Stat260/CS294: Spectral Graph Methods

Lecture 8 - 02/17/2015

Lecture: Expanders, in theory and in practice (1 of 2)

Lecturer: Michael Mahoney

Scribe: Michael Mahoney

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

8 Introduction: Expanders, in theory and in practice (1 of 2)

Expander graphs, also called *expanders*, are remarkable structures that are widely-used in TCS and discrete mathematics. They have a wide range of applications:

- They reduce the need for randomness and are useful for derandomizing randomized algorithms so, if random bits are a valuable resource and thus you want to derandomized some of the randomized algorithms we discussed before, then this is a good place to start.
- They can be used to find good error-correcting codes that are efficiently encodable and decodable—roughly the reason is that they spread things out.
- They can be used to provide a new proof of the so-called PCP theorem, which provides a new characterization of the complexity class NP, and applications to the hardness of approximate computation.
- They are a useful concept in data analysis applications, since expanders look random, or are empirically quasi-random, and it is often the case that the data, especially when viewed at large, look pretty noisy.

For such useful things, it is somewhat surprising that (although they are very well-known in computer science and TCS in particular due to their algorithmic and complexity connections) expanders are almost unknown outside computer science. This is unfortunate since:

- The world is just a bigger place when you know about expanders.
- Expanders have a number of initially counterintuitive properties, like they are very sparse and very well-connected, that are typical of a lot of data and thus that are good to have an intuition about.
- They are "extremal" in many ways, so they are a good limiting case if you want to see how far you can push your ideas/algorithms to work.
- Expanders are the structures that are "most unlike" low-dimensional spaces—so if you don't know about them then your understanding of the mathematical structures that can be used to describe data, as well as of possible ways that data can look will be rather limited, *e.g.*, you might think that curved low-dimensional spaces are good ideas.

Related to the comment about expanders having extremal properties, if you know how your algorithm behaves on, say, expanders, hypercubes (which are similar and different in interesting ways), trees (which we won't get to as much, but will mention), and low-dimensional spaces, they you probably have a pretty good idea of how it will behave on your data. That is very different than knowing how it will behave in any one of those places, which doesn't give you much insight into how it will behave more generally; this extremal property is used mostly by TCS people for algorithm development, but it can be invaluable for understanding how/when your algorithm works and when it doesn't on your non-worst-case data.

We will talk about expander graphs. One issue is that we can define expanders both for degreehomogeneous graphs as well as for degree-heterogeneous graphs; and, although many of the basic ideas are similar in the two cases, there are some important differences between the two cases. After defining them (which can be done via expansion/conductance or the leading nontrivial eigenvalue of the combinatorial/normalized Laplacian), we will focus on the following aspects of expanders and expander-like graphs.

- Expanders are graphs that are very well-connected.
- Expanders are graphs that are sparse versions/approximations of a complete graph.
- Expanders are graphs on which diffusions and random walks mix rapidly.
- Expanders are the metric spaces that are least like low-dimensional Euclidean spaces.

Along the way, we might have a chance to mention a few other things, e.g.: how big λ_2 could be with Ramanujan graphs and Wigner's semicircle result; trivial ways with d_{max} to extend the Cheeger Inequality to degree-heterogeneous graphs, as well as non-trivial ways with the normalized Laplacian; pseudorandom graphs, converses, and the Expander Mixing Lemma; and maybe others.

Before beginning with some definitions, we should note that we can't draw a meaningful/interpretable picture of an expander, which is unfortunate since people like to visualize things. The reason for that is that there are no good "cuts" in an expander—relatedly, they embed poorly in low-dimensional spaces, which is what you are doing when you visualize on a two-dimensional piece of paper. The remedy for this is to compute all sorts of other things to try to get a non-visual intuition about how they behave.

8.1 A first definition of expanders

Let's start by working with *d*-regular graphs—we'll relax this regularity assumption later. But many of the most extremal properties of expanders hold for degree-regular graphs, so we will consider them first.

Definition 1. A graph G = (V, E) is d-regular if all vertices have the same degree d, i.e., each vertex is incident to exactly d edges.

Also, it will be useful to have the following notion of the set of edges between two sets S and T (or from S to T), both of which are subsets of the vertex set (which may or may not be the complement of each other).

Definition 2. For $S, T \subset V$, denote

$$E(S,T) = \{ (u,v) \in E | u \in S, v \in T \}.$$

Given this notation, we can define the expansion of a graph. (This is slightly different from other definitions I have given.)

Definition 3. The expansion or edge expansion ratio of a graph G is

$$h(G) = \min_{S:|S| \le \frac{n}{2}} \frac{E(S,S)}{|S|}$$

Note that this is slightly different (just in terms of the scaling) than the edge expansion of G which we defined before as:

$$\phi\left(G\right) = \min_{S \subset V : |S| \le \frac{|V|}{2}} \frac{E\left(S,S\right)}{d|S|}.$$

We'll use this today, since I'll be following a proof from HLW, and they use this, and following their notation should make it easier. There should be no surprises, except just be aware that there is a factor of d difference from what you might expect.

(As an aside, recall that there are a number of extensions of this basic idea to measure other or more fine versions of this how well connected is a graph:

- Different notions of boundary—e.g., vertex expansion.
- Consider size-resolved minimum—in Markov chains and how good communities are as a function of size.
- Different denominators, which measure different notions of the "size" of a set S:
 - Sparsity or cut ratio: $\min \frac{E(S,\bar{S})}{|S| \cdot |\bar{S}|}$ —this is equivalent to expansion in a certain sense that we will get to.
 - Conductance or NCut—this is identical for *d*-regular graphs but is more useful in practice and gives tighter bounds in theory if there is degree heterogeneity.

We won't deal with these immediately, but we will get back to some later. This ends the aside.)

In either case above, the expansion is a measure to quantify how well-connected is the graph. Given this, informally we call a *d*-regular graph G an *expander* if $h(G) \ge \epsilon$ where ϵ is a constant. More precisely, let's define an expander:

Definition 4. A graph G is a (d, ϵ) -expander if it is d-regular and $h(G) \ge \epsilon$, where ϵ is a constant, independent of n.

Alternatively, sometimes expansion is defined in terms of a sequence of graphs:

Definition 5. A sequence of d-regular graphs $\{G_i\}_{i \in \mathbb{Z}^+}$ is a family of expander graphs if $\exists \epsilon > 0$ s.t. $h(G_i) \ge \epsilon, \forall i$.

If we have done the normalization correctly, then $h(G) \in [0, d]$ and $\phi(G) \in [0, 1]$, where large means more expander-like and small means that there are good partitions. So, think of the constant ϵ as d/10 (and it would be 1/10, if we used $\phi(G)$ normalization). Of course, there is a theory/practice issue here, e.g., sometimes you are given a single graph and sometimes it can be hard to tell a moderately large constant from a factor of $\log(n)$; we will return to these issues later.

8.2 Alternative definition via eigenvalues

Although expanders can be a little tricky and counterintuitive, there are a number of ways to deal with them. One of those ways, but certainly not the only way, is to compute eigenvectors and eigenvalues associated with matrices related to the graph. For example, if we compute the second eigenvalue of the Laplacian, then we have Cheeger's Inequality, which says that if the graph G is an expander, then we have a (non-tight, due to the quadratic approximation) bound on the second eigenvalue, and vice versa. That is, one way to test if a graph is an expander is to compute that eigenvalue and check.

Of central interest to a lot of things is λ_2^{LAP} , which is the Fiedler value or second smallest eigenvalue of the Laplacian. Two things to note:

- If we work with Adjacency matrices rather than Laplacians, then we are interested in how far λ_2^{ADJ} is from d.
- We often normalized things so as to interpret them in terms of a random walk, in which case the top eigenvalue = 1 with the top eigenvector being the probability distribution. In that case, we are interested in how far λ_2 is from 1.

(Since I'm drawing notes from several different places, we'll be a little inconsistent on what the notation means, but we should be consistent within each class or section of class.)

Here is Cheeger's Inequality, stated in terms of h(G) above.

• If $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of the Laplacian (not normalized, i.e. D - A) of a *d*-regular graph *G*, then:

$$\frac{\lambda_2}{2} \le h(G) \le \sqrt{2d\lambda_2}$$

The \sqrt{d} in the upper bound is due to our scaling.

Alternatively, here is Cheeger's Inequality, stated in terms of h(G) for an Adjacency Matrix.

• If $d = \mu_1 \ge \mu_2 \ge \ldots \ge \mu_n$ are the eigenvalues of the Adjacency Matrix A(G) of d-regular graph G, then:

$$\frac{d-\mu_2}{2} \le h(G) \le \sqrt{2d(d-\mu_2)}$$

Therefore, the expansion of the graph is related to its spectral gap $(d - \mu_2)$. Thus, we can define a graph to be an expander if $\mu_2 \leq d - \epsilon$ or $\lambda_2 \geq \epsilon$ where λ_2 is the second eigenvalue of the matrix L(G) = D - A(G) where D is the diagonal degree matrix. Slightly more formally, here is the alternate definition of expanders:

Definition 6. A sequence of d-regular graphs $\{G_n\}_n \in \mathbb{N}$ is a family of expander graphs if $|\lambda_i^{ADJ}| \leq d - \epsilon$, i.e. if all the eigenvalues of A are bounded away from d

Remark. The last requirement can be written as $\lambda_2^{LAP} \ge c, \forall n, i.e.$, that all the eigenvalues of the Laplacian bounded below and away from c > 0.

In terms of the edge expansion $\phi(G)$ we defined last time, this definition would become the following.

Definition 7. A family of constant-degree expanders is a family of graphs $\{G_n\}_{n\in\mathbb{N}}$ s.t. each graph in G_n is d-regular graph on n vertices and such that there exists an absolute constant ϕ^* , independent of n, s.t. $\phi(G_n) \ge \phi^*$, for all n.

8.3 Expanders and Non-expanders

A clique or a complete graph is an expander, if we relax the requirement that the *d*-regular graph have a fixed *d*, independent of *n*. Moreover, $G_{n,p}$ (the random graph), for $p \gtrsim \frac{\log(n)}{n}$ is also an expander, with *d* growing only weakly with *n*. (We may show that later.) Of greatest interest—at least for theoretical considerations—is the case that *d* is a constant independent of *n*.

8.3.1 Very sparse expanders

In this case, the idea of an expander, *i.e.*, an *extremely* sparse and *extremely* well-connected graph is nice; but do they exist? It wasn't obvious until someone proved it, but the answer is YES. In fact, a typical *d*-regular graph is an expander with high probability under certain random graph models. Here is a theorem that we will not prove.

Theorem 1. Fix $d \in \mathbb{Z}^+ \geq 3$. Then, a randomly chosen d-regular graph is an expander w.h.p.

Remark. Clearly, the above theorem is false if d = 1 (in which case we get a bunch of edges) or if d = 2 (in which case we get a bunch of cycles); but it holds even for d = 3.

Remark. The point of comparison for this should be if $d \gtrsim \frac{\log(n)}{n}$. In this case, "measure concentration" in the asymptotic regime, and so it is plausible (and can be proved to be true) that the graph has no good partitions. To understand this, recall that one common random graph model is the Erdos-Renyi $\mathcal{G}_{n,p}$ model, where there are n vertices and edges are chosen to exist with probability p. (We will probably describe this ER model as well as some of its basic properties later; at a minimum, we will revisit it when we talk about stochastic blockmodels.) The related $\mathcal{G}_{n,m}$ model is another common model where graphs with n vertices and m edges are chosen uniformly at random. An important fact is that if we set p such that there are on average m edges, then $\mathcal{G}_{n,m}$ is very similar (in strong senses of the word) to $\mathcal{G}_{n,p}$ —if $p \geq \log n/n$. (That is the basis for the oft-made observation that $\mathcal{G}_{n,m}$ and $\mathcal{G}_{n,p}$ are "the same.") However, for the above definition of expanders, we require in addition that d is a constant. Importantly, in that regime, the graphs are sparse enough that measure hasn't concentrated, and they are not the same. In particular, if p = 3/n, $\mathcal{G}_{n,p}$ usually generates a graph that is not connected (and there are other properties that we might return to later). However, (by the above theorem) $\mathcal{G}_{n,m}$ with corresponding parameters usually yields a connected graph with very high expansion. We can think of randomized expander construction as a version of $\mathcal{G}_{n,m}$, further constrained to *d*-regular graphs.

Remark. There are explicit deterministic constructions for expanders—they have algorithmic applications. That is an FYI, but for what we will be doing that won't matter much. Moreover, later we will see that the basic idea is still useful even when we aren't satisfying the basic definition of expanders given above, e.g., when there is degree heterogeneity, when a graph has good small but no good large cuts, etc.

8.3.2 Some non-expanders

It might not be clear how big is big and how small is small—in particular, how big can h (or λ) be. Relatedly, how "connected" can a graph be? To answer this, let's consider a few graphs.

- Path graph. (For a path graph, $\mu_1 = \Theta(1/n^2)$). If we remove 1 edge, then we can cut the graph into two 50-50 pieces.
- Two-dimensional $\sqrt{n} \times \sqrt{n}$ grid. (For a $\sqrt{n} \times \sqrt{n}$ grid, $\mu_1 = \Theta(1/n)$.) Here, you can't disconnect the graph by removing 1 edge, and the removal of a constant number of edges can only remove a constant number of vertices from the graph. But, it is possible to remove \sqrt{n} of the edge, i.e., an $O(\frac{1}{\sqrt{n}})$ fraction of the total, and split the graph into two 50-50 pieces.
- For a 3D grid, $\mu_1 = \Theta(1/n^{2/3})$.
- A k-dimensional hypercube is still better connected. But it is possible to remove a very small fraction of the edges (the edges of a dimension cut, which are $\frac{1}{k} = \frac{1}{\log(n)}$ fraction of the total) and split half the vertices from the other half.
- For a binary tree, e.g., a complete binary tree on n vertices, $\mu_1 = \Theta(1/n)$.
- For a $K_n K_n$ dumbbell, (two expanders or complete graphs joined by an edge) $\mu_1 = \Theta(1/n)$.
- For a ring on n vertices, $\mu_1 = \Theta(1/n)$.
- Clique. Here, to remove a p fraction of the vertices from the rest, you must remove $\geq p(1-p)$ fraction of the edges. That is, it is very well connected. (While can take a complete graph to be the "gold standard" for connectivity, it does, however, have the problem that it is dense; thus, we will be interested in sparse versions of a complete that are similarly well-connected.)
- For an expander, $\mu_1 = \Theta(1)$.

Remark. A basic question to ask is whether, say, $\mu_1 \sim \Theta(\text{poly}(1/n))$ is "good" or "bad," say, in some applied sense? The answer is that it can be either: it can be bad, if you are interested in connectivity, e.g., a network where nodes are communication devices or computers and edges correspond to an available link; or it can be good, either for algorithmic reasons if e.g. you are interested in divide and conquer algorithms, or for statistical reasons since this can be used to quantify conditional independence and inference.

Remark. Recall the quadratic relationship between $d - \lambda_2$ and h. If $d - \lambda_2$ is $\Theta(1)$, then that is not much difference (a topic which will return to later), but if it is $\Theta(1/n)$ or $\Theta(1/n^2)$ then it makes a big difference. A consequence of this is that by TCS standards, spectral partitioning does a reasonably-good job partitioning expanders (basically since the quadratic of a constant is a constant), while everyone else would wonder why it makes sense to partition expanders; while by TCS standards, spectral partitioning does *not* do well in general, since it has a worst-case approximation factor that depends on n, while everyone else would say that it does pretty well on their data sets.

8.3.3 How large can the spectral gap be?

A question of interest is: how large can the spectral gap be? The answer here depends on the relationship between n, the number of nodes in the graph and d, the degree of each node (assumed to be the same for now.) In particular, the answer is different if d is fixed as n grows or if d grows with n as n grows. As as extreme example of the latter case, consider the complete graph K_n on n vertices, in which case d = n - 1. The adjacency matrix of K_n is $A_{K_n} = J - I$, where J is the all-ones matrix, and where $I = I_n$ is the diagonal identity matrix. The spectrum of the adjacency matrix of K_n is $\{n - 1, -1, \ldots, -1\}$, and $\lambda = 1$. More interesting for us here is the case that d is fixed and n is large, in which case $n \gg d$, in which case we have the following theorem (which is due to Alon and Boppana).

Theorem 2 (Alon-Boppana). Denoting $\lambda = \max(|\mu_2|, |\mu_n|)$, we have, for every d-regular graph:

$$\lambda \ge 2\sqrt{d-1} - o_n(1)$$

So, the eigengap $d - \mu_2$ is not larger than $d - 2\sqrt{d-1}$. For those familiar with Wigner's semicircle law, note the similar form.

The next question is: How tight is this? In fact, it is pretty close to tight in the following sense: there exists constructions of graphs, called Ramanujan graphs, where the second eigenvalue of L(G)is $\lambda_1(G) = d - 2\sqrt{d-1}$, and so the tightness is achieved. Note also that this is of the same scale as Wigner's semicircle law; the precise statements are somewhat different, but the connection should not be surprising.

8.4 Why is *d* fixed?

A question that arises is why is d fixed in the definition, since there is often degree variability in practice. Basically that is since it makes things harder, and so it is significant that expanders exist even then. Moreover, for certain theoretical issues that is important. But, in practice the idea of an expander is still useful, and so we go into that here.

We can define expanders: i.t.o. boundary expansion; or i.t.o. λ_2 . The intuition is that it is well-connected and then get lots of nice properties:

- Well-connected, so random walks converge fast.
- Quasi-random, meaning that it is empirically random (although in a fairly weak sense).

Here are several things to note:

- Most theorems in graph theory go through to weighted graphs, if you are willing to have factors like $\frac{w_{max}}{w_{min}}$ —that is a problem if there is very significant degree heterogeneity or heterogeneity in weights, as is common. So in that case many of those results are less interesting.
- In many applications the data are extremely sparse, like a constant number of edges on average (although there may be a big variance).
- There are several realms of d, since it might not be obvious what is big and what is small:

- d = n: complete (or nearly complete) graph.
- $-d = \Omega(\text{polylog}(n))$: still dense, certainly in a theoretical sense, as this is basically the asymptotic region.
- $-d = \Theta(\text{polylog}(n))$: still sufficiently dense that measure concentrates, *i.e.*, enough concentration for applications; Harr measure is uniform, and there are no "outliers"
- $-d = \Theta(1)$: In this regime things are very sparse, $G_{nm} \neq G_{np}$, so you have a situation where the graph has a giant component but isn't fully connected; so 3-regular random graphs are different than G_{np} with $p = \frac{3}{n}$.

You should think in terms of $d = \Theta(\text{polylog}(n))$ at most, although often can't tell $O(\log n)$ versus a big constant, and comparing trivial statistics can hide what you want.

- The main properties we will show will generalize to degree variability. In particular:
 - High expansion \rightarrow high conductance.
 - Random walks converge to "uniform" distribution \rightarrow random walks converge to a distribution that is uniform over the edges, meaning proportional to the degree of a node.
 - Expander Mixing Property \rightarrow Discrepancy and Empirical Quasi-randomness

So, for theoretical applications, we need $d = \Theta(1)$; but for data applications, think i.t.o. a graph being expander-like, i.e., think of some of the things we are discussing as being relevant for the properties of that data graph, if: (1) it has good conductance properties; and (2) it is empirically quasi-random. This happens when data are extremely sparse and pretty noisy, both of which they often are.

8.5 Expanders are graphs that are very well-connected

Here, we will describe several results that quantify the idea that expanders are graphs that are very well-connected.

8.5.1 Robustness of the largest component to the removal of edges

Here is an example of a lemma characterizing how constant-degree graphs with constant expansion are very sparse graphs with extremely good connectivity properties. In words, what the following lemma says is that the removal of k edges cannot cause more that $O\left(\frac{k}{d}\right)$ vertices to be disconnected from the rest. (Note that it is always possible to disconnect $\frac{k}{d}$ vertices after removing k edges, so the connectivity of an expander is the best possible.)

Lemma 1. Let G = (V, E) be a regular graph with expansion ϕ . Then, after an $\epsilon < \phi$ fraction of the edges are adversarially removed, the graph has a connected component that has at least $1 - \frac{\epsilon}{2\phi}$ fraction of the vertices.

Proof: Let d be the degree of G. Let $E' \subseteq E$ be an arbitrary subset of $\leq \epsilon |E| = \epsilon d \frac{|V|}{2}$ edges. Let C_1, \ldots, C_m be the connected components of the graph $(V, E \setminus E')$, ordered s.t.

$$|C_1| \ge |C_2| \ge \cdots |C_m|.$$

In this case, we want to prove that

$$|C_1| \ge |V| \left(1 - \frac{2\epsilon}{\phi}\right)$$

To do this, note that

$$|E'| \ge \frac{1}{2} \sum_{ij} E(C_i, C_j) = \frac{1}{2} \sum_i E(C_i, V \setminus C_i).$$

So, if $|C_1| \leq \frac{|V|}{2}$, then

$$|E'| \ge \frac{1}{2} \sum_{i} d\phi |C_i| = \frac{1}{2} d\phi |V|,$$

which is a contradiction if $\phi > \epsilon$. On the other hand, if $|C_1| \ge \frac{|V|}{2}$, then let's define S to be $S = C_2 \cup \ldots \cup C_m$. Then, we have

$$|E'| \ge E(C_1, S) \ge d\phi |S|,$$

which implies that

$$|S| \le \frac{\epsilon}{2\phi} |V|,$$

and so $|C_1 \ge \left(1 - \frac{\epsilon}{2\phi}\right)|V|$, from which the lemma follows.

\diamond

8.5.2 Relatedly, expanders exhibit quasi-randomness

In addition to being well-connected in the above sense (and other senses), expanders also "look random" in certain senses.

One direction For example, here I will discuss connections with something I will call "empirical quasi-randomness." It is a particular notion of things looking random that will be useful for what we will discuss. Basically, it says that the number of edges between any two subsets of nodes is very close to the expected value, which is what you would see in a random graph. Somewhat more precisely, it says that when λ below is small, then the graph has the following quasi-randomness property: for every two disjoint sets of vertices, S and T, the number of edges between S and T is close to $\frac{d}{n}|S| \cdot |T|$, i.e., what we would expect a random graph with the same average degree d to have. (Of course, this could also hide other structures of potential interest, as we will discuss later, but it is a reasonable notion of "looking random" in the large scale.) Here, I will do it in terms of expansion—we can generalize it and do it with conductance and discrepancy, and we may do that later.

We will start with the following theorem, called the "Expander Mixing Lemma," which shows that if the spectral gap is large, then the number of edges between two subsets of the graph vertices can be approximated by the same number for a random graph, *i.e.*, what would be expected on average, so it looks empirically random. Note that $\frac{d}{n}|S| \cdot |T|$ is the average value of the number of edges between the two sets of nodes in a random graph; also, note that $\lambda \sqrt{|S| \cdot |T|}$ is an "additive" scale factor, which might be very large, e.g., too large for the following lemma to give an interesting bound, in particular when one of S or T is very small. **Theorem 3** (Expander Mixing Lemma). Let G = (V, E) be a d-regular graph, with |V| = nand $\lambda = \max(|\mu_2|, |\mu_n|)$, where μ_i is the *i*-th largest eigenvalue of the (non-normalized) Adjacency Matrix. Then, for all $S, T \subseteq V$, we have the following:

$$\left| |E(S,T)| - \frac{d}{n} |S| \cdot |T| \right| \le \lambda \sqrt{|S| \cdot |T|}.$$

Proof. Define χ_S and χ_T to be the characteristic vectors of S and T. Then, if $\{v_j\}_{j=1}^n$ are orthonormal eigenvectors of A_G , with $v_1 = \frac{1}{\sqrt{n}}(1, \ldots, 1)$, then we can write the expansion of χ_S and χ_T in terms of those eigenvalues as: $\chi_S = \sum_i \alpha_i v_i$ and $\chi_T = \sum_j \beta_j v_j$. Thus,

$$|E(S,T)| = \chi_S^T A \chi_T$$

= $\left(\sum_i \alpha_i v_i\right) A\left(\sum_j \beta_j v_j\right)$
= $\left(\sum_i \alpha_i v_i\right) \left(\sum_j \mu_j \beta_j v_j\right)$
= $\sum_i \mu_i \alpha_i \beta_i$ since the v_i 's are orthonormal

Thus,

$$E(S,T)| = \sum \mu_i \alpha_i \beta_i$$

= $\mu_1 \alpha_1 \beta_1 + \sum_{i \ge 2} \mu_i \alpha_i \beta_i$
= $d \frac{|S| \cdot |T|}{n} + \sum_{i \ge 1} \mu_i \alpha_i \beta_i$,

where the last inequality is because, $\alpha_1 = \langle \chi_S, \frac{\overrightarrow{1}}{\sqrt{n}} \rangle = \frac{|S|}{\sqrt{n}}$ and (similarly) $\beta_1 = \frac{|T|}{\sqrt{n}}$, and $\mu_1 = d$. Hence,

$$\begin{aligned} \left| |E(S,T)| - \frac{d}{n} |S| \cdot |T| \right| &= \left| \sum_{i=2}^{n} \mu_{i} \alpha_{i} \beta_{i} \right| \\ &\leq \sum_{i \geq 2} |\mu_{i} \alpha_{i} \beta_{i}| \\ &\leq \lambda \sum_{i \geq 1} |\alpha_{i}| |\beta_{i}| \\ &\leq \lambda ||\alpha||_{2} ||\beta||_{2} = \lambda ||\chi_{S}||_{2} ||\chi_{T}||_{2} = \lambda \sqrt{|S| \cdot |T|} \end{aligned}$$

Other direction There is also a partial converse to this result:

Theorem 4 (Bilu and Linial). Let G be a d-regular graph, and suppose that

$$\left| E(S,T) - \frac{d}{n} |S| \cdot |T| \right| \le \rho \sqrt{|S| \cdot |T|}$$

holds \forall disjoint S,T and for some $\rho > 0$. Then

$$\lambda \le O\left(\rho\left(1 + \log(\frac{d}{\rho})\right)\right)$$

8.5.3 Some extra comments

We have been describing these results in terms of regular and unweighted graphs for simplicity, especially of analysis since the statements of the theorems don't change much under generalization. Important to note: these results can be generalized to weighted graphs with irregular number of edges per nodes using discrepancy. Informally, think of these characterizations as intuitively defining what the interesting properties of an expander are for real data, or what an expander is more generally, or what it means for a data set to look expander-like.

Although we won't worry too much about those issues, it is important to note that for certain, mostly algorithmic and theoretical applications, the fact that $d = \Theta(1)$, etc. are very important.

8.6 Expanders are graphs that are sparse versions/approximations of a complete graph

To quantify the idea that constant-degree expanders are sparse approximations to the complete graph, we need two steps:

- 1. first, a way to say that two graphs are close; and
- 2. second, a way to show that, with respect to that closeness measure, expanders and the complete graph are close.

8.6.1 A metric of closeness between two graphs

For the first step, we will view a graph as a Laplacian and vice versa, and we will consider the partial order over PSD matrices. In particular, recall that for a symmetric matrix A, we can write

$$A \succeq 0$$

to mean that

$$A \in PSD$$

(and, relatedly, $A \succ 0$ to mean that it is PD). In this case, we can write $A \succeq B$ to mean that $A-B \succeq 0$. Note that \succeq is a partial order. Unlike the real numbers, where every pair is comparable, for symmetric matrices, some are and some are not. But for pairs to which it does apply, it acts like a full order, in that, e.g.,

$$A \succeq B$$
 and $B \succeq C$ implies $A \succeq C$
 $A \succeq B$ implies that $A + C \succeq B + C$,

for symmetric matrices A, B, and C.

By viewing a graph as its Laplacian, we can use this to define an inequality over graphs. In particular, for graphs G and H, we can write

 $G \succeq H$ to mean that $L_G \succeq L_H$.

In particular, from our previous results, we know that if G = (V, E) is a graph and H = (V, F) is a subgraph of G, then $L_G \succeq L_H$. This follows since the Laplacian of a graph is the sum of the Laplacians of its edges: i.e., since $F \subseteq E$, we have

$$L_G = \sum_{einE} L_e = \sum_{e \in F} L_e + \sum_{einE \setminus F} L_e \preceq \sum_{e \in F} L_e = L_H,$$

which follows since $\sum_{e \in E \setminus F} L_e \succeq 0$.

That last expression uses the additive property of the order; now let's look at the multiplicative property that is also respected by that order.

If we have a graph G = (V, E) and a graph H = (V, E'), let's define the graph $c \cdot H$ to be the same as the graph H, except that every edge is multiplied by c. Then, we can prove relationships between graphs such as the following.

Lemma 2. If G and H are graphs s.t.

$$G \succeq c \cdot H$$

then, for all k we have that

$$\lambda_k(G) \ge c\lambda_k(H).$$

Proof: The proof is by the min-max Courant-Fischer variational characterization. We won't do it in detail. See DS, 09/10/12.

 \diamond

From this, we can prove more general relationships, e.g., bounds if edges are removed or rewieghted. In particular, the following two lemmas are almost corollaries of Lemma 2.

Lemma 3. If G is a graph and H is obtained by adding an edge to G or increasing the weight of an edge in G, then, for all i, we have that $\lambda_i(G) \leq \lambda_i(H)$.

Lemma 4. If $G = (V, E, W_1)$ is a graph and $H = (V, E, W_2)$ is a graph that differs from G only in its weights, then

$$G \succeq \min_{e \in E} \frac{w_1(e)}{w_2(e)} H.$$

Given the above discussion, we can use this to define the notion that two graphs approximate each other, basically by saying that they are close if their Laplacian quadratic forms are close. In particular, here is the definition.

Definition 8. Let G and H be graphs. We say that H is a c-approximation to H if

$$cH \succeq G \succeq \frac{1}{c}H.$$

As a special case, note that if $c = 1 + \epsilon$, for some $\epsilon \in (0, 1)$, then we have that the two graphs are very close.

8.6.2 Expanders and complete graphs are close in that metric

Given this notion of closeness between two graphs, we can now show that constant degree expanders are sparse approximations of the complete graph. The following theorem is one formalization of this idea. This establishes the closeness; and, since constant-degree expanders are very sparse, this result shows that they are sparse approximations of the complete graph. (We note in passing that it is know more generally that every graph can be approximated by a complete graph; this graph sparsification problem is of interest in many areas, and we might return to it.)

Theorem 5. For every $\epsilon > 0$, there exists a d > 0 such that for all sufficiently large n, there is a d regular graph G_n that is a $1 \pm \epsilon$ approximation of the complete graph K_n

Proof: Recall that a constant-degree expander is a *d*-regular graph whose Adjacency Matrix eigenvalues satisfy

$$|\alpha_i| \le \epsilon d,\tag{1}$$

 \diamond

for all $i \ge 2$, for some $\epsilon < 1$. We will show that graphs satisfying this condition also satisfy the condition of Def. 8 (with $c = 1 + \epsilon$) to be a good approximation of the complete graph.

To do so, recall that

$$(1-\epsilon) H \preceq G \preceq (1+\epsilon) H$$

means that

$$(1-\epsilon) x^T L_H x \le x^T L_G x \le (1+\epsilon) x^T L_H x$$

Let G be the Adjacency Matrix of the graph whose eigenvalues satisfy Eqn. (1). Given this, recall that the Laplacian eigenvalues satisfy $\lambda_i = d - \alpha_i$, and so all of the non-zero eigenvalues of L_G are in the interval between $(1 - \epsilon) d$ and $(1 + \epsilon) d$. I.e., for all x s.t. $x \perp \vec{1}$, we have that

$$(1-\epsilon) x^T x \le x^T L_G x \le (1+\epsilon) x^T x$$

(This follows from Courant-Fischer or by expanding x is an eigenvalue basis.)

On the other hand, for the complete graph K_n , we know that all vectors x that are $\perp \vec{1}$ satisfy

$$x^T L_{K_n} x = n x^T x.$$

So, let H be the graph

$$H = \frac{d}{n}K_n,$$

from which it follows that

$$x^T L_H x = dx^T x.$$

Thus, the graph G is an ϵ -approximation of the graph H, from which the theorem follows.

For completeness, consider G - H and let's look at its norm to see that it is small. First note that

$$(1-\epsilon) H \preceq G \preceq (1+\epsilon) H$$
 implies that $-\epsilon H \preceq G - H \preceq \epsilon H$

Since G and H are symmetric, and all of the eigenvalues of ϵH are either 0 or d, this tells us that

$$\|L_G - L_H\|_2 \le \epsilon d.$$

8.7 Expanders are graphs on which diffusions and random walks mix rapidly

We will have more to say about different types of diffusions and random walks later, so for now we will only work with one variant and establish one simple variant of the idea that random walks on expander graphs mix or equilibrate quickly to their equilibrium distribution.

Let G = (V, E, W) be a weighted graph, and we want to understand something about how random walks behave on G. One might expect that if, e.g., the graph was a dumbbell graph, then random walks that started in the one half would take a very long time to reach the other half; on the other hand, one might hope that if there are no such bottlenecks, e.g., bottlenecks revealed by the expansion of second eigenvalue, than random walks would mix relatively quickly.

To see this, let $p_t \in \mathbb{R}^n$, where *n* is the number of nodes in the graph, be a probability distribution at time *t*. This is just some probability distribution over the nodes, *e.g.*, it could be a discrete Dirac δ -function, *i.e.*, the indicator of a node, at time t = 0; it could be the uniform distribution; or it could be something else. Given this distribution at time *t*, the transition rule that governs the distribution at time t + 1 is:

• To go to p_{t+1} , move to a neighbor with probability ~ the weight of the edge. (In the case of unweighted graphs, this means that move to each neighbor with equal probability.) That is, to get to p_{t+1} from p_t , sum over neighbors

$$p_{t+1}(u) = \sum_{v:(u,v)\in E} \frac{W(u,v)}{d(v)} p_t(v)$$

where $d(v) = \sum_{u} W(u, v)$ is the weighted degree of v.

As a technical point, there are going to be bottlenecks, and so we will often consider a "lazy" random walk, which removed that trivial bottleneck that the graph is bipartite thus not mixing (i.e. the stationary distribution doesn't exist) and only increases the mixing time by a factor of two (intuitively, on expectation in two steps in the "lazy" walk we walk one step as in the simple random walk)—which doesn't matter in theory, since there we are interested in polynomial versus exponential times, and in practice the issues might be easy to diagnose or can be dealt with in less aggressive ways. Plus it's nicer in theory, since then things are SPSD.

By making a random walk "lazy," we mean the following: Let

$$p_{t+1}(u) = \frac{1}{2}p_t(u) + \frac{1}{2}\sum_{v:(u,v)\in E}\frac{W(u,v)}{d(v)}p_t(v).$$

That is, $p_{t+1} = \frac{1}{2} (I + AD^{-1}) p_t$, and so the transition matrix $W_G = A_G D_G^{-1}$ is replaced with $W_G = \frac{1}{2} (I + A_G D_G^{-1})$ —this is an asymmetric matrix that is similar in some sense to the normalized Laplacian.

Then, after t steps, we are basically considering W_G^t , in the sense that

$$p_0 \to p_t = W p_{t-1} = W^2 p_{t-2} = \dots = W^t p_t.$$

Fact. Regardless of the initial distribution, the lazy random walk converges to $\pi(i) = \frac{d(i)}{\sum_j d(j)}$, which is the right eigenvector of W with eigenvalue 1.

Fact. If $1 = \omega_1 \ge \omega_2 \ge \cdots \ge \omega_n \ge 0$ are eigenvalues of W, with $\pi(i) = \frac{d(i)}{\sum_j d(j)}$, then ω_2 governs the rate of convergence to the stationary distribution.

There are a number of ways to formalize this "rate of mixing" result, depending on the norm used and other things. In particular, a very good way is with the total variation distance, which is defined as:

$$\|p - q\|_{TVD} = \max_{S \subseteq V} \left\{ \sum_{v \in S} p_v - \sum_{v \in S} q_v \right\} = \frac{1}{2} \|p - q\|_1.$$

(There are other measures if you are interested in mixing rates of Markov chains.) But the basic point is that if $1 - \omega_2$ is large, *i.e.*, you are an expander, then a random walk converges fast. For example:

Theorem 6. Assume G = (V, E) with |V| = n is d-regular, A is the adjacency matrix of G, and $\hat{A} = \frac{1}{d}A$ is the transition matrix of a random walk on G, i.e., the normalized Adjacency Matrix. Also, assume $\lambda = \max(|\mu_2|, |\mu_n|) = \alpha d$ (recall μ_i is the *i*-th largest eigenvalue of A, not \hat{A}). Then

$$||\hat{A}^t p - u||_1 \le \sqrt{n\alpha^t},$$

where u is the stationary distribution of the random walk, which is the uniform distribution in the undirected d-regular graph, and p is an arbitrary initial distribution on V.

In particular, if $t \geq \frac{c}{1-\alpha} \log\left(\frac{n}{\epsilon}\right)$, for some absolute constant c independent of n, then $\|u - \hat{A}^t p\| \leq \epsilon$.

Proof. Let us define the matrix $\hat{J} = \frac{1}{n} \vec{1} \vec{1}^{\top}$, where, as before, $\vec{1}$ is the all ones vector of length n. Note that, for any probability vector p, we have

$$\hat{J}p = \frac{1}{n}\vec{1}\vec{1}^{\top}p$$
$$= \frac{1}{n}\vec{1}\cdot 1$$
$$= u.$$

Now, since $\hat{A} = \frac{1}{d}A$ we have $\hat{\mu}_i = \mu_i/d$, where $\hat{\mu}_i$ denotes the *i*th largest eigenvalue of \hat{A} , and the eigenvectors of \hat{A} are equal to those of A. Hence, we have

$$\begin{split} \left\| \hat{A}^t - \hat{J} \right\|_2 &= \max_{w: \|w\|_2 \le 1} \| (\hat{A}^t - \hat{J}) w \|_2 \\ &= \sigma_{\max} \left(\hat{A}^t - \hat{J} \right) \\ &= \sigma_{\max} \left(\sum_{i=1}^n \hat{\mu}_i^t v_i v_i^\top - \frac{1}{n} \vec{1} \vec{1}^\top \right) \\ &\stackrel{(a)}{=} \sigma_{\max} \left(\sum_{i=2}^n \hat{\mu}_i^t v_i v_i^\top \right) \\ &= \max\{ |\hat{\mu}_2^t|, |\hat{\mu}_n^t| \} \\ &= \alpha^t, \end{split}$$

where (a) follows since $v_1 = \frac{1}{\sqrt{n}}\vec{1}$ and $\hat{\mu}_1 = 1$. Then,

$$\begin{split} \left\| \hat{A}^t p - u \right\|_1 &\leq \sqrt{n} \left\| \hat{A}^t p - u \right\|_2 \\ &\leq \sqrt{n} \left\| \hat{A}^t p - \hat{J} p \right\|_2 \\ &\leq \sqrt{n} \left\| \hat{A}^t - \hat{J} \right\|_2 \left\| p \right\|_2 \\ &\leq \sqrt{n} \alpha^t, \end{split}$$

which concludes the proof.

This theorem shows that if the spectral gap is large (i.e. α is small), then we the walk mixes rapidly.

This is one example of a large body of work on rapidly mixing Markov chains. For example, there are extensions of this to degree-heterogeneous graphs and all sorts of other thigns Later, we might revisit this a little, when we see how tight this is; in particular, one issue that arises when we discuss local and locally-biased spectral methods is that how quickly a random walk mixes depends on not only the second eigenvalue but also on the size of the set achieving that minimum conductance value.