A "TRIPLE BIRTHDAY PROBLEM" WITH AN APPLICATION TO TIME TO FAILURE IN LATIN SQUARE GENERATION

by

Thomas M. Sellke
Purdue University

Technical Report #92-39

Department of Statistics
Purdue University

August 1992

★

A "TRIPLE BIRTHDAY PROBLEM" WITH AN APPLICATION TO TIME TO FAILURE IN LATIN SQUARE GENERATION

by

Thomas M. Sellke
Purdue University

Abstract

Consider an $n \times n \times n$ cube divided into n^3 unit cubes. Sample unit cubes with replacement (according to the uniform distribution on the n^3 unit cubes) until a cube is obtained whose perpendicular projection onto some side of the big cube is the same as that of some previously chosen cube. Let T be the number of unit cubes that you have to sample until this happens. This paper derives an asymptotic ($n \to \infty$) formula for $P\{T > k+1 | T > k\}$. A corollary is that $P\{n^{-1}T > t\} \overset{n\to\infty}{\longrightarrow} \exp(-\frac{3}{2}t^2)$ for $t > 0$.

1. Introduction.

Suppose an $n \times n \times n$ cube is divided up into n^3 unit cubes (or "boxes") in the obvious way. Randomly sample unit cubes *with* replacement (according to the uniform distribution on the n^3 cubes) until you get a unit cube whose perpendicular projection onto some side of the big cube is the same as that of some previously chosen cube. If T is the number of unit cubes you have to sample until this happens, what is the distribution of T?

Here is why this is a "triple birthday problem." If one looks only at projections onto the n^2 squares of a fixed face of the big cube, the time N until a match occurs (i. e., until some square is obtained twice) is precisely the match time for a birthday problem with n^2 equally likely possible birthdays. By the standard birthday problem argument

$$P\{N > k\} \;=\; \prod_{\ell\,=\,1}^{k-1} \frac{n^2 - \ell}{n^2} \;, \tag{1.1}$$

and using the approximation $\log(1-x) \;\approx\; -x$, one gets

$$\log P\{N > k\} \;\approx\; -\frac{1}{n^2} \sum_{\ell\,=\,1}^{k-1} \ell \;=\; -\frac{k(k-1)}{2n^2} \;\approx\; -\frac{k^2}{2n^2}$$

when n is large. Now let N_1, N_2, and N_3 be the match times for the three distinct projections of unit cubes onto faces of the big cube. Then

$$T \;=\; \min(N_1, N_2, N_3) \;,$$

so that T is the time of the first match in any of three identical but dependent birthday problems.

Even though N_1, N_2, and N_3 are dependent, intuition suggests that they may be approximately independent for large n. Thus, one expects

$$\begin{aligned} P\{T > k\} \;&=\; P\{N_1 > k, N_2 > k, N_3 > k\} \\ &\stackrel{?}{\approx}\; P\{N_1 > k\}^3 \;\approx\; \exp\left(-\frac{3k^2}{2n^2}\right). \end{aligned} \tag{1.2}$$

This note will derive approximations (with explicit error bounds) for $P\{T > k+1 \mid T > k\}$ which in turn imply (1. 2) for large n. Our final estimate for $P\{T > k+1 \mid T > k\}$ is

$$1 - n^{-3}\left[(3n-2)k - 2\binom{k}{2}\frac{3}{n+2} + \binom{k}{3}d_n + 36\binom{k}{3}\frac{(n-1)^3}{n^6(n+2)}\frac{P\{T > k-2\}}{P\{T > k\}}\right] \quad , \tag{1.3}$$

where

$$d_n \;=\; \frac{6(n-1)}{n^4 + 2n^3 - 6n^2 - 14n + 14} \quad .$$

If the conditional probabilities $P\{T > k+1 \mid T > k\}$ are estimated first for $k = 1$, then $k = 2$, then $k = 3$, etc. , then the ratio [whose reciprocal appears in (1.3)]

$$\frac{P\{T > k\}}{P\{T > k-2\}} \;=\; P\{T > k-1 \mid T > k-2\}P\{T > k \mid T > k-1\}$$

can be approximated by the product of previous estimates.

Here is the application to Latin square generation. Start with an $n \times n$ square divided up into n^2 unit squares. Randomly sample unit squares *without* replacement. Each time a square is selected, randomly choose a number from $\{1, 2, \cdots, n\}$ (with replacement) and put it into the chosen unit square. Then check to see whether the Latin square property has been violated, i. e. , whether the number in the last square equals a number in another square of the same row or column. Let $\widetilde{T}$ be the time until failure of the Latin square property. Then with N_1 and T as above,

$$P\{\widetilde{T} > k\} \;=\; P\{T > k \mid N_1 > k\} \;=\; \frac{P\{T > k\}}{P\{N_1 > k\}} \quad . \tag{1.4}$$

The argument for (1.4) is as follows. The cube face for the N_1 projection corresponds to the $n \times n$ square in the Latin square story. The other cube dimension corresponds to the number written into a unit square. Conditioning on $N_1 > k$ corresponds to the requirement that squares are sampled without replacement in the Latin square story. With these correspondences between the problems, $T > k$ means that the Latin square property has not yet been violated.

Since we have an exact expression for $P\{N_1 > k\}$, an approximation for $P\{T > k\}$ can be plugged into (1.4) to yield an approximation for $P\{\widetilde{T} > k\}$. If $P\{T > k\} \approx \exp\left(-\frac{3k^2}{2n^2}\right)$, then

$$P\{\widetilde{T} > k\} \;\approx\; \exp\left(-\frac{k^2}{n^2}\right) \quad .$$

If the Latin square story is modified to have the sampling of squares done *with* replacement, and if "failure" occurs when a square is chosen a second time as well as when the Latin square property is violated by the number in a new square, then the time until failure is precisely T.

2. A Formula for $P\{T > k+1 \mid T > k\}$ in Terms of "Expected Exclusion Overlap".

Let

$$S_n \;=\; \{(a_1, a_2, a_3) : a_i \in \{1, 2, \cdots, n\}\} \;\; .$$

Let $X_1, X_2, \cdots$ be independent and uniformly distributed on S_n. In terms of the X_i's,

$$T \;=\; \inf\{k : X_k \text{ shares at least two coordinates with some } X_j,\, j < k\} \;\; .$$

From now on, the elements of S_n will be referred to as boxes, since I find that this terminology fits nicely with the geometric problem description of the introduction. Also, let $G_k \;=\; \{T > k\}$, with the mnemonic that G_k means that $X_1, \cdots, X_k$ are a "good" choice of boxes for Latin square generation.

Suppose now that $G_k \;=\; \{T > k\}$ has occurred. The next box chosen, namely X_{k+1}, is equally likely to be any of the n^3 boxes in S_n. Let $M_k \;=\; M_k(X_1, \cdots, X_k)$ be the number of choices for box X_{k+1} which would cause $G_{k+1} \;=\; \{T > k+1\}$ to fail. Then obviously

$$P\{G_{k+1} \mid X_1, \cdots, X_k\} \;=\; \frac{n^3 - M_k}{n^3}$$

and

$$P\{G_{k+1} \mid G_k\} \;=\; \frac{n^3 - E(M_k \mid G_k)}{n^3} \;\; . \tag{2.1}$$

Thus, the problem of estimating $P\{T > k+1 \mid T > k\} \;=\; P\{G_{k+1} \mid G_k\}$ reduces to estimating $E(M_k \mid G_k)$

There are $1 + 3(n-1) \;=\; 3n - 2$ boxes which share at least two coordinates with X_1. For example, if $X_1 \;=\; (1,1,1)$, then these boxes are $(1,1,1)$ itself and all boxes of the form $(j,1,1)$, $(1,j,1)$, or $(1,1,j)$, $j \neq 1$. Let us describe this by saying that X_1 *excludes* $3n-2$ boxes as choices for X_{k+1} which are compatible with event G_{k+1}. Thus,

$$M_k \;\leq\; (3_n - 2)k \;, \tag{2.2}$$

since each of $X_1, \ldots, X_k$ excludes $3n-2$ boxes. However, a box may be excluded by one X_i, by two X_i's, or by three X_i's when G_k occurs. One could say that "exclusion overlap" occurs if a box is excluded by two or three X_i's.

Write $X_i \sim X_j$ if boxes X_i and X_j have exactly one coordinate in common. (Despite what the notation may suggest, " $\sim$ " is clearly not an equivalence relation, since $X_1 \sim X_2$ and $X_2 \sim X_3$ do not imply $X_1 \sim X_3$.) If $X_1 \sim X_2$, then there are exactly two boxes excluded by both X_1 and X_2. For example, if $X_1 = (1,1,1)$ and $X_2 = (1,2,2)$, then boxes $(1,1,2)$ and $(1,2,1)$ are excluded by both X_1 and X_2 .

At the risk of some confusion, write $X_i \sim X_j \sim X_l$ if each pair of boxes from among the three share one coordinate, *which is a different coordinate for each different pair.* Thus, $X_1 \sim X_2 \sim X_3$ would occur if $X_1 = (1,1,1)$, $X_2 = (1,2,2)$, and $X_3 = (3,1,2)$, since X_1 and X_2 share the first coordinate, X_1 and X_3 share the second coordinate, and X_2 and X_3 share the third coordinate. The reason for paying attention to this quirky event $X_i \sim X_j \sim X_l$ is that it means that X_i, X_j, and X_l all exclude the same box, which would be (1,1,2) in the last example.

A simple accounting argument shows

$$M_k \;=\; (3n-2)k - 2\sum_{i<j} I\{X_i \sim X_j\} + \sum_{i<j<l} I\{X_i \sim X_j \sim X_l\} \;. \tag{2.3}$$

Since the boxes $X_1, \ldots, X_i$ are exchangeable, given G_k,

$$\begin{aligned} E(M_k|G_k) \;=\; (3n-2)k \;& - 2\binom{k}{2} P\{X_1 \sim X_2|G_k\} \\ & + \binom{k}{3} P\{X_1 \sim X_2 \sim X_3|G_k\} \;. \end{aligned} \tag{2.4}$$

Plugging (2.4) into (2.1) gives

$$P\{G_{k+1}|G_k\} \;=\; 1 - n^{-3}\left[(3n-2)k - 2\binom{k}{2} P\{X_1 \sim X_2|G_k\} + \binom{k}{3} P\{X_1 \sim X_2 \sim X_3|G_k\}\right] \;. \tag{2.5}$$

The next section will obtain crude bounds on the conditional probabilities in (2.5) which are sufficient to prove the weak convergence

$$P\left\{\frac{T}{n} \;>\; t\right\} \;\stackrel{n\to\infty}{\longrightarrow}\; e^{-\frac{3}{2}t^2}, \qquad t > 0 \;. \tag{2.6}$$

Later sections will derive much better estimates for these probabilities.

3. Weak Convergence of $\frac{T}{n}$

The unconditional probability of $X_1 \sim X_2$ is

$$P\{X_1 \sim X_2\} \;=\; \frac{3(n-1)^2}{n^3} \;<\; \frac{3}{n} \;. \tag{3.1}$$

(Argument: For any possible choice of X_1, there are $(n-1)^2$ choices for X_2 which agree with X_1 in the first coordinate but not in the second or third coordinates.)

We immediately get the bound

$$P\{X_1 \sim X_2 | G_k\} \;<\; P\{G_k\}^{-1}\, \frac{3}{n} \;. \tag{3.2}$$

Likewise,

$$P\{X_1 \sim X_2 \sim X_3\} \;=\; n^3 \times 3! \times \left(\frac{n-1}{n^3}\right)^3 \;=\; 6\,\frac{(n-1)^3}{n^6} \;<\; \frac{6}{n^3} \;. \tag{3.3}$$

(Argument: For $X_1 \sim X_2 \sim X_3$ with common exclusion box (1,1,1), the forms of X_1, X_2, and X_3 must be $(i,1,1)$, $(1,j,l)$, and $(1,1,l)$ in some order, with $i \neq 1$, $j \neq 1$, and $l \neq 1$. This has probability $3!(n-1/n^3)^3$. Multiply by n^3 because there are n^3 possible "triple exclusion" boxes.) Thus

$$P\{X_1 \sim X_2 \sim X_3 | G_k\} < P\{G_k\}^{-1}\, \frac{6}{n^3} \;. \tag{3.4}$$

Now, let's get a crude lower bound on $P\{G_k\}$, which will give a crude upper bound on $P\{G_k\}^{-1}$. Plugging (2.2) into (2.1) yields

$$P\{G_{k+1} | G_k\} \;\geq\; \frac{n^3 - (3n-2)k}{n^3} \;. \tag{3.5}$$

Since $ln(1-x) \geq -(2\ ln\ 2)x > -2x \qquad for \quad 0 < x < \frac{1}{2}$,

$$ln\, P\{G_k\} \;=\; \sum_{l=1}^{k-1} ln\, P\{G_{l+1} | G_l\} \;\geq\; -\frac{6}{n^2} \sum_{l=1}^{k-1} l \;\geq -\frac{3k^2}{n^2} \qquad for \quad k < \frac{n^3}{6} \tag{3.6}$$

and so

$$P\{G_k\}^{-1} \;\leq\; exp\left(3\,\frac{k^2}{n^2}\right) \qquad for \quad k < \frac{n^3}{6} \;. \tag{3.7}$$

Plugging (3.7) into (3.2) and (3.4) and then using the resulting bounds in (2.5) yields for $k < \frac{n^3}{6}$ that

$$\begin{aligned} P\{G_{k+1} | G_k\} \;&\leq\; 1 - n^{-3}\left[(3n-2)k - 2\binom{k}{2}\frac{3}{n}\, exp\left(3\frac{k^2}{n^2}\right)\right] \\ &\leq\; 1 - \frac{(3n-2)k}{n^3} \;+\; \frac{3k^2}{n^4}\; exp\left(3\frac{k^2}{n^2}\right) \end{aligned} \tag{3.8}$$

and

$$P\{G_{k+1}|G_k\} \geq 1 - n^{-3}\left[(3n-2)k + \binom{k}{3}\frac{6}{n^3}\ exp\left(3\frac{k^2}{n^2}\right)\right]$$
$$\geq 1 - \frac{3k}{n^2} - \frac{k^3}{n^6}\ exp\left(3\frac{k^2}{n^2}\right) \tag{3.9}$$

Now fix $t > 0$. If we multiply over $k \leq tn$ in (3.8) and in (3.9), then easy estimates of the logarithms give upper and lower bounds on $ln\ P\{T > tn\}$ which in turn imply the weak convergence (2.6).

4. Good Estimates of $P\{T > k+1|T > k\}$

The previous section showed that the conditional probability terms in (2.5) are asymptotically negligible as far as weak convergence of $\frac{T}{n}$ is concerned. This section will obtain reasonably good approximations for these conditional probabilities, and Section 5 will do better yet. It will be seen that the events $X_1 \sim X_2$ and $X_1 \sim X_2 \sim X_3$ are approximately independent of the event $G_k = \{T > k\}$. Let's start with several basic lemmas.

Lemma 4.1: Given event G_k, each of the n^3 boxes in S_n has probability exactly $\frac{k}{n^3}$ of being one of $X_1, ..., X_k$.

Proof: Just use symmetry. ∎

Lemma 4.2 For $2 < k \leq \frac{n^2}{12}$, $P\{G_k|G_{k-2}\} > \frac{1}{2}$.

Proof: For $k \leq \frac{n^2}{12}$, (2.2) implies

$$M_k \leq \frac{n^3}{4}\ .$$

Plugging this into (2.1) yields

$$P\{G_{k+1}|G_k\} \geq \frac{3}{4} \qquad for \qquad k \leq \frac{n^2}{12}\ .$$

Since $k \leq \frac{n^2}{12}$ implies $k-1 \leq \frac{n^2}{12}$ and $k-2 \leq \frac{n^2}{12}$,

$$P\{G_k|G_{k-2}\} = P\{G_{k-1}|G_{k-2}\}\ P\{G_k|G_{k-1}\} \geq \frac{9}{16} > \frac{1}{2}\ ,$$

provided $k \leq \frac{n^2}{12}$. ∎

Remark: From now on, we will focus attention on $k \leq \frac{n^2}{12}$, since Lemma 4.2 applies to such $k's$. Note from (2.6) that $P\{T > \frac{n^2}{12}\}$ is very small when n is large.

First we need some more notation. Define the events

$$A = \{X_1 = (1,1,1) \text{ and } X_2 = (1,2,2)\} \tag{4.1}$$

$$B = \{X_1 = (1,1,1) \text{ and } X_2 = (2,2,2)\} \ . \tag{4.2}$$

By the same sort of symmetry which implies Lemma 4.1,

$$P\{X_1 \sim X_2 | G_k\} = 3n^3(n-1)^2 \ P\{A|G_k\} \tag{4.3}$$

since, given $\{X_1 \sim X_2\} \cap G_k$, X_1 and X_2 are equally likely to be any of the $3n^3(n-1)^2$ possible ordered pairs of boxes in S_n for which $X_1 \sim X_2$. Likewise

$$P\{\text{not } X_1 \sim X_2 | G_k\} = n^3(n-1)^3 \ P\{B|G_k\} \ . \tag{4.4}$$

Also, it is obvious that, unconditionally,

$$P(A) = P(B) = n^{-6} \ . \tag{4.5}$$

Let G^*_{k-2} be the event that no two of $X_3, ..., X_k$ share two or more coordinates. Thus, G^*_{k-2} is just like G_{k-2}, except that it refers to $X_3, ..., X_k$ rather than $X_1, ..., X_{k-2}$. Also, denote $X_3, ..., X_k$ by $X^*_3, ..., X^*_k$, so that a general X^*_i refers to one of the X_i's other than X_1 or X_2.

Let R_{12} be the set of eight boxes all of whose coordinates are either 1 or 2. Let Q_{12} be the set of $12(n-2)$ boxes which have exactly two coordinates which are 1 or 2, like (1,7,1) or (3,1,2), for example.

Proposition 4.6 For $k \leq \frac{n^2}{12}$,

$$\left| P\{A|G_k\} - P\{B|G_k\} \right| < \frac{4k}{n^2} P(A) = \frac{4k}{n^8} \ . \tag{4.6}$$

Proof. I claim that

$$\left| P\{G_k | G^*_{k-2} \cap A\} - P\{G_k | G^*_{k-2} \cap B\} \right| < \frac{2k}{n^2} \ . \tag{4.7}$$

Explanation: If no X^*_i's are in R_{12} or Q_{12}, then G_k will be true for both A and B. If (2,2,7), say, is one of the X^*_i's, then G_k might be true under A but G_k will be violated under B. By Lemma 4.1, the probability that one of the X^*_i's is (2,2,7) is exactly $\frac{k-2}{n^3}$. The number of boxes like (2,2,7) whose choice as an X^*_i box would make G_k false under B but *possibly* true under

A is $2(n-2)+2 = 2n-2 < 2n$. The probability, given G^*_{k-2}, that at least one of these boxes is an X^*_i is $\leq \frac{(2n-2)(k-2)}{n^3} < \frac{2k}{n^2}$, by Lemma 4.1. Thus, (using the fact that events A and B are independent of G^*_{k-2})

$$P\{G_k|G^*_{k-2} \cap A\} - P\{G_k|G^*_{k-2} \cap B\} < \frac{2k}{n^2} .$$

Likewise, there are precisely $2(n-2)$ possible choices for X^*_i boxes which would make G_k false under A but possibly true under B, which gives us the other direction in (4.7).

Multiply through in (4.7) by

$$P\{G^*_{k-2} \cap A\} = P\{G^*_{k-2} \cap B\} = P\{G_{k-2}\}P\{A\} ,$$

and divide through by $P\{G_k\}$ to get

$$\left| P\{A|G_k\} - P\{B|G_k\} \right| \leq \frac{P\{G_{k-2}\}}{P\{G_k\}} \frac{2k}{n^2} P\{A\} .$$

Applying (4.5) and Lemma 4.2 finishes the proof of Proposition 4.6. ∎

There are $3n^3(n-1)^2$ possible "Type A" choices for X_1 and X_2 for which $X_1 \sim X_2$ [cf. (4.3)]. All have the same conditional probability, given G_k, namely $P\{A|G_k\}$. There are $n^3(n-1)^3$ possible "Type B" choices for X_1 and X_2 for which $X_1 \sim X_2$ is *not* true but $G_2 = \{T > 2\}$ is true [cf. (4.4)]. Given G_k, all have the same conditional probability, namely $P\{B|G_k\}$. Proposition 4.6 says that, given G_k with $k \leq \frac{n^2}{12}$, the common conditional probability of "Type A" choices for X_1 and X_2 differs by less than $\frac{4k}{n^8}$ from the common conditional probability for "Type B" choices. Since there are a total of $n^3(n-1)^3 + 3n^3(n-1)^2$ possible choices for X_1 and X_2 compatible with G_k, their average conditional probability must be the reciprocal of this number. It follows for $k \leq \frac{n^2}{12}$ that

$$\left| P\{A|G_k\} - \frac{1}{n^3(n-1)^3 + 3n^3(n-1)^2} \right| < \frac{4k}{n^8} \tag{4.8}$$

and

$$\left| P\{B|G_k\} - \frac{1}{n^3(n-1)^3 + 3n^3(n-1)^2} \right| < \frac{4k}{n^8} . \tag{4.9}$$

By (4.8) and (4.3),

$$\begin{aligned} &\left| P\{X_1 \sim X_2 | G_k\} - \frac{3n^3(n-1)^2}{n^3(n-1)^3 + 3n^3(n-1)^2} \right| \\ &= \left| P\{X_1 \sim X_2 | G_k\} - \frac{3}{n+2} \right| < 3n^3(n-1)^2 \frac{4k}{n^8} < \frac{12k}{n^3} . \end{aligned} \tag{4.10}$$

Let's state this formally:

Proposition 4.11: If $k \leq \frac{n^2}{12}$, then

$$\left| P\{X \sim X_2 | G_k\} - \frac{3}{n+2} \right| < \frac{12k}{n^3} .$$

Now let's do something similar for $P\{X_1 \sim X_2 \sim X_3 | G_k\}$. Define events C_1, C_2, ... to be all the different possible choices for $X_1\ X_2\ X_3$ consistent with $G_3 = \{T > 3\}$ for which X_1 = (1,1,1) and for which all coordinates of X_2 and X_3 are 1's, 2's and 3's. For example, one *might* take

$$C_1 = \{X_1 = (1,1,1), X_2 = (2,2,2), X_3 = (3,3,3)\}$$

and

$$C_2 = \{X_1 = (1,1,1), X_2 = (1,2,2), X_3 = (1,3,3)\}$$

etc. Any C_i of course has unconditional probability $P(C_i) = n^{-9}$. I'm not sure how many such C_i's there are. There surely aren't more than 300, but I don't much care. What I will show is that all these possible events C_i have about the same conditional probability, given G_k. Then we'll use the fact that the conditional probability, given G_k, of *any* possible choice of $X_1\ X_2\ X_3$ is equal to the conditional probability of *some* C_i. It will follow that

$$P\{X_1 \sim X_2 \sim X_3 | G_k\} \approx \frac{\text{no. of } X_1 \sim X_2 \sim X_3 \text{ choices for } X_1,\ X_2,\ X_3}{\text{total no. of } G_3 \text{ choices for } X_1,\ X_2,\ X_3}$$

Denote $X_4, .., X_k$ by $\widehat{X}_4, .., \widehat{X}_k$, so that a generic $\widehat{X}_i$ refers to one of the X_i's other than X_1, X_2 or X_3. Let $\widehat{G}_{k-3}$ be the same as G_{k-3} except in terms of $\widehat{X}_4, .., \widehat{X}_k$ instead of $X_1, .., X_{k-3}$. Let $\widehat{R}_{123}$ be the set of all boxes in S_n all of whose coordinates are either 1, 2, or 3. Let $\widehat{Q}_{123}$ be the set of boxes in S_n exactly two of whose coordinates are either 1, 2 or 3, like (7, 1, 3) or (2, 2, 4) for example.

So, now let C_1 and C_2 be any two distinct C_i's (not necessarily the ones in the example above).

Proposition 4.12: If C_1 and C_2 are any two events C_i as described above and $k \leq \frac{n^2}{12}$, then

$$\left| P\{C_1|G_k\} - P\{C_2|G_k\} \right| < \frac{12k}{n^2} P\{C_1\} = \frac{12k}{n^{11}} . \tag{4.12}$$

Proof: The argument is the same as for Proposition 4.6. We start by showing

$$\left| P\{G_k|\widehat{G}_{k-3} \cap C_1\} - P\{G_k|\widehat{G}_{k-3} \cap C_2\} \right| \leq \frac{6k}{n^2} . \tag{4.13}$$

If G_k is to be true under C_1 but not true under C_2, then there must be at least one $\widehat{X}_i$ in a particular set of $< 6n$ boxes, all of which are in $\widehat{R}_{123} \cup \widehat{Q}_{123}$. By Lemma 4.1, the probability of at least one $\widehat{X}_i$ landing in this set of $< 6n$ boxes is less than

$$\frac{(6n)(k-3)}{n^3} < \frac{6k}{n^2} ;$$

likewise for the other direction in (4.13). Now, multiply through in (4.13) by

$$P\{\widehat{G}_{k-3} \cap C_1\} = P\{\widehat{G}_{k-3} \cap C_2\} = P(G_{k-3})\,P(C_1) = P(G_{k-3})n^{-9}$$

and divide through by $P(G_k)$. Applying Lemma 4.2 finishes the proof of (4.12). ∎

By (3.3), there are $3!n^3(n-1)^3$ ways of choosing X_1, X_2 and X_3 so that $X_1 \sim X_2 \sim X_3$. By straightforward but slightly tedious counting, the total number of ways of choosing X_1, X_2 and X_3 so that $G_3 = \{T > 3\}$ holds is

$$n^3(n-1)^2(n^4 + 2n^3 - 6n^2 - 14n + 14) .$$

Let

$$d_n = \frac{6(n-1)}{n^4 + 2n^3 - 6n^2 - 14n + 14} . \tag{4.14}$$

Note that d_n is about $6n^{-3}$ for large n. Combining these numbers with Proposition 4.12 yields

Proposition 4.15: If $k \leq \frac{n^2}{12}$, then for any event C_i

$$\left| P\{C_i|G_k\} - \frac{1}{n^3(n-1)^2(n^4 + 2n^3 - 6n^2 - 14n + 14)} \right| < \frac{12k}{n^{11}}$$

and

Proposition 4.16: If $k \leq \frac{n^2}{12}$, then

$$\left| P\{X_1 \sim X_2 \sim X_3|G_k\} - d_n \right| < \frac{72k(n-1)^3}{n^8} < \frac{72k}{n^5} .$$

Propositions 4.10 and 4.16 applied to (2.5) yield

Proposition 4.17: If $k \le \frac{n^2}{12}$, then $P\{G_{k+1}|G_k\}$ is greater than

$$1 - n^{-3}\left[(3n-2)k - 2\binom{k}{2}\ \frac{3}{n+2} + \binom{k}{3}\ d_n + 12\,\frac{k^3}{n^3} + 12\,\frac{k^4}{n^5}\right] \tag{4.16}$$

and is less than

$$1 - n^{-3}\left[(3n-2)k - 2\binom{k}{2}\ \frac{3}{n+2} + \binom{k}{3}\ d_n - 12\,\frac{k^3}{n^3} - 12\,\frac{k^4}{n^5}\right] \quad . \tag{4.17}$$

Remark: The logarithm of the ratio between the upper and lower bounds for $P(G_{\alpha n})$ from Proposition 4.17 will be about

$$6\,\alpha^4 n^{-2} + 4.8\alpha^5 n^{-3}.$$

Thus, the ratio of the bounds for $P(G_k)$ is close to 1 even well out into the upper tail when n is large.

Another Remark: The weak convergence in (2.6) is of course not enough to establish

$$E\left(\frac{T}{n}\right) \ \to\ \int_0^\infty\ e^{-\frac{3}{2}t^2}\ dt\ =\ \sqrt{\frac{\pi}{6}} \quad \text{as}\ \ n \to \infty\ . \tag{4.18}$$

or

$$E\left(\frac{\tilde{T}}{n}\right) \ \to\ \int_0^\infty\ e^{-t^2}\ dt\ =\ \frac{\sqrt{\pi}}{2} \quad \text{as}\ \ n \to \infty\ . \tag{4.19}$$

The convergence results (4.18) and (4.19) do follow easily, however, from Proposition 4.17, together with formulas (1.1) and (1.4) in the case of (4.19).

5. Better Estimates of $P\{T > k+1|T > k\}$.

The inequality (4.7) in the previous section was rather crude. We obtained one direction by summing over all possible choices for X_i^* boxes which would violate G_k in the presence of B but not necessarily in the presence of A, and the other direction by summing over all possible choices for X_i^* boxes which would violate G_k in the presence of A but not necessarily in the presence of A. We couldn't do much better at the time because all we had to bound probabilities was Lemma 4.1. In particular, we had no good handle on the probability, given G_{k-2}^*, that at least two X_i^*'s or at least three X_i^*'s would land in $R_{12} \cup Q_{12}$, or on how they would look if they did. However, now our results in Section 4 give us information on such things, so we can revise (4.7) in light of these results to bootstrap our way to a better estimate. By paying attention to

the number of X_i^*'s in R_{12} and Q_{12} and to where they land, we will get a lot of cancellation in our estimates of the difference

$$P\{G_k|G_{k-2}^* \cap A\} - P\{G_k|G_{k-2}^* \cap B\} \quad . \tag{5.1}$$

The same ideas apply to the improvement of (4.13), although we won't carry this out.

Decompose $P\{G_k|G_{k-2}^* \cap A\}$ as

$$\begin{aligned} P\{G_k|G_{k-2}^* \cap A\} \;=\; & P\{G_k \text{ and no } X_i^* \text{ in } R_{12} \text{ or } Q_{12}|G_{k-2}^* \cap A\} \\ & + \; P\{G_k \text{ and at least one } X_i^* \text{ in } R_{12}|\; G_{k-2}^* \cap A\} \\ & + \; P\{G_k \text{ and exactly one } X_i^* \text{ in } Q_{12}, \text{ no } X_i^* \text{ in } R_{12} \;|G_{k-2}^* \cap A\} \\ & + \; P\{G_k \text{ and at least two } X_i^*\text{'s in } Q_{12}, \text{ no } X_i^* \text{ in } R_{12} \;|\{G_{k-2}^* \cap A\} \quad . \end{aligned} \tag{5.2}$$

$P\{G_k|G_{k-2}^* \cap B\}$ has the same decomposition, with B in place of A. The first term of the "A" expansion (5.2) and the first term in the "B" expansion both equal

$$P\{\text{no } X_i^* \text{ in } R_{12} \text{ or } Q_{12}|G_{k-2}^*\} \quad ,$$

so these terms cancel in (5.1). The third terms in the A and B expansions are also equal. This follows from the fact that A and B are independent of the X_i^*'s and from

$$\begin{aligned} & P\{G_k|A \cap G_{k-2}^* \cap \text{exactly one } X_2^* \text{ in } Q_{12}, \text{ no } X_i^* \text{ in } R_{12}\} \\ = \; & P\{G_k|B \cap G_{k-2}^* \cap \text{exactly one } X_2^* \text{ in } Q_{12}, \text{ no } X_i^* \text{ in } R_{12}\} \\ = \; & \frac{1}{2} \quad . \end{aligned}$$

The explanation for this last equality is that the single X_i^* in Q_{12} is conditionally equally likely to be any of the Q_{12} boxes. G_k will be true under A (respectively B) for half of these Q_{12} boxes. So anyway, the third terms cancel in (5.1).

Now let's consider the second terms. The second term in the B expansion is 0, since any box in R_{12} violates G_k under B. There are exactly two boxes in R_{12}, namely (2,2,1) and (2,1,2), which are not excluded by A. Thus, the second A term in (5.2) is bounded above by

$$P\{\text{at least one } X_i^* \text{ is (2,2,1) or (2,2,2)}|G_{k-2}^*\} \;\leq\; \frac{2(k-2)}{n^3} \quad . \tag{5.3}$$

This second A term in (5.2) is bounded below by

$$\begin{aligned}
& 2(k-2)\,P\{X_3^* = (2,2,1),\ \text{no other } X_i^*\text{'s in } R_{12}\cup Q_{12}\ |G_{k-2}^*\} \\
= \ & 2(k-2)\,P\{X_3^* = (2,2,1)\mid G_{k-2}^*\} \\
& - \ 2(k-2)\,P\{X_3^* = (1,1,1),\ \geq \text{ one other } X_i^* \text{ in } R_{12}\cup Q_{12}\ |G_{k-2}^*\} \\
\geq \ & \frac{2(k-2)}{n^3} - 2(k-2)(k-3)\,P\{X_3^* = (1,1,1) \text{ and } X_4^* \text{ in } R_{12}\cup Q_{12}|G_{k-2}^*\} \\
\geq \ & \frac{2(k-2)}{n^3} - 2(k-2)(k-3)(6n-9)\,P\{A|G_{k-2}\} \\
& - \ 2(k-2)(k-3)(3n-5)\,P\{B|G_{k-2}\}
\end{aligned} \tag{5.4}$$

[Explanation for the last inequality: There are $(6n-9)$ choices for X_4^* in $R_{12}\cup Q_{12}$ which have one coordinate in common with (1,1,1) and all these choices are equally likely, given $X_3^* = (1,1,1)$ and G_{k-2}^*. Thus

$$\begin{aligned}
P\{X_3^* \ &= (1,1,1), X_3^* \sim X_4^*, \text{and } X_4^* \text{ in } R_{12}\cup Q_{12}|G_{k-2}^*\} \\
&= (6n-9)\,P\{X_3^* = (1,1,1), X_4^* = (1,2,2)|G_{k-2}^*\} \\
&= (6n-9)\,P\{A|G_{k-2}\}\ .
\end{aligned}$$

Similarly, there are $(3n-5)$ choices for X_4^* in $R_{12}\cup Q_{12}$ which have no coordinates in common with (1,1,1).] But by (4.8) and (4.9), with $k-2$ in place of k,

$$P\{A|G_{k-2}^*\} \ < \ \frac{1}{n^3(n-1)^3} + \frac{4k}{n^8} \tag{5.5}$$

and (5.6)

$$P\{B|G_{k-2}^*\} \ < \ \frac{1}{n^3(n-1)^3} + \frac{4k}{n^8}\ . \tag{5.6}$$

Plugging these inequalities into (5.4) yields that

$$\frac{2(k-2)}{n^3} \ - \ \frac{18k^2}{n^3(n-1)^2} \ - \ \frac{72k^3}{n^7} \tag{5.7}$$

is a lower bound for the second A term in (5.2). Thus, this second term in (5.2) is always less than $\frac{2(k-2)}{n^3}$, but only a little less if $\frac{k}{n^2}$ is small.

The fourth and last term of the A expansion (5.2) is less than

$$
\begin{aligned}
&\binom{k}{2}\ P\{G_k;\ X_3^* \text{ and } X_4^* \text{in}\ \ Q_{12} | G_{k-2}^* \cap A\} \\
&= \binom{k}{2}\ P\{G_k;\ X_3^* \text{ and } X_4^* \text{in}\ \ Q_{12},\ X_3^* \sim X_4^* | G_{k-2}^* \cap A\} \\
&\quad + \binom{k}{2}\ P\{G_k;\ X_3^* \text{ and } X_4^* \text{in}\ \ Q_{12},\ \text{not}\ \ X_3^* \sim X_4^* | G_{k-2}^* \cap A\} \\
&= \binom{k}{2} 2(n-2)(10n-21) P\{X_3^* = (1,1,3),\ X_4^* \ = \ (1,2,4)\ | G_{k-2}^*\} \qquad (5.8) \\
&\quad + \binom{k}{2} 2(n-2)(5n-7) P\{X_3^* = (1,1,3),\ X_4^* \ = \ (2,3,2)\ | G_{k-2}^*\} \\
&= \binom{k}{2} 2(n-2)(10n-21) P\{A | G_{k-2}\} \\
&\quad + \binom{k}{2} 2(n-2)(5n-7)\ P\{B | G_{k-2}\}\ .
\end{aligned}
$$

The explanation for the second equality in (5.8) is that there are $2(n-2)(10n-21)$ ways of choosing X_3^* and X_4^* in Q_{12} so that $X_3^* \sim X_4^*$ and so that G_k is not necessarily false in the presence of A. These $2(n-2)(10n-21)$ ways are conditionally equiprobable. Likewise, there are $2(n-2)(5n-7)$ allowable choices for X_3^* and X_4^* in Q_{12} *without* $X_3^* \sim X_4^*$.

An inclusion-exclusion argument shows that a *lower* bound for the fourth A term in (5.2) is

$$
\begin{aligned}
&\binom{k}{2}\ P\{G_k;\ X_3^* \text{ and } X_4^* \text{in}\ \ Q_{12} | G_{k-2}^* \cap A\} \\
&- \ 3\binom{k}{3}\ P\{G_k;\ X_3^*, X_4^*, X_5^*, \text{ in }\ Q_{12}, | G_{k-2}^* \cap A\} \\
&> \quad \binom{k}{2}\ 2(n-2)(10n-21)\ P(A | G_{k-2}) \qquad (5.9) \\
&\qquad + \binom{k}{2} 2(n-2)\ (5n-7)\ P(B | G_{k-2}) \\
&\qquad - \ 3\binom{k}{3} 8(n-2)^2\ (15n-28)\ P\,(C_{max} | G_{k-2}) \quad ,
\end{aligned}
$$

where C_{max} is the C_i in Section 4 (with $k-2$ in place of k) maximizing $P(C_i | G_{k-2})$. By Proposition 4.15,

$$
P\,(C_{max} | G_{k-2}) \ < \ \frac{1}{n^3(n-1)^6} + \frac{12k}{n^{11}}\ ,
$$

since the denominator of d_n in (4.14) is greater than $(n-1)^4$ for $n \geq 3$. Thus, a *lower* bound

for the fourth A term in (5.2) is

$$\begin{aligned} & \binom{k}{2}\ 2(n-2)(10n-21)\ P(A|G_{k_2}) \\ + \quad & \binom{k}{2}\ 2(n-2)(5n-7)\ P(B|G_{k_2}) \\ & - \frac{60k^3}{n^3(n-1)^3} - \frac{720\,k^4}{n^8}\ . \end{aligned} \tag{5.11}$$

The same arguments as in the previous paragraph show that the fourth term in the B version of (5.2) is bounded above by

$$\begin{aligned} & \binom{k}{2}\ 2(n-2)(6n-9)\ P(A|G_{k-2}) \\ + \quad & \binom{k}{2}\ 2(n-2)(9n-21)\ P(B|G_{k-2}) \end{aligned} \tag{5.12}$$

and bounded below by

$$\begin{aligned} & \binom{k}{2}\ 2(n-2)(6n-9)\ P(A|G_{k_2}) \\ + \quad & \binom{k}{2}\ 2(n-2)(9n-21)\ P(B|G_{k_2}) \\ & - \frac{60k^3}{n^3(n-1)^3} - \frac{720\,k^4}{n^8}\ . \end{aligned} \tag{5.13}$$

Thus, by Proposition 4.6 and (5.5), the difference between the fourth term in (5.2) and the corresponding B term is bounded in absolute value by

$$\frac{16k^3}{n^6} + \frac{2k^2}{n^3(n-1)^2} + \frac{8k^3}{n^7} + \frac{60k^3}{n^3(n-1)^3} + \frac{720k^4}{n^8}\ . \tag{5.14}$$

Putting the above results [including the bounds in (5.3), (5.7) and (5.14)] together yields for $k \le \frac{n^2}{12}$ that

$$\begin{aligned} & \left| P\{G_k|G^*_{k-2} \cap A\} - P\{G_k|G^*_{k-2} \cap B\} - 2\frac{k-2}{n^3} \right| \\ \le\ & \frac{18k^2}{n^3(n-1)^2} + \frac{72k^3}{n^7} + \frac{16k^3}{n^6} + \frac{2k^2}{n^3(n-1)^2} + \frac{8k^3}{n^7} + \frac{60k^3}{n^3(n-1)^3} + \frac{720k^4}{n^8} \\ <\ & \frac{k^2}{n^4}\left[\frac{20}{n-2} + \frac{80k}{n^3} + \frac{76k}{(n-1)(n-2)} + \frac{720k^2}{n^4}\right] \\ <\ & \begin{cases} 14 \cdot \frac{k^2}{n^4} & \text{if } n \ge 20 \text{ and } k \le \frac{n^2}{n^4} \\ 0.86 \cdot \frac{k^2}{n^4} & \text{if } n \ge 1000 \text{ and } k \le \frac{n^2}{100}\ . \end{cases} \end{aligned} \tag{5.15}$$

[The inequality $(n-1)^2 > n(n-2)$ was used in going from the second line to the third line in (5.15).]

For the sake of keeping the formulas simple, let's assume $n \geq 20$ and $k \leq \frac{n^2}{12}$, keeping in mind that the constant 14 can be replaced by a smaller value if $n > 20$ and $k < \frac{n^2}{12}$, or in general by the expression in square brackets in (5.15). Multiplying through in (5.15) by

$$P\{G_k\}^{-1}\, P\{G^*_{k-2} \cap A\} \;=\; P\{G_k\}^{-1}\, P\{G_{k-2}\}\, n^{-6} \;<\; \frac{2}{n^6}$$

yields

Proposition 5.16. If $n \geq 20$ and $k \leq \frac{n^2}{12}$, then

$$\left| P\{A|G_k\} \;-\; P\{B|G_k\} \;-\; \frac{2(k-2)\,P\{G_{k-2}\}}{n^9\; P\{G_k\}} \right| \;<\; 14 \cdot \frac{2k^2}{n^{10}} \;.$$

Now recall from Section 4 that there are $3n^3(n-1)^2$ choices (equiprobable, given G_k) for X_1 and X_2 with $X_1 \sim X_2$ and $n^3(n-1)^3$ choices (equiprobable, given G_k) for X_1 and X_2 with no common coordinates. It is, of course, still true that the average conditional probability of these $n^3(n-1)^3 + 3n^3(n-1)^2$ choices must be the reciprocal of this number. From this and from Proposition 5.17 we get for $n \geq 20$ and $k \leq \frac{n^2}{12}$ that

$$\left| P\{A|G_k\} \;-\; \frac{1}{n^3(n-1)^3 + 3n^3(n-1)^2} \;-\; \frac{2(k-2)(n-1)P\{G_{k-2}\}}{n^9(n+2)P\{G_k\}} \right| \;<\; 14 \cdot \frac{2k^2}{n^{10}} \;. \quad (5.17)$$

Now by (5.17) and (4.3) we get

Proposition 5.18. If $n \geq 20$ and $k \leq \frac{n^2}{12}$, then

$$\left| P\{X_1 \sim X_2|G_k\} \;-\; \frac{3}{n+2} \;-\; \frac{6\,(k-2)\,(n-1)^3\;\; P\{G_{k-2}\}}{n^6(n+2)\; P\{G_k\}} \right| \;<\; 14 \cdot \frac{6k^2}{n^5} \;,$$

where the 14 can be replaced by the expression in square brackets in (5.15), providing $k \leq \frac{n^2}{12}$.

Propositions 5.18 and 4.14 applied to formula (2.5) yield the following improvement (at least for large n) of Proposition 4.15.

Proposition 5.19. If $n \geq 20$ and $k \leq \frac{n^2}{12}$, then the difference between $P\{G_{k+1}|G_k\}$ and

$$1 \;-\; n^{-3}\left[(3n-2)k - 2\binom{k}{2}\,\frac{3}{n+2} \;+\; \binom{k}{3}\,d_n \;+\; 36\binom{k}{3}\,\frac{(n-1)^3\;\; P\{G_{k-2}\}}{n^6\,(n+2)\; P\{G_k\}}\right]$$

is bounded in absolute value by

$$14 \cdot \frac{6k^4}{n^8} + \frac{12k^4}{n^8} \quad .$$

Remark. Again, the 14 in the bound may be replaced by the expression in square brackets in (5.15) when $k \leq \frac{n^2}{12}$ holds. Also, the ratio of probabilities in the estimate of $P\{G_{k+1}|G_k\}$ is equal to the reciprocal of

$$P\{G_k|G_{k-2}\} \quad = \quad P\{G_{k-1}|G_{k-2}\}\, P\{G_k|G_{k-1}\} \quad ,$$

which can be estimated by the product of our estimates for $P\{G_{k-1}|G_{k-2}\}$ and $P\{G_k|G_{k-1}\}$.

A Final Remark: The methods of this section could be used to improve Proposition 4.15, which would cause the $\frac{12k^4}{n^5}$ term in the bound in Proposition 5.19 to be replaced by a smaller order term. One could also use Proposition 5.16 in place of Proposition 4.6 to replace the first term of (5.14) by something of smaller order in n. Finally, one could replace the last two terms in (5.14) by terms of smaller order by working harder to get more cancellation between the fourth A and B terms in (5.2). The result of all this would presumably be to replace the bound in Proposition 5.19 by a bound of the form $c\,\frac{k^4}{n^9}$ for some constant c.