CMPSCI 711: More Advanced Algorithms

Graphs 10: Correlation Clustering

Andrew McGregor

- A 3 approximation for correlation clustering in complete graphs.

  [Ailon, Charikar, Newman, J.ACM 08]

- Emulating algorithm in small space and limited number of passes.

  [Ahn, Cormode, Guha, McGregor, Wirth ICML 16]

Last Compiled: March 20, 2018

Correlation Clustering

- Let G be a complete graph on n nodes where edges are colored red or blue. Given a clustering, say an edge e is unhappy if

$$(e \text{ is red and endpoints in different clusters})$$
$$\text{or } (e \text{ is blue and endpoints in same clusters})$$

- Problem: Find clustering minimizing number of unhappy edges.
- Say a triangle with exactly two red edges is bad. Let $\mathcal{B}$ be set of bad triangles. Minimum number of unhappy edges is at least

$$\min\{\sum_{e\in E} x_e : x_e \in \{0,1\} \text{ and } \sum_{e\in T} x_e \geq 1 \text{ for all } T \in \mathcal{B}\}$$

since, if we interpret $x_e = 1$ as making e unhappy, we need to disappoint at least one edge in each bad triangle.

Correlation Clustering Algorithm

We say node u is a friend of node v is the edge $\{u, v\}$ is red.

Non-Streaming Version:

- Randomly order nodes: $v_1, v_2, \ldots, v_n$. Mark each as uncovered.
- For $i = 1$ to n: If v_i uncovered, let v_i and it's uncovered friends be a cluster. Cover these nodes and say "v_i was chosen as a pivot."

Emulating in the Streaming Model:

- Preprocess: Randomly order nodes: $v_1, v_2, \ldots, v_n$.
- First Pass: Store all red edges incident to $\{v_1, \ldots, v_{\sqrt{n}}\}$. Emulate the first $\sqrt{n}$ steps of the algorithm.
- Second Pass: Store all red edges that have both endpoints uncovered at end of first pass. Emulate remaining steps of the algorithm.

Will show a) achieves factor 3 approx and b) the streaming algorithm uses $\tilde{O}(n^{1.5})$ space. Can also get a $O(\log \log n)$ pass, $\tilde{O}(n)$ space algorithm.

Analyzing Approximation Ratio: Part 1

- Let $T \in \mathcal{B}$ with nodes $\{a, b, c\}$. Define events

  B_T = node in T chosen as a pivot and other nodes uncovered at time

  $B^a_{b,c} = B_T \cap \{\text{pivot}=a\}$ $\quad B^b_{a,c} = B_T \cap \{\text{pivot}=b\}$ $\quad B^c_{a,b} = B_T \cap \{\text{pivot}=c\}$
- Charge an unhappy edge to bad triangle formed by it and the pivot when it was made unhappy. Each $T \in \mathcal{B}$ charged at most once.
- Let $z_T = \mathbb{P}[B_T]/3$. Expected cost is $\sum_{t\in\mathcal{B}} \mathbb{P}[B_T] = 3\sum_{t\in\mathcal{B}} z_T$.
- $\mathbb{P}\left[B^a_{b,c}\right] = \mathbb{P}\left[B^b_{a,c}\right] = \mathbb{P}\left[B^c_{a,b}\right] = z_T$, e.g.,

$$\mathbb{P}\left[B^a_{b,c}\right] = \mathbb{P}\left[B^a_{b,c} | B_T\right]\mathbb{P}[B_T] = \mathbb{P}[B_T]/3 = z_T \ .$$

- Since $B^a_{b,c} \cap B^{a'}_{b,c} = \emptyset$ for $\{a, b, c\}, \{a', b, c\} \in \mathcal{B}$,

$$\sum_{a:\{a,b,c\}\in\mathcal{B}} \mathbb{P}\left[B^a_{b,c}\right] \leq 1$$

- And so for any $e = \{b, c\}$

$$\sum_{T\in\mathcal{B}:e\in T} z_T = \sum_{a:\{a,b,c\}\in\mathcal{B}} \mathbb{P}\left[B^a_{b,c}\right] \leq 1$$

Analyzing Approximation Ratio: Part 2

- Using LP duality:

$$\begin{aligned}
\textsc{opt} &\geq \min\{\sum_{e\in E} x_e : x_e \in \{0,1\} \text{ and } \sum_{e\in T} x_e \geq 1 \text{ for all } T \in \mathcal{B}\} \\
&\geq \min\{\sum_{e\in E} x_e : x_e \geq 0 \text{ and } \sum_{e\in T} x_e \geq 1 \text{ for all } T \in \mathcal{B}\} \\
&= \max\{\sum_{T\in\mathcal{B}} y_T : y_T \geq 0 \text{ and } \sum_{T\in\mathcal{B}:e\in T} y_T \leq 1 \text{ for all } e \in E\} \\
&\geq \sum_{T\in\mathcal{B}} z_T
\end{aligned}$$

  where the last line follows since $\sum_{T\in\mathcal{B}:e\in T} z_T \leq 1$.

- Hence, expected cost is

$$\sum_{t\in\mathcal{B}} \mathbb{P}[B_T] = 3\sum_{t\in\mathcal{B}} z_T \leq 3\textsc{opt} \ .$$

Space Analysis

Algorithm stores $\sqrt{n} \times n$ edges in first pass. Next lemma implies every uncovered node has $\tilde{O}(n^{0.5})$ friends after first pass with high probability. Hence, $\tilde{O}(n^{1.5})$ edges are stored in the second pass.

Lemma

After r iterations, every uncovered node has $< 10(\log n)n/r$ friends whp.

- Let $\alpha = 10(\log n)n/r$. Fix a node v and define event,

  $B_i =$ "v uncovered and has at least α uncovered friends after i iterations"

- Note that $\mathbb{P}[B_i | B_{i-1} \cap \ldots B_1] \leq 1 - \frac{\alpha}{n-i+1} \leq \exp(-\alpha/n)$ and so,

$$\mathbb{P}[B_r] = \mathbb{P}[B_r \cap B_{r-1} \cap \ldots \cap B_1] \leq \exp(-\alpha/n)^r \leq 1/n^{10}$$

- Hence, the union bound implies that with probability at least $1 - 1/n^9$, every uncovered node has less than α friends.