Random Cayley Digraphs and the Discrete Logarithm

Extended Abstract

Jeremy Horwitz[1] and Ramarathnam Venkatesan[2]

[1] Stanford University, Stanford, CA 94305, USA
horwitz@cs.stanford.edu
[2] Microsoft Research, Redmond, WA 98052, USA
venkie@microsoft.com

Abstract. We formally show that there is an algorithm for DLOG over all abelian groups that runs in expected optimal time (up to logarithmic factors) and uses only a small amount of space. To our knowledge, this is the first such analysis. Our algorithm is a modification of the classic Pollard rho, introducing explicit randomization of the parameters for the updating steps of the algorithm, and is analyzed using random walks with limited independence over abelian groups (a study which is of its own interest). Our analysis shows that finding cycles in such large graphs over groups that can be efficiently locally navigated is as hard as DLOG.

1 Introduction

The Discrete Logarithm Problem (DLOG) defined over abelian groups plays a fundamental role in cryptography as a basis for many primitives (*e.g.*, Diffie-Hellman key exchange, DSS, and ElGamal signatures). The algorithms to find DLOG fall into two types: the generic, black-box, exponential-time algorithms that use only the group structure (*e.g.*, baby-step giant-step and Pollard rho) and the domain-specific subexponential algorithms (*e.g.*, index calculus methods), which are not yet known to exist for groups over elliptic curves. Because of its generality and that it uses a very small amount of space, Pollard rho [8] is practically and theoretically important.

Surprisingly, there is no formal analysis of the classic Pollard rho without random-oracle assumptions. The standard analysis is heuristic: it approximates the rho walk with a totally random walk (*i.e.*, a walk which at every step randomly and independently jumps to another group element) and then infers the existence of a cycle of length $\sqrt{p}$ using the birthday paradox. But, in reality, the walk is far from random: the algorithm only makes a *deterministic walk* (which is crucial for Floyd's algorithm to find a cycle using only a small amount of space) on a 3-regular directed graph over $\mathbb{Z}_p^\times$ that is constructed semi-randomly. By using a random oracle for the moves to the *neighboring* nodes, Teske [11, 12] has analyzed both the original Pollard rho as well as more general k-regular graphs (for $k \geq 3$); for $k \geq 6$ she derives an $O(\sqrt{|G|})$ bound for finite abelian groups

using a result of Hildebrand. Lack of independence between moves creates difficulty in analysis, especially since the move from a node z depends on (the label of) the node z. Earlier, Bach [2] studied Pollard rho for factoring and showed that the probability of success for the rho method is $c(\log^2 p)/p$ (for some $c > 0$), which is only slightly better than the obvious bound of $1/p$.

We explicitly introduce randomness by slightly modifying the algorithm and then base our treatment on random walks on Cayley graphs over abelian groups. Recall that a k-regular Cayley digraph (directed graph) on a group G has a set S of k generators. Its set of nodes is G and its edges are formed by connecting every α in G by a directed edge to αg_i, for every $g_i \in S$. To solve for $y = g^x$ in G, we construct S with equal number of random powers of y and g, and construct a navigation function $h_E(\alpha)$ (for $\alpha \in G$) which maps into $\{1, 2, \ldots, k\}$ (for $k = O(\log p)$), by picking a random polynomial over a suitable extension of $\mathbb{F}_2$ and truncating its output. We start at some $z_0 \in G$ and move from z_i to z_{i+1} by multiplying z_i by the generator in S with index $h_E(z_i)$. Finally, we look for a collision in the z_is.

Firstly, we show that our modified algorithm, which is a *random* walk *with limited independence* on a *random* Cayley graph (*i.e.*, S is a random k-subset of the group), finds the DLOG in optimal time (up to logarithmic factors). We note that a random choice of generators is important for two reasons: first, to show that the rho algorithm produces a nontrivial relationship (Theorem 1). Second, to guarantee the existence of Cayley graphs over *any* abelian group with an underlying Markov chain that rapidly mixes (without randomization, no such universal construction is known); the rapid-mixing property in turn is crucial for removing the dependence on a random-oracle assumption. This complements the result of Shoup [10] who showed that *generic* algorithms for DLOG must take at least $\sqrt{|G|}$ steps. It would be interesting to know if random walks exploiting specific group properties yield faster algorithms.

This analysis also allows us to show that finding nontrivial cycles (*i.e.*, smaller than the group order) in random Cayley graphs over an abelian group G of order p is as hard as solving DLOG over G. These graphs are *succinctly presented* in the sense that they are defined by simple rules for moving from a node to its neighbors; they are, however, too huge to be explicitly stored. Our succinct graphs have girth (*i.e.*, the length of shortest cycle) $O(\log p)$; however, to computationally efficient algorithms, the girth appears to be exponential in $\log p$. This allows for the construction of secure hash functions. A significantly longer version of this paper (including experimental results which exhibit practical run times and parallel our theoretical results) will be available from the authors.

2 Preliminaries and Statement of Results

In this section we present relevant definitions, motivation, and statements of our results. Our study is from the point of view of path finding or navigating in exponentially large graphs that have simple rules for moving from one node to another. We assume the constraint that one has a limited amount of memory.

2.1 Cayley Digraphs

In view of the Pohlig-Hellman result on DLOG [7], we consider only prime-order groups; we denote the order of the group discussed in this paper by p. In such a group, every element except the identity is a generator. For notions related to graph theory and random walks, we refer the reader to [4].

Let G be a multiplicative abelian group of order p and $S = \{g_1, g_2, \ldots, g_{2n}\} \subseteq G$ (we write $2n$ since $|S|$ will always be even). A *Cayley digraph generated by* S is denoted by $\mathcal{G}(G, S) = \mathcal{G} = (V, E)$ and has the set of nodes $V = G$ and the set of edges $E = \{(g, gg_i) \ : \ g \in G, g_i \in S\}$. (Most papers study undirected versions where, if $g \in S$, then $g^{-1} \in S$, and may additionally assume that the unit $1 \in S$ (*i.e.*, all nodes have self loops); we cannot assume either of these conditions.) A *path* of length t is a sequence $(v_0, v_1, \ldots, v_t)$ with every $(v_i, v_{i+1}) \in E$. A path is a *cycle* if it also satisfies $v_t = v_0$. In this paper, our main parameter is $2n = O(\log p)$, where p is large enough to make DLOG hard, while path lengths t can be exponentially large in $2n$. Since G is abelian, paths (and cycles) of length t admit succinct representations of size $O(2n \log t)$ as: given a path (or cycle), we write it as $X = (x_1, x_2, \ldots, x_{2n}) \in \mathbb{N}^{2n}$ where x_i is the number of the edges of the form (g, gg_i) in the path. Since $g^p = 1$ for any $g \in G$, cycles occur in $\mathcal{G}$ trivially; we will be interested only in *nontrivial* cycles having length $t < p$. We assume that all our paths and cycles are nontrivial and have length $t \leq \Lambda$ for a fixed constant $\Lambda = (\log^{O(1)} p)\sqrt{p} = o(p)$. Having $t = o(p)$ avoids wraparound problems even when we add the lengths of a constant number of paths.

Succinct Graphs. We say that $\mathcal{G} = \mathcal{G}(G, S)$ is a *random Cayley digraph over* G if the elements of S are picked from G randomly and independently. By a navigation algorithm for a graph (V, E), we mean some algorithm to compute $f(u, i) = v$, where v is the ith ordered neighbor (under some predefined ordering) of the node u. If the graph is d-regular, then it can be edge colored, for example, with d colors, and we set $f(u, i) = v$ if the edge (u, v) has the ith color. A graph is *succinctly presented* (or *succinct*) if there is a navigation algorithm $f(u, i)$ that runs in time $|u|^{O(1)}$, where $|u|$ is the length of its label. We note that Cayley graphs over $\mathbb{Z}_p^\times$ are succinct because one can take the standard binary representation of integers as the label and compute $f(\alpha, i)$ as αg_i, where g_i is the ith generator in S. Another example is the k-dimensional hypercube with vertex set $\mathbb{Z}_2^k$ with vertices connected if and only if they differ in exactly one co-ordinate.

2.2 Limited Independence

A sequence of random variables $z_0, z_1, \ldots, z_t$ is called *m-wise independent* if any subsequence of at most m variables is independent; in our case they will be uniformly distributed. A 2-wise independent sequence is also called a *pair-wise independent* sequence. A function $f(x)$ is *m-wise independent* if, for any sequence of inputs $\{z_i\}_{i=0}^t$, the sequence $\{f(z_i)\}_{i=0}^t$ is m-wise independent. We will randomly choose polynomials of degree $m - 1$ defined over an extension

field of $\mathbb{F}_2$; notice that these polynomials are m-wise independent. Indeed, given $\boldsymbol{z} = (z_0, z_1, \ldots, z_{m-1})$ with distinct x_i and $\boldsymbol{y} = (y_0, y_1, \ldots, y_{m-1})$, one can find a polynomial with $f(z_i) = y_i$: solve the equation $\boldsymbol{y} = V\boldsymbol{f}$ (where $V = (z_i^j)_{0 \le i,j < m}$ is a Vandermonde (and, hence, invertible) matrix) for $\boldsymbol{f} = (f_0, f_1, \ldots, f_{m-1})$ and set $f(x) = \sum_{i=0}^{m-1} f_i x^i$. We note that if we truncate each of the outputs of $f(z_i)$ (to some number of least-significant bits), we will still have an m-wise independent sequence. To see this, note that in this case we are given only the truncated bits of entries in y and we may arbitrarily extend them to fully specify a vector y and proceed as before. By incrementing if necessary, we shall assume that m is even. We now recall the following tail inequality [3] for the sum Z of a sequence of m-wise independent variables taking values in $[0, 1]$. Set $\mu := \mathrm{E}[Z]$ and let $a > 0$. Then, we have $\Pr[|Z - \mu| \ge a] \le 8((m\mu + m^2)/(a^2))^{m/2}$.

2.3 Finding Cycles in Succinct Graphs and DLOG

While finding paths and cycles efficiently in the usual graphs is well-understood, finding paths and cycles in succinct graphs using only small space may be hard (though, in some cases, such as hypercubes, this is trivial). Indeed, one may view the classic Pollard rho for solving $y = g^x$ as a method both to define (using y and g) a succinctly presented graph together with its navigation algorithm h and to find a cycle in the succinct graph (then solve a linear equation to find DLOG). Our modification to Pollard rho differs only in the definition step and is aimed at bounding the run time and the success probability in the cycle-finding step.

Pollard Rho Algorithm. Let $g \neq 1$ be fixed. Given $y \in G = \langle g \rangle$, the task is to find x such that $y = g^x$. The algorithm (in some simple way) partitions G into three approximately equal-sized sets T_1, T_2, and T_3 (taking care that $1 \notin T_3$). Now, define the navigation algorithm $h_\rho \colon G \to G$ as: $h_\rho(z) = zg$ for $z \in T_1$, $h_\rho(z) = zy$ for $z \in T_2$, and $h_\rho(z) = z^2$ for $z \in T_3$.

Starting with some fixed $z_0 = g^r$, construct a sequence $\{z_i\}_{i=0}^t$ with $z_{i+1} = h_\rho(z_i)$ until a collision occurs (*i.e.*, $z_u = z_v$ for some $u \neq v$). Then use Floyd's algorithm to find a cycle, which yields a relationship of the form $bx = a + rc \bmod p$.

Remark 1. It is crucial that h above is deterministic if one wants to preserve the main advantages of small space and being able to avoid exhaustive search over the entire group. As noted earlier, in standard analysis for the rho method, one treats the z_is as if they were random and independent (equivalently, one treats the graph as a complete graph and the navigation function h as if it were chosen randomly from the set of all functions from G to G) and uses the birthday paradox to bound $t = O(\sqrt{p})$. Also, we note that there is no formal guarantee for the probability that b^{-1} exists $(\mathrm{mod} p)$, which is required to finally discover x.

Cayley Rho Algorithm. Fix a cyclic group G (of order p) and a generator $g \in G$ with respect to which we will solve DLOG. Where $2n$ is the size of $S \subseteq G$ (the set of generators for the Cayley graph), we, for convenience, assume that $2n$ is a power of 2. (Experiments show that when $2n$ is at least $4 \log_2 p$, the Cayley rho algorithm performs better than the Pollard rho; further details appear in the full version of this paper.) We fix an extension field $E/\mathbb{F}_2$ with $[E : \mathbb{F}_2] = 3\lceil \log p \rceil$ (unless otherwise stated, we always mean the base-2 logarithm) and set $d := \nu\lceil \log p \rceil$, where ν is a small constant. Define H to be the set of all degree-d polynomials from E to E. Let $y = g^x$ be given. We construct an algorithm $\mathcal{C}(y)$ as follows:

1. **Defining the succinct graph**: Randomly choose $r_1, r_2, \ldots, r_n \in \mathbb{Z}_p$ and $s_1, s_2, \ldots, s_n \in \mathbb{Z}_p$. Then let $(g_1, g_2, \ldots, g_{2n})$ be a random permutation of $(g^{r_1}, g^{r_2}, \ldots, g^{r_n}, y^{s_1}, y^{s_2}, \ldots, y^{s_n})$. Let $S := \{g_1, g_2, \ldots, g_{2n}\}$ and $\mathcal{G} = \mathcal{G}(G, S)$ be the *random* Cayley graph generated by S over G.
   Initializing the navigation algorithm: We randomly choose and fix a polynomial $h' \colon E \to E$ from H.
   Computing $h(\alpha)$: Given $\alpha \in G$, we use a standard $\lceil \log p \rceil$-bit binary representation of α and pad it with a suitable prefix of zeros to get $\alpha' \in E$. Define $h_E \colon E \to \{1, 2, \ldots, 2n\}$ so $h_E(\alpha')$ is the $\log_2(2n)$ least-significant bits of binary representation of $h'(\alpha')$. Define $h \colon G \to G$ by $h(\alpha) = \alpha g_c$, where $c := h_E(\alpha')$.
2. As in Pollard rho, we can use a procedure $\mathcal{A}(\mathcal{G})$ which outputs a cycle $X = (x_1, x_2, \ldots, x_{2n})$ in $\mathcal{G}$ (*i.e.*, $\prod g_i^{x_i} = 1$). If the cycle is trivial, we repeat the entire algorithm; else we solve a linear equation (described below). (In case the equation cannot be solved (*i.e.*, it is $0x = 0$), $\mathcal{C}$ must be restarted.)

By abusing notation we may write $h \in \mathsf{H}$ or $h_E \in \mathsf{H}$ (really only $h' \in \mathsf{H}$).

2.4 Notation for Walks

Throughout this paper, we utilize a number of functions (particularly h, h_ρ, h_E, and h_2) to describe our random walks, primarily to simplify our analysis and to make our notation more convenient for both the authors and the readers.

The transition function $h \colon G \to G$ is most similar to a standard transition function for a Markov chain: it takes as input the current state and returns the next state. (The method for its construction is explained in Section 2.3.) The function $h_\rho \colon G \to G$ represents the Pollard rho transition function, which we only mention in a referential context. We use $h_E \colon E \to \{1, 2, \ldots, 2n\}$ (as described in Section 2.3) as an intermediate construction en route to building h. We overload h_E to allow $h_E \colon G \to \{1, 2, \ldots, 2n\}$ (where, in these instances, h_E appropriately pads a natural binary representation of its input with zeroes in order to apply h_E as usual (as described in Section 2.3)). A technical necessity used only in Section 6, $h_2 \colon G \times \mathbb{N} \to G$ is constructed from a function h_2' (which is randomly chosen from a set of bivariate polynomials) just as h is constructed from h'. h_2 is constructed so that when $\gamma_i = 0$, $h_2(z_i, \gamma_i) = h(z_i)$. The probability (over

choice of h_2) that there is *no* collision in the walk defined from h_2 is equal to the probability (over choice of h) that there is *no* collision in the walk defined from h. This result is discussed in greater detail in Lemma 13.

2.5 Main Results

Theorem 1 (Near-Optimal Convergence). *Let the Cayley rho algorithm $\mathcal{C}$ take $\widetilde{O}(\sqrt{p})$ (*i.e., *$O(\sqrt{p})$ up to factors of $\log p$) moves on the graph. Then, (a) the probability (over the random choices made by $\mathcal{C}$) of a cycle of length $\widetilde{O}(\sqrt{p})$ occurring is a positive constant and (b) when the cycle-finding algorithm $\mathcal{A}$ returns successfully, $\mathcal{C}$ solves* DLOG *with probability at least $(2n^2)^{-1}$; thus the expected number of calls to the cycle-finding algorithm $\mathcal{A}$ is at most $2n^2$.*

Corollary 1 (DLOG $\preceq$ Cycle Finding). *Finding cycles in random Cayley graphs over G is as hard as solving* DLOG *on G.*

The corollary follows from part (b) of the theorem, since it applies to *any* cycle-finding algorithm $\mathcal{A}$. To prove Theorem 1(a), we use the next theorem.

Theorem 2 (Rapid Mixing). *Let $\mathcal{G}$ be a random Cayley digraph over an abelian group G of prime order p and let $z_0 \in G$ be arbitrary. Starting from z_0, let the endpoint of a t-step (totally independent) random walk be z_t. If $t \geq 2 \log p$, then, for any $\alpha \in G$, $|\Pr[z_t = \alpha] - 1/p| \leq p^{-2}$.*

Rapid mixing of Cayley graphs is well-studied; however, we could not find a reference for the case of Cayley digraphs with both $O(\log p)$ generators and no self loops that states the required bound ($O(p^{-2})$ rather than $O(1)$) on the deviation from the uniform. However, our proof is simple, and all the required Markov chain properties are derived directly from Lemma 2. Yet, the theorem is insufficient for us to prove results unconditionally; if we assumed that the navigation function is a purely random function, then we would get the result using the above theorem. It is simple to show, using elementary matrix methods, the following: starting at an arbitrary z_i, if a purely random walk on an expander converges to an almost-uniform distribution $\mu\colon G \to [0,1]$ in τ steps (*i.e.*, the node $z_{i+\tau}$ is almost-uniformly distributed), then, for any $t > \tau$, the distribution of z_{i+t} remains almost-uniformly distributed. This need not be true when the walk steps are correlated. However, using that G is abelian, we can show that the walk remains almost-uniformly distributed. This result appears to be the first of its type and is of interest by itself.

3 Proof of Theorem 1(b)

Proof. Let $\mathcal{A}(\mathcal{G})$ find a cycle of length $t = o(p)$. From this cycle, we get an equation of the form $z_0 = z_0 \prod_{i=1}^{2n} g_i^{w_i}$, for some initial node $z_0 \in G$ and $0 \leq w_i \leq t$, where $\sum_{i=1}^{2n} w_i = t$. From the definition of the g_i, we see that $\prod_{i=1}^{n} g^{-r_i w_i} = \prod_{i=1}^{n} y^{s_i w_{n+i}}$. Hence, $-\sum_{i=1}^{n} r_i w_i = x \sum_{i=1}^{n} s_i w_{n+i} \pmod p$, which yields x

unless $\sum_{i=1}^{n} r_i w_i = \sum_{i=1}^{n} s_i w_{n+i} = 0 \pmod p$. The probability that we *cannot* find x (because the aforementioned sums are zero) is bounded above by $1 - \frac{1}{n^2+1}$ (from Lemma 1), so we expect to rerun $\mathcal{A}$ at most $n^2 + 1 \leq 2n^2$ times. □

Lemma 1. *Let $k_1, k_2, \ldots, k_{2n} \in \mathbb{Z}_p$ be such that for $i \neq j$, $k_i \neq \pm k_j \pmod p$. Fix $t = o(p)$ and randomly choose $\sigma \in S_{2n}$. An adversary, given the k_i and t (but not σ), chooses $0 \leq w_i \leq t$ (not all zero) and we say the adversary wins if $\sum_{i=1}^{n} k_{\sigma(i)} w_i = \sum_{j=n+1}^{2n} k_{\sigma(j)} w_j = 0 \pmod p$. Then the probability (over choices of σ) that the adversary wins is at most $1 - \frac{1}{n^2+1}$.*

4 The Markov Chain Induced by G

We define our random walk on $\mathcal{G}$ as follows: starting at an initial node z_0, one picks, uniformly at random, one of the outgoing edges (say, $(z_0, z_0 g_i)$) and moves to the opposite node (*i.e.*, $z_1 := z_0 g_i$). Then we iterate this step, using independent coin flips at each node. The induced Markov chain (which we denote by **MC**) has the transition matrix M with entries $m_{\alpha\beta} = 1/2n$ if there is an edge from the node α to node β (else it is zero); the adjacency matrix $A(\mathcal{G})$ has $a_{\alpha\beta} = 2n m_{\alpha\beta}$. Our graphs are directed and we must work out many of their properties from scratch. We point out that existing literature on rapid mixing cannot be directly used for a variety of reasons: our graphs are directed; we cannot add self loops to guarantee aperiodicity; we need to derive quantitative bounds on the deviation (from the uniform distribution); and, most importantly, our walks are not entirely independent. Here, matrix theory cannot be applied at all, and we utilize a probabilistic argument that capitalizes on the abelian property and shows (in this case) that if a purely random walk is convergent, then so is the related limited-independence random walk.

4.1 Conventions and Markov Chain Preliminaries

Conventions we use include denoting the walk by $z_0, z_1, \ldots, z_t$ and defining a function $\boldsymbol{c}$: $\{0, 1, \ldots, t-1\} \to \{1, 2, \ldots, 2n\}$ so $z_{i+1} = z_i g_{\boldsymbol{c}(i)}$. Notice that the random walk is completely specified by $\boldsymbol{c}$; as such, we often refer to $\boldsymbol{c}$ as a walk.

Let $\Omega_t := \{(x_1, x_2, \ldots, x_{2n}) \in \mathbb{N}^{2n} \ : \ \sum_{i=1}^{2n} x_i = t\}$. For $1 \leq j \leq 2n$, set $y_j := \left|\boldsymbol{c}^{-1}(j)\right|$. In other words, the random walk induced by $\boldsymbol{c}$ picks each generator g_i a total y_i times during the t-step random walk. Notice that there is a well-defined map $\boldsymbol{c} \mapsto Y = (y_1, y_2, \ldots, y_{2n})$, which we will write as $\psi(\boldsymbol{c}) = Y$. Let $\lambda(Y) = \Pr_{\boldsymbol{c}}[\psi(\boldsymbol{c}) = Y]$ and $\mu(Y) = |\Omega_t|^{-1} = \binom{t+2n-1}{2n-1}^{-1}$.

The group S_{2n} of permutations of $\{1, 2, \ldots, 2n\}$ acts on Ω_t and we denote its orbits by $T_1, T_2, \ldots, T_N$. We note that $Y = (y_1, y_2, \ldots, y_{2n})$ and Y' both belong to the same orbit T_j if and only if Y is a permutation of Y' (*i.e.*, $Y' = (y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(2n)})$ for some $\sigma \in S_{2n}$). Clearly this induces an equivalence relation, and we write $Y \sim Y'$ if and only if $Y, Y' \in T_j$ for some j. As usual, we say that T_j is the orbit of Y. An important fact here is that if $Y \sim Y'$, then

$\lambda(Y) = \lambda(Y')$, since the sequence $\{\boldsymbol{c}(i)\}_{i=0}^{t-1}$ and $\{\sigma(\boldsymbol{c}(i))\}_{i=0}^{t-1}$ have the same λ probability for any $\sigma \in S_{2n}$.

We first prove a preliminary lemma that is analogous to a result of Erdős and Rényi [6], who showed that random subproducts of the (uniformly-chosen) generators are almost-uniformly distributed. Our method allows one to quantify the dependence of the quality of this distribution in terms of the walk length, as well as to show many properties of the random $\mathcal{G}$.

We fix $g \neq 1$ so that $G = \langle g \rangle$. Recall that we represent a path X of length t_X by a $2n$-tuple of nonnegative integers $X = (x_1, x_2, \ldots, x_{2n})$ such that $\sum_i x_i = t_X$. We say that two distinct nonzero paths X and Y are *linearly correlated* if, as vectors, they are scalar multiples of each other (*i.e.*, $t_X Y = t_Y X \pmod p$). Otherwise, they are said to be *linearly uncorrelated*. For example, $X \neq Y$ will be linearly uncorrelated if $t_X = t_Y$ or if they are binary vectors. In addition, if $\max\{t_X^2, t_Y^2\} < p$, it is sufficient that $t_X Y = t_Y X$ holds over $\mathbb{Z}$. Note that if two vectors are linearly *independent* over $\mathbb{F}_p$, they will be linearly *uncorrelated* in our sense, but the converse need not hold.

We consider pairs of linearly uncorrelated paths and conclude that random Ss induce a pairwise-independent function on them. For a given random S, with $g_i := g^{\alpha_i}$, we define a mapping ϕ_S to take a path X to the node $\prod_i g^{\alpha_i x_i}$. Without loss of generality, we may assume that the starting point of the walk is unity. As such, $\phi_S(X)$ is the endpoint of the walk specified by X. We will heavily rely on Corollary 2 below, which is immediate from Lemma 2.

Remark 2. We will assume that S is formed by picking $2n$ elements randomly and independently from G. These need not be distinct, so S can be a multi-set. Our main analysis requires only a lower bound on the size of S. By a tiny increase in the number of elements picked, one can be assured that S has $2n$ elements with probability at least $1 - p^{-3}/4$. Constructing S can be viewed as randomly choosing an $S \in \mathsf{S} := \{S \subseteq G \,:\, |S| = 2n\}$.

Also, note that the next lemma allows the case $1 \in S$.

Lemma 2 (Pairwise Independence). *In a Cayley digraph over a group of prime order p, let X and Y be two arbitrary distinct nonzero linearly uncorrelated paths of lengths at most Λ. Then, the mapping $\phi_S(X) := \prod g^{\alpha_i x_i}$ is a pairwise-independent mapping,* i.e.*, for any $a, b \in \mathbb{Z}_p$,*

$$\Pr_S\left[\phi_S(X) = g^a \wedge \phi_S(Y) = g^b\right] = \Pr_S\left[\phi_S(X) = g^a\right] \Pr_S\left[\phi_S(Y) = g^b\right] = \frac{1}{p} \cdot \frac{1}{p} \ .$$

Corollary 2. *(a) On $A \subseteq \Omega_t$, for equal-length ($t \leq \Lambda$) paths, the mapping $\phi_S(X)$ is pairwise-independent. (b) The restriction of $\phi_S(X)$ to the set $B := \{(x_1, x_2, \ldots, x_{2n}) \,:\, x_i \in \{0, 1\},$ not all zero$\}$ is a pairwise-independent map. In this case, $X \mapsto \phi_S(X)$ is a subset-product map on nonempty sets of generators.*

Corollary 3. *Let $\mathsf{S}_0 \subseteq \mathsf{S}$ be such that $1 - \frac{|\mathsf{S}_0|}{|\mathsf{S}|} \leq \varepsilon$ (with $\varepsilon \leq p^{-2}$). Then,*

$$1 - 2\varepsilon(1 - 1/p) \leq \Pr_S[\phi_S(X) = g^a | S \in \mathsf{S}_0] / \Pr_S[\phi_S(X) = g^a] \leq 1 + 2\varepsilon \ .$$

4.2 Properties of MC

Notice that (unless $S = \{1\}$) the elements of S generate G and, thus, the Cayley digraph $\mathcal{G}$ is strongly connected (*i.e.*, **MC** is irreducible). For any irreducible Markov chain, by the Perron-Frobenius theorem, the adjacency matrix has 1 as the maximal eigenvalue; additionally, this eigenvalue has multiplicity one. To guarantee a stationary distribution of the chain, it must also be aperiodic (stated as Lemma 3). Note that the group structure imposes that the in-degree and the out-degree of any node are the same (both equal to $|S|$), making M doubly stochastic (*i.e.*, every column sums to one, as does every row). Hence, if **MC** has a stationary distribution, it must be the uniform distribution. In addition to allowing us to conclude that **MC** has a unique, uniform stationary distribution, the proof of the following lemma also yields a $\Theta(\log |G|)$ bound for both the diameter and the girth of almost every graph.

Lemma 3. ***MC*** *is aperiodic for all but a negligible fraction of choices of* S*.*

5 Rapid Mixing (Proof of Theorem 2)

We recall standard definitions. The *boundary* of a $D \subseteq V$ is the set $\partial D = \{v \in V \ : \ v \notin D$ and v has incoming edge from some node in $D\}$.

If $U \subseteq V$ and for every subset W of U we have $|\partial W| \geq \varepsilon |W|$, then U is then called *ε-expanding*. We call the subgraph induced by an ε-expanding subset an *ε-expanding graph*. The entire graph $\mathcal{G} = (V, E)$ is called an *ε-expander* if every subset of size at most $\frac{|V|}{2}$ is ε-expanding.

Normally, ε is taken to be a constant as the size of $\mathcal{G}$ grows; one shows that on such expanders a random walk rapidly mixes in the sense that it reaches a distribution exceptionally close to its stationary (uniform) distribution in $O(\log p)$ steps. Cayley graphs and general expanders are the subject of extensive literature and the reader may wish to consult [1, 5, 9] as well as the short survey in the full version of this paper.

5.1 Outline of the Proof of Theorem 2

First, Lemma 4(a) will allow us to conclude that almost all choices of the set S of generators are *good* in the sense that:

(†) for a sufficiently large walk length t, for $\alpha \in G$ and $A \subseteq \Omega_t$ with $|A| \geq p^5$, $\left| \frac{|\phi_S^{-1}(\alpha) \cap A|}{|A|} - \frac{1}{p} \right| < \frac{1}{p^2}$.

Thus, if we pick a random Y from A, the endpoint $\phi_S(Y)$ will be almost-uniformly distributed. However, a random walk $\boldsymbol{c}$ does not induce a uniform distribution on the tuples $Y \in A$ for arbitrary A, but, if A is an orbit in Ω_t under the action of S_{2n}, the induced distribution $Y \mapsto \Pr_{\boldsymbol{c}}[Y | Y \in A]$ is indeed uniform (within a fixed orbit A). To use (†), we will need $|A| \geq p^5$; however, there are many small orbits (*e.g.*, the orbit of $Y = (s, s, \ldots, s)$). Fortunately, Lemma 5 will help complete the proof by showing the following property:

(‡) with overwhelming probability, a random walk $\boldsymbol{c}$ generates a $\psi(\boldsymbol{c}) = Y$ whose orbit under S_{2n} has, for *every* S, size at least p^5.

5.2 Proof of Theorem 2

Lemma 4. *Fix $\alpha \in G$. (a) If $A \subseteq \Omega_t$ with $|A| \geq p^5$, then, for all but a p^{-2} fraction of $S \in \mathsf{S}$, $|\Pr_{X\in A}[\phi_S(X) = \alpha] - 1/p| < p^{-2}$. (b) If B is as defined in Corollary 2, and if $2n \geq 8 \log p$, then $|\Pr_{X\in B}[\phi_S(X) = \alpha] - 1/p| < p^{-2}$.*

Lemma 4(a) shows that almost all S satisfy (†), so now we address (‡):

Lemma 5. *If $2n$ is a constant multiple of $\log p$, then there is a $t = O(\log p)$ such that, for a t-step random walk $\boldsymbol{c}$, we have*

$$\Pr_{\boldsymbol{c}}[Y := \psi(\boldsymbol{c}) \text{ has an orbit of size no more than } p^5] = o(p^{-2}) \ .$$

Now we use these lemmata to prove Theorem 2. Notice that, for a random walk $\boldsymbol{c}$, $\Pr_{\boldsymbol{c}}[Y|Y \in T_j] = 1/|T_j|$, since for any two $Y, Y' \in T_j$, we have $Y \sim Y'$ and $\lambda(Y) = \lambda(Y')$. Now arrange the T_j in increasing order by size, and pick the smallest $L \in \mathbb{N}$ so that $|T_L| > p^5$. Now, for *every* "good" (see (†)) S:

$$\begin{aligned}
\Pr_Y[\phi_S(Y) = \alpha] &= \sum_{j=1}^{L} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \\
&\quad + \sum_{j=L+1}^{N} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \\
&\leq \sum_{j=1}^{L} \Pr_Y[Y \in T_j] + \sum_{j=L+1}^{N} \left[\frac{1}{p} + \left(\frac{1}{p^2}\right)\right] \Pr_Y[Y \in T_j] \\
&\leq o(p^{-2}) + \left(\frac{1}{p} + \frac{1}{p^2}\right) \sum_{j=L+1}^{N} \Pr_Y[Y \in T_j] \leq o(p^{-2}) + \left(\frac{1}{p} + \frac{1}{p^2}\right) \ .
\end{aligned}$$

We complete the proof of Theorem 2 by noticing that, for every good S, we have a similar lower bound: $\Pr_Y[\phi_S(Y) = \alpha] \geq \sum_{j=L+1}^{N} \Pr_Y[\phi_S(Y) = \alpha | Y \in T_j] \Pr_Y[Y \in T_j] \geq \left(\frac{1}{p} - \frac{1}{p^2}\right) \sum_{j=L+1}^{N} \Pr_Y[Y \in T_j] \geq \left(\frac{1}{p} - \frac{1}{p^2}\right)\left(1 - o(p^{-2})\right)$. □

5.3 Rapid Mixing with Limited Independence

In this section, we will denote by w a lower bound on the local-independence parameter of the hash functions so that $\boldsymbol{c}$ will be w-wise independent (and hence, $d \geq w$). Our analysis is applicable to any $\boldsymbol{c}$ that is w-wise independent. For example, $\boldsymbol{c}(r)$ may depend only on r, $\boldsymbol{c}(r)$ may depend only on z_r, or $\boldsymbol{c}(r)$ may depend on both, possibly with additional parameters. Indeed, we use this fact in Section 6.

We need to compute $\Pr[z_j = \alpha | z_i]$. For convenience, we write $\overline{g_r} = g_{\boldsymbol{c}(r)}$, so that the sequence of generators chosen for the walk is $\overline{g_0}, \overline{g_1}, \ldots, \overline{g_{t-1}}$. We denote the intervals of integers as $[a, b] = \{a, a+1, \ldots, b\}$, $(a, b] = \{a+1, a+2, \ldots, b\}$, etc., and we denote the shift by m of an interval $I = [a, b)$ by $I + m := [a+m, b+$

m). For notational convenience, we define, for an interval I, $\boldsymbol{\pi}(I) := \prod_{r\in I} \overline{g_r}$. Let τ be such that the distribution after τ steps of random walk is within p^{-2} of the uniform. Set $L := \lfloor t/\tau \rfloor - 1$. We will see that, for an overwhelming fraction of hash functions (or, equivalently, $\boldsymbol{c}$), the following *cancellation property* holds for some $A_i \subseteq [t-(i+1)\tau, t-i\tau)$ (for $1 \le i < L$): $\boldsymbol{\pi}(A_i)\boldsymbol{\pi}([t-i\tau, t-(i-1)\tau)\setminus A_{i-1}) = 1$ (where $A_0 := \emptyset$). Hence,

$$\begin{aligned} z_t &= z_\tau \boldsymbol{\pi}([\tau, t)) = z_\tau \boldsymbol{\pi}([\tau, t-2\tau))\boldsymbol{\pi}([t-2\tau, t-\tau)\setminus A_1) \\ &= z_\tau \boldsymbol{\pi}([\tau, t-3\tau))\boldsymbol{\pi}([t-3\tau, t-2\tau)\setminus A_2) \\ &= \cdots = z_\tau \boldsymbol{\pi}([\tau, t-L\tau))\boldsymbol{\pi}([t-L\tau, t-(L-1)\tau)\setminus A_{L-1}) \ . \end{aligned}$$

That is, the walk beyond τ steps repeatedly introduces a multiplicative factor of $1 \in G$ via subproducts over small (*i.e.*, of length at most 2τ) intervals; this does not mean that the z_is repeat, since the terms in the subproducts need not be consecutive (*i.e.*, the A_i need not be intervals). To be precise, τ is defined to be the minimal value so that if $\mu_\tau\colon G \to [0,1]$ is the distribution of the node z_τ, then $|\mu_\tau(\alpha) - 1/p| < p^{-2}$ (for all α and all starting points z_0 for the walk). The exact values for μ_τ may depend on the starting point or the independence parameter of h, but μ_τ is well-defined without knowing these, up to the additive p^{-2} error term. We call μ_τ (up to this error term) the distribution after τ steps.

Our basic parameters will be w, τ and Δ; here Δ (see Lemma 6) is a lower bound on the length of a walk during which every $g_i \in S$ (alternatively, some constant fraction of S) will almost surely be chosen at least once. First we have three simple lemmata:

Lemma 6. *Let $J = [s, s+\Delta) \subseteq [0, \Lambda]$ be given. Then there is a set $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ of size at least $|\mathsf{H}|\left(1 - p^{-3}/4\right)$ for whose members the following hold:*

(a) if $\Delta \ge 5(2n)^2$ and $2n\sqrt{6} \ge w \ge 2n \ge 4\log p$, then $\{\overline{g_j} \ : \ j \in J\} = S$ and

(b) if $\Delta \ge \left(\frac{20}{3}\right) w$ and $w > 2n + 3 + 4\log p$, then there exists a $B \subseteq J$ such that $S' := \{\overline{g_j} \ : \ j \in B\}$ has at least $\frac{1}{4}|S|$ elements.

Lemma 7. *Let $\alpha \in G$ be arbitrary. If $2n > 8\log p$, then, for every $J = [s, s+\Delta) \subseteq [0, \Lambda]$ such that $\{\overline{g_j} \ : \ j \in J\} = S$, $\mathsf{S}_{\text{GOOD}} := \{S \ : \ \exists A \subseteq J \text{ s.t. } \boldsymbol{\pi}(A) = \alpha\}$ has probability at least $1 - p^{-3}/4$. Additionally, the conclusion holds under the weaker requirement that $S' := \{\overline{g_j} \ : \ j \in J\}$ has at least $8\log p$ elements.*

Lemma 8. *Let $s \le \Lambda$ and $\alpha \in G$ be arbitrary. Recall that μ_τ is the probability distribution (defined up to $O(p^{-2})$ error terms) after τ steps of a totally independent random walk. Let H_{GOOD} be any set containing at least a $1 - p^{-3}/4$ fraction of H (and assume the degree of the polynomials d is at least $\tau + \Delta$). Set $I := [s, s+\tau)$ and $J := [s+\tau, s+\tau+\Delta)$. Then, for any $A \subseteq J$, there is a ζ such that $|\zeta| \le p^{-2}$, for which*

$$\begin{aligned} \Pr_{\boldsymbol{c}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J\setminus A)] &= \Pr_{h_E \in \mathsf{H}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J\setminus A)] \\ &= \mu_\tau(\alpha\boldsymbol{\pi}(J\setminus A)) = \Pr_{h_E \in \mathsf{H}_{\text{GOOD}}}[\boldsymbol{\pi}(I) = \alpha\boldsymbol{\pi}(J\setminus A)] + \zeta \ . \end{aligned}$$

Remark 3. Lemma 6 holds for every interval $J \subseteq [s, \Lambda]$ and Lemma 8 holds for every interval $(I \cup J) \subseteq [s, \Lambda]$, both because we consider paths of every possible length when performing the run-time analysis of the Cayley rho.

Lemma 9 (Rewind). *Let $\frac{3}{20}\Delta' \geq w \geq 2n \geq 32 \log p$ and $d \geq \Delta' + \tau$. Let $i + \tau < j \leq t \leq \Lambda$ and let $\alpha \in G$ be arbitrary. Then there exist $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, each of probability at least $1 - p^{-3}/4$, such that the following holds over $h_E \in \mathsf{H}_{\text{GOOD}}$, $S \in \mathsf{S}_{\text{GOOD}}$: $|\Pr_{h,S}[z_j = \alpha] - \Pr_{h,S}[z_{i+\tau} = \alpha]| \leq p^{-2}$.*

Lemma 10. *Put $\Delta' = \Delta + \tau$. Let i, j, k, ℓ be such that $i + \Delta' < j \leq \Lambda$ and $k + \Delta' < \ell \leq \Lambda$, and $[i, j] \cap [k, \ell] \neq \emptyset \Rightarrow (|i - k| > \Delta'$ and $|j - \ell| > \Delta')$. Let $\alpha, \beta \in G$ be arbitrary. If $d \geq 2\Delta'$, then there are sets $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, both of probability at least $1 - p^{-3}/4$, such that $|\Pr[z_j = \alpha | z_\ell = \beta; z_i, z_k] - \Pr[z_j = \alpha | z_i]| \leq p^{-2}$ when the probabilities are viewed over $h_E \in \mathsf{H}_{\text{GOOD}}$ and $S \in \mathsf{S}_{\text{GOOD}}$.*

Now we consider the case when one of the walks is too short to guarantee that it mixes to a uniform distribution.

Lemma 11. *Let $\Delta \geq \Delta' + \tau$, with Δ' as in Lemma 9 and let i, j, k, ℓ be such that $\ell < k + \Delta \leq \Lambda$ and $i + \Delta < j \leq \Lambda$, and $[i, j] \cap [k, \ell] \neq \emptyset \Rightarrow (|i - k| > \Delta$ and $|j - \ell| > \Delta)$. Let $\alpha, \beta \in G$ be arbitrary. If $d \geq 2(\tau + \Delta)$ and $|S| \geq 2\Delta$, then there are sets $\mathsf{H}_{\text{GOOD}} \subseteq \mathsf{H}$ and $\mathsf{S}_{\text{GOOD}} \subseteq \mathsf{S}$, both of probability at least $1 - p^{-3}/4$, such that $|\Pr[z_j = \alpha | z_\ell = \beta; z_i, z_k] - \Pr[z_j = \alpha | z_i]| \leq p^{-2}$ when the probabilities are viewed over $h_E \in \mathsf{H}_{\text{GOOD}}$ and $S \in \mathsf{S}_{\text{GOOD}}$.*

6 Run Time of Cayley Rho

Let $z_0, z_1, \ldots, z_t \in G$ denote the sequence produced by the Cayley rho algorithm $\mathcal{C}$. Define the random variables Y_{ij} to be 0 when $z_i \neq z_j$ and 1 otherwise (for $i, j \in \{0, 1, \ldots, t\}$). Then the number of collisions is $Y := \sum_{i<j} Y_{ij}$. Put $\mu := E_{S,h}[Y]$ and $\sigma^2 = E_{S,h}[(Y - \mu)^2]$. We wish to bound $\varrho := \Pr[Y = 0] \leq \Pr[|Y - \mu| \geq \mu] \leq \frac{\sigma^2}{\mu^2} = \frac{E_{S,h}[Y^2]}{\mu^2} - 1$. In this section we prove

Lemma 12. *There is a $v = O(\log p)$ such that if $t \geq 4v\xi\sqrt{p}$, then $\varrho \leq \xi^{-2}$.*

As $\Delta' + \tau = O(\log p)$, we may choose v so that $v \geq 2(\Delta' + \tau)$, where Δ' and τ are as defined in the previous section. We also will assume in this section that $d \geq \chi(2n)$ (for some constant χ) and that $m < 2n$; we try to prove the result in terms of t, optimal up to a constant factor. A path is called *short* if its length is at most v; otherwise the path is called *long*.

In the proof of the lemma, a technical issue stems from the fact that if $Y_{a,a+L} = 1$ corresponds to a cycle, then $Y_{b,b+qL} = 1$ for every $b \geq a$ and every $q \in \mathbb{N}$. Thus, in the equation $\mathrm{E}_{S,h}\left[Y^2\right] = \sum \mathrm{E}_{S,h}\left[Y_{ij} Y_{k\ell}\right]$, if the shortest cycle corresponds to $Y_{a,a+L} = 1$ (*i.e.*, L is minimal), then, for each $q = 1, 2, \ldots, \lfloor t/L \rfloor$ and $b \geq a$, further cycles occur so that we get $Y_{b,b+qL} = 1$. This will make a significant contribution to $E_{S,h}\left[Y^2\right]$, but these correlations are due to the

existence of cycles *after* the first cycle occurs, and we need only to find an upper bound for the probability of the absence of cycles.

To this end, we now construct a new walk which coincides with $\boldsymbol{c}$ up to the first collision. As $h(\alpha)$ (for $\alpha \in G$) was constructed from $h_E(y)$ (for $y \in E$) in Section 2.3, so we construct $h_2(\alpha, \gamma)$ (for $\alpha \in G$ and $\gamma \in \{0, 1, \ldots, \Lambda^2\}$) from $h_E(\gamma \circ \alpha)$ ($\circ$ denotes concatenation; γ will be prefixed with the necessary leading zeroes for it to occupy the $2\lceil \log_2 p \rceil$ most-significant bits of the input to h_E). We construct h_2 so $h_2(\alpha, 0) = h(\alpha)$ for all $\alpha \in G$. Given a random walk $\boldsymbol{c}$ and an h_2 constructed from a given h_E, we define a *modified random walk* $\widetilde{\boldsymbol{c}}$ as follows: $\widetilde{\boldsymbol{c}}(0) := h_2(\widetilde{z}_0, \gamma_0)$ and $\widetilde{\boldsymbol{c}}(i) := h_2(\widetilde{z}_i, \gamma_i)$, where $\gamma_i = |\{s \in \{0, 1, \ldots, i\} \ : \ \exists s' < s \ \text{ s.t. } \ \widetilde{z}_s = \widetilde{z}_{s'}\}| < \Lambda^2$ and the $\widetilde{z}_i$ are defined so $\widetilde{z}_{i+1} = \widetilde{z}_i g_{\widetilde{\boldsymbol{c}}(i)}$ and $\widetilde{z}_0 = z_0$ (the z_0 associated with the original walk $\boldsymbol{c}$). Now we have a simple lemma:

Lemma 13. *Let h_E be a random polynomial of degree $d \geq m$. Fix z_0. Then the following hold: (a) for any $t \leq \Lambda$, if $\gamma_t = 0$, then the modified random walk $\widetilde{\boldsymbol{c}}$ agrees with $\boldsymbol{c}$ and they both generate the same sequence $\{z_i\}_{i=0}^{t}$; (b) the following three are m-wise independent functions of their input (α, γ, and α, respectively): $h_\gamma(\alpha) := h_E(\gamma \circ \alpha)$, $h_\alpha(\gamma) := h_E(\gamma \circ \alpha)$ and $h_E(0 \circ \alpha) = h_E(\alpha)$; (c) the modified walk $\{\widetilde{\boldsymbol{c}}_i\}_{i=0}^{t-1}$ is an m-wise independent sequence; and (d) the probability (over h_E) that there is no collision is the same for both $\widetilde{\boldsymbol{c}}$ and $\boldsymbol{c}$.*

We point out that after the first collision (*i.e.*, when $\gamma_i > 0$), the walks $\widetilde{\boldsymbol{c}}$ and $\boldsymbol{c}$ can be markedly different. The modified walk is likely hard to implement in full generality without increasing time or space requirements significantly.

For every fixed $S \in \mathsf{S}$, when the path from z_i to z_j is long, $\Pr_h[Y_{ij} = 1]$ is only approximately $1/p$ (within p^{-2} error). However,

Lemma 14. *If z_i and z_j are endpoints of a short path, then $E_{S,h}[Y_{ij}] = 1/p$.*

Proof. Let $L := j - i \leq v$ be the length of the path $X \in \Omega_L$ from z_i to z_j. Then,

$$\begin{aligned}\mathrm{E}_{S,h}[Y_{ij}] &= \Pr_{S,\mathbf{c}}[Y_{ij} = 1] = \Pr_{S, X=\psi(\boldsymbol{c})}[\phi_S(X) = 1] \\ &= \sum_{\overline{X} \in \Omega_L} \underbrace{\Pr_S[\phi_S(\overline{X}) = 1 | X = \overline{X}]}_{\text{Lemma 2}} \Pr_{\boldsymbol{c}}[X = \overline{X}] = \sum_{\overline{X} \in \Omega_L} \frac{1}{p} \Pr_{\boldsymbol{c}}[X = \overline{X}] = \frac{1}{p} \ .\end{aligned}$$

Notice that $\Pr_S[\phi_S(\overline{X}) = 1 | X = \overline{X}]$ is well-defined and that Lemma 2 can be applied to compute it. This is consistent with the intuitive observation that on short distances, the Cayley rho walk appears independent. □

Now define $U := \{(i, j, k, \ell) \ : \ 0 \leq i < j \leq t \ \text{ and } \ 0 \leq k < \ell \leq t\}$; each element of U represents a pair of paths: one from i to j and one from k to ℓ. Recall that $v > 2(\Delta' + \tau)$, where Δ' and τ are as in the previous section. Define K to be the set of tuples $(i, j, k, \ell) \in U$ containing entries that satisfy the assumptions of Lemmata 10 and 11; $\overline{K}$ will denote $U \setminus K$. Thus, $\overline{K}$ contains path pairs that are (a) both *short* (*i.e.*, of length at most v) or (b) one of the paths is

long and the other one is short but the short one has its end points within a short distance from the endpoints of the long path. Thus $\overline{K}$ contains at most $\binom{t+1}{2}(2v)^2 + \binom{t+1}{2}(2v)^2 = 8v^2\binom{t+1}{2}$ many 4-tuples.

Lemma 15. *If* $(i,j,k,\ell) \in \overline{K}$*, then* $|\Pr[Y_{ij}=1|Y_{k\ell}=1] - 1/p| = O(p^{-2})$.

Proof. We use the basic relations $\Pr[A|B] = \sum_i \Pr[A|BC_i]\Pr[C_i|B]$ (where $\{C_i\}$ is a partition) and $\Pr[A|B] \le \Pr[A]/\Pr[B]$. Now, let C_1 be the event $(h,S) \in \mathsf{H}_{\text{GOOD}} \times \mathsf{S}_{\text{GOOD}}$ and C_2 its complement. Recall that $\Pr[C_2] = O(p^{-3})$. Then, $\Pr_{S,h}[Y_{k\ell}=1|Y_{ij}=1]$ equals

$$\underbrace{\Pr_{S,h}[Y_{k\ell}=1|Y_{ij}=1;C_1]}_{\text{Lemma 10}} \Pr_{S,h}[C_1|Y_{k\ell}=1] + \Pr_{S,h}[Y_{ij}=1|Y_{k\ell}=1;C_2] \Pr_{S,h}[C_2|Y_{k\ell}=1] \ ,$$

which is no more than $\left(\frac{1}{p}+O(p^{-2})\right)\cdot 1 + 1\cdot\frac{\Pr_{S,h}[C_2]}{\Pr_{S,h}[Y_{k\ell}=1]} \le \frac{1}{p}+O(p^{-2})+\frac{O(p^{-3})}{1/p} = \frac{1}{p}+O(p^{-2})$. □

Now we can finish our proof of the lower bound for the probability that a cycle exists (Lemma 12). We notice that $\mathrm{E}_{S,h}[Y^2] = \sum_{\substack{i<j\\k<\ell}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] = \sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] + \sum_K \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}]$ and proceed to bound each term.

$$\sum_K \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] = \sum_K \underbrace{\Pr[Y_{ij}=1|Y_{k\ell}=1]}_{\text{Lemma 15}} \Pr[Y_{k\ell}=1] \le \sum_K \left(\frac{1}{p}+O(p^{-2})\right)\frac{1}{p}$$
$$= |K|\left(p^{-2}+O(p^{-3})\right) = \left(\binom{t+1}{2}^2 - \left|\overline{K}\right|\right)\left(p^{-2}+O(p^{-3})\right) \ .$$

In the complementary range, $\sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}Y_{k\ell}] \le \sum_{\overline{K}} \mathrm{E}_{S,h}[Y_{ij}] = \sum_{\overline{K}} \frac{1}{p} = \left|\overline{K}\right|\frac{1}{p}$. Finally, by Lemma 14, $\mathrm{E}_{S,h}[Y] = \sum Y_{ij} = \sum_{i<j}\frac{1}{p} = \binom{t+1}{2}\frac{1}{p}$. Combining the results, we see that

$$\varrho \le \frac{E_{S,h}[Y^2]}{E_{S,h}[Y]^2} - 1 \le \left(\binom{t+1}{2}\frac{1}{p}\right)^{-2}\left[\left|\overline{K}\right|\frac{1}{p} + \frac{\binom{t+1}{2}^2 - \left|\overline{K}\right|}{p^2}\left(1+O\left(\frac{1}{p}\right)\right)\right] - 1 \ ,$$

which is less than $16v^2p/t^2$. Hence, when $t \ge 4v\xi\sqrt{p}$, we get $\varrho \le \xi^{-2}$. □

We thank Prasad Tetali for generous discussions related to random walks on directed graphs.

References

1. N. Alon and Y. Roichman, "Random Cayley Graphs and Expanders." *Random Structures and Algorithms*, **5**:271–284, 1994.
2. E. Bach, "Toward a Theory of Pollard's Rho Method." *Information and Computation*, **90**(2):139–155, 1991.

3. M. Bellare and J. Rompel, "Randomness-Efficient Oblivious Sampling." *Symposium on Foundations of Computer Science (FOCS '94)*:276–287, 1994.
4. B. Bollobas, *Modern Graph Theory*, Graduate Texts in Mathematics **184**. Springer-Verlag, Berlin, 1998.
5. A. Broder and E. Shamir, "On the Second Eigenvalue of Random Regular Graphs." *Symposium on the Foundations of Computer Science (FOCS '87)*:286–294, 1987.
6. P. Erdős and A. Rényi, "Probabilistic Methods in Group Theory." *Journal d'Analyse Mathématique*, **14**:127–138, 1965.
7. S.C. Pohlig and M.E. Hellman, "An Improved Algorithm for Computing Logarithms over GF(p) and Its Cryptographic Significance." *IEEE Transactions on Information Theory*, **24**:106–110, 1978.
8. J.M. Pollard, "Monte Carlo Methods for Index Computation (mod p)." *Mathematics of Computation*, **32**(143):918–924, 1978.
9. Y. Roichman, "On Random Random Walks." *Annals of Probability*, **24**(2):1001–1011, 1996.
10. V. Shoup, "Lower Bounds for Discrete Logarithms and Related Problems." *Advances in Cryptology: EUROCRYPT '97 (LNCS 1233)*:256–266, 1997.
11. E. Teske, "Speeding Up Pollard's Rho Method for Computing Discrete Logarithms." *Algorithmic Number Theory Symposium III: ANTS-III (LNCS 1423)*:541–554, 1998.
12. E. Teske, "On Random Walks for Pollard's Rho Method." *Mathematics of Computation*, **70**:809–825, 2001.