Lecture 4: Graph Limits and Graphons

36-781, Fall 2016

3 November 2016

Abstract

Contents

1 Reprise: Convergence of Dense Graph Sequences

Let's review where we were by the end of last lecture.

Two graphs f and g are **isomorphic**, $f \simeq g$, when there is an invertible mapping ϕ from $V(f)$ to $V(g)$, such that $(i,j) \in E(f)$ if and only if $(\phi(i), \phi(j)) \in E(g)$. In words, the two graphs are identical up to the names of the nodes.

Fix two graphs f and g, with k and n nodes respectively. The number of subgraph isomorphism of the motif f in g, $\mathrm{Iso}(f,g)$, is the number of k-node (induced) subgraphs of g which are isomorphic to f. The isomorphism density of f is the number of isomorphisms divided by the number of k-node subgraphs in g:

$$t_{iso}(f,g) \equiv \frac{\mathrm{Iso}(f,g)}{n(n-1)\ldots(n-(k-1))} \tag{1}$$

We'll agree to define $t_{iso}(f,g)$ to be 0 if $k > n$. Equivalently, if we write $G[k]$ for the induced subgraph we get by randomly sampling k distinct nodes from g, we have

$$t_{iso}(f,g) = \Pr(f \simeq G[k]) \tag{2}$$

It seems fairly reasonable a priori to look at the density of motifs as telling us a lot about what the network is like — it includes the density of edges, of triangles, of k-stars, of k-cycles, etc., and so encodes all sorts of facts about the topology of the graph and how it's put together. Accordingly, we *define* a sequence of graphs $g_1, g_2, \ldots g_m, \ldots$ to be convergent when, for all motifs f,

$$t_{iso}(f, g_m) \tag{3}$$

converges as $m \to \infty$.

Notice that for any fixed f, the $t_{iso}(f, g_m)$ is just a sequence of real numbers, and we know what it is for one of those to converge[1], so our formula is at least well-defined. We also observed that if the sequence is constant, i.e. if all the $g_m = g_1$, then all the motif densities are constant, so there are at least *some* convergent graph sequences. We didn't show that there were non-trivial convergent sequences, but there are.

We also asserted that while convergence of all the motif densities was what we'd like, calculating $Iso(f, g)$ is hard[2]. We therefore introduced a weaker notion, that of an **injective homomorphism**, namely an invertible map ϕ from $V(f)$ to $V(g)$ where $(i, j) \in E(f)$ implies $(\phi(i), \phi(j)) \in V(g)$. Isomorphisms preserve both edges and non-edges; these homomorphisms preserve the edges of f, but not necessarily the non-edges. We can thus define

$$t_{injective}(f, g) \equiv \Pr\left(f \precsim G[k]\right) \tag{4}$$

where for two graphs f and g, $f \precsim g$ is to be read as "f and g have the same number of nodes, and, in some numbering of the nodes, $E(f) \subseteq E(g)$". This will prove much easier to work with, but it encodes the same information as t_{iso}. This is because $f \precsim g$ if and only if there is some f' where $f \precsim f'$ and $f' \simeq g$. Therefore

$$t_{injective}(f, g) = \sum_{f' \gtrsim f} t_{iso}(f', g) \tag{5}$$

This implies that knowledge of all the isomorphism densities gives us knowledge of all the injection densities. On the other hand, since this is a linear and invertible system of equations, knowledge of all the injection densities gives us all the isomorphism densities. Thus, all the isomorphism densities will converge along a sequence of graphs g_m if and only if all the injection densities also converge along g_m.

Finally, we went from injection homomorphisms to simple homomorphisms. A **homomorphism** from f to g is a mapping ϕ from $V(f)$ to $V(g)$ such that if $(i, j) \in E(f)$, then $(\phi(i), \phi(j)) \in E(g)$. Not only does the mapping not have to preserve non-edges, it's not required to be invertible, so it can map multiple nodes in

[1] Cauchy: the sequence $x_1, x_2, \ldots$ converges when, for any $\epsilon > 0$, there's an $N(\epsilon)$ such that $m, n > N(\epsilon)$ implies $|x_m - x_n| \leq \epsilon$. Notice that this criterion doesn't require us to give the limit to which the sequence converges.

[2] Indeed, the problem of deciding whether an *arbitrary* f is isomorphic to an induced subgraph of an *arbitrary* g is NP-complete.

f to the same node in g. Write $\mathrm{Hom}(f,g)$ for the number of such homomorphisms. Then

$$t_{hom}(f,g) \equiv \frac{\mathrm{Hom}(f,g)}{n^k} \tag{6}$$

Probabilistically, write $G'[k]$ for the graph we obtain by sampling k nodes *with replacement* from $V(g)$, and taking their induced subgraph. If $G'[k]$ contains multiple copies of a given node, there is no edge between those copies, but they otherwise have the same set of neighbors. Then

$$t_{hom}(f,g) = \Pr\left(f \precsim G'[k]\right) \tag{7}$$

Because sampling with and without replacement are so similar, it should be plausible that $t_{injective}$ and t_{hom} will be close. In fact, if k is held fixed while n is allowed to grow, you can show that the probability of a k sample drawn with replacement contains any repeated nodes is at most $k^2/2n$ (Exercise 1). Therefore

$$|t_{hom}(f,g) - t_{injective}(f,g)| \leq \frac{k^2}{2n} \tag{8}$$

Thus, if we're dealing with a *growing* sequence of graphs graphs, i.e., with $|V(g_m)| \to \infty$ as $m \to \infty$, then $t_{injective}$ converges if and only if t_{hom} also converges.

Summing up, we began by wanting the density of motif isomorphisms to converge, for all motifs. This is equivalent to having the density of injections converge. And that, in turn, is equivalent to having the density of mere homomorphisms converge, at least for a growing number of nodes. So our final criterion is:

Definition 1 *A sequence of graphs g_m, with $|V(g_m)| \to \infty$, **converges** when, for all motifs f, $t_{hom}(f,g_m)$ converges.*

2 Calculating Homomorphism Densities

Being able to go from isomophism densities to homomorphism densities would be an idle mathematical curiosity, unless it were actually easy to calculate homomorphism densities. Here is the trick for doing so.

Start with our favorite graph g with n nodes. Put the nodes in some arbitrary order, and write down its adjacency matrix $\mathbf{a}$. We now squash this into the unit square, getting a function $w_g : [0,1] \times [0,1] \mapsto \{0,1\}$, as follows:

$$w_g(u,v) \equiv a_{\lceil nu \rceil \lceil nv \rceil} \tag{9}$$

We can now calculate many properties of the graph as integrals with respect to w_g. For instance, the density of edges is

$$\int_0^1 \int_0^1 w_g(u_1,u_2) du_1 du_2 \tag{10}$$

To see this, think of drawing two independent and uniformly distributed random numbers between 0 and 1, say U_1 and U_2, then using them to get two nodes, namely $\lceil nU_1 \rceil$ and $\lceil nU_2 \rceil$, and then looking at whether or not there's an edge between those nodes in g.

On the other hand, the density of triangles is

$$\int_0^1 \int_0^1 \int_0^1 w_g(u_1, u_2) w_g(u_2, u_3) w_g(u_3, u_1) du_1 du_2 du_3 \tag{11}$$

Again, we can think of this as sampling three nodes, with replacement, and seeing whether they form a triangle.

You should be able to convince yourself that for an arbitrary k-node motif f, See Exercise 4

$$t_{hom}(f, g) = \int_{[0,1]^k} \prod_{(i,j) \in E(f)} w_g(u_i, u_j) du_{1:k} \tag{12}$$

We've agreed to say that a sequence of graphs $g_1, g_2, \ldots g_m, \ldots$ converge when $t_{hom}(f, g_m)$ converges for all f. By extension, we'll say that the sequence of functions w_{g_m} **graph-converges** when $\int_{[0,1]^k} \prod_{(i,j) \in E(f)} w_{g_m}(u_i, u_j) du_{1:k}$ converges for all f. Notice that this is *not* necessarily the same as the functions converging "pointwise", i.e., as saying that $w_{g_m}(u, v)$ converges for each fixed $(u, v) \in [0, 1]^2$.

3 Graphons

One of the goals of the first part of lecture 3 was to prime you for the notion that when a sequence of mathematical objects converges, the limit is not necessarily an object of the same sort, but me be a member of some broader class which contains the objects in the sequence as special cases. Thus

- The limit of a sequence of rational numbers may be an arbitrary real number;
- The limit of a sequence of discrete probability vectors can be an point on the simplex;
- The limit of a sequence of empirical probability distributions (= mixture of delta functions) can be a continuous probability density.

By now you should be ready to accept the idea that the limit of a sequence of graphs need not be another graph, but something else[3]. That something else is called a

[3] There is actually a subtle mathematical issue here. Suppose we start with a space $\mathcal{X}$, and have some notion of a sequence of objects from it, say $\{x_n\}$, converging, but there are convergent sequences where the limit is not in $\mathcal{X}$. We can then switch to the space $\mathcal{X}'$ of all convergent $\mathcal{X}$-valued sequences. We define the distance between two such sequences as the limit of the distances between $\{x_n\}$ and $\{y_n\}$ as $n \to \infty$, and identify sequences whose distance tends to 0. The original points in $\mathcal{X}$ can then be embedded in $\mathcal{X}'$ as the constant-valued sequences, the original notion of convergent sequence carries over, and all limits are now points in the space, i.e., the limit of a sequence of points in $\mathcal{X}$ is an equivalence class of convergent $\mathcal{X}$-valued sequences. For example, real numbers can be viewed as equivalence classes of convergent sequences of rational numbers. We therefore don't have to *posit* the existence of limiting objects, we can *construct* them from the objects we started with. Of course, there may be other ways of defining the limiting objects, but they *can* always be defined through this operation of "completion".

graphon.

The clue to seeing what sort of thing a graphon is comes from the last section, where we calculated homomorphism densities as integrals of functions over the unit square (Eq. 12). For each f, $t_{hom}(f, g_m)$ is just a number, so for the sequence to converge, there must be some limiting number it converges *to*; it's natural to write

$$t_{hom}(f, g_\infty) \equiv \lim_{m\to\infty} t_{hom}(f, g_m) \tag{13}$$

but it's important to remember that we've not established what kind of thing g_∞ is yet, that the right-hand side is well-defined and the left-hand side is just a notationally-convenient abbreviation for it. Now, by Eq. 12,

$$t_{hom}(f, g_\infty) = \lim_{m\to\infty} \int_{[0,1]^k} \prod_{(i,j)\in E(f)} w_{g_m}(u_i, u_j) d\, u_{1:k} \tag{14}$$

We are going to *define* the graph-limit of a sequence of functions w_{g_m} as the w_∞ for which

$$\lim_{m\to\infty} \int_{[0,1]^k} \prod_{(i,j)\in E(f)} w_{g_m}(u_i, u_j) d\, u_{1:k} = \int_{[0,1]^k} \prod_{(i,j)\in E(f)} w_\infty(u_i, u_j) d\, u_{1:k} \tag{15}$$

If this were a proper mathematics class, at this point we'd need to do a couple of things.

1. Prove that there exists a w_∞ which satisfies that equation for any given motif f.
2. Prove that the *same* w_∞ can satisfy that equation simultaneously for all f.
3. Characterize the properties of these w_∞.

Since this is not a proper mathematics class, I will just *assert* that such limiting functions exist, and that they are precisely the symmetric functions from $[0,1]^2$ to $[0,1]$.

Definition 2 *Any function $w : [0,1] \times [0,1] \mapsto [0,1]$ for which $w(u,v) = w(v,u)$ is a **graphon function**, and*

$$t_{hom}(f, w) \equiv \int_{[0,1]^k} \prod_{(i,j)\in E(f)} w(u_i, u_j) d\, u_{1:k}$$

*Two graphon functions are **equivalent** when $t_{hom}(f, w_1) = t_{hom}(f, w_2)$ for all f. A **graphon** is an equivalence class of graphon functions, and we define $t_{hom}(f, g)$ as $t_{hom}(f, w)$ for $w \in g$. Any function in the equivalence class is a **presentation** (or **representation**) of the graphon.*

Notice that $t_{hom}(f, w)$ is the same for all $w \in g$, so that $t_{hom}(f, g)$ is well-defined.

There are many equivalent representations of the same graphon:

Proposition 1 *Let $\phi : [0,1] \mapsto [0,1]$ be an invertible function which preserves length, so that $\int_{[a,b]} dx = \int_{\phi([a,b])} dx$. If $w(u,v)$ is a graphon function, then so is $w \circ \phi = w(\phi(u), \phi(v))$, and the two graphon functions are equivalent.*

PROOF: Exercise 6.

Notice that the map $\phi(x) = 1 - x$ is invertible and length-preserving, so there are always at least *two* representations of any graphon. But so is $\phi(x) = 1 - x$ for $x < 0.5$, $\phi(x) = x - 0.5$ for $x \geq 0.5$, etc., *ad infinitum*.

Naturally, people often loosely speak of a *particular* function w as "the graphon", but it's often important to keep the two notions separate. (Think of the distinction between a vector as a geometric object, drawn between two points in space, and all the different the column matrices of numbers which represent that vector in particular coordinate systems.)

We may now saw that the limit of a sequence of graphs is a graphon, and have that mean something. Specifically:

Definition 3 *A sequence of graphs g_m converges on a graphon g_∞ when $t_{hom}(f, g_m) \to t_{hom}(f, g_\infty)$.*

We similarly define convergence for sequences of graphons through convergence of their homomorphism densities.

4 The Cut Metric

We have defined convergence for sequences of graphs, and for sequences of graphons, through convergence of all their homomorphism densities. We haven't needed, yet, to define a distance between graphs or a distance between graphons, but it would be nice to have a distance where convergence, as we've defined it, coincides with the distance going to zero. That is, it would be nice to **metrize** our notion of convergence. There is, as it happens, just such a metric, known as the **cut metric**[4]. It is customary to build this up from graphs, but it may be easier to grasp it starting from the way it works with graphons, and then go down to the graph case.

Start with any two graphon functions w_1 and w_2. Let's define

$$||w_1 - w_2||_\square \equiv \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} w_1(u,v) - w_2(u,v) du\, dv \right| \tag{16}$$

Now, as in Proposition 1, let ϕ be an invertible, length-preserving map of the unit interval into itself, and let $w \circ \phi(u,v) = w(\phi(u), \phi(v))$. Then we define

$$d_\square(w_1, w_2) \equiv \inf_\phi ||w_1 \circ \phi - w_2||_\square \tag{17}$$

$d_\square$ will be zero between equivalent graphon functions, and so we can extend $d_\square$ to graphons in the obvious way — graphons are close iff they have close representations.

[4] It's not the only one; Diaconis and Janson (2008) offers another one, based directly on weighted differences in homomorphism densities.

$d_\square$ also gives us a metric between graphs; we just define $d_\square(g_1, g_2)$ to be $d_\square(w_{g_1}, w_{g_2})$. Finally, if for a single graph g we consider

$$||g||_\square \equiv ||w_g||_\square = \sup_{S,T \subseteq [0,1]} \int_{S \times T} w_g(u,v) du\, dv \tag{18}$$

This is at least

$$\frac{1}{n^2} \max_{S,T \subseteq 1:n} \sum_{i \in S} \sum_{j \in T} a_{ij} \tag{19}$$

The sum over dyads is the number of edges which "cross the cut" between the set of nodes S and the set of nodes T, hence this is the **cut norm** of the graph, and the root of the name **cut metric**.

4.1 Connection to Homomorphism Densities

Proposition 2 *For any two graphon functions w_1, w_2, and any finite simple graph f,*

$$|t_{hom}(f, w_1) - t_{hom}(f, w_2)| \leq |E(f)| d_\square(w_1, w_2) \tag{20}$$

PROOF (after Borgs *et al.* (2006, Lemma 4.4)): If we can prove

$$|t_{hom}(f, w_1) - t_{hom}(f, w_2)| \leq |E(f)| \, ||w_1 - w_2||_\square \tag{21}$$

we'll be done, since $t_{hom}(f, w_1) = t_{hom}(f, w_1 \circ \phi)$, and $d_\square(w_1, w_2) = \inf_\phi ||w_1 \circ \phi - w_2||_\square$. So let's focus on Eq. 21.

$$\begin{aligned}
& t_{hom}(f, w_1) - t_{hom}(f, w_2) && (22)\\
&= \int_{[0,1]^k} \prod_{(i,j) \in E(f)} w_1(u_i, u_j) du_{1:k} - \int_{[0,1]^k} \prod_{(i,j) \in E(f)} w_2(u_i, u_j) du_{1:k} \\
&= \int_{[0,1]^k} \left(\prod_{(i,j) \in E(f)} w_1(u_i, u_j) - \prod_{(i,j) \in E(f)} w_2(u_i, w_j) \right) du_{1:k} && (23)\\
&= \int_{[0,1]^k} \sum_{(i,j) \in E(f)} \left(\prod_{(i',j') \leq (i,j)} w_1(u_{i'}, u_{j'}) \right) \left(w_1(u_i, u_j) - w_2(u_i, u_j) \right) \left(\prod_{(i',j') > (i,j)} w_2(u_{i'}, u_{j'}) \right) du_{1:k} && (24)\\
&= \sum_{(i,j) \in E(f)} \int_{[0,1]^k} \left(\prod_{(i',j') \leq (i,j)} w_1(u_{i'}, u_{j'}) \right) \left(w_1(u_i, u_j) - w_2(u_i, u_j) \right) \left(\prod_{(i',j') > (i,j)} w_2(u_{i'}, u_{j'}) \right) du_{1:k} && (25)
\end{aligned}$$

In the next-to-last line, I've imposed an order on the pairs (i, j) in $E(f)$, but it should be clear that every ordering of the edges will give the same answer. In the last line,

I've simply exchanged the order of summation and integration. Now

$$\begin{aligned}
&\int_{[0,1]^k} \left(\prod_{(i',j') \leq (i,j)} w_1(u_{i'}, u_{j'}) \right) \left(w_1(u_i, u_j) - w_2(u_i, u_j) \right) \left(\prod_{(i',j') > (i,j)} w_2(u_{i'}, u_{j'}) \right) du_{1:k} \\
&= \int_{[0,1]^2} h(u_i)(w_1(u_i, u_j) - w_2(u_i, u_j))h'(u_j) du_i du_j \qquad (26)
\end{aligned}$$

for some functions h, h' which are guaranteed to be ≤ 1, because they're integrals (over all the us except u_i and u_j) of products of w_1 and w_2. Therefore

$$\begin{aligned}
&\int_{[0,1]^2} h(u_i)(w_1(u_i, u_j) - w_2(u_i, u_j))h'(u_j) du_i du_j \\
&= \left| \int_{[0,1]^2} h(u_i)(w_1(u_i, u_j) - w_2(u_i, u_j))h'(u_j) du_i du_j \right| \qquad (27) \\
&\leq \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} w_1(u_i, u_j) - w_2(u_i, u_j) du_i du_j \right| \qquad (28) \\
&= d_{\square}(w_1, w_2) \qquad (29)
\end{aligned}$$

Substituting into Eq. 25, we get

$$|t_{hom}(f, w_1) - t_{hom}(f, w_2)| \leq |E(f)| d_{\square}(w_1, w_2) \qquad (30)$$

□

We can now connect this metric to the question of whether a sequence of graphs g_m converges to a graphon g. Proposition 2 tells us that if $d_{\square}(g_m, g) \to 0$ as $m \to \infty$, then g_m converges to g. To go in the other direction, and show that g_m converging to g implies $d_{\square}(g_m, g)$ requires a more subtle argument, resting on how the distribution of induced subgraphs drawn from g_m stabilize. We will accordingly look at this later, when we consider the connection between graphons and network models; if you're impatient, see Borgs *et al.* (2006).

5 Dense vs. Sparse Graph Sequences

Recall that a sequence of graphs is **dense** when its density, the ratio of actual edges to possible edges, doesn't go to zero:

$$\frac{|E(g_m)|}{|V(g_m)|^2} = O(1) \qquad (31)$$

If, on the other hand, the density vanishes asymptotically,

$$\frac{|E(g_m)|}{|V(g_m)|^2} = o(1) \qquad (32)$$

then the graph sequence is **sparse**.

Our definition of convergence for a graph sequence applies, strictly speaking, to both dense and sparse graph sequences, but it's only interesting for the dense ones. This is because every sparse graph sequence which converges has for its limit the empty, all-0 graphon. To see this, notice that the edge density is just $t_{hom}(K_2, g_m)$, where K_2 is the complete graph on 2 nodes. Since, in a sparse graph sequence, the edge density goes to zero, and we're positing that g_m converges, we must have that $\int_{[0,1]^2} w_\infty(u_1, u_2) du_1 du_2 = 0$. But since w_∞ is non-negative, this means that $w_\infty = 0$ (almost everywhere). See Exercise 7.

Of course, just because sparse graph sequences don't have interesting limits in terms of homomorphism densities doesn't mean that there mightn't be other useful notions of convergence for them. We will come back to this issue, which has loomed increasingly large in recent work on network modeling, towards the end of the course. For now, however, we'll confine ourselves to dense graphs.

6 Exercises

1. Show that if k objects are drawn with replacement from a size n, the probability that one or more objects are repeated is at most $k^2/2n$.

2. Convince yourself that when $g \simeq g'$,

$$t_{iso}(f, g) = t_{iso}(f, g') \tag{33}$$
$$t_{injection}(f, g) = t_{injection}(f, g') \tag{34}$$
$$t_{hom}(f, g) = t_{hom}(f, g') \tag{35}$$

   That is, show that all the variant motif densities are the same for isomorphic graphs.

3. Using Exercise 2, show that a sequence $g_1, g_2, \ldots$ is convergent if and only if every sequence $g'_1, g'_2, \ldots$, where every $g_m \simeq g'_m$, is also convergent.

4. Eq. 9 defines w_g using the adjacency matrix of g, and so a *particular* ordering of the nodes in g. Show that the right-hand side of Eq. 12 is always the same regardless of the ordering of the nodes used to get w_g; equivalently, show that the right-hand side doesn't change when we substitute $w_{g'}$ for w_g, for any $g' \simeq g$.

5. Is this true?

$$t_{iso}(f, g) = \int_{[0,1]^k} \left(\prod_{(i,j) \in E(f)} w_g(u_i, u_j) \prod_{(i,j) \notin E(f)} 1 - w_g(u_i, u_j) \right) du_{1:k} \tag{36}$$

6. Prove Proposition 1.

7. §5 showed that if a sequence of sparse graphs converges, then it converges to the empty graphon, $w(u,v) = 0$.

   (a) (Easy) Show that if a sequence of graphs converges to the empty graphon, the sequence is sparse.

   (b) (Harder) Set $\rho_m \equiv \frac{|E(g_m)|}{|V(g_m)|^2}$. Can you find an upper bound for $t_{hom}(f,g)$ in terms of ρ_m, and possibly also of $|V(f)|$ and $|E(f)|$? *Hint:* Consider the product generalization of Hölder's inequality,

$$\int \left| \prod_{s=1}^{r} h_i(x) \mu(dx) \right| \leq \prod_{s=1}^{r} \left| \int h_i^r(x) \mu(dx) \right|^{1/r} \tag{37}$$

   (c) (Harder) Does every sparse graph sequence converge to the empty graphon? That is, does every sparse graph sequence converge?

References

Borgs, Christian, Jennifer T. Chayes, László Lovász, Vera T. Sós, Balázs Szegedy and Katalin Vesztergombi (2006). "Graph Limits and Parameter Testing." In *Proceedings of the 38th Annual ACM Symposium on the Theory of Computing [STOC 2006]*, pp. 261–270. New York: ACM. URL `http://research.microsoft.com/en-us/um/people/jchayes/Papers/TestStoc.pdf`.

Diaconis, Persi and Svante Janson (2008). "Graph Limits and Exchangeable Random Graphs." *Rendiconti di Matematica e delle sue Applicazioni*, **28**: 33–61. URL `http://arxiv.org/abs/0712.2749`.