A Generalized Cover Time for Random Walks on Graphs

Cyril Banderier[1] and Robert P. Dobrow[2]

[1] Algorithms Project. INRIA (Rocquencourt), France.
`Cyril.Banderier,at,inria.fr,`
`http://algo.inria.fr/banderier/`
[2] Clarkson University (Potsdam, NY), USA.
`dobrowb@clarkson.edu,`
`http://www.clarkson.edu/~dobrowb/`

Abstract. Given a random walk on a graph, the cover time is the first time (number of steps) that every vertex has been hit (covered) by the walk. Define the *marking time* for the walk as follows. When the walk reaches vertex v_i, a coin is flipped and with probability p_i the vertex is *marked* (or colored). We study the time that every vertex is marked. (When all the p_i's are equal to 1, this gives the usual cover time problem.) General formulas are given for the marking time of a graph. Connections are made with the generalized coupon collector's problem. Asymptotics for small p_i's are given. Techniques used include combinatorics of random walks, theory of determinants, analysis and probabilistic considerations.

1 Introduction

The following problem was submitted during a supper at the meeting Analysis of Algorithms in June 1999, at Barcelona:

Conjecture 1 (Supper Conjecture) *Imagine m guests around a table, someone has the water carafe and decides to pour some water in his glass with probability p. Then he gives the water carafe randomly to his right or left neighbor. This one does the same, and so on. Call $T(p)$ the number of carafe moves before everyone has got some water. What can one say about the average time $E(T(p))$? In particular, is it true that $pE(T(p)) \to mH_m$ when $p \to 0$? (H_m is the m-th harmonic number).*

One will show that this conjecture is indeed true!

In fact, we will tackle the problem for a slightly more general problem: think about a dinner where some people are more or less not in speaking terms with some others and so they give the carafe preferentially to their friends!

The problem we consider was motivated by the game Trivial Pursuit. Players move pieces around a game board answering questions on various topics (e.g., history, sports, etc.). On certain positions in the game board, if a player answers a question correctly he gets a colored piece. And in our simplified version of the game, when a player gets all the colored pieces he wins. Assume that for each topic there is some probability of answering a question correctly. How long will a typical Trivial Pursuit game take to play?

The reader should be able to see the connection with the following model. Consider the usual (discrete) random walk on a connected directed graph G with vertex set $V = \{v_1, \ldots, v_m\}$. When the random walk reaches vertex v_i, that vertex is *marked* with probability p_i, $i = 1, \ldots, m$. We are interested in the *marking time* $T(\{p_1, \ldots, p_m\})$, the first time that all the vertices have been marked.

When all the p_i's are equal to 1, this is the usual cover time problem. When the graph is the complete graph and the random walk moves to any vertex uniformly at random, the problem reduces to the classical coupon collector's problem. (E.g., How many packs of Pokemon cards[1] should you buy in order to collect the full set of Pokemon characters?) It is well know that the expected time to cover a graph with m vertices is mH_m, where $H_k := \sum_{i=1}^{k} i^{-1}$ is the kth Harmonic number. The generalized coupon collector's problem asks for the time to cover the complete graph when the transition probabilities (weights) for the underlying random walk are not constant (see the survey [2] for a lot of applications). There are classical results established either by combinatorics (inclusion/exclusion principles [9], shuffle product [5]) or probability (martingales [8], tails estimates [7]). We also give some references for the coupon collector problem for arbitrary graphs [4]; some articles deal with asymptotic considerations [3].

Our problem (which involves two levels of randomness!) seems quite new and of course allows us to rederive/improve previous results (in the peculiar case p=1). When all the p_i's are equal (say, to p), a nice but finally predictable result is that $pE(T(p))$ is a rational number (as usual for Markovian process with rational transition probabilities). Of course, in principle all the questions concerning the random walk can be addressed by solving appropriate linear equations using the transition matrix for the random walk Markov chain but such an approach quickly becomes infeasible for even small values of m. On the complete graph, in the case of equal marking probabilities ($p_i = p$ for all i), and of probabilities π_i for the coupons, a natural conjecture is that $E[T(p)] \sim (m/p)H_m$ when $p \to 0$. The reason is because for small p, one would expect the time to mark a particular vertex to be Geometric with parameter p (and thus mean $1/p$). A quick (but erroneous) probabilistic reasoning consists in seeing the marking probability p as a change of time-scale and thus one gets (for any graph) $E(T(p)) = E(T(1))/p$. This is false (even asymptocally)! However, we will show this intuition to be true for regular graphs and prove a stronger result: For an arbitrary graph G, one has $pE(T(p)) \to K$ as $p \to 0$, where K is the expected cover time for the generalized coupon collector's problem where the set of weights is the stationary distribution for the random walk. This result will be proven in Section 4.

Notation 1 (Graph) *Through all the paper, G is a directed connected graph with vertices $v_1, \ldots, v_m$ and with a transition matrix A (that is, a_{ij} is the probability of a transition from v_i to v_j). The stationary distribution of this graph is noted $\pi_1, \ldots, \pi_m$. The probability to mark the vertex v_i whenever it is visited is*

[1] In the USA, kids have a great craze for these cards. They trade them and stare at them a lot... :-)

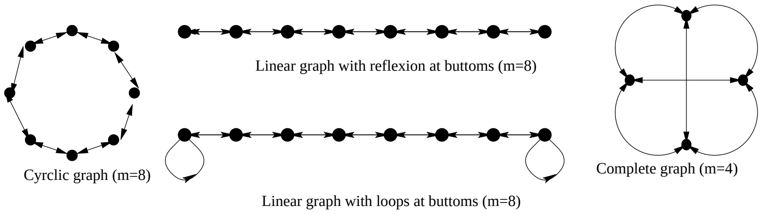

Fig. 1. Some of the oriented connected graphs considered in our examples.

noted p_i ($i = 1, \ldots, m$), and one sets $q_i := 1 - p_i$. When all the p_i's are equal, one simply notes p the probability of marking a vertex. $\widetilde{G}$ is a complete graph with transition matrix $\widetilde{A}$ (related in some sense to A, as explained later).

Notation 2 (Random variables) *T (resp. $\widetilde{T}$) is the "waiting time" random variable that represents the first time when all the vertices of the graph G (resp. $\widetilde{G}$) have been marked. The random variable X_i (resp. $\widetilde{X}_i$) gives the first time the i-th vertex of the graph G (resp. $\widetilde{G}$) is marked.*

Notation 3 (Operators) *One notes $[z^n]S$ the coefficient of z^n in a given series S. For any subset α of $\{1, \ldots, m\}$, $|\alpha|$ stands for the number of elements in α and one notes*

ζ_α the substitution $u_i \leftarrow 0$ for $i \in \alpha$ and u_i stays unchanged for $i \notin \alpha$,
σ_α the substitution $u_i \leftarrow q_i$ for $i \in \alpha$ and $u_i \leftarrow 1$ for $i \notin \alpha$.

The same notation will be used with respect to any other set of formal variables (for example with $t_1, \ldots, t_m$ instead of $u_1, \ldots, u_m$). We will also denote the identity matrix as Id *and U as the diagonal matrix with diagonal elements $u_1, \ldots, u_m$.*

Examples. To illustrate the above notations: $[z^2]\,(1 + 4z^2 + z^3) = 4$. For $m = 4$ and $\alpha = \{1, 2, 4\}$, one has $\zeta_\alpha\,(3u_1 + u_1u_2 + u_3^3 + u_4) = u_3^3$ and $\sigma_\alpha\,(3u_1 + u_1u_2 + u_3^3 + u_4) = 3q_1 + q_1q_2 + 1 + q_4$.

In the sequel, all the graphs considered have m vertices ($m > 1$) and are directed and irreducible: there is a sequence of connected edges linking any pair of vertices. First, we establish a combinatorial formula for the marking time in section 2. One will explain in section 3 a probabilistic intuition which allows us to simplify the problem and thus to prove the Supper Conjecture in Section 4, another proof is given in Section 5. The last sections deal with peculiar cases.

2 Marking Time on any Graph

We first give a generating function-based formula (based on finite differences) for $E(T(p))$, the expected *marking time*. For the *cover time* on the *complete* graph, another approach can be found in [5]. We refer the reader to Figure 4 for an example of our approach on two graphs of size 3.

Theorem 1 (General formula for average marking time) *The average time for marking all the vertices of an arbitrary graph G is given by*

$$E(T(\{p_1,\ldots,p_m\})) = \sum_{\alpha\neq\emptyset} (-1)^{|\alpha|+1} \sigma_\alpha \sum_{i=1}^{m} u_1 (\mathrm{Id} - AU)^{-1}_{1i},$$

where A, U, the p_i's, α and the σ_α's are defined as in Notation 1 and 3.

Proof. For the transition matrix A, the entry $A^n_{i,j}$ of A^n, gives the probability of moving from vertex v_i to v_j in n steps. With the matrix U defined as in Notation 3, the coefficient of the monomial $u_1^{k_1}\cdots u_m^{k_m}$ in $(AU)^n_{i,j}$ gives the probability that the random walk moves from v_i to v_j in n steps such that vertex v_t is visited k_t times, $t = 1,\ldots,m$. Thus the probability generating function for the walks on the graph (beginning in v_1), where z encodes the length of the walk and the u_i's encode the number of times the walk visits v_i is

$$F(z,u_1,\ldots,u_m) = u_1 \sum_{i=1}^{m}\sum_{k=0}^{\infty} z^k (AU)^k_{1,i} = u_1 \sum_{i=1}^{m} (\mathrm{Id} - zAU)^{-1}_{1,i}.$$

Recall that p_i (respectively, $q_i := 1 - p_i$) is the probability to mark (respectively, not to mark) the vertex v_i.

Taking into account the fact that the walk has visited all the vertices marking them at least once leads to the substitution $u_i^n \leftarrow 1 - q_i^n$. This justifies the introduction of the difference operator $\Delta_i f(u_i) := f(1) - f(q_i)$. So $F^+ := \Delta_1\Delta_2\ldots\Delta_m F$ gives the probability generating function of the walks that marked *all* the vertices and one has in fact

$$F^+(z) = \sum_{\alpha\subseteq\{1,\ldots,m\}} (-1)^{|\alpha|} \sigma_\alpha F(z,u_1,\ldots,u_m),$$

i.e., F^+ is the sum of F evaluated the set σ_α (defined in Notation 3). Therefore the probability generating function for $T(\{p_1,\ldots,p_m\})$ is $(1-z)F^+(z)$, so

$$E(T(\{p_1,\ldots,p_m\})) = \frac{\partial}{\partial z}_{|z=1} (1-z)F^+(z).$$

A change of variable $1 - z = t$ and a local development in $t = 0$ gives

$$\frac{\partial}{\partial z}_{|z=1} \sigma_\alpha \left(\frac{(1-z)u_1}{\mathrm{Id} - zAU}\right)_{ij} = -\sigma_\alpha \left(\mathrm{Id} - AU\right)^{-1}_{ij}$$

Note that $\mathrm{Id} - A$ is never invertible (whereas $\mathrm{Id} - zA$ is always invertible), so one has to deal apart with the substitution $\sigma_\emptyset$. □

Note that setting the p_i's to 1 gives a formula for the coupon collector problem.

3 The Probabilistic Intuition

We restrict here the discussion to the case when all the p_i's are small (equivalently, you can think p small with $p := \max p_i$). One argues that the random walk behaves, for small values of p, like a walk on the complete graph (*with loops*). We obtain a new Markov chain that is "equivalent" to the original Markov Chain, in the sense that they both have the same (limiting) stationary distributions.

Let $\{\pi_i\}_{i=1}^m$ be the stationary distribution for the random walk on G. Note that π_i gives the probability that the walk is in vertex v_i in stationarity (i.e., after a "long time"). When p is small, the time to mark the graph is high. And the proportion of time n that the walk is in vertex v_i will be $\pi_i n + O(\sqrt{n})$ for large n, with probability close to 1. (This can be shown, for instance, by a central limit theorem for Markov chains, or from results on large deviations.)

Thus, for "long enough" walks over G (when p is small, all walks are "long enough"), the average length of time to mark all the vertices should be very nearly the average length of time to mark the vertices of the complete graph (*with loops*), but with transition probabilities corresponding to the stationary distribution for the original graph. Let $\widetilde{G}$ denote the complete graph on m vertices with loops. Define its transition matrix $\widetilde{A}$ with the transition probabilities $\widetilde{a}_{i,j} := \pi_j$ (for $i, j = 1, \ldots, m$). Consider a random walk process for the general marking problem that begins in the vertex v_i with probability π_i. For such a process define $\widetilde{X}_k$ to be the first time that vertex $\widetilde{v}_k$ is marked. Then the $\widetilde{X}_k$ are geometric random variables with parameter $\pi_k p_k$ and $P(\widetilde{X}_k = n) = \pi_k p_k (1 - \pi_k p_k)^{n-1}$, but there are not independent (as $\widetilde{X}_i \neq \widetilde{X}_j$ for $i \neq j$).

Observe that $T = \max(X_1, \ldots, X_m)$ and $\widetilde{T} = \max(\widetilde{X}_1, \ldots, \widetilde{X}_m)$. Consider now a new process on the graph. Instead of one "random walker," consider m particles at each of the vertices all moving simultaneously to neighboring vertices according to the same transition mechanism, but independently of each other. If we let Y_i (and $\widetilde{Y}_i$) denote the first time that vertex v_i is marked then observe that the Y_i has the same distribution as X_i but the Y_i's are independent and the X_i's are not. Define now $Z = \max(Y_1, \ldots, Y_m)$ and $\widetilde{Z} = \max(\widetilde{Y}_1, \ldots, \widetilde{Y}_m)$. Our intuition is as follows: As the $\widetilde{Y}_i' s$ and the Y_i's will behave similarly for small p, one should have that $E(\widetilde{Z}) \approx E(Z)$. As simultaneous markings (for the Y process) occur with probability $O(p^2)$, one has that $E(Z) \approx E(T)$. Also $E(\widetilde{Z}) \approx E(\widetilde{T})$ and thus $E(T) \approx E(\widetilde{T})$. Thus the study of the expected marking time on the graph $\widetilde{G}$ should answer our question as to the expected marking time on the graph G. We make the above rigorous in the next section.

4 Algebraic and Combinatorial Proof

Theorem 2 (Closed form formula with stationary distribution) *Let $\widetilde{G}$ be a graph with transition probabilities $\widetilde{a}_{i,j} := \pi_j$. The expected time $E(\widetilde{T})$ for marking the whole graph $\widetilde{G}$ is $1/p$ times the time needed for visiting all the vertices*

$$E(\widetilde{T}) = \frac{1}{p} E(\textit{coupon collector on } \widetilde{G}),$$

and an inclusion-exclusion formula holds

$$E(\widetilde{T}) = \frac{1}{p} \sum_{\alpha \neq \emptyset} \frac{(-1)^{m-|\alpha|}}{-\zeta_\alpha(\pi_1 + \pi_2 + \ldots + \pi_m)},$$

where the ζ_α's are defined as in Notation 3.

Proof. Consider the $\pi_i' s$, the q_i's, and the p_i's as formal variables, then

$$\text{Prob}(\widetilde{T} \leq n) = \{p_1^{>0}\} \ldots \{p_m^{>0}\}(\pi_1 p_1 + \pi_1 q_1 + \ldots + \pi_m p_m + \pi_m q_m)^n,$$

where $\{p_1^{>0} S\}$ (with $S \in \mathbb{R}[[p_1, \ldots, p_m, q_1, \ldots, q_m]]$) stands for the sum of monomials in p_1 of positive exponent in S. Thus, we have

$$\text{Prob}(\widetilde{T} \leq n) = \sum_{\alpha \subseteq \{1,\ldots,m\}} (-1)^{|\alpha|} (\zeta_\alpha(\pi_1 p_1 + \pi_1 q_1 + \ldots + \pi_m p_m + \pi_m q_m))^n,$$

where the substitutions ζ_α are defined as in Notation 3. Multiplying by z^n and summing for $n \geq 0$, when all the p_i's are equal to p, with the substitutions ζ_α now taken to act on the π_i's, leads to

$$\sum \text{Prob}(\widetilde{T} \leq n) z^n = \sum_{\alpha \subseteq \{1,\ldots,m\}} \frac{(-1)^{|\alpha|}}{1 - z(q + p\zeta_\alpha(\pi_1 + \pi_2 + \ldots + \pi_m))}.$$

Multiplying by $1 - z$ and differentiating in $z = 1$ gives the expected time

$$E(\widetilde{T}) = \sum_{\alpha \neq \emptyset} (-1)^{|\alpha|} \frac{1}{p(-1 + \zeta_\alpha(\pi_1 + \pi_2 + \ldots + \pi_m))}.$$

□

	graph G	graph $\widetilde{G}$
transition matrix	$A = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$	$\widetilde{A} = \begin{pmatrix} 3/8 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/4 \end{pmatrix}$
stationary distribution	$(3/8, 3/8, 1/4)$	$(3/8, 3/8, 1/4)$
	$U = \begin{pmatrix} u_1 & 0 & 0 \\ 0 & u_2 & 0 \\ 0 & 0 & u_3 \end{pmatrix}$	$U = \begin{pmatrix} u_1 & 0 & 0 \\ 0 & u_2 & 0 \\ 0 & 0 & u_3 \end{pmatrix}$
generating matrix of the walks	$(\text{Id} - zAU)^{-1}$	$(\text{Id} - z\widetilde{A}U)^{-1}$
expected covering time	$95/12 \simeq 7.91$	$29/5 \simeq 5.80$
expected marking time (closed form)	$\dfrac{435 + 494p + 187p^2 + 24p^3}{75p + 60p^2 + 9p^3}$	$29/5p^{-1}$
expected marking time (asymptotics)	$E(T(p)) = 29/5p^{-1} + O(1)$	$E(\widetilde{T}(p)) = 29/5p^{-1}$

Fig. 2. Example of our approach on a graph of size 3. Most of the results in the literature are about the (expected) covering time of the $\widetilde{G}$-column (and other higher moments). In our situation (the G-column), the lack of independence is the main difficulty. Our paper shows that in order to get the last entries in the G-column, one can proceed as follow: Consider the first 2 entries of the G-column, then deal with $\widetilde{G}$-column whose last entries gives the wanted result (for G).

Proposition 1 *The marking time T can be approximated by $\widetilde{T}$:*

$$E(\widetilde{T}) - E(T) = O(1), \qquad (p \to 0).$$

Proof. In order to prove that

$$\sum_{\alpha\neq\emptyset}(-1)^{|\alpha|+1}\sigma_\alpha\sum_{i=1}^{m}u_1(\mathrm{Id}-\widetilde{A}U)^{-1}_{1i} - \sum_{\alpha\neq\emptyset}(-1)^{|\alpha|+1}\sigma_\alpha\sum_{i=1}^{m}u_1(\mathrm{Id}-AU)^{-1}_{1i} = O(1)$$

where the substitutions are defined in Notation 3, it is sufficient to check that

$$[p^{-1}]\sigma_\alpha\left(\mathrm{Id}-\widetilde{A}U\right)^{-1}_{ij} - [p^{-1}]\sigma_\alpha\left(\mathrm{Id}-AU\right)^{-1}_{ij} = O(1). \tag{1}$$

In fact, one has $(1-zA)^{-1} = \frac{\widetilde{A}}{1-z} + O(1)$. This is established by writing $A = P + R$, which decomposes A as an eigenprojection on the eigensubspace related to its eigenvector π, plus the projection R on the supplementary subspace (as it is well known $P^n = P = \widetilde{A}$ and $PR = RP = 0$), thus $\sum z^n(P+R)^n = \frac{\widetilde{A}}{1-z} + (\mathrm{Id}-zR)^{-1}$ where $(\mathrm{Id}-zR)^{-1}$ is actually regular in $z=1$ (as, by Perron–Frobenius theory, 1 is an eigenvalue of multiplicity 1), so $(1-zR)^{-1} = O(1)$.

So this sets the case of the equality (1) when $\alpha = \{1,\dots,m\}$ (simply consider $z = 1-p$). Theq other cases appear as a perturbation of two noninvertible matrices (namely $\mathrm{Id}-A$ and $\mathrm{Id}-\widetilde{A}$), which gives

$$(\mathrm{Id}-A+\epsilon B)^{-1} = \frac{\widetilde{A}}{-\epsilon\lambda'(0)} + O(1) = (\mathrm{Id}-\widetilde{A}+\epsilon\widetilde{B})^{-1}$$

where B is A with its i-th column set to 0 whenever $i \in \alpha$ (so $A - \epsilon B = \sigma_\alpha AU$), and where $\lambda(\epsilon)$ is the perturbation of the eigenvalue 1, so $\lambda'(0) = \sum_{i\in\alpha}\pi_i$. See [6] for the analicity of the perturbations of projections, eigenvalues, etc. □

For people with probabilistic affinities, we give below another approach which also proves that $E(T) = E(\widetilde{T}) + O(1)$ as $p \to 0$.

5 Analytical and Probabilistic Proof

The following proposition (identical to Proposition 1) is the key point

Proposition 2 (Equivalence of expected times) *The marking time for the graph G and $\widetilde{G}$ have the same first order asymptotics, namely*

$$E(T) = E(\widetilde{T}) + O(1).$$

Proof. The main idea is the following

$$T \approx (T \text{ without its tails}) \approx (\widetilde{T} \text{ without its tails }) \approx \widetilde{T},$$

where $\approx$ means here that the expectations have the same first order asymptotics as $p \to 0$. Note that

$$P(T \leq n) = \sum_{\lambda=(n_1,\dots,n_m)} p_\lambda c_\lambda,$$

where the sum is over all nonnegative m-tuples $\lambda = (n_1, \dots, n_m)$ such that $\sum_k n_k = n$, p_λ is the probability that the walk takes n steps and visits each vertex v_k n_k times, $k = 1, \dots, m$, and c_λ is the probability that all vertices have been marked by such a tour, that is $c_\lambda = \prod_{j=1}^m (1-(1-p)^{n_j})$. Define the "central interval" I as

$$I := \left[\frac{1}{p|\ln p|}, \frac{|\ln p|}{p}\right].$$

Further define the "multidimensional box" B as

$$B := \prod_{i=1}^{m} [n\pi_i - \sqrt{n}\ln n, n\pi_i + \sqrt{n}\ln n].$$

For small p and for large n, by a classic result in large deviation theory, one has

$$n\pi_j - \sqrt{n}\ln n < n_j < n\pi_j + \sqrt{n}\ln n$$

(that is, $\lambda \in B$) with probability $1 - \exp(-c\ln^2 n)$ (with $c > 0$, see [1]). As $\mathrm{Prob}(T = n)$ is the probability that a success occurs exactly at the n-th step, one has

$$E(T) = \sum_{n \geq 0} (1 - \mathrm{Prob}(T \leq n)).$$

This sum can be split as follows (with a lot of abusive but natural notations!)

$$E(T) = \sum_{n<I} + \sum_{n\in I, \lambda\in B} + \sum_{n\in I, \lambda\notin B} + \sum_{n>I, \lambda\in B} + \sum_{n>I, \lambda\notin B} \tag{2}$$

where, for example, $n < I$ means for n before the "central interval" I. The first sum is bounded by the length of the interval of summation, which is $o(1/p)$ as $p \to 0$, the sums for $\lambda \notin B$ are bounded by $\exp(-c\ln^2 n)$ and so is the remaining sum for $n > I$.

One now focuses on the sum over I and B. By a limit theorem on Markov Chains [1], p_λ follows a multidimensional Gaussian law $g(\lambda)$, thus

$$E(T) = \sum_{n\in I} \left(1 - \sum_{\lambda\in B} g(\lambda)c_\lambda\right) + o(p^{-1}). \tag{3}$$

The same scheme can be applied to $\widetilde{G}$, with the same central interval I and box B since the two random walks have the same stationary distribution. Thus $E(T) = E(\widetilde{T}) + o(p^{-1}) = E(\widetilde{T}) + O(1)$, since $E(\widetilde{T})$ and $E(T)$ are *rational fractions* in p (see Theorem 1 and 2). $\square$

The Theorem below is the main result of the paper and answers precisely to the Supper Conjecture. The following "integral formula" is well known (cf. [5], where it is established for the complete graph by means of shuffle products and Laplace transform), but we give an alternate derivation here, valid for any graph.

Theorem 3 (First order asymptotics. Integral form) *For any graph G with a stationary distribution $(\pi_1, \dots, \pi_m)$, the expected marking time is*

$$E(T) = \frac{K}{p} + O(1), \qquad \textit{where} \quad K = \int_0^\infty (1 - \prod_{j=1}^m (1 - \exp(-\pi_j x)))\, dx.$$

Proof. With the same notations as above, the starting point is

$$c_\lambda = c_{(n_1,\dots,n_m)} = \prod_{j=1}^m (1 - (1-p)^{n_j}) = \prod_{j=1}^m (1 - e^{-p' n_j}),$$

where $p' := -\ln(1-p)$. Thus

$$E(T) = \sum_{n \in I} \left(1 - \sum_{\lambda \in B} g(\lambda) \prod_{j=1}^m 1 - \exp(-p' n_j \pi_j) \right) + O(1).$$

In the box B, $n_j = n\pi_j + \epsilon_j \sqrt{n} \ln n$ (where $|\epsilon_j| < 1$), so

$$c_\lambda = \prod_{j=1}^m (1 - \exp(-p' n \pi_j - p' \epsilon_j \sqrt{n} \ln n)), \text{ and thus}$$

$$E(T) = \sum_{n \in I} \left(1 - \sum_{\lambda \in B} g(\lambda) \prod_{j=1}^m (1 - \exp(-p' n \pi_j - p' \epsilon_j \sqrt{n} \ln n)) \right) + O(1).$$

Now, majoring $\sum g(\lambda)$ by $1 - \exp(-c \ln^2 n)$ yields

$$E(T) = \sum_{n \in I} \left(1 - \prod_{j=1}^m 1 - \exp(-p' n \pi_j - p' \epsilon_j \sqrt{n} \ln n) \right) + O(1) + o(p^{-1}).$$

The sums being for $p^{-1} |\ln p| \le n \le p^{-1} |\ln p|$, this leads to

$$\sum \exp(-p'(O(n))(1 - \exp(-p' o(n))) \le \sum (1 - \exp(-p' o(n))$$
$$\le \sum p' o(n) \le p' p^{-1} |\ln p| o(p^{-1} |\ln p|) = o(p^{-1}),$$

and completing the tails gives

$$E(T) = \sum_{n \in I} \left(1 - \prod_{j=1}^m 1 - \exp(-p' n \pi_j) \right) + O(1) + o(p^{-1}).$$
$$= \sum_{n \ge 0} (1 - \prod_{j=1}^m (1 - \exp(-p' n \pi_j))) + O(1) + o(p^{-1}). \tag{4}$$

Indeed the first introduced part is $\le p^{-1}/|\ln p| = o(p^{-1})$ and the last introduced part is $\le \sum_{n > p^{-1} |\ln p|} 2^m \exp(-p' m \min(\pi_i))^n = o(p^{-1})$.

Set $f(x) := 1 - \prod_{j=1}^m (1 - \exp(-p' x \pi_j)$. As $f'(x)$ has a fast enough decay to 0 near ∞, the Euler-Maclaurin formula gives:

$$\sum_{n=0}^{\infty} f(n) - \int_0^{\infty} f(x)dx = \frac{f(0)+f(\infty)}{2} + \int_0^{\infty} (x - \lfloor x \rfloor - 1/2) f'(x) dx = O(1).$$

Applying this to the formula (4) leads to

$$E(T) = \frac{1}{p'} \int_0^{\infty} (1 - \prod_{j=1}^{m} (1 - \exp(-\pi_j x)))\, dx + O(1) \text{ when } p' \to 0.$$

Since $p' = p + o(p)$, one has $E(T) = K/p + O(1)$. □

The "integral formula" allows us to compute $E(T)$ to any precision in linear time (as all the integrated functions have a nice behavior), whereas the formulae of Theorem 1 and 2 are impracticable because they comprise 2^m summands.

6 What about Balanced or Regular Graphs?

The most common model for random walk on a graph (the unweighted case) is to suppose that at a particular vertex the walk moves with probability proportional to the degree of that vertex. It is well known that the stationary distribution for such a walk is related to the degree of the vertices. Call a graph "balanced" if the outdegree of each vertex is equal to its indegree and if each outgoing edge is equally likely. For balanced graphs, there is a simple relation between the stationary distribution (the left eigenvector associated to eigenvalue 1 of the transition matrix of the graph) and the degrees of edges:

Proposition 3 (Stationary distribution for balanced graphs) *For balanced graphs, one has $\pi_i = \frac{N_i}{\sum_1^m N_i}$ (N_i is the number of incoming edges of vertex v_i).*

Proof. Let N_{ij} be the number of edges from v_i to v_j, and note N_{*j} the number of incoming edges to v_j and N_{i*} the number of outgoing edges from v_i. The left eigenvector (for eigenvalue 1) satisfies

$$(\pi_1, \pi_2, \ldots, \pi_m) \begin{pmatrix} N_{11}/N_{1*} & \ldots & N_{1m}/N_{1*} \\ \vdots & \vdots & \vdots \\ N_{m1}/N_{m*} & \ldots & N_{mm}/N_{m*} \end{pmatrix} = (\pi_1, \pi_2, \ldots, \pi_m);$$

hence $\sum_i \pi_i \frac{N_{ij}}{N_{i*}} = \pi_j$. This equation is indeed satisfied by $\pi_i := \frac{N_{*i}}{\sum_1^m N_{*i}}$ when $N_{i*} = N_{*i}$. □

Theorem 2 rewrites in these cases:

Corollary 1 (Balanced graphs) *For balanced graphs with $N := \sum N_i$ edges, one has*

$$E(T) = \frac{N}{p} \sum_{\alpha \neq \emptyset} \frac{(-1)^{m-|\alpha|}}{-\zeta_\alpha(N_1 + N_2 + \ldots + N_m)} + O(1),$$

where the substitutions ζ_α operate on the $\{N_j\}$.

Proof. Direct consequence of Theorem 2 and Proposition 3. □

There are several classes of graphs for which the formula can be simplified. The more interesting one, the class of regular graphs (graphs whose *all* vertices have the same number of incoming and outgoing edges), is the object of the following corollary.

Corollary 2 (Regular graphs) *When all the π_i's are equal (in particular if G is a regular graph and outgoing edges are chosen uniformly at random), one has $E(T) = \frac{mH_m}{p} + O(1)$.*

Proof.

$$E(\widetilde{T}) = \sum_{i=1}^{m} (-1)^i \binom{m}{i} \frac{1}{p(-1 + \frac{m-i}{m})} = \frac{m}{p} \sum_{i=1}^{m} (-1)^{i+1} \binom{m}{i} \frac{1}{i} = \frac{mH_m}{p}.$$

The last equality follows as iteration of forward finite difference operators and Euler's transform. Equivalently, one could use Theorem 3. Here the integral becomes $\int_0^\infty (1 - \prod_{j=1}^m (1 - \exp(-x/m)))\, dx$, which simplifies to mH_m. A third proof of this formula is probabilistic and considers the maximum of i.i.d. geometric random variables. Thus when $\pi_i = 1/m$ (for $i = 1, \dots, m$), one has

$$\begin{aligned} \mathrm{Prob}(\widetilde{Z} = n) &= \sum_i \mathrm{Prob}(\widetilde{X}_i = n) \prod_{j \neq i} \mathrm{Prob}(\widetilde{X}_j < n) \\ &= m \left(p/m(1-p/m)^{n-1}\right) (1 - (1-p/m)^{n-1})^{m-1}. \end{aligned}$$

Since $E(\widetilde{T}) = \sum_{n \geq 1} n \mathrm{Prob}(\widetilde{T} = n)$, one is interested by the coefficient of the lowest term in the Laurent development at $p = 0$ of

$$\sum_{n \geq 1} nz(1 - \frac{z}{m})^{n-1} (1 - (1 - \frac{z}{m})^{n-1})^{m-1} = \sum_{k=0}^{m-1} (-1)^k \binom{k}{m-1} \frac{z}{(1 - (1 - z/m)^{k+1})^2}.$$

As $(1 - (1 - z/m)^{k+1})^2 = O(z^2)$ as $z \to 0$, the valuation of the development in Laurent series of $z\left(1 - (1 - z/m)^{k+1}\right)^{-2}$ is -1. Thus the coefficient of the lowest term of $F_m(z)$ is the sum of the residues of the rational fractions, then

$$\mathrm{Res}(z = 0) = m \sum_{k=0}^{m-1} (-1)^k \binom{k}{m-1} \frac{m}{(k+1)^2} = mH_m.$$

7 Examples

The formula for the expected time in the coupon collector problem reduces to $\frac{m^2 - m + 2}{2}$ for the cyclic graph C_m, to $m^2 - 2m + 2$ for the line graph L_m (with reflection) and to $(m-1)H_{m-1} + 1$ for the complete graph without loops K_m.

For L_m, there is an equivalence between the coupon collector problem and the random walks (with jumps $+1, -1$) from 0 to m of height $\leq m$, the last ones being linked to continued fraction theory and hence to a quotient of Chebyshev polynomials.

We apply below the formula of Theorem 1 for all nonisomorphic unoriented connected graphs without multiplicity having at most 4 vertices. We give for each graph, the stationary distribution, the average time for the classical coupon collector problem and, in the last column, the coefficient of p^{-1} (the leading coefficient) in the asymptotics of the average time for marking the whole graph:

size	graph	stationary distributions	expected cover time	expected marking time
$m=1$	K_1	(1)	1	1
$m=2$	K_2	(1/2, 1/2)	2	3
$m=3$	K_3	(1/3, 1/3, 1/3)	4	$3H_3$=11/2
$m=3$	L_3	(1/4, 1/2, 1/4)	5	19/3
$m=4$	L_4	(1/6, 1/3, 1/3, 1/6)	10	99/10
$m=4$	K_4	(1/4, 1/4, 1/4, 1/4)	13/2	$4H_4$=25/3
$m=4$	C_4	(1/4, 1/4, 1/4, 1/4)	7	$4H_4$=25/3
$m=4$	Y_4	(1/5, 2/5, 1/5, 1/5)	10	110/10
$m=4$	T_4	(1/8, 3/8, 1/4, 1/4)	163/15	369/35
$m=4$	Q_4	(3/10, 1/5, 1/5, 3/10)	69/10	62/7

The graphs K_m, L_m and C_m are defined as above whereas T_4 is the triangle with a tail, Y_4 is the Y shape, Q_4 is the square with one diagonal. We recall that, as seen in previous sections, for the complete graph with loops, the coupon collector time is mH_m and the marking time is mH_mp^{-1}. This is finally the behavior of all regular graphs, those for which the Supper Conjecture holds.

Acknowledgement. The first of the two authors would like to thank Philippe Flajolet for a lot of helpful remarks and does not resist quoting one of his pseudo-reformulations of our problem "You are a PhD student and you attempt to prove a theorem; your chance of success at each attempt is p (p very close to 0, of course). Let T be the time it takes you to get the theorem right..."

References

1. Aldous (David J.) and Fill (James). – *Reversible Markov Chains and Random Walks on Graphs.* – Book in preparation. Available at http://www.stat.berkeley.edu/users/aldous/book.html.
2. Boneh (Arnon) and Hofri (Micha). – The coupon-collector problem revisited. *Comm. Statist. Stochastic Models*, vol. 13, n° 1, 1997, pp. 39–66.
3. Csörgő (Sándor). – A rate of convergence for coupon collectors. *Acta Sci. Math. (Szeged)*, vol. 57, n° 1-4, 1993, pp. 337–351.
4. Feige (Uriel). – Collecting coupons on trees, and the cover time of random walks. *Comput. Complexity*, vol. 6, n° 4, 1996/97, pp. 341–356.
5. Flajolet (Philippe), Gardy (Danièle), and Thimonier (Loÿs). – Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, vol. 39, n° 3, 1992, pp. 207–229.
6. Kato (Tosio). – *Perturbation theory for linear operators.* – Springer-Verlag, 1995.
7. Nath (Harmindar B.). – Waiting time in the coupon-collector's problem. *Austral. J. Statist.*, vol. 15, 1973, pp. 132–135.
8. Sen (Pranab Kumar). – Invariance principles for the coupon collector's problem: a martingale approach. *Ann. Statist.*, vol. 7, n° 2, 1979, pp. 372–380.
9. von Schelling (Hermann). – Coupon collecting for unequal probabilities. *Amer. Math. Monthly*, vol. 61, 1954, pp. 306–311.