

# Cover Times of Random Walks

R. Teal Witter

September 2, 2019

## **Abstract**

The cover time of a graph is the number of steps a discrete random walk takes to visit every vertex at least once. We begin by investigating the cover time of small and structured graphs but quickly find that our precise strategy fails for larger, unstructured ones. To bound the cover time on arbitrary graphs, we apply more general probabilistic arguments and an important electrical network identity. Finally, we describe the limiting distribution of the cover time of complete graphs with self-edges and apply our analysis to an intuitive example.

# Contents

<b>1</b>	<b>Introduction and Background</b>	<b>3</b>
1.1	Definitions . . . . .	3
1.2	Electrical Networks . . . . .	4
1.3	Outline . . . . .	5
<b>2</b>	<b>Cover Times of Structured Graphs</b>	<b>5</b>
2.1	Small Example . . . . .	5
2.2	Complete Graph . . . . .	6
2.3	Line Graph . . . . .	8
2.4	Simple Cycle . . . . .	10
<b>3</b>	<b>Cover Time Bounds</b>	<b>11</b>
3.1	Commute Time Identity . . . . .	11
3.2	Matthews Method . . . . .	12
3.3	Matthews Method on Binary Trees . . . . .	14
3.4	Spanning Tree Argument . . . . .	16
<b>4</b>	<b>Cover Time Distributions</b>	<b>17</b>
4.1	Complete Graph . . . . .	17
4.2	Birthday Problem . . . . .	20
<b>5</b>	<b>Appendix</b>	<b>23</b>
5.1	Birthday Problem R Code . . . . .	23

# 1 Introduction and Background

To introduce the concept of a cover time, we begin by describing two intuitive examples. Consider a coupon collector who receives one of  $r$  distinct coupons in the mail everyday. For example if  $r = 3$ , the collector may receive coupon 1 today, coupon 3 tomorrow, coupon 1 (again) the next day, and finally coupon 2 the following day. For arbitrary  $r$ , we ask how many days until our collector has collected at least one of every coupon? (To solve the coupon collector problem we will frame it in terms of the complete graph with self-edges.)

Next consider a particular case of coupon collecting called the birthday problem. Assuming all 365 days of the year, excluding February 29, are equally likely, how many people would we need to ensure each birthday is represented by at least one person? A lower bound is certainly 365 but it is possible to get an arbitrarily large population without ‘covering’ every birthday.

Other applications of cover times include universal travel sequences, graph connectivity algorithms, and protocol testing.

In this paper, we use a variety of strategies like first step analysis and draw on several related topics such as electrical networks to rigorously investigate cover times.

## 1.1 Definitions

We consider an undirected, connected graph  $G$  without multi-edges. Let  $V(G)$  be the set of vertices and let  $E(G)$  be the set of edges on  $G$ . Then  $n$  is the number of vertices  $|V(G)|$ . We say  $i \sim j$  for  $i$  and  $j$  in the vertex set  $V(G)$  if  $(i, j)$  is in  $E(G)$ . Assume that each vertex is accessible from every other vertex. That is, for  $x$  and  $y$  there exists a path  $x = v_0 \sim v_1 \sim \dots \sim v_n = y$ .

Let  $C_{i,j} = C_{j,i}$  be a non-negative weight assigned to the edge between vertices  $i$  and  $j$ . We will mostly consider unweighted graphs, meaning  $C_{i,j} = 1$  if  $i \sim j$  and  $C_{i,j} = 0$  otherwise.

Define the weight of vertex  $i$  as

$$C_i = \sum_{j:i \sim j} C_{i,j}.$$

In the unweighted case,  $C_i$  is simply the degree of vertex  $i$ .

We consider a random walk  $(X_t)$  on  $G$ . Call  $X_t$  the vertex that the walk is on at time  $t$ . The random walk begins at vertex  $X_0$ . At each step, the walk at vertex  $i$  moves to neighboring vertex  $j$  with probability  $C(i, j)/C(i)$ .

Let  $T_j$  be the number of steps until the first visit to  $j$  after time  $t = 0$ . Formally,

$$T_j = \min \{t > 0 : X_t = j\}.$$

Let  $E_i[T_j] = E[T_j | X_0 = i]$ . This is the expected time of the first visit to  $j$  given that the random walk started in vertex  $i$ . Then  $E_j[T_j]$  is the return time to  $j$  since  $t$  is strictly positive. Now define the hitting time  $t_{\text{hit}}$  for the graph be the maximum expected time between two vertices on  $G$ . Formally,

$$t_{\text{hit}} = \max_{i,j \in V(G)} E_i[T_j].$$

Let the cover time random variable  $\tau_{\text{cov}}$  be the first time all vertices have been visited by the random walk  $(X_t)$ . Formally,  $\tau_{\text{cov}}$  is the minimum time that for all  $i \in V(G)$  there exists  $t \leq \tau_{\text{cov}}$  such that  $X_t = i$ . We define the cover time as the expected value of  $\tau_{\text{cov}}$  from the worst initial vertex. Formally,

$$t_{\text{cov}} = \max_{j \in V(G)} E_j[\tau_{\text{cov}}].$$

## 1.2 Electrical Networks

There are many connections between electrical networks and random walks [3]. Physical properties like voltage, current, and resistance all have probabilistic interpretations. In this section, we explore electrical networks to build intuition about random walks.

So that the reader may more easily distinguish between vertices, consider  $x$  and  $y$  in  $V(G)$ . Recall that  $C_{x,y}$  is the weight of the edge between vertices  $x$  and  $y$ ,  $C_x$  is the sum of all  $C_{x,y}$  where  $x$  and  $y$  share an edge, and the probability  $P_{x,y}$  that we go from  $x$  to  $y$  is  $C_{x,y}/C_x$ . For the rest of this section, we will call  $C_{x,y}$  the conductance between  $x$  and  $y$ . Define the resistance  $R_{x,y}$  between  $x$  and  $y$  as  $1/C_{x,y}$ . (We will soon justify the electrical network interpretation of a random walk using Ohm's Law.)

We now consider the distribution of the random walk on the vertices of our graph. Define  $\pi_x$  as  $C_x/C$  where  $C = \sum_y C_y$ . We will show that  $\pi$ , the vector of  $\pi_i$  where  $i$  runs over all vertices, is the stationary distribution of the random walk. Then  $\pi_x$  must be the long term proportion of time spent in vertex  $x$ . We first use our definitions to show that  $\pi$  satisfies detailed balance,

$$C_x P_{x,y} = C_x \frac{C_{x,y}}{C_x} = C_{x,y} = C_{y,x} = C_y \frac{C_{y,x}}{C_y} = C_y P_{y,x}.$$

Dividing by  $C$  gives  $\pi_x P_{x,y} = \pi_y P_{y,x}$ , or detailed balance. We next use detailed balance to show that  $\pi$  is the stationary distribution. Summing over  $x$ ,

$$\sum_x \pi_x P_{x,y} = \sum_x \pi_y P_{y,x} = \pi_y \sum_x P_{y,x} = \pi_y.$$

(We get the last equality because we know that when leaving vertex  $y$  the random walk goes with probability 1 to some vertex  $x$ .) We rely on [6] to show that  $\pi P = \pi$  where  $P$  is the matrix of transition probabilities and how it follows that  $\pi$  is the stationary distribution of the random walk. Then  $\pi_x$  is indeed the long term proportion of time spent in vertex  $x$ .

We now introduce voltage, current, and effective resistance. Consider arbitrary  $a$  and  $b$  in  $V(G)$ . This is equivalent to setting the voltage to be 1 at  $a$  and 0 at  $b$ . We may think of voltage  $v_x$  as the probability starting from  $a$  we get to  $x$  before  $b$ . (We have not proved this yet but we will by the end of this subsection.) With this interpretation it makes sense that  $v_a$ , the probability we get to  $a$  before  $b$ , is 1 (we are already at  $a$ ) and  $v_b$ , the probability we get to  $b$  before  $a$ , is 0 (we are again already in  $a$ ). We do not prove this but the voltage at every vertex  $x$  in  $V(G) \setminus \{a, b\}$  is determined by the boundary values at  $a$  and  $b$ .

By Ohm's Law, the current between two vertices is the product of the difference in their voltages and the conductance of their edge. We may think of current as the net number of movements through an edge. Kirchoff's Current Law says that the sum

of current on any vertex is 0. Probabilistically, every time we enter a vertex we must also leave it. Using Kirchoff's Current Law and the definition of current, we justify the probabilistic interpretation of voltage as follows

$$\begin{aligned}\sum_y i_{x,y} &= 0 = \sum_y (v_x - v_y) C_{x,y} \\ v_x \sum_y C_{x,y} &= \sum_y C_{x,y} v_y \\ v_x &= \sum_y \frac{C_{x,y}}{C_x} v_y = \sum_y P_{x,y} v_y.\end{aligned}$$

It follows that  $v_x = \sum_y P_{x,y} v_y$  using total probability on the likelihood of reaching  $a$  before  $b$  from  $x$ .

### 1.3 Outline

Section 2 primarily relies on [2]. We see that while in principle it is always possible to exactly calculate the cover time of graph, it is in general very difficult to do so. For the remainder of this section, we investigate structured graphs for which exact cover time analysis is tractable: the complete graph, line graphs, and simple cycles.

Section 3 deals with the large or unstructured graphs for which we must use cover time bounds. We begin by establishing an important identity between commute times and effective resistance in Section 3.1. (Our electrical network analysis is an extension of [4] and we use [7] for the next two subsections.) We take our strategy for finding the cover time of the complete graph from Section 2.2 and apply it to general graphs in Section 3.2. In Section 3.3, we use both the commute time identity and the Matthews Method to bound the cover time of a binary tree. In Section 3.4, we follow [1] and use the idea of a spanning tree as well as the commute time identity to develop another bound for general graphs.

Up until Section 4, we only consider expected cover time. We now delve into the distribution of the cover time on a complete graph with self-edges. We draw on [5] and find that the cover time random variable approaches a Poisson distribution as the number of vertices goes to infinity. In Section 4.2, we apply Section 4.1 to the birthday problem. (The R code for our computations and Fig. 8 can be found in Section 5.1.)

## 2 Cover Times of Structured Graphs

### 2.1 Small Example

In principle, it is always possible to find the exact cover time of a graph using first step analysis. We demonstrate the strategy on a small graph suggested by [2].

Fig. 1 shows our small graph and the first step analysis for the random walk  $(X_t)$  started from the top right vertex. The black circle indicates the current vertex of the random walk and the empty circles indicate the visited vertices.

For a graph with  $n$  vertices, there are  $n$  possible choices for the current vertex and  $2^{n-1}$  possible combinations of visited and unvisited vertices for each other one. Of course, some cases are inaccessible by a random walk. But even on Fig. 1 where we

exploit symmetry, there are almost a prohibitive number. It is easy to see that the number of cases grows exponentially with the number of vertices.

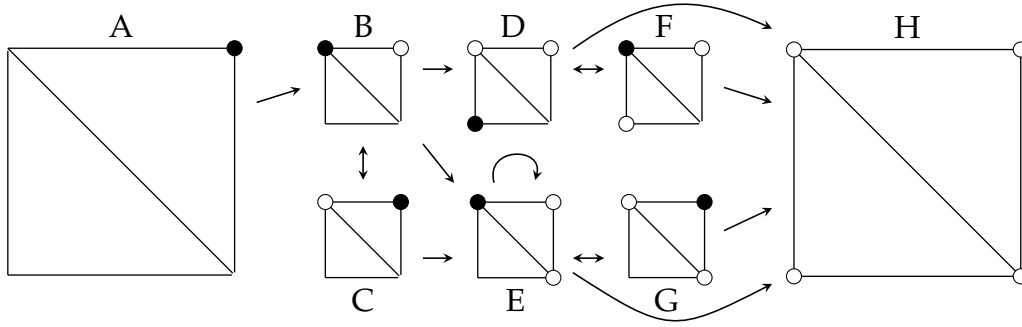


Figure 1: First step analysis applied to a small graph with  $n = 4$ .

If we let the label of each graph in Fig. 1 denote the cover time from that case, we have the following system of equations:

$$\begin{aligned}
 t_{\text{cov}} = A &= 1 + B & B &= 1 + \frac{1}{2}C + \frac{1}{2}D \\
 C &= 1 + \frac{1}{2}B + \frac{1}{2}E & D &= 1 + \frac{1}{2}F + \frac{1}{2}H \\
 E &= 1 + \frac{1}{3}B + \frac{1}{3}G + \frac{1}{3}H & F &= 1 + \frac{1}{2}D + \frac{1}{2}G \\
 G &= 1 + \frac{1}{2}E + \frac{1}{2}H & H &= 0
 \end{aligned}$$

With a little linear algebra, we find that  $t_{\text{cov}} = 43/5$ . We leave it as an exercise to the reader to verify that the time to visit all vertices from the other (i.e. symmetrically distinct) vertex is  $43/6$ .

For large graphs, the first step analysis as an approach to finding the expected cover time is obviously intractable. The rest of this section deals with particularly symmetric graphs for which it is possible to find a solution for an arbitrary number of vertices  $n$ . However, we must rely on cover time bounds for the vast majority of cases.

## 2.2 Complete Graph

We call the graph with an edge between each pair of vertices the complete graph. Note that a complete graph with  $n$  vertices has  $\binom{n}{2}$  edges. (We exclude self-edges.)

Our goal is to find the expected number of steps until we have visited all  $n$  vertices in the complete graph. Observe that the random walk can start at any vertex without loss of generality because the graph is symmetric. The strategy is to write the cover time  $t_{\text{cov}}$  in terms of the expected value of more simple random variables. Let  $X_i$  be the random variable that represents the number of steps to go from  $i - 1$  to  $i$  distinct vertices.

We write the cover time in terms of  $X_i$  and use linearity of expectation.

$$t_{\text{cov}} = E[X_1 + X_2 + X_3 + \cdots + X_n] = \sum_{i=1}^n E[X_i]$$

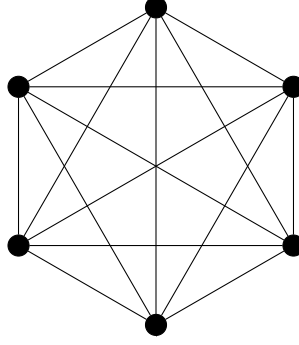


Figure 2: The complete graph with six vertices (and 15 edges).

The random variable  $X_i$  takes value  $k$  when  $k - 1$  steps lead us to already visited vertices and the  $k^{\text{th}}$  step is to a previously unvisited vertex. Then  $X_i$  is certainly geometric and  $E[X_i] = 1/p_i$ .

We now find the probability  $p_i$  that we go to a previously unvisited vertex is  $(n - i + 1)/(n - 1)$ . This is because there are  $n - i + 1$  unvisited vertices and a total of  $n - 1$  adjacent vertices. It follows that  $E[X_i] = (n - 1)/(n - i + 1)$ . Since we start at a vertex,  $E[X_1] = 0$ . Then

$$\begin{aligned} t_{\text{cov}} &= 0 + \sum_{i=2}^n \frac{n-1}{n-i+1} \\ &= (n-1) \left( \frac{1}{n-1} + \frac{1}{n-2} + \cdots + 1 \right). \end{aligned}$$

A natural interpretation of the cover time on a complete graph is the coupon collector problem we introduced in Section 1. There are  $r$  distinct coupons one of which arrives in the mail everyday. We want to know how many days until the collector has collected each coupon. The strategy is to think of every coupon as a vertex on the complete graph. We then move from vertex to vertex with uniform probability. The difference between the coupon collector problem and the complete graph is that we now have self-edges whereas before we had to leave our current vertex at each step. That is, we can now stay in the same vertex (provided the collector receives the same coupon two days in a row). We apply our approach to the complete graph and substitute the  $r$  possible coupons we can find for the  $n - 1$  vertices we could move to. Then

$$t_{\text{cov}} = r \left( \frac{1}{r-1} + \frac{1}{r-2} + \cdots + 1 \right).$$

Another more subtle application of the cover time on a complete graph is the cover time on an  $n$ -star. An  $n$ -star is a graph with one central vertex and  $n - 1$  adjacent vertices each with a single edge to the center. A random walk begins at the central vertex and moves with uniform probability to one of the  $n - 1$  adjacent vertices. Whereas in the complete graph and coupon collecting problems we visited a (possibly) new vertex in one step, a random walk on the  $n$ -star visits a (possibly) new vertex in two steps: one step to each leaf and one step back. We apply our approach to the complete graph and conclude that

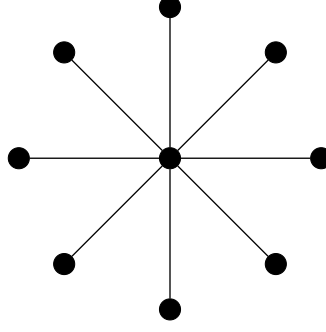


Figure 3: The nine-star.

$$t_{\text{cov}} = 2(n-1) \left( \frac{1}{n-1} + \frac{1}{n-2} + \cdots + 1 \right).$$

### 2.3 Line Graph

A line graph is a set of vertices connected in a line. (Fig. 4 is an example with  $n = N + 1$ .)

Our goal is to determine the cover time of a line graph. Let  $(X_t)$  be the random walk started at vertex  $i$ . Notice that a random walk will have covered all vertices exactly when it has visited both endpoints. We can use this observation to break up the cover time into the time it takes to reach one of the endpoints and the time it takes to reach the other endpoint.

We will begin by considering the time a random walk takes to visit either one of the endpoints from vertex  $i$ . From  $X_t = i$ , we move with equal probability to  $i + 1$  and  $i - 1$  until we reach either 0 or  $N$ .



Figure 4: A line graph with a vertex at each integer from 0 to  $N$ .

Call  $e_i$  the expected number of steps until we reach either end of our random walk from state  $i$ . Our strategy is to write and solve a recurrence relation that uses what we know about the random walk. If  $i = 0, N$ , we are already at the endpoint and  $e_i = 0$ . Otherwise, we take a step. There's a half probability our step is to the right and a half probability it is to the left.

$$\begin{aligned} e_i &= 1 + \frac{1}{2}(e_{i+1}) + \frac{1}{2}(e_{i-1}) \\ 2e_i &= 2 + e_{i+1} - e_{i-1} \\ (e_{i+1} - e_i) &= (e_i - e_{i-1}) - 2 \end{aligned}$$

While the final line may look more complicated than what we started with, notice that the structure of the equality lends itself to telescoping.

To solve the telescoping equations, observe that the walk is symmetric: we could flip the walk over the vertical axis without changing  $e_i$ . This makes sense because our



intuition tells us that the only identifying feature of  $i$  is its respective distances to the endpoints. The logical conclusion is that  $e_i = e_{N-i}$ . In particular, we have already seen that  $e_0 = e_N = 0$  and it naturally follows that  $e_1 = e_{N-1}$ . We use these observations to find  $e_1$ :

$$\begin{aligned}
(e_2 - e_1) &= (e_1 - e_0) - 2 \\
(e_3 - e_2) &= (e_1 - e_0) - 4 \\
&\vdots \\
(e_N - e_{N-1}) &= (e_1 - e_0) - 2(N-1) \\
(0 - e_1) &= (e_1 - 0) - 2(N-1) \\
e_1 &= N-1.
\end{aligned}$$

(Note that the vertical dots denote inductive reasoning.) We now know  $e_1$ . However, our goal is to find the expected number of steps from an arbitrary point  $i$ . We use  $e_1$  to find the general solution.

$$\begin{aligned}
(e_2 - e_1) &= (e_1 - e_0) - 2 \\
(e_3 - e_2) &= (e_1 - e_0) - 4 \\
&\vdots \\
(e_i - e_{i-1}) &= (e_1 - e_0) - 2(i-1)
\end{aligned}$$

Remember that  $e_0 = 0$ ,  $e_1 = N-1$ , and  $\sum_{i=1}^{i-1} i = (i-1)i/2$ . We add all  $i-1$  equations:

$$\begin{aligned}
(e_i - e_1) &= (i-1)e_1 - 2[1 + 2 + \cdots + i-1] \\
e_i &= i(N-1) - 2 \frac{(i-1)i}{2} \\
&= i(N-1-i+1) = i(N-i)
\end{aligned} \tag{1}$$

Theorem 1 follows from Eq. (1) and the observation that any line graph with  $N+1$  vertices can be relabeled so that the left endpoint is at 0 and the right endpoint is at  $N$ .

**Theorem 1.** *The expected number of steps from a state to either end of a line graph is the product of the distances from that state to each endpoint.*

We have the expected time until we reach one endpoint of our line graph. Now we want to find the expected time from one endpoint to the other  $E_0[T_N]$ . By symmetry,  $E_0[T_N] = E_N[T_0]$ .

We write

$$E_0[T_N] = E_0[T_1] + E_1[T_2] + \cdots + E_i[T_{i+1}] + \cdots + E_{N-1}[T_N].$$

where  $E_i[T_{i+1}]$  is the expected time from state  $i$  to  $i+1$ . Since there is only one edge incident to 0, we always move from 0 to 1. From state  $i \neq 0$ , we take one step in each

direction with equal probability. In half the cases, we have arrived at  $i + 1$ . In the other half of cases, we are in  $i - 1$ . It will take us  $E_{i-1}[T_i]$  steps back to  $i$  and then another  $E_i[T_{i+1}]$  steps to  $i + 1$ . Formally,

$$\begin{aligned} E_i[T_{i+1}] &= 1 + \frac{1}{2}(E_{i-1}[T_i] + E_i[T_{i+1}]) + \frac{1}{2}0 \\ &= 2 + E_{i-1}[T_i]. \end{aligned}$$

We can use  $E_0[T_1] = 1$  to solve for  $E_i[T_{i+1}]$ :

$$\begin{aligned} E_i[T_{i+1}] &= E_{i-1}[T_i] + 2 = E_{i-2}[T_{i-1}] + 4 \\ &= E_0[T_1] + 2(i - 1) = 2i - 1. \end{aligned}$$

Then

$$E_0[T_N] = 1 + 3 + \cdots + 2i - 1 + \cdots + 2(N - 1) - 1 = N^2$$

By Theorem 1, for the random walk from state  $i$  on the line graph,

$$t_{\text{cov}} = i(N - i) + N^2.$$

## 2.4 Simple Cycle

A simple cycle is a connected graph where each vertex has an edge to exactly two other vertices. (Fig. 5 is an example.)

Consider a random walk on a simple cycle as in Fig. 5. (Note that the starting vertex is arbitrary since the graph is symmetric.) Let  $(X_i)$  be the walk on the number line (between  $-n$  and  $n$ ) where  $X_0 = 0$ . The cover time  $t_{\text{cov}}$  of the simple cycle is the expected number of steps until we visit  $n$  distinct integers on the number line.

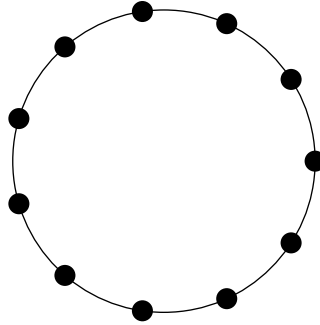


Figure 5: A cycle graph with  $n = 11$  vertices.

Let  $\{a, a + 1, \dots, b - 1, b\}$  denote the set of vertices we have visited. Since the walk is on a line, we can write  $[a, b]$  for the range of the walk where  $a$  is the smallest and  $b$  is the largest vertex we have visited. Notice that if  $a \leq -n$  or  $b \geq n$ , we have covered all  $n$  vertices on the simple cycle and conclude the random walk.

In a slight abuse of notation, let  $T_k$  be the number of steps until we reach  $k$  distinct vertices. Formally,  $T_k = \min\{t : b - a + 1 = k\}$ . Then by telescoping,

$$t_{\text{cov}} = E[T_n] = E[(T_2 - T_1) + (T_3 - T_2) + \cdots + (T_k - T_{k-1}) + \cdots + (T_n - T_{n-1})]$$

At time  $T_k$ , we are either at integer  $a$  or  $b$ . Then  $E[T_{k+1} - T_k]$  is the first time we reach  $a - 1$  or  $b + 1$ . Notice that we can shift our endpoints to 0 and  $b + 1 - (a - 1) = b - a + 2 = k + 1$ . The expected number of steps to either  $a - 1$  or  $b + 1$  is  $1(k + 1 - 1) = k$  by Theorem 1.

By linearity of expectation,

$$\begin{aligned} t_{\text{cov}} &= (E[T_2] - E[T_1]) + \cdots + (E[T_{k+1}] - E[T_k]) + \cdots + (E[T_n] - E[T_{n-1}]) \\ &= 1 + 2 + \cdots + k - 1 + \cdots + n - 1 = \frac{1}{2}n(n-1). \end{aligned}$$

### 3 Cover Time Bounds

#### 3.1 Commute Time Identity

The commute time between vertices  $a$  and  $b$  is the expected number of steps for a random walk starting in  $a$  to visit  $b$  and then return to  $a$ . Following [4], we use the electrical network analysis in Section 1.2 to prove the commute time identity so we may use it to help us bound cover times.

Define the probability  $P_a(T_b < T_a)$  that starting at  $a$  we reach  $b$  before returning to  $a$ . (Recall that  $T_a > 0$ .) Now define the effective conductance  $C_{\text{eff}}$  as  $i_a/v_a$  and the effective resistance  $R_{\text{eff}}$  as  $1/C_{\text{eff}}$ .

With a voltage of 1 at  $a$  and 0 at  $b$ ,

$$\begin{aligned} C_{\text{eff}} = i_a/1 &= \sum_y (v_a - v_y) C_{a,y} = \sum_y (v_a - v_y) \frac{C_{a,y}}{C_a} C_a \\ &= C_a \left( \frac{v_a}{C_a} \sum_y C_{a,y} - \sum_y \frac{C_{a,y}}{C_a} v_y \right) = C_a (1 - \sum_y P_{a,y} v_y). \end{aligned}$$

Recall  $\sum_y P_{a,y} v_y$  is the probability we reach  $a$  before  $b$  so the complement is the probability we reach  $b$  before returning to  $a$ .

Then

$$C_{\text{eff}} = C_a P_a(T_b < T_a)$$

and

$$R_{\text{eff}} = \frac{1}{C_a P_a(T_b < T_a)}.$$

We use the fact that  $\pi_a = C_a/C$  and multiply by  $C$ ,

$$R_{\text{eff}} C = \frac{1}{\pi_a P_a(T_b < T_a)}.$$

Recall that  $P_a(T_b < T_a)$  is the probability, starting from  $a$ , that we get to  $b$  before returning to  $a$ . We do independent trials, starting from  $a$ , until we reach  $b$  before  $a$ . Each “failure” resets the process so  $P_a(T_b < T_a)$  is geometrically distributed. Then the expected value  $1/P_a(T_b < T_a)$  is the expected number of times we return to  $a$  before we finally reach  $b$  (and again return to  $a$ ).

Recall that  $\pi_a$  is the stationary probability that  $(X_t) = a$ . By a similar argument,  $1/\pi_a$  is the expected return time to  $a$ . If we condition on reaching  $b$  before turning to  $a$ ,

$$\frac{1}{\pi_a} = E_a[T_a] = E_a[T_a | T_b < T_a] P(T_b < T_a) + E_a[T_a | T_a < T_b] P(T_a < T_b).$$

The expected commute time between  $a$  and  $b$  is the random sum expectation,

$$E_a[T_b] + E_b[T_a] = E_a[T_a | T_b < T_a] + E_a[T_a | T_a < T_b] \left( \frac{1}{P(T_b < T_a)} - 1 \right).$$

Observe that  $1/P(T_b < T_a) - 1 = (1 - P(T_b < T_a))/P(T_b < T_a)$ , which simplifies to  $P(T_a < T_b)/P(T_b < T_a)$ . Multiplying by the probability we reach  $b$  before returning to  $a$ ,

$$E_a[T_a | T_b < T_a] P(T_b < T_a) + E_a[T_a | T_a < T_b] P(T_a < T_b) = \frac{1}{\pi_a}$$

and the commute identity follows

$$E_a[T_b] + E_b[T_a] = \frac{1}{\pi_a P_a(T_b < T_a)} = R_{\text{eff}} C. \quad (2)$$

## 3.2 Matthews Method

We have seen that it is in general very hard to calculate the cover time of a random walk on a graph. The exceptions are symmetric or small graphs. In lieu of an exact solution, we want to bound the cover time.

The Matthews Method is an example of a remarkably good bound [7]. In fact, it is tight for the complete graph: the expected time between any two vertices is  $n - 1 = t_{\text{hit}}$  and Theorem 2 gives us the main result of Section 2.2.

An added benefit is that the proof uses strategies we employed on the simple cycle and follows an intuitive structure.

**Theorem 2** (Matthews Upper Bound). *Let  $(X_t)$  be a random walk on a graph with  $n$  vertices. Then*

$$t_{\text{cov}} \leq t_{\text{hit}} \left( 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} \right).$$

*Proof.* Without loss of generality, label the vertices  $\{1, 2, \dots, n\}$  such that  $n$  is the vertex from which it takes the most steps to cover the remaining vertices. Let  $\sigma$  be a uniform permutation of the unvisited  $n - 1$  vertices. The strategy is to look for the vertices in order  $\sigma$ .

Let  $T_k$  be the first time that vertices  $\sigma_1, \sigma_2, \dots, \sigma_k$  have all been visited. For ease of notation, set  $L_k = X_{T_k}$  to be the last state among  $\sigma_1, \sigma_2, \dots, \sigma_k$  to be visited the first time they are all visited.

We begin by writing the cover time in a clever way (similar to the strategy we employed on the cycle graph):

$$\begin{aligned} t_{\text{cov}} &= E_n[T_{n-1}] \\ &\leq E_n[T_1 + (T_2 - T_1) + \cdots + (T_k - T_{k-1}) + \cdots + (T_{n-1} - T_{n-2})]. \end{aligned}$$

Our goal is to bound the cover time in terms of the hitting time. Consider the first term of the sum. Recall that the hitting time  $t_{\text{hit}}$  is the maximum expected time from any vertex to any other. It follows that for any choice of  $\sigma_1$ , the expected time from  $n$  to  $\sigma_1$  is less than or equal to  $t_{\text{hit}}$ . Formally, for  $s$  an arbitrary state such that  $1 \leq s \leq n-1$ ,  $E_n[T_1] = E_n[T_s] \leq t_{\text{hit}}$ .

Now consider the  $k^{\text{th}}$  term of the sum for  $1 < k \leq n-1$ . By total expectation,

$$E_n[T_k - T_{k-1}] = \sum_{i=1}^k E_n[T_k - T_{k-1} | L(k) = \sigma_i] P(L_k = \sigma_i).$$

Assume that the last vertex  $L_k$  to be visited among the first  $k$  vertices is not  $\sigma_k$ . Then the first time we visit the first  $k-1$  vertices is also the first time we visit the first  $k$  vertices (since we must have already visited  $\sigma_k$ ). Formally,  $E_n[T_k - T_{k-1} | L(k) \neq \sigma_k] = 0$ . Then

$$E_n[T_k - T_{k-1}] = 0 + E_n[T_k - T_{k-1} | L(k) = \sigma_k] P(L_k = \sigma_k).$$

Since  $\sigma$  is a random permutation, the probability that we visit  $\sigma_k$  last is  $1/k$ . Now we find the expected time between  $T_{k-1}$  and  $T_k$  given that last vertex we visit is  $\sigma_k$ . Consider  $r$  and  $s$  where  $1 \leq r \neq s \leq n-1$  such that  $L_{k-1} = r$  and  $L_k = s$ . Then  $E_n[T_k - T_{k-1} | L(k) = \sigma_k] = E_r[T_s] \leq t_{\text{hit}}$ . So

$$E_n[T_k - T_{k-1}] \leq t_{\text{hit}} \frac{1}{k}.$$

Putting it all together,

$$t_{\text{cov}} \leq t_{\text{hit}} \left( 1 + \frac{1}{2} + \cdots + \frac{1}{k} + \cdots + \frac{1}{n-1} \right).$$

□

The same strategy for the Matthews upper bound can be used to find a lower bound. Instead of looking for all  $n$  vertices, we search for  $A \subseteq V(G)$ . The Matthews lower bound will always hold. However, we can tighten the bound by picking  $A$  such that the hitting time is large between any pair of vertices in  $A$ . We use ‘large’ in a relative sense and offer an illustrative example in Section 3.3.

**Theorem 3** (Matthews Lower Bound). *Let  $A \subseteq V(G)$ . Define  $t_{\min}^A = \min_{a,b \in A, a \neq b} E_b[T_a]$ . Let  $(X_t)$  be a random walk on a graph with  $n$  vertices. Then*

$$t_{\text{cov}} \geq t_{\min}^A \left( 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{|A|-1} \right).$$

*Proof.* Without loss of generality, label the vertices  $\{1, 2, \dots, x\}$  such that  $x = |A|$  and vertex  $x$  is the vertex from which it takes the most steps to cover the other vertices. Let  $\sigma$  be a uniform permutation of the vertices in  $A$ . The strategy is to look for the vertices in order  $\sigma$ .

Let  $T_k$  be the first time that vertices  $\sigma_1, \sigma_2, \dots, \sigma_k$  have all been visited. For ease of notation, set  $L_k = X_{T_k}$ .

We again write the cover time in a clever way:

$$t_{\text{cov}} \geq \mathbb{E}_x[T_1 + (T_2 - T_1) + \dots + (T_k - T_{k-1}) + \dots + (T_{|A|} - T_{|A|-1})].$$

Our goal is to bound the cover time in terms of  $t_{\min}^A$ . Consider the first term of the sum. With probability  $1/|A|$ ,  $\sigma_1 = x$  and  $T_1 = 0$ . Otherwise, the expected time from  $x$  to  $T_1$  is less than or equal to  $t_{\min}^A$ . Then

$$\mathbb{E}_x[T_1] \geq \frac{1}{|A|} 0 + \frac{|A|-1}{|A|} t_{\min}^A = \left(1 - \frac{1}{|A|}\right) t_{\min}^A$$

Now consider the  $k^{\text{th}}$  term of the sum for  $k > 1$ . By total expectation,

$$\mathbb{E}_x[T_k - T_{k-1}] = \sum_{i=1}^k \mathbb{E}_x[T_k - T_{k-1} | L(k) = \sigma_i] P(L_k = \sigma_i).$$

Assume that the last vertex  $L_k$  to be visited among the first  $k$  vertices is not  $\sigma_k$ . Then  $\mathbb{E}_x[T_k - T_{k-1} | L(k) \neq \sigma_k] = 0$ . It follows that

$$\mathbb{E}_x[T_k - T_{k-1}] = 0 + \mathbb{E}_x[T_k - T_{k-1} | L(k) = \sigma_k] P(L_k = \sigma_k).$$

Since  $\sigma$  is a random permutation,  $P(L_k = \sigma_k) = 1/k$ . Now find the expected time between  $T_{k-1}$  and  $T_k$  given that last vertex we visit is  $\sigma_k$ . Consider  $r, s \in A$  such that  $L_{k-1} = r$  and  $L_k = s$ . Then  $\mathbb{E}_x[T_k - T_{k-1} | L(k) = \sigma_k] = \mathbb{E}_r[T_s] \geq t_{\min}^A$ . So

$$\mathbb{E}_x[T_k - T_{k-1}] \geq t_{\min}^A \frac{1}{k}.$$

Putting it all together,

$$\begin{aligned} t_{\text{cov}} &\geq t_{\min}^A \left(1 - \frac{1}{|A|} + \frac{1}{2} + \dots + \frac{1}{k} + \dots + \frac{1}{|A|-1} + \frac{1}{|A|}\right) \\ &= t_{\min}^A \left(1 + \frac{1}{2} + \dots + \frac{1}{k} + \dots + \frac{1}{|A|-1}\right). \end{aligned}$$

□

### 3.3 Matthews Method on Binary Trees

Following [7], consider an unweighted, balanced binary tree as in Fig. 6 with a single root  $\rho$  and  $k$  levels. A vertex has two children if it is distance less than  $k$  from the root and no children if it is a leaf at distance  $k$  from the root. There are  $n = 1 + 2 + 4 + \dots + 2^k = 2^{k+1} - 1$  vertices.

By Eq. (2), the commute time between the root and a leaf  $a$  is

$$CR_{\rho,a} = 2(n-1)k.$$

Note that resistance adds in series so effective resistance is just the sum of the resistances along a path. Then  $R_{x,y} = i$  where  $i$  is the length of the unique path between  $x$  and  $y$ .

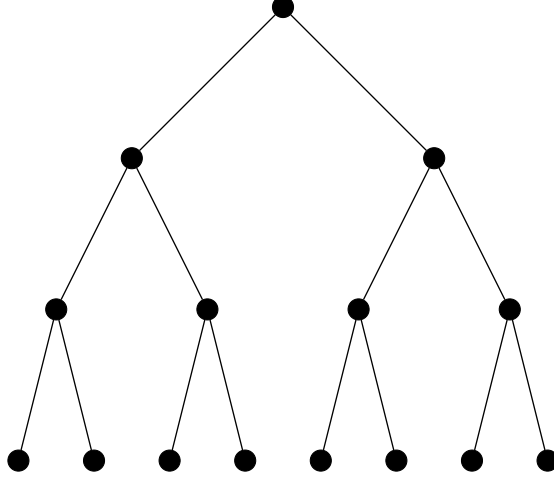


Figure 6: A binary tree with  $k = 3$  has  $n = 15$ .

We first use the Matthews Upper Bound. The maximum hitting distance is between two vertices whose most recent common ancestor is the root. For this pair, the hitting time is the same as the commute time to the root by symmetry. Consider  $a$  and  $b$  two leaves whose most recent common ancestor is the root. Then

$$E_a[T_b] = E_a[T_\rho] + E_\rho[T_a] = 2(n-1)k$$

and Theorem 2 gives

$$\begin{aligned} t_{\text{cov}} &\leq 2(n-1)k \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) \\ &\sim \left(2 - \frac{2}{n}\right) \cdot nk \log n \sim \left(2 - \frac{2}{n}\right) nk \log 2^{k+1} - 1 \\ &\sim (2 + o(1)) \cdot nk^2 \log 2 \end{aligned}$$

We now use the Matthews Lower Bound. We want to choose a set  $A \subseteq V(G)$  with a large hitting time between any pair of vertices. To find such a set, fix a level  $h$  in our binary tree. Then let  $A$  be a set of  $2^h$  leaves so that each vertex at level  $h$  has a unique descendent in  $A$ . As  $h$  increases, the distance between the leaves in  $A$  decreases. Thus we want to choose a smaller value of  $h$  so that our lower bound will be bigger (and better).

For  $a, b \in A$ , the hitting time between  $a$  and  $b$  is the commute time from  $a$  to their nearest common ancestor at level  $h' < h$ . By Eq. (2), we have

$$E_a[T_b] = E_a[T_{h'}] + E_{h'}[T_a] = 2(n-1)(k-h').$$

In the worst case,  $h' = h - 1$ . Then by Theorem 3,

$$\begin{aligned} t_{\text{cov}} &\geq 2(n-1)(k-h+1) \left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{h-1}-1}\right) \\ &\geq (2 + o(1)) \cdot n(k-h)h \log 2. \end{aligned}$$

If  $h = \lfloor k/2 \rfloor$ ,

$$t_{\text{cov}} \geq \frac{1}{4}(2 + o(1)) \cdot nk^2 \log 2$$

then there is a factor of 4 gap between the upper bound and lower bound of the cover time of a binary tree!

### 3.4 Spanning Tree Argument

Following [1], we introduce the concept of a spanning tree to create another upper bound for the cover time of a random walk.

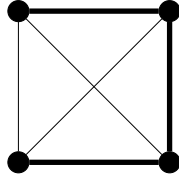


Figure 7: A spanning tree (in bold) of the complete graph with  $n = 4$ .

A tree is a connected graph with no cycles (i.e. there is exactly one path between every pair of points.) A spanning tree  $\mathcal{T}$  for a connected graph  $G$  is a tree with  $V(\mathcal{T}) = V(G)$  and  $E(\mathcal{T}) \leq E(G)$ . Fig. 7 shows an example spanning tree where the bold edges indicate the edges in  $E(\mathcal{T})$  and the light edges indicate the edges in  $E(G)$  but not  $E(\mathcal{T})$ . A consequence of the definition is that  $|E(\mathcal{T})| = n - 1$ .

In Section 3.1, we fixed vertices  $a$  and  $b$  and called the effective resistance between them  $R_{\text{eff}}$ . We now consider  $R_{x,y}$  the effective resistance between vertices  $x$  and  $y$ . Recall that  $C$  is the sum of the conductances of every vertex in our graph.

**Theorem 4** (Spanning Tree Upper Bound). *For any spanning tree  $\mathcal{T}$  on a weighted graph,*

$$t_{\text{cov}} \leq C \sum_{x,y \in E(\mathcal{T})} R_{x,y} \leq C \sum_{x,y \in E(\mathcal{T})} 1/C_{x,y}.$$

*In particular for an unweighted graph,*

$$t_{\text{cov}} \leq 2|E(G)|(n - 1).$$

*Proof.* Given a spanning tree  $\mathcal{T}$ , there exists a vertex  $v$  with only one edge. Consider the path  $v_0, v_1, \dots, v_{2n-2}$  where  $v = v_0 = v_{2n-2}$  that traverses each edge in the spanning tree once in each direction and covers every vertex. We construct this path by starting at  $v$  and walking along all  $n - 1$  edges to the other vertex with only one edge and then coming back to  $v$ .

The expected time of this particular path on the original graph is certainly greater than or equal to the cover time. That is,

$$t_{\text{cov}} \leq \sum_{i=0}^{2n-3} E_{v_i}[T_{v_{i+1}}].$$



We rewrite the sum in terms of commute times,

$$t_{\text{cov}} \leq \sum_{x,y \in E(\mathcal{T})} E_x[T_y] + E_y[T_x].$$

Using Eq. (2), we have  $E_x[T_y] + E_y[T_x] = C R_{x,y}$  and,

$$t_{\text{cov}} \leq C \sum_{x,y \in E(\mathcal{T})} R_{x,y}.$$

Observe that if  $x \sim y$ , the effective conductance between  $x$  and  $y$  is at least  $C_{x,y}$  since each additional path only increases the flow. It follows that the effective resistance between  $x$  and  $y$  is at most  $1/C_{x,y}$  so

$$t_{\text{cov}} \leq C \sum_{x,y \in E(\mathcal{T})} 1/C_{x,y}.$$

If the graph is unweighted, then  $C_{x,y} = 1$  and  $C = 2|E(G)|$ . Then

$$t_{\text{cov}} \leq 2|E(G)|(n-1).$$

□

## 4 Cover Time Distributions

### 4.1 Complete Graph

Following [5], we consider the distribution of the cover time of a random walk on the complete graph with self-edges. Call  $r$  the number of steps since we began the random walk. So that the notation works out more easily, set  $r = 1$  when we begin our walk at  $X_0$ . Let  $A_i$  be the event that vertex  $i$  is not visited and  $N_n$  the number of unvisited vertices. Then

$$P(A_i) = (1 - 1/n)^r$$

and

$$E[N_n] = n(1 - 1/n)^r.$$

If  $r/n \rightarrow c$ , then

$$E[N_n]/n = (1 - 1/n)^r = [(1 - 1/n)^n]^{r/n} \rightarrow (e^{-1})^c \rightarrow e^{-c}.$$

We compute the variance of  $N_n$  by finding  $E[N_n^2]$  and  $E[N_n]^2$ . Let  $\mathbb{1}_m$  be the indicator random variable for event  $A_m$ . That is,  $\mathbb{1}_m = 1$  if vertex  $m$  is unvisited 0 otherwise. We first find  $E[N_n^2]$

$$E[N_n^2] = E \left[ \left( \sum_{m=1}^n \mathbb{1}_m \right)^2 \right] = E \left[ \sum_{m=1}^n \sum_{k=1}^n \mathbb{1}_m \mathbb{1}_k \right] = \sum_{1 \leq k, m \leq n} P(A_k \cap A_m).$$

The second equality holds because we distribute terms. The third equality holds because  $\mathbb{1}_m \mathbb{1}_k = \mathbb{1}_{m \cap k}$  and the expectation of an indicator random variable is the event's probability. Then we find  $E[N_n]^2$

$$E[N_n]^2 = E \left[ \left( \sum_{m=1}^n \mathbb{1}_m \right) \left( \sum_{k=1}^n \mathbb{1}_k \right) \right] = \sum_{1 \leq k, m \leq n} P(A_k) P(A_m).$$

Putting both together, we find the variance

$$\begin{aligned} \text{Var}[N_n] &= E[N_n^2] - E[N_n]^2 \\ &= \sum_{1 \leq k, m \leq n} P(A_k \cap A_m) - P(A_k)P(A_m) \\ &= n(n-1) [P(A_k \cap A_m) - P(A_k)P(A_m)] + n [P(A_k \cap A_m) - P(A_k)P(A_m)] \\ &= n(n-1) [(1-2/n)^r - (1-1/n)^{2r}] + n [(1-1/n)^r - (1-1/n)^{2r}]. \end{aligned}$$

where  $k \neq m$  in the first term and  $k = m$  in the second.

As  $n \rightarrow \infty$ ,

$$\begin{aligned} \text{Var}(N_n/n) &= \text{Var}(N_n)/n^2 \\ &= (1-2/n)^r - (1-1/n)^{2r} + o(n) \\ &\rightarrow e^{-2c} - (e^{-c})^2 = 0. \end{aligned}$$

By Chebyshev's Inequality,

$$P(|N_n/n - E[N_n/n]| \geq \epsilon) \leq \text{Var}(N_n/n)/\epsilon^2 \rightarrow 0.$$

Thus we conclude that  $N_n/n$  converges in probability to  $e^{-c}$ .

The Poisson approximation of the binomial implies that if  $n \rightarrow \infty$  and  $r/n \rightarrow c$ , then the number of visits to each vertex will approach a Poisson distribution with mean  $c$ . We can use this to intuit our latest result: the proportion of unvisited vertices approaches  $e^{-c}$ .

We prove the following theorem and then use it to gain insight into the distribution of the cover time variable  $\tau_{\text{cov}}$ .

**Theorem 5.** *If  $ne^{-r/n} \rightarrow \lambda \in [0, \infty)$  as  $n \rightarrow \infty$ , then the number of unvisited vertices approaches a Poisson distribution with mean  $\lambda$ .*

*Proof.* Observe that the chance a particular set of  $k$  vertices are unvisited is

$$(1 - k/n)^r.$$

Define  $p_m(r, n)$  as the probability that exactly  $m$  of  $n$  vertices are unvisited when  $r$  steps have been taken. Then the probability that all vertices have been visited is the complement of the probability that at least one vertex has not been visited.

By the principle of inclusion-exclusion,

$$\begin{aligned} p_0(r, n) &= \binom{n}{0} 1^r - \binom{n}{1} (1 - 1/n)^r + \dots \\ &= \sum_{k=0}^n (-1)^k \binom{n}{k} (1 - k/n)^r. \end{aligned} \tag{3}$$

If we consider the location of the  $m$  unvisited vertices, and multiply probability that those  $m$  vertices are unvisited and the rest are visited,

$$p_m(r, n) = \binom{n}{m} \left(1 - \frac{m}{n}\right)^r p_0(r, n - m).$$

Our strategy is to show that  $p_m(r, n) \rightarrow \frac{\lambda^m}{m!} e^{-\lambda}$ . The first step is to prove that

$$\binom{n}{m} \left(1 - \frac{m}{n}\right)^r \rightarrow \lambda^m / m! \quad (4)$$

We bound the expression from the top and the bottom and use the squeeze theorem to prove the most recent limit. We will first show the upper bound. Notice that  $(1 - x) \leq e^{-x}$  and recall our assumption that  $ne^{r/n} \rightarrow \lambda$ . Then

$$\binom{n}{m} \left(1 - \frac{m}{n}\right)^r \leq \frac{n^m}{m!} e^{mr/n} = \frac{(ne^{r/n})^m}{m!} \rightarrow \frac{\lambda^m}{m!}. \quad (5)$$

We now turn to the more tricky lower bound. Notice that  $\binom{n}{m} \geq (n - m)^m / m!$ . Then

$$\begin{aligned} \binom{n}{m} \left(1 - \frac{m}{n}\right)^r &\geq \left(1 - \frac{m}{n}\right)^r \frac{(n - m)^m}{m!} \left(\frac{n^m}{n^m}\right) \\ &\geq n^m \left(1 - \frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^m \frac{1}{m!}. \end{aligned}$$

As  $n \rightarrow \infty$ ,  $(1 - \frac{m}{n})^m \rightarrow 1$  and  $1/m!$  is constant. We use the Taylor series to bound the rest. (The following equality is valid for  $-1 \leq t \leq 1$  but we only need  $0 \leq t \leq 1$  because  $m/n \rightarrow 0$  and  $m, n$  are counts of vertices with  $m \leq n$ .) We apply

$$\begin{aligned} \log(1 - t) &= 0 - t - t^2/2 - t^3/3 - \dots \\ &\geq -t - t^2/2(1 + 1/2 + 1/4 + \dots) = -t - t^2 \end{aligned}$$

to

$$\begin{aligned} \log \left( n^m \left(1 - \frac{m}{n}\right)^r \right) &= m \log n + r \log \left(1 - \frac{m}{n}\right) \\ &\geq m \log n - r \frac{m}{n} - r \left(\frac{m}{n}\right)^2. \end{aligned}$$

We use our assumption that  $ne^{-r/n} \rightarrow \lambda$  to solve for  $r$

$$\begin{aligned} ne^{-r/n} &\rightarrow \lambda \\ \log n - \frac{r}{n} &\rightarrow \log \lambda \\ \frac{(n \log n - n \log \lambda) - r}{n} &\rightarrow 0 \end{aligned}$$

then

$$r = (n \log n - n \log \lambda) + o(n).$$

Since  $r(m/n)^2 \rightarrow 0$  and

$$\begin{aligned} \frac{rm}{n} &= m \log n - m \log \lambda + o(n)m \\ m \log n - \frac{rm}{n} &\rightarrow m \log \lambda \end{aligned}$$

we have

$$\log \left( n^m \left( 1 - \frac{m}{n} \right)^r \right) \geq m \log n - \frac{rm}{n} - \frac{rm}{n^2} \rightarrow \log \lambda^m.$$

It follows that

$$\lim_{n \rightarrow \infty} n^m \left( 1 - \frac{m}{n} \right)^r \geq \lambda^m.$$

Now we put the pieces back together. The most recent limit and Eq. (5) imply Eq. (4). By Eq. (4) and Eq. (3),

$$p_0(r, n) \rightarrow \sum_{k=0}^{\infty} (-1)^k \lambda^k / k! = e^{-\lambda}.$$

Notice  $p_0(r, n - m)$  also goes to  $e^{-\lambda}$  since  $m$  is fixed and  $(n - m)e^{r/(n-m)} \rightarrow \lambda$ . Finally, the most recent limit and the equation for  $p_m(r, n)$  imply

$$p_m(r, n) \rightarrow \frac{\lambda^m}{m!} e^{-\lambda}.$$

Thus  $N_n$  approaches a Poisson distribution with mean  $\lambda$  as claimed. □

We use Theorem 5 to find the distribution of the cover time variable  $\tau_{\text{cov}}$ . Set  $r = n \log n + nx$ . Then

$$ne^{-r/n} = ne^{-\log n - x} = nn^{-1}e^{-x} = e^{-x}.$$

Using the fact that  $\tau_{\text{cov}} \leq m$  if and only if the first  $m$  steps of the random walk visit all  $n$  vertices and  $ne^{-r/n} \rightarrow e^{-x}$  as the assumption of Theorem 5 (with  $e^x$  in the role of  $\lambda$ ),

$$P(N_n = 0) = P(\tau_{\text{cov}} \leq r) = P(\tau_{\text{cov}} - n \log n \leq nx) \rightarrow e^{-e^{-x}}. \quad (6)$$

## 4.2 Birthday Problem

Consider a complete graph (with self-edges) and 365 vertices. We may think of the cover time variable of this graph as the number of people until every birthday is represented. We call this the birthday problem.

We use Eq. (6) to calculate the probability that all birthdays are represented in populations of size  $r = 1825$  and  $r = 2190$ , respectively. The number of vertices  $n$  is 365 so  $n \log n = 2153$ . It follows that

$$\begin{aligned} P(\tau_{\text{cov}} \leq 1825) &= P\left(\frac{\tau_{\text{cov}} - 2153}{365} \leq \frac{-328}{365}\right) \\ &\approx e^{-e^{-0.89863}} \approx 0.08576 \end{aligned}$$

but

$$\begin{aligned} P(\tau_{\text{cov}} \leq 2190) &= P\left(\frac{\tau_{\text{cov}} - 2153}{365} \leq \frac{37}{365}\right) \\ &\approx e^{-e^{-0.10137}} \approx 0.40511. \end{aligned}$$

There appears to be a critical threshold where it quickly becomes very likely that all birthdays have been represented. Fig. 8 verifies our intuition.

Cover Times of the Birthday Problem (1000 trials)

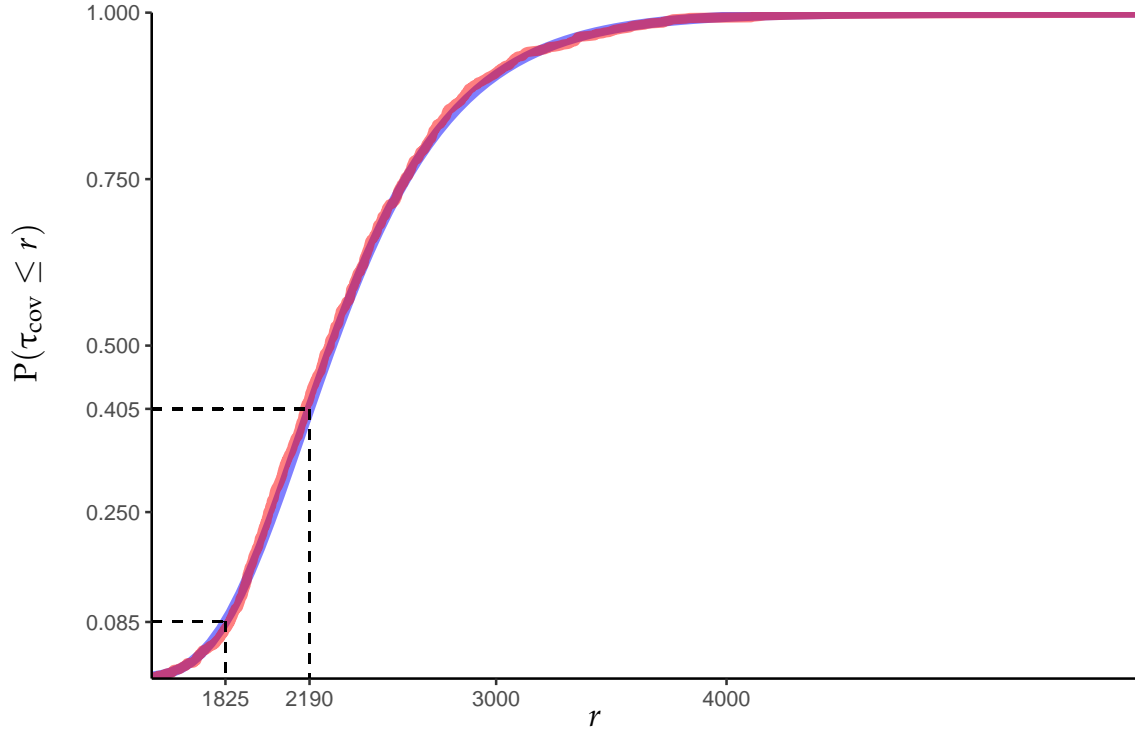


Figure 8: The cumulative distribution of the birthday problem. The red line denotes the empirical distribution from 1000 trials. The blue line denotes the distribution from Eq. (6). (Our code to generate the figure appears in Section 5.1.)

## References

- [1] David Aldous and James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs*. 2002.
- [2] Gunnar Blom, Lars Holst, and Dennis Sandell. *Problems and Snapshots from the World of Probability*. Springer, 1994.
- [3] Béla Bollobás. *Modern Graph Theory*. Springer, 1998.
- [4] Peter G. Doyle and J. Laurie Snell. *Random Walks and Electric Networks*. Mathematical Assn of Amer, 1984.
- [5] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2011.
- [6] Samuel Karlin. *A First Course in Stochastic Processes*. Academic Press, 2014.
- [7] David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. Cambridge University Press, 2017.

## 5 Appendix

### 5.1 Birthday Problem R Code

```
# R. Teal Witter
# Cover Time of Birthdays
library(tidyverse)
library(ggplot2)

trials <- 1000          # Number of trials
n <- 365                # Number of 'vertices' to cover
cover_times <- c()
for (i in 1:trials) {
  x <- 0
  birthdays <- 1:n # Unvisited birthdays
  # Count number of people until every birthday is covered
  while (length(birthdays) > 0) {
    # Uniform birthdays
    birthday <- sample(1:n, 1, replace=TRUE)
    # Remove birthday
    birthdays <- birthdays[birthdays != birthday]
    # Count steps
    x <- x+1
  }
  cover_times[i] <- x
}

# Append percentiles
cover_times_analyzed <- data.frame(cover_times) %>%
  mutate(percentile=ecdf(cover_times)(cover_times))

# Analytical result of cover time CDF
cdf <- function(r) {
  x <- (r - n * log(n)) / n
  return (exp(-exp(-x)))
}

# Start of x axis
start_x <- min(cover_times_analyzed$cover_times)

cdf(2190) # Probability all birthdays represented with 2190 people
cdf(1825) # Probability all birthdays represented with 1825 people

# Figure of birthday cover times
cover_times_analyzed %>%
  ggplot() + theme_classic() +
  # Analytical cover times
```

```

stat_function(fun = cdf, color="blue", alpha=.5, size=2) +
# Empirical cover times
geom_line(aes(x=cover_times_analyzed$cover_times,
              y=cover_times_analyzed$percentile),
          color="red", alpha=.5, size=2) +
ylab("") + xlab("") +
# Dashed lines to indicate points of interest
geom_segment(aes(x=2190, y=cdf(2190), xend=2190, yend=0),
             linetype="dashed") +
geom_segment(aes(x=1825, y=cdf(1825), xend=1825, yend=0),
             linetype="dashed") +
geom_segment(aes(x=start_x, y=cdf(1825), xend=1825, yend=cdf(1825)),
             linetype="dashed") +
geom_segment(aes(x=start_x, y=cdf(2190), xend=2190, yend=cdf(2190)),
             linetype="dashed") +
scale_x_continuous(expand = c(0, 0),
                   breaks=c(1825, 2190, 3000, 4000)) +
scale_y_continuous(expand = c(0, 0),
                   breaks=c(.085, .25, .405, .5, .75, 1))

```