Personalized Influencer Detection: Topic and Exposure-Conformity Aware

Zekarias T. Kefato
University of Trento
Via Sommarive, 9 I-38123 Povo
Trento, Italy
Email: zekarias.kefato@unitn.it

Alberto Montresor
University of Trento
Via Sommarive, 9 I-38123 Povo
Trento, Italy
Email: alberto.montresor@unitn.it

ABSTRACT

Influence analysis has been the subject of intense investigation, particularly since the emergence of online social networks. While early studies have focused on identifying global influencers, recently the attention has been enlarged to consider influence from the most diverse perspectives; e.g., topic-based influence propagation has been extensively analyzed and several important results have been obtained. A perspective that has not received sufficient attention is the study of personalized influence analysis; personalized here means the capability of influencing a specific user. Guo et al. have identified this as an important problem, noting that other kinds of influencers (such as the global ones) are not necessarily good individual influencers. The goal of this paper is to fill this gap by proposing a mechanism to identify personalized influencers. Unlike prior studies that heavily rely on *topic similarities*, we incorporate interaction histories, referred as *exposure conformity*, in learning influence propagation probabilities. We propose a random-walk algorithm to tackle the problem. We have empirically validated the effectiveness of our algorithm by comparing it against several widely-used baseline techniques. Finally, our findings contribute additional insights in support of the need for personalized influence analysis. Our findings also show interesting properties with respect to influence depending on the position of the individuals we are personalizing on.

Keywords

Personalized Influence Analysis, Social Networks, Algorithms

1. INTRODUCTION

In online social networks (OSN) such as Twitter, Facebook, LinkedIn and Weibo, *influence analysis* is one of the most important problems, at the core of fundamental applications like viral marketing and recommendation. A growing number of studies have tackled this problem from different perspectives. Beginning from the early days of OSNs, many studies [15, 10, 6, 19, 1] have focused on detecting global influencers; that is, the most influential users in the whole network. Even though interesting results have been obtained, recent studies [20, 8, 3, 14, 4] have proposed alternative and complementary approaches centered around *topics*. Starting from the assumption that OSN users are neither interested nor specialized in all the topics of discussion, these works claim that it is more realistic and effective to identify *topic-based influencers*. Users consume information that is relevant to a limited number of topics, and their ability to win others' attention is stronger in their specialties rather than in some other topic for which they have a vague knowledge.

An important challenge in influence analysis has been the definition of *influence* by itself [2], which may have different meanings depending on the type of network. We consider that a user in an OSN has been influenced by other users if she has been *exposed* to their content and acted in response (*conform*). For example, a user on Twitter and Facebook may be exposed via her feed; and could respond (conform) in the form of replying, commenting, sharing, or re-tweeting. We assume such forms of reactions as an evidence that she might have been influenced; and this reaction is referred to as *exposure conformity* in this study.

Prior studies [15, 10, 6, 19, 20, 8, 3, 14, 4] are mostly focused on detecting influencers for the whole network or for a particular topic. Little emphasis have been given to the notion of personalized influencer detection. Nonetheless, it had been argued by Guo et al. [12] that detecting personalized influencers is a critical task, as global influencers and immediate neighbors are not necessarily the best influencers for a given user.

The approach of Guo et al. [12], while being the first to tackle the influencer detection problem, is not topic- and exposure conformity-aware. Moreover, their algorithm can neither be directly adapted to the topic-aware variant [16]. The contribution of this paper is to present a novel solution to the personalized influencer detection problem, where both topics and exposure conformity are taken into consideration.

The motivation for this work is given by two contradictory findings. On one side, Weng et al. [20] show some evidence for the presence of "homophily", i.e. the tendency of individuals to associate and bond with similar others. Cha et al. [7], on the other hand, show a small presence of homophily in a larger dataset. In an attempt to unify these two results, we combine *topic similarity* (based on the idea that people interested in similar topics are probably connected in the network), and *exposure conformity* (Assuming that explicit

.

interactions enables us to uncover informative and hidden user preferences from equally relevant information sources on a topic). Our *TOpic and exposure-conformity-aware Personalized Influencers Detection* (TopID) algorithm identifies influencers in two phases:

- In the first phase, the algorithm infers a topic profile for every user by utilizing the content that each user has generated. This profile models the degree of interest that each user has across T topics as a probability distribution. The profile is further used to compute *topic similarity* between every pair of connected users. Then it computes the so called *exposure conformity* based on interaction histories.
- In the second phase, first *exposure conformity* and *topic similarity* are combined together to estimate influence propagation probabilities between connected users; then, the influence score of every user upon a given user and a given topic is computed. Depending on this score, users will be ranked to identify the top-K influencers of a user on the specified topic.

As a first step towards validating the performance of our algorithm, we evaluated TopID in the topic-wide influencers detection task. Towards this end, we pick one of the well-known techniques, TwitterRank [20], and compare our method against it. The main difference between our method and the one of Weng et al.'s is exposure conformity; and we show that indeed accounting for exposure conformity leads to better performance. After establishing the performance improvement in the topic-wide setting, we show that our approach consistently gives improved results in the personalized influencers detection as well. This is in comparison with a set of widely used baseline techniques and the personalized version of TwitterRank [20].

The rest of the paper is organized as follows. Section 2 introduces the terminology and presents the problem statement. Section 3 discusses the proposed algorithm. Section 4 reports the experiments and results. Finally Section 5 discusses related works and we conclude in Section 6.

2. PROBLEM STATEMENT

In this section we define the main concepts discussed in the paper, i.e. the interaction network, the exposure conformity and the topic profiles, and we formally state the problem.

Definition 1 (Interaction Network). *An interaction network is a graph $G = (V, E)$ where the vertex set V represents users and an edge $(u, v) \in E$ means that the user u follows v.*

The set of followers of a user v is denoted by $\mathcal{F}_v^{\leftarrow} = \{u|(u, v) \in E\}$, while the set of users followed by v is denoted by $\mathcal{F}_v^{\rightarrow} = \{u|(v, u) \in E\}$.

We consider a set of topics $\Theta = \{1, \ldots, T\}$, where T is a parameter of our approach. Each user is associated to a *topic profile* defined as follows:

Definition 2 (Topic profile). *The topic profile function $\phi : V \in [0, 1]^T$ associates each user v with a T-tuple $\phi(v)$ showing the distribution of the degree of interest of user v with respect to each of the topic $t \in \Theta$, for both posting and consuming contents related to t. Being a distribution, $\sum_{t=1}^{T} \phi_t(v) = 1$, where $\phi_t(v)$ denotes the t-th entry of $\phi(v)$.*

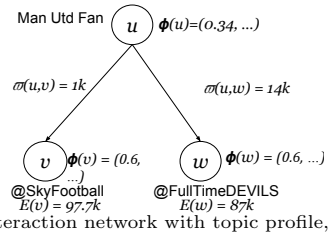

Figure 1: An interaction network with topic profile, exposure conformity, and exposure statistics

Whenever a user v generates some content, she *exposes* all of her followers $\mathcal{F}_v^{\leftarrow}$ to it. A user $u \in \mathcal{F}_v^{\leftarrow}$ *conforms* to the exposure if she responds to it via some interaction mechanism provided by the OSN, such as retweets in Twitter and post sharing in Facebook.

Based on the following consideration, we define *exposure conformity* as follows:

Definition 3 (Exposure Conformity). *The function $\varpi : E \rightarrow \mathbb{Z}^+$ represents the degree of exposure conformity between pair of users; in particular, $\varpi(u, v)$ represents the number of times that the follower u has conformed to the exposure from the followed user v.*

Figure 1 shows an example interaction network with the exposure conformity weights associated to edges, and the topic profiles associated to vertices.

Our goal is to identify the top influencers with respect to a given user $u \in V$ and a given topic $t \in \Theta$. In such sense, our approach is *personalized*: it provides information about the most likely users to contact on a particular topic so as to propagate influence towards the user in consideration.

This problem has not to be confused with the identification of the top influencers with respect to a given topic, called *topic-wide influencers detection* [20, 14, 3, 8, 4]. As we shall validate later in Section 4, topic-wide influencers are not necessarily the best influencers for a specific user.

The TopID problem (TOpic and exposure-conformity-aware Personalized Influencers Detection) can be identified as follows:

Problem 1 (TopID). *Given the input constituted by an interaction network $G = (V, E)$, an exposure conformity function ϖ and a topic profile function ϕ, and given a user u and a topic t, the goal is to identify a scoring function $\psi_{u,t} : V \rightarrow \mathbb{R}^+$ that associates all users with an* influence score *measuring their ability to influence user u with respect to topic t.*

By ordering the nodes in non-increasing scoring values $\psi_{u,t}$, it is then possible to obtain the top-K influencers for node u, where K is a parameter of the desired output.

All the notations and symbols that are used in the paper are presented in Table 1.

3. THE TOPID FRAMEWORK

In this section, we present the framework that we use to compute the scoring function. The framework is composed of two phases. In the first one, *pre-processing*, the input datasets are transformed into the exposure conformity and topic profile functions necessary to our computation. In the second phase, *scoring*, the TopID algorithm first computes the influence propagation probabilities and then computes the scoring function.

Notation	Description
$\mathcal{F}_v^{\leftarrow}$	Users who follow v
$\mathcal{F}_v^{\rightarrow}$	Users who v follows
Θ	The set $\{1, \ldots, T\}$ of topics
$\phi_t(v)$	The interest of user v in topic t, obtained from the topic profile $\phi(v)$
$\varpi(u, v)$	Exposure conformity of user u with respect to the exposure from v
$\psi_{u;t}(v)$	Influence score of user v upon user u according to topic t
$\pi_t(u, v)$	The similarity between the topic profile of users u and v according to topic t
$\rho_t(u, v)$	The influence propagation probability from v to u (the probability that u will be influenced by v) with respect to t
Ω^S	A sample space of the set of all possible activation paths from the set S to the whole network
$LID_u(S)$	Local degree of influence of the set S of users upon u

Table 1: The list of symbols and notations used in the paper

3.1 Phase 1: Pre-processing

In the pre-processing phase, two tasks are performed: (1) computing a topic profile for every user and (2) computing the exposure conformity for every edge.

Topic profiling: OSN platforms provide means for users to generate posts. We assume that users primarily post on a limited number of relevant topics, and we employ a topic-profiling process so as to identify such topics.

For each user v, the topic profile is computed based on the content (tweets, posts) that they have generated. Each node is associated to a collection of relevant words extracted from such content.

We adopt a widely-used technique [20, 14, 4] called Latent Dirichlet Allocation (LDA), an unsupervised machine learning algorithm that is used to classify collection of words in a corpus into a set of topics, represented as a probability distribution [5]. As a result, LDA associates the content generated by each user v to a topic-proportion vector $\phi(v) \in [0, 1]^T$, such that $\phi_t(v)$ can be considered the probability for user v to show interest on topic t. For example, for a finite number of topics T, consider the interaction network in Figure 1. The topic profile of u is $\phi(u) = (0.34, ...)$, that corresponds to a 34% probability of interest on the first topic.

Once we profiled each user with regards to the topics, our next goal is to compute the topic similarity between users. Topic similarity has been regarded as an important factor in measuring influence [20, 14]. The main reason for this being the notion of *homophily* from social theory. According to this theory, it is the like-mindedness of a group of users who led them to form different kinds of links in social networks [17]. The like-mindedness can be measured in terms of characteristics, attributes, interest, and other OSN user behaviors. It has been argued that homophily is one of the factors for users to form interconnection or interaction links [9, 20]. In other words, connected users are likely to share interest in similar topics. Furthermore, some papers [14] have also argued that a user shows interest in others content when she finds it to be relevant and unique to her information need.

Therefore, partially agreeing to this line of research, we consider topic similarity as one of the important influence factors. We compute topic similarity as Weng et al. [20], i.e. the topic similarity between two users u and v for a topic t is denoted as $\pi_t(u, v)$ and is given by:

$$\pi_t(u, v) = 1 - |\phi_t(u) - \phi_t(v)| \tag{1}$$

Exposure Conformity: Weng et al. [20] focused on topic similarities to compute influence strength, motivated by an evidence of homophily in their (relatively small) Twitter dataset (6,748 Singapore users), where 72.4% of the twitterers are in a reciprocal relation with more than 80% of their followers. Later studies [7, 1], however, have shown that in a significantly larger dataset (54,981,152 users and 1,963,263,821 links) there is much smaller reciprocity (10%). For this reason, Cha et al. [7] have focused on three influence metrics, namely in-degree influence, retweet influence and mention influence.

Exposure conformity is a metric that enables us to identify an unified measure of influence based on the above findings, capable to account for all of the observed interaction histories among users. Hence we compute it as the sum of the measures of three kinds of interaction, namely retweets, mentions and replies, along a directed edge (u, v), stored in $\varpi(u, v)$.

3.2 Phase 2: Scoring

Once the desired input is obtained, thanks to well-known techniques, in this section we compute the influence score function. In order to do so, we need first to provide a model for influence propagation probability, by combining topic similarity and exposure conformity.

Influence Propagation Probability: Let $\rho_t(u, v)$ denote the influence propagation probability from v to u according to topic t; in other words, the probability that u will be influenced by v when v post some content related to topic t. It is computed as:

$$\begin{aligned} \rho'_t(u, v) &= \frac{\varpi(u, v) + c}{1 + E(u, v)} \cdot \pi_t(u, v) \\ \rho_t(u, v) &= \frac{\rho'_t(u, v)}{\sum_{w \in \mathcal{F}_u^{\rightarrow}} \rho'_t(u, w)} \end{aligned} \tag{2}$$

Here, $E(u, v)$ is the total number of times that u has been *exposed* to the posts from v (as a simplified assumption we consider $E(u, v) = E(v)$, where $E(v)$ is the number of posts of v) and c is a small constant. The constant is included to prevent vanishing probabilities; i.e., if u and v have never interacted, $\varpi(u, v)$ would be 0, and hence $\rho_t(u, v) = 0$. Moreover, the fact that we have not observed any interaction so far does not necessarily mean that they will never interact.

The notion in the above formulation is that, each time a user posts some content, all of her followers are exposed to it. But only those who reacted are leaving traces of influence propagation. This model allows us to capture influence propagation probability based on users topic similarity and the degree of actual activity footprint of users.

To better understand TopID's computational model, let us consider the scenario where all the users in $\mathcal{F}_u^{\rightarrow}$ have the same topic profile for a particular topic. Consider the example network in Figure 1 and assume that *Football-fb* is one of the topics. In addition, we have two actual Twit-

Algorithm 1 TopID algorithm

Input: G: The interaction network
Input: Φ: the topic profile function
Input: ϖ: the exposure conformity function
Input: u: The personalized user u
Input: Θ: The set of T topics
Input: α: The teleportation probability
Output: $\psi_{u;t}(V)$: Each user's $v \in V \setminus u$ influence score on u according to the topics $t \in \Theta$.

1: **procedure** TopID
2: **for** $(w, v) \in E$ **do**
3: **for** $t \in \Theta$ **do**
4: $\rho'_t(w, v) = \frac{\varpi(w,v)+c}{1+E(w,v)} \cdot \pi_t(w, v)$
5: $\psi_{u;t} = \tau_{u;t}$ ▷ Initialization
6: **repeat**
7: $\psi'_{u;t} = \psi_{u;t}$
8: **for** $v \in V$ **do**
9: **for** $t \in \Theta$ **do**
10: $\psi_{u;t}(v) = \alpha \cdot \tau_{u;t} + (1-\alpha) \sum_{w \in \mathcal{F}_v^{\rightarrow}} \psi'_{u;t}(w) \cdot \rho_t(w, v)$
11: **until** *Convergence*
12: **return** $\psi_{u;\Theta}$

ter users (v-@SkyFootball and w-@FullTimeDEVILS) and one ideal Twitter user (u- Manchester United fan). @SkyFootball is one of the main stream media user accounts dedicated to football news. @FullTimeDEVILS is a user account dedicated to the Manchester United Football Club (United) related news and it is among the popular United fans hot discussion/debate bases. Our ideal user u represents a typical United fan who actively engages in discussion related to his club. Assume that v and w have the same topic profile for the first topic - *fb*, *i.e.* $\phi_{fb}(v) = \phi_{fb}(w) = 0.6$. This leads us to compute the same topic similarities for $\pi_{fb}(u, v) = \pi_{fb}(u, w) = 0.74$. Therefore the influence propagation probability solely depends on ϖ, and hence $\rho'_{fb}(u, v) \approx 0.008$ and $\rho'_{fb}(u, w) \approx 0.119$. (For simplicity we have ignored the constant term c in the numerator and 1 in the denominator.) Clearly we see that even though the two users v and w are equally relevant in terms of the topic football for the ideal user u, we notice a significant difference between their influence strength upon u. The main reason for this comes due to the exposure conformity that captures a user's preference from two equally relevant information sources due to a deeper level of interest. Even though the topic football is relevant for this user, his particular interest in football can be further uncovered by the frequency of interactions he made with users that are equally relevant for this topic.

The other obvious scenario is where the users in $\mathcal{F}_u^{\rightarrow}$ have completely different topic profiles for a topic. In these kinds of cases, apparently the exposure conformity serves to bias the influence propagation probability.

In general, regarding the topic similarity between u and the users in $\mathcal{F}_u^{\rightarrow}$, one of the following two conditions hold:

1. The topic similarity between u and a subset of the users $U = \{w_1, \ldots, w_j\} \subseteq \mathcal{F}_u^{\rightarrow}$ for a given topic t is the same, *i.e.* $\pi_t(u, w_1) = \ldots = \pi_t(u, w_j)$.

2. The topic similarity between u and a subset of the users $U' = \{w_1, \ldots, w_p\} \subseteq \mathcal{F}_u^{\rightarrow}$ for a given topic t is different, *i.e.* $\pi_t(u, w_1) \neq \ldots \neq \pi_t(u, w_p)$.

In the former case, a user $w \in U$ *is most likely* to propagate influence on the user u compared to any $w_j \in U \setminus w$. Then w should be the user who u has the most frequent interaction with relative to U, *i.e.* $\nexists w_j \in U \setminus w : \varpi(u, w_j) > \varpi(u, w)$. In the latter case, the influence propagation probability from $w' \in U'$ to u *is strongly biased*, compared to any $w_p \in U' \setminus w'$. Then w' is the user who u has the most frequent interaction with relative to U', *i.e.* $\nexists w_p \in U' \setminus w' : \varpi(u, w_p) > \varpi(u, w')$.

From these observations we realize that TopID's computation scheme does not rely on a strong homophily assumption. Rather it strives to strike a balance between retrospective analysis and homophily.

The pseudo code for computing the influence propagation probability presented in this subsection is given in Algorithm 1, lines 2–4.

Score function computation: We now illustrate how the influence propagation probabilities are used to compute the influence score of all the users upon a personalized user according to different topics. The score in turn is used to obtain the top-K influencers of the given user for a given topic. That is, given a user $u \in V$ and a topic $t \in \Theta$, we compute every user's $v \in V \setminus u$ influence score on u relative to t, denoted by $\psi_{u;t}(v)$, and rank them accordingly. Recall that this is the scoring function that we need to compute in order to address Problem 1. The scoring is based on the random walk model similar to the topic-sensitive variant of PageRank-like algorithms [13, 14, 20].

Consider a random item ι pertinent to a topic t, randomly propagating along the edges in the network starting from some source user. During its propagation, it jumps from a user v to w with probability $\rho_t(w, v)$, i.e., the probability that w will be influenced by v (the probability that w adopts item ι from v).

In terms of the random walk model, after the item ι has arrived at user v, w makes a transition to v with probability $\rho_t(w, v)$ and adopt the item. Then a user v is considered to be a likely influencer of her follower w, i.e., $w \in \mathcal{F}_v^{\leftarrow}$, relative to t, if v has a strong influence propagation power upon w so that w is frequently convinced to adopt random items from v that are pertinent to t. The same principle is applied to the followers of w: if w manages to frequently convince a follower x, i.e., $x \in \mathcal{F}_w^{\leftarrow}$, to adopt items pertinent to t, w in turn is likely to be the influencer of x according t. At this point, consider a chain of follower relations, *i.e.* $\{(w, v), (x, w)\} \subseteq E$. In an intuitive sense, we assume that if v is a most likely influencer to w, and w to x, then by transitivity v is most likely to influence x. These sorts of chained influence propagations lead us to the recursive formulation in Equation 3.

$$\psi_t(v) = \sum_{w \in \mathcal{F}_v^{\leftarrow}} \psi_t(w) \cdot \rho_t(w, v) \tag{3}$$

In some sense, Equation 3 enables us to compute the influence score of v for any topic t.

Similar to topic-sensitive PageRank like algorithms, in Equation 3 v's influence according to t is computed recursively as a weighted sum of her followers influence score according to t. Nevertheless, our aim is not to merely compute whether a user v is influential according to t, but rather to compute whether v is an influencer of a given user u, according to t. Hence following the same line of argument as above, if several items pertinent to t that originate from v

directly or indirectly propagate to u better than any other source, v is a likely influencer of u based on t. Subsequently we reformulate Equation 3 as in Equation 4.

$$\psi_{u;t}(v) = \sum_{w \in \mathcal{F}_v^{\leftarrow}} \psi_{u;t}(w) \cdot \rho_t(w, v) \quad (4)$$

The above formulation suffers from the well-known problem of "dangling nodes", which could trap influence in our case. Following the familiar trick [18] of "teleportation" we give the complete scoring function as follows:

$$\psi_{u;t}(v) = \alpha \cdot \tau_{u;t}(v) + (1 - \alpha) \sum_{w \in \mathcal{F}_v^{\leftarrow}} \psi_{u;t}(w) \cdot \rho_t(w, v) \quad (5)$$

Where $\tau_{u;t}$ is the teleportation vector and it is 1 if $u = v$ and 0 otherwise, implying that the only teleportation that we have is back to u. α is the teleportation probability, and it controls the decision that a user w has to make. That is whether to adopt an item from (make a transition to) a user v that she follows or to teleport to a random user.

Executing the TopID algorithm according to the following scoring function then allows us to obtain an influence score of each $v \in V \setminus u$ on u according to a given topic $t \in \Theta$, $\phi_{u;t}(v)$. Algorithm 1, lines 5–10 give the high-level procedure followed to compute $\psi_{u;\Theta}(V)$. Afterwards we rank each user to obtain the most likely top-K influencers of u in every topic t. In the following section we empirically validate the effectiveness of our algorithm.

4. EXPERIMENTS AND RESULTS

In this section we evaluate the performance of TopID and compare it against the following four baselines: (1) TwitterRank influencers (2) TopID topic-wide influencers – TopID-TW (3) Local IN-Degree rAnk – *Linda* (4) Global IN-Degree rAnk – *Ginda*. For the first baseline, we have utilized our own implementation of TwitterRank both for personalized and topic-wide influencers. In the second case, we consider topic-wide influencers detected by TopID, *i.e.* without personalizing on any user. The third and fourth are scorings based on in-degree and are among the most popular influence measures that are used in the literature [12, 7, 11, 1]. In the case of *Linda* we consider the set of users that a given user follows, which we have ranked according to their in-degree. Whereas for *Ginda* we take a set of users globally ranked according to their in-degree.

The goal of the evaluation is to measure the effectiveness of the influencers that we have detected in diffusing or spreading influence. This can be achieved by seeding an influence diffusion from the set of influencers and measure the final spread throughout the network according to some diffusion model. For diffusing the information, we adopt the interaction network given in Definition 1, opportunely transposed to reverse the direction of the edges (from an users to her followers instead of the opposite). We call this the *diffusion network*.

4.1 Evaluation Metrics

One of the techniques that is used to evaluate the effectiveness of algorithms like TopID is a metric based on repeated simulations of influence diffusion [14]. In this study we adopt a similar strategy and use the following two metrics: (1) *Influence Spread (IS)* (2) *Local Influence Degree (LID)*. Since our focus is on personalized influencers detection, we shall consider the latter one that is suitable for this detection. The former one (*IS*), first introduced in [15], estimates a given users influence power via Monte Carlo experiments. For our purpose we use it to evaluate the effectiveness of influencers in spreading influence under the topic-wide setting. That is, in a given experiment we simulate influence diffusion starting from the set of detected influencers using discrete time independent cascade diffusion model [15]. In each simulation we keep track of the number of activated users, i.e. the users influenced during the simulation of influence diffusion starting from a specific influence source. Then from a set of R simulations we take the expected number of activated users.

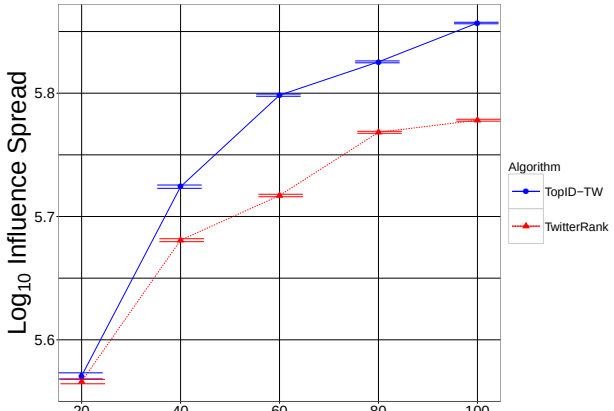

Figure 2: Influence spread performance comparison: TopID topic-wide (TopID-TW) vs TwitterRank

Based on this metric, we compare our algorithm against TwitterRank [20] and the results are reported in Figure 2. The result shows the mean of the expected influence spread averaged over 10 topics in log-scale along with error margins computed at 99% confidence interval. We see that in fact TopID gives an improved performance from the topic-wide perspective. This experiment is carried out to back the case for exposure conformity in general. Let us now consider the evaluation metric intended for the personalized setting.

Local Influence Degree (LID): *LID*, proposed by Guo et.al. [12], is a metric for personalized influence maximization. *LID* is computed based on a set of activated (live) paths from the set S of influencers to the parents of the personalized user. An active path H denotes a path of influence propagation over the *diffusion network*. Suppose the set of all active paths starting from a set of users S in the *diffusion network* is denoted by Ω^S, again let the set of all activation paths from S that do not contain u be $\Omega^S \setminus u$. Then $LID_u(S)$ of a set S on u is computed by combining the influence propagation information from S to the active parents of u and the influence propagation probabilities from the active parents to u (Figure 4):

$$LID_u(S) = \sum_{h \in \Omega^S \setminus u} P(H = h) \left(1 - \prod_{v \in h} (1 - \rho(u, v))\right)$$

where $P(H = h)$ is the probability of the event $h \in \Omega^S \setminus u$. Since this space is exponential, the *de facto* standard to approximate *LID* is using Monte Carlo as follows.

$$LID_u(S) \approx \frac{1}{R} \sum_{i=1}^{R} \left(1 - \prod_{v \in h_i, h_i \in \Omega^S \setminus u} (1 - \rho(u, v))\right)$$

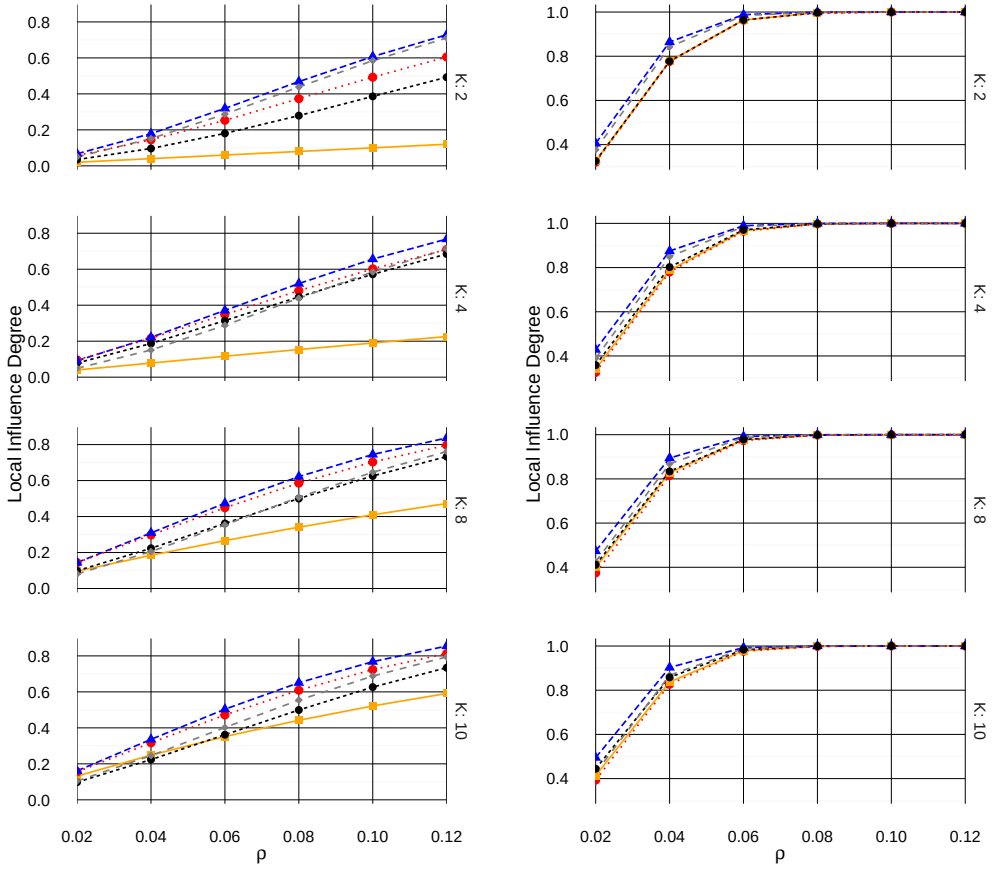

Figure 3: Local influence degree on the *High-in* (left column) and *High-out* (right column) users.

We use $LID_u(S)$ to evaluate the influence degree of a set S of users on a user u that we are personalizing on.

The next task is to pick the users to personalize on, since we do not intend to fully personalize. For this purpose we have used the following two criteria : (i) A user with the highest in-degree and non-zero out-degree – *High-in* (ii) A user with the smallest in- and highest out-degree – *High-out*. The criteria is applied on a set of users that we have filtered based on the number of keywords (keywords that summarize a user's posts), *i.e.*, users with at least 10 keywords. This is to increase the chance of getting users who are interested in at least one topic. We also considered other kinds of users, for example random and highest in-degree and out-degree users, and obtained similar results. Due to space constraint we focus only on the two types of users in Table 2.

4.2 Dataset and Experimental Settings

We exploit a publicly-available dataset released for the

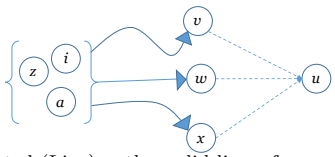

Figure 4: Activated (Live) paths, solid lines, from a set $S = \{a, q, z\}$ of users to parents $\{v, w, x\}$ of u, and $\{v, w, x\}$ are activated users. Dotted arrows are edges in the diffusion network

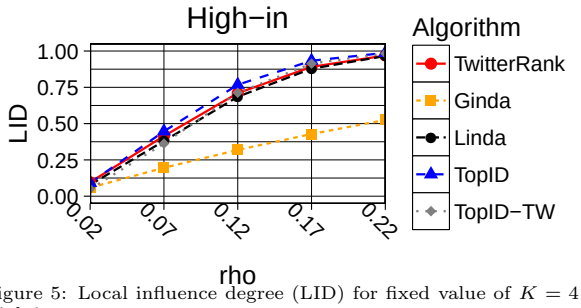

Figure 5: Local influence degree (LID) for fixed value of $K = 4$ and High-in user

2012 KDDCup[1], one the most used in information diffusion studies. It is collected from Tencent Weibo, one of the largest micro-blogging platforms in China and contains 2,320,895 users and 50,655,143 follower relations.

The dataset contains a set of keywords extracted for each user. Based on documents generated from these keywords, we have extracted 10 topics, and for each user we have computed the topic profile across these topics. We utilize the LDA implementation of Graphlab Create, a machine-learning library based on C++ and Python [2]. For all the remaining tasks, the main algorithm and the evaluation methods discussed above, we have implemented them in C++.

[1]http://www.kddcup2012.org/c/kddcup2012-track1
[2]https://turi.com/products/create/

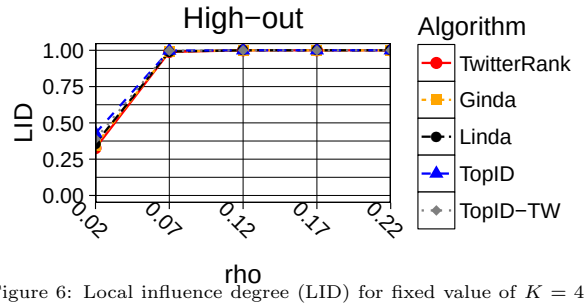

Figure 6: Local influence degree (LID) for fixed value of $K = 4$ and High-out user

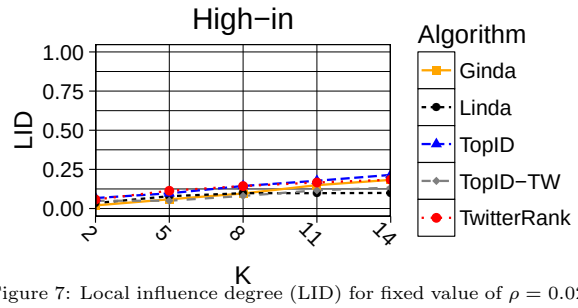

Figure 7: Local influence degree (LID) for fixed value of $\rho = 0.02$ and High-in user

For the teleportation probability, α, in Equation 5 we have used $\alpha = 0.15$. For the Monte Carlo experiments we consider the most widely adopted value for R, $R = 10,000$. For the top-K influencers, we consider $K \in \{20, 40, 60, 80, 100\}$ in the topic-wide setting; and $K \in \{2, 4, 8, 10\}$ in the personalized setting. The selected personalized users' details are given in Table 2. All the results reported below are summarized over the top-3 topics of the personalized user. The distribution over the top-3 topics (topic profile) of each user is shown in the last column of Table 2.

Finally we evaluate performance according to different values of ρ and K. That is, with different values of influence diffusion probabilities $\rho \in \{0.02, 0.04, 0.06, 0.08, 0.10, 0.12\}$ and different number of influential users $\in \{2, 4, 8, 10\}$. Furthermore we study the effect of each parameter independently, *i.e.*, for different values of one parameter we fix the value of the other. Here, the intent is to clearly understand the interplay between ρ, K and the *local influence degree* on a given user. For this reason, we fix the value of K at $K = 4$ and analyze the effect of different values of $\rho \in \{0.02, 0.07, 0.12, 0.17, 0.22\}$. On the other side, for a fixed value of $\rho = 0.02$ we analyze the effect of different values of $K \in \{2, 5, 8, 11, 14\}$

4.3 Results Discussion

In our first experiment we evaluate the performance of TopID and the baselines across different values of ρ and K. Figure 3 shows the results that we have obtained for *High-in* and *High-out* users. In both cases we observe that TopID gives the best results; moreover, as we have anticipated in earlier sections, we can clearly observe that the topic-wide (TopID-TW) influencers are not necessarily influencers of a given user.

The next experiment is carried out by fixing one parameter and observing the effect of the other parameter on different values. In Figures 5 and 6 we report the results for

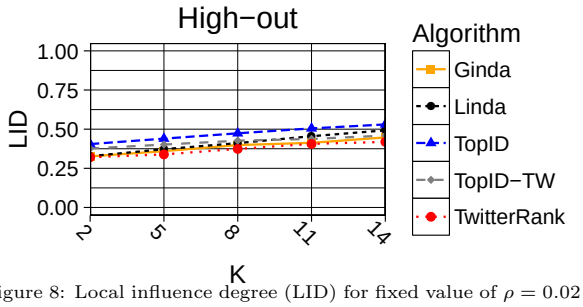

Figure 8: Local influence degree (LID) for fixed value of $\rho = 0.02$ and High-out user

fixed $K = 4$ and in Figures 7 and 8 for $\rho = 0.02$.

In general, in all of our experiments we observe that users with large out-degree are more susceptible to influence than those with small. We also observe that ρ has a significantly greater impact than K in increasing the local influence degree. Increasing ρ causes the *LID* to quickly increase towards one, whereas increasing K has a very small effect.

User	In-degree	Out-degree	Topic-profile(top-3)
High-in	429,570.0	77.0	$\langle 0.52, 0.17, 0.14 \rangle$
High-out	0.0	1242.0	$\langle 0.28, 0.13, 0.11 \rangle$

Table 2: Personalized users' properties, User, In-degree (#Followers), Out-degree(#Users they follow), Topic-Profile (top-3 topics)

5. RELATED WORK

Influence analysis in social networks has been studied extensively [15, 10, 6, 19, 20, 8, 3, 14, 4]. Usually the goal is to identify the most influential users that maximize influence spread throughout the network. The most notable application of such task is viral marketing, where the aim is to seed the marketing campaign from a small number of users, aka influencers, and spread influence through the word-of-mouth effect. In order to effectively execute such a campaign, several approaches have been proposed. A number of studies [15, 10, 6, 19] have addressed this as an influence maximization problem, specially following the seminal work by Kempe et al. [15], in which influence maximization has been formalized as discrete stochastic optimization problem.

Follow up studies [6, 19, 10] have proposed different ways of improving the simple and effective greedy algorithm proposed in [15]. More recent studies [8, 3, 4, 12, 16] have casted the traditional influence maximization problem into different dimensions. Examples of these dimensions include topic-aware influence maximization, and personalized influence maximization. In the topic-aware influence maximization, the goal is to identify influencers according to specify topics, whereas in personalized influence maximization the identification process becomes sensitive to each user. That is, the influence maximization is carried out for each user independently.

The study by Guo et al. [12] is the first to address the personalized influence maximization problem. They formalized the problem in a similar manner as the standard influence maximization, and in a suitable way for personalized setting. They have proposed two kinds of efficient algorithms, the first is a greedy algorithm that is based on Monte Carlo simulation. The second is an online algorithm called *Local Cascade Algorithm* (LCA) based on a "local cascade com-

munity". LCA employs a trick that allows them to identify influencers in an on-line fashion without compromising the effectiveness achieved in their first algorithm.

Very recently, a study by Li et al. [16] have addressed the problem of personalized influential topic search. In this study the goal is to answer a query forwarded by an OSN user. They tackle the problem based on a summary information that captures the context of the user in the social network and topic-representatives and propagation indexes for efficient query answering. In our study we address a somewhat different problem and a completely different approach as described in the paper.

6. CONCLUSIONS

In this study, we have presented a novel algorithm called TopID whose goal is to detect personalized influencers. The algorithm is based on two fundamental notions, namely topic awareness and exposure conformity. Inspired by prior (but partially contradicting) findings, we strive to find a balance and unify them. Topic awareness aims at capturing relevant topics for each user and the similarity between users over different topics, whereas exposure conformity aims at capturing the interaction histories between connected users.

Following a rather strong evidence for exposure conformity from previous studies, TopID tries not to rely merely on topic similarity. In fact, we are able to validate our assumptions through several empirical evaluations and different kinds of users, and overall achieve consistently improved results compared to a set of baselines. In future studies we would like to expand our work by considering interaction histories among groups of "coherent" users, not just between pairs of connected users.

7. REFERENCES

[1] A. S. Badashian and E. Stroulia. Measuring user influence in github: The million follower fallacy. In *Proc. of the* 3^{rd} *Int. Workshop on CrowdSourcing in Software Engineering*, CSI-SE '16. ACM, 2016.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proc. of the* 4^{th} *ACM Int. Conf. on Web Search and Data Mining*, WSDM '11. ACM, 2011.

[3] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *Proc. of the* 12^{th} *IEEE Int. Conf. on Data Mining*, ICDM '12. IEEE Computer Society, 2012.

[4] B. Bi, Y. Tian, Y. Sismanis, A. Balmin, and J. Cho. Scalable topic-specific influence analysis on microblogs. In *Proc. of the* 7^{th} *ACM Int. Conf. on Web Search and Data Mining*, WSDM '14. ACM, 2014.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *Proc. of the* 25^{th} *Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14. Society for Industrial and Applied Mathematics, 2014.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. of the* 10^{th} *Int. Conf. on Weblogs and Social Media*, ICWSM '10, 2014.

[8] S. Chen, J. Fan, G. Li, J. Feng, K.-l. Tan, and J. Tang. Online topic-aware influence maximization. *Proc. VLDB Endow.*, 8(6):666–677, Feb. 2015.

[9] W. Chen, L. V. S. Lakshmanan, and C. Castillo. *Information and Influence Propagation in Social Networks*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2013.

[10] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proc. of the* 15^{th} *ACM Int. Conf.*, KDD '09. ACM, 2009.

[11] E. Dubois and D. Gaffney. The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. In *Proc. of the* 10^{th} *Int. Conf. on Weblogs and Social Media*, ICWSM '10, 2014.

[12] J. Guo, P. Zhang, C. Zhou, Y. Cao, and L. Guo. Personalized influence maximization on social networks. In *Proc. of the* 22^{th} *ACM Int. Conf. on Information & Knowledge Management*, CIKM '13. ACM, 2013.

[13] T. H. Haveliwala. Topic-sensitive pagerank. In *Proc. of the* 11^{th} *Int. Conf. on World Wide Web*, WWW '02. ACM, 2002.

[14] J. Herzig, Y. Mass, and H. Roitman. An author-reader influence model for detecting topic-based influencers in social media. In *Proc. of the* 25^{th} *ACM Conf. on Hypertext and Social Media*, HT '14. ACM, 2014.

[15] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the* 9^{th} *ACM Int. Conf.*, KDD '03. ACM, 2003.

[16] J. Li, C. Liu, J. X. Yu, Y. Chen, T. Sellis, and J. S. Culpepper. Personalized influential topic search via social network summarization. *IEEE Transactions on Knowledge and Data Engineering*, 28(7), July 2016.

[17] C. Lozares, J. M. Verd, I. Cruz, and O. Barranco. Homophily and heterophily in personal networks. from mutual acquaintance to relationship intensity. *Quality & Quantity*, 48(5), 2014.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proc. of the* 7^{th} *Int. World Wide Web Conf.*, 1998.

[19] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proc. of the 2014 ACM Int. Conf.*, SIGMOD '14. ACM, 2014.

[20] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proc. of the* 3^{rd} *ACM Int. Conf. on Web Search and Data Mining*, WSDM '10. ACM, 2010.