WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/130817

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

Speed and concentration of the covering time for structured coupon collectors

Journal:	*Applied Probability Journals*
Manuscript ID	AP-17568.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Falgas-Ravry, Victor; Umeå Universitet Teknisk-Naturvetenskaplig Fakultet, Department of Mathematics and Mathematical Statistics Larsson, Joel; University of Warwick, Mathematics Institute Markström, Klas; Umeå Universitet Teknisk-Naturvetenskaplig Fakultet, Department of Mathematics and Mathematical Statistics
2010 Mathematics Subject Classification (Primary):	60C05
2010 Mathematics Subject Classification (Secondary):	05C80, 05B40, 60G99
Keywords:	Coupon collector, Combinatorial probability, Concentration inequalities

SCHOLARONE™
Manuscripts

Applied Probability Trust (4 December 2019)

SPEED AND CONCENTRATION OF THE COVERING TIME FOR STRUCTURED COUPON COLLECTORS

VICTOR FALGAS-RAVRY,* *Department of Mathematics and Mathematical Statistics, Umeå Universitet. Research partially supported by a grant from the Kempe foundation.*

JOEL LARSSON,** *Mathematics Institute, Warwick University. Supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 639046).*

KLAS MARKSTRÖM,*** *Department of Mathematics and Mathematical Statistics, Umeå Universitet. Research supported by a grant from the Swedish Research Council (Vetenskapsrådet).*

Abstract

Let V be an n-set, and let X be a random variable taking values in the powerset of V. Suppose we are given a sequence of random coupons $X_1, X_2, \ldots$, where the X_i are independent random variables with distribution given by X. The covering time T is the smallest integer $t \geq 0$ such that $\bigcup_{i=1}^{t} X_i = V$. The distribution of T is important in many applications in combinatorial probability, and has been extensively studied. However the literature has focused almost exclusively on the case where X is assumed to be symmetric and/or uniform in some way.

In this paper we study the covering time for much more general random variables X; we give general criteria for T being sharply concentrated around its mean, precise tools to estimate that mean, as well as examples where T fails to be concentrated and when structural properties in the distribution of X allow for a very different behaviour of T relative to the symmetric/uniform case.

* Email address: victor.falgas-ravry@math.umu.se
** Email address: joel.larsson@warwick.ac.uk
** Email address: klas.markstrm@math.umu.se

Keywords: Coupon collector; concentration inequalities; combinatorial probability

2010 Mathematics Subject Classification: Primary 60C05
Secondary 05B40;05C80;60G99

1. Introduction

In this paper, we study the random covering problem in a general setting: we are interested in the distribution of the covering time T for general distributions of the random covering variable X. With the exception of a result of Aldous [4], discussed later, this is as far as we are aware the first time the covering problem is studied in this generality. However, the question is a natural one: there are many applications where the covering variable is 'non-uniform' in a way which puts it outside the current literature on covering problems. Also, a common drawback of many of the existing exact results about covering processes is that the expressions obtained often involve a large number of summands and are hard to evaluate directly; this was pointed out for example by Sellke [55] and Adler and Ross [2].

Our own focus is on simple, easy-to-use concentration inequalities for the covering time which can be applied in a straightforward way. The basic questions we seek to answer: how does the distribution of X affect the covering time? Can we exploit 'structure' in the choice of X to 'speed up' or 'slow down' the covering? And when can we guarantee that T is sharply concentrated?

Our paper is structured as follows. In Section 2, we gather together elementary bounds for the covering time, and identify the range of possible speeds of the covering process, giving examples going over the entire spectrum. We follow on in Section 3 with the main results of this paper, namely general structure theorems giving sufficient conditions for the covering time of an arbitrary random covering variable to be sharply concentrated. These are stated in Section 3.1 and proved in Section 3.3–3.6. In Section 4 we discuss 'fast' coverage by structured random variables. Finally in Section 5 we give some applications of our results to the connectivity of random graphs, continuum percolation, random graph colourings, the unsatisfiability threshold for k-SAT and the appearance of perfect matchings in random graphs. We end with some questions and remarks.

1.1. Definitions

Let V be a finite set; usually we shall take $V = [n] := \{1, 2, \dots, n\}$. Let X be a random variable taking values in the power-set of V. A random variable X taking values in the power-set of V is referred to as a *random covering variable*, or *random COUPON.* We say that X is *exchangeable* if the law of X is invariant under every permutation of V. We call X *transitive* if the law of X is invariant under the action of a transitive subgroup of $\mathrm{Sym}(V)$. We say X is *balanced* if for every $v, v' \in V$ we have $\mathbb{P}(v \in X) = \mathbb{P}(v' \in X)$. Finally, X is *k-uniform* if $|X| = k$ with probability 1.

We consider an infinite sequence $\mathbf{X} = \{X_1, X_2, \dots\}$ of i.i.d. random covering variables $X_i \sim X$. We view this as a sequence of random coupons received by a coupon collector; we refer to X_i as the i^{th} coupon, and to the collector as the X-coupon collector. We set $C_t = C_t(\mathbf{X}) = \bigcup_{i \le t} X_i$ to be the collection of elements of V covered by the union of the first t coupons $X_1, X_2, \dots, X_t$, and define the *covering time* $T = T(\mathbf{X})$ to be

$$T = \inf\{t : \ C_t = V\}.$$

This quantity T is sometimes referred to as the *waiting time* in the literature. Note that T could be infinite if, for example, X almost surely does not cover (contain) some element $v \in V$. We also define

$$T_{\frac{1}{2}} = T_{\frac{1}{2}}(\mathbf{X}) = \inf\left\{t : \ \mathbb{P}(C_t = V) \ge \frac{1}{2}\right\},$$

to be the earliest time by which we have at least a fifty percent chance of having covered V, and for a subset $A \subseteq V$ we let $\tau_A = \tau_A(\mathbf{X})$ be the least t such that $A \subseteq C_t$ if it exists, and infinite otherwise. For $v \in V$, we let $d_v(t)$, the *degree* of v at time t, denote the number of sets X_i with $i \le t$ containing v.

Our aim in this paper is to prove concentration results for the covering time T in a general setting, i.e. for arbitrary random covering variables X. We shall also consider applications where $V \subseteq \mathbb{R}^d$ is a compact set and X takes values among the compact subsets of V, and define V_t, T and $T_{\frac{1}{2}}$ analogously to the discrete case. In this continuous setting, we shall use $|A|$ to denote the Lebesgue measure of a set A. For a sequence of events $(\mathcal{A}_n)_{n \in \mathbb{N}}$, we say that $\mathcal{A}_n$ holds *with high probability* (*w.h.p.*) if

$$\lim_{n \to \infty} \mathbb{P}(\mathcal{A}_n) = 1.$$

Also, we say that a sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ is *sharply concentrated* around $f(n)$ if $(Y_n/f(n))_{n\in\mathbb{N}}$ converges to 1 in probability, i.e. if $\forall\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n/f(n) - 1| > \varepsilon) = 0.$$

We recall here the standard Landau notation for asymptotic behaviour. For functions $f, g :\ \mathbb{N} \to \mathbb{R}_{\geq 0}$, we say that $f = O(g)$ if there exists $C > 0$ such that $f(n) \leq Cg(n)$ for all but finitely many n. We write $f = o(g)$ to denote that $\lim_{n\to\infty} f(n)/g(n) = 0$. Finally, we use $f = \Omega(g)$ to denote $g = O(f)$, we write $f = \theta(g)$ if both $f = O(g)$ and $f = \Omega(g)$ hold, and use $f = \omega(g)$ or $f \gg g$ to denote $g = o(f)$.

1.2. Some examples

We give below some examples of random covering variables X, illustrating the definitions of exchangeable, transitive and balanced above.

Our first example is that of the quintessential 'nice' random covering variable: the k-uniform, exchangeable random coupon variable, which was the focus of most of the previous work on coupon collecting.

Example 1. Let X be a k-set of $V = [n]$ selected uniformly at random, for some $k :\ 1 \leq k \leq n$; X is k-uniform and exchangeable.

Next we give three examples of 'structured' coupon collectors, of the kind that motivate our work in this paper.

Example 2. Let G be a graph on n vertices. Let X be the random coupon obtained by selecting a vertex x of $V = V(G)$ uniformly at random and taking as the coupon the closed neighbourhood of x in G, $\bar{\Gamma}(x) := \{y \in V(G) :\ xy \in E(G)\} \cup \{x\}$. Here X is balanced if and only if the graph G is regular, and transitive if and only if the graph G has a transitive automorphism group.

Example 3. Let $V = Q_d$ be the discrete d-dimensional hypercube $\{0,1\}^d$, and let X be a k-dimensional subcube of Q_d chosen uniformly at random, for some $k :\ 0 \leq k \leq d$. This random covering variable X is transitive and 2^k-uniform but not exchangeable, and, as described in Section 5, underlies the random SAT problem.

Example 4. Let V be the square of area n, $[0, \sqrt{n}]^2 \subset \mathbb{R}^2$, and let X be the intersection of V with the disc of radius r about a uniformly chosen random point $x \in V$.

This random covering variable X is neither uniform nor balanced, due to boundary effects; it is relevant to problems of coverage in random geometric graph theory (see Section 5).

1.3. Motivation for coupon collecting

The problem of determining the covering time of a set by a union of random subsets is of fundamental importance in several areas of mathematics, most notably in probability theory, discrete mathematics and mathematical statistics. This importance is illustrated both by the age of the problem — in its simplest form, the covering problem can be traced back to de Moivre [44] in 1711 — and by the many appellations it has amassed through the years. It has been studied by a large number of mathematicians from a variety of backgrounds and under a variety of names: matrix occupancy [15], allocation of particles in complexes [59], committee problem [40], chromosome problem [56], urn-sampling [55] or urn-occupancy problem [16], the Dixie cup problem [46], and, perhaps most famously, the coupon collector problem [6].

The ubiquitous nature of the covering problem is due to its wide range of applications. It is linked to the study of random walks [3], colouring [10] and degree sequences [41] in graph theory. In Section 5 we also give applications of the coupon collector problem to the connectivity of random graphs. The performance analysis of many exploration or optimisation algorithms in theoretical computer science involves a solution to a covering problem [47, 58], while the unsatisfiability threshold for SAT corresponds to the cover time of a hypercube by random subcubes [34, 39]. The 'reverse' coupon collector problem — estimating the size of V given C_t — is important to IP trace-back algorithms [54] and the study of biological diversity [48, 45] amongst other applications, while the study of the degrees $d_v(t)$, $v \in V$, is central to hashing and load balancing [52]. There are further applications in population genetics [36, 51], evolutionary algorithms for fitness selection [49] and disordered system physics [28].

1.4. Previous work on coupon collecting

Most of the previous work on covering problems in the spirit of the present paper focused on the case where X is an exchangeable, k-uniform random covering variable. The case $k = 1$, known as the *coupon collector's problem* has received by far the most attention. It can be traced back to de Moivre [44], who computed the probability $\mathbb{P}(V_t = V)$ exactly. Laplace [12] later generalised de Moivre's result to the k-uniform case for $k \geq 1$.

The second half of the twentieth century saw great activity on the problem, with many results replicated independently by researchers. Pólya [50] gave an expression for the expected covering time T in the k-uniform exchangeable case. Feller's textbook [20] included a computation of $\mathbb{E}T$ in the special case $k = 1$. Still in the case $k = 1$, Newman and Shepp [46] computed the expected time necessary for m-coverage of V (covering every point at least m times). Erdős and Rényi [16] computed the asymptotic distribution of the m-coverage time for $m \geq 1$; as their result is of particular relevance to this paper, we state it below:

Theorem 1. (Erdős–Rényi.) *Let $V = [n]$, and let X be the random coupon obtained by selecting a singleton from V uniformly at random. Denote by T^m be the time at which every point of V has been covered by at least m of the coupons $X_1, \ldots, X_{T^m}$. Then for every $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(T^m < n\log n + (m-1)n\log\log n + xn\right) = e^{-e^{-x}}.$$

In particular T^m, is sharply concentrated around

$$n\log n + (m-1)n\log\log n.$$

Continuing work on the 1-uniform exchangeable case, Baum and Billingsley [6] proved results on the asymptotic distribution of the size of C_t (the number of coupons collected by time t) as a function of t; Holst [26] later generalised their result to unbalanced 1-uniform random covering variables X. A number of researchers worked on the distribution of the degrees $(d_v(t))_{v\in V}$, such as Eicker, Siddiqui and Mielke [15], Mikhailov [43], Barbour and Holst [5] and Khakimullin and Enatskaya [35], all of whom dealt with the k-uniform case with $k > 1$ as well. A number of the papers cited above also deal with the unbalanced, 1-uniform case; let us mention in addition the work of Papanicolaou, Kokolakis and Boneh [47], who gave an expression for the expected covering time when X is a randomly chosen 1-uniform random covering variable.

In the exchangeable k-uniform case with $k > 1$, several researchers [40, 24, 57] computed, like Laplace, the expected covering time, giving closed-form formulae. Vatutin and Mikhailov [59] determined the asymptotic distribution of the number of degree zero (i.e. uncovered) vertices, which in turn gives results on the distribution of the covering time.

Recently Ferrante and Frigo [21] gave an expression for the expected covering time when X is a k-uniform covering random variable with different $v \in V$ receiving different weights. In a different direction, improving results of Sellke, Ivchenko [30] computed the asymptotic distribution of the covering time when $n \to \infty$ and X is a fixed (i.e. not depending on n) non-uniform exchangeable random variable; similar results were also obtained by Johnson and Sellke [32], while a closed-form expression for the expectation of T appeared in Adler and Ross [2].

Finally, Aldous [4] proved a general abstract result about covering times, in connection with random walks on graphs. To state his result, we need one more definition. Given a random covering variable X and an X-coupon collector, we let $B = B(\mathbf{X})$ denote the set of "holdouts", which is to say the last subset of V to be covered: $B = C_T \setminus C_{T-1}$ (if the coupon collector does not cover V, we set B to be the collection of never-covered elements of V). Recall also that τ_A is the cover time of $A \subseteq V$.

Theorem 2. (Aldous.) *Suppose $\mathbb{E}T = \omega(1)$. Then T is sharply concentrated around its expectation if and only if*

$$\frac{\mathbb{E}_B(\mathbb{E}\tau_B)}{\mathbb{E}T} = o(1),$$

where $\mathbb{E}_B$ is expectation over B.

One way to think of the quantity $\mathbb{E}_B(\mathbb{E}\tau_B)$ is that we let the coupon collector process run until V is covered at time T, then throw away the last set C_T that was collected, and continue the process. What is the expected time before V is covered again?

The power of Aldous's theorem is its generality and the necessity and sufficiency of its hypothesis for the concentration of the covering time. However as Aldous observed "[w]ithout any structure being imposed [...] it is not clear how to estimate [$\mathbb{E}_B(\mathbb{E}\tau_B)$] in order to use these results". Indeed, computing

$$\mathbb{E}_B(\mathbb{E}\tau_B) = \sum_{A \subseteq V} \mathbb{P}(A = B)\mathbb{E}\tau_A$$

requires us to estimate both the probability that a given set A is the "holdout" and to compute its expected covering time, both of which may be non-trivial tasks.

Several surveys have been written on coupon collectors, random allocation, urn occupancy problems, etc. Amongst others, let us mention the book of Johnson and Kotz [33] and Kolchin, Sevast'yanov and Chistyakov [38], the surveys of Ivanov, Ivchenko and Medvedev [29] and Kobza, Jacobson and Vaughan [37], and the papers of Holst [27], Stadje [57], Flajolet, Gardy and Thimonier [22] and McKay and Skerman [41].

2. Preliminaries: thresholds and elementary bounds

2.1. Coarse threshold

It follows from a simple application of the Bollobás–Thomason threshold theorem [7] that a covering process as we have defined it will always have a coarse threshold:

Proposition 1. (Coarse threshold.) *Let X be a covering random variable for a set V. Then*

$$\mathbb{P}(C_t = V) = \begin{cases} o(1) & \textit{if } t \ll T_{\frac{1}{2}} \\ 1 - o(1) & \textit{if } t \gg T_{\frac{1}{2}}. \end{cases}$$

Thus the covering time T is w.h.p. of the same order as $T_{\frac{1}{2}}$. In the present work, however, we are interested in a much sharper form of concentration than the one guaranteed by Proposition 1: we want the covering time T to be *sharply concentrated*, i.e. we want that $T/T_{\frac{1}{2}} \to 1$ in probability. As we shall see in the next subsection, we cannot in general guarantee this kind of sharp concentration. A question of crucial interest is then what conditions are necessary or sufficient to have sharp concentration for T — and how the value of $\mathbb{E}T$ may be computed in such cases.

2.2. Elementary bounds

Let X be a covering random variable for an n-set set V. For each $v \in V$, let $q_v = \mathbb{P}(v \in X)$, and set $q_\star = \min_v q_v$. We have the following elementary bounds on the location of $T_{\frac{1}{2}}$ and probable location of T.

Proposition 2.

$$\frac{\log(2)}{-\log(1-q_\star)} \leq T_{\frac{1}{2}} \leq \frac{\log(2n)}{-\log(1-q_\star)}.$$

What is more, for any fixed $\varepsilon > 0$

$$\mathbb{P}\left(T \leq (1+\varepsilon)\frac{\log n}{-\log\left(1-q_\star\right)}\right) \geq 1 - n^{-\varepsilon}.$$

Proof. For $t \geq T_{\frac{1}{2}}$ we have

$$\frac{1}{2} \leq \mathbb{P}(C_t = V) \leq \inf_{v \in V} \mathbb{P}(v \in C_t) = 1 - (1 - q_\star)^t ,$$

from which the claimed lower bound on $T_{\frac{1}{2}}$ follows. For the upper bound, $t \leq T_{\frac{1}{2}}$ implies

$$\frac{1}{2} \leq \mathbb{P}(C_t \neq V) \leq \sum_{v \in V} \mathbb{P}(v \notin C_t) = \sum_{v \in V} (1 - q_v)^t \leq n(1 - q_\star)^t.$$

Finally, for the 'what is more' statement, note that for $t \geq (1 + \varepsilon) \cdot \frac{\log n}{-\log(1 - q_\star)}$, the expected number of vertices not yet collected is

$$\mathbb{E}|V \setminus C_t| = n(1 - q_\star)^t \leq n^{-\varepsilon},$$

whence by Markov's inequality with probability at least $1 - n^{-\varepsilon}$ we have $C_t = V$ and $T \leq t$. □

Note that if X is balanced then $q_v = \frac{\mu}{n}$ for all $v \in V$, where $\mu := \mathbb{E}|X|$. In particular if $\mu = o(n)$ then the bounds above can be rewritten as

$$\frac{n \log 2}{(1 + o(1))\mu} = \frac{\log 2}{-\log\left(1 - \frac{\mu}{n}\right)} \leq T_{\frac{1}{2}} \leq \frac{\log(2n)}{-\log\left(1 - \frac{\mu}{n}\right)} = \frac{n \log n}{(1 + o(1))\mu}.$$

Perhaps surprisingly, these elementary bounds are essentially sharp. As we shall show in the next section, the covering time T for the exchangeable k-uniform coupon collector (Example 1) is sharply concentrated around the value $\frac{\log n}{-\log\left(1 - \frac{k}{n}\right)}$; in particular if $k = o(n)$, $T_{\frac{1}{2}} = (1 + o(1)) \frac{n \log n}{k}$. We think of this as 'slow coverage'. On the other hand, there are instances of 'fast coverage', discussed in greater detail in Section 4. We give here a simple example.

Example 5. (*Coupon collector with lottery.*) Set $V = [n]$, and $p = c/n$ for some $c = c(n) \in [0, n]$. Let X be with probability $1 - p$ a singleton from V chosen uniformly at random, and with probability p the entire set V.

Note that X is an exchangeable random covering variable, with expected size

$$\mathbb{E}|X| = (1 - p) + pn = 1 + c - \frac{c}{n}.$$

Proposition 3. *Let X and V be as in Example 5. Assume $c = o(n)$ and c is bounded away from 0. Then:*

1. $T_{\frac{1}{2}} = (1+o(1))\frac{n\log 2}{c}$*;*

2. $\mathbb{E}T = (1+o(1))\frac{n}{c}$*;*

3. $\lim_{n\to\infty}\mathbb{P}\left(T > \frac{xn}{c}\right) = e^{-x}$ *for any fixed* $x \geq 0$*.*

Proof. By Theorem 1, w.h.p. the 1-uniform exchangeable coupon collector does not cover $[n]$ in time less than $\frac{1}{2}n\log n$. Thus for time $t < \frac{1}{2}n\log n$, w.h.p. $T \leq t$ if and only if we have 'won the lottery' by time T, that is, if $X_i = [n]$ for some $i \leq t$. This event occurs with probability $1-(1-p)^t$.

To obtain part *1* of the proposition, we observe that if $t \geq T_{\frac{1}{2}}$ then

$$\frac{1}{2} + o(1) \leq 1-(1-p)^t,$$

yielding $t \geq (1+o(1))\frac{\log 2}{\log(1-p)} = (1+o(1))\frac{n\log 2}{c}$, and we show similarly that if $t \leq T_{\frac{1}{2}}$ then $t \leq (1+o(1))\frac{n\log 2}{c}$ to conclude.

For part *2*, let T' be the time at which we first 'win the lottery' by receiving all of V as our coupon. We have

$$\mathbb{E}T' = \sum_t tp(1-p)^{(t-1)} = \frac{1}{p} = \frac{n}{c}.$$

Since $T \leq T'$, we have that $\mathbb{E}T \leq \mathbb{E}T'$. Now from our estimates for the probability of winning the lottery by time t above, w.h.p. we have $T' = o(n\log n)$. Thus by Theorem 1, w.h.p. $T = T'$, and

$$\begin{aligned}\mathbb{E}T &\geq \sum_t t\mathbb{P}(T'=t|T'=T)\mathbb{P}(T'=T)\\ &\geq \sum_t t\left(\mathbb{P}(T'=t) - o(1)\right)\\ &= (1+o(1))\mathbb{E}(T'),\end{aligned}$$

whence we are done.

Finally for part *3*, we simply note that w.h.p. $T = T'$, and that $\mathbb{P}(T' > \frac{xn}{c}) = (1-p)^{\frac{xn}{c}} = e^{-x(1+O(n^{-1}))} \to e^{-x}$. □

Proposition 3 shows two things. First of all, the lower bound on $T_{\frac{1}{2}}$ in Proposition 2 is essentially sharp; indeed taking $c = c(n)$ tending to infinity slowly, we have in Example 5 that $\mu = \mathbb{E}|X| = c(1+o(1))$, and $T_{\frac{1}{2}} = (1+o(1))\frac{n\log 2}{\mu}$. Further, by varying the value of $c = c(n)$ from $\Omega(\frac{1}{\log n})$ to $o(n)$, we can get $T_{\frac{1}{2}}$ to take asymptotically any value between the bounds from Proposition 2.

Secondly, we cannot in general expect T to be sharply concentrated: part *3* of Proposition 3 shows that we do not get sharper concentration than the Bollobás–Thomason-type concentration guaranteed by Proposition 1. With this in mind, we next focus on conditions on X which guarantees sharp concentration of the covering time T and/or 'slow coverage'.

3. General concentration results

Let $V = [n]$ and X be a random coupon variable for V. In this section we prove general results establishing (simple, easily checkable) sufficient conditions for sharp concentration of the covering time $T(\mathbf{X})$. We also include some results in the special case where the random coupon variable X is balanced, transitive or exchangeable. Before stating our results, we need to introduce some notation.

Our proof strategy involves approximating the discrete-time process of collecting coupons by a continuous-time process. Instead of the coupon collector drawing a random coupon X at integer time points, she draws a random coupon X (from the same distribution) at times given by a Poisson process with parameter 1.

The times at which any given coupon is drawn will then be a thinned Poisson process, and the Poisson processes associated with different coupons will be independent. The times at which any particular element $x \in V$ is drawn will also be a thinned Poisson process, though the Poisson processes associated with different elements $x, y \in V$ will not in general be independent. Working in the continuous rather than in the discrete setting will greatly simplify calculations.

For $S \subseteq [n]$, set $h(S) = \mathbb{P}(X = S)$. For every S with $h(S) > 0$, start a Poisson process $\mathcal{P}_S$ with intensity $h(S)$. Each time an event occurs in $\mathcal{P}_S$, the coupon collector draws the coupon S. Let $q_x := \sum_{S \ni x} h(S)$ be the total intensity of all coupons covering x, and let $q_{xy} := \sum_{S \ni x,y} h(S)$ be the total intensity of all coupons covering x and y simultaneously. Equivalently, $q_x = \mathbb{P}(x \in X)$ and $q_{xy} = \mathbb{P}(x, y \in X)$. Note that $\sum_{x \in V} q_x = \sum_S |S| \cdot h(S) = \mathbb{E}X =: \mu$.

Let $Z_{x,t}$ be the indicator event of the element x not being covered at time t, and let $Z_t := \sum_{x\in V} Z_{x,t}$. An element x has not been covered by time t if its associated Poisson process with intensity q_x has had no events in the time interval $[0,t]$. The probability of this occurring is $e^{-q_x t}$, and so the first two moments of Z_t are:

$$\mathbb{E}Z_t = \sum_{x\in V} e^{-q_x t} \quad \text{and} \quad \mathbb{E}Z_t^2 = \sum_{x,y\in V} e^{-(q_x+q_y-q_{xy})t}.$$

Many of our proofs will use the second moment method to show concentration of Z_t, which implies concentration of $T(\mathbf{X})$.

3.1. Results

Definition 1. A coupon collector has the *first moment property* if there exists $T^-(n)$ and $T^+(n)$ such that $T^- = (1+o(1))T^+$, $\mathbb{E}Z_{T^-} \to \infty$ and $\mathbb{E}Z_{T^+} \to 0$.

This holds trivially for balanced coupon collectors, but may fail if X is far from balanced — for instance, if some elements of V occur very rarely. Our first result gives us sufficient conditions for the first moment property to hold.

For any $\alpha \in \mathbb{R}$, let $\|\mathbf{q}\|_\alpha$ be the α-Hölder mean of the vector of intensities $\mathbf{q} = (q_x)$, i.e. $\|\mathbf{q}\|_\alpha := \left(\frac{1}{n}\sum_x q_x^\alpha\right)^{\frac{1}{\alpha}}$ (with the usual convention that for $\alpha = 0$, $\|\mathbf{q}\|_0$ is the geometric mean $(\prod_x q_x)^{1/n}$).

Theorem 3. *Set $q_\star := \min_x q_x$. If any of the following conditions is satisfied, then first moment property holds.*

1. *There exists $t \gg q_\star^{-1}$ such that $\mathbb{E}Z_t \gg 1$*
2. *There exists $\alpha = o(\log n)$ such that $(\alpha+2)\|\mathbf{q}\|_{-\alpha} \le \log n \cdot q_\star$*
3. *There exists $A_r = \exp(\omega(r))$, which does not depend on n, such that for any $r > 0$ and all sufficiently large n, the number of $y \in V$ satisfying $q_y < rq_\star$ is at least A_r.*

If $\frac{q_x}{q_y} \le \frac{1}{2}\log n$ for all x, y, condition *2* is met trivially with $\alpha = 0$; in fact it can be shown that the factor $\frac{1}{2}$ can be replaced by any positive number. So in particular Theorem 3 applies to 'almost balanced' random coupon variables X.

Our second result gives w.h.p. bounds on T when correlations are bounded.

Theorem 4. *If the first moment property holds and there exists $C = C(n)$ such that*

1. *the coupons have C-bounded correlation, i.e. $q_{xy} \le C q_x q_y$ for all $x \ne y$, and*

2. *$C\bar{q} = o(\frac{1}{\log n})$, where $\bar{q} = \mathbb{E}\big[q_\theta \big| \theta$ not covered at time $T^-\big]$ is the expected size of q_θ for θ drawn uniformly at random from V,*

then $T^- \le T(\mathbf{X}) \le T^+$ w.h.p.

The parameter $\bar{q}$ should be viewed as the 'speed' of covering at time T^-, and it is usually hard to compute exactly. However in order to apply Theorem 4 it is enough to give an upper bound on $\bar{q}$. The simplest such bound, namely $\bar{q} \le \max_x q_x$, can easily be improved. For instance, it is straight-forward to show that $\bar{q} \le \|\mathbf{q}\|_{-\alpha}$ for any finite α and all n sufficiently large. This makes condition *2* easy to check in many situations.

We can obtain further results when X is assumed to be balanced. The next theorem tells us that if either the coupons are 'small' (size $o(n)$) or the pairwise correlations between the elements of V are 'not too strong' then we have sharp concentration for T.

Theorem 5. *Let X be a balanced random coupon variable with $\mu := \mathbb{E}|X|$.*

1. *If there exists t such that $\sum_{x,y}(e^{q_{xy}t} - 1) = o(n^2)$ and $\sum_x e^{-q_x t} = \omega(1)$, then $T(\mathbf{X}) \ge t$ w.h.p.*

2. *If there exists $1 \ll \beta(n) < n$ and $q = o\left(\frac{\mu}{n \log \beta}\right)$ with $q_{xy} \le q$ for all but at most $\frac{1}{\beta}n^2$ 'bad' pairs (x, y), then $T(\mathbf{X}) \ge \frac{n}{\mu}(\log \beta - \omega(1))$ w.h.p.*

   In particular, if $q = o\left(\frac{\mu}{n \log n}\right)$ and there are at most $n^{1+o(1)}$ such 'bad' pairs, then
   $T(\mathbf{X}) = (1 \pm o(1))\frac{n \log n}{\mu}$ w.h.p.

3. *If all coupons have size at most M, then $T(\mathbf{X}) \ge \frac{n}{\mu}(\log n - \log M - \omega(1))$ w.h.p., for $\omega(1)$ tending to infinity arbitrarily slowly.*

   In particular, if $M = n^{o(1)}$, then $T(\mathbf{X}) = (1 + o(1))\frac{n \log n}{\mu}$ w.h.p.

4. *If $q_{xy} = q_{x'y'}$ for all $x \ne y, x' \ne y'$, and all coupons have size at most M, and T^- is such that $T^- = \frac{n}{\mu} \cdot \min\big(o(\frac{n}{M}), \log n - \omega(1)\big)$, then $T^- \le T(\mathbf{X})$ w.h.p.*

   In particular, if $M = o(\frac{n}{\log n})$, then $T(\mathbf{X}) = (1 + o(1))\frac{n \log n}{\mu}$ w.h.p.

There are examples where the w.h.p. lower bounds given by this theorem are sharp, while the upper bounds given by the first moment method are not; see for instance Example 6 in Section 4. Note also that unlike Theorem 2, Theorem 5 also locates the threshold.

For balanced random covering variables we also have good control for both concentration and the covering time when X satisfies an 'almost negative correlation' condition. Here below we say that a function $m = m(n)$ is *sub-polynomial* in n if $m = n^{o(1)}$.

Theorem 6. *Let $\delta > 0$ be fixed. Let X be a balanced covering random variable for an n-set V, with $\mathbb{P}(x \in X) = c$ for some $c \in (0, 1-\delta)$. Suppose further that we have* almost negative correlations*, namely that there exist $\eta = o(1/\log n)$ and $b = b(n)$ sub-polynomial in n such that for any $x \in V$*

$$\mathbb{P}(x, y \notin X) \leq (1-c)^2(1+\eta).$$

holds for all but at most b elements y. Then, w.h.p., $T(\mathbf{X}) = (1+o(1)) \cdot \frac{\log n}{-\log(1-c)}$.

Note that if $\eta = 0$ then the correlation condition is the same as the commonly used pairwise negative correlation condition. Recently a substantial theory for negatively correlated random variables has been developed and numerous common examples have been shown to have this and even stronger correlation properties, see [8].

Corollary 3.1. *Suppose that X is balanced, has pairwise negative correlation, and $\mathbb{P}(x \in X) = c \leq 1 - \delta$ for a fixed $\delta > 0$. Then w.h.p. $T(\mathbf{X}) = (1+o(1))\frac{\log n}{-\log(1-c)}$.*

If $c = o(1)$ then the equality above may be rewritten as $T(\mathbf{X}) = (1+o(1))\frac{n \log n}{\mathbb{E}|X|}$.

We next give conditions implying sharp concentration for the covering time of an exchangeable random variable X around the same value as a *uniform* exchangeable random variable with the same mean coupon size.

Theorem 7. *Let X be an exchangeable random coupon variable, for $V = [n]$, with maximum coupon size M, average coupon size μ and mean square coupon size χ. If any of the four conditions below holds, then w.h.p. $T(\mathbf{X}) = (1+o(1))\frac{\log n}{-\log(1-\frac{\mu}{n})}$ (which in the case $\mu = o(n)$ can be rewritten as $T = (1+o(1))\frac{n \log n}{\mu}$).*

1. *$M = o(\sqrt{n \log n})$;*

2. $M = o(n)$ *and* $M = o(\sqrt{\mu n \log n})$*;*

3. $M = o(n)$ *and* $\chi = o(\mu n \log n)$*;*

4. $\mu < (1-\delta)n$ *for some* $\delta > 0$ *and* $\chi = (1 + o(\frac{1}{\mu n \log n}))\mu^2$*.*

Note that the theorem includes the case when X is k-uniform for $k = cn$. Roughly speaking, the conditions in the theorem move from small coupons, with no other assumptions, to larger coupons where successively stronger size concentration is needed.

In some applications it is useful to have more accurate information about the sharpness of the concentration. We thus include a final result on the cover time T for the k-uniform exchangeable coupon collector in the sublinear case $k = o(n)$, in the spirit of the theorem of Erdős and Rényi (Theorem 1) mentioned in the introduction.

Theorem 8. *If* $k = o(n)$*, then the covering time* T *for a* k*-uniform exchangeable coupon collector is sharply concentrated around* $\frac{n \log n}{k}$*. More precisely, we have* $\mathbb{P}\Big(|T - \frac{n\log n}{k}| > \frac{cn}{k}\Big) \to e^{-c}$ *as* $n \to \infty$*.*

3.2. Continuous-time approximation of the coupon collector

In this subsection, we formalize our approximation of the discrete-time coupon collector by a continuous-time process. As described above, for every subset $S \subseteq V$ with $h(S) = \mathbb{P}(X = S) > 0$, we start at time $t = 0$ a Poisson process $\mathcal{P}_S$ with intensity $h(S)$. Our continuous coupon collector receives S as a coupon each time an event occurs in $\mathcal{P}_S$. List the coupons in the order they are received by the continuous collector as $S_1, S_2, S_3, \ldots$ The distribution of the sequence $\mathbf{S} = (S_n)_{n \in \mathbb{N}}$ is identical to that of the sequence of coupons $\mathbf{X}$ received by the (discrete-time) X-coupon collector. Furthermore, the time t_m at which the continuous coupon collector receives his m^{th} coupon is sharply concentrated around m for $m = \omega(1)$. Indeed, by a standard bound on the Poisson distribution, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|t_m - m| \geq \varepsilon m\right) = O\left(\frac{1}{\sqrt{m\varepsilon^2}} e^{-\frac{m\varepsilon^2}{2}}\right).$$

In particular, provided the covering time for the continuous coupon collector t_T is large (grows with n), we have that w.h.p. $t_T = (1 + o(1))T$. Thus it is enough to prove w.h.p. bounds on t_T to establish w.h.p. bounds on T. We shall thus in a slight abuse of notation identify t_T with T in the rest of the paper, and prove bounds for the covering time via the continuous coupon collector. In particular we shall set $T = \inf\{t : \ Z_t = 0\}$.

3.3. Proofs: concentration of the covering time

It will be useful to consider the function $f(t) = \log(\mathbb{E}Z_t)$. The first two derivatives of f are

$$f'(t) = -\frac{\sum_x q_x e^{-q_x t}}{\sum_x e^{-q_x t}} \leq 0, \quad f''(t) = \frac{1}{2} \cdot \frac{\sum_{x,y} (q_x - q_y)^2 e^{-(q_x+q_y)t}}{\sum_{x,y} e^{-(q_x+q_y)t}} \geq 0,$$

from which we can see that f is a decreasing convex function. In particular for any $t \geq 0$,

$$f(t) - tf'(t) \leq f(0) = \log n. \tag{1}$$

Similarly, for any $t > 0$ and $\Delta < t$,

$$f(t-\Delta) - f(t) \geq -\Delta f'(t) \qquad \text{and} \qquad f(t) - f(t+\Delta) \geq -\Delta f'(t). \tag{2}$$

Finally, note that $f'(t) = -\mathbb{E}[q_\theta | \theta \text{ not covered at time } t]$, where θ is chosen uniformly at random from V. The following lemma gives the basic first and second moment bounds on the covering time.

Lemma 1. *Let $T = T(\mathbf{X})$ be the covering time for a coupon collector $\mathbf{X}$, and let (q_x) and (q_{xy}) be its associated single and pairwise intensities.*

1. *If $t = t(n)$ is such that $\sum_x e^{-q_x t} \to 0$, as $n \to \infty$, then $T \leq t$ w.h.p.*
2. *If $t = t(n)$ is such that $\frac{\sum_{x \neq y} (e^{q_{xy} t} - 1) \cdot e^{-(q_x+q_y)t}}{\sum_{x,y} e^{-(q_x+q_y)t}} = o(1)$ and $\sum_x e^{-q_x t} \to \infty$, then $T \geq t$ w.h.p.*

Proof. We divide the proof into two parts.

Part 1. Suppose $t = t(n)$ satisfies the lemma's assumption. By Markov's inequality $\mathbb{P}(t \geq T) = \mathbb{P}(Z_t > 0) \leq \mathbb{E}Z_t = \sum_x e^{-q_x t} \to 0$, so $t < T$ w.h.p.

Part 2. Suppose $t = t(n)$ satisfies our assumption. Then $\mathbb{E}Z_t \to \infty$, and

$$\begin{aligned}\frac{\mathrm{Var}[Z_t]}{\mathbb{E}[Z_t]^2} &= \frac{\sum_{x,y} (e^{q_{xy} t} - 1) \cdot e^{-(q_x+q_y)t}}{\sum_{x,y} e^{-(q_x+q_y)t}} \\ &< \frac{\sum_{x \neq y} (e^{q_{xy} t} - 1) \cdot e^{-(q_x+q_y)t}}{\sum_{x,y} e^{-(q_x+q_y)t}} + \frac{1}{\sum_x e^{-q_x t}} = o(1) + o(1),\end{aligned}$$

so by Chebyshev's inequality $Z_t = (1+o(1))\mathbb{E}Z_t \to \infty$ w.h.p., so that w.h.p. $Z_t > 0$ and $T > t$. □

Proof of Theorem 4. By the first moment property, $\mathbb{E}Z_{T^+} \to 0$, from which it is immediate by Lemma 1 part 1 that w.h.p. $T \le T^+$. To establish the lower bound on T, we shall consider the set of 'rare' coupons $U = \{x \in V : q_x \le 2\bar{q}\}$. Let Y_t be the number of $x \in U$ for which x is uncovered at time t.

Claim 1. *$\mathbb{E}Y_t \to \infty$ for any $t \le T^-$*

Proof. We bound $\bar{q}$ from below to get:

$$\bar{q} \ge \frac{\sum_{x\in V\setminus U} q_x e^{-q_x T^-}}{\sum_{x\in V} e^{-q_x T^-}} \ge 2\bar{q}\frac{\sum_{x\in V\setminus U} e^{-q_x T^-}}{\sum_{x\in V} e^{-q_x T^-}} = 2\bar{q}\cdot\left(1 - \frac{\mathbb{E}Y_{T^-}}{\mathbb{E}Z_{T^-}}\right).$$

Dividing both sides by $\bar{q}$ gives us $1 \ge 2\big(1 - \frac{\mathbb{E}Y_{T^-}}{\mathbb{E}Z_{T^-}}\big)$, which implies $\mathbb{E}Y_{T^-} \ge \frac{1}{2}\mathbb{E}Z_{T^-}$. Since by assumption *1* $\mathbb{E}Z_{T^-} \to \infty$, and since $\mathbb{E}Y_t$ is decreasing in t, we must have that $\mathbb{E}Y_t \to \infty$ for any $t \le T^-$, as claimed. □

Now, as observed after inequality (2),

$$f'(t) = -\mathbb{E}[q_\theta | \theta \text{ not covered at time } t],$$

and in particular $f'(T^-) = -\bar{q}$. By assumption *1* $f(T^-) \to \infty$, so inequality (1) gives

$$T^- \cdot \bar{q} \le f(T^-) - T^- f'(T^-) \le \log n \tag{3}$$

We are now in a position to apply part 2 of Lemma 1 to the *restriction* of the coupon collector to the set of rare coupons U (i.e. the coupon collector with covering variable $X \cap U$). For any $x \ne y$, we have that $q_{xy}t \le Cq_xq_yt$ by assumption *1.* If $x, y \in U$ this quantity is at most $4C(\bar{q})^2T^-$. By inequality (3) and our assumption *2*, $4C(\bar{q})^2T^- \le 4C\bar{q}\log n = o(1)$. Thus

$$\begin{aligned}&\frac{\sum_{x,y\in U: x\ne y}(e^{q_{xy}t} - 1)\cdot e^{-(q_x+q_y)t}}{\sum_{x,y\in U} e^{-(q_x+q_y)t}} \\ &\le (e^{4C\bar{q}\log n} - 1)\cdot\frac{\sum_{x,y\in U: x\ne y} e^{-(q_x+q_y)t}}{\sum_{x,y\in U} e^{-(q_x+q_y)t}} = o(1).\end{aligned}$$

Since by Claim 1 $\mathbb{E}Y_t \to \infty$, we have by Lemma 1 part 2 that $T^- \le \inf\{t : Y_t = 0\}$ w.h.p. Since by construction $Y_t \le Z_t$, this gives $T^- \le T$ w.h.p., as required. □

3.4. Proofs: sharp transition for $\mathbb{E}Z_t$

Proof of Theorem 3. Let $T^* = T^*(n)$ be the unique real for which $\mathbb{E}Z_{T^*} = 1$.

1. By our assumption, we can find $\Delta = \Delta(n)$ such that $T^* \gg \Delta \gg \frac{1}{\min_x q_x}$. We will show that $T^- := T^* - \Delta$ and $T^+ := T^* + \Delta$ have the desired properties. By definition of Δ, we have $T^- = (1+o(1))T^+$. Now $-\Delta f'(T^*) \geq \Delta \min_x q_x \gg 1$, so by inequality (2) we have $f(T^*-\Delta) - f(T^*)$ and $f(T^*) - f(T^*+\Delta)$ both tending to infinity. Since $f(T^*) = 0$, this implies that $\mathbb{E}Z_{T^*-\Delta} \to \infty$ and $\mathbb{E}Z_{T^*+\Delta} \to 0$, as required.

2. Let $\alpha(n) = o(\log n)$ be as in the assumption. Pick $1 \ll c \ll \log n/(\alpha+2)$, and set $t^* = (\log n - c)/\|\mathbf{q}\|_{-\alpha}$. By assumption, $t^* \gg \min_x q_x$.

   For any $x \in V$, we have
$$q_x t^* \geq \frac{(\alpha+2)\|\mathbf{q}\|_{-\alpha} t^*}{\log n} = \alpha + 2 - \frac{(\alpha+2)c}{\log n} = \alpha + 2 - o(1).$$

   Now the function $z \mapsto e^{-z^{-1/\alpha}}$ is convex over those z satisfying $z^{-1/\alpha} \geq \alpha+1$. We can therefore apply Jensen's inequality as follows:
$$\mathbb{E}Z_{t^*} = \sum_{x\in V} e^{-q_x t^*} = \sum_{x \in V} e^{-(q_x t_*)^{-\alpha\cdot(-1/\alpha)}} \geq n \exp(-\|\mathbf{q}\|_{-\alpha} t^*) = e^c.$$

   But $e^c \to \infty$, so this gives us a $t^* \gg \min_x q_x$ such that $\mathbb{E}Z_{t^*} \gg 1$. We are then done by part *1*.

3. Let the function A_r be as in the assumption. Let $R = R(n)$ be the largest r such that there are at least A_r elements y with $q_y \leq r \min_x q_x$; R is finite for every n, but by assumption tends to infinity as $n \to \infty$. We can therefore find $t^* = t^*(n)$ satisfying
$$\frac{1}{\min_x q_x} \ll t^* \ll \frac{\log A_R}{R \min_x q_x}.$$

   We now bound $\mathbb{E}Z_{t^*}$ from below:
$$\mathbb{E}Z_{t^*} \geq \sum_{\substack{y \in V:\\ q_y \leq R\min_x q_x}} e^{-q_y t^*} \geq A_R e^{-R \min_x q_x t^*} \gg 1,$$

   by the choice of t^*. We are then done by part *1* .

□

3.5. Proofs: balanced coupons

Proof of Theorem 5. Since X is balanced, $q_x = \mu/n$ for all $x \in V$. We will show that we can apply part *2* of Lemma 1 provided *1* holds, and then that each of conditions *2*–*4* implies *1*. The 'in particular' statements in *2*–*4* combine the lower bound given by those special cases with the upper bound on T from Proposition 2.

1. Since $q_x = q_y$ for all x, y, we have
$$\frac{\sum_{x,y}(e^{q_{xy}t}-1)\cdot e^{-(q_x+q_y)t}}{\sum_{x,y} e^{-(q_x+q_y)t}} = \frac{\sum_{x,y}(e^{q_{xy}t}-1)}{n^2} = \frac{o(n^2)}{n^2} = o(1).$$
We can therefore apply part *2.* of Lemma 1 to conclude that $T(\mathbf{X}) \geq T^-$.

2. Set $t = n(\log\beta - \omega(1))/\mu$ for some $\omega(1)$ tending to infinity arbitrarily slowly. Note
$$Z_t = ne^{-\log\beta+\omega(1)} \geq e^{\omega(1)} \to +\infty.$$
Let E be the set of exceptional pairs (x, y) with $q_{xy} > q$. Since $q_{xy} \leq q_x = \frac{\mu}{n}$ for any x, y, we have:
$$\sum_{(x,y)\notin E} e^{q_{xy}t} \leq n^2 e^{qt} \leq n^2 + o(n^2), \qquad \text{and}$$
$$\sum_{(x,y)\in E} e^{q_{xy}t} \leq \frac{n^2}{\beta}\cdot e^{\frac{\mu t}{n}} = \frac{n^2}{\beta}\cdot e^{\log\beta-\omega(1)} = o(n^2),$$
Together, these bound give that $\sum_{x,y} e^{q_{xy}t} = n^2 + o(n^2)$.

3. Fix $x \in V$, and consider the sum $\sum_{y\in V} q_{xy}$. Each subset $X \subseteq V$ containing x contributes $h(X)$ to $|X|$ terms of the sum. Thus
$$\sum_{y\in V} q_{xy} = \sum_{X\ni x} |X|h(X) \leq M \sum_{X\ni x} h(X) = Mq_x.$$
Furthermore, for every y, $q_{xy} \leq q_x = \frac{\mu}{n}$. We ask therefore: which choices of $\tilde{q}_{xy}$, subject to the constraints $\sum_{y\in V} \tilde{q}_{xy} \leq M\frac{\mu}{n}$ and $0 \leq \tilde{q}_{xy} \leq \frac{\mu}{n}$, maximize the expression $\sum_{y\in V} e^{\tilde{q}_{xy}t} - 1$? Since $z \mapsto e^{zt} - 1$ is an increasing function for $t > 0$, the optimal $\tilde{q}_{xy}$ must satisfy $\sum_{y\in V} \tilde{q}_{xy} = M\frac{\mu}{n}$. By the Karamata inequality the maximum of the sum is then attained when M of the $q_{\tilde{x}}y$ are equal to $\frac{\mu}{n}$ and the rest are equal to 0. Thus
$$\sum_{y\in V}(e^{q_{xy}t}-1) \leq \sum_{y\in V}(e^{\tilde{q}_{xy}t}-1) < M\cdot e^{\frac{\mu t}{n}}.$$

Setting $t = \frac{n(\log n - \log M - \omega(1))}{\mu}$ for an arbitrary $\omega(1)$ tending to infinity, and summing over all x, we get

$$\sum_{x,y\in V}(e^{q_{xy}t}-1) < Mn\cdot e^{\frac{\mu t}{n}} = Mn\cdot e^{\log n-\log M-\omega(1)} = o(n^2).$$

In addition our choice of t ensures $\mathbb{E}Z_t = Me^{\omega(1)} \to +\infty$.

4. If $q_{xy} = q$ for all $x \neq y$ and some q, then

$$M\mu \geq \sum_{x,y} q_{xy} = \mu + \sum_{x\neq y} q_{xy} = \mu + n(n-1)q,$$

so $q \leq \frac{(M-1)\mu}{(n-1)n}$. For $t \leq \frac{n(\log n - \omega(1))}{\mu}$ with $t = o\big(\frac{n^2}{M\mu}\big)$, we have that $qt = o(1)$ and $q_x t \leq \log n - \omega(1)$, whence $Z_t \to \infty$ and

$$\begin{aligned}\sum_{(x,y):\ x\neq y}(e^{q_{xy}t}-1) + \sum_x (e^{q_x t}-1)\\ < n^2(e^{o(1)}-1) + ne^{\log n - \omega(1)} = o(n^2).\end{aligned}$$

In all cases, condition *1* is satisfied. □

Proof of Theorem 6. Since X is balanced, we have that $t_0 = \frac{\log n}{-\log(1-c)}$ is a first-moment threshold for the expected number of uncovered vertices $\mathbb{E}Z_t = n(1-c)^t$. In particular we have that for any fixed $\varepsilon > 0$ the covering time $T = T(\mathbf{X})$ satisfies $T < (1+\varepsilon)\frac{\log n}{-\log(1-c)}$ w.h.p. We turn our attention to the variance of Z_t to show concentration of its value just below the first-moment threshold t_0.

$$\begin{aligned}\mathbb{E}[Z_t^2] &= \sum_{x,y}\mathbb{P}(x,y\notin C_t) = \sum_{x,y}(1-\mathbb{P}(x,y\in X))^t\\ &\leq n\big((1-c)^{2t}(1+\eta)^t(n-b) + (1-c)^t b\big)\\ &< n^2(1-c)^{2t}\left((1+\eta)^t + b\left(\frac{1}{n(1-c)^t}\right)\right).\end{aligned}$$

Now for $\varepsilon > 0$ fixed and $t \leq (1-\varepsilon)\frac{\log n}{-\log(1-c)}$, our assumptions on b and η tell us that the above is at most

$$(\mathbb{E}Z_t)^2\left(e^{\frac{\eta\log n}{-\log(1-c)}} + \frac{b}{n^\varepsilon}\right) = (\mathbb{E}Z_t)^2(1+o(1)).$$

Chebyshev's inequality is then enough to give us concentration of Z_t about its (large, non-zero) mean for these values of t. In particular w.h.p. $T > (1-\varepsilon)\frac{\log n}{-\log(1-c)}$. Thus w.h.p. $T = (1+o(1))\frac{\log n}{-\log(1-c)}$, as required. □

3.6. Proofs: exchangeable coupons

In the case where X is an exchangeable random variable, we exhibit a (natural) coupling between the process of covering V by X with the classical coupon collector problem (covering by singletons chosen uniformly at random), which allows us to determine (up to a small error) the expectation of the covering time T as well as, in the case where $|X| = o(n)$ holds w.h.p., to prove that T is concentrated around its mean. We note that a similar coupling appears in a work of Sellke [55], though it is used for a different purpose.

We begin by proving Theorem 8. Let $k = k(n)$ be a sequence of natural numbers. Set $V = V(n) = [n]$, and let $X = X(n)$ be the random covering variable for V obtained by selecting a k-set from V uniformly at random. Let also $Y = Y(n)$ be the classical random coupon variable for V, namely the random covering variable obtained by selecting a singleton from V uniformly at random.

Proof of theorem 8. We couple the k-uniform coupon sequence $\mathbf{X}$ to the sequence of coupons received by the Y-coupon collector, $\mathbf{Y} = (Y_i)_{i=1}^{\infty}$. For natural numbers $a \le b$, set $C_Y[a,b] := \bigcup_{i\in[a,b]} Y_i$. Let $a_0 = 0$, and define a_i, $i \ge 1$, recursively to be the least integer such that $|C_Y[a_{i-1}{+}1, a_i]| = k$. Next, let $X_i = C_Y[a_{i-1}{+}1, a_i]$. Clearly, the X_i obtained are independent random sets, uniformly distributed among the k-sets in V, so $(X_i)_{i=1}^{\infty} \sim \mathbf{X}$. Furthermore, the integers $\ell_i := a_i - a_{i-1}$ are i.i.d. random variables.

This coupling between the coupon collectors enables us to relate $T(\mathbf{X})$ to $T(\mathbf{Y})$. For any natural number t, we have that

$$\bigcup_{j=1}^{t} X_j = \bigcup_{i=1}^{a_t} Y_i,$$

so $T(\mathbf{X}) \le t$ if and only if $T(\mathbf{Y}) \le a_t$. Conversely, $T(\mathbf{X}) > t$ if and only if $T(\mathbf{Y}) > a_t$. In other words,

$$\sum_{i=1}^{T(\mathbf{X})-1} \ell_i = a_{T(\mathbf{X})-1} < T(\mathbf{Y}) \le a_{T(\mathbf{X})} = \sum_{i=1}^{T(\mathbf{X})} \ell_i. \tag{4}$$

At this point, it is straightforward to get an estimate for $\mathbb{E}T(\mathbf{X})$ in terms of the (well–known) expectations of $T(\mathbf{Y})$ and ℓ_1, via an application of Wald's inequality. To obtain sharp concentration for $T(\mathbf{X})$ we need only do a little more work. Let $S_m := \sum_{i=1}^{m} \ell_i$. We shall use the following lemma, establishing sharp concentration for S_m, together with the Erdős–Rényi sharp concentration theorem for $T(\mathbf{Y})$ to deduce we have the desired sharp concentration for $T(\mathbf{X})$.

Lemma 2. *If $k = o(n)$, then for all $c > 0$ and $m > \frac{n}{k}$ the following inequality holds:*

$$\mathbb{P}\left(|S_m - \mathbb{E}S_m| > c \cdot k\sqrt{\frac{m}{n}}\right) < 4 \cdot e^{-c}.$$

Proof of Lemma 2. For $0 \le i \le k-1$, let τ_i be the time it takes for the singleton collector to draw the $(i+1)^{\text{th}}$ distinct coupon after she has collected i distinct coupons. Clearly, $\tau_i \sim \text{Geom}(\frac{n-i}{n})$. and has moment-generating function

$$M_{\tau_i}(\lambda) := \mathbb{E}[e^{\lambda\tau_i}] = \frac{(1-\frac{i}{n})e^{\lambda}}{1-\frac{i}{n}e^{\lambda}}.$$

Note that $\ell_1 = \sum_{i=0}^{k-1}\tau_i$. Since $S_m = \sum_{i=1}^{m}\ell_i$ is the sum of m independent copies of ℓ_1, its moment generating function is given by

$$M_{S_m}(\lambda) = \left(\prod_{i=0}^{k-1}\frac{(1-\frac{i}{n})e^{\lambda}}{1-\frac{i}{n}e^{\lambda}}\right)^m.$$

Applying Markov's inequality to the random variable $\exp(\lambda S_m)$, for some λ: $\lambda \neq 0,\ \lambda = o(1)$ to be specified later, gives

$$\begin{aligned}
&\mathbb{P}\left(e^{\lambda S_m} > e^{\lambda\mathbb{E}S_m + c}\right) < \frac{M_{S_m}(\lambda)}{\exp(\lambda\mathbb{E}S_m + c)} = \\
=&\frac{\exp\left(m\sum_{i=0}^{k-1}(\lambda + \log(1-\frac{i}{n}) - \log(1-\frac{i}{n}e^{\lambda}))\right)}{\exp\left(m\left(\sum_{i=0}^{k-1}\frac{\lambda}{1-\frac{i}{n}}\right)+c\right)} \\
=&\exp\left(m\sum_{i=0}^{k-1}\left(\lambda + \log\left(1-\frac{i}{n}\right) - \log\left(1-\frac{i}{n}e^{\lambda}\right) - \frac{\lambda}{1-\frac{i}{n}}\right)\right) \\
\le&\exp\left(mk\left[\lambda + \log\left(1-\frac{k}{n}\right) - \log\left(1-\frac{k}{n}e^{\lambda}\right) - \frac{\lambda}{1-\frac{k}{n}}\right] + c\right),
\end{aligned} \tag{5}$$

where the last inequality holds since the summands are non-decreasing in i (this can be checked e.g. by computing the derivative of a summand with respect to i). We use a Taylor expansion of degree $d = \lceil -\log|\lambda| \rceil$ to estimate the quantity inside the square brackets.

$$\begin{aligned}
&\lambda + \log\left(1-\frac{k}{n}\right) - \log\left(1-\frac{k}{n}e^{\lambda}\right) - \frac{\lambda}{1-\frac{k}{n}} \\
\le&\sum_{j=1}^{d}(e^{j\lambda} - j\lambda - 1)\frac{k^j}{jn^j} + \frac{k^{d+1}}{(d+1)(n-k)^{d+1}}
\end{aligned} \tag{6}$$

Note that $(e^{j\lambda} - j\lambda - 1) = (\frac{1}{2} + o(1)) \cdot (j\lambda)^2$, since $j\lambda = o(1)$, whereas $\frac{k^{d+1}}{(n-k)^{d+1}} \ll (e^{-2})^d \cdot \frac{k}{n} \leq \frac{\lambda^2 k}{n}$, since $\frac{k}{n-k} = o(1)$ by assumption. The right hand side of inequality (6) can thus be bounded by

$$\left(\frac{1}{2} + o(1)\right) \cdot \lambda^2 \sum_{j=1}^{d} \frac{jk^j}{n^j} + o\left(\frac{\lambda^2 k}{n}\right) = \frac{(1+o(1))\lambda^2 k}{2n}.$$

Applying this bound to the right-hand side of inequality (5) gives us the following:

$$\mathbb{P}\left(e^{\lambda S_m} > e^{\lambda \mathbb{E} S_m + c}\right) < \exp\left(\frac{(1+o(1))m\lambda^2 k^2}{2n} + c\right).$$

Letting $\lambda = \pm\frac{1}{k}\sqrt{\frac{n}{m}}$ we obtain

$$\mathbb{P}\left(S_m - \mathbb{E}S_m > \quad ck\sqrt{\frac{m}{n}}\right) < e^{\frac{1}{2}+o(1)-c}, \text{ and}$$
$$\mathbb{P}\left(S_m - \mathbb{E}S_m < -ck\sqrt{\frac{m}{n}}\right) < e^{\frac{1}{2}+o(1)-c}.$$

Thus for n sufficiently large, the probability that S_m diverges from its expectation by more than $ck\sqrt{\frac{m}{n}}$ is at most $2e^{(\frac{1}{2}+o(1))-c} < 4e^{-c}$. □

Equation (4) can also be formulated as

$$S_{T(\mathbf{X})-1} < T(\mathbf{Y}) \leq S_{T(\mathbf{X})}. \tag{7}$$

Lemma 2 gives us that $|S_m - \mathbb{E}S_m| < \sqrt{mk}$ with probability $1 - O(e^{-\sqrt{n/k}}) = 1 - o(1)$. Since each ℓ_i is independent from $T(\mathbf{X})$ (how long it takes to collect one k-set tells us nothing about how many k-sets are needed to cover the entire set of coupons), we can use the lemma with $m = T(\mathbf{X})$ to bound the right-hand side of inequality (7), and $m = T(\mathbf{X}) - 1$ for the left-hand side. (The lemma requires that $T(\mathbf{X}) > \frac{n}{k}$, which holds w.h.p. by the first moment method.) This gives us that, w.h.p.,

$$k(T(\mathbf{X}) - 1) - \sqrt{k(T(\mathbf{X}) - 1)} < T(\mathbf{Y}) \leq kT(\mathbf{X}) + \sqrt{kT(\mathbf{X})}.$$

By Theorem 1, $T(\mathbf{X}) < \frac{2n\log n}{k}$ holds w.h.p., and $|T(\mathbf{Y}) - n\log n| < cn$ holds with probability at least $1 - e^{-c} + o(1)$. Applying the triangle inequality, we see that

$$\begin{aligned}\left|T(\mathbf{X}) - \frac{n\log n}{k}\right| &\leq \left|\frac{T(\mathbf{Y})}{k} - \frac{n\log n}{k}\right| + \left|T(\mathbf{X}) - \frac{T(\mathbf{Y})}{k}\right| \\ &\leq c\cdot\frac{n}{k} + \frac{\sqrt{T(\mathbf{X})}}{k} + 1 \leq c\cdot\frac{n}{k} + \frac{\sqrt{2n\log n}}{k\sqrt{k}} \\ &= (c + o(1))\cdot\frac{n}{k}\end{aligned}$$

holds with probability at least $1 - e^{-c} + o(1)$. The theorem follows. □

We now turn our attention to Theorem 7. Suppose that we have an exchangeable random covering variable W for the set $V = [n]$. Let $\mu = \mathbb{E}|W|$, let M be the maximum value that $|W|$ takes with strictly positive probability, and let $\chi = \mathbb{E}[|W|^2]$.

Coupling the W-coupon sequence $\mathbf{W} = (W_1, W_2, \dots)$ with the singleton coupon sequence $Y_1, Y_2, \dots$ as in the proof of Theorem 8, we get the following analogue of Equation (4):

$$\sum_{i=1}^{T(\mathbf{W})-1} \ell_i < T(\mathbf{Y}) \le \sum_{i=1}^{T(\mathbf{W})} \ell_i, \tag{8}$$

where ℓ_i is the least integer such that $C_Y[\ell_1 + \cdots + \ell_{i-1} + 1, \ell_1 + \cdots + \ell_i] = |W_i|$. Applying Wald's inequality, we get that

$$\frac{\mathbb{E}T(\mathbf{Y})}{\mathbb{E}\ell_1} \le \mathbb{E}T(\mathbf{W}) < 1 + \frac{\mathbb{E}T(\mathbf{Y})}{\mathbb{E}\ell_1}. \tag{9}$$

In particular if $\mathbb{E}\ell_1 = o(n \log n)$, we have $\mathbb{E}T(\mathbf{W}) = (1 + o(1))\frac{n \log n}{\mathbb{E}\ell_1}$. An inconvenient aspect of this expression is that it remains in terms of $\mathbb{E}\ell_1$, the expected number of single coupon we need to draw in order to see $|W|$ distinct coupons. However if $M = o(n)$, note that for any $m \le M$ the expected number of single coupons we need to draw in order to see m distinct coupons is

$$\sum_{i=0}^{m-1} \frac{n}{n-i} = (1 + o(1))n \log\left(\frac{n}{n-m}\right) = (1 + o(1))m, \tag{10}$$

and thus $\mathbb{E}\ell_1 = (1 + o(1))\mathbb{E}|W|$. Together with (9), (10) establishes the following:

Proposition 4. *For the W-collector with maximum coupon size M and mean coupon size μ, the following hold:*

1. *if $M = o(n)$, $\mathbb{E}T(\mathbf{W}) = (1 + o(1))\frac{n \log n}{\mu}$;*
2. *if $\mathbb{E}\ell_1 = o(n \log n)$, $\mathbb{E}T(\mathbf{W}) = (1 + o(1))\frac{n \log n}{\mathbb{E}\ell_1}$;*
3. *if $\mathbb{E}\ell_1 = \Omega(n \log n)$, $\mathbb{E}T(\mathbf{W}) = O(1)$.*

Theorem 7, which we now prove gives conditions for the covering time $T(\mathbf{W})$ to be sharply concentrated around its expected value.

Proof of Theorem 7. We first prove that if any of conditions *1–3* holds, then w.h.p. $T(\mathbf{W}) = (1+o(1))\frac{n\log n}{\mu}$. Note that *1–3* give us $M = o(n)$, whence $\mathbb{E}\ell_1 \le (1+o(1))M = o(n\log n)$. As in Theorem 8, having sandwiched $T(\mathbf{Y})$ between two sums of independent identically distributed random variables $S_{T(\mathbf{W})-1} := \sum_{i=1}^{T(\mathbf{W})-1} t_i$ and $S_{T(\mathbf{W})} := \sum_{i=1}^{T(\mathbf{W})} t_i$, the crux of the proof lies in showing these two (random) sums are concentrated around their respective means. Indeed, provided we can show that w.h.p. $S_{T(\mathbf{W})} = (1+o(1))T(\mathbf{W})\mathbb{E}\ell_1$ and $S_{T(\mathbf{W})-1} = (1+o(1))(T(\mathbf{W})-1)\mathbb{E}\ell_1$, we have that w.h.p.

$$(1+o(1))\frac{n\log n}{\mathbb{E}\ell_1} = (1+o(1))\frac{T(\mathbf{Y})}{\mathbb{E}\ell_1} \le T(\mathbf{W}), \text{ and}$$
$$T(\mathbf{W}) \le (1+o(1))\frac{T(\mathbf{Y})}{\mathbb{E}\ell_1} + 1 = (1+o(1))\frac{n\log n}{\mathbb{E}\ell_1}$$

by appealing to Theorem 1 (and the fact that $\mathbb{E}\ell_1 = o(n\log n)$ by (10). Let us therefore establish the concentration we require.

We use the following generalized Chernoff bound, see e.g. Theorems 2.8 and 2.9 in [9].

Lemma 3. (Generalized Chernoff bound.) *Let $(U_i)_{i=1}^t$ be a sequence of independent, identically distributed non-negative integer-valued random variables, with $U := U_1 \le M$ with probability* 1. *Let $\varepsilon > 0$ be fixed. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^t U_i - t\mathbb{E}U\right| \ge \varepsilon\mathbb{E}U\right) \le 2\exp\left(-\frac{\varepsilon^2 t^2(\mathbb{E}U)^2}{2t\mathbb{E}[U^2] + 2Mt\mathbb{E}U/3}\right).$$

We apply the Lemma to $|W|$. Suppose condition *2* holds. Then $M = o(n)$ and thus $\mathbb{E}\ell_1 = (1+o(1))\mu$. For any fixed $\varepsilon > 0$ and $t = (1+o(1))\frac{n\log n}{\mu}$ we have that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^t |W_i| - t\mu\right| \ge \varepsilon t\mu\right) &\le 2\exp\left(-\frac{\varepsilon^2 t\mu^2}{2\chi + 2M\mu/3}\right)\\ &\le 2\exp\left(-\varepsilon^2(1+o(1))\frac{n\log n\mu}{2M^2 + 2M\mu/3}\right)\\ &= \exp\left(-\varepsilon^2(1+o(1))\frac{n\log n\mu}{2M^2}\right) = o(1),\end{aligned}$$

where the last equality used the fact that $M = o(n\log n)$. Thus for t around the expected value of $T(\mathbf{W})$, the sum $S_t = \sum_{i=1}^t \ell_i$ is w.h.p. concentrated around its mean $(1+o(1))\mu t$. It follows that if *2* is satisfied then w.h.p. $T(\mathbf{W}) = (1+o(1))\frac{T(\mathbf{Y})}{\mu}$, as desired. Since condition *2* implies *1* this also establishes that *1* is sufficient for $T(\mathbf{W})$

to be sharply concentrated around $\frac{n\log n}{\mu}$. For condition *3*, we use the same argument as for *2* but use the assumption $\chi = o(n\log n\mu)$ to bound χ instead of the bound $\chi \leq M^2$.

For conditions *4*, we show that we can truncate W; for $\varepsilon > 0$ fixed, Chebyshev's inequality implies

$$\mathbb{P}\left(\big||W| - \mu\big| > \varepsilon\mu\right) \leq \frac{\chi - \mu^2}{\varepsilon^2\mu^2} = o\left(\frac{1}{n\log n\mu}\right).$$

Thus the expected number of coupons with size differing from μ by more than $\varepsilon\mu$ which occur by time $t = (1+o(1))\frac{n\log n}{\mathbb{E}\ell_1} \leq (1+o(1))\frac{n\log n}{\mu}$ is $o(1)$. By Markov's inequality w.h.p. no such coupon is seen by that time, and we can couple/sandwich the W-coupon collectors between two k-uniform exchangeable coupon collectors X^- and X^+, collecting coupons of size $k_- = (1-\varepsilon)\mu$ and $k_+ = (1+\varepsilon)\mu$ respectively, in such a way as to have $T(\mathbf{X}^-) \leq T(\mathbf{W}) \leq T(\mathbf{X}^+)$.

We then split into two cases. If $\mu = o(n)$, then by Theorem 8 w.h.p. these two sandwiching coupon collectors finish at times $T(\mathbf{X}^-) = (1+o(1))\frac{n\log n}{(1-\varepsilon)\mu}$ and $T(\mathbf{X}^+) = (1+o(1))\frac{n\log n}{(1+\varepsilon)\mu}$ respectively. Since $\varepsilon > 0$ was arbitrary we deduce that $T(\mathbf{W}) = (1+o(1))\frac{n\log n}{\mu}$ as desired. If on the other hand $\mu = cn$ for some $c \in (0,1)$, then by Corollary 3.1 w.h.p. these two sandwiching coupon collectors finish at times $T(\mathbf{X}^-) = (1+o(1))\frac{\log n}{-\log(1-c(1-\varepsilon))}$ and $T(\mathbf{X}^+) = (1+o(1))\frac{\log n}{-\log(1-c(1+\varepsilon))}$ respectively (provided we picked ε sufficiently small so that $c(1+\varepsilon) < 1$ and $c(1-\varepsilon) > 0$). Since $\varepsilon > 0$ was arbitrary we deduce that $T(\mathbf{W}) = (1+o(1))\frac{\log n}{-\log(1-c)}$ as desired. □

4. Fast coverage

Let V be an n-set, and let X be a random covering variable for V with average coupon size $\mu = \mathbb{E}|X|$. If $\mu < (1-\delta)n$ for some fixed $\delta > 0$ and X is exchangeable and uniform, then w.h.p. the covering time $T(\mathbf{x})$ for the X-coupon collector satisfies $T(\mathbf{X}) = (1+o(1))\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}$ (Corollary 3.1). However if we replace the 'exchangeable' assumption by 'transitive', $T(\mathbf{X})$ can be sharply concentrated on a strictly smaller value. For a balanced, not necessarily uniform X with average coupon size $\mu < (1-\delta)n$, we say that the X-coupon collector is *fast* if there exists a strictly positive constant $\eta > 0$ such that w.h.p. $T(\mathbf{X}) < (1-\eta)\frac{\log n}{\left(1-\frac{\mu}{n}\right)}$. In this section, we briefly discuss fast coverage. We have already seen one example of a fast coupon collector in Example 5. We now give a second example of a fast collector which demonstrates a different way of getting fast coverage.

Example 6. [Coupon collecting on a smaller set] Let $V = [k] \times [n]$. For every $i \in [n]$, let $X = [k] \times \{i\}$ with probability $\frac{1}{n}$.

The covering variable X in the example above is transitive and k-uniform. Set $N = |V|$ and $k = n^{\alpha} = N^{\frac{\alpha}{1+\alpha}}$. Provided $k = o(N)$ (i.e. provided $\alpha = O(1)$), the covering time of a exchangeable k-uniform coupon collector on an N-set is w.h.p. concentrated around $(1 + o(1))\frac{N \log N}{k}$. However the X-coupon collector is really collecting from a smaller set of size n: we may identify each of the coupons $[k] \times \{i\}$ with a singleton $\{x_i\}$. We can then couple the X-collector on V with a 1-uniform exchangeable coupon collector $\mathbf{X}'$ on the set $\{x_1, x_2, \dots x_n\}$. By Theorem 1, the covering time $T(\mathbf{X})$ is thus w.h.p. concentrated around $T(\mathbf{X}') = (1 + o(1))n \log n = \left(\frac{1}{1+\alpha} + o(1)\right) \frac{N \log N}{k}$. Thus for any $\alpha > 0$, the X-coupon collector finishes collecting earlier than one would expect knowing only the mean-size of its coupons.

4.1. Sufficient conditions for fast coverage

We have given two instances of fast coverage so far. In Example 5, fast coverage occurred because though the average coupon size was small, there was a small chance of ‘winning the lottery’ and receiving a very large coupon. In Example 6, fast coverage occurred because X was structured in such a way that the problem of covering $V = [kn]$ with k-sets was actually equivalent to the problem of covering a much smaller set $V' = [n]$, which could be achieved more rapidly (and also entailed having some very large pairwise correlations q_{xy}).

We can restate these two ‘speeding up’ properties in a formal way.

Theorem 9. *Let V be an n-set. Let X be a transitive coupon variable for V with average coupon size $\mu = o(n)$. Then if any of the following conditions are satisfied, X is fast:*

1. *there exist some $\varepsilon > 0$ and $C \geq 1 + \varepsilon$ such that $\mathbb{P}[|X| \geq C\mu] \geq \frac{1+\varepsilon}{C}$;*

2. *there exists $1 \ll n' \leq \frac{n}{\mu}$, and a partition of V into n' subsets $V = \sqcup_{i=1}^{n'} V_i$ such*

that $\mathbb{P}(V_i \subseteq X) \geq (1+\varepsilon)\frac{\log n'}{\log n}\left(-\log\left(1-\frac{\mu}{n}\right)\right)$ *for every* $i \in [n']$.

Proof. Suppose condition *1* is satisfied. Let $\eta > 0$ be a fixed positive number to be fixed later. We say that coupons of size at least $C\mu$ are *large*, and we call other coupons *small*. We couple X with a transitive $C\mu$-uniform covering variable Y, by setting Y to be a $C\mu$-subset of X chosen uniformly at random if X is large, and to be the empty set otherwise. By Proposition 2, w.h.p. the Y-collector will need at most $(1+\eta)\frac{\log n}{-\log\left(1-\frac{C\mu}{n}\right)}$ non-empty coupons to cover V. Set $p = \frac{1+\varepsilon}{C}$. Let t be an integer with

$$\left(\frac{1+\eta}{1-\eta}\right)\left(\frac{1}{p}\right)\frac{\log n}{-\log\left(1-\frac{C\mu}{n}\right)} \leq t \leq (1-\eta)\,\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}.$$

Since the left hand side is at most $\frac{1+\eta}{1-\eta}\frac{1+o(1)}{1+\varepsilon}\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}$, picking η sufficiently small relative to ε and n sufficiently large, we can always do this. We claim that w.h.p. the Y-collector will have covered all of V by time t. Indeed, the probability that $Y \neq \emptyset$ is, by assumption, at least p. By a standard Chernoff bound, the probability that at least $(1-\eta)pt$ of the first t coupons of the Y-coupon collectors are non-empty is at least $1 - e^{-\frac{\eta^2 pt}{3}} = 1 - o(1)$. (Here we use the fact that $pt = \Omega\left(\frac{\log n}{-\log\left(1-\frac{C\mu}{n}\right)}\right) \to \infty$ as $n \to \infty$.) Thus w.h.p. by time t we have seen at least $(1-\eta)pt$ non-empty Y-coupons; since, by our choice of t, this is at least $(1+\eta)\frac{\log n}{-\log\left(1-\frac{C\mu}{n}\right)}$, whence w.h.p. these non-empty Y-coupons cover all of V. Our coupling of Y with X then implies that w.h.p. $T(\mathbf{X}) \leq t$. Since by definition $t \leq (1-\eta)\,\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}$, we conclude that X is fast.

For the second part of the theorem, suppose condition *2* is satisfied. We define a random covering variable Z for $[n']$ as follows: set $Y = \{i : V_i \subseteq X\}$. Set $p = (1+\varepsilon)\frac{\log n'}{\log n}\left(-\log\left(1-\frac{\mu}{n}\right)\right)$. Let $\eta > 0$ be chosen sufficiently small so that $1+\varepsilon > \frac{1+\eta}{1-\eta}$. Let t be an integer with

$$\frac{(1+\eta)\log n'}{p} \leq t \leq (1-\eta)\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}.$$

By our choice of η, and for n sufficiently large, we can always pick such a t. We claim that w.h.p. the Y-collector will have covered all of $[n']$ by time t. Indeed by condition *2* the expected number of $i \in [n']$ not covered by the Y-coupon collector by time t is

$$\sum_{i\in[n']}(1-\mathbb{P}(i \in Y))^t \leq n'(1-p)^t \leq e^{-\eta\log(n')} = o(1),$$

so that by Markov's inequality w.h.p. the Y-coupon collector has covered $[n']$ by time t. By the coupling of Y with X, and the fact that $\bigcup_i V_i = V$, it follows that w.h.p. $T(\mathbf{X}) \leq t$. Since we chose $t \leq (1-\eta)\frac{\log n}{-\log\left(1-\frac{\mu}{n}\right)}$, we conclude that X is fast. □

Theorem 9 leaves a number of interesting questions open. To begin with, are there other, subtler ways of being fast than either winning the lottery or collecting a smaller coupon set? In particular, are there conditions on the pairwise intensities $(q_{xy})_{x,y\in V}$ which imply fast coverage? Furthermore, Theorem 9 says nothing on what the probable value of $T(\mathbf{X})$ actually is. In cases where X is fast, can we determine good bounds for $\mathbb{E}T$? With its ties to the k-SAT problem (see the next section), this is one of the most important open problems related to this paper.

5. Applications

5.1. Connectivity in random graphs

We consider the discrete time multigraph process $(G_t)_{t\geq 0}$ obtained by starting with the empty graph G_0 on $V = [n]$ and at each time step $t \geq 1$ selecting an edge uv uniformly at random and adding it to G_{t-1} to form G_t. We associate $n/2$ coupon collectors $\mathbf{X}^i$ to this process, $1 \leq i \leq \frac{n}{2}$. The i^{th} such collector aims to cover each i-set A with an edge from A to $V \setminus A$. Since each edge uv connects $2\binom{n-2}{i-1}$ i-sets to their complements in V, the i^{th} collector is $2\binom{n-2}{i-1}$-uniform and balanced, and aims to cover a set of size $\binom{n}{i}$. By Proposition 2, we thus have that her covering time $T(\mathbf{X}^i)$ will be w.h.p. at most $(1+o(1))t_i$ where

$$t_i = \frac{\log\binom{n}{i}}{-\log\left(1-\frac{2\binom{n-2}{i-1}}{\binom{n}{i}}\right)} = \frac{\log\binom{n}{i}}{-\log\left(1-\frac{i(n-i)}{\binom{n}{2}}\right)}.$$

For $i = o(n)$, $t_i = t_1 - \frac{n\log i}{2} + o(n)$, while for $i = \theta(n)$ $t_i = O(t_1/\log n)$. Further by Proposition 2 we know that for any fixed $\eta > 0$ we have that $T(\mathbf{X}^i) > (1+\eta)t_i$ with probability at most $n^{-\eta}$. Also in the case $i = 1$ the collector's random coupon variable is in fact exchangeable and 2-uniform. By Theorem 8, for any $x > 0$

$$\mathbb{P}(T(\mathbf{X}^1) > t_1 + \frac{xn}{2}) \leq e^{-x}(1+o(1)), \text{ and}$$
$$\mathbb{P}(T(\mathbf{X}^1) < t_1 - \frac{xn}{2}) \leq e^{-x}(1+o(1)).$$

Thus by the union bound we have that for any $x = x(n) > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_i T(\mathbf{X}^i) > t_1 + \frac{xn}{2}\right)\\
\leq &\sum_i \mathbb{P}\left(T(\mathbf{X}^i) > t_i\left(1 + \frac{\log i + x}{\log n}\right)(1 + o(1))\right)\\
\leq &(1 + o(1))\sum_{i \geq 1} e^{-\log i + x}\\
\leq &(1 + o(1))e^{-x}\log n
\end{aligned}$$

In particular, setting $x = \varepsilon \log n$, the inequality above together with our bound on $\mathbb{P}(T(\mathbf{X}^1) < t_1 - xn)$ establishes the following:

Theorem 10. *Let $\varepsilon > 0$ be fixed. Then*

$$\begin{aligned}
&\mathbb{P}\left(G_t \textit{ is connected}\right) \leq n^{-\varepsilon + o(1)} && \textit{for } t \leq \frac{n\log n}{2}(1 - \varepsilon),\\
&\mathbb{P}\left(G_t \textit{ is connected}\right) \geq 1 - n^{-\varepsilon + o(1)} && \textit{for } t \geq \frac{n\log n}{2}(1 + \varepsilon).
\end{aligned}$$

It is easy to relate G_t to the *size model* $G_{n,m}$ of random graphs obtained by selecting m-distinct edges uniformly at random and adding them to the empty graph on n vertices. Indeed Markov's inequality shows that for $t = O(n \log n)$, w.h.p. G_t contains only $O(\frac{t^2}{n^2}) = O((\log n)^2)$ repeated edges, so one can couple G_t with with $G_{n,m}$ up to the connectivity threshold for G_t in such a way that $G_{n,t-O((\log n)^2)} \subseteq G_t \subseteq G_{n,t}$. In this way, Theorem 10 allows us to recover (a slightly weaker form of) the classical results of Erdős and Rényi [17] on the connectivity threshold for $G_{n,m}$: w.h.p. $G_{n,m}$ becomes connected at size $m = (1 + o(1))\frac{n\log n}{2}$.

5.2. Covering a square with random discs

We return to Example 4. Let V be the torus obtained by identifying the opposite sides of the square of area n $[0, \sqrt{n}]^2 \subset \mathbb{R}^2$, and let X be the intersection of V with the disc of radius $r = r(n)$ about a uniformly chosen random point $x \in V$. Draw a sequence $\mathbf{X} = (X_1, X_2, \ldots)$ of independent random subsets of V distributed according to X. When does their union w.h.p. cover V? This is known as a *coverage* problem, and is a continuous analogue of the coupon collector problem. Coverage problems have been widely studied in random geometric graph theory, with motivation coming from applications to wireless networks, especially sensor networks (see the introduction of [53] for a history of coverage problems).

We discretise the problem and apply our results to show sharp concentration of the covering time $T = T(\mathbf{X})$ in the case where $r(n)$ is of order $o(\sqrt{n})$ and bounded away from 0 (so the measure of X is $O(\pi r^2) = o(n)$). Tile V with squares of side length s, where $s = s(r, n)$ is chosen so that $s = o(r)$, $s = n^{o(1)}$, and $\sqrt{n}/s \in \mathbb{N}$. Let $\mathcal{T}$ denote the collection of all the tiles; by construction, $|\mathcal{T}| = n/s^2$ Given a disc D of radius r in V, we let I_- to be the collection of tiles wholly contained inside D, and I_+ to be the collection of tiles having non-empty intersection with D. The random variable X gives rise, via I_- and I_+, to two random variables X_- and X_+ taking values among the subsets of $\mathcal{T}$.

For any D as above, it is easy to show (see e.g. Lemma 8 of [18]) that the boundary of D meets at most $\frac{18\pi r}{s}$ tiles; thus $|I_-|$ and $|I_+|$ are both within $\frac{18\pi r}{s}$ of $\frac{|D|}{s^2} = \frac{\pi r^2}{s^2}$. Both of X_- and X_+ are clearly balanced random covering variables for $\mathcal{T}$.

By part *3* of Theorem 5 their covering times $T(\mathbf{X}_-)$ and $T(\mathbf{X}_+)$ are therefore w.h.p. concentrated around $\frac{\log(ns^{-2})}{-\log\left(1-\frac{\pi r^2}{n}\right)} = (1 + o(1))\frac{n \log n}{\pi r^2}$. Since by construction of the random variable X_- and X_+ we have that $T(\mathbf{X}_-) \leq T(\mathbf{X}) \leq T(\mathbf{X}_+)$, we deduce that w.h.p. the covering time for the torus V satisfies $T(\mathbf{X}) = (1 + o(1))\frac{n \log n}{\pi r^2}$.

It is easy to adapt the argument above to show that the covering time does not change significantly if instead of a torus we try to cover a square S of area n with discs of radius r centred at uniformly chosen random points in S. The random covering variables we use are no longer quite balanced: there are $O(\frac{r\sqrt{n}}{s^2})$ tiles within distance r of the boundary of S, each of which is covered with probability at least $\frac{\pi r^2}{2n}(1 + o(1))$, and $O(\frac{r^2}{s^2})$ tiles within distance r of a corner of S, each of which is covered with probability at least $\frac{\pi r^2}{4n}(1 + o(1))$. The first moment method shows both of these sets of 'boundary tiles' are w.h.p. covered by the time we have drawn $(1 + \varepsilon)\frac{n \log n}{\pi r^2}$ discs, while the 'central tiles' at distance at least r from the boundary are w.h.p. covered by that time by our result for the torus. This yields the following well-known result on covering processes (see [25]).

Theorem 11. *Let V be a square or torus of area n. Let X be the intersection of V with a disc of radius r about a uniformly chosen random point in V, where $r = r(n)$ is* *bounded away from* 0 *and satisfied* $r(n) = o(\sqrt{n})$*. Then w.h.p. the covering time* T *of the continuous* X*-coupon collector on* V *satisfies* $T(\mathbf{X}) = (1+o(1))\frac{n\log n}{\pi r^2}$.

As an example of how to apply part *2* of Theorem 5, we can instead let the radius be a random variable R satisfying $F(r) := \mathbb{P}(R \geq r) = \Theta(r^{-\alpha})$ (as $r \to \infty$) for some fixed $\alpha > 0$. We now want to show that for all but $n^{1+o(1)}$ pairs of tiles x, y, $q_{xy}/q_x = o(1/\log n)$ (uniformly in x and y).

A necessary condition for a disc C covering x to also cover y is that its radius is at least $d(x,y)/2$. The distribution of the area of C, conditional on $x \in C$, is the size-biased version of πR^2. Hence

$$\mathbb{P}(\text{radius}(C) \geq r) = \Theta(r^2\mathbb{P}(R \geq r)) = \Theta(r^{2-\alpha}),$$

and it follows that $q_{xy}/q_x = O(d(x,y)^{2-\alpha})$. If $\alpha = 2+\varepsilon$ for some $\epsilon > 0$, then we can apply part *2* of Theorem 5 with pairs x, y considered 'bad' if $d(x,y) < (\log n)^{2/\varepsilon}$. For each x there are $O((\log n)^{4/\varepsilon}) = n^{o(1)}$ y's within that distance, and for x, y further apart $q_{xy}/q_x = O((\log n)^{-2}) = o(1/\log n)$. So the conditions of part *2* are satisfied, and the cover time is concentrated around its mean. If on the other hand $\alpha \leq 2$, the probability that $R > \sqrt{2n}$ (in which case that disc will cover the entire torus) is at least $\Omega(n^{-1})$. This is analogous to Example 5 (coupon collector with lottery) and the cover time will not be sharply concentrated.

More generally, our argument for Theorem 11 in the torus adapts immediately to any balanced random covering variable X taking values among the compact subsets of V and satisfying with probability 1 *1* $|X| \leq \varepsilon|V|$, and *2* $|\partial X| \leq \varepsilon|X|$, where $|\partial X|$ denotes the measure (length) of the boundary of X and $\varepsilon = o(1)$. For such X, we again have

$$T(\mathbf{X}) = (1+o(1))\frac{|V|\log|V|}{\mathbb{E}|X|}.$$

Thus we may replace 'disc' in the results above by e.g. 'ellipse', 'annulus', 'square', 'polygon', or even let X be given by a probability distribution on a finite collection of shapes having the same Lebesgue measure and satisfying the required isoperimetric inequality. These are special cases of a celebrated result of Janson [31].

5.3. Covering the edges of a graph by spanning trees, and matroids by bases

Let G be a connected edge-transitive graph, on n vertices, of minimum degree d and let X be a spanning tree of G drawn uniformly at random from the set of all such trees. Our goal is now to cover the edge set E of G with the edges of trees from X.

It is well known that the random spanning tree is pairwise negatively correlated, with respect to the edges, in fact it satisfies the even stronger negative correlation property of being a Rayleigh measure on E, see [8].

So, from Theorem 5 we can conclude that the covering time T is sharply concentrated around $\frac{(nd/2)\log(nd/2)}{n-1}$, as long as $d \gg 1$.

Covering the edge set of a graph is a special case of the problem of covering the ground set of a matroid by random drawn bases of the matroid. In [19] it was shown that a large class of matroids, the *balanced* matroids, which contain the class of cycle matroids of a graph, have pairwise negative correlation. In the same way as for trees we can conclude that if a balanced matroid of size n has rank r then the covering time T is sharply concentrated around $n\log(n)/r$, as long as $\log(r) = o(\log(n))$.

5.4. Random k-SAT

The Random Boolean Satisfiability (SAT) problem is the following. Given n Boolean variables $x_1, x_2, \ldots x_n$ and an integer sequence $k = k(n)$, we form a random clause $C = l_1 \vee l_2 \vee \ldots \vee l_k$ by selecting a k-subset $\{y_1, y_2, \ldots y_k\}$ of literals uniformly at random, setting $l_i = y_i$ with probability $1/2$ and $l_i = \neg y_i$ otherwise, independently for each i, and taking C to be the join of the literals l_i. We now consider a sequence of independent, identically distributed random clauses $C_1, C_2, \ldots$, with distribution given by C, and define a sequence of logical formulae in conjunctive normal form $F_t = \bigwedge_{i=1}^{t} C_i$ for $t = 0, 1, \ldots$. For $n \to \infty$, the random k-SAT problem asks whether or not there exists w.h.p˙ an assignment of truth values to the variables $x_1, \ldots, x_n$ such that the logical formula F_t is satisfied. The random k-SAT problem is of fundamental importance to theoretical computer science and has been extensively studied (see [11]).

Here we note that this problem is equivalent to determining the covering time of a coupon collector problem. The space of satisfying assignments for a formula consisting of t clauses involving n variables can be viewed as the complement of the union of t subcubes of $\{0,1\}^n$. If each of those t clauses involves exactly k distinct literals (that is, if we are working with an instance of k-SAT), then each of those t subcubes has dimension $n-k$. In particular, we can couple the sequence of i.i.d. clauses $C_1, C_2, \ldots$ with a sequence of independent coupons $X_1, X_2, \ldots$, with $X_i \sim X$, where X is the random coupon given by selecting an $(n-k)$-dimensional subcube of the n-dimensional discrete hypercube $V = \{0,1\}^n$ uniformly at random. The formula F_t is then satisfiable if and only if the X-coupon collector has failed to cover V by time t.

The random variable X is 2^{n-k}-uniform and transitive. Proposition 2 thus gives some elementary upper bounds on the satisfiability threshold $T(\mathbf{X})$ for F_t: for any $\varepsilon > 0$, w.h.p.

$$T(\mathbf{X}) \leq (1+\varepsilon)\frac{\log 2^n}{-\log(1-2^{-k})} = (1+\varepsilon)n\frac{\log 2}{-\log(1-2^{-k})}.$$

For $k(n)$ large enough this bound is in fact an equality, as first proven in [23]. Using Theorem 5 we can obtain the same result.

Theorem 12. *Let $k = \log_2 n + \omega(n)$, where $\omega(n) \to \infty$, then w.h.p. $T(\mathbf{X}) = (1+o(1))n2^k \log 2$*

Proof. Let $N = 2^n$ be the number of vertices in the hypercube Q_n, our base set. We shall show condition *1* in Theorem 5 is satisfied to deduce the claimed sharp concentration result for $T(\mathbf{X})$. Checking that *1* holds is a matter of simple computations. Most of the estimates needed here are standard so we only sketch the argument. We note first of all that in our setting, for any pair of vertices x and y at Hamming distance i in the hypercube Q_n,

$$q_{xy} = \frac{\binom{n-i}{k}}{2^k\binom{n}{k}}$$

By symmetry, condition *1* is equivalent to

$$S = \sum_{y \neq 0}(\exp(tq_{0,y}) - 1) = o(N).$$

Now the threshold for k-satisfiability we shall obtain from Theorem 5 (which is the first moment threshold) is $N \log N/(N/2^k) = 2^k n \log 2$. We therefore let $t =$

$2^k n\log 2\cdot(1-\delta_n)$ for some $\delta_n = o(1)$ to be determined later. Now

$$S = \sum_y [-1+\exp(n\log 2\cdot(1-\delta_n)\cdot 2^k q_{0,y})]$$
$$= \sum_{i=1}^{n-k}\binom{n}{i}\Big[-1+\exp\Big(n\log 2\cdot(1-\delta_n)\cdot\frac{\binom{n-i}{k}}{\binom{n}{k}}\Big)\Big]$$

Let a_i be the i:th term of this sum. We deal separately with the three cases $i \ge \frac{n}{2}-\frac{n}{k}$, $\frac{n\ln k}{k-1} \le i < \frac{n}{2}-\frac{n}{k}$ and $i < \frac{n\ln k}{k-1}$. In the first case, we change the summation index so that $i = \frac{n}{2}-j$. Note that

$$\frac{\binom{n-i}{k}}{\binom{n}{k}} \le \big(1-\frac{i}{n}\big)^k = 2^{-k}\big(1+\frac{2j}{n}\big)^k \le 2^{-k}\exp\Big(\frac{2jk}{n}\Big)$$

Summing over all j such that $-\frac{n}{2} \le j < \frac{n}{k}$ we get that

$$2^{-n}\sum_{i=\frac{n}{2}-\frac{n}{k}}^{n-k} a_i = 2^{-n}\sum_{j=k-\frac{n}{2}}^{\frac{n}{k}}\binom{n}{\frac{n}{2}-j}\Big[-1+\exp\Big(n\log(2)\cdot\frac{\binom{\frac{n}{2}-j}{k}}{\binom{n}{k}}\Big)\Big]$$
$$\le -1+\exp(2^{-\omega(n)}e^2) = o(1)$$

In the second case a convexity argument shows that

$$2^{-n}\sum_{i=\frac{n\ln k}{k-1}}^{\frac{n}{2}-\frac{n}{k}} a_i \le \exp\Big(\log n - \frac{2n}{k^2}+o(1)\Big) = o(1)$$

Finally, for $i \le \frac{n\log k}{k-1}$, coarser bounds suffice: $\binom{n-i}{k}/\binom{n}{k} \le 1$ and $\log\binom{n}{i} \le 2i\log(n/i)$. Thus

$$2^{-n}\sum_{i=1}^{\frac{n\log k}{k-1}} a_i \le 2^{-n}\sum_{i=1}^{\frac{n\log k}{k-1}}\exp\Big(2i\log\big(\frac{n}{i}\big)+n\log(2)(1-\delta_n)\Big)$$
$$= n\exp\Big(n\Big[\frac{2\log(k)^2}{k}-\log(2)\delta_n\Big]\Big),$$

which is $o(1)$ provided $\log(2)n\delta_n - \frac{2n\log(k)^2}{k} - \log(n) \to \infty$; this is satisfied for instance if we choose $\delta_n = (\log n)^{-\frac{1}{2}}$.

Together these three cases, and the choice of δ_n above give that $2^{-n}\sum_{i=1}^{n-k} a_i = o(1)$, or in other words that $S = o(N)$, and condition *1* is satisfied (since $S = o(N)$ and $\sum_x e^{-q_x t} = e^{N\delta_n} = \omega(1)$). The result is then immediate from Theorem 5. □

For constant k the simple first moment bound does not give the correct value for the satisfiability threshold. For $k = 3$ our simple upper bound is $T(\mathbf{X}) \le (5.190\ldots + o(1))\,n$

and it has been shown that $T(\mathbf{X}) \leq 4.506n$, [14]. Heuristics based on spin-glass theory has lead to the conjecture that the correct threshold is $4.267\ldots n$, see [42]. The well known satisfiability conjecture states that for each k there exists a constant c_k such that the threshold for random k-SAT is $c_k n$. Recently a proof of this conjecture for sufficiently large values of k has been announced [13]. It is also known [11] that as k increases the threshold location scales as $2^k \ln 2 - 1/2(1+\ln 2) + o_k(1)$, thus matching to leading order the bound given by the coupon collector.

6. Concluding remarks

Another natural coupon collector problem is the q-colourability of the uniform random graph. Here the set V we are covering is the set of all strings of length n over the alphabet $[q]$. Each string is interpreted as a vertex colouring of an n vertex graph. For each edge e in the complete graph on n vertices we create a coupon consisting of all colourings in which the endpoints of e have the same colour. The covering time T for this coupon process now corresponds to the threshold for a uniform random graph of size T on n vertices ceasing to be q-colourable.

This coupon collector process is not balanced: colourings with more unequal colour class sizes induce more monochromatic edges and are therefore easier to cover. However, most colourings have almost balanced color class sizes, and restricting our attention to balanced colourings leads to a transitive coupon collector.

Denote by X the random coupon variable associated with the process; X has size $q^{-1}|V|$ and is transitive and uniform, but is very much non-exchangeable: there are both strong positive and strong negative correlations between the various colourings, so that our Theorems 5 and 6 do not apply. For $q = 2$, it is known that the covering time $T(\mathbf{X})$ is *not* sharply concentrated. This stands in contrast with the situation for $q \geq 3$: in [1] it was proven that the chromatic number of a random graph with edge probability $p = \frac{c}{n}$ has two possible values, and for all but a discrete sequence of values for c w.h.p. only one value. This result would follow directly from a sharp threshold result for the coupon collector process described above.

A natural question is then whether any transitive, cn-uniform random coupon variable X with $c > 0$ sufficiently small has sharp concentration of $T(\mathbf{X})$, i.e. whether random q-colouring threshold for large q is determined by general coupon collector results (as opposed to specific structural features of the random colouring setting).

In a different direction, much remains to be done on the case of fast coupon collectors, as remarked at the end of Section 4. The k-SAT problem for small k gives us an example of a transitive, uniform and linear-sized coupon collector which is fast. The difficulty of that problem suggests the rigorous study of fast coupon collectors will be hard in general. Nevertheless we feel that the following problems are well-motivated, and for μ small enough may prove tractable.

Problem 1. Let X be a μ-uniform transitive random covering variable for an n-set V.

1. Give estimates for the value of $T_{\frac{1}{2}}$ in terms of μ and the pairwise intensities $q_{xy} = \mathbb{P}(\{x, y\} \subseteq X)$, $x, y \in V$;
2. Give sufficient conditions for $T(\mathbf{X})$ to be sharply concentrated about $T_{\frac{1}{2}}$.

References

[1] D. Achlioptas and A. Naor. (2005). The two possible values of the chromatic number of a random graph. *Ann. of Math. (2)*, 162(3):1335–1351.

[2] I. Adler and S.M. Ross. (2001). The coupon subset collection problem. *Journal of Applied Probability*, pp 737–746.

[3] D.J. Aldous. (1989). An introduction to covering problems for random walks on graphs. *Journal of Theoretical Probability*, 2(1):87–89.

[4] D.J. Aldous. (1991). Threshold limits for cover times. *Journal of Theoretical Probability*, 4(1):197–211.

[5] A.D. Barbour and L. Holst. (1989). Some applications of the stein-chen method for proving poisson convergence. *Advances in Applied Probability*, pp 74–90.

[6] L.E. Baum and P. Billingsley. (1965). Asymptotic distributions for the coupon collector's problem. *The Annals of Mathematical Statistics*, pp 1835–1839, 1965.

[7] B. Bollobás and A.G. Thomason. (1987). Threshold functions. *Combinatorica*, 7(1):35–38.

[8] J. Borcea, P. Brändén, and T. Liggett. (2009). Negative dependence and the geometry of polynomials. *J. Amer. Math. Soc.*, 22(2):521–567.

[9] Fan R.K. Chung and Linyuan Lu. (2006). *Complex graphs and networks*, volume 107. American mathematical society Providence.

[10] V. Chvátal. (1991). Almost all graphs with 1.44n edges are 3-colorable. *Random Structures & Algorithms*, 2(1):11–28.

[11] A. Coja-Oghlan. (2014). The asymptotic k-SAT threshold. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pp 804–813, New York, NY, USA. ACM.

[12] P.S. de Laplace. (1774). Mémoire sur les suites récurro-récurrentes et sur leurs usages dans la théorie des hasards. *Mém. Acad. Roy. Sci. Paris*, 6:353–371.

[13] J. Ding, A. Sly, and N. Sun. (2015). Proof of the satisfiability conjecture for Large k. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp 59–68.

[14] O. Dubois, Y. Boufkhad, and J. Mandler. (2000). Typical random 3-SAT formulae and the satisfiability threshold. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '00, pp 126–127.

[15] P.J. Eicker, M.M. Siddiqui, and P.W. Mielke. (1972). A matrix occupancy problem. *The Annals of Mathematical Statistics*, 43(3):988–996.

[16] P. Erdős and A. Rényi. (1961). On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 6(1-2):215–220.

[17] P. Erdős and A. Rényi. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61.

[18] V. Falgas-Ravry and M. Walters. (2012). Sharpness in the k-nearest-neighbours random geometric graph model. *Advances in Applied Probability*, 44(3):617–634.

[19] T. Feder and M. Mihail. (1992). Balanced matroids. In *Proceedings of the Twenty-fourth Annual ACM Symposium on Theory of Computing*, STOC '92, pp 26–38. ACM.

[20] W. Feller. (1950). *An Introduction to Probability Theory and Its Applications: Volume One*. John Wiley & Sons.

[21] M. Ferrante and N. Frigo. (2012). A note on the coupon-collector's problem with multiple arrivals and the random sampling. *arXiv preprint arXiv:1209.2667*.

[22] P. Flajolet, D. Gardy, and L. Thimonier. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229.

[23] A. Frieze and N. Wormald. (2005). Random k-SAT: a tight threshold for moderately growing k. *Combinatorica*, 25(3):297–305.

[24] A.M. Gittelsohn. (1969). An occupancy problem. *The American Statistician*, 23(2):11–12.

[25] Peter Hall. (1988). *Introduction to the theory of coverage processes*. John Wiley & Sons Incorporated.

[26] L. Holst. (1977). Some asymptotic results for occupancy problems. *The Annals of Probability*, pp 1028–1035.

[27] L. Holst. (1986). On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review*, 54(1):15–27.

[28] T. Huillet. (2003). Sampling problems for randomly broken sticks. *Journal of Physics A: Mathematical and General*, 36(14):3947.

[29] V.A. Ivanov, G.I. Ivchenko, and Y.I. Medvedev. (1985). Discrete problems in probability theory. *Journal of Soviet Mathematics*, 31(2):2759–2795.

[30] G.I. Ivchenko. (1998). How many samples does it take to see all the balls in an urn? *Mathematical Notes*, 64(1):49–54.

[31] S. Janson. (1986). Random coverings in several dimensions. *Acta Mathematica*, 156(1):83–118.

[32] B.C. Johnson and T.M. Sellke. (2010). On the number of iid samples required to observe all of the balls in an urn. *Methodology and Computing in Applied Probability*, 12(1):139–154.

[33] N.L. Johnson and S. Kotz. (1977). *Urn models and their application: an approach to modern discrete probability theory*. Wiley New York.

[34] A.C. Kaporis, L.M. Kirousis, Y.C. Stamatiou, M. Vamvakari, and M. Zito. (2001). Coupon collectors, q-binomial coefficients and the unsatisfiability threshold. In *Theoretical Computer Science*, pp 328–338. Springer.

[35] E.R. Khakimullin and N.Y. Enatskaya. (1997). Limit theorems for the number of empty cells. *Discrete Mathematics and Applications*, 7:209–220.

[36] J.F.C. Kingman. (1978). Random partitions in population genetics. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 361(1704):1–20.

[37] J.E. Kobza, S.H. Jacobson, and D.E. Vaughan. (2007). A survey of the coupon collector's problem with random sample sizes. *Methodology and Computing in Applied Probability*, 9(4):573–584.

[38] V.F. Kolchin, B.A. Sevast'yanov, and V.P. Chistyakov. (1978). *Random allocations*. Wiley, New York.

[39] J. Larsson, and K. Markström. (2019). Biased random k-SAT *arXiv preprint arXiv:1906.05127.*

[40] N. Mantel and B.S. Pasternack. (1968). A class of occupancy problems. *The American Statistician*, 22(2):23–24.

[41] B.D. McKay and F. Skerman. (2013). Degree sequences of random digraphs and bipartite graphs. *arXiv preprint arXiv:1302.2446*.

[42] M. Mézard and R. Zecchina. (2002). Random K-satisfiability problem: From an analytic solution to an efficient algorithm. *Phys. Rev. E*, 66:056126.

[43] V.G. Mikhailov. (1978). An estimate of the rate of convergence to the poisson distribution in group allocation of particles. *Theory of Probability & Its Applications*, 22(3):554–562.

[44] A. Moivre. (1711). De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus. *Phil. Trans. Roy. Soc. London A*, 27:213–264.

[45] P. Neal and J. Moriary. (2009). Sampling efficiency and biodiversity. *The University of Manchester Probability and Statistics Group Research Report*, 9.

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

[47] V.G. Papanicolaou, G.E. Kokolakis, and S. Boneh. (1998). Asymptotics for the random coupon collector problem. *Journal of computational and applied mathematics*, 93(2):95–105.

[48] G.P. Patil and C. Taillie. (1977). Diversity as a concept and its implications for random environments. *Bulletin de l'Institut International de Statistique*, 4:497–515.

[49] R. Poli. (2005). Tournament selection, iterated coupon-collection problem, and backward-chaining evolutionary algorithms. In *Foundations of Genetic Algorithms*, pp 132–155. Springer.

[50] G. Pólya. (1930). Eine Wahrscheinlichkeitsaufgabe in der Kundenwerbung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 10(1):96–97.

[51] A. Poon, B.H. Davis, and L. Chao. (2005). The coupon collector and the suppressor mutation estimating the number of compensatory mutations by maximum likelihood. *Genetics*, 170(3):1323–1332.

[52] M. Raab and A. Steger. (1998). Balls into Bins – a simple and tight analysis. In *Randomization and Approximation Techniques in Computer Science*, pp 159–170. Springer.

[53] A. Sarkar and M. Haenggi. (2013). Secrecy coverage. *Internet Mathematics*, 9(2-3):199–216.

[54] S. Savage, D. Wetherall, A. Karlin, and T. Anderson. (2001). Network support for ip traceback. *Networking, IEEE/ACM Transactions on*, 9(3):226–237.

[55] T.M. Sellke. (1995). How many iid samples does it take to see all the balls in a box? *The Annals of Applied Probability*, 5(1):294–309.

[56] D.A. Sprott. (1969). A note on a class of occupancy problems. *The American Statistician*, 23(2):12–13.

[57] W. Stadje. (1990). The collector's problem with group drawings. *Advances in Applied Probability*, pp 866–882.

[58] S. Vasudevan, D. Towsley, D. Goeckel, and R. Khalili. (2009). Neighbor discovery in wireless networks and the coupon collector's problem. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pp 181–192. ACM.

[59] V.A. Vatutin and V.G. Mikhailov. (1983). Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory of Probability & Its Applications*, 27(4):734–743.