Graphons and Cut Distance

Graph Schemas

- We want to understand the properties not just of individual graphs, but of general graph schemas; often given by a constructive process or a sequence of distributions for each potential graph size n.
- For instance, we don't usually care about the properties of a given individual sample from $G(n, p)$, we care about understanding what properties we should expect asymptotically as $n \to \infty$.
- Have a sense that graphs of radically different sizes sampled from a fixed schema should be similar; how do we formalize this?

Testing by Sampling

- Lets say we have two graphs, G and H, and we want to say they are ”similar” even though they’re of radically different sizes.
- One idea is to compare whether the two have certain statistics that are close.
- Number of subgraphs is a basic statistic, although the bigger of the two should a priori have more subgraphs of any given type, so we should normalize somehow.

Distribution of Subgraphs

k-Sample Distribution

Given a graph G, and $|G| \geq k \geq 2$, define a probability distribution $\sigma_{G,k}$ on all graphs of size k by:

- For each ordered length k subset of vertices:
- Count which subgraph is induced by restricting to those k vertices.
- When done, normalize to probability 1.

In the case $k > |G|$, we trivially extend the definition by putting all the probability mass on the graph with no vertices.

- If $k = 2$, we are just calculating the density of edges.
- (Example on board)
- $\sigma_{\cdot,k}$ gives us a way to compare between graphs of wildly different sizes.

Sampling Distance

- We can compare distributions $\sigma_{G,k}$ and $\sigma_{H,k}$ by examining the maximum amount they differ on any subgraph. Formally, we define the *variation distance*:

$$d_{\text{var}}(\sigma_{G,k}, \sigma_{H,k}) = \sup_{X \text{a graph of size } k} |\sigma_{G,k}(X) - \sigma_{H,k}(X)|$$

- The selection of k was arbitrary, so we encode examining all k simultaneously by adding together the variation distances for each k in a convergent sum; this is the *sampling distance*:

$$\delta_{\text{samp}}(G, G') = \Sigma_{k=1}^{\infty} \frac{1}{2^k} d_{\text{var}}(\sigma_{G,k}, \sigma_{G',k})$$

- Sampling distance provides provides a way to compare two graphs of different scales; if they are close enough in the sampling distance, they should be close in any statistic that depends continuously on subgraph densities.
- (Example on board)

Limits in the Sampling Distance

- Let's imagine sampling from $G(n, p)$ as $n \to \infty$: with high probability, their subgraph densities will converge for any fixed k.
- This means that with high probability, this sequence is Cauchy in the sampling distance.
- Clearly there is no finite graph this converges to, since the subgraph distributions would differ for large k. What does the completion of this space look like?

Graphons

Definition

A *graphon* is a symmetric, measurable function $[0,1]^2 \to [0,1]$.

- Every graph is a graphon: turn the adjacency matrix into a pixel-picture (example on board.)
- We can think of graphons as edge-weighted graphs on a Continuum of vertices.
- Measurability allows us to calculate graph properties via integration; for instance, the "density of triangles" in a graphon W is given by

$$\int_{[0,1]^3} W(x,y)W(y,z)W(z,x)dxdydz$$

  This allows us to extend sampling distance to graphons in the natural way.
- Sampling distance becomes only a pseudometric in the space of graphons. We can mod out by equivalence (known as weak isomorphism) if we want.

Graphons as Random Graph Schemas

We can sample a graph of arbitrary size from a graphon as follows:

- Uniformly sample points $X = \{x_1, \ldots, x_n\}$ from $[0, 1]$.
- Restrict your graphon to $X \times X$ to get a $[0, 1]$-edge-weighted graph.
- Sample a simple graph from your weighted graph by including each edge with probability equal to its weight.

You can show that by taking $n \to \infty$, with high probability this creates a sequence converging to the original graphon in the sampling distance.

Cut Distance of Finite Graphs

Definition

Let G, G' be two graphs with node set $\{1, \ldots, n\}$. For subsets S, T of $\{1, \ldots, n\}$, let $e_G(S, T)$ be the number of edges starting in S and ending in T (edges in $S \cap T$ counted twice.) Define their *cut distance* as

$$d_{\text{cut}}(G, G') = \max_{S,T \subseteq V(G)} \frac{|e_G(S, T) - e_{G'}(S, T)|}{n^2}$$

- This is an equivalent (pseudo)metric that's often easier to calculate or work with.
- (Example on board.)
- We usually don't care about labels, so can generalize the cut distance by minimizing over relabelings. We can also extend it to graphs of different sizes by blowing up the number of vertices of each to their least common multiple.

Cut Distance of Graphons

Definition

Let W, W' be graphons. Their cut distance is defined as

$$d_{\text{cut}}(W, W') = \inf_{\phi} \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} W(x,y) - W'(\phi(x), \phi(y)) dxdy \right|$$

where ϕ is taken over all measure preserving transformations of $[0,1]$ (which plays the role of a relabeling.)

This is a strict translation of cut distance for finite graphs to the continuous case, except the addition of measurability.
FACT: The inf and sup in the above computation are actually achieved.

Grab-bag

- The space of graphons in either the cut or sampling distance is compact; this implies that it is totally bounded. This turns out to be closely related to the Szemeredi Regularity Lemma
- Total boundedness implies that for any fixed cut distance, there exists a finite sample of representatives whose Voronoi cells are smaller than that size. The Regularity Lemma is a result about approximating graphs by "nearby" quasirandom graphs.
- Graphons are also useful in problems in extremal graph theory. A classic extremal graph theory problem: what is the largest edge-density of a graph that includes no triangle?
- Just like moving from rationals to reals allows us to apply limiting techniques to optimization problems over the rationals, moving from finite graphs to graphons allows us to do the same with optimizing over subgraph density related properties.

...There's a catch

This entire discussion is not very useful unless the graph sequences involved have edge density > 0.

Sampling doesn't tell us much in the non-dense case, since almost all sampled graphs will be trivial. Non-dense graph sequences converge to the trivial graphon.

There are notions of distance and limit-representing objects in the case of bounded degree, but it's an active area of research.

Some of the real-world graphs that we care about seem closer to the bounded degree case than the dense case, so this isn't the whole story.

Summary

- We can meaningfully complete the space of graphs with respect to subgraph sampling.
- Graphons provide representing objects for certain graph generation algorithms and extremal graph property solutions.
- Working in the space of graphons makes things easier?
- This whole story is a bit different if we don't assume non-zero edge density.

End