ECIS 2015 Research-in-Progress Papers ECIS 2015 Proceedings

Spring 5-29-2015

A Community-based Approach for Improving the Profile Information Quality in Social Networks

Diana Hristova
University of Regensburg, diana.hristova@wiwi.uni-regensburg.de

Jian Ma
City University of Hong Kong, isjian@cityu.edu.hk

Recommended Citation
Hristova, Diana and Ma, Jian, "A Community-based Approach for Improving the Profile Information Quality in Social Networks" (2015). *ECIS 2015 Research-in-Progress Papers.* Paper 18.
ISBN 978-3-00-050284-2
http://aisel.aisnet.org/ecis2015_rip/18

A COMMUNITY-BASED APPROACH FOR IMPROVING THE PROFILE INFORMATION QUALITY IN SOCIAL NETWORKS

Research in Progress

Hristova, Diana, University of Regensburg, diana.hristova@ur.de[1]

Jian Ma, City University of Hong Kong, isjian@cityu.edu.hk

Abstract

Social networks are becoming increasingly popular and also increasingly valuable platforms. However, since their value is strongly influenced by the information quality of user profiles, assuring the correctness, currency and completeness (3C) of profile information is important for their success. In this paper we present an integrated approach for improving the 3C in social networks. Our approach is based on the idea that, if a user is active in a community which differs from the information community determined by her/his profile information, then there is a quality deficiency which needs to be improved. To identify the exact information quality problem, past community structure is considered. Thus, our approach is applicable to many different social networks and also considers the temporal dynamics of the network. Social network companies can apply it to achieve high-quality profile information, which is essential for many applications (e.g. head hunting on LinkedIn). We demonstrate this contribution, by applying it to the research social network ScholarMate. The initial results show that the activity and information communities of the user do not always overlap and that our approach effectively addresses information quality problems in real-world social networks.

Keywords: Information Quality, Social Networks, Currency, Completeness, Correctness, Community.

1 Introduction and Background

In the past years and with the development of mobile technologies, the popularity of social networks (SN)[2] has increased tremendously (Statista, 2015) and so has their value. In 2012, Facebook held its IPO for a total valuation of more than $100 billion (Nasdaq, 2012). Companies use SN for many different purposes such as market research, or recruitment (Hoffman and Fodor, 2010; Heidemann *et al.*, 2012) and thus the investments in SN activities have rapidly increased in the last years (ExactTarget, 2014). However, the success of these investments and thus the value of SN is threatened by the low information quality (IQ) of profile information (Pike *et al.*, 2013). According to a survey by Breitbarth (2012), only 50.5 % of the participants described their LinkedIn profile as complete (LinkedIn, 2012). Similarly, Abel *et al.* (2010) found that the average Tweeter user completed his/her profile only up to 48.9 %. Moreover, in a survey by Emedia (2007), 31% of the respondents stated that due to security reasons they provide fake information about themselves in SN (cf. Consumer Reports, 2012). Also, it was estimated that about 4% of the Tweeter accounts are fake (DeMicheli and Stroppa, 2013; cf. also FACEBOOK INC., 2014). Finally, according to another survey, only 44% of the respondents always keep their social media profile up-to-date (Software Advice, 2014).

[1] This research was funded by DAAD Doktorandenstipendium.

[2] We mean by a social network the whole online social network.

These IQ problems influence the value of SN negatively (Krombholz *et al.*, 2012). Chen (2013), states that "…the level of self-disclosure…" (p. 661) (i.e. completeness of information) is important for the SN value and for the success of personalised user services and advertising (cf. also Xu *et al.* (2013) and Chen and Sharma (2015)). Moreover, correct profile information is considered to be critical for the success of marketing initiatives (Alt and Reinhold, 2012; Xu *et al.*, 2013). Finally, current profile information is vital for the success of recruitment initiatives in professional SN (Probst and Görz, 2013). For example, many professional SN such as Xing (XING, 2013) offer a premium service, which is especially suitable for recruiters and which success is threatened by the low currency of profile information (Probst and Görz, 2013). We define profile information as any field in a SN profile which the user can fill-in upon registration and update afterwards. This includes information such as birthday, education, contact details, or interests (cf. Amichai-Hamburger and Vinitzky, 2010), but excludes any content generated or received by the user such as likes, groups, events, or news feed.

IQ is a multidimensional concept (Wang and Strong, 1996) and the dimensions completeness, currency and correctness, which we call the 3C, are among the most important ones (Agarwal and Yiliyasi, 2010; Lee *et al.*, 2002). In the IQ literature, completeness represents weather a certain attribute value is missing or not (Cai and Ziad, 2003; Batini and Scannapieco, 2006; Even and Shankaranarayanan, 2007). Correctness, which is also often referred to as accuracy (Batini and Scannapieco, 2006; Even and Shankaranarayanan, 2007; Fisher *et al.*, 2009), is defined as the degree to which an entered attribute value corresponds to its real-world counterpart at the time of entrance. Finally, currency is defined as the degree to which a correctly entered (or updated) attribute value "…*still* corresponds to the current value of its real-world counterpart…" (Heinrich *et al.*, 2009, p.5) at the time of analysis. While the definition of completeness is straightforward, currency and correctness are related issues (Heinrich and Klier, 2015). Indeed, violating both dimensions has the same implications, but different causes. An outdated attribute value was correct at some point in the past, while an incorrect attribute value was never correct. In the following we define the 3C for profile information in SN. Since profile information may consist of multiple attribute values (e.g. interests), the definitions apply also to a set of attribute values.

Definition 1: A (set of) attribute value(s) in a user's profile is ***correct and current*** *if it/they coincide with the real-world (set of) attribute value(s) of the user at the time of analysis. A (set of) attribute value(s) in a user's profile is/are* ***incorrect*** *if it/they was/were incorrect at the time of entrance. An attribute value in a user's profile is* ***outdated*** *if it/they was/were correct at the time of entrance, but do not coincide with the real-world (set of) attribute value(s) of the user at the time of analysis.*

Definition 2: A (set of) attribute value(s) for a given attribute in a user's profile in a SN is/are ***complete****, if no relevant attribute value(s) is/are missing from the set.*

There is a large body of behavioural research literature devoted to identifying the sources of 3C problems in profile information. One of the main reasons for users to provide both incomplete (e.g. no education information) and/or incorrect information (e.g. wrong name) is that they have privacy concerns (Son and Kim, 2008; Krasnova *et al.*, 2009; Abdesslem *et al.*, 2012; Krombholz *et al.*, 2012; Chen, 2013). Such concerns may be due to the use of their information by third parties (e.g. for marketing or recruitment) or because of fear of possible harassment (Krasnova *et al.*, 2009). A second reason is that some users use SN to present themselves in a better light than in reality (Krasnova *et al.*, 2009). For example, only positive information may be revealed or the profile information may be exaggerated as opposed to reality (e.g. higher grade). In addition, the completeness of profile information is affected if users are too introverted to provide information about themselves (Chen, 2013) and the correctness is influenced by the intentional generation of fake profiles. The latter is done, for example for gathering data, harassing real users, or even for creating false business reputation (Krombholz *et al.*, 2012). Finally, low currency profile may be due to the lack of social pressure (i.e. the number of profile visitors) to keep the information up-to-date (Probst and Görz, 2013).

To improve 3C in SN, these approaches propose developing practices for changing the user's attitude. For example, campaigns that encourage openness may address introversion (Chen, 2013; Chen and

Sharma, 2015). Also educating users about security risks and introducing better security features may address privacy concerns (Son and Kim, 2008; Chen, 2013; Chen and Sharma, 2015). However, this can be very time-consuming, and also difficult to implement, as it requires different solutions for the different factors influencing low IQ. To address these issues, we develop an automated approach for improving all 3C in SN profile information. We use the community structure of the SN, where a community is defined as a set of at least two nodes that have similar characteristics at a certain point in time (Fortunato, 2010). Our approach extends ideas from other fields, not directly focusing on the 3C.

The first C, correctness, can be connected to the literature regarding the identification of fake or Sybil user accounts. Most approaches (Cao *et al.*, 2012; Fire *et al.*, 2014) are based on the idea that fake users tend to be friends with other fake users instead of with real users and thus there are very few relations between fakes and non-fakes. The structure of these relations can be used to identify fake and thus incorrect profiles. For example, Fire *et al.* (2014) use common friends, “same family”-information, common activities such as posts and tags, and other indicators on Facebook to determine the strength of the connection between two users and thus to distinguish fakes from non-fakes. The second C, currency, is related to the literature on link prediction in dynamic SN. A link can be any kind of a relationship such as friendship, common information, or common activities and the link prediction methods identify such potential relationships. Their main idea is that nodes which share similar neighborhoods today (e.g. a common friends) are more likely to be connected in the future (Lu *et al.*, 2010; Bliss *et al.*, 2014). For example, Bliss *et al.* (2014) predict the future reciprocal replies between Tweeter users by using current neighborhood structures and user characteristics such as word similarity. Soundarajan and Hopcroft (2012) and Li *et al.* (2014) go one step further by considering the community structure of the SN in addition to the neighborhood characteristics. The third C, completeness, is related to the works dealing with the identification of missing profile information. One such approach (Mislove *et al.*, 2010) uses the common profile information (e.g. same major, college, year for students) and the friends’ network in Facebook to predict missing attribute values, by assuming that users are friends with users who have the same attribute values. Another related work is the one by Li *et al.* (2012) who develop an approach for determining the missing home location of Tweeter users. They use the people followed by the user, by assuming that people follow people who live close to them. The literature on link prediction discussed above can also be applied to identify missing attribute values (Soundarajan and Hopcroft, 2012).

The presented approaches are based on the idea that common neighborhoods or communities as well as common information should be used to identify 3C problems in SN. We borrow this idea by comparing the user’s community with respect to her/his profile information (e.g. interests) with the community defined by her/his activities (e.g. shares likes, etc.). In case they differ, there should be a violation of one of the 3C. Due to currency, our approach also addresses the temporal dynamics of the SN. We contribute to the literature by presenting an automated approach for all 3C, which can be applied to a SN that provides some activity functionality. In the next section we present our approach which is evaluated in Section 3 based on the research SN ScholarMate (www.scholarmate.com). Finally, in Section 4 main conclusions are drawn and paths for future research are discussed.

2 Methodology

2.1 Basic concepts

We begin by introducing some basic concepts, each of which is time-dependent as we consider the dynamics of the network. Each user in the SN is represented by a node and denoted by $N_j, j \in \{1, \dots, n_t\}$ where n_t is the number of users in the network at time t. There are three types of edges in the SN representing the different relationships between the users. The first edge type is an undirected *friend edge* which represents a friendship between two users. The second edge type is the undirected, attribute-specific *information edge*. Its label is the (set) of common attribute value(s) between two friends for a given attribute and at a certain point in time. For example, if two friends share the interests $I1$ and $I2$, then the *information edge* between them will be assigned the label $\{I1, I2\}$. The third edge type is the weighted, directed, attribute-specific *activity edge* which, as opposed to the *friend* and *information edge*, is determined over a period of time. Its weight is the total number of activities (e.g. shares, likes, invitation for events), for a given attribute (e.g. interests), of one user on the profile of one of her/his friends during the considered time period. We exclude private communication from the activities (i.e. messages), as it cannot be automatically extracted and has been shown not to bring much additional information (Jones *et al.*, 2013). The attribute for a certain activity is identified with text classification methods such as rule-based or SVM classifiers (Aggarwal and Zhai, 2012). For example, if the user liked a post of one of her/his friends about football, this will be classified with the attribute "interests". Thus, we also exclude activities without text (e.g. pictures without description). The *activity edge* is directed, because a user may be very active on the profile of another user, but the reverse must not be true. In order to determine an appropriate weight for the activity edge, we need to consider the quality of the relationship between the users. In particular, if two users are very close to each other, then they may automatically be more active on each other's profile, regardless of their profile information and this will bias the estimations about the profile IQ. To account for this, we consider the literature regarding the measurement of the tie strength between two users (Xiang *et al.*, 2010; Jones *et al.*, 2013; Luarn *et al.*, 2015). According to it, both profile similarities and common activities influence the tie strength in SN, where the former has been empirically shown to be non-significant (Jones *et al.*, 2013; Luarn *et al.*, 2015). Thus, we account for the tie strength, by dividing the number of activities for a given attribute over the total number of activities (regardless of the attribute) of the user on the profile of the other user.

Based on these edge types, we define *information* and *activity communities*. An *information community* consists of a set of nodes which are connected with *information edges* for the same (set of) attribute values of an attribute. It is denoted by $I^{S_m(H)}{}_t, m \in \{1, .., k_t(H)\}$ where $S_m(H)$ stands for the (set of) attribute value(s) of the attribute H which are shared among all users in the community, t is the point in time and $k_t(H)$ is the total number of *information communities* in the network for the attribute H. If the attribute H takes only one value (e.g. the city of residence), then one user will belong to only one *information community* for H, otherwise *information communities* for a given attribute may overlap. *Activity communities* are user-specific and attribute-specific and are denoted by $A_{t,\Delta t}(N_j, H)$ for the user N_j and the attribute H. They consist of the friends of the user on whose profile s/he was particularly active regarding the attribute H over the time period $[t - \Delta t, t]$ where $t - \Delta t$ is no earlier than the time of registration of the user in the SN. To derive the *activity community* we only consider the *activity edges* originating at the user and ignore the *activity edges* ending at the user which results in an undirected graph. Based on these concepts, in the next subsection we present our approach.

2.2 Improving 3C in SN

To illustrate the main idea of our approach, we consider the three cases in Figure 1 where we omitted the labels and the weights of the edges for simplicity. The user of interest is represented by the node N_1 and in all three cases belongs to the *information community* $I^{S_1(H)}{}_{t_1} = \{N_1, N_2, N_5\}$ (e.g. $S_1(H)$ is the same set of interests, H stands for the attribute "interests"), but to different *activity communities*. In Figure 1 a), which represents the present situation, the *activity community* $A_{t_1,\Delta t_1}(N_1, H) \;= \{N_3, N_4\}$ of the user does not overlap with her/his *information community*. This suggests that the attribute value(s) $S_1(H)$ in the profile of N_1 do(es) not coincide with the real-world attribute value(s) at t_1 and is/are thus either incorrect or outdated (cf. Definition 1). For example, if a user shares the same profile interests with users on which profile s/he is not active, then s/he may either have changed the interest or never have had them. To determine the exact IQ violation, we need to consider the past community structure. If in the past the *information* and *activity communities* overlapped, as presented in Figure 1 b), then the information is outdated and thus currency is violated. If this was not the case, then the information is incorrect and thus correctness is violated.

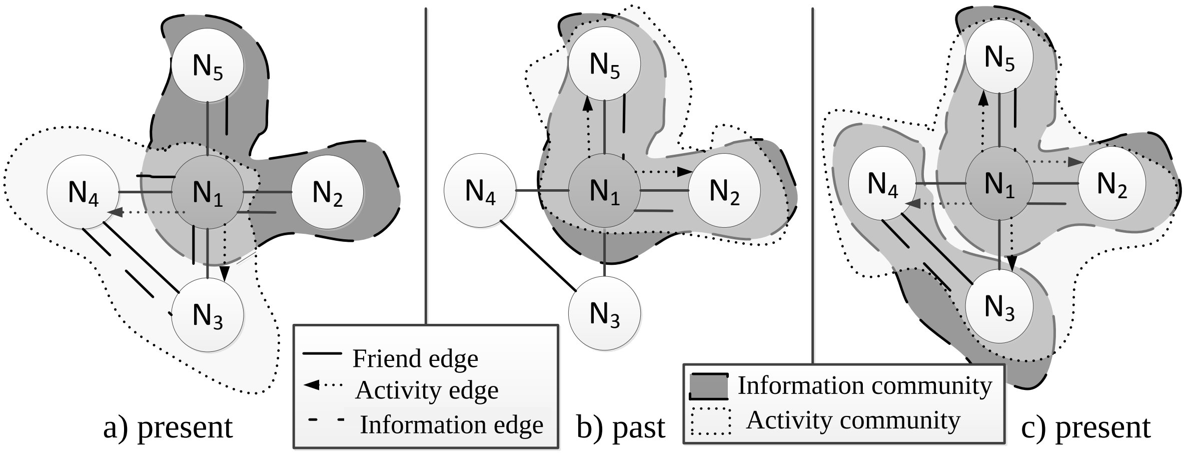


Figure 1. Possible information and activity communities

Here arises the question of how to define “the past”. To answer this question we consider the definition of currency from the IQ literature discussed above. According to it, currency of a (set of) attribute values depends on three points in time: the time t_0 at which the (set of) attribute value(s) was entered (or updated), the time $t_0 + \Delta t$ at which the (set of) real-word attribute value(s) changed, and the time t_1 at which the analysis takes place. Since t_1 is known and t_0 can be determined based on the activity log provided by most SN (e.g. Facebook, Xing), only determining Δt is an issue. To resolve it, we can use existing literature. For example, in the literature on link prediction (Nguyen *et al.*, 2011; Bliss *et al.*, 2014), standard time intervals are used (e.g. one day) and the data for each time interval is iteratively analysed. In our case this implies comparing the past and present *activity communities* for each time interval from t_0 to t_1 with the corresponding *information community*. However, this may be too inefficient for realistic applications. An alternative approach would be to determine the most probable point for $t_0 + \Delta t$. This can be done based on the IQ literature, for example by using historical data or expert estimations (Heinrich and Klier, 2015). In the following we will denote the present activity community of node N_1 by $A_{t_1,t_1-(t_0+\Delta t)}(N_1, H)$ and the past activity community by $A_{t_0+\Delta t,\Delta t}(N_1, H)$.

As opposed to Figure 1 a), in Figure 1 c), which also represents the present situation, the *activity community* of the user $A_{t_1,t_1-(t_0+\Delta t)}(N_1, H) = \{N_2, N_3, N_4, N_5\}$ overlaps with her/his *information community*, so there is no violation of currency and correctness. However, the *activity community* also overlaps with an *information community* to which the user **does not** belong ($I^{S_2(H)}{}_{t_1} = \{N_3, N_4\}$). This implies that there are friends of the user on whose profile s/he is active (regarding the given attribute), who share common information between each other regarding this attribute and do not share this information with the user. Thus, it is natural to assume that the common attribute value(s) $S_2(H)$ between the nodes N_3 and N_4 may be missing from the profile of N_1 (cf. Definition 2). To make the idea clearer, consider again the example with the interests of the user. If user N_1 is very active (regarding the attribute “interests”) on the profiles of users N_3 and N_4 who have the same interest, then it is natural that user N_1 also shares this interest. To sum up, correctness and currency are violated if the user’s *information community* does not overlap with her/his present *activity community*. Completeness is violated if the user’s present *activity community* overlaps with an *information community* to which s/he does not belong. In addition, more than one IQ dimension can be violated as presented in Figure 2 a). There the user’s *activity community* does not overlap with her/his *information community* (correctness or currency), but overlaps with the *information community* of her/his friends (completeness).

Based on the above ideas, correctness and currency are validated only by considering the communities of the user, while completeness is validated by additionally considering the communities of the user’s friends. This implies that an important condition for the reliable validation of completeness is that the profile information of the user’s friends satisfies the 3C. However, this must not be the case as presented in Figure 2 b). In it, the profile information of the users N_3 and N_4 is either incorrect or outdated and it is also incomplete. However, since the claim that the profile information of user N_1 is incomplete is based on the fact that the profile information of both N_3 and N_4 is correct, current, and complete, we cannot make a statement regarding the completeness of the profile information of N_1. To solve this issue, before improving the completeness of the profile information of N_1, we first need to improve the 3C of the profile information of the users N_3 and N_4. However, the completeness of the profile information of N_4 cannot be improved unless N_6 and N_7 satisfy the 3C. Following this line of thoughts, we traverse the network until we reach a set of nodes which satisfy the 3C, improving the correctness and currency of all the nodes on the way. In Figure 2 b), these nodes are N_8, N_9, N_{10}, N_{11}. Then we improve the completeness of the nodes N_3 and N_6 and recalculate the community structure, since the profile information changes and so do communities. Afterwards we do the same for the next level of nodes and so on until the *information community* for improving the completeness of N_1 consists only of nodes satisfying the 3C. Based on this, we define for each of the 3C a property for validating it. In all cases an overlap must consist of at least two nodes including the user.

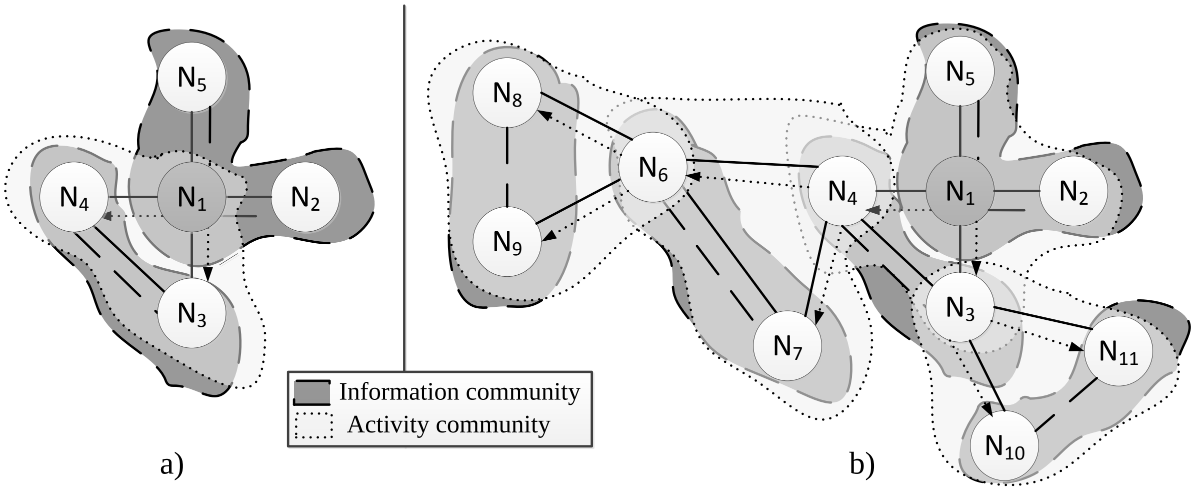


Figure 2. a) Violation of more than one IQ dimension; b) Network aspect of completeness

Property 1 (Outdatedness): A node N_j is said at time t_1 and for the time span Δt to contain **outdated** attribute value(s) $S_m(H)$ of the attribute H, if the user belongs to an *information community* $I^{S_m(H)}{}_{t_1}$, which does not overlap with the present *activity community* $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$ and which existed in the past and overlapped with the past *activity community* $A_{t_0+\Delta t,\Delta t}(N_1,H)$ for the attribute H.

Property 2 (Incorrectness): A node N_j is said at time t_1 for the time span Δt to contain **incorrect** attribute value(s) $S_m(H)$ of the attribute H, if the user belongs to an *information community* $I^{S_m(H)}{}_{t_1}$, which does not overlap with the present *activity community* $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$ for the attribute H and the attribute value(s) $S_m(H)$ is/are not outdated.

Property 3 (Incompleteness): A node N_j is said at time t_1 for the time span Δt to contain **incomplete** information for the attribute value(s) $S_m(H)$ of the attribute H, if the user's present *activity community* $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$ for the attribute H overlaps with an *information community* $I^{S_m(H)}{}_{t_1}$ to which the user does not belong. The nodes in the *information community* $I^{S_m(H)}{}_{t_1}$ must contain correct, current and complete information regarding the attribute values in $S_m(H)$.

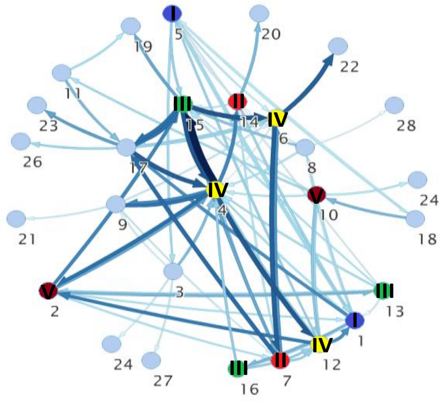
Based on these properties, Figure 3 presents our algorithm for improving 3C in SN. We assume that the *information* and *activity communities* were already identified by an approach for community detection (Fortunato, 2010; Aggarwal, 2011). The first five steps of the algorithm consist of checking whether the profile information of the user satisfies the 3C and removing the attribute value(s) which are incorrect and/or outdated. Steps 6-10 deal with improving the completeness of the attribute value(s). In Step 7 we start at the node of interest and traverse the network until a node is reached which is incomplete according to Property 3, by improving the currency and correctness of all nodes on the way. Then, the completeness of this node is improved and duplicates are cleared (Step 8). If the attribute cannot have more than one attribute value in the profile (e.g. city of residence) and this is the case after Step 7, the oldest value is removed (Step 9). Finally, the community structure is updated and the next level of nodes is considered (Step 10). This is done until the node of interest is reached again and it is incomplete according to Property 3. At that point its completeness can be improved as well. During these steps, many nodes other than the node of interest are addressed and their IQ is improved. These nodes are automatically skipped in Steps 1-5 saving time by avoiding double-checks.

```
Algorithm for improving 3C in SN

Input: Network at time $t_1$ consisting of the nodes $N_1,..,N_{n_t}$; Network information communities $I^{S_m(H)}{}_{t_1}, m \in \{1,..,k_{t_1}(H)\}$ for all attributes $H$; For each attribute $H$ the maximum number of possible attribute value(s) $L(H) > 0$; Time span $\Delta t$; For each node $N_j, j \in \{1,\dots,n_{t_1}\}$, for each attribute $H$, the present activity community $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$ and the past activity community $A_{t_0+\Delta t,\Delta t}(N_1,H)$;

Output: A network where all nodes contain correct, current and complete attribute values;

$Complete(H) = NULL, Correct(H) = NULL, Current(H) = NULL, \forall H$; //initialise the sets of users with complete, correct and current profile information for the attribute $H$ respectively;

For each attribute $H$ and each node $N_j, j \in \{1,\dots,n_{t_1}\}$
 1: If $N_j \in Complete(H)$ and $N_j \in Correct(H)$ and $N_j \in Current(H), j \leftarrow j+1$
 2: If each information community $I^{S_m(H)}{}_{t_1}$ containing $N_j$ overlaps with $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$, then $N_j$ is correct
    and current and should be added to both $Correct(H)$ and $Current(H)$
 3: If $N_j \notin Correct(H)$ or $N_j \notin Current(H)$, identify the sets of attribute value(s) for which the profile
    information of $N_j$ satisfies Property 1 or Property 2 and remove them from the profile of $N_j$, add $N_j$ to both
    $Correct(H)$ and $Current(H)$
 4: If $A_{t_1,t_1-(t_0+\Delta t)}(N_1,H)$ does not overlap with any of the information communities which do not contain $N_j$,
    then $N_j$ is complete and should be added to $Complete(H)$
 5: If $N_j \in Complete(H), N_j \in Correct(H), N_j \in Current(H), j \leftarrow j+1$
 6: $Node \Leftarrow N_j$
 7: While $Node$ does not satisfy Property 3
        a. Identify the sets of attribute value(s) $S_m(H)$ for which the profile information of $Node$ satisfies
           Property 1 or Property 2 and remove them from the profile of $Node$, set $Correct(H) =$
           $Correct(H) \cup Node, Current(H) = Current(H) \cup Node$
        b. Identify the information communities $I^{S_m(H)}{}_{t_1}, m \in \{1,..,k_t(H)\}$ for the attribute $H$ to which $Node$
           does not belong and which overlap with the activity community of $Node$ for the attribute $H$ (the
           overlap must contain at least two nodes)
        c. For each non-empty $I^{S_m(H)}{}_{t_1}$ from Step 7.b. and each $\widetilde{N}_l$ from $I^{S_m(H)}{}_{t_1}$, set $Node \Leftarrow \widetilde{N}_l$
    End
 8: Complete the profile information of $Node$, add $Node$ to $Complete(H)$, clear duplicates
 9: If $L(H) = 1$, and the user's profile contains more than one attribute value for $H$, remove the oldest attribute
    value
10: If $Node \neq N_j$, update the community structure and go to Step 7, else $j \leftarrow j+1$.

End
```

Figure 3. Algorithm for improving 3C in SN

3 Evaluation

We evaluated our approach by applying it to the research SN ScholarMate, which has more than 2 million members and is the largest research SN in Asia. The data was provided by the owner of ScholarMate, IRIS Systems (Shenzhen). There are several reasons for choosing this SN. First, it is used for the assignment of reviewers to projects for funding based on the reviewers' research interests (Silva *et al.*, 2013). Thus, it is crucial that the research interests on a user's profile satisfy the 3C, as otherwise a project may be assigned an inappropriate reviewer leading to the funding of unsuccessful projects. Second, due to its narrow specialisation, it allows for a good preliminary evaluation without requiring additional assumptions and methods. In particular we concentrate on the attribute "research interest(s)" and assume that the activities for these attribute are the likes, comments and shares of research publications. We determined the *activity community* of a user as the five users on whose profile s/he has been the most active for the last 2 years and determined the *information communities* by applying overlapping clustering (N'Cir *et al.*, 2015). The main idea of overlapping clustering is that, as opposed to traditional clustering approaches, it allows an element to belong to more than one cluster. Since a researcher can work on more than one set of research interests, this is a reasonable method to use. We concentrated on one of the most active users in the network (after removing test profiles) and based on the literature (Bliss *et al.*, 2014) extracted all the friends of this user at the first and second level. The results are presented in Figure 4. They show that the *information* and *activity* communities do not always overlap and that the *information community* of the users suffers from low IQ which needs to be improved before improving their profile information. To validate our results, we plan to ask the users for the quality of the profile information by an already developed application.



Node	Information Community	Activity community	Property?		Information community	
			1,2	3	Outdated/incorrect	Missing
12	IV (4,6,12)	1,2,4,5,10	x	✓	Not relevant	I, V
1	I (1,5)	4,5,8,16,17	x	x	Not relevant	Not relevant
5	I (1,5)	1,3,6,15,18	x	x	Not relevant	Not relevant
2	V (2,10)	1,4,7,13,15	✓	✓	V	III
10	V (2,10)	4,11,12,15,25	✓	✓	V	IV
13	III (13,15,16)	1,2,4,5,15	x	✓	Not relevant	I
15	III (13,15,16)	4,6,7,17,19	✓	✓	III	IV
4	IV (4,6,12)	2,9,12,14,15	x	x	Not relevant	Not relevant
6	IV (4,6,12)	4,7,17,22	x	x	Not relevant	Not relevant

Figure 4. SN and the results from the first iteration of our approach on ScholarMate

4 Conclusion, Limitations and Future Research

In this paper we present an integrated approach for improving the currency, correctness, and completeness of profile information in SN. Our approach considers the overlap between the *information* and *activity communities* of the user and also the past community structure. Thus, it can be applied to any SN providing some interaction functionality. Moreover, the dynamic changes of the SN are considered when improving the currency of the profile information. In addition, our approach is very efficient, because when improving the IQ of one user, it also improves the IQ of other connected users. We conducted an evaluation on the research SN ScholarMate and received some very promising results. In particular, it is evident that the *information* and *activity communities* of the users do not always overlap in reality and that there is an IQ problem in real-world SN addressed by our approach.

The presented approach has some limitations. To begin with, since *activity communities* are user-specific, it is not straightforward to apply traditional techniques for community detection. Thus they need to be accordingly modified by future research. A similar problem exists with *information communities, where* we proposed a method based on overlapping clustering. They are defined on a network level, but a community is formed only if all nodes contain certain attribute value(s). Finally, to confirm the results, a more extensive evaluation including other SN should take place.

References

Abdesslem, F.B., Parris, I. and Henderson, T. (2012), “Reliable online social network data collection”, in Abraham, A. (Ed.), *Computational Social Networks,* Springer, pp. 183–210.

Abel, F., Henze, N., Herder, E. and Krause, D. (2010), “Interweaving Public User Profiles on the Web”, in Bra, P. de, Kobsa, A. and Chin, D. (Eds.), *User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science,* Vol. 6075, Springer, pp. 16-27.

Agarwal, N. and Yiliyasi, Y. (2010), “Information quality challenges in social media”, in *Proceedings of the International Conference on Information Quality (ICIQ),* University of Arkansas at Little Rock (UALR).

Aggarwal, C.C. (2011), *Social Network Data Analytics,* 1st ed., Springer.

Aggarwal, C.C. and Zhai, C. (2012), *Mining text data,* Springer Science & Business Media.

Alt, R. and Reinhold, O. (2012), “Social-Customer-Relationship-Management (Social-CRM)”, *Business & Information Systems Engineering (BISE),* 54 (5), 281–286.

Amichai-Hamburger, Y. and Vinitzky, G. (2010), “Social network use and personality”, *Computers in human behavior,* 26 (6), 1289–1295.

Batini, C. and Scannapieco, M. (2006), *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications),* Springer.

Bliss, C.A., Frank, M.R., Danforth, C.M. and Dodds, P.S. (2014), “An evolutionary algorithm approach to link prediction in dynamic social networks”, *Journal of Computational Science,* 5 (5), 750–764.

Breitbarth, W. (2012), “2012 Power Formula for LinkedIn Success User Survey”, available at: http://www.powerformula.net/2347/linkedin-infographic-want-to-know-what-others-are-doing/ (upon request) (accessed 20 March 2015).

Cai, Y. and Ziad, M. (2003), “Evaluating completeness of an information product”, in *Proceedings of AMCIS 2003,* Paper 294.

Cao, Q., Sirivianos, M., Yang, X. and Pregueiro, T. (2012), “Aiding the Detection of Fake Accounts in Large Scale Social Online Services”, in *NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 197–210.

Chen, R. (2013), “Living a private life in public social networks: An exploration of member self-disclosure”, *Decision Support Systems,* 55 (3), 661–668.

Chen, R. and Sharma, S.K. (2015), “Learning and self-disclosure behavior on social networking sites: the case of Facebook users”, *European Journal of Information Systems (EJIS),* 24 (1), 93–106.

Consumer Reports (2012), “State of the Net survey”, available at: http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm#editor (accessed 20 March 2015).

DeMicheli, C. and Stroppa, A. (2013), “Twitter and the underground market”, available at: http://nexa.polito.it/nexacenterfiles/lunch-11-de_micheli-stroppa.pdf (accessed 20 March 2015).

Emedia (2007), “RapidResearch Social Networking Sites Survey”, available at: http://www.realwire.com/release_detail.asp?ReleaseID=6671 (accessed 20 March 2015).

Even, A. and Shankaranarayanan, G. (2007), “Utility-driven assessment of data quality”, *ACM SIGMIS Database,* 38 (2), 75–93.

ExactTarget (2014), *2014 State of Marketing*, available at: http://content.exacttarget.com/en/StateOfMarketing2014?ls=Website&lss=StateofMarketing2014&lssm=Corporate&camp=701A0000000g98RIAQ (accessed 22 November 2014).

FACEBOOK INC. (2014), “Facebook Annual Report”, available at: http://investor.fb.com/secfiling.cfm?filingID=1326801-14-7 (accessed 20 March 2015).

Fire, M., Kagan, D., Elyashar, A. and Elovici, Y. (2014), “Friend or foe? Fake profile identification in online social networks”, *Social Network Analysis and Mining*, 4 (1), 1–23.

Fisher, C.W., Lauria, E.J.M. and Matheus, C.C. (2009), “An accuracy metric: Percentages, randomness, and probabilities”, *Journal of Data and Information Quality (JDIQ)*, 1 (3), p. 16.

Fortunato, S. (2010), “Community detection in graphs”, *Physics Reports*, 486 (3), 75–174.

Heidemann, J., Klier, M. and Probst, F. (2012), “Online social networks: A survey of a global phenomenon”, *Computer Networks*, 56 (18), 3866–3878.

Heinrich, B. and Klier, M. (2015), “Metric-based data quality assessment—Developing and evaluating a probability-based currency metric”, *Decision Support Systems*, 72, 82–96.

Heinrich, B., Klier, M. and Kaiser, M. (2009), “A procedure to develop metrics for currency and its application in CRM”, *Journal of Data and Information Quality (JDIQ)*, 1 (1), p. 5.

Hoffman, D.L. and Fodor, M. (2010), “Can you measure the ROI of your social media marketing?”, *Sloan Management Review*, 52 (1), 40–49.

Jones, J.J., Settle, J.E., Bond, R.M., Fariss, C.J., Marlow, C. and Fowler, J.H. (2013), “Inferring tie strength from online directed behavior”, *PloS one*, 8 (1), e52168.

Krasnova, H., Günther, O., Spiekermann, S. and Koroleva, K. (2009), “Privacy concerns and identity in online social networks”, *Identity in the Information Society*, 2 (1), 39–63.

Krombholz, K., Merkl, D. and Weippl, E. (2012), “Fake identities in social media: A case study on the sustainability of the Facebook business model”, *Journal of Service Science Research*, 4 (2), 175-212.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002), “AIMQ: a methodology for information quality assessment”, *Information & management*, 40 (2), 133–146.

Li, F., He, J., Huang, G., Zhang, Y. and Shi, Y. (2014), “A Clustering-based Link Prediction Method in Social Networks”, in *2014 International Conference on Computational Science*, Vol. 29, pp. 432–442.

Li, R., Wang, S., Deng, H., Wang, R. and Chang, K.C.-C. (2012), “Towards social user profiling: unified and discriminative influence model for inferring home locations”, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1023–1031.

LinkedIn (2012), “LinkedIn’s New Requirements for a 100% Complete Profile”, available at: http://www.linkedin-makeover.com/2012/02/20/linkedins-new-requirements-for-a-100-complete-profile/ (accessed 20 March 2015).

Lu, Z., Savas, B., Tang, W. and Dhillon, I.S. (2010), “Supervised link prediction using multiple sources”, in *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pp. 923–928.

Luarn, P., Chiu, Y.-P. and Jansen, J. (2015), “Key variables to predict tie strength on social network sites”, *Internet Research*, 25 (2).

Mislove, A., Viswanath, B., Gummadi, K.P. and Druschel, P. (2010), "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*, New York, New York, USA, pp. 251–260.

N'Cir, C.-E.B., Cleuziou, G. and Essoussi, N. (2015), "Overview of Overlapping Partitional Clustering Methods", in Celebi, M.E. (Ed.), *Partitional Clustering Algorithms*, Springer, pp. 245–275.

Nasdaq (2012), "FACEBOOK INC (FB) IPO", available at: http://www.nasdaq.com/markets/ipos/company/facebook-inc-673740-69138 (accessed 24 March 2015).

Nguyen, N.P., Dinh, T.N., Xuan, Y. and Thai, M.T. (2011), "Adaptive algorithms for detecting community structure in dynamic social networks", in *2011 Proceedings IEEE INFOCOM*.

Pike, J.C., Bateman, P.J. and Butler, B. (2013), "Dialectic Tensions of Information Quality: Social Networking Sites and Hiring", *Journal of Computer-Mediated Communication*, 19 (1), 56–77.

Probst, F. and Görz, Q. (2013), "Data Quality Goes Social: What Drives Data Currency In Online Social Networks?", in *2013 European Conference on Information Systems (ECIS 2013)*.

Silva, T., Guo, Z., Ma, J., Jiang, H. and Chen, H. (2013), "A social network-empowered research analytics framework for project selection", *Decision Support Systems*, 55 (4), 957–968.

Software Advice (2014), "Why Your Social Recruiting Tactics Are Wrong", available at: http://new-talent-times.softwareadvice.com/survey-why-social-recruiting-tactics-wrong-0714/ (accessed 20 March 2015).

Son, J.-Y. and Kim, S.S. (2008), "Internet users' information privacy-protective responses: A taxonomy and a nomological model", *MIS Quarterly*, 503–529.

Soundarajan, S. and Hopcroft, J. (2012), "Using community information to improve the precision of link prediction methods", in *Proceedings of the 21st International Conference Companion on World Wide Web*.

Statista (2015), "Number of social network users worldwide from 2010 to 2018 (in billions). based on eMarketer and the American Marketing Association", available at: http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed 23 March 2015).

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information Systems (JMIS)*, 12 (4), 5–33.

Xiang, R., Neville, J. and Rogati, M. (2010), "Modeling relationship strength in online social networks", in *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*.

XING (2013), "Annual Report 2013", available at: https://corporate.xing.com/fileadmin/IR/XING_AG_results_FY_2013.pdf (accessed 20 March 2015).

Xu, C., Visinescu, L. and Kim, D. (2013), "Disclose Intimately, Honesty, Heavily, Positively and Intentionally: An Exploration of Self-Disclosure in Social Networking Sites", in *34th International Conference on Information Systems (ICIS)*, Paper 70.