

Chapter 3

Probabilistic Analysis II: Random Graph Theory

We will see another general probabilistic tool: *random graphs*. We review how to analyze some basic properties of them while presenting an example of how to use them to model an interesting algorithmic question.

3.1 Power of Two Choices in Balls and Bins

We now review another variant of balls-and-bins experiments, aptly named the *power of two choices*. The problem we consider is as follows.

Problem 3.1. We are given a collection of m balls and n bins (so the number of balls can be different from that of the bins here). For each ball, we *independently* sample *two* distinct bins chosen *uniformly at random*; then place this ball in the bin which has the smaller load between the two (breaking ties arbitrarily). What is the asymptotically maximum load of a bin with probability at least $2/3$?

As before, our goal is to find the smallest function $T(n)$, asymptotically, such that

$$\Pr(\text{the load of every bin is at most } T(n)) \geq \frac{2}{3}.$$

Before the solution, let us go over a bit of practical motivation. The balls-and-bins experiment is a natural model of *hashing* and *task allocation*, e.g., choosing a server/processor (a bin) for an arriving task (a ball) in an online fashion. In some of these cases, we cannot readily know the loads of *all* servers in advance when a task arrives to choose a server with minimal load to assign this task to. For instance, when the tasks are coming from many parallel streams, no single stream can keep track of the loads on the servers (without a costly coordination across streams) nor is it feasible for it to “query” the load of every server when a new task arrives (without incurring a huge communication cost). The method in [Problem 3.1](#) then gives a “lightweight” way of implementing an approximately load-balanced solution with no coordination between the streams and minimal communication (only querying load of two bins at a time).

It was proved by [\[Mit01\]](#) that the right answer to [Problem 3.1](#) for $m = n$ is a load of

$$\log \log n \pm \Theta(1),$$

which is exponentially better than the original approach in [Problem 2.1](#) (also notice how sharp this bound is). This is a beautiful but intricate proof that we will not cover in these notes;

instead, we are going to prove a weaker version of this result which asymptotically gets the same bound (with a worse leading constant) and requires a slightly smaller number of balls than bins. This other proof is due to [KLMadH92] for a different problem of parallel hashing.

Theorem 3.2 ([KLMadH92]). *In Problem 3.1 for $m = n/100$ balls, the maximum load is $O(\log \log n)$ with high probability.*

Remark. The term ‘with high probability’ in Theorem 3.2 is used to refer to a probability which is at least $1 - 1/n^c$ for some constant $c \geq 1$; in most cases, one does not even explicitly specify this constant and just leaves it to be more than one^a.

As an example, if $n = 10^4$ (which is not even that large) and $c = 4$ (again not large), then the probability of this event happening is at most 10^{-16} – you can think of this number as follows: if every person on earth tries this event once every hour, *in expectation*, it still takes more than a century for the event to happen even once in the entire world!

^aAs shown in Problem 2.11, we can often increase this constant c to any other larger constant with minimal loss in the other parameters; as such, we can be more relaxed in this definition.

Remark. Morally speaking, the assumption of $m = n/100$ in Theorem 3.2 instead of $m = n$ is not that consequential. We can just think of the experiment with $m = n$ as 100 repetitions of the process with $m = n/100$ restarted each time, and so the maximum load remains $O(\log \log n)$ balls.

We should note that this is not entirely accurate because after each restart, the initial loads are not the same, but that should only help us because we are sending the balls to less loaded bins. Nevertheless, we will not attempt to formalize this and stick to the $m = n/100$ case for simplicity.

We will prove this theorem in the rest of this chapter. Interestingly, this can be proven using basic ideas from **random graph theory**, a beautiful area of discrete probability theory.

Consider the following distribution $\mathbb{G}(m, n)$ on n -vertex graphs:

Definition 3.3. We define the **random graph distribution** $\mathbb{G}(m, n)$ as follows. Sample an n -vertex graph $G = (V, E)$ with vertices $V = [n]$ by sampling m pairs of vertices uniformly at random and independently (with repetition) and add them as edges to G .

The distribution $\mathbb{G}(m, n)$ is one standard random graph distribution¹ and in random graph theory, we study properties of the graphs sampled from this and similar distributions. For instance, we can prove the following results.

Lemma 3.4. *For $m = n/100$, the largest connected component in a graph $G = (V, E)$ sampled from $\mathbb{G}(m, n)$ has $O(\log n)$ vertices with high probability.*

Lemma 3.5. *For $m = n/100$, any graph $G = (V, E)$ sampled from $\mathbb{G}(m, n)$ with high probability satisfies the following property: for any set $S \subseteq V$ of vertices, there are at most $5 \cdot |S|$ edges between the vertices in S .*

¹Another standard form—called the Erdős–Rényi random graphs—is a distribution wherein every pair of vertices is sampled as an edge independently with some given probability p .

We will prove both these properties in the next section. For now, let us see what any of these has to do with [Problem 3.1](#).

3.1.1 Proof of [Theorem 3.2](#) (assuming [Lemmas 3.4](#) and [3.5](#))

We can think of the balls-and-bins experiment of [Problem 3.1](#) as sampling a graph $G = (V, E)$ from $\mathbb{G}(m, n)$: the vertices in V correspond to the bins and each edge of E corresponds to a ball and shows which two bins are sampled randomly for it². Can we *upper bound* the maximum load of a bin in this trial of the balls-and-bins experiment if we are only given the graph G ? Yes, we are going to do that exactly.

Throughout this proof, we fix $m = n/100$, which corresponds to the setting of the problem we are interested in. We define the following process for a graph $G \sim \mathbb{G}(n/100, n)$.

Iterative peeling process: Peel all vertices in the graph G with degree at most 20 (i.e., remove them and their edges) and call them group B_1 ; then, again peel all vertices with degree at most 20 now (after we removed the vertices in B_1 , the degree of some other vertices can drop), and call them B_2 , and continue like this until either all vertices have been peeled or no vertex can be peeled.

Let $B_1, B_2, \dots$, be these groups and define B_∞ to be the last group that contains all vertices that could not have been peeled at all.

The following lemma is the key to the proof.

Lemma 3.6. *Consider a balls-and-bins trial, its corresponding graph G , and its iterative peeling process. For any $i \geq 1$, $i \neq \infty$, the bins in B_i have load at most $21 \cdot i$.*

Proof. We prove this inductively. Consider the base case of $i = 1$. These are bins with degree at most 20 in G ; this means that at most 20 balls ever even considered picking them and so even if those balls all enter these bins, the load of these bins is at most $20 < 21 \cdot 1$.

Consider the induction step for $i + 1$ (refer to [Figure 3.1](#) for an illustration of this proof). All the bins in $B_{\leq i}$ have load at most $21 \cdot i$ by induction and a bin $v \in B_{i+1}$ only has ≤ 20 edges to vertices *not* in $B_{\leq i}$ by definition. Consider any ball between $B_{\leq i}$ and v . If the load of v ever becomes more than $21 \cdot i$, then that ball chooses its other endpoint in $B_{\leq i}$ as its lower load bin by induction hypothesis. This means that the balls between $B_{\leq i}$ and v can only increase the load of v to $21 \cdot i + 1$. The remaining balls may increase the load by another 20 at most, so the load of v will be at most $21i + 1 + 20 = 21 \cdot (i + 1)$, proving the induction hypothesis. $\square$

We are now ready to conclude the proof of [Theorem 3.2](#). We argued that the corresponding graph G of a trial is sampled from $\mathbb{G}(n/100, n)$ and thus, with high probability, it satisfies both properties of [Lemma 3.4](#) and [Lemma 3.5](#) (we can do a union bound on the events of both lemmas). Conditioned on this, we now argue that the number of groups $B_1, B_2, \dots$, can only be $\log \log n + O(1)$ and $B_\infty = \emptyset$.

Consider any connected component C of G and we know that its size is $|C| = O(\log n)$ (as we conditioned on the event of [Lemma 3.4](#)). Moreover, we know that the average degree of vertices in C is at most 10 (as we conditioned on the event of [Lemma 3.5](#) and because the sum of the

²Note that this is a *lossy* “mapping” as in, even though each trial of the balls-and-bins experiment defines a fixed graph G , given the graph only, we cannot recover *all* the information about the trial necessarily; in particular we are losing the information about the ordering of the arrival of balls now.

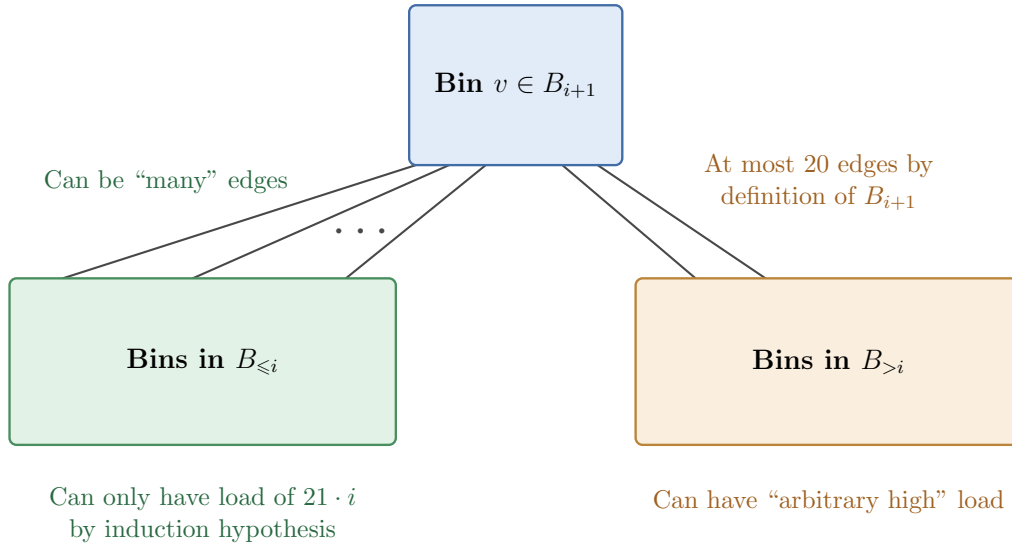


Figure 3.1: An illustration of the proof of the induction step. The edges/balls between v and $B_{>i}$ can (potentially) always choose v to assign their ball, but they are few and do not increase the load of v by much. The edges between v and $B_{≤i}$ will only choose v to assign when its load is “small” as otherwise they can choose $B_{≤i}$ instead by induction hypothesis.

degrees is twice the number of edges³). This means that at most $|C|/2$ vertices in C can have degree more than 20 otherwise average degree itself becomes more than 10, a contradiction⁴. So, in the first iteration of the peeling process, the size of the connected component C shrinks to at most half of its size. But, again, even in this new component, the average degree is at most 10 (again by conditioning on the event of Lemma 3.5 that holds for *every* subset). So, in the next iteration, again size of C shrinks by another factor of half. This means that after at most $\log |C|$ iterations of the peeling process, this entire component is peeled.

In conclusion, as the size of each connected component of G is $O(\log n)$, we get that after $\log O(\log n) = \log \log n + O(1)$ iterations of the peeling process, no vertex remains in the graph (and thus $B_\infty = \emptyset$). But then, by Lemma 3.6, we get that the maximum load of every bin is at most $21 \cdot \log \log n + O(1)$. As the events we conditioned on happen with high probability, we obtain this $O(\log \log n)$ bound with high probability also. This concludes the proof of Theorem 3.2.

3.2 Random Graph Theory: Proofs of Lemmas 3.4 and 3.5

We now prove the two results from random graph theory (Lemmas 3.4 and 3.5) that we used earlier in the proof of Theorem 3.2.

There is a general strategy applicable to proving many random graph theory results: show that for the property to be false, a certain *witness* should exist in the graph; then, prove that each witness happens with such a low probability in the graph that you can union bound over all its possible choices (we will shortly see two examples to help with clarifying this vague strategy). The heart of these arguments then usually is in finding the “right” witness.

We will use this approach in proving both lemmas.

³This is the so-called handshaking lemma: each edge contributes to degree of two vertices.

⁴This can be proven by an application of Markov bound; do you see how?

3.2.1 Lemma 3.4: Size of Connected Components in Random Graphs

We start with Lemma 3.4 that bounds the size of connected components in $\mathbb{G}(n/100, n)$.

Lemma (Restatement of Lemma 3.4). *The largest connected component in a graph $G = (V, E)$ sampled from $\mathbb{G}(n/100, n)$ has $O(\log n)$ vertices with high probability.*

Proof. Consider a connected component C in a graph G . Then, since C has a spanning tree, we know that there are at least $|C| - 1$ different edges inside C in G . Thus, for any $k \geq 1$, if we have a connected component of size k in G , we necessarily have a set S of k vertices with at least $k - 1$ edges. We use such an S as our *witness* and prove that for some $k_0 = \Theta(\log n)$, with high probability, there is no witness of size $k \geq k_0$ in G .

Fix an integer $k \geq 1$. We have,

$$\begin{aligned} \Pr(\exists \text{ a set of size } k \text{ with } \geq k - 1 \text{ different edges}) &\leq \sum_{\substack{S \subseteq V \\ |S|=k}} \Pr(S \text{ has } \geq k - 1 \text{ edges}) \\ &\quad \text{(by union bound)} \\ &\leq \sum_{\substack{S \subseteq V \\ |S|=k}} \sum_{\substack{F \subseteq \binom{S}{2} \\ |F|=k-1}} \Pr(\text{all edges of } F \text{ appear in } G) \end{aligned}$$

again, by union bound, because some fixed $k - 1$ pairs inside $\binom{S}{2}$ should all be edges in G ; here, the notation $\binom{S}{2}$ for a set (and not a number) S means the set of all pairs in S . Continuing,

$$\leq \sum_{\substack{S \subseteq V \\ |S|=k}} \sum_{\substack{F \subseteq \binom{S}{2} \\ |F|=k-1}} \left(\frac{n/100}{\binom{n}{2}} \right)^{k-1}$$

(each edge appears in G with prob. $\leq m/\binom{n}{2}$ and edges are negatively correlated; see below (*))

$$\begin{aligned} &= \binom{n}{k} \cdot \binom{\binom{k}{2}}{k-1} \cdot \left(\frac{n/100}{\binom{n}{2}} \right)^{k-1} \\ &\quad \text{(by the bound on number of choices of } S \text{ and } F) \\ &= \binom{n}{k} \cdot \binom{\frac{k \cdot (k-1)}{2}}{k-1} \cdot \left(\frac{n}{50 \cdot n \cdot (n-1)} \right)^{k-1} \\ &\leq \left(\frac{e \cdot n}{k} \right)^k \cdot \left(\frac{e \cdot k \cdot (k-1)}{2 \cdot (k-1)} \right)^{k-1} \cdot \left(\frac{1}{50 \cdot (n-1)} \right)^{k-1} \\ &\quad \text{(as } \binom{a}{b} \leq (e \cdot a/b)^b \text{ by Fact 2.10)} \\ &\leq \left(\frac{e \cdot n}{k} \right)^k \cdot (2k)^{k-1} \cdot \left(\frac{1}{40 \cdot n} \right)^k \cdot (40n) \\ &\quad \text{(we used } (e/2) \leq 2 \text{ and } n-1 \geq 4n/5 \text{ as } n \rightarrow \infty) \\ &\leq \left(\frac{2e}{40} \right)^k \cdot 40n \leq \left(\frac{1}{2} \right)^k \cdot 40n. \end{aligned}$$

In the equation (*), the edges are *negatively correlated* meaning that the probability of their joint event is upper bounded by what would have happened if they were independent,⁵ because we

⁵A series of events $E_1, \dots, E_t$ are called **negatively correlated** if

$$\Pr(E_1 \wedge E_2 \wedge \dots \wedge E_t) \leq \prod_{i=1}^t \Pr(E_i).$$

have a fixed budget of m samples and if an edge is sampled, it can only decrease the probability that another edge is also chosen. Let $k_0 = 10 \log n$. Then, for any $k \geq k_0$, we have,

$$\Pr(\exists \text{ a set of size } k \text{ with } \geq k-1 \text{ edges}) \leq (1/2)^k \cdot 40n \leq (1/2)^{k_0} \cdot 40n = (1/n)^{10} \cdot 40n \leq 1/n^8,$$

as $n \rightarrow \infty$. Taking another union bound over $n - k_0 + 1$ choices of k in $[k_0 : n]$, we get that

$$\begin{aligned} \Pr(\exists \text{ a connected component of size } \geq 10 \log n) \\ \leq \Pr(\exists \text{ a set } S \text{ of size } \geq 10 \log n \text{ with } \geq |S| - 1 \text{ edges}) \leq 1/n^7, \end{aligned}$$

concluding the proof. $\square$

3.2.2 Lemma 3.5: Induced Edge Counts in Random Graphs

Let us now prove Lemma 3.5.

Lemma (Restatement of Lemma 3.5). *Any graph $G = (V, E)$ sampled from $\mathbb{G}(n/100, n)$ with high probability satisfies: for any set $S \subseteq V$ of vertices, there are at most $5 \cdot |S|$ edges between the vertices in S .*

Proof. This time, we simply take the witness to be a set S with more than $5 \cdot |S|$ edges inside. Note that these edges need not be distinct, i.e., we consider each parallel edge separately. However, the way we count this is to say that at least $5|S|$ out of m sampled edges should be inside S . So, for any fixed $S \subseteq V$ with size $|S| = k$,

$$\Pr(S \text{ has } \geq 5k \text{ edges inside}) \leq \binom{n/100}{5k} \cdot \left(\frac{\binom{k}{2}}{\binom{n}{2}} \right)^{5k};$$

here, first term chooses $5k$ of the $n/100$ samples and the second term is the probability that these samples are all in S . We can now do a union bound over all choices of S with a fixed size $k \geq 1$ and have,

$$\begin{aligned} \Pr(\exists \text{ a set } S \text{ of size } k \text{ with } \geq 5k \text{ edges inside}) &\leq \binom{n}{k} \cdot \binom{n/100}{5k} \cdot \left(\frac{\binom{k}{2}}{\binom{n}{2}} \right)^{5k} \\ &\quad (\text{the first term choose } k \text{ vertices of } S \text{ from } n \text{ vertices of } V, \text{ and the rest is from above}) \\ &\leq \left(\frac{e \cdot n}{k} \right)^k \cdot \left(\frac{e \cdot n}{500 \cdot k} \right)^{5k} \cdot \left(\frac{k \cdot (k-1)}{n \cdot (n-1)} \right)^{5k} \\ &\quad (\text{using } \binom{a}{b} \leq (e \cdot a/b)^b \text{ inequality}) \\ &\leq \left(\frac{e \cdot n}{k} \right)^k \cdot \left(\frac{(k-1)}{100 \cdot (n-1)} \right)^{5k} \\ &\quad (\text{as } e/500 < 1/100) \\ &\leq \left(\frac{n}{k} \right)^k \cdot \left(\frac{k}{n} \right)^{5k} \left(\frac{e}{10^{10}} \right)^k \\ &\quad (\text{as } (k-1)/(n-1) \leq k/n \text{ and } 100^5 = 10^{10}) \\ &\leq \left(\frac{k}{n} \right)^{4k} \cdot \left(\frac{1}{10} \right)^{9k}; \quad (\text{as } e < 10) \end{aligned}$$

now note that either $1 \leq k \leq n^{0.1}$ or $n^{0.1} < k \leq n$; in the first case:

$$\left(\frac{k}{n} \right)^{4k} \cdot \left(\frac{1}{10} \right)^{9k} \leq \left(\frac{1}{n^{0.9}} \right)^{4k} \cdot 1 \leq \frac{1}{n^{3.6}};$$

in the second case:

$$\left(\frac{k}{n}\right)^{4k} \cdot \left(\frac{1}{10}\right)^{9k} \leq 1 \cdot \left(\frac{1}{10}\right)^{n^{0.1}} \ll \frac{1}{n^{100}},$$

for $n \rightarrow \infty$. A union bound over n choices of k implies that

$$\Pr(\exists \text{ a set } S \text{ of vertices with } \geq 5|S| \text{ edges inside}) \leq \frac{1}{n^2},$$

concluding the proof. □

This concludes our study of power of two choices in balls-and-bins experiments.

Remark. In the proofs of both [Lemma 3.4](#) and [Lemma 3.5](#) we have been *extremely* relaxed with the constants and inequalities; despite this, we still obtained the desired bounds (this is common in “these types of probabilistic proofs” where we leave ‘these types’ part entirely undefined on purpose: you know it, when you see it!). In principle, we could have tried to optimize the bounds and that would have, in the limit, allowed us to prove the same result as in [Theorem 3.2](#) for $m = (1 - \varepsilon) \cdot n$ for any constant $\varepsilon > 0$.

Nevertheless, when $m = n$, this approach would not work for a key reason: a random graph with n random edges with high probability has a “giant” connected component with $\Theta(n)$ vertices (as opposed to $O(\log n)$ in [Lemma 3.4](#)) and that would completely break the arguments used in this proof. The bound on the size of this connected component is related to the notion of “giant component” in random graphs that is a fascinating topic in random graph theory but is beyond the scope of this text^a.

^aJust for some context: when $m = (1 + \varepsilon) \cdot n/2$ for a constant $\varepsilon > 0$, then with high probability there is a single connected component of size $\Omega(n)$ —called the giant component—with all other components being of size only $O(\log n)$. Interesting, when m is very close to $n/2$, the size of the largest connected component is instead concentrated at $\sim n^{2/3}$ vertices (which is still enough to break the arguments used in our proof).