Probabilistic Method and Random Graphs

Lecture 6. Hashing and Random Graphs [1]

Xingwu Liu

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

[1] The slides are mainly based on Chapter 5 of the textbook *Probability and Computing* and Lectures 12&13 of Ryan O'Donnell's lecture notes of *Probability and Computing*.

Preface

Questions, comments, or suggestions?

A recap of Lecture 5

Joint distribution of bin loads

$\Pr(X_1 = k_1, \ldots X_n = k_n) = \frac{m!}{k_1!k_2!\cdots k_n!n^m}$

Poisson approximation theorem

- $(X_1^{(m)}, X_2^{(m)}, \ldots X_n^{(m)}) \sim (Y_1^{(\mu)}, Y_2^{(\mu)}, \ldots Y_n^{(\mu)} | \sum Y_i^{(\mu)} = m)$
- $\mathbb{E}[f(X_1^{(m)}, \ldots X_n^{(m)})] \leq e\sqrt{m}\mathbb{E}[f(Y_1^{(m)}, \ldots Y_n^{(m)})]$
  - $\Pr(\mathcal{E}(X_1^{(m)}, \ldots X_n^{(m)})) \leq e\sqrt{m}\Pr(\mathcal{E}(Y_1^{(m)}, \ldots Y_n^{(m)}))$
  - $e\sqrt{m}$ can be improved to 2, if f is monotonic in m

Applications

- For the coupon collector's problem, $\lim_{n\to\infty} \Pr(X > n\ln n + cn) = 1 - e^{-e^{-c}}$
- Max load: $L(n, n) > \frac{\ln n}{\ln\ln n}$ with high probability

Application: Hashing

Used to look up records, protect data, find duplications ...

Membership problem: password checker

Binary search vs Hashing

Hash table (1953, H. P. Luhn @IBM)

Hash functions: efficient, deterministic, uniform, non-invertible

Random: coin tossing, SUHA

SHA-1 (broken by Wang et al., 2005)

Bins&Balls model

Efficiency

Search time for m words in n bins: expected vs worst.

Space: $\geq 256m$ bits if each word has 256 bits.

Potential wasted space: $\frac{1}{e}$ in the case of $m = n$.

Trade space for time. Can we improve space-efficiency?

Information Fingerprint

Fingerprint

Succinct identification of lengthy information

Fingerprint hashing

Fingerprinting $\rightsquigarrow$ sorting fingerprints (rather than original data)
$\rightsquigarrow$ binary search.

Trade time for space

Performance

False positive: due to loss of information

No other errors

Partial correction using white lists

False positive

Probability of a false positive: m words, b bits

Fingerprint of an acceptable differs from that of a bad: $1 - \frac{1}{2^b}$.
Probability of a false positive: $1 - \left(1 - \frac{1}{2^b}\right)^m \geq 1 - e^{-\frac{m}{2^b}}$.

Determine b

For a constant c, false positive $< c \Rightarrow e^{-\frac{m}{2^b}} \geq 1 - c$.
So, $b \geq \log_2 \frac{-m}{\ln(1-c)} = \Omega(\ln m)$.
If $b \geq 2 \log_2 m$, false positive $< \frac{1}{m}$.

2^{16} words, 32-bit fingerprints, false positive $< 2^{-16}$.
Save a factor of 8 if each word has 256 bits.

Can more space be saved while getting more time-efficient?

Bloom Filter

1970, CACM, by Burton H. Bloom.

Used in Bigtable and HBase.

Basic idea

Hash table + fingerprinting
Illustration

False positive is the only source of errors.

False positive: m words, n-bit array, k mappings

A specific bit is 0 with probability $\left(1 - \frac{1}{n}\right)^{km} \approx e^{-\frac{km}{n}} \triangleq p$.
Resonable to assume that a fraction p of bits are 0.
By Poisson approximation and Chernoff bounds.
False positive probability: $f \triangleq \left(1 - \left(1 - \frac{1}{n}\right)^{km}\right)^k \approx \left(1 - e^{-\frac{km}{n}}\right)^k$

Determine k for fixed m, n

Objective

Minimize f.
Dilemma of k: chances to find a 0-bit vs the fraction of 0-bits.

Optimal k

$\frac{d \ln f}{dk} = \ln\left(1 - e^{-\frac{km}{n}}\right) + \frac{km}{n}\frac{e^{-\frac{km}{n}}}{1-e^{-\frac{km}{n}}}$.
$\frac{d \ln f}{dk}\Big|_{k=\frac{n}{m}\ln 2} = 0$.
$f\big|_{k=\frac{n}{m}\ln 2} = 2^{-k} \approx 0.6185^{n/m}$.
$f < 0.02$ if $n = 8m$, and $f < 2^{-16}$ if $n = 23m$, saving 1/4 space

Remark

Fix n/m, the $\#$bits per item, and get a constant error probability.
In fingerprint hashing, $\Omega(\ln m)$ bits per item guarantee a constant error probability

An Introduction to Random Graphs

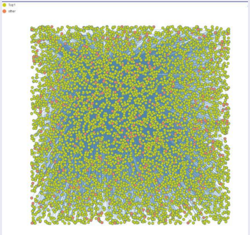
Motivation of studying random graphs

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 up to milllions

What do they look like?

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



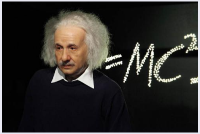
Looking through statistical lens

Part of the statistics

- How dense are the edges, $m = O(n)$ or $\Theta(n^2)$?
- Is it connected?
    - If not connected, the distribution of component size
    - 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 necessarily or just happen to appear?
- 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 (algorithmical)

- 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 the best candidate

Definition of random graphs

Intuition: stochastic experiments

- God plays a dice, resulting in a random number
- God plays an amazing toy, resulting in a random graph
  - Amazing toy: a big 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

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

Stochastic process:

input: n and $p \in [0, 1]$
output: indicators E_{ij}
for $i = 1 \cdot \cdot n$
for $j = i + 1 \cdot \cdot n$
$E_{ij} \leftarrow$ Bernoulli(p)

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

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.

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 $Bin\left(\binom{n}{2}, \frac{1}{2}\right)$ 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: Chernoff bound + Union Bound

Let D_i be the degree of vertex i.
$\Pr(D_i > \frac{n}{2} + \sqrt{n \ln n}) \leq e^{-(2\sqrt{\ln n})^2/2} = n^{-2}$.
Likewise, $\Pr(D_i < \frac{n}{2} - \sqrt{n \ln n}) \leq n^{-2}$.
By union bound, $\Pr(\frac{n}{2} - \sqrt{n \ln n} \leq D_i \leq \frac{n}{2} - \sqrt{n \ln n} \text{ for all } i) \geq 1 - \frac{2(n+1)}{n^2} = 1 - O(\frac{1}{n})$

Another generative model of random graphs

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

Randomly *independently* assign m edges among n vertices.
Equiv: All n-vertex m-edge graphs, uniformly distributed.

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

$\mathcal{G}_{n,m} \sim \mathcal{G}_{n,p} | (m \text{ edges exist})$
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 $\mathcal{G}_{n,p}$.

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

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.

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

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

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

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

Continued proof

For odd l

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

For even l, likewise

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

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, $\lim_{n\to\infty} \Pr(E) = 1 - e^{-e^{-c}}$

Reference

Lectures 12&13 of the CMU lecture notes by Ryan O'Donnell.