Graph Sparsification

- Sparse graphs:
  Graphs with less # of edges
  $\Rightarrow$ less space to store the graphs
  $\Rightarrow$ less processing time

- IDEA: Given an undirected weighted graph $G=(V,E,w)$, where $|V|=n$, $|E|=n$, the goal is to output the graph H, a subgraph of G with fewer edges, where H may be reweighted, while preserving "interesting quantities"

Interesting quantities:
  e.g. extremal (min, max) cuts, eigenvalues, random walk properties.
  (typically captured by graph Laplacian)

Graph Sparsification w.r.t. cuts I

- For simplicity, let's assume we are given an undirected unweighted graph G. The goal is to approximate G by a sparse graph while preserving the cut size for all possible cuts with small errors.
- More precisely, a cut $(S, \bar{S})$ is a partition of V into two subsets S, $\bar{S} = V \setminus S$. Let $E_G(S,\bar{S})$ denote the set of edges crossing the cut $(S,\bar{S})$ in G, i.e., $E_G(S,\bar{S}) = \{u,v \in E(G) \mid u \in S \text{ and } v \notin S\}$. The capacity (or size) of the cut $(S,\bar{S})$ is denoted by $|E_G(S,\bar{S})|$. If G is weighted by a weight function $w: E \to \mathbb{R}$, then $|E_G(S,\bar{S})| = \sum_{e \in E_G(S,\bar{S})} w(e)$.

Goal: Construct a graph $H = (V, E')$, where $E' \subset E$ and $|E'| \ll |E|$, and H is potentially reweight by a function $w': E' \to \mathbb{R}$ s.t. $\forall U \subseteq V$

$$|E_H(U,\bar{U})| = (1 \pm \epsilon)\,|E_G(U,\bar{U})|$$

for a small fixed $\epsilon > 0$. Such a graph H is call a "cut sparsifier" of G

Cut Sparsifier of a Complete Graph

- Given a complete graph $G = K_n$, how can we construct a cut sparsifier H of K_n? (note that $|E(K_n)| = O(n^2)$.)

- Sub sampling: Consider the following proces:
  - Sample (keep) every edge independenty with some probability p. Then,

$$\mathbb{E}[\,|E(H)|\,] = p\,|E(K_n)| \quad , \text{ and}$$

$$\forall U \subseteq V(K_n),\ \mathbb{E}[\,|E_H(U,\bar{U})|\,] = p\,|E_{K_n}(U,\bar{U})|$$

  - Let's assign weight $\frac{1}{p}$ to each edge of H so that

$$\forall U \subseteq V(K_n),\ \mathbb{E}[\,|E_H(U,\bar{U})|\,] = \cancel{p}\,\frac{1}{\cancel{p}}\,|E_{K_n}(U,\bar{U})|.$$

  - So in expectation, cuts are preserved !! Let's analyze how likely the cut capacity is close to its expectation.

Theorem [Chernoff-Heffding Concentration Bound]
Let $X = \sum_{i \in [n]} X_i$, where each $X_i \in [0,1]$ is an indicator random variable, and $(X_i : i \in [n])$ are independently distributed. Then,

- $\forall t > 0$, $\Pr\{|X - \mathbb{E}[X]| > t\} \le e^{-2t^2/n}$
- $\forall 0 < \varepsilon < 1$, $\Pr\{X < (1-\varepsilon)\mathbb{E}[X]\} \le e^{-\varepsilon^2 \mathbb{E}[X]/2}$
- $\forall 0 < \varepsilon < 1$, $\Pr\{X > (1+\varepsilon)\mathbb{E}[X]\} \le e^{-\varepsilon^2 \mathbb{E}[X]/3}$
- $\forall t > 2e\mathbb{E}[X]$, $\Pr\{X > t\} \le 2^{-t}$.

Analysis of Subsampling

- To simplify the analysis, we consider H in the unweighted version.

For a subset $U \subseteq V$, let $q = |U|$, $C_H = |E_H(U, \bar{U})|$, and $C_G = |E_G(U, \bar{U})| \geq \frac{qn}{2}$, thus $\mathbb{E}[C_H] = pC_G \geq \frac{pqn}{2}$. Using the concentration bound from above,

$$\Pr\{ C_H > (1+\varepsilon)\mathbb{E}[C_H] \} \leq e^{-\varepsilon^2 \mathbb{E}[C_H]/3} = e^{-\frac{\varepsilon^2 p C_G}{3}} \leq e^{-\frac{\varepsilon^2 pqn}{6}}$$

- Suppose we want the RHS to be at most $1/n^{dq}$ for some fixed $d > 1$ (say $d=5$)

Then, we shoud set $p \geq \frac{6d\log n}{\varepsilon^2 n}$

$$\Rightarrow e^{-\frac{\varepsilon^2 pqn}{6}} \leq e^{-\cancel{\varepsilon^2}\left(\frac{\cancel{6}d\log n}{\cancel{\varepsilon^2}\cancel{n}}\right)\frac{q\cancel{n}}{\cancel{6}}} = e^{-dq\log n} = n^{-dq} = \frac{1}{n^{dq}}$$

- Note that a similar bound applies to deviation in the other direction, we get

$$\Pr\{ C_H \notin (1 \pm \varepsilon)\mathbb{E}[C_H] \} \leq \frac{2}{n^{dq}}$$

- Also, note that the failure probability above is for a single cut. The probability for any cut to fail is obtained by the following analysis.

$$\Pr\{\text{any cut fails}\} \leq \Pr\{\text{any cut fails independently}\}$$

$$= \sum_{1<q<N} (\#\text{ of cut of size } q) \cdot \Pr\{\text{cut of size } q \text{ fails}\}$$

$$\leq \sum_{1<q<N} \binom{n}{q} \cdot \frac{2}{n^{dq}}$$

$\theta(n^q)$

$$\leq \sum_{1<q<N} \cancel{n^q} \cdot \frac{2}{n^d \cdot \cancel{n^q}} \leq \frac{2n}{n^d} = \frac{2}{n^{d-1}}.$$

- The subsampling will fail with probability at most $2/n^{d-1}$, if $p \geqslant \frac{6d\log n}{\varepsilon^2 n}$

- If we set $d=5$ then, the downsampling will success with high probability (i.e., the success probability is at least $1-2/n^4$), and the graph sparsifier H will have $p|E(K_n)| = \frac{(6\times 5)\log n}{\varepsilon^2 n} \times \Theta(n^2) = \Theta\left(\frac{n\log n}{\varepsilon^2}\right)$

$$\Rightarrow \mathbb{E}\left[|E[H]|\right] = \tilde{\Theta}\left(\frac{n}{\varepsilon^2}\right)$$

- Clearly, the sub sampling of H takes time $O(m)$.

Theorem 1: There exists a randomized construction of a cut sparsifier H of a complete graph K_n in time $O(m)$, where $|E(H)| = \tilde{\Theta}\left(\frac{n}{\varepsilon^2}\right)$, with high probability of success.