A Mathematical Introduction to Data Science **December 2, 2014**

Homework 6. Cheeger Inequalities and Spectral Clustering

Instructor: Yuan Yao *Due: Tuesday November 25, 2014*

The problem below marked by * is optional with bonus credits.

1. *Spectral Bipartition*: Consider the 376-by-475 matrix X of character-event for A Dream of Red Mansions, e.g. in the Matlab format

   `http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/hongloumeng376.mat`

   with a readme file:

   `http://www.math.pku.edu.cn/teachers/yaoy/data/hongloumeng/readme.m`

   Construct a weighted adjacency matrix for character-cooccurance network $A = XX^T$. Define the degree matrix $D = \mathrm{diag}(\sum_j A_{ij})$. Check if the graph is connected.

   (a) Find the second smallest generalized eigenvector of $L = D - A$, i.e. $(D - A)f = \lambda_2 f$ where $\lambda_2 > 0$;

   (b) Sort the nodes (characters) according to the ascending order of f, such that $f_1 \leq f_2 \leq \ldots \leq f_n$, and construct the subset $S_i = \{1, \ldots, i\}$;

   (c) Find an optimal subset S^* such that the following is minimized

   $$\alpha_f = \min_{S_i} \left\{ \frac{|\partial S_i|}{\min(|S_i|, |\bar{S}_i|)} \right\}$$

   where $|\partial S_i| = \sum_{x \sim y, x \in S_i, y \in \bar{S}_i} A_{xy}$ and $|S_i| = \sum_{x \in S_i} d_x = \sum_{x \in S_i, y} A_{xy}$.

   (d) Check if $\lambda_2 > \alpha_f$;

   (e) Quite often people find a suboptimal cut by $S^+ = \{i : f_i \geq 0\}$ and $S^- = \{i : f_i < 0\}$. Compute its Cheeger ratio

   $$h_{S^+} = \frac{|\partial S^+|}{\min(|S^+|, |S^-|)}$$

   and compare it with α_f, λ_2.

   (f) You may further recursively bipartite the subgraphs into two groups, which gives a recursive spectral bipartition.

2. *Directed Graph Laplacian*: Consider the following dataset with Chinese (mainland) University Weblink during 12/2001-1/2002,

   `http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/univ_cn.mat`

   where `rank_cn` is the research ranking of universities in that year, `univ_cn` contains the webpages of universities, and `W_cn` is the link matrix from university i to j.

Define a PageRank Markov Chain

$$P = \alpha P_0 + (1-\alpha)\frac{1}{n}ee^T, \quad \alpha = 0.85$$

where $P_0 = D_{out}^{-1}A$. Let $\phi \in \mathbb{R}_+^n$ be the stationary distribution of P, i.e. PageRank vector. Define $\Phi = \mathrm{diag}(\phi_i) \in \mathbb{R}^{n\times n}$.

(a) Construct the normalized directed Laplacian

$$\vec{\mathcal{L}} = I - \frac{1}{2}(\Phi^{1/2}P\Phi^{-1/2} + \Phi^{-1/2}P^T\Phi^{1/2})$$

(b) Use the second eigenvector of $\vec{\mathcal{L}}$ to bipartite the universities into two groups, and describe your algorithm in detail;

(c) Try to explain your observation through directed graph Cheeger inequality.

3. **Chung's Short Proof of Cheeger's Inequality*:

Chung's short proof is based on the fact that

$$h_G = \inf_{f\neq 0}\sup_{c\in\mathbb{R}} \frac{\sum_{x\sim y}|f(x)-f(y)|}{\sum_x |f(x)-c|d_x} \tag{1}$$

where the supreme over c is reached at $c^* \in median(f(x) : x \in V)$. Such a claim can be found in Theorem 2.9 in Chung's monograph, Spectral Graph Theory. In fact, Theorem 2.9 implies that the infimum above is reached at certain function f. From here,

$$\begin{aligned}
\lambda_1 &= R(f) = \sup_c \frac{\sum_{x\sim y}(f(x)-f(y))^2}{\sum_x (f(x)-c)^2 d_x}, && (2)\\
&\geq \frac{\sum_{x\sim y}(g(x)-g(y))^2}{\sum_x g(x)^2 d_x}, \quad g(x) = f(x)-c && (3)\\
&= \frac{(\sum_{x\sim y}(g(x)-g(y))^2)(\sum_{x\sim y}(g(x)+g(y))^2)}{(\sum_{x\in V} g^2(x)d_x)((\sum_{x\sim y}(g(x)+g(y))^2)} && (4)\\
&\geq \frac{(\sum_{x\sim y}|g^2(x)-g^2(y)|)^2}{(\sum_{x\in V} g^2(x)d_x)((\sum_{x\sim y}(g(x)+g(y))^2)}, \quad \text{Cauchy-Schwartz Inequality} && (5)\\
&\geq \frac{(\sum_{x\sim y}|g^2(x)-g^2(y)|)^2}{2(\sum_{x\in V} g^2(x)d_x)^2}, \quad (g(x)+g(y))^2 \leq 2(g^2(x)+g^2(y)) && (6)\\
&\geq \frac{h_G^2}{2}. && (7)
\end{aligned}$$

Is there any step wrong in the reasoning above? If yes, can you remedy it/them?