3

INEQUALITIES FOR RANDOM WALKS ON TREES*

LUC DEVROYE and AMINE SBIHI
McGill University

3.1. INTRODUCTION

On a tree with n vertices, a random walk is carried out by visiting neighbors of the current vertex with equal probability. Thus, each neighbor of a given current vertex i with degree d_i has probability $1/d_i$ of becoming the current node. In this chapter, we are interested in general inequalities for $\mathbf{E}T$, where the cover time T is the time needed to visit all the vertices. Obviously, $\mathbf{E}T$ depends on the structure of the tree and the starting point. In particular, we will study inequalities that are uniformly valid over all starting points and over large subclasses of trees.

Random walks on trees were studied by Moon in 1973 and by Göbel and Jagers in 1974. Both papers are largely confined to the problem of determining the first and second moments of the random variable

$$T_{i,j} = \min\{n \geq 1: X_n = j | X_0 = i\},$$

the time needed for a random walk (X_n) started at vertex i to reach j.

*Research sponsored by NSERC grant A3456 and FCAR grant EQ-1679.

Random Graphs, Volume 2. Edited by Alan Frieze and Tomasz Łuczak.
 ISBN: 0471-57292-6.

Nevertheless, some of the key results in our proof are taken from these papers. Random walks on graphs were also studied by Göbel and Jagers (1974). In a profound paper, Mazo (1982) has shown that the average of all $\mathbf{E}T_{i,j}$ is at least $n - 1$ for any graph with n nodes. For a recent survey of results on the cover time T, see Aldous (1989b). Turning to lower bounds, we have for example:

1. For a stationary random walk on a graph, $\mathbf{E}T \geq cn \log n$ for a universal constant c (Aldous, 1989b).
2. For trees, $\mathbf{E}T \geq (1 + o(1))n \log n$ [Kahn, Linial, Nisan, and Saks (1989)]. For trees with maximal degree bounded by Δ, $\mathbf{E}T \geq (n \log^2 n)/(600 \log^2 2\Delta)$ [Zuckerman (1989)].
3. Bounds involving the structure of the graph are given by Broder and Karlin (1989).
4. For a cube, $\mathbf{E}T \geq (n-1)H_{n-1} \sim n \log n$, where $H_n = \sum_{i=1}^{n} 1/i$ is the harmonic function. A detailed analysis, including the asymptotic distribution of the normalized cover time, is given in Matthews (1987, 1988b, 1989).
5. For distance-regular graphs, $\mathbf{E}T \geq (n-1)\log n$ and for symmetric graphs ($\mathbf{E}T_{i,j} = \mathbf{E}T_{j,i}$ for all i, j), $\mathbf{E}T \geq \frac{1}{2} n \log n - O(n \log\log n)$ [Devroye and Sbihi (1990)].

Upper and lower bounds and asymptotics for other graphs can be found in Broder and Karlin (1989), Kahn et al. (1989) (k-regular graphs), Aldous (1983) (finite groups, rapidly mixing chains), Devroye and Sbihi (1990), and Aldous (1989a) (vertex-transitive graphs). Aleliunas, Karp, Lipton, Lovász and Rackoff (1979) proved that $\mathbf{E}T \leq 2r(n-1)$, where r is the number of edges in the graph. Thus, for trees, since $r = n - 1$, we have $\mathbf{E}T \leq 2(n-1)^2$. This bound is attainable, modulo a constant 2, for chains (linear graphs). Though the constant seems suboptimal, we will not bother to improve the Aleliunas inequality here. Rather, we will concentrate on the universal lower bound. Our main result is:

Theorem 3.1. Let $\mathbf{T}_n$ be the collection of all trees with n vertices and let t denote a member of $\mathbf{T}_n$. Then

$$\inf_{t \in \mathbf{T}_n} \mathbf{E}T \geq 2n \log n - O(n \log\log n).$$

Note that Theorem 3.1 improves on the result of Kahn et al. (1989); it also uses a totally different method of proof. Stars (i.e., trees in which $n - 1$ nodes are connected to a central node) have $\mathbf{E}T = 2(n-1)H_{n-1} - 1$ if we start at the central node, so the coefficient 2 in the bound of Theorem 3.1 cannot be bettered. Because stars are rather unnatural trees, one might wonder what happens for more "natural" trees. We will provide the reader with a variety of inequalities for $\mathbf{E}T$ that involve some shape parameters of the tree, such as the diameter, the number of leaves, or the maximal degree. For example, for trees having bounded degree Δ, we have:

Theorem 3.2. Let $\mathbf{T}_n$ denote the collection of all trees with n nodes and maximal degree Δ with $\Delta \geq 3$ possibly depending upon n. If $\Delta = n^{o(1)}$, we have

$$\inf_{t \in \mathbf{T}_n} \mathbf{E}T \geq \frac{1 + o(1)}{2 \log(\Delta - 1)} n \log^2 n .$$

By considering complete $(\Delta - 1)$-ary trees, we will see that this bound has the right dependence on n and Δ. Also, the $o(1)$ term is explicitly derived in the proof of Theorem 3.1; it depends on Δ and n only. The inequality is not valid when Δ increases polynomially in n, but in those cases, more complicated but valid lower bounds are attainable by taking the minimum of expressions (3.7) and (3.8). Theorem 3.2 improves the lower bound of Zuckerman (1989) in two respects: It has a smaller constant out front and it betters Zuckerman's bound by a factor of about $300 \log \Delta$.

The random walk discussed in this chapter has some relevance in computer networks, where the network is fixed and some messages are randomly routed. $\mathbf{E}T$ is the expected time at which all the nodes of the network have received the message. In social networks, the model might be used to study the spread of gossip or disease in a fixed network of friends, although a random walk on a graph including cycles seems perhaps slightly more appropriate.

We employ a recent result of Matthews (1988a) regarding the time needed by a Markov chain to visit all of a finite collection of subsets of its state space.

3.2. AN INEQUALITY OF MATTHEWS

We begin with an inequality of Matthews (1988a), presented here in a form convenient to us.

Lemma 3.1. (Matthews's inequality). Let $X_0, X_1, X_2, \ldots$ be a Markov chain on a countable state space $\{a_0, a_1, \ldots\}$. If T is the time needed for the Markov chain to visit the collection $A = \{a_0, a_1, \ldots, a_k\}$, then for $X_0 = a_0$,

$$\mathbf{E}T \geq \mu(A) H_k,$$

where H_k is the kth harmonic number, and

$$\mu(A) \triangleq \min_{1 \leq j \leq k} \min_{0 \leq i \leq k;\, i \neq j} \mathbf{E}T_{i,j}$$

and $\mathbf{E}T_{i,j}$ is the expected time needed to reach state a_j when starting from a_i.

In our problem, we will get the most out of this by identifying large subsets of vertices on which $\mu(A)$, as defined before, is large. The requirement that both $\mu(A)$ and the size $|A|$ of the subset A are large is, in general, contradictory, so we will often have to strike a suitable balance. A particularly interesting set is the set A of all leaves. It is clear that a random walk on a tree covers all the vertices if and only if it covers all the leaves. By the strong Markov property,

$$\mathbf{E}_s T = \mathbf{E}_s T_A + \sum_{k \in A} \mathbf{P}_s(X_{T_A} = k) \mathbf{E}_k T \geq \min_{k \in A} \mathbf{E}_k T \tag{3.1}$$

for any vertex s, where T_A and X_{T_A} are, respectively, the first hitting time and the hitting place of A, and $\mathbf{E}_s$ is the conditional expectation given $X_0 = s$. One also has $\mathbf{E}_s T_A \leq D^2/4$ [see Pearce (1980)], where D is the diameter of the tree. Thus,

$$\mathbf{E}_s T \leq \max_{k \in A} \mathbf{E}_k T + D^2/4.$$

Example. Random walks on cubes. The D-dimensional cube graph has $n = 2^D$ vertices, each representing a different binary sequence of length D. Two vertices are connected by an edge if and only if the corresponding sequences differ in only one place. For any pair of vertices (i, j), it can be shown [see, for example, Göbel and Jagers (1974)] that

$$\mathbf{E}T_{i,j} = \sum_{m=1}^{k} \left(\sum_{l=m}^{D} \binom{D}{l} \Big/ \binom{D-1}{m-1} \right)$$

if the Hamming distance between i and j is k. It is easy to see that this expected time is a monotone function of k. Its minimal value is $2^D - 1 = n - 1$ and its maximal value is $2^D(1 + O(1/D))$ as $D \to \infty$. By Matthews's inequality, $\mathbf{E}T \geq (n-1)H_{n-1}$. Also, by an upper bound symmetric to the lower bound of Lemma 3.1 (Matthews, 1988), $\mathbf{E}T \sim n \log n$. For more details, see Matthews (1988a, 1989), Aldous (1983), Diaconis (1988), or Devroye and Sbihi (1989).

3.3. FIRST PASSAGE TIMES

For all connected graphs with r edges, it is known that $\mathbf{E}T_{i,i} = 2r/d_i$, where d_i is the degree of node i and $T_{i,i}$ is the time of first return to i when starting a random walk from i [Göbel and Jagers (1974)]. Moon (1973) has shown that when i and j are neighboring vertices in a tree, then $\mathbf{E}T_{i,j} = 2N - 1$, where N is the size of the component containing i when edge (i, j) is deleted. From this, we easily obtain the exact value of $\mathbf{E}T_{i,j}$ in trees when $i \neq j$: Arguing as in Moon (1973), we see that

$$\mathbf{E}T_{i,j} = \sum_{m=1}^{k} \mathbf{E}T_{a_{m-1}, a_m},$$

where $a_0 = i$, $a_k = j$, and $a_1, \ldots, a_{k-1}$ are the vertices on the unique path from i to j. Consider the subtree S_m containing a_{m-1} when edge (a_{m-1}, a_m) is deleted. Then

$$\mathbf{E}T_{i,j} = \sum_{m=1}^{k} (2|S_m| - 1). \tag{3.2}$$

Corollary 3.1. *j* is a leaf. If $n \geq 3$ and i, j are leaves, then $\mathbf{E}T_{i,j} \geq 2(n-1)$.

Corollary 3.2. *i* and *j* are at distance *k*. $\mathbf{E}T_{i,j} \geq k^2$. This comes from (3.2) with $|S_1| \geq 1$, $|S_{m+1}| \geq |S_m| + 1$, and thus $|S_m| \geq m$. Hence, $\mathbf{E}T_{i,j} \geq \sum_{m=1}^{k}(2m - 1) = k(k+1) - k = k^2$.

3.4. PROVING THEOREM 3.1

We split the proof of Theorem 3.1 into several lemmas. The first lemma follows directly from Matthews's lower bound and Corollary 3.1.

Lemma 3.2. Assume that a tree has L leaves and n vertices, $n \geq 3$. Then for a random walk started at an arbitrary node,

$$\mathbf{E}T \geq 2(n-1)H_{L-1}.$$

Any random walk started at a nonleaf node reaches the set A of leaf nodes at some vertex i and needs to visit all other leaves starting at the given leaf i. Thus, regardless of where we begin the walk,

$$\mathbf{E}T \geq \inf_{i \in A} \sup_{\substack{j \in A \\ j \neq i}} \mathbf{E}T_{i,j}. \tag{3.3}$$

Fix a leaf i. If $d(i,j)$ is the distance between nodes i and j, we have

$$\max_{\substack{j \in A \\ j \neq i}} d(i,j) \geq (L-1)^{-1} \sum_{\substack{j \in A \\ j \neq i}} d(i,j) \geq (n-1)/(L-1).$$

Thus, by (3.3) and Corollary 3.2, $\mathbf{E}T \geq ((n-1)/(L-1))^2$. This combined, with Lemma 3.2 shows

$$ET \geq \inf_{\substack{n-1 \geq u \geq 2 \\ u \text{ integer}}} \max\left(2(n-1)H_{u-1}, \left(\frac{n-1}{u-1}\right)^2\right) \sim n \log n.$$

Here we used the fact that the supremum is attained for $u \sim \sqrt{n/\log n}$. However, this bound falls short of the best possible bound by a factor of 2. This necessitates the search for a stronger lower bound as a function of L.

Lemma 3.3. For a tree with n nodes and L leaves, and for an arbitrary starting node, we have $\mathbf{E}T \geq 2(n-1)^2/L$.

Proof. Our proof is based on inequality (3.3). We take an arbitrary leaf and label it 1 for convenience. Consider next the maximal path from 1 to any leaf, and label the nodes $1, 2, \ldots, l$. (Thus, the maximal path length is $l-1$.) We will show that $\mathbf{E}T_{1,l} \geq 2(n-1)^2/L$.

Attached to the maximal path are zero or more nonempty trees whose roots are thus at distance 1 from the maximal path. For the ith such tree, define n_i (the cardinality), l_i (the number of leaves), and d_i (the distance from l to the neighbor on the maximal path of the root of the tree). We have

$$\sum_i n_i = n - l, \tag{3.4}$$

$$\sum_i l_i = L - 2, \tag{3.5}$$

$$d_l \geq 1 + \frac{n_i - 1}{l_i}. \tag{3.6}$$

Inequality (3.6) is seen by noting that the height of the ith tree is greater than or equal to $(n_i - 1)/l_i$ and that the height plus 1 cannot exceed d_i, for otherwise l is not the furthest leaf from 1. Note that $l_i > 0$ and $n_i > 0$ for all i. From Moon's formula (3.2),

$$\begin{aligned}
\mathbf{E}T_{1,l} &= 2\sum_i d_i n_i + (l-1)^2 \\
&\geq 2\sum_i \left(1 + \frac{n_i - 1}{l_i}\right) n_i + (l-1)^2 \\
&\geq 2\sum_i \frac{n_i^2}{l_i} + (l-1)^2 \\
&= 2\sum_i \sum_{k=1}^{l_i} \left(\frac{n_i}{l_i}\right)^2 + (l-1)^2 \\
&\geq 2\frac{(\Sigma_i \Sigma_{k=1}^{l_i} n_i/l_i)^2}{\Sigma_i \Sigma_{k=1}^{l_i} 1} + (l-1)^2 \quad \text{(Cauchy–Schwarz inequality)} \\
&= 2\frac{(\Sigma_i n_i)^2}{L-2} + (l-1)^2 = 2\frac{(n-l)^2}{L-2} + (l-1)^2 \\
&\geq 2\frac{(n-1)^2}{L}.
\end{aligned}$$

In the last step, we simply noted that the quadratic in l is minimal when $l = 2(n-1)/L$. □

The lower bound on $\inf_{i \in A} \sup_{j \in A, j \neq i} \mathbf{E}T_{i,j}$ shown in Lemma 3.3 can be attained by octopus trees; these are trees with L leaves at the end of tentacles of length $(n-1)/L$ each, where by assumption $n-1$ is a multiple of L.

Proof of theorem 3.1. By Lemmas 3.2 and 3.3, for $n \geq 3$,

$$\mathbf{E}T \geq \inf_{\substack{n-1 \geq u \geq 2 \\ u \text{ integer}}} \max\left(2(n-1)H_{u-1}, \frac{2(n-1)^2}{u}\right).$$

The supremum is reached for a sequence $u^* \sim n/\log n$. Thus, $\mathbf{E}T \geq 2n \log n - O(n \log\log n)$. □

3.5. TREES WITH BOUNDED DEGREE: PROOF OF THEOREM 3.2

Set $d = \Delta - 1$ where Δ is the maximal degree. The edges of the tree will be oriented away from the starting node, which we shall call the root. As is customary with rooted trees, we can define the height of a node as the distance (path length) from a node to the furthest leaf in its subtree. We take an integer s, which depends in a controlled manner on n, and consider the set A_s consisting of all vertices of height s. With each vertex $a_i \in A_s$, associate one leaf b_i at distance s in the subtree of a_i. Let $N = |A_s|$ be the number of (a_i, b_i) pairs. We will see later that by our choice of s, we have $N > 1$.

We first claim that for any b_i and for any a not in the subtree rooted at a_i,

$$\mathbf{E}T_{a,b_i} \geq s\left(2n - \frac{2d^{s+1}}{d-1} - 1\right).$$

Indeed, the subtree rooted at each such a_i has at most $1 + d + d^2 + \cdots + d^s = (d^{s+1}-1)/(d-1)$ vertices. Next, apply Moon's identity (3.2). Remark that the N subtrees rooted at the a_is are disjoint. Thus, by choosing the node a on the unique path from a_i to a_j,

$$\mathbf{E}T_{b_i,b_j} \geq s\left(2n - \frac{2d^{s+1}}{d-1} + s\right).$$

Let us consider the set A of all vertices b_i. We can apply Lemma 3.1

and conclude that, no matter where we start the random walk,

$$\mathbf{E}T \geq s\left(2n - \frac{2d^{s+1}}{d-1} + s\right)H_{N-1}. \tag{3.7}$$

This bound is the backbone of the proof. However, it is only useful when $d^s = o(n)$ (which we will take care of later) and when N is at least equal to a polynomially increasing function of n.

A second lower bound can be obtained from Lemma 3.3. If L is the number of leaves in the tree, we know that $\mathbf{E}T \geq 2(n-1)^2/L$. From the preceding construction, we also have $L \leq Nd^{s+1}/(d-1)$, so that

$$\mathbf{E}T \geq \frac{2(n-1)^2(d-1)}{Nd^{s+1}}. \tag{3.8}$$

This lower bound is a decreasing function of N and should be useful for small values of N. We are free to choose s, but not N. It is to our advantage to choose $s \sim \log n/(2\log d)$ and $s \leq \log n/(2\log d)$. Note that $d^{s+1} \leq d\sqrt{n}$. Combining (3.7) and (3.8) we thus obtain

$$\begin{aligned}\mathbf{E}T &\geq \inf_{N \text{ integer}} \max\left(s\left(2n - \frac{2d^{s+1}}{d-1} + s\right)H_{N-1}, \frac{2(n-1)^2(d-1)}{Nd^{s+1}}\right)\\ &\geq \inf_{N \text{ integer}} \max\left(s\left(2 - \frac{2d}{(d-1)\sqrt{n}}\right)n\log N, \frac{2(n-1)^2(d-1)}{Nd^{s+1}}\right).\end{aligned}$$

The lower bound is of the form $\inf_{N \text{ integer}} \max(a\log N, b/N)$ for positive numbers a, b. For $N \geq (b/a)/\log(b/a)$, we have $\mathbf{E}T \geq a\log(b/a) - a\log\log(b/a)$, while for $N \leq (b/a)/\log(b/a)$, we have $\mathbf{E}T \geq a\log(b/a)$. Asymptotically, $b \sim n^{3/2+o(1)}$ and $a \sim n\log n/\log d$. Thus, $b/a \sim n^{1/2+o(1)}$ and

$$\mathbf{E}T \geq \frac{(1+o(1))n\log^2 n}{2\log d}. \quad \square$$

3.6. OPTIMALITY OF THE BOUNDS: STARS AND COMPLETE *d*-ARY TREES

A *star* is a tree in which one node is connected to the $n-1$ other nodes. It is easy to see that the expected waiting time between the first time k leaves have been visited and the first time $k+1$ leaves have been visited is $2(n-1)/(n-1-k)$, when $k \geq 1$, and is 1 when $k=0$, provided that the random walk is started at the central node. Summing this yields

$$\mathbf{E}T = 1 + 2(n-1)H_{n-2},$$

where we assume that $n \geq 2$ and $H_0 = 0$. This is asymptotic to $2n \log n$.

Consider finally a *complete d-ary tree* with $n = \Sigma_{i=0}^{r} d^i$ nodes; that is, it has $r+1$ full levels of nodes. Also, the maximal degree is $\Delta = d+1$. From Moon's formula (3.2), it is easy to see that $\mathbf{E}T_{i,j}$ is maximal when both i and j are leaves, and by symmetry, $\mathbf{E}T_{i,j} = \mathbf{E}T_{j,i}$ when i and j are leaves. It is known that for any tree $\mathbf{E}T_{i,j} + \mathbf{E}T_{j,i} = 2(n-1)d(i,j)$. By the preceding symmetry, $\max_{i,j} \mathbf{E}T_{i,j} = (n-1)\max_{i,j} d(i,j) = 2(n-1)r$. By Matthews' upper bound (similar to the lower bound of Lemma 3.1, with max instead of min) and the fact that $n = (d^{r+1}-1)/(d-1)$ and the number of leaves $L = (n(d-1)+1)/d$, no matter how $d \geq 2$ varies with n, we see that starting at a leaf,

$$\mathbf{E}T \leq 2r(n-1)H_{L-1}.$$

By adding an r^2 term to the preceding bound, one obtains for the walk started at any node (see Section 3.2)

$$\mathbf{E}T \leq \frac{2n \log^2 n}{\log d} + O(n \log_d n).$$

Modulo a constant factor of 4, this agrees with the lower bound of Theorem 3.2, for any fixed d (Δ). We conclude that the dependence of the lower bound on n and Δ in Theorem 3.2 is optimal.

REFERENCES

Aldous, D. J. (1983). On the time taken by random walks on finite groups to visit every state. *Z. Wahrsch. Verw. Gebiete* **62** 361–393.

Aldous, D. (1989a). An introduction to covering problems for random walks on graphs. *J. Theoret. Probab.* **2** 87–89.

Aldous, D. (1989b). Hitting times for random walks on vertex-transitive graphs. *Proc. Cambridge Philos. Soc.* **106** 179–191.

Aldous, D. J. (1989c). Lower bounds for covering times for reversible Markov chains and random walks on graphs. *J. Theoret. Probab.* **2** 91–100.

Aleliunas, R., Karp, R. M., Lipton, R. J., Lovász, L., and Rackoff, C. (1979). Random walks, universal transversal sequences, and the complexity of maze problems. In *Proceedings of the 20th Annual Symposium on the Foundations of Computer Science*. IEEE, 218–223.

Broder, A. Z. and Karlin, A. R. (1989). Bounds on the cover time. *J. Theoret. Probab.* **2** 101–120.

Devroye, L. and Sbihi, A. (1990). Random walks on highly symmetric graphs. *J. Theoret. Probab.* **3** 497–514.

Diaconis, P. (1988). *Group Representation in Probability and Statistics. Lecture Notes* **11**. IMS, Hayward, CA.

Göbel, F. and Jagers, A. A. (1974). Random walks on graphs. *Stochastic Process. Appl.* **2** 311–336.

Kahn, J. D., Linial, N., Nisan, N., and Saks, M. E. (1989). On the cover time of random walks on graphs. *J. Theoret. Probab.* **2** 121–128.

Matthews, P. (1987). Mixing rates for a random walk on the cube. *SIAM J. Algebraic and Discrete Methods* **8** 746–752.

Matthews, P. (1988a). Covering problems for Brownian motion on spheres. *Ann. Probab.* **16** 189–199.

Matthews, P. (1988b). Covering problems for Markov chains. *Ann. Probab.* **16** 1215–1228.

Matthews, P. (1989). Some sample path properties of a random walk on the cube. *J. Theoret. Probab.* **2** 129–146.

Mazo, J. E. (1982). Some extremal Markov chains. *Bell Syst. Tech. J.* **61** 2065–2080.

Moon, J. W. (1973). Random walks on random trees. *J. Australian Math. Soc.* **15** 42–53.

Pearce, L. H. (1980). Random walks on trees. *Discrete Mathematics* **30** 269–276.

Zuckerman, D. (1989). Covering times of random walks on bounded degree trees and other graphs. *J. Theoret. Probab.* **2** 147–157.