Complex Networks and Statistical Learning
Homework 5

Chenghua Liu
liuch18@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

1 Community Detection

1.1 "Networks", Exercise 14.1

Consider a "line graph" consisting of n nodes in a row :

a) Show that if we divide the network into two parts by cutting any single edge, such that one part has r nodes and the other has $n-r$, the modularity, Eq. (7.58), takes the value

$$Q = \frac{3 - 4n + 4rn - 4r^2}{2(n-1)^2}$$

b) Hence show that when n is even the optimal such division, in terms of modularity, is the division that splits the network exactly down the middle.

Solution:

(a) According to the definition $Q = \frac{1}{2m}\sum b_{ij} = \frac{1}{2m}\sum\left(a_{ij} - \frac{k_i k_j}{2m}\right)\delta_{g_i,g_j}$, we have

$$B = \begin{pmatrix} -\frac{1}{2m} & 1-\frac{2}{2m} & -\frac{2}{2m} & -\frac{2}{2m} & \cdots & -\frac{1}{2m} \\ 1-\frac{2}{2m} & -\frac{4}{2m} & 1-\frac{4}{2m} & -\frac{4}{2m} & \cdots & -\frac{2}{2m} \\ -\frac{2}{2m} & 1-\frac{4}{2m} & -\frac{4}{2m} & 1-\frac{4}{2m} & \cdots & -\frac{2}{2m} \\ -\frac{2}{2m} & -\frac{4}{2m} & 1-\frac{4}{2m} & -\frac{4}{2m} & \cdots & -\frac{2}{2m} \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots \\ -\frac{2}{2m} & -\frac{4}{2m} & -\frac{4}{2m} & -\frac{4}{2m} & \cdots & -\frac{2}{2m} \\ -\frac{1}{2m} & -\frac{2}{2m} & -\frac{2}{2m} & -\frac{2}{2m} & \cdots & -\frac{1}{2m} \end{pmatrix}$$

Supposed that the first r nodes are in a community, and the last $n-r$ in the other community, then we can split the matrix B into 4 parts and only compute k

$$B = \begin{pmatrix} B_{1:r,1:r} & 0 \\ 0 & B_{(r+1):n,(r+1):n} \end{pmatrix}$$

Taking the sum of the matrix B, we have the following terms

- two ends in the main diagonal: $2\left(-\frac{1}{2m}\right)$
- number of positive ones in total: $2(n-2)$
- number of $-\frac{2}{2m}$ in the first row: $2(r-1)\left(-\frac{2}{2m}\right)$
- number of $-\frac{4}{2m}$ in the first r rows: $(r-1)^2\left(-\frac{4}{2m}\right)$
- number of $-\frac{2}{2m}$ in the last column: $2(n-r-1)\left(-\frac{2}{2m}\right)$
- number of $-\frac{4}{2m}$ in the last $n-r$ rows: $(n-r-1)^2\left(-\frac{4}{2m}\right)$

Substitute $m = n-1$, add up all the above six items and simplify the equation, we have

$$\begin{aligned} Q =& \frac{\left(\frac{-1}{n-1} + 2\cdot n - 4 - \frac{2\cdot(r-1)\cdot 2}{2\cdot(n-1)} - \frac{(r-1)^2\cdot 4}{2\cdot(n-1)} - \frac{2\cdot(n-r-1)\cdot 2}{2\cdot(n-1)} - \frac{(n-r-1)^2\cdot 4}{2\cdot(n-1)}\right)}{2\cdot(n-1)} \\ =& \frac{(4r-4)n - 4r^2 + 3}{2(n-1)^2} \end{aligned}$$

(b)

We take the derivative of Q with r, we have

$$Q'(r) = \frac{4n-8r}{2(n-1)^2} \quad Q'(r) = 0 \Rightarrow r = \frac{n}{2}$$

And note that $Q''(r) < 0$, which imples it is concave .So $r = \frac{n}{2}$ maxmize the $Q(r)$. Therefore, the modularity reaches tha maximum value is the division splits the network exactly down the middle.

2 PageRank

Consider a random walk formulation of topic- specific PageRank:

$$x_t = (1-\gamma)AD^{-1}x_t + \frac{\gamma}{n}t$$

where t is a probability distribution over topic- specific nodes, let x_t be the topic-specific PageRank vector, and x_v be the personalized PageRank vector of node v. Derive a formula of x_t using x_v and t as input (your answer should not include A and D).

Solution:

According to the equation, we have

$$x_t = \left(I - (1-\gamma)AD^{-1}\right)^{-1}\frac{\gamma}{n}t$$

Note that $t = \sum_{v=1}^{n} t_v e_v$. Therefore,

$$\begin{aligned} x_t &= \left(I - (1-\gamma)AD^{-1}\right)^{-1} \frac{\gamma}{n} \sum_{v=1}^{n} t_v e_v \\ &= \sum_{v=1}^{n} t_v \left(I - (1-\gamma)AD^{-1}\right)^{-1} \frac{\gamma}{n} e_v \\ &= \sum_{v=1}^{n} t_v x_v \end{aligned}$$

3 Percolation

3.1 "Networks", Exercise 15.1

Consider a site percolation process in which nodes are removed uniformly at random from a random 4 -regular network (i.e., a configuration model where all nodes have degree 4). You can assume the network is large.

a) Give an expression for the size S of the giant percolation cluster as a fraction of total network size.

b) Find the critical occupation probability ϕ_c.

c) Find the value of ϕ at which $S = 1$. This implies that the giant cluster fills the whole network. How can this happen, given that the most it can fill is the whole of the giant component?

Solution:

(a) In a random 4-regular network, the degree distribution and the excess degree distribution are respectively

$$p_k = 1[k=4], \quad q_k = \frac{(k+1)p_{k+1}}{\langle k \rangle} = 1[k=3]$$

Recall the equation:

$$u = 1 - \phi + \phi \sum_{k=0}^{+\infty} q_k u^k = 1 - \phi + \phi u^3$$

If $\phi \leq \phi_{\mathrm{c}}$, the only solution is $u = 1$. Therefore,

$$S = \phi \left(1 - \sum_{k=0}^{+\infty} p_k u^k\right) = \phi \left(1 - u^4\right) = 0$$

If $\phi > \phi_{\mathrm{c}}$, we have

$$\phi = \frac{1-u}{1-u^3} = \frac{1}{1+u+u^2}, \quad u = \frac{\sqrt{4\phi^{-1} - 3} - 1}{2}$$

Therefore,

$$
\begin{aligned}
S &= \phi \left(1 - \sum_{k=0}^{+\infty} p_k u^k \right) = \phi \left(1 - u^4 \right) \\
&= \phi \left[1 - \left(\frac{\sqrt{4\phi^{-1} - 3} - 1}{2} \right)^4 \right]
\end{aligned}
$$

(b)

In a random 4-regular network, $\langle k \rangle = 4 \Rightarrow \langle k^2 \rangle = 4^2$. So the critical occupation probability is

$$
\phi_{\mathrm{c}} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} = \frac{4}{4^2 - 4} = \frac{1}{3}
$$

(c)

When $S = 1$, we have the solution that $\phi = 1$. This can happen if and only if the network is connected.

4 Contagion Process

4.1 "Networks", Exercise 16.1

Consider the bond percolation model of an epidemic in Section 16.3.1 and suppose that for a particular value of ϕ the giant cluster occupies a fraction S of the network. What is the probability of an epidemic outbreak if the disease starts simultaneously at c different nodes, chosen uniformly and independently at random from the whole network? Note that this probability tends exponentially to 1 as c gets larger. The chances of avoiding an epidemic become slim when a disease starts at many points simultaneously.

Solution:

Note that an epidemic outbreak happens if at least one of the c chosen nodes lies in the giant cluster. Therefore, the probability of an epidemic outbreak is $1-(1-S)^c$. This probability tends exponentially to 1 as c gets larger.

5 Spectral Graph Theory

Write down a formal proof of Cheeger's Inequality by filling out the missing parts in our proof sketch.

Solution:

First of all, let's review some notations and state the Cheeger's Inequality again.

The normalized adjacency matrix is

$$
\mathscr{A} \triangleq D^{-1/2} A D^{-1/2}
$$

where A is the adjacency matrix of G and $D = \mathrm{diag}\{d(i)\}$ is the degree matrix. For a graph G (with no isolated vertices)

$$D^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{d(1)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{d(2)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{d(n)}} \end{pmatrix}$$

and $d(i)$ is the degree of vertex i.Then, we define The normalized Laplacian matrix is

$$\begin{aligned} \mathscr{L} &\triangleq I - \mathscr{A} \\ &= D^{-1/2}(D-A)D^{-1/2} \\ &= D^{-1/2}L_G D^{-1/2} \end{aligned}$$

Where L_G is the (unnormalized) Laplacian.When $S \subseteq V$, we define $\delta(S) \triangleq \{(u,v) \in E : u \in S, v \notin S\}$ as the set of edges with exactly one endpoint in S, and $\mathrm{vol}(S) = \sum_{i \in S} d(i)$. The conductance of S is defined as

$$\phi(S) = \frac{|\delta(S)|}{\min(\mathrm{vol}(S), \mathrm{vol}(V-S))}$$

and the conductance of G is defined as $\phi(G) = \min_{S \subseteq V} \phi(S)$. Let $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ denote the eigenvalues of $\mathscr{L}$.

Denote x_2 to be the eigenvector associated with λ_2. Its Raleigh quotient $R(x_2) = \frac{x_2^T \mathscr{L} x_2}{x_2^T x_2}$ is simply λ_2.

The Cheeger's Inequality is say that

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

We'll prove the first inequality, and start the proof of the second inequality.

* * *

Recall that

$$\lambda_2 = \min_{x:<x, D^{1/2}e>=0} \frac{x^T \mathscr{L} x}{x^T x} = \min_{x:<x, D^{1/2}e>=0} \frac{x^T D^{-1/2} L_G D^{-1/2} x}{x^T x}$$

Consider the change of variables obtained by setting $y = D^{-1/2}x$ and $x = D^{1/2}y$:

$$\lambda_2 = \min_{y:<D^{1/2}y, D^{1/2}e>=0} \frac{y^T L_G y}{\left(D^{1/2}y\right)^T \left(D^{1/2}y\right)} = \min_{y:<D^{1/2}y, D^{1/2}e>=0} \frac{y^T L_G y}{y^T D y}$$

The minimum is being taken over all y such that $< D^{1/2}y, D^{1/2}e >= 0$. That is, over y such that:

$$\left(D^{1/2}y\right)^T D^{1/2}e = 0 \Longleftrightarrow y^T D e = 0 \Longleftrightarrow \sum_{i \in V} d(i)y(i) = 0$$

Hence, we have that

$$\lambda_2 = \min_{y:\sum_{i\in V} d(i)y(i)=0} \frac{\sum_{(i,j)\in E}(y(i)-y(j))^2}{\sum_{i\in V} d(i)y(i)^2}$$

Now let S^* be such that $\phi(G)=\phi(S^*)$, and try defining

$$\hat{y}(i) = \begin{cases} 1, & i\in S^* \\ 0, & \text{else} \end{cases}$$

It would be great if λ_2 was bounded by $\frac{|\delta(S^*)|}{\sum_{i\in S^*} d(i)} = \frac{|\delta(S^*)|}{\mathrm{vol}(S^*)}$. However, there are two problems. We have $\sum_{i\in V} d(i)\hat{y}(i) \neq 0$; moreover $\frac{|\delta(S^*)|}{\mathrm{vol}(S^*)}$ might not be $\phi(S^*)$, as we want the denominator to be $\min\{\mathrm{vol}(S^*), \mathrm{vol}(V-S^*)\}$. Hence, we redefine

$$\hat{y}(i) = \begin{cases} \frac{1}{\mathrm{vol}(S^*)}, & i\in S^* \\ -\frac{1}{\mathrm{vol}(V-S^*)}, & \text{else.} \end{cases}$$

Now we notice that:

$$\sum_{i\in V} d(i)\hat{y}(i) = \frac{\sum_{i\in S^*} d(i)}{\mathrm{vol}(S^*)} - \frac{\sum_{i\notin S^*} d(i)}{\mathrm{vol}(V-S^*)} = 1-1=0$$

Thus, this is a feasible solution to the minimization problem defining λ_2, and we have that the only edges contributing anything nonzero to the numerator are those with exactly one endpoint in S^*. Thus:

$$\begin{aligned}
\lambda_2 &\le \frac{|\delta(S^*)|\left(\frac{1}{\mathrm{vol}(S^*)}+\frac{1}{\mathrm{vol}(V-S^*)}\right)^2}{\sum_{i\in S^*} d(i)\left(\frac{1}{\mathrm{vol}(S^*)}\right)^2 + \sum_{i\notin S^*} d(i)\left(\frac{1}{\mathrm{vol}(V-S^*)}\right)^2} \\
&= \frac{|\delta(S^*)|\left(\frac{1}{\mathrm{vol}(S^*)}+\frac{1}{\mathrm{vol}(V-S^*)}\right)^2}{\frac{1}{\mathrm{vol}(S^*)}+\frac{1}{\mathrm{vol}(V-S^*)}} \\
&= |\delta(S^*)|\left(\frac{1}{\mathrm{vol}(S^*)}+\frac{1}{\mathrm{vol}(V-S^*)}\right) \\
&\le 2|\delta(S^*)|\max\left\{\frac{1}{\mathrm{vol}(S^*)}, \frac{1}{\mathrm{vol}(V-S^*)}\right\} \\
&= \frac{2|\delta(S^*)|}{\min\{\mathrm{vol}(S^*), \mathrm{vol}(V-S^*)\}} \\
&= 2\phi(G)
\end{aligned}$$

This completes the proof of the first inequality. To get the second, the idea is to suppose we had a y with

$$R(y) \equiv \frac{\sum_{(i,j)\in E}(y(i)-y(j))^2}{\sum_{i\in V} d(i)y(i)^2}$$

Claim 5.0.1. *We'll be able to find a cut $S\subset \mathrm{supp}(Y) \triangleq \{i\in V : y(i)\neq 0\}$ with $\frac{\delta(S)}{\mathrm{vol}(S)} \le \sqrt{2R(y)}$*

Proof: Without loss of generality, we assume $-1 \le y(i) \le 1$, as we can scale y if not. Our trick (from Trevisan) is to pick $t \in (0,1]$ uniformly at random, and let $S_t = \{i \in V : y(i)^2 \ge t\}$. Notice that:

$$\mathbb{E}\left[\mathrm{vol}\left(S_t\right)\right] = \sum_{i \in V} d(i) \Pr\left[i \in S_t\right] = \sum_{i \in V} d(i) y(i)^2$$

and assuming that $(i,j) \in E \implies y(i)^2 \le y(j)^2$,

$$\mathbb{E}\left[\left|\delta\left(S_t\right)\right|\right] = \sum_{(i,j)\in E} \Pr\left[(i,j) \in \delta\left(S_t\right)\right] = \sum_{(i,j)\in E} \Pr\left[y(i)^2 < t \le y(j)^2\right] = \sum_{(i,j)\in E} \left(y(j)^2 - y(i)^2\right)$$

Rewriting the above using difference of squares and using Cauchy-Schwarz,

$$\begin{aligned}
\sum_{(i,h)\in E} (y(j)-y(i))(y(j)+y(i)) &\le \sqrt{\sum_{(i,j)\in E} (y(j)-y(i))^2 \sum_{(i,j)\in E} (y(j)+y(i))^2} \\
&\le \sqrt{\sum_{(i,j)\in E} (y(j)-y(i))^2} \sqrt{2 \sum_{(i,j)\in E} (y(j)^2+y(i)^2)} \\
&= \sqrt{\sum_{(i,j)\in E} (y(j)-y(i))^2} \sqrt{\sum_{i\in V} 2d(i)y(i)^2} \\
&= \sqrt{2R(y)} \sqrt{\sum_{i\in V} d(i)y(i)^2}
\end{aligned}$$

This gives that

$$\frac{\mathbb{E}\left[\left|\delta\left(S_t\right)\right|\right]}{\mathbb{E}\left[\mathrm{vol}\left(S_t\right)\right]} \le \sqrt{2R(y)} \implies \mathbb{E}\left[\left|\delta\left(S_t\right)\right| - \sqrt{2R(y)}\,\mathrm{vol}\left(S_t\right)\right] \le 0$$

This means that there exists a t such that

$$\frac{\left|\delta\left(S_t\right)\right|}{\mathrm{vol}\left(S_t\right)} \le \sqrt{2R(y)}$$

$\square$

We have proved that, for any vector $y \in \mathbb{R}^n$ with $\sum_{i\in V} d(i)y(i) = 0$, we can find $S_t \subseteq \mathrm{supp}(y) = \{i \in V : y(i) \ne 0\}$ such that $\frac{|\delta(S_t)|}{\mathrm{vol}(S_t)} \le \sqrt{2R(y)}$. We also saw that $\lambda_2 = \min R(y)$. The issue is that we may have $\mathrm{vol}\left(S_t\right) > \mathrm{vol}\left(V - S_t\right)$. To fix this, we will modify y so that $\mathrm{vol}(\mathrm{supp}(y)) \le m($ recall that $\mathrm{vol}(V) = 2m)$. The idea is to pick c such that the two sets $\{i : y(i) < c\}$ and $\{i : y(i) > c\}$ both have volume at most m, then find S_t for both of them and take the best one.

Claim 5.0.2. *Let $z = y - ce$, where $e \in \mathbb{R}^n$ is the vector of all ones. Then*

(i) $z^T D z \ge y^T D y$.

(ii) $z^T L_G z = y^T L_G y$.

(iii) *Let* $z_+(i) = \max(0, z(i))$ *and* $z_-(i) = \min(0, z(i))$. *Then* $\min\left(R\left(z_+\right), R\left(z_-\right)\right) \leq R(z) \leq R(y)$ *and* $\operatorname{supp}\left(z_+\right), \operatorname{supp}\left(z_-\right)$ *both have volume at most* m.

proof:

(i) Let $f(c) = (y - ce)^T D(y - ce) = \sum_{i \in V} d(i)(y(i) - c)^2$. We have $f'(c) = \sum_{i \in V}(-2y(i)d(i) + 2cd(i)) = 2c \sum_{i \in V} d(i)$, by $\sum_i y(i)d(i) = 0$ Also, $f''(c) = 2\sum_i d(i) > 0$, so that f is minimized when $f'(c) = 0 \Longleftrightarrow c = 0$, so that $z^T Dz \geq y^T Dy$, as desired.

(ii) Indeed,

$$\begin{aligned} z^T L_G z = \sum_{(i,j) \in E} (z(i) - z(j))^2 &= \sum_{(i,j) \in E} ((y(i) - c) - (y(j) - c))^2 \\ &= \sum_{(i,j) \in E} (y(i) - y(j))^2 = y^T L_G y \end{aligned}$$

(iii) Note that

$$z^T Dz = \sum_{i \in V} d(i) z(i)^2 = \sum_{i \in V} d(i) z_+(i)^2 + \sum_{i \in V} d(i) z_-(i)^2 = z_+^T D z_+ + z_-^T D z_-$$

and

$$z^T L_G z \geq z_+^T L_G z_+ + z_-^T L_G z_-$$

if we can show that $(z(i) - z(j))^2 \geq \left(z_+(i) - z_+(j)\right)^2 + \left(z_-(i) - z_-(j)\right)^2$ for all i, j. This follows since if $z(i)$ and $z(j)$ have the same sign, then clearly $(z(i) - z(j))^2 = \left(z_+(i) - z_+(j)\right)^2 + \left(z_-(i) - z_-(j)\right)^2$ (where one of the two terms is zero), while if $z(i)$ and $z(j)$ have opposite signs then

$$\begin{aligned} (z(i) - z(j))^2 &= z(i)^2 - 2z(i)z(j) + z(j)^2 \\ &\geq z(i)^2 + z(j)^2 \\ &\geq \left(z_+(i) - z_+(j)\right)^2 + \left(z_-(i) - z_-(j)\right)^2, \end{aligned}$$

□

Now, we can finish the proof of Cheeger's inequality.We can find $S_+ \subseteq \operatorname{supp}\left(z_+\right), S_- \subseteq \operatorname{supp}\left(z_-\right)$ with

$$\begin{aligned} \min\left(\phi\left(S_+\right), \phi\left(S_-\right)\right) = \min\left(\frac{|\,\delta\left(S_+\right)}{\operatorname{vol}\left(S_+\right)}, \frac{|\,\delta\left(S_-\right)}{\operatorname{vol}\left(S_-\right)}\right) &\leq \min\left(\sqrt{2R\left(z_+\right)}, \sqrt{2R\left(z_-\right)}\right) \\ &\leq \sqrt{2R(y)} \end{aligned}$$

so that $\phi(G) \leq \min\left(\phi\left(S_+\right), \phi\left(S_-\right)\right) \leq \min \sqrt{2R(y)} = \sqrt{2\lambda_2}$, as desired.

6 Reference

[1] ORIE 6334: David P. Williamson,Bridging Continuous and Discrete Optimization

[2] MATH 867: Artem Novozhilov,Topics in Applied Mathematics: Mathematics of Networks