Thu 3/21

Last time: Graph Entropy

\[ G = (V,E), \quad H(G) = \min_{X \text{ uniform on } V} \ I(X;Y) \]
\[ Y \subseteq V \text{ independent set} \]
\[ X \in Y. \]

Subadditivity: \[ H(G_1 U G_2) \leq H(G_1) + H(G_2) \]

Monotonicity: \[ H(G_1) \leq H(G_1 U G_2) \]

Disjoint union: \[ G_1, \ldots, G_k \text{ connected components of } G, \]
\[ H(G) = \sum_{i \in [k]} \rho_i H(G_i), \quad \rho_i = \frac{|V(G_i)|}{|V(G)|}. \]

Today: Applications in Combinatorics.

(for lower bounds)

[ Intuition: A combinatorial process that produces a "high entropy" graph from "small entropy" ones needs a lot of steps by subadditivity.]

Quick example: What is the smallest number of bipartite graphs that is needed to cover the complete graph \( K_n \)?
Construction (upper bound):

Represent each vertex by an $l$-bit string, $l = \lceil \log n \rceil$.

The $i$th bipartite graph connects every two vertices whose binary representation differs at the $i$th coordinate.

$\Rightarrow l$ graphs are sufficient.

Lower bound: Let $G_1, \ldots, G_l$ be the bipartite graphs.

$H(G_1 \cup \ldots \cup G_l) = H(K_n) = \log n$

But $\{ H(G_i) \leq 1$ 

$H(G_1 \cup \ldots \cup G_l) \leq \sum_{i \in [l]} H(G_i) \leq l$. 

$\Rightarrow l \geq \lceil \log n \rceil$. 

[Similar bound can be shown by a chromatic number argument].
Perfect hash function family:

Setting: A database where each "file" is an element of $\mathbb{N}$ (e.g., a $\log N$-bit string).

A hash function maps a file to a much smaller domain, i.e., $h: \mathbb{N} \rightarrow [b]$, $b \ll \mathbb{N}$.

If $b \ll \mathbb{N}$, a lot of "collision" must occur, i.e., $x, y$ s.t. $h(x) = h(y)$.

To work around this, we use a "family" of hash functions $H = \{ h_1, \ldots, h_t \}, h_i: \mathbb{N} \rightarrow [b]$.

We can hope to design a "nice" family that can differentiate every set of up to $k$ files.

That is, $\forall S \subseteq \mathbb{N}, |S| = k$, $\exists h \in H$ s.t. $h$ is injective on $S$. 

$$(k \leq b)$$
This is called a "k-perfect hash family".

**Question:** How small can \( t = |\mathcal{H}| \) be?

**Upper bound:** \( t = O(k \log(N_k)) \) is possible,
where \( b \geq k^2 \).

**Proof:** Pick each function \( h_i : [N] \rightarrow [b] \)
uniformly randomly and independently.

Fix \( S \subseteq [N], |S| = k \).

\[
\Pr [ h_i \text{ is injective on } S ] = 1 - \frac{b-1}{b} \cdot \frac{b-2}{b} \cdots \frac{b-k+1}{b} \\
\geq \left(1 - \frac{k}{b}\right)^k \\
\geq \left(1 - \frac{1}{k}\right)^k \\
\geq \frac{1}{4}.
\]

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

Union bound on all \( S \):

\[
\Pr (\text{family is not perfect}) \leq \binom{N}{k} (\frac{3}{4})^t \\
\leq \left(\frac{Ne}{k}\right)^k (\frac{3}{4})^t = 2^O(k \log(N_k)) - \Omega(t) \\
< 1 \text{ for some } t = O(k \log \frac{N}{k}). \quad \square
\]
Lower bounds:

* \( t \geq \frac{\log N}{\log b} \) (\( \forall k \geq 2 \))

Proof: For every \( x_1, x_2 \in \mathbb{N} \), we must have

\( (h_1(x_1), \rightarrow h_t(x_1)) \neq (h_1(x_2), \rightarrow h_t(x_2)) \).

By the pigeonhole principle,

\[ N \leq b^t \Rightarrow t \geq \frac{\log N}{\log b} \]

* Stronger lower bound via graph entropy:

Theorem: \( t \geq \frac{b^{k-1}}{\log t} \left( \log (N-k+2) \right) \left( \log \left( \frac{N-k+2}{b-k+2} \right) \right) \).

(Fredman, Komlós '84)

Proof (assuming \( b | N \)).

Define \( G = (V, E) \).

\[ V = \left\{ (D, x) : D \subseteq \mathbb{N}, |D| = k-2, x \in \mathbb{N} \setminus D \right\} \]

\[ E = \left\{ \forall (D, x_1), (D, x_2) : \forall D, x_1 \neq x_2 \right\} \]

\( \Rightarrow \) a clique of size \( N-k+2 \) for each \( D \).
\[ H(G) = H(\text{each component}) = \log(N-k+2) \]

Now we construct \( \{G_h\}, h \in \mathcal{H}, \text{ s.t. } G = \bigcup_{h \in \mathcal{H}} G_h \). \[ V(G_h) = V(G) \]

\[ E(G_h) = \left\{ (\mathcal{D}, x_1), (\mathcal{D}, x_2) : h \text{ is injective on } \mathcal{D} \cup \{x_1, x_2\} \right\} \]

[For every \((\mathcal{D}, x_1, x_2)\), \( \exists h \in \mathcal{H} \) injective on \( \mathcal{D} \cup \{x_1, x_2\} \)]

\[ G = \bigcup_{h \in \mathcal{H}} G_h. \]

All we need to do is to upper bound \( H(G_h) \).

* If \( h \) is not injective on \( \mathcal{D} \)

\[ G_h = \text{empty} \Rightarrow H(G_h) = 0. \]

* If not, \[ G_h \text{ is } (b-k+2) \text{-partite with parts } \]

\[ A_i : = \{ (\mathcal{D}, x) : h(x) = i \} \]

\((i \notin h(\mathcal{D}))\) subadditivity of \( \sum H(G) \)

\[ \Rightarrow H(G_h) \leq \log(b-k+2) \Rightarrow t > \frac{\log(N-k+2)}{\log(b-k+2)} \]
Improving the bound:

Observation: each $G_h$ has a large number of isolated vertices of $G_h$

$p := \Pr(\text{uniformly random vertex is isolated})$

Disjoint union $\Rightarrow H(G) = p H(E) + (1-p) H(D_1 \cup D_2 \cup \cdots)$

$\leq (1-p) \log (b-k+2)$.

Bounding $p$:

$(S, x)$ is isolated iff $h$ is not injective on $D \cup \{x\}$

$\Rightarrow p = \Pr_S(h \text{ is not injective on } S)$

Claim: To lower bound $p$, wlog we can assume $|h^{-1}(1)| = |h^{-1}(2)| = \ldots = |h^{-1}(b)| = \frac{N}{b}$. 
Proof. Suppose \(|h^{-1}(1)| > |h^{-1}(2)|\), and \(h(x) = 1\). Show that by making \(h(x) = 2\), \(P\) only gets larger.

\[
Pr\left(h \text{ injective on } S \right) = Pr(x \in S) Pr(h \text{ inj. on } S | x \in S) + Pr(x \notin S) Pr(h \text{ inj. on } S | x \notin S)
\]

Can only decrease \(\left|h^{-1}(h(x))\right|\) by making \(h(x) = 2\), remain unchanged.

\[
1 - p = Pr\left(h \text{ injective on } S \right) = 1 - \frac{b-1}{b} = \frac{b-2}{b} = \frac{b-k+2}{b}
\]

\[
\Rightarrow \frac{H(G)}{\max_H(G_h)} \geq \frac{\log(N-k+2)}{(1-p) \log(b-k+2)}
\]

\[
= \frac{b^{k-1}}{b(b-1) - (b-k+2)} \cdot \frac{\log(N-k+2)}{\log(b-k+2)}
\]

\[
\square
\]