Counting Connected Graphs Asymptotically

Remco van der Hofstad*
Joel Spencer†

August 24, 2005

Abstract

We find the asymptotic number of connected graphs with k vertices and $k-1+l$ edges when k, l approach infinity, reproving a result of Bender, Canfield and McKay. We use the *probabilistic method*, analyzing breadth-first search on the random graph $G(k,p)$ for an appropriate edge probability p. Central is analysis of a random walk with fixed beginning and end which is tilted to the left.

1 The Main Results

In this paper, we investigate the number of graphs with a given complexity. Here, the *complexity* of a graph is its number of edges minus its number of vertices plus one. For $k, l \geq 0$, we write $C(k,l)$ for the number of labeled connected graphs with k vertices and complexity l.

The study of $C(k,l)$ has a long history. Cayley's Theorem gives the exact formula for the number of trees, $C(k,0) = k^{k-2}$. The asymptotic formula for the number of unicyclic graphs, $C(k,1)$, has been given by Rényi [10] and others. Wright [12] gave the asymptotics of $C(k,l)$ for l arbitrary but fixed and $k \to \infty$, and also studied the asymptotic behavior of $C(k,l)$ when $l = o(k^{1/3})$ in [13].

The asymptotics of $C(k,l)$ for all $k, l \to \infty$ were found by Bender, Canfield, and McKay [3]. The proof in [3] is based on an explicit recursive formula for $C(k,l)$. In this paper, we give an alternate, and substantially different, derivation of the Bender, Canfield, McKay results. Our argument is Erdős Magic, using the study of the random graph $G(k,p)$ to find the asymptotics of the strictly enumerative $C(k,l)$. The critical idea, given in Theorem 1.1 below, involves an analysis of a breadth-first search algorithm on $G(k,p)$. Similar methods, with somewhat weaker results in our cases, were employed recently by Coja-Oghlan, Moore and Sanwalani [4]. We can also use the results and methodology in this paper to find local statistics on the joint distribution of the size and complexity of the dominant component of $G(n,p)$ in the supercritical regime, which we defer to a future publication [6]. Further, while computational issues are not addressed in our current work, these methods may be used to efficiently generate a random connected graph of given size and complexity [1]. The idea of using the random graph to study $C(k,l)$ has appeared previously. In [8], a reformulation in terms of random graphs was used to prove upper and lower bounds on $C(k,l)$, extending the upper bound in [2]. The idea in [8] is that the expected value of the number of connected components with k vertices and complexity l can be explicitly written in terms of $C(k,l)$. Bounds on the random number of such components then imply bounds on $C(k,l)$. In

*Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. rhofstad@win.tue.nl

†Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012, U.S.A. spencer@cims.nyu.edu

[11], a more sophisticated analysis was performed, where the connected component of a given node in the random graph was explored using breadth-first search. This analysis allows us to rewrite the asymptotics of $C(k,l)$ for l fixed in terms of Brownian excursions. Interestingly, this identifies the Wright constants c_l for the asymptotics $C(k,l) \sim c_l k^{k-2} k^{3l/2}$ in terms of moments of the mean distance from the origin of a Brownian excursion. These moments were also investigated in [7], but the connection to the Wright constants had not been made before.

Quite recently, Pittel and Wormald [9] had yet another approach to the enumeration of connected graphs. They found asymptotics for the 2-core with a given number of vertices and edges and were then able to give the asymptotics of $C(k,l)$ for all $k,l \to \infty$. The plethora of methodologies employed is, in our minds, a reflection of the fundamental nature and depth of this problem.

In this paper, we will use the breadth-first search representation of connected components in $G(k,p)$ for $C(k,l)$, choose p appropriately, and analyse the resulting problem using probabilistic means. The critical identity is Theorem 1.1, which rewrites the $C(k,l)$ in terms of a k-step conditioned random walk with steps that are Poisson random variables minus one, and where the parameter of the Poisson steps varies with time. The main work in this paper then lies in the study of this random walk.

We note that when $l \geq k \ln k$ (and even somewhat less) the asymptotics of $C(k,l)$ are trivial as asymptotically almost all graphs on k vertices and $k-1+l$ edges will be connected. Thus, while our methods extend further, we shall restrict ourselves to finding the asymptotics of $C(k,l)$ with $l \leq k \ln k$ and $l \to \infty$. It will be convenient to subdivide the possible l into three regimes:

1. Very Large: $l \gg k$ and $l \leq k \ln k$;

2. Large: $l = \Theta(k)$;

3. Small: $l \ll k$ and $l \to \infty$.

The main ideas of this paper are given in Section 1, where we state the main results Theorems 1.2 and 1.5. The consequences of our main results for $C(k,l)$ are formulated in Section 2. The proofs of Theorems 1.2 and 1.5 are given in Section 3.

We employ the following fairly standard probability notation. $\mathrm{BIN}[n,p]$ denotes the Binomial Distribution with parameter n and probability of success p. $\mathrm{Po}(\lambda)$ denotes the Poisson Distribution with mean λ. $I[A]$ denotes the indicator random variable of an event A.

1.1 Tilted Balls into Bins

Let $k \geq 2$ be an integer. In this section, we define a process of placing $k-1$ balls into k bins with a tilted distribution, which makes it more likely that balls are placed in the left bins. In Section 1.2 below, we will find an identity between this bin process and $C(k,l)$.

Let $p \in (0,1]$. We have $k-1$ balls $1 \leq j \leq k-1$ and k bins $1 \leq i \leq k$. We place the j^{th} ball into bin T_j, where T_j is a random variable with distribution given by

$$\Pr[T_j = i] = \frac{p(1-p)^{i-1}}{1-(1-p)^k}. \tag{1.1}$$

This is a truncated geometric distribution. Note that the larger p is the more the T_j are tilted to the left. We shall call p the *tilt* of the distribution. The T_j are independent and identically distributed. Let Z_i, $1 \leq i \leq k$, denote the number of balls in the i^{th} bin. Set $Y_0 = 1$ and $Y_i = Y_{i-1} + Z_i - 1$ for $1 \leq i \leq k$. Note that $Y_k = 0$ as there are precisely $k-1$ balls. Let TREE be the event that

$$Y_t > 0 \quad \text{for} \quad 1 \leq t \leq k-1, \qquad Y_k = 0, \tag{1.2}$$

or, alternatively,

$$Z_1 + \ldots + Z_t \geq t \quad \text{for} \quad 1 \leq t \leq k-1, \qquad Z_1 + \ldots + Z_k = k-1. \tag{1.3}$$

(Note that $Y_k = 0$ and $Z_1 + \ldots + Z_k = k - 1$ hold trivially when we place $k-1$ balls. In a later stage, we will also consider Poisson placement of balls, which explains why we add these restrictions in the event TREE.) We note that we use the term TREE because there is a natural bijection between placements of $k-1$ balls into k bins satisfying TREE and trees on k vertices. Alternatively, one may consider $Y_0, Y_1, \ldots, Y_k$, with $Y_0 = 1, Y_k = 0$, as a walk with fixed endpoints, or a bridge. The condition TREE can then be interpreted as saying that the bridge is an *excursion.* In certain limiting situations the bridge approaches a biased Brownian bridge, where the bias depends on the parameter p.

Definition 1. *Set*

$$M = \binom{k}{2} - \sum_{j=1}^{k-1} T_j \tag{1.4}$$

in the probability space in which the T_j are independent with distribution given by (1.1). Set M^* equal the same random variable but in the above probability space conditioned on the event TREE.

We can give an alternative definition of M as follows:

$$\binom{k}{2} - \sum_{j=1}^{k-1} T_j = \sum_{i=1}^{k-1} (Y_i - 1), \tag{1.5}$$

which can be seen by noting that both sides of (1.5) increase by one when one ball is moved one position to the left and decrease by one when one ball is moved one position to the right. Since one can get from any placement to any other placement via a series of these moves, the two sides of (1.5) must differ by a constant. However, when $T_j = j$ for $1 \leq j \leq k-1$, we have $Y_i = 1$ for $1 \leq i \leq k-1$ and $Y_k = 0$ and so the sides are equal for this placement of balls.

1.2 The Critical Identity

The main idea of our approach is given in Theorem 1.1 below. Note that this result is exact, there are no asymptotics.

Theorem 1.1. *For all $k, l \in \mathbb{N}$, $p \in (0, 1]$,*

$$A_1 A_2 A_3 = C(k, l) p^{k+l-1} (1-p)^{\binom{k}{2}-(k+l-1)}, \tag{1.6}$$

where

$$A_1 = (1 - (1-p)^k)^{k-1}, \tag{1.7}$$
$$A_2 = \Pr[\text{TREE}], \tag{1.8}$$
$$A_3 = \Pr[\text{BIN}[M^*, p] = l]. \tag{1.9}$$

Proof. The right hand side of (1.6) is the probability that $G(k,p)$ is connected and has complexity l. We show that the left hand side of (1.6) also gives this probability. Designate a root vertex v and label the other vertices $1 \leq j \leq k-1$. We analyze breadth-first search on $G(k,p)$, starting with root v. (More precisely, the queue is initially $\{v\}$. In Stage 1 we pop v off the queue and add to the queue the neighbors of v. Each successive stage we pop a vertex off the queue and add to the queue its neighbors that haven't already been in the queue. The process stops when the queue is empty.)

Each non-root j flips a coin k times, where heads occurs with probability p. The i^{th} flip being heads has the following meaning. *If* the breadth-first search reaches the stage in which the i^{th} vertex is "popped" and if at that moment vertex j has not yet entered the queue, then j is adjacent to the popped vertex. (On an intuitive level breadth-first search is often viewed as vertices already "in" searching for new neighbors. Here, however, we view the vertices that are "out" as trying to get in! As the graph is undirected the two approaches yield the same result, but this change in viewpoint is absolutely central to our analysis.) To get all vertices, it is necessary that each j has at least one head. This happens with probability A_1. Conditioning on that, we let T_j be that first i when j had a head. So T_j has the truncated geometric distribution of (1.1). While the process continues Y_t is the size of the queue. The condition that the process does not terminate before stage k is precisely that no $Y_t = 0$ for $1 \le t \le k-1$, which is TREE, so this gives A_2. Now the only $\{w_1, w_2\}$ whose adjacency has not been determined are those for which (letting w_1 be the first one popped) w_2 was in the queue when w_1 was popped. There are precisely $\sum_{t=0}^{k-1}(Y_t - 1)$ of such pairs, i.e., we add the size of the queue minus the popped vertex over each stage, except for the last stage. Since we are conditioning on TREE, the random variable $\sum_{t=0}^{k-1}(Y_t - 1)$ has distribution M^*. We now look at those pairs, each is adjacent with independent probability p and to have complexity l, we need to have exactly l such pairs adjacent, so that the probability of this event equals A_3. □

Our approach to finding the asymptotics of $C(k,l)$ will be to find the asymptotics of A_2, A_3. This we shall be able to do when, critically, p has the appropriate value. We will let p depend on l and k, and the choice of p is described in more detail in Section 1.3. Looking ahead, we shall assume

$$k^{-3/2} \ll p \le 10\frac{\ln k}{k}. \tag{1.10}$$

It will be convenient to subdivide the possible p into three regimes:

1. Very Large: $k^{-1} \ll p \le 10\frac{\ln k}{k}$;
2. Large: $p = \Theta(k^{-1})$;
3. Small: $k^{-3/2} \ll p \ll k^{-1}$.

In each of these cases, we will write $p = \frac{c}{k}$, where $c = c(k) \to 0$ when p is small, and $c = c(k) \to \infty$ when p is very large.

The remainder of this paper is organized as follows. In Section 1.3, we define how to choose p appropriately, and we show that the above three regimes for p correspond to the three regimes of l given earlier. In Section 1.4, we investigate two walk problems, and relate the probability of TREE to the probability that these two walks do not revisit their starting point 0. In Section 1.5, we show that both M, and, more importantly, M^* obey a central limit theorem. In Section 2, we state the consequences of our results concerning Pr[TREE] and the asymptotic normality of M^* for $C(k,l)$ in the three cases above. In Section 3, we prove our main results.

1.3 The Choice of Tilt

Let μ, σ^2 denote the mean and variance of M. Both of these have closed forms as a function of p. We have the exact calculation

$$\mu = (k-1)[\frac{k}{2} - E[T_1]] = (k-1)\Big[\frac{k}{2} - \frac{1-(k+1)p(1-p)^k - (1-p)^{k+1}}{p(1-(1-p)^k)}\Big]. \tag{1.11}$$

We choose p to satisfy the equation

$$p\mu = l. \tag{1.12}$$

We can show from Calculus that $\mu = \mu(p)$ is an increasing function of p and so (1.12) will have a unique solution. The asymptotics depends on the regime. Writing $p = \frac{c}{k}$, we have for p satisfying (1.10),

$$\boxed{p\mu \sim f_1(c)k \qquad \text{and} \qquad \sigma^2 \sim f_2(c)k^3,} \tag{1.13}$$

with

$$f_1(c) = c\left[\frac{1}{2} - \frac{1-(c+1)e^{-c}}{c(1-e^{-c})}\right], \tag{1.14}$$

and, setting $\kappa = c(1-e^{-c})^{-1}$,

$$f_2(c) = \kappa\left[e^{-c}[-c^{-1} - 2c^{-2} - 2c^{-3}] + 2c^{-3}\right] - \left(\kappa[e^{-c}(-c^{-1} - c^{-2}) + c^{-2}]\right)^2. \tag{1.15}$$

For the first equality in (1.13) to hold, we use that (1.10) implies that $1-(1-p)^k \sim 1 - e^{-pk} = 1 - e^{-c}$. The second equality in (1.13) is similar.

In particular, for p small, and using that $f_1(c) = \frac{c^2}{12} + O(c^3)$ and $f_2(c) \sim \frac{1}{12}$ when $c \to 0$,

$$p\mu = \frac{k^3p^2}{12} + O(k^4p^3 + 1) \qquad \text{and} \qquad \sigma^2 \sim \frac{k^3}{12}. \tag{1.16}$$

For the error term in the first equality in (1.16), we need to use an improvement of (1.11), where, for p small, we keep track of the precise errors.

For p very large, on the other hand, now using that $f_1(c) \sim \frac{c}{2}$ and $f_2(c) \sim c^{-2}$ when $c \to \infty$, as well as $(1-p)^k \sim 0$, we obtain

$$p\mu \sim \frac{pk^2}{2} \qquad \text{and} \qquad \sigma^2 \sim \frac{k}{p^2}. \tag{1.17}$$

We see that the three regimes of p do indeed correspond to the three regimes of l, as we show now. Indeed, for l small, we have that $p\mu \sim \frac{k^3p^2}{12} = l$ is equivalent to

$$p \sim k^{-3/2}\sqrt{12l}. \tag{1.18}$$

On the other hand, for l large, if $l \sim k\beta$, then

$$p \sim \frac{c}{k} \qquad \text{with} \qquad f_1(c) = \beta, \tag{1.19}$$

while for l very large,

$$p \sim 2lk^{-2}. \tag{1.20}$$

The asymptotics in (1.13) with $p \sim \frac{c}{k}$ can be found by approximating $k^{-1}T_j$ by the continuous truncated exponential distribution over $[0,1]$, which has density $ce^{-cx}/(1-e^{-c})$.

1.4 Two Walks

We define two basic walks. In our applications, the Z_i, Z_i^R below will be random variables of various sorts and so ESC, ESC_L, ESC^R, ESC_L^R become events.

Definition 2. *Let $Z_1, Z_2, \ldots$ be nonnegative integers. The* leftwalk *is defined by the initial condition $Y_0 = 1$ and the recursion $Y_i = Y_{i-1} + Z_i - 1$ for $i \geq 1$. The escape event, denoted* ESC, *is that $Y_i > 0$ for all $i \geq 1$. The event* ESC_L, *or escape until time L, is that $Y_i > 0$ for $1 \leq i \leq L$.*

It shall often be convenient to count the bins "from the right." Let Z_i^R, $1 \le i \le k$, denote the number of balls in the $k-i+1$-st bin. Set $Y_0^R = 0$ and $Y_i^R = Y_{i-1}^R + 1 - Z_i^R$, so that $Y_i^R = Y_{k-i}$. Then TREE becomes

$$Y_t^R > 0 \quad \text{for} \quad 1 \le t \le k-1, \qquad Y_k^R = 0, \tag{1.21}$$

or, alternatively,

$$Z_1^R + \ldots + Z_t^R \le t-1 \quad \text{for} \quad 1 \le t \le k-1, \qquad Z_1^R + \ldots + Z_k^R = k-1. \tag{1.22}$$

We shall generally use the superscript R when examining bins from the right. In particular, we set $i^R = k - i$, so that bin i^R is the i^{th} bin from the right.

Definition 3. *Let $Z_1^R, Z_2^R, \ldots$ be nonnegative integers. The* rightwalk *is defined by the initial condition $Y_0^R = 0$ and the recursion $Y_i^R = Y_{i-1}^R + 1 - Z_i^R$ for $i \ge 1$. The escape event, denoted* ESC^R, *is that $Y_i^R > 0$ for all $i \ge 1$. The event ESC_L^R, or escape until time L, is that $Y_i^R > 0$ for $1 \le i \le L$.*

We allow Z_i, Z_i^R to be defined only for $1 \le i \le L$ in which case Y_i, Y_i^R are defined for $0 \le i \le L$ and $\mathrm{ESC}_L, \mathrm{ESC}_L^R$ are well defined. Indeed, our main results will be for these walks of length L, the infinite walks shall be a convenient auxiliary tool.

When $k-1$ balls are placed into k bins with tilt p and Z_i is the number of balls in bin i, the event ESC_L is that $Y_t > 0$ for $1 \le t \le L$. Letting Z_i^R be the number of balls in bin i^R, the event ESC_L^R is that $Y_t^R > 0$ for $1 \le t \le L$.

Definition 4. *Given $L < \frac{1}{2}k$, we call bins with $1 \le i \le L$ the left side; bins with $1 \le i^R \le L$ the right side; and all other bins the middle.*

Now consider the tilted balls into bins formulation of Section 1.1. Set

$$\lambda = (k-1)\frac{p}{1-(1-p)^k} \qquad \text{and} \qquad \lambda^R = (k-1)\frac{p(1-p)^{k-1}}{1-(1-p)^k}, \tag{1.23}$$

so that λ, λ^R are the expected number of balls in the leftmost and rightmost bin respectively. When $p = \frac{c}{k}$, the asymptotics of λ and λ^R are given by

$$\boxed{\lambda \sim \frac{c}{1-e^{-c}} \qquad \text{and} \qquad \lambda^R \sim \frac{ce^{-c}}{1-e^{-c}}.} \tag{1.24}$$

In particular, for p very large, $\lambda \to \infty$ and $\lambda^R \sim 0$, while for p small, $\lambda = 1 + \frac{pk}{2}(1+o(1))$ and $\lambda^R = 1 - \frac{pk}{2}(1+o(1))$.

Theorem 1.2. *Let* ESC *be given by Definition 2, with all $Z_i \sim \mathrm{Po}(\lambda)$. Let ESC^R be given by Definition 3, with all $Z_i^R \sim \mathrm{Po}(\lambda^R)$. Let p be in the range given by (1.10). Then*

$$\Pr[\mathrm{TREE}] \sim \Pr[\mathrm{ESC}]\Pr[\mathrm{ESC}^R]. \tag{1.25}$$

We may naturally interpret Theorem 1.2 as saying that the event TREE is asymptotically equal to the probability that the left and right sides satisfy the conditions imposed by TREE. For i small, Y_i behaves like a leftwalk with Z_i being roughly $\mathrm{Po}(\lambda)$ and Y_i^R behaves like a rightwalk with Z_i^R being roughly $\mathrm{Po}(\lambda^R)$. The proof of Theorem 1.2 is deferred to Section 3.

The left and right walks with Z_i, Z_i^R independent Poisson have been well studied. Let $Z \sim \mathrm{Po}(1+\varepsilon)$. Let $Z_i \sim Z$ for all i, and the Z_i are mutually independent. Consider the leftwalk as given by Definition 2.

Theorem 1.3. $\Pr[\mathrm{ESC}] = y$ *where y is the unique real number in $(0,1)$ such that $e^{-(1+\varepsilon)y} = 1 - y$. Further, if $y = y(\varepsilon)$, then $y \le 2\varepsilon$ for all positive ε and $y \sim 2\varepsilon$ as $\varepsilon \to 0^+$.*

Proof. We use that there is a bijection between random walks with i.i.d. steps with distribution $\mathrm{Po}(\lambda)-1$ and Galton-Watson trees with offspring distribution $\mathrm{Po}(\lambda)$. This bijection is such that random walks that never return to the origin are mapped to branching process configurations where the tree is infinite. For the latter, we have that the probability is the survival probability of the branching process. The extinction probability x satisfies

$$e^{-\lambda(x-1)} = x. \tag{1.26}$$

Therefore, for the survival probability $y = 1 - x$, we obtain

$$e^{-(1+\varepsilon)y} = 1 - y. \tag{1.27}$$

The inequality $y(\varepsilon) \le 2\varepsilon$ and the asymptotics $y(\varepsilon) \sim 2\varepsilon$ are elementary calculus exercises. □

Next let $Z^R \sim \mathrm{Po}(1-\varepsilon)$. Let $Z_i^R \sim Z^R$ all i, independent. Consider the rightwalk as given by Definition 3. Then we can identify the probability of ESC^R exactly as follows:

Theorem 1.4. $\Pr[\mathrm{ESC}^R] = \varepsilon$.

Proof. Consider an infinite walk starting at zero with step size $1-P$ where P is Poisson with mean $1-\varepsilon$. Here, $\varepsilon \in (0,1)$ but we do *not* assume $\varepsilon \to 0$. We claim $\Pr[\mathrm{ESC}^R] = \varepsilon$ precisely. Take an infinite random walk, $0 = Y_0, Y_1, Y_2, \ldots$ and let W_n be the number of i, where $0 \le i < n$, for which $S_t = Y_{i+t} - Y_i$ for $t \ge 0$ never returns to zero, i.e., the number of i, $0 \le i < n$ for which $Y_j > Y_i$ for all $j > i$.

For each i this has probability α of occurring so that by linearity of expectation $E[W_n] = n\alpha$. Let V_n be the minimum of Y_j for $j \ge n$. Then, by definition $W_n = \max[V_n, 0]$. Indeed, for each $0 \le j < V_n$, let $i = i(j)$ be the maximal i, $0 \le i < n$ for which $Y_i = j$. These are precisely the i for which the walk beginning at time i has the desired property. Thus $n\alpha = E[\max[V_n, 0]]$. So far everything is exact and now it follows from the fact that the random walk has positive drift that

$$\lim_{n\to\infty} \frac{E[\max[V_n, 0]]}{n} = \varepsilon. \tag{1.28}$$

□

Suppose $p = \frac{c}{n}$ with $c > 0$ fixed. Then $\Pr[\mathrm{ESC}] = y$ where $e^{-\lambda y} = 1 - y$ by (1.27) and λ is given by (1.23), so that $\Pr[\mathrm{ESC}] \sim 1 - e^{-c}$. Theorem 1.4 gives $\Pr[\mathrm{ESC}^R] = 1 - \lambda^R$ where λ^R is given by (1.23), so that $\Pr[\mathrm{ESC}^R] \sim \frac{1-(c+1)e^{-c}}{1-e^{-c}}$. The asymptotics of Theorem 1.2 are then that for $p = \frac{c}{k}$,

$$\boxed{\Pr[\mathrm{TREE}] \sim 1 - (c+1)e^{-c}.} \tag{1.29}$$

For p small

$$\Pr[\mathrm{TREE}] \sim \frac{p^2k^2}{2}, \tag{1.30}$$

while for p very large

$$\Pr[\mathrm{TREE}] \sim 1. \tag{1.31}$$

1.5 The Limiting Gaussian

In this section, we give an asymptotic normal law for M^* and the consequent asymptotics of A_3. For M, by the fact that the T_j are independent, Esseen's Inequality (or the Lindeberg-Feller central limit theorem, or any of a wide variety of standard probability tools) gives that M is asymptotically Gaussian with mean μ given in (1.11) and variance σ^2. Therefore, for any fixed real u,

$$\Pr[M \le \mu + u\sigma] \sim \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \tag{1.32}$$

Theorem 1.5. *Let M^* be given by Definition 1. Then for any fixed real u*

$$\Pr[M^* \le \mu + u\sigma] \sim \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \tag{1.33}$$

Here, importantly, μ is given by (1.11), the expectation of the unconditioned M. Theorem 1.5 then has the natural interpretation that conditioning on TREE does not change the asymptotic distribution of M. The proof of Theorem 1.5 is deferred to Section 3.

We next use Theorem 1.5 to determine the asymptotics of A_3. For this, we define σ_Y by

$$\sigma_Y^2 = p\mu + p^2\sigma^2. \tag{1.34}$$

Proposition 1.6. *With p given by (1.12) and σ_Y given by (1.34), whenever $p^2\sigma^2 = O(p\mu)$,*

$$\Pr[\mathrm{BIN}[M^*, p] = l] \sim \sigma_Y^{-1}(2\pi)^{-1/2}. \tag{1.35}$$

Proof. For integral $m \ge 0$, we define $f(m) = \Pr[\mathrm{BIN}[m,p] = l]$. We calculate $f(m+1)/f(m) = (m+1)(1-p)/(m-l+1)$. This quantity is one at $m = \mu - 1$, greater than one for $m < \mu - 1$ and less than one for $m > \mu - 1$ so the function $f(m)$ is unimodal, hitting a maximum at $m = \lfloor \mu \rfloor$. Stirling's Formula gives $f(\lfloor \mu \rfloor) \sim (2\pi l)^{-1/2}$.

We write

$$\Pr[\mathrm{BIN}[M^*, p] = l] = E[f(M^*)] = f(\lfloor \mu \rfloor) E[g(M^*)].$$

where we set $g(m) = f(m)/f(\lfloor \mu \rfloor) \le 1$. Let K be a large positive constant, and let A_K be the event $|M^* - \mu| \le K\sigma$. We split

$$E[g(M^*)] = E[g(M^*)I[A_K]]E[g(M^*)I[\overline{A}_K]].$$

As $g(m) \le 1$ uniformly

$$E[g(M^*)I[\overline{A}_K]] \le \Pr[\overline{A}_K] \sim \Pr[|N| > K],$$

where N is the standard normal distribution. Parametrising $m = \mu + x\sigma$, Stirling's Formula yields

$$g(m) \sim e^{-\frac{x^2}{2}\frac{p^2\sigma^2}{p\mu}},$$

uniformly over $|x| \le K$. (Roughly, $\mathrm{BIN}[\mu + x\sigma, p] \sim \mathrm{BIN}[\mu, p] + xp\sigma$ and so we want $\mathrm{BIN}[\mu, p]$ to be $(xp\sigma)^{1/2}/(p\mu)$ standard deviations off the mean. Note that when $p^2\sigma^2 = o(p\mu)$, we simply have $g(m) \sim 1$.) Thus

$$E[g(M^*)I[A_K]] \sim \int_{-K}^{K} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{-\frac{x^2}{2}\frac{p^2\sigma^2}{p\mu}} dx.$$

Now letting $K \to \infty$, and using that $\Pr[|N| > K] \to 0$, as well as

$$E[g(M^*)] \sim \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}\sigma_Y^2 l^{-1}} = l^{1/2}\sigma_Y^{-1}, \tag{1.36}$$

we arrive at

$$E[f(M^*)] \sim (2\pi l)^{-1/2} E[g(M^*)] \sim (2\pi)^{-1/2}\sigma_Y^{-1}.$$

□

Again we can look at the asymptotics (we won't need finer expressions) in the different regimes.

1. When l is small, then $p\mu = l \sim p^2\sigma^2$ and
$$\sigma_Y^2 \sim 2l. \tag{1.37}$$
2. When l is large, say $l \sim k\beta$, then $\sigma^2 p^2 \sim c^2 f_2(c)k$ with c satisfying $f_1(c) = \beta$ as in (1.19) so that
$$\sigma_Y^2 \sim l(1 + c^2 f_2(c)\beta^{-1}). \tag{1.38}$$
3. When l is very large, then $\sigma^2 p^2 = o(p\mu)$ and
$$\sigma_Y^2 \sim l. \tag{1.39}$$

2 Asymptotics for $C(k,l)$

We now use the results in the previous section, in particular Theorems 1.1, 1.2 and 1.5, to derive the asymptotics for $C(k,l)$. Indeed, for p given by (1.12), the asymptotics of all terms in (1.6) are known except $C(k,l)$. Hence we can solve for the asymptotics of $C(k,l)$. While Theorem 1.1 and the auxiliary results allow us to find the asymptotics of $C(k,l)$ in theory, some of the technical work can be challenging. Here we indicate some of the major cases.

It shall be helpful not to use the precise p given by (1.12). Recall that Theorem 1.1 holds for any value of p. For the moment let us write $\mu = \mu(p)$, $\sigma = \sigma(p)$, $A_2 = A_2(p)$ and $A_3 = A_3(p)$ to emphasize this dependence.

Lemma 2.1. *Let p_0 be the value of p satisfying (1.12), i.e., $p_0\mu(p_0) = l$. Let p be such that $p \sim p_0$ and $p\mu(p) = l + o(l^{1/2})$. Then $A_2(p) \sim A_2(p_0)$, $\sigma(p) \sim \sigma(p_0)$ and $A_3(p) \sim A_3(p_0)$.*

Proof. The asymptotics of $A_2(p) = \Pr[\text{TREE}]$ are given by (1.29)– (1.31) and clearly have this property. Similarly the formulae (1.13),(1.16)– (1.17) show that $\sigma(p) \sim \sigma(p_0)$. An examination of the proof of Proposition 1.6 gives that the asymptotics (1.35) of $\Pr[\text{BIN}[M^*, p_0] = l]$ apply to $\Pr[\text{BIN}[M^*, p] = l]$ as long as $l - p\mu(p) = o(l^{1/2})$ as the integral (1.36) remains the same. □

2.1 l Small

Theorem 2.2. *When $l = o(k^{1/2})$ and $l \to \infty$,*
$$C(k,l) \sim k^{k-2}k^{3l/2}(e/12l)^{l/2}(3\pi^{-1/2})l^{1/2}. \tag{2.1}$$

Proof. Setting $p = k^{-3/2}\sqrt{12l}$ we have from (1.16) that $p\mu = l + O(k^4p^3+1)$ and $O(k^4p^3) = O(k^{-1/2}l^{3/2}) = o(l^{1/2})$ as $l = o(k^{1/2})$. Let p_0 satisfy $p_0\mu(p_0) = l$. Then $p \sim p_0$ by (1.16), and $p\mu = l + o(l^{1/2})$. Lemma 2.1 then gives $A_2(p) \sim A_2(p_0)$ and $A_3(p) \sim A_3(p_0)$.

We start with the exact formula
$$C(k,l) = A_1A_2A_3p^{-(k+l-1)}(1-p)^{-\binom{k}{2}+(k+l-1)}. \tag{2.2}$$
By (1.30), we have
$$A_2 = \Pr[\text{TREE}] \sim \frac{1}{2}(kp)^2 = 6lk^{-1}. \tag{2.3}$$
We further have $p^2\sigma^2 + p\mu \sim 2l$ so that Proposition 1.6 and (1.37) give
$$A_3 = \Pr[\text{BIN}[M^*, p] = l] \sim (2\pi)^{-1/2}(2l)^{-1/2}. \tag{2.4}$$

We have to be quite careful with the asymptotics of the exact formula $A_1 = [1-(1-p)^k]^{k-1}$ of Theorem 1.1. We have

$$1-(1-p)^k = pk\left(1-\frac{pk}{2}+\frac{p^2k^2}{6}-\frac{p^3k^3}{24}+O(p^4k^4)\right), \tag{2.5}$$

and

$$\ln\left(1-\frac{pk}{2}+\frac{p^2k^2}{6}-\frac{p^3k^3}{24}+O(p^4k^4)\right) = -\frac{pk}{2}+\frac{p^2k^2}{6}-\frac{p^3k^3}{24}-\frac{p^2k^2}{8}+\frac{p^3k^3}{12}-\frac{p^3k^3}{24}+O(p^4k^4) = -\frac{pk}{2}+\frac{l}{2k}+o(k^{-1}), \tag{2.6}$$

The p^3k^3 terms cancel and our assumption $l = o(k^{1/2})$ implies that the error term $O(p^4k^4) = o(k^{-1})$. Hence the error term in $\ln A_1$ is $o(1)$. This gives

$$A_1 \sim p^{k-1}k^{k-1}e^{-pk^2/2}e^{l/2}. \tag{2.7}$$

We further have

$$p^{k+l-1} = p^{k-1}(12l)^{l/2}k^{-3l/2}, \tag{2.8}$$

and the asymptotics

$$(1-p)^{k^2/2} \sim e^{-k^2p/2}, \tag{2.9}$$

while

$$(1-p)^{k+l-1} \sim 1. \tag{2.10}$$

Then Theorem 1.1 puts everything together and yields (2.1). □

2.2 l Large

In this section, we take l such that $\beta \equiv l/k$ is uniformly bounded and uniformly positive, and investigate the scaling of $C(k,l)$ in this range. We state the result uniformly in $\beta \in [\varepsilon, \varepsilon^{-1}]$, since we cannot fix β due to the fact that $l = \beta k$ needs to be an integer. The main result is as follows:

Theorem 2.3. *Let $\varepsilon > 0$ be arbitrary and fixed. Then as $k \to \infty$ and $l = l(k) \in [\varepsilon k, \varepsilon^{-1}k]$,*

$$C(k,k\beta) \sim A \cdot B^k \cdot k^{(1+\beta)k} \cdot k^{-3/2}, \tag{2.11}$$

where $\beta = l/k$, c is the solution to

$$e^{-c} = \frac{2(\beta+1)-c}{2(\beta+1)+c}, \tag{2.12}$$

$$A = \frac{c(c-2\beta)}{\sqrt{8\pi\beta(1+c^2f_2(c)/\beta)}}e^{-c(\beta/2+1)}, \tag{2.13}$$

and

$$B = \frac{2}{c^\beta\sqrt{4(\beta+1)^2-c^2}}. \tag{2.14}$$

Proof. Let l satisfy that $l = l(k) \in [\varepsilon k, \varepsilon^{-1}k]$. Then the p of (1.12) satisfies

$$p = \frac{c}{k} + O(k^{-2}), \tag{2.15}$$

where c is the solution to (2.12). Changing p by an additive $O(k^{-2})$ term changes $p\mu(p)$ by $O(1)$. Lemma 2.1 allows us to set $p = \frac{c}{k}$ with A_2, A_3 the same as for that p given by (1.12). We get $C(k,\beta k)$ from the equation

$$A_1A_2A_3 = C(k,\beta k)p^{k+\beta k-1}(1-p)^{\binom{k}{2}-k-\beta k+1} \tag{2.16}$$

Here, taking care to note that the asymptotics $(1-p)^k \sim e^{-c}$ are *not* sufficiently precise to give the asymptotics of A_1, we find

$$A_1 \sim \Big(1 - e^{-c}(1 - \frac{c^2}{2k})\Big)^{k-1} \sim \Big(\frac{2c}{2(\beta+1)+c}\Big)^{k-1} e^{((\beta+1)/c-1/2)c^2/2}, \tag{2.17}$$

while

$$A_2 \sim 1-(c+1)e^{-c}, \qquad \text{and} \qquad A_3 = \frac{1}{\sqrt{2\pi k\beta(1+c^2 f_2(c)/\beta)}}. \tag{2.18}$$

Further,

$$p^{k(1+\beta)-1} = \Big(\frac{c}{k}\Big)^{k(1+\beta)-1}, \tag{2.19}$$

while, taking care to estimate $\ln(1-p)$ as $-ck^{-1} - \frac{1}{2}c^2k^{-2}$,

$$(1-p)^{\binom{k}{2}-k(1+\beta)+1} \sim e^{-kc/2-c^2/4+c(\beta+1.5)}, \tag{2.20}$$

Solving and employing uniformity of convergence we obtain Theorem 2.3. □

2.3 l Very Large

As a third example suppose $l = \lfloor ck \ln k \rfloor$. We prove the following result:

Theorem 2.4. *For $l = \lfloor ck \ln k \rfloor$ with $c > \frac{1}{2}$,*

$$C(k,l) \sim \binom{\frac{k(k-1)}{2}}{k+l-1}. \tag{2.21}$$

This has the interpretation that the proportion of graphs with k vertices and $k+l-1$ edges which are connected is asymptotically one, or that the probability that a random graph with k vertices and $k+l-1$ edges is connected is asymptotically one. As such, this is immediate from a classic results of Erdős and Rényi [5].

Proof. We again start from (2.2). Then (1.20) gives that $p \sim 2lk^{-2}$, which implies $A_1 \sim A_2 \sim 1$. (Note that $A_1 \sim 1$ fails for $c < \frac{1}{2}$.) Further, Proposition 1.6 with the asymptotics in (1.39) gives $A_3 \sim (2\pi l)^{-1/2}$. It shall be convenient to rewrite this as $A_3 \sim (2\pi(k+l-1))^{-1/2}$. We conclude that

$$C(k,l) \sim (2\pi(k+l-1))^{-1/2} p^{-(k+l-1)}(1-p)^{-\binom{k}{2}+(k+l-1)} = (2\pi B)^{-1/2} p^{-B}(1-p)^{-A+B}, \tag{2.22}$$

where we abbreviate $A = \binom{k}{2}, B = k+l-1$. However, this is not a sufficiently precise approximation of p to give the asymptotics of $C(k,l)$. Rather, in the region $p \sim 2lk^{-2}$, the exact expression (1.11) can be rewritten as follows:

Lemma 2.5.

$$\mu = \binom{k}{2} - (k-1)p^{-1} + \frac{k(k+1)(1-p)^k}{1-(1-p)^k}. \tag{2.23}$$

Proof. This is a simple calculation. □

By Lemma 2.5, and using that $1-(1-p)^k \sim 1$, we obtain that

$$\mu = \binom{k}{2} - (k-1)p^{-1} + O(k^2(1-p)^k). \tag{2.24}$$

As we have required from (1.12) that $p\mu = l$, we have

$$p\binom{k}{2} = k + l - 1 + O(pk^2(1-p)^k). \tag{2.25}$$

We note that

$$\binom{A}{B} p^B (1-p)^{A-B} = \Pr[\mathrm{BIN}[A,p] = B] \sim (2\pi B)^{-1/2}, \tag{2.26}$$

where the latter equality holds by the local central limit theorem for the binomial distribution whenever $B - pA = o(\sqrt{p(1-p)A})$. Note that by (2.25), with $A = \binom{k}{2}$ and $B = k + l - 1$,

$$B - pA = (k + l - 1) - p\binom{k}{2} = O(pk^2(1-p)^k) = o(\sqrt{p(1-p)A}) = o(k\sqrt{p}), \tag{2.27}$$

precisely when

$$(1-p)^k \sqrt{pk^2} = o(1). \tag{2.28}$$

At this stage we need only the weaker assumption $l \sim ck \ln k$ with $c > \frac{1}{4}$. As $p \sim 2lk^{-2}$, $(pk^2)^{1/2} = k^{-1/2+o(1)}$. Also $\ln((1-p)^k) = k\ln(1-p) \sim -kp \sim 2c \ln k$, so that $(1-p)^k = k^{-2c+o(1)}$ and (2.28) holds. Therefore, in this case, Theorem 1.1 gives

$$C(k,l) p^B (1-p)^{A-B} \sim (2\pi B)^{-1/2}, \tag{2.29}$$

and thus we deduce

$$C(k,l) \sim \binom{A}{B} = \binom{\frac{k(k-1)}{2}}{k + l - 1}. \tag{2.30}$$

$\square$

We note that in principle it is possible to extend the above asymptotics to other l for which $\frac{l}{k} \to \infty$, using Lemma 2.5 and more precise local central limit theorems for $\Pr[\mathrm{BIN}[A,p] = B]$.

3 The Technical Theorems

In this section, we prove Theorems 1.2 and 1.5. The values λ, λ^R, the expected number of balls in the first and last bins respectively, are given by (1.23). We start in Section 3.1 with the easy case where p is large and very large. The remaining Sections 3.2–3.8 are devoted to the hard case where $k^{-3/2} \ll p \ll k^{-1}$.

3.1 The Easy Case: p Very Large and p Large

We note that the arguments for the "hard case" apply to the cases where p is large and very large as well. However, many of the subtleties of the hard case can be avoided when $p = \Omega(k^{-1})$. Here we give, without full details, a simpler argument that works in these important cases.

First suppose $p \gg k^{-1}$. Let FAIL_t be the event $Y_t \leq 0$. For example, FAIL_1 is the event $Z_1 = 0$ which has probability $e^{-(1+o(1))\lambda}$ which approaches zero. The event FAIL_k is the event $Z_1^R > 0$ so $\Pr[\mathrm{FAIL}_k] \leq E[Z_1^R] = \lambda^R \to 0$. In general, as each ball is dropped independently, $Z_1 + \ldots + Z_t$ has distribution $\mathrm{BIN}[k-1, \alpha]$ where $\alpha = \Pr[T_j \leq i]$ as given by the distribution (1.1). (Near the right side it is easier to work with $\Pr[Y_t^R \leq 0]$.) Chernoff bounds give that $\sum_{t=1}^{k} \Pr[\mathrm{FAIL}_t] \to 0$ and so $\Pr[\mathrm{TREE}] \to 1$, giving Theorem 1.2. Conditioning on an event that holds with probability $1 - o(1)$

cannot effect an asymptotic Gaussian distribution and so Theorem 1.5 follows immediately for the very large case.

We next proceed with the case where p is large. Set $p = \frac{c}{k}$. Note λ, λ^R are given by (1.23). We split bins into left, right and middle by Definition 4. We set $L = \lfloor \ln^2 k \rfloor$ for definiteness, though a fairly wide range of L would do. With FAIL_t as above, Chernoff bounds give $\sum_{L<t\le k-L} \Pr[\mathrm{FAIL}_t] = o(1)$. With probability $1-o(1)$, no bin on the left nor right side has more than $\ln^2 k$ balls so that the total number of balls on the left and right side is less than $\ln^4 k$. Thus, $\Pr[\mathrm{TREE}]$ is within $o(1)$ of the probability that both sides have less than $\ln^4 k$ balls and that the leftwalk satisfies ESC_L and that the rightwalk satisfies ESC_L^R.

Let $Z_i^* \sim \mathrm{Po}(\lambda)$, $1 \le i \le L$, be independent. Let $Z_i^{R*} \sim \mathrm{Po}(\lambda^R)$, $1 \le i \le L$, be independent. Placing balls into the left and right sides with these distributions, with probability $1-o(1)$, both left and right sides get less than $\ln^4 k$ balls. Both $\Pr[\mathrm{ESC}_L], \Pr[\mathrm{ESC}_L^R]$ are within $o(1)$ of $\Pr[\mathrm{ESC}], \Pr[\mathrm{ESC}^R]$ for the infinite walks and, as they are now independent, $\Pr[\mathrm{ESC}_L \wedge \mathrm{ESC}_L^R]$ would be within $o(1)$ of $\Pr[\mathrm{ESC}]\Pr[\mathrm{ESC}^R]$. For any fixed nonnegative integers $x_1, \ldots, x_L; x_1^R, \ldots, x_L^R$ the probability that $Z_i = x_i$, $1 \le i \le L$ and $Z_i^R = x_i^R$, $1 \le i \le L$ approaches the same probability with the Z_i, Z_i^R replaced by the independent Poissons Z_i^*, Z_i^{R*}. Hence, $\Pr[\mathrm{TREE}]$ is within $o(1)$ of $\Pr[\mathrm{ESC}]\Pr[\mathrm{ESC}^R]$, giving Theorem 1.2.

We next proceed with the central limit theorem Theorem 1.5. The proof of this result is more subtle, and we need to show that both mean and variance are not affected by the conditioning. Consider any fixed nonnegative integers $x_1, \ldots, x_L; x_1^R, \ldots, x_L^R$ so that, with x_i balls in bin i and x_i^R balls in bin i^R the events ESC_L and ESC_L^R both hold. Set $m_L = x_1 + \ldots + x_L$, $m_R = x_1^R + \ldots + x_L^R$ and further assume $m_L < \ln^4 k$ and $m_R < \ln^4 k$. Let M^{**} be the distribution of $\binom{k}{2} - \sum T_x$ where we assume that all remaining balls are placed in the middle bins with the truncated geometric distribution. Thus, the law of M^{**} is the law of M conditioned on some fixed values of m_L and m_R satisfying that $m_L < \ln^4 k$ and $m_R < \ln^4 k$. Let $\mu^{**} = E[M^{**}]$. Then, the following proposition shows that the conditioning does not affect the mean too much:

Proposition 3.1. *For any fixed m_L and m_R satisfying that $m_L < \ln^4 k$ and $m_R < \ln^4 k$, the equality $\mu^{**} - \mu = O(k \ln^4 k)$ holds.*

Proof. Lets call the distributions of M^{**} and M fixededge and unrestricted respectively. There are two differences between these distributions. First, the $m_L + m_R$ balls are explicitly placed in the fixededge distribution. The difference in expectation for any particular ball can be at most k so the total difference for these less than $\ln^4 k$ balls is less than $k \ln^4 k$. For the other balls, the distinction is between the truncated geometric and the unrestricted distribution. Let Y_T, Y_U be the placement of a single ball in these two distributions. Consider the experiment of selecting Y_T from the unrestricted distribution and then reassigning it with the truncated geometric if it did not land in a middle bin. With this linkage we have $Y_T \neq Y_U$ only when the reassignment is made which occurs with probability $O(k^{-1} \ln^2 k)$ for each ball. When it does occur for a given ball, the values are, as always, within k. Hence, the difference in the expectations by reassigning one ball is $O(\ln^2 k)$. The total difference for all (at most k) of these balls is then $O(k \ln^2 k)$. Thus, $\mu^{**} - \mu = O(k \ln^4 k) + O(k \ln^2 k)$, giving Proposition 3.1. □

Next we claim that M^{**} satisfies the asymptotic Gaussian (1.32). We may write $M^{**} = \alpha - \sum T_j^{**}$ where α is a constant which depends on the fixed placement, the sum ranges over those j for which ball j goes into the middle, and T_j^{**} has the distribution of T given by (1.1) conditioned on it being in the middle. We claim M^{**} has variance $\sim \sigma^2$ with σ^2 given by (1.13). For M, M^{**} the variance comes from the independent T_j, T_j^{**} respectively. There are $k-1$ and $\sim k$ terms respectively. The variance of each T_j and each T_j^{**} is $\sim f_2(c)k^2$. An easy way to see this is that $k^{-1}T_j^{**}$ has the asymptotic continuous distribution on $[0,1]$ with density $e^{-cx}/(1-e^{-c})$, which is the asymptotic law of $k^{-1}T_j$ when $k \to \infty$.

From Esseen's Inequality, M^{**} is asymptotically Gaussian with mean μ^{**} and variance $\sim \sigma^2$. Since $\mu - \mu^{**} = O(k \ln^4 k) = o(k^{3/2})$, M^{**} is asymptotically Gaussian with the original μ, σ^2.

Finally, we consider M^*. In the unconditioned placement of balls the probability that either $m_L > \ln^4 k$ or $m_R > \ln^4 k$ was $o(1)$. We are now conditioning on TREE but we have already shown that, in this regime, Pr[TREE] is bounded away from zero. Hence, in the conditioned placement of balls the probability that either $m_L > \ln^4 k$ or $m_R > \ln^4 k$ is still $o(1)$. We have also shown that $\Pr[\text{TREE}] = (1 + o(1)) \Pr[\text{ESC}_L \wedge \text{ESC}_L^R]$. Therefore, conditioning on TREE is equivalent to conditioning on $\text{ESC}_L \wedge \text{ESC}_L^R$, which only changes the law of the balls placed in the left and right bins. Therefore, excluding $o(1)$ probability, M^* is a combination of distributions M^{**}, where the laws of m_L and m_R are the conditional laws given $\text{ESC}_L \wedge \text{ESC}_L^R$. For any fixed value of m_L and m_R, satisfying $m_L < \ln^4 k$ and $m_R < \ln^4 k$, the corresponding M^{**} is asymptotically Gaussian with the same mean and variance. Hence, M^* is as well. This completes the proof of Theorem 1.5 in the case when p is large. □

3.2 The Hard Case

In Sections 3.2–3.8, we study the general case where $pk^{3/2} \to \infty$ and $pk \to 0$. Our arguments can be made considerably simpler when p is not too close to the lower bound $k^{-3/2}$. When we present the general results, we will indicate the simplification when $p = k^{-1.4}$. These simplifications actually work down to $k^{-3/2}$ times a polylog factor.

We split the k bins into left, middle and right sides as given by Definition 4. We carefully choose L so that

$$(kp)^{-2} \ll L \ll k^{-1/2} p^{-1} \tag{3.1}$$

For example, when $p = k^{-1.4}$, we set $L = k^{0.85}$, far away from both bounds of (3.1).

Note that the lower bound of (1.10) on p allows us to do this. Also note that $k^{-1/2}p^{-1} \ll k$ so that

$$L \ll k. \tag{3.2}$$

We set

$$\varepsilon = \frac{pk}{2} = o(1), \tag{3.3}$$

since p is small. A careful analysis of (1.1) gives that

$$\Pr[T_j = i] = \frac{1}{k}(1 + \varepsilon + o(\varepsilon)) \quad \text{for} \quad 1 \le i \le L, \tag{3.4}$$

and

$$\Pr[T_j = i^R] = \frac{1}{k}(1 - \varepsilon + o(\varepsilon)) \quad \text{for} \quad 1 \le i \le L. \tag{3.5}$$

Roughly speaking, each bin on the left side will get $\text{Po}(1+\varepsilon)$ balls, while the bins on the right side will get $\text{Po}(1-\varepsilon)$ balls. It shall turn out that the event TREE is dominated by the events of (1.3) for $1 \le t \le L$ and the events of (1.22) for $1 \le t \le L$.

3.3 Scaling for Small Bias Walks

Mathematical physicists well understand that walks with a bias $\varepsilon = o(1)$ are naturally scaled by time ε^{-2}. Up to time $O(\varepsilon^{-2})$ the walk behaves as if it had zero drift and afterwards the drift takes over. Propositions 3.2–3.3 below investigate the probability of never returning to the starting point, and are quite natural. We write $\Pr^*_\varepsilon$ for the law where each bin $1, 2, \ldots$ receives a Poisson$(1 + \varepsilon)$ number of balls.

Proposition 3.2. *If $\varepsilon \to 0^+$ and $L \to \infty$ is such that $L \gg \varepsilon^{-2}$, then* $\Pr^*_\varepsilon[\mathrm{ESC}_L] \sim 2\varepsilon$.

Proof. As $\Pr^*_\varepsilon[\mathrm{ESC}] \sim 2\varepsilon$ it suffices to show $\Pr^*_\varepsilon[\neg\mathrm{ESC} \wedge \mathrm{ESC}_L] = o(\varepsilon)$.

In the simpler case when $p = k^{-1.4}$ so $\varepsilon = k^{-0.4}/2$ and $L = k^{0.85}$, we can bound $\Pr^*_\varepsilon[\mathrm{ESC}_L]$ by the sum over $t > L$ of $\Pr^*_\varepsilon[Z_1 + \ldots + Z_t < t]$. Here $Z_1 + \ldots + Z_t \sim \mathrm{Po}(t(1+\varepsilon))$. Basic Chernoff bounds show that this probability is so low and drops so fast that summed over all $t > L$ it is $o(\varepsilon)$. Indeed, it is of the form $\exp[-k^{c+o(1)}]$ for some positive constant c. Now we extend the proof to the small p's for which $pk^{3/2} \to \infty$.

Consider the infinite walk and let W be the number of $t \geq L$ such that $Y_t \leq \frac{L\varepsilon}{2}$. Parametrize $t = Lx$. Then

$$\Pr{}^*_\varepsilon[Y_t \leq \frac{L\varepsilon}{2}] \leq \Pr{}^*_\varepsilon[\mathrm{Po}(Lx(1+\varepsilon)) \leq Lx + \frac{Lx\varepsilon}{2}]. \tag{3.6}$$

Basic Chernoff bounds give that this is at most $\exp[-(Lx\varepsilon)^2/8(Lx(1+\varepsilon))] \leq \exp[-L\varepsilon^2 x/16]$. Since $L\varepsilon^2 \gg 1$, this probability is $o(1)$ for every $x \in (0,1)$ fixed. Therefore, by linearity of expectation, $E[W] = o(L)$. Let B be the event that $Y_t = 0$ for some $t \geq L$. Then we claim that $E[W|B] \geq 0.48L$.

Indeed, consider the first such $t \geq L$ with $Y_t = 0$. Conditionally on the history up to time t, we have $\Pr^*_\varepsilon[Y_{t+s} \leq L\varepsilon/2] \geq 0.99$ for all $0 \leq s \leq L(0.49)$. As $E[W] \geq E[W|B]\Pr^*_\varepsilon[B]$, we deduce that $\Pr^*_\varepsilon[B] = o(1)$. Now ESC_L is an increasing event and B is a decreasing event, so that by the FKG inequality

$$\Pr{}^*_\varepsilon[\mathrm{ESC}_L \wedge B] \leq \Pr{}^*_\varepsilon[\mathrm{ESC}_L]\Pr{}^*_\varepsilon[B], \tag{3.7}$$

so that $\Pr^*_\varepsilon[\mathrm{ESC}_L \wedge \neg\mathrm{ESC}] = \Pr^*_\varepsilon[\mathrm{ESC}_L \wedge B] = o(\varepsilon)$. Since, by Theorem 1.3, we have $\Pr{}^*_\varepsilon[\mathrm{ESC}] \sim 2\varepsilon$, Proposition 3.2 follows. □

The next proposition gives a similar result for ESC^R_L. In its statement, we let $\Pr^*_{R,\varepsilon}$ denote the probability law where each bin $1, 2, \ldots, \infty$ receives a Poisson$(1-\varepsilon)$ number of balls.

Proposition 3.3. *If $\varepsilon \to 0^+$ and $L \to \infty$ is such that $L \gg \varepsilon^{-2}$, then* $\Pr^*_{R,\varepsilon}[\mathrm{ESC}^R_L] \sim \varepsilon$.

Proof. Similar to the proof of Proposition 3.2. □

We further require two small extensions:

Corollary 3.4. *Let $\varepsilon \to 0^+$ and $L \gg \varepsilon^{-2}$. Let $\lambda_1, \ldots, \lambda_L$ be such that all $\lambda_i = 1 + \varepsilon + o(\varepsilon)$. Let $Z_i \sim \mathrm{Po}(\lambda_i)$ be independent and consider the leftwalk defined by Definition 2. Then* $\Pr[\mathrm{ESC}_L] \sim 2\varepsilon$.

Proof. For any fixed $\delta > 0$ we can sandwich this model between one in which all $\lambda_i = 1 + \varepsilon(1-\delta)$ and one in which all $\lambda_i = 1 - \varepsilon(1-\delta)$. From Proposition 3.2 we bound $\Pr[\mathrm{ESC}_L]$ between $\sim 2\varepsilon(1-\delta)$ and $\sim 2\varepsilon(1+\delta)$. As δ can be arbitrarily small this gives Corollary 3.4. □

Corollary 3.5. *Let $\varepsilon \to 0^+$ and $L \gg \varepsilon^{-2}$. Let $\lambda^R_1, \ldots, \lambda^R_L$ be such that all $\lambda^R_i = 1 - \varepsilon + o(\varepsilon)$. Let $Z^R_i \sim \mathrm{Po}(\lambda^R_i)$ be independent and consider the rightwalk defined by Definition 3. Then* $\Pr[\mathrm{ESC}^R_L] \sim \varepsilon$.

Proof. Similar as the proof of Corollary 3.4, now using Proposition 3.3 instead of Proposition 3.2. □

3.4 Poisson versus Fixed

It will often be convenient to start with a Poisson number of balls with parameter k, rather than with precisely k balls. Indeed, in the Poisson case, the number of balls per bin are independent Poisson random variables, which is often quite convenient in the analysis. Therefore, sometimes we wish to compare the probability that an event holds when we use a Poisson number of balls to the probability that the event holds when we use a fixed number of balls. In this section, we prove a result that allows us to compare these probabilities, and which will in particular allow us to convert a probability for the Poisson law into a statement for the probability of the event for a fixed number of balls.

We first introduce some notation that allows us to make this comparison. Consider an event A that depends on a nonnegative integer variable X. Let $g(\lambda)$ be $\Pr[A]$ when X has a Poisson distribution with mean λ. Let $f(m)$ be $\Pr[A]$ when $X = m$. (As an important example, drop X balls into bins $1, \ldots, L$ with left-tilt p.) These are related by the equality

$$g(\lambda) = \sum_{m=0}^{\infty} f(m) \Pr[\mathrm{Po}(\lambda) = m]. \tag{3.8}$$

Here we want to go from asymptotics of g to asymptotics of f. We would naturally want to say that $g(m)$ and $f(m)$ are quite close. This is true when f and g are *increasing* or *decreasing.*

We say A is increasing if f, g are increasing; decreasing if f, g are decreasing and monotone if one of those hold. For balls into bins models, an event A is increasing when A keeps on holding when extra balls are added. An event A is decreasing when A^c is increasing. In particular, ESC_L, ESC_L^R are increasing and decreasing respectively. When A is monotone and g is relatively smooth the following result allows us to derive the asymptotics of f from those of g:

Lemma 3.6. *Let $\lambda_1, \lambda_2 \to \infty$ with $\lambda_2 = \lambda_1 + \omega\lambda_1^{1/2}$ where $\omega \to \infty$. Suppose $g(\lambda_1) \sim g(\lambda_2)$. Then:*
If A is increasing, then $f(\lambda_1) \leq (1 + o(1))g(\lambda_2)$.
If A is decreasing, then $f(\lambda_2) \leq (1 + o(1))g(\lambda_1)$.
If A is increasing, then $f(\lambda_2) \geq (1 + o(1))g(\lambda_1)$.
If A is decreasing, then $f(\lambda_1) \geq (1 + o(1))g(\lambda_2)$.

Proof. Assume A is increasing. Truncating (3.8) to $m \geq \lambda_1$ gives $g(\lambda_2) \geq f(\lambda_1) \Pr[\mathrm{Po}(\lambda_2) \geq \lambda_1]$. Chebyschev's Inequality gives that the probability is $1 - o(1)$, giving the first part of Lemma 3.6. Now we show the third part. Calculation gives that for $j \geq \lambda_2$, $Pr[\mathrm{Po}(\lambda_2) = j] \gg Pr[\mathrm{Po}(\lambda_1) = j]$. Now consider the expansion (3.8) for both $\lambda = \lambda_1$ and $\lambda = \lambda_2$. We bound

$$\sum_{m \geq \lambda_2} f(m) \Pr[\mathrm{Po}(\lambda_1) = m] \ll \sum_{m \geq \lambda_2} f(m) \Pr[\mathrm{Po}(\lambda_2) = m] \leq g(\lambda_2) \sim g(\lambda_1), \tag{3.9}$$

so that

$$g(\lambda_1) \sim \sum_{m < \lambda_2} f(m) \Pr[\mathrm{Po}(\lambda_1) = m] \leq f(\lambda_2). \tag{3.10}$$

Statements two and four are similar. □

In application we will deal with situations in which $g(\lambda)$ is asymptotically constant in an interval around λ_0 of width $\gg \sqrt{\lambda_0}$. In that case $f(m) \sim g(m)$ for all m in that interval.

3.5 The probability of TREE in the left, right and middle bins

In this section, we investigate the probabilities of TREE in the left, right and middle bins. The main results are Propositions 3.7, 3.8 and 3.9. In Sections 3.6–3.7, these results, as well as Corollaries 3.4 and 3.5, will be combined to prove Theorem 1.2.

We first use Lemma 3.6 together with the results in Corollaries 3.4 and 3.5 to investigate the probabilities of ESC_L and of ESC_L^R:

Proposition 3.7. *Let $\varepsilon \to 0^+$ and $L \gg \varepsilon^{-2}$. Let $m = L(1+\varepsilon+o(\varepsilon))$. Let $p_1, \ldots, p_L$ all have $p_i = \frac{1}{L}+o(\frac{\varepsilon}{L})$. Let $f(m)$ denote the probability of ESC_L when precisely m balls are placed in bins $1, \ldots, L$ according to this distribution. Then $f(m) \sim 2\varepsilon$.*

Proof. Let $g(\lambda)$ denote the probability when the number of balls is Poisson with mean λ. Corollary 3.4 gives that $g(\lambda) \sim 2\varepsilon$ for any λ for which $\lambda = L(1+\varepsilon+o(\varepsilon))$. But $L\varepsilon \gg \sqrt{L}$ as $L \gg \varepsilon^{-2}$. Thus in this range $f(m) \sim g(m)$ by Lemma 3.6. □

Proposition 3.8. *Let $\varepsilon \to 0^+$ and $L \gg \varepsilon^{-2}$. Let $m = L(1-\varepsilon+o(\varepsilon))$. Let $p_1^R, \ldots, p_L^R$ have all $p_i^R = \frac{1}{L}+o(\frac{\varepsilon}{L})$. Let $f(m)$ denote the probability of ESC_L^R when precisely m balls are placed in bins $1, \ldots, L$ according to this distribution. Then $f(m) \sim \varepsilon$.*

Proof. Similar to that of Proposition 3.7, now using Corollary 3.5 instead of Corollary 3.4. □

The following result will be used to show that most placement of balls which are good on the left and right sides are also good in the middle. This will be a crucial step in order to show that the probability of TREE is asymptotic to the probability of $\mathrm{ESC}_L \wedge \mathrm{ESC}_L^R$.

Proposition 3.9. *Let M balls be placed uniformly in bins $1, \ldots, M$, let Z_i be the number of balls in bin i, and define a walk by $Y_0 = 0$, $Y_i = Y_{i-1} + Z_i - 1$ for $1 \le i \le M$. Set MIN equal to the minimum of Y_i, $0 \le i \le M$. Assume $M, s \to \infty$. Then*

$$\Pr[\mathrm{MIN} < -s\sqrt{M}] = o(1). \tag{3.11}$$

Proof. The proof makes essential use of Lemma 3.6. First suppose all $Z_i \sim \mathrm{Po}(1)$, independent. As $s \to \infty$, $\Pr[Y_M < -s\sqrt{M}] = o(1)$. Let F_i be the event that $Y_i < -s\sqrt{M}$, while $Y_j \ge -s\sqrt{M}$ for all $j < i$. If F_i occurs, then, by the strong Markov property,

$$\Pr[Y_M < -s\sqrt{M} | F_i] \ge \Pr[Y_{M-i} \le 0] \ge c, \tag{3.12}$$

where $c > 0$ uniformly in M, i. Therefore, since the F_i are disjoint and $\bigvee F_i = \{\mathrm{MIN} < -s\sqrt{M}\}$, we obtain that $\Pr[\bigvee F_i] = o(1)$. In the terminology of Lemma 3.6, we have $g(M) = o(1)$ as the total number of balls is Poisson with mean M. Since the event $\{\mathrm{MIN} < -s\sqrt{M}\}$ is decreasing, we also obtain that $f(M) = o(1)$, where $f(M) = \Pr[\mathrm{MIN} < -s\sqrt{M}]$. Indeed, in this simple case, this can also be obtained directly by truncating (3.8), and thus noting that $g(M) \ge f(M) \Pr[\mathrm{Po}(M) \le M]$, so that $f(M) = o(1)$. □

Corollary 3.10. *Assume $M, s \to \infty$ and that $A, B \ge s\sqrt{M}$. Let $M + B - A$ balls be placed uniformly in bins $1, \ldots, M$. Let Z_i be the number of balls in bin i, and define a walk by $Y_0 = A$, $Y_i = Y_{i-1} + Z_i - 1$ for $1 \le i \le M$ so that $Y_M = B$. Set MIN equal to the minimum of Y_i, $0 \le i \le M$. Then*

$$\Pr[\mathrm{MIN} \le 0] = o(1). \tag{3.13}$$

Proof. When $A = B$ this is simply Theorem 3.9 with the walk raised by A. If $A < B$, then ignore the first $B - A$ balls so that now the walk goes from A to A. If $A > B$, then we add $A - B$ fictitious balls so now the walk goes from A to A and then we lower the walk by $A - B$ so it goes from B to B. In both cases we have only increased the probability that the walk hits zero. In both cases we have reduced to the $A = B$ case and so (3.13) holds. □

3.6 A simple upper bound on Pr[TREE]

In this section, we combine Corollaries 3.4 and 3.5 to prove the upper bound on Pr[TREE] in Theorem 1.2. To obtain this upper bound, it will be useful to relate the problem of a fixed number of balls to a Poisson number of balls. This relation is stated in Proposition 3.11, and will also be instrumental in the remainder of the proof of Theorem 1.2, as well as in the proof of Theorem 1.5.

Recall that $M = \binom{k}{2} - \sum_j T_j$. Let $\Pr^*_{\vec\lambda}$ be the law where the number Z_i of balls in bin i is a Poisson random variable with mean λ_i. Later we will choose λ_i appropriately. We write

$$\Lambda = \sum_{i=1}^{k} \lambda_i, \tag{3.14}$$

and, in this section, use the notation $\Pr_{\vec p}$ to denote the law of Z_i, where $k-1$ balls are put into k bins, and the probability to put the j^{th} ball into the i^{th} bin is equal to, for $i = 1, \ldots, k$,

$$p_i = \Pr[T_j = i] = \frac{\lambda_i}{\Lambda}. \tag{3.15}$$

Note that we recover (1.1) when we choose

$$\lambda_i^* = (k-1)\frac{p(1-p)^i}{1-(1-p)^k}, \tag{3.16}$$

for which

$$\Lambda = k - 1. \tag{3.17}$$

Therefore, for this choice, we have that $\Pr_{\vec p} = \Pr$. We will write $\Pr^* = \Pr^*_{\vec\lambda^*}$ when we use $\lambda_1^*, \ldots, \lambda_k^*$ in (3.16). However, later on, it will be convenient to work with more general choices of $\lambda_1, \ldots, \lambda_k$.

The laws of TREE under $\Pr^*_{\vec p}$ and $\Pr_{\vec\lambda}$ are related as follows:

Proposition 3.11. *For every $\lambda_1, \ldots, \lambda_k$, and every random variable X,*

$$E_{\vec p}[I[\text{TREE}]X] = \frac{E^*_{\vec\lambda}[I[\text{TREE}]X]}{\Pr^*_{\vec\lambda}[\text{Po}(\Lambda) = k-1]}, \tag{3.18}$$

where the tilt $p_1, \ldots, p_k$ is related to the parameters of the Poisson distribution $\lambda_1, \ldots, \lambda_k$ by (3.15).

Proof. This result is classical when we note that $\text{TREE} = \text{TREE} \wedge \{\sum_{i=1}^k Z_i = k-1\}$, and the fact that $\sum_{i=1}^k Z_i$ has law $\text{Po}(\Lambda)$. Therefore, the claim is identical to the statement that

$$E_{\vec p}[I[\text{TREE}]X] = E^*_{\vec\lambda}\Big[I[\text{TREE}]X \Big| \sum_{i=1}^{k-1} Z_i = k-1\Big]. \tag{3.19}$$

□

We continue by using Proposition 3.11 to prove a simple bound for the probability of TREE which is useful in the course of the proof:

Proposition 3.12. *For $L = o(k)$, and with λ_i given by (3.16) and p_i by (3.15), then*

$$\Pr[\text{TREE}] \le (1+o(1)) \Pr{}^*[\text{ESC}_L] \Pr{}^*[\text{ESC}_L^R]. \tag{3.20}$$

The same conclusion holds when $\vec\lambda = (\lambda_1, \ldots, \lambda_k)$ satisfies that $\Lambda = k + o(\sqrt{k})$, and p_i is defined by (3.15).

Proof. We first use that for $\vec{p}$ as in (3.15), we have $\Pr_{\vec{p}} = \Pr$, while for $\vec{\lambda}$ as in (3.16), we have $\Pr^* = \Pr^*_{\vec{\lambda}}$.

We use Proposition 3.11 with $X = 1$,

$$\Pr[\text{TREE}] = \frac{\Pr^*[\text{TREE}]}{\Pr^*[\text{Po}(\Lambda) = k-1]}. \tag{3.21}$$

Let μ_L, μ_R be the expected number of balls in the first L and the last L bins respectively. From (3.4–3.5), we obtain that $\mu_L = L(1+\varepsilon+o(\varepsilon))$ and $\mu_R = L(1-\varepsilon+o(\varepsilon))$. Let m_L, m_R be the actual number of balls in the first L and the last L bins respectively. Then we use that

$$\Pr^*[\text{TREE}] \le \sum_{A,B} \Pr^*[\text{ESC}_L \wedge \{m_L = A\}] \Pr^*[\text{ESC}_L^R \wedge \{m_R = B\}] \Pr^*\Big[\sum_{i=L+1}^{k-L-1} Z_i = k-1-A-B\Big], \tag{3.22}$$

since we omit the requirements on the middle bins imposed by TREE. However, uniformly in A, B,

$$\begin{aligned} \Pr^*\Big[\sum_{i=L+1}^{k-L-1} Z_i = k-1-A-B\Big] &= \Pr^*[\text{Po}(\Lambda - \mu_L - \mu_R) = k-1-A-B] \\ &\le \Pr^*\Big[\text{Po}(\Lambda - \mu_L - \mu_R) = \lfloor \Lambda - \mu_L - \mu_R \rfloor\Big] \\ &\sim \frac{1}{\sqrt{2\pi(\Lambda - \mu_L - \mu_R)}}, \end{aligned} \tag{3.23}$$

since $\Pr^*\Big[\text{Po}(\lambda) = l\Big]$ is maximal when $l = \lfloor \lambda \rfloor$.

Since $\Lambda = k + o(\sqrt{k})$ and $\mu_L + \mu_R = o(k)$, we have that

$$\Lambda - \mu_L - \mu_R = k + o(k), \tag{3.24}$$

so that

$$\Pr^*\Big[\text{Po}(\Lambda - \mu_L - \mu_R) = \lfloor \Lambda - \mu_L - \mu_R \rfloor\Big] \sim \frac{1}{\sqrt{2\pi k}} \sim \Pr^*[\text{Po}(\Lambda) = k-1]. \tag{3.25}$$

Performing the sums over A, B gives that

$$\Pr[\text{TREE}] \le (1+o(1)) \Pr^*[\text{ESC}_L] \Pr^*[\text{ESC}_L^R]. \tag{3.26}$$

□

Application of Corollaries 3.4 and 3.5 leads to the following upper bound on Pr[TREE], when we use that $L \gg \varepsilon^{-2}$:

Corollary 3.13. *For $pk^{3/2} \to \infty$ and $pk \to 0$,*

$$\Pr[\text{TREE}] \le (1+o(1)) 2\varepsilon^2. \tag{3.27}$$

The same conclusion holds for $\Pr_{\vec{p}}[\text{TREE}]$ when p_i is given by (3.15), and $\lambda_i = 1+\varepsilon+o(\varepsilon)$ for every $i = o(k)$, and $\lambda_i = 1-\varepsilon+o(\varepsilon)$ for every i such that $k - i = o(k)$, while $\Lambda = k + o(\sqrt{k})$.

□

3.7 The Hard Case: Pr[TREE]

In this section, we complete the proof of Theorem 1.2. By Proposition 3.12, it suffices to prove a lower bound.

We place $k-1$ balls into bins $1, \ldots, k$ with left-tilt p as given by (1.1). Recall that m_L, m_R are the actual number of balls in the first L and the last L bins respectively. Note that m_L, m_R have Binomial distributions with $k-1$ coin flips (the balls) and probability of success $\frac{\mu_L}{k-1}$ and $\frac{\mu_R}{k-1}$ respectively. With foresight, we fix ω such that

$$\omega\sqrt{L} \ll \sqrt{k} \quad \text{and} \quad \omega \ll \varepsilon\sqrt{L} \quad \text{and} \quad \omega \to +\infty. \tag{3.28}$$

The assumed bounds in (3.2) allow us to find such ω. We say that placement of balls is *normal* if $|m_L - \mu_L| < \omega\sqrt{L}$ and $|m_R - \mu_R| < \omega\sqrt{L}$.

We shall naturally refer to a partial placement of balls into the left and right sides, leaving the placement into the middle bins undetermined, as normal if it meets the above criteria. We first prove an extension of Theorem 1.2, which will also be useful in proving Theorem 1.5.

Theorem 3.14. *Assume that $pk^{3/2} \to \infty$ and $pk \to 0$. Then, with probability $\sim 2\varepsilon^2$, the event* TREE *occurs* and *the placement is normal. Consequently,* Pr[TREE] $\sim 2\varepsilon^2$.

Clearly, Theorem 1.2 is a consequence of Theorem 3.14. We first describe a simple example. When $p = k^{-1.4}$ and $L = k^{0.85}$, we set $\omega = k^{0.001}$. Now the probability of a placement not being normal is $o(\varepsilon^2)$ and so may be ignored. We now extend the proof to all p's with $pk^{3/2} \to \infty$:

Proof. Let NICE denote the event $\mathrm{ESC}_L \wedge \mathrm{ESC}_L^R \wedge \{m_L, m_R \text{ normal}\}$. From Propositions 3.7–3.8,

$$\Pr[\mathrm{ESC}_L | m_L = A] \sim 2\varepsilon, \qquad \text{and} \qquad \Pr[\mathrm{ESC}_L^R | m_R = B] \sim \varepsilon, \tag{3.29}$$

for every normal A and B. Thus

$$\begin{aligned} \Pr[\mathrm{NICE}] &= \sum_{A,B \text{ normal}} \Pr[\mathrm{ESC}_L \wedge \mathrm{ESC}_L^R | m_L = A, m_R = B] \Pr[m_L = A, m_R = B] \\ &\sim 2\varepsilon^2 \Pr[m_L, m_R \text{ normal}] \sim 2\varepsilon^2, \end{aligned} \tag{3.30}$$

where we use that $\Pr[m_L, m_R \text{ normal}] \sim 1$.

We effectively need to show that there is "no middle sag," that such paths do not usually hit zero somewhere in the middle. When $p = k^{-1.4}$ and $L = k^{0.85}$ simple Chernoff bounds give that $\Pr[Y_i \leq 0]$ is exceeding small for any middle i. Summing over all middle i the probability that some middle i has $Y_i \leq 0$ is $o(\varepsilon^2)$ and so may be ignored. However, the argument for all p's with $p \gg k^{-3/2}$ is surprisingly delicate. We will show Pr[TREE|NICE] $= 1 - o(1)$. We shall do this in two steps.

We shall first extend the paths from L to a larger L' defined below and then complete the path.

Let L' satisfy

$$k^{-1/2}p^{-1} \ll L' \ll k. \tag{3.31}$$

and let ω' satisfy

$$\omega'\sqrt{L'} \ll \sqrt{k} \quad \text{and} \quad \omega' \ll \sqrt{L'} \quad \text{and} \quad \omega' \to +\infty. \tag{3.32}$$

Let $m_{L'}$, $m_{R'}$ denote the actual number of balls in the first L' and the last L' bins respectively and let $\mu_{L'}$, $\mu_{R'}$ be the expected number of such balls. We say that a placement of balls is L'-normal if $|m_{L'} - \mu_{L'}| < \omega'\sqrt{L'}$ and $|m_{R'} - \mu_{R'}| < \omega'\sqrt{L'}$.

Let NICE$'$ denote the event $\mathrm{ESC}_{L'} \wedge \mathrm{ESC}_{L'}^R \wedge \{m_{L'}, m_{R'}\ L' - \text{normal}\}$. The arguments yielding (3.30) hold for these values. That is, NICE and NICE$'$ each hold with probability $\sim 2\varepsilon^2$.

Corollaries 3.4–3.5 give that $\mathrm{ESC}_L \wedge \mathrm{ESC}_L^R$ has probability $\sim 2\varepsilon^2$. Thus the probability that ESC_L and ESC_L^R occur, but that m_L, m_R are not both normal is $o(\varepsilon^2)$. For NICE$'$ to hold and NICE to fail these would all need occur. Hence $\Pr[\mathrm{NICE}' \wedge \mathrm{NICE}] \sim 2\varepsilon^2$. That is,

$$\Pr[\mathrm{NICE}'|\mathrm{NICE}] = 1 - o(1). \tag{3.33}$$

Now we want to show $\Pr[\mathrm{TREE}|\mathrm{NICE}'] = 1 - o(1)$. It suffices to show this conditioning on explicit normal values $m_{L'}, m_{R'}$. Set $A = 1 + m_{L'} - L'$ and $B = L' - m_{R'}$. We now consider the middle bins as those not amongst the first or last L' bins. In the middle we are placing balls with left-tilt p and considering a walk that begins at A and ends at B. Our normality assumption and (3.31) imply

$$A \sim B \sim L'\varepsilon \gg \sqrt{k}. \tag{3.34}$$

We claim with probability $1 - o(1)$, the walk will not hit zero. Removing the tilt moves balls to the right, which makes it more likely that the walk does hit zero. Therefore, it suffices to show this when the balls are placed with uniform probability in each bin. This is precisely Corollary 3.10, where $M = k - 2L'$. Note that our selection (3.31) of L' has assured $M \sim k$ and $A, B \gg \sqrt{k}$, so that the conditions of Corollary 3.10 are met.

We conclude that conditioning on NICE$'$ and any particular normal $m_{L'}, m_{R'}$ the event TREE holds with probability asymptotic to one. Hence

$$\Pr[\mathrm{TREE}|\mathrm{NICE}'] = 1 - o(1). \tag{3.35}$$

Combining this with (3.33) gives

$$\Pr[\mathrm{TREE}|\mathrm{NICE}] = 1 - o(1). \tag{3.36}$$

Combined with (3.30), $\Pr[\mathrm{TREE} \wedge \mathrm{NICE}] \sim 2\varepsilon^2$, the first part of Theorem 3.14. The upper bound (3.27) completes the proof. □

3.8 The Hard Case: Asymptotic Gaussian

In this section, we prove Theorem 1.5. This proof relies on the rewrite in Proposition 3.11. We therefore only need to investigate the law of M under the measure $\Pr^*_{\vec{\lambda}}$ for an appropriate choice of $\vec{\lambda}$. For this, we note that we can rewrite

$$\sum_{j=1}^{k-1} T_j = \sum_{i=1}^{k} iZ_i, \tag{3.37}$$

where Z_i is the number of balls placed in the i^{th} bin. Recall that Z_i is $\mathrm{Po}(\lambda_i)$, where λ_i is defined in (3.16).

Define, for every $t \in \mathbb{R}$,

$$\lambda_{i,t} = \lambda_i e^{t(i-k/2)}, \tag{3.38}$$

and write $\Pr^*_t = \Pr^*_{\vec{\lambda}_t}$ for the law of the process when Z_i is $\mathrm{Po}(\lambda_{i,t})$ for all $i = 1, \ldots, k$. We also write E^*_t for the expectation w.r.t. $\Pr^*_t$. Note that $E^* = E^*_0$. We also note that in this case, with

$$p_{i,t} = \frac{\lambda_{i,t}}{\Lambda_t}, \qquad \text{where} \qquad \Lambda_t = \sum_{i=1}^{k} \lambda_{i,t}, \tag{3.39}$$

we have that

$$p_{i,t} = \frac{p_t(1-p_t)^{i-1}}{1-(1-p_t)^k}, \tag{3.40}$$

where

$$1 - p_t = (1-p)e^t, \qquad \text{so that} \qquad p_t = p + (1-p)(e^t - 1). \tag{3.41}$$

Therefore, the probabilities for the different bins still follow the tilted distribution in (1.1).

The proposition below gives an explicit equality for the moment generating function of $M - E^*[M]$ conditionally on TREE:

Proposition 3.15. *The equality*

$$E_0^*(e^{-t(M-E^*[M])}|\text{TREE}) = e^{\sum_{i=1}^k \lambda_i \left[e^{t(i-k/2)}-1-t(i-k/2)\right]} \frac{\Pr_t^*[\text{TREE}]}{\Pr_0^*[\text{TREE}]} \tag{3.42}$$

holds.

Proof. When TREE holds, then

$$Z_1 + \ldots + Z_k = k-1. \tag{3.43}$$

Therefore, when TREE holds, and using (3.37),

$$M = \binom{k}{2} - \sum_{j=1}^{k-1} T_j = -\sum_{i=1}^{k} (i-k/2) Z_i. \tag{3.44}$$

Similarly, since

$$\sum_{i=1}^{k} E^*[Z_i] = \sum_{i=1}^{k} \lambda_i = k-1, \tag{3.45}$$

we also have that

$$E^*[M] = E^*\Big[\binom{k}{2} - \sum_{j=1}^{k-1} T_j\Big] = -\sum_{i=1}^{k}(i-k/2)E^*[Z_i], \tag{3.46}$$

so that we arrive at the equality that, when TREE holds,

$$M - E^*[M] = -\sum_{i=1}^{k}(i-k/2)(Z_i - E^*[Z_i]). \tag{3.47}$$

Therefore, we can write out

$$\begin{aligned}
&E_0^*(e^{-t(M-E^*[M])}|\text{TREE}) \qquad (3.48)\\
&= \frac{1}{\Pr_0^*[\text{TREE}]} e^{-t\sum_{i=1}^k (i-k/2)\lambda_i} \sum_{\vec{z}\in\mathbb{N}^k} e^{t\sum_{i=1}^k (i-k/2)z_i} \prod_{i=1}^{k} e^{-\lambda_i}\frac{\lambda_i^{z_i}}{z_i!} I[\text{TREE}]\\
&= \frac{1}{\Pr_0^*[\text{TREE}]} e^{-t\sum_{i=1}^k (i-k/2)\lambda_i} \sum_{\vec{z}\in\mathbb{N}^k} \prod_{i=1}^{n} e^{-\lambda_i}\frac{[e^{t(i-k/2)}\lambda_i]^{z_i}}{z_i!} I[\text{TREE}]\\
&= \frac{1}{\Pr_0^*[\text{TREE}]} e^{-t\sum_{i=1}^k (i-k/2)\lambda_i} e^{\sum_{i=1}^k(\lambda_{i,t}-\lambda_i)} \sum_{\vec{z}\in\mathbb{N}^n} \prod_{i=1}^{n} e^{-\lambda_{i,t}}\frac{\lambda_{i,t}^{z_i}}{z_i!} I[\text{TREE}]\\
&= e^{\sum_{i=1}^k \lambda_i[e^{t(i-k/2)}-1-t(i-k/2)]}\frac{\Pr_t^*[\text{TREE}]}{\Pr_0^*[\text{TREE}]}.
\end{aligned}$$

□

We now formulate a corollary of Proposition 3.15. It is statement, we write $\Pr_t$ for the measure where the tilt is given by (3.39), and where $\lambda_{i,t}$ is given by (3.38).

Corollary 3.16. *Let $t_k = tk^{-3/2}$ and assume that p satisfies $pk^{3/2} \to \infty$ and $p \ll k^{-1}$. Then, for every $t \in \mathbb{R}$ fixed,*

$$\Pr_{t_k}[\mathrm{TREE}] = 2\varepsilon^2(1 + o(1)). \tag{3.49}$$

Consequently, for Pr *and conditionally on* TREE, *the random variable $k^{-3/2}(M - E^*[M])$ converges weakly to a normal distribution with variance $\frac{1}{12}$.*

We conclude from Corollary 3.16 that we obtain the central limit theorem 'for free' from the scaling of the probability of TREE, which holds for all $t \in \mathbb{R}$ fixed. As a consequence, we obtain that Theorem 1.5 holds. Therefore, we are left to prove Corollary 3.16.

Proof. The equality in (3.49) follows by Theorem 3.14, using the extensions in Corollaries 3.4–3.5, together with the fact that $p_{t_k} = p(1 + o(1))$ under the assumptions in Corollary 3.16. Therefore, we obtain from Theorem 3.14 that

$$\Pr_{t_k}[\mathrm{TREE}] \sim \Pr[\mathrm{TREE}]. \tag{3.50}$$

To prove the asymptotic normality of $k^{-3/2}(M - E^*[M])$ conditionally on TREE, we start by using Proposition 3.11, which implies that

$$E[e^{-t_k(M-E^*[M])} I[\mathrm{TREE}]] = \frac{E^*[e^{-t_k(M-E^*[M])} I[\mathrm{TREE}]]}{\Pr^*[\mathrm{Po}(\Lambda) = k-1]}. \tag{3.51}$$

Then we use Proposition 3.15 to obtain that

$$E\big[e^{-t_k(M-E^*[M])}|\mathrm{TREE}\big] = e^{\sum_{i=1}^k \lambda_i [e^{t_k(i-k/2)} - 1 - t_k(i-k/2)]} \frac{\Pr_{t_k}[\mathrm{TREE}]}{\Pr[\mathrm{TREE}]} \frac{\Pr^*[\mathrm{Po}(\Lambda_{t_k}) = k-1]}{\Pr^*[\mathrm{Po}(\Lambda) = k-1]}. \tag{3.52}$$

Furthermore, using that $\lambda_i - 1 = O(\varepsilon)$ uniformly in i, we can compute the exponent of the first factor as

$$\begin{aligned}
\sum_{i=1}^k \lambda_i [e^{t_k(i-k/2)} - 1 - t_k(i-k/2)] &= \frac{1}{2}\sum_{i=1}^k \lambda_i (i-k/2)^2 t_k^2 + O\Big(\sum_{i=1}^k |i-k/2|^3 t_k^3\Big) \\
&= \frac{1}{2}\sum_{i=1}^k (i-k/2)^2 t_k^2 + O\Big(\sum_{i=1}^k |i-k/2|^3 t_k^3\Big) + O(\varepsilon) \\
&= \sum_{i=0}^{k/2} (i-k/2)^2 t^2 k^{-3} + O(k^{-1/2} + \varepsilon) = \frac{t^2}{24} + O(k^{-1/2} + \varepsilon).
\end{aligned} \tag{3.53}$$

Moreover, since

$$\begin{aligned}
\Lambda_{t_k} &= \sum_{i=1}^k \lambda_{i,t_k} = \sum_{i=1}^k \lambda_i + \sum_{i=1}^k \lambda_i [e^{t_k(i-k/2)} - 1] \\
&= k - 1 + \sum_{i=1}^k \lambda_i t_k (i-k/2) + O\Big(\sum_{i=1}^k \lambda_i \big[t_k(i-k/2)\big]^2\Big) \\
&= k - 1 + \sum_{i=1}^k t_k(i-k/2) + \sum_{i=1}^k (\lambda_i - 1) t_k (i-k/2) + O\Big(\sum_{i=1}^k \lambda_i \big[t_k(i-k/2)\big]^2\Big) = k + o(k^{1/2}),
\end{aligned} \tag{3.54}$$

the local central limit theorem for $\Pr^*[\mathrm{Po}(\Lambda_{t_k}) = k-1]$ remains valid and we obtain

$$\Pr^*[\mathrm{Po}(\Lambda_{t_k}) = k-1] \sim \Pr^*[\mathrm{Po}(\Lambda) = k-1]. \tag{3.55}$$

Together with (3.50), we conclude that

$$E_0\Big(e^{-\frac{t}{k^{3/2}}(M-E^*[M])}|\mathrm{TREE}\Big) \sim e^{t^2/24}. \tag{3.56}$$

Since $e^{t^2/24}$ is the moment generating function of a Gaussian random variable with mean 0 and variance 1/12, this completes the proof of Corollary 3.16. $\square$

Acknowledgement

This project was initiated during visits of both authors to Microsoft Research in July 2004. We thank Benny Sudakov and Michael Krivelevich for useful discussions at the start of this project. J.S. thanks Nitin Arora for help with the asymptotics of $C(k,l)$. The work of RvdH was supported in part by the Netherlands Organisation for Scientific Research (NWO), and was performed in part while visiting the Mittag-Leffler Institute in November 2004.

References

[1] N. Arora and J. Spencer. Generating Random Connected Graphs. In preparation.

[2] B. Bollobás. The evolution of sparse graphs. In *Graph Theory and Combinatorics (Cambridge 1983)* Academic, New York, 248–289, (1984).

[3] E. Bender, E. Rodney Canfield and B.D. McKay. The Asymptotic Number of Labeled Connected Graphs with a Given Number of Vertices and Edges. *Random Structures & Algorithms* **2**: 127-169, (1990).

[4] A. Coja-Oghlan, C. Moore and V. Sanwalani. Counting Connected Graphs and Hypergraphs via the Probabilistic Method. Proc. 8th Intl. Workshop on Randomization and Computation (RANDOM '04), 322–333.

[5] P. Erdős and A. Renyi. On the evolution of random graphs *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5**: 17–61, (1960).

[6] R. van der Hofstad and J. Spencer. The number of edges and vertices of the giant component. In preparation.

[7] G. Louchard. The Brownian excursion: a numerical analysis. *Computers and Mathematics with Applications* **10**: 413–417, (1984).

[8] T. Łuczak. On the number of sparse connected graphs. *Random Structures & Algorithms* **2**: 171–173, (1990).

[9] B. Pittel and N.C. Wormald Counting connected graphs inside-out. *J. Combin. Theory Ser. B* **93**(2): 127–172, (2005).

[10] A. Rényi. On connected graphs I. *Publ. Math. Inst. Hungarian Acad. Sci.,* **4**(159): 385-388, (1959).

[11] J. Spencer. Enumerating Graphs and Brownian Motion. *Communications on Pure and Appl. Math.* **50**: 293–296, (1997).

[12] E.M. Wright. The number of connected sparsely edges graphs. *Journ. of Graph Theory* **1**: 317–330, (1977).

[13] E.M. Wright. The number of connected sparsely edges graphs. III. *Journ. of Graph Theory* **4**: 393–407, (1980).