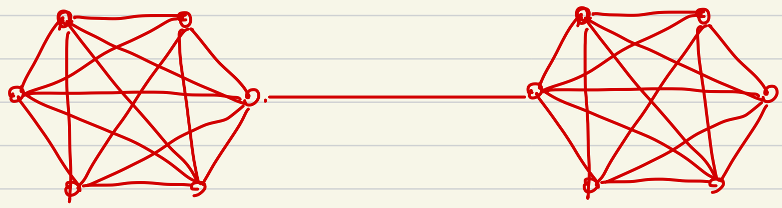
- The random subsampling of edges works great only if the input graph has min cuts of large size. This can be seen as the number of edges decreass by a factor of $\Theta = \left(\frac{\ln n}{C_G^*}\right)$ in the sparsifier.
- Now, let's think about how the random subsampling could go wrong.



The above graph is called "dumbell graph" (two cliques of $n/2$ vertices connected by a single edge). Obsere that the size of the min cut in such graph is $C_G^* = 1$. This means if we want the random subsampling to success w.h.p. we should set $\frac{(d+2)\ln n}{\varepsilon^2 C_G^*} \leq p \leq 1$, but this is impossible because $\frac{(d+2)\ln n}{\varepsilon^2 C_G^*} = \frac{(d+2)\ln n}{\varepsilon^2} \geq (d+2)\ln n \geq 1$.

Moreover, the number of edges in the sparsifier is $\Theta\left(\frac{\ln n}{C_G^*} n^2\right) = \Theta\left(n^2 \cdot \ln n\right)$, so it still quite dense.

Cut Sparsifier by Edge Connectivity

- For dumbbell graph, intuitively, we would like to get rid of most edges in the graph, but keep the edge in the middle. This means we shold not subsample all edges with the same probability p.

IDEA: Subsample the edges non-uniformly: Important edges should be chosen with high probability.

- To quantify how importance an edge is, we introduce the notion of edge-connectivity:

For an edge e, we let k_e be the minimum capacity of a cut containing e, i.e.,

$$k_e := \min\{E_G(U,\bar{U}) : U \subset V \text{ and } e \in E_G(U,\bar{U})\}$$

Note that edges with high connectivity only appear in cuts with many other edges, so they are not extremely important.

- In the dumbbell graph, the clique edges have connectivity $n/2 - 1$, where as the single edge in the middle has connectivity 1.

- For ease of discussion, for $U \subseteq V$, we let $c_H(U) = c_H(E_G(U,\bar{U}))$ to denote the capacity of cut $E_H(U,\bar{U})$, and $c_G(U) = |E_G(U,\bar{U})|$.

The subsampling process (scheme)

1. $\forall e \in E(G)$, compute the edge connectivity k_e.
2. For $i = 1, 2, \dots, \rho$
3. For each $e \in E(G)$, subsample e with probability $1/k_e$, and increase weight of e by k_e/ρ

- Remark that, after the subsampling, $\forall e \in E(G)$, $\mathbb{E}[w_e] = \sum_{i=1}^{\rho} \frac{1}{k_e} \cdot \frac{k_e}{\rho} = 1$.

Moreover, $\forall U \subseteq V$, $\mathbb{E}[C_H(U)] = \sum_{e \in E_G(U, \bar{U})} 1 = |E_G(U, \bar{U})| = C_G(U)$

So, in expectation, all cuts are preserved!!

Now, let's analyze how likely the cut capacity is close to its expectation.

Analysis of the Subsampling process:

- Consider some cut $E_G(U, \bar{U})$ for $U \subseteq V$. The weight of that cut after subsampling is the r.v.

$$C_H(U) := \sum_{i=1}^{\rho} \sum_{e \in E_G(U, \bar{U})} \frac{k_e}{\rho} X_{i,e},$$

where $X_{i,e}$ is an i.r.v. s.t.

$$X_{i,e} := \begin{cases} 1, & \text{if } e \text{ is chosen with probability } \frac{1}{k_e} \text{ at the } i\text{th round} \\ 0, & \text{otherwise} \end{cases}$$

Note that here we cannot apply the Chernoff bound directly as it is designed to work for any sum of independent random variables taking values in $[0, 1]$, but here we have different coefficients k_e.

Analysis trick: group edges by the edge connectivity

- To handle the different coefficients, we partition the edges into groups whose connectivities are roughly the same.
- Formally, we set $E = E_1 \cup \cdots\cdots \cup E_{\log n}$, where

$$E_i = \{e \in E : 2^{i-1} \leq k_e < 2^i\}$$

Now, instead of proving concentration for $C_H(U)$, we will prove concentration for $C_H(E_G(U,\bar{U}) \cap E_i)$.

- So, let $F = E_G(U,\bar{U}) \cap E_i$. We call such a set a <u>cut-induced set</u>. Note that there could be some other cut U' s.t. $F = E_G(U,\bar{U}) \cap E_i = E_G(U',\bar{U}') \cap E_i$. We are going to focus on the smallest such cut because we want the sampling error of $C_H(F)$ to be small relative to $|E_G(U,\bar{U})|$. That is hardest when $|E_G(U,\bar{U})|$ is small. So, define

$$q(F) = \min\{|E_G(U,\bar{U})| : U \subseteq V \text{ and } E_G(U,\bar{U}) \cap E_i = F\}$$

- Let's now analyze the sampling error of $C_H(F)$
  - Recall that $C_H(F) = \sum_{i=1}^{p} \sum_{e \in F} \frac{k_e}{p} X_{i,e}$. Because we know $\forall e \in F,\ k_e < 2^i$, we can rewrite

$$C_H(F) = \sum_{i=1}^{p} \sum_{e \in F} \frac{2^i}{p} Y_e \quad , \text{ where}$$

$$Y_e = \begin{cases} k_e/2^i, & \text{if } e \text{ is chosen with probability } 1/k_e \\ 0, & \text{otherwise} \end{cases}$$

Note that each $Y_e \in [0,1]$. Define $Y := \sum_{i=1}^{\rho} \sum_{e \in F} Y_e$, and we can apply the Chernoff bound to Y.

Also, note $Y = c_H(F) \times \frac{\rho}{2^i} \Rightarrow \mathbb{E}[Y] = \frac{\rho}{2^i} \mathbb{E}[c_H(F)] = \frac{\rho}{2^i} c_G(F)$
$= \frac{\rho}{2^i} |F|$

$$\Pr\{ c_H(F) > (1+\delta) \mathbb{E}[c_H(F)] \} =$$

$$\Pr\{ Y > (1+\delta) \mathbb{E}[Y] \} \leq e^{-\frac{\delta^2 \mathbb{E}[Y]}{3}}$$

For technical reason, we define $\delta = \frac{\varepsilon q(F)}{|F| \log n}$ (the sampling error to be small!)

$$\Rightarrow \Pr\{ Y > (1+\delta) \mathbb{E}[Y] \} \leq e^{-\frac{\delta^2 \mathbb{E}[Y]}{3}}$$
$$\leq e^{-\left(\frac{\varepsilon^2 q(F)^2}{|F|^2 \log^2 n}\right) \cdot \frac{\rho}{2^i} \frac{\cancel{|F|}}{3}}$$
$$= e^{-\frac{\varepsilon^2 \rho q(F)^2}{3 \cdot 2^i \log^2 n |F|}}$$

Since any cut C satisfying $C \cap E_i = F$ must have $|C| \geq |F|$, so $q(F) \geq |F|$.

$$\Rightarrow \Pr\{ Y > (1+\delta) \mathbb{E}[Y] \} \leq e^{-\frac{\varepsilon^2 \rho q(F)}{3 \cdot 2^i \log^2 n} \cdot \frac{q(F)}{|F|}}$$
$$\leq e^{-\frac{\varepsilon^2 \rho q(F)}{3 \cdot 2^i \log^2 n} \cdot \frac{\cancel{|F|}}{\cancel{|F|}}}$$
$$= e^{-\frac{\varepsilon^2 \rho q(F)}{3 \cdot 2^i \log^2 n}}$$

<u>Claim 1</u>: Let $F \subseteq E_i$ be a cut-induced set. Then,

$$\Pr\{ c_H(F) \notin (1 \pm \delta) \mathbb{E}[c_H(F)] \}$$
$$= \Pr\{ |c_H(F) - \mathbb{E}[c_H(F)]| > \delta \mathbb{E}[c_H(F)] \}$$
$$= \Pr\{ |c_H(F) - \mathbb{E}[c_H(F)]| > \frac{\varepsilon q(F)}{\log n} \}$$
$$\leq 2 \cdot e^{-\frac{\varepsilon^2 \rho q(F)}{3 \cdot 2^i \log^2 n}} = \frac{2}{e^{\frac{\varepsilon^2 \rho q(F)}{3 \cdot 2^i} \log^2 n}}$$

- At this point, we have shown that any induced-cut set fails with certain probability relative to the size of the set.
- Indeed, if we can show that, w.h.p., any cut-induced set F, $|w(F)-|F|| \leq \frac{\varepsilon q(F)}{\log n}$. Then, $\forall U \subseteq V$, we have that

$$|C_H(U) - |E_G(U,\bar{U})|| \leq \sum_{i=1}^{\log n} |C_H(E_G(U,\bar{U}) \cap E_i) - |E_G(U,\bar{U}) \cap E_i||$$

$$\leq \sum_{i=1}^{\log n} \frac{\varepsilon q(E_G(U,\bar{U}) \cap E_i)}{\log n} \text{ (by union bound)}$$

$$\leq \sum_{i=1}^{\log n} \frac{\varepsilon |E_G(U,\bar{U})|}{\log n} \quad (\forall C \text{ s.t. } F = C \cap E_i, \; q(F) \leq C$$

$$= \varepsilon |E_G(U,\bar{U})|$$

- Remark that to prove this desired result, we need to show that, w.h.p., the second inequality holds
  - Fix any $i = 1, 2, \ldots, \log n$. Let $F^1, F^2, \ldots$ be all the cut-induced subsets of E_i, ordered s.t. $q(F^1) \leq q(F^2) \leq \ldots$.
  - Define (by Claim 1)

$$p_j := \Pr\left\{ |C_H(F^j) - \mathbb{E}[C_H(F^j)]| > \frac{\varepsilon q(F^j)}{\log n} \right\}$$

$$\leq 2 e^{-\left(\frac{\varepsilon^2 \rho q(F^j)}{3 \cdot 2^i \log^2 n}\right)}$$

- Next, we will use a union bound to show that all of the F^j's are concentrated. However, we need to be clever because there can be exponentially many of them.

- Consider the first n^2 cut-induced sets $F^1, \dots, F^{n^2}$.
Since $\forall e \in F^j$, e belongs to E_i, we have $k_e \geq 2^{i-1}$.
This means any cut containing the edge e is of size at least 2^{i-1}, and therefore $q(F^j) \geq 2^{i-1}$.
(from cut capacity)

By plugging $\rho = 100 \log^3 n / \varepsilon^2$, we have

$$p_j \leq 2 e^{-\left(\frac{100 \log^3 n}{\varepsilon^2} \cdot \frac{\varepsilon^2 \cancel{2^{i-1}}}{3 \cdot \frac{\cancel{2^i}}{2} \log^2 n}\right)} \leq 2 \cdot e^{-(16 \log n)} \leq 2 n^{-16}.$$

It follows that

$$\sum_{j=1}^{n^2} p_j \leq \sum_{j=1}^{n^2} 2 n^{-16} \leq \frac{2}{n^{14}}.$$

- Now, consider the remaining cut-induced sets F^j for $j > n^2$. In fact, we need one more key ingredient stated in the following theorem:

Theorem 2: Let $G = (V, E)$ be a graph. Let $B \subseteq E$ be arbitrary and let $K \leq \min\{k_e : e \in B\}$. Then, for every real $\alpha \geq 1$,

$$|\{E_G(U, \bar{U}) \cap B : U \subseteq V \text{ and } |E_G(U, \bar{U})| \leq \alpha K\}| < n^{2\alpha}.$$

If we apply Theorem 2 with $B = E_i$, implying $K = 2^{i-1}$, then the theorem states that

$$\forall \alpha \geq 1, |\{\text{cut induced set } F \subseteq E_i : q(F) \leq \alpha 2^{i-1}\}| < n^{2\alpha}$$

*** So, the number of F^j with $q(F^j) \leq \alpha 2^{i-1}$ is less than $n^{2\alpha}$.

- By the ordering of the F^j's, it must follow that

$$q(\underline{F^{n^{2\alpha}}}) > \alpha 2^{i-1}$$

$$F^{n^2}, F^{n^{2}+1}, \ldots$$

- By substituting $\alpha = \frac{\ln j}{2\ln n}$ (for technical reason), we get

$$q(F^j) > \frac{\ln j}{2\ln n} 2^{i-1}$$

$\Rightarrow$ For $j > n^2$,
$$P_j \leq 2e^{-\frac{100\log^{3}n}{\cancel{\varepsilon^2}} \cdot \frac{\cancel{\varepsilon^2}}{3\cdot \cancel{2^i}\log^2 n} \cdot \frac{\ln j}{2\cancel{\ln n}} \cdot \cancel{2^{i-1}}}$$
$$< j^{-8}$$

$$\Rightarrow \sum_{j>n^2} P_j \leq \sum_{j>n^2} j^{-8} \leq \int_{n^2}^{\infty} j^{-8}\, dj = -\frac{j^{-7}}{7}\Big|_{j=n^2}^{\infty}$$
$$< n^{-14}.$$

$\Rightarrow$ So, by a union bound,

$$\Pr\left\{ c_H(E_G(U,\bar{U}) \cap E_i) - |E_G(U,\bar{U}) \cap E_i| > \frac{\varepsilon\, q(E_G(U,\bar{U}) \cap E_i)}{\log n} \right\}$$
$$\leq \sum_{j=1}^{n^2} P_j + \sum_{j>n^2} P_j \leq \frac{2}{n^{14}} + \frac{1}{n^{14}} < \frac{1}{n^2}.$$

- Hence, $\forall U \subseteq V$

$$\Pr\left\{ c_H(U) \notin (1 \pm \varepsilon) \left| E_G(U,\bar{U}) \right| \right\} \leq$$

$$\sum_{i=1}^{\log n} \Pr\left\{ c_H(E_G(U,\bar{U}) \cap E_i) - \left| E_G(U,\bar{U}) \cap E_i \right| > \frac{\varepsilon\, q(E_G(U,\bar{U}) \cap E_i)}{\log n} \right\}$$

$$\leq \sum_{i=1}^{\log n} \frac{1}{n^2} \leq \frac{1}{n}$$

- We conclude here that the subsampling by edge connectivity success with high probability if the sampling process takes $\rho = \frac{100 \log^3 n}{3}$ rounds.

- Now, let's consider the number edges in the sparsifier H obtained by the sampling process.
  - $|E(H)| \leq \sum_{i=1}^{\rho} \sum_{e \in E(G)} X_{i,e}$
  - $\mathbb{E}\left[|E(H)| \right] = \rho \sum_{e \in E(G)} 1/k_e = O\left(\frac{\log^3 n}{\varepsilon^2}\right) \cdot \sum_{e \in E(G)} 1/k_e$

Fact 3: For any graph $G = (V, E)$, $\sum_{e \in E(G)} 1/k_e \leq n - 1$.

$$\Rightarrow \mathbb{E}\left[|E(H)| \right] = O\left(\frac{n \log^3 n}{\varepsilon^2} \right)$$

Theorem 4: The subsampling process by edge connectivity can produce a cut sparsifier H of an arbitrary graph G in $O\left(\frac{m \log^3 n}{\varepsilon^2}\right)$ time with $|E(H)| = O\left(\frac{n \log^3 n}{\varepsilon^2}\right)$ w.h.p.