4.3 Algorithms for nets and regularity partitions

The results above can be applied to the computation of (weak) regularity partitions in the "property testing" model for huge dense graphs.

If we face a large network (think of the internet) the first challenge is to obtain information about it. Often, we don't even know the number of nodes. In one of the models, we assume that we are able to randomly sample small subgraphs. The theory of this, a sort of a statistics where we work with graphs instead of numbers, is called *property testing*. It is simple to describe a reasonably realistic sampling process: we select independently a number k of random nodes, and determine the edges between them, to get a random induced subgraph. (We have to assume, of course, that we have methods to select a uniformly distributed random node of the graph, and to determine whether two nodes are adjacent.) So we see a random k-node graph with a certain distribution.

It turns out that this sample contains enough information to determine many properties and parameters of the graph, with some error of course. This error can be made arbitrarily small with high probability if we choose the sample size k sufficiently large, depending on the error bound (and only on the error bound, not on the graph!).

ε-nets and ε-covers in metric spaces. We need to survey some notions and elementary arguments about packing and covering with respect to the 2-neighborhood metric d. Let $\varepsilon > 0$. A set of nodes $S \subseteq V$ is an *ε-cover*, if $d(v,S) \le \varepsilon$ for every point $v \in V$. (Here, as usual, $d(v,S) = \min_{s\in S} d(s,v)$.) We say that S is an *ε-packing*, if $d(u,v) \ge \varepsilon$ for every $u,v \in S$. An *ε-net* is a set that is both an ε-packing and an ε-cover. It is clear that a maximal ε-packing must be an ε-covering (and so, and ε-net).

It is clear that if S is an ε-cover, then the Voronoi cells of S have diameter at most 2ε.

An *average ε-cover* is a set $S \subseteq V$ such that $\sum_{v\in V} d(v,S) \le \varepsilon n$. An *average ε-net* is an average (2ε)-cover that is also an ε-packing. (It is useful to allow this relaxation by a factor of 2 here.) For every average ε-cover, a maximal subset that is an ε-packing is an average ε-net.

We will need the following simple lemma.

Lemma 4.3.1 *Let $T, S \subseteq V$. Then there is a subset $S' \subseteq S$ such that $|S'| \le |T|$ and $\overline{d}(S') \le \overline{d}(S') + 2\overline{d}(T)$.*

Proof. For every point in T choose a nearest point of S, and let S' be the set of points chosen this way. Clearly $|S'| \le |T|$. For every $x \in V$, let $y \in S$ and $z \in T$ be the points nearest to x. Then $d(z,S) \le d(z,y) \le d(x,z) + d(x,y) = d(x,T) + d(x,S)$, and hence, by its definition, S' contains a point y' with $d(z,y') \le d(x,T) + d(x,S)$. Hence

$$d(x,S') \le d(x,y') \le d(x,z) + d(z,y') \le 2d(x,T) + d(x,S).$$

Averaging over x, the lemma follows. □

Algorithms. We assume, as before, that we can generate independent, uniformly distributed random points from V. The 2-neighborhood metric can be expressed as

$$d(i,j) = \frac{1}{n^2}|A^2(\mathbf{e}_i - \mathbf{e}_j)|_1 = \frac{1}{n^2}\sum_k |(A^2)_{ik} - (A^2)_{jk}| = \frac{1}{n^2}\sum_k \Big|\sum_r (A_{ir} - A_{jr})A_{rk}\Big|$$

(where A is the adjacency matrix of the graph G). We can write this as

$$d(i,j) = \mathsf{E}_k\big|\mathsf{E}_r((A_{ir} - A_{jr})A_{rk})\big|,$$

where k and r are chosen randomly and uniformly from V. This shows that we can approximately compute $d(i,j)$ by random sampling from V.

How to construct ε-nets? If V is not too large, then we can go through the nodes in any order, and build up the ε-net S, by adding a new node v to S if $d(v,S) \ge \varepsilon$. In other words, S is a greedily selected maximal ε-packing, and therefore an ε-net.

It can be proved that no ε-packing in a graph G can be larger than $K(\varepsilon) = 2^{1000(1/\varepsilon^2)\log(1/\varepsilon)}$. The proof is a combination of Lemma 4.0.2(a) with the proof of Theorem 3.3.2, and not described in this course.

Average ε-nets can be constructed by a simple randomized algorithm even if V is large. We build up S randomly: at each step we generate a new random node, and add it to S if its distance from S is at least ε (otherwise, we throw it away). For some pre-specified $A \ge 1$, we stop if for A/ε consecutive steps no new point has been added to S. The set S formed this way is trivially an ε-packing, but not necessarily maximal. However, it is likely to be an average (2ε)-cover. Indeed, let t be the number of points x with $d(x,S) > \varepsilon$. These points are at distance at most 1 from S, and hence

$$\sum_{x\in V} d(x,S) \le (n-t)\varepsilon + t < n\varepsilon + t.$$

So if S is not an average (2ε)-cover, then $t > \varepsilon n$. The probability that we have not hit this set for A/ε steps is

$$\Big(1 - \frac{t}{n}\Big)^{A/\varepsilon} < e^{-\frac{t}{n}\frac{A}{\varepsilon}} < e^{-A}.$$

The time cost of this algorithm is at most $K(\varepsilon)A/\varepsilon$.

We can modify this procedure to find an almost optimal ε-cover. Consider the smallest average ε-cover T (we don't know which points belong to T, of course), and construct a (possibly much larger) average 2ε-cover S by the procedure described above. We know that $|T| \le |S| \le K(\varepsilon)$ with high probability. By Lemma 4.3.1, S contains an average (6ε)-cover S' of size $|S'| \le |T|$. We can try all subsets of S to find S'.

This clearly inefficient last step is nevertheless useful in the property testing model. Then the point is that the time bound depends only on ε, not on the size of V.

Theorem 4.3.2 *There is a randomized algorithm in the property testing model that computes, in time that depends only on ε, a set $S \subseteq V$ that is not larger than then smallest average ε-cover, such that with probability at least $1-\varepsilon$, the set S satisfies $\overline{d}(S) \leq 6\varepsilon$.*

Once this set S is computed, we can consider the task of computing a weak regularity partition solved: we can take the Voronoi cells of the set S, with respect to the metric d, as the classes of this partition. This partition $\mathcal{P}$ satisfies $d_\square(G, G_\mathcal{P}) \leq 8\varepsilon^{1/2}$.

Note, however, that "computing the partition $\mathcal{P}$" does not mean that we compute a list of all nodes, specifying the partition class it belongs to: we assume that the graph is very large, so such a list would be too long. What we want is a way to determine about any given node which partition class it belongs to, in other words, which element of S is closest to it. Since the metric d is computable by sampling, this can be done in the property testing model. (Of course, using sampling we can only compute an approximation of the metric d, with high probability. We don't discuss the issue of estimating the errors here.)