ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/301444819

Social Circle Discovery in Ego-Networks by Mining the Latent Structure of User Connections and Profile Attributes

Conference Paper · August 2015

DOI: 10.1145/2808797.2809303

CITATIONS
11

READS
326

3 authors:

Georgios Petkos
Information Technologies Institute (ITI)
24 PUBLICATIONS **724** CITATIONS

SEE PROFILE

Symeon Papadopoulos
The Centre for Research and Technology, Hellas
256 PUBLICATIONS **4,720** CITATIONS

SEE PROFILE

Ioannis (Yiannis) Kompatsiaris
The Centre for Research and Technology, Hellas
1,023 PUBLICATIONS **14,035** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Envisage View project

beAWARE View project

All content following this page was uploaded by Symeon Papadopoulos on 12 May 2017.

The user has requested enhancement of the downloaded file.

Social Circle Discovery in Ego-Networks by Mining the Latent Structure of User Connections and Profile Attributes

Georgios Petkos, Symeon Papadopoulos, Yiannis Kompatsiaris
Information Technologies Institute, Centre for Research and Technology Hellas Thessaloniki, Greece
{gpetkos,papadop,ikom}@iti.gr

***Abstract*—Online Social Networks (OSN) allow their users to organize their friends into groups, also known as *social circles*. These social circles can be used to better manage who has access to users' posted content and also to control the content posted from other users that they view. Unfortunately, these social circles are generated manually and this can be a laborious process for users with more than a few friends. In this paper, we propose an approach for automatically generating social circles that takes into account both the profile information of the friends to be grouped, as well as the social network connectivity between them, while it allows multiple membership of friends in social circles. The approach is based on an adaptation of the widely used Latent Dirichlet Allocation model and, despite the fact that it does not explicitly model social network connectivity, as other state of the art methods do, it manages to achieve results that are competitive and even better than those obtained from such methods, at a considerable lower computational cost.**

I. Introduction

According to statistics reported on February 2014[1], the average number of friends per Facebook user is 338, with the median being 200. In another very popular OSN, Twitter, the most recently reported average number of followers that a user has is 208[2]. Considering these numbers, it is clear that, in many cases, a user is likely to find it difficult to perceive their audience and subsequently to control their privacy settings. Indeed a number of studies have shown that OSN users face a number of challenges with respect to perceiving their audience and controlling their privacy settings. For instance, in [1], 65 Facebook users were asked to carefully examine their profiles and it was found that all of them identified at least one sharing violation, i.e. they were all sharing content with people that they really would not like to.

A tool that has been offered by OSN services to their users as a means of improving privacy control is friend lists. Friend lists, or *social circles* as they are often referred to in related works [3], provide the possibility of creating sets of friends that can be used for defining the users to which posted content will be visible and also for defining the users from which posted content will be displayed in the user's timeline. Moreover, friends lists can be a very useful tool that allows for a clearer perception of the audience of a user. Facebook in particular, offers three predetermined lists to which the user can add his/her friends. The first includes close friends, the second includes acquaintances, friends with which the user does not need to stay in close touch, and the third includes friends with which the user does not really want to share content. Additionally, Facebook allows the creation of custom lists that include friends that are grouped together due to some specific characteristic or context. Unfortunately though, the maintenance of these lists is performed manually by the users and is therefore a tedious process for a user with more than a few friends. Indeed, Facebook has stated that only roughly 5% of users have created at least one list of friends.

The goal of this paper is to provide a method for automatically generating social circles. In order to perform the task of social circle generation, there are two types of information that can be taken into account. The first is the set of any available profile properties that characterize the friends of the user. Clearly, the members of groups of friends are likely to share common profile properties, e.g., they may go to the same school. The second is the set of social network connections between the friends of the user. That is, members of the same circle are likely to be relatively densely linked to each other in the OSN. Nevertheless, as will be experimentally demonstrated in Section IV, neither of these sources of information is likely to provide very accurate results on its own and it is beneficial to combine them. The proposed approach takes into account both sources of information through a computationally simple and theoretically grounded framework.

A. Problem formulation

The problem of *social circle discovery* can be formalized as follows: Given a set of friends of a user $F = \{f_1, f_2, ...f_X\}$, with each friend f_i being represented by a set of Y attributes (e.g., which school they attended, where they reside, what languages they speak, etc.), i.e., $fp_i = \{f_{i1}, f_{i2}, ..., f_{iY}\}$, and given the set of explicit OSN relationships between the friends in F, which can be organized in a graph $G = (F, E)$, where F is the set of vertices of G and E is the set of explicit OSN connections between them, the goal is to recommend a set of circles $C = \{C_1, C_2, ..., C_K\}$, each of which contains a number of friends, i.e. $C_i = \{f_k | f_k \in F\}$. The problem is naturally treated as a clustering problem, albeit, it is recognized that a friend of a user may belong at the same time to multiple social circles. For instance, an OSN friend of a user may belong at the same time to the "school friends" social circle as well as to the "relatives" social circle. Therefore, the clustering approach to be adopted is required to be capable of producing overlapping clusters.

[1] http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics

[2] http://news.yahoo.com/twitter-statistics-by-the-numbers-153151584.html

B. Contributions

The proposed approach employs a very common probabilistic model with this property, the Latent Dirichlet Allocation (LDA) [2]. LDA is typically applied for topic detection in textual corpora. In particular, it detects a number of topics, each of which is modelled as a distribution over words and each document is assigned a probability distribution over the detected topics. Effectively, the set of detected topics also defines a soft clustering of documents, in which cluster membership for some document is given by its posterior distribution of topics. In our scenario, the documents represent the friends of the user and the produced topics are interpreted as the social circles. We take into account both the profile properties of the friends and the OSN links between them with a straightforward adaptation of LDA. In particular, **we represent each user with a document** whose tokens comprise a) the profile properties of the user, b) the set of identifiers (ids) of neighbouring users, and c) his/her own id. Thus, the vocabulary of the corpus processed by LDA consists of two types of elements, the first is the set of properties and the second consists of user ids. Eventually, the extracted topics will be distributions over both user properties and user ids, representing at the same time both the dominant properties of the users in a circle as well as who the core members of the circle are[3]. As will be experimentally demonstrated, the proposed approach achieves results competitive to state-of-the-art methods of considerably higher computational complexity [3].

The rest of this paper is structured as follows. Section II discusses a number of related works. Subsequently, Section III presents in detail the proposed approach and Section IV presents experimental results that demonstrate the value of the proposed method. Finally, Section V concludes this paper.

II. Related work

Before proceeding with discussing existing automatic methods for social circle detection, it is useful to mention that there are a number of studies that examine the criteria that users actually consider when organizing their OSN friends in groups; such criteria would also be useful for automatic methods. One such study [4] identifies six relevant criteria:

- Existence of cliques, involving groups of friends being highly connected to each other.
- Tie strength, which in fact may have various aspects: closeness, emotional intensity of relationship, level of trust and frequency of communication.
- Temporal episodes. Some groups tend to emerge from people that are all present in a significant event.
- Geographical locations and spatial proximity.
- Functional roles. Some links in a social network tend to form because they have some particular use or provide some specific service to the user. For instance, one participant of the study reported that she had added as contact a person she had encountered through a classified advertisement, so that the link could be used as a bookmark for future communication regarding the transaction.
- Organizational boundaries. For instance, people that work in the same company.

Looking at this list one can note that indeed, both connectivity (first criterion) and profile features (reflected in different ways in the rest of the criteria) are actually considered by users when they group their friends.

Interestingly, the authors of [4] also argue that the Structural Clustering Algorithm for Networks (SCAN) [5], a graph clustering algorithm, is very suitable for capturing most of these factors, and they perform some relevant experimental analysis supporting this claim. Moreover, they identified that users have difficulty grouping particular friends and they found that this was because these friends either had a weak association with any of the groups or they had strong associations with multiple groups. Since SCAN could successfully identify such people as outliers or hubs, it could be useful not only for grouping friends but also for identifying friends to which users should pay more attention when building their privacy policy. SCAN is an algorithm that we will use in our experiments.

Nevertheless, [4] concludes that there are important pitfalls when attempting to build a completely automated method for grouping friends. The reason is that they found that different people considered different grouping criteria to varying degrees. Thus, they claim, it is necessary to consider the different prioritization of the users and this is hard to achieve in a completely automated manner.

A further study that examines the mechanisms by which users group their OSN friends is presented in [6]. They identify specific types of groups of OSN friends:

- General friends, with some specific sub-categories: location-based, generic friends, close friends and friends of friends.
- College friends, either general college friends or college club/group friends.
- Friends from other education: high school friends and grade school friends.
- Other categories such as family, work, church and "dont know" friends, people that the user hardly knows or has never met in person.

This list indicates that social circles may form due to various factors and contexts, and hence any type of available information has to be taken into account for carrying out the task.

Let us now look at specific methods that are applicable to the problem. Naturally, there have been a number of methods that take into account only OSN connectivity and are based on graph clustering techniques [21], [22]. For instance, [19] compared a number of community detection algorithms and found that Infomap [23] performed best: the produced communities matched the ground-truth communities (provided by the users) better than other competing algorithms. A similar study [24] employed the method of [25] for discovering social circles. More generally, there is a large variety of graph

[3] Interestingly, as discussed in Section II, there have been extensions of LDA in which interactions between documents, in the form of a graph, are taken into account; however, these approaches are much more complex, in the sense that they explicitly compute a generative distribution for the graph.

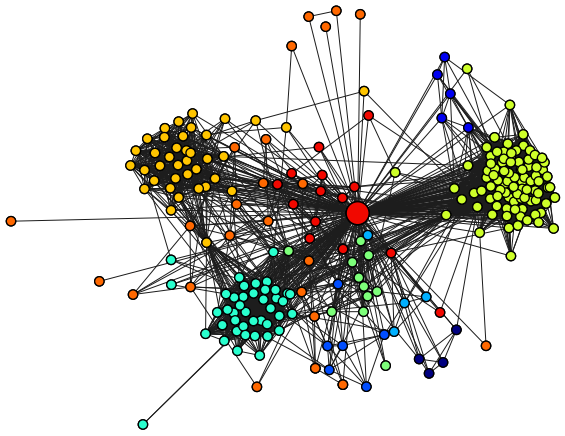


Fig. 1. Example Facebook ego-network from the dataset used in the experimental study (Section IV). The network comprises 239 nodes and 8,896 edges. 10 communities were discovered using the OSLOM algorithm [11] and are illustrated with different colors (although some nodes are assigned to multiple communities only a single color per node is used). One may observe that many nodes are naturally clustered into communities (largely corresponding to the user-defined social circles), while there are others, for which this assignment is not very clear. For such nodes, additional information beyond connectivity is necessary to correctly assign them to social circles.

clustering techniques that can be used for the problem. For instance, the OSLOM (Order Statistics Local Optimization Method) algorithm [11] seems appropriate, since it can detect overlapping communities, and at the same time it can also detect hierarchies of communities. Moreover, OSLOM is able to take into account edge directions and weights. An example community structure produced by OSLOM for one of the ego-networks used in our experiments is presented in Figure 1. Other overlapping community detection algorithms, such as COPRA (Community Overlap PRopagation Algorithm) [12] and Demon [16], based on label propagation, could also be used. Another candidate algorithm is Louvain [13], which relies on the maximization of the "modularity" measure. Other popular methods include the Girvan-Newman algorithm [14], the Clique Percolation Method [17] and the Laplacian Eigenmaps [15].

A probabilistic approach that takes into account only network connectivity and that allows for a node to belong to multiple groups relies on the Mixed Membership Stochastic Blockmodels [7]. This approach defines a generative model for the network formation process, in which the block-to-block (group-to-group) interactions play a central role.

More sophisticated methods consider not only the structure of the friend network but also the profile attributes of friends. Instances of such methods are presented in [34], [35]. However, these methods cannot assign items to multiple groups. Other probabilistic methods, including the methods presented in [26] and [3], are capable of assigning friends to multiple circles. Both operate by building a generative probabilistic model. Of particular interest to the work presented here are methods that are based on LDA. These extend LDA by adding to it a generative process for the network structure. Those include the Block-LDA method [10], Relational Topic Models [9] and Topic Link LDA [8]. Nevertheless, all these approaches invest resources in order to explicitly model the network structure formation and, as we will see later, competitive results for the task at hand can be obtained without this overhead.

Other recent approaches also take into account the strength of interactions [27]. Utilizing tie strength is a meaningful option for the task at hand and could be used even for producing very general circles like those suggested by Facebook (close friends, acquaintances and restricted). That is, tie strength between the user and some friend could be computed and the resulting value could be used to assign the friend in one of these circles. In fact, there is a number of methods that focus on characterizing the tie strength between a pair of friends. [18] presents an approach in which tie strength between a pair of users is predicted using a simple linear regression model. Their model takes into account a total of 74 features, including the number of inbox messages exchanged, the number of "social wall" messages exchanged, the number of photos in which both users are present, the number of days since the last communication, various network measures, etc. [19] presents a similar approach but with the number of predictors reduced to 14, leading to a faster system, as the respective OSN API needs to be queried far fewer times. A further approach along the same lines was proposed by [20].

III. Proposed method

One of the key requirements of social circle discovery is the capability to assign friends in more than one circles. Probabilistic methods such as [3] and the LDA variants [8], [10], [9] discussed in Section II meet this requirement. However, all these methods dedicate significant modelling capacity in order to explicitly model the network formation process. The modelling complexity associated with these methods comes at a significant computational cost. In particular tasks, having an explicit model of how the connectivity patterns emerge is useful, however, for the task at hand, i.e., detection of social circles, such a model may not be required. In fact, as the experiments indicate, the proposed method is able to achieve results comparable to the state-of-the-art method [3] (that explicitly models network connectivity) without explicitly modelling network formation and with much lower complexity. In the following, we will first revisit the main principles of the LDA topic modelling framework, which constitutes the basis of the proposed method, and we will then present the specifics of social circle discovery using LDA to capture the social circle structure that is associated with a user's friends' connectivity and their profile attributes.

A. Topic modelling using LDA

Latent Dirichlet Allocation (LDA) [2] is a generative model, typically used in the context of modelling text documents, as a means of explaining a set of new observations (documents) based on a set of unobserved latent variables (topics). Some necessary math notation that is typically used to present the key LDA concepts and process is the following:

- A word corresponds to a basic token in an examined document. Note that a word comes from a finite size vocabulary V and that a word is represented by its index w_m in the vocabulary.
- A document $\mathbf{w}$ denotes a sequence of N words, i.e. $\mathbf{w} = \{w_1, w_2, ..., w_N\}$.

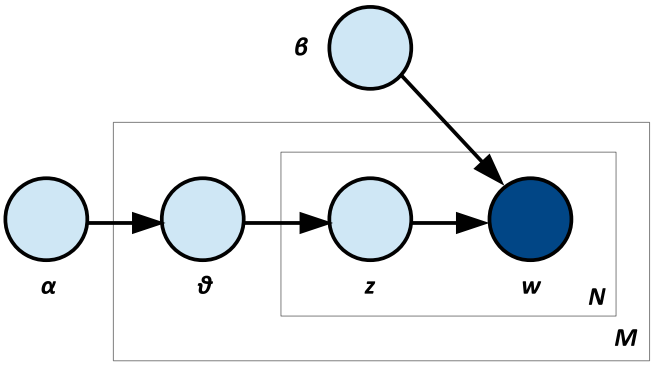


Fig. 2. Graphical model representation of LDA.

- A corpus $\mathbf{D}$ denotes a collection of M documents, i.e. $\mathbf{D} = \{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_M}\}$

LDA defines a generative model of a corpus of documents, in which each document is represented by a distribution over topics, and each topic is represented by a distribution over terms. The generative process for a document $\mathbf{w}$ in a corpus $\mathbf{D}$ is described by the following steps:

1) Choose $N_i \sim \text{Poisson}(\xi)$
2) Choose $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\text{Dir}(\alpha)$ is the Dirichlet distribution for parameter α
3) For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$ and $i \in \{1, \ldots, M\}$
   a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$
   b) Choose a word $w_{i,j} \sim \text{Multinomial}(\beta, z_{i,j})$

Here, α is a Dirichlet prior on the per-document topic distributions, β is the parameter of a Dirichlet prior on the per-topic word distribution, θ_i is the topic distribution for document i, z_{ij} is the topic for the j^{th} word in document i, and w_{ij} is the specific word. A graphical model representation of LDA is presented in Figure 2. Please note that only w is shaded, meaning that only the words of each document are observed and all other parameters are estimated.

The LDA model defines the structure of the joint probability distribution of words, topics and parameters but does not specify how inference and learning of the model parameters are performed. In the original paper [2] a Variational Bayesian method was utilized, but other approaches have employed different techniques such as Expectation Propagation [28]. There are different quantities that can be inferred from the joint distribution once learning has taken place, but for our purposes the most important ones are the posterior distribution of topics given a document $p(z|\mathbf{w})$ and the distribution of words per topic $p(\mathbf{w}|z)$.

B. LDA-based connectivity and profile attribute modelling

Having reviewed LDA, let us now return to the problem of social circle detection. With reference to the notation and problem definition that were presented in the introduction, one straightforward way to utilize LDA taking into account only the friends' profile properties would be to treat each friend f_i as a document $\mathbf{w_i}$, where $\mathbf{w_i}$ includes the friend's properties f_{p_i}. The identified topics would then be interpreted as the social circles. After training the model, the topic distribution for each friend would provide a circle membership assignment.

Let us now consider what we should do if we were to consider only connectivity information. A common idea in graph clustering approaches, such as those mentioned in Section II is that communities tend to have rather many *internal* connections, i.e. between their members, compared to connections to the rest of the network. This means that nodes that belong to the same community are likely to share a number of common neighbours. In fact, for several communities, it is likely that some nodes will be very central and that most of the other nodes in the community will be connected to them. These nodes that belong to the same community, just like documents that are related to a given topic will share the vocabulary that is specific to the topic, i.e. a "vocabulary of common neighbours". Thus, one can apply LDA for social circle detection taking into account only connectivity information by treating again each friend f_i as a document w_i, but this time w_i will include the friend's neighbours in the graph, as well as its own id. Formally, we set $\mathbf{w_i} = i \cup \{j|(i,j) \in E\}$.

In order to leverage both sources of information, i.e. profile attributes and friends' connectivity, we may also combine them using the same framework. Formally, the document w_i for a friend f_i would be $w_i = f_{p_i} \cup i \cup \{j|(i,j) \in E\}$. The vocabulary for the corpus would then include both user attributes and friends' ids and therefore the posterior distribution of words for a specific social circle would provide both the dominant properties of the users in the community, as well as the dominant users, at the same time.

Considering that combining the two types of features is likely to improve the results obtained from any of the two types of information alone, it is interesting to examine whether further features could be included to further improve the performance. One such possibility is to add features obtained from a network-based community detection algorithm that is executed in a preliminary stage. In particular, one could extract the communities of the ego-network, obtain the community ids to which each friend belongs, and then include these ids as additional words in the document for each friend.

C. LDA-based social circle discovery

To perform social circle discovery, the posterior distributions of topics given a document (or, posterior distributions of circles given a friend) need to be interpreted in order to produce a final assignment of friends to circles. Various approaches are possible, e.g., a possibility would be to examine the entropy of the distribution in order to estimate how concentrated the probability mass is and based on that to decide on the number of circles to which the circle is to be assigned. Nevertheless, we opt for a simpler approach. Each friend is assigned to a circle if the corresponding posterior probability for that circle is above the threshold $\frac{1}{K}$, where K is the number of topics/circles.

An important decision for the performance of LDA is the selection of the number of topics K to be produced. In order to select K automatically, we utilize the corrected Akaike Information Criterion (AICc) [29].

$$AICc = -2LL + 2p + \frac{2p(p+1)}{nA - p - 1} \quad (1)$$

where LL denotes the log-likelihood of the model, nA is the number of words in the corpus (not the size of the vocabulary, it is rather the sum of words in all documents) and p is the set of parameters of the model. Note that the number of parameters is $p = K(M-1) + N(K-1)$, where K is again the number of topics, N is the number of documents and M is the size of the vocabulary. Models with different values of K are fitted and the one with the lowest $AICc$ is selected. Note that as K increases, the log-likelihood increases and therefore the first term of AICc will decrease; however, it will be penalized by the other factors that grow as K increases.

Finally, it should be noted that, due to the fact that LDA is utilized, the proposed approach can directly take into account only properties of friends that are represented by categorical variables; it is not straightforward to take into account properties of friends that are represented by numerical or ordinal variables. In principle, such a scenario could be handled either by quantization and ignoring the order of the variables, or by extending the LDA model to explicitly take into account such variables. In our experiments, we do not have numerical or ordinal variables, all data are already quantized (and anonymized) by the data providers, so this is not an issue and plain LDA is utilized.

IV. Experiments

A. Dataset

We evaluate the proposed method on two datasets that were also used for the evaluation of the state-of-the-art method [3] on the problem. The first dataset consists of the required profile and network data for 10 Facebook users; that is, the networks around these 10 users and the profile attributes of all the users that appear in these networks. These 10 users also provided a manually produced assignment of their friends to social circles, effectively providing ground truth for the evaluation of the automatically produced social circles. The second dataset is similar but concerns 973 users from Twitter and the ground truth has been obtained by fetching lists that the users had already created. The data has been made publicly available by the authors of [3][4]. It is also noted that the datasets are completely anonymized; for instance, the actual name of the school that a user has attended is not provided, this is rather replaced by some arbitrary id.

Before reporting on the experimental results, we present some preliminary exploratory results on the data. These results indicate that perfect results cannot be expected based only on either network structure or profile information. We first look at the ground truth for a single Facebook user in the dataset and compute the distribution of profile attributes in each social circle. Then, we compute the Kullback-Leibler divergence between each pair of distributions p_j and p_k.

$$KL(p_j, p_k) = \sum_{w_i} p_j(w_i) log \frac{p_j(w_i)}{p_k(w_i)} \quad (2)$$

Low values of $KL(p_j, p_k)$ indicate that the two distributions are similar. Figure 4 shows in the form of a heatmap all pairs of KL divergences for the set of ground truth topics for the randomly selected user.

[4]http://snap.stanford.edu/data/

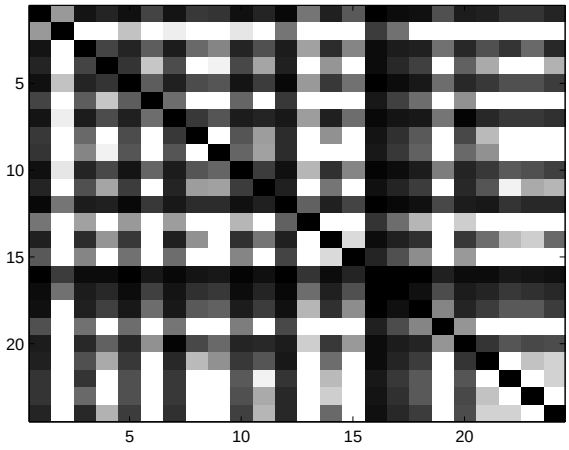


Fig. 3. Heatmap representing the KL-Divergence of word distributions between the circles of a randomly chosen Facebook user, as computed by Equation 2. Dark cells indicate that the KL divergence between the corresponding pair of distributions is small.

As illustrated in Figure 4, the distribution for each social circle is quite similar to the distributions of at least some other social circles. Therefore, a method that is based only on the profile attributes of the friends of a user will have to distinguish between distributions that are quite similar to each other. Similar heatmaps are obtained for the other users in the dataset as well.

Furthermore, we explore how separated the ground truth circles are, considering only network information. In particular, we utilize the well known modularity measure [30]. Given a grouping of the nodes of a graph in K sets, the modularity is computed by Equation 3.

$$Q = \sum_{i=1}^{K} (e_{ii} - a_i^2) \quad (3)$$

where e_{ii} is the percentage of edges in group i and a_i is the percentage of edges with at least one end in group i. Modularity ranges between -1 and +1, with positive numbers indicating that the number of edges within groups exceeds the number expected on the basis of chance. In general, positive values will indicate easier to cluster graphs. We compute the contribution to modularity of each social circle of the 10 Facebook users. That is, we compute the summand in the previous equation for each social circle. This gives an indication of how separated each social circle is from the rest. The distribution is shown in Figure 4. As it can be seen, there are only a few social circles with a clearly positive score, meaning that they are very well separated. There are also a few with a clearly negative score, but most of them have a score very close to zero. This indicates that most of the social circles are not very well separated from the rest, considering only network connectivity. In short, these explorative results indicate a) the difficulty of the task and b) the fact that neither of the sources of information (profiles and network connectivity) alone are likely to result in perfect results and therefore it may be useful to combine them.

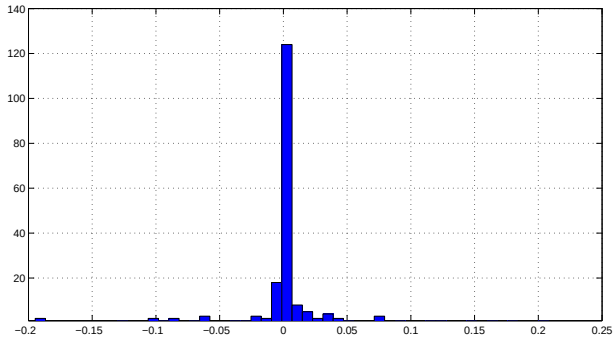


Fig. 4. Histogram representing the distribution of modularity contributions for individual circles in the Facebook dataset.

B. Experimental setup

In order to be able to directly compare our results to those reported in [3], the same evaluation metric, Balanced Error Rate (BER) is utilized (please note that the evaluation script that was utilized was provided by the authors of [3]). In particular, assume $C = \{C_1, C_2, ..., C_K\}$ is the set of automatically produced circles and that $\bar{C} = \{\bar{C}_1, \bar{C}_2, ..., \bar{C}_{\bar{K}}\}$ is the set of ground truth circles. The BER between a produced circle C_i and a ground truth circle $\bar{C}_i$ is given by Equation 4.

$$BER(C_i, \bar{C}_i) = \frac{1}{2}\left(\frac{|C_i \setminus \bar{C}_i|}{|\bar{C}_i|} + \frac{|C_i^c \setminus \bar{C}_i^c|}{|\bar{C}_i^c|}\right) \quad (4)$$

where the superscript c denotes the set complement operator. In [3] it is mentioned that this measure is preferable to e.g. 0/1 loss, which assigns extremely low error to trivial predictions. In order to find the correspondence between the circles in C and the circles in $\bar{C}$, the optimal match (i.e. the one with the minimum sum of BER scores) is computed via linear assignment. This can be considered as the total matching score between the automatically produced C and the reference $\bar{C}$ circles structure, is denoted by σ and is computed with the help of Equation 5.

$$\sigma(C, \bar{C}) = \max_{f: C \to \bar{C}'} \frac{1}{|f|} \sum_{C \in \mathrm{dom}(f)} (1 - BER(C, f(C))) \quad (5)$$

where f is the (potentially partial) correspondence function between C and $\bar{C}$. It should be noted that this matching does not require the number of produced circles to be the same as the number of ground truth circles. That is, if the number of produced circles is smaller than the number of ground truth circles, some ground truth circles will not be matched and if the number of ground truth circles is smaller than the number of ground truth circles then some produced circles will not be matched; in both cases there is no penalty. To compute σ over a set of ego-networks, we average over the obtained values.

We evaluate and compare five state-of-the-art approaches, with four variants of the proposed approach:

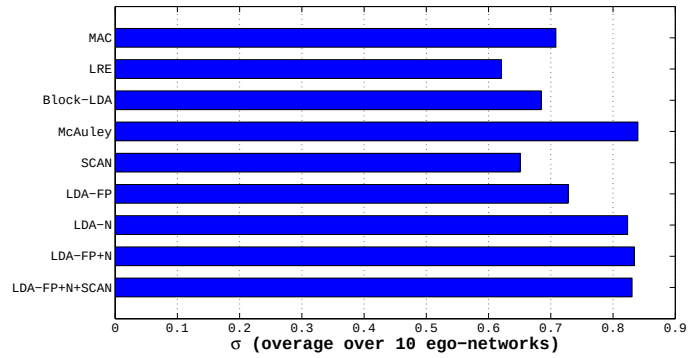
- Multi-Assignment Clustering (MAC) proposed by [31], which operates on the vectors of profile attributes (i.e. does not use connectivity information) and assigns each corresponding node to multiple clusters.
- Low-Rank Embedding (LRE) [32], in which both the node attributes and the connections between them are projected into a feature space, where classical clustering techniques are applicable.
- Block-LDA [10], a generative model-based technique that jointly considers profile attributes of users and their connections.
- SCAN community detection [5]. SCAN depends on two parameters, μ and ϵ that control the minimum number of nodes that a detected community should have and how close detected communities should be to a fully connected subgraph. We report the best results obtained by varying these two parameters.
- Social circle discovery method by McAuley and Leskovec [3], which takes into account both the users' profile attributes and their connections.
- LDA using only the set of friends' profile attributes to construct the documents. We will refer to this method as LDA-FP.
- LDA using only the network structure to construct the documents. That is, each document consists of the ids of the corresponding user's neighbours and the user's own id. We will refer to this method as LDA-N.
- LDA using both the set of friends' profile features and the network structure. We will refer to this method as LDA-FP+N.
- LDA using the set of friends' profile features, the network structure and the ids of communities assigned to friends by the SCAN algorithm. We will refer to this method as LDA-FP+N+SCAN.

The results of the first three methods [31], [32], [10], as well as of the method by McAuley and Leskovec [3] have been obtained from [3].

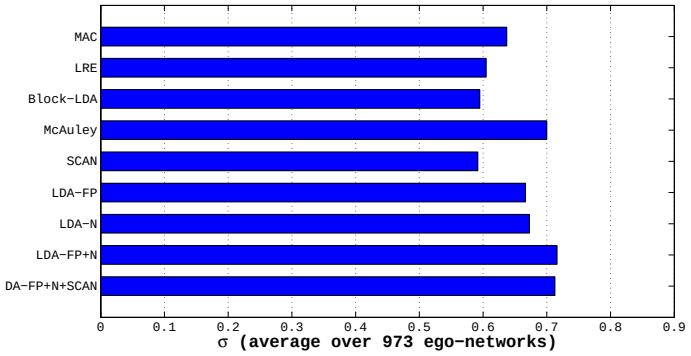
C. Results

Figure 5 shows the performance of the tested methods on the Facebook (top) and the Twitter (bottom) ego-networks. The performance is measured in terms of average σ, therefore higher values correspond to better social circle accuracy.

The first thing to note is that for both OSNs the proposed methods that combine profile and network information are competitive to the method of McAuley and Leskovec [3]. More particularly, for the Facebook ego-networks, [3] reports a score of 0.84, while LDA-FP+N achieves a score of 0.8343 and LDA-FP+N+SCAN achieves a score of 0.8406. For the Twitter ego-networks, [3] report a BER score of 0.70, while LDA-FP+N achieves a score of 0.7161 and LDA-FP+N+SCAN 0.7127. That is, LDA-FP+N and LDA-FP+N+SCAN slightly outperform the method of McAuley and Leskovec on the Twitter ego-networks. Also, these two methods perform significantly better than the other competing approaches. It should also be mentioned that the performance of all methods on the Twitter ego-networks is lower compared to the performance achieved on the Facebook ego-networks. According to [3], this is due to the fact that many Twitter circles have not been maintained since they were created (hence they may not reflect



(a) Facebook



(b) Twitter

Fig. 5. Results on the Facebook (top) and Twitter (bottom) datasets. MAC denotes the Multi-assignment clustering method described in [31], LRE is the Low-Rank Embedding method proposed in [32], Block-LDA the method proposed in [10], while McAuley denotes the state-of-the-art approach described in [3], LDA-FP denotes LDA using only the friends' properties to construct the documents, LDA-N denotes LDA using only the neighbours of each friend to construct the documents, LDA-FP+N is a combination of the previous two and LDA-FP+N+SCAN is an extension of LDA-FP+N in which we add community identifiers as computed by SCAN

reality very accurately), whereas the Facebook circles were created on demand for the purposes of the study.

It can also be noticed that LDA-FP, that utilizes only profile information, and SCAN and LDA-N that take into account only network information are outperformed by the methods that take into account both (LDA-FP+N and LDA-FP+N+SCAN). Nevertheless, for the Facebook dataset in particular, LDA-N is quite close to the performance of LDA-FP+N and LDA-FP+N+SCAN; instead, on the Twitter ego-networks, LDA-N achieves somewhat lower performance.

Finally, Table I presents the top words for random sample of topics produced by the LDA-FP+N approach. Please note that due to the fact that data are anonymized, profile properties appear with a category description and a numerical value identifier, whereas friends are identified by a single number. It can be noticed that there are circles that are primarily characterized by profile attributes, others that are primarily characterized by user ids (i.e. connectivity) and others by both.

V. Conclusions

In this paper, we presented an approach for grouping the friends of an OSN user in a set of social circles. The proposed approach utilizes LDA at its core, treating the set of friends to be grouped as documents and interprets the topics produced as social circles. In order to take into account both the profile information of the friends to be clustered as well as the OSN connections between them we proposed to include in the document representing each friend, not only his / her profile properties, but also the ids of the other friends to which he / she is connected in the OSN. As compared to previous approaches that attempt to cluster relational dataset, i.e. datasets that include datapoints linked to each other, the proposed approach does not explicitly model network connectivity. Instead, network connectivity is partly modelled jointly with the profile properties distribution of each circle. This allows for a simpler model that is easier to train and compute, but with performance comparable, and even better than other state of the art methods.

Regarding future work, we intend to further improve the proposed approach by also considering the strength of ties between the users. We also intend to consider alternatives to LDA, such as the Hierarchical Dirichlet Process [33].

TABLE I. Top words for a random sample of produced social circles using the LDA-FP+N method.

Circle	Top words
1	locale_127 gender_78 educ_type_53 educ_type_55 work_end_157 gender_77
2	67 122 200 21 277 188
3	locale_127 gender_78 educ_type_54 gender_77 school_id_52 educ_year_66
4	53 242 346 249 80 94
5	gender_78 332 locale_127 work_employer_50 339 324
6	educ_type_53 gender_78 educ_concentration_14 educ_type_55 school_50 locale_127

Acknowledgements

This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

References

[1] M. Madejski, M. Johnson, and S. Bellovin, *A study of privacy settings errors in an online social network*, 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE, 2012.

[2] D. Blei, A. Ng, and M. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 (March 2003), 993-1022, 2003.

[3] J. Leskovec and J. Mcauley, *Learning to Discover Social Circles in Ego Networks*, Advances in Neural Information Processing Systems 25, 2012.

[4] S. Jones, and E. ONeill, *Feasibility of structural network clustering for group-based privacy control in social networks*, Proceedings of the 6th Symposium on Usable Privacy and Security, 2010.

[5] X. Xu, N. Yuruk, Z. Feng and T. Schweiger, *SCAN: a structural clustering algorithm for networks*, In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07). ACM, New York, NY, USA, 824-833, 2007.

[6] P. Kelley, R. Brewer, Y. Mayer, L. Cranor, and N. Sadeh, *An Investigation into Facebook Friend Grouping*. Human-Computer Interaction INTERACT 2011. Lecture Notes in Computer Science Volume 6948, pp. 216-233, 2011.

[7] E. Airoldi, D. Blei, S. Fienberg and E. Xing, *Mixed Membership Stochastic Blockmodels*, Journal of Machine Learning Research 9(Sep), 1981–2014, 2008.

[8] Y. Liu, A. Niculescu-Mizil and W. Gryc, *Topic-link LDA: joint models of topic and author community*. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). ACM, New York, NY, USA, 665-672, 2009.

[9] J. Chang and J. Boyd-Graber, *Relational topic models for document networks*, In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics AISTATS, 2009.

[10] R. Balasubramanyan and W. Cohen, *Block-LDA: Jointly modeling entity-annotated text and entity-entity links*, In ICML 2010 Workshop on Topic Models: Structure, Applications, Evaluation, and Extensions, 2010.

[11] A. Lancichinetti, F. Radicchi, J. Ramasco, and S. Fortunato, *Finding statistically significant communities in networks*. PLOS one 6, no. 4 (2011), e18961, 2011.

[12] S. Gregory, *Finding overlapping communities in networks by label propagation*, New Journal of Physics 12, no. 10, 2010.

[13] V. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics, 2008.

[14] M. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical review E 69, no. 2, 2004.

[15] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural computation 15, no. 6, 2003.

[16] M. Coscia, G. Rossetti, F. Giannotti and D. Pedreschi, *Demon: a local-first discovery method for overlapping communities*, In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 615-623. ACM, 2012.

[17] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature 435, 814, 2005.

[18] E. Gilbert, and K. Karahalios, *Predicting Tie Strength with Social Media*, In Proceedings of the 27th International Conference on Human Factors in Computing Systems, 2009.

[19] R. Fogues, J. Such, and A. Espinosa, *BFF: A tool for eliciting tie strength and user communities in social networking services*, Information Systems Frontiers 16.2, 225-237, 2014.

[20] T. Spiliotopoulos, D. Pereira and I. Oakley, *Predicting Tie Strength with the Facebook API*, Proceedings of the 18th Panhellenic Conference on Informatics, 1-5, 2014.

[21] S. Fortunato, *Community Detection in Graphs*, Physics Reports, 486 (3-5), 75-174, 2010.

[22] S. Papadopoulos, Y. Kompatsiaris, A. Vakali and P. Spyridonos, *Community Detection in Social Media*, Data Mining and Knowledge Discovery 24, no. 3, 515-554, Springer, 2012.

[23] M. Rosvall, and C. Bergstrom, *Maps of Random Walks on Complex Networks Reveal Community Structure*, Proceedings of the National Academy of Sciences (PNAS), 105, no. 4, 1118-1123, 2008.

[24] F. Adu-Oppong, C. Gardiner, and P. Tsang, *Social Circles: Tackling Privacy in Social Networks*, Symposium on Usable Privacy and Security (SOUPS), 2008.

[25] N. Mishra, R. Schreiber, I. Stanton and R. Tarjan, *Clustering Social Networks*, In 5th International Workshop on Algorithms and Models for the Web-Graph, LNCS volume 4863, 5667. Springer, 2007.

[26] J. Yang, J. McAuley, and J. Leskovec, *Community Detection in Networks with Node Attributes*, ICDM '13, 2013.

[27] H. Dev, M. Ali and T. Hashem, *User Interaction Based Community Detection in Online Social Networks*, Database Systems for Advanced Applications. LNCS volume 8422, 296-310. 2014.

[28] T. Minka and J. Lafferty, *Expectation-propagation for the generative aspect model*, In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, 2002.

[29] C. Hurvich and C. Tsai, *Regression and time series model selection in small samples*, Biometrika 76, 297307, 1989.

[30] M. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences (PNAS) 103, no. 23, 2006.

[31] A. Streich, M. Frank, D. Basin and J. Buhmann, *Multi-asignment clustering for boolean data*, Journal of Machine Learning Research 13, no. 1, 459-489, 2012.

[32] T. Yoshida, *Towards finding hidden communities based on user profiles*, In ICDM Workshops, 2010.

[33] Y. Teh, M. Jordan, M. Beal, and D. Blei, *Hierarchical Dirichlet Processes*, Journal of the American Statistical Association 101, 15661581, 2006.

[34] Y. Ruan, D. Fuhry and S. Parthasarathy, *Efficient community detection in large networks using content and links*, In Proceedings of the 22nd international conference on World Wide Web (WWW '13), 1089-1098, 2013.

[35] Y. Zhou, H. Cheng and J. Xu Yu, *Graph clustering based on structural/attribute similarities*, Proceedings of the VLDB Endowment 2, no. 1 (August 2009), 718-729, 2009.

View publication stats