Probabilistic Method and Random Graphs

Lecture 8. Random Graphs[1]

Xingwu Liu

Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China

[1]Based on Lecture 13 of Ryan O'Donnell's lecture notes of *Probability and Computing*.

Questions, comments, or suggestions?

A recap of Lecture 7

Poisson approximation theorem

$$\mathbb{E}\left[f\left(X_1^{(m)}, \ldots X_n^{(m)}\right)\right] \leq e\sqrt{m}\mathbb{E}\left[f\left(Y_1^{(m)}, \ldots Y_n^{(m)}\right)\right]$$

- $\Pr\left(\mathcal{E}\left(X_1^{(m)}, \ldots X_n^{(m)}\right)\right) \leq e\sqrt{m}\Pr\left(\mathcal{E}\left(Y_1^{(m)}, \ldots Y_n^{(m)}\right)\right)$
- $e\sqrt{m}$ can be improved to 2, if f is monotonic in m

Applications

- Max load: $L(n,n) > \frac{\ln n}{\ln \ln n}$ with high probability
- Max load: $L(n,n) = \Theta\left(\frac{\ln n}{\ln \ln n}\right)$ with high probability

A recap of Lecture 7

Hashing

- Hash table: accurate, time-efficient, space-inefficient
- Info. fingerprint: small error, time-inefficient, space-efficient
- Bloom filter: small error, time-efficient, more space-efficient

Type	Space	Time	Error rate
Hash table	$\geq 256m$	Constant	0
Information fingerprint	$m \lg_2 \frac{m}{c}$	$\ln m$	c
Bloom filter	$m \frac{-\ln c}{\ln 2}$	Constant	c

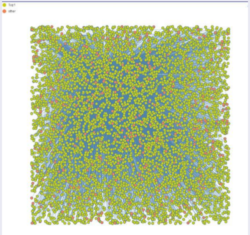
Motivation of studying random graphs

Gigantic graphs are ubiquitous

- Web link network: Teras of vertices and edges
- Phone network: Billions of vertices and edges
- Facebook user network: Billions of vertices and edges
- Human neural networks: 86 Billion vertices, $10^{14} - 10^{15}$ edges
- Network of Twitter users, wiki pages ...: size $\geq$ milllions

What do they look like?

- Impossible to draw and look
- What's meant by 'look like'?



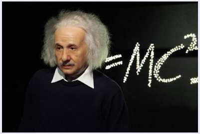
Looking through statistical lens

Examples of the statistics

- How dense are the graphs, $m = O(n)$ or $\Theta(n^2)$?
- Is it connected?
  - If not connected, how big are the components?
  - If connected, diameter
- What's the degree distribution?
- What's the girth? How many triangles are there?

Feasible for a single graph?

Yes, but not of the style of a **scientist**



Scientists' concerns

Interconnection

- Do the features appear inevitably or accidentally?
- Do various gigantic graphs have common statistical features?
- What accounts for the statistical difference between them?

Prediction

- What will a newly created gigantic graph be like?
- How is one statistical feature, given some others?

Exploitation (algorithmic)

- How do the features help algorithms? Say, routing, marketing
- What properties of the graphs determine the performance?

Key to solution

Modelling gigantic graphs: random graphs are a good candidate

Definition of random graphs

Intuition: stochastic experiments

- God plays a dice, resulting in a random number from 1 to 6
- God plays an amazing toy, resulting in a random graph
  - Amazing toy: a huge dice with a graph on each facet

Axiomatic definition of random graphs

Random graph with n vertices

- Sample space: all graphs on n vertices
- Events: every subset of the sample space is an event
- Probability function: any normalized non-negative function on the sample space

An example

$\mathcal{G}_n$: uniform random graph on n vertices

The probability function has equal value on all graphs

Simple questions on $\mathcal{G}_n$

Random variable $X : G \mapsto$ the number of edges of G

- What's $\mathbb{E}[X]$?
- What's $Var[X]$?

Tough? Not easy, at least.
Big names appeared!

A generative model of random graphs

$\mathcal{G}_{n,p}$, Erdös-Rényi model

Stochastic process:
Input: n and $p \in [0,1]$
Output: indicators $E_{ij}, 1 \leq i < j \leq n$
for $i = 1 \cdot \cdot n$
    for $j = i+1 \cdot \cdot n$
        $E_{ij} \leftarrow$ Bernoulli(p)

In one word:
$\mathcal{G}_{n,p}$ is an n-vertex graph the existence of each of whose edges is independently determined by tossing a p-coin.

Proposed in 1959 by Gilbert (1923-2013, American coding theorist and mathematician). Motivated by phone networks.

Erdös&Rényi get the naming credit due to extensive work

An example: $p = \frac{1}{2}$

Uniform distribution over n-vertex graphs

$\mathcal{G}_{n,\frac{1}{2}} \sim \mathcal{G}_n$, the axiomatic definition
What does it look like?

The number of edges

In $\mathcal{G}_{n,\frac{1}{2}}$, the number of edges has $\mathrm{Bin}\left(\binom{n}{2}, \frac{1}{2}\right)$ distribution.
Expectation: $\frac{n(n-1)}{4}$.
Variance: $\frac{n(n-1)}{8}$.
The expected degree of vertex i: $\frac{n-1}{2}$

Homogeneous degree distribution

Concentration theorem

In $\mathcal{G}_{n+1,\frac{1}{2}}$, all vertices have degree between $\frac{n}{2} - \sqrt{n \ln n}$ and $\frac{n}{2} + \sqrt{n \ln n}$ w.h.p.

Proof: Hoeffding's Inequality + Union Bound

Let D_i be the degree of vertex i.
$\Pr(D_i > \frac{n}{2} + \sqrt{n \ln n}) \leq e^{-2(\sqrt{n \ln n})^2/n} = n^{-2}$.
Likewise, $\Pr(D_i < \frac{n}{2} - \sqrt{n \ln n}) \leq n^{-2}$. So,

$$\Pr\left(\left|D_i - \frac{n}{2}\right| \geq \sqrt{n \ln n}\right) \leq \frac{2}{n^2},$$

$$\Pr\left(\bigcup_{i=1}^{n+1}\left(\left|D_i - \frac{n}{2}\right| \geq \sqrt{n \ln n}\right)\right) \leq \frac{2(n+1)}{n^2} = O(\frac{1}{n}),$$

$$\Pr\left(\bigcap_{i=1}^{n+1}\left(\left|D_i - \frac{n}{2}\right| < \sqrt{n \ln n}\right)\right) \geq 1 - O(\frac{1}{n}).$$

Another generative model of random graphs

$\mathcal{G}_{n,m}$

Randomly *independently* assign m edges among n vertices.
Equiv: uniform distribution over all n-vertex m-edge graphs

Proposed by Erdös&Rényi in 1959, and
independently by Austin, Fagen, Penney and Riordan in 1959.
Hard to study, due to dependency among edges.
Can we decouple the edges? Yes, sort of.

Decoupling the edges

$\mathcal{G}_{n,m} \sim \mathcal{G}_{n,p} | (m \text{ edges exist})$, for any $p \in (0,1)$.
Recall the Poisson Approximation Theorem

Both are called Erdös-Rényi model.
$\mathcal{G}_{n,p}$ is more popular.

Application of the decoupling

Probability of having isolated vertices

In random graph $\mathcal{G}_{n,m}$ with $m = \frac{n \ln n + cn}{2}$, the probability that there is an isolated vertex converges to $1 - e^{-e^{-c}}$.

Proof (By myself)

Basically, follow the proof of the theorem about coupon collecting. It is reduced to $\mathcal{G}_{n,p}$ with $p = \frac{\ln n + c}{n}$.

Problem reduction

In $\mathcal{G}_{n,p}$ with $p = \frac{\ln n + c}{n}$, the probability that there is an isolated vertex converges to $1 - e^{-e^{-c}}$.

Proof

E_i: the event that vertex v_i is isolated in $\mathcal{G}_{n,p}$.
E: the event that at least one vertex is isolated in $\mathcal{G}_{n,p}$.

$$\Pr(E) = \Pr(\cup_{i=1}^{n} E_i)$$
$$= -\sum_{k=1}^{n}(-1)^k \sum_{1\le i_1<i_2<\ldots<i_k\le n} \Pr(\cap_{j=1}^{k} E_{i_j}).$$

By Bonferroni inequalities,
$\Pr(E) \le -\sum_{k=1}^{l}(-1)^k \sum_{1\le i_1<\ldots<i_k\le n} \Pr(\cap_{j=1}^{k} E_{i_j})$, for odd l.

$$\Pr(\cap_{j=1}^{k} E_{i_j}) = (1-p)^{(n-k)k+\frac{k(k-1)}{2}} = (1-p)^{nk-\frac{k(k+1)}{2}}.$$

$$\Pr(E) \le -\sum_{k=1}^{l}(-1)^k \binom{n}{k} (1-p)^{nk-\frac{k(k+1)}{2}}, \text{for odd } l$$

$$\binom{n}{k}(1-p)^{nk-\frac{k(k+1)}{2}} > \frac{(n-k)^k}{k!}(1-p)^{nk-\frac{k(k+1)}{2}} \overset{n\to\infty}{=} \frac{e^{-ck}}{k!}.$$
$$\binom{n}{k}(1-p)^{nk-\frac{k(k+1)}{2}} < \frac{n^k}{k!}(1-p)^{nk-\frac{k(k+1)}{2}} \overset{n\to\infty}{=} \frac{e^{-ck}}{k!}$$

Continued proof

For odd l

$$\overline{\lim}_{n\to\infty} \Pr(E) \le -\sum_{k=1}^{l} \frac{(-e^{-c})^k}{k!} = 1 - \sum_{k=0}^{l} \frac{(-e^{-c})^k}{k!}$$

For even l, likewise

$$\underline{\lim}_{n\to\infty} \Pr(E) \ge -\sum_{k=1}^{l} \frac{(-e^{-c})^k}{k!} = 1 - \sum_{k=0}^{l} \frac{(-e^{-c})^k}{k!}$$

Altogether

Let l go to infinity. We have
$\underline{\lim}_{n\to\infty} \Pr(E) = \overline{\lim}_{n\to\infty} \Pr(E) = 1 - e^{-e^{-c}}$.
So, $\lim_{n\to\infty} \Pr(E) = 1 - e^{-e^{-c}}$

Reflection on $\mathcal{G}_{n,p}$

Homogeneity in degree

Degree of each vertex is Bin$(n-1, p)$.
Highly concentrated, as proven

Dense for constant p

$m = \Theta(n^2)$ whp.
Billions of vertices with zeta edges, too dense

Unfit for real-world networks

Heterogeneous in degree distribution.
Sort of sparse

Remark

$\mathcal{G}_{n,p}$-type randomness does appear in big graphs

Szemerédi Regularity Lemma

Tool in extremal graph theory by Endre Szemerédi in 1970's

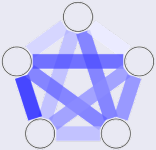
Hungarian-American (1940-)
Doctor vs Mathematician
Gelfond vs Gelfand

Szemerédi's Regularity Lemma

$\forall \epsilon, m > 0, \exists M > m$ such that any graph G with at least M vertices has an ϵ-regular k-partition, where $\exists m \leq k \leq M$.

Remark

Every large enough graph can be partitioned into a bounded number of parts which pairwise are like random graphs.



$$M = \left. m^{m^{m^{\cdot^{\cdot^{\cdot^{m}}}}}} \right\} d$$

$$\epsilon^{-\frac{1}{16}} \leq d = O(\epsilon^{-5})$$