Lecture 6: Mid-Semester Review - Prediction in Random Graphs

Radu Balan

Department of Mathematics, AMSC, CSCAMM and NWC
University of Maryland, College Park, MD

March 9, 2017

Models and Data Sets

The overarching problem is the following:

Main Problem

Given a dynamical graph-based data set, discover if data can be explained as a structured data graph, or just as a random graph.

To do so, we need to understand: (1) how to generate dynamical graphs; (2) how to analyze these graphs.
Random graphs: two main classes, $\mathcal{G}_{n,p}$ and $\Gamma^{n,m}$.
Structured graphs: weighted graphs or percolation graphs $\rightarrow$ sequence of nested graphs.
What to look for:

1. complete subgraphs (cliques)
2. connectivity (number and size of connected components)
3. spectral gap and optimal partitions (Cheeger constant)

Sequence of Nested Graphs

We fix the number of vertices n. Sequence: $(G_m)_{0\leq m\leq M}$ of graphs $G_m=(\mathcal{V},\mathcal{E}_m)$, where each G_m has exactly n vertices, $|\mathcal{V}|=n$, and m edges, $|\mathcal{E}_m|=m$. Additionally we require $\mathcal{E}_m\subset\mathcal{E}_{m+1}$ (nestedness).

Examples: see movies

- Quasi-Regular percolation graph : PercGraph_n100N10d2_sig0.100000_lp2.000000.mp4
- Vertices are permuted randomly : PercGraph_scrambled_n100N10d2_sig0.100000_lp2.000000.mp4
- Edges are permuted randomly : PercGraph_random_n100N10d2_sig0.100000_lp2.000000.mp4

Random Graphs

The *Erdöd-Rényi class* $\mathcal{G}_{n,p}$ of random graphs: the number of vertices is fixed to n, and each edge is selected independently with probability p. The probability mass function, $P(G)$ for a graph G with n vertices and m edges is

$$P(G) = p^m (1-p)^{n(n-1)/2-m} \quad , \quad \frac{n(n-1)}{2} = \binom{n}{2}.$$

Random Graphs

The *Erdöd-Rényi class* $\mathcal{G}_{n,p}$ of random graphs: the number of vertices is fixed to n, and each edge is selected independently with probability p. The probability mass function, $P(G)$ for a graph G with n vertices and m edges is

$$P(G) = p^m (1-p)^{n(n-1)/2-m} \quad , \quad \frac{n(n-1)}{2} = \binom{n}{2}.$$

The class $\Gamma^{n,m}$ is the set of all graphs with n vertices and exactly m edges. In this class, the graph probability distribution is uniform:

$$P(G) = 1 / \binom{n(n-1)/2}{m}.$$

Distribution of Cliques
Expected Values

Let X_q denote the number of q-cliques in a random graph G. Then the expectation of X_q in $\mathcal{G}_{n,p}$ class is

$$\mathbb{E}[X_q] = \binom{n}{q} p^{q(q-1)/2}$$

The expectation of X_q in the class $\Gamma^{n,m}$ is approximated by the above formula for $p = \frac{2m}{n(n-1)}$:

$$\mathbb{E}[X_q] \approx \binom{n}{q} \left(\frac{2m}{n(n-1)}\right)^{q(q-1)/2} \sim \theta_q \frac{m^{q(q-1)/2}}{n^{q(q-2)}}$$

$$\mathbb{E}[X_3] \sim \theta \frac{m^3}{n^3} \quad , \quad \mathbb{E}[X_4] \sim \theta \frac{m^6}{n^8}$$

3-Cliques and 4-cliques

Thresholds

Theorem

Let $m = m(n)$ be the number of edges in $\Gamma^{n,m}$.

1. *If $m \gg n$ (i.e. $\lim_{n\to\infty} \frac{m}{n} = \infty$) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ 3-clique] \to 1.$
2. *If $m \ll n$ (i.e. $\lim_{n\to\infty} \frac{m}{n} = 0$) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ 3-clique] \to 0.$

Theorem

Let $m = m(n)$ be the number of edges in $\Gamma^{n,m}$.

1. *If $m \gg n^{4/3}$ (i.e. $\lim_{n\to\infty} \frac{m}{n^{4/3}} = \infty$) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ 4-clique] \to 1.$
2. *If $m \ll n^{4/3}$ (i.e. $\lim_{n\to\infty} \frac{m}{n^{4/3}} = 0$) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ 4-clique] \to 0.$

q-Cliques
Thresholds

Theorem

Let $p = p(n)$ be the edge probability in $\mathcal{G}_{n,p}$. Let $q \geq 3$ be and integer.

1. *If $p \gg \frac{1}{n^{2/(q-1)}}$ (i.e. $\lim_{n\to\infty} n^{2/(q-1)} p = \infty$) then $\lim_{n\to\infty} Prob[G \in \mathcal{G}_{n,p}$ has a $q - clique] \to 1$.*
2. *If $p \ll \frac{1}{n^{2/(q-1)}}$ (i.e. $\lim_{n\to\infty} n^{2/(q-1)} p = 0$) then $\lim_{n\to\infty} Prob[G \in \mathcal{G}_{n,p}$ has a $q - clique] \to 0$.*

q-Cliques
Thresholds

Theorem

Let $p = p(n)$ be the edge probability in $\mathcal{G}_{n,p}$. Let $q \geq 3$ be and integer.

1. *If* $p \gg \frac{1}{n^{2/(q-1)}}$ *(i.e.* $\lim_{n\to\infty} n^{2/(q-1)} p = \infty$*) then* $\lim_{n\to\infty} Prob[G \in \mathcal{G}_{n,p}\ has\ a\ q-clique] \to 1$.
2. *If* $p \ll \frac{1}{n^{2/(q-1)}}$ *(i.e.* $\lim_{n\to\infty} n^{2/(q-1)} p = 0$*) then* $\lim_{n\to\infty} Prob[G \in \mathcal{G}_{n,p}\ has\ a\ q-clique] \to 0$.

Theorem

Let $m = m(n)$ be the number of edges in $\Gamma^{n,m}$. Let $q \geq 3$ be and integer.

1. *If* $m \gg n^{2(q-2)/(q-1)}$ *(i.e.* $\lim_{n\to\infty} \frac{m}{n^{2(q-2)/(q-1)}} = \infty$*) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ q-clique] \to 1$.
2. *If* $m \ll n^{2(q-2)/(q-1)}$ *(i.e.* $\lim_{n\to\infty} \frac{m}{n^{2(q-1)/(q-1)}} = 0$*) then* $\lim_{n\to\infty} Prob[G \in \Gamma^{n,m}\ has\ a\ q-clique] \to 0$.

3-Cliques and 4-Cliques
Behavior at the threshold

In general we obtain a "coarse threshold". Recall a Poisson process X with parameter λ has p.m.f. $Prob[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$.

Theorem

In $\mathcal{G}_{n,p}$,

1. *For $p = \frac{c}{n}$, X_3 is asymptotically Poisson with parameter $\lambda = c^3/6$.*
2. *For $p = \frac{c}{n^{2/3}}$, X_4 is asymptotically Poisson with parameter $\lambda = c^6/24$.*

3-Cliques and 4-Cliques

Behavior at the threshold

In general we obtain a "coarse threshold". Recall a Poisson process X with parameter λ has p.m.f. $Prob[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$.

Theorem

In $\mathcal{G}_{n,p}$,

1. *For $p = \frac{c}{n}$, X_3 is asymptotically Poisson with parameter $\lambda = c^3/6$.*
2. *For $p = \frac{c}{n^{2/3}}$, X_4 is asymptotically Poisson with parameter $\lambda = c^6/24$.*

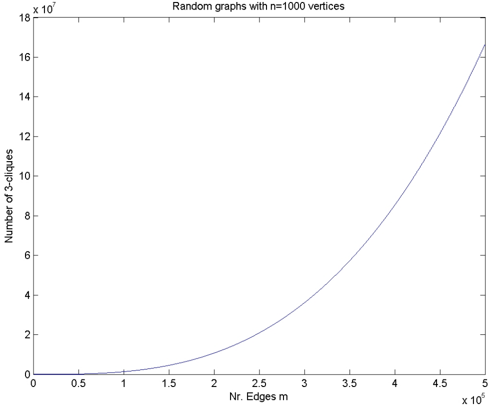
Theorem

In $\Gamma^{n,m}$,

1. *For $m = cn$, X_3 is asymptotically Poisson with parameter $\lambda = 4c^3/3$.*
2. *For $m = cn^{4/3}$, X_4 is asymptotically Poisson with parameter $\lambda = 8c^6/3$.*

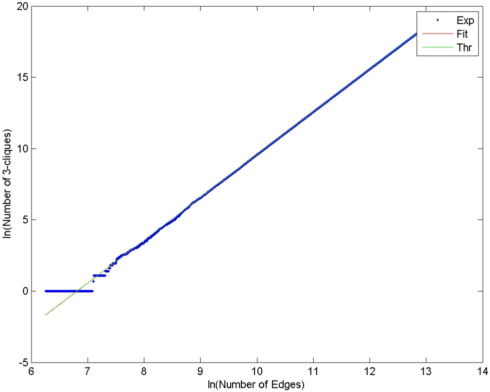
Numerical Results

3-cliques for random graph with $n = 1000$ vertices



Numerical Results

3-cliques for random graph with $n = 1000$ vertices



Connectivity

Strong threshold

Theorem

1. *Let* $m = m(n)$ *satisfies* $m \ll \frac{1}{2} n \log(n)$. *Then*

$$\lim_{n\to\infty} Prob[G \in \Gamma^{n,m} \text{ is connected}] = 0$$

2. *Let* $m = m(n)$ *satisfies* $m \gg \frac{1}{2} n \log(n)$. *Then*

$$\lim_{n\to\infty} Prob[G \in \Gamma^{n,m} \text{ is connected}] = 1$$

Connectivity
Strong threshold

Theorem

1. *Let* $m = m(n)$ *satisfies* $m \ll \frac{1}{2} n \log(n)$. *Then*
$$\lim_{n \to \infty} Prob[G \in \Gamma^{n,m} \text{ is connected}] = 0$$
2. *Let* $m = m(n)$ *satisfies* $m \gg \frac{1}{2} n \log(n)$. *Then*
$$\lim_{n \to \infty} Prob[G \in \Gamma^{n,m} \text{ is connected}] = 1$$
3. *Assume* $m = \frac{1}{2} n \log(n) + tn + o(n)$, *where* $o(n) \ll n$. *Then*
$$\lim_{n \to \infty} Prob[G \in \Gamma^{n,m} \text{ is connected}] = e^{-e^{-2t}}$$

Connectivity
Strong threshold

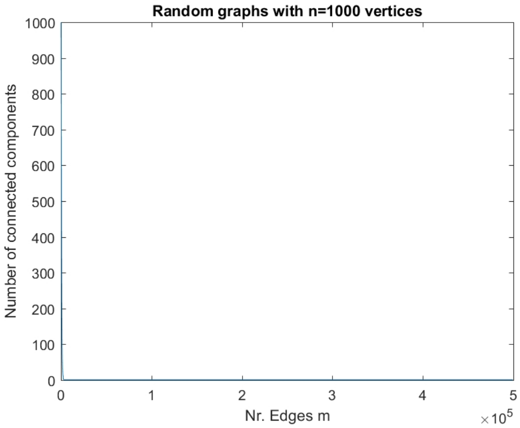
Theorem

1. *Let* $m = m(n)$ *satisfies* $m \ll \frac{1}{2}n\log(n)$. *Then*
$$\lim_{n\to\infty} Prob[G \in \Gamma^{n,m} \textit{ is connected}] = 0$$
2. *Let* $m = m(n)$ *satisfies* $m \gg \frac{1}{2}n\log(n)$. *Then*
$$\lim_{n\to\infty} Prob[G \in \Gamma^{n,m} \textit{ is connected}] = 1$$
3. *Assume* $m = \frac{1}{2}n\log(n) + tn + o(n)$, *where* $o(n) \ll n$. *Then*
$$\lim_{n\to\infty} Prob[G \in \Gamma^{n,m} \textit{ is connected}] = e^{-e^{-2t}}$$

In this case $\frac{1}{2}n\log(n)$ is known as a *strong threshold*.

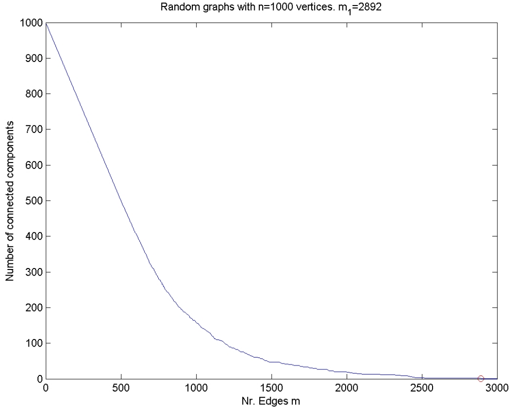
Numerical Results

Connectivity for random graph with $n = 1000$ vertices



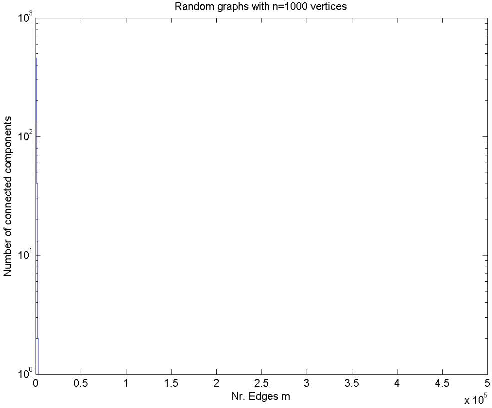
Numerical Results

Connectivity for random graph with $n = 1000$ vertices



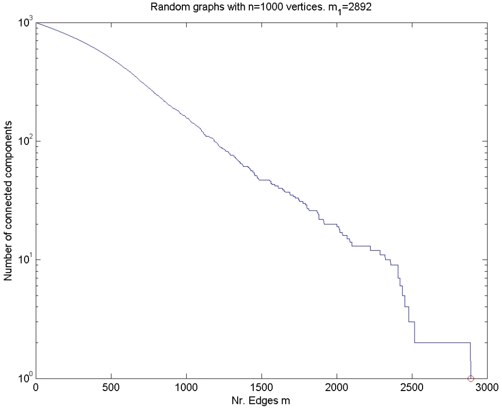
Numerical Results

Connectivity for random graph with $n = 1000$ vertices



Numerical Results

Connectivity for random graph with $n = 1000$ vertices



Graph Laplacians

$\Delta, L, \tilde{\Delta}$

Recall the Laplacian matrices:

$$\Delta = D - A \quad , \quad \Delta_{ij} = \begin{cases} d_i & \text{if} \quad i = j \\ -1 & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad \text{otherwise} \end{cases}$$

$$L = D^{-1}\Delta \quad , \quad L_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \text{ and } d_i > 0 \\ -\frac{1}{d(i)} & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad \text{otherwise} \end{cases}$$

$$\tilde{\Delta} = D^{-1/2}\Delta D^{-1/2} \quad , \quad \tilde{\Delta}_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \text{ and } d_i > 0 \\ -\frac{1}{\sqrt{d(i)d(j)}} & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad \text{otherwise} \end{cases}$$

Graph Laplacians
$\Delta, L, \tilde{\Delta}$

Recall the Laplacian matrices:

$$\Delta = D - A \quad , \quad \Delta_{ij} = \begin{cases} d_i & \text{if} \quad i = j \\ -1 & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad otherwise \end{cases}$$

$$L = D^{-1}\Delta \quad , \quad L_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \text{ and } d_i > 0 \\ -\frac{1}{d(i)} & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad otherwise \end{cases}$$

$$\tilde{\Delta} = D^{-1/2}\Delta D^{-1/2} \quad , \quad \tilde{\Delta}_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \text{ and } d_i > 0 \\ -\frac{1}{\sqrt{d(i)d(j)}} & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \quad otherwise \end{cases}$$

Remark: $D^{-1}, D^{-1/2}$ are the pseudoinverses.

Eigenvalues of Laplacians

$\Delta, L, \tilde{\Delta}$

What do we know about the set of eigenvalues of these matrices for a graph G with n vertices?

1. $\Delta = \Delta^T \geq 0$ and hence its eigenvalues are non-negative real numbers.
2. $eigs(\tilde{\Delta}) = eigs(L) \subset [0, 2]$.
3. 0 is always an eigenvalue and its multiplicity equals the number of connected components of G,

$$dim\, ker(\Delta) = dim\, ker(L) = dim\, ker(\tilde{\Delta}) = \#connected\ \ components.$$

Eigenvalues of Laplacians

$\Delta, L, \tilde{\Delta}$

What do we know about the set of eigenvalues of these matrices for a graph G with n vertices?

1. $\Delta = \Delta^T \geq 0$ and hence its eigenvalues are non-negative real numbers.
2. $eigs(\tilde{\Delta}) = eigs(L) \subset [0, 2]$.
3. 0 is always an eigenvalue and its multiplicity equals the number of connected components of G,

$$dim\,ker(\Delta) = dim\,ker(L) = dim\,ker(\tilde{\Delta}) = \#connected\ \ components.$$

Let $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$ be the eigenvalues of $\tilde{\Delta}$. Denote

$$\lambda(G) = \max_{1 \leq i \leq n-1} |1 - \lambda_i|.$$

Note $\sum_{i=1}^{n-1} \lambda_i = trace(\tilde{\Delta}) = n$. Hence the average eigenvalue is about 1. $\lambda(G)$ is called *the absolute gap* and measures the spread of eigenvalues away from 1.

The spectral absolute gap
$\lambda(G)$

The main result in [8]) says that for connected graphs w/h.p.:

$$\lambda_1 \geq 1 - \frac{C}{\sqrt{\text{Average Degree}}} = 1 - \frac{C}{\sqrt{p(n-1)}} = 1 - C\sqrt{\frac{n}{2m}}.$$

Theorem (For class $\mathcal{G}_{n,p}$)

Fix $\delta > 0$ and let $p > (\frac{1}{2} + \delta)log(n)/n$. Let $d = p(n-1)$ denote the expected degree of a vertex. Let $\tilde{G}$ be the giant component of the Erdös-Rényi graph. For every fixed $\varepsilon > 0$, there is a constant $C = C(\delta, \varepsilon)$, so that

$$\lambda(\tilde{G}) \leq \frac{C}{\sqrt{d}}$$

with probability at least $1 - Cn\exp(-(2-\varepsilon)d) - C\exp(-d^{1/4}log(n))$.

Connectivity threshold: $p \sim \frac{log(n)}{n}$.

The spectral absolute gap
$\lambda(G)$

The main result in [8] says that for connected graphs w/h.p.:

$$\lambda_1 \geq 1 - \frac{C}{\sqrt{\text{Average Degree}}} = 1 - \frac{C}{\sqrt{p(n-1)}} = 1 - C\sqrt{\frac{n}{2m}}.$$

Theorem (For class $\Gamma^{n,m}$)

Fix $\delta > 0$ and let $m > \frac{1}{2}(\frac{1}{2}+\delta)n\,log(n)$. Let $d = \frac{2m}{n}$ denote the expected degree of a vertex. Let $\tilde{G}$ be the giant component of the Erdös-Rényi graph. For every fixed $\varepsilon > 0$, there is a constant $C = C(\delta, \varepsilon)$, so that

$$\lambda(\tilde{G}) \leq \frac{C}{\sqrt{d}}$$

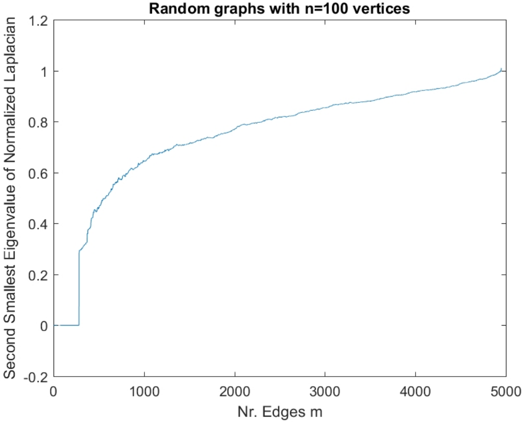
with probability at least $1 - Cn\exp(-(2-\varepsilon)d) - C\exp(-d^{1/4}log(n))$.

Connectivity threshold: $m \sim \frac{1}{2}n\,log(n)$.

Numerical Results

λ_1 for random graphs

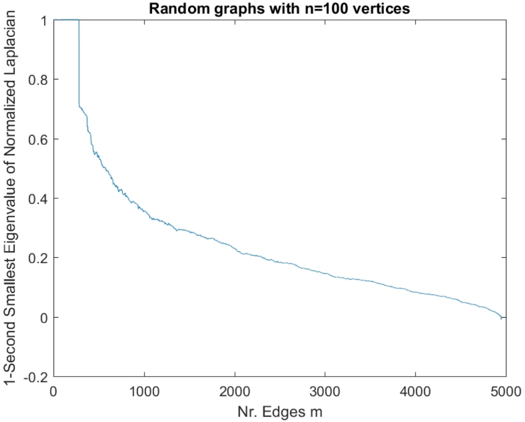
Results for $n = 100$ vertices: $\lambda_1(\tilde{G}) \approx 1 - \frac{C}{\sqrt{m}}$.



Numerical Results

$1-\lambda_1$ for random graphs

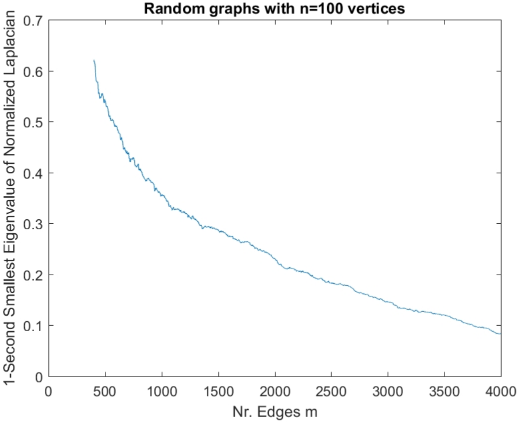
Results for $n = 100$ vertices: $\lambda_1(\tilde{G}) \approx 1 - \frac{C}{\sqrt{m}}$.



Numerical Results

$1 - \lambda_1$ for random graphs

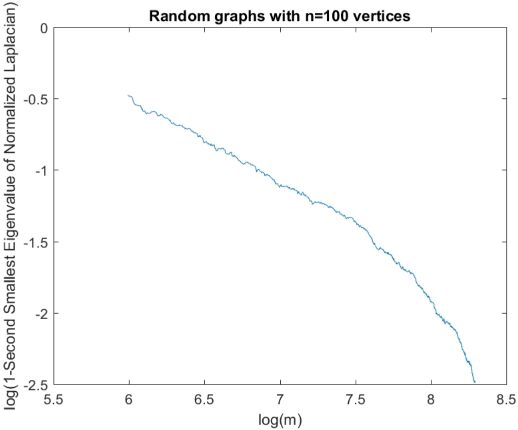
Results for $n = 100$ vertices: $\lambda_1(\tilde{G}) \approx 1 - \frac{C}{\sqrt{m}}$.



Numerical Results

$log(1-\lambda_1)$ vs. $log(m)$ for random graphs

Results for $n = 100$ vertices: $\lambda_1(\tilde{G}) \approx 1 - \frac{C}{\sqrt{m}}$.



The Cheeger constant

h_G and Optimal Partitions

Fix a graph $G = (\mathcal{V}, \mathcal{E})$ with n vertices and m edges. The *Cheeger constant* h_G is the optimum value of

$$h_G = \min_{S \subset \mathcal{V}} \frac{|E(S, \bar{S})|}{min(vol(S), vol(\bar{S}))}$$

where

1. For two disjoint sets of vertices A abd B, $E(A, B)$ denotes the set of edges that connect vertices in A with vertices in B:
$$E(A, B) = \{(x, y) \in \mathcal{E} \quad , \quad x \in A \ , \ y \in B\}.$$
2. The *volume* of a set of vertices is the sum of its degrees:
$$vol(A) = \sum_{x \in A} d_x.$$
3. For a set of vertices A, denote $\bar{A} = \mathcal{V} \setminus A$ its complement.

The Cheeger inequalities

h_G and λ_1

Theorem

For a connected graph

$$2h_G \geq \lambda_1 > 1 - \sqrt{1 - h_G^2} > \frac{h_G^2}{2}.$$

Equivalently:

$$\sqrt{2\lambda_1} > \sqrt{1 - (1 - \lambda_1)^2} > h_G \geq \frac{\lambda_1}{2}.$$

Proof of upper bound reveals a "good" initial guess of the optimal partition:

1. Compute eigenpair (λ_1, g^1) for the second smallest eigenvalue;
2. Form the partition:

$$S = \{k \in \mathcal{V} \ , \ g_k^1 \geq 0\} \ , \ \bar{S} = \{k \in \mathcal{V} \ , \ g_k^1 < 0\}$$

Min-cut Problems
Weighted Graphs

The Cheeger inequality holds true for weighted graphs, $G = (\mathcal{V}, \mathcal{E}, W)$.

- $\Delta = D - W$, $D = diag(w_i)_{1 \leq i \leq n}$, $w_i = \sum_{j \neq i} w_{i,j}$
- $\tilde{\Delta} = D^{-1/2} \Delta D^{-1/2} = I - D^{-1/2} W D^{-1/2}$
- $eigs(\tilde{\Delta}) \subset [0, 2]$
- $h_G = \min_S \frac{\sum_{x \in S, y \in \bar{S}} W_{x,y}}{min(\sum_{x \in S} W_{x,x}, \sum_{y \in \bar{S}} W_{y,y})}$
- $2h_G \geq \lambda_1 \geq 1 - \sqrt{1 - h_G^2}$
- Good initial guess for optimal partition: Compute the eigenpair (λ_1, g^1) associated to the second smallest eigenvalue of $\tilde{\Delta}$; set:

$$S = \{k \in \mathcal{V} \ \ , \ \ g_k^1 \geq 0\} \ \ , \ \ \bar{S} = \{k \in \mathcal{V} \ \ , \ \ g_k^1 < 0\}$$

References

- B. Bollobás, **Graph Theory. An Introductory Course**, Springer-Verlag 1979. **99**(25), 15879–15882 (2002).
- F. Chung, **Spectral Graph Theory**, AMS 1997.
- F. Chung, L. Lu, The average distances in random graphs with given expected degrees, Proc. Nat.Acad.Sci. 2002.
- F. Chung, L. Lu, V. Vu, The spectra of random graphs with Given Expected Degrees, Internet Math. **1**(3), 257–275 (2004).
- R. Diestel, **Graph Theory**, 3rd Edition, Springer-Verlag 2005.
- P. Erdös, A. Rényi, On The Evolution of Random Graphs
- G. Grimmett, **Probability on Graphs. Random Processes on Graphs and Lattices**, Cambridge Press 2010.

C. Hoffman, M. Kahle, E. Paquette, Spectral Gap of Random Graphs and Applications to Random Topology, arXiv: 1201.0425 [math.CO] 17 Sept. 2014.

J. Leskovec, J. Kleinberg, C. Faloutsos, Graph Evolution: Densification and Shrinking Diameters, ACM Trans. on Knowl.Disc.Data, **1**(1) 2007.