Assessing significance in a Markov chain without mixing

Maria Chikina[a], **Alan Frieze**[b], **and Wesley Pegden**[b,1]

[a]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15213; and [b]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved January 24, 2017 (received for review October 21, 2016)

We present a statistical test to detect that a presented state of a reversible Markov chain was not chosen from a stationary distribution. In particular, given a value function for the states of the Markov chain, we would like to show rigorously that the presented state is an outlier with respect to the values, by establishing a *p* value under the null hypothesis that it was chosen from a stationary distribution of the chain. A simple heuristic used in practice is to sample ranks of states from long random trajectories on the Markov chain and compare these with the rank of the presented state; if the presented state is a 0.1% outlier compared with the sampled ranks (its rank is in the bottom 0.1% of sampled ranks), then this observation should correspond to a *p* value of 0.001. This significance is not rigorous, however, without good bounds on the mixing time of the Markov chain. Our test is the following: Given the presented state in the Markov chain, take a random walk from the presented state for any number of steps. We prove that observing that the presented state is an ε-outlier on the walk is significant at $p=\sqrt{2\varepsilon}$ under the null hypothesis that the state was chosen from a stationary distribution. We assume nothing about the Markov chain beyond reversibility and show that significance at $p\approx\sqrt{\varepsilon}$ is best possible in general. We illustrate the use of our test with a potential application to the rigorous detection of gerrymandering in Congressional districting.

Markov chain | mixing time | gerrymandering | outlier | *p* value

The essential problem in statistics is to bound the probability of a surprising observation under a null hypothesis that observations are being drawn from some unbiased probability distribution. This calculation can fail to be straightforward for a number of reasons. On the one hand, defining the way in which the outcome is surprising requires care; for example, intricate techniques have been developed to allow sophisticated analysis of cases where multiple hypotheses are being tested. On the other hand, the correct choice of the unbiased distribution implied by the null hypothesis is often not immediately clear; classical tools like the t test are often applied by making simplifying assumptions about the distribution in such cases. If the distribution is well-defined but is not be amenable to mathematical analysis, a p value can still be calculated using bootstrapping if test samples can be drawn from the distribution.

A third way for p value calculations to be nontrivial occurs when the observation is surprising in a simple way and the null hypothesis distribution is known but where there is no simple algorithm to draw samples from this distribution. In these cases, the best candidate method to sample from the null hypothesis is often through a Markov chain, which essentially takes a long random walk on the possible values of the distribution. Under suitable conditions, theorems are available that guarantee that the chain converges to its stationary distribution, allowing a random sample to be drawn from a distribution quantifiably close to the target distribution. This principle has given rise to diverse applications of Markov chains, including to simulations of chemical reactions, Markov chain Monte Carlo statistical methods, protein folding, and statistical physics models.

A persistent problem in applications of Markov chains is the often unknown rate at which the chain converges with the stationary distribution (1, 2). It is rare to have rigorous results on the mixing time of a real-world Markov chain, which means that, in practice, sampling is performed by running a Markov chain for a "long time" and hoping that sufficient mixing has occurred. In some applications, such as in simulations of the Potts model from statistical physics, practitioners have developed modified Markov chains in the hopes of achieving faster convergence (3), but such algorithms have still been shown to have exponential mixing times in many settings (4–6).

In this article, we are concerned with the problem of assessing statistical significance in a Markov chain without requiring results on the mixing time of the chain or indeed, any special structure at all in the chain beyond reversibility. Formally, we consider a reversible Markov chain $\mathcal{M}$ on a state space Σ, which has an associated label function $\omega:\Sigma\to\Re$. (The definition of Markov chain is recalled at the end of this section.) The labels constitute auxiliary information and are not assumed to have any relationship to the transition probabilities of $\mathcal{M}$. We would like to show that a presented state σ_0 is unusual for states drawn from a stationary distribution π. If we have good bounds on the mixing time of $\mathcal{M}$, then we can simply sample from a distribution of $\omega(\pi)$ and use bootstrapping to obtain a rigorous p value for the significance of the smallness of the label of σ_0. However, such bounds are rarely available.

We propose the following simple and rigorous test to detect that σ_0 is unusual relative to states chosen randomly according to π, which does not require bounds on the mixing rate of $\mathcal{M}$.

The $\sqrt{\varepsilon}$ test. Observe a trajectory $\sigma_0,\sigma_1,\sigma_2\ldots,\sigma_k$ from the state σ_0 for any fixed k. The event that $\omega(\sigma_0)$ is an ε-outlier among $\omega(\sigma_0),\ldots,\omega(\sigma_k)$ is significant at $p=\sqrt{2\varepsilon}$ under the null hypothesis that $\sigma_0\sim\pi$.

Here, we say that a real number α_0 is an ε-outlier among $\alpha_0,\alpha_2,\ldots,\alpha_k$ if there are, at most, $\varepsilon(k+1)$ indices i for which

Significance

Markov chains are simple mathematical objects that can be used to generate random samples from a probability space by taking a random walk on elements of the space. Unfortunately, in applications, it is often unknown how long a chain must be run to generate good samples, and in practice, the time required is often simply too long. This difficulty can preclude the possibility of using Markov chains to make rigorous statistical claims in many cases. We develop a rigorous statistical test for Markov chains which can avoid this problem, and apply it to the problem of detecting bias in Congressional districting.

Author contributions: M.C., A.F., and W.P. performed research and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: wes@math.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617540114/-/DCSupplemental.

$\alpha_i \le \alpha_0$. In particular, note for the $\sqrt{\varepsilon}$ test that the only relevant feature of the label function is the ranking that it imposes on the elements of Σ. In *SI Text*, we consider the statistical power of the test and show that the relationship $p \approx \sqrt{\varepsilon}$ is best possible. We leave as an open question whether the constant $\sqrt{2}$ can be improved.

Roughly speaking, this kind of test is possible, because a reversible Markov chain cannot have many local outliers (Fig. 1). Rigorously, the validity of the test is a consequence of the following theorem.

Theorem 1.1. Let $\mathcal{M} = X_0, X_1, \ldots$ be a reversible Markov chain with a stationary distribution π, and suppose the states of $\mathcal{M}$ have real-valued labels. If $X_0 \sim \pi$, then for any fixed k, the probability that the label of X_0 is an ε-outlier from among the list of labels observed in the trajectory $X_0, X_1, X_2, \ldots, X_k$ is, at most, $\sqrt{2\varepsilon}$.

We emphasize that Theorem 1.1 makes no assumptions on the structure of the Markov chain beyond reversibility. In particular, it applies even if the chain is not irreducible (in other words, even if the state space is not connected), although in this case, the chain will never mix.

In Detecting Bias in Political Districting, we apply the test to Markov chains generating random political districting for which no results on rapid mixing exist. In particular, we show that, for various simple choices of constraints on what constitutes a “valid” Congressional districting (e.g., that the districts are contiguous and satisfy certain geometric constraints), the current Congressional districting of Pennsylvania is significantly biased under the null hypothesis of a districting chosen at random from the set of valid districting. (We obtain p values between $\approx 2.5 \cdot 10^{-4}$ and $\approx 8.1 \cdot 10^{-7}$ for the constraints that we considered.)

One hypothetical application of the $\sqrt{\varepsilon}$ test is the possibility of rigorously showing that a chain is not mixed. In particular, suppose that Research Group 1 has run a reversible Markov chain for n_1 steps and believes that this was sufficient to mix the chain. Research Group 2 runs the chain for another n_2 steps, producing a trajectory of total length $n_1 + n_2$, and notices that a property of interest changes in these n_2 additional steps. Heuristically, this observation suggests that n_1 steps were not sufficient to mix the chain, and the $\sqrt{\varepsilon}$ test quantifies this reasoning rigorously. For this application, however, we must allow X_0 to be distributed not exactly as the stationary distribution π but as some distribution π' with total variation distance to π that is small, as is the scenario for a “mixed” Markov chain. In *SI Text*, we give a version of Theorem 1.1, which applies in this scenario.

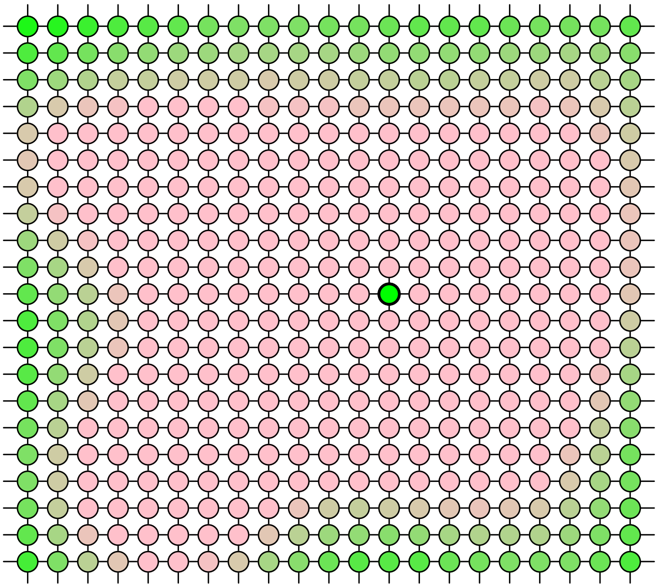

Fig. 1. This schematic illustrates a region of a potentially much larger Markov chain with a very simple structure; from each state seen here, a jump is made with equal probability to each of four neighboring states. Colors from green to pink represent labels from small to large, respectively. It is impossible to know from this local region alone whether the highlighted green state has unusually small label in this chain overall. However, to an unusual degree, this state is a local outlier. The $\sqrt{\varepsilon}$ test is based on the fact that no reversible Markov chain can have too many local outliers.

One area of research related to this manuscript concerns methods for perfect sampling from Markov chains. Beginning with the Coupling from the Past (CFTP) algorithm of Propp and Wilson (7, 8) and several extensions (9, 10), these techniques are designed to allow sampling of states exactly from the stationary distribution π without having rigorous bounds on the mixing time of the chain. Compared with the $\sqrt{\varepsilon}$ test, perfect sampling techniques have the disadvantages that they require the Markov chain to possess a certain structure for the method to be implementable and that the time that it takes to generate each perfect sample is unbounded. Moreover, although perfect sampling methods do not require rigorous bounds on mixing times to work, they will not run efficiently on a slowly mixing chain. The point is that for a chain that has the right structure and that actually mixes quickly (despite an absence of a rigorous bound on the mixing time), algorithms like CFTP can be used to rigorously generate perfect samples. However, the $\sqrt{\varepsilon}$ test applies to any reversible Markov chain, regardless of the structure, and has running time k chosen by the user. Importantly, it is quite possible that the test can detect bias in a sample even when k is much smaller than the mixing time of the chain, which seems to be the case in the districting example discussed in Detecting Bias in Political Districting. Of course, unlike perfect sampling methods, the $\sqrt{\varepsilon}$ test can only be used to show that a given sample is not chosen from π; it does not give a way for generating samples from π.

Definitions

We remind the reader that a Markov chain is a discrete time random process; at each step, the chain jumps to a new state, which only depends on the previous state. Formally, a Markov chain $\mathcal{M}$ on a state space Σ is a sequence $\mathcal{M} = X_0, X_1, X_2, \ldots$ of random variables taking values in Σ (which correspond to states that may be occupied at each step), such that, for any $\sigma, \sigma_0, \ldots, \sigma_{n-1} \in \Sigma$,

$$\begin{aligned}\mathbf{Pr}(X_n = \sigma | X_0 = \sigma_0, X_1 = \sigma_1, \ldots, X_{n-1} = \sigma_{n-1}) \\ = \mathbf{Pr}(X_1 = \sigma | X_0 = \sigma_{n-1}).\end{aligned}$$

Note that a Markov chain is completely described by the distribution of X_0 and the transition probabilities $\mathbf{Pr}(X_1 = \sigma_1 | X_0 = \sigma_0)$ for all pairs $\sigma_0, \sigma_1 \in \Sigma$. Terminology is often abused, so that the Markov chain refers only to the ensemble of transition probabilities, regardless of the choice of distribution for X_0.

With this abuse of terminology, a stationary distribution for the Markov chain is a distribution π, such that $X_0 \sim \pi$ implies that $X_1 \sim \pi$ and therefore, that $X_i \sim \pi$ for all i. When the distribution of X_0 is a stationary distribution, the Markov chain $X_0, X_1, \ldots$ is said to be stationary. A stationary chain is said to be reversible if, for all i, k, the sequence of random variables $(X_i, X_{i+1}, \ldots, X_{i+k})$ is identical in distribution to the sequence $(X_{i+k}, X_{i+k-1}, \ldots, X_i)$. Finally, a chain is reducible if there is a pair of states σ_0, σ_1, such that σ_1 is inaccessible from σ_0 via legal transitions and irreducible otherwise.

A simple example of a Markov chain is a random walk on a directed graph beginning from an initial vertex X_0 chosen from some distribution. Here, Σ is the vertex set of the directed graph. If we are allowed to label the directed edges with positive reals and if the probability of traveling along an arc is proportional to the label of the arc (among those leaving the present vertex), then any Markov chain has such a representation, because the transition probability $\mathbf{Pr}(X_1 = \sigma_1 | X_0 = \sigma_0)$ can be taken as the label of the arc from σ_0 to σ_1. Finally, if the graph is undirected, the corresponding Markov chain is reversible.

Detecting Bias in Political Districting

A central feature of American democracy is the selection of Congressional districts in which local elections are held to directly elect national representatives. Because a separate election is held in each district, the proportions of party affiliations of the slate of representatives elected in a state do not always match the proportions of statewide votes cast for each party. In practice, large deviations from this seemingly desirable target do occur.

Various tests have been proposed to detect "gerrymandering" of districting, in which a district is drawn in such a way as to bias the resulting slate of representatives toward one party, which can be accomplished by concentrating voters of the unfavored party in a few districts. One class of methods to detect gerrymandering concerns heuristic "smell tests," which judge whether districting seems generally reasonable in its statistical properties (11, 12). For example, such tests may frown on districting in which difference between the mean and median votes on district by district basis is unusually large (13).

The simplest statistical smell test, of course, is whether the party affiliation of the elected slate of representatives is close in proportion to the party affiliations of votes for representatives. Many states have failed this simple test spectacularly, such as in Pennsylvania; in 2012, 48.77% of votes were cast for Republican representatives and 50.20% of votes were cast for Democrat representatives in an election that resulted in a slate of 13 Republican representatives and 5 Democrat representatives.

Heuristic statistical tests such as these all suffer from lack of rigor, however, because of the fact that the statistical properties of "typical" districting are not rigorously characterized. For example, it has been shown (14) that Democrats may be at a natural disadvantage when drawing electoral maps, even when no bias is at play, because Democrat voters are often highly geographically concentrated in urban areas. Particularly problematic is that the degree of geographic clustering of partisans is highly variable from state to state: what looks like gerrymandered districting in one state may be a natural consequence of geography in another.

Some work has been done in which the properties of valid districting are defined (which may be required to have roughly equal populations among districts, districts with reasonable boundaries, etc.), so that the characteristics of a given districting can be compared with what would be typical for valid districting of the state in question, by using computers to generate random districting (15, 16); there is discussion of this in ref. 13. However, much of this work has relied on heuristic sampling procedures, which do not have the property of selecting districting with equal probability (and more generally, distributions that are not well-characterized), undermining rigorous statistical claims about the properties of typical districts.

In an attempt to establish a rigorous framework for this kind of approach, several groups (17–19) have used Markov chains to sample random valid districting for the purpose of such comparisons. Like many other applications of real-world Markov chains, however, these methods suffer from the completely unknown mixing time of the chains in question. Indeed, no work has even established that the Markov chains are irreducible (in the case of districting, irreducibility means that any valid districting can be reached from any other by a legal sequence of steps), even if valid districting was only required to consist of contiguous districts of roughly equal populations. Additionally, indeed, for very restrictive notions of what constitutes valid districting, irreducibility certainly fails.

As a straightforward application of the $\sqrt{\varepsilon}$ test, we can achieve rigorous p values in Markov models of political districting, despite the lack of bounds on mixing times of the chains. In particular, for all choices of the constraints on valid districting that we tested, the $\sqrt{\varepsilon}$ test showed that the current Congressional districting of Pennsylvania is an outlier at significance thresholds ranging from $p \approx 2.5 \cdot 10^{-4}$ to $p \approx 8.1 \cdot 10^{-7}$. Detailed results of these runs are in *SI Text*.

A key advantage of the Markov chain approach to gerrymandering is that it rests on a rigorous framework, namely comparing the actual districting of a state with typical (i.e., random) districting from a well-defined set of valid districting. The rigor of the approach thus depends on the availability of a precise definition of what constitutes valid districting; in principle and in practice, the best choice of definition is a legal question. Although some work on Markov chains for redistricting (in particular, ref. 19) has aimed to account for complex constraints on valid districting, our main goal in this manuscript is to illustrate the application of the $\sqrt{\varepsilon}$ test. In particular, we have erred on the side of using relatively simple sets of constraints on valid districting in our Markov chains, while checking that our significance results are not highly sensitive to the parameters that we use. However, our test immediately gives a way of putting the work, such as that in ref. 19, on a rigorous statistical footing.

The full description of the Markov chain that we use in this work is given in *SI Text*, but its basic structure is as follows: Pennsylvania is divided into roughly 9,000 census blocks. (These blocks can be seen on close inspection of Fig. 2.) We define a division of these blocks into 18 districts to be a valid districting of Pennsylvania if districts differ in population by less than 2%, are contiguous, are simply connected (districts do not contain holes), and are "compact" in ways that we discuss in *SI Text*; roughly, this final condition prohibits districts with extremely contorted structure. The state space of the Markov chain is the set of valid districting of the state, and one step of the Markov chain

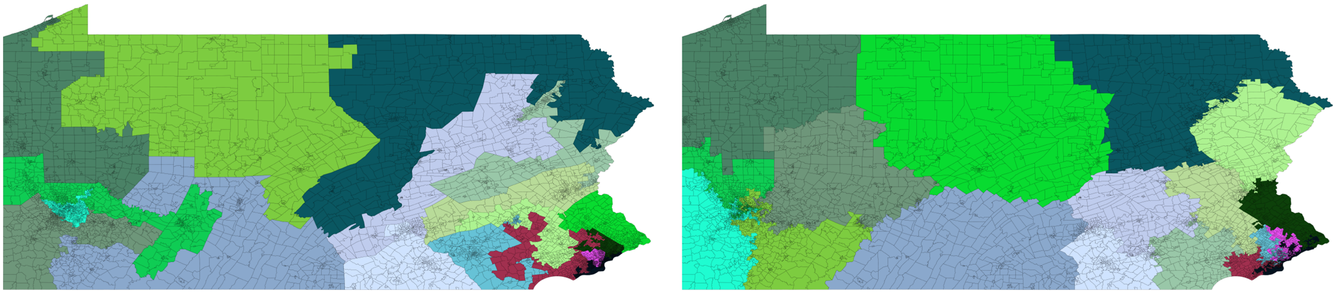

Fig. 2. (*Left*) The current districting of Pennsylvania. (*Right*) Districting produced by the Markov chain after 2^{40} steps. (Detailed parameters for this run are given in *SI Text*.)

consists of randomly swapping a precinct on the boundary of a district to a neighboring district if the result is still a valid districting. As we discuss in *SI Text*, the chain is adjusted slightly to ensure that the uniform distribution on valid districting is indeed a stationary distribution for the chain. Observe that this Markov chain has a potentially huge state space; if the only constraint on valid districting was that the districts have roughly equal population, there would be 10^{10000} or so valid districtings. Although contiguity and especially, compactness are severe restrictions that will decrease this number substantially, it seems difficult to compute effective upper bounds on the number of resulting valid districtings, and certainly, it is still enormous. Impressively, these considerations are all immaterial to our very general method.

Applying the $\sqrt{\varepsilon}$ test involves the choice of a label function $\omega(\sigma)$, which assigns a real number to each districting. We have conducted runs using two label functions: ω_{var} is the (negative) variance of the proportion of Democrats in each district of the districting (as measured by 2012 presidential votes), and ω_{MM} is the difference between the median and mean of the proportions of Democrats in each district; ω_{MM} is motivated by the fact that this metric has a long history of use in gerrymandering and is directly tied to the goals of gerrymandering, whereas the use of the variance is motivated by the fact that it can change quickly with small changes in districtings. These two choices are discussed further in *SI Text*, but an important point is that our use of these label functions is not based on an assumption that small values of ω_{var} or ω_{MM} directly imply gerrymandering. Instead, because Theorem 1.1 is valid for any fixed label function, these labels are tools used to show significance, which are chosen because they are simple and natural functions on vectors that can be quickly computed, seem likely to be different for typical versus gerrymandered districtings, and have the potential to change relatively quickly with small changes in districtings. For the various notions of valid districtings that we considered, the $\sqrt{\varepsilon}$ test showed significance at p values in the range from 10^{-4} to 10^{-5} for the ω_{MM} label function and the range from 10^{-4} to 10^{-7} for the ω_{var} label function (see Fig. S1 and Table S1).

As noted earlier, the $\sqrt{\varepsilon}$ test can easily be used with more complicated Markov chains that capture more intricate definitions of the set of valid districtings. For example, the current districting of Pennsylvania splits fewer rural counties than the districting in Fig. 2, *Right*, and the number of county splits is one of many metrics for valid districtings considered by the Markov chains developed in ref. 19. Indeed, our test will be of particular value in cases where complex notions of what constitute valid districting slow the chain to make the heuristic mixing assumption particularly questionable. Regarding mixing time, even our chain with relatively weak constraints on the districtings (and very fast running time in implementation) seems to mix too slowly to sample π, even heuristically; in Fig. 2, we see that several districts still seem to have not left their general position from the initial districting, even after 2^{40} steps.

On the same note, it should also be kept in mind that, although our result gives a method to rigorously disprove that a given districting is unbiased—e.g., to show that the districting is unusual among districtings X_0 distributed according to the stationary distribution π—it does so without giving a method to sample from the stationary distribution. In particular, our method cannot answer the question of how many seats Republicans and Democrats should have in a typical districting of Pennsylvania, because we are still not mixing the chain. Instead, Theorem 1.1 has given us a way to disprove $X_0 \sim \pi$ without sampling π.

Proof of Theorem 1.1

We let π denote any stationary distribution for $\mathcal{M}$ and suppose that the initial state X_0 is distributed as $X_0 \sim \pi$, so that in fact, $X_i \sim \pi$ for all i. We say σ_j is ℓ-small among $\sigma_0, \dots, \sigma_k$ if there are, at most, ℓ indices $i \neq j$ among $0, \dots, k$, such that the label of σ_i is, at most, the label of σ_j. In particular, σ_j is 0-small among $\sigma_0, \sigma_1, \dots, \sigma_k$ when its label is the unique minimum label, and we encourage readers to focus on this $\ell = 0$ case in their first reading of the proof.

For $0 \leq j \leq k$, we define

$$\rho^k_{j,\ell} := \mathbf{Pr}\,(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k)$$

$$\rho^k_{j,\ell}(\sigma) := \mathbf{Pr}(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k \mid X_j = \sigma).$$

Observe that, because $X_s \sim \pi$ for all s, we also have that

$$\rho^k_{j,\ell}(\sigma) = \\ \mathbf{Pr}\,(X_{s+j} \text{ is } \ell\text{-small among } X_s, \dots, X_{s+k} \mid X_{s+j} = \sigma). \qquad \textbf{[1]}$$

We begin by noting two easy facts.

Observation 4.1.

$$\rho^k_{j,\ell}(\sigma) = \rho^k_{k-j,\ell}(\sigma).$$

Proof. Because $\mathcal{M} = X_0, X_1, \dots$ is stationary and reversible, the probability that $(X_0, \dots, X_k) = (\sigma_0, \dots, \sigma_k)$ is equal to the probability that $(X_0, \dots, X_k) = (\sigma_k, \dots, \sigma_0)$ for any fixed sequence $(\sigma_0, \dots, \sigma_k)$. Thus, any sequence $(\sigma_0, \dots, \sigma_k)$ for which $\sigma_j = \sigma$ and σ_j is a ℓ-small corresponds to an equiprobable sequence $(\sigma_k, \dots, \sigma_0)$, for which $\sigma_{k-j} = \sigma$ and σ_{k-j} is ℓ-small. □

Observation 4.2.

$$\rho^k_{j,2\ell}(\sigma) \geq \rho^j_{j,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma).$$

Proof. Consider the events that X_j is an ℓ-small among $X_0, \dots, X_j$ and among $X_j, \dots, X_k$. These events are conditionally independent when conditioning on the value of $X_j = \sigma$, and $\rho^j_{j,\ell}(\sigma)$ gives the probability of the first of these events, whereas applying Eq. **1** with $s = j$ gives that $\rho^{k-j}_{0,\ell}(\sigma)$ gives the probability of the second event.

Finally, when both of these events happen, we have that X_j is 2ℓ-small among $X_0, \dots, X_k$. □

We can now deduce that

$$\rho^k_{j,2\ell}(\sigma) \geq \rho^j_{j,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma) = \rho^j_{0,\ell}(\sigma) \cdot \rho^{k-j}_{0,\ell}(\sigma) \\ \geq \left(\rho^k_{0,\ell}(\sigma)\right)^2. \qquad \textbf{[2]}$$

Indeed, the first inequality follows from Observation 4.2, the equality follows from Observation 4.1, and the final inequality follows from the fact that $\rho^k_{j,\ell}(\sigma)$ is monotone nonincreasing in k for fixed j, ℓ, σ.

Observe now that $\rho^k_{j,\ell} = E\,\rho^k_{j,\ell}(X_j)$, where the expectation is taken over the random choice of $X_j \sim \pi$.

Thus, taking expectations in Eq. **2**, we find that

$$\rho^k_{j,2\ell} = \mathbf{E}\rho^k_{j,2\ell}(\sigma) \geq \mathbf{E}\left(\left(\rho^k_{0,\ell}(\sigma)\right)^2\right) \\ \geq \left(\mathbf{E}\rho^k_{0,\ell}(\sigma)\right)^2 = \left(\rho^k_{0,\ell}\right)^2, \qquad \textbf{[3]}$$

where the second of the two inequalities is the Cauchy–Schwartz inequality.

For the final step in the proof, we sum the left- and right-hand sides of Eq. **3** to obtain

$$\sum_{j=0}^{k} \rho^k_{j,2\ell} \geq (k+1)\left(\rho^k_{0,\ell}\right)^2.$$

If we let ξ_j $(0 \leq i \leq k)$ be the indicator variable that is one whenever X_j is 2ℓ-small among $X_0, \ldots, X_k$, then $\sum_{j=0}^{k} \xi_j$ is the number of 2ℓ-small terms, which is always, at most, $2\ell + 1$. Therefore, linearity of expectation gives that

$$2\ell + 1 \geq (k+1)\left(\rho_{0,\ell}^{k}\right)^2, \qquad \textbf{[4]}$$

giving that

$$\rho_{0,\ell}^{k} \leq \sqrt{\frac{2\ell + 1}{k+1}}. \qquad \textbf{[5]}$$

Theorem 1.1 follows, because if X_i is an ε-outlier among $X_0, \ldots, X_k$, then X_i is necessarily ℓ-small among $X_0, \ldots, X_k$ for $\ell = \lfloor \varepsilon(k+1) - 1 \rfloor \leq \varepsilon(k+1) - 1$, and then, we have $2\ell + 1 \leq 2\varepsilon(k+1) - 1 \leq 2\varepsilon(k+1)$. □

ACKNOWLEDGMENTS. We thank John Nagle, Danny Sleator, and Dan Zuckerman for helpful conversations. M.C. was supported, in part, by NIH Grants 1R03MH10900901A1 and U545U54HG00854003. A.F. was supported, in part, by National Science Foundation (NSF) Grants DMS1362785 and CCF1522984 and Simons Foundation Grant 333329. W.P. was supported, in part, by NSF Grant DMS-1363136 and the Sloan Foundation.

1. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7(4):457–472.
2. Gelman A, Rubin DB (1992) A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, eds Bernardo JM, Berger JO, Dawid AP, Smith AFM (Clarendon, Gloucestershire, UK), pp 625–631.
3. Swendsen RH, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett* 58(2):86–88.
4. Borgs C, Chayes J, Tetali P (2012) Tight bounds for mixing of the Swendsen–Wang algorithm at the Potts transition point. *Probab Theory Relat Fields* 152(3-4):509–557.
5. Cooper C, Frieze A (1999) Mixing properties of the Swendsen-Wang process on classes of graphs. *Random Struct Algorithms* 15(3-4):242–261.
6. Gore VK, Jerrum MR (1999) The Swendsen–Wang process does not always mix rapidly. *J Stat Phys* 97(1-2):67–86.
7. Propp JG, Wilson DB (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct Algorithms* 9(1-2):223–252.
8. Propp J, Wilson D (1998) Coupling from the past: A user's guide. *Microsurveys in Discrete Probability*, eds Aldous DJ, Propp J (American Mathematical Society, Providence, RI), Vol 41, pp 181–192.
9. Fill JA (1997) An interruptible algorithm for perfect sampling via Markov chains. Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing (Association for Computing Machinery, New York), pp 688–695.
10. Huber M (2004) Perfect sampling using bounding chains. *Ann Appl Probab* 14(2): 734–753.
11. Wang SS-H (2016) Three tests for practical evaluation of partisan gerrymandering (December 28, 2015). *Stanford Law Rev* 68:1263–1321.
12. Nagle JF (2015) Measures of partisan bias for legislating fair elections. *Elect Law J* 14(4):346–360.
13. McDonald MD, Best RE (2015) Unfair partisan gerrymanders in politics and law: A diagnostic applied to six cases. *Elect Law J* 14(4):312–330.
14. Chen J, Rodden J (2013) Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quart J Polit Sci* 8(3):239–269.
15. Cirincione C, Darling TA, O'Rourke TG (2000) Assessing South Carolina's 1990s Congressional districting. *Polit Geogr* 19(2):189–211.
16. Rogerson PA, Yang Z (1999) The effects of spatial population distributions and political districting on minority representation. *Soc Sci Comput Rev* 17(1):27–39.
17. Fifield B, Higgins M, Imai K, Tarr A (2015) *A New Automated Redistricting Simulator Using Markov Chain Monte Carlo. Working Paper. Technical Report*. Available at imai.princeton.edu/research/files/redist.pdf. Accessed October 21, 2016.
18. Chenyun Wu L, Xiaotian Dou J, Frieze A, Miller D, Sleator D (2015) Impartial redistricting: A Markov chain approach. arXiv:1510.03247.
19. Vaughn C, Bangia S, Dou B, Guo S, Mattingly J (2016) *Quantifying Gerrymandering@Duke*. Available at https://services.math.duke.edu/projects/gerrymandering. Accessed October 21, 2016.
20. Ansolabehere S, Palmer M, Lee A (2014) *Precinct-Level Election Data, Harvard Dataverse, v1*. Available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21919. Accessed October 21, 2016.
21. Edgeworth FY (1897) Miscellaneous applications of the calculus of probabilities. *J R Stat Soc* 60(3):681–698.
22. Levin DA, Peres Y, Wilmer EL (2009) *Markov Chains and Mixing Times* (American Mathematical Society, Providence, RI).
23. Aldous D, Fill J (2002) *Reversible Markov Chains and Random Walks on Graphs*. Available at https://www.stat.berkeley.edu/~aldous/RWG/book.html. Accessed October 21, 2016.
24. Frieze A, Karoński M (2015) *Introduction to Random Graphs* (Cambridge Univ Press, Cambridge, UK).