

SEPARATING EFFECT FROM SIGNIFICANCE IN MARKOV CHAIN TESTS

MARIA CHIKINA, ALAN FRIEZE, JONATHAN MATTINGLY, AND WESLEY PEGDEN

ABSTRACT. We give qualitative and quantitative improvements to theorems which enable significance testing in Markov Chains, with a particular eye toward the goal of enabling strong, interpretable, and statistically rigorous claims of political gerrymandering. Our results can be used to demonstrate at a desired significance level that a given Markov Chain state (e.g., a districting) is extremely unusual (rather than just atypical) with respect to the fragility of its characteristics in the chain. We also provide theorems specialized to leverage quantitative improvements when there is a product structure in the underlying probability space, as can occur due to geographical constraints on districtings.

1. MOTIVATION

At its core, this note discusses improvements on a number of theorems for significance testing in Markov Chains. The improvements to the Theorem statements are both qualitative and quantitative to enable strong, easily interpretable statistical claims, and include extensions to settings where more structural assumptions lead to huge improvements in the bounds. This class of theorems is particular interest because they do not assume that the chain has converged to equilibrium. This can be of huge practical importance.

Yet, this tells only part of the story. The development of this class of algorithms and these particular extensions have been directly motivated by a question of great contemporary interest; detecting and quantifying gerrymandering.

The definiteness and correctness provided by these theorem provide substantial weight in a legal setting. The basic recipe in the gerrymandering context is the following. One starts a reversible Markov change from a particular redistricting map which claims to be typical among maps one which the Markov chain's invariant distribution is concentrated.

Date: October 24, 2019.

Research supported in part by NSF grant DMS-1362785.

Research supported in part by NSF grant DMS-1363136 and the Sloan foundation.

Operationally, this allows one to rigorously assess the likelihood of choosing a particular map if one was only considered a specific collection of non-partisan considerations. These methods (and theorems) have been used successfully by one of the authors in Gerrymandering court cases in Pennsylvania and North Carolina.

The first part of this article gives new results along these lines, extending the work in [CFP] to allow separation of effect size from the quantification of statistical significance. The second part, in Section 7, develops versions of some of these results in a special setting with a particular structure on the probability space motivated by recent legal proceedings. In particular, in balancing the federal one-person-one-vote mandate with the “keep counties whole” provision of the North Carolina Constitution, the North Carolina courts ruled in *Stephenson v. Bartlett* that a particular algorithm should be used to “cluster” the counties into independent county groups which are districted separately. This gives a product structure to the underlying probability space which can be exploited in theorems designed to take advantage of it.

2. INTRODUCTION

Consider a reversible Markov Chain \mathcal{M} whose state-space Σ is endowed with some labeling $\omega : \Sigma \rightarrow \mathbb{R}$, and for which π is a stationary distribution. \mathcal{M} , π , ω , and a fixed integer k determine a vector

$$p_0^k, p_1^k, \dots, p_k^k$$

where for each i , p_i^k is the probability that for a k -step π -stationary trajectory X_0, \dots, X_k , the minimum ω value occurs at X_i . In other words, p_i^k is the probability that if we choose X_0 randomly from the stationary distribution π and take k steps in \mathcal{M} to obtain the trajectory X_0, X_1, \dots, X_k , that we observe that $\omega(X_i)$ is the minimum among $\omega(X_0), \dots, \omega(X_k)$. Note that if we adopted the convention that we break ties among the values $\omega(X_0), \dots, \omega(X_k)$ randomly, we would have that $p_0^k + \dots + p_k^k = 1$, for any \mathcal{M} , π , and k .

At first glance, it might be natural to assume that we must have something like $p_i^k \approx \frac{1}{k+1}$ for all $0 \leq i \leq k$. But this is actually quite far from the truth; [CFP] showed that for some \mathcal{M} , π , k , we can have p_0^k as large as essentially $\frac{1}{\sqrt{2\pi k}}$.

As shown in [CFP], this is essentially the worst possible behavior for p_0^k . In particular, we can generalize the vector $\{p_i^k\}$ defined above as possible: let us define, given \mathcal{M} , π , k , and ε , the vector

$$p_{0,\varepsilon}^k, p_{1,\varepsilon}^k, \dots, p_{k,\varepsilon}^k$$

where each $p_{i,\varepsilon}^k$ is the probability that $\omega(X_i)$ is among the smallest ε values in the list $\omega(X_0), \dots, \omega(X_k)$. Then in [CFP] we proved:

Theorem 2.1. *Given a reversible Markov chain \mathcal{M} with stationary distribution π , an $\varepsilon > 0$, $k \geq 0$, and With $p_{i,\varepsilon}^k$ defined as above, we have that*

$$p_{0,\varepsilon}^k \leq \sqrt{2\varepsilon}.$$

Note that the example from [CFP] realizing $p_0^k \approx \frac{1}{\sqrt{2\pi k}}$ shows that this theorem is best possible, up to constant factors.

One important application of Theorem 2.1 is that it characterizes the statistical significance associated to the result of a natural test for gerrymandering of political districtings. In particular, consider the following general procedure to evaluate a districting of a state:

Local Outlier Test

- (1) Beginning from the districting being evaluated,
- (2) Make a sequence of random changes to the districting, while preserving some set of constraints imposed on the districtings.
- (3) Evaluate the partisan properties of each districting encountered (e.g., by simulating elections using past voting data).
- (4) Call the original districting “carefully crafted” or “gerrymandered” if the overwhelming majority of districtings produced by making small random changes are less partisan than the original districting.

Naturally, the test described above can be implemented so that it precisely satisfies the hypotheses of Theorem 2.1. For this purpose, a (very large) set of comparison districtings are defined, to which the districting being evaluated belongs. For example, the comparison districtings may be the districtings built out of Census blocks (or some other unit) which are contiguous, equal in population up to some specified deviation, or include other constraints. A Markov chain \mathcal{M} is defined on this set of districtings, where transitions in the chain correspond to changes in districtings. (For example, a transition may correspond to randomly changing the district assignment of a randomly chosen Census block which currently borders more than one district, subject to the constraints imposed on the comparison set.) The “random changes” from Step 2 will then be precisely governed by the transition probabilities of the Markov chain \mathcal{M} . By designing \mathcal{M} so that the uniform distribution π on the set of comparison districtings Σ is a stationary distribution for \mathcal{M} , Theorem 2.1 gives an upper bound on the false-positive rate (in other words, global statistical significance) for the “gerrymandered” declaration when it is made in Step 4.

Apart from its application to gerrymandering, Theorem 2.1 has a simple informal interpretation for the general behavior of reversible Markov chains, namely: *typical (i.e., stationary) states are unlikely to change in a consistent way under a sequence of chain transitions*, with a best-possible quantification of this fact (up to constant factors).

Also, in the general setting of a reversible Markov chain, the theorem leads to a simple quantitative procedure for asserting rigorously that σ_0 is atypical with respect to π without knowing the mixing time of \mathcal{M} : simply observe a random trajectory $\sigma_0 = X_0, X_1, X_2 \dots, X_k$ from σ_0 for any fixed k . If $\omega(\sigma_0)$ is an ε -outlier among $\omega(X_0), \dots, \omega(X_k)$, then this is statistically significant at $\sqrt{2\varepsilon}$ against the null hypothesis that $\sigma_0 \sim \pi$.

This quantitative test is potentially useful because $\sqrt{2\varepsilon}$ converges quickly enough to 0 as $\varepsilon \rightarrow 0$; in particular, it is possible to obtain good statistical significance from observations which can be made with reasonable computational resources. Of course, faster convergence to 0 would be even better, but, as already noted, $p \approx \sqrt{\varepsilon}$ is roughly a best possible upper bound.

Unknown to the authors at the time of the publication of [CFP], a 1989 paper of Besag and Clifford described a test related to that based on Theorem 2.1, which has essentially a one-line proof, which we discuss in Section 4:

Theorem 2.2 (Besag and Clifford serial test). *Fix any number k and suppose that σ_0 is chosen from a stationary distribution π , and that ξ is chosen uniformly in $\{0, \dots, k\}$. Consider two independent trajectories Y_0, Y_1, \dots and Z_0, Z_1, \dots in the reversible Markov Chain \mathcal{M} (whose states have real-valued labels) from $Y_0 = Z_0 = \sigma_0$. If we choose σ_0 from a stationary distribution π of \mathcal{M} , then for any k we have that*

$$\Pr(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\sigma_0), \omega(Y_1), \dots, \omega(Y_\xi), \omega(Z_1), \dots, \omega(Z_{k-\xi})) \leq \varepsilon.$$

Here, a real number a_0 is an ε -outlier among a_0, \dots, a_k if

$$\#\{i \in \{0, \dots, k\} \mid a_i \leq a_0\} \leq \varepsilon(k+1).$$

In particular, the striking thing about Theorem 2.2 is that it achieves a best-possible dependence on the parameter ε . (Notice that ε would be the correct value of the probability if, for example, the Markov chain is simply a collection of independent random samples.) The sacrifice is in Theorem 2.2's slightly more complicated intuitive interpretation, which would be: *typical (i.e., stationary) states are unlikely to change in a consistent way under two sequences of chain transitions of random complementary lengths*. In particular, in applications of these statistical tests to aspects of public policy, it is desirable to have tests with simple, intuitive interpretations. To enable better significance testing in this sphere, one goal of the present note is to prove a theorem enabling Markov chain significance testing which is intuitively interpretable in the sense of Theorem 2.1, while having linear dependence on ε , as in Theorem 2.2.

One common feature of the tests based on Theorem 2.1 and 2.2 is the use of randomness. In particular, the probability space at play in these theorems includes both the random choice of σ_0 assumed by the null hypothesis and the random steps taken by the Markov chain from σ_0 . Thus the measures of “how (globally) unusual” σ_0 is with respect to its performance in the local outlier test and “how sure” we are that σ_0 is unusual in this respect are intertwined in the final p -value. In particular, the effect size and the statistical significance are not explicitly separated.

To further the goal of simplifying the interpretation of the results of these tests, our approach in this note will also show that tests like these can be efficiently used in a way which separates the measure of statistical significance from the question of the magnitude of the effect. In particular, recalling the probabilities $p_{0,\varepsilon}^k, \dots, p_{k,\varepsilon}^k$ defined previously, let us define the probability $p_{0,\varepsilon}^k(\sigma_0)$ to be the probability that

on a trajectory $\sigma_0 = X_0, X_1, \dots, X_k$, $\omega(\sigma_0)$ is among the smallest ε fraction of the list $\omega(X_0), \dots, \omega(X_k)$. Now we make the following definition:

Definition 2.3. With respect to k , the state σ_0 is an (ε, α) -outlier in \mathcal{M} if, among all states in \mathcal{M} , $p_{0,\varepsilon}^k(\sigma_0)$ is in the largest α fraction of the values of $p_{0,\varepsilon}^k(\sigma)$ over all states $\sigma \in \mathcal{M}$, weighted according to π .

In particular, being an (ε, α) -outlier measures the likelihood of σ_0 to fail the local outlier test, ranked against *all other states* $\sigma \sim \pi$ of the chain \mathcal{M} . For example, fix $k = 10^9$. If σ_0 is a $(10^{-6}, 10^{-5})$ -outlier in \mathcal{M} and π is the uniform distribution, this means that among *all* states $\sigma \in \mathcal{M}$, σ_0 is more likely than all but a 10^{-5} fraction of states to have an ω -value in the bottom 10^{-6} values $\omega(X_0), \omega(X_1), \dots, \omega(X_{10^9})$. Note that the probability space underlying the “more likely” claim here just concerns the choice of the random trajectory X_1, \dots, X_{10^9} from \mathcal{M} .

Note that whether σ_0 is a (ε, α) -outlier is a deterministic question about the properties of σ_0, \mathcal{M} , and ω . Thus it is a deterministic measure (defined in terms of certain probabilities) of the extent to which σ_0 is unusual (globally, in all of \mathcal{M}) with respect to its local fragility in the chain.

The following theorem enables one to assert statistical significance for the property of being an (ε, α) -outlier. In particular, while tests based on Theorems 2.1 and 2.2 take as their null hypothesis that $\sigma_0 \sim \pi$, the following theorem takes as its null hypothesis merely that σ_0 is not an (ε, α) -outlier.

Theorem 2.4. Consider m independent trajectories

$$\begin{aligned} \mathcal{T}^1 &= (X_0^1, X_1^1, \dots, X_k^1), \\ &\vdots \\ \mathcal{T}^m &= (X_0^m, X_1^m, \dots, X_k^m) \end{aligned}$$

of length k in the reversible Markov Chain \mathcal{M} (whose states have real-valued labels) from a common starting point $X_0^1 = \dots = X_0^m = \sigma_0$. Define the random variable ρ to be the number of trajectories \mathcal{T}^i on which σ_0 is an ε -outlier.

If σ_0 is not an (ε, α) -outlier, then

$$(1) \quad \Pr \left(\rho \geq m \sqrt{\frac{2\varepsilon}{\alpha}} + r \right) \leq e^{-\min(r^2 \sqrt{\alpha/2\varepsilon}/3m, r/3)}.$$

In particular, apart from separating measures of statistical significance from the quantification of a local outlier, Theorem 2.4 connects the intuitive Local Outlier Test tied to Theorem 2.1 (which motivates the definition of a (ε, α) -outlier) to the better quantitative dependence on ε in Theorem 2.2.

To compare the quantitative performance of Theorem 2.4 to Theorems 2.1 and 2.2, consider the case of a state σ_0 for which a random trajectory $\sigma_0 = X_0, X_1, \dots, X_k$ is likely (say with some constant probability p') to find σ_0 an ε' -outlier. For Theorem

2.1, significance at $p \approx \sqrt{2\varepsilon}$ would be obtained¹, while using Theorem 2.2, one would hope to obtain significance of $\approx \varepsilon'$. Applying Theorem 2.4, we would expect to see ρ around $m \cdot p'$. In particular, we could demonstrate that σ_0 is an (ε', α) outlier for $\alpha = \frac{3\varepsilon}{(p')^2}$ (a linear dependence on ε) at a p -value which can be made arbitrarily small (at an exponential rate) as we increase the number of observed trajectories m . As we will see in Section 5, the exponential tail in (1) can be replaced by a binomial tail. In particular, the following special case applies:

Theorem 2.5. *With $\mathcal{T}^1, \dots, \mathcal{T}^m$ as in Theorem 2.4, we have that if σ_0 is not an (ε, α) outlier, then*

$$\Pr(\sigma_0 \text{ an } \varepsilon\text{-outlier on all of } \mathcal{T}^1, \dots, \mathcal{T}^m) \leq \left(\frac{2\varepsilon}{\alpha}\right)^{m/2}.$$

Theorem 2.5 also has advantages from the standpoint of avoiding the need to correct for multiple hypothesis testing, as we discuss in Section 3.

To prove Theorem 2.4, we will prove the following, which has a quantitative dependence on ε which is nearly as strong as in Theorem 2.2, while eliminating the need for the random choice of ξ there.

Theorem 2.6. *Consider two independent trajectories Y_0, \dots, Y_k and Z_0, \dots, Z_k in the reversible Markov Chain \mathcal{M} (whose states have real-valued labels) from a common starting point $Y_0 = Z_0 = \sigma_0$. If we choose σ_0 from a stationary distribution π of \mathcal{M} , then for any k we have that*

$$\Pr(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\sigma_0), \omega(Y_1), \dots, \omega(Y_k), \omega(Z_1), \dots, \omega(Z_k)) < 2\varepsilon.$$

Note that Theorem 2.6 is equivalent to the statement that the probabilities $p_{i,\varepsilon}^k$ always satisfy

$$(2) \quad p_{k,\varepsilon}^{2k} < 2\varepsilon.$$

Remark 2.7. As in the case of Theorem 2.1, it seems like an interesting question to investigate the tightness of the constant 2; we will see in Section 7 that there are settings where the impact of this constant is inflated to have outside-importance. We point out here that at least for the case of $k = 1$, $\varepsilon = 1/3$, $\rho_{1,\frac{1}{3}}^2$ can be at least as large as $\frac{1}{2}$, showing that the constant 2 in (2) cannot be replaced by a constant less than $\frac{3}{2}$, in general. To see this, consider, for example, a bipartite complete graph $K_{n,n}$, where the labels of the vertices of one side are $1, \dots, n$ and the other are $n+1, \dots, 2n$. For the Markov chain given by the random walk on this undirected graph, we have that $\rho_{1,\frac{1}{3}}^2 = \frac{1}{2}$. Note that for this example, it is still the case that $\rho_{k,\varepsilon}^{2k} \rightarrow \varepsilon$ as $k \rightarrow \infty$, leaving open the possibility that the 2 in (2) can be replaced with an expression asymptotically equivalent to 1.

¹Multiple tests have limited utility here or with Theorem 2.2 since there is no independence (the null hypothesis $\sigma_0 \sim \pi$ is not being resampled). In particular, multiple runs might be done merely until a trajectory is seen on which σ_0 is indeed an ε' outlier (requiring $1/p'$ runs, on average), in conjunction with multiple hypothesis testing.

The following theorem is the analog of Theorem 2.4 obtained when one uses an analog of Besag and Clifford's Theorem 2.2 in place of 2.6 in the proof. This version pays the price of using a random k instead of a fixed k for the notion of an (ε, α) -outlier, but has the advantage that the constant 2 is eliminated from the bound. (Note that as in Theorem 2.4, the notion of (ε, α) -outlier used here is still just defined with respect to a single path, although Theorem 2.2 depends on using two independent trajectories.)

Theorem 2.8. *Consider m independent trajectories*

$$\mathcal{T}^1 = (X_0^1, X_1^1, \dots, X_{k_1}^1),$$

$$\vdots$$

$$\mathcal{T}^m = (X_0^m, X_1^m, \dots, X_{k_m}^m)$$

in the reversible Markov Chain \mathcal{M} (whose states have real-valued labels) from a common starting point $X_0^1 = \dots = X_0^m = \sigma_0$, where each of the lengths k_i are independently drawn random numbers from a geometric distribution. Define the random variable ρ to be the number of trajectories \mathcal{T}^i on which σ_0 is an ε -outlier.

If σ_0 is not an (ε, α) -outlier with respect to k drawn from the geometric distribution, then

$$(3) \quad \Pr\left(\rho \geq m\sqrt{\frac{\varepsilon}{\alpha}} + r\right) \leq e^{-\min(r^2\sqrt{\alpha/\varepsilon}/3m, r/3)}.$$

Again, there is an analogous version to Theorem 2.5, where 2ε is replaced by ε .

In their paper, Besag and Clifford also describe a parallel test, which we will discuss in Section 6. In particular, in Section 6 we will describe a test which generalizes Besag and Clifford's serial and parallel tests in a way which could be useful in certain parallel regimes.

Finally, we consider an interesting case in the analysis of districtings that arises when the districting problem can be decomposed into several non-interacting districting problems; for example, for the districting for the state Senate of North Carolina, the state is divided into 29 "county clusters", each corresponding to a prescribed number of districts based on their populations, so that a districting of the whole state is obtained by non-interacting districting processes in these different county clusters. In this case, the probability space of random districtings is really a product space, and this structure can be exploited in a strong way for the statistical tests developed in this manuscript. We develop results for this setting in Section 7.

3. MULTIPLE HYPOTHESIS CONSIDERATIONS

When applying Theorem 2.4 directly, one cannot simply run m trajectories, observe the list $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$ where each ε_i is the minimum ε_i for which σ_0 is an ε_i -outlier on \mathcal{T}^i , and then, post-hoc, freely choose the parameters α and ε in Theorem 2.4 to achieve some desired trade-off between α and the significance p .

The problem, of course, is that in this case one is testing multiple hypotheses (infinitely many in fact; one for each possible pair ε and α) which would require a multiple hypothesis correction.

One way to avoid this problem is to essentially do a form of cross validation, were a few trajectories are run for the purposes of selecting suitable ε and α , and then discarded from the set of trajectories from which we obtain significance.

A simpler approach, however, is to simply set the parameter $\varepsilon = \varepsilon_{(t)}$ as the t th-smallest element of the list $\varepsilon_1, \dots, \varepsilon_m$ for some fixed value t . The case $t = m$, for example, corresponds to taking ε as the maximum value, leading to the application of Theorem 2.5.

The reasons this avoids the need for a multiple hypothesis correction is that we can order our hypothesis events by containment. In particular, when we apply this test with some value of t , we will always have $\rho = t$. Thus the significance obtained will depend just on the parameter $\varepsilon_{(t)}$ returned by taking the t -th smallest ε_i and on our choice of α (as opposed to say, the particular values of the other ε_i 's which are not the t -th smallest). In particular, regardless of how we wish to trade-off the values of α and p we can assert from our test, our optimum choice of α (for our fixed choice of t) will depend just on the value $\varepsilon_{(t)}$. In particular, we can view α as a function $\alpha(\varepsilon_{(t)})$, so that we when applying Theorem 2.4 with with $\varepsilon = \varepsilon_{(t)}$, we are evaluating the single-parameter infinite family of hypotheses $H_{\varepsilon_{(t)}, \alpha(\varepsilon_{(t)})}$, and we do not require multiple hypothesis correction since the hypotheses are nested; i.e., since

$$(4) \quad \varepsilon_{(t)} \leq \varepsilon'_{(t)} \implies H_{\varepsilon_{(t)}, \alpha(\varepsilon_{(t)})} \subseteq H_{\varepsilon'_{(t)}, \alpha(\varepsilon_{(t)})}.$$

Indeed, (4) implies that

$$\Pr \left(\bigcup_{\varepsilon_{(t)} \leq \beta} H_{\varepsilon_{(t)}, \alpha(\varepsilon_{(t)})} \right) = \Pr(H_{\beta, \alpha(\beta)}),$$

which ensures that when applying Theorem 2.4 in this scenario, the probability of returning a p -value $\leq p_0$ for any fixed value p_0 will indeed be at most p_0 .

4. PROOF BACKGROUND

We begin this section by giving the proof of Theorem 2.6. In doing so we will introduce some notation that will be useful throughout the rest of this note. To make things as accessible as possible, we give every detail of the proof.

In this manuscript, a Markov Chain \mathcal{M} on Σ is specified by the transition probabilities $\{\pi_{\sigma_1, \sigma_2} \mid \sigma_1, \sigma_2 \in \Sigma\}$ of a chain. A *trajectory* of \mathcal{M} is a sequence of random variables X_0, X_1, \dots required to have the property that for each i and $\sigma_0, \dots, \sigma_i$, we have

$$(5) \quad \Pr(X_i = \sigma_i \mid X_{i-1} = \sigma_{i-1}, X_{i-2} = \sigma_{i-2} \dots, X_0 = \sigma_0) = \pi_{\sigma_i, \sigma_{i-1}}.$$

In particular, the Markov property of the trajectory is that the conditioning on X_{i-2}, X_{i-3}, \dots is irrelevant once we condition on the value of X_{i-1} . Recall that π is a stationary distribution if $X_0 \sim \pi$ implies that $X_1 \sim \pi$ and thus also that $X_i \sim \pi$ for all $i \geq 0$; in this case we that the trajectory X_0, X_1, \dots is π -stationary. The Markov Chain \mathcal{M} is *reversible* if any π -stationary trajectory X_0, \dots, X_k is equivalent in distribution to its reverse X_k, \dots, X_0 .

We say that a_j is ℓ -small among a_0, \dots, a_s if there are at most ℓ indices $i \neq j$ among $0, \dots, s$ such that $a_i \leq a_j$. The following simple definition is at the heart of the proofs of Theorems 2.1, 2.6, 2.2.

Definition 4.1. Given a Markov Chain \mathcal{M} with labels $\omega : \Sigma \rightarrow \mathbb{R}$ and stationary distribution π , we define for each $\ell, j \leq k$ a real number $\rho_{j,\ell}^k$, which is the probability that for a π -stationary trajectory X_0, X_1, \dots, X_k , we have that $\omega(X_j)$ is ℓ -small among $\omega(X_0), \dots, \omega(X_k)$.

Observe that (5) implies that all π -stationary trajectories of a fixed length are all identical in distribution, and in particular, that the $\rho_{j,\ell}^k$'s are well-defined.

Next observe that if the sequence of random variables X_0, X_1, \dots is a π -stationary trajectory for \mathcal{M} , then so is any interval of it. For example,

$$(X_{k-j}, \dots, X_k, \dots, X_{2k-j})$$

is another stationary trajectory, and thus the probability that $\omega(X_k)$ is ℓ -small among $\omega(X_{k-j}), \dots, \omega(X_{2k-j})$ is equal to $\rho_{j,\ell}^k$. In particular, since

$$(\omega(X_k) \text{ is } \ell\text{-small among } \omega(X_{k-j}), \dots, \omega(X_{2k-j}))$$

follows from

$$(\omega(X_k) \text{ is } \ell\text{-small among } \omega(X_0), \dots, \omega(X_{2k}))$$

for all $j = 0, \dots, k$, we have that

$$(6) \quad \rho_{k,\ell}^{2k} \leq \rho_{j,\ell}^k.$$

We also have that $\sum_{j=0}^k \rho_{j,\ell}^k \leq \ell + 1$. Indeed, by linearity of expectation, this sum is the expected number of indices $j \in 0, \dots, k$ such that $\omega(X_j)$ is ℓ -small among $\omega(X_0), \dots, \omega(X_k)$. Thus, averaging the left and right sides of (6) over j from 0 to k , we obtain

$$(7) \quad \rho_{k,\ell}^{2k} \leq \frac{\ell + 1}{k + 1} < 2 \cdot \frac{\ell + 1}{2k + 1}.$$

Line (7) already gives the theorem, once we make the following trivial observation:

Observation 4.2. *Under the hypotheses of Theorem 2.6, we have that*

$$Y_k, Y_{k-1}, \dots, Y_1, \sigma_0, Z_1, Z_2, \dots, Z_k$$

is a π -stationary trajectory.

This is an elementary consequence of the definitions, but since we will generalize this statement in Section 6, we give all the details here:

Proof of Observation 4.2. Our hypothesis is that Y_1, Y_2, \dots, Y_k and Z_1, Z_2, \dots, Z_k are independent trajectories from a common state $Y_0 = Z_0 = \sigma_0$ chosen from the stationary distribution π . Stationarity implies that

$$(Z_0, Z_1, \dots, Z_k) \sim (X_k, X_{k+1}, \dots, X_{2k}).$$

Similarly, stationarity and reversibility imply that

$$(Y_k, Y_{k-1}, \dots, Y_0) \sim (X_0, X_1, \dots, X_k).$$

Finally, our assumption that Y_1, Y_2, \dots and Z_1, Z_2, \dots are independent trajectories from σ_0 is equivalent to the condition that, for any $s_0, y_1, z_1, y_2, z_2, \dots, y_k, z_k \in \Sigma$, we have for all $j \geq 0$ that

$$(8) \quad \Pr(Z_j = z_j \mid Z_{j-1} = z_{j-1}, \dots, Z_1 = z_1, Z_0 = Y_0 = s_0, Y_1 = y_1, \dots, Y_k = y_k) \\ = \Pr(Z_j = z_j \mid Z_{j-1} = z_{j-1}, \dots, Z_1 = z_1, Z_0 = s_0)$$

Of course, since \mathcal{M} is a Markov Chain, this second probability is simply

$$\Pr(Z_j = z_j \mid Z_{j-1} = z_{j-1}) = \Pr(X_{k+j} = z_j \mid X_{k+j-1} = z_{j-1}).$$

In particular, by induction on $j \geq 1$,

$$(Y_k, Y_{k-1}, \dots, Y_0 = Z_0, Z_1, \dots, Z_j) \sim (X_0, X_1, \dots, X_k, X_{k+1}, \dots, X_{k+j}),$$

and in particular

$$(9) \quad (Y_k, \dots, \sigma_0, \dots, Z_k) \sim (X_0, \dots, X_k, \dots, X_{2k}).$$

□

Pared down to its bare minimum, this proof of Theorem 2.6 works by using that $\rho_{k,\ell}^{2k}$ is a lower bound on each $\rho_{j,\ell}^k$, and then applying the simple inequality

$$(10) \quad \sum_{j=0}^k \rho_{j,\ell}^k \leq \ell + 1.$$

The proof of Theorem 2.2 of Besag and Clifford is in some sense even simpler, using only (10), despite the fact that Theorem 2.2 has better dependence on ε (on the other hand, it is not directly applicable to (ε, α) -outliers in the way that we will use Theorem 2.6). Recall from Definition 4.1 that the $\rho_{j,\ell}^k$'s are fixed real numbers associated to a stationary Markov Chain. If ℓ, k are fixed and ξ is chosen randomly from 0 to k , then the resulting $\rho_{\xi,\ell}^k$ is a random variable uniformly distributed on the set of real numbers $\{\rho_{0,\ell}^k, \rho_{1,\ell}^k, \dots, \rho_{k,\ell}^k\}$. In particular, Theorem 2.2 is proved by writing that the probability that $\omega(\sigma_0)$ is ℓ -small among $\omega(\sigma_0), \omega(Y_1), \dots, \omega(Y_\xi), \omega(Z_1), \dots, \omega(Z_{k-\xi})$ is given by

$$\frac{1}{k+1} (\rho_{0,\ell}^k + \rho_{1,\ell}^k + \dots + \rho_{k,\ell}^k) \leq \frac{\ell+1}{k+1},$$

where the inequality is from (10). Note that we are using an analog of Observation 4.2 to know that for any j , $Y_j, \dots, Y_1, \sigma_0, Z_1, Z_{k-j}$ is a π -stationary trajectory.

5. GLOBAL SIGNIFICANCE FOR LOCAL OUTLIERS

We now prove Theorem 2.4 from Theorem 2.6.

Proof of Theorem 2.4. For a π -stationary trajectory X_0, \dots, X_k , let us define $p_{j,\varepsilon}^k(\sigma)$ to be the probability that $\omega(X_j)$ is in the bottom ε fraction of the values $\omega(X_0), \dots, \omega(X_k)$, *conditioned* on the event that $X_j = \sigma$.

In particular, to prove Theorem 2.4, we will prove the following claim:

Claim: If σ_0 is not an (ε, α) -outlier, then

$$(11) \quad p_{0,\varepsilon}^k(\sigma_0) \leq \sqrt{\frac{2\varepsilon}{\alpha}}.$$

Let us first see why the claim implies the theorem. Recall the random variable ρ is the number of trajectories \mathcal{T}^i from σ_0 on which σ_0 is observed to be an ε -outlier with respect to the labeling ω . The random variable ρ is thus a sum of m independent Bernoulli random variables, which each take value 1 with probability $\leq \sqrt{\frac{2\varepsilon}{\alpha}}$ by the claim. In particular, by Chernoff's bound, we have

$$(12) \quad \Pr\left(\rho \geq (1 + \delta)m\sqrt{\frac{2\varepsilon}{\alpha}}\right) \leq e^{-\min(\delta, \delta^2)m\sqrt{\frac{2\varepsilon}{\alpha}}/3},$$

giving the theorem. (Note the key point of the claim is that α is *inside* the square root in (11), while a straightforward application of Theorem 2.1 would give an expression with α outside the square root.)

To prove (11), consider a π -stationary trajectory $X_0, \dots, X_k, \dots, X_{2k}$ and condition on the event that $X_k = \sigma$ for some arbitrary $\sigma \in \Sigma$. Since \mathcal{M} is reversible, we can view this trajectory as two independent trajectories X_{k+1}, \dots, X_{2k} and $X_{k-1}, X_{k-2}, \dots, X_0$ both beginning from σ . In particular, letting A and B be the events that $\omega(X_k)$ is an ε -outlier among the lists $\omega(X_0), \dots, \omega(X_k)$ and $\omega(X_k), \dots, \omega(X_{2k})$, respectively, we have that

$$(13) \quad p_{0,\varepsilon}^k(\sigma)^2 = \Pr(A \cap B) \leq p_{k,\varepsilon}^{2k}(\sigma).$$

Now, the assumption that the given $\sigma_0 \in \Sigma$ is not an (ε, α) -outlier gives that for a random $\sigma \sim \pi$, we have that

$$(14) \quad \Pr(p_{0,\varepsilon}^k(\sigma) \geq p_{0,\varepsilon}^k(\sigma_0)) \geq \alpha.$$

Line 13 gives that $p_{0,\varepsilon}^k(\sigma)^2 \leq p_{k,\varepsilon}^{2k}(\sigma)$, and Theorem 2.6 gives that $p_{k,\varepsilon}^{2k} \leq 2\varepsilon$. Thus taking expectations with respect to a random $\sigma \sim \pi$, we obtain that

$$\mathbf{E}_{\sigma \sim \pi} (p_{0,\varepsilon}^k(\sigma)^2) \leq \mathbf{E}_{\sigma \sim \pi} (p_{k,\varepsilon}^{2k}(\sigma)) = p_{k,\varepsilon}^{2k} \leq 2\varepsilon.$$

On the other hand, we can use (14) to write

$$\mathbf{E}_{\sigma \sim \pi} (p_{0,\varepsilon}^k(\sigma)^2) \geq \alpha \cdot p_{0,\varepsilon}^k(\sigma_0)^2,$$

so that we have

$$p_{0,\varepsilon}^k(\sigma_0)^2 \leq \frac{2\varepsilon}{\alpha}.$$

□

The proof of Theorem 2.8 is quite similar:

Proof of Theorem 2.8. For a π -stationary trajectory X_0, \dots, X_k and a real number μ , let us define $p_{0,\varepsilon}^\mu(\sigma)$ to be the probability that $\omega(X_j)$ is in the bottom ε fraction of the values $\omega(X_0), \dots, \omega(X_k)$, *conditioned* on the event that $X_0 = \sigma$, where the length k is chosen from a geometric distribution with mean μ supported on $0, 1, 2, \dots$; i.e., $k = t$ with probability $\frac{1}{\mu+1}(1 - \frac{1}{\mu+1})^t$.

To prove Theorem 2.8, it suffices to prove that if σ_0 is not an (ε, α) -outlier with respect to k drawn from the geometric distribution with mean μ , then

$$(15) \quad p_{0,\varepsilon}^\mu(\sigma_0) \leq \sqrt{\frac{\varepsilon}{\alpha}}.$$

To prove (15), suppose that k_1 and k_2 are independent random variables which are geometrically distributed with mean μ , and consider a π -stationary trajectory

$$X_0, \dots, X_{k_1}, \dots, X_{k_1+k_2}$$

of random length k_1+k_2 , and condition on the event that $X_{k_1} = \sigma$ for some arbitrary $\sigma \in \Sigma$. Since \mathcal{M} is reversible, we can view this trajectory as two independent trajectories $X_{k_1}, X_{k_1+1}, \dots, X_{k_1+k_2}$ and $X_{k_1}, X_{k_1-1}, X_{k_1-2}, \dots, X_0$ both beginning from $X_{k_1} = \sigma$, of random lengths k_2 and k_1 , respectively. In particular, letting A and B be the events that $\omega(X_{k_1})$ is an ε -outlier among the lists $\omega(X_0), \dots, \omega(X_{k_1})$ and $\omega(X_{k_1}), \dots, \omega(X_{k_1+k_2})$, respectively, we have that

$$(16) \quad p_{0,\varepsilon}^\mu(\sigma)^2 = \Pr(A \cap B) \\ \leq \Pr(\omega(X_{k_1}) \text{ is an } \varepsilon\text{-outlier among } \omega(X_0), \dots, \omega(X_{k_1+k_2}) \mid X_{k_1} = \sigma)$$

where, in this last expression, k_1 and k_2 are random variables. Now, the assumption that the given $\sigma_0 \in \Sigma$ is not an (ε, α) -outlier gives that for a random $\sigma \sim \pi$, we have that

$$(17) \quad \Pr(p_{0,\varepsilon}^\mu(\sigma) \geq p_{0,\varepsilon}^\mu(\sigma_0)) \geq \alpha.$$

Thus we write

$$(18) \quad \alpha \cdot p_{0,\varepsilon}^\mu(\sigma_0)^2 \leq \mathbf{E}_{\sigma \sim \pi} (p_{0,\varepsilon}^\mu(\sigma)^2) \\ \leq \Pr(\omega(X_{k_1}) \text{ is an } \varepsilon\text{-outlier among } \omega(X_0), \dots, \omega(X_{k_1+k_2})),$$

where the last inequality follows from line (16).

On the other hand, considering the righthand side of Line (18), we have that conditioning on any value for the length $\ell = k_1 + k_2$ of the trajectory, k_1 is uniformly distributed in the range $\{0, \dots, \ell\}$. This is ensured by the geometric distribution, simply because for any ℓ and any $x \in (0, \dots, \ell)$, we have that the probability

$$\Pr(k_1 = x \text{ AND } k_2 = \ell - x) = \frac{1}{\mu+1}(1 - \frac{1}{\mu+1})^x \frac{1}{\mu+1}(1 - \frac{1}{\mu+1})^{\ell-x} = \left(\frac{1}{\mu+1}\right)^2 \left(1 - \frac{1}{\mu+1}\right)^\ell$$

is independent of x . In particular, conditioning on any particular value for the length $\ell = k_1 + k_2$, we have that the probability that $\omega(X_{k_1})$ is an ε -outlier on the trajectory is at most ε , since X_{k_1} is a uniformly randomly chosen element of the trajectory $X_0, \dots, X_{k_1+k_2}$; note that this part of the proof is exactly the same as

the proof of Theorem 2.2. In particular, for the righthand-side of line (18), we are writing

$$\begin{aligned}
 (19) \quad \alpha \cdot p_{0,\varepsilon}^\mu(\sigma_0)^2 &\leq \Pr(\omega(X_{k_1}) \text{ is an } \varepsilon\text{-outlier among } \omega(X_0), \dots, \omega(X_{k_1+k_2})) \\
 &\leq \max_{\ell} \Pr(\omega(X_{k_1}) \text{ is an } \varepsilon\text{-outlier among } \omega(X_0), \dots, \omega(X_{k_1+k_2}) \mid k_1 + k_2 = \ell) \\
 &\leq \varepsilon.
 \end{aligned}$$

This gives line (15) and completes the proof. \square

We close this section by noting that in implementations where m is not enormous, it may be sensible to use the exact binomial tail in place of the Chernoff bound in (12). In particular, this gives the following versions:

Theorem 5.1. *With ρ as in Theorem 2.4, we have that if σ_0 is not an (ε, α) outlier, then*

$$(20) \quad \Pr(\rho \geq K) \leq \sum_{k=K}^m \binom{m}{k} \left(\frac{2\varepsilon}{\alpha}\right)^{k/2} \left(1 - \sqrt{\frac{2\varepsilon}{\alpha}}\right)^{m-k}.$$

Theorem 5.2. *With ρ as in Theorem 2.8, we have that if σ_0 is not an (ε, α) outlier, then*

$$(21) \quad \Pr(\rho \geq K) \leq \sum_{k=K}^m \binom{m}{k} \left(\frac{\varepsilon}{\alpha}\right)^{k/2} \left(1 - \sqrt{\frac{\varepsilon}{\alpha}}\right)^{m-k}.$$

6. GENERALIZING THE BESAG AND CLIFFORD TESTS

Theorem 2.4 is attractive because it succeeds at separating statistical significance from effect size, and at demonstrating statistical significance for an intuitively-interpretable deterministic property of state in the Markov Chain. This is especially important when public-policy decisions must be made by non-experts on the basis of such tests.

In some cases, however, these may not be important goals. In particular, one may simply desire a statistical test which is as effective as possible at disproving the null hypothesis $\sigma \sim \pi$. This is a task at which Besag and Clifford's Theorem 2.2 excels.

In their paper, Besag and Clifford also prove the following result, to enable a test designed to take efficient advantage of parallelism:

Theorem 6.1 (Besag and Clifford parallel test). *Fix numbers k and m . Suppose that σ_0 is chosen from a stationary distribution π of the reversible Markov Chain \mathcal{M} , and suppose we sample a trajectory X_1, X_2, \dots, X_k from $X_0 = \sigma_0$, and then branch to sample $m - 1$ trajectories $Z_1^s, Z_2^s, \dots, Z_k^s$ ($2 \leq s \leq m$) all from the state $Z_0^s = X_k$. Then we have that*

$$\Pr(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\sigma_0), \omega(Z_k^2), \omega(Z_k^3), \dots, \omega(Z_k^m)) \leq \varepsilon.$$

Proof. For this theorem it suffices to observe that $\sigma_0, Z_k^2, \dots, Z_k^m$ are *exchangable* random variables—that is, all permutations of the sequence $\sigma_0, Z_k^2, \dots, Z_k^m$ are identical in distribution. This is because if σ_0 is chosen from π and then the Z_k^i 's are chosen as above, the result is equivalent in distribution to the case where X_k is chosen from π and then each Z_k^i is chosen (independently) as the end of a trajectory X_k, Z_1^i, \dots, Z_k^i , and $\sigma_0 = Y_k$ is chosen (independently) as the end of a trajectory X_k, Y_1, \dots, Y_k . Here we are using that reversibility implies that (X_k, Y_1, \dots, Y_k) is identical in distribution to $(\sigma_0, X_1, \dots, X_k)$. \square

With an eye towards finding a common generalization of Besag and Clifford's serial and parallel tests, we define a *Markov outlier test* as a significance test with the following general features:

- The test begins from a state σ_0 of the Markov Chain which, under the null hypothesis, is assumed to be stationary;
- random steps in the Markov chain are sampled from the initial state and/or from subsequent states exposed by the test;
- the ranking of the initial state's label is compared among the labels of some (possibly all) of the visited states; it is an ε -outlier if it's label is among the bottom ε of the comparison labels. Some function $\rho(\varepsilon)$ assigns valid statistical significance to the test results, as in the above theorems.

In particular, such a test may consist of single or multiple trajectories, may branch once or multiple times, etc. In this section, we prove the validity of a parallelizable Markov outlier test with best possible function $\rho(\varepsilon) = \varepsilon$, but for which it is natural to expect the ε -power of the test—that is, its tendency to return small values of ε when σ_0 truly is an outlier—surpasses that of Theorems 2.2 and 6.1. In particular, we prove the following theorem:

Theorem 6.2 (Star-split test). *Fix numbers m and k . Suppose that σ_0 is chosen from a stationary distribution π of the reversible Markov Chain \mathcal{M} , and suppose that ξ is chosen randomly in $\{1, \dots, k\}$. Now sample trajectories X_1, \dots, X_ξ and $Y_1, \dots, Y_{k-\xi}$ from σ_0 , and then branch and sample $m-1$ trajectories $Z_1^s, Z_2^s, \dots, Z_k^s$ ($2 \leq s \leq m$) all from the state $Z_0^s = X_\xi$. Then we have that*

$$\Pr \left(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\sigma_0), \omega(X_1), \dots, \omega(X_{\xi-1}), \right. \\ \omega(Y_1), \dots, \omega(Y_{k-\xi}), \\ \omega(Z_1^2), \dots, \omega(Z_k^2) \\ \vdots \\ \left. \omega(Z_k^m), \dots, \omega(Z_k^m) \right) \leq \varepsilon.$$

In particular, note that the set of comparison random variables used consists of all random variables exposed by the test *except* X_ξ .

To compare Theorem 6.2 with Theorems 6.1 and 2.2, let us note that it is natural to expect the ε -power of a Markov chain significance test to depend on:

- (a) How many comparisons are generated by the test, and
- (b) how far typical comparison states are from the state being tested, where we measure distance to a comparison state by the number of Markov chain transitions which the test used to generate the comparison.

If unlimited parallelism is available, then the Besag/Clifford parallel test is essentially optimal from these parameters, as it draws an unlimited number of samples, whose distance from the initial state is whatever serial running time is used. Conversely, in a purely serial setting, the Besag/Clifford test is essentially optimal with respect to these parameters.

But it is natural to expect that even when parallelism is available, the number n of samples we desire will often be significantly greater than the parallelism factor ℓ available. In this case, the Besag/Clifford parallel test will use n comparisons at distance $d \approx \ell t/n$, where t is the serial time used by the test. In particular, the typical distance to a comparison can be considerably less than t when ℓ compares unfavorably with n .

On the other hand, Besag/Clifford serial test generates comparisons whose typical distance is roughly $t/2$, but cannot make use of parallelism beyond $\ell = 2$. For an apples-to-apples comparison, it is natural to consider the case of carrying out their serial test using only every d th state encountered as a comparison state for some d . This is equivalent to applying the test to the d th-power of the Markov chain, instead of applying it directly. (In practical applications, this is a sensible choice when comparing the labels of states is expensive relative to the time required to carry out transitions of the chain.) Now if ℓ is a small constant, we see that with $t \cdot d$ steps, the BC parallel test can generate roughly n comparisons all at distance d from the state being tested, the serial test could generate comparisons at distances $d, 2d, 3d, \dots, kd$ (measured in terms of transitions in \mathcal{M}), where these distances occur with multiplicity at most 2, and $k = \max(\xi, n - \xi) \geq n/2$. In particular, the serial test generates a similar number of comparisons in this way but at much greater distances from the state we are evaluating, making it more likely that we are able to detect that the input state is an outlier.

Consider now the star-split test. Again, to facilitate comparison, we suppose the test is being applied to the d th power of \mathcal{M} . If serial time $t \approx sd$ is to be used, then we will branch into $\ell - 1$ trajectories after $\xi \cdot \mathcal{M}^d$ chain, where ξ is randomly chosen from $\{0, \frac{s}{2}\}$. Thus comparisons used lie at a set of distances $d, 2d, \dots, (\xi + \frac{s}{2})d$ similar to the case of the Besag/Clifford serial test above. But now the distances $d, 2d, \dots, (\xi d - 1)d$ will have multiplicities at most 2 in the set of comparison distances, while the distances $(\xi + 1)d, (\xi + 2)d, \dots, (\xi + \frac{s}{2})d$ all have multiplicity at least $\ell - 1$. In particular, the test allows us to make more comparisons to more distance states, essentially by a factor of the parallelism factor being used. In particular, it is natural to expect performance to improve as ℓ increases. Moreover, the star-split test is equivalent to the Besag/Clifford serial test for $\ell \leq 2$, and essentially equivalent to their parallel test in the large ℓ limit. (To make this latter correspondence exact, one can apply Theorem 6.2 to the d th power of a Markov chain \mathcal{M} , and take $k = 1$.)

We now turn to the task of proving Theorem 6.2. Unlike Theorems 2.1, 2.6, and 2.2, the comparison states used in Theorems 6.1 and 6.2 cannot be viewed as a single trajectory in \mathcal{M} . This motivates the natural generalization of the notion of a π -stationary trajectory as follows:

Definition 6.3. Given a reversible Markov Chain \mathcal{M} with stationary distribution π and an undirected tree T , a π -stationary T -projection is a collection of random variables $\{X_v\}_{v \in T}$ such that:

- (i) for all $v \in T$, $X_v \sim \pi$;
- (ii) for any edge $\{u, v\}$ in T , if we let T_u denote the vertex-set of the connected component of u in $T \setminus \{u, v\}$ and $\{\sigma_w\}_{w \in T}$ is an arbitrary collection of states, then

$$\Pr \left(X_v = \sigma_v \mid \bigwedge_{w \in T_u} X_w = \sigma_w \right) = \pi_{\sigma_u, \sigma_v}.$$

In analogy to the case of π -stationary trajectories, Definition 6.3 easily gives the following, by induction:

Observation 6.4. For fixed π and T , if $\{X_w\}_{w \in T}$ and $\{Y_w\}_{w \in T}$ are both π -stationary T -projections, then the two collections $\{X_w\}_{w \in T}$ and $\{Y_w\}_{w \in T}$ are equivalent in distribution. \square

This enables the following natural analog of Definition 4.1:

Definition 6.5. Given a Markov Chain \mathcal{M} with labels $\omega : \Sigma \rightarrow \mathbb{R}$ and stationary distribution π , we define for each ℓ , each undirected tree T , each vertex subset $S \subset T$ and each vertex $v \in S$ a real number $\rho_{v, \ell}^{T, S}$, which is the probability that for a π -stationary T -projection $\{X_w\}_{w \in T}$, we have that $\omega(X_v)$ is ℓ -small among $\{\omega(X_w)\}_{w \in S}$.

Observe that as in (10) we have for any tree T and any vertex subset S of T , we have that

$$(22) \quad \sum_{w \in S} \rho_{w, \ell}^{T, S} \leq \ell + 1.$$

The following Observation, applied recursively, gives the natural analog of Observation 4.2. Again the proof is an easy exercise in the definitions.

Observation 6.6. Suppose that T is an undirected tree, v is a leaf of T , $T' = T \setminus v$, and $\{X_w\}_{w \in T'}$ is a π -stationary T' -projection. Suppose further that X_v is a random variable such that for all $\{\sigma_w\}_{w \in T}$ we have that

$$(23) \quad \Pr \left(X_v = \sigma_v \mid \bigwedge_{w \in T'} (X_w = \sigma_w) \right) = \pi_{\sigma_u, \sigma_v},$$

where u is the neighbor of v in T . Then $\{X_w\}_{w \in T}$ is a π -stationary T -projection. \square

We can rephrase the proof of Theorem 6.1 in this language. Let T be the tree consisting of m paths of length k sharing a common endpoint and no other vertices, and let S be the leaves of T . By symmetry, we have that $\rho_{w,\ell}^{T,S}$ is constant over $w \in S$. On the other hand, Observation 6.6 gives that under the hypotheses of Theorem 6.1, $\sigma_0, X_1, \dots, X_k$, and the Z_i^s 's are a π -stationary T -projection, with obvious assignments (e.g., σ_0 corresponds to a leaf of T ; X_k corresponds to the center). In particular, (22) implies that $\rho_{v,\ell}^{T,S} \leq \frac{\ell+1}{n}$, which gives the theorem.

On the other hand, the definitions makes the following proof easy as well, using the same simple idea as Besag and Clifford's Theorem 2.2.

Proof of Theorem 6.2. Define T to be the undirected tree with vertex set $\{v_0\} \cup \{v_j^s \mid 1 \leq s \leq m, 1 \leq j \leq k\}$, with edges $\{v_0, v_1^s\}$ for each $1 \leq s \leq m$ and $\{v_j^s, v_{j+1}^s\}$ for each $1 \leq s \leq m, 1 \leq j \leq k-1$. Now we let S consist of all vertices of T except the center v_0 , and let S_j denote the set of m vertices in S at distance j from v_0 . By symmetry, we have that $\rho_{v,\ell}^{T,S}$ is constant in each S_j ; in particular, we have that

$$\rho_{v_1^s,\ell}^{T,S} = \frac{1}{n} \sum_{s=1}^m \rho_{v_j^s,\ell}^{T,S}$$

and together with (22) this gives that

$$(24) \quad \sum_{j=1}^k \rho_{v_j^s,\ell}^{T,S} \leq \frac{\ell+1}{n}.$$

Now if we let

$$W_{v_0} = X_k, \\ W_{v_j^s} = \begin{cases} X_{\xi-j} & s = 1, 1 \leq j < \xi \\ \sigma_0 & s = 1, j = \xi \\ Y_{j-\xi} & s = 1, j > \xi \\ Z_j^s & 2 \leq s \leq m, 1 \leq j \leq k, \end{cases}$$

then $\{W_w\}_{w \in T}$ is a π -stationary T -projection under the hypotheses of Theorem 6.2, by recursively applying Observation 6.6. Moreover, as ξ is chosen randomly among $\{1, \dots, k\}$, the probability that $\omega(\sigma_0) = \omega(W_{v_\xi^1})$ is ℓ -small among $\{\omega(W_w)\}_{w \in S}$ is given by

$$\frac{1}{k} \left(\rho_{v_1^1,\ell}^{T,S} + \dots + \rho_{v_k^1,\ell}^{T,S} \right) \leq \frac{\ell+1}{kn},$$

where the inequality is from (24), giving the Theorem. \square

7. THE PRODUCT SPACE SETTING

The appeal of the theorems developed thus far in this paper is that they can be applied to any reversible Markov chain without any knowledge of its structure. However, there are some important cases where additional information about the

structure of the stationary distribution of a chain *is* available, and can be exploited to enable more powerful statistical claims.

In this section, we consider the problem of evaluating claims of gerrymandering with a Markov Chain where the probability distribution on districtings is known to have a product structure imposed by geographical constraints. For example, the North Carolina Supreme Court has ruled in *Stephenson v. Bartlett* that districtings of that state must respect groupings of counties determined by a prescribed algorithm. In particular a set of explicit rules (nearly) determine a partition of the counties of North Carolina into county groupings whose populations are each close to an integer multiple of an ideal district size (see [CHTHM] for recent results on these rules), and then the districting of the state is comprised of independent districtings of each of the county groupings.

In this way, the probability space of uniformly random districtings is a product space, with a random districting of the whole state equivalent to collection of random independent districtings of each of the separate county groupings. We wish to exploit this structure for greater statistical power. In particular, running trajectories of length k in each of d clusters generates a total of k^d comparison maps with only $k \cdot d$ total Markov chain steps. To take advantage of the potential power of this enormous comparison set, we need theorems which allow us to compare a given map not just to a trajectory of maps in a Markov chain (since the k^d maps do not form a trajectory) but to the product of trajectories. This is what we show in this section.

Formally, in the product space setting, we have a collection $\mathcal{M}^{[d]}$ of d Markov Chains $\mathcal{M}_1, \dots, \mathcal{M}_d$, each \mathcal{M}_i on state space Σ_i (each corresponding to one county grouping in North Carolina, for example). We are given a label function $\omega : \Sigma^{[d]} \rightarrow \mathbb{R}$, where here $\Sigma^{[d]} = \Sigma_1 \times \dots \times \Sigma_d$. In the first theorem in this section, which is a direct analog of the Besag and Clifford test, we consider a $\sigma_0 \in \Sigma^{[d]}$ distributed as $\sigma_0 \sim \pi^{[d]}$, where here $\pi^{[d]}$ indicates the product space of stationary distributions π_i of the \mathcal{M}_i . (In the gerrymandering case, $\pi^{[d]}$ is a random map chosen by randomly selecting a map for each separate county cluster.) In the tests discussed earlier in this paper, a state $\sigma_0 \sim \mathcal{M}$ is evaluated by comparing a state σ_0 to other states on a trajectory containing σ_0 . In the product setting, we compare σ_0 against a product of one trajectory from each \mathcal{M}_i .

In particular, given the collection $\mathcal{M}^{[d]}$, a state $\sigma_0 = (\sigma_0^1, \dots, \sigma_0^d) \in \Sigma^{[d]}$, and $\mathbf{j} = (j_1, \dots, j_d)$, $\mathbf{k} = (k_1, \dots, k_d)$, we define the *trajectory product* $\mathbf{X}_{\sigma_0, \mathbf{j}, \mathbf{k}}$ which is obtained by considering, for each i , a trajectory $X_0^i, \dots, X_{k_i}^i$ in \mathcal{M}_i conditioned on $X_{j_i}^i = \sigma_0^i$. $\mathbf{X}_{\sigma_0, \mathbf{j}, \mathbf{k}}$ is simply the set of all d -tuples consisting of one element from each such trajectory.

We define the *stationary trajectory product* $\mathbf{X}_{\pi^{[d]}, \mathbf{k}}$, analogously, except that the trajectories used are all stationary, instead of conditioning on $X_{j_i}^i = \sigma_0^i$.

Theorem 7.1. *Given reversible Markov Chains $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_d$, fix any number k and suppose that $\sigma_0^1, \dots, \sigma_0^d$ are chosen from stationary distributions π_1, \dots, π_d of $\mathcal{M}_1, \dots, \mathcal{M}_d$, and that ξ_1, \dots, ξ_d are chosen uniformly and independently in*

$\{0, \dots, k\}$. For each $s = 1, \dots, d$, consider two independent trajectories Y_0^s, Y_1^s, \dots and Z_0^s, Z_1^s, \dots in the reversible Markov Chain \mathcal{M}_s from $Y_0^s = Z_0^s = \sigma_0^s$. Let $\omega : \mathcal{M}_1 \times \dots \times \mathcal{M}_d \rightarrow \mathbb{R}$ be a label function on the product space, write $\sigma_0 = (\sigma_0^1, \dots, \sigma_0^d)$, and denote by $\mathbf{Z}_{\sigma_0, k}$ the (random) set of all vectors (a_1, \dots, a_d) such that for each i , $a_i \in (\sigma_0^i, Y_1^i, \dots, Y_{\xi_i}^i, Z_1^i, \dots, Z_{k-\xi_i}^i)$. Then we have that

$$(25) \quad \Pr(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\mathbf{x}), \mathbf{x} \in \mathbf{Z}_{\sigma_0, k}) \leq \varepsilon.$$

Proof. Like the proof of Theorem 2.2, this proof is very simple; it is just a matter of digesting notation. First observe that $\mathbf{Z}_{\sigma_0, k}$ is simply a trajectory product $\mathbf{X}_{\sigma_0, \xi, \mathbf{k}}$, where where $\mathbf{k} = (k, \dots, k)$ and ξ is the random variable (ξ_1, \dots, ξ_d) .

In particular, under the hypothesis that $\sigma_0^i \sim \pi_i$ for all i , $\mathbf{Z}_{\sigma_0, k}$ is in fact a stationary trajectory product $\mathbf{X}_{\pi^{[d]}, \mathbf{k}}$. In particular, by the random, independent choice of the ξ_i 's, the probability in (25) is equivalent to the probability that the label of a random element of the a stationary trajectory product is among ε smallest labels in the stationary trajectory product; this probability is at most ε . \square

The following is an analog of Theorem 2.6 for the product space setting.

Theorem 7.2. *Given reversible Markov Chains $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_d$, fix any number k and suppose that $\sigma_0^1, \dots, \sigma_0^d$ are chosen from stationary distributions π_1, \dots, π_d of $\mathcal{M}^1, \dots, \mathcal{M}^d$. For each $s = 1, \dots, d$, consider two independent trajectories Y_0^s, Y_1^s, \dots and Z_0^s, Z_1^s, \dots in the reversible Markov Chain \mathcal{M}^s from $Y_0^s = Z_0^s = \sigma_0^s$. Let $\omega : \mathcal{M}_1 \times \dots \times \mathcal{M}_d \rightarrow \mathbb{R}$ be a label function on the product space, write $\sigma_0 = (\sigma_0^1, \dots, \sigma_0^d)$, and denote by $\mathbf{Z}_{\sigma_0, k}$ the (random) set of all vectors (a_1, \dots, a_d) such that for each i , $a_i \in (\sigma_0^i, Y_1^i, \dots, Y_k^i, Z_1^i, \dots, Z_k^i)$. Then we have that*

$$(26) \quad \Pr(\omega(\sigma_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\mathbf{x}), \mathbf{x} \in \mathbf{Z}_{\sigma_0, k}) \leq 2^d \cdot \varepsilon.$$

Proof. First consider d independent stationary trajectories $X_0^i, X_1^i, X_2^i, \dots$ for each $i = 1, \dots, d$, and define $\mathbf{X}_{\pi, k}$ to be the collection of all $(k+1)^d$ d -tuples (a_1, \dots, a_d) where, for each i , $a_i \in \{X_0^i, \dots, X_k^i\}$.

In analogy to Definition 4.1, we define $\rho_{\mathbf{j}, \ell}^k$ for $\mathbf{j} = (j_1, j_2, \dots, j_k)$ to be the probability that for $X_{\mathbf{j}} = (X_{j_1}^1, \dots, X_{j_d}^d) \in \mathbf{X}_{\pi, k}$, we have that $\omega(X_{\mathbf{j}})$ is ℓ -small among the ω -labels of all elements of $\mathbf{X}_{\pi, k}$.

Observe that for $\mathbf{k} = (k, \dots, k)$, we have in analogy to equation (6) that

$$(27) \quad \rho_{\mathbf{k}, \ell}^{2k} \leq \rho_{\mathbf{j}, \ell}^k$$

for any $\mathbf{j} = (j_1, \dots, j_d)$. And of course we have that

$$\sum_{\mathbf{j}} \rho_{\mathbf{j}, \ell}^k \leq \ell + 1.$$

Thus averaging both sides of (27) gives that

$$(28) \quad \rho_{\mathbf{k}, \ell}^{2k} \leq \frac{\ell + 1}{(k+1)^d} \leq 2^d \frac{\ell + 1}{(2k+1)^d}.$$

Now observe that the the statement that

$$\omega(\boldsymbol{\sigma}_0) \text{ is an } \varepsilon\text{-outlier among } \omega(\mathbf{x}), \mathbf{x} \in \mathbf{Z}_{\boldsymbol{\sigma}_0, k}$$

equivalent to the statement that

$$\omega(\boldsymbol{\sigma}_0) \text{ is an } \ell\text{-small among } \omega(\mathbf{x}), \mathbf{x} \in \mathbf{Z}_{\boldsymbol{\sigma}_0, k}$$

for $\ell = \varepsilon \cdot (2k + 1)^d - 1$; thus (28) gives the theorem, since $\rho_{\mathbf{k}, \ell}^{2k}$ is precisely the probability that this second statement holds. \square

The presence of the 2^d in (26) is now potentially more annoying than the constant 2 in (2.6), and it is natural to ask whether it can be avoided. However, using the example from Remark 2.7, it is easy to see that an exponential factor $(\frac{3}{2})^d$ may really be necessary, at least if $k = 1$. Whether such a factor can be avoided for larger values of k is an interesting question. However, as we discuss below, this seemingly large exponential penalty is actually likely dwarfed by the quantitative benefits of the product setting, in many real-world cases.

7.1. Illustrative product examples. The fact the estimate in Theorem 7.1 looks like original Theorem 2.2, hides the power in the product version. More misleading is the fact that Theorem 7.2 has a 2^d which seems to make the theorem degrade with increasing d .

Let us begin by considering the simplest example we are looking for the single extreme outlier across the entire product space. Let us further assume that this global extreme is obtained by choosing each of the extreme element in each part of the product space. An example of this comes for the Gerrymandering application where one is naturally interested in the seat count. Each of the product coordinates represents the seats from a particular geographic region. In some states such as North Carolina judicial rulings break the problem up into the product measure required by Theorem 7.1 and Theorem 7.2 by stipulating that particular geographic regions must be redistricted independently.

For illustrative purposes, lets assume that there are L different outcomes in each of the d different factors of the product space. Hence the chance of getting the minimum in any of the d different components is $1/L$. However, getting the minimum in the whole product space requires getting the minimum in each of the components and so is $1/L^d$. Hence in this setting one can take $\epsilon = 1/L^d$ in Theorem 7.1 and Theorem 7.2. Thus even in Theorem 7.2 as long as $L > 2$, one has a significant improvement as d grows.

Now lets consider a second slightly more complicated example which builds on the proceeding one. Let us equip each \mathcal{M}_i with a function ω_i and decide that we are interested in the event

$$(29) \quad \mathcal{E}(\delta) = \left\{ \{\xi_i\}_1^d : \sum_{i=1}^d \omega_i(\xi_i) \leq \delta \right\}.$$

Then one can take

$$\epsilon = \frac{|\mathcal{E}(\delta)|}{L^d}$$

in Theorem 7.2 and 2^d times this in Theorem 7.2, where $|\mathcal{E}(\delta)|$ is simply the number of elements in the set $\mathcal{E}(\delta)$. This can lead to a significant improvement in the power of the test in the product case over the general case when $|\mathcal{E}(\delta)|$ grows slower than L^d .

There remains the task of calculating $|\mathcal{E}(\delta)|$. In the gerrymandering examples we have in mind, this can be done efficiently. When counting seat counts, the map ω_i is a many-to-one map with a range consisting of a few discrete values. This means that one can tabulate exactly the number of samples which produce a given value of ω_i . Since we are typically interested extreme values of

$$\omega(\xi) = \sum_{i=1}^d \omega_i(\xi_i),$$

there are often only a few partitions of each value of ω made from possible values of ω_i . When this true, the size of \mathcal{E} can be calculated exactly efficiently.

For example, let us assume there are d geographical regions which each needs to be divided into 4 districts. Furthermore each party always wins at least one seat in each geographical region; hence, the only possible outcomes are 1, 2 or 3 seats in each region for a given party. If ω_i counts the number of seats for the party of interest in geographic region i , let us suppose for concreteness that we want are interested in $\delta = 2d$. To calculate $|\mathcal{E}(\delta)|$, we need to only keep track of the number of times 1, 2 or 3 seats is produced in each geographic region. We can then combine these numbers by summing over all of the ways the numbers 1, 2 and 3 can add numbers between d and $2d$. (The smallest $\omega(\xi)$ can be given our assumptions is d .) This is a straightforward calculation for which there exist fast algorithms which leverage the hierarchical structure. Namely, group each region with another and calculate the combined possible seat counts and their frequencies. Continuing up the tree recursively one can calculate $|\mathcal{E}(\delta)|$ in only logarithmically many levels.

It is worth remarking, that not all statistics of interest fall as neatly into this framework which enables simple and efficient computation. For instance, calculating the ranked marginals used in [HSLGBRM] requires choosing some representation of the histogram, such as a fixed binning, and would yield only approximate results.

7.2. Towards an (ϵ, α) -outlier theorem for product spaces. In general, the cost of making a straightforward translation of Theorems 2.4 or 2.8 to the product-space setting are surprisingly large: in both cases, the square root is replaced by a 2^d th root, according to the natural generalization of the proofs of those theorems.

Accordingly, in this section we point out simply that by using a more complicated definition of (ϵ, α) -outliers for the product space setting, an analog of Theorem 2.8 is then easy. In particular, let us define

$$(30) \quad p_{\mathbf{U}, \epsilon}^{\mathbf{k}}(\sigma_0) := \Pr(\omega(\sigma_0) \text{ an } \epsilon\text{-outlier in } \mathbf{X}_{\sigma_0, \mathbf{j}, \mathbf{k}}),$$

where $\mathbf{j} = (j_1, \dots, j_d)$ is chosen randomly with respect to the uniform distributions $j_i \sim \text{Unif}[0, k_i]$ (here $\mathbf{k} = (k_1, \dots, k_d)$).

Now we define a state σ_0 to be an (ε, α) -outlier with respect to a distribution \mathbf{k} if among all states in $\Sigma^{[d]}$, we have that $p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma_0)$ is in the the largest α fraction of the values of $p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma)$ over *all* states $\sigma \in \mathcal{M}^{[d]}$, weighted according to π .

Theorem 7.3. *We are given Markov Chains $\mathcal{M}_1, \dots, \mathcal{M}_d$. Suppose that σ_0 is not an (ε, α) -outlier with respect to \mathbf{k} . Then*

$$p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma_0) \leq \frac{\varepsilon}{\alpha}.$$

Proof. This follows immediately from the definitions. From the definition of (ε, α) -outlier given above for the product setting, we have that if σ_0 is not an (ε, α) -outlier, then for a random $\sigma \sim \pi$,

$$\Pr\left(p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma) \geq p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma_0)\right) \geq \alpha.$$

Thus we can write

$$\mathbf{E}_{\sigma \sim \pi} p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma) \geq \alpha \cdot p_{\mathbf{U}, \varepsilon}^{\mathbf{k}}(\sigma_0).$$

And of course this expectation is just the probability that a random element of $\mathbf{X}_{\pi, \mathbf{k}}$ is an ε -outlier on $X_{\pi, \mathbf{k}}$, which is at most ε . \square

Of course this kind of trivial proof would be possible in the general non-product space setting also, but the sacrifice is that (ε, α) -outliers cannot be defined with respect to the endpoints of trajectories, which appears most natural. Whether theorems analogous to 2.4 and 2.8 are possible in the product space setting without an explosive dependence on the dimension d seems like a very interesting question.

REFERENCES

- [CFP] M. Chikina, A. Frieze, W. Pegden. Assessing significance in a Markov Chain without mixing, in *Proceedings of the National Academy of Sciences* **114** 2860–2864.
- [CHTHM] D. Carter, Z. Hunter, D. Teague, G. Herschlag, J. Mattingly. Optimal Legislative County Clustering in North Carolina, arXiv e-prints, arXiv:1908.11801
- [HSLGBRM] Herschlag, G., Kang, H. S., Luo, J., et al. 2018, arXiv e-prints, arXiv:1801.03783

DEPARTMENT OF COMPUTATIONAL AND SYSTEMS BIOLOGY, UNIVERSITY OF PITTSBURGH, 3078 BIOMEDICAL SCIENCE TOWER 3, PITTSBURGH, PA 15213, U.S.A.

E-mail address, email: `mchikina@pitt.edu`

E-mail address, email: `alan@random.math.cmu.edu`

E-mail address, email: `jonm@math.duke.edu`

E-mail address, email: `wes@math.cmu.edu`

DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PA
15213, U.S.A.