

---

# Statistical Mechanics of Semi-Supervised Clustering in Sparse Graphs

---

**Greg Ver Steeg** and **Aram Galstyan**  
USC Information Sciences Institute  
Marina del Rey, CA 90292, USA  
{gregv,galstyan}@isi.edu

**Armen Allahverdyan**  
Yerevan Physics Institute  
Yerevan 375036, Armenia  
aarmen@mail.yerphi.am

The problem of detecting modules, or clusters, in networks has recently attracted a considerable interest both in statistical physics and computer science communities. In addition to algorithmic development, recent research has addressed the *feasibility* of detecting clusters, assuming they are present in the network. Consider the so called *planted bi-partition* model, where the nodes are partitioned into two equal-sized groups, and each link within a group and between the groups is present with probabilities  $p$  and  $r$ , respectively. It is known that in dense networks where the average connectivity scales linearly with the number of nodes  $N$  (e.g.,  $p$  and  $r$  are constants), the clusters in the planted partition model can be recovered with asymptotically *perfect* accuracy for any  $p - r > N^{-1/2+\epsilon}$ .

The situation is significantly different for sparse graphs, where the average connectivity remains finite in the asymptotic limit  $N \rightarrow \infty$ . Indeed, recently it was shown [3] that planted partition models (of arbitrary topology) with finite connectivities are characterized by a phase transition from *detectable* to *undetectable* regimes as one increases the overlap between the clusters, with the transition point depending on the actual degree distribution of the partitions. From the perspective of statistical inference, this type of phase transition between detectable and undetectable regimes is undesirable, as it signals inference instabilities – large fluctuations in accuracy in response to relatively small shifts in the parameters.

We have previously shown that this instability can be avoided if one uses prior knowledge about the underlying group structure [1]. Namely, it was demonstrated that knowing the correct cluster assignments for arbitrarily small but *finite* fraction of nodes destroys the criticality and moves the detection threshold to its intuitive (dense-network limit) value  $p = r$ . This can be viewed as a *semi-supervised* version of the problem, as opposed to an unsupervised version where the only available information is the observed graph structure.

In this paper, we study semi-supervised graph clustering when the background knowledge is expressed as pair-wise constraints in the form of *must-links* (*cannot-links*), which imply that a pair of nodes must be (cannot be) assigned to the same group. These types of constraints are more realistic in scenarios where it is easier to assess similarity of two objects than to label them individually. The goal of this paper is to theoretically examine the impact of semi-supervision on clustering accuracy. We focus on bi-cluster graphs, and study the impact of semi-supervision for varying constraint density and overlap between the clusters.

Our results suggest that when the density of the constraints is sufficiently small, their only impact is to shift the detection threshold while preserving the criticality. Intuitively, we can see that the link constraints form connected components of nodes whose relative group assignment is known. We can view these components as nodes in a renormalized graph without link constraints. Clustering on our constrained graph corresponds to solving an unsupervised clustering problem on this renormalized graph, and, therefore, we expect a critical threshold. On the other hand, when the density of link constraints is above the bond percolation threshold, we expect that the size of the largest component of nodes connected by constraints is of  $O(N)$ . Indeed, we find that in this case, the threshold is once again shifted to the limit  $p = r$ . A more detailed discussion is provided in [2].

## References

- [1] A. E. Allahverdyan, G. Ver Steeg, and A. Galstyan, *Community Detection with and without Prior Information*, Europhys. Lett. **90**, 18002 (2010).
- [2] G. Ver Steeg, A. Galstyan, and A. E. Allahverdyan, *Statistical Mechanics of Semi-Supervised Clustering in Sparse Graphs*, arXiv:1101.4227, (2011).
- [3] J. Reichardt and M. Leone, Phys. Rev. Lett. **101**, 078701 (2008).