Semantic Clustering of Social Networks using Points of View

Juan David Cruz* — **Cécile Bothorel*** — **François Poulet****

**Département LUSSI, Télécom – Bretagne*
{juan.cruzgomez, cecile.bothorel}@telecom-bretagne.eu

***Université de Rennes 1 – IRISA*
francois.poulet@irisa.fr

ABSTRACT. *Classic algorithms for community detection in social networks use the structural information to identify groups in social networks, i.e., how clusters are formed according to the topology of the relationships. However, these methods do not take into account any semantic information which could guide the clustering process, and which may add elements to do further analyses. The method we propose, uses in a conjoint way, the semantic information from the social network, represented by the points of view, and its structural information. This information integrates the relationships, expressed by the edges on one hand, and the implicit relations deduced from the semantic information on the other hand.*

RÉSUMÉ. *Les algorithmes classiques de détection de communautés dans les réseaux sociaux utilisent l'information structurelle pour détecter des groupes, i.e la topologie du graphe de relations. Toutefois, ils ne prennent en compte aucune information externe qui peut guider le processus et aider à la réalisation des analyses du réseau selon différentes perspectives. La méthode proposée utilise de façon conjointe, l'information sémantique du réseau social, représentée par des points de vue, et son information structurelle. Elle permet la combinaison entre les relations sociales explicites, les arêtes du graphe social, et les relations implicites, dites sémantiques, correspondant par exemple à des intérêts ou des usages similaires.*

KEYWORDS: *Social Network Analysis, Socio–semantic Networks, Communities Detection, Graph Clustering, Self Organizing Maps*

MOTS-CLÉS : *Analysis de Réseaux Sociaux, Réseaux Socio-sémantiques, Détection de Communautés, Clustering des graphes, Self Organizing Maps*

1. Introduction

A social network is composed of a group of actors linked according to different types of relationships. However, a social network contains more information than only links and actors: since actors can be persons and organizations, they have additional semantic information which enriches that network. By using the semantic information, social network analyses could be performed from different perspectives, not only from the structural one.

Hence, we propose a method which combines information from the network topology and from the actors, or the network's semantic information. This semantic information can be divided into subsets of information, called *points of view*. Then, a point of view can be defined as an ensemble of features which represents a state of the network under a given perspective and can be used to guide, in this case, the communities detection process.

For example, in a social network composed of the employees from an enterprise, and their links being whether they have sent messages to each other or not, it is possible to define a point of view as the type of projects of each employee has been involved into. Thus, using this point of view could be used to find communities of people which have met before and have worked in projects with similar profiles. This could be used to find experts or to create work teams in the organization.

The paper is organized as follows. In Section 2 is presented some previous work in community detection in social networks, in Section 3 we present the definition of the point of view of social networks; in Section 4 we present the proposed clustering method. In Section 5 some experiments and preliminary results are presented before the conclusion and future work.

2. Related Work

Several methods have been developed to find clusters in a graph, or which is equivalent, to find communities in a social network. In general, those methods have been defined as optimization problems where the objective function is the maximization of some quality index. The indices measure the quality of a partition $\mathbf{C}$ based on the number of edges within the cluster and the number of inter–cluster edges.

(Gaetler, 2005) and (Brandes *et al.,* 2008) define three quality indices: the *coverage*, which measures the weight of all the intracluster edges compared to the weight of all edges within the graph; the *conductance*, which is based on the observation that if a cluster is well connected, then, a large number of edges have to be removed in order to bisect it, and the *performance*, which defines the quality of a given cluster based on the "correctness" of the classification of a pair of nodes. Additionally, another index, the modularity Q, proposed by (Newman *et al.,* 2004), compares the fraction of the edges within each cluster with the fraction of edges among clusters, i.e., the intracluster edges density versus the inter-cluster sparsity. This index is the most commonly used in the different clustering methods as presented by (Fortunato, 2010).

The classic graph clustering algorithms aim to find groups optimizing one of the indices shown above. These approaches can find better partitions when the adjacency matrix of the graph is sparse (Fortunato, 2010).

The algorithm proposed by (Newman, 2001) iteratively finds and removes the edge with the highest betweenness score. This process allows to find groups which are loosely connected between them and with well connected nodes within the group. The main drawback of this approach is the complexity of the calculation of the betweenness, the general algorithm will take $O\left(mn^2\right)$ for m edges and n nodes, its cost for huge graphs is prohibitive.

The fast unfolding algorithm, proposed by (Blondel *et al.,* 2008), is an agglomerative algorithm to find communities. In the first step each node is assigned to one community and the initial modularity is calculated. Then, each node i is removed from its community and moved iteratively to each community. After each movement the modularity gain is calculated, and i will be assigned to the community giving the largest positive gain. If no positive gain is possible, i remains in its original community. This process is applied iteratively until no further improvement can be achieved and no individual move will improve the modularity. This algorithm is executed in linear time for sparse graphs (Blondel *et al.,* 2008).

(Du *et al.,* 2007) present an algorithm to detect communities in large–scale social networks. Their method is based on the enumeration of all the maximal cliques, i.e., a complete subgraph which is not contained in any other complete subgraph. After all the maximal cliques are enumerated, they generate kernels associated to those cliques and then, perform the community detection by assigning nodes to each kernel. After this, they try to optimize the modularity obtained by moving nodes accordingly.

Most of the classic algorithms find disjunct partitions. However, most of the social networks from the real world may contain actors belonging to more than one community. For example, (Pizzuti, 2009) presents a method for detecting overlapped communities. This method uses a genetic algorithm with a fitness function which minimizes the relation between the edges within each group and the edges connecting nodes outside each group.

Other clustering methods, such as Markov Clustering, Iterative Conductance Cutting and geometric minimum spanning tree, are discussed in (Brandes *et al.,* 2008), and some methods for evaluating communities are presented by (Kwak *et al.,* 2009) and by (Günter *et al.,* 2003). In general classical methods take only into account the structural configuration of the graph: they do not use any information associated to the nodes.

3. Defining the Point of View of Socio–Semantic Networks

Socio–semantic networks contain an important amount of information, coming not only from their topology, but from the different contexts in which such networks have evolved. This information can be seen as features associated to the actors and to the relationships in the network, and give more elements to analyze a network from different perspectives.

3.1. *Some Notations*

Given an undirected graph $G\left(V, E\right)$ representing a social network, where V is the non-empty set of vertices, representing actors and E is the set of edges representing the relationships among them. Let v_i and v_j be two vertices from V and let $e\left(x, y\right)$ be the edge defined by the vertices x and y. Thus, if $e\left(v_i, v_j\right) \in E$ then v_i and v_j are neighbors. A partition $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ is a partition of the set V into k non–empty disjoint subsets C_i. Let F_V be the set of features of the actors of the social network, which can be represented by a matrix of size $|V| \times |F_V|$.

3.2. *Representation of Point of View*

Given a semantic network $\mathcal{S} = \langle G, F_V \rangle$, let $F_V^* \in \mathcal{P}\left(F_V\right) \setminus F_V$, where $\mathcal{P}\left(A\right)$ is the powerset of the set A, be a set of features to define the point of view PoV. Thus, for each vertex $v_i \in V$ there is assigned a binary vector ξ_i of size $|F_V^*| = f$. If the vertex i has the feature $p, 1 \leq p \leq f$ from F_V^*, then $\xi_{i,p} = 1$ or 0, otherwise. Then, each binary vector ξ can be defined as $\xi_i = v_i \times F_V^*$, where $v_i \in V$. Then, a point of view is defined as the set of the union of all instances derived from the set F_V^*:

$$PoV_{F_V^*} = \bigcup_{i=1}^{|V|} \xi_i \qquad [1]$$

Note that different nodes could have the same instance ξ.

4. Using the Point of View to Influence the Clustering

By merging the semantical and the structural information it is possible to guide the graph clustering process by adding information related to the similarity of the nodes in a real context. To do this, the community detection process is divided into two phases. During the first one, the point of view is clustered using Kohonen maps (Kohonen, 1997) to obtain groups based on the similarity of the node features. Thus, the groups found in the first phase are used to change the weight of the edges in the graph, and then, in the second phase, a classic community detection algorithm is used.

4.1. *Phase 1: Semantic Clustering*

Given a point of view derived from a set F_V^* defined as in Section 3, each node can be characterized by its vector of features, or an instance ξ of the point of view. Using each each instance ξ is an input pattern for the training, the SOM will group the nodes according to the similarities of their features.

The SOM network $\mathcal{N}$ has been implemented using a square lattice of $l \times l$ neurons, where $l = |F_V^*|$ is the number of features in the point of view. The initial values of the weights of the SOM are randomly selected.

The complexity of this algorithm is proportional to the number of features in the point of view, the number of instances and the size of the neural network. It can be expressed as $T = O\left(f^3 \cdot n\right)$, where n is the number of nodes of the graph and f the number of features in the point of view. The outcome of the algorithm is a partition $\mathcal{C}_{SOM}$ (recall partition definition from Section 3.1) of the nodes assigned to the neurons.

4.2. *Phase 2: Structural Clustering and Community Detection*

Once the semantic partition $\mathcal{C}_{SOM}$ has been found it is possible to begin the second phase of the proposed method. During this phase we use a classic graph clustering algorithm to find communities, specifically, the fast unfolding algorithm, proposed by (Blondel *et al.,* 2008) and presented in Section 2. This algorithm uses the modularity Q, presented by (Newman *et al.,* 2004) as quality measure.

Before the execution of the fast unfolding algorithm, we include the information from the phase 1. This is performed by changing the weights of the edges according to the partition $\mathcal{C}_{SOM}$. Thus, for each pair of neighbor vertices $v_i, v_j, \forall i \neq j \in V$, the weight of the edge $e\left(v_i, v_j\right)$ is changed according with the Euclidean distance of the PoV instances corresponding to each node by:

$$w_{ij} = 1 + \alpha \left(1 - d\left(\mathcal{N}_{ij}\right)\right) \delta_{ij} \qquad [2]$$

where $\alpha \geq 1$ is a constant value, $d\left(\mathcal{N}_{ij}\right)$ is the distance between the neurons i and j, and $\delta_{ij} = 1$ if v_i and v_j belong to the same partition in $\mathcal{C}_{SOM}$, $\delta_{ij} = 0$ otherwise.

After the weights are changed according to Equation 2, a partition $\mathcal{C}_{SOM \rightarrow FU}$ is found using the Fast Unfolding algorithm. This partition contains the final set of communities, which has both, the semantic information and the structural information.

Since our approach adds a preprocessing layer in order to find a semantic partition $\mathcal{C}_{SOM}$, the complexity is given in function of the complexity of the semantic clustering, as explained in Section 4.1, plus the complexity of the fast unfolding algorithm which has been reported to be linear in the number of nodes for sparse adjacency matrices (Blondel *et al.,* 2008). Thus, the overall complexity of our method for a defined F_V^* is $O\left(|F_V^*|^3 \cdot |V|\right)$.

5. Preliminar Experiments

Some preliminary experiments were developed using one graph and two points of view generated from the semantic information contained in the data–set, which is an extract from Twitter. The graph used in experiments is composed of 5389 nodes and 27347 edges, and has an initial modularity of -2.5192×10^{-3}.

For the experimentation we compare the clustering results of two classic clustering algorithms and our proposed method. The first classic algorithm is the SOM, it finds partitions based only on the semantic information, denoted by $\mathcal{C}_{SOM}$, the second classic algorithm, fast unfolding, it finds communities using only the structure of the graph, denoted by $\mathcal{C}_{FU}$, and our proposed method denoted by $\mathcal{C}_{SOM \to FU}$.

To measure the result of the experiments, we use the average Euclidean distance within the groups obtained calculating for each pair of nodes from a group, the distance between the instance of the point of view assigned to each one and the modularity Q to evaluate, from a structural perspective, the obtained partition. The idea is to minimize the distance within groups and to maximize the modularity.

The points of view used in the experiments are:

1) **Time zone division:** it is composed of 33 features representing the different world time zones, including the non–standard ones, registered in the data set. These time zones can be regarded as the general geographical distribution of the friends of some Twitter user. Thus, each instance is described by the presence of friends in each time zone.

2) **User profile:** The first feature indicates if the user follows more people, or in the twitter environment, has more friends, than followers. Users who have more followers than friends are usually people, or organizations, which have a lot of people interested in their updates and messages. This is the case of politicians and public figures. The next three features indicate the user behavior according to the number of messages sent. Thus, the features are: below the mean, between the mean plus three standard deviations and, over mean plus three standard deviations. In this data set nearly 82% of users are below the mean of the messages sent.

Experiments were executed using a graph of 5389 nodes and 27347 edges extracted from a Twitter data set, composed of ~ 204000 nodes and ~ 326000 edges.

The result of the experiments are reported in Table 1. The average intracluster distance found by our proposed method is less than the average intracluster distance found by the graph based algorithm.

For the point of view PoV_1, the modularity obtained by the classical graph clustering algorithm an by our approach is very similar. This is due to the structure of the

PoV	**Experiment**	**Final Q**	**Average Intracluster Distance**	**Standard dev.**
PoV_1	$\mathcal{C}_{SOM}$	-0.0075	0.3697	0.1059
	$\mathcal{C}_{FU}$	0.5728	1.8091	1.3584
	$\mathcal{C}_{SOM \to FU}$	0.5747	1.1947	0.8489
PoV_2	$\mathcal{C}_{SOM}$	-0.2991	0	0
	$\mathcal{C}_{FU}$	0.5728	0.7100	0.6565
	$\mathcal{C}_{SOM \to FU}$	0.6351	0.5507	0.5577

Table 1. *Results of the experiments performed comparing the result of the classic algorithms versus the proposed method.*

point of view, which uses information associated with the geographic localization of the friends of each actor. We may think here that friendship tends to be similar when considering friends.

For this point of view PoV_2, the SOM clustered the nodes into six groups, each one expressing one of the possible instances. Creating a graph from the SOM clustering will produce better semantic clusters, however, the modularity is worst than the one from the original graph. This shows that the SOM groups are totally unrelated with the structure of the graph.

In the case of the graph based clustering and the PoV based clustering the results are different. The performance of the PoV based algorithm was better according to the modularity and the average intracluster distance.

6. Conclusion and Future Work

The classic community detection algorithms use information only from the network structure and do not take into account the semantic information, which could be used to influence the clustering process.

Assigning the weights derived from the results of the semantic clustering to the edges, the semantic information is included into the community detection process and the two types of informations are merged to find and visualize a social network from a selected point of view.

Regarding the execution time of our method, the complexity is higher than the complexity for the graph based one. Today, this imposes some restrictions in the number of features. The sensibility of the execution time to the number of features is high because of the SOM training.

The high number of dimensions may mislead the SOM training because of the Hughes effect (Hughes, 1968), also known as the curse of dimensionality, and how the semantic distance is measured. Hence, we will study the statistical properties of the points of view to try to reduce this effect.

For future work we will also continue the study of the influence of the point of view in the community detection process including the definition of points of view from the graph's edges. Additionally, we plan to work on the development of a visualization algorithm for hierarchical social networks.

7. References

Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E., « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, n° 10, p. P10008 (12pp), 2008.

Brandes U., Gaetler M., Wagner D., « Engineering graph clustering: Models and experimental evaluation », *Journal of Experimental Algorithmics*, vol. 12, p. 1-26, 2008.

Du N., Wu B., Pei X., Wang B., Xu L., « Community detection in large-scale social networks », *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, New York, NY, USA, p. 16-25, 2007.

Fortunato S., « Community detection in graphs », *Physics Reports*, vol. 486, n° 3-5, p. 75 - 174, 2010.

Gaetler M., *Network Analysis: Methodological Foundations*, Springer Berlin / Heidelberg, chapter Clustering, p. 178 - 215, 2005.

Günter S., Bunke H., « Validation indices for graph clustering », *Pattern Recognition Letters*, vol. 24, n° 8, p. 1107-1113, 2003.

Hughes G. F., « On the mean accuracy of statistical pattern recognizers », *IEEE Transactions on Information Theory*, vol. 14, n° 1, p. 55-63, 1968.

Kohonen T., *Self-Organizing Maps*, Springer, 1997.

Kwak H., Choi Y., Eom Y.-H., Jeong H., Moon S., « Mining communities in networks: a solution for consistency and its evaluation », *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, ACM, New York, NY, USA, p. 301-314, 2009.

Newman M. E., « Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. », *Physical Review. E, Statistical Nonliner and Soft Matter Physics*, vol. 64, p. 7, July, 2001.

Newman M. E. J., Girvan M., « Finding and evaluating community structure in networks », *Physical Review. E, Statistical Nonliner and Soft Matter Physics*, vol. 69, n° 2, p. 026113, Feb, 2004.

Pizzuti C., « Overlapped community detection in complex networks », *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, ACM, New York, NY, USA, p. 859-866, 2009.