Topologically Ordered Graph Clustering via Deterministic Annealing

Fabrice Rossi[1] and Nathalie Villa[2]

1– Institut TELECOM, TELECOM ParisTech, LTCI - UMR CNRS 5141
46, rue Barrault, 75013 Paris – France

2– Institut de Mathématiques de Toulouse, Université de Toulouse,
UMR CNRS 5219, 118 route de Narbonne, 31062 Toulouse cedex 9 – France

Abstract. This paper proposes an organized generalization of Newman and Girvan's modularity measure for graph clustering. Optimized via a deterministic annealing scheme, this measure produces topologically ordered graph partitions that lead to faithful and readable graph representations on a 2 dimensional SOM like planar grid.

1   Introduction

Large and complex graphs are natural ways of describing real world systems that involve interactions between objects: persons and/or organizations in social networks, articles in citation networks, web sites on the world wide web, proteins in regulatory networks, etc. [8]. However, the complexity of real world graphs limits the possibilities of exploratory analysis: even one hundred nodes are difficult to display in a meaningful way. As pointed out in [10], a possible solution for this problem is provided by graph clustering methods [12]: the graph is first clustered into a simplified graph and then rendered via standard graph visualization methods [4]. We explore in this paper a more direct approach, inspired by our previous work on social network analysis [1]. We build a topologically ordered clustering of the graph which is displayed on a 2 dimensional Self-Organizing Map like planar grid. Our method is based on an extension of the well known modularity measure [10]: Section 2 recalls its definition and introduces its topologically organized generalization. Section 3 presents the deterministic annealing scheme used to optimize the organized modularity. Finally, Section 4 gives some experimental results obtained on a real world coauthorship network.

2   Graph clustering quality measures

2.1   Modularity

Given a non oriented graph of size N described by its (symmetric) weight matrix, W, we denote $k_i = \sum_j W_{ij}$ the degree of node i. The total weight of the graph is denoted $m = \frac{1}{2}\sum_{i,j} W_{ij}$. Let us consider a partition of the graph into C clusters given by the $N \times C$ assignment matrix M such that $M_{ik} \in \{0, 1\}$ and $\sum_k M_{ik} = 1$ ($M_{ik} = 1$ when the node i is assigned to cluster C_k). A quality measure for such a partition is the **modularity** [10] given by $Q(M) = \frac{1}{2m}\sum_{i,j}\sum_k M_{ik}M_{jk}(W_{ij} - P_{ij})$ where P is a size N square symmetric matrix

defined by $P_{ij} = \frac{k_i k_j}{2m}$. The rationale of the measure is to compare the weight of a link between two nodes in a cluster, W_{ij}, to a simple random model, P_{ij}, in which the weights are proportional to the degrees of the nodes and independent of the clusters. A good partition tends to cluster nodes that are more connected that one expects based solely on the degrees of the nodes: this corresponds to $W_{ij} > P_{ij}$ and to higher values of $Q(M)$.

2.2 Organized modularity

We propose a generalization of the modularity inspired by the Self-Organizing Map (SOM) principle. Our goal is to cluster a graph into a high modularity partition that can be represented faithfully on a two dimensional plane. As in the SOM, the positions of the clusters in the representation space is chosen by the user, e.g., on a regular grid. Then a clustered graph is displayed using a glyph for each cluster at the specified position and segments between glyphs that summarize the edges between the clusters' elements. For instance, a partition in C clusters $(C_k)_{1 \leq k \leq C}$, can be represented by C squares with areas proportional to $|C_k|$ and with a segment of thickness s_{kl}, proportional to $\sum_{i \in C_k, j \in C_l} W_{ij}$, between squares k and l.

Such a representation is faithful and readable only for a well chosen partition. Firstly the clusters must have a limited substructure: the subgraph with nodes in C_k should be as complete as possible. Secondly, the connection between nodes from different clusters must be as limited as possible. Finally, long distance connections must be prohibited to avoid segment crossing as this is a well known major readability issue in graph drawing [2, 4]. The two first criteria are controlled by modularity optimization. We propose to take into account the last one via an **organized** version of the modularity.

We assume given a prior structure in $\mathbb{R}^2$ which is represented by a symmetric C by C matrix S of prior similarities between clusters. For instance $S_{kl} = \exp(-\sigma \|\mathbf{x}_k - \mathbf{x}_l\|^2)$ where $\mathbf{x}_k$ is the prior position of cluster C_k in $\mathbb{R}^2$. Then we propose to maximize $O(M) = \frac{1}{2m} \sum_{i,j} \sum_{k,l} M_{ik} S_{kl} M_{jl} (W_{ij} - P_{ij})$. In the standard modularity, the term $W_{ij} - P_{ij}$ is taken in account only when i and j belong to the same cluster. In the organized version, this term is always taken in account, but with a weight S_{kl} equal to the prior similarity between C_k (with $i \in C_k$) and C_l (with $j \in C_l$). This favors connected clusters to be close in the prior structure. If there are indeed significant connections between nodes in two clusters C_k and C_l (i.e., $W_{ij} - P_{ij} > 0$), then the value of $O(M)$ will be higher if S_{kl} is high than if it is low. This is similar to the SOM principle in which a prototype has to be close to observations assigned to its unit but also, to a lesser extent, to observations assigned to neighboring units in the prior structure.

In the rest of the paper, we denote $B_{ij} = \frac{1}{2m} (W_{ij} - P_{ij})$ for $i \neq j$ and $B_{ii} = 0$. It is clear that maximizing $O(M)$ is equivalent to maximizing $F(M) = \sum_{i \neq j} \sum_{k,l} M_{ik} S_{kl} M_{jl} B_{ij}$.

3 Deterministic Annealing

The main difficulty with (organized) modularity maximization is that it is a discrete optimization problem. Following [7] for modularity, we propose to solve it by deterministic annealing [11]. Deterministic annealing tries to solve the complex combinatorial problem of maximizing F via an analysis of the Gibbs distribution obtained as the asymptotic regime of a classical simulated annealing (or via the principal of maximum entropy). In our case, the Gibbs distribution for temperature $\frac{1}{\beta}$ is $P(M) = \frac{1}{Z_P} \exp(\beta F(M))$ where the normalization constant Z_P is given by $Z_P = \sum_M \exp(\beta F(M))$, where the sum is taken over all partitions into C clusters.

The main idea is to compute the expectations of the assignments at a fix temperature with respect to the Gibbs distribution, i.e., the $\mathbb{E}(M_{ik})$, and then to decrease the temperature while tracking the evolution of the expectations. Unfortunately, F is not linear with respect to M and computing Z_P and P is therefore difficult. Following previous work on similar topics, we approximate P by a distribution that factorizes (see e.g., [5]). This corresponds to approximating the interaction between say M_{ik} and all the other variables via a *mean field* E_{ik}. More precisely, we consider the bi-linear cost function $U(M, E) = \sum_i \sum_k M_{ik} E_{ik}$ where E is a N by C matrix of partial assignment costs. For a fixed temperature $\frac{1}{\beta}$, we look for a mean field E that gives a distribution $R(M, E) = \frac{1}{Z_R(E)} \exp(\beta U(M, E))$ close to $P(M)$, in the sense that the Kullback-Leibler divergence $KL(R|P)$ between R and P is minimal, with $KL(R|P) = \sum_M R(M, E) \ln \frac{R(M,E)}{P(M)}$. At a minimum, the gradient of $KL(R|P)$ with respect to E is zero. This leads to the following classical mean field equations:

$$\frac{\partial \mathbb{E}_R(F(M))}{\partial E_{jl}} = \sum_k \frac{\partial \mathbb{E}_R(M_{jk})}{\partial E_{jl}} E_{jk}, \ \forall j, l. \tag{1}$$

They are obtained using the main consequence of the mean field approximation, namely the independence between M_{ik} and M_{jl} for $i \neq j$ under the distribution R, i.e., the fact that $\mathbb{E}_R(M_{ik} M_{jl}) = \mathbb{E}_R(M_{ik}) \mathbb{E}_R(M_{jl})$ for $i \neq j$.

To solve the mean field equations, we use a EM like approach. We consider the $\mathbb{E}_R(M_{ik})$ fixed and solve the equations for E_{jl} (maximization phase). Then we compute the new values of the $\mathbb{E}_R(M_{ik})$ (expectation phase). This latter phase leads to the very simple standard deterministic annealing update rule:

$$\mathbb{E}_R(M_{ik}) = \frac{\exp(\beta E_{ik})}{\sum_l \exp(\beta E_{il})}. \tag{2}$$

Moreover, the independence property recalled above gives

$$\mathbb{E}_R(F(M)) = \sum_{i \neq j} \sum_{k,l} \mathbb{E}_R(M_{ik}) S_{kl} \mathbb{E}_R(M_{jl}) B_{ij}.$$

Then, some straightforward calculations show that equation (1) is fulfilled if the

mean field is given by

$$E_{jk} = 2 \sum_{i \neq j} \sum_{l} \mathbb{E}_R \left(M_{il} \right) S_{kl} B_{ij}, \tag{3}$$

or, in matrix notations, $E = B\mathbb{E}_R \left(M \right) S$, using the symmetry of B and S.

Finally, given an annealing schedule, i.e., an increasing series $(\beta_l)_{1 \leq l \leq L}$, the organized modularity maximization algorithm proceeds as follows:

1. initialize $\mathbb{E}_R \left(M \right)$ randomly (with $\mathbb{E}_R \left(M_{ik} \right) \in [0, 1]$ and $\sum_k \mathbb{E}_R \left(M_{ik} \right) = 1$)

2. for $l \in \{1, \ldots, L\}$:

    (a) compute E using equation (3)

    (b) compute $\mathbb{E}_R \left(M \right)$ using equation (2) with $\beta = \beta_l$

    (c) go back to (a) until convergence

3. threshold $\mathbb{E}_R \left(M \right)$ into a partition M

4   Experiments

We report results obtained on the coauthorship network of scientists working on network theory and experiment compiled by M. Newman in May 2006 [9][1]. More precisely, we work on the largest connected component of the network, which has 379 nodes. We compare several organization algorithms:

- the *kernel SOM* as described in [13, 1]. The experiments were conducted with two kernels: the heat kernel [6], $K_\gamma = e^{-\gamma L}$ and the generalized inverse of the Laplacian [3], $K = L^+$. The first kernel SOM will be denoted by SOM-HK in the following and the second one by SOM-GI.

- the optimization of the *organized modularity by deterministic annealing*, as described in the previous sections. This algorithm will be denoted by DA-OM.

In addition, deterministic annealing was used to build an ordinary (not organized) clustering on the basis of the optimization of the modularity (as described in [7]): in the range of 8 to 16 clusters, the modularity stays slightly above 0.8 with no major variation. Above 16 clusters, the modularity starts to decrease: even if the optimization algorithm manages to produce some empty clusters, convergence to a global optimum is impaired by the use of a large number of clusters. As a consequence, we have limited our exploration to a maximum of 16 clusters. As an example of the general tendency, we provide here the results obtained on a 4×4 squared grid (i.e., 16 clusters).

Each organization algorithm involves some parameters. We used a grid search approach to optimize them. For DA-OM, we specified the prior structure via a

---

[1]The data are available at `http://www-personal.umich.edu/~mejn/netdata/`

Gaussian similarity $S_{kl} = \exp(-\sigma\|\mathbf{x}_k - \mathbf{x}_l\|^2)$: the σ parameter was optimized by the grid search. This value plays a similar role as the initial radius of neighborhood which was also optimized. In addition, the γ of heat kernel was also tuned.

To compare the methods, we compute two quality measures: the modularity (to measure the quality of the clustering) and the percentage of crossing edges on the grid (to measure the visual quality of the organization). For each method, the grid search produced Pareto optimal points for these two quantities. They are reported in Table 1. Deterministic annealing for optimization

	DA-OM		SOM-GI	SOM-HK	
Modularity	**0.836**	0.825	0.816	0.771	0.024
% of edge crossing	**0.019**	0.016	0.005	xxx	xxx

Table 1: Comparison of two quality measures for the organization algorithms

of the organized modularity provides good solutions either on a classification point of view (high modularity) and on a visual point of view (small number of edge crossing). The solutions obtained by SOM with generalized inverse kernel (SOM-GI) are also quite good for both quality measures, with a tendency to produce better organized mapping (in the sense of a higher number of crossing edges) but worse modularity. On the contrary, SOM with heat kernel provides much worse solutions and seems to fail to find a good tradeoff between clustering and organization, at least on that example.

Finally, Figure 1 presents the obtained organizations of the collaboration network by the two best methods (DA-OM and SOM-GI).
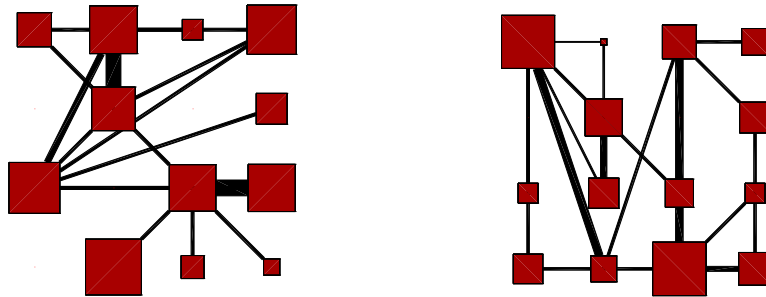

Fig. 1: Organizations obtained by DA-OM (left) and SOM-GI (right, for the optimum with the best modularity). The square surfaces are proportional to cluster sizes and the edges' width to the sum of weights between two clusters

Both methods give satisfactory simplifications of the network that are well-organized on the grid. Interestingly, the DA-OM algorithm produces empty clusters: this is a consequence of the use of the modularity measure which doesn't systematically increase with the number of clusters. This means in practice that

deterministic annealing leads to an automatic optimal choice of the number of clusters, for both organized and standard modularities[2]. In addition, the computational cost of DA-OM is less important than the one of SOM-GI. Indeed the pseudo inverse of the Laplacian will generally be a full matrix leading to a cost of $O(N^2C)$ per iteration of the kernel SOM. On the contrary, DA-OM can leverage the sparsity of the matrix B to obtain a complexity of $(m + N)C$ per iteration of the deterministic annealing algorithm [7] in the case of modularity and $(m + N)C + NC^2$ for its organized version.

The approach proposed in this paper seems therefore to provide a good solution for computing a topologically organized clustering of a graph. We are currently investigating its application to larger graphs as well as the properties of the organized modularity measure.

References

[1] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7–9):1257–1273, March 2008.

[2] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs.* Prentice Hall, 1999.

[3] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, pages 355–369, March 2007.

[4] I. Herman, G. Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[5] T. Hofmann and J. M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14, January 1997.

[6] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[7] S. Lehmann and L. K. Hansen. Deterministic modularity optimization. *European Physical Journal B*, 60(1):83–88, November 2007.

[8] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[9] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 2006.

[10] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004.

[11] K. Rose. Deterministic annealing for clustering, compression,classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, November 1998.

[12] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, August 2007.

[13] N. Villa and F. Rossi. A comparison between dissimilarity som and kernel som for clustering the vertices of a graph. In *Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 07)*, Bielefeld (Germany), September 2007.

---

[2]One interesting feature of the (organized) modularity it that it starts decreasing when a graph dependent threshold on the number of clusters is reached.