1 Recap

- Graph Entropy: Given $G = (V, E)$, we define $H(G) = \min I(X; Y)$ over joint distributions $(X, Y)$ where $X$ is a uniformly random vertex in $V$, and $Y$ is an independent subset of $V$ that contains $X$.
- $H(G_1 \cup G_2) \leq H(G_1) + H(G_2)$.
- $H(G_1) \leq H(G_1 \cup G_2)$
- If $G_1, ..., G_k$ are connected component of $G$, $H(G) = \sum_{i \in [k]} \rho_i H(G_i)$ where $\rho_i := \frac{|V(G_i)|}{|V(G)|}$.

2 Number of bipartite graphs to cover the complete graph

Suppose that we have the complete graph $K_n = (V, \binom{V}{2})$. We want to cover $K_n$ by $l$ bipartite graphs, $G_1, ..., G_l$ in a sense that

- For each $i$, $G_i = (V, E_i)$ is a bipartite graph.
- For each $(u, v) \in \binom{V}{2}$, $(u, v) \in E_i$ for some $i$. In other words, $K_n = G_1 \cup ... \cup G_l$.

Question: What is the minimum number $l$ of bipartite graphs needed to cover $K_n$?

Construction: Identify each vertex with a binary string of length $\lceil \log n \rceil$. The $i$th bipartite graph connects every two vertices whose binary representations differ at the $i$th position. It is easy to see that these $\lceil \log n \rceil$ bipartite graphs cover all the pairs.

Lower bound: In the previous lecture, we saw that

- $H(K_n) = \log n$
- $K_n = G_1 \cup ... \cup G_l$ implies $H(K_n) \leq \sum_{i} H(G_i)$
- $H(G_i) \leq 1$

Therefore, $\log n = H(K_n) \leq \sum_{i} H(G_i) \leq l \leq \lceil \log n \rceil$. Generally, given a graph $G = (V, E)$, the same upper and lower bound techniques work to show that $H(G) \leq l \leq \lceil \log \chi(G) \rceil$ (for the upper bound, identify each color with a binary string). $\log \chi(G)$, which is always at least $H(G)$, gives one intuition about $H(G)$, even though the difference can be made arbitrarily large.
3 Perfect Hash Families

Setting: A database where each file is an element of $[N]$. A hash function maps a file to a much smaller domain; $h : [N] \to [b]$ where $b \ll N$.

Suppose we have a hash family $\mathcal{H} = \{h_1, ..., h_t\}$ where for each $i$, $h_i : [N] \to [b]$ is a hash function. Our goal is to design $\mathcal{H}$ such that it can differentiate between up to $k$ files ($k < b$). In other words,

$$\forall S \subseteq [N], |S| = k : \exists h \in \mathcal{H} \text{ such that } h \text{ is injective on } S$$

If we think $\mathcal{H}$ as an $N \times t$ matrix (each row corresponds to a file $x$, each column corresponds to a hash function $h_i$, and $\mathcal{H}(x, h_i)$), we require that for every choice of $k$ rows ($x_1, ..., x_k$), there exists a column $h_i$ such that $h_i(x_1), ..., h_i(x_k)$ are pairwise distinct. We call $\mathcal{H}$ $k$-perfect hash family if the above condition is satisfied. The question is, how small can $t$ be?

3.1 Upper bound

Claim 3.1. Assume $b \geq k^2$. Then $t = O(k \log \frac{N}{k})$ suffices.

Proof. Pick each $h_i : [N] \to [b]$ uniformly and independently at random. Fix $S \subseteq [N], |S| = k$.

$$Pr[h_i \text{ is injective on } S] = 1 \cdot \frac{b - 1}{b} \cdot \frac{b - k + 1}{b} \geq (1 - \frac{k}{b})^k \geq (1 - \frac{1}{k})^k \geq \frac{1}{4}$$

$$\Rightarrow Pr[\forall i, h_i \text{ is not injective on } S] \leq \left(\frac{3}{4}\right)^t$$

$$\Rightarrow Pr[\mathcal{H} \text{ is not } k\text{-perfect}] \leq \left(\frac{N}{k}\right)^t \leq \left(\frac{Ne}{k}\right)^t \left(\frac{3}{4}\right)^t = 2^O(k \log(N/k)) = \Omega(t)$$

The probability can be made less than 1 for some $t = O(k \log \frac{N}{k})$. \hfill \Box

3.2 Lower bound

Claim 3.2. For all $k \geq 2$, $t \geq \frac{\log N}{\log b}$.

Proof. It follows from the pigeonhole principle: $\forall x_1 \neq x_2 \in [N]$, we must have $(h_1(x_1), ..., h_t(x_1)) \neq (h_2(x_1), ..., h_t(x_2))$. Therefore, $N \leq b^t \Rightarrow t \geq \frac{\log N}{\log b}$. \hfill \Box

There is a stronger lower bound due to Fredman Komlős in 1984.

Theorem 3.3. $t \geq \frac{\log N}{b(b-1) \ldots (b-k+2)} \log(b(b-1) \ldots (b-k+2))$

Proof. Assume $b|N$. Define $G = (V, E)$ such that

- $V = \{(D, x) : D \subseteq [N], |D| = k - 2, x \in [N] - D\}$.
- $E = \{((D, x_1), (D, x_2)) : \forall D, x_1 \neq x_2\}$. 

2
Given a $k$-perfect hash family $\mathcal{H}$, we construct $\{G_h\}$ such that $G = \cup_{h \in \mathcal{H}} G_h$. The construction is as the following.

- $V(G_h) = V(G)$.
- $E = \{((D, x_1), (D, x_2)) : h$ is injective on $D \cup \{x_1, x_2\}\}$.

Every $\{(D, x_1), (D, x_2)\} \in E(G)$ is covered by $G_h$ where $h$ is injective on $D \cup \{x_1, x_2\}$, so $G = \cup_{h \in \mathcal{H}} G_h$.

Now we want to argue that each $H(G_h)$ is small. Fix $h$. For a choice of $D$,

- If $h$ is not injective on $D$, $H(G_{h,D}) = 0$ where $G_{h,D}$ indicates the connected component of $G_h$ corresponding to $D$.
- If $h$ is injective on $D$, $G_{h,D}$ is $(b - k + 2)$-partite. This can be shown by defining $A_i := \{(D, x) : h(x) = i\}$ for each $i \notin h(D)$. Since $h$ is injective there are exactly $b - k + 2$ choices of $i$, and there is no edge between $(D, x_1)$ and $(D, x_2)$ if $h(x_1) = h(x_2)$. From the last lecture, $H(G_{h,D}) \leq \log(b - k + 2)$.

In any case, $H(G_h, D) \leq \log(b - k + 2)$ and $H(G_h) \leq \log(b - k + 2)$. Together with $H(G) = \log(N - k + 2)$, we can conclude that $t \geq \frac{\log(N - k + 2)}{\log(b - k + 2)}$.

To get a better bound, we want to show that $G_h$ has a large fraction of isolated vertices. Define $p$ the probability that a uniform random vertex of $G_h$ is isolated. Let $E$ be the set of isolated vertices. The same argument shows that $H(G_h - E) \leq \log(b - k + 2)$ as well, so we have

$$H(G_h) = pH(E) + (1 - p)H(G_h - E) \leq (1 - p)\log(b - k + 2)$$

Therefore, an upper bound of $1 - p$ is needed to achieve a better lower bound on $t$. $(D, x)$ is isolated if and only if $h$ is not injective on $D \cup \{x\}$, so $p$ is the probability over uniformly chosen $(k - 1)$-subset $S$ that $h$ is not injective on $S$.

**Claim 3.4.** *Without loss of generality, we can assume that $|h^{-1}(1)| = \ldots = |h^{-1}(b)| = \frac{N}{b}$. In other words, maximum $p$ (minimum $1 - p$) is achieved by $|h^{-1}(1)| = \ldots = |h^{-1}(b)| = \frac{N}{b}$.*

**Proof.** Assume that $|h^{-1}(1)| > |h^{-1}(2)| + 1$. Take any $x$ such that $h(x) = 1$ and change $h$ such that $h(x) = 2$.

$$p = \Pr_S[h \text{ is injective on } S] = \Pr(x \in S) \Pr_S[h \text{ is injective on } S|x \in S] + \Pr(x \notin S) \Pr_S[h \text{ is injective on } S|x \notin S]$$

Since we only changed $h(x)$, the second term does not change. The first term increases since given that $x \in S$, $S - \{x\}$ needs to be disjoint from $h^{-1}(h(x))$ and the size of it became smaller. 

\[ \square \]
Now, $1 - p \leq 1 \cdot \frac{b-1}{b} \cdot \frac{b-k+2}{b} \cdots$ and $H(G_h) \leq (1 - p) \log(b - k + 2)$ Therefore,

$$t \geq \frac{H(G)}{\max_h H(G_h)} \geq \frac{\log(N - k + 2)}{(1 - p) \log(b - k + 2)} \geq \frac{b^{k-1}}{b(b-1)\cdots(b-k+2)} \frac{\log(N - k + 2)}{\log(b - k + 2)}$$