1

Thu 3/21

Last time: Graph Entropy

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

Subadditivity: $H(G_1 \cup G_2) \leq H(G_1) + H(G_2)$

Monotonicity: $H(G_1) \leq H(G_1 \cup G_2)$

Disjoint union: $G_1, \dots, G_k$ connected components of G,

$$H(G) = \sum_{i \in [k]} p_i \, H(G_i) \quad , \quad p_i \triangleq \frac{|V(G_i)|}{|V(G)|}.$$

Today: Applications in Combinatorics.
(for lower bounds)

[Intuition: A combinatorial process that ~~ex~~ 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?

②

* Construction (upper bound):

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

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 $G_1, \dots, G_\ell$ be the bipartite graphs.

$$H(G_1 \cup \dots \cup G_\ell) = H(K_n) = \log n$$

But
$$\begin{cases} H(G_i) \leq 1 \\ H(G_1 \cup \dots \cup G_\ell) \leq \sum_{i \in [\ell]} H(G_i) \leq \ell. \end{cases}$$

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

$\square$

[similar bound can be shown by a chromatic number argument].

Perfect hash function family:

Setting: A database where each "file" is an element of $[N]$ (eg, a $\lfloor \log N \rfloor$-bit string).

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

~~Obviously cannot existing~~
[Instead of storing each file $x \in [N]$, an indexing database can store the hash $h(x)$.]

In order to search for x in the database, search for $h(x)$ in the hash table.

If $b \ll n$, 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 $\mathcal{H} = \{ h_1, \ldots, h_t \}$, $h_i : [N] \to [b]$.

We can hope to design a "nice" family that can differentiate every set of up to k files. $\boxed{(k \leq b)}$

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

- This is called a "k-perfect hash family".
- <u>Question:</u> How small can $t = |\mathcal{H}|$ be?
- Upper bound: $t = O\left(k \log(N/k)\right)$ is possible, where $b \geq k^2$.

<u>Proof</u> Pick each function $h_i : [N] \to [b]$ uniformly randomly and independently.

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

$$\Pr\left[h_1 \text{ is injective on } S\right] = 1 \cdot \frac{b-1}{b} \cdot \frac{b-2}{b} \cdots \frac{b-k+1}{b}$$

$$\geq \left(1 - k/b\right)^k \geq \left(1 - 1/k\right)^k \geq 1/4.$$

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

Union bound on all S:

$$\Pr\left(\text{family is not perfect}\right) \leq \binom{N}{k}\left(3/4\right)^t$$

$$\leq \left(\frac{Ne}{k}\right)^k \left(3/4\right)^t = 2^{O\left(k \log\left(\frac{N}{k}\right)\right) - \Omega(t)}$$

$$< 1 \quad \text{for some } t = O\left(k \log N/k\right).$$

$\square$

Lower bounds:

* $t \geq \log N / \log b \quad (\forall k \geq 2)$

Proof: For every distinct $x_1, x_2 \in [N]$, we must have

$$(h_1(x_1), \dots, h_t(x_1)) \neq (h_1(x_2), \dots, h_t(x_2)).$$

By the pigeonhole principle,

$$N \leq b^t \Rightarrow t \geq \log N / \log b. \qquad \square$$

* Stronger lower bound via graph entropy:

Theorem: (Fredman, Komlós '84)

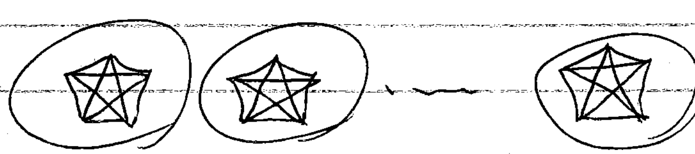
$$t \geq \frac{b^{k-1}}{b(b-1)\cdots(b-k+2)} \cdot \frac{\log(N-k+2)}{\log(b-k+2)}.$$

Proof (assuming $b \mid N$).

Define $G = (V, E)$.

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

$$E = \{ (D, x_1), (D, x_2) : \forall D, x_1 \neq x_2 \}$$



a clique of size $N-k+2$ for each D.

$\Rightarrow H(G) = H(\text{each component}) = \log(N-k+2)$.

Now we construct $\{G_h\}$, $h \in \mathcal{H}$, s.t. $G = \bigcup_{h \in \mathcal{H}} G_h$.

$$\begin{cases} V(G_h) = V(G) \\ E(G_h) = \{ (\mathcal{D}, x_1), (\mathcal{D}, x_2) : h \text{ is injective on } \mathcal{D} \cup \{x_1, x_2\} \} \end{cases}$$

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

$$\Downarrow$$

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

all we need to do is to upper bound $H(G_h)$.

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

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

* If not, ~~[G_h is k-partite] since injectivity information only depends on~~

G_h is $(b-k+2)$-partite with parts

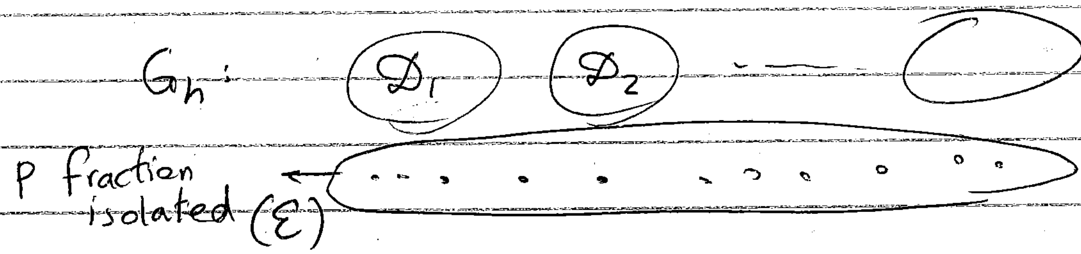
$A_i := \{ (\mathcal{D}, x) : h(x) = i \}$
$(i \notin h(\mathcal{D}))$

$$\Rightarrow H(G_h) \leq \log(b-k+2) \xRightarrow{\text{subadditivity of } H(G)} t \geq \frac{\log(N-k+2)}{\log(b-k+2)}.$$

Improving the bound:

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

$p :=$ Pr(uniformly random vertex of G_h is isolated)



Disjoint union $\Rightarrow H(G_h) = \underbrace{p H(\mathcal{E})}_{0} + (1-p) H(\mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots)$

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

Bounding p:

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

$$\Rightarrow p = \Pr_{\substack{S \\ |S|=k-1}} (\ h \text{ is not injective on } S\ )$$

Claim: To lower bound p, wlog we can assume $|h^{-1}(1)| = |h^{-1}(2)| = \cdots = |h^{-1}(b)| = N/b$.

8

(of Claim)
<u>Proof</u>: Suppose $|h^{-1}(1)| > |h^{-1}(2)|$,

and $h(x)=1$. Show that by making $h(x)=2$,

P only gets larger.

can only ~~increase~~ decreasing by $|h^{-1}(h(x))|$.

$$\Pr_S(h \text{ injective on } S) = \Pr(x \in S)\Pr(h \text{ inj. on } S \mid x \in S) + \Pr(x \notin S)\Pr(\quad " \quad \mid x \notin S)$$

remain unchanged.

$\square$

Now, $1-p = \Pr_S(h \text{ injective on } S) = 1 \cdot \frac{b-1}{b} \frac{b-2}{b} \cdots \frac{b-k+2}{b}$

~~$\Rightarrow H(G_h) \leq H$~~

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

$$= \frac{b^{k-1}}{b(b-1)\cdots(b-k+2)} \frac{\log(N-k+2)}{\log(b-k+2)}.$$

$\square$