Discrete Mathematics and Theoretical Computer Science DMTCS vol. **15:2**, 2013, 259–270

Coupon collecting and transversals of hypergraphs

Marcel Wild[1] Svante Janson[2†] Stephan Wagner[1‡] Dirk Laurie[1]

[1] *University of Stellenbosch, South Africa*
[2] *University of Uppsala, Sweden*

received 21st Nov. 2012, revised 28th June 2013, accepted 2nd Sep. 2013.

The classic Coupon-Collector Problem (CCP) is generalized to a setting where coupons can serve more than one purpose. We show how the expected number of coupons that needs to be drawn can be determined by means of enumerating transversals of hypergraphs, where coupons can be drawn either with or without replacement. Only basic probability theory is needed for this purpose. The transversal counting can be done efficiently by a recently introduced algorithm that encodes all possible transversals in an efficient way. Our results are illustrated by applications to, amongst others, chess and roulette.

Keywords: coupon collector, transversal

1 Introduction

In the popular game of Roulette a small metal bullet is spun and stopped at random on one of the $w = 37$ numbers $0, 1, 2, \ldots, 36$. Apart from 0 each one of these numbers has several properties. For example 13 is at the same time odd, black, in the second dozen, the first column; see Figure 1. We will show how to compute the expected time to encounter, in successive draws at random, all these properties: even, odd, red, black, 1–18, 19–36, 1st 12, 2nd 12, 3rd 12, 1st column, 2nd column, 3rd column.

Our general setting is as follows. Let W be a set whose w many elements will be viewed as "coupons". Let $\mathcal{G} = \{G_1, \cdots, G_h\}$ be any family of nonempty (not necessarily distinct) subsets. Thus $\bigcup \mathcal{G} \subseteq W$. By definition G_i contains exactly the coupons c of the i-th *goal* (purpose, property, etc.) Put another way, each fixed coupon $c \in W$ is *multipurpose* in the sense that it can serve many goals according to the sets G_i that contain c. If $\bigcup \mathcal{G} = W$, then every coupon has at least one goal. It is convenient to imagine the w many coupons as being located in an urn.

In a *length n trial* a set of n coupons is picked at random one by one, and all occuring goals are recorded. For any picked coupon some of its goals may have occured already and are not again taken into

†Email: `svante.janson@math.uu.se`
‡Email: `{mwild,swagner,dpl}@sun.ac.za`. S. W. was supported financially by the National Research Foundation of South Africa under grant number 70560.

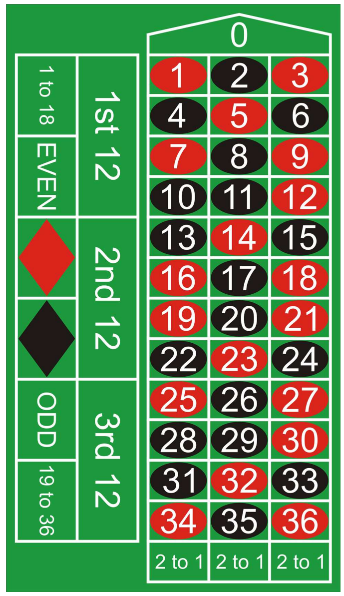

Fig. 1: Roulette seen as a multi-goal coupon collector's problem.

account. In a trial *with replacement* each coupon is put back into the urn after its goals have been ticked off. Thus at each moment every fixed coupon is drawn with probability $\frac{1}{w}$. In a trial *without replacement* no drawn coupons are put back. Then necessarily $n \leq w$. Again at each moment every coupon remaining in the urn has the same probability to be drawn (namely $\frac{1}{r}$, where r is the number of coupons remaining). A trial is *successful* if all h goals show up, i.e., if each of the sets G_i contains at least one of the coupons that have been drawn. We call a successful trial *sharply successful* if the h goals are only completed in the *last* draw.

The *Generalized Coupon-Collector Problem* (GCCP) is to calculate the expected length ℓ of a sharply successful trial. We shall use the notation $\ell = \ell_r(\mathcal{G})$ for a GCCP with *replacement*, and $\ell = \ell_{nr}(\mathcal{G})$ for a GCCP *without replacement*.

Coming back to Roulette, which serves as an example to illustrate our results, the coupons are the numbers $0, 1, \ldots, 36$, and they constitute the set W. The G_i in this example are the sets of all red numbers, all black numbers, all even numbers, etc. As we have seen, all coupons have several properties, except for 0, which has none (so $\bigcup \mathcal{G}$ is a proper subset of W in this case). Using the method of §4 it turns out that the expected length of a sharply successful trial in this GCCP with replacement is

$$\frac{54728027202913}{7600186994400} \approx 7.201.$$

If after each "drawing" one prevents the spinning wheel from delivering the same number again (so that we have a GCCP *without* replacement) the corresponding number is obviously smaller; in fact it is

$$\frac{65774035502891}{10043104242600} \approx 6.549.$$

Notice that we can also model our multipurpose coupons $c_i \in W$ with *different* drawing probabilities p_i as follows (for simplicity we only focus on drawings *with* replacement). If without great loss of generality all p_i's are rational, say $p_i = m_i/(wM)$, replace each c_i by m_i copies $c_i', c_i'', \cdots$ which all have exactly the same goals as c_i. Let W' be the new wM-element set of coupons and let $\mathcal{G}'$ match $\mathcal{G}$ in the obvious way. Then $\ell_{nr}(\mathcal{G}')$ is the expected length of a sharply succesful trial with *original* coupons $c_i \in W$ if they were subject to the drawing probabilities p_i.

There is a simple yet natural situation where the coupons in W *already* furnish (but do not "have") potentially different drawing probabilities. Namely, suppose that each $c \in W$ has exactly *one* of h goals which we then refer to as its *type*, and that different coupons can have the same type. Then $\mathcal{G}^* = \{G_1, \cdots, G_h\}$ is a *partition* of W and $p_i = |G_i|/|W|$ is the probability for drawing a type i coupon. This matches the "classic" Coupon-Collector Problem (CCP) except that in the latter framework there is no $\mathcal{G}^*$ but simply an unbounded supply of coupons. Each belongs to exactly one of h types, the i-th type being drawn with probability p_i. The expected length $\ell(p_1, \cdots, p_h)$ of a sharply successful trial is known to be [DB62, p.269]

$$\ell(p_1,\ldots,p_h) = \sum_{1\le i\le h} \frac{1}{p_i} - \sum_{1\le i\le j\le h} \frac{1}{p_i+p_j} + \sum_{1\le i<j<k\le h} \frac{1}{p_i+p_j+p_k} - \cdots \pm \frac{1}{p_1+\cdots+p_h}. \quad (1)$$

In particular, if $p_1 = p_2 = \cdots = p_h = \frac{1}{h}$ (call this the *homogeneous* CCP) then (1) can be shown [Fel57, Example IX.3(d)] to simplify to

$$\ell\left(\frac{1}{h}, \cdots, \frac{1}{h}\right) = hH(h), \quad (2)$$

where $H(h) := 1 + \frac{1}{2} + \cdots + \frac{1}{h}$ is the harmonic number.

For instance, setting $h = 6$ in (2) one finds that a die has to be thrown 14.7 times on average until all numbers have shown up. The CCP is very classical and has been studied by many authors from different perspectives, see for example [Pól30, FGT92, Daw91, Pin80]. It can also be found in various textbooks, such as those of Feller [Fel57, Example IX.3(d)], Blom, Holst and Sandell [BHS94, 7.5–7.6, 15.4], Motwani and Raghavan [MR95, §3.6], and Flajolet and Sedgewick [FS09, Example II.11]. Boneh and Hofri [BH97] provide a survey with a focus on computational methods. Recent extensions can be found in [FHL02] and [AOR03]. As an application, the CCP can be used for testing randomness [Fel57, Footnote 19, p. 59], [Knu97, 3.3.2.E].

Even though our setting is more general than the CCP, whose formula (1) is intimidating enough, the present article does not feature subtle probability arguments, but is rather based on counting transversals of set systems, exploiting an algorithm that has recently been introduced in [Wil13]. The connection to coupon collecting is discussed in §2, and §3 actually deals with counting transversals. Surprisingly perhaps, our approach to the GCCP appeals more to the GCCP *without replacement*. Only afterwards in §4 we tackle the GCCP with replacement.

A numerical evaluation of our method pitted against the inclusion-exclusion approach (1), as well as further examples, follow in §5 and §6.

2 The GCCP without replacement

In this and the next section all trials are assumed to be *without* replacement. Our approach to the GCCP is mathematically straightforward; the main point is that there is an efficient way to realize it algorithmically, as will be explained in the next section.

We shall use the notation $[h] := \{1, 2, \ldots, h\}$ for positive integers h. Recall that $G_i \subseteq W$ is the set of coupons of the i-th goal ($i \in [h]$). The *hypergraph* (= set system) $\mathcal{G} = \{G_1, \ldots, G_h\}$ fully determines all aspects of the GCCP. Specifically, $X \subseteq W$ is a *transversal* (or *hitting set*) of $\mathcal{G}$ if $X \cap G_i \neq \emptyset$ for all $i \in [h]$. Such a set X of coupons displays each goal at least once, and so each permutation of X corresponds to a successful trial. Conversely, each successful trial uses a set X of coupons that is a transversal of $\mathcal{G}$. Therefore, if

$$\tau_k := \text{number of } k\text{-element transversals of } \mathcal{G}$$

for some fixed $k \in \{0, 1, \ldots, w\}$, then exactly $k!\tau_k$ trials among the $w(w-1)\cdots(w-k+1)$ many length k trials are successful. Now let q_k be the probability that a length k trial is successful. In particular $q_0 = 0$ and $q_w = 1$, and generally

$$q_k = \frac{k!\tau_k}{w(w-1)\cdots(w-k+1)} \left(= \frac{\tau_k}{\binom{w}{k}} \right). \tag{3}$$

Note that

$$s_k := q_k - q_{k-1} \qquad (k \in [w]) \tag{4}$$

is the probability that a length k trial is sharply successful. Therefore $\ell_{nr}(\mathcal{G})$ can be found by calculating the numbers τ_k:

Theorem 1 *For drawings without replacement, the expected length of a sharply successful trial is*

$$\ell_{nr}(\mathcal{G}) = \sum_{k=1}^{w} k s_k = \sum_{k=0}^{w-1} (1 - q_k) = w - \sum_{k=1}^{w-1} q_k = w - \sum_{k=1}^{w-1} \frac{\tau_k}{\binom{w}{k}}. \tag{5}$$

3 Counting transversals

In this section all trials are still *without* replacement. In the previous section, we have seen that the GCCP can be reduced to the task of counting transversals, which will be illustrated by means of a simple example now. Consider a set $W = \{c_1, \ldots, c_8\}$ of eight coupons, each one of which serves between one and three goals from among G_1, G_2, G_3, G_4 according to Table 1.

For instance, the trials c_1, c_3, c_5 and c_6, c_2, c_8, c_7 are successful. The first is sharply successful, the second is not. In order to calculate the expected length of a sharply successful trial, we put $\mathcal{G}_1 := \{G_1, G_2, G_3, G_4\}$ and aim to count the τ_k many k-element transversals of $\mathcal{G}_1$ ($k \in [8]$).

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
G_1	X	X	X					
G_2			X			X	X	X
G_3		X		X	X	X		
G_4	X		X	X		X		X

Tab. 1: Toy problem with 8 coupons and 4 goals

In order to do so we shall encode the *transversal hypergraph*

$$\mathcal{T}r(\mathcal{G}_1) := \{X \subseteq W : \ X \text{ is transversal of } \mathcal{G}_1\}$$

in a compact way, i.e. not by listing the transversals one by one. This is the main idea behind the transversal e-algorithm that was recently introduced in [Wil13]. We do not repeat the full details of its implementation here, but rather focus on its *output* only, which is a sequence of $\{0, 1, 2, e\}$-valued rows.

Note first that each subset $X \subseteq W$ can be encoded as a bitstring of length 8, where each 1 indicates an element of the set. For instance, the set $\{c_1, c_3, c_6, c_7\}$ would be encoded as $(1, 0, 1, 0, 0, 1, 1, 0)$. In order to obtain a more compact representation, we introduce additional symbols: 2 is a "don't care" symbol, which indicates that the corresponding entry can be 0 or 1. For example, consider the $\{0, 1, 2\}$-valued row (ignore the boldface for the moment)

$$r := (2, 1, 0, 2, 2, \mathbf{1}, 2, \mathbf{2}).$$

Each of the 2's stands for a 0 or a 1, so this row r encodes a total of $2^5 = 32$ length 8 bitstrings (including, for example, $(0, 1, 0, 1, 1, 1, 0, 0)$ and $(1, 1, 0, 0, 1, 1, 0, 1)$), or equivalently 32 subsets $X \subseteq W$. Because $\{c_2, c_6\} \subseteq X$ for each $X \in r$, each $X \in r$ is a transversal of $\mathcal{G}$. Similarly, if

$$r' := (2, 1, 0, 2, 2, \mathbf{0}, 2, \mathbf{1}),$$

then all 16 members $X \in r'$ contain c_2 and c_8 and are thus transversals of $\mathcal{G}$. Note that the sets represented by r and r' are disjoint: all $X \in r$ contain c_6, while no $X \in r'$ does. Using r and r' is clearly more efficient than listing $32 + 16 = 48$ bitstrings individually, but one can compress things even more by introducing another symbol: by definition, a string of symbols $ee \cdots e$ (not necessarily on contiguous positions) means that any 0-1-pattern with *at least one* 1 is allowed. In other words, only $00 \cdots 0$ is forbidden. Now note that r and r' only differ in two positions, indicated by boldface digits. Putting ee in those two positions, we cover both r (all of whose elements contain c_6 and possibly also c_8) and r' (whose elements do not contain c_6, but necessarily c_8). This gives us

$$r \cup r' = (2, 1, 0, 2, 2, e, 2, e). \tag{6}$$

It turns out that the whole transversal hypergraph $\mathcal{T}r(\mathcal{G}_1)$ can be written as a disjoint union of five such $\{0, 1, 2, e\}$-valued rows (Table 2), which are generated by the aforementioned e-algorithm. For instance, r_3 is the row in (6). Note that a row may contain several e-blocks, which are then notationally distinguished (writing e, e' or $e_1, e_2, \ldots$) as in row r_5.

It is shown in Theorem 3 of [Wil13] how generally for an h-element hypergraph $\mathcal{G} \subseteq \mathcal{P}([w])$ its transversal hypergraph $\mathcal{T}r(\mathcal{G})$ can be represented as a union of R disjoint $\{0, 1, 2, e\}$-valued rows in time $O(Rh^2w^2)$. As mentioned earlier, the details of the e-algorithm that performs this task can be found in [Wil13] as well. Let us only emphasize that it does not make use of a merging process as we used to reduce r and r' to a single row in (6) (which only served to show that $0, 1, 2, e$ are more powerful than $0, 1, 2$ alone), but rather generates the $\{0, 1, 2, e\}$-valued rows *from scratch*. By the disjointness of rows R is always bounded by $N := |\mathcal{T}r(\mathcal{G})|$, and in practice often $R \ll N$. For instance, in our example $R = 5$ while $N = 120 + 16 + 48 + 6 + 9 = 199$ (see Table 2).

Once such a list of rows encoding all possible transversals has been generated, counting them is not difficult any more. Note again that the rows generated by the e-algorithm are mutually disjoint. For the

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	
r_1	2	e	1	e	e	e	2	2	$\|r_1\| = 120$
r_2	1	0	0	2	2	1	2	2	$\|r_2\| = 16$
r_3	2	1	0	2	2	e	2	e	$\|r_3\| = 48$
r_4	e	1	0	e	2	0	1	0	$\|r_4\| = 6$
r_5	1	0	0	e	e	0	e'	e'	$\|r_5\| = 9$

Tab. 2: $\mathcal{T}r(\mathcal{G}_1)$ as disjoint union of $\{0, 1, 2, e\}$-valued rows

example in Table 2, we have e.g. $r_3 \cap r_4 = \emptyset$ because each $X \in r_3$ has $X \cap \{c_6, c_8\} \neq \emptyset$, and each $Y \in r_4$ has $Y \cap \{c_6, c_8\} = \emptyset$. Accordingly, it suffices to count the transversals represented by each row individually. Each 2 contributes a factor 2, and each e-block of length ℓ yields a factor $2^\ell - 1$. Thus for example $|r_3| = 2^3 \cdot (2^2 - 1) = 48$ or $|r_5| = (2^2 - 1) \cdot (2^2 - 1) = 9$.

Calculating

$$\mathsf{Card}(r, k) := |\{X \in r : \ |X| = k\}| \tag{7}$$

for an arbitrary $\{0, 1, 2, e\}$-valued row r of length w, and any $k \in [w]$, is only slightly more subtle than getting $|r|$, and can be done in time $O(kw^2 \log^2 w)$ [Wil13, Theorem 1]. One only needs to replace each 0 by a factor 1, each 1 by a factor x, each 2 by a factor $1 + x$, and each e-block of length ℓ by $(1 + x)^\ell - 1$ to obtain a polynomial associated with each row whose coefficients are the $\mathsf{Card}(r, k)$. For example, row r_1 gives us

$$\mathsf{pol}(r_1, x) = x(1 + x)^3((1 + x)^4 - 1) = 4x^2 + 18x^3 + 34x^4 + 35x^5 + 21x^6 + 7x^7 + x^8.$$

The full table that we obtain for our toy problem looks as follows:

$k =$	1	2	3	4	5	6	7	8		$\|r_i\|$
$\mathsf{Card}(r_1, k) =$	0	4	18	34	35	21	7	1		120
$\mathsf{Card}(r_2, k) =$	0	1	4	6	4	1	0	0		16
$\mathsf{Card}(r_3, k) =$	0	2	9	16	14	6	1	0		48
$\mathsf{Card}(r_4, k) =$	0	0	2	3	1	0	0	0		6
$\mathsf{Card}(r_5, k) =$	0	0	4	4	1	0	0	0		9
$\tau_k =$	0	7	37	63	55	28	8	1		199

For instance, the transversals counted by $\mathsf{Card}(r_5, 4) = 4$ are

$$\{c_1, c_4, c_5, c_7\}, \ \{c_1, c_4, c_5, c_8\}, \ \{c_1, c_4, c_7, c_8\}, \text{ and } \{c_1, c_5, c_7, c_8\}.$$

Having the τ_k's we can evaluate the probability q_k of having a successful trial of length k by Formula (3). For example, we have $q_2 = \tau_2 / \binom{8}{2} = \frac{7}{28} = \frac{1}{4}$. More generally we obtain the following table:

q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
0	$\frac{1}{4}$	$\frac{37}{56}$	$\frac{9}{10}$	$\frac{55}{56}$	1	1	1

Hence (5) gives

$$\ell_{nr}(\mathcal{G}_1) = 8 - q_7 - \cdots - q_2 - q_1 = \frac{449}{140} \approx 3.2. \tag{8}$$

4 The GCCP with replacement

Without further mention all trials in this section are *with* replacement. Let t'_n be the number of successful length n trials, i.e. trials where all goals of coupons have occured at some point (so $t'_0 = 0$). Thus

$$q'_n := \frac{t'_n}{w^n} \qquad (n \geq 0) \tag{9}$$

is the probability that a length n trial is successful, and

$$s'_n := q'_n - q'_{n-1} \qquad (n \geq 1) \tag{10}$$

is the probability that a length n trial is sharply successful. The expected length of a sharply successful trial is

$$\ell_r(\mathcal{G}) = \sum_{n=1}^{\infty} n s'_n. \tag{11}$$

As to calculating the numbers t'_n, observe that no matter how coupons c_i are repeated in a length n trial, the underlying set of (distinct) coupons must be a k-element transversal X of $\mathcal{G}$, for some $k \leq n$. For a fixed k-element set of coupons $X \subseteq W$ the number of length n trials with underlying set X equals the number of ways to distribute n distinct balls (corresponding to the positions in the trial) to k distinct buckets (corresponding to the coupons) in such a way that no bucket stays empty. It is well known that this number (the number of surjections from an n-element set to a k-element set) is $k!S(n,k)$, where $S(n,k)$ is the Stirling number of the second kind. Accordingly, recalling that τ_k is the number of k-element transversals of $\mathcal{G}$, we deduce that

$$t'_n = \begin{cases} 1!\,\tau_1 S(n,1) + 2!\,\tau_2 S(n,2) + \cdots + n!\,\tau_n S(n,n), & n \leq w, \\ 1!\,\tau_1 S(n,1) + 2!\,\tau_2 S(n,2) + \cdots + w!\,\tau_w S(n,w), & n > w. \end{cases} \tag{12}$$

Fortunately the infinite sum in (11) can be evaluated as a finite sum. Moreover, one can express it in terms of the probabilities q_k that a length k-trial *without* replacement is successful. Recall from (3) that $\tau_k = q_k \binom{w}{k}$.

Theorem 2 *For drawings with replacement the expected length of a sharply successful trial is*

$$\ell_r(\mathcal{G}) = w \sum_{k=0}^{w-1} \frac{1-q_k}{w-k} = w \left(H(w) - \sum_{k=1}^{w-1} \frac{q_k}{w-k} \right) = wH(w) - \sum_{k=1}^{w-1} \frac{\tau_k}{\binom{w-1}{k}}.$$

Proof: Drawing with replacement yields an infinite sequence of coupons. If we ignore all repetitions and only keep the *new* coupons, i.e., the coupons that have not occured earlier in the sequence, we obtain a subsequence of distinct coupons. With probability 1, every coupon occurs eventually, so the subsequence will contain all w coupons, and by symmetry they can come in any order with the same probability $1/w!$; hence, the subsequence of new coupons is the same as drawing without replacement.

If we draw with replacement we stop when we first have attained all goals, i.e, when the trial is sharply successful. Since repeated coupons do not help (or hinder), it is clear that we will always stop at a point when we get a new coupon. Moreover, by the argument above, the probability that we stop when we get the k-th new coupon is precisely the probability s_k given in (4) that a trial with k drawings without replacement is sharply successful. The positions of the new coupons are, by symmetry, stochastically independent of the sequence of values of the new coupons. Hence, provided that we stop at the k-th new coupon, the expected number of coupons drawn equals the expected number e_k of drawings required to get k distinct coupons (ignoring their values), which is known to be

$$e_k = \sum_{i=1}^{k} \frac{w}{w+1-i}.$$

See e.g. [Fel57, Example IX.3(d)] for this well-known fact; the standard argument is that when we have got j distinct coupons, the probability that the next coupon is new is $(w-j)/w$, and thus the expected waiting time for the next new coupon is $w/(w-j)$. Recalling that $q_0 = 0$ and $q_w = 1$, we deduce:

$$\begin{aligned}
\ell_r(\mathcal{G}) &= \sum_{k=1}^{w} s_k e_k = \sum_{k=1}^{w} \left((q_k - q_{k-1}) \sum_{i=1}^{k} \frac{w}{w+1-i} \right) \\
&= (q_1 - q_0)\frac{w}{w} + (q_2 - q_1)\left(\frac{w}{w} + \frac{w}{w-1}\right) + \cdots + (q_w - q_{w-1})\left(\frac{w}{w} + \frac{w}{w-1} + \cdots + \frac{w}{1}\right) \\
&= \frac{w}{w}(q_w - q_0) + \frac{w}{w-1}(q_w - q_1) + \cdots + \frac{w}{2}(q_w - q_{w-2}) + \frac{w}{1}(q_w - q_{w-1}) \\
&= wH(w) - w\sum_{k=1}^{w-1} \frac{q_k}{w-k} = w\sum_{k=0}^{w-1} \frac{1-q_k}{w-k}.
\end{aligned}$$

From $\tau_k = q_k\binom{w}{k}$, the rightmost formula in the Theorem follows. □

For instance, for our running example Theorem 2 yields $\ell_r(\mathcal{G}_1) = \frac{59}{15} \approx 3.9$ as opposed to $\ell_{nr}(\mathcal{G}_1) \approx 3.2$ from (8). Notice that $\ell_r(\mathcal{G}) = wH(w)$ if and only if all q_k $(k < w)$ are equal to 0, which is the classical case where each coupon has only one goal (and all these goals are distinct). The other extreme $\ell_r(\mathcal{G}) = w\frac{1}{w} = 1$ occurs if and only if all $q_k = 1$ $(1 \le k \le w)$, which means that every coupon fulfils every goal.

Remark 1 The key tool of our proof of Theorem 2 is the fact that $q_k - q_{k-1}$ (the probability of a length k trial without replacement being sharply successful) is also the probability that a trial with replacement is sharply successful when the k-th distinct coupon is drawn. This fact can also be used to compute the variance (and in principle also higher moments) of the trial length: to this end, note that the expectation of the square of the number of coupons needed to collect k distinct coupons is (by the same argument as before, decomposing into k independent geometrically distributed random variables)

$$\sum_{i=1}^{k} \frac{(i-1)w}{(w+1-i)^2} + \left(\sum_{i=1}^{k} \frac{w}{w+1-i}\right)^2.$$

Now repeating the argument of the proof of Theorem 2 yields the following expression for the variance:

$$\sum_{k=0}^{w-1} (1-q_k)\left(\frac{w(w+k)}{(w-k)^2} + \frac{2w^2}{w-k}(H(w) - H(w-k))\right) - \ell_r(\mathcal{G})^2.$$

In our toy example, this yields a variance of $\frac{836}{225} \approx 3.7$. For drawings without replacement, the situation is much simpler, and the variance is

$$\sum_{k=0}^{w-1} (2k+1)(1-q_k) - \ell_{nr}(\mathcal{G})^2,$$

which equals $\frac{18339}{19600} \approx 0.9$ in our example. This ends Remark 1.

Let us finally consider a related problem: how many goals will be fulfilled once n coupons have been drawn? Referring to Table 1, the probability that goal 1 does *not* belong to a randomly drawn coupon is $a_1 = \frac{5}{8}$. Similarly define a_2, a_3, a_4, so $a_i = 1 - m_i/w$ if goal i is served by m_i coupons.

Coupled to a random drawing of coupons, set the random variable $X_i := 1$ if goal i comes up, and $X_i := 0$ otherwise. Hence the expected value of X_i in a length n trial is $E_n[X_i] = 1 - a_i^n$. For drawing without replacement, the corresponding formula is

$$E_n[X_i] = 1 - \frac{\binom{w-m_i}{n}}{\binom{w}{n}} = 1 - \frac{(w-m_i)!\,(w-n)!}{(w-n-m_i)!\,w!},$$

interpreted as 1 if $m_i + n > w$.

By linearity of expectation, one calculates that $e_n = \sum_{i=1}^{h} E_n[X_i]$ is the expected number of goals gathered in a length n trial. For instance, $e_4 \approx 3.7$ for drawing with replacement in our running example.

5 The nonhomogenous CCP: Pitting the e-algorithm against inclusion-exclusion

In the introduction, we gave the formula

$$\ell(p_1,\dots,p_h) = \sum_{1\le i\le h} \frac{1}{p_i} - \sum_{1\le i\le j\le h} \frac{1}{p_i+p_j} + \sum_{1\le i<j<k\le h} \frac{1}{p_i+p_j+p_k} - \cdots \pm \frac{1}{p_1+\cdots+p_h}.$$

for the expected length of a sharply successful trial with h single-purpose coupons whose probabilities are $p_1, p_2, \dots, p_h$. Boneh and Hofri [BH97, p. 43] emphasize the computational difficulty to evaluate this formula as h increases, and then go on to use integration for approximation. Recall that for rational p_i's in the (classic) CCP, say

$$p_1 = \frac{1}{10},\ p_2 = \frac{2}{10},\ p_3 = \frac{3}{10},\ p_4 = \frac{4}{10},$$

our approach uses $W = [10]$ and the partition

$$\mathcal{G}^* = \{\{1\},\{2,3\},\{4,5,6\},\{7,8,9,10\}\}.$$

Because the sets in $\mathcal{G}^*$ are disjoint, we can do with a *single* $\{0,1,2,e\}$-valued row

$$r = (1,\ \ e_2, e_2,\ \ e_3, e_3, e_3,\ \ e_4, e_4, e_4, e_4).$$

h	$\ell_r(\mathcal{G}^*)$	exclusion	incl-excl.
10	68.9846	0	0.2
15	150.606	0	7.7
27	474.463	0.3	43193
50	1600.38	4.1	-
100	6338.75	72	-
150	14215.1	455	-
200	25229.5	1829	-
400	100667	96272	-

Tab. 3: Total time in seconds taken when computing $\ell_r(\mathcal{G}^*)$ by the e-algorithm (exclusion) and by the inclusion-exclusion algorithm.

One computes the numbers $\tau_k = \mathsf{Card}(r,k)$ $(k \in [10])$ as we have seen in §3 and from them $\ell_r(\mathcal{G}^*)$ according to Theorem 2. Table 3 compares the e-algorithm with the inclusion-exclusion approach (1) on instances $(p_1, \dots, p_h)$ of the particular but natural type

$$p_1 = \frac{1}{w}, \quad p_2 = \frac{2}{w}, \quad \dots, \quad p_h = \frac{h}{w} \quad \left(\text{hence } w = 1 + \cdots + h = \frac{h(h+1)}{2}\right)$$

which is uniquely defined by h (= first column in Table 3). As to inclusion-exclusion, we used a standard Gray-code in order to more economically generate the subsets of $[h]$ one by one from their predecessors, and also used that for common denominator probabilities one can simplify the terms in (1); say

$$\frac{1}{p_i + p_j + p_k} = \frac{1}{\frac{i}{w} + \frac{j}{w} + \frac{k}{w}} = \frac{w}{i + j + k}.$$

The value of $\ell_r(\mathcal{G}^*)$ is rounded to 6 digits albeit Mathematica, provided with the numbers τ_k $(k \in [w])$, delivered the *exact* value as a fraction of two very large integers. For instance $h = 400$ gives $w = 80200$ and 3108 sec of the 96272 sec total time were spent on plugging $\tau_1, \tau_2, \dots, \tau_{80200}$ into the formula of Theorem 2. As is apparent, inclusion-exclusion (formula (1)) cannot compete.

For the particular p_i's considered one can show [DB62, p.269] that $\ell_r(\mathcal{G}^*)$ is asymptotically equal to $\left(\frac{4\pi}{\sqrt{3}} - 6\right)\binom{h+1}{2}$ as $h \to \infty$. Already for $h = 15$ the latter gives the tight approximation 150.624 to the true (rounded) value 150.606.

6 Information spreading and the expected time to dominate a chess board

In many GCCP applications the goals of a coupon c are *other coupons*, namely those that c wishes to "influence" in some way. More succinctly, we may consider a graph G with vertex set W as a group of people whose friendship relations are reflected by the edges of G. Suppose members $c \in W$ are phoned at random from outside W and told a piece of information. If c shares the news with all his friends, what is the expected number $\ell_r(\mathcal{G})$ of phone calls necessary before the whole of W is informed? (The *minimum* number of phone calls necessary is called the *domination number* of G.) What is the analogous number $\ell_{nr}(\mathcal{G})$ when nobody is phoned twice? The method presented in the previous sections can provide answers to these questions.

A nice illustrative example of the graph framework is the problem to determine the expected number ℓ_{nr}(queens) of queens it takes when they are placed on a chessboard at random until the queens dominate the board, i.e., all 64 squares (coupons) are occupied or threatened. If occupied squares can still be drawn (without effect apart from increasing the trial's length), let ℓ_r(queens) be the corresponding number. We also define ℓ_{nr}(rooks), ℓ_{nr}(kings), ... in an analogous fashion. One obtains the following results (rounded to four decimals):

$$\begin{array}{lcllcl} \ell_{nr}(\text{queens}) & = & 11.8402 & \ell_r(\text{queens}) & = & 15.2945 \\ \ell_{nr}(\text{rooks}) & = & 15.0045 & \ell_r(\text{rooks}) & = & 17.1308 \\ \ell_{nr}(\text{kings}) & = & 30.4091 & \ell_r(\text{kings}) & = & 42.4282 \end{array}$$

If one does not consider a square occupied by a queen as threatened by her (after all, an unthreatened knight can capture her), the numbers ℓ_{nr}(queens) and ℓ_r(queens) grow to ℓ^*_{nr}(queens)$= 12.7094$ respectively ℓ^*_r(queens) $= 16.3149$.

Similar GCCP applications e.g. to trading card games such as *Magic: The Gathering*, and much more, are conceivable. One may also further want to generalize to problems where a certain number α_i of coupons in class G_i needs to be collected to fulfil the task. We hope to do so in a future publication.

References

[AOR03] Ilan Adler, Shmuel Oren, and Sheldon M. Ross. The coupon-collector's problem revisited. *J. Appl. Probab.*, 40(2):513–518, 2003.

[BH97] Arnon Boneh and Micha Hofri. The coupon-collector problem revisited—a survey of engineering problems and computational methods. *Comm. Statist. Stochastic Models*, 13(1):39–66, 1997.

[BHS94] Gunnar Blom, Lars Holst, and Dennis Sandell. *Problems and snapshots from the world of probability*. Springer-Verlag, New York, 1994.

[Daw91] Brian Dawkins. Siobhan's problem: the coupon collector revisited. *Amer. Statist.*, 45(1):76–82, 1991.

[DB62] F. N. David and D. E. Barton. *Combinatorial chance*. Hafner Publishing Co., New York, 1962.

[Fel57] William Feller. *An introduction to probability theory and its applications. Vol. I.* John Wiley and Sons, Inc., New York, 1957. 2nd ed.

[FGT92] Philippe Flajolet, Danièle Gardy, and Loÿs Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Appl. Math.*, 39(3):207–229, 1992.

[FHL02] Dominique Foata, Guo-Niu Han, and Bodo Lass. Les nombres hyperharmoniques et la fratrie du collectionneur de vignettes. *Sém. Lothar. Combin.*, 47:Article B47a, 20, 2001/02.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[Knu97] Donald E. Knuth. *The art of computer programming. Vol. 2*. Addison-Wesley Publishing Co., Reading, Mass., third edition, 1997.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, Cambridge, 1995.

[Pin80] N. Pintacuda. Coupons collectors via the martingales. *Boll. Un. Mat. Ital. A (5)*, 17(1):174–177, 1980.

[Pól30] George Pólya. Eine Wahrscheinlichkeitsaufgabe in der Kundenwerbung. *Z. Angew. Math. Mech.*, 10(1):96–97, 1930.

[Wil13] Marcel Wild. Counting or producing all fixed cardinality transversals. *Algorithmica*, to appear, 2013.