Probability & Computing

Bounded Differences & Geometric Inhomogeneous Random Graphs
Recall: Concentration

Concentration Inequalities

- Bound the probability for a random variable to deviate from its expectation
- Markov: generally applicable, but not very strong
- Chebychev: stronger, but requires knowledge about variance
- Chernoff: even stronger, but requires knowledge about moment generating functions
  (simpler variants work, e.g., for sums of independent random variables)

Example

Today: similarly strong but beyond sums of independent Bernoulli random variables

- \( k \) balls distributed uniformly at random over \( n \) bins
- Random variable \( X \) counts empty bins
- Let \( X_i = 1 \{ \text{Bin } i \text{ is empty} \} \) for \( i \in [n] \) \( \Rightarrow \) \( X = \sum_{i=1}^{n} X_i \)
- Concentration: \( \Pr[X \geq \mathbb{E}[X] + 5\sqrt{k}] \)
  - Markov: \( \Pr[X \geq \mathbb{E}[X] + 5\sqrt{k}] \leq \frac{\mathbb{E}[X]}{\mathbb{E}[X]+5\sqrt{k}} = 1 - \frac{5\sqrt{k}}{\mathbb{E}[X]+5\sqrt{k}} \quad \xrightarrow{n \to \infty} \quad 1 \quad \checkmark \)
  - Chebychev: tedious...
  - Chernoff: \( \times \) (our Bernoulli random variables are not independent)
Recall: Concentration

Concentration Inequalities
- Bound the probability for a random variable to deviate from its expectation
- Markov: generally applicable, but not very strong
- Chebychev: stronger, but requires knowledge about variance
- Chernoff: even stronger, but requires knowledge about moment generating functions (simpler variants work, e.g., for sums of independent random variables)

Example
Today: similarly strong but beyond sums of independent Bernoulli random variables
- $k$ balls distributed uniformly at random over $n$ bins
- Random variable $X$ counts empty bins
- Let independent $Y_j \sim \mathcal{U}([n])$ for $j \in [k]$ denote the bin of the $j$-th ball
  \[ X = f(Y_1, \ldots, Y_k) = \sum_{i \in [n]} 1_{\{\#j: Y_j = i\}} \] (summands not independent, but the $Y_j$ are)
  \[ = \sum_{i \in [n]} \max_{j \in [k]} \left\{2 - |\{Y_j, i\}|\right\} \] (“not” a sum Bernoulli random variables)

Can we show concentration for some arbitrary function of independent random variables?
... under certain conditions!
Method of Bounded Differences

Aka ... Bounded differences inequality, McDiarmid’s inequality, Azuma-Hoeffding inequality

Idea If changing one of the random inputs of $f(X_1, ..., X_k)$ does not change $f(\cdot)$ much
then a lot has to go wrong for $f(\cdot)$ to deviate from its expected value

Definition: A function $f: S^n \to \mathbb{R}$ satisfies the bounded differences condition ("Lipschitz condition") with parameters $\Delta_i$, if $|f(X_1, ..., X_i, ..., X_n) - f(X_1, ..., X_i', ..., X_n)| \leq \Delta_i$ for all $i \in [n]$ and $X_i, X_i' \in S$.

Theorem: Let $X_1, ..., X_n$ be independent random variables taking values in a set $S$. Let $f: S^n \to \mathbb{R}$ satisfy the bounded differences condition with parameters $\Delta_i$. Then, for $\Delta = \sum_{i \in [n]} \Delta_i^2$:

$$\Pr[|f - \mathbb{E}[f]| \geq t] \leq 2e^{-2t^2/\Delta}.$$  (write $f$ for $f(X_1, ..., X_n)$)

Lemma: $\Pr[f \geq \mathbb{E}[f] + t] \leq e^{-2t^2/\Delta}$. Also for $\Pr[f \leq \mathbb{E}[f] - t]$

Cor. $\mathbb{E}[f] \leq g(n): \Pr[f \geq cg(n)] \leq e^{-2((c-1)g(n))^2/\Delta}$.
Application: Balls into Bins

- \( k \) balls distributed uniformly at random over \( n \) bins
- Random variable \( X \) counts empty bins
- Let independent \( Y_j \sim \mathcal{U}([n]) \) for \( j \in [k] \) denote the bin of the \( j \)-th ball, and \( X = f(Y_1, \ldots, Y_k) \)

Bounded differences condition

Intuition: How much can the number of empty bins change if we move a ball from one bin to another?
- A ball is moved from an almost empty bin to...
  - ... an empty bin \( \Rightarrow +1 - 1 \Rightarrow \Delta_i = 0 \)
  - ... a non-empty bin \( \Rightarrow +1 \Rightarrow \Delta_i = 1 \)
- A ball is moved from a not almost empty bin to...
  - ... an empty bin \( \Rightarrow -1 \Rightarrow \Delta_i = 1 \)
  - ... a non-empty bin \( \Rightarrow \Delta_i = 0 \)

Concentration via bounded differences

\[
\Delta = \sum_{i=1}^{k} \Delta_i^2 \leq \sum_{i=1}^{k} 1^2 = k \Rightarrow \Pr[f \geq \mathbb{E}[f] + 5\sqrt{k}] \leq e^{-2(5\sqrt{k})^2/k} = e^{-50}
\]

Much better than Markov’s → 1
Application: The Factory

- Products are distributed uniformly at random over boxes on a conveyor belt
  - \( n \) products \( \frac{m}{n/k} \) boxes \( k = \log \log(n) \)
  
- A camera scans \( k + 1 \) consecutive boxes simultaneously

- Problem: Empty box in view \( \Rightarrow \) reflection blinds camera, products remain unscanned

- Question: How many products avoid quality assurance?  
  \[ \text{Show: } o(n) \text{ with prob. } 1 - O\left(\frac{1}{n}\right) \]

Formalize

- \textit{chain}: consecutive sequence of non-empty boxes
- \textit{short chain}: incl. max. chain of length \( \leq k \) \( \Rightarrow \) \textit{exactly} products in short chains unscanned

- \( X_i = \) number of products in box \( i \), \( Y_i = \) indicator whether box \( i \) is in a short chain

- Then \( X = \sum_{i=1}^{m} X_i \cdot Y_i \) is the number of unscanned products

- Problem: Dependencies (between \( X_i \)’s, between \( X_i \) and \( Y_i \))

- Solution: Relax dependencies and compute upper bound instead
Application: The Factory

- Products are distributed uniformly at random over boxes on a conveyor belt
  - \( n \) products, \( m = n/k \) boxes, \( k = \log \log(n) \)
  - \( E_k(i) = 1 \)

- A camera scans \( k + 1 \) consecutive boxes simultaneously

- Problem: Empty box in view \( \Rightarrow \) reflection blinds camera, products remain unscanned

- Question: How many products avoid quality assurance?  
  \( \text{Show: } o(n) \text{ with prob. } 1 - O\left(\frac{1}{n}\right) \)

Relax and bound

- \( X_i = \) number of products in box \( i \), \( Y_i = \) indicator whether box \( i \) is in a short chain

- Then \( X = \sum_{i=1}^{m} X_i \cdot Y_i \) is the number of unscanned products

- \( E_k(i) = \) number of empty boxes in box \( i \) and \( k \) closest (assuming \( k \) even)

| Box \( i \) in short chain \( \Rightarrow \) \( E_k(i) > 0 \) |
| \( Y_i' = \) indicator whether \( E_k(i) > 0 \Rightarrow Y_i \leq Y_i' \) |

| \( X = \sum_{i=1}^{m} X_i \cdot Y_i \leq \sum_{i=1}^{m} X_i \cdot Y_i' =: X' \) |
Expectation of $X'$ (for $n$ large enough)

\[
\mathbb{E}[X'] = \sum_{i=1}^{m} \mathbb{E}[X_i \cdot Y_i']
\]

(law of total expectation)

\[
= \sum_{\ell=0}^{k+1} \mathbb{E}[X_i \cdot Y_i' \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=0}^{k} \mathbb{E}[X_i \cdot Y_i' \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=1}^{k} \mathbb{E}[X_i \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

Expected number of products in box $i$, knowing that exactly $\ell$ boxes are empty

- Box $i$ empty? $\Rightarrow X_i = 0$
- Else: $n$ products distributed u.a.r. over $m' = m - \ell$ boxes

\[
\mathbb{E}[X_i \mid E_k(i) = \ell] = \frac{n}{m'} \leq 2 \log \log(n)
\]

\[
m' \geq \frac{n}{\log \log(n)} - \log \log(n)
\]

(for $n$ large enough)

- $n$ products
- $m = n/k$ boxes, $k = \log \log(n)$
- $X' = \sum X_i \cdot Y_i'$
- $X_i$, products in box $i$
- $E_k(i)$, number empty boxes in box $i$ and $k$ closest
- $Y_i'$, indicator $E_k(i) > 0$
Expectation of $X'$ (for $n$ large enough)

\[
\mathbb{E}[X'] = \sum_{i=1}^{m} \mathbb{E}[X_i \cdot Y'_i]
\]

(law of total expectation)

\[
= \sum_{\ell=0}^{k+1} \mathbb{E}[X_i \cdot Y'_i \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=0}^{k} \mathbb{E}[X_i \cdot Y'_i \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=1}^{k} \mathbb{E}[X_i \mid E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
\leq \sum_{\ell=1}^{k} 2 \log \log(n) \cdot \Pr[E_k(i) = \ell]
\]

\[
= 2 \log \log(n) \sum_{\ell=1}^{k} \Pr[E_k(i) = \ell]
\]

\[
\leq \Pr[\text{“Exists an empty box among } k + 1 \text{”}]
\]

\[
\leq (k + 1) \cdot \Pr[\text{“A given box is empty”}]
\]

\[
\leq 2k \left(1 - \frac{1}{m}\right)^n \leq 2k \cdot e^{-k} = 2 \frac{\log \log(n)}{\log(n)}
\]

- $n$ products
- $m = n/k$ boxes, $k = \log \log(n)$
- $X' = \sum X_i \cdot Y'_i$
- $X_i$, products in box $i$
- $E_k(i)$, number empty boxes in box $i$ and $k$ closest
- $Y'_i$, indicator $E_k(i) > 0$
Expectation of $X'$ (for $n$ large enough)

\[
\mathbb{E}[X'] = \sum_{i=1}^{m} \mathbb{E}[X_i \cdot Y_i']
\]

(law of total expectation)

\[
= \sum_{\ell=0}^{k+1} \mathbb{E}[X_i \cdot Y_i' | E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=0}^{k} \mathbb{E}[X_i \cdot Y_i' | E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
= \sum_{\ell=1}^{k} \mathbb{E}[X_i | E_k(i) = \ell] \cdot \Pr[E_k(i) = \ell]
\]

\[
\leq \sum_{\ell=1}^{k} 2 \log \log(n) \cdot \Pr[E_k(i) = \ell]
\]

\[
= 2 \log \log(n) \sum_{\ell=1}^{k} \Pr[E_k(i) = \ell]
\]

\[
\leq 2 \log \log(n) \cdot 2 \frac{\log \log(n)}{\log(n)}
\]

\[
= 4 \frac{\log \log(n)^2}{\log(n)}
\]

\[
= \sum_{i=1}^{m} 4 \frac{\log \log(n)^2}{\log(n)} = m \cdot 4 \frac{\log \log(n)^2}{\log(n)} = \frac{n}{\log \log(n)} \cdot 4 \frac{\log \log(n)^2}{\log(n)} = n \cdot 4 \frac{\log \log(n)^2}{\log(n)} = o(n)
\]
Concentration of $X$ (for $n$ large enough)

**Bounded Differences**
- View $X$ as a function $f(Z_1, \ldots, Z_n)$ of independent rand. var.
  where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product
- Bounded differences condition:
  - Worst change in number of products in short chains when moving a single product from one box to another
  - Consider chain of $2k + 1$ boxes containing all $n$ products and one box contains only one of them

\[ |f(\ldots, Z_j, \ldots) - f(\ldots, Z'_j, \ldots)| \leq \Delta_j \text{ for all } j \text{ and } Z_j, Z'_j \]

\[ \Rightarrow X = 0, \text{ since no short chain and, thus, no products in short chains} \]

\[ \Rightarrow X = n, \text{ since all products in short chains now} \]

- $n$ products
- $m = n/k$ boxes, $k = \log \log(n)$
- $X_j = \sum X_i \cdot Y_i$
- $X_i$, products in box $i$
- $Y_i$, indicator $i$ in short chain

\[ \mathbb{E}[X] \leq \mathbb{E}[X'] \leq 4n \frac{\log \log(n)}{\log(n)} \]
Concentration of $X$  
(for $n$ large enough)

**Bounded Differences**
- View $X$ as a function $f(Z_1, \ldots, Z_n)$ of independent rand. var.
  where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product
- Bounded differences condition: $\Delta_j \leq n$
- Bounded differences inequality:
  \[ \Delta = \sum_{j=1}^{n} \Delta_j^2 \leq \sum_{j=1}^{n} n^2 = n^3 \quad g(n) = 4n^{\frac{\log \log(n)}{\log(n)}} \]
  \[
  \Pr \left[ X \geq c4n^{\frac{\log \log(n)}{\log(n)}} \right] \leq \exp \left( -\frac{2(c - 1)^2 \left( 4n^{\frac{\log \log(n)}{\log(n)}} \right)^2}{n^3} \right) \\
  \text{This bound is useless, since worst-case changes are too big} \\
  = \exp \left( -\Theta \left( \frac{\log \log(n)^2}{n \log(n)^2} \right) \right) \rightarrow 1 \quad n \to \infty
  \]
  
  But this case (all products in few boxes) is super unlikely...

- $n$ products
- $m = n/k$ boxes, $k = \log \log(n)$
- $X = \sum X_i \cdot Y_i$
- $X_i$, products in box $i$
- $Y_i$, indicator $i$ in short chain
- $\mathbb{E}[X] \leq \mathbb{E}[X'] \leq 4n^{\frac{\log \log(n)}{\log(n)}}$

| $f(\ldots, Z_j, \ldots) - f(\ldots, Z'_j, \ldots)$ | $\leq \Delta_j$ for all $j$ and $Z_j, Z'_j$ |

Function $f(Z_1, \ldots, Z_n)$:
- $Z_1, \ldots, Z_n$ independent
- bounded differences $\Delta_j$
- $\Delta = \sum_{j=1}^{n} \Delta_j^2$
- $g(n) \geq \mathbb{E}[f]$
- $\Pr[f \geq cg(n)] \leq e^{-2((c-1)g(n))^2/\Delta}$. 

---

7  Maximilian Katzmann, Stefan Walzer – Probability & Computing  
Institute of Theoretical Informatics, Algorithm Engineering & Scalable Algorithms
Method of Typical Bounded Differences

**Definition:** A function $f : S^n \to \mathbb{R}$ satisfies the **typical bounded differences condition** with respect to
- an event $A \subseteq S^n$ and
- parameters $\Delta_i^A \leq \Delta_i$ for $i \in [n],$

if $|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X'_i, \ldots, X_n)| \leq \begin{cases} \Delta_i^A, & \text{if } (X_1, \ldots, X_i, \ldots, X_n) \in A, \\ \Delta_i, & \text{otherwise} \end{cases}$

for all $i \in [n]$ and $X_i, X'_i \in S.$

- $\Delta_i^A$ is worst-case change, assuming $A$ held before the change

**Theorem:** Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $S$, let $A \subseteq S^n$ be an event, and let $f : S^n \to \mathbb{R}$ satisfy the typical bounded differences condition w.r.t. $A$ and parameters $\Delta_i^A \leq \Delta_i.$ Then, for $g(n) \geq \mathbb{E}[f]$, for all $\varepsilon_i \in (0, 1]$ and $\Delta = \sum_{i \in [n]} (\Delta_i^A + \varepsilon_i (\Delta_i - \Delta_i^A))^2,$ $\Pr[f \geq cg(n)] \leq e^{-((c-1)g(n))^2/(2\Delta)} + \Pr[\neg A] \sum_{i \in [n]} \frac{1}{\varepsilon_i}.$

Method of Typical Bounded Differences

**Theorem:** Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $S$, let $A \subseteq S^n$ be an event, and let $f : S^n \to \mathbb{R}$ satisfy the typical bounded differences condition w.r.t. $A$ and parameters $\Delta_i^A \leq \Delta_i$. Then, for $g(n) \geq \mathbb{E}[f]$, for all $\epsilon_i \in (0, 1]$ and $\Delta = \sum_{i \in [n]} (\Delta_i^A + \epsilon_i (\Delta_i - \Delta_i^A))^2$: $\Pr[f \geq cg(n)] \leq e^{-((c-1)g(n))^2/(2\Delta)} + \Pr[\neg A] \sum_{i \in [n]} \frac{1}{\epsilon_i}$.

- Function of independent random variables as before
- $A$ is the good, typical event that should be very likely to occur
- $\Delta$ is sum of squared worst-case changes as before
  - We still consider general worst-case changes as before
  - But we can use the $\epsilon_i$ to mitigate the worst-case effects
  - And focus on the worst-case changes, assuming $A$ held before the change
- But we have to pay for the mitigation!
  - With the probability that the good event $A$ does not occur
  - Multiplied with the inverse mitigators

The more we need to mitigate, the higher the price!
Not too bad if $A$ is very likely to occur!
Application: The Factory (2nd Try)

- View $X$ as a function $f(Z_1, \ldots, Z_n)$ of independent rand. var. where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product
- Bounded differences condition: $\Delta_j \leq n$
  - When all $n$ products fall into $2k + 1 = O(\log \log(n))$ boxes
  - But expected number of products in a single box $i$:
    $$E[B_i] = \frac{n}{m} = \frac{n}{\log \log(n)} = \log \log(n)$$
  - And, thus, expected number in sequence of $2k + 1$ boxes
    $$E[S] = \sum_{i=1}^{2k+1} E[B_i] = O(\log \log(n)^2) \leq \delta \log(n) =: g(n) \text{ (for any } \delta > 0 \text{ and sufficiently large } n)$$
- So typically a sequence should contain way fewer than $n$ products
- Typical event $A = \{\text{“Every sequence of } 2k + 1 \text{ boxes contains } O(\log(n)) \text{ products”}\}$
- See $S$ as sum of independent Bernoulli rand. var. (whether $j$-th product is in sequence)
- Chernoff: For $g(n) \geq E[S]$: $Pr[S \geq (1 + \varepsilon)g(n)] \leq e^{-\varepsilon^2/3 \cdot g(n)} = e^{-\varepsilon^2/3 \cdot \delta \log(n)} = n^{-\delta \varepsilon^2/3}$
- Union bound over $\leq n$ sequences: $Pr[\neg A] \leq n^{-\delta \varepsilon^2/3 + 1} \leq n^{-\lambda}$ (for arbitrarily large $\lambda$)
Application: The Factory (2nd Try)

- View $X$ as a function $f(Z_1, \ldots, Z_n)$ of independent rand. var. where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product.

- Bounded differences condition: $\Delta_j \leq n$

- Typical event $A = \{"Every sequence of $2k+1$ boxes contains $O(\log(n))$ products"\}, $\Pr[\neg A] \leq n^{-\lambda}$ (for arbitrary $\lambda$).

- Typical bounded differences condition:
  - Worst change in $f$ when moving a product from one box to another, assuming $A$ held before the move.
  - Moving one product empties at most one box $\Rightarrow$ at most two new short chains.
  - Assuming $A$, these short chains combined contain $O(\log(n))$ products $\Rightarrow \Delta_j^A = O(\log(n))$.

$n$ products
$m = n/k$ boxes, $k = \log \log(n)$
$X = \sum X_i \cdot Y_i$
$X_i$, products in box $i$
$Y_i$, indicator $i$ in short chain

$\mathbb{E}[X] \leq \mathbb{E}[X'] \leq 4n \frac{\log \log(n)}{\log(n)}$
Application: The Factory (2nd Try)

- View $X$ as a function $f(Z_1, \ldots, Z_n)$ of independent random variables, where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product.
- Bounded differences condition: $\Delta_j \leq n$
- Typical event $A = \{\text{"Every sequence of } 2k+1 \text{ boxes contains } O(\log(n)) \text{ products"}\}$, $\Pr[\neg A] \leq n^{-\lambda}$ (for arbitrary $\lambda$)
- Typical bounded differences condition: $\Delta^A_j = O(\log(n))$
- Typical bounded differences inequality:
  $$\Delta = \sum_{j=1}^{n} (\Delta^A_j + \epsilon_j (\Delta_j - \Delta^A_j))^2 \leq \sum_{j=1}^{n} (\Delta^A_j + \epsilon_j \Delta_j)^2$$
  Mitigators, arbitrary $\epsilon_j \in (0, 1]$!
  $$\leq \sum_{j=1}^{n} (O(\log(n)) + \epsilon_j n)^2$$
  $$= \sum_{j=1}^{n} (O(\log(n)) + 1)^2$$
  $$= O(n \log(n)^2) \quad \text{Much better than } n^3 \text{ from before!}$$

- $n$ products
- $m = n/k$ boxes, $k = \log \log(n)$
- $X = \sum X_i \cdot Y_i$
- $X_i$, products in box $i$
- $Y_i$, indicator $i$ in short chain

$$\mathbb{E}[X] \leq \mathbb{E}[X'] \leq 4n \frac{\log \log(n)}{\log(n)}$$

Function $f(Z_1, \ldots, Z_n)$:
- $Z_1, \ldots, Z_n$ independent
- typical event $A$
- bounded differences $\Delta^A_j \leq \Delta_j$

$$\Delta = \sum_{j=1}^{n} (\Delta^A_j + \epsilon_j (\Delta_j - \Delta^A_j))^2$$

$$g(n) \geq \mathbb{E}[f]$$

$$\Pr[f \geq cg(n)] \leq e^{-(c-1)g(n)^2/(2\Delta)} + \Pr[\neg A] \sum_{j=1}^{n} \frac{1}{\epsilon_j}$$
Application: The Factory (2nd Try)

- View $X$ as a function $f(Z_1,\ldots,Z_n)$ of independent rand. var.
  where $Z_j$ for $j \in [n]$ denotes the box of the $j$-th product
- Bounded differences condition: $\Delta_j \leq n$
- Typical event $A = \{"Every sequence of $2k+1$ boxes contains $O(\log(n))$ products"\}, $\Pr[\neg A] \leq n^{-\lambda}$ (for arbitrary $\lambda$)
- Typical bounded differences condition: $\Delta^A_j = O(\log(n))$
- Typical bounded differences inequality:
  \[
  \Delta = O(n \log(n)^2) \quad g(n) = 4n \frac{\log \log(n)}{\log(n)} \quad \varepsilon_j = \frac{1}{n}
  \]

  $\Pr[X \geq c4n \frac{\log \log(n)}{\log(n)}] \leq \exp\left(-\Omega\left(n \frac{\log \log(n)^2}{\log(n)^2}\right)\right)$

  $= O(1/n)$

  $\Pr[\neg A] \sum_{j=1}^{n} \frac{1}{\varepsilon_j} \leq n^{-\lambda} \cdot n^2 = O(1/n)$ for $\lambda = 3$

  $\Pr[f \geq cg(n)] \leq e^{-(c-1)g(n)^2/(2\Delta)}$
Geometric Inhomogeneous Random Graphs

Motivation
- Average-case analysis: analyze models that represent the real world
- Models seen so far
  - Erdős-Rényi random graphs: simple but no locality
  - Random geometric graphs: locality but no heterogeneity (all vertices roughly same degree)

Not realistic: celebrities are very-high-degree vertices in social networks

Realistic representation: power-law distribution

Pareto distribution: \( X \sim \text{Par}(\alpha, x_{\text{min}}) \)

\[
f_X(x) = \begin{cases} \alpha x_{\text{min}} \cdot x^{-(\alpha+1)}, & \text{if } x \geq x_{\text{min}} \\ 0, & \text{otherwise} \end{cases}
\]

Idea
- Add Pareto distribution to RGGs


YouTube

konect.cc/plot/degree.a.youtube-links.full.png

You Tube

(most vertices small degree)

(few vertices very high degree)
Geometric Inhomogeneous Random Graphs

Definition
- Consider \( n \) vertices
- For each vertex \( v \) independently:
  - Draw a position \( x_v \) uniformly on \( \mathbb{T}^d \)
  - Draw a weight \( w_v \) from \( \text{Par}(\tau - 1, 1) \) for \( \tau \in (2, 3) \implies f_{w_v}(w) = (\tau - 1)w^{-\tau} \)
- Connect \( u \) and \( v \) with an edge, iff
  \[
  \text{dist}(x_u, x_v) \leq \left( \frac{\lambda w_u \cdot w_v}{n} \right)^{1/d} \leq \text{\text{L}_\infty\text{-norm}} \implies \text{const. controls the avg. degree}
  \]
- For \( d = 1 \), linear relation between distance and weight
  \[
  x \leq \lambda \frac{w_v \cdot y}{n} \iff y \geq \frac{n}{\lambda w_v} x
  \]

“Power-Law Exponent”
Geometric Inhomogeneous Random Graphs

**Definition**
- Consider $n$ vertices
- For each vertex $v$ independently:
  - Draw a position $x_v$ uniformly on $\mathbb{T}^d$
  - Draw a weight $w_v$ from $\text{Par}(\tau - 1, 1)$ for $\tau \in (2, 3) \Rightarrow f_{w_v}(w) = (\tau - 1)w^{-\tau}$
- Connect $u$ and $v$ with an edge, iff
  $$\text{dist}(x_u, x_v) \leq \left(\frac{\lambda w_u w_v}{n}\right)^{1/d} L_{\infty} \text{-norm}$$
  const. controls the avg. degree
  For $d = 1$, linear relation between distance and weight $y = w_u, x = \text{dist}(x_u, x_v)$
  $$x \leq \lambda \frac{w_v y}{n} \iff y \geq \frac{n}{\lambda w_v} x$$

"Power-Law Exponent"
Geometric Inhomogeneous Random Graphs

Definition

- Consider \( n \) vertices
- For each vertex \( v \) independently:
  - Draw a position \( x_v \) uniformly on \( \mathbb{T}^d \)
  - Draw a weight \( w_v \) from \( \text{Par}(\tau - 1, 1) \) for \( \tau \in (2, 3) \) ⇒ \( f_{w_v}(w) = (\tau - 1)w^{\tau} \)
- Connect \( u \) and \( v \) with an edge, iff
  \[
  \frac{\text{dist}(x_u, x_v)}{n} \leq \left( \frac{w_u \cdot w_v}{n} \right)^{1/d}
  \]
    - \( L_\infty \)-norm const. controls the avg. degree
- For \( d = 1 \), linear relation between distance and weight
  \[
  x = \text{dist}(x_u, x_v), \quad y = w_u, \quad x = \frac{w_v \cdot y}{n}
  \]
    - \( x \leq \lambda \frac{w_v \cdot y}{n} \Leftrightarrow y \geq \frac{n}{\lambda w_v} x \)
- The lower \( w_v \), the steeper the wedge
  - The lower the degree

“Power-Law Exponent”
Expected Degree ($d = 1$)

- Consider vertex $v$ with weight $w_v$
- We want to compute $\mathbb{E}[\text{deg}(v) \mid w_v]$
- Consider $X_u$ for $u \in V \setminus \{v\}$ indicating whether \{u, v\} $\in E$

\[
\text{deg}(v) = \sum_{u \in V \setminus \{v\}} X_u
\]

\[
\mathbb{E}[\text{deg}(v) \mid w_v] = \sum_{u \in V \setminus \{v\}} \mathbb{E}[X_u \mid w_v]
\]

This is not the area of the shape, since weights are not distributed uniformly!

$\Rightarrow$ Use law of total probability to account for that.

GIRG
- $n$ independent vertices
- $x_v \sim \mathcal{U}([0, 1])$
- $w_v \sim \text{Par}(\tau - 1, 1)$ for $\tau \in (2, 3)$
  \[f_{w_v}(w) = (\tau - 1)w^{-\tau}\]
- $u, v$ adjacent iff
  \[\text{dist}(x_u, x_v) \leq \lambda \frac{w_u \cdot w_v}{n}\]
Expected Degree ($d = 1$)

- Consider vertex $v$ with weight $w_v$
- We want to compute $\mathbb{E}[\text{deg}(v) \mid w_v]$
- Consider $X_u$ for $u \in V \setminus \{v\}$ indicating whether $\{u, v\} \in E$

\[ \text{deg}(v) = \sum_{u \in V \setminus \{v\}} X_u \]

\[ \mathbb{E}[\text{deg}(v) \mid w_v] = \sum_{u \in V \setminus \{v\}} \mathbb{E}[X_u \mid w_v] \]

\[ = \Theta(n \Pr[\{u, v\} \in E \mid w_v]) \]

\[ = \Theta(n \int_1^{\infty} \Pr[u \in N(v) \mid w_u = w, w_v] f_{w_u}(w) dw) \]

\[
\begin{align*}
\text{Case 1: } w &\leq \frac{n}{2 \lambda w_v} \Rightarrow \lambda \frac{w \cdot w_v}{n} \leq \frac{1}{2} \Rightarrow 2\lambda \frac{w \cdot w_v}{n} = \Theta(\frac{w \cdot w_v}{n}) \\
\Pr[x_u \in [\frac{1}{2} - \lambda \frac{w \cdot w_v}{n}, \frac{1}{2} + \lambda \frac{w \cdot w_v}{n}]]
\end{align*}
\]
Expected Degree \((d = 1)\)

- Consider vertex \(v\) with weight \(w_v\)
- We want to compute \(\mathbb{E}[\deg(v) \mid w_v]\)
- Consider \(X_u\) for \(u \in V \setminus \{v\}\) indicating whether \(\{u, v\} \in E\)

\[
\deg(v) = \sum_{u \in V \setminus \{v\}} X_u
\]

\[
\mathbb{E}[\deg(v) \mid w_v] = \sum_{u \in V \setminus \{v\}} \mathbb{E}[X_u \mid w_v]
\]

\[
= \Theta(n \Pr[\{u, v\} \in E \mid w_v])
\]

w.l.o.g. \(x_v = \frac{1}{2}\)

\[
= \Theta(n \int_{1}^{\infty} \Pr[u \in N(v) \mid w_u = w, w_v] f_{w_u}(w) \, dw)
\]

\[
= \Pr[x_u \in [\frac{1}{2} - \lambda \frac{w \cdot w_v}{n}, \frac{1}{2} + \lambda \frac{w \cdot w_v}{n}]] = 2\lambda w \cdot w_v = \Theta(\frac{w \cdot w_v}{n}) = 1
\]

Case 1: \(w \leq \frac{n}{2\lambda w_v} \Rightarrow \lambda \frac{w \cdot w_v}{n} \leq \frac{1}{2}\)

Case 2: \(w > \frac{n}{2\lambda w_v} \Rightarrow \lambda \frac{w \cdot w_v}{n} > \frac{1}{2} \Rightarrow 1\)
Expected Degree ($d = 1$)

- Consider vertex $v$ with weight $w_v$
- We want to compute $\mathbb{E}[\deg(v) \mid w_v] = \begin{cases} \Theta(n), & \text{if } w_v \geq \frac{n}{2\lambda} \\ \end{cases}$
- Consider $X_u$ for $u \in V \setminus \{v\}$ indicating whether $\{u, v\} \in E$
- $\deg(v) = \sum_{u \in V \setminus \{v\}} X_u$
- $\mathbb{E}[\deg(v) \mid w_v] = \sum_{u \in V \setminus \{v\}} \mathbb{E}[X_u \mid w_v]$
  - $= \Theta(n \Pr[\{u, v\} \in E \mid w_v])$, w.l.o.g. $x_v = \frac{1}{2}$
  - $= \Theta(n \int_1^\infty \Pr[u \in N(v) \mid w_u = w, w_v] f_{w_u}(w) \, dw)$
  - $= \Theta\left(n \left( \int_{\frac{n}{2\lambda w_v}}^\infty w_{w_v} \cdot f_{w_u}(w) \, dw + \int_{\frac{n}{2\lambda w_v}}^\infty 1 \cdot f_{w_u}(w) \, dw \right) \right)$
  - $= \Theta\left(n \left( \int_{\frac{n}{2\lambda w_v}}^\infty \frac{1}{n} \cdot f_{w_u}(w) \, dw \right) \right)$

If $w_v \geq \frac{n}{2\lambda}$, then $\frac{n}{2\lambda w_v} \leq 1$

$= \Theta(n)$

**GIRG**
- $n$ independent vertices
- $x_v \sim U([0, 1])$
- $w_v \sim \text{Par}(\tau - 1, 1)$ for $\tau \in (2, 3)$
  - $f_{w_v}(w) = (\tau - 1)w^{-\tau}$
- $u, v$ adjacent iff $\text{dist}(x_u, x_v) \leq \lambda \frac{w_u \cdot w_v}{n}$

$= \Theta(\frac{n^{R\infty \frac{1}{n}} \Pr[w_u \geq \frac{n}{2\lambda w_v}])$

$= \Theta(\frac{n}{R\infty \frac{n}{2\lambda w_v}} f_{w_u}(w) \, dw)$

$= \Pr[w_u \geq \frac{n}{2\lambda w_v}]$

$= \Pr[w_u \geq 1] = 1$
Expected Degree (d = 1)

- Consider vertex $v$ with weight $w_v$
- We want to compute $\mathbb{E}[\deg(v) \mid w_v] = \begin{cases} \Theta(n), & \text{if } w_v \geq \frac{n}{2\lambda} \\ \end{cases}$
- Consider $X_u$ for $u \in V \setminus \{v\}$ indicating whether $\{u, v\} \in E$
- $\deg(v) = \sum_{u \in V \setminus \{v\}} X_u$
- $\mathbb{E}[\deg(v) \mid w_v] = \sum_{u \in V \setminus \{v\}} \mathbb{E}[X_u \mid w_v]$
  - $= \Theta(n \Pr[\{u, v\} \in E \mid w_v])$
  - w.l.o.g $x_v = \frac{1}{2}$

If $w_v < \frac{n}{2\lambda}$

- $= \Theta(n \int_1^{\infty} \Pr[u \in N(v) \mid w_u = w, w_v] f_{w_u}(w) dw)$
- $= \Theta\left(n \left(\int_1^{\frac{n}{2\lambda w_v}} \frac{w \cdot w_v}{n} f_{w_u}(w) dw + \Pr[w_u \geq \frac{n}{2\lambda w_v}]\right)\right)$
  - (via CDF of Par)
  - $= \Theta\left(\left(\frac{n}{2\lambda w_v}\right)^{-(\tau - 1)}\right)$
  - $= \Theta\left(\left(\frac{2\lambda w_v}{n}\right)^{\tau - 1}\right)$
  - $< 1$
  - $= O\left(\frac{w_v}{n}\right)$

**GIRG**

- $n$ independent vertices
- $x_v \sim \mathcal{U}([0, 1])$
- $w_v \sim \text{Par}(\tau - 1, 1)$ for $\tau \in (2, 3)$
  - $f_{w_v}(w) = (\tau - 1)w^{-\tau}$
- $u, v$ adjacent iff
  - $\text{dist}(x_u, x_v) \leq \lambda \frac{w_u \cdot w_v}{n}$
Expected Degree \((d = 1)\)

- Consider vertex \(v\) with weight \(w_v\)
- We want to compute \(E[\text{deg}(v) \mid w_v] = \Theta(n)\), if \(w_v \geq \frac{n}{2\lambda}\)
- We want to compute \(E[\text{deg}(v) \mid w_v]\)

\[
\begin{align*}
\text{deg}(v) &= \sum_{u \in V \setminus \{v\}} X_u \\
E[\text{deg}(v) \mid w_v] &= \sum_{u \in V \setminus \{v\}} E[X_u \mid w_v] \\
&= \Theta(n \text{Pr}[\{u, v\} \in E \mid w_v])
\end{align*}
\]

w.l.o.g \(x_v = \frac{1}{2}\)

If \(w_v < \frac{n}{2\lambda}\)

\[
\begin{align*}
&= \Theta(n \int_1^n \text{Pr}[u \in N(v) \mid w_u = w, w_v] f_{w_u}(w)dw) \\
&= \Theta\left(\left(n \int_1^n \frac{w \cdot w_v}{n} f_{w_u}(w)dw + \text{Pr}[w_u \geq \frac{n}{2\lambda w_v}]\right)\right) \\
&= \Theta\left(n \int_1^n \frac{w \cdot w_v}{n} f_{w_u}(w)dw\right) + O(w_v) \\
&= \Theta\left(n \frac{w_v}{n} \int_1^n \frac{w \cdot w_v}{n} f_{w_u}(w)dw\right) + O(w_v) \\
&= \Theta\left(w_v \int_1^n w^{-(\tau - 1)}dw\right) + O(w_v)
\end{align*}
\]

\[
\begin{align*}
&= \Theta\left(w_v \left[\frac{1}{-(\tau - 2)} w^{-(\tau - 2)}\right]_{\frac{n}{2\lambda w_v}}^n + O(w_v)\right) \\
&= \Theta\left(w_v \left[w^{-(\tau - 2)}\right]_{\frac{n}{2\lambda w_v}}^1 + O(w_v)\right) \\
&= \Theta\left(w_v \left[1 - \left(\frac{n}{2\lambda w_v}\right)^{-(\tau - 2)}\right] + O(w_v)\right) \\
&= \Theta\left(w_v \left[1 - \left(\frac{n}{2\lambda w_v}\right)^{-(\tau - 2)}\right] \right) + O(w_v)
\end{align*}
\]

\(\leq 1\) and \(O(1)\)
Are GIRGs Realistic?

**Structural Properties**
- Heterogeneity: \( \text{deg}(v) \approx w_v, w_v \sim \text{Par}(\tau - 1, 1) \) → power-law degree distribution ✓ (also works with other weight distributions)
- Locality (not seen here) ✓

**Algorithmic Properties**
- Setup: GIRGs with varying degrees of heterogeneity and locality (each dot is a graph)


![Graph](image)

What we considered just now

Basically random geometric graphs

GIRGs without geometry / ER with weights
Are GIRGs Realistic?

**Structural Properties**
- Heterogeneity: \( \text{deg}(v) \approx w_v \), \( w_v \sim \text{Par}(\tau - 1, 1) \) \( \sim \) power-law degree distribution ✓
- Locality (not seen here) ✓

**(also works with other weight distributions)**

**Algorithmic Properties**
- Setup: GIRGs with varying degrees of heterogeneity and locality (each dot is a graph)
- Measure algorithmic properties on GIRGs and real graphs
  - Bidirectional breadth-first-search
  - Diameter computation via BFS
  - Vertex cover kernel size
  - Louvain clustering algorithm
  - Number of maximal cliques
  - Chromatic number kernel size

\[ \text{kernel size } n^x: x = \]

Use GIRGs for average-case analysis!
Vertex Cover Approximation

Vertex Cover
- Given undirected graph $G = (V, E)$ (induced subgraph)
- Find a smallest $S \subseteq V$ such that $G[V \setminus S]$ is edgeless
- NP-complete

Vertex Cover Approximation
- Find a small vertex cover $S'$ fast
- Approximation ratio: $r = |S'|/|S|
- NP-hard to approximate with $r < \sqrt{2}$
- Believed to be NP-hard for $r < 2 - \varepsilon$ for const. $\varepsilon$

Practice
- Simple approximation algorithm repeatedly takes/deletes vertex of largest degree
- Close to optimal ratios on real graphs

"Vertex Cover on Complex Networks", Da Silva, Gimenez-Lugo, Da Silva, IJMPC 2013
Analysis on GIRGs

(based on) "Efficiently Approximating Vertex Cover on Scale-Free Networks with Underlying Hyperbolic Geometry", Bläsius, Friedrich, K., Algorithmica 2023

Keep it simple
- Consider vertices in order of decreasing degree in original graph
- Consider vertices in order of decreasing weight

Learn from the Model
- Once high-degree vertices are taken/removed, remaining vertices have roughly equal weight/degree
- Greedy algorithm picks vertices at random
- Improve quality by solving small separated components exactly
- Two variants
  - Search and solve small components after each greedily taken vertex
  - Take greedy until red line, solve small components exactly, take rest greedy too

This variant yields an upper bound on the quality of the other
Analysis on GIRGs – Approximation Ratio

**Theorem:** Let $G$ be GIRG with $n$ vertices and $m$ edges. Then, an approximate vertex cover $S'$ of $G$ can be computed in time $O(m \log(n))$ such that the approximation ratio is $(1 + o(1))$ asymptotically almost surely.

**Proof**

**Approximation Ratio**

- Differentiate greedily taken vertices $S'_g$ from ones in exactly solved components $S'_e$.
- For each small component, the optimal solution $S$ cannot contain fewer vertices than $S'_e$ does
  \[ |S'_e| \leq |S| \]
  \[ \Rightarrow r = \frac{|S'|}{|S|} = \frac{|S'_e| + |S'_g|}{|S|} \leq \frac{|S| + |S'_g|}{|S|} = 1 + \frac{|S'_g|}{|S|} \]
- $|S| = \Omega(n)$ with prob $1 - o(1)$

"Greed is Good for Deterministic Scale-Free Networks", Chauhan et al. FSTTCS 2016

Remains to show: $|S'_g| = o(n)$
Analysis on GIRGs – Greedy Vertices $\geq t$

Lemma: Let $G$ be a GIRG with $n$ vertices, let $t = \omega(1)$, and let $N_{w \geq t}$ be the number of vertices with weight at least $t$. Then, $N_{w \geq t} = o(n)$ with probability $1 - O(1/n)$.

Proof

- Consider random variable $X_v = 1_{\{w_v \geq t\}}$
- $N_{w \geq t}$ is the sum of independent Bernoulli random variables
- Expectation $\mathbb{E}[N_{w \geq t}] = \sum_{v \in V} \mathbb{E}[X_v] = n \Pr[w_v \geq t]$
  (via CDF of Par) $= nt^{-(\tau - 1)}$
  ($t = \omega(1), \tau \in (2, 3)) = o(n)$
- Since there is a $g(n) \in o(n) \cap \Omega(\log(n))$ with $g(n) \geq \mathbb{E}[N_{w \geq t}]$, Chernoff gives concentration

GIRG
- $n$ independent vertices
- $w_v \sim \text{Par}(\tau - 1, 1)$ for $\tau \in (2, 3)$
Analysis on GIRGs – Greedy Vertices $< t$

- After (the $o(n)$) vertices with weight $\geq t$ are removed, the graph decomposes into several components
  - Components of size $\leq \log \log(n)$ are solved exactly
  - Larger components are assumed to be taken greedily (need to show: these are $o(n)$)
- Hard to determine how likely it is for a vertex to be in a large component
- Make use of geometry! Overestimate components by counting how many vertices are geometrically very close

**Idea**
- Discretize ground space into cells such that edges cannot span empty cells
- Use empty cells as delimiters between components
- Regard chains of non-empty cells as one component
- Count all vertices that are in chains containing $> \log \log(n)$ vertices (also potentially counting small components)
- When does a chain contain too many vertices?
Analysis on GIRGs – Greedy Vertices $< t$

- After (the $o(n)$) vertices with weight $\geq t$ are removed, the graph decomposes into several components
  - Components of size $\leq \log \log(n)$ are solved exactly
  - Larger components are assumed to be taken greedily (need to show: these are $o(n)$)
- Hard to determine how likely it is for a vertex to be in a large component
- Make use of geometry! Overestimate components by counting how many vertices are geometrically very close

**Case 1** Too many cells in long chains, say $> k$ cells

- Unlikely, if cells are small
- Proof via method of bounded differences!
  Total number of cells in long chains does not change much ($\leq 2k + 1$) when one cell moves from empty to non-empty (or vice versa)
- Use Poissonization to get rid of dependencies
Analysis on GIRGs – Greedy Vertices $< t$

- After (the $o(n)$) vertices with weight $\geq t$ are removed, the graph decomposes into several components
  - Components of size $\leq \log \log(n)$ are solved exactly
  - Larger components are assumed to be taken greedily (need to show: these are $o(n)$)
- Hard to determine how likely it is for a vertex to be in a large component
- Make use of geometry! Overestimate components by counting how many vertices are geometrically very close

**Case 2** Short chains ($\leq k$ cells) contain too many vertices

- Unlikely, if cells are small
- Proof via method of *typical* bounded differences!
  - Imagine cells as boxes on conveyor belt
  - Imagine vertices as products
  - Typically not many vertices in few cells
  $\leadsto$ w.h.p., $o(n)$ vertices in large components $\checkmark$
Conclusion

Method of Bounded Differences
- Concentration for function of independent random variables
- Bounded differences ("Lipschitz") condition
  - What is the worst that can happen when changing one input?
  - Chernoff-like bound, weakened by sum of squared worst changes
  - Useless if worst changes are too large

Method of Typical Bounded Differences
- Define typical event, distinguish worst changes depending on whether event occurred
- Use mitigators to weaken impact of general worst changes
- Pay with probability that typical event does not occur, multiplied with inverse mitigators

Geometric Inhomogeneous Random Graphs
- Pretty realistic graph model (heterogeneity, locality)
- Not too hard to analyze
- Used for average-case analysis (e.g. vertex cover approximation) (not discussed in lecture)