Clustering with Prior Information

Armen E. Allahverdyan
Yerevan Physics Institute
Yerevan, Armenia
aarmen@yerphi.am

Aram Galstyan
USC / ISI
Marina del Rey, CA, USA
galstyan@isi.edu

Greg Ver Steeg
USC / ISI
Marina del Rey, CA, USA
gregv@isi.edu

A fundamental issue in clustering concerns one's ability (and limitation) to detect clusters, assuming they are built-in to the model that generates the data [1, 4]. Results for the *planted partition* graph models suggest that clusters can be recovered with arbitrary accuracy if sufficient data (link density) is available [2]. More recently, this problem of cluster detectability has been addressed theoretically for *sparse* graphs, by formulating it through a certain Ising–Potts Hamiltonian [6]. It was shown that clustering in the sparse planted partition model is characterized by a phase transition from detectable to undetectable regimes as one increases the overlap between the clusters [6]. Specifically, for sufficiently large inter–cluster coupling, the underlying (planted) cluster structure has no impact on the optimal (minimum–energy) configuration of the Hamiltonian.

Here we examine the cluster–detection problem in semi–supervised settings, when one has some background knowledge about the clusters. Generally speaking, such information can be in form of pair-wise constraints (via must– and cannot links), or, alternatively, via known cluster assignments for a fraction of nodes. Here we consider the latter scenario defined on a *planted bisection* graph model [2]. Namely, consider two clusters containing N–nodes each. Each pair of nodes within the same cluster is linked with probability α/N, where α is the average within–cluster connectivity. Also, each pair of nodes in different clusters is linked with probability γ/N, where γ is inter–cluster connectivity. We assign weights J and K to within–cluster and inter–cluster links, and define the following (min–cut) Hamiltonian [3, 6]:

$$H = -\sum\nolimits_{i<j}^{N} J_{ij} s_i s_j - \sum\nolimits_{i<j}^{N} \bar{J}_{ij} \bar{s}_i \bar{s}_j - \sum\nolimits_{i,j}^{N} K_{ij} s_i \bar{s}_j, \tag{1}$$

where we made the bi–cluster nature of the network explicit by introducing separate spin variables $s_i = \pm 1$ and $\bar{s}_i = \pm 1$ ($i = 1, \dots, N$) for two clusters. Here J_{ij} and $\bar{J}_{ij}$ are iid random variables which assume zero with probability $1 - \frac{\alpha}{N}$ and $J > 0$ with probability $\frac{\alpha}{N}$. Likewise, K_{ij} identically and independently are equal to zero with probability $1 - \frac{\gamma}{N}$ and to $K > 0$ with probability $\frac{\gamma}{N}$.

The optimal clustering corresponds to the spin configuration that minimizes the above Hamiltonian. To simplify the analysis, we assume equipartition, so that the (1) will be studied under the constraint $\sum_{i=1}^{N} s_i + \sum_{i=1}^{N} \bar{s}_i = 0$. Thus, detecting the sign of a given spin s_i at zero temperature (so as to exclude all thermal fluctuations) we can conclude to which cluster the corresponding node belongs: all spins having equal signs belong to the same cluster. The error probability $p_{\rm e}$ for the cluster assignment is

$$p_{\rm e} = (1 - |m|)/2, \quad m \equiv [\langle s_i \rangle_{T=0}]_{\rm av} = -[\langle \bar{s}_i \rangle_{T=0}]_{\rm av}, \tag{2}$$

where m is the (single–cluster) magnetization, $\langle \dots \rangle_{T=0}$ is the zero-temperature Gibbsian average, i.e. the average over all configurations of spins having in the thermodynamic limit the minimal energy given by (1), and where $[\dots]_{\rm av}$ is the average over the bi-graph structure.

The above formulation refers to unsupervised clustering. In the semi–supervised case, we assume that for some (randomly distributed) nodes their cluster assignment is known in advance. We introduce this knowledge into the model by *fixing* the spins situated at those nodes to corresponding values. Thus, (1) is modified as

$$\widetilde{H} = H - \sum\nolimits_{i=1}^{N} f_i s_i - \sum\nolimits_{i=1}^{N} \bar{f}_i \bar{s}_i, \tag{3}$$

where f_i (resp. $\bar{f}_i$) are identically and independently distributed random variables that are equal to 0 with probability $1-\rho$ and to ∞ (resp. $-\infty$) with probability ρ.

We study the above model within the Bethe–Peierls approximation. Let $P(h)$ ($\bar{P}(h)$) denote the probability of an internal (*cavity*) field acting on an s ($\bar{s}$) spin. Then we have according to the zero temperature cavity method [5]:

$$P(h) = \sum_{n,m=0}^{\infty} \frac{\gamma^m e^{-\gamma}}{m!} \frac{\alpha^n e^{-\alpha}}{n!} \int \hat{p}(f)\mathrm{d}f \int \prod\nolimits_{k=1}^{m} P(h_k)\mathrm{d}h_k \int \prod\nolimits_{l=1}^{n} \bar{P}(g_k)\mathrm{d}g_k$$
$$\times \delta\left(h - f - \sum\nolimits_{k=1}^{m} \phi[h_k, J] - \sum\nolimits_{k=1}^{n} \phi[g_k, K] \right), \tag{4}$$

where $\phi[a,b] \equiv \mathrm{sign}(a)\min[\,|a|, b\,]$, and where g_k (resp. h_k) are the fields acting on the s-spin from $\bar{s}$-spin (resp. from other s-spins). These fields naturally enter with weight $\frac{\gamma^m e^{-\gamma}}{m!}$ (resp. $\frac{\alpha^n e^{-\alpha}}{n!}$), which is the degree distribution of the corresponding Erdös–Rényi network. Also, (4) $\hat{p}(f) = \rho\delta(f-\infty) + (1-\rho)\delta(f)$ is the distribution of the frozen (supervising) field acting on s-spins.

Due to (1–3) and the complete inversion symmetry between the two clusters, we can take $\bar{P}(g) = P(-g)$, and then (4) is worked out via the Fourier representation of the delta-function yielding $P(h) = \rho\delta(h-\infty) + (1-\rho)\widetilde{P}(h)$, where $\widetilde{P}(h)$ refers to those s-spins, which were not directly frozen by infinitely strong random fields:

$$\begin{aligned} \widetilde{P}(h) \quad &= \quad e^{-\alpha-\gamma} \int \frac{\mathrm{d}z}{2\pi}\, e^{izh}\, \exp\Big\{ \alpha\rho\, e^{-izJ} + \gamma\rho\, e^{izK} \\ &+ \quad (1-\rho) \int \mathrm{d}g \widetilde{P}(g) \Big[\alpha e^{-iz\,\mathrm{sign}(g)\,\min[|g_1|,J]} + \gamma e^{iz\,\mathrm{sign}(g)\,\min[|g|,K]} \Big] \Big\}. \end{aligned} \tag{5}$$

Since J and K are integers, we look for the solution for Eq. 5 in the following form:

$$\widetilde{P}(h) = \sum\nolimits_{n=-\infty}^{\infty} c_n \delta(h-n), \tag{6}$$

The average magnetization (hence, the error probability) is expressed throuhg the coefficients c_n as

$$m = \int \mathrm{d}h\, \widetilde{P}(h)\, \mathrm{sign}(h) = \sum_{n=0}^{\infty} [c_n - c_{-n}] \tag{7}$$

Another relevant quantity is the so called Edwards–Anderson order parameter given by[5]:

$$q = [\,\langle s_i \rangle_{T=0}^2\,]_{\mathrm{av}} = \int \mathrm{d}h\, \widetilde{P}(h)\, \mathrm{sign}^2(h) = 1 - c_0 \tag{8}$$

where $[\ldots]_{\mathrm{av}}$ is now the average over the bi-graph structure and the random fields.

Let us first focus on the *unweighted* ($J = K = 1$) and *unsupervised* ($\rho = 0$) scenario that was studied in Ref. [6]. In this case, the order parameters can be shown to satisfy the following system of transcendental equations:

$$1 - q = e^{-(\alpha+\gamma)q} I_0[\sqrt{(\alpha+\gamma)^2 q^2 - (\alpha-\gamma)^2 m^2}], \tag{9}$$

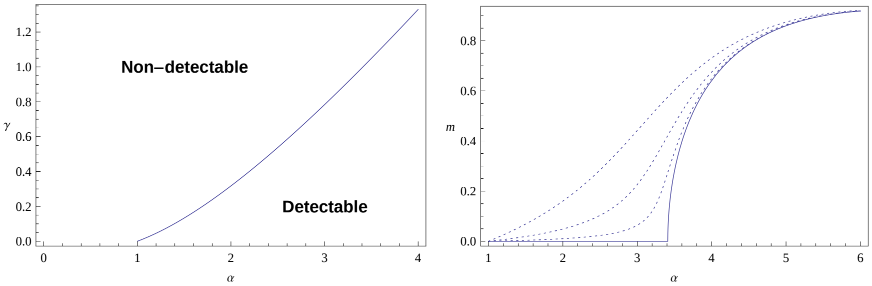
$$m = -2e^{-(\alpha+\gamma)q} \sum\nolimits_{n=1}^{\infty} I_n(x) \sinh\left[n \mathrm{atanh} \frac{(\gamma-\alpha)m}{(\alpha+\gamma)q} \right]. \tag{10}$$

where $I_n(x)$ is the modified Bessel function. Eq. (10) predicts a second-order transition, where m is the order-parameter. In the vicinity of this transition we use Taylor expansion of (9, 10) over m to obtain

$$1 - q = e^{-(\alpha+\gamma)q} I_0[(\alpha+\gamma)q], \tag{11}$$

$$1 = (\alpha-\gamma)(1-q) \left(1 + \frac{I_1[(\alpha+\gamma)q]}{I_0[(\alpha+\gamma)q]} \right). \tag{12}$$

We have $m > 0$ ($m = 0$) if the RHS of (12) is larger (smaller) than its LHS. Thus, Eq. (12) determines the detection threshold, above of which the method is capable of detecting clusters with *better* *than random* probability of error (2). The critical line on the (α, γ) plane is shown in Figure 1(a): It starts from $(\alpha = 1, \gamma = 0)$, since (11) predicts a percolation bound for q: $q = 0$ $(q > 0)$ for $\alpha + \gamma < 1$ $(\alpha + \gamma > 1)$. Naturally, close the percolation bound $\alpha = 1$, even very small inter–cluster coupling γ nullifies m. Fig. 1(a) also shows that at the detection threshold $\alpha > \gamma$. Furthermore, it can be shown that the difference $\alpha - \gamma$ at the threshold grows as $\sqrt{2\pi(\alpha + \gamma)}$ for a large $\alpha + \gamma$; see (11). Thus, the ratio $\frac{\alpha - \gamma}{\alpha + \gamma}$ converges to zero for a large $\alpha + \gamma$. In this *weak* sense, the detection threshold converges to $\alpha = \gamma$ for large $\alpha + \gamma$, while for any finite α the unsupervised clustering detection threshold lies below the line $\alpha = \gamma$.



(a) (b)

Figure 1: (a) The phase diagram for $J = K = 1$. The line on the (α, γ) plane indicates second-order phase-transition from $m = 0$ (no clustering detection) to $m > 0$ (clustering detection); (b) Normal curve: magnetization m versus α for $\gamma = 1$. m undergoes second-order phase-transition at $\alpha = 3.4$. Dashed curves: remnant [semi-supervised] magnetization$\widetilde{m}$ versus α for $\gamma = 1$. From top to bottom: $\rho = 0.2,\ 0.05,\ 0.01$.

Under semi-supervised setting, we still employ (6) and obtain (9, 10), but now in the RHS of these equations one should substitute $m \to \rho + (1 - \rho)m$ and $q \to \rho + (1 - \rho)q$. Expanding over a small m we get

$$\begin{aligned} 1 - q &= e^{-(\alpha+\gamma)(\rho+[1-\rho]q)} I_0[(\alpha + \gamma)(\rho + [1 - \rho]q)], \\ m &= \rho(\alpha - \gamma)(1 - q) \left[1 + \frac{I_1[(\alpha + \gamma)(\rho + [1 - \rho]q)]}{I_0[(\alpha + \gamma)(\rho + [1 - \rho]q)]}\right]. \end{aligned} \tag{13}$$

Now $m > 0$ for any $\alpha - \gamma > 0$. This is the average-connectivity threshold, which for the considered unweighted scenario is the only possible definition of clustering. Thus, *any* generic semi-supervising leads to the theoretically best possible threshold $\alpha = \gamma$; see Fig. 1(b).

The above analysis can be generalized to the planted partition graphs with integer weights, although the resulting equations are rather involved. Here we overview our main results obtained for two particular (but important) cases (the details are forthcoming). First, consider $2J = K = 2$. While for the previous unweighted situation, any amount of semi-supervision (as quantified by ρ) sufficed for shifting the clustering threshold to a ρ-independent value, here the detection threshold starts to depend on ρ, and the smallest threshold is achieved for $\rho \to 0$. This is illustrated in see Table 1. To understand this seemingly counterintuitive observation, note that detection threshold is achieved as a balance between the inter–cluster links [with the average connectivity γ and weight $K = 2$] that—due to negatively frozen spins—exert negative fields on the test spin and the within–cluster links [with the average connectivity α and weight $J = 1$] that exert positive fields. A larger ρ facilitates the negative fields, since they have twice larger weight, which explains why vanishing semi-supervising $\rho \to 0$ facilitates a lower detection threshold.

Now consider rather paradoxical aspect of the semi-supervised detection threshold: it is *smaller* than the value deduced from balancing the cumulative weights of within–cluster and inter–cluster links, which yields $\alpha J = \gamma K$. Indeed, according to Table 1 (where $\gamma = 1$) we have $\alpha = 1.5$ (reached for $\rho \to 0$) versus the weight-balancing value $\alpha = 2$. This result seemingly contradicts the intuition we got so far: *i)* a rough intuition about Hamiltonian (1) is that it is based on defining a cluster via the within–cluster weight being larger than the inter–cluster weight. *ii)* The unsupervised threshold is well above the weight-balancing prediction (see last column in Table 1). *iii)* In the unweighted case

ρ	0.3	0.1	0.05	0.005	0.0005	0
α	1.6812	1.5976	1.5617	1.5173	1.5119	4.9122
q	0.86734	0.7871	0.7518	0.7087	0.7036	0.8373
$m\vert_{\alpha=2}$	0.0401	0.0182	0.0102	0.0016	10^{-4}	–

Table 1: Weighted situation: $2J = K = 2$. For $\gamma = 1$ and various semi-supervising degrees ρ we list the clustering threshold α and the value of q at this threshold.

$(J = K)$ the semi-supervising just reduces the detection threshold towards $\alpha = \gamma$, which coincides with the weight-balancing value.

To understand this effect, we turn to the physical picture of the threshold, where positively and negatively acting links driven by the semi-supervised (frozen) spins compensate each other. At the weight-balancing point $\alpha J = \gamma K$ (with $J < K$) fewer (but stronger) inter–cluster links have the same weight as more numerous (but weaker) within–cluster links. Since the within–cluster links are more numerous, their overall effect on a (randomly chosen) test spin is more deterministic and hence capable of building up a positive m at $\alpha J = \gamma K$. Thus, the actual threshold is reached for $\alpha J < \gamma K$.

We alo calculate m at the weight-balancing value $\alpha J = \gamma K$, since this is the semi-supervising benefit of those who would insist on the weight-balancing definition of the threshold; see Table 1. Note finally that for large values of γ both unsupervised and semi-supervised thresholds converge to $\alpha J = \gamma K$, since now fluctuations are irrelevant from the outset.

All these effects turn upside-down for $2K = J = 2$. Now the threshold is minimized for the maximal semi-supervising $\rho \to 1$, and the semi-supervised detection threshold α is always larger than the weight-balancing value $\gamma K/J$. These results are explained by "inverting" the above arguments developed for $J < K$.

In summary, we have demonstrated analytically that any small (but finite) amount of semi–supervision suppresses the phase transition in cluster detectability for the planted–bisection model, by shifting the detection threshold to its lowest possible value. For graphs where the links within and across the clusters have different weights, we found that semi–supervision leads to a detection threshold that depends on ρ. Furthermore, if $J < K$, then for $\rho \to 0_+$, the detection threshold converges to a value lower (better) from the one obtained via balancing within–cluster and inter–cluster weights. This suggests that for weighted graphs a small [but generic] semi-supervising can be employed for defining the very clustering structure. This definition is non-trivial, since it performs better than the weight-balancing definition. Note also that for weighted graphs the very notion of the detection threshold is not clear *a priori*, in contrast to unweighted networks, where the *only* possible definition goes via the connectivity balance $\alpha = \gamma$. To illustrate this unclarity, consider a node connected to one cluster via few heavy links, and to another cluster via many light links. To which cluster this node should belong *in principle*? Our (speculative) answer is that the proper cluster assignment in this case can be defined via semi-supervising.

References

[1] M. Ackerman and S. Ben-David, "Clusterability: A Theoretical Study", Proceedings of AISTATS- 09, JMLR: W&CP 5, pp. 1-8, (2009).

[2] A. Condon and R. M. Karp *Algorithms for Graph Partitioning on the Planted Partition Model*, Random Structures and Algorithms **18**, pp.116–140, 2001.

[3] Y. Fu and P.W. Anderson, "Application of statistical mechanics to NP-complete problems in combinatorial optimisation", J. Phys. A **19**, 1605 (1986).

[4] U. von Luxburg, S. Ben-David, "Towards a statistical theory for clustering", PASCAL Workshop on Statistics and Optimization of Clustering, London, U.K., (2005).

[5] M. Mezard and G. Parisi, "Mean-field theory of randomly frustrated systems with finite connectivity", Europhys. Lett. **3**, 1067 (1987).

[6] J. Reichardt and M. Leone, "(Un)detectable Cluster Structure in Sparse Networks", Phys. Rev. Lett. **101**, 078701 (2008).