The Strength of Considering Tie Strength in Social Interest Profiling

Asma Chader*, Hamid Haddadou, Leila Hamdad
and Walid-Khaled Hidouci

Laboratoire de la Communication dans les Systèmes Informatiques (LCSI), Ecole Nationale Supérieure d'Informatique (ESI), BP 68M, 16309, Oued-Smar, Algiers, Algeria
Email: aa_chader@esi.dz; h_haddadou@esi.dz; l_hamdad@esi.dz; w_hidouci@esi.dz
**Corresponding Author*

Received 24 September 2019; Accepted 21 May 2020;
Publication 28 July 2020

Abstract

With the emergence of social networking platforms and great amount of generated content, analyzing people interactions and behaviour raises new opportunities for several applications such as user interest profiling. In this context, this paper highlights the importance of considering relationship strength to infer more refined and relevant interests from user's direct neighbourhood. We propose WeiCoBSP, a Weight-aware Community-Based Social Profiling approach that leverages strength of ego-friend and friend-friend relationships. The former, describing connections with the profiled user, allows to identify most relevant people from whom to infer worthwhile interests. The latter qualifies connections among user's neighbourhood and enables depicting the most realistic community structure of the network. We present an empirical evaluation performed on real world co-authorship networks, validating our approach. Experimental results demonstrate the ability of WeiCoBSP to infer user's interest accurately, improving greatly the unweighted CoBSP process but also results of experiments assessing separately ego-friend and friend-friend relationships strength.

Journal of Web Engineering, Vol. 19_3-4, 457–502.
doi: 10.13052/jwe1540-9589.19345

Keywords: Social profiling, user profile, relationship strength, weighted social networks, egocentric networks.

1 Introduction

The popularity of social networking platforms is worldwide ever increasing especially with the growing usage of mobile devices and mobile social networks. A recent survey by Statista[1] reveals 2.32 billion and 321 million monthly active users on Facebook and twitter as of fourth quarter 2018,[2,3] with about 55 million status updates made every day on Facebook and 500 million tweets sent on twitter.[4,5] This great volume of social generated content has motivated a lot of research aimed at better understanding people interactions and behavior. In particular, several studies were interested in user social profiling, which is the focus of this paper, to enable a wide range of applications, such as personalization, recommendation and targeted advertising [36]. In the context of social networks, old applications become even more challenging (e.g., serving ads without queries in targeted advertising) and several new tasks such as recommending friends, hashtags or location-based recommendations emerged, which underlines how much is crucial inferring the unobserved user's attributes or interests in current social network analysis.

A straightforward way to infer unobserved attributes is to utilize user-generated content such as tweets, comments and likes or location check-ins [36]. In addition, recent studies increasingly exploit structural information and activities from user's neighborhood (i.e., social relationships) to discover his interests. Such approaches differentiate between personal interests (related to user's generated content) and social ones, inferred from user's neighborhood [29]. These models are based on "social influence" and "homophily" described as the tendency of people to associate with similar others. Indeed, people are more likely to befriend those sharing same characteristics (as age, education, social class, racial/ethnic group, interests or geographic location) and, on the other hand, social relationships influence

[1] https://www.statista.com

[2] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[3] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[4] https://www.omnicoreagency.com/twitter-statistics/

[5] https://www.omnicoreagency.com/facebook-statistics/

people such that friends increasingly resemble one another over time [19]. Leveraging information of social relationships can be useful to alleviate the missing attributes problem especially in cold start situation (empty profile for new users) or sparse user profiles (missing interests due to the lack of user's activity). In fact, many social media users, preserving their privacy, disclose only few information publicly. In addition, there is a rise of passive use of social media platforms. As reported by [36], four out of ten Facebook users as well as a significant portion of Twitter ones browse only information without generating any content and these users are potential target users to aforementioned applications.

The existing approaches can be categorized along different axes, based on relationships they use (full social graph or direct relations), on how profiles are represented (keyword, concept or semantic network profiles [36]) or based on methods used for data collection (inside vs. outside the target platform) among others. This paper is interested on using information from direct social relations of the profiled user (aka. user's egocentric network) which has been proved efficient to infer relevant interests [41]. Moreover, a keyword representation of profiles (vector of weighted terms) is adopted with focus on profiles of user's interests.

To fully exploit the information contained in social relations, we are working in our research team on different approaches taking into account the relationships strength [7]. In this paper we present WeiCoBSP, a Weight-aware Community-Based Social Profiling approach that leverages strength of both ego-friend and friend-friend relationships to infer user's interests. To the best of the authors knowledge, there is no direct literature exploring the effect of the two types of relationships strength in social profiling. This approach is based on CoBSP (Community-Based Social Profile) process proposed by [41]; admitting that social groups via particular affinities with the user (e.g., family, sports club, etc.) are more significant to describe him than individuals (e.g., the most central ones). We aim to go one-step further by investigating tie strength together with community structure around the user to produce a more relevant social profile. In fact, the CoBSP approach, designed to operate on binary representation of the user's egocentric network, assumes that all friends are equally significant to the profiled user as well as to each other. This assumption is too restrictive: in analyzing social relationships, not only the structure of the network is important but also the strength of connections; different strengths interpret different association and similarity degrees among users and thus communities extracted under the binary correspondence of the network are often different and less representative of the real community

structure. Moreover, the study of tie strength speculates that "the stronger the tie connecting two individuals, the more similar they are" [13]. Accordingly, the ego-friend strength are integrated to enhance the distinguishable extent of profiled user ties and select the most relevant communities as a source of information. We believe that relationships with stronger affinity or to which the user is more dedicated, may reveal more valuable information about his interests. The friend-friend strength, for its part, characterizes the relationships among user's friends. We aim at modelling egocentric network as close to real network as possible in order to depict the most accurate community structure around the user.

Results from an extensive empirical evaluation of WeiCoBSP on real world co-authorship networks (collected from DBLP computer science bibliography) are reported. Different experiments are also conducted to assess separately the effect of each of ego-friend and friend-friend relationships strength. These latter suggest that the proposed approach (leveraging both relationships strength) provides a considerably higher accuracy of interests' prediction compared to unweighted CoBSP as well as to experiments based on solely one relationship type (ego-friend or friend-friend); which demonstrates the relevance of prior premises.

The remainder of the paper is organized as follows: Section 2 introduces our problem and the associated preliminaries and Section 3 the related work review. Section 4 first presents the existing work on user profiling from egocentric social network communities and then describes our proposition to extend it for weighted networks. The performance of the proposed approach is studied in Section 5, which presents and discusses experimental results and findings. Finally, Section 6 draws conclusions and gives some perspectives to our work.

2 Motivation and Problem Abstraction

This section first introduces the egocentric network and user profile models and then formulates our problem.

An egocentric (or simply ego) network generally consists of a single user, so-called ego, together with individuals to whom this ego is connected (alters) and all the connections among them. Other possible models consider either only the ego-alter ties as in [3, 34], discussed in literature review (see Section 3) or only alter-alter ties as in the baseline method [41] which discards ego-alter ties since it assumes a binary network where all alters are equally significant to ego. In this paper, we adopt the former model

and assume the ego network to be weighted and undirected. The egocentric network is described as follows: For a given user *u*, so called ego, for whom we aim to build social profile, we consider the non-oriented weighted graph $G = (V,E,E')$ where V denotes the set of individuals directly connected to u (user u is not in V), E is the set of relationships among V nodes and E' denotes the set of relationships between the user u and V nodes. For each node $v \in V, Wv \in E'$ denotes its weight to the user u and for two nodes $v,v' \in V$, $Wvv' \in E$ denotes the weight of the edge between them, which can take on only positive values. The network community structure is denoted by $C = \{C1, C2, C3 \ldots\}$, where $Ci \in C$ denotes a community (for simplicity, Ci is denoted as C if there is no confusion).

Regarding profile representation, the user profile is composed of a set of weighted user interests; the weight of each interest $i \in I$ for a user $v \in V$ (denoted $w(i,v)$) indicates the importance of the interest i with respect to user v. Formally,

$$P(v) = \{(i, w(i,v)),\ i \in I, v \in V\} \tag{1}$$

where I denotes the set of user interests, and V denotes the set of users.

Each user's profile, as modelized in [41], is composed of a "user" and "social" dimensions. Each of them is represented as a vector of weighted interests. In the user dimension $U(v)$, interests are computed based on the user's own activities (e.g. shared information, tweets). On the other hand, interests in the social dimension $S(v) = P(G)$, are extracted using information gathered from people in the user's egocentric network G.

Based on the above definitions, we formulate now our problem. Considering the increasing passive use of online social networks along with the wish of users to preserve their privacy, many user's attributes are missing in their profiles. We study herein the user interests' inference in egocentric network through relationship strength. Thus, in our setting, for a user u, whose interests are unknown, we are given his ego network $G = (V,E,E')$ with available interests profiles from its alter (their user dimensions, i.e. the set $\{U(v), v \in V\}$) and we aim to predict u's interests by leveraging community structure C and relationship strength (both ego-alter, E', and alter-alter, E, sets) to produce his social dimension $S(u)$.

To infer user's interests, the community-based model [41], upon which our approach is built, relies on the striking observation that communities around the user are more significant to describe him since they represent groups of people having particular affinities with him (e.g., family, sports club). However, this work takes a simplistic assumption that all friends

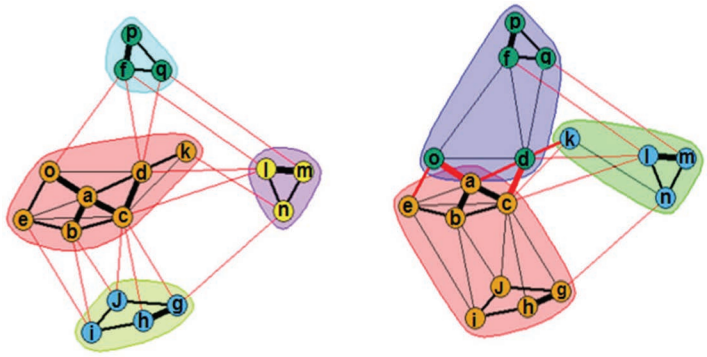


Figure 1 Communities detected on weighted and binary representation of a network.

are equally related to ego user as well as to each other (i.e. they ignored the strength of relationships between users), which will result in not fully capturing the richness of the social information in the user's egocentric network. On the one hand, because all user friends are obviously not equally significant to him. For instance, multiple collaborations or citations in co-authorship network reflect in most cases authors shared interests; such relations are more likely to provide more relevant information about the profiled user than occasional collaborations. Thus, modelling real networks as simple binary may increase the level of noise in the learned profile with lot of meaningless interests inferred from less relevant people. On the other hand, this issue affects, in particular, the community extraction phase on which the profiling process is completely based. Considering only binary relations between user's friends cannot depict correctly the community structure of the egocentric network, highly dependent on link weights [12, 27]. Communities based only on binary representation of social connections are quite different and less accurate than those extracted by including link weights. To illustrate this, Figure 1 shows an example of detected communities on a weighted network and its binary representation (using *greedy optimization of modularity* method [9]), where different structures can be seen (four communities detected on weighted network versus only three on binary one). We can also observe that some nodes, such as *d*,*c* or *a*,*o*, showing greater tie strength and assigned to the same community in weighted version (which makes sense since they reflect closer relationships), cannot be correctly depicted on binary network.

Consequently, if the networks are treated as binary, much of useful knowledge will be lost, which probably leads to degradation in performance. In this paper, we propose to integrate the evidence of the relationship strength between users to further enhance the distinguishable extent of ties and improve the community-based social profile inference.

Once the network model is chosen to be weighted, we try to address the following challenges (i) how to change metrics and algorithms describing different steps of the CoBSP process to deal with those networks directly? And (ii) How to choose relevant communities in the user's weighted network from whom significant interests will be derived?

3 Related Work

Social user profiling has drawn extensive interest as it raises new opportunities in many applications like personalized search, targeted advertisement and recommendation [36]. Surveys and literature reviews such as [1, 36] provided an overview of the methods used in user profiling. Several studies focus on user generated content (also called user-centric data) such as tweets [8, 35, 40], user preferences (e.g. comments or likes) [4, 17] and location check-ins [44] to predict user profile attributes (e.g. gender, location, political affiliation). However, inferring interests from user centric-data requires users to be continuously generating content. Hence, those methods fail to profile passive users who consume information without creating any content. In another aspect, the scientific literature outlines many studies that, addressing this deficiency, exploit the relationship information (i.e., the social graph) in user profiling [10, 23, 26, 41]. The intuition behind such approaches is the tendency of individuals to associate with similar others, the so-called 'homophily' concept. Earlier studies along this line [15, 16, 26] were designed to operate on a full social graph, which is often hard to obtain and too costly to process. Accordingly, other approaches were interested in identifying users' attributes from his direct neighborhood relationships. Most of them were conducted on Twitter [36] and consider only user-friend's connections [3, 34, 36]. Still in direct neighborhood based methods, but considering connections among friends too, [41] describes a community-based algorithm to infer user's attributes via user-groups affinities. [23] attempts to learn attributes by searching for the reasons behind link formation within the users' egocentric network. The underlying assumption is that social connections are discriminatively correlated with user attributes (e.g., employer) through relationship type (e.g., colleague).

Similarly, [25] proposes a social-aware topic model by reconciling the observed user characteristics and social network structure to discover the latent reasons behind social connections and further extract users' potential attributes.

While in the future we plan to integrate insights from diverse research directions, particularly relationships type [23], we focus herein on works directly related to ours. The study that motivated our approach [41] describes a community-based algorithm to infer user's attributes in his egocentric network. Authors represent each user profile with a personal and social dimension and infer social attributes utilizing the combined interests of communities. To do so, the proposed process, named CoBSP (Community-Based Social Profile) starts with a community detection step followed by community profiling where interests of each community are determined from its members to finally combine all interests and construct the social dimension of the profiled user. This process will be presented later in Section 4.1.

A drawback of this approach is that a sparse or small network could lead to misinterpretations in the user's modelling process [28]. Thus, their second work [28] extends the previous community-based algorithm by integrating more relationships so as to generate more significant communities. A snowball sampling technique was adopted to identify and add user's distance-2 neighbors (friends of a friend) into the egocentric network. More recent studies [5, 29, 30] were interested in dynamic characteristics of the user's social networks. They integrate temporal criteria to the existing community-based approach and consider the evolution of both relationships and shared information in the network. Their time-aware approach showed good performance on Facebook and DBLP data but in [29], they found that it does not produce the same effectiveness on Twitter dataset.

Previous works studying the CoBSP [5, 28–30, 41] have mainly focused on binary friendship relations (e.g., friends or not) and treat equally all relationships of the profiled user. In reality, social connections are not merely binary; they have associated weights that record people's relation strength: close friends, acquaintances, teammates, family or work colleagues. Our objective is to go one-step further by investigating tie strength to emphasize the discriminative resolution of user's relationships in order to produce a more relevant social profile. The next section will present the existing CoBSP process and our weight-aware approach leveraging both community structure and tie strength of users.

4 Proposition: Weight-aware Egocentric-network Based Social Profiling

4.1 Baseline Approach: Community-based User Profiling Process

Before detailing the CoBSP process, it is worth noting that this latter performs on a binary alter-alter model of egocentric network. Thus, for a given ego user *u*, it considers the sub graph $G(u) = (V,E)$ where *V* is the set of individuals directly connected to *u* and *E* is the set of relationships between *V* nodes, without any associated strength. The generation of interest profiles can be seen as a four-stage process described as follows. (Further details can be found in [41])

Stage 1 – *Community detection*: this first step consists on extracting communities from the egocentric network. As individuals usually belong to multiple communities at once, and structure of egocentric network evolves over time, this step is performed by applying iLCD algorithm [6] that performs very well with overlap as well as dynamics of network [41].

Stage 2 – *Community profiling*: In this second stage, the sematic profile of each community, $c \in C$, found in the previous step is computed. Given that the "user dimensions" of people around the ego (profiled user) are already calculated and represented in a vector of weighted interests, the profile of a community *c* aggregates interests of all members forming this community. Following ideas used in the TF-IDF measure [41], each interest *i* is assigned a score according to its frequency in the "user dimensions" of *c* members: the set $U(v_c), v_c \in c$ and in other communities' profiles. The set $I(c)$ contains, therefore, all the community c's weighted interests.

Stage 3 – *Interest weight calculation*: this step consists in computing the final weight of interests in the community profile. The weight of each interest *i* in the community *c* called $w(i,c)$ depends on structural score of the community (centrality measure) and semantic score (calculated in previous stage) by a parameter $\alpha \in [0, 1]$ controlling the contribution of each score (α is set empirically) as presented in Equation (2):

$$w(i, c) = a\ Struct_{score}(c) + (1 - a)\ Semantic_{score}(i, c) \tag{2}$$

Stage 4 – *Social dimension derivation*: In the final stage, the social dimension $S(u)$ of the user's profile is derived by combining the weights calculated in the previous phase for each interest $i \in I$. At the end of the third step, an interest i may appear in several community profiles and have, thus, different associated scores. This final step consists in computing a single weight for each interest i in $S(u)$; to do so, the combination function Lin-CombMNZ [14] is adopted [41].

4.2 Proposed Approach: Weight-aware Community-based User Profiling Process

Having provided an overview of CoBSP process, this section presents the proposed weight-aware approach, so-called WeiCoBSP, which leverages the strength of both ego-alter and alter-alter ties to infer user's interest. Different from the baseline by its model of egocentric network, this extension affects all the stages of the process since a different model of egocentric network is adopted. In the following, we will go through the four stages and describe at each step the main changes and differences between our proposed WeiCoBSP and the existing CoBSP processes.

A global view of the profiling process, which expects in input the weighted egocentric network $G = (V,E',E)$ and the alter user dimensions (set $U(v), v \in V$), is presented in Algorithm 1.

Algorithm 1 Weighted Community-based Social Profile (WeiCoBSP)

Require: User's u egocentric weighted network G = (V,E′,E),
"User dimensions" of people in G,{U(v),v ∈ V}
Ensure: S(u): User's u social dimension
C ← Weighted Overlapping Community Detection (G)
for each c in C **do**
  I(c) ← Semantic Score (i,c) using Equation (5)
  Weighted Structural Score (c) ← Centrality (c,G) using Equation (8)
  for each i in I(c) **do**
    w(i,c) ← Interest Final Score Calculation (i,c) using Equation (11)
  end for
end for
for each i in I(C) **do**
  W(i,S(u)) ← Score Combination (i, C) using Equation (13)
end for
return S(u)

4.2.1 Stage 1: Weighted community detection

The first step (Line 1 in algorithm) consists on extracting overlapping communities in the weighted egocentric network *G*. Since many networks are naturally weighted (in particular Online Social Networks), the community discovery in weighted networks has attracted increasing attention [2, 18, 24, 27, 42]. Some of proposed algorithms are designed to detect overlapping communities (where a node can belong simultaneously to several groups) and/or deals with the network dynamics [2, 18, 42]. The iLCD algorithm [6] used in the original process does not take tie strength into consideration. Thus, in our work the OSLOM algorithm [18], which considers the link weights, and handles both overlapping communities and network dynamics, is used. After this first stage, the set *C* contains all extracted communities.

4.2.2 Stage 2: Community profiling

As for the original approach, in the second phase of the algorithm (Lines 3,4), the profile of each community *c* in *C* is computed. We recall that $I(c)$ denotes the set of community *c*'s interests. Each interest *i* in $I(c)$ is weighted according to two scores: its semantic score in the community *c* and the structural score of *c*. The profiles of all nodes in *V* are assumed to be already calculated and represented in a vector of weighted interests.

In our extended process, the semantic score calculation remains unchanged. For each interest *i* in a community *c*, its semantic score depends on the weight of this interest for all members of the community as well as its recurrence in other communities' profiles. Inspired by the well-known TF-IDF (Term Frequency/Inverse Document Frequency) weighting measure, the semantic score for an interest *i* in the community *c* will be computed following two steps:

In the first, is calculated the average of weights of this interest for all members forming the community to get its W_{Stf} score, standing for *semantic tf weight* as follows (Equation (3)):

$$W_{Stf}(i,c) = \frac{\sum_{v_c=1}^{m} W(i,U(v_c))}{m},$$
$$W(i,U(v_c)) = \begin{cases} w(i,v_c), & \text{if } (i,w(i,v_c)) \in U(v_c) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $U(v_c)$ is the "user dimension" of the node $v_c \in c$, $w(i,v_c)$ represents the weight of the interest *i* in $U(v_c)$ and *m* is the number of users in community *c*.

The second step is, for its part, analogous to the IDF used in Information Retrieval systems to determine the relevance of terms in a specific

document [38]. Herein, it is about finding out the specificity of each community regarding other ones. In other words, it consists of looking, for each interest $i \in I(c)$, whether this interest is common or rare across all communities. We assume that rarer an interest is, the more representative of the intrinsic affinity between members of the community, it will be. This second score, so called W_{SIdf} score, standing for *semantic idf weight*, is computed by the following Equation (4):

$$W_{Sidf}(i,c) = log\frac{|C|}{|\{c \in C \ : i \in I(c)\}|} \tag{4}$$

where $|C|$ is the total number of communities in the user's egocentric network and $|\{c \in C: i \in I(C)\}|$ the number of communities where the interest i appears.

Finally, the semantic score of the interest i in the community c is computed as the product of the two W_{Stf} and W_{Sidf} scores:

$$Semantic_{score}(i,c) = W_{Stf}(i,c) \times W_{Sidf}(i,c) \tag{5}$$

Here, a high sematic score is reached by a high interest frequency (and weight) among the community members and a low global frequency of the interest in the whole egocentric network (i.e. in other communities).

On the other hand, the structural score of communities (Line 4 in algorithm) must be generalized to cover the weights in the egocentric network G.

Following the original work, we consider structural score as centrality measure of the community and extend it to apply to our weighted graph G. Centrality is regarded as one of the most important tools to explore actor roles in social networks. In our context, it may be one of the commonly used measures as degree or proximity.

As reported in [41], Everett and Borgatti [11] proposed extensions of usual individual based centrality measures to the group's level. For instance, the degree centrality considered here, is defined for a community c in a graph G as the number of people outside the group that are connected to at least one c member (denoted $|N(c)|$) normalized by the number of people that are not in c, ($|V| - |c|$). Note that different ties to the same member of c community are only counted once [11]. The expression of group degree centrality is given below:

$$CentralityDegree(c,G) = \frac{|N(c)|}{|V| - |c|} \tag{6}$$

In the other side, some network centrality measures based on individuals have been generalized for weighted networks [27, 32]. The simplest extension of a node degree is the sum of all the weights along its edges [27]. However, authors in [32] showed that the latter proposition gives only a crude measure of the node's real involvement in the network. Accordingly, they proposed a more realistic extension of degree centrality by combining both the number of ties (neglected in the previous extension) that a focal node has, and the average weight of these ties (strength) and using a tuning parameter to set the relative importance between them. They formalize it as:

$$WeightedDegree(n, G) = Deg(n)^{(1-\beta)} \times Stren(n)^{\beta} \tag{7}$$

where *Deg*(*n*) is the traditional degree centrality of the node *n* (based solely on the number of ties) and *Stren*(*n*) is the total weight of *n*'s ties. Note that β is a positive parameter that can set according to the research setting and data. If this parameter is between 0 and 1, then having a high degree is taken as favorable, i.e. it increases the value of the measure, whereas setting it above 1 decreases the value of the measure in favor of a greater concentration of node strength.

To compute the structural score, we deem that both the number of connections and their associated strength are indicators of the community's centrality in the egocentric network. Hence, based on a combination of the two extensions, i.e. the group degree and the weighted degree centralities, presented in formulas (6) and (7), the weighted degree centrality for a community *c* in *C* is defined as follows (Equation (8)):

$$WeightedStructural_{score} = \frac{|N(c)|^{(1-\beta)} \times W(c)^{(\beta)}}{|V| - |c|} \tag{8}$$

where |*N*(*c*)| denotes the group extended degree centrality and *W*(*c*) the group extended strength centrality similarly computed, i.e. the total weight of people outside the community *c* that are connected to at least one *c* member (Equation (9)).

$$W(c) = \sum_{v \in c, v' \in (V \setminus c)} W v v' \tag{9}$$

where $W_{vv'} \in E$ denotes the strength of the tie between *v* and *v*′ in the ego network. Here note that like for group degree, multiple connections are only counted once. We can use either the maximum or the average of these connections weights. Regarding the tuning parameter, a value of β between

0 and 1 allows us to consider both number and strength of links. Moreover, if the parameter is set to 0, the outcomes of the measures are solely based on the number of ties and conversely, if it equals to 1, the measure is based on ties strength only and the number of ties is disregarded.

4.2.3 Stage 3: Interest weight calculation

Once both semantic and structural scores are computed for all communities, the next step (line 6 in algorithm) consists on computing the final interests' profile of each community. As for individuals in the egocentric network, the profile of a community is composed of a set of weighted interests.

$$P(c) = \{(i, w(i,c)),\ i \in I(c),\ c \in C\} \tag{10}$$

where $w(i,c)$, the weight of the interest i in the community profile $P(c)$, is computed as in the original CoBSP process by an adjusted combination of the semantic and structural scores as in Equation (11) below:

$$w(i,c) = a\mathit{WeightedStructural}_{\text{score}}(c) + (1-\alpha)\ \mathit{Semantic}_{\text{score}}(i,c) \tag{11}$$

where $\alpha \in [0, 1]$ is a tuning parameter to evaluate the impact of structural score compared to semantic one.

4.2.4 Stage 4: Social dimension derivation

The last phase (Lines 9 to 11 in Algorithm 1) consists in deriving the social dimension $S(u)$ of the user's profile by computing the final weight of each interest $i \in S(u)$ (called $w(i,S(u))$. As each community of the user's egocentric network is addressed separately in the previous steps, an interest i may appear in different community profiles. Now we investigate how to obtain a single weight from the different interest i weights $w(i,c)$ of each community c where $i \in I(c)$.

To combine the weight of interests of communities, authors in [41] apply the linear function LIN-CombMNZ [14], a variant of the CombMNZ function commonly used in information retrieval to merge search engine results. Following their analogies with information retrieval where users' interests are seen as documents and communities as search engines, the relationship strength of each community, treated as a whole, with the ego user is considered as an importance weight attached to search engine.

Thus, we focus, at a first time, on how to formulate the strength of a community $c \in C$ with the profiled user from relationship strength of all its members. Note that, this stage considers only the ego-alter tie strength

whereas earlier steps (community detection and centrality measures) were based on alter-alter strengths.

For each community $c \in C$, its strength to the ego user, denoted Str_C, is computed by a combination between the number of links the community has with the ego (size of the community, denoted $|c|$) and the sum of those links strength (denoted $W(c)$). Each of them normalized by whether the total number of links or the total strength of all users in the egocentric network, see Equation (12) below. This definition results from the intuitive idea of considering the sum of weights between ego and all the community c members as its strength and an analogy we did with the degree centrality in weighted networks. We deem that the number of ties between the ego user and community c has a significant value to add to the tie strength (i.e. the presence of many ties might be considered) to measure the involvement of the community in the user's egocentric network.

Hence, for each community $c \in C$, its strength is formally defined as:

$$\begin{aligned} Str_c &= \frac{|c|^{(1-\gamma)} \times W(c)^{\gamma}}{|E'|^{(1-\gamma)} \times {W_T}^{\gamma}} \\ W(c) &= \sum_{v \in c} Wv, \quad W_T = \sum_{v \in V} Wv \end{aligned} \tag{12}$$

where $W_v \in E'$ denotes the tie strength between ego and node v and γ is a damping factor to relativize the importance of community size or strength. In the Equation above, a value of $\gamma = 1$ meets our basic idea to consider the sum of weights as structural score, while setting $\gamma = 0.5$ implies that the same importance is attributed to number of ties and tie weights in contributing to the centrality measurement. We describe in Section 5 the parametric study that enabled us to find the fittest value of γ and evaluate, as well, the effect of both tie weight and number of links to estimate the strength of the whole community.

Once the strength of communities with the ego user is computed, we describe in following the interest weight combination to derive the social dimension $S(u)$ of the profiled user.

As already mentioned, in the original unweighted approach, authors apply the LIN-CombMNZ function to compute the final weight of each interest in the social dimension. In this function, each score given by a search engine to the document is multiplied by a coefficient that relativizes its contribution in the final score. This coefficient varies between 1 and the number of merged systems. Thus, if n systems are merged, the highest score is privileged

and multiplied by *n*; the second is multiplied by $n-1$, etc. Finally, all these adjusted scores are summed to obtain the final score of the document.

For each interest $i \in I$, the combination should take into account two different aspects; the first is the strength of communities having *i* as interest with the ego user and the second is the weight of this interest in each community profile. In fact, if a community has a high weight for the interest *i*, the combination for all communities should, in turn, generate a high weight for *i*; which makes sense as this interest may be the affinity between the user and this community.

Hence, in our WLIN-CombMNZ function, for Weighted LINCombMNZ, we keep the same linear combination and multiply the score assigned to the interest in community *c* (using Equation (11)) by its relationship strength with ego user. Thus, both communities having highest score for the interest as aforestated, and those with strongest relations with the profiled user are privileged. Indeed, we deem that these communities are more relevant and might be assigned higher rank than others.

To do so, before multiplying by relationship strength Str_{Cj}, communities are ordered increasingly $(W(i, C_{j-1}) < W(i, C_j))$ according to their scores for the interest. If there are *n* communities in the user's egocentric network, the community which has the highest score for the interest is privileged and its score is multiplied by $n (j = n$ in Equation (13)), the second score is multiplied by $n-1, \ldots,$ the lowest weight for the interest is not privileged and multiplied by 1. Note that, unlike previous steps, communities are denoted C_j because, at this stage, computation implicates several communities at once (those concerning the interest *i*). Formally, for each interest, its combined weight *W*(*i*,*S*(*u*)) is calculated as below:

$$W(i, S(u)) = \sum_{j=1}^{n} W(i, C_j) \times j \times Str_{Cj} \tag{13}$$

where *W*(*i*,*Cj*) is the weight of the interest *i* in community *Cj*'s profile as in Equation (10) and Str_{Cj} is the weight associated to C_j as computed in Equation (12).

5 Experiment

In this section, the performance of our weight-aware approach is empirically evaluated. The profiled user is described by two dimensions [41]. The user dimension which consists on the user's real profile and the social dimension

which is built from his egocentric network using the two approaches (CoBSP and WeiCoBSP). Our evaluation is directed at determining the approach that provides a more relevant social profile (i.e. the closest to the user's real profile).

5.1 Dataset Description

To enable evaluation, a large set of real world egocentric networks was collected from DBLP computer science bibliography. We first survey several publicly available egocentric network datasets [20, 22, 31, 33], but find them not suitable for our experiments. None of them includes both users' attributes and link weights. [31] and [33], provide a Facebook weighted network originate from an online community for students at University of California, Irvine. However, user attributes are omitted for privacy concern. [20] collects users' ego networks from Facebook, Google+ and twitter and [22] collects ego networks with users' attribute and relationship types from LinkedIn. However, all these datasets are binary networks where ties are either present or absent without any associated weights. We also investigate methods to derive link's weight from available information in the datasets, but to no avail . Tie strength can be calculated by considering topological and/or semantic properties in the social network. In our study, available user attributes (semantic properties) cannot be used to predict tie strength since this latter is used in profiling process. Topological characteristics (number of common neighbors, neighborhood overlap...) cannot neither be used because (1) egocentric networks include only nodes directly adjacent to ego user and relationships among them. Thus, for each node we have only a partial knowledge of his neighborhood (i.e. only those relations with the ego related nodes) and (2) structural properties of the network (degree centrality for instance) are also used in our proposed profiling process. Thus, a new dataset was collected from DBLP to enable experiments. We construct co-authorship networks where an author's egocentric network is composed of his co-authors and the set of the weighted relationships between them. DBLP was selected because it is a publicly available database that provides a comprehensive list of research papers with several metadata (publication date, venue, authors...) so we can address the aforementioned aspects. Authors' profiles can be easily built by analyzing keywords from the titles of their publications [39, 41] whereas authors' links can be weighted by a measure of strength of their collaboration; for instance frequency of co authorship, duration of relationship, etc.

To avoid identical data sources and inherent biases in evaluation results and associated interpretations, the real and social dimensions are built from two distinct and not connected networks: DBLP and ResearchGate. Using the two approaches (WeiCoBSP and CoBSP), the social dimension of the ego user is built from his egocentric weighted network and co-author's publications in DBLP. On the other hand, the real profile, which serves as a ground truth, is derived from explicit interests the ego user filled in his ResearchGate profile independently of his publications in DBLP. The advantage of this demarche is twofold. We avoid the bias of using the same data (author's publications titles) to build the user and the social dimensions and we evaluate our proposed approach against realistic author's interests. In our experiments, we also consider authors who have a sufficient number of both co-authors and real interests indicated in ResearchGate, so that we get consistent and realistic data for evaluation. Indeed, for an author with small egocentric network, the community detection algorithm might return very few or no communities. This makes him not relevant to experiments on community-based approaches. In addition, we consider authors who have as many as possible interests in their ResearchGate profiles to build the most realistic user dimension for evaluation. Subsequently, we retain authors with at least 50 co-authors and that have more than six interests in their ResearchGate profile. The identification of these authors is conducted manually. A set of 75 egocentric networks was collected for this experiment. The studied authors have an average of 95 co-authors distributed between 50 and 214 co-authors, and an average of 19 interests indicated in their ResearchGate profiles. By analyzing these egocentric networks, we reached 7094 author's profiles in DBLP. Figure 2 shows the degree distribution (number of co-authors) of all the 75 studied authors. Only few authors have a high number of co-authors (more than 150) while many have a low number of co-authors (between 50 and 150); which suggests that degree distribution follows a power law as for all authors in DBLP [43] and social networks in general.

In another side, we analyzed the number of common keywords between the publication titles of the profiled user in DBLP and the interests filled in his ResearchGate profile. We do so to evaluate the compatibility of data, since different sources are used to build social and user profiles. Figure 3 shows the number of authors regarding keywords number in their ResearchGate profiles. The 75 authors' profiles of our dataset have between 7 and 56 interests. Figure 4 shows the percentage of ResearchGate keywords found in publications titles obtained from DBLP, sorted by this percentage. For almost all users, the majority of keywords in their ResearchGate profile also exists

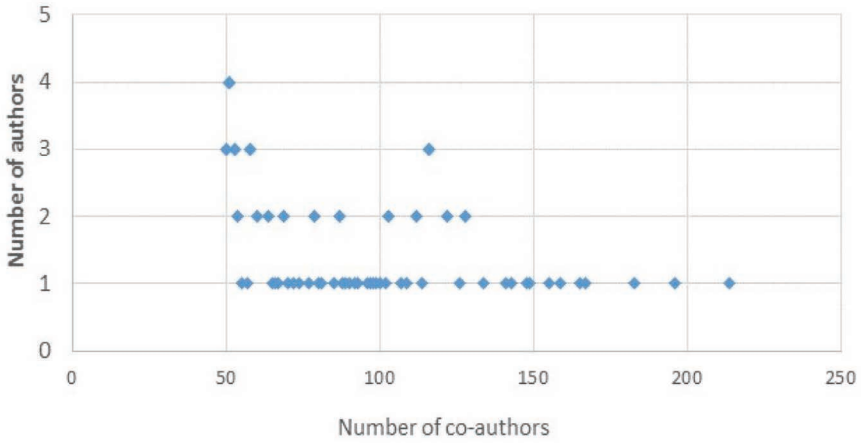


Figure 2 Distribution of the number of authors for each co-author (for all 75 authors studied in this experiment).

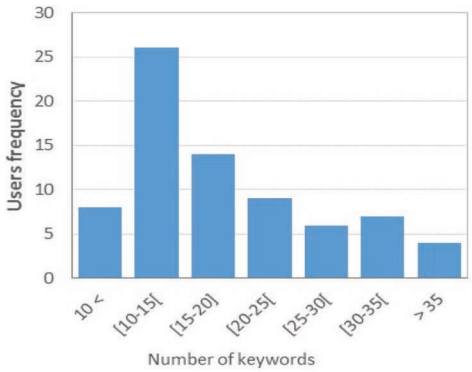


Figure 3 Number of users according to number of keywords in ResearchGate profiles.

in the DBLP publication titles. The average percentage of common keywords is 65.35% with a minimum value of 14.81% and a maximum value of 100%.

5.2 Egocentric Network and Profiles Construction

The whole DBLP dataset is available as one big XML file that may be too costly to process and parse (the file's current size is about 2.3 GB).[6] As we need only a few facts (list of co-authors and list of publications), we

[6]https://dblp.org/xml/

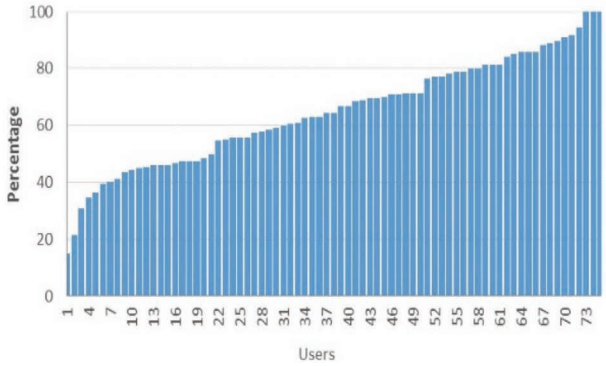


Figure 4 Percentage of users' ResearchGate keywords found in their DBLP publications titles.

make use of the DBLP provided web API (XML services). A comprehensive documentation of this API is provided in [21]. The example in Figure 5 illustrates data records for the author name "Walid Khaled HIDOUCI". The Figure shows three XML files returned when looking for the list of his co-authors (Figure 5(a)), the list of publications of this author (Figure 5(b)) and the details about one of its publications (Figure 5(c)).

The user dimension (real profile) is built from the explicit interests indicated in ResearchGate profile. Figure 6 gives an example of an author's profile in the ResearchGate social network where can be seen his function, affiliation as well as the explicit list of interests that the author filled in his profile.

The following sub-sections describe the process of building the weighted egocentric network from DBLP and the process of building authors' profiles from both DBLP and ResearchGate.

5.2.1 Weighted egocentric network building process

To build the weighted egocentric network of the selected author, we first look for relationships between his co-authors to connect them. Then, we determine the strength of each connection between a pair of authors (including those with the central author) by using the count attribute that specifies the number of shared publications.

The weight of the link between two authors is estimated on the basis of two factors:

- Frequency of co-authorship: authors that frequently collaborate should have a higher weight.

Figure 5 Sample of XML files returned by the DBLP XML API for the author Walid Khaled HIDOUCI.

- Total number of authored articles: to indicate how exclusive or nonexclusive the co-authorship relation is.

We compute the weight of the co-authorship link between two nodes u and v in the egocentric network (denoted W_{uv}) as the proportion of shared papers relative to the total authored papers of each one. Thus, weights are assigned directly in terms of number of collaborations, and inversely regarding other authors involved. With this metric, a relationship between a pair of nodes having few collaborations with other nodes is considered more important. For each $u,v \in V$, W_{uv} is computed by following Equation (14):

$$W_{uv} = \frac{2 \times N_{uv}}{N_u + N_v} \tag{14}$$

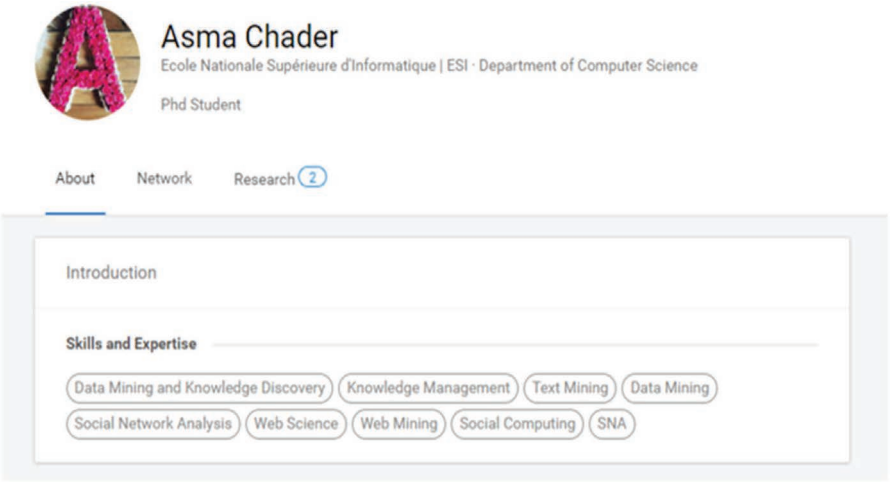


Figure 6 An author's explicit interests from his profile on ResearchGate.

where N_{uv} is the number of co-authored papers, and N_u, N_v represent, respectively, the total number of author's u and v publications.

5.2.2 Profiles building process

To build the social profile, publication titles are extracted using the DBLP XML API [21]. On the contrary, the real profile from ResearchGate are built manually since there is no API available to extract data from this platform.

Social profile construction

For each community in the co-authorship network, interests are detected by mining texts in publication titles of all its members. After collecting publication titles, keywords were analyzed to extract the interests using text-mining techniques. As a part of data pre-processing, we filter out empty words and apply a stemming algorithm to reduce vocabulary variability. As most publications are written in English, the *porter*-stemming algorithm [37] is used. The resulting terms are weighted relatively to their frequency. Finally, the terms from all publications are collected to constitute the interests set of the community.

Real profile construction

To build author's profile as a ground truth, we analyze his list of interests shared via ResearchGate. We first collect keywords from his profile and then

extract semantic units by applying the same text-mining techniques as in the construction of social interests. At last, each extracted element gets assigned a weight which reflects its frequency in the set of all found interests.

5.3 Evaluation Protocol

In order to validate our proposition, we apply our weight-aware method under a parametric study; which aims to assess the effect of tuning parameters involved in our calculation and deduce their appropriate values. Hence, experiments are done for different combination of values of parameters used in Equations (2), (8), (11) and (12), where the value of each parameter is ranged between 0 and 1.

Furthermore, to fairly compare our approach against the existing CoBSP and ensure that external variables to the profiling process do not alter the results of comparative study, the same community detection algorithm is applied for both approaches. The OSLOM algorithm [18] which performs on both weighted and unweighted networks is used.

Results are evaluated using the precision, the recall and the F-measure metrics as commonly done in related work [5, 25, 28, 29, 41]. The precision represents the proportion of relevant found interests (i.e. interests that truly belong to user's profile) in relation to the total number of found interests (Equation (15)). The recall (Equation (16)) represents the proportion of relevant found interests compared to the total number of real interests (user profile). Finally, the F-measure is the synthetic indicator that combines precision and recall via their harmonic mean (Equation (17)). Results will be presented by the average mean precision (respectively recall, F-measure) of all studied authors. In our experimentation context, for a given author, these evaluation metrics are computed as below:

$$Precision = \frac{N(I_{su})}{N(I_s)} \tag{15}$$

$$Recall = \frac{N(I_{su})}{N(I_u)} \tag{16}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

where $N(I_{su})$ is the number of relevant interests in the social dimension (i.e. also present in the real user profile), $N(I_s)$ is the total number of interests in the social dimension and $N(I_u)$ is the total number of interests in the user dimension (real profile).

Finally, let us note that the number of interests computed in the social dimension can be too large since we use text from publication titles. Thus, only top *X* interests for all built social profiles are considered. In fact, like in adaptive systems, we are interested only on the most relevant interests to the profiled user. Therefore, we utilize precision $P@5$, $P@10$, $P@20$ and $P@30$ respectively for the mean precision values at the first 5, 10, 20 and 30 returned interests. With respect to *n* profiled users (75 in this experiment), it will be calculated by:

$$P@X = \frac{\sum_{i=1}^{n} P_i@X}{n}, \quad P_i@X = \frac{N(I_{su})@X}{X} \tag{18}$$

where $N(I_{su})@X$ is the number of relevant interests in the social dimension when considering top *X* returned interests. Similarly, to compute recall at top *X* interests, we use the following Equation (19):

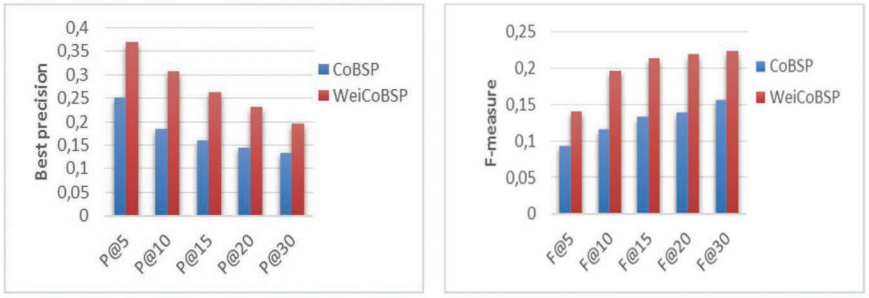
$$R@X = \frac{\sum_{i=1}^{n} R_i@X}{n}, \quad R_i@X = \frac{N(I_{su})@X}{N(Iu)} \tag{19}$$

5.4 Results and Discussion

In this section, we present the results of our evaluations. Three different experiments were conducted. In the first, we compare our proposed approach against the existing CoBSP with respect to the parametric study to evaluate the effectiveness of considering tie strength to infer more relevant user interests. The other experiments attempt to further understand the effect of tie strength, they were particularly interested in the contribution that each of ego-alter and alter-alter tie strength can have in the profiling process. More specifically, in the second experiment, the unweighted profiling process (i.e. the existing CoBSP) is applied on communities extracted by considering weighted egocentric networks. In so doing, the ego-alter weights are completely ignored and the outcomes of the process are solely based on the alter-alter strength. Conversely, the third experiment is based on ego-alter strength only, the communities are those extracted considering a binary version of the ego network (i.e. where all the ties with a weight greater than 0 are set to present) then our weight-aware version of CoBSP is applied.

5.4.1 General results

We first studied the performance of our weight-aware method and existing CoBSP under parametric study. From the results shown in Figures 7–9, we



(a) Best precision (b) Best F-measure

Figure 7 Comparison of the best results of WeiCoBSP and CoBSP approaches when varying top returned interests.

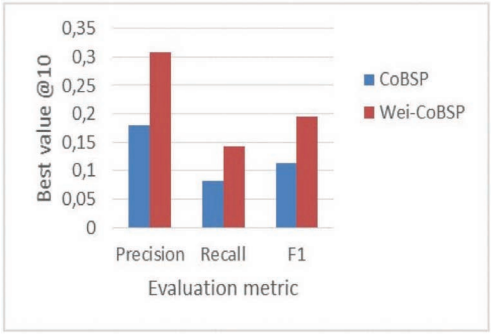


Figure 8 Comparison of best metric values considering top10 interests with best parameters for each.

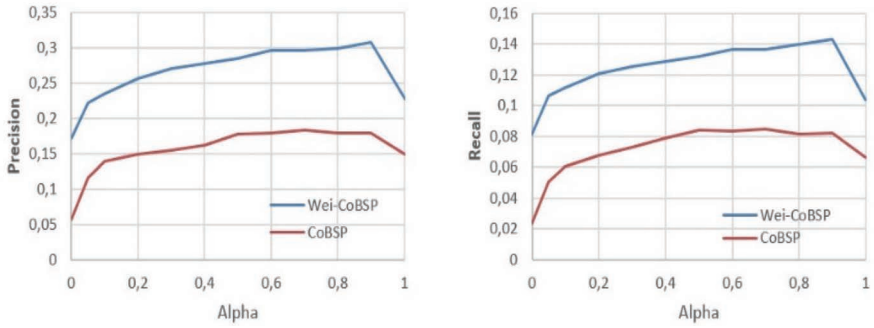


Figure 9 Comparison of average precision and average recall according to α for all users with best parameter combination ($\beta = 0.3, \gamma = 0.3$ for WeiCoBSP).

can observe that our method performs the best with significant improvement upon the baseline on all metrics . This improvement shows the effectiveness of our proposition and confirms our premise that strongest relationships may reveal more relevant information about the profiled user.

Figure 7 shows the overall performance comparison in terms of best precision and F-measure when varying the top *X* returned interests, $X \in \{5, 10, 20, 30\}$ and where each metric is averaged across all users. For our weight-aware method, these results are obtained when applying the optimum values of β and γ ($\beta = \gamma = 0.3$). From this figure, we can see that on both metrics and over all values of *X* considered WeiCoBSP significantly outperforms the existing CoBSP. The two approaches achieve their best precision of 37.07% and 25.07% respectively at the top 5 returned interest, *P*@5, while the best F-measure values are observed when $X > 15$ (see Figure 7(b)). Beyond this value ($X = 15$), the variation of F-measure is quite small. It is slightly increased with, for example, +0,005 between *F*@15 and *F*@20 for both CoBSP and WeiCoBSP. Regarding improvement, we can observe that the best results of the weight-aware process outperforms those obtained by CoBSP of 12.4 and 8% in terms of mean precision (*P*@10) and F-measure (*F*@20) representing an improvement of 48% and 57.72% (these are relative improvements) respectively.

Hereafter, all presented results are computed at the top 10 returned interests with respect to the ground-truth real profile. This value is observed to offer a good compromise between precision and recall as well as to ensure significant values of these metrics. In fact, for users having few interests indicated in their ResearchGate profile (e.g. less than 15, that represent significant portion (around 45%) of our dataset, see Figure 3) results @15 and above, can be less meaningful as the number of interests derived in the social profile is increased while the number of interests in the real profile remains less than 15. In following, we first compare our weight-aware approach with the existing CoBSP and then investigate WeiCoBSP specifically.

The results are first analyzed with respect to α values since this latter is the only parameter involved in existing CoBSP process. Figure 8 presents the comparison of best precision, recall and F-measure considering the top 10 interests. For the CoBSP approach, the best precision (0.18) is observed when $\alpha \in \{0.7, 0.9\}$ and the best recall (0.083) when $\alpha \in [0.5, 0.8]$. In comparison, our WeiCoBSP approach achieves its best precision of 0.308 when (α, β, γ) equals to (0.8, 0.3, 0.3) and (0.9, 0.3, 0.3) while its best recall value (0.143) is observed when $\alpha = 0.9$, $\beta = 0.3$ and $\gamma = 0.3$.

Figure 9 depicts the results of each process by the average value of precision (respectively recall) for all users (β and γ are still fixed at 0.3, their optimum value when considering all α values).

For all values of α, the proposed weight-aware method outperforms the CoBSP profiling process, represented by the blue curve, the highest results are obtained when $\alpha \in [0.5, 0.8]$ for CoBSP process and when $\alpha \in [0.6, 0.9]$ for WeiCoBSP, in which interval we observe an average gain of 0.119 (respectively 0.054) in terms of precision (respectively recall). These positive gain values show once again the benefits of considering relationship strength when inferring social profiles.

5.4.2 Parametric study

To illustrate the effect of varying different tuning parameters and deduce their most accurate values, we analyzed at a second stage the performance of our weight-aware profiling process WeiCoBSP with respect to β and γ. Note that these parameters are not involved in CoBSP calculations. This means that results of this latter will never vary whatever the values of β and γ; hence its representation as a straight line in different graphs throughout this section.

We first studied the effect of β parameter used in communities' structural score calculation to control the contribution of connections' number compared to their strength. We remind that when β is set to one of the benchmark values 0 or 1, the outcome is only based on one measure (only degree or only strength, respectively) while ranging β between 0 and 1 allows to consider both measures and positively value each of them; which implies that increments in number of links and strength increase the outcome. In contrast, a value of β above than 1 would negatively value the number of links. We are not interested at this latter case in our study since we deem important to leverage both number and strength of links. Yet, we are still testing the value $\beta = 1.5$ only as confirmation.

Figure 10 presents the results in terms of average precision and recall of WeiCoBSP process when β varies with respect to α values where our method achieves its best performance ($\alpha \in [0.6, 0.9]$) and by fixing γ to 1.0. (Results according to other values of γ parameter are also comparable, and can be seen in Figure 14 concluding the parametric study).

We observe that all WeiCoBSP precision curves follow roughly the same distribution. The optimal results are always achieved when both links number and strength are taken into account. Similar results in terms of recall are also observed as can be seen in the plot on the right side of Figure 10. Hence, we focus hereafter on precision rates. The best precision results, 29.73%

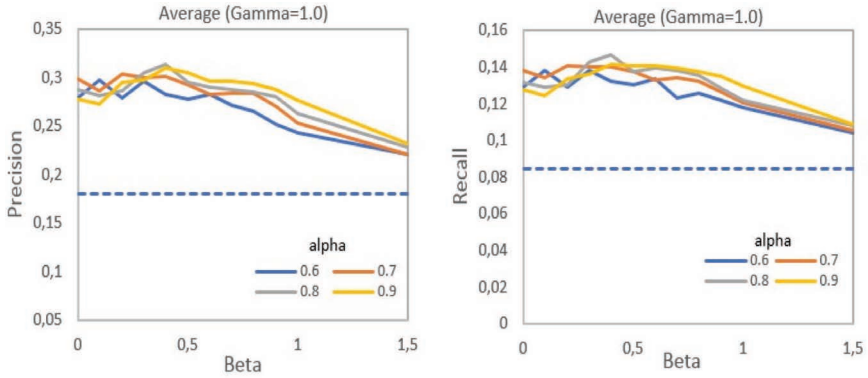


Figure 10 Comparison of the average precision (left) and average recall (right) according to β variation for $\alpha \in [0.6, 0.9]$ ($\gamma = 1.0$ for WeiCoBSP).

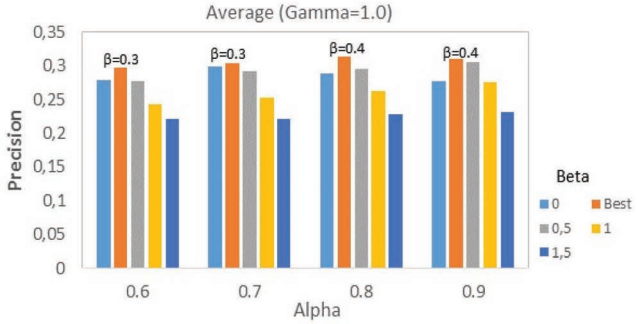


Figure 11 WeiCoBSP results according to different benchmark values of β, with $\alpha \in [0.6, 0.9]$ and $\gamma = 1.0$.

and 30.4% for $\alpha\{0.6, 0.7\}$ and 0.313, 0.309 for $\alpha \in 0.8, 0.9$ are obtained when β equals to 0.3 and 0.4 respectively, with an average gain of 8,04% (representing a substantial improvement of 35,64%) upon the worst observed results and a best (respectively average) gain of 13.3% (respectively 12.6%) compared to CoBSP process. This supports our hypothesis that number of ties (the presence of many ties) as well as ties with greater strength have a part to play in quantifying the involvement of communities in the egocentric network. Moreover, a considerable improvement is observed when setting β between 0.3 and 0.5, which implies that number of ties is relatively more important.

Figure 11 shows, particularly, results according to different benchmark values of β along with the best observed result (represented by the orange

bar in different groups of the chart with the corresponding β value indicated above) according to α values.

As expected, the precision decreases significantly when the number of links is not considered (i.e. when $\beta = 1$) and attains its worst rates if this latter is negatively valued (i.e. when $\beta = 1.5$). We also found that by setting β to 0.5, which attributes the same importance to number and strength of community ties, the WeiCoBSP method performs similarly as (slightly better than) when setting $\beta = 0$ (i.e. when strength is disregarded) for smaller values of α. On the other hand, greater α values showed a better performance when $\beta = 0.5$ and a decreasing difference in precision between benchmark values $\beta = 0$ and $\beta = 1$ results as α increases. For instance, the precision reaches 0.305 at $\alpha = 0.9$ with an insubstantial loss of 0.4% compared to the best observed result and a gain of 2.85% over benchmark values 0 and 1 where quite similar precisions of 27.73%, 27.6% are respectively observed. These last observations suggest that when the contribution of the structural score compared to the semantic one is quite small ($\alpha \in [0.1, 0.5]$), a β lower than 0.5 might be more suitable. Such a β will increase the importance of the number of links in structural score calculation. Conversely, when the contribution of structural score increases ($\alpha > 0.6$), number and strength of links tend to relatively have comparable impact. This can be explained as due to a loss compensated by the gain. In fact, when varying the parameter value, a certain number of users is not correctly profiled anymore, while there is a comparable amount of correctly profiled users at present.

In essence, we conclude that when computing structural score of communities, both number of ties and ties strength should be considered (and positively valued) with slightly more weight given to ties number as WeiCoBSP achieves exemplary results when β equals to 0.3 or 0.4.

Similarly, the variations of WeiCoBSP results was studied according to γ. This parameter controls the importance between the number of ties and associated weights in contributing to estimate the strength of each community to ego user when computing the final score of interests in the social profile (Equation (12), stage 4 of the algorithm). It is worth noting that the same benchmark values and associated interpretations as in in β parameter variation are applicable. We present the results of WeiCoBSP in terms of average precision, recall and F-measure according to γ (horizontal axis) when $\alpha \in [0.6, 0.9]$ (different curves) and β is fixed to its optimum value (i.e. $\beta = 0.4$) as shown in Figure 12.

Based on Figure 12, we observe that the curves are almost flat over the range of γ parameter value. This suggests that variation of γ, compared to β,

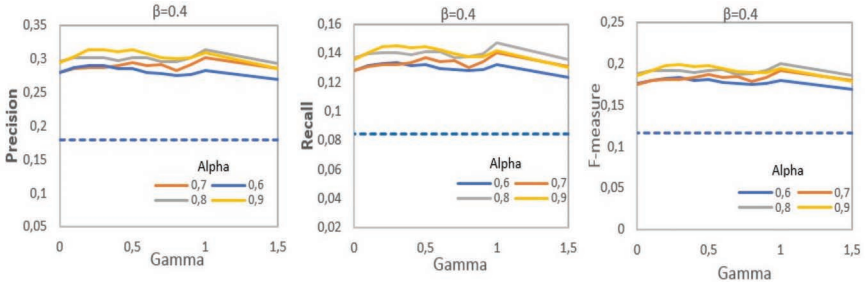


Figure 12 Comparison of the average precision (left), recall (Middle) and F-measure (right) according to γ variation for $\alpha \in [0.6, 0.9]$ ($\beta = 0.4$ for WeiCoBSP).

has relatively low impact over the effectiveness of our weight-aware process. We remind that in this experiment, we are not studying the relevance of community-ego connections in the WeiCoBSP process, but the contribution of number of links and their strength to characterize these connections.

As can be seen from the figure, the best precision and recall rates are generally reached when γ is set low where both number and strength are considered (with a higher contribution from links number) and when $\gamma = 1.0$ which disregards number of links in favor to strength. The worst rates are always observed when $\gamma = 1.5$ which negatively values the number of links. In more detail, WeiCoBSP achieves its best precision of 0.313 when $\gamma = 0.3$ and $\gamma = 1.0$ by setting α to 0.9 and 0.8 respectively, with an improvement of 9.81% ($\gamma = 0.3$, $\alpha = 0.9$) and 6.82% ($\gamma = 1.0$, $\alpha = 0.8$) over results achieved when $\gamma = 1.5$ (0.293 and 0.285). In terms of recall, higher improvements of 11.92% and 8.46% with respect to the same values of γ and α are observed.

Results according to different benchmark values of γ, presented in Figure 13, show less accurate performance when considering only links number ($\gamma = 0$) regardless α value and evaluated measure.

In this chart, we also represent results by setting γ to 0.3 since this value, along with $\gamma = 1$, produced the best performance. Looking more in depth, we observe, however, that $\gamma = 1$ outperforms $\gamma = 0.3$ in terms of improvement. Indeed, comparing best results when $\gamma = 0.3$ (i.e. for $\alpha \in 0.6, 0.9$) against those obtained by setting $\gamma = 1$ and vice versa for $\gamma = 1$ (i.e. $\alpha \in \{0.7, 0.8\}$), we found that results achieved by setting γ to 1 improves the ones of $\gamma = 0.3$ by around 2.1% in average. This behavior is also reported for lower α, notably when β takes its best values between 0.3 and 0.5.

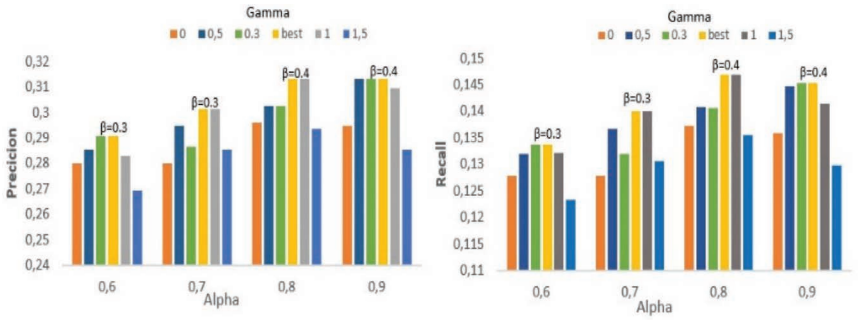


Figure 13 WeiCoBSP results according to different benchmark values of γ, with $\alpha \in [0.6, 0.9]$.

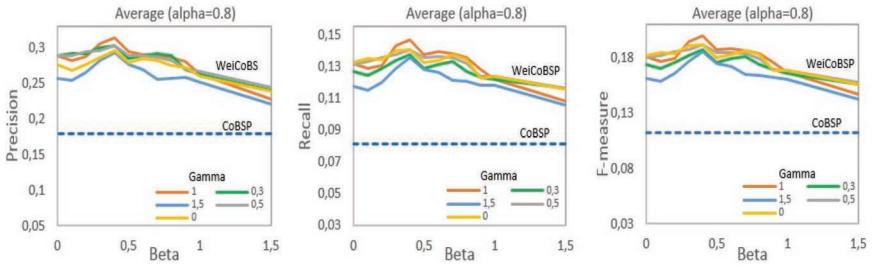


Figure 14 Comparison of average precision, recall and F-measure according to β and γ parameters when fixing $\alpha = 0.8$.

To conclude our parametric study, we plotted performance of our WeiCoBSP process according to β and γ parameters while fixing α value. Figure 14 presents the results according to β (horizontal axis), γ (different curves) when $\alpha = 0.8$, in terms of average precision, recall and F-measure.

We can globally conclude that both number and strength of links are valuable to characterize the communities in the egocentric network. Regarding community-community connections, β parameter used to compute communities' structural score seems to have more impact on WeiCoBSP performance. As already mentioned (and can be seen from Figure 14), best results are achieved when β is set relatively low $(0.3 = \beta < 0.5)$ regardless γ value. This suggests that number of links have higher contribution than strength. Although, links strength should not be neglected since considerable improvements are recorded over results setting β to 0. Regarding ego-community connections, even though γ parameter affects the performance, its impact remains relatively low compared to β one. Best results are observed when γ equals 0.3 and 1(represented respectively by orange and green curves in

the figure). These two values, even though corresponding to almost opposed settings, produced the most accurate results. However, setting γ to 1 gives relatively better performance, notably in terms of recall; which aligns with our previous observations from experiment where β parameter was fixed. This demonstrates the promising ability of strength to correctly characterize the ego-community connections.

5.4.3 Alter-alter strength: community structure impact

We have argued in the motivation section that ignoring link weights leads to not correctly depict the community structure of the egocentric network and, therefore, negatively affects the performance of profiling process.

To assess this, we conduct an experiment where we exclusively studied the impact of community structure on CoBSP prediction ability. Hence, we applied the unweighted profiling process on communities extracted from the weighted egocentric network; which we called WeiComCoBSP. The only difference with the existing CoBSP lies on the first stage (community detection), in which the OSLOM algorithm [18] is applied while taking into consideration the strength of ties. The remaining stages do not change. Moreover, the structural score of communities is computed based on links number only (as in the original process) to exclude any other effect of links strength. Note also that ego-alter connections are discarded here (all alters are equally significant to ego) which allows to particularly study the effect of alter-alter strength on social profile inference.

Figure 15 outlines the results of WeiComCoBSP process (represented by the orange curve in the plot) in terms of average precision and recall compared to CoBSP and WeiCoBSP, both illustrated in Figure 15, as the

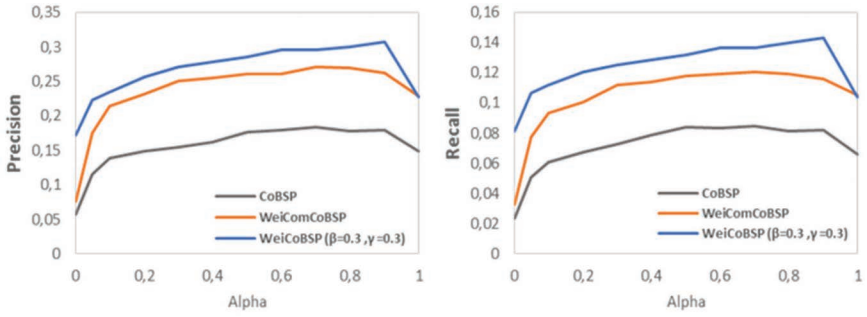


Figure 15 Comparison of WeiComCoBSP results in terms of precision (left) and recall (right) according to α, ($\beta = 0.3$, $\gamma = 0.3$ for WeiCoBSP).

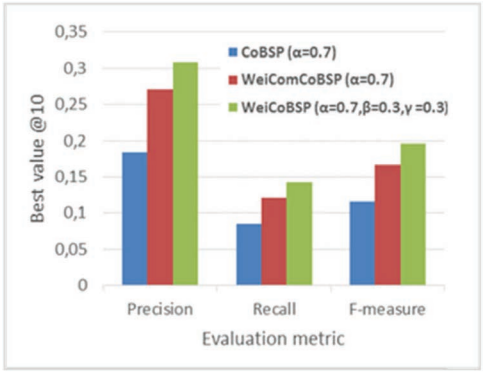


Figure 16 Comparison of best performance results for CoBSP, WeiComCoBSP and WeiCoBSP with $\alpha = 0.7$ ($\beta = 0.3$, $\gamma = 0.3$ for WeiCoBSP).

α parameter varies. For WeiCoBSP, these results are obtained with β and γ equal to 0.3. The best results achieved by each approach are illustrated in Figure 16.

As we would expect, link weights and ensuing community structure indeed affect the profiling process performance. WeiComCoBSP substantially outperforms the existing CoBSP independently of α values on all measures. We observe a significant improvement of 50.15% and 46.97% (both values are average) in terms of precision and recall, respectively. Based on Figure 16, presenting the best performance results, we observe that, like CoBSP, the best WeiComCoBSP precision and recall rates, 0.27 and 0.12 respectively, are achieved when $\alpha = 0.7$. They represent gains of 0.087 and 0.036 interpreting 47.10% and 42.52% respective improvements upon CoBSP results. This figure shows, moreover, that WeiCoBSP achieves the most accurate performance among the three approaches; it improves results obtained by WeiComCoBSP with 13.82% and 18.83% in terms of precision and recall respectively. Finally, as can be seen from Figure 15, WeiCoBSP achieved a higher improvement in terms of precision (compared to recall) upon CoBSP while, conversely, a higher recall improvement of 27.35% (against 20.89% for precision) is achieved upon WeiComCoBSP.

As follows from the figures shown above and related observations, ties strength may carry crucial information that influence the organization of the ego network. Hence, taking into consideration this strength allows a more accurate determination of communities and enhances the profiling process. However, despite this significant improvements upon CoBSP, WeiComCoBSP performance is still inferior compared to WeiCoBSP one.

This indicates that even ego-alter connections influence the social profile inference and supports our prior assumption that both alter-alter and ego-alter connections (with their strength) should be leveraged. The contribution of ego-alter connections will be the subject of the next experiment.

5.4.4 Ego-alter strength

As already mentioned, this last experiment aims at assessing the impact of ego-alter connections on the profiling process. These connections were totally discarded in the CoBSP approach based on an alter-alter model of ego networks. Thus, neither number of connections each community has with ego user nor their strength were studied. The combination between number and strength of links adopted in our approach allows us to evaluate both. Indeed, if the tuning parameter is set to 0, the outcomes are only based on links number which enables to study separately their effect. For other values, alternative outcomes based on both the number of ties and tie weights are attained. To evaluate ego-alter connections influence without considering other factors depending on strength (namely community structure and centrality measure in structural score), our weight-aware profiling process is applied on communities extracted from the unweighted egocentric network. The binary version of structural score (based on links number only) is also kept in order to exclude any effect of alter-alter strength. Herein, the first stages of the process remains unchanged. The difference between CoBSP and present experiment, so-called WeiEgoCoBSP, lies on the fourth (and last) stage where each interest gets assigned a final score according to its scores in different communities. This final score is computed based on the strength of each community (taken as a whole) to ego user.

Based on prior results of our parametric study, this experiment is conducted with respect to four values of γ parameter (used to control the importance between the number of ties and associated weights when computing the strength of communities): $\gamma = 0$ to study the effect of links number and $\gamma \in \{0.3, 0.5, 1.0\}$ that combine both number and strength of links (with 0.3 and 1 values, WeiCoBSP achieved its best performance). Figure 17 shows results in terms of average precision and recall of WeiEgoCoBSP when fixing $\gamma = 1$, CoBSP and WeiCoBSP when α varies. For WeiCoBSP, these results are obtained when β and γ equal to 0.3.

For all values of α WeiEgoCoBSP outperforms the unweighted CoBSP process except when $\alpha = 1$ where a slight loss of 0.8% and 0.42% in terms of precision and recall respectively is observed. WeiEgoCoBSP achieved an average improvement of 35.37% for precision and 42.79% in terms of recall

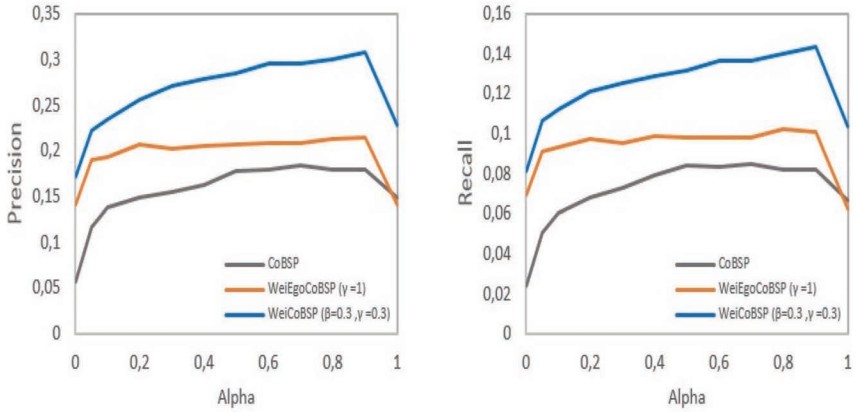


Figure 17 Comparison of WeiEgoCoBSP results in terms of precision (left) and recall (right) according to α, ($\beta = 0.3$, $\gamma = 0.3$ for WeiCoBSP, $\gamma = 1$ for WeiEgoCoBSP).

upon CoBSP results. These values ensure that ego-alter connections play also an important role in building the user social profile as alter-alter ones do. Once again, our weighted WeiCoBSP approach that leverages both types of connections has the best performance.

Based on Figure 17, we observe also that, apart from extreme α values (i.e. $\alpha = 0$ and $\alpha = 1$), results of WeiEgoCoBSP are quite constant. Besides, taking a closer look at results when $\alpha = 0$, which considers only semantic score of communities to describe interests in the social profile, a closer rates to those obtained by WeiCoBSP are observed with significant improvements (gains of 8.4% and 4.52% for precision and recall) over CoBSP results. This may be helpful in case of isolated communities in the egocentric network (i.e. communities having few or no connections with other ones). In fact, their structural score based on centrality measure could be very low which penalizes them. Such isolated communities as well as results according to size and density of user's ego network will be studied in future work. Regarding γ parameter variation, we drew globally similar conclusions as from the parametric study. Ties strength has the dominant influence on ego-community connections characterization. Corresponding plots as well as CoBSP results are depicted in Figure 18.

Overall, best results are achieved when $\gamma = 1$ which disregards number of links in favor to strength. As can be seen from the plots, better improvement upon other γ values are observed when $\alpha < 0.7$, beyond this value, results are almost comparable particularly when $\gamma = 0.3$ which achieved, as well, accurate performance in the parametric study.

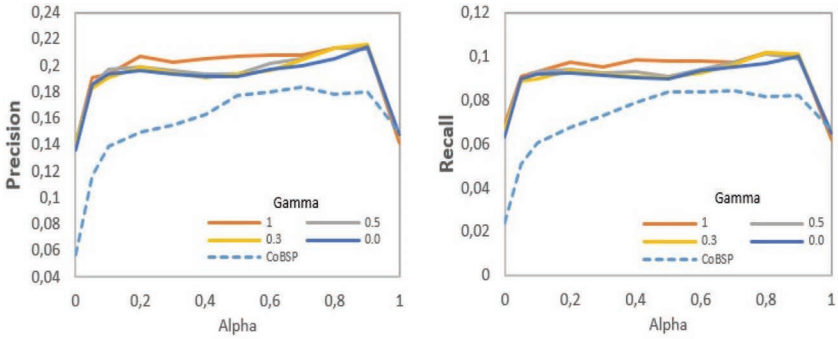


Figure 18 Comparison of WeiEgoCoBSP results according to α and γ, in terms of precision and recall.

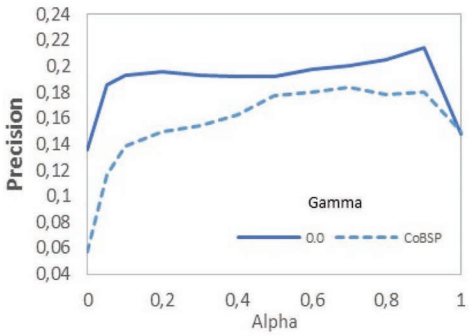


Figure 19 Comparison of WeiEgoCoBSP and CoBSP results when fixing γ to 0 (only links number considered).

Finally, comparing results of WeiEgoCoBSP when fixing γ to 0 (only links number considered) against CoBSP ones, illustrated in Figure 19, we can clearly see an important improvement when considering ego-alter connection even if without strength. Average gains of 3.54% and 1.86% over CoBSP in terms of precision and recall are observed.

5.4.5 Discussion

Throughout this results section, our proposed weight-aware approach showed a promising potential to accurately infer user social profile from his egocentric network. This demonstrates that relationships strength holds valuable information and proves how much is worthy leveraging them. This section discusses key findings and provides our final thoughts about the used method and the limitations.

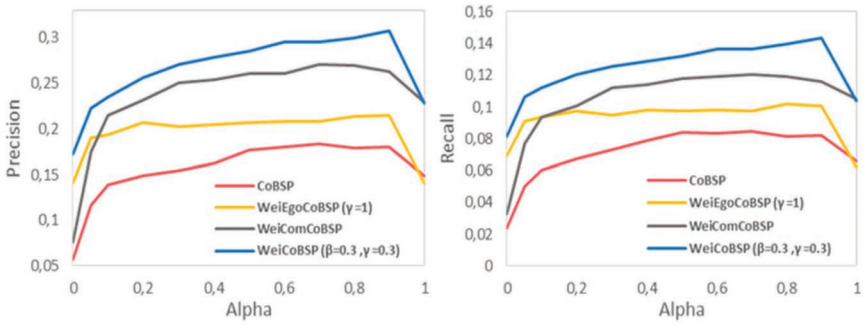


Figure 20 Comparison of all approaches results in terms of average precision (left) and recall (right) according to α ($\beta = 0.3$, $\gamma = 0.3$ for WeiCoBSP, $\gamma = 1$ for WeiEgoCoBSP).

To enable evaluation, the performance of our weight aware method and existing CoBSP were first studied under parametric study, and then we investigate specifically the ego-alter and alter-alter connections to further assess their influence. In the parametric study, the fittest values for β and γ parameters were examined. Both parameters are used to control the contribution of ties number compared to strength: the former in the communities' structural score and the latter in ego-communities strength calculations.

Figure 20 sums up the performance results for all approaches with respect to α values, for weight-aware methods (WeiCoBSP and WeiEgoCoBSP), the best configuration of γ and/or β parameters is used. As can be seen from the figure (and already shown in experiments), our weight-aware WeiCoBSP performs the best among the compared algorithms, followed by WeiComCoBSP based on weighted community structure then WeiEgoCoBSP which focuses on ego-community connections and finally the existing CoBSP. WeiCoBSP achieves this by leveraging both alter-alter (like WeiComCoBSP) and ego-alter (like WeiEgoCoBSP) relationships strength. Our results also suggest that, in the former, both number and strength of links should be considered to calculate structural score of communities (with higher contribution from number of links) while in ego-community connections, links strength has the dominant influence. Besides, when comparing results of the two last experiments (assessing alter-alter and ego-alter relationships influence), WeiComCoBSP generally outperforms WeiEgoCoBSP except for very low α values (≤ 0.05). This indicates that alter-alter connections through community structure have greater impact on profile building process than ego-alter ones. These results confirm as well our findings from

the parametric study that γ parameter effect was relatively low compared to β one.

On the basis of findings presented in our experiments, work on the remaining issues is continuing. As previously mentioned, we intend at a first place to explore the ability of ego-alter connections to alleviate the problem of isolated communities in the egocentric network where the community has very few (or no) connections with others. Besides, it would be interesting to study results according to the size and density of users' ego networks. The network density describes the interactions among user's alters (or coauthors in our experiments context) and may affect the performance of the community-based process especially in case of sparse networks, as reported by [41]. Moreover, co-authorship networks such as DBLP can be relatively less connected compared to online social networks; which reinforces our interest to study the proposed weight-aware approach against the existing CoBSP and investigate the contribution of (ego-alter and alter-alter) ties strength with respect to network density. Another observation one can make regarding our experiment results is the low rates of performance metrics. This can be explained by the fact that inferred interests are compared to optional data from ResearchGate profiles that may be obsolete or partly incompatible [30]. In addition, the keyword based method used to extract interest from the publication titles can lead to noisy and/or incompatible interests with those extracted from ResearchGate. Such methods are often criticized for lacking semantic information and failing to capture relationships among words [36]. For instance, a user may be interested in social network analysis in general, which is filled in his ResearchGate profile but extracted keywords from publications titles include "profiling" or "community" that are not obviously linked to social network analysis. To address this problem, a concept-based profile representation instead of keywords one can be adopted [36]. This model is increasingly used to represent user profile in microblogging networks (such as Twitter). Besides describing relationships between concepts, user interests can also be linked to pre-existing knowledge bases (e.g. WordNet,[7] DBpedia[8]) which can be useful for dealing with polysemy as well. Another interesting approach would be the closed frequent keywords sets model proposed in [39] to extract topics from publication titles. In this approach, authors form keywords-sets from substrings of the title's phrases and maintain

[7] https://wordnet.princeton.edu/

[8] https://wiki.dbpedia.org/

the relative ordering of keywords; which preserves, hence, the underlying semantics [39].

6 Conclusion and Perspectives

This paper brings forward a social profiling approach, so called WeiCoBSP, towards weighted egocentric networks. We suggest taking into account relationship strength to infer more refined and accurate interests and improve the effectiveness of the existing community-based approaches. Our starting point was the identification of a couple of issues ensuing from assumed binary model in these approaches, namely, their incapacity to distinguish most relevant people to the profiled user from others and to correctly depict the real community structure of user's ego network. In our work, we aimed to address these issues by leveraging strength of both ego-friend and friend-friend relationships. The former allows us to identify relevant people from whom to infer worthwhile interests while the latter, qualifying connections among people in user's neighborhood, enables extraction of the most realistic community structure in the egocentric network.

To validate our approach, an extensive empirical evaluation is performed on real world co-authorship networks (DBLP/ResearchGate). Experimental results show the ability of WeiCoBSP to infer user's interest accurately, improving greatly the unweighted CoBSP performance (57.72% improvement in terms of F-measure). Moreover, parametric study and experiments assessing separately ego-friend and friend-friend relationships strength led us conclude that the latter through community structure have a relatively higher impact on profile building process. Hence, evaluation findings confirm the efficiency of the proposed approach to address challenges stated in the motivation section, that is, how to change metrics and algorithms describing different steps of CoBSP process to deal with weighted networks? And how to choose relevant people in the weighted network from whom significant interests will be derived?

In future work, we would like to evaluate our approach in larger sets of egocentric networks from different social platforms (e.g., Facebook and LinkedIn). It will be also interesting to incorporate other factors or social features into the model to further enhance profiling accuracy. Finally, although the many research efforts studying user profiling in social networks, it is still very challenging to collaboratively model both user-generated content and social relationships, thus in a long term perspective, we plan to further explore this research direction.

References

[1] Ahmad Abdel-Hafez and Yue Xu. A survey of user modelling in social media websites. *Computer and Information Science*, 6(4):59–71, 2013.

[2] Sajid Yousuf Bhat and Muhammad Abulaish. Hoctracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1019–1013, 2014.

[3] Parantapa Bhattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghosh, and Krishna P Gummadi. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 357–360. ACM, 2014.

[4] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 131–140. ACM, 2013.

[5] Marie-Françoise Canut, Sirinya On-At, André Péninou, and Florence Sédes. Time-aware egocentric network-based user profiling. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 569–572. IEEE, 2015.

[6] Remy Cazabet, Frederic Amblard, and Chihab Hanachi. Detection of overlapping communities in dynamical social networks. In *2010 IEEE Second International Conference on Social Computing*, pages 309–314. IEEE, 2010.

[7] Asma Chader, Hamid Haddadou, and Walid-Khaled Hidouci. All friends are not equal: weight-aware egocentric network-based user profiling. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 482–488. IEEE, 2017.

[8] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768. ACM, 2010.

[9] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[10] Raissa Yapan Dougnon, Philippe Fournier-Viger, Jerry Chun-Wei Lin, and Roger Nkambou. Inferring social network user profiles using a partial social graph. *Journal of Intelligent Information Systems*, 47(2):313–344, 2016.

[11] Martin G Everett and Stephen P Borgatti. The centrality of groups and classes. *The Journal of Mathematical Sociology*, 23(3):181–201, 1999.

[12] Ying Fan, Menghui Li, Peng Zhang, Jinshan Wu, and Zengru Di. The effect of weight on community structure of networks. *Physica A: Statistical Mechanics and its Applications*, 378(2):583–590, 2007.

[13] Mark S Granovetter. The strength of weak ties. In *Social Networks*, pages 347–367. Elsevier, 1977.

[14] Gilles Hubert, Yannick Loiseau, and Josiane Mothe. Etude de différentes fonctions de fusion de systèmes de recherche d'information. *Le document numérique dans le monde de la science et de la recherche (CIDE'10)*, pages 199–207, 2007.

[15] David Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *ICWSM*, 2013.

[16] Xiangnan Kong, Xiaoxiao Shi, and Philip S Yu. Multi-label collective classification. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 618–629. SIAM, 2011.

[17] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[18] Andrea Lancichinetti, Filippo Radicchi, Jose J Ramasco, and Santo Fortunato. Findingt' statistically significant communities in networks. *PloS One*, 6(4):e18961, 2011.

[19] Sang Yup Lee. Homophily and social influence among online casual game players. *Telematics and Informatics*, 32(4):656–666, 2015.

[20] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, pages 539–547, 2012.

[21] Michael Ley. Dblp: some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500, 2009.

[22] Rui Li and Kevin Chen-Chuan Chang. Egonet-uiuc: A dataset for ego network research. *arXiv preprint arXiv:1309.4157*, 2013.

[23] Rui Li, Chi Wang, and Kevin Chen-Chuan Chang. User profiling in an ego network: coprofiling attributes and relationships. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 819–830. ACM, 2014.

[24] Zongqing Lu, Yonggang Wen, and Guohong Cao. Community detection in weighted networks: Algorithms and applications. In *2013 IEEE*

International Conference on Pervasive Computing and Communications (PerCom), pages 179–184. IEEE, 2013.

[25] Chao Ma, Chen Zhu, Yanjie Fu, Hengshu Zhu, Guiquan Liu, and Enhong Chen. Social user profiling: A social-aware topic modeling perspective. In *International Conference on Database Systems for Advanced Applications*, pages 610–622. Springer, 2017.

[26] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 251–260. ACM, 2010.

[27] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[28] Sirinya On-At, Marie-Françoise Canut, Andre Péninou, and Florence Sèdes. Deriving user's profile from sparse egocentric networks: Using snowball sampling and link prediction. In *Ninth International Conference on Digital Information Management (ICDIM 2014)*, pages 80–85. IEEE, 2014.

[29] Sirinya On-At, Arnaud Quirin, Andre Péninou, Nadine Baptiste-Jessel, Marie-Françoise Canut, and Florence Sèdes. Taking into account the evolution of users social profile: Experiments on twitter and some learned lessons. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–12. IEEE, 2016.

[30] Sirinya On-at, Arnaud Quirin, Andre Péninou, Nadine Baptiste-Jessel, Marie-Françoise Canut, and Florence Sèdes. A parametric study to construct time-aware social profiles. In *Trends in Social Network Analysis*, pages 21–50. Springer, 2017.

[31] Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, 2013.

[32] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[33] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155–163, 2009.

[34] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 430–438. ACM, 2011.

[35] Guangyuan Piao and John G Breslin. Exploring dynamics and semantics of user interests for user modeling on twitter for link recommendations. In *Proceedings of the 12th International Conference on Semantic Systems*, pages 81–88. ACM, 2016.

[36] Guangyuan Piao and John G Breslin. Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction*, 28(3):277–329, 2018.

[37] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[38] Gerard Salton and Robert Kenneth Waldstein. Term relevance weights in on-line information retrieval. *Information Processing & Management*, 14(1):29–35, 1978.

[39] Kumar Shubankar, AdityaPratap Singh, and Vikram Pudi. A frequent keyword-set based algorithm for topic modeling and clustering of research papers. In *2011 3rd Conference on Data Mining and Optimization (DMO)*, pages 96–102. IEEE, 2011.

[40] Patrick Siehndel and Ricardo Kawase. Twikime!: user profiles that make sense. In *Proceedings of the 2012th International Conference on Posters & Demonstrations TrackVolume 914*, pages 61–64. CEUR-WS. org, 2012.

[41] Dieudonne Tchuente, Marie-Francoise Canut, Nadine Jessel, André Péninou, and Florence Sèdes. A community-based algorithm for deriving users' profiles from egocentrics' networks: experiment on facebook and dblp. *Social Network Analysis and Mining*, 3(3):667–683, 2013.

[42] Jierui Xie and Boleslaw K Szymanski. Towards linear time overlapping community detection in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 25–36. Springer, 2012.

[43] Yi Zeng, Yiyu Yao, and Ning Zhong. Dblp-sse: A dblp search support engine. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 626–630. IEEE Computer Society, 2009.

[44] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 295–304. ACM, 2015.

Biographies

Asma Chader is a PhD student at Ecole Nationale Supérieure d'Informatique (ESI), Algiers, Algeria. She received her Engineering and Master's degree in 2015 from the same school. Her research interests include data mining, social network analysis and sentiment analysis.

Hamid Haddadou is a Lecturer at Ecole Nationale Supérieure d'Informatique (ESI), Algiers, Algeria. He is the head of Applied Mathematics team at the "Laboratoire de Communication dans les Systèmes Informatique", LCSI (Laboratory of Communication in Computer Systems). He had his PhD, and Magister degree in Mathematics at TheăUniversity of Science and Technology -ŰăHouari Boumediene USTHB (Algiers, Algeria). His research interests include mainly networks modelling, image processing and multiscale mathematic modelling.

Leila Hamdad is a Lecturer at Ecole Nationale Supérieure en Informatique (ESI), Algiers, Algeria. She is member of the Laboratoire de Communication dans les Systèmes Informatique, LCSI, ESI (Laboratory of Communication in Computer Systems) in Applied mathematics team. She had her PhD on Computer Science in the same school and a Magister degree in Mathematics at TheăUniversity of Science and Technology –ăHouari Boumediene USTHB (Algiers, Algeria). Her topics of interest are related to data mining, machine learning, spatial statistics and parallel computing.

Walid-Khaled Hidouci is currently Associate Professor at Ecole Nationale Supérieure dŠInformatique (ESI), Algiers, Algeria since 1993. His areas of interest mainly concern database systems, data structures, operating systems and parallel programming. Since 2010, he has been leading the "Advanced Databases" (BDA) team at the "Laboratoire de Communication dans les Systèmes Informatique", LCSI, ESI (Laboratory of Communication in Computer Systems).