Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 1659–1664, February 1996
Mathematics

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on April 25, 1995.

The cutoff phenomenon in finite Markov chains

PERSI DIACONIS

Department of Mathematics, Harvard University, Cambridge, MA 02138

Contributed by Persi Diaconis, December 4, 1995

ABSTRACT **Natural mixing processes modeled by Markov chains often show a sharp cutoff in their convergence to long-time behavior. This paper presents problems where the cutoff can be proved (card shuffling, the Ehrenfests' urn). It shows that chains with polynomial growth (drunkard's walk) do not show cutoffs. The best general understanding of such cutoffs (high multiplicity of second eigenvalues due to symmetry) is explored. Examples are given where the symmetry is broken but the cutoff phenomenon persists.**

Section 1. Introduction

Markov chains are widely used as models and computational devices in areas ranging from statistics to physics. A chain starts at a beginning state x in some finite set of states $\mathcal{X}$. At each time, it moves from its current state (say z) to a new state y with probability $P(z, y)$. Thus, after two steps, the chain goes from x to y with probability $P^2(x, y) = \Sigma P(x, z) P(z, y)$. One feature of Markov chains is a limiting stationary distribution $\pi(y)$. Under mild conditions, no matter what the starting state, after many steps the chance that the chain is at y is approximately $\pi(y)$. [In symbols, $P^k(x, y) \to \pi(y)$.] A good example to keep in mind is repeated shuffling of a deck of 52 cards. For most methods of shuffling cards, the stationary distribution is uniform, $\pi(y) = 1/52!$, the limit result says that repeated shuffles mix the cards up.

It is important to know how long the chain takes to reach stationarity. As explained below, it takes about seven ordinary riffle shuffles to adequately mix 52 cards. The familiar overhand shuffle (small clumps of cards dropped from one hand to another) takes about 2500 shuffles (1). A quantitative notion of "close to stationarity" uses the variation distance

$$\|P_x^k - \pi\| = \max_A |P^k(x, A) - \pi(A)|$$

with $P^k(x, A) = \Sigma_{y \in A} P^k(x, y)$ denoting the chance that the chain started at x is in the set A after k steps. The maximum is over all subsets A of $\mathcal{X}$. Thus, if $\|P_x^k - \pi\|$ is small, the stationary probability is a good approximation, uniformly. As an example, in shuffling cards, A might be the set of arrangements where the ace of spades is in the top 1/3 of the deck. Then, $\pi(A) = 1/3$ and one is asking that the chance $P_x^k(A)$ be about 1/3. With the chain and starting state specified one has a well posed math problem: given a tolerance $\varepsilon > 0$, how large should k be so that $\|P_x^k - \pi\| < \varepsilon$?

A surprising recent discovery is that convergence to stationarity shows a sharp cutoff; the distance $\|P_x^k - \pi\|$ stays close to its maximum value at 1 for a while, then suddenly drops to a quite small value and then tends to zero exponentially fast.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

As an example, consider the Gilbert–Shannon–Reeds model for riffle shuffling cards. A deck of n cards is cut into two piles according to a symmetric binomial distribution. Then the two piles are riffled together by the following rule: if one pile has A cards and the other has B cards, drop the next card from the A pile with probability $A/A+B$ (and from the B pile with probability $B/A+B$). The dropping is continued until both piles have been run through, using a new A, B at each stage. This specifies $P(x, y)$ for all arrangements x, y.

Following earlier work by Gilbert, Shannon, Reeds, and Aldous, a definitive analysis of the riffle shuffle chain was produced in joint work with David Bayer (2). Table 1 shows the distance to stationarity for 52 cards as the number of shuffles k varies. The distance to stationarity thus stays essentially at its maximum value of one up to five shuffles. Then it rapidly tends from one to zero. The final numbers decrease by a factor of 1/2, and this exponential decay continues forever.

The data in Table 1 are derived from a closed form expression for the chance of being in any arrangement after any number of shuffles. This formula has connections with combinatorics, cohomology theory, Lie algebras, and other subjects developed in refs. 2 and 3. The formula can be used to give sharp approximations to the distance for any deck size:

THEOREM 1. *Let* $P(x, y)$ *result from the Gilbert–Shannon–Reeds distribution for riffle shuffling* n *cards (2). Let* $k = (3/2)\log_2 n + \theta$. *Then,*

$$\|P_x^k - \pi\| = 1 - 2\Phi\left(-2^{-\theta}/4\sqrt{3}\right) + O\left(1/\sqrt{n}\right)$$

with $\Phi(z) = \int_{-\infty}^{z} (e^{-t^2/2}/\sqrt{2\pi})dt$.

Theorem 1 shows that a graph of the distance to stationarity versus k appears as shown in Fig. 1. There is a sharp cutoff at $(3/2)\log n$; the distance tends to 0 exponentially past this point. It tends to one doubly exponentially before this point.

This cutoff phenomenon has been proved to occur in dozens of different classes of Markov chains. Classical analysis of such chains was content with a bound on the second eigenvalue. This determines the exponential rate and so answers the question: "how close is the chain to stationarity after a million shuffles." Modern workers realize that questions of practical relevance require determining if seven steps suffice.

Section 2 describes examples where the cutoff phenomenon can be proved. These include Ehrenfests' urn, a variety of shuffles, and random walks on matrix groups both finite and compact. Section 3 gives examples where there is no cutoff. This includes the classical drunkard's walk and random walk on graphs with polynomial growth. Section 4 gives my best understanding of what causes cutoff phenomena in terms of repeated eigenvalues and symmetry. Section 5 gives some new examples where the symmetry is broken but the phenomenon persists. The final section describes how things change if different notions of distance are used.

Table 1. Distance to stationarity for repeated shuffles of 52 cards

	k									
	1	2	3	4	5	6	7	8	9	10
$\|P^k - \pi\|$	1.000	1.000	1.000	1.000	0.924	0.624	0.312	0.161	0.083	0.041

To close this introduction, here is a definition of cutoffs: let P_n, π_n be Markov chains on sets $\mathcal{X}_n$. Let a_n, b_n be functions tending to infinity with b_n/a_n tending to zero. Say the chains satisfy an a_n, b_n cutoff if for some starting states x_n and all fixed real θ, with $k_n = \lfloor a_n + \theta b_n \rfloor$, then

$$\|P_n^{k_n} - \pi_n\| \to c(\theta)$$

with $c(\theta)$ a function tending to 0 for θ tending to infinity and tending to 1 for θ tending to minus infinity. For example, in *Theorem 1* above, $\mathcal{X}_n$ is the set of arrangements of n cards, P_n is the Gilbert–Shannon–Reeds distribution, π_n is the uniform distribution on $\mathcal{X}_n$, $a_n = (3/2)\log_2 n$, $b_n = 1$.

Section 2. Examples of the Cutoff Phenomena

A. The Ehrenfests' Urn. In a decisive discussion of a paradox in statistical mechanics (how can entropy increase be compatible with a system returning to its starting state?) P. and T. Ehrenfest introduced a simple model for diffusion. Their model involved two urns and d balls. To start, all the balls are in urn 2. Each time, a ball is randomly chosen and moved to the other urn. It is intuitively clear that after a large number of switches any one of the balls is equally likely to be in either urn. The state of the chain is determined by the number of balls in urn 1. The stationary distribution is the binomial: $\pi(j) = \binom{d}{j}/2^d, 0 \le j \le d$. To state a precise result, a periodicity problem must be ruled out. Let the chain remain in its current state with probability $1/(d+1)$. Thus, the state space (number of balls in urn 1) is $\{0, 1, \ldots, d\}$. The transition probabilities are $P(i, i-1) = i/(d+1)$, $P(i, i) = 1/(d+1)$, $P(i, i+1) = (d-i)/(d+1)$, $0 \le i \le d$. A cutoff for the Ehrenfest chain starting at 0 was proved in refs. 4 and 5; the cutoff occurs at $(1/4)d \log d$:

THEOREM 2. *For the Ehrenfest chain on* $\{0, 1, \ldots, d\}$ *started at* 0, *if* $k = 1/4(d+1)(\log d + \theta)$ *with* $\theta > 0$, *then*

$$\|P_0^k - \pi\| \le \frac{1}{\sqrt{2}}\{e^{e^{-\theta}} - 1\}^{1/2}.$$

Conversely, if $k = (1/4)d(\log d - \theta)$, *the total variation distance tends to* 1 *for* d *and* θ *large.*

Returning to the Ehrenfests' story, suppose entropy is measured by $I(k) = \sum_{j=0}^{d} P_0^k(j) \log P_0^k(j)$, this is clearly increasing in k [the chain tends to stationarity, $P_0^k(j) \to \pi(j)$]. However, it is equally certain that the chain will return to zero (infinitely often as time goes on). The theory described above quantifies this: it takes about $(1/4)d \log d$ steps for the chain to reach maximum entropy but about 2^d steps to have a reasonable chance of returning to 0. If d is large—e.g., Avogadro's number—we will not observe such returns. Kac (6) gives a masterful development of this point.

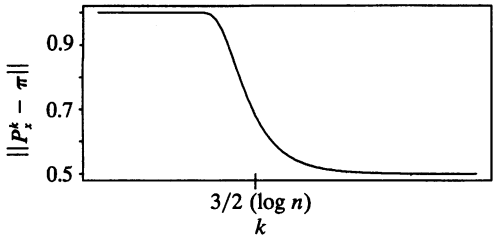

FIG. 1. The cutoff phenomenon for repeated riffle shuffles of n = 52 cards.

The shape of the cutoff for the Ehrenfests' chain is determined in ref. 7. By the same methods used there, it can be proved that for the chain started at $d/2$, order d steps are necessary and suffice to achieve stationarity. Further, there is no cutoff: for $k = \theta d$

$$\|P_{d/2}^k - \pi\| \sim f(\theta)$$

with f a continuous function of θ on $(0, \infty)$.

A different model of diffusion was introduced earlier by Bernoulli and Laplace (39): There were n black balls and n red balls distributed between two urns. Initially the colors were segregated. At each stage, a ball was chosen randomly from each urn and the two balls exchanged. In ref. 5 it was shown that the associated Markov chain has a cutoff at $(d/4)(\log d + \theta)$ as above. The sharp results for the Bernoulli–Laplace model were crucial in studying a collection of related processes: the exclusion processes. Here, one has a graph (such as an $n \times n$ grid) and k particles with at most one per vertex. At each time, a particle is chosen at random and then a neighboring vertex is chosen at random. If the neighboring vertex is unoccupied, the particle moves there. If the vertex is occupied, the system stays as it was. Good bounds for the rate of convergence were achieved in refs. 8 and 9 by comparison with the Bernoulli–Laplace model. However, the available technique is too crude to determine if there is a cutoff for the exclusion process. It is certainly natural to conjecture such cutoffs.

B. Random Transpositions. This is perhaps the earliest problem where a sharp cutoff was demonstrated. Picture n cards labeled $1, 2, \ldots, n$. Initially, they are in a row on the table. At each time, the left and right hands randomly choose cards (so, left = right with probability $1/n$). Then the two cards are switched. This is a Markov chain on the set of all $n!$ permutations. It has a uniform stationary distribution. There is a cutoff at $(1/2)n \log n$:

THEOREM 3. *For the random transposition chain and any starting* x, *for* $k = (1/2)n(\log n + \theta)$ *with* $\theta > 0$,

$$\|P_x^k - \pi\| \le Ae^{-\theta/2}$$

with A *a universal constant (10). Conversely, if* $k = (1/2)n(\log n - \theta n$, *then the distance to stationarity tends to* 1 *for* n *and* θ *large.*

Theorem 1 was proved by using detailed knowledge of the character theory of the permutation group. The techniques are fairly general. They work for random walk on any finite group provided the underlying measure is concentrated on a union of conjugacy classes. For example, Hildebrand (11) worked with "random transvections" in the $n \times n$ matrices with elements in a finite field. He showed a cutoff at $n + \theta$. The method also works for compact groups. Rosenthal (12) and Porod (13) have demonstrated sharp cutoffs for several natural walks on the orthogonal and other classical groups.

The method also works for less symmetric problems: Flatto *et al.* (14) showed a cutoff at $n(\log n + \theta)$ for "transpose random with top." Lulov (15) studied the following problem. Take n even; randomly transpose a pair of cards, then a different pair, and so on until $n/2$ pairs have been exchanged. This all counts as one shuffle. Lulov showed that the variation distance is small after three shuffles (but not after two).

The character theory method can be used to get less precise results for complex random walks. In ref. 16 the random transpositions result is combined with a comparison technique to get good results for general random walks—e.g., randomly transpose

top two or randomly move top to bottom; this takes order $n^3 \log n$ to get random. In a remarkable piece of work, Gluck (18) shows that for essentially any small generating conjugacy class of any finite group of Lie type, order rank (G) steps are necessary and sufficient for convergence to stationarity. None of the examples in this paragraph have had cutoffs proved, although again it is natural to conjecture cutoffs in all of them.

Returning to the problem of random transpositions, there is now a different method which leads to a completely different proof of the cutoff phenomenon. This is the method of strong stationary times introduced in joint work with Aldous (19) and Fill (20). Its successful application to transpositions is due to Broder and Matthews (21).

C. Library and List Management Problems. Imagine n folders (or computer files) are used from time to time, the i^{th} folder being used proportion $w(i)$ of the time. It makes sense to keep popular folders near the top of the pile. If the weights are unknown, one common sense procedure is the move to the top rule: after a folder is used it is replaced on top. This leads to a Markov chain on the set of $n!$ arrangements. Such chains have been studied by computer scientists, geneticists, library scientists, and probabilists. Fill (22) contains a good overview.

In this section, I study a generalization of this chain in which the folders are removed in groups. Thus let $[n] = \{1, 2, \ldots, n\}$. For $s \subseteq [n]$, there are weights $w(s) \geq 0$, $\Sigma\, w(s) = 1$. Suppose that the weights separate points in the sense that for each i, j there is an s with $w(s) > 0$, and i in s, j in s^c or vice versa. A Markov chain proceeds on the $n!$ arrangements (thought of as arrangements of cards) by choosing s from $w(s)$, removing the cards with labels in s, keeping them in their same relative order, and moving them to the top. This chain has a unique stationary distribution π. The main result is a simple bound on the distance to stationarity. To describe this, define separation parameters

$$\theta(i,j) = \sum_{\substack{i \in s, j \in s^c \\ \text{or} \\ j \in s, i \in s^c}} w(s).$$

This is the chance of items i and j being separated in a single move. The following useful bound holds.

THEOREM 4. *Let* P(x, y) *be the weighted set to top chain. Then for any starting state* x *and all* k,

$$\|\mathrm{P}_{\mathrm{x}}^{\mathrm{k}} - \pi\| \leq \sum_{\mathrm{i<j}} (1 - \theta_{\mathrm{ij}})^{\mathrm{k}}.$$

Example 1: Let $w\{i\} = 1/n$, $1 \leq i \leq n$, $w(s) = 0$ otherwise. The chain becomes the simple "random to top" chain studied in refs. 19 and 23. Then $\theta_{ij} = 2/n$ and the bound becomes $\|P_x^k - \pi\| \leq \binom{n}{2}(1 - \frac{2}{n})^k$. Thus, if $k = n \log n + cn$, the bound becomes $e^{-2c}/2$. Arguments in ref. 19 show that the variation distance is essentially 1 if $k = n \log n - cn$. Thus, there is a sharp cutoff at $n \log n$. In joint work with Fill and Pitman (23) the exact shape of the cutoff is determined.

Example 2: This uses weights like Zipf's law. Again, $w(s) = 0$ if s has 2 or more elements. Fix a parameter $t \geq 0$. Define

$$w(i) = Z/(n + 1 - i)^t,\ 1 \leq i \leq n,\ Z^{-1} = \sum_1^n (n + 1 - i)^{-t}. \quad \textbf{[2.1]}$$

Given k, define $c = c(n, k)$ by

$$\begin{aligned} t = 0 \quad & k = n(\log n + c) \\ 0 < t < 1 \quad & k = [n/(1-t)](\log n - \log\log n + c) \\ t = 1 \quad & k = n \log n(\log n - \log\log n + c) \\ t > 1 \quad & k = (n^t/\zeta(t))(\log n - \log\log n + c) \end{aligned}$$

$$\zeta(t) = \sum_{j=1}^{\infty} 1/j^t. \quad \textbf{[2.2]}$$

By using classical calculus estimates, it is straightforward to plug into *Theorem 4* to prove

COROLLARY 1. *Let* P(x, y) *be weighted random to top with weights (Eq.* **2.1***). Let* c = c(n, k) *be defined by Eq.* **2.2***. Then, there is a positive continuous function* a(t) *so that*

$$\|\mathrm{P}_{\mathrm{x}}^{\mathrm{k}} - \pi\| \leq \mathrm{a(t)e}^{-\mathrm{c}} + \mathrm{o(1)}.$$

The result is sharp in that given k and c as above with $c < 0$, there is a starting state x such that

$$\|P_x^k - \pi\| \geq f(x, c) + o(1)$$

with $f(t, c)$ continuous and tending to 1 as $c \to -\infty$ for each fixed t. These results show sharp cutoffs of various esoteric types. There are many other simple choices of weights for which similar bounds can be obtained.

Proof of Theorem 4: The argument proceeds by coupling (see refs. 4 or 24). Let the original chain start out in arrangement x_0. Let a second chain start off at x_* chosen according to the stationary distribution π. Each time, choose a subset according to weights $w(s)$ and move the same objects to the top in both arrangements (they may well be in different relative orders). Let T be the first time that every pair $i \neq j$ has been separated. This T is a coupling time: the two arrangements are identical at T. Indeed for each i and j, the last time they are separated they are in the same relative order in both lists. Thus, at time T, each pair of items is in the same relative order, so the two arrangements are equal. The chance that items i and j have not been separated in k steps is at most $(1 - \theta_{ij})^k$. Hence the theorem.

We note in closing that the arguments of this section have been extended to a variety of related problems in an elegant and comprehensive way by Dobrow and Fill (25). One curious note: for the chains analyzed in *Theorem 4*, all the eigenvalues are known (13). It seems impossible to use these eigenvalues to get results comparable to those given above. This shows the power of the coupling method.

D. Further Examples. Cutoffs have been proved in several other examples. Belsley (27) and Silver (28) have shown that some versions of the widely used Metropolis algorithm have cutoffs. D'Aristotle (29), Belsley (27), Greenhalgh (30), and Hora (31) have shown that cutoffs occur for a variety of walks on algebraic objects, such as the subspaces of a fixed dimension over a finite field. A different line of work, so-called random–random walks, show that cutoff phenomena are generic; for example, Aldous and Diaconis (19) show that most Markov chains reach stationarity in two steps! There, the notion of "most" involves a probability distribution over all Markov chains. In fact, Markov chains that are actually run involve fairly sparse transition matrices. A remarkable body of work by Dou, Hildebrand, and Wilson (32–34) shows that cutoffs are generic, even when the support is restricted. The next section provides some examples where cutoffs do not occur.

Section 3. Problems Without a Cutoff

The simplest natural chain without a cutoff is simple random walk on the integers mod n: picture n places around a circle. A particle hops from its current place to a neighboring place (or stays fixed) with probability 1/3. Thus, $P(i, i + 1) = P(i, i) = P(i, i - 1) = 1/3$. The stationary distribution is uniform: $\pi(i) = 1/n$. It is easy to show (24) that for n large,

$$\|P_n^k - \pi\| \sim C\left(\frac{k}{n^2}\right)$$

with $C(\theta)$ a positive decreasing continuous function of θ which is 1 at zero and zero at infinity. Thus, the transition to stationarity happens gradually at about n^2 steps. Essentially,

the same behavior occurs for Markov chains on low-dimensional regular grids. Here are two examples.

Picture a convex set in the plane. Let $\mathcal{X}$ be the lattice points inside the set (points with integer coordinates). A Markov chain proceeds on $\mathcal{X}$ as follows: if the chain is at z, pick one of the four neighboring lattice points of z at random. If the new point is in $\mathcal{X}$, the chain moves there. If the new point is outside $\mathcal{X}$, the chain stays at z. Assuming that the chain is connected, the stationary distribution is uniform in $\mathcal{X}$. In joint work with Saloff-Coste (35), it is shown that this chain takes order γ^2 steps to get random where γ is the diameter of $\mathcal{X}$ (the largest distance between two points of $\mathcal{X}$). Further, it is shown that there is no cutoff; roughly put, if the chain goes $10\gamma^2$ steps, it is close to random. If it goes $(1/10)\gamma^2$ steps, it is far from random.

Such Markov chains arise in a variety of applied problems, such as choosing contingency tables with fixed row and column sums. Many examples can be found in refs. 16 and 35. These papers also contain other approaches to analysis. The technical tools developed in ref. 35 allow similar conclusions for chains with state spaces that have polynomial growth: order γ^2 steps are necessary and suffice for convergence; there is no cutoff.

The Markov chains on the discrete circle and repeated shuffling of cards are examples of random walk on finite groups. Consider a finite group G and a symmetric set of generators S of G. Suppose that the identity is in S (to rule out periodicity problems). A Markov chain begins at the identity and proceeds by repeatedly picking elements of S and multiplying. This chain converges to the uniform stationary distribution $\pi(x) = 1/|G|$.

For a specific example, consider the set of 3×3 matrices which are upper triangular, have ones on the diagonal, and have entries which are integers modulo n. This is the finite Heisenberg group. For S, take the following set of five matrices:

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&\pm 1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&\pm 1\\0&0&1\end{pmatrix}.$$

Following earlier work by Zack (36), it is shown in refs. 16 and 37 that this walk takes order n^2 steps to get random, and there is no cutoff. Here, the diameter is n. The techniques give the same conclusion for any finite nilpotent group supposing only that $|S|$ and the degree of nilpotency stay bounded as the group gets large: order (diameter)^2 steps are necessary and suffice for convergence, and there is no cutoff.

In contrast, random walks with growing number of generators on nilpotent groups with growing degree of nilpotency are expected to show a cutoff. For example, consider the $d \times d$ upper triangular matrices with ones on the diagonal and integer entries modulo n. Let $E(i, j)$ be such a matrix with a one in position (i, j) and zeros elsewhere. Then, a generating set is $S = \{ID, E(1, 2)^{\pm}, E(2, 3)^{\pm}, E(d-1, d)^{\pm}\}$. Stong (17) has analyzed this walk with d large; the results are not quite sharp enough to determine if cutoffs exist, but they are conjectured. Random walk on product groups G^n with G fixed and growing n can be proved to have cutoffs. One such example, the hypercube, is discussed in the next section.

Section 4. What Makes It Cutoff?

This section outlines my best understanding of what causes cutoff phenomena. The argument is an extension of the following simple idea: cutoff phenomena occur because of high multiplicity of the second eigenvalue. To elaborate, it is useful to work with a special class: reversible Markov chains. These are chains $P(x, y)$ and stationary distributions $\pi(x)$ satisfying $\pi(x)P(x, y) = \pi(y)P(y, x)$. When $\pi(x) = 1/|\mathcal{X}|$, this says the matrix $P(x, y)$ is symmetric. More generally, it says the chain run forward is the same as the chain run backward. Irreducible reversible chains have real eigenvalues β_i, $0 \le i \le |\mathcal{X}| - 1$, with $\beta_0 = 1 > \beta_1 \ge \beta_2 \cdots \ge \beta_{|\mathcal{X}|-1} \ge -1$. Each eigenvalue has a corresponding eigenvector V_i. If P is considered as an $|\mathcal{X}| \times |\mathcal{X}|$ matrix and V_i is a column vector, then $PV_i = \beta_i V_i$. If the eigenvectors are normed such that $\|V_i\| = 1$ in the inner product $\langle V, W\rangle = \Sigma V(x)W(x)\pi(x)$, then an easy application of the Cauchy–Schwarz inequality shows

$$4\|P_x^k - \pi\|^2 \le \sum_{i=1}^{|\mathcal{X}|-1} V_i(x)^2 \beta_i^{2k}. \qquad [4.1]$$

This bound is both the key to our present understanding and a main method of proof for cutoff phenomena. The bound is reasonably sharp so that the right side serves as a useful surrogate for the left side. It turns out that for many examples the first term $V_i^2(x)\beta_i^{2k}$ is all that matters; it dominates the rest and determines the final behavior.

Let us consider two examples: the Ehrenfests' urn and random walk on the discrete circle. For the urn (*Theorem 2* in Section 2) the eigenvalues are $\beta_i = 1 - (2i/d+1)$, $0 \le i \le d$. Thus $\beta_1 = 1 - (2/d+1)$. The eigenvectors can be identified with a classical family of orthogonal polynomials (Krawtchouck polynomials). In particular, $V_1(x) = 2(x - d/2)/(d)^{1/2}$. Thus, the lead term in Eq. **4.1** is $[4(x - d/2)/d][1-(2/d+1)]^{2k}$. Consider how large k has to be to make this small. For $x = 0$, the lead term becomes $d[1 - (2/d + 1)]^{2k}$. For $k = (1/4)(d + 1)(\log d + \theta)$, this is about $e^{-\theta}$ when d is large. Thus, there is a cutoff. The cutoff behavior occurs because of the leading d. When $x = (d/2) + (d)^{1/2}$, the lead term becomes $4[1 - (2/d + 1)]^{2k}$. When $k = \theta(d + 1)$, the lead term is about $e^{-\theta}$. Thus, there is no cutoff.

As a second example, consider the lead example of section 3—simple random walk on the integers mod n. Now the eigenvalues are $(1/3) + (2/3)\cos(2\pi j/n) \sim 1 - (4\pi^2 j^2/3n^2)$. Thus, $\beta_1 \sim 1 - (4\pi^2/3n^2)$. The eigenvector is bounded. Thus, the lead term is essentially $[1 - (4\pi^2/3n^2)]^{2k}$. When $k = \theta n^2$, this is essentially $e^{-4\pi^2\theta/3}$. Again, there is no cutoff.

There is more to do in making these arguments rigorous, but the lead term behavior can be shown to determine things. See ref. 24 (chapter 3) for all details.

The random transpositions walk exhibits instructive behavior: the 2nd eigenvalue is $[1 - (2/n)]$ and the eigenfunction is bounded. However, the 2nd eigenvalue occurs with multiplicity $(n-1)^2$. Thus, in the upper bound, one must choose k large enough to make $(n-1)^2[1 - (2/n)]^{2k}$ small. This leads to $k = (1/2)n(\log n + \theta)$ and a cutoff. For general random walk on a finite group G, the eigenvalues are associated with the various representations of the group (see, e.g., ref. 24 for background and examples). If the second eigenvalue is associated with a representation of dimension d, the lead term in the bound will be of form $d\beta_1^{2k}$. In examples, there is a sequence of groups $G(n)$, say, and β_1 is close to 1, say, $\beta_1 = 1 - (1/f(n))$. If d grows with n, the lead term is $d(n)(1 - (1/f(n))^{2k}$. This is small for $k = (f(n)/2)(\log d(n) + \theta)$.

These heuristics thus predict a cutoff for random walks where the size of the representations grow with the size of the group. For example, the permutation group on n letters has its smallest representation (other than the trivial and alternating representations) of dimension $(n-1)$. The heuristics also predict no cutoff for random walk on abelian groups where all representations have degree $d = 1$. Then, the lead term is of the form $(1 - 1/f(n))^{2k}$ and $k = \theta f(n)$ steps are required to make this small.

Of course, some groups have many one-dimensional representations, as well as many higher-dimensional representations. The Heisenberg group (mod n) discussed in Section 3 gives an example. This group has size n^3. It has n^2 one-dimensional representations and $n-1$ representations of dimension n. For the walk described, the largest eigenvalue occurs at a one-

dimensional representation, so there is no cutoff. One can construct walks for which the largest eigenvalue occurs at an n-dimensional representation. Then, there would be a cutoff.

An instructive example is random walk on the group of binary n-tuples. A natural walk picks a coordinate at random and changes it to its opposite (mod 2). If the identity is added to this generating set (to avoid parity problems) the walk can be shown to get random for $k = (1/4)n(\log n + \theta)$. It thus shows a cutoff, even though the group is abelian. The apparent contradiction to the heuristic is resolved by noting that the walk has a large symmetry group (the symmetry group acts to permute the generators). This forces high multiplicity of the 2nd eigenvalue $\{1 - [2/(n + 1)]$ with multiplicity $n\}$. In the next section the symmetry is broken, and a cutoff is still shown.

The discussion above has focused on upper bounds on the variation distance through Eq. **4.1** with a claim that there are matching lower bounds. Often, lower bounds are easy to obtain. A systematic method for random walk on groups which often seems to work is outlined in ref. 24 and developed in ref. 26. It uses the lead term to suggest useful test functions.

Some of the examples (such as the Gilbert–Shannon–Reeds method of shuffling or the library examples of Section 3) are not reversible yet still show sharp cutoffs. This simply emphasizes our lack of understanding.

I close this section with a question: let $G(n)$ be a naturally occurring sequence of groups and let $S(n)$ be a sequence of generating sets. Does the crucial eigenvalue β_1 arise from a representation close to the trivial representation? Thus, for random walk on the integers (mod n), β_1 occurs at the representation. Thus, for random walk on the integers (mod n), β_1 occurs at the representation $e^{2\pi i/n}$. For random transpositions, β_1 occurs at the $(n - 1)$-dimensional representation: the classical Riemann–Lebesgue Lemma is one version of this. (The Fourier transform of an L^1 function tends to zero.) Gluck's (18) remarkable analysis of random walk on finite groups of Lie type proves a version. A natural conjecture is that for natural generating sets, the eigenvalues will decrease with the dimension of the associated representation. It is not easy to see how this can be made precise, but it seems to occur in all natural examples.

Section 5. Breaking Symmetry

At present writing, proof of a cutoff is a difficult, delicate affair, requiring detailed knowledge of the chain, such as all eigenvalues and eigenvectors. Most of the examples where this can be pushed through arise from random walk on groups, with the walk having a fair amount of symmetry. It is natural to wonder if the cutoff occurs for less symmetric chains. In this section, I break the symmetry for the natural walk on the hypercube and show that a cutoff persists.

Let $\mathcal{X}$ be the set of binary n-tuples (so $|\mathcal{X}| = 2^n$). Define a Markov chain on $\mathcal{X}$ by $P(x, y) = p_0$ if $x = y$, $P(x, y) = p_i$ if $x = y$ except in the i^{th} coordinate, $P(x, y) = 0$ otherwise. Here p_i are positive weights summing to 1. The chain has a simple intuitive description: pick i, $0 \le i \le n$ with probability p_i. If 0 is chosen, the chain stays fixed. If i is chosen, the i^{th} coordinate is changed to its opposite. When $p_i = 1/(d + 1)$ this becomes the nearest neighbor walk described above. In all cases, it has a uniform stationary distribution $\pi(x) = 1/2^n$. A one-parameter family of weights modeled after Zipf's law will now be studied.

For $0 < a < \infty$ let $p_i = p_i^a = Z/(1 + i)^a$, $0 \le i \le n$ with $Z^{-1} = \Sigma_{i=0}^n 1/(1 + i)^d$. Let k and $c = c(n, a, k)$ be related as follows:

$$\begin{aligned} a = 0 \quad & k = (1/4)n(\log n + c) \\ 0 < a < 1 \quad & k = (n/4(1 - a))(\log n - \log\log n + c) \\ a = 1 \quad & k = (n \log n/4)(\log n - \log\log n + c) \\ 1 < a < \infty \quad & k = (n^a/4\zeta(a))(\log n - \log\log n + c), \\ & \zeta(a) = \Sigma_{j=1}^{\infty} 1/j^a. \end{aligned}$$

THEOREM 5. *For nearest neighbor random walk on the binary* n*-tuples with weights* p_i^a, *let* k *be as above with* $C > 0$. *There is a positive continuous function* A(a) *such that for any starting state* x

$$\|P_x^k - \pi\| \le A(a)\,(e^{e^{-c}} - 1)^{1/2}.$$

Conversely, with k *and* c *as above with* $C > 0$, *there is an* x *such that*

$$\|P_x^k - \pi\| \ge A(a, c),$$

with A(a, c) *positive and tending to one as* c *tends to* $-\infty$.

Remarks. The upper bound tends to zero like $e^{-c/2}$ when c is large. The lower bound shows that all of these walks have cutoffs. When the weights tend to zero exponentially fast (as $1/2^i$) a similar analysis shows that the time to stationarity is $1/\min p_i$, and there is no cutoff. If weights are chosen at random a similar analysis shows that the time to stationarity is order n^2, and there is no cutoff.

Proof. The chain described is a random walk on the group $\mathcal{X}$. The basic upper bound Eq. **4.1** becomes

$$4\|P_x^k - \pi\|^2 \le \sum_{x \ne 0} (1 - 2x\cdot p)^{2k}$$

$$\le \left\{ \prod_{i=1}^{n} (1 + e^{-4kp_i}) - 1 \right\} + e^{-4kp_0} \prod_{i=1}^{n} (1 + e^{-4kp_i})$$

with $x\cdot p = x_1p_1 + \cdots + x_np_n$. Consider the product $\pi(1 + e^{-4kp_i}) = e^{\Sigma \log(1+e^{-4kp_i})}$. From the definition of p_i and k, we see that in each of the cases $k = S/4p_{i^*}$, where $p_{i^*} = \min p_i$, and $S = \{\log n - \log\log n + c\}$. It follows that $-4kp_i < S$ whence $e^{4kp_i} < e^{-c} \log n/n$. Now $\log(1 + x) \le x + x^2/2$ for $0 < x \le 1/2$. Thus, $\Sigma \log(1 + e^{-4kp_i}) \le \Sigma e^{-4kp_i} + ((\log n)^2/n)(e^{-2c}/2)$. We next study Σe^{-4kp_i}. Using the definitions, consider $-4kp_{n+1-j} = -S/[1 - j/(n + 1)]^a$. Now $(1 - x)^{-a} \ge 1 + ax$ on $[0, 1]$ by convexity. So, $-4kp_{n+1-j} \le -[1 + (aj/n + 1)]S$. Thus,

$$\Sigma e^{-4kp_i} \le e^{-S} \Sigma e^{-\frac{ajs}{(n+1)}} = \frac{e^{-c}\log n}{n+1} \cdot \frac{1 - e^{-ans/(n+1)}}{1 - e^{-as/(n+1)}} \sim \frac{e^{-c}}{a}.$$

This bound, with an easier bound on the second product, yields the stated upper bound. For the lower bound, use a test function of form $S(x) = \Sigma w_i(-1)^{x_i}$ with $w_i = \pm(1 - 2p_i)^k W$, $W = (w_1^2 + \cdots + w_n^2)^{1/2}$, and the sign chosen to make $w_i(1 - 2p_i)^k \ge 0$. Proceeding as in ref. 24, it is straightforward to compute the mean and variance of $S(x)$ under the distributions P_x^k and $\pi(x)$. Then Chebyshev's inequality shows that these two distributions are separated. Further details are omitted.

Theorem 5 shows that the cutoff phenomenon has a certain robustness. The fact that it was discovered at all suggests that it may be the rule rather than an exception.

Section 6. Other Distances

All of the results above have been stated for the total variation distance. This widely used distance has the following equivalent versions:

$$\|P - \pi\| = \max_{A \subset \mathcal{X}} |P(A) - \pi(A)|$$

$$= \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - \pi(x)| = \max_{\|f\| \le 1} |P(f) - \pi(f)|.$$

The first version is a direct probabilistic interpretation. The second version is 1/2 the ℓ^1 norm which is convenient for computation. The third version shows that $\|P - \pi\|$ is the usual

norm topology of measures as the dual of the bounded measurable functions.

It is natural to consider the ℓ^2 norm as a measure of distance. This has mathematical convenience but lacks a direct probabilistic interpretation. Further, it needs to be normalized in terms of the problem at hand. For example, consider a space of $2n$ points. Let $\pi(x) = 1/2n$. Let $P(x) = 1/n$ on the first n points and 0 on the last n points. Then $\|P - \pi\|_2 = \{\Sigma|P(i) - \pi(i)|^2\}^{1/2} = 1/(2n)^{1/2}$ tends to zero, but the two measures are not close. For the uniform measure π, the distance $n\|P - \pi\|_2^2$ is sensible, but in general it is not simple to see how to norm the ℓ^2 distance nor compare from problem to problem. The best available version is the χ_2 distance $\chi(P, \pi) = \Sigma_x [P(x) - \pi(x)]^2/\pi(x)$. This satisfies $\|P - \pi\| \leq \chi(P, \pi) \leq \sup(P(x) - \pi(x)/\pi(x))^2$. As shown in ref. 8 the χ_2 and maximum relative error are essentially equivalent for Markov chains. Further, as shown by Su (38), the entropy distance $\text{Ent}(\pi, P) = \Sigma_x P(x) \log P(x)/\pi(x)$ satisfies $\text{Ent}(\pi, P) \leq \log(1 + \chi(P, \pi))$. We see that many "sensible" distances are equivalent; when one is small, they all are. See Su (38) for further study of how the choice of distance affects the cutoff phenomena.

In practical problems, one may be interested in only one feature of a chain. The total variation may be large because of an unrelated different feature. For example, consider the Gilbert–Shannon–Reeds measured in *Theorem 1*. It takes $(3/2)\log_2 n$ shuffles to get random uniformly. If one is only interested in the large cycles of a permutation then ref. 3 shows that one shuffle is enough! No one knows how many shuffles are enough to have the four bridge hands dealt from 52 cards approximately equally likely (although seven shuffles suffice). Fill (22) studies a specific feature (average search cost) of the library problems studied in *Theorem 3*. They find cutoffs at different times from the variation cutoffs.

The total variation studies reported here have been important in pointing to a new phenomenon which is believed to be widespread. The careful work required to prove variation cutoffs often leads to a more or less complete understanding of the chain such that essentially any natural question can be answered.

The term cutoff phenomena first appeared in joint work with David Aldous, the main developer of the modern quantitative theory of Markov chains. Proofs of the first results in this subject were done jointly with Mehrdad Shahshahani and R. L. Graham. All of my recent work is a joint effort with Laurent Saloff-Coste. I thank them and a generation of graduate students whose work has allowed the present survey.

1. Pemantle, R. (1989) *J. Theor. Prob.* **2,** 37–50.
2. Bayer, D. & Diaconis, P. (1992) *Ann. Appl. Prob.* **2,** 294–313.
3. Diaconis, P., McGrath, M. & Pitman, J. (1995) *Combinatorica* **15,** 11–29.
4. Aldous, D. (1983) *Springer Lect. Notes Math.* **986,** 243–297.
5. Diaconis, P. & Shahshahani, M. (1987) *Siam J. Math. Anal.* **18,** 208–218.
6. Kac, M. (1947) *Am. Math. Mon.* **54,** 369–391.
7. Diaconis, P., Graham, R. & Morrison, J. (1990) *Random Struct. Algorithms* **1,** 51–72.
8. Diaconis, P. & Saloff-Coste, L. (1993) *Ann. Appl. Prob.* **3,** 696–730.
9. Quastel, J. (1992) *Comm. Pure Appl. Math.* **45,** 623–679.
10. Diaconis, P. & Shahshahani, M. (1981) *Z. Wahrsch. Verw. Geb.* **57,** 159–179.
11. Hildebrand, M. (1992) *J. Algebraic Combinatorics* **1,** 133–150.
12. Rosenthal, J. (1994) *Ann. Prob.* **22,** 398–423.
13. Porod, U. (1995) *Prob. Theor. Related Fields* **101,** 277–289.
14. Flatto, L., Odlyzko, A. & Wales, D. (1985) *Ann. Prob.* **13,** 151–178.
15. Lulov, N. (1995) Ph.D. thesis (Harvard Univ., Cambridge, MA).
16. Diaconis, P. & Saloff-Coste, L. (1994) *Geom. Funct. Anal.* **4,** 1–36.
17. Stong, R. (1995) *Ann. Prob.* **23,** 2250–2279.
18. Gluck, D. (1995) *Adv. Math.*, in press.
19. Aldous, D. & Diaconis, P. (1986) *Am. Math. Mon.* **93,** 155–177.
20. Diaconis, P. & Fill, J. (1990) *Ann. Prob.* **18,** 1483–1522.
21. Matthews, P. (1988) *J. Theor. Prob.* **1,** 411–423.
22. Fill, J. (1995) *Theor. Comp. Sci.*, in press.
23. Diaconis, P., Fill, J. & Pitman, J. (1992) *Comb. Prob. Comp.* **1,** 135–155.
24. Diaconis, P. (1986) *Group Representations in Probability and Statistics* (IMS, Hayward, CA).
25. Dobrow, R. & Fill, J. (1995) *Ann. Appl. Prob.* **5,** 20–36.
26. Saloff-Coste, L. (1994) *Math. Zeit.* **217,** 641–677.
27. Belsley, E. (1993) Ph.D. thesis (Harvard Univ., Cambridge, MA).
28. Silver, J. (1995) Ph.D. thesis (Harvard Univ., Cambridge, MA).
29. D'Aristotle, A. (1995) *J. Theor. Prob.* **8,** 321–346.
30. Greenhalgh, A. (1987) Ph.D. dissertation (Stanford Univ., Stanford, CA).
31. Hora, A. (1995) *Random Walks on Association Schemes and Their Critical Times to Reach Equilibrium,* technical report (Okayama University, Okayama, Japan).
32. Dou, C. (1992) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
33. Dou, C. & Hildebrand, M. (1994) *Ann. Prob.*, in press.
34. Wilson, D. (1995) *Prob. Theor. Related Fields*, in press.
35. Diaconis, P. & Saloff-Coste, L. (1995) *Theor. Prob.*, in press.
36. Zack, M. (1989) Ph.D. thesis (Univ. of California at San Diego, La Jolla, CA).
37. Diaconis, P. & Saloff-Coste, L. (1995) *J. Fourier Anal. Applications*, Kahane Special Issue, 189–208.
38. Su, F. (1995) Ph.D. thesis (Harvard Univ., Cambridge, MA).
39. Feller, W. (1968) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), Vol. 1, 3rd Ed., p. 378.