Density-based community detection in social networks

Kumar Subramani, Alexander Velkov, Irene Ntoutsi, Peer Kröger, Hans-Peter Kriegel
Institute for Informatics, Ludwig-Maximilians-Universität München, Munich, Germany
kumar.subramani@consol.de,alexvelkov@gmail.com, {ntoutsi, kroeger, kriegel}@dbs.ifi.lmu.de
Homepage: http://www.dbs.ifi.lmu.de

***Abstract*—This paper deals with community detection in social networks using density-based clustering. We compare two well-known concepts for community detection that are implemented as distance functions in the algorithms SCAN [1] and DENGRAPH [2], the structural similarity of nodes and the number of interactions between nodes, respectively, in order to evaluate advantages and limitations of these approaches. Additionally, we propose to use a hierarchical approach for clustering in order to get rid of the problem of choosing an appropriate density threshold for community detection, a severe limitation of the applicability and usefulness of the SCAN and DENGRAPH algorithms in real life applications. We conduct all experiments on data sets with different characteristics, particularly Twitter data and Enron data.**

I. Introduction

Social networks have become a trend in this modern era. Therefore its awareness is spreading rapidly attracting a lot of new users. Social networking also helps commercial organisations gain new contacts and clients. Thus, it enables one to establish relationships on a global level even being at home. There are numerous ways to define a social network. According to Barker a social network is defined as *"Individuals or groups linked by some common bond, shared social status, similar or shared functions, or geographic or cultural connection. Social networks form and discontinue on an ad hoc basis depending on specific need and interest."* [3]. From this definition, it is not necessary that the members of a social network know each other or share anything outside their network. But, in reality, members of such networks try to seek a bond with friends, family or colleagues from the real world.

Finding *communities* in social networks is an important challenge for a broad range of applications many of them related to personalized services and marketing. Similar to the definition of social networks the notion of a social network community can be perceived through various properties of the network. Such a property might be a set of members interacting extensively within a group and have a looser connection to the other members of the network. For example fans of a football club or a group of philatelists may form a community in a social network by posting textual messages or photos regarding these topics. Social networks "produce" a lot of data regarding their users and their interactions that can be used for community mining. Analyzing such type of data is challenging due to the quantity of the generated data and the different patterns that might exist in it. One of the core interests in mining social network communities lies in the analysis of the relationships between the members of these networks. These relationships can be interactions, friendships, followers, common hobbies, geographical locations, etc.

Detecting communities in real-world social networks can be achieved by choosing an objective function that captures the intuition of an underlying community. The number of common friends/ followers relationships or the number of common hobbies between the members in a social network are some examples for an objective function. One can apply an algorithm to extract the members who have a larger semantic resemblance to the chosen objective function and hence detect a community for the application of interest. This idea of community detection when translated to a social network graph will result in a set of nodes which belong to a community as defined by the objective function.

One way to detect communities is by applying *clustering* on the social network graph. Clustering can be informally defined as partitioning of set of objects into groups so that objects within the same group are similar to each other and less similar to objects of other groups. A lot of clustering methods have been proposed [4] like the partitioning methods (e.g k-Means), the density-based methods (e.g DBSCAN) and the grid-based methods (e.g STING). Density-based methods do not require an apriori knowledge about the resulting number of clusters nor do they make any assumption about the shape of the resulting clusters. Moreover, they are insensitive to outliers and noise. For all these reasons, we focus on density-based methods hereafter.

Recently, density-based methods have been used for community detection in social networks, e.g. DENGRAPH [2] and SCAN [1]. Both these methods build upon the well known density based clustering algorithm DBSCAN by incorporating an appropriate distance function in the clustering process. In particular, DENGRAPH [2] introduces the interactions-based distance that calculates the aggregated number of interactions between two users in a social network, while SCAN [1] introduces the structure-based distance that calculates the shared neighbors between two users. Since the distance functions have different semantics, the resulting clusters/ communities have also different semantics. However, it is not clear what are the limitations and the benefits of each approach, since so far there is no systematic evaluation of both approaches.

Moreover, it is not clear on what datasets these approaches are expected to work. In this paper, we evaluate both approaches and point out their pros and cons for community detection. Moreover, we experiment with datasets of different characteristics and derive conclusions on the type of datasets that better fit each distance function. Another important decision in density-based clustering is the choice of the density threshold. It is not straightforward to chose an appropriate density threshold and the above methods do not provide any clue towards this direction. Here we propose a solution for choosing the appropriate density threshold which is based on the density-based hierarchical algorithm OPTICS [5].

In summary, the contributions of this work are as follows: First, we provide an evaluation of the different distance measures for density-based community discovery in social networks. We try to infer the salient features of the different measures and to understand the type of network datasets that better fit each distance function. Second, we propose to use OPTICS as a way to deal with the problem of choosing the appropriate density threshold for community detection.

The remainder is organized as follows: In Section II we review the related work. In Section III we describe our problem settings and the different distance functions. In Section IV we present two case studies on real datasets, namely the Enron dataset and the Twitter dataset. Conclusions and outlook are presented in Section V.

II. Related work

A. Density-based clustering algorithms

Clustering is one of the most well studied areas in data mining and as such a large number of algorithms have been proposed. Partitioning methods like K-Means [6] require the number of clusters to be discovered as an input to the algorithm. On the contrary, in density-based methods the number of clusters is not predefined, rather it is revealed by the algorithms based on the data set characteristics. Here clusters are defined as areas of high density surrounded by areas of low density, the so called noise. Clusters of arbitrary shape can be discovered and also, outliers and noise do not affect the outcome of the algorithms. Such kind of methods are more appropriate for our case, since it is difficult to predict the number of clusters/ communities that exist in a social network. So, we focus on density-based clustering algorithms hereafter.

DBSCAN [7] is the pioneer work in this area. Here the density is defined locally in the neighborhood of each point. DBSCAN requires two input parameters: the distance ϵ for defining the neighborhood of a point and the minimum number of points μ in the neighborhood for characterizing a point as core, i.e., as located in a dense region. A cluster is defined as a set of connected points that meet the density parameters ϵ and μ. The major drawback of DBSCAN is that it cannot detect nested clusters especially over a space with variable density.

To deal with this limitation, the algorithm OPTICS [5] has been proposed which can accentuate clusters with variable density. OPTICS does not provide a clustering result, rather it outputs an augmented ordering of the data set, representing its density-based clustering structure. The augmented cluster-ordering of every point in the data set using OPTICS is equivalent to the density-based clustering but for a broader range of densities. So OPTICS simulates DBSCAN for an infinite number of distance parameters ϵ_i, $0 \leq \epsilon_i \leq \epsilon$.

B. Density-based clustering in social networks

The idea of density-based clustering has recently been applied to social network analysis. The algorithms DENGRAPH [2] and SCAN [1] have been proposed that introduce two different distance functions for community detection and integrate these functions into the DBSCAN algorithm.

DENGRAPH [2] is an incremental density-based community discovery algorithm. It discovers density-based clusters over a network of interacting nodes. The distance function is defined in terms of the interactions between the nodes/members of the network. The intuition is that the closeness of two members in a network is captured by the number of interactions between these members. In a social network, this means that people that interact to each other form a community.

SCAN [1] is a density-based clustering algorithm that detects clusters, hubs and outliers in a network. The idea here is to use the structure of each node, i.e., its neighborhood nodes, for similarity assessment. So, two nodes are considered to be similar if their structures are similar, that is if they share many common nodes. A node is assigned to a cluster according to how many neighbors out of all its neighbors it shares with other nodes in the cluster.

III. Density-based community detection

We model a social network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; $\mathcal{V}$ is the set of nodes corresponding to people in the network and $\mathcal{E}$ is the set of edges corresponding to interactions between these people. The semantics of the interactions depend on the network itself. Also, depending on the network, the edges might be weighted or not.

In order to detect communities, it is important to specify an appropriate similarity or distance function between nodes/ members of the network (cf. Section III-A). Another important decision is the clustering algorithm (cf. Section III-B) for extracting the clusters/ communities of similar nodes.

A. Distance functions

To assess the similarity or distance between two nodes in the network, we employ two well known distance functions from the literature, the interactions-based distance and the structure-based distance.

1) Interactions-based distance function: The interactions-based distance function was introduced in DENGRAPH [2] and is based on the aggregated number of interactions between two members of a social network. The original distance function refers to a directed graph, so here we modify the definition for our undirected graph case.

Definition 1 (Interactions-based distance function): Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected weighted graph. The interactions-based distance between two nodes $A, B \in \mathcal{V}$ is defined as:

$$dist(A,B) = \begin{cases} 0 & , \text{ if } \quad A = B \\ (I_{A,B})^{-1} & , \text{ if } \quad I_{A,B} > 0 \\ c & , \text{ if } \quad I_{A,B} = 0 \end{cases}$$

where $I_{A,B}$ is the sum of interactions between A and B, irrespective of the initiator A or B. The higher the number of reciprocal interactions between two nodes, the smaller their distance is. The notion of interactions is network dependent; in the experimental section we will define the semantics of the interactions for each data set.

The distance takes values in the $[0-1]$ range and can be defined when we observe interactions between two nodes. Note though that might not exist interactions between all pairs of nodes and thus, the distance between two nodes might be undefined. To model such kind of cases, we set the distance to a constant c (outside the valid range of the distance values) (We used $c = 2$ for the experiments).

2) Structure-based distance function: The structure-based distance function was introduced in SCAN [1] and is based on the shared neighbors between two nodes. Originally it was introduced as a similarity function, so here we modify the definition as follows:

Definition 2 (Structure-based distance function): Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected unweighted graph. The structure-based distance between two nodes $A, B \in \mathcal{G}$ is defined as:

$$dist(A,B) = \begin{cases} 0 & , \text{ if } \quad A = B \\ 1 - \sigma(A,B) & , \text{ otherwise} \end{cases}$$

where

$$\sigma(A,B) = \frac{|\Gamma(A) \cap \Gamma(B)|}{\sqrt{|\Gamma(A)||\Gamma(B)|}}$$

and

$$\Gamma(A) = \{A\} \cup \{B \in \mathcal{V} \mid (A,B) \in \mathcal{E}\}$$

In the above definition, $\Gamma(A)$ is the structure of node A consisting of all its neighbor nodes, i.e., all those nodes in $\mathcal{G}$ with whom A is connected via an edge. The similarity function $\sigma()$ relies on the number of common neighbors between two nodes; the higher the number of neighbors they share, the higher their similarity is. This number is normalized with the geometric mean, so its values lie in the $[0-1]$ range. The extreme value 0 is achieved when A and B are connected to each other but they are not connected to any other node in the graph, so $\sigma(A,B) = 1$. The extreme value 1 is achieved when there are no shared neighbors between A and B and thus, $\Gamma(A) \cap \Gamma(H) = \emptyset$.

B. Density-based community detection

As already mentioned, we opt for density-based clustering algorithms since they do not require the number of clusters as input to the algorithm, they are insensitive to noise and outliers and also they can discover clusters of arbitrary shapes. In particular, we experimented with the DBSCAN algorithm [7]. DBSCAN requires as input the distance threshold ϵ for computing the neighborhood of each point and the density threshold μ for deciding whether a point is core. The parameter ϵ is crucial for DBSCAN and it is not always easy to be defined. To this end, we used the algorithm OPTICS [5] which provides an ordering of the data set and thus can guide the ϵ selection process.

So, we use both DBSCAN and OPTICS accompanied with the distance measures we discussed before in order to detect communities in a social network and the appropriate density threshold level where these communities exist.

IV. Case Studies

We experimented with the density-based clustering algorithm DBSCAN and the two distance functions, interactions-based distance and structure-based distance, in two real social network datasets, the Enron email dataset and the Twitter dataset. We used OPTICS for the selection of the appropriate density threshold parameters for DBSCAN. Hereafter, we present our findings for each dataset and we summarize the findings at the end of the section.

Both distance functions and their incorporation into DBSCAN and OPTICS where implemented in Weka [8].

A. Enron dataset

The Enron dataset [9] contains around 517,431 emails from 151 users. distributed in 3500 folders. It can be modeled as an undirected graph where the *nodes* stand for the employees and the *edges* represent interactions between them. Replies or forwarded emails between two employees A, B count as interactions between them. The sum of all interactions is represented by $I_{A,B}$ and comprises the weight of edge $\mathcal{E}(A,B)$. Note that there might be no edges between every pair of nodes, this is because not all nodes/ employees interact with each other.

1) Interactions-based communities: We first applied DBSCAN with the interactions-based distance (cf. Definition 1) using μ = 3 and ϵ = 0.3. The resulting clustering is depicted in Figure 1; there is a single "grey" community consisting of a set of people connected through interactions.

The result strongly depends on the choice of the ϵ parameter. We used OPTICS for choosing the appropriate value for ϵ; the cluster ordering is depicted in Figure 2. The members in the right extreme position in the dent depicted by OPTICS are "benasante", "lsmith", "erosenberg". If we start DBSCAN with a smaller ϵ, e.g., ϵ = 0.2, these members are not going to be part of the cluster since $0.2 < 0.25$ (their smallest interactions distance).

2) Structure-based communities: We also applied DBSCAN with the structure-based distance (cf. Definition 2) using μ = 3 and ϵ = 0.5. The resulting clusters are depicted in Figure 3 (left). Recall that in this case the edges are unweighted and the distance relies on the neighbors of each node. Five communities were discovered (depicted in different colors) which represent cohesive groups in the network. The *green community* is the most cohesive one consisting of people that interact mostly with people within their group rather than other people. For example, "jonalstad" and "dougcebryk" are members of the group because they interact mainly or

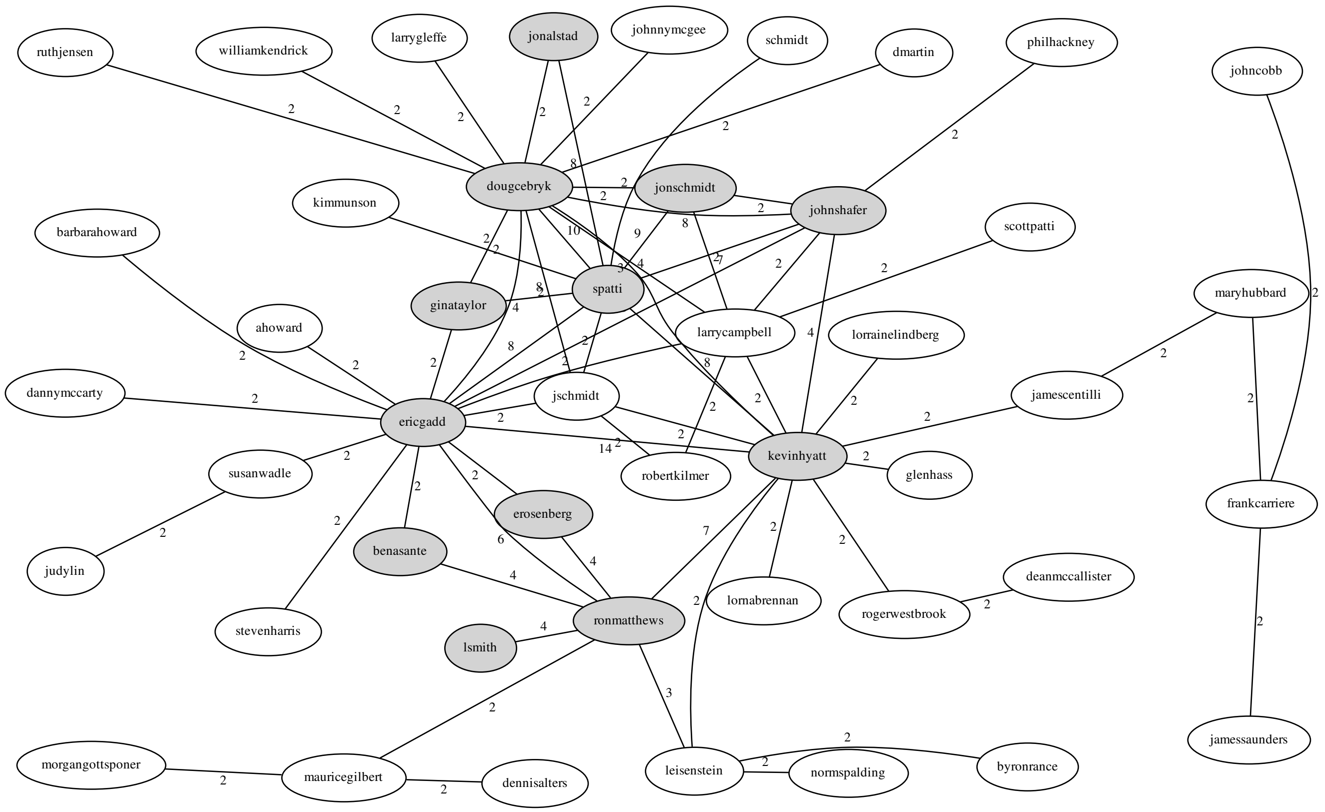


Fig. 1. Enron communities extracted with DBSCAN and the interactions-based distance function using μ = 3, ϵ = 0.3.

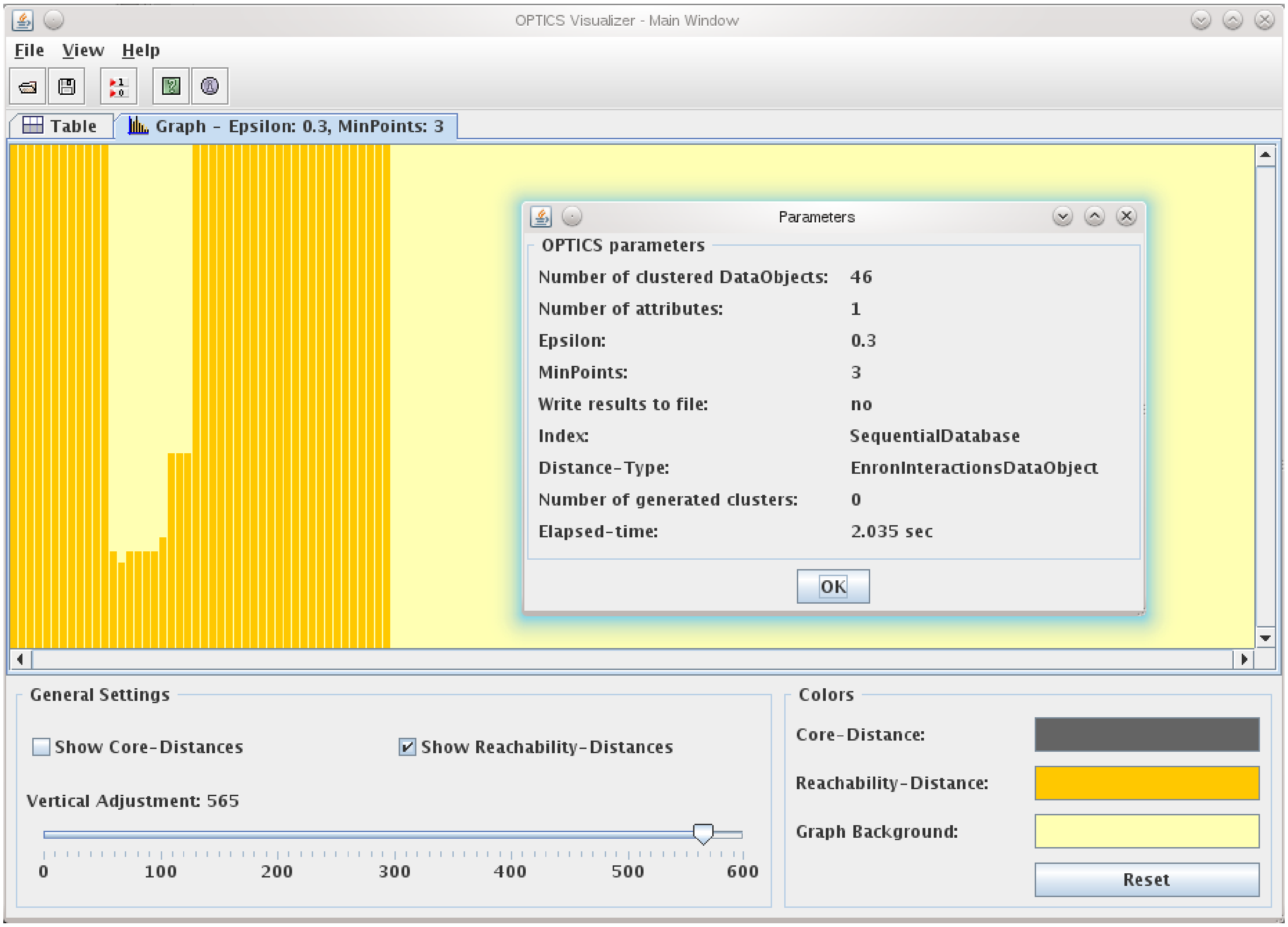
exclusively with enough members of the group. The node "ruthjensen" though, is not a part of the group even if it interacts only with "dougcebryk" from the group; the reason is it does not interact with enough people within the group. Contrary to this intuition, "scottpatti" is a member of the group. This is because, the structural distance between "scottpatti" and "larrycampel" does not exceed ϵ, whereas the structure distance between "ruthjensen" and "dougcebryk" exceeds ϵ. In particular, "dougcebryk" has more neighbors than "larrycampel" which directly affects the structure distance score.

The boundary of the community strongly depends on ϵ. If we increase ϵ (to e.g., 0.7), then peripheral nodes like "ruthjensen" will be part of the green community. This way though, the cohesiveness of the community is reduced, since such kind of members do not communicate with any other member of this community. So the choice of ϵ is important for deriving intuitive communities. Again we use OPTICS to guide us towards the appropriate ϵ selection. Figure 3 (right) shows the resulting clusters using DBSCAN for a smaller ϵ value i.e 0.4. Using a smaller ϵ we see that "ericgadd" and "kevinhyatt" among others disappeared from the cluster although they had many interactions with the members of the "green" community. The reason is that they had interactions with a lot of nodes outside the group also. This influenced the structural distance measure between them and the members of the green group. To be more specific, the geometric mean in this calculation increases resulting in a smaller σ, and thus, in a bigger distance. Therefore by reducing the ϵ value, they do not fit into the ϵ-neighborhood of the members of the green group.

B. Twitter dataset

Twitter is one of the most popular microblogging platforms nowadays with more than 175 million users. People, called twitters, can broadcast their messages, called tweets, to the public or they can address specific people in the network through retweets and replies. A retweet (RT) is an act of reposting a message from another Twitter user. A reply (@) is a direct message to another Twitter. The Twitter dataset was crawled as part of the "Web Profile Monitoring" project [10]. Twitter dataset can be modeled as an undirected graph where the *nodes* stand for the Twitters and the *edges* represent interactions between them. An interaction between two members A and B is defined as a reply or a retweet between them. The aggregated number of interactions is denoted by $I_{A,B}$. Note that there might be no edges between every pair of nodes, this is because not all twitters interact with each other.

1) Interactions-based communities: We first applied DBSCAN with the interactions-based distance (cf. Definition 1) using μ = 3 and ϵ = 0.05. The resulting clustering is depicted



OPTICS parameters

Number of clustered DataObjects:	46
Number of attributes:	1
Epsilon:	0.3
MinPoints:	3
Write results to file:	no
Index:	SequentialDatabase
Distance-Type:	EnronInteractionsDataObject
Number of generated clusters:	0
Elapsed-time:	2.035 sec

□ Show Core-Distances ☑ Show Reachability-Distances

Vertical Adjustment: 565

Fig. 2. Enron cluster ordering using OPTICS with the interactions distance function and $\mu = 3$, $\epsilon = 0.3$.

in Figure 4 (left); all disconnected sets of nodes are clusters except for the ones in grey color. Every disconnected subset is detected as a cluster; this is due to the small ϵ we chose. Thus why the cluster around "SerenaWilliams" has been merged with the cluster around "KimKardashian". This is the so called Single Link Effect [5] phenomenon, which comprises one of the major drawbacks of DBSCAN. Note that in this case only the central nodes, namely "SerenaWilliams" and "KimKardashian", are core nodes whereas the rest of them are border nodes since the density in their neighborhood is less than the threshold μ=3.

We used OPTICS to choose a suitable ϵ for DBSCAN; the cluster ordering is depicted in Figure 4 (right) where all dents represent regions with different densities. By lowering ϵ, the number of points inside the clusters decreases, so not all nodes in a disconnected set of nodes are assigned to a cluster. So, the detected communities might shrink or grow depending on the choice of ϵ.

2) Structure-based communities: We also applied DBSCAN with the structure-based distance (cf. Definition 2) using $\mu = 3$ and $\epsilon = 0.45$. The resulting clusters are depicted in Figure 5 (left) in different colors. Contrary to the intuition, none of the big subsets is recognized as a community. The reason is the large number of neighbors around the center of the set. For example, "SerenaWilliams" has more than 10 neighbours in her neighborhood and only a few of these neighbors interact with each other, thus the number of common neighbors is reduced.

The small sets of nodes detected as communities coloured in *grey* bear a larger structural similarity measure to each other in comparison to the bigger sets only because of the smaller number of neighbours around the central nodes. The clusters in *yellow* and *blue* are better representatives for a set of structurally similar nodes since the members have communicated within the community which has resulted in a more cohesive structure. The first two dents in the cluster ordering depicted in Figure 5 (right) represent the yellow and blue communities. One can observe that their reachability distance is lower than the rest of the dents. This implies that these regions are much denser which corresponds to the idea that these nodes are structurally more similar to each other. By choosing a smaller value for ϵ one may find only these

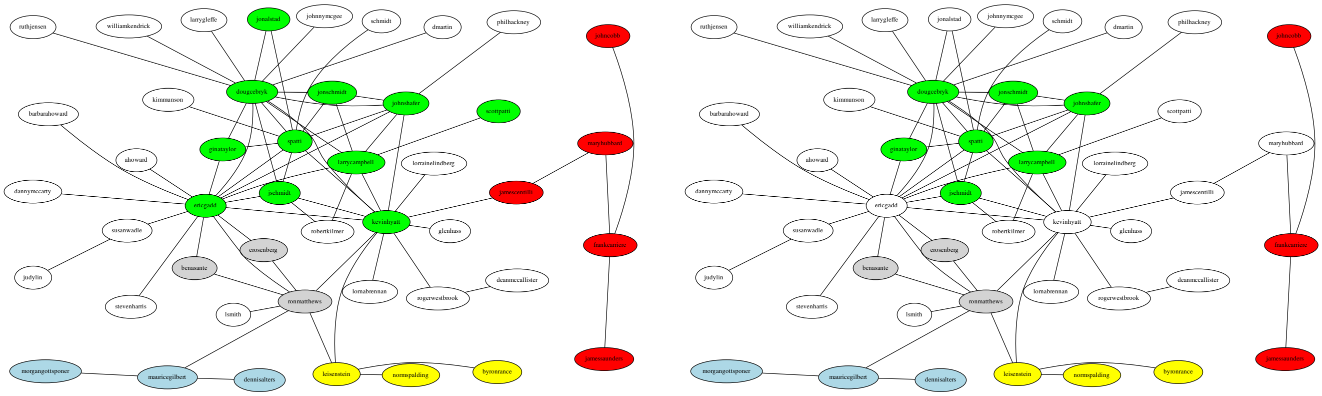
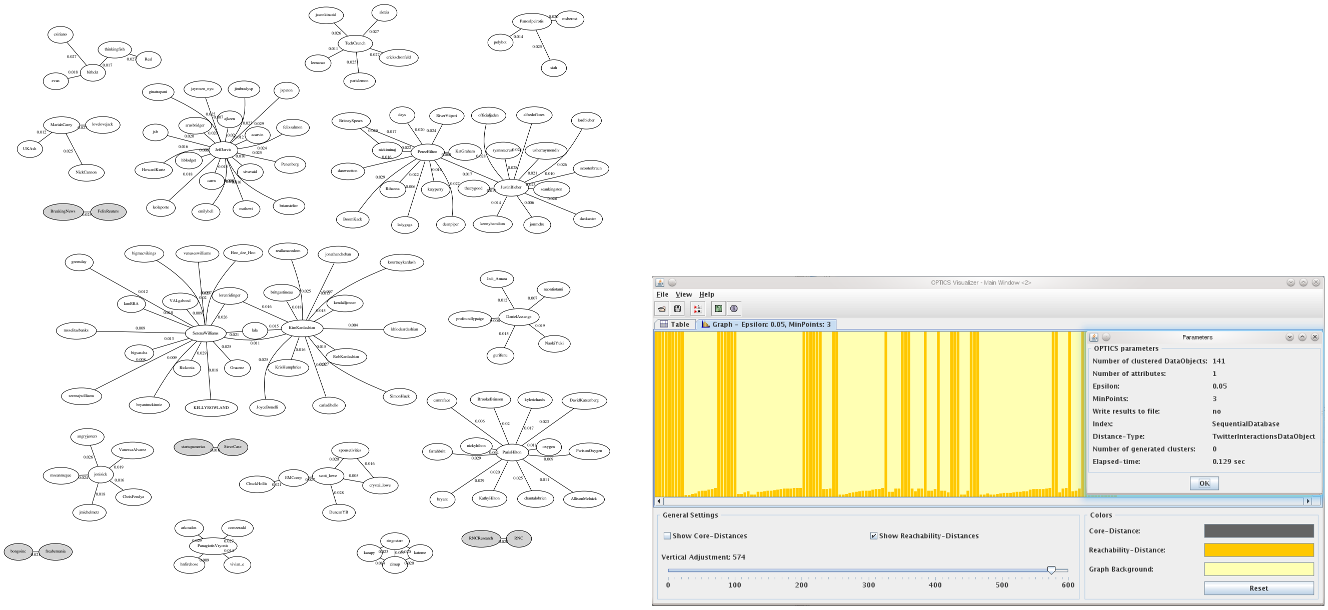


Fig. 3. Enron communities extracted with DBSCAN and the structure-based distance function using μ = 3 and ϵ = 0.5 (left), ϵ = 0.4 (right)



OPTICS Visualizer - Main Window <2>

Parameters

OPTICS parameters

Number of clustered DataObjects:	141
Number of attributes:	1
Epsilon:	0.05
MinPoints:	3
Write results to file:	no
Index:	SequentialDatabase
Distance-Type:	TwitterInteractionsDataObject
Number of generated clusters:	0
Elapsed-time:	0.129 sec

General Settings

□ Show Core-Distances ☑ Show Reachability-Distances

Vertical Adjustment: 574

Colors

Core-Distance: ■

Reachability-Distance: 🟨

Graph Background: 🟨

Fig. 4. Twitter communities extracted with DBSCAN and the interactions-based distance function using μ = 3, ϵ = 0.05.(left). The corresponding cluster ordering extracted with OPTICS (right)

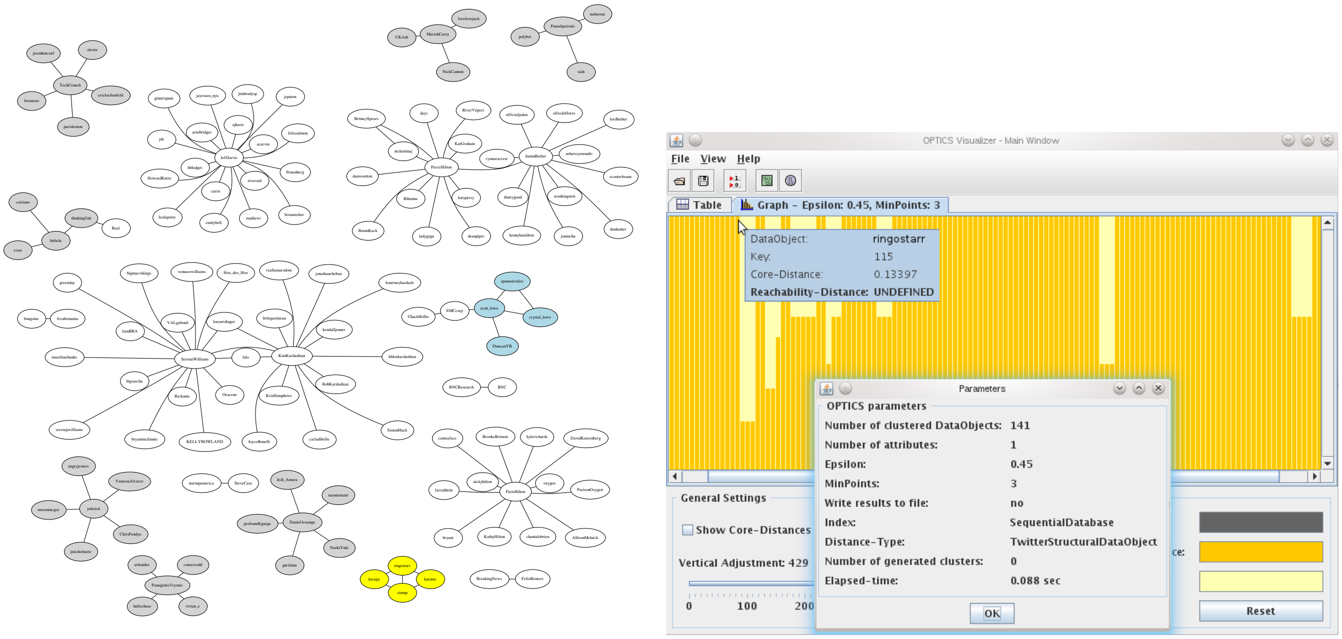
two communities.

The yellow cluster in the graph is represented by the first dent in the cluster ordering. As shown in Figure 4 (left) the member “ringostarr” is detected as a core point since the ϵ and μ values are satisfied. Being the first point of the community detected its reachability distance is undefined which expresses the fact that this node is not reachable from the preceding ones in the cluster ordering. The core distance which is evident from Figure 5 (right) is 0.13397 which is the distance to his μ^{th} (μ=3) nearest neighbour which is either “karapy” or “katome”. This distance is set as the reachability distance for the rest of the members in the community.

C. Concluding observations

In the previous subsections we have applied DBSCAN and OPTICS using both interactions-based and structure-based distance functions on two datasets of different characteristics, the Enron dataset and the Twitter dataset.

Our first finding is that the notion of density is not so intuitive in social networks, and as a result it is difficult to choose a density threshold that produces meaningful clusters. The difficulty lies in the fact that the density is defined by a distance function. Neither by combining existing distance functions nor by defining a new one will help us to fully understand the resulting notion of the density in a social network dataset. On the contrary, the augmented cluster ordering of OPTICS helped us to understand the effect of applying a specific distance function as well as the effect of varying the density threshold. In our experiments, the usage of OPTICS helped us to choose the ϵ parameter for DBSCAN. By varying ϵ, the size and the shape of the communities change. These changes are more easily perceived through OPTICS by



OPTICS Visualizer - Main Window

DataObject: ringostarr
Key: 115
Core-Distance: 0.13397
Reachability-Distance: UNDEFINED

Parameters

OPTICS parameters

Number of clustered DataObjects:	141
Number of attributes:	1
Epsilon:	0.45
MinPoints:	3
Write results to file:	no
Index:	SequentialDatabase
Distance-Type:	TwitterStructuralDataObject
Number of generated clusters:	0
Elapsed-time:	0.088 sec

Fig. 5. Twitter communities extracted with DBSCAN and the structure-based distance function using μ = 3 and ϵ = 0.45 (left). The corresponding cluster ordering by OPTICS (right)

observing the wider or deeper dents in the cluster ordering visualization. This novel idea of applying OPTICS for social network analysis, allows one to get deeper insights into the density of the given dataset under a specific distance function and to chose appropriate ϵ values.

Our second finding is related to the special characteristics of each dataset: The Enron dataset contained no disconnected sets of nodes and was structurally more cohesive in comparison to the Twitter dataset. The Twitter dataset contained a set of disconnected nodes which had a star network structure. This type of structure contained nodes that were connected to a central node and had no other connections and thus it was not cohesive. These differences in the social networks structures occur because of the inherent restrictions inside them. For example employees communicate through emails more often with colleagues in the same department than with the others in the company which leads to this cohesiveness in the Enron dataset. Twitter users broadcast tweets, this is the main idea behind this network. This fact results in connected star networks around the social network accounts w.r.t. interactions.

The application of the *interactions*-based distance function on the *Enron* data graph resulted in a single cluster. We filtered this graph in order to get only a minimum of 3 interactions between its members. Since the edges of the graph are defined according to the interactions between the nodes they connect, applying DBSCAN returned the same cluster. This is due to the fact that the chosen ϵ was smaller than the filtering value. Hence we see that the value of ϵ is directly connected to the interactions distance measure. When using the function on the Twitter dataset it resulted in communities who were disconnected from each other as seen in the graph. Variation of the ϵ value in this case resulted in a change of the size of the same communities. It is interesting to note that if two such disconnected communities happened to have a link between them at some point of time and if the weight of this link fits to the selected ϵ value then the two communities will be detected as a single community by DBSCAN (Single Link Effect [5]). Such a phenomena can occur in the full dataset of twitter This means that the semantic of different communities may be lost in this process of merging. In the worst case a large portion of the dataset will be identified as a single community. This can be accounted due to bad choice of ϵ or due to the single link effect of DBSCAN.

The application of the *structural*-based distance function on the *Enron* dataset found multiple discrete communities with a cohesive structure. This means that the people interacted more often inside their group than with the members outside the group. The notion of ϵ is more complex than in the case of interactions-based distance function. We can perceive two properties which are resulting from the definition of the structural distance function:

1) The choice of ϵ controls a node's membership to a community w.r.t. to its number of connections outside the group.
2) The choice of ϵ also has an influence on the membership of a node to the community depending on the difference between the number of neighbours and the size of the community.

On the very first sight these two properties which are responsible for the formation of communities are not easily understood in comparison to the interactions function, because of its definition is more complex and non intuitive. The fact that the periphery nodes join a community more under the influence of the second property acts as an indicator for an unsuitable choice of ϵ. Since Enron as a dataset is inherently cohesive in its nature the usage of the structural similarity distance function to detect communities in it makes sense.

A small number of communities were detected in the Twitter dataset. The majority of sets of nodes forming a star network were not detected as communities which directly reflects upon the structure of these sets of nodes as non cohesive. There were star networks with a smaller size that were detected as communities because of a non optimal ϵ value. Shifting the ϵ to a more suitable value resulted in finding truly cohesive communities. The cluster ordering of OPTICS made the choice of this ϵ value easier. Nevertheless the structural similarity distance function is rather not appropriate to be used in combination with DBSCAN for community detection in social networks like Twitter which are inherently not cohesive in nature.

V. Conclusion

Community detection is of great interest to many practical applications of social networks. This paper deals with community detection in social networks using the density based clustering paradigm and evaluates two well-known concepts for community detection based on structural similarity of nodes and the number of interactions between nodes, respectively, implemented in the distance functions of the algorithms SCAN [1] and DENGRAPH [2]. For the evaluation we used Twitter and Enron data. Both differ significantly in their network characteristics. Our analysis showed that a community definition is rather volatile and depends on the application scenario. whether an absolute definition for a community in social networks can be made at all, is an intriguing question raised by our analysis.

We examined in a systematic way the salient features of communities detected by the two distance functions. Our empirical results demonstrate that determining the clustering structure of a social network is intricate. The novel usage of OPTICS [5] in combination with the two distance functions helped us to get an insight into the cluster structure and the notion of density in these data sets. Furthermore, OPTICS eased the choice of the global parameter ϵ to overcome the density threshold problems of density based algorithms. Thus, the knowledge gained through the empirical analysis w.r.t. to the notion of density and the choice of ϵ will act as a basis for the future researchers to have a better idea about this in a social network scenario.

The Twitter data set manifested a non-cohesive, star network structure. Since Twitter acts more as a broadcast microbolgging platform it can induce such non-cohesive structures in its underlying data. When using the interactions distance function communities were found. However by varying the ϵ value one could also merge two clusters into one and hence cause a loss of its existential semantics. Using the structural similarity distance function resulted only in a very few meaningful (structurally cohesive) communities. Therefore, this distance function can be seen as a non-optimal choice to detect communities in a data set like Twitter. The Enron data set was inherently cohesive in comparison to the Twitter data. This could arise due to the restrictions present in them. A social network which inherently restricts interactions between its users leads naturally to a cohesive network. The structural similarity distance function was hence more suitable and meaningful in comparison to the interactions distance function for this data set. This was further affirmed when structurally cohesive communities (clusters) were found when using the former distance function on this data set. Furthermore, when using this function the entry of the *hubs* into the cohesive community acted as an additional indicator for the choice of an optimal ϵ value.

In summary, we can say that the structural similarity distance function is a good choice to detect communities in a cohesive network but at the same time non intuitive in its computation which makes its usage difficult. On the contrary, the interactions distance function was more simple and intuitive in its application. However, it was confined to detect meaningful communities in both data sets.

In the future, we plan to combine the two orthogonal concepts of similarity used here in order to get the best of two worlds. Combining these two notions in one distance function properly, however, is not a trivial task. However, the results in this paper showed that this in combination with hierarchical density based clustering may lead to interesting insights into the community structure of social networks.

References

[1] 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*. ACM, 2007, pp. 824–833.

[2] T. Falkowski, A. Barth, and M. Spiliopoulou, "Dengraph: A density-based community detection algorithm," 2007.

[3] R. Barker, *The social work dictionary*. NASW press Washington, DC, 2003.

[4] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan Kaufmann Pub, 2011.

[5] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '99. New York, NY, USA: ACM, 1999, pp. 49–60. [Online]. Available: http://doi.acm.org/10.1145/304182.304187

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, pp. 264–323, September 1999. [Online]. Available: http://doi.acm.org/10.1145/331499.331504

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, vol. 1996. Portland: AAAI Press, 1996, pp. 226–231.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[9] W. Cohen, "Enron email data set," urlwww-2.cs.cmu.edu/ enron/, [Online; accessed 01-Sptember-2011].

[10] V. Georgiev, "Web profile monitoring." Department of Computer Science Database Systems Group, LMU Munich, 2011.