Probing the robustness of the clustering

David Gfeller[1], Jean-Cédric Chappelier[2], and Paolo De Los Rios[1]

[1] *Laboratoire de Biophysique Statistique, SB/ITP and*

[2] *School of Computer and Communication Sciences,*

Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland

(Dated: May 4, 2005)

Keywords: analysis of complex networks, clustering

Many complex systems can be represented as complex networks. Nevertheless the very large size of such systems often does not allow any possible mental or graphical representation. It has been pointed out that a complex network consists in many cases of several densely interconnected clusters, with only a few connections to the rest of the network. The identification of these clusters is therefore a key point to understand the relevant organization of the network and to reduce the size and complexity of the system. Here we introduce a new method to probe the robustness of the clusters. This is particularly important since often clustering algorithms produce a hard clustering even for the nodes that "lie between" clusters, whose classification is very arguable.

Our method is based on the introduction of noise over the weight of the edges. After running the clustering algorithm for several realizations of the noise, we can assess a probability on the edges of connecting two nodes belonging to the same cluster. This probability represents how much the grouping of two nodes in a cluster is reasonable. For example nodes connected together by edges with probability close to 1 can be assigned to the same cluster with high confidence. On the other hand, nodes connected with a probability close to 0.5 are rather *unstable* with respect to the cluster structure.

Furthermore we introduce a general measure of the robustness of the clusters, defined as the *clustering entropy*. It allows to discriminate networks with well-defined clusters from networks that, although topologically very similar, do not have a significant cluster structure.

We tested our method on several real world examples. In a linguistic network based on the synonymy relation, the unstable nodes corresponded to ambiguous words. In sociological networks they represented individuals connected to different groups. In all those cases, the clustering entropy of the real network was much lower than the one of a randomized network where the degree of the nodes had been conserved. We finally note that the method can be applied with any clustering algorithm, provided that the algorithm allows weights on the edges.