Point of View Based Clustering of Socio-Semantic Networks

Juan David CRUZ[1]

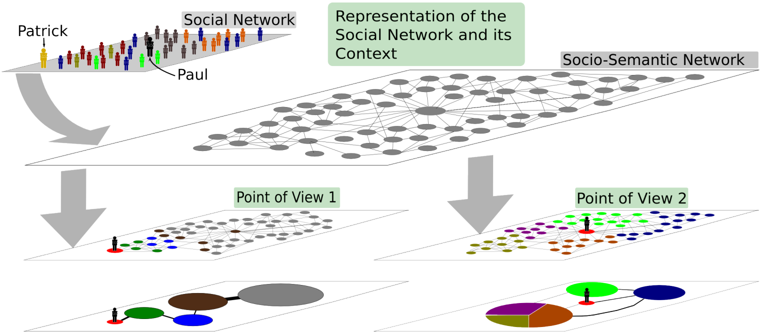
Cécile BOTHOREL[1]

François POULET[2]

Séminaire ComplexNetworks – LIP6

Outline

Social Networks and Points of View



It is possible to obtain different partitions from different points of view

Social Networks and Points of View (Example)

The following examples use information of a social network created from a Twitter data set.

- **Point of view 1:** The time zone distribution of the neighbors of each actor in the network.
  - For each actor in the network there is an unique time zone value wich represents the meridian in it is located. For the whole network there is a finite set Z of existent time zones.
  - Given an actor a, its neighbors can be assigned to one or more of the time zones contained in Z.
  - Let a_Z be the assignation vector of a over the time zone set. Thus, $a_{Z_i} = 1$ iff a has a neighbor in the time zone i, 0 otherwise.
  - Example: $Z = \{-8, -5, 0, +1, +3\}$, $a_Z = [0, 1, 0, 1, 1]$.

Social Networks and Points of View (Example)

The following examples use information of a social network created from a Twitter data set.

- **Point of view 2:** The messaging profile of each actor in the network.
  - Each actor in the network sends messages over the network to inform or comment something.
  - Each actor has a number of followers and a number of persons being followed by him (friends).
  - $Z_0 = 1$ if the actor has more friends than followers.
  - $Z_1 = 1$ if the number of messages (n) sent by actor is less than the total average. $Z_2 = 1$ if $\mu \leq n < 3\sigma$. $Z_3 = 1$ if $n \geq 3\sigma$.

Motivation

- Socio–semantic networks contains both:
  - The social graph (structural information)
  - Semantic information represented by the features of the vertices and the edges.
- By the combination of both it is possible to make analyses from different perspectives.
- Given this information, how to identify communities derived from the conjoint use of it?
- It is necessary to measure the quality of the partitions found using this information in two levels:
  - The quality of the graph clustering
  - The quality of the semantic information within the communities

Quality Measures

Type	Objective	Examples
Similarity	Reduce the distance between the members of the same group while the distance between groups is increased.	Manhattan L_1 **Euclidean** L_2 Chebyshev L_∞
Quality	Increase the number of edges within each community while the number of edges between communities is reduced. In general: $index(\mathbf{C}) = \frac{f(\mathbf{C})+g(\mathbf{C})}{N(G)}$ [1]	Coverage Conductance Performance **Modularity**

Graph Clustering Algorithms

Several graph clustering algorithms have been developed, among others:

- Newman [2] (Modularity optimization)
- Fast unfolding [3] (Modularity optimization)
- Maximal cliques enumeration and kernel generation [4] (Modularity optimization)
- Genetic algorithm for detecting communities in large graphs [5] (Fitness function based on modularity)
- Genetic algorithm for detecting overlapped communities [6] (Fitness function based on internal edges vs. outgoing edges)

General Notation

- Given an undirected graph $G(V,E)$ with a set V of vertices and E of edges:
  - Let $\mathscr{C} = \{C_1, C_2, \ldots, C_k\}$ be a partition which is a division of the set V into non–empty, disjoint subsets C_i.
  - Let$\mathbf{F}_V$ be the set of features of the actors of the social network.
  - Let $\mathbf{F}_E$ be the set of features associated to each edge.
- Let $F_V \in \mathscr{P}(\mathbf{F}_V) \setminus \mathbf{F}_V$, where $\mathscr{P}(A)$ is the powerset of the set A.
- Each vertex $v_i \in V$ there is assigned a binary vector (instance) ξ_i of size $\|F_V\| = f$ and defined by:

$$\xi_i = v_i \times F_V$$

The Representation of a Point of View

- The point of view is the set of all the instances derived from a given F_V:

$$PoV_{F_V} = \bigcup_{i=1}^{\|V\|} \xi_i$$

	Point of View			
Nodes	**Feature 1**	**Feature 2**	...	**Feature f**
Node 1	1	0	...	0
Node 2	0	1	...	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
Node n	1	0	...	1

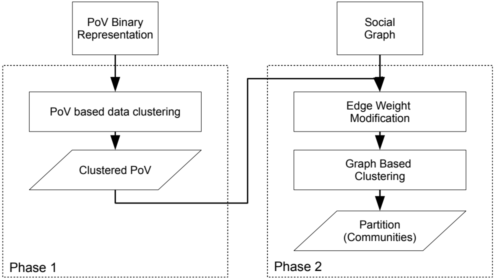
- The assignation of features to each node in the network

- We will use a simple example to show the different steps of the algorithm.
- For this example we use:
  - An undirected graph G with 29 nodes and 90 edges.
  - A point of view composed of view of three features:
  -

	Feature 1	Feature 2	Feature 3
1	0	0	0
2	0	0	1
⋮	⋮	⋮	⋮
29	1	1	0

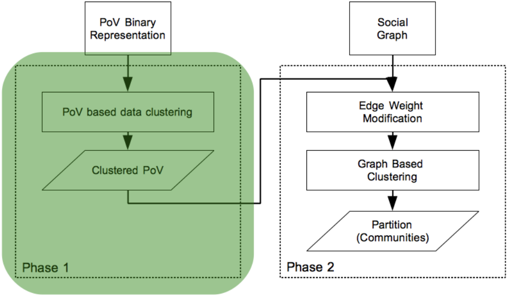
General Architecture

- Guide the community detection algorithm according to semantic information.



- Use of clustering techniques from different domains.

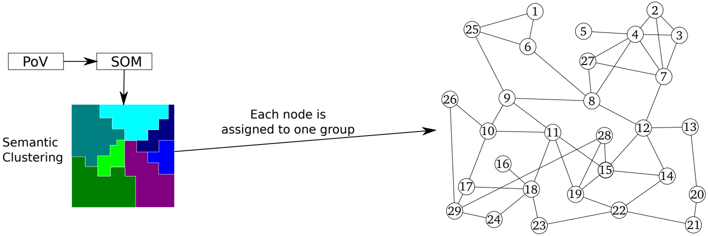
Semantic Clustering



Semantic Clustering

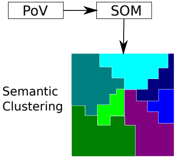
- Clustering of the defined point of view: search nodes with similar instances of features.
- Use of Self–Organizing Maps (SOM): non–supervised machine learning method [7].
- The proximity between the input vector (instance) and the weight vector of the network is measured with the Euclidean distance.
- The SOM algorithm will find some number of groups.

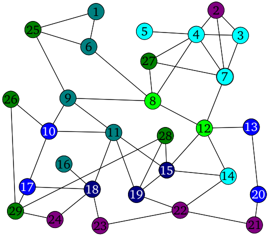
Example



- The SOM will group the nodes according to their instances, i.e., according to their semantic similarity.

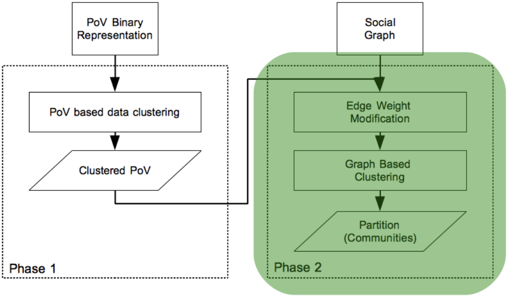
Example





- The SOM will group the nodes according to their instances, i.e., according to their semantic similarity.
- Note that there are nodes which are semantically close but not are not even neighbors.

Weights Assignation and Community Detection



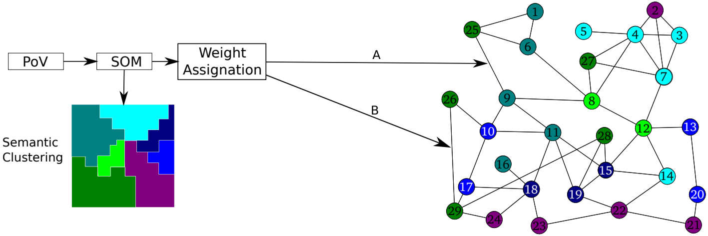
Weights Assignation and Community Detection

- Given the trained SOM network $\mathscr{N}$ and a graph $G(V,E)$:
- For each $e(i,j) \in E$, the weight will be changed according to:

$$w_{ij} = 1 + \alpha\left(1 - d\left(\mathscr{N}_{ij}\right)\right)\delta_{ij}$$

  where $\alpha \geq 1$ is constant parameter, $d\left(\mathscr{N}_{ij}\right)$, is the distance between the node i and the node j in the SOM network and $\delta_{ij} = 1$ if i,j belong to the same group in the SOM network.
- After the weights are set, a classic graph clustering algorithm (the fast unfolding algorithm [3]) is used.

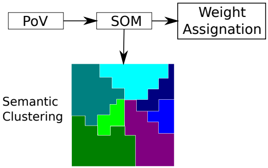
Example

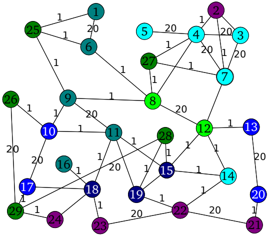


Using equation 2 with $\alpha = 19$:

- Case A $e(9,25)$:
  - Node 9 belongs to a different group than 25.
  - $w_{9,25} = 1$
- Case B $e(26,29)$:
  - Node 26 belongs to the same group than 29.
  - $w_{26,29} = 20$: The distance between the node 26 and the node 29 in the SOM network is 0.

Example

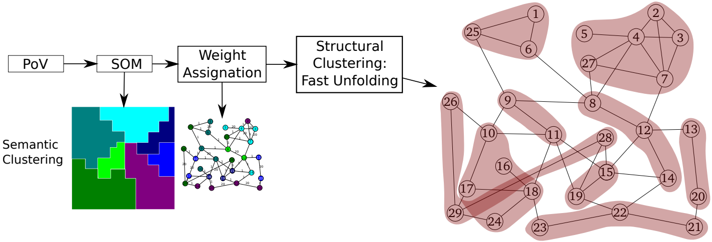




Using equation 2 with $\alpha = 19$:

- Case A $e(9, 25)$:
  - Node 9 belongs to a different group than 25.
  - $w_{9,25} = 1$
- Case B $e(26, 29)$:
  - Node 26 belongs to the same group than 29.
  - $w_{26,29} = 20$: The distance between the node 26 and the node 29 in the SOM network is 0.

Example



- After the weights are changed, the fast unfolding algorithm is used to find the communities.
- This algorithm is influenced by the assignation of weights according to the semantic clustering.
- This way structural and semantic information are used to find communities.
- The final communities are those surrounded in red.

Experiments Configuration I

- In each experiment three algorithm were compared:
  - SOM
  - Fast unfolding
  - Our method
- Performed in two levels:
  - The final modularity: to measure the quality of the partition.
  - The average intra–cluster Euclidean distance: to measure the quality of the semantic clustering.
- The experiment were executed using a graph of 5389 nodes and 27347 edges extracted from a Twitter data set. The initial modularity of this graph is -2.5192×10^{-3}

Experiments Configuration II

- The experiments were performed using two different points of view.
- Point of View 1:
  - Composed of 33 features. Each feature represents a time zone from the Twitter data set.
  - A feature will be set to 1 if the node has at least one friend in the time zone represented by the feature.
  - Distances vary from 0 to $\sqrt{32}$

Experiments Configuration III

- Point of View 2:
  - Composed of 4 features representing the messaging profile of each user.
  - The first feature is set to 1 if the user has more friends than followers.
  - The next three features indicate the user behavior according to the number of messages sent: below the mean, between the mean plus three standard deviations and, over mean plus three standard deviations.
  - Distances vary from 0 to $\sqrt{3}$

TELECOM Bretagne

Case Twitter – Point of View 1

Experiment	Final Q	Avg. Intracluster Distance
SOM Graph	-7.5×10^{-3}	0.3697
Graph based clustering	0.5728	1.8091
PoV based Clustering	0.5747	1.1947

- The average intracluster distance found by our proposed method is less than the average intracluster distance found by the graph based algorithm.
- The modularity obtained is very similar: the point of view uses information associated with the localization of people's friends.
- The modularity of the graph from the SOM clustering is not very different from the modularity of the original graph.
- SOM groups are close to the structure of the non–clustered graph.

Case Twitter – Point of View 2

Experiment	Final Q	Avg. Intracluster Distance
SOM Graph	-0.2991	0
Graph based clustering	0.5728	0.7100
PoV based Clustering	0.6351	0.5507

- The SOM clustered the nodes into six groups, each one expressing one of the possible instances described above. This explains the average distance found.
- Creating a graph from the SOM clustering will produce better semantic clusters, however, the modularity is worst than the one from the original graph.
- The SOM groups are totally unrelated with the structure of the graph.
- Regarding the modularity and the average intracluster distance, the performance of the PoV based algorithm was better.

Conclusions and Future Work I

- The classic community detection algorithms do not take into account the semantic information to **influence the clustering process**.
- Changing the weights according to the results of the semantic clustering, the semantic information is **included into the community detection process**.
- The two types of informations are **merged** to find and visualize a social network **from a selected point of view**.

Conclusions and Future Work II

- Future work
  - Make tests over the obtained partitions: rand index, robustness tests...
  - Study the case of overlapping communities.
  - Include the features of the edges into the point of view generation.
  - Development of a visualization algorithm for representing the PoV and the transition between two points of view.

Thanks for your attention

Questions?

Contact: juan.cruzgomez@telecom-bretagne.eu

For Further Reading I

- M. Gaetler, *Network Analysis: Methodological Foundations*, ch. Clustering, pp. 178 – 215. Springer Berlin / Heidelberg, 2005.
- M. E. Newman, "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.
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008 (12pp), 2008.

For Further Reading II

- N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, (New York, NY, USA), pp. 16–25, ACM, 2007.
- M. Lipczak and E. Milios, "Agglomerative genetic algorithm for clustering in social networks," in *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, (New York, NY, USA), pp. 1243–1250, ACM, 2009.

For Further Reading III

- C. Pizzuti, "Overlapped community detection in complex networks," in *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, (New York, NY, USA), pp. 859–866, ACM, 2009.
- T. Kohonen, *Self-Organizing Maps*.
  Springer, 1997.