Omnibus Sequences, Coupon Collection, and Missing Word Counts

Sunil Abraham
Oxford University
s.abraham@gmail.com

Greg Brockman
Massachusetts Institute of Technology
gregory.brockman@gmail.com

Stephanie Sapp
University of California, Berkeley
sapp.stephanie@gmail.com

Anant P. Godbole
East Tennessee State University
godbolea@etsu.edu

July 21, 2011

Abstract

In this paper, we study the properties of k-*omnisequences* of length n, defined to be strings of length n that contain all strings of smaller length k embedded as (not necessarily contiguous) *subsequences*. We start by proving an elementary result that relates our problem to the classical coupon collector problem. After a short survey of relevant results in coupon collection, we focus our attention on the number M of strings (or words) of length k that are *not* found as subsequences of an n string, showing that there is a gap between the probability threshold for the emergence of an omnisequence and the zero-infinity threshold for $\mathbb{E}(M)$.

1 Introduction

One English translation of Leo Tolstoy's novel *War and Peace* has the following notable property: it contains this paragraph as a subsequence. Ignoring punctuation and special fonts, if one were to write just the letters and spaces that appear in the book as a string, then there would be a subsequence

of that string that is identical to the string of letters and spaces in this paragraph. The full property is more general than that — *War and Peace* contains as a subsequence *any* possible string of up to nine hundred fifty letters and spaces. That includes valid English passages such as the first nine hundred fifty characters of President Obama's Inaugural Address, as well as a string of q's of the same length. *War and Peace* is thus a tome that is *nine hundred fifty-omnibus (or omni)* over the twenty seven character alphabet $\{a, b, c, \ldots, z, SPACE\}$.

Of course, such a text is not at all hard to create by design. Consider repeating the string $S =$"$abcd \ldots xyzSPACE$" 950 times back to back. Clearly one could then find as a subsequence any possible string $(b_1, \ldots, b_k)$ of length at most 950, by choosing the first desired letter b_1 from the first repetition of S, the second required letter b_2 from the second repetition, and so on. In general, the shortest string that contains all the a^k k-words over an a letter alphabet is of length ak; simply write the alphabet k times back to back as done above and note that a string of length $\leq ak - 1$ necessarily contains a letter ξ that is represented at most $k - 1$ times, making the k-string $\xi\xi \ldots \xi$ impossible to obtain as a subsequence.

In this paper, we will study the k-omni behavior of properties of random n-strings over an alphabet of size a. Section 2 explores connections between omnisequences and the coupon collector problem; the key link between the two is given by what we term the "waddle lemma." Section 3 focuses on deriving conditions under which a random sequence is almost never or almost always k-omni. Additionally, we compute exact probabilities for a sequence to be k-omni when its length is *at* the threshold value. Section 4 is devoted to a quick review of some of the deeper properties of coupon collection. We continue, in Section 5, by deriving a "zero-infinity" threshold for the expected number of missing k-sequences, uncovering the fact that this threshold is not the same as that for the emergence of the omni property. A more detailed analysis is then undertaken. We end with potential applications and a list of open problems in Section 6.

2 A Twist on Coupon Collection

THE CLASSIC COUPON COLLECTOR PROBLEM: Let $H(1..a)$ be defined as the sum $1 + \frac{1}{2} + \ldots + \frac{1}{a}$, i.e., $H(1..a)$ is the ath partial sum of the harmonic series. Suppose that a "coupon collector" wishes to collect

one of each of a toys that are found in cereal boxes. Let the associated waiting time be denoted by $W_{1,a}$. It is well known (see, e.g., Feller[8]) that $\mathbb{E}(W_{1,a}) = a(1+\frac{1}{2}+\ldots\frac{1}{a}) = aH(1..a) \approx a[\log a+\gamma+o(1)]$ coupons, since the first purchase yields the first new toy; the expected waiting time until the second new toy is purchased is the mean of a geometric random variable with parameter $\frac{a-1}{a}$, which equals $\frac{a}{a-1}$; the third toy takes on average $\frac{a}{a-2}$ new purchases, and so on. In addition the variance of the waiting time in the coupon collector problem is given by $\mathbb{V}(W_{1,a}) = (a^2\sum_{i=1}^{a-1} 1/i^2 - aH(1..a-1)) < \frac{\pi^2}{6}a^2$.

It turns out that the omnisequence problem is inextricably linked to the coupon collector problem. The following elementary lemma in really not surprising in hindsight.

Lemma 2.1. *(The Basic "Waddle[1] Lemma".) A sequence S is k-omni if and only if there exists a set P of completed sets of coupons (1-omni substrings of S) in succession such that $|P| \geq k$.*

Proof. Sufficiency is easy to establish. Consider necessity. Suppose there exist $m < k$ successive 1-omni substrings of S. Let these be as "tight" as possible, so that the last letter in any substring is the first occurrence of that letter. Let these letters be $a_1, a_2, \ldots, a_m$ and let $A = (a_1 a_2 \ldots a_m c \ldots c)$, $(k-m$ cs), where c is any letter not in the string after the letter a_m in the mth string. Then A is not a subsequence of the string. Contradiction. □

At this juncture, it should be clear how to algorithmically find any given length k string in a k-omnisequence S. One can proceed greedily: read the omnisequence from left to right, and when the next desired letter is found, record its position. The above proof shows that this algorithm will always yield the desired string precisely when S is k-omni.

Similarly, we can design a greedy algorithm to determine the maximum k for which a given string S is k-omni. Simply read across S from left to right, recording each time a new 1-omni substring (complete coupon collection) is obtained. The total number of such substrings will be the desired k. Applying such an algorithm to one of several English translations of *War and Peace*, the second-named author's computer demonstrated the novel to be 950-omni but not 951-omni.

[1] "Research Experiences for Undergraduates" (REU) groups are close knit social/mathematical entities, and team members often develop their own vernacular. In 2008, the first named author decided to call 1-omni strings (or completed sets of coupons) *waddles*, to reflect the fact that one must waddle from one letter to the other to attain a coupon collection, and that the required gait is quite distinct from a run.

3 Threshold Behavior and Behavior at the Threshold

Consider rolling a fair die with a sides and recording the sequence of rolls obtained. Using the fact that we are looking at *successive renewals* of the k required waddles, the expected number of rolls $\mathbb{E}(W_{k,a})$ needed before the recorded sequence is k-omni on $[a]$ equals $aH(1..a)k$, since the mean waiting time for a single waddle is $aH(1..a)$. By independence, moreover, $\mathbb{V}(W_{k,a}) = \mathbb{V}(\sum_{i=1}^{k} W_{1,a}) = \sum_{i=1}^{k} \mathbb{V}(W_{1,a}) = k(a^2 \sum_{i=1}^{a-1} 1/i^2 - aH(1..a-1)) < k\frac{\pi^2}{6}a^2$. Setting $W = W_{k,a}$ for simplicity, we note that $\mathbb{V}(W) = o(\mathbb{E}(W))^2$, not just for fixed a as $k \to \infty$ but also in general if at least one of a, k tends to infinity. This is our signal that W will be tightly concentrated around its mean; Chebychev's inequality easily leads us to the following result:

Theorem 3.1. *Let $r > 0$ be a constant, and fix $a \geq 2$. Let $n = rk$, where n, k are both integers. Then*

$$\lim_{k\to\infty} \mathbb{P}(\text{Sequence of length } n \text{ is } k \text{ omni}) = \begin{cases} 0, & \text{if } r < aH(1..a) \\ 1, & \text{if } r > aH(1..a) \end{cases}$$

Proof. We provide just a proof of the second part of the result; the first is proved similarly. Let $n = kaH(1..a) + \varphi(k)\sqrt{k}a$, where $\varphi(k) \to \infty$ is any sequence such that $\varphi(k) = o(\sqrt{k})$. In other words, n is smaller than $(1+\varepsilon)aH(1..a) \cdot k$ for any $\varepsilon > 0$. We have

$$\begin{aligned} \mathbb{P}(\text{not omni}) &= \mathbb{P}(W > kaH(1..a) + \varphi(k)\sqrt{k}a) \\ &= \mathbb{P}(W - E(W) \geq \varphi(k)\sqrt{k}a) \\ &\leq \frac{\mathbb{V}(W)}{\varphi^2(k)ka^2} \\ &\leq \frac{\pi^2}{6\varphi^2(k)} \to 0, \end{aligned}$$

as asserted. If $a \to \infty$ for fixed length words, the above proof may be modified by letting $n = kaH(1..a) + a\varphi(a)$ where $\varphi(a) \to \infty$ can grow at an arbitrarily slow rate as long as $\varphi(a) = o(H(1..a))$. In general, though, we may take $n = (1+\varepsilon)aH(1..a)k$ as long as at least one of a, k tend to infinity. □

We now explore behavior at some threshold values of n, e.g., when $n = aH(1..a)k + O(1)$. Let $P(n,k,a)$ denote the probability that a sequence of length n on an alphabet $[a]$ is k-omni, and let $N(n,k,a)$ be the number of k-omni sequences of length n on $[a]$. In the binary case, when $2H(1..2) = 3$, we have

Theorem 3.2. *$P(3k-1,k,2) = \frac{1}{2}$ for each k. Furthermore, for constant c, as $k \to \infty$, $P(3k+c,k,2) \to \frac{1}{2}$.*

Proof. We provide a constructive count of $N(n,k,2)$. By Lemma 2.1, a string is k-omni precisely when it contains at least k successive 1-omni substrings. For any string $S = (s_1, s_2, \ldots, s_n)$, let $S_{i..j}$ denote the substring $(s_i, \ldots, s_j)$. Given a k-omni string S, let $\{i_j\}_{j=0}^{m}$ be defined as follows: $i_0 = 0$, and for $j > 1$, i_j is the smallest integer such that $S_{i_{j-1}+1..i_j}$ is 1-omni. Now define the sequence $\{i'_j\}_{j=1}^{m}$ by $i'_j = i_j - i_{j-1}$; that is, each i'_j gives the length of the relevant 1-omni substring of S.

Now suppose $i'_1 + i'_2 + \ldots + i'_k = t$ for some fixed t. Since $i'_1, i'_2, \ldots, i'_k \geq 2$, elementary combinatorics gives that there are $\binom{t-k-1}{k-1}$ solutions to this equation. For each solution $(i_1, i_2, \ldots, i_k)$, there are precisely $2^{k+(n-t)}$ choices for S, since each 1-omni substring can be independently chosen to be of the form $11\ldots10$ or $00\ldots01$, and the remaining $n-t$ elements of S can then be chosen arbitrarily. Thus there are a total of

$$N(n,k,2) = \sum_{t=2k}^{n} \binom{t-k-1}{k-1} 2^{n+k-t}$$

possible k-omnisequences of length n, and the probability that a given sequence of length n is k-omni is

$$\begin{aligned} P(n,k,2) &= \frac{N(n,k,2)}{2^n} \\ &= \sum_{t=2k}^{n} \binom{t-k-1}{k-1} 2^{k-t} \\ &= \frac{1}{2^k} \sum_{t=0}^{n-2k} \binom{t+k-1}{k-1} 2^{-t}. \end{aligned}$$

Since $\sum_{t=0}^{k-1} \binom{t+k-1}{k-1} 2^{-t} = 2^{k-1}$ (see, e.g., Gould [12]), $P(3k-1,k,2) = \frac{1}{2}$. On the other hand, it is not hard to see that if c is constant (or indeed if $c = o(\sqrt{k})$), $P(3k+c,k,2) \to P(3k-1,k,2) = \frac{1}{2}$ as $k \to \infty$. This proves the theorem. □

4 Old and Recent Results on Coupon Collection

Our intent in this section is to provide a quick review of some classical and recent work on coupon collection, keeping potential applications to omnisequences in mind at all times. In Section 5, we will use the groundwork laid down in this section to make progress beyond Theorem 3.1. We start with a review of several approaches to coupon collection.

(i) Perhaps the most natural way to view the coupon collector problem is as an occupancy problem is which we place n balls in a urns so that the a^n possibilities are equiprobable. This is the *classical* approach detailed, e.g., in Feller [8], and which yields, e.g., an exact expression for p_b, the probability that exactly b of the a coupons have not been collected, which is the $k = 1$ version of the problem studied in Section 5, namely *number of missing k-words*, i.e., words that are not found as a subsequence of a given n-string.

(ii) Consider next the *the waiting time* approach mentioned at the beginning of Section 2: The waiting time $W_{1,a}$ for the completion of a collection of a coupons is expressed as the sum of a geometric random variables with declining success probabilities 1, $(a-1)/a, (a-2)/a, \ldots, 1/a$. This representation enabled us to quickly discover the fact that

$$\mathbb{E}(W_{1,a}) = aH(1..a) = a(\log a + \gamma + o(1)) \ \ (a \to \infty),$$

can be used to compute generating function, moments, etc., and was the basis of the method employed by Erdős and Rényi [7] to prove the extreme value limit theorem

$$\mathbb{P}\left(\frac{W_{1,a} - a\log a}{a} \le x\right) \to \exp\{-e^{-x}\} := \Psi_1(x) \ \ (a \to \infty). \tag{1}$$

Finally, and of relevance to us, the geometric representation has been used (see [13] for references) to work out the asymptotics for the distribution of the number of missing coupons, the waiting time for the b^{th} coupon, etc.

(iii) Third, the *Poisson embedding* approach is at the basis of the exposition in Aldous [3], who uses a heuristic to correctly "guess" several answers — both to the coupon collector problem and to various generalizations such as waiting times until most coupons are collected; until each coupon is collected $A + 1$ times; until each coupon is collected once when these are not equally likely to occur; until each subset in a class is hit; etc. Lars Holst's

important paper [13] shows how we may embed the placement of balls in urns (or, equivalently the drawing of balls from urns) into a Poisson process, so that many classical "quota-related" occupancy problems such as the birthday problem, coupon collector problem, and occupancy count problem can be recast in terms of order statistics from a gamma distribution. In addition, this method enables one to provide easier solutions to the problem of multiple coupon collection ($A+1$ coupons of each kind).

(iv) *Poisson Approximation* is another possibility that allows one to go beyond waiting time analyses. Now, if n is large compared to a, then a coupon being missing would be a rare occurrence. The number of missing coupons, or the number of empty boxes, ought to have a Poisson distribution. The Stein-Chen method of Poisson approximation, [6], enables one to quantify closeness to a Poisson distribution in an appropriate metric. We find in Chapter 6 of [6], or in the paper [5] that features many examples related to occupancy, that the total variation distance between the distribution of the number of boxes with $m \geq 2$ or more balls (birthday coincidences), or the number of empty boxes (missing coupons) and appropriately defined Poisson distributions, is small under a set of conditions that permit large expected values. Now we shall see in Section 5 that it is a rare occurrence for words to not be embedded in alphabet strings. Can the count of such words have a Poisson distribution?

(v) Significant progress has been made in recent years towards a fine-resolution understanding of coupon collection, using methods from *Analytic Combinatorics* ([9], [10], [11], [18], [24]). It is undeniable, as we shall see later, that it is precisely results of this nature that will help one understand and establish the link between that which we know (waddle counts; coupons) and that which we seek to know (missing word counts).

There are many variations on the basic coupon collection theme; see, e.g., [3], [2], [18], [17]. The generalization of particular interest to us, is coupon collection until $A+1$ copies of each coupon are scored, $A \geq 1$. Note that a (minimal) collection with $A+1$ copies of each coupon might decompose into anywhere between one and $A+1$ waddles. In the case that $A=1$, the problem goes by the name of the "double dixie cup" problem and was first studied in [21]. Simpler proofs of several of the results in [21] were given in [13] and [18]. In [13], it was proved that with $V = V_{a,A+1}$ denoting the waiting time until $A+1$ copies of each coupon are obtained and

$$V^* = V^*_{a,A+1} = \frac{V}{a} - \log a - A \log\log a + \log A!,$$

that as $a \to \infty$,

(i) $V^*_{a,1} \ldots, V^*_{a,m}$ are asymptotically independent;

(ii) $\mathbb{P}(V^*_{a,A+1} \leq u) \to \exp\{-e^{-u}\}$; and

(iii) $\mathbb{E}(V_{a,A+1}) = a\,(\log a + A \log\log a + \gamma - \log A! + o(1))$.

The above discussion illustrates that an important auxiliary variable would be counts of coupons of different types. Several questions may be asked. Perhaps the first is how many coupons have been collected precisely once at the end of a successful coupon collecting quest. Myers and Wilf [18] solve this problem. Among their key results is the fact that on average $H(1..a) \approx \log a$ coupons have been collected precisely once at the end of a minimal completed coupon collection. This implies that roughly $\log a$ coupons need to be collected a second time by the double dixie cup collector; the rest have already been collected twice as part of the first collection. Thus, heuristically, the additional waiting time until these singletons turn into doubles is $a \cdot \log\log a$, as seen above. This fact reveals a key difference between the waiting time until each coupon is collected k times and the waiting time until the sequence is k-omni, for which the expected value is $kaH(1..a)$. A problem inverse to that in [18] was tackled by Badus et al. [4], who asked how many copies there are of the r^{th} new coupon to be collected. Zeilberger [24] gave a simpler proof of a closed form formula, first derived in [10], for the generating function $\sum_{i=1}^{\infty} E(Y_i)t^i$, where Y_i is the number of coupons that have been collected precisely i times. The multivariate generating function of $\mathbb{P}(Y_1 = y_1, \ldots, Y_r = y_r, W_{1,a} = w)$ was also derived in [10]. By rephrasing the problem in terms of a coupon collector and his ordered infinite sequence of younger brothers to whom duplicates are passed on sequentially, Foata and Zeilbeger [11] derive results about the expected numbers of missing coupons in the collections of younger brothers, when p brothers have complete collections. A simpler proof of these results, for $p = 1$, is provided in [1].

How do the above results shed light on omnisequences? We have seen that the normalized waiting time for the coupon collector follows asymptotically a Gumbel distribution as the number of coupons gets large. A generalization to unequal coupon probabilities is given by Neal [20]. A further generalization is provided by Martinez [16], who proves a ratio limit theorem for the waiting time until $A+1$ copies of a coupons are obtained. For equally likely coupons, there are a host of approximations for small values of a; these are of the normal, saddlepoint, and lognormal types, and a good summary may be found in [15]. The point to emphasize is that the situation is complicated even for a single waddle if the coupon size is small. On the other hand, if

k is allowed to get large, then the following result on the waiting time for a sequence to become k-omni follows easily from the central limit theorem.

Theorem 4.1. *Let $W_{k,a}$ be the waiting time until a sequence $\{X_n\}_{n=1}^{\infty}$ of i.i.d. letters uniformly generated from $\{1, 2, \ldots, a\}$ becomes k-omni. Then*

$$\mathbb{P}\left(\frac{W_{k,a} - kaH(1..a)}{\sqrt{k}S} \leq x\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-u^2/2\} du \ \ (k \to \infty),$$

where S denotes the standard deviation for the waiting time $W_{1,a}$ until a single coupon collection is obtained.

The above result can be used to deduce, for example, that as $k \to \infty$ (a fixed) the probability $P(kaH(1..a) + \sqrt{k}, k, a)$ that a string of length $kaH(1..a) + \sqrt{k}$ is k-omni satisfies

$$P(kaH(1..a) + \sqrt{k}, k, a) = \mathbb{P}\left(\frac{W_{k,a} - kaH(1..a)}{\sqrt{k}S} \leq \frac{1}{S}\right) \to \Phi\left(\frac{1}{S}\right),$$

where Φ is the standard normal distribution function; for $a = 4$, we get $S = 3.8$ and a limiting value of approximately 0.603. Also, Theorem 4.1 reveals that the probability $P(kaH(1..a) + O(1), k, a)$ is asymptotically 0.5 for all a, thus extending Theorem 3.2.

The situation is different if k is held fixed and we allow a to tend to infinity. By conditioning on the $a!^k$ orders in which letters could be generated so as to yield an omnibus sequence, we see that $W_{k,a}$ can be written as the sum of ak independent geometric waiting times, with precisely k having success probability j/a, $1 \leq j \leq a$. Thus, we recognize that

$$\mathbb{P}\left(\frac{W_{k,a} - ka\log a}{a} \leq x\right)$$

represents the distribution function of the sum of k identical copies of the normalized single waddle times $(W_{1,a} - a\log a)/a$. The next result follows easily from the Erdős-Rényi result (1):

Theorem 4.2.

$$\mathbb{P}\left(\frac{W_{k,a} - ka\log a}{a} \leq x\right) \to \Psi_k(x) \ \ (a \to \infty),$$

where Ψ_k is the distribution function of the sum of k independent Gumbel variables.

Unfortunately, the representation of Ψ_k, as given by Nadarajah [19], is not amenable to easy analysis.

We end this section with a possibly new, possibly folklore, result. The original solution for the expected waddletime when coupons are present in unequal proportions *appears* to be the one originally given by Von Schelling [23]. Another expression is in Section 3 of [17] and the distribution is described in Theorem 2.2 of [20]. Below we offer the above-mentioned alternative expression for the expectation.

Assume first for specificity that $\mathbb{P}(R) = 1/2, \mathbb{P}(B) = 1/3, \mathbb{P}(G) = 1/6$, and conditioning on the order in which the three colors appear (these are $RBG, RGB, BRG, BGR, GRB, GBR$ with respective probabilities $\frac{1}{2} \cdot \frac{1/3}{1/3+1/6} \cdot 1$, $\frac{1}{6}$, $\frac{1}{4}$, $\frac{1}{12}$, $\frac{1}{10}$, $\frac{1}{15}$), we need to find the conditional expectation of the waiting time given the order of the first appearance of the colors. Assume that the order is RBG. The waiting time is then clearly $1 + x + 6$, where x is the additional waiting time until the B appears. It might appear that this waiting time ought to be shorter than if one were waiting for a G after an R. *But it isn't.* The conditional distribution computation reveals that in fact $x = 2$, the same as the waiting time for either B or G, given that R appeared first. Thus, in this example,

$$\begin{aligned}\mathbb{E}(W_{1,3}) &= \frac{1}{3} \cdot [1+2+6] + \frac{1}{6} \cdot [1+2+3] + \frac{1}{4} \cdot [1+1.5+6] + \\ &\quad \frac{1}{12} \cdot [1+1.5+2] + \frac{1}{10} \cdot [1+1.2+3] + \frac{1}{15} \cdot [1+1.2+2] = 7.3,\end{aligned}$$

and, in general we have the following

Alternative Expression for Expected Waddletime *Let balls be independently thrown into boxes labeled* $1, 2, \ldots, a$ *so that any ball hits box* j *with probability* p_j*. Then the expected value of the time* $W = W_{1,p_1 \ldots, p_a}$ *until all boxes are nonempty satisfies*

$$\mathbb{E}(W) = \prod_{j=1}^{a} p_j \sum_{\pi \in \mathcal{S}_{|a|}} q_{\pi(1)} \ldots q_{\pi(a-1)} \left(1 + q_{\pi(1)} + \ldots + q_{\pi(a-1)}\right),$$

where

$$q_{\pi(j)} = \frac{1}{1 - \sum_{i=1}^{j} p_{\pi(i)}}$$

and $\pi = (\pi_1 \ldots, \pi_a)$ *varies over all the permutations in the symmetric group* $\mathcal{S}_{|a|}$ *on* $\{1, 2, \ldots, a\}$.

The above expression yields an expected 1-omni time of around 2250 until each of the letters A through Z are randomly obtained, if we generate the letters according to the frequency with which they actually appear in "normal" English text. Also, the same basic technique can be used to derive a direct expression for the expected collection time for (say) two copies of each coupon in the non uniform case.

5 Missing Word Counts

We now change our approach to the omnibus problem. Instead of considering only sequences that contain all possible length k strings, consider strings that do not necessarily attain them all. Given a sequence S of length n on $[a]$, define the *number of missing k-sequences* of S to be the number of distinct k-sequences on $[a]$ that cannot be obtained as a subsequence of S. Denote this quantity by $M = M_{k,a,n} = M_{k,a,n}(S)$, so that $M_{k,a,n}(S) = \sum_{T \in [a]^k} I_{k,a,n}(S,T)$ where the indicator variable $I_{k,a,n}(S,T) = I(T)$ equals 1 iff the word T is not a subsequence of S. The following result is critical and is in marked contrast to the situation when words have to occur as strings, where, e.g., with $a = 2$ and $n = 3$, the probability that 11 is missing as a string is $5/8$, whereas the corresponding probability for 10 is $1/2$.

Lemma 5.1. *The probability that a k-sequence is missing in a random string of length n is equal to the probability that any other k-sequence is missing in that string.*

Proof. Say S is length n, and let $T = (t_1, t_2, \ldots, t_k)$ be any word. Then T is missing if and only if for some $0 \leq j \leq k-1$ we make "j-fold progress" towards the attainment of T, i.e., the first j letters of T can be found in S as a subsequence, but not the first $j+1$. Let us choose the spots where these j letters are to appear for the first time in $\binom{n}{j}$ ways. Label the spots as $i_1, \ldots, i_j$. Now the letters prior to i_1 cannot contain the letter t_1, the letters in between i_1 and i_2 must be devoid of a t_2, etc. It follows that

$$\mathbb{P}(I(T) = 1) = \sum_{j=0}^{k-1} \binom{n}{j} \left(\frac{1}{a}\right)^j \left(\frac{a-1}{a}\right)^{n-j},$$

which is merely the cumulative binomial probability $B(n, k-1, 1/a)$. The above expression is dependent on only n and k, but not on what sequence

T is. Notice that, for example, when $a = 2$ and $T = 11\ldots1$, we should interpret the above equation as saying that T is missing if and only if the sequence S contains at most $(k-1)$ 1's. □

5.1 The Gap

We now calculate the asymptotics of $\mathbb{E}(M_{k,a,n})$, the expected number of words of length k that are missing in a random n-string, as $k \to \infty$ and $n/k = r$ is held constant. By linearity of expectation,

$$E(M_{k,a,n}) = a^k \sum_{j=0}^{k-1} \binom{n}{j} \left(\frac{1}{a}\right)^j \left(\frac{a-1}{a}\right)^{n-j}. \tag{2}$$

Now for $n \geq ak$, the maximum term in the sum (2) is the one corresponding to $j = k-1$. This is easy to see by taking ratios of consecutive terms, and can be made precise by the following inequality from Barbour et al. [6]:

$$\mathrm{Bi}(n,p)\{0,\ldots,m{-}1\} \leq \frac{(n-m)p}{(n-1)p-(m-1)}\mathrm{Bi}(n,p)\{m{-}1\}, \;\; m < np{+}(1{-}p),$$

where

$$\mathrm{Bi}(n,p)\{A\} = \sum_{j\in A} \binom{n}{j} p^j (1-p)^{n-j}.$$

This leads to

$$a^k \mathrm{Bi}(n,\frac{1}{a})\{k{-}1\} \leq E(M_{k,a,n}) \leq \frac{a^k}{\left(1-\frac{ak}{n}\right)}\mathrm{Bi}(n,\frac{1}{a})\{k{-}1\} \leq 4a^k \mathrm{Bi}(n,\frac{1}{a})\{k{-}1\}$$

if, e.g., we take $n \geq \frac{8}{9}kaH(1..a)$. Thus,

$$E(M_{k,a,n}) \sim \Lambda \cdot \frac{1}{a^{n-k}} \cdot \binom{n}{k-1} (a-1)^{n-k+1}$$

for some constant Λ. Applying Stirling's approximation with $n = rk$, we see that

$$E(M_{k,a,n}) \sim \frac{\Lambda(a-1)\sqrt{r}}{(r-1+o(1))\sqrt{2\pi(r-1)k}} \left(\frac{(a-1)^{r-1}r^r}{a^{r-1}(r-1)^{r-1}}\right)^k. \tag{3}$$

Now we have seen that previous asymptotic results are all couched in terms of alphabet sizes that grow to infinity. On the other hand, omnibus behavior is best appreciated for long words from a fixed size alphabet. Accordingly, we ask what happens to $\mathbb{E}(M)$ as $k \to \infty$, and find from (3) that with

$$D(a,r) = \frac{(a-1)^{r-1} r^r}{a^{r-1}(r-1)^{r-1}},$$

and a fixed, $\mathbb{E}(M) \to 0$ as $k \to \infty$ if $D(a,r) \leq 1$, and $\mathbb{E}(M) \to \infty$ ($k \to \infty$) if $D(a,r) > 1$.

Notice the similarity to Theorem 3.1. Holding the ratio $n/k = r$ constant, we again find that there is a threshold value of r at which there is a sudden change in the asymptotics. However, these threshold values are not equal to one another. Recall, e.g., that for k-omni strings, the threshold ratio (prior to which the probability of a string being k-omni was 0, beyond which it was 1) is $2H(1..2) = 3$ for $a = 2$. However, again for $a = 2$, we can show that $D(2,r) = 1$ when $r \approx 4.403$. What is going on? It appears that for values of n between $3k$ and $4.403k$, sequences are omni with high probability, and yet the expected number of missing sequences is huge, much like the two-valued random variable X that takes on values 0 and n^2 with probabilities $1 - 1/n$ and $1/n$ respectively: $\mathbb{E}(X)$ is large even though X equals zero most of the time. It appears that M is *similarly not concentrated around its mean*. Specifically, rare non-omni sequences tend to have unaccomplished waddles that lead to very large numbers of missing words. We return to this question in the next section, but for now demonstrate the fact that there is a negligible "gap" when the alphabet size is large. In other words, as $a \to \infty$, the difference between these threshold values grows without bound, but their ratio converges to one:

Theorem 5.2. *Given a, let $r(a)$ be the real solution to $D(a, r(a)) = 1$. Then as $a \to \infty$, $\frac{r(a)}{aH(1..a)} \to 1$.*

Proof. We show that for large a, $a(\log a + \log\log a) < r(a) < a(\log a + \log\log a + 2)$. Since also $aH(1..a) \sim a \log a$, $a \to \infty$, the result will follow immediately via the squeeze theorem. Set $r'(a) = a(\log a + \log\log a + c)$ for a large and c constant. Then

$$\begin{aligned} D(a, r'(a)) &= \left(\frac{a-1}{a}\right)^{r'(a)-1} \left(\frac{r'(a)}{r'(a)-1}\right)^{r'(a)} (r'(a) - 1) \\ &\sim e^{-r'(a)/a} \cdot e \cdot r'(a)(1 + o(1)). \end{aligned}$$

Thus

$$\begin{aligned} D(a, r'(a)) &\sim e^{-(\log a + \log\log a + c)} \cdot e \cdot a(\log a + \log\log a + c) \\ &= \frac{1}{a \log a} \cdot e^{-c+1} \cdot a(\log a + \log\log a + c) \\ &\sim e^{1-c} \end{aligned}$$

Thus if $c = 0$, $D(a, r'(a)) > 1$, and if $c = 2$, $D(a, r'(a)) < 1$. But $D(a, r(a)) = 1$, and the result follows by monotonicity. □

5.2 Understanding the Gap

A central question is the following: How many successive waddles does a random sequence of length n contain? We seek, in other words, to understand the level crossing time

$$\begin{aligned} \tau &= \inf\{t : W_{t,a} > n\} \\ &= \inf\{t : W_{1,a,1} + W_{1,a,2} + \ldots + W_{1,a,t} > n\}, \end{aligned}$$

where the $W_{1,a,j}$'s are i.i.d. random variables with distribution equal to that of a single waddle-time; if $\tau = t$ then the sequence is $(t-1)$-omni (there are $t-1$ waddles).

Now if there are $r < k$ waddles, then a rather naïve lower bound for the number $M_{k,a}$ of missing words of length k is a^{k-r-1}, as follows. Since there are r waddles, let a_0 be a letter not contained among the letters after the rth waddle is accomplished. Furthermore, let $a_1, \ldots, a_r$ be the last letters in the r successfully completed coupon collections. Then we see that none of the words $a_1 a_2 \ldots a_r a_0 x_1 x_2 \ldots x_{k-r-1}$ are contained in the string, where the x_j's are arbitrary. Thus even $\sqrt{k}$ fewer waddles than required would lead to at least $a^{\sqrt{k}-1}$ missing words.

Let n be fixed. We invoke the basic renewal equations from Section XIII.6 in Feller[8], that state that the number N_n of disjoint occurrences, among the first n trials, of a recurrent event $\mathcal{E}$ with mean μ and variance σ^2, satisfies

$$\mathbb{E}(N_n) \sim \frac{n}{\mu}; \;\; \mathbb{V}(N_n) \sim \frac{n\sigma^2}{\mu^3}.$$

We thus see that n random keystrokes on an a letter keyboard are expected to contain $n/aH(1..a)$ disjoint sets of strings that do not miss any letter, and that the variance of this quantity is of order $n\pi^2/6aH^3(1..a)$. Moreover, N_n

is tightly concentrated around its mean, as evidenced, e.g., by Chebychev's inequality or the Azuma-Hoeffding martingale inequality ([22]) that yields, since altering one of the keystrokes $X_1, \ldots, X_n$ can change N_n by at most one,

$$\mathbb{P}\left(\left|N_n - \frac{n}{aH(1..a)}\right| > \lambda\right) \leq 2\exp\{-\lambda^2/2n\}, \tag{4}$$

so that for fixed a, the number of waddles is concentrated in an interval of width $\sqrt{n\varphi(n)}$ around its expected value — which is of order $\Theta(n)$, where $\varphi(n)$ may tend to infinity arbitrarily slowly. How then can we get *significantly* fewer waddles than expected? To fix our ideas, we recall from (3) that for $a = 2$ we expect $(27/16)^k$ missing words if $n = 3k$, the threshold value for the sequence to be omni, and $\mathbb{E}(M_{k,a}) = (256/216)^k$ if $n = 4k$. These values are derived from the linearity of expectation, and provide little insight into what *causes* words to be missing, the correlations between the presence or absence of words, etc. Now, setting $a = 2$ and $n = 4k$ in (4), we see that for k large enough,

$$\mathbb{P}(N_n < k/2) \leq \mathbb{P}(|N_n - 1.33k| \geq 0.83k) \preceq (0.916)^k. \tag{5}$$

Now the actual probability of having a shortfall of $0.83k$ or more waddles is certainly smaller than that given by (5), but such a shortfall would, as discussed above, lead to at least $2^{k/2}$ missing words — and, making believe that (5) is sharp, an expected value of at least $(\sqrt{2} \cdot 0.916))^k \approx (1.3)^k$ for the number of missing words. Now, we know this is false (the correct expected value is $(256/216)^k = (1.18)^k$) but we believe the above crude analysis *does* add value.

To give a more specific example, we compute the probability that a sequence of length $kaH(1..a)$ has fewer than $k - \sqrt{k}$ waddles. By Theorem 4.1, this converges to some constant B, and leads to the conclusion that $\mathbb{E}(M_{k,a}) \geq B \cdot a^{\sqrt{k}}$ which certainly tends to infinity.

Fleshing out the relationship between unaccomplished coupon collections and missing word counts clearly remains a key problem that warrants deeper further investigation.

6 Applications and Open Problems

We believe that omnisequences have a large number of potential applications. Below are some of our thoughts on the matter.

Cryptography: Omnisequences could provide a potential method for cryptography. For example, suppose that Alice and Bob meet and exchange one-time pads of randomly generated letters (or even an innocuous looking copy of *War and Peace*). The encryption process for a message then becomes to greedily find the position of the desired letters within the pad. For example, given a pad of "abfpodod...," the ciphertext of "food" would be "3,5,7,8." The decryption process simply involves reading across the pad and recording the letters that appear in the relevant positions. Notice that both the encryption and decryption process are exceedingly simple and require very little computational resources; more complicated schema can certainly be employed. Our results show that if we want a random pad to be able to encrypt any message of length k, it should have length of at least $26H(1..26)k \approx 100k$. (Of course, a disadvantage of this cryptographic scheme is that only about 1% of the letters in the pad will actually be used.) This is essentially a variation of (or perhaps identical to) schemes that have actually been employed in the past.

Randomness tests: The results of the Coupon Collector problem have been used to analyze the randomness of data samples, such as in Kendall and Babington Smith [14]. The related but distinct results we have derived for k-omnisequences could be applied to randomness tests. The following examples are specific cases of a general agenda that is under investigation: (i) We know that the length of a k-omnisequence is tightly concentrated around its expected value for large k and for large a as evidenced by Theorems 4.1 and 4.2. Using the former as a more tractable example, testing the hypothesis that the letters are independently and uniformly generated can be accomplished using the central limit theorem of Theorem 4.1, and the power of the test under any specific alternative may be evaluated using empirical or semi-theoretical means. (ii) In a slightly different vein, suppose the data that we are able to observe consists of a single waddle. Can this small sample be used to give better estimates of the coupon probabilities p_i than the classical MLE estimate? Can we test the hypothesis that $p_i = 1/a \forall i$?

Derivation of identities: Omnisequences are a combinatorial structure that provide for multiple ways of counting any one event. In the process of doing this research, the authors stumbled upon a number of combinatorial identities, some perhaps not noticed before. For example, in Lemma 5.1 it

was shown that

$$\sum_{i=k}^{n} \binom{n}{i} (a-1)^{n-i},$$

is equal to the total number of n-sequences not missing a word T, which can also be calculated as

$$\sum_{i=k}^{n} \binom{i-1}{k-1} a^{n-i}(a-1)^{i-k}$$

as follows: Let the ith element of S, S_i, be the first appearance in S of the last letter, T_k, of T, given that letters $T_1, \ldots, T_{k-1}$ have appeared sequentially in S. Now choose the positions of the relevant terms of the subsequence in $\binom{i-1}{k-1}$ ways. Consider when T_j, appears in S; each subsequent term prior to T_{j+1} in S has $a-1$ choices, namely not T_{j+1}. The $n-i$ elements after S_i have a choices. Hence

$$\sum_{i=k}^{n} \binom{i-1}{k-1} a^{n-i}(a-1)^{i-k} = \sum_{i=k}^{n} \binom{n}{i} (a-1)^{n-i}.$$

A second such identity can be derived by considering the total number of (minimal) 1-omnisequences of length n on $[a]$. First of all, we can construct such a sequence in the following manner. Let $\alpha_1\alpha_2\ldots\alpha_a$ be a permutation of a, denoting the order in which the letters first appear; the first letter in the omnisequence is thus α_1 and the last is α_n. The remaining $n-a$ letters can then be placed with restriction that the letters between α_i and α_{i+1} may acquire any of the values $\alpha_1, \alpha_2, \ldots, \alpha_i$. We note that this construction will always yield a distinct 1-omnisequence, and furthermore every 1-omnisequence can be constructed in this way. Now if there are $l_i - 1$ letters, $l_i \geq 1$, between α_i and α_{i+1}, then we note that we can create $1^{l_1-1}2^{l_2-1}\cdots(a-1)^{l_{a-1}-1}$ 1-omnisequences. Furthermore, we had $a!$ ways of creating the original permutation. Hence we calculate the total number of 1-omnisequences of length n as $a!\sum_{l_1+\ldots+l_{a-1}=n-1} 1^{l_1-1}2^{l_2-1}\cdots(a-1)^{l_{a-1}-1} = a\sum_{l_1+\ldots+l_{a-1}=n-1} 1^{l_1}2^{l_2}\cdots(a-1)^{l_{a-1}}$.

Alternatively, we can consider fixing the last letter of our 1-omnisequence (which can be done in a ways). Suppose that we want to have $l_1, l_2, \ldots, l_{a-1}$ copies of each of the remaining first, second, ..., $a-1$st letters in our 1-omnisequence. Since the arrangement of these letters is arbitrary, we have that there are $\binom{n-1}{l_1,l_2,\ldots,l_{a-1}}$ sequences we can construct in this way. Again, this

construction provides a distinct 1-omnisequence, and all 1-omnisequences of length n can be constructed in this manner. Thus we obtain that there are $a \sum_{l_1+\ldots+l_{a-1}=n-1} \binom{n-1}{l_1,l_2,\ldots,l_{a-1}}$ such omnisequences. Combining our results, we find that

$$\sum_{l_1+\ldots+l_{a-1}=n-1} 1^{l_1} 2^{l_2} \cdots (a-1)^{l_{a-1}} = \sum_{l_1+\ldots+l_{a-1}=n-1} \binom{n-1}{l_1, l_2, \ldots, l_{a-1}}.$$

Linguistics: Using a random i.i.d. non-uniform model based on the frequencies of letters and spaces, one can calculate that the expected length of a 1-omnisequence in English is about 2250. However, our experiments with various text samples have shown that this is very rarely achieved and the real value for many authors is probably more like 4000. In language, letters are of course not randomly distributed. Rather, they follow some weighted distribution (even this is, of course, a simplified model of language). We note that our results can be thus be extended to languages using a weighted version of the Coupon Collector problem, such as is provided by Hermann Von Schelling [23]. In any case, this provides for some very interesting analysis and could conceivably be put to work checking, say, the degree of relationship between two languages, or testing hypotheses regarding disputed authorship (e.g., William Shakespeare vs. Francis Bacon, or Christopher Marlowe, or Edward de Vere).

Open Questions: Questions for further investigation have been mentioned throughout the paper, but here are a few others that we consider to be central.

(i) What is the relationship between the number of waddles in a non-omnibus sequence and the number of missing k-words?

(ii) Can we approximate the distribution $\mathcal{L}(M_{k,a}(S))$ of missing words in an n string?

(iii) What is the variance of $M_{k,a}(S)$?,

and, last but certainly not least,

(iv) What are the general properties of two dimensional $n \times n$ arrays over $[a]$ that contain all $k \times k$ arrays as submatrices? (we call such arrays "omnimosaics.")

7 Acknowledgment

The research of all four authors was supported by NSF-REU Grant 0552730 and conducted at ETSU during the summer of 2008 when SA, GB, and SS were students at Oxford, Harvard and Hopkins respectively. The research of AG was further supported at JHU by the Acheson J. Duncan Fund for the Advancement of Research in Statistics. We appreciate the thoughtful comments of the two anonymous referees, which have improved this paper substantially.

References

[1] I. Adler, S. Oren, and S. Ross (2003), "The coupon collector's problem revisited," *J. Applied Probability* **40**, 513–518.

[2] I. Adler and S. Ross (2001), "The coupon subset collection problem," *J. Applied Probability* **38**, 737–746.

[3] D. Aldous (1989), *Probability Approximations via the Poisson Clumping Heuristic,* Springer Verlag, New York.

[4] A. Badus, A. Godbole, E. LeDell, and N. Lents (2003), "Some contributions to the coupon collector problem," Extended Abstract, 2003 Permutation Patterns Conference, Dunedin, New Zealand. Manuscript in preparation.

[5] A. Barbour and L. Holst (1989), "Some applications of the Stein-Chen method for proving Poisson convergence," *Advances in Applied Probability* **21**, 74–90.

[6] A. Barbour, L. Holst, and S. Janson (1992), *Poisson Approximation,* Oxford University Press.

[7] P. Erdős and A. Rényi (1961), "On a classical problem of probability theory," *Magy. Tud. Akad. Mat. Kutató Int. Közl.* **6**, 215-220.

[8] W. Feller (1968), *An Introduction to Probability Theory and its Applications, Volume 1*, John Wiley, New York.

[9] P. Flajolet and R. Sedgewick (2009), *Analytic Combinatorics,* Cambridge University Press.

[10] D. Foata, G. Han, and B. Lass (2001), "Les nombres hyperharmoniques et la fratrie du collectionneur de vignettes," *Sém. Lothar. Combin.* **47**, Article B47A.

[11] D. Foata and D. Zeilberger (2003), "The collector's brotherhood problem using the Newman-Shepp symbolic method," *Algebra Universalis* **49**, 387–395.

[12] H.W. Gould (1972), *Combinatorial Identities*, Morgantown, WV.

[13] L. Holst (1986), "On Birthday, Collectors', Occupancy and Other Classical Urn Problems," *International Statistical Review* **54**, 15–27.

[14] M. Kendall and B. Babington Smith (1938), "Randomness and random sampling numbers," *J. Royal Statist. Society* **101**, 147–166.

[15] D. Kuonen (2001), *Computer-Intensive Statistical Methods: Saddlepoint Approximations with Applications in Bootstrap and Robust Inference,"* Doctoral Dissertation, École Polytechnique Fédérale, Lausanne.

[16] S. Martinez (2004), "Some bounds on the coupon collector problem," *Random Structures and Algorithms* **25**, 208–226.

[17] R. May (2008), "Coupon collecting with quotas," *The Electronic Journal of Combinatorics* **15**, Paper No. N31.

[18] A. Myers and H. Wilf (2003), "Some new aspects of the coupon collector's problem," *SIAM J. Discrete Mathematics* **17**, 1–17.

[19] S. Nadarajah (2008), "Exact distribution of the linear combination of p Gumbel random variables," *International Journal of Computer Mathematics* **85**, 1355–1362.

[20] P. Neal (2008), "The generalised coupon collector problem," *J. Applied Probability* **45**, 621–629.

[21] D. Newman and L. Shepp (1960), "The double dixie cup problem," *The American Mathematical Monthly* **67**, 58–61.

[22] J. Steele (1997), *Probability Theory and Combinatorial Optimization,* SIAM, Philadelphia.

[23] H. Von Schelling (1954), "Coupon collecting for unequal probabilities," *The American Mathematical Monthly* **61**, 306–311.

[24] D. Zeilberger (2001), "How many singles, doubles, triples, etc. should the coupon collector expect?" Unpublished manuscript available at Prof. Zeilberger's website.