves remarkably better performance, showing the effectiveness of our methodology.

10 CONCLUSION AND FUTURE WORK

Motivated by the increasing attention paid to the role of influential users in OSNs, we are the first ones, to the best of our knowledge, to propose a novel, general framework for predicting the most influential members of Facebook groups. Furthermore, following the described framework, we implemented a practical and viable methodology that has been tested over a robust experimental campaign. Such a campaign is based on the collection of interactions from among about 800,000 users from 18 Facebook's groups belonging to different categories (Education, Politics, News, Entertainment, and Sport). Achieved results are striking: By properly tuning the system parameters, when predicting 10 influencers, the median number of them who turned out to be real ones was 3—taking into account the groups of all the categories–, while it was 5 when taking into account specific groups. Moreover, we have also gained interesting insight into this new domain of investigation. In particular, we found out that, for the Influencers Prediction process, the defined KNN prediction algorithm provided the highest accuracy results among the most commonly used

(similar) algorithms. Furthermore, we found that the best accuracy of the `Prediction` task was obtained when the dimension of the vector representing the results provided by the set of centrality measures taken into account was reduced to 1 through the PCA method. Finally, for what concerns the size of the training set, the best accuracy was obtained using the interactions collected in 1 month and considering a granularity of 1 week for the temporal network representing them.

As future work, we plan to enhance the proposed framework by investigating the effect of more sophisticated prediction models and other data dimensionality reduction techniques on the accuracy of the prediction. Another interesting research direction would be to study cross-group correlation phenomena. Finally, since the proposed framework has been designed to be modular and general, it provides good applicability and portability to other scenarios as well. Future work will concern the application of the proposed framework to other relevant scenarios.

Overall, the complexity of the domain, the novelty of the tackled problem, and the quality and viability of achieved results, other than being interesting on their own, also pave the way for further research in the highlighted directions.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers who, with their comments, helped to improve the quality of the manuscript.The findings achieved herein are solely the responsibility of the authors.

REFERENCES

[1] D. Aha and D. Kibler. 1991. Instance-based learning algorithms. *Machine Learning* 6 (1991), 37–66.

[2] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, Sri Devi Ravana, Ejaz Ahmed, Ghulam Mujtaba, Muhammad Usman Shahid Khan, and Samee U. Khan. 2018. Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Computing Surveys* 51, 1 (2018), 16.

[3] Chainarong Amornbunchornvej, Ivan Brugere, Ariana Strandburg-Peshkin, Damien R. Farine, Margaret C. Crofoot, and Tanya Y. Berger-Wolf. 2018. Coordination event detection and initiator identification in time series data. *ACM Transactions on Knowledge Discovery from Data* 12, 5 (2018), 53.

[4] Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341.

[5] Joel Backaler. 2018. Business to consumer (B2C) influencer marketing landscape. In *Digital Influence*. Palgrave Macmillan, Cham, 55–68.

[6] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. Social influence in social advertising: Evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 146–161. DOI: https://doi.org/10.1145/2229012.2229027

[7] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, 65–74.

[8] Prithwish Basu, Amotz Bar-Noy, Ram Ramanathan, and Matthew P. Johnson. 2010. Modeling and analysis of time-varying graphs. *CoRR abs/1012.0260*. Retrieved from https://arxiv.org/abs/1012.0260.

[9] Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. 2012. Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks. In *Proceedings of the International Symposium on String Processing and Information Retrieval*. Liliana Calderón-Benavides, Cristina González-Caro, Edgar Chávez, and Nivio Ziviani (Eds.). Springer, Berlin, 111–117.

[10] Ranran Bian, Yun Sing Koh, Gillian Dobbie, and Anna Divoli. 2019. Identifying top-k nodes in social networks: A survey. *ACM Computing Surveys* 52, 1 (2019), 1–33.

[11] Paolo Boldi and Sebastiano Vigna. 2014. Axioms for centrality. *Internet Mathematics* 10, 3–4 (2014), 222–262.

[12] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. CRC press.

[13] Duncan Brown and Nick Hayes. 2008. *Influencer Marketing—Who Really Influences Your Customers?* Butterworth-Heinemann.

[14] Phil E. Brown and Junlan Feng. 2011. Measuring user influence on Twitter using modified k-shell decomposition. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*.
[15] Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. 2012. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems* 27, 5 (2012), 387–408.
[16] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2011. Crawling Facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. ACM, 52.
[17] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. Vol. 10. Association for the Advancement of Artificial Intelligence, 10–17.
[18] Patrali Chatterjee. 2011. Drivers of new product recommending and referral behaviour on social network sites. *International Journal of Advertising* 30, 1 (2011), 77–101.
[19] Stefano Cresci, Roberto Di Pietro, and Maurizio Tesconi. 2019. Semantically-aware statistical metrics via weighting kernels. In *Proceedings of the 2019 IEEE International Conference on Data Science and Advanced Analytics*, Lisa Singh, Richard D. De Veaux, George Karypis, Francesco Bonchi, and Jennifer Hill (Eds.). IEEE, 51–60. DOI: https://doi.org/10.1109/DSAA.2019.00019
[20] Andrea De Salve, Marco Dondio, Barbara Guidi, and Laura Ricci. 2016. The impact of user's availability on on-line ego networks: A Facebook analysis. *Computer Communications* 73 (2016), 211–218.
[21] Andrea De Salve, Barbara Guidi, and Paolo Mori. 2018. Predicting the availability of users' devices in decentralized online social networks. *Concurrency and Computation: Practice and Experience* 30, 20 (2018), e4390.
[22] P. Alex Dow, Lada A. Adamic, and Adrien Friggeri. 2013. The anatomy of large Facebook cascades. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. Vol. 1. Association for the Advancement of Artificial Intelligence, 145–154.
[23] Fredrik Erlandsson, Piotr Bródka, Anton Borg, and Henric Johnson. 2016. Finding influential users in social media using association rule learning. *Entropy* 18, 5 (2016), 164.
[24] Afonso Ferreira. 2004. Building a reference combinatorial model for MANETs. *IEEE Network* 18, 5 (2004), 24–29.
[25] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2009. A walk in Facebook: Uniform sampling of users in online social networks. *CoRR arXiv:0906.0060*. Retrieved from https://arxiv.org/abs/0906.0060.
[26] Ke Gu, Linyu Wang, and Bo Yin. 2019. Social community detection and message propagation scheme based on personal willingness in social network. *Soft Computing* 23, 15 (2019), 6267–6285.
[27] Behnam Hajian and Tony White. 2011. Modelling influence in a social network: Metrics and evaluation. In *Proceedings of the 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust/2011 IEEE 3rd International Conference on Social Computing*. IEEE, 497–500.
[28] Carleen Hawn. 2009. Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care. *Health Affairs* 28, 2 (2009), 361–368.
[29] Petter Holme and Jari Saramäki. 2012. Temporal networks. *Physics Reports* 519, 3 (2012), 97–125.
[30] Ahmed Rageh Ismail. 2015. Leveraging the potential of word of mouth: The role of love, excitement and image of fashion brands. *Journal of Global Fashion Marketing* 6, 2 (2015), 87–102.
[31] Imrul Kayes, Xiaoning Qian, John Skvoretz, and Adriana Iamnitchi. 2012. How influential are you: Detecting influential bloggers in a blogging community. In *Proceedings of the International Conference on Social Informatics*. Springer, 29–42.
[32] John G. Kemeny. 1959. Mathematics without numbers. *Daedalus* 88, 4 (1959), 577–591.
[33] David Kempe, Jon Kleinberg, and Amit Kumar. 2002. Connectivity and inference problems for temporal networks. *Journal of Computer and System Sciences* 64, 4 (2002), 820–842.
[34] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 137–146.
[35] Vassilis Kostakos. 2009. Temporal graphs. *Physica A: Statistical Mechanics and Its Applications* 388, 6 (2009), 1007–1023.
[36] Peter Langfelder, Bin Zhang, and Steve Horvath. 2007. Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics* 24, 5 (2007), 719–720.
[37] Linyuan Lü, Tao Zhou, Qian-Ming Zhang, and H. Eugene Stanley. 2016. The H-index of a network node and its relation to degree and coreness. *Nature Communications* 7 (2016), Article 10168. DOI: https://doi.org/10.1038/ncomms10168
[38] Francesca Magno and Fabio Cassia. 2018. The impact of social media influencers in tourism. *Anatolia* 29, 2 (2018), 288–290.
[39] Adrija Majumdar, Debashis Saha, and Parthasarathi Dasgupta. 2015. An analytical method to identify social ambassadors for a mobile service provider's brand page on Facebook. In *Proceedings of the 2015 Applications and Innovations in Mobile Computing*. IEEE, 117–123.
[40] Vincenzo Nicosia, John Tang, Cecilia Mascolo, Mirco Musolesi, Giovanni Russo, and Vito Latora. 2013. Graph metrics for temporal networks. In *Temporal Networks*. P. Holme, J. Saramäki (Eds.). Springer, 15–40.

[41] Róbert Pálovics and András A. Benczúr. 2015. Temporal influence over the Last. fm social network. *Social Network Analysis and Mining* 5, 1 (2015), 4.
[42] Watcharapan Ponchai, Bunthit Watanapa, and Kaneungjit Suriyathumrongkul. 2015. Finding characteristics of influencer in social network using association rule mining. In *Proceedings of the 10th International Conference on e-Business*.
[43] S. Sharma, R. Rabade, and N. Mishra. 2014. Survey of influential user identification techniques in online social networks. In *Survey of Influential User Identification Techniques in Online Social Networks*. Vol. 235. S. Thampi, A. Abraham, S. Pal, and J. Rodriguez (Eds.). Springer, Cham, 359–370.
[44] Amir Afrasiabi Rad and Morad Benyoucef. 2011. Towards detecting influential users in social networks. In *Proceedings of the International Conference on E-Technologies*. Springer, 227–240.
[45] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor DSouza. 2015. Klout score: Measuring influence across multiple social networks. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2282–2289.
[46] Fabián Riquelme and Pablo González-Cantergiani. 2016. Measuring user influence on Twitter: A survey. *Information Processing & Management* 52, 5 (2016), 949–975.
[47] Thomas S. Robertson. 1967. The process of innovation and the diffusion of innovation. *The Journal of Marketing* 31, 1 (1967), 14–19.
[48] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 18–33.
[49] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy. 1999. Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, 11, 5 (1999), 1188–1193.
[50] Nemanja Spasojevic, Zhisheng Li, Adithya Rao, and Prantik Bhattacharyya. 2015. When-to-post on social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2127–2136.
[51] Karthik Subbian, Charu Aggarwal, and Jaideep Srivastava. 2016. Mining influencers using information flows in social streams. *ACM Transactions on Knowledge Discovery from Data* 10, 3 (2016), 1–28.
[52] K. Subbian and P. Melville. 2011. Supervised rank aggregation for predicting influencers in Twitter. In *Proceedings of the 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing*. IEEE, 661–665.
[53] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 807–816.
[54] Alexandru Topirceanu, Mihai Udrescu, and Radu Marculescu. 2018. Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution. *Scientific Reports* 8, 1 (2018), 10871.
[55] Thomas W. Valente. 1996. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory* 2, 2 (1996), 163–164.
[56] Brian E. Weeks, Alberto Ardèvol-Abreu, and Homero Gil de Zúñiga. 2017. Online influence? Social media use, opinion leadership, and political persuasion. *International Journal of Public Opinion Research* 29, 2 (2017), 214–239.
[57] Leila Weitzel, Paulo Quaresma, and José Palazzo M. de Oliveira. 2012. Measuring node importance on Twitter microblogging. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*. ACM, 11.
[58] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. ACM, 261–270.
[59] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. 2010. Turank: Twitter user ranking based on user-tweet graph analysis. In *Proceedings of the International Conference on Web Information Systems Engineering*. Springer, 240–253.
[60] Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2015. Who influenced you? Predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data* 9, 3 (2015), 25.
[61] Shiwen Zhang, Yaping Lin, Qin Liu, Junqiang Jiang, Bo Yin, and Kim-Kwang Raymond Choo. 2017. Secure hitch in location based social networks. *Computer Communications* 100 (2017), 65–77. DOI: https://doi.org/10.1016/j.comcom.2017.01.011
[62] J. Zhou, F. Liu, and H. Zhou. 2018. Understanding health food messages on Twitter for health literacy promotion. *Perspectives in Public Health* 138, 3 (2018), 173–179.
[63] Yang Zhou and Ling Liu. 2015. Social influence based clustering and optimization over heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data* 10, 1 (2015), 2.
[64] Kechen Zhuang, Haibo Shen, and Hong Zhang. 2017. User spread influence measurement in microblog. *Multimedia Tools and Applications* 76, 3 (2017), 3169–3185.

Received November 2019; revised September 2020; accepted December 2020