HDSR
HARVARD DATA SCIENCE REVIEW

Volume 0 Issue 0
DOI: 00.000
ISSN: 2644-2353

Recombination: A Family of Markov Chains for Redistricting

Daryl DeFord[†], Moon Duchin[‡,*], Justin Solomon[◇]

[†] Washington State University

[‡] Tufts University

[◇] Massachusetts Institute of Technology

Abstract. *Redistricting* is the problem of partitioning a set of geographic units into a fixed number of subsets called *districts*, subject to a list of rules and priorities. These districts are used for elections, making their delineation highly consequential. It has been hard for quantitative researchers to orient to an application domain in which rule vagueness can be a feature rather than a bug—but law and policy often prefer a reasonable range of values to a paradigm of optimization. In recent years, the use of randomized methods to sample from the vast space of districting plans has been gaining traction in U.S. courts of law for identifying partisan gerrymanders, and it is now emerging as a promising assessment tool for legislatures and independent commissions, even before districts are enacted. In this article, we set up redistricting as a graph partition problem and introduce a new family of Markov chains called *recombination* (or ReCom). We focus on the use of spanning trees for recombination, an idea introduced by our research group in 2018 and now in wide use in the redistricting field. ReCom is a large-step random walk on the space of graph partitions, in contrast with commonly used Flip walks, which change the assignment label of one or a few nodes at a time. Important points of comparison concern the speed of convergence to stationarity, the form of the target distribution, and the characteristics of samples that can be obtained in practical time. We use real-world data to demonstrate advantages of spanning tree ReCom and its relative weighting of plans, and we give a broader exposition of both the challenges of this approach and the analytical tools that it enables. We close with a short case study of race and redistricting in the Virginia House of Delegates.

Keywords: redistricting, gerrymandering, Markov chains, sampling

[*] moon.duchin@tufts.edu

Media Summary

Districts that hold plurality elections are the favored American device for converting votes into political representation. But control of the district lines can confer a surprising degree of power over the outcomes, even under real-world conditions—abusing this power is called 'gerrymandering.' Computational methods are quickly gaining traction as a way to understand the magnitude of gerrymandering and to tease its effects apart from the mere consequences of the system. Proposed districting plans can now be put in context by comparison to large samples of alternative valid plans, holding the political geography constant. In this article, we rethink the question of what makes a good sample for this application. We present and motivate a novel sampling method using spanning trees and use real data to show the method in action. Though the universe of plausible districting plans is forbiddingly vast, our heuristic tests suggest that we can now use randomized techniques to construct a *representative sample* in a reasonable time.

1. Introduction

In many countries, geographic regions are divided into electoral districts, such as when states are divided into districts that elect individual members to the U.S. House of Representatives. The task of drawing district boundaries, or *redistricting*, is fraught with technical, practical, political, and even philosophical challenges, and the ultimate choice of a districting plan has major consequences in terms of which groups are able to elect their candidates of choice. Even the best-intentioned map-drawers have a formidable task in drawing plans whose structure promotes basic fairness principles set out in law or widely held in public opinion. The ease of achieving an agenda through control of redistricting makes it common for line-drawers to *gerrymander*, or to design plans specifically skewing elections toward that preferred outcome, such as favoring or disfavoring a political party, demographic group, or collection of incumbents.

One fundamental technical challenge in the study of redistricting is to contend with the sheer number of possible ways to construct districting plans. State geographies admit enormous numbers of divisions into contiguous districts; even when winnowing down to districting plans that satisfy criteria set forth by legislatures, commissions, or voter referenda, the number remains far too large to enumerate all possible plans in a state. The numbers are easily in the range of googols rather than billions, as we will explain here.

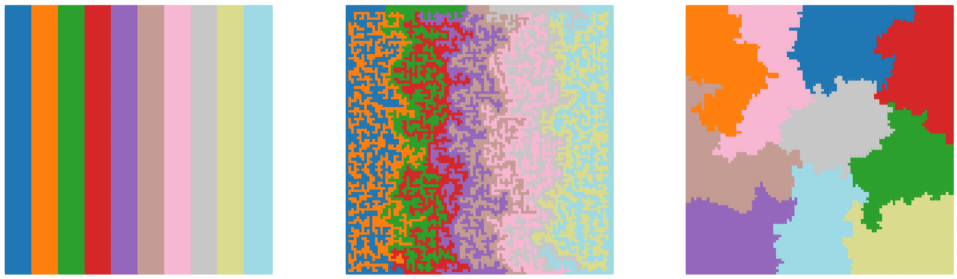
Recent methods for analyzing and comparing districting plans attempt to do so by placing a plan in the context of valid alternatives—that is, those that cut up the same jurisdiction by the same rules and with the structural features of the geography and the pattern of voting held constant. Modern computational techniques can generate large and diverse *ensembles* of comparison plans, even if building the full space of alternatives is out of reach. These ensembles contain *samples* from the full space of plans, aiming to help compare a plan's properties to the range of possible designs. More than this, we need some assurance of *representative sampling*; that is, we need to relate the sampling distribution to the rules and priorities articulated by redistricters.

In one powerful application, ensembles have been used to conduct *outlier analysis*, or to argue that a proposed plan has properties that are extreme relative to the comparison statistics of alternative plans. Arguments like this have featured in a string of recent legal challenges to partisan gerrymanders (Pennsylvania, North Carolina, Michigan, Wisconsin, Ohio), all of which were successful at the district court or state supreme court level. Outliers also received significant attention from the U.S. Supreme Court (culminating in *Rucho v. Common Cause*, 2019), but a 5–4 majority declared that it was too hard for a federal court to decide how much is *too much* of an outlier.

Outside of federal courts, the method is very much alive not only in state-level legal challenges but as a screening step for the initial adoption of plans, and we expect numerous states to employ ensemble analysis this year (2021) when new plans are enacted around the country. These methods can help clarify the influence of each individual state's political geography—the physical geography, demographics, and pattern of voting—as well as the tradeoffs between possible rules and criteria. But the inferences that can be drawn from ensembles rely heavily on the distributions from which the ensembles are sampled.

In the past 5 years, researchers have turned to *Markov chain Monte Carlo*, or MCMC, for sampling. MCMC methods offer strong underlying theory and heuristics, in the form of mixing theorems and convergence diagnostics. The idea is to form a random walk on the state space and collect states visited by the walker to build a sample. The first and most natural approach to defining a random walk on graph partitions is to simply reassign one node at a time through a random process in a Flip process. In a districting context, this amounts to changing the labeling of individual geographic units along district borders. In the standard MCMC paradigm, we would modify the basic Flip step to adjust the ultimate stationary distribution. The traditional ways of doing that are explored here.

In contrast, we define a new family of random walks called *recombination* (or ReCom) Markov chains on the space of partitions, focusing on one variant based on a step that *fuses* two districts and randomly *repartitions* them to form a new plan by cutting an edge of a spanning tree. We argue that spanning tree ReCom has favorable properties that make it well suited to the study of redistricting—in particular, it is tied to a new way of thinking about district 'compactness' that represents a major conceptual and practical improvement on the status quo.



Initial Partition | 1,000,000 Flip steps | 100 ReCom steps

FIGURE 1. Comparison of the basic Flip proposal to the spanning tree ReCom proposal to be described later. Each Markov chain was run from the initial partition of a 100×100 grid into 10 parts (left). Flip chains produce winding, fractal-like districts (center), while ReCom favors 'compact,' or geometrically tame, partitions (right). A typical plan drawn from uniform distribution will have wild boundaries of the type seen in Flip chains, simply because there are far more non-compact than compact partitions.

Applied MCMC relies on convergence heuristics to raise confidence in the findings, along with an explanation of the sampling distribution that is targeted. We present evidence that ReCom converges efficiently to approximately the spanning tree distribution: a way of weighting plans that

comports with traditional districting criteria with little or no parameter tuning by the user. The description of that target distribution is found in Section 5.

In shifting to spanning tree-based sampling methods to overcome the limitations of Flip-based chains, we were led to a new point of view on compactness that has significant independent value. Recombination implements compactness in a soft stochastic fashion, rather than selecting and manually weighting or thresholding a score. We make the case that this spanning tree count favored by recombination amounts to a new kind of compactness, better suited to the needs of redistricting: it draws on latent cluster structure in geographical networks rather than treating redistricting as a Euclidean geometry problem where ideal districts are circles and squares.

1.1. **Contributions.** First and foremost, we intend this article to be readable as an introduction to computational redistricting. We compare Markov chain approaches for benchmarking the behavior of districting plans, with an eye to efficiency and replicability. After some preliminaries, we:

- lay out the practical setup for implementing Markov chains for redistricting (Section 3);
- define the Flip and ReCom random walks on the space of graph partitions (Section 4);
- discuss spanning trees and compactness, describing the distribution targeted by ReCom (Section 5);
- offer experimental comparisons on real and idealized data, reviewing various classical techniques to consider the most promising variants of Flip chains (Section 6); and
- provide a model analysis of racial gerrymandering in the Virginia House of Delegates (Section 7).

To aid reproducibility of our work, open-source implementations of ReCom are available online (Voting Rights Data Institute, 2018, 2020). The first paper to employ spanning tree recombination was a 2018 report on the Virginia House of Delegates case study that was written for reform advocates, legislators, and the general public (DeFord et al., 2018), whose findings we summarize in Section 7. The present authors and our collaborators have applied recombination chains in numerous theoretical and applied projects to date (Angulu et al., 2020; Caldera et al., 2020; DeFord & Duchin, 2019; DeFord et al., 2020; Najt et al., 2019; Weighill & Rodden, 2021), and several of the other computational redistricting research teams have now adopted spanning tree methodology (Autry et al., 2020; Benade & Procaccia, 2020; Carter et al., 2019; McCartan & Imai, 2020).

One step abstracted from redistricting, recombination chains allow sampling from the space of graph partitions for which the pieces are balanced with respect to some function on the nodes (in this case, the districts have nearly equal population). Since balanced graph partitions appear in a large variety of settings, we expect that spanning tree recombination will find diverse applications.

1.2. **Distinctive Features of the Redistricting Problem.** The mathematization of redistricting that we study here is the problem of *sampling from the state space of balanced partitions of a graph into a fixed number of connected subgraphs.* (The graph formulation of redistricting is laid out in Section 3.1.) Although this sounds similar to successful settings for standard Markov chain methods, some essential features of the redistricting problem combine to present great challenges that can cause classical techniques to fail.

Non-uniform sampling. It is crucial to understand that sampling *uniformly* from all valid partitions is not a goal that fits the application to redistricting, nor has uniform sampling ever been credibly attempted in any legal application. (For example, Wesley Pegden's expert work (Pegden, 2017), based on the rigorous theorems from Chikina et al. (2017), bounds the probability that a plan was chosen from the uniform distribution but does not rely on even approximately uniform sampling

to do so. Jonathan Mattingly's expert work (Mattingly, 2017) targets a prescribed nonuniform distribution.)

Although the uniform distribution over graph partitions might seem to be the canonical choice, nonuniform sampling is needed for two fundamental reasons: an ensemble drawn from the uniform distribution would be regarded as prohibitively non-compact in the application domain, and there are in any case obstructions to uniform sampling at the practical scale of redistricting problems. Put simply, there is no hope to accomplish near-uniform sampling on a state-sized problem using a practical algorithm, and even if a uniform sampler could be implemented, it would produce samples with features that make them unusable for redistricting.

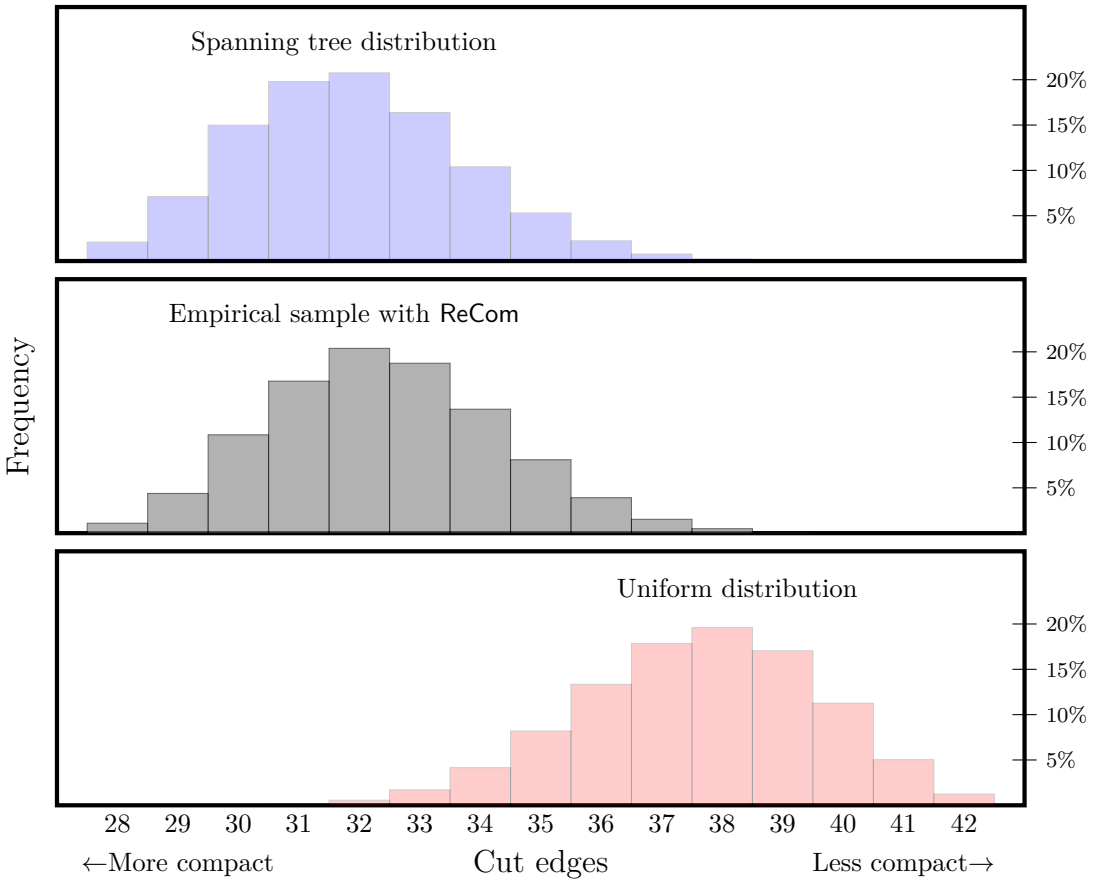


FIGURE 2. Ensemble statistics for plans partitioning the 7×7 grid into 7 districts of 7 units each, based on a complete enumeration of all 158,753,814 valid plans. We use a compactness statistic called *cut edges*, explained further in the text, for the comparison. The uniform distribution is shown in red; the nonuniform (spanning tree) distribution is shown in blue. An empirical sample of 100,000 plans from a recombination chain is shown in gray, confirming success in approximately targeting the spanning tree distribution.

A uniform sample will be dominated by wildly shaped districts (as illustrated in Figure 1 and in Section 5.1). If any reasonable shape score is specified and a threshold is set, then the vast bulk of a uniform sample will consist of plans close to the worst-allowable score, as in Figure 9 and Appendix Figure 14. (For instance, the so-called *Polsby-Popper score* is the most used in redistricting litigation. It is just an isoperimetric ratio—area by perimeter squared—so the pictures make clear that typical uniform samples will have scores near zero.) This makes a uniform ensemble poorly suited to draw usable comparisons. In terms of tractability, it is known that the existence of an efficient uniform sampler, even for planar graphs of bounded degree, would imply RP = NP (Najt et al., 2019). Like most complexity results, this is proved by building stylized graph *gadgets* that do not resemble naturalistic examples, but the experiments here (and the experimental evidence in a range of other papers discussed here) corroborate the slow convergence of Flip walks—and their uniform variants—in practice. Since we cannot (and should not) target the uniform distribution, we must specify an alternative target. When ReCom is run with the bipartition method described later, it approximately targets a closed-form expression called the *spanning tree distribution*, so that plans are weighted according to the number of spanning trees of their districts. Similar spanning tree counts appear widely in the network science literature and are commonly used to pick out clusters or 'communities' in graphs. In Section 5.1, we explain why this favors plans that look plump and compact to the eye.

Limits to analogies from statistical physics. The motivating intuition for a range of MCMC applications comes from statistical physics, where Markov chain methods have been successfully applied for many decades. These physics-style analyses traditionally seek to explore the behavior of so-called Hamiltonian energy functions associated to labelings of lattice nodes with values representing physical quantities. A fundamental example is the *Ising model*, where each node of a lattice is assigned a 'spin' and the associated Hamiltonian is the sum over the edges of ± 1 according to whether the endpoint values agree. In this case, as in many statistical physics models, it is easy to sample uniformly, by assigning spins independently at each node—this is used repeatedly in building the theory. But the preceding discussion should alert us to a likely source of problems: approaches that leverage uniform sampling will transfer poorly to redistricting.

In our setting we require that the pieces with a common label be connected and have a prescribed total weight—the *contiguity* and *balance* constraints of redistricting. These are large-scale properties of each district, which cannot be validated within a local neighborhood of a reassigned node. The fact that so much structure is non-local creates a long-range dependence that is not a common feature of statistical physics problems, and this impedes the effectiveness of Markov chains that act by making local changes to the districting plan like those in the Flip family. And even for computations that can be evaluated locally in the redistricting context, the number of neighboring states can be prohibitively large.

Relatedly, the space of districting plans exhibits a surprising rigidity that is not present in the motivating problems. Techniques that are meant to accelerate Flip chains encounter combinatorial obstructions: The sequence of changes that would be needed to travel between qualitatively distinct plans becomes exponentially unlikely at scale. For instance, MCMC practitioners have had great success with a suite of techniques that work by varying a 'temperature' parameter, alternating between (1) hot: transiting the state space quickly by loosening or setting aside other constraints, and (2) cool: tightening the constraints to improve to states with specified features. But in our

setting (Section 6.3), the high-temperature regime turns out to produce plans that are fractal-shaped and quite rigid. Tame plans (low temperature) are rare and well-separated. This makes the temperature variation techniques less effective than one would expect, and can even cause temperature variation to produce near-loops returning close to their starting position rather than exploring the state space effectively. One view of this phenomenon can be found in Abrishami et al. (2020, figure 12) where MDS plots show that the annealing procedure does not allow the Markov chain to move a large distance through the state space. We will provide a visual example in Figure 10.

1.3. **Review of Computational Approaches to Redistricting.** Computational methods for generating districting plans have been proposed since at least the work of Weaver, Hess, and Nagel in the 1960s (Nagel, 1965; Weaver & Hess, 1963). Like several of the modern approaches, Nagel's algorithm works by incrementally improving districting plans in some metric while taking into account criteria like population balance, compactness, and partisan balance. Many basic elements that are still relevant for modern computational redistricting approaches were already in place in that work. Contiguity is captured using a graph structure or "touchlist" (see our Section 3.1); quantitative criteria are extracted from redistricting rules (see our Section 3.2); a greedy hill-climbing strategy improves plans from an initial configuration; and randomization is used to improve the results. A version of the Flip step (called "the trading part") even appears in Nagel's optimization procedure. Their particular stochastic algorithm made use of hardware available at the time: "[R]un the same set of data cards a few times with the cards arranged in a different random order each time."

Since this initial exploration, computational redistricting has co-evolved with the development of modern algorithms and computing equipment. In the following, we highlight a few incomplete but representative examples; see Altman and McDonald (2010), Cirincione et al. (2000), Ricca et al. (2013), and Tasnádi (2011) for broader surveys; only selected recent work is cited here.

Optimization. Perhaps the most common redistricting approach discussed in the technical literature is the *optimization* of districting plans. Optimization algorithms are designed to extremize objective functions measuring plan properties, while satisfying some set of constraints. Most commonly, algorithms proposed for this task maintain contiguity and population balance of the districts and try to maximize the 'compactness' through some measure of shape (Jin, 2017; Kim, 2011). Many authors have followed Weaver and Hess by using Voronoi or power diagrams with some variant of k-means (Cohen-Addad et al., 2017, 2018; Fryer Jr & Holden, 2011; Levin & Friedler, 2019); there is a lineage of approaches through integer programming with fluid-flow constraints to impose contiguity (Buchanan et al., 2019); and there is even a partial differential equations approach with a volume-preserving curvature flow (Jacobs & Walch, 2018).

Optimization algorithms have not so far become a significant element of reform efforts around redistricting practices, partly because of the difficulty of using them in assessment of proposed plans that take many criteria into account besides those reflected in the objective function. Moreover, most formulations of global optimization problems for full-scale districting plans are likely computationally intractable to solve, as most of the above-listed authors clearly acknowledge.

Assembly. Here, a randomized process is used to create a plan from scratch, and this process is repeated to create a collection of plans that will be used as a basis for comparison. Note that an optimization algorithm with some stochasticity could be run repeatedly as an assembly algorithm, but

generally the goals of assembly algorithms are to produce diversity while the goals of optimization algorithms are to find one or a few best examples.

The most basic assembly technique is to use a greedy *agglomerative* strategy, such as starting from k random choices among the geographical units as the seeds of districts and growing outward by adding neighboring units until the jurisdiction has been filled up and the plan is complete, or combining the units by successive merges until a plan has the required number of districts, then possibly trading units to rebalance. These are colorfully called "Petri dish" methods in Duchin and Spencer (2021). Typically, these algorithms abandon a plan and restart if they reach a dead-end configuration (one that cannot be completed into a valid plan), which can happen often. Examples include Chen and Rodden (2013, 2016), Haas et al. (2020), and Magleby and Mosesson (2018). We are not aware of any theory to characterize the support and qualitative properties of the sampling distributions that result from these procedures.

Random walks. A great deal of mathematical attention has recently focused on random walk approaches to redistricting. These methods use a step-by-step modification procedure to begin with one districting plan and incrementally transform it. Examples include Chikina et al. (2020), Chikina et al. (2017), and Fifield, Higgins, et al. (2020), Herschlag et al. (2020), Herschlag et al. (2017). An evolutionary-style variant with the same basic step can be found in W. Cho and Liu (2016) and Liu et al. (2016). The use of random walks for sampling is well developed across scientific domains in the form of MCMC techniques. This is what the bulk of the present paper will consider.

We emphasize that while many of the techniques used in litigation have been Flip-based, they inevitably involve customizations, such as carefully tuned constraints and weighting, crossover steps, and more. The experiments here are not intended to reproduce the precise setup of any of these implementations (in part because the detailed specifications and code are not always public). Many of the drawbacks, limitations, and subtleties of working with flip chains are well known to practitioners but not yet present in the literature.

Benchmarking sampling techniques is challenging but fundamentally important. For instance, Fifield, Imai, et al. (2020) offer a complete enumeration of partitions in a very small problem, with 70 rather than thousands of units. The logic in that paper is heavily premised on uniform sampling, but our Figure 2 illustrates benchmarking with respect to an alternative target distribution. (Similar reweighting of a test set for benchmark purposes is presented in Carter et al. (2019).) These demonstrations should be read with caution because it is unclear if complete enumerations will ever be possible on a large enough geography for all the relevant phenomena of realistic redistricting problems to become apparent. Nonetheless, having spent a great deal of time attempting to compare different implementations of redistricting samplers, we are convinced that this is valuable, especially because alternative implementations in code will be all but impossible to fully vet. In particular, anyone claiming to use spanning tree ReCom should demonstrate its alignment with the spanning tree distribution on the largest available validation data sets.

2. Markov Chains

A Markov chain is simply a process for moving between positions in a *state space* according to a transition rule under which the probability of arriving at a particular position at time $n+1$ depends only on the position at time n. That is, it is a random walk without memory. A basic but powerful example of a Markov chain is the simple random walk on a graph: from any node, the process chooses a neighboring node uniformly at random for the next step. More generally, one could take

a weighted random walk on a graph, imposing different probabilities on the incident edges. One of the fundamental facts in Markov chain theory is that any Markov chain can be accurately modeled as a (not necessarily simple) random walk on a (possibly directed) graph. Markov chains are used for a huge variety of applications, from Google's PageRank algorithm to speech recognition to modeling phase transitions in physical materials. In particular, MCMC is a class of statistical methods that are used for sampling, with a vast and fast-growing literature and a long track record of modeling success, including in a range of social science applications. See the classic survey by Diaconis (2009) for definitions, an introduction to Markov chain theory, and a lively guide to applications.

The theoretical appeal of Markov chains comes from the convergence guarantees that they provide. The fundamental theorem says that for any ergodic Markov chain there exists a unique stationary distribution, and that iterating the transition step causes any initial state or probability distribution to converge to that steady state. The number of steps that it takes to pass a threshold of closeness to the steady state is called the *mixing time*; in applications, it is extremely rare to be able to rigorously prove a bound on mixing time; instead, scientific authors often appeal to a suite of heuristic convergence tests and diagnostics, as we do here.

This article is devoted to investigating Markov chains for a *global* exploration of the universe of valid redistricting plans. From a mathematical perspective, the gold standard would be to define Markov chains for which we can (1) characterize the stationary distribution π and (2) compute the mixing time. In most scientific applications, the stationary distribution is specified in advance through the choice of an objective function and a Metropolis–Hastings or Gibbs sampler that weights states according to their scores. From a practical perspective in redistricting, confirming mixing to a distribution with a simple closed-form description is neither necessary nor sufficient. Over the last several years, our research group has reoriented to what we view as a domain-specific gold standard: $(1')$ explanation of the distributional design and the weight that it places on particular kinds of districting plans, matched to the law and practice of redistricting, and $(2')$ convergence heuristics and sensitivity analysis that give researchers confidence in the robustness and replicability of their techniques. Though far from completing that research program, the use of spanning trees has opened up many fruitful directions for exploration.

Stronger sampling and convergence theorems are available for *reversible* Markov chains, those for which the steady-state probability of being at state P and transitioning to Q equals the probability of being at Q and transitioning to P for all pairs P, Q from the state space. In particular, a sequence of elegant theorems from the 1980s to now (Besag & Clifford, 1989; Chikina et al., 2020; Chikina et al., 2017) shows that samples from reversible Markov chains admit conclusions about their likelihood of having been drawn from a stationary distribution π long before the sampling distribution approaches π. For redistricting, this theory enables what we might call *local* search: While only sampling a relatively small neighborhood, we can draw conclusions about whether a plan has properties that are typical of random draws from π. Importantly, these techniques can circumvent the mixing and convergence issues, but they must still contend with issues of distributional design and sensitivity to user choice.

In applications, MCMC runs are often carried out with burn time (i.e., discarding the first m steps) and subsampling (collecting every r samples after that to create the ensemble). If r is set to match the mixing time, then the draws will be approximately uncorrelated and the ensemble will be distributed according to the steady-state measure. Experiments in the present article can be interpreted as exploring the choice of a suitable design for a Flip chain—for instance, Appendix

Figures 13 and 15 show that the subsampling parameter would have to be well into the millions to achieve approximate independence for a Flip chain in Virginia.

Though the possibility of pseudo-convergence is always a caveat, the experiments also lend support to the use of ReCom chains with no burn-in or subsampling. (For a discussion of burn time, pseudo-convergence, and the applicability of the Markov chain central limit theorems to the $m = 0, r = 1$ case, see Geyer (2011).)

Some of the performance obstructions described here have led researchers to use extremely fast and/or parallelized implementations, serious computing (or supercomputing) power, and various highly tuned or hybrid techniques that sometimes sacrifice the Markov property entirely or make external replicability impossible. In contrast, on full-scale problems, a ReCom chain with run length in the tens of thousands of steps produces ensembles that pass many tests of quality, both in terms of convergence and in distributional design. Depending on the details of the data, this can be run in a matter of hours on a standard laptop. Indeed, since this article was first drafted, there is a new implementation in the high-performance language Julia that can get to millions of steps within minutes (Voting Rights Data Institute, 2020).

3. Setting Up the Redistricting Problem

Before providing the technical details of Flip and ReCom, we set up the analysis of districting plans as a *discrete* problem and explain how Markov chains can be designed to produce plans that comply with the rules of redistricting.

3.1. Redistricting as a Graph Partition Problem.

The earliest understanding of pathologies that arise in redistricting was largely contour-driven. Starting with the original 'gerrymander,' whose salamander-shaped boundary inspired a famous 1812 political cartoon, irregular district boundaries on a map were understood to be signals that unfair division had taken place. Several contemporary authors now argue for replacing the focus on contour-based compactness with discrete compactness (Duchin & Tenner, 2018), and in practice the vast majority of algorithmic approaches discussed here adopt the discrete model for the problem overall. There are many reasons for this shift in perspective. In practice, a district is an aggregation of a finite number of census blocks (defined by the Census Bureau every 10 years) or precincts (defined by state, county, or local authorities, and aligned to census geography only once every 10 years). District boundaries extremely rarely cut through census blocks and typically preserve precincts,[1] making it reasonable to compare a proposed plan to alternatives built from block or precinct units. Furthermore, these discretizations give ample granularity; for instance, most states have several thousand precincts and several hundred thousand census blocks.

From the discrete perspective, our basic object is the *dual graph* to a geographic partition of the state into units. We build this graph $G = (V, E)$ by designating a vertex for each geographic unit (e.g., block or precinct) and placing edges in E between those units that are geographically adjacent; Figure 3 shows an example of this construction on the counties of Iowa. With this formalism, a districting plan is a partition of the nodes of V into subsets that induce connected components of G. This way, redistricting can be understood as an instance of graph partitioning, a well-studied problem in combinatorics, applied math, and network science (Nascimento & de Carvalho, 2011;

[1]For example, the current Massachusetts plan splits just 1.5% of precincts. But measuring the degree of precinct preservation is very difficult in most states because precincts change frequently and may in some cases be adjusted to match the districts rather than the other way around.

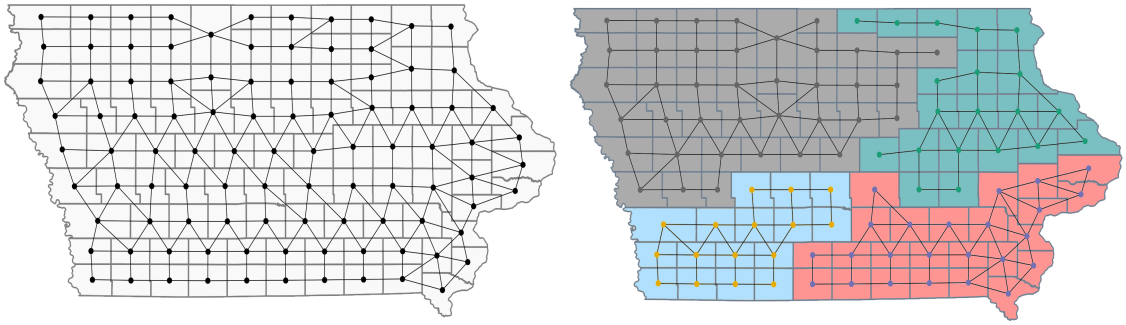


FIGURE 3. Dual graph of Iowa's counties together with the current Iowa congressional districts. Iowa is currently the only state whose congressional districts are made of whole counties.

Schaeffer, 2007). Equivalently, a districting plan is an assignment of each node to one of k districts via a labeling map $V \to \{1, \ldots, k\}$. The nodes (and sometimes the edges) of G are decorated with assorted data, especially the population associated to each vertex, which is crucial for plan validity. Other attributes for vertices may include the assignment of the unit to a municipality or a vector of its demographic statistics. Relevant attributes attached to edges might include the length of the boundary shared between the two adjacent units.

Generating seed plans. To run our Markov chains, we need a valid initial state—or *seed*—in addition to the proposal method that transitions from state to state. Although in some situations we may want to start chains from the currently enacted plan, we will need other seed plans if we wish to demonstrate that our ensembles are adequately independent of starting point. Thus, it is useful to be able to construct starting plans that are at least contiguous and tolerably population balanced. Agglomerative methods (see Section 1.3) or spanning tree methods (see Section 4.3) can be used for generation of seed plans, and both are implemented in our codebase.

3.2. **Sampling from the Space of Valid Plans.** Increasing availability of computational resources has fundamentally changed the analysis and design of districting plans by making it possible to explore the space of valid districting plans much more efficiently and fully. It is now clear that any literal reading of the requirements governing redistricting permits an enormous number of potential plans for each state, far too many to build by hand or to consider systematically. The space of valid plans only grows if we account for the many possible readings of the criteria.

To illustrate this, consider the redistricting rule present in 10 states that dictates that state House districts should nest perfectly inside state Senate districts, either two-to-one (AK, IA, IL, MN, MT, NV, OR, WY) or three-to-one (OH, WI). One tight interpretation of this mandate would be to fix the House districts in advance and admit only those Senate plans that group appropriate numbers of adjacent House districts. If even this narrow interpretation is applied to Minnesota, for example, a perfect matching analysis indicates that there are still 6,156,723,718,225,577,984, or over 6×10^{18}, ways to form valid state Senate plans just by pairing the current House districts (Caldera

et al., 2020). The actual choice left to redistricters, who in reality control House and Senate lines simultaneously, is far more open, and 10^{100} seems to us to be a modest estimate.

Operationalizing the rules. Securing *operational* versions of rules and priorities governing the redistricting process requires a sequence of modeling decisions, with major consequences for the properties of the ensemble. Constitutional and statutory provisions governing redistricting are never precise enough to admit a single unambiguous mathematical interpretation. We briefly survey the operationalization of important redistricting rules:

- **Population balance.** For each district, we can limit its percentage deviation from the ideal size (state population divided by k, the number of districts). The case law around tolerated population deviation is thorny and still evolving (Hebert et al., 2010, ch. 1).[2] Excessively tight requirements for population balance can spike the rejection rate of the Markov chain and impede its efficiency or even disconnect the search space entirely. Even for Congressional districts, which are often balanced to near-perfect equality in enacted plans, a precinct-based ensemble with $\leq 1\%$ deviation can still provide a good comparator, because those plans typically can be quickly tuned by a mapmaker at the block level without breaking their other measurable features.
- **Contiguity.** Most states require district contiguity by law, and it is the standard practice even when not formally required. But even contiguity has subtleties in practice, because of water, corner adjacency, and the presence of pieces that are themselves disconnected. Unfortunately, this means that contiguity must be handled by building and cleaning dual graphs for each state on a case-by-case basis.
- **Compactness.** Many states have a 'compactness' rule in law indicating a loose preference for regular district shapes, but few attempt a definition, and several of the conventional definitions are naive and problematic. There are several standard scores in litigation, especially an isoperimetric score ("Polsby-Popper") and a comparison to the circumscribed circle ("Reock"), each one applied to single districts. It is easy to critique these scores, which are readily seen to be under-defined, unstable, and inconsistent (Bar-Natan et al., 2020; Barnes & Solomon, 2020; DeFord, Lavenant, et al., 2019; Duchin & Tenner, 2018; Zhang et al., 2020). In practice, compactness is almost everywhere ruled by the proverbial eyeball test. We will handle compactness in a mathematically natural manner for a discrete model: we count the number of *cut edges* in a plan, that is, the number of edges in the dual graph whose endpoints belong to different districts (see Section 5). This gives a notion of the discrete perimeter of a plan, and it corresponds well to informal visual standards of regular district shapes (the eyeball test that is used in practice much more heavily than any score). The cut edges count is closely (inversely) correlated to the number of spanning trees of the districts.
- **Splitting rules.** Many states express a preference for districting plans that 'respect' or 'preserve' areas that are larger than the basic units of the plan, such as counties, municipalities, and (often underdefined) *communities of interest*. There is no consensus on best practices for quantifying the relationship of a plan to a sparse set of geographical boundary curves. Simply counting the number of units split (e.g., counties touching more than one district) or employing an entropy-like splitting score are two alternatives that have been used in prior studies (DeFord & Duchin, 2019; Mattingly, 2017). See Duchin and Spencer (2021) for a comparison of several alternatives.

[2]For years, the basis of apportionment has been the raw population count from the decennial census, but there are clear moves to change to a more restrictive population basis, such as by citizenship.

- **Voting Rights Act (VRA).** The Voting Rights Act of 1965 is standing federal law that requires districts to be drawn to provide qualifying minority groups with the opportunity to elect candidates of choice (Hebert et al., 2010, ch. 3-5). Since the VRA legal test involves assessing "the totality of the circumstances," including local histories of discrimination and patterns of racially polarized voting, this is extraordinarily difficult to model in a Markov chain. A new attempt to operationalize the core notion of *effective districts*, built collaboratively with data scientists and a voting rights attorney, can be found in Becker et al. (2021).
- **Neutrality.** Often state rules will dictate that certain considerations should not be taken into account in the redistricting process, such as partisan data or incumbency status. This is easily handled in algorithm design by not recording or inputting associated data, like election results or incumbent addresses.

Finally, most of these criteria are subject to an additional decision:

- **Aggregation and combination.** Many standard metrics used to analyze districting plans (as described above) are computed on a district-by-district basis, without specifying a scheme to aggregate scores across districts that would make plans mutually comparable. If, for instance, we use an L^∞ or sup norm to summarize the compactness scores of the individual districts, then all but the worst district can be altered with no penalty. Choosing L^1 or L^2 aggregation takes all scores into account, but to some extent allows better districts to cover for worse ones. Pegden (2017) has argued for L^{-1} aggregation (i.e., adding reciprocals) to heavily penalize the worst abuses for scores measured on a $[0, 1]$ scale. A modeler with multiple objective functions must also decide whether to combine them into a fused objective function, whether to threshold them at different levels, how to navigate a Pareto front of possible trade-offs, and so on.

Our discussion in Section 6 provides details of how we approach these decisions in our experiments.

4. The Flip and Recombination Chains

4.1. **Notation.** Given a dual graph $G = (V, E)$, a k–partition of G is a collection of disjoint subsets $P = \{V_1, V_2 \ldots, V_k\}$ such that $\bigsqcup V_i = V$. The V_i are thought of as 'districts' and the partition P as a 'districting plan' on the graph G. The full set of k–partitions of G will be denoted $\mathcal{P}_k(G)$.

We may abuse notation by using the same symbol P to denote the labeling function $P : V \to \{1, \ldots, k\}$. That is, $P(u) = i$ means that $u \in V_i$ for the plan P. In a further notational shortcut, we will sometimes write $P(u) = V_i$ to emphasize that the labels index districts. This labeling function allows us to represent the set of cut edges in the plan as $\partial P = \{(u, v) \in E : \ P(u) \neq P(v)\}$. We denote the set of boundary nodes by $\partial_V P = \{u \in e : e \in \partial P\}$. In the dual graphs derived from real-world data, our nodes are weighted with populations or other demographic data, which we represent with functions $w : V \to \mathbb{R}$.

This notation allows us to express constraints on the districts efficiently. For example, contiguity can be enforced by requiring that the induced subgraph on each V_i is connected. The cut edge count used here as a measure of compactness is written $|\partial P|$. A condition that bounds population deviation can be written as

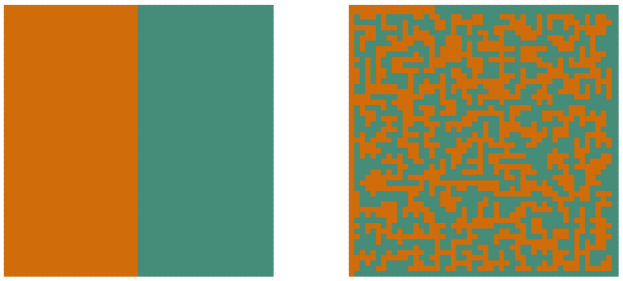
$$(1 - \varepsilon)\frac{\sum_V w(v)}{k} \leq |V_i| \leq (1 + \varepsilon)\frac{\sum_V w(v)}{k}.$$

For a given analysis or experiment, once the constraints have been set and fixed, we will make use of a function $C : \mathcal{P}_k(G) \mapsto \{\texttt{True}, \texttt{False}\}$ to denote the validity check. This avoids cumbersome notation to make explicit all of the individual constraints.

Next, we set out proposal methods for comparison, that is, procedures for transitioning between states of $\mathcal{P}_k(G)$ according to a probabilistic rule. Formally, each X_P is a $[0,1]^{\mathcal{P}_k(G)}$-valued random variable with coordinates summing to one, describing the transition probabilities. Since $\mathcal{P}_k(G)$ is a gigantic but finite state space, the proposal distribution can be viewed as a stochastic matrix with rows and columns indexed by the states P, such that the (P, Q) entry $X_P(Q)$ is the probability of transitioning from P to Q in a single move. The resulting process is a Markov chain: each successive state is drawn according to X_P, where P is the current state. Since these matrices are too large to build, we may prefer to think of the proposal distribution as a stochastic algorithm for modifying the assignment of some subset of V. This latter perspective does not require computing transition probabilities explicitly, but rather leaves them implicit in the algorithm.

In this section, we introduce the main Flip and ReCom proposals analyzed in the paper and describe some of their qualitative properties, with particular attention to the spanning tree method.

(A) Sequence of four flip steps



(B) Before and after 500,000 flip steps

FIGURE 4. Flip steps. A single node on the boundary changes assignment at each move, preserving contiguity. This is illustrated schematically on a 5×4 grid and then the end state of a long run is depicted on a 50×50 grid.

4.2. **Flip Proposals.** At its simplest, a Flip proposal changes the assignment of a single node at each step in the chain in a manner that preserves the contiguity of the plan. See Figure 4 for a sequence of steps in this type of Markov chain and a randomly generated 2–partition of a 50×50 grid, representative of the types of partitions generated by Flip and its variants. This procedure provides a convenient vehicle for exploring the complexity of the partition-sampling problem.

To implement Flip, we must decide how to select a node whose assignment will change, for which we define an intermediate process called Node Choice. To ensure contiguity, it is intuitive to begin by choosing a vertex of $\partial_V P$ or an edge of ∂P, but because degrees vary, this can introduce nonuniformity to the process. To construct a *reversible* flip chain we follow Chikina et al. (2017) and instead sample uniformly from the set of (node, district) pairs (u, V_i) where $u \in \partial_V P$ and there exists a cut edge $(u, v) \in \partial P$ with $P(v) = V_i$. This procedure amounts to making a uniform choice among the partitions that differ only by the assignment of a single boundary node. Pseudocode for this method is presented in Algorithm 1. The associated Markov chain has transition probabilities given by

$$X_P(Q) = \begin{cases} \frac{1}{|\{(v,P(w)):(v,w)\in\partial P\}|} & |\{P(u) \neq Q(u) : u \in \partial P\}| = 1 \text{ and } |\{P(u) \neq Q(u) : u \notin \partial P\}| = 0; \\ 0 & \text{otherwise.} \end{cases}$$

This can be interpreted as a simple random walk on $\mathcal{P}_k(G)$ where two partitions are connected if they differ at a single boundary node. Thus, the Markov chain is reversible. Its stationary distribution is nonuniform, since each plan is weighted proportionally to the number of (node, district) pairs in its boundary. Evaluating this steady state is further complicated by the fact that each of these potential neighbors may fail constraint checks governed by θ.

Algorithm 1: `Node Choice`

Input: Dual graph $G = (V, E)$ and current partition P

Output: A new partition Q

Select: A (node, district) pair (u, V_i) uniformly from $\{(v, P(w)) : (v, w) \in \partial P\}$

Define: $Q(v) = \begin{cases} V_i & \text{if } u = v \\ P(v) & \text{otherwise.} \end{cases}$

Return: Q

Algorithm 2: `Flip`

Input: Dual graph $G = (V, E)$ and the current partition P

Output: A new partition Q

Initialize: *Allowed* = `False`

while *Allowed* = `False` **do**

  Q = Node Choice(G, P)

  Allowed $= C(Q)$

end

Return: Q

At each step, the Node Choice algorithm grows one district by a node and shrinks another. One can quickly verify that a Node Choice step maintains contiguity in the district that grows but may break contiguity in the district that shrinks. In fact, after many steps it is likely to produce a plan with no contiguous districts at all. To address this, we adopt a rejection sampling approach, only accepting contiguous proposals. This produces our basic Flip chain (see Algorithm 2 for pseudocode and Figures 1 and 4 for visuals). The rejection setup does not break reversibility of the associated Markov chain, since it now amounts to a simple random walk on the restricted state space.

Rejection sampling is practical because it is far more efficient to evaluate whether or not a particular plan is permissible than to determine the full set of adjacent plans at each step. Both the size of the state space and the relatively expensive computations that are required at the scale of real-world dual graphs contribute to this issue. If the proposal fails contiguity or another constraint check, we simply generate new proposed plans from the previous state until one passes the check.

These methods have the advantage of explainability in court and step-by-step efficiency for computational purposes, since each new proposed plan is only a small perturbation of the previous one. The same property that allows this apparent computational advantage, however, also makes it

difficult for Flip-type proposals to explore the space of permissible plans efficiently. Figure 1 shows that after 1 million steps the structure of the initial state is still clearly visible, and we will present evidence that one billion steps is enough to improve matters significantly, but not to the point of approximate convergence of the ensemble. Thus, the actual computational advantage is less clear, as it may take a substantially larger number of steps of the chain to provide reliable samples. This issue is exacerbated when legal criteria impose strict constraints on the space of plans, which may easily cause disconnectedness of the state space under this proposal. A user can choose to ensure connectivity by relaxing even hard legal constraints during the run and winnowing to a valid sample later, which requires additional choices and tuning.

Researchers have attempted to address this slow-mixing issue in practice, including using simulated annealing or parallel tempering in Fifield, Higgins, et al. (2020) and Herschlag et al. (2020), Herschlag et al. (2017) and a Swendsen-Wang variant in Fifield, Higgins, et al. (2020) that changes the assignments of several nodes at a time. However, we will show in Section 6 that on the scale of real-world problems, these fixes are not immediately sufficient to overcome the fundamental barrier to successful sampling that is caused by the combination of extremely slow mixing and the domination of distended shapes.

4.3. **ReCom Proposals.** The performance obstructions for the Flip chain motivate the move to a new Markov chain on partitions, which changes the assignment of many vertices at once while preserving contiguity. Our new proposal is more computationally costly than Flip at each step in the Markov chain, but this tradeoff is net favorable thanks to superior convergence and distributional design.

In maximum generality, each step of these chains will merge some ℓ out of k districts and then repartition them into new connected pieces. We call this procedure *recombination* (ReCom), motivated by the biological metaphor of recombining genetic information. This general procedure is summarized in Algorithm 6 in the Appendix.

The focus of the present paper is recombination with a spanning tree method of bipartitioning, which we now describe. Recall that a *spanning tree* of a graph is a connected subgraph containing all n vertices but only $n-1$ of the edges, so that there are no cycles in the subgraph. This form of recombination fuses two adjacent districts (i.e., $\ell = 2$), draws a spanning tree of the merged subgraph, and cuts it to form two new districts.

Figure 5 shows a schematic of a single step with this proposal; the middle image shows a spanning tree of the 5×4 grid.

- First, draw a spanning tree uniformly at random from among all of the spanning trees of the merged region. The implementation used in the experiments here employs the loop-erased random walk method of Wilson's algorithm (Wilson, 1996). Wilson's algorithm is notable in that it samples uniformly from all possible spanning trees in polynomial time. The GerryChain package Voting Rights Data Institute (2018) also has an implementation of minimal spanning trees with randomized weights, which is faster and qualitatively similar.
- Next, seek an edge to cut from the spanning tree so that the complementary components have population balance within the permitted tolerance. For an arbitrary spanning tree, it is not always possible to find such an edge, in which case we draw a new tree; this is another rejection step in our implementation. In practice, the rejection rate is low enough that this chain runs efficiently. If there are multiple edges that could be cut to generate partitions with the desired tolerance, we sample uniformly from among them.

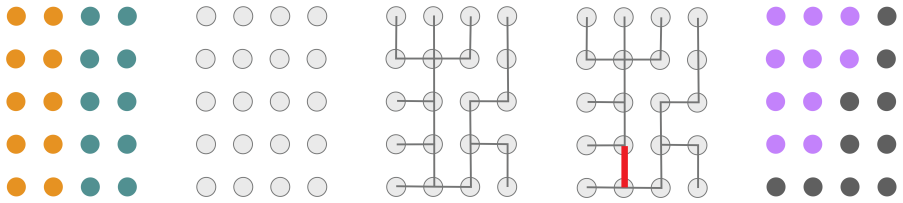


FIGURE 5. A schematic of a ReCom spanning tree step for a small grid with $k = 2$ districts that are merged ($\ell = 2$) and resplit. Deleting the indicated edge from the spanning tree leaves two connected components with an equal number of nodes.

Pseudocode for this technique is provided in Algorithm 3.

Algorithm 3: `ReCom` (Spanning tree bipartitioning)

Input: Dual graph $G = (V, E)$, the current partition P, population tolerance ε
Output: The next partition Q

Select: $(u, v) \in \partial P$ uniformly
Set $W_1 = P(u)$ and $W_2 = P(v)$
Form the induced subgraph H of G on the nodes of $W_1 \cup W_2$.
Initialize: Cuttable = `False`
while *Cuttable = False* **do**
  Sample a spanning tree T of H
  Let EdgeList = []
  for *edge in T* **do**
    Let $T_1, T_2 = T \setminus edge$
    if $|T_1| - |T_2| < \varepsilon |T|$ **then**
      Add edge to EdgeList
      Cuttable = True
    end
  end
end
Select cut uniformly from EdgeList
Let $R = T \setminus cut$
Define $Q(v) = \begin{cases} R(v) & v \in H \\ P(v) & \text{otherwise} \end{cases}$
Return: Q

A similar spanning tree approach to creating initial seeds is available: draw a spanning tree for the entire graph G, then recursively seek edges to cut that leave one complementary component of appropriate population for a district.

5. Distributional Design

In Section 6, we will conduct experiments that demonstrate the behavior of the chains. First, we discuss what is known about their target distributions, and about prospects for approximate sampling from the target in a reasonable time.

5.1. **Spanning Trees and Compactness.** As we have discussed already, 'compactness' is a vague but important term of art in redistricting: compact districts are those with tamer or plumper shapes. This can refer to having high area relative to perimeter, shorter boundary length, fewer spikes or necks or tentacles, and so on. In the experimental treatment in the current paper, we focus on the discrete perimeter as a way to measure compactness, and we refer the reader to Duchin and Tenner (2018) for a deeper discussion of how this fits with the literature in law, political science, and geography.

Recall from Section 4.1 that for a plan P that partitions a graph $G = (V, E)$, we denote by $\partial P \subset E$ its set of cut edges, or the edges of G whose endpoints are in different districts of P. A slight variant is to count the number of *boundary nodes* $\partial_V P \subset V$ (those nodes at the endpoint of some cut edge, representing geographic units on the frontier of a district). There is a great deal of mathematical literature connected to concepts of combinatorial perimeter, from the Min Cut problem to the Cheeger constant to expander graphs. Although we focus on the discrete compactness scores here, a dizzying array of compactness metrics has been proposed in connection to redistricting, and the analysis here—that Flip must contend with serious compactness problems—would apply to any reasonable score, as the visuals throughout this article illustrate.

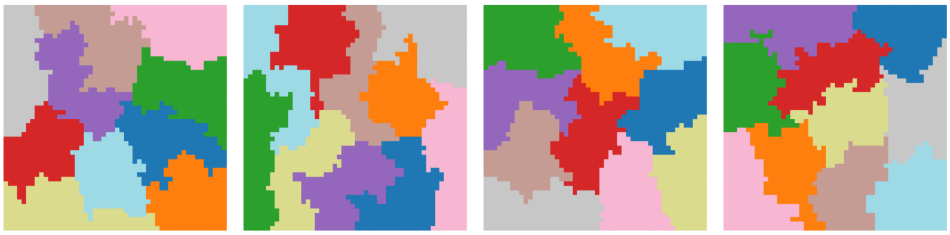


Figure 6. ReCom and compactness. The recombination proposal tends to produce compact districts. Each of these plans was selected after only 100 ReCom steps starting from the same vertical-stripes partition. Unlike the Flip samples, these partitions have relatively short boundaries in addition to displaying low correlation with the initial state. The picture is suggestive that the performance of ReCom is more keyed to the number of districts than the number of constituent units.

The reason that the uniform distribution is so dominated by non-compact districts is a simple matter of counting: there are far more chaotic than regular partitions. As an illustration, consider bipartitioning an $n \times n$ square grid into pieces of nearly the same number of nodes. If the budget of edges you are allowed to cut is roughly n, there are polynomially many ways to bipartition, but the number grows exponentially as you relax the limitation on the boundary size. This exponential

growth also explains why the imposition of any strict limit on boundary length will leave almost everything at or near the limit.

As an alternative, let us define $\mathsf{sp}(G)$ to be the number of spanning trees of the graph G, and $\mathsf{sp}(P) = \prod_i \mathsf{sp}(V_i)$ to be the product of the number of spanning trees of the parts of a partition. We will aim to sample from the distribution in which the probability of selecting partition P is proportional to $\mathsf{sp}(P)$. We first explain why this is desirable and then how ReCom is tailored to this goal.

Why spanning tree weighting is desirable. First, the mathematics: Kirchhoff's matrix-tree theorem tells us that the precise number of spanning trees of any graph G on N nodes is $\mathsf{sp}(G) = \det(\Delta')$, where Δ' is any $(N-1)\times(N-1)$ minor of the combinatorial Laplacian Δ of G. Equivalently, $\mathsf{sp}(G)$ is $\frac{1}{N}$ times the product of the nonzero eigenvalues of the Laplacian. For instance, for an $n \times n$ grid, the number of spanning trees is asymptotic to $C^{n^2} = C^N$, where C is a constant whose value is roughly 3.21 (Temperley, 1972). There are deep theorems suggesting that squares are optimal—i.e., $n \times n$ subgraphs of grids have more spanning trees than any other subgraphs with n^2 nodes (Chinta et al., 2010; Kenyon, 2000). For more discussion, see the open questions in Section 8.1. This means that if a block-like district is altered by creating a simple 'neck' or 'tentacle' with just two or three nodes, it will reduce the number of possible spanning trees by roughly a factor of C^2 or C^3, making the district 10 or 30 times less likely to appear when selection is proportional to $\mathsf{sp}(P)$.

The long snaky districts that are observed in the Flip ensembles are nearly trees themselves, and therefore have a dramatically lower $\mathsf{sp}(P)$ because they admit far fewer spanning trees than their plumper cousins. For example, the initial partition of the 50×50 grid in Figure 4 has a spanning tree score of roughly 10^{1210} while the final partition scores roughly 10^{282}. That means that the tame partition is over 10^{900} times preferred by spanning-tree weighting, while the uniform distribution weights them exactly the same. This explains why districts with a greater number of spanning trees are more compact to the eye, assuming that the building-block units are roughly comparable.

For another point of view on the spanning tree distribution, consider the vast literature on clustering and so-called 'community structure' in graphs and networks. Combinatorial methods are very frequently used to identify clusters with high internal connectivity and a low number of exterior connections; for instance, spanning tree methods to find clusters are exposited in Kleinberg and Tardos (2006, Section 4.7). From this point of view, ReCom is an efficient way to produce diverse examples of balanced partitions that draw out the latent cluster structure in a geographical network.

Redistricting is performed by assigning units of census geography to districts, so the analysis of the district shapes will succeed better and more naturally if it takes that discrete structure into account while still conforming to the eyeball test. Spanning tree weighting favors districts that are well clustered and well separated, and it has the added bonus of admitting fast algorithms. Crucially, it operates without user-defined thresholds, which is an excellent fit for legal applications, since the law itself prefers ranges to thresholds.

Why ReCom targets the spanning tree distribution. For a partition P to be proposed in a ReCom chain, we must have selected a spanning tree of G that restricts to each district as a spanning tree of that district. This means that the probability of selecting a partition P will be roughly proportional to $\mathsf{sp}(P)$. (The idea that one can cut spanning trees to create partitions, and that the resulting distribution will have factors proportional to the number of trees in a block, is a very natural one and appears for instance in this ArXiv note.) We can think of this relationship between

spanning trees of G and districting plans as projection by the edge deletion map that sends trees to partitions.

To illustrate, consider the $k = 2$ case. The number of ways for a recombination step to produce a bipartition of a graph G into subgraphs H_1 and H_2 is the number of spanning trees of H_1 times the number of spanning trees of H_2 times the number of edges between H_1 and H_2 that exist in G.

New work of Cannon et al. (Cannon et al., 2020), building on the research in the present paper, introduces a variant of ReCom that is proven to have precisely this stationary distribution. By slightly modifying the method of selecting district pairs and adding a correction term to the acceptance probability, the authors create a modified chain whose steady state is proportional to $\mathsf{sp}(P)$ and establish detailed balance, which means that the chain is *reversible*. Long runs with reversible ReCom show that its convergence speed is significantly slower (both in terms of step count and runtime) than the unweighted version described in the present paper, but that it obtains very similar summary statistics on grids and on real data at scale.

Figure 2 shows that ReCom succeeds at approximating the exact spanning tree distribution on a small grid with respect to the cut edges count. This explicit comparison is not possible on a full-scale problem because of the lack of a complete enumeration of plans, but we note that the ability of ReCom to target the sp distribution will likely get *better* on large problems because the discrepancy is driven by boundary effects between districts, and those effects become relatively smaller as the districts grow large.

5.2. **Complexity and Efficiency.** We turn to tractability considerations next. In the study of computational complexity, P $\subseteq$ RP $\subseteq$ NP are complexity classes (polynomial time, randomized polynomial time, and nondeterministic polynomial time); it is widely believed that P = RP, and RP $\neq$ NP. Recent theoretical work of DeFord–Najt–Solomon (Najt et al., 2019) shows that flip and uniform flip procedures mix exponentially slowly on several families of graphs, including planar triangulations of bounded degree. That paper also shows that even approximately uniform sampling is intractable, in the sense that an efficient solution would imply RP = NP. Thus, methods that target the uniform distribution may face complexity obstructions, particularly with respect to worst-case scenarios. This should trigger increased scrutiny of the quality of sampling. We remark that in Fifield, Imai, et al. (2020), the authors attempt to approximate uniform sampling by reweighting a Gibbs sample with normalizing coefficients that are imposed after the sample is collected. For this to succeed, the Gibbs chain would need to be run for long enough to accept significant numbers of exponentially unlikely proposals into the ensemble. Because the sample size needed would to obtain good estimates would therefore explode, this scheme does not circumvent the complexity obstructions to uniform sampling.

Our experiments in Section 6 highlight some of these challenges in a practical setting by showing that Flip ensembles continue to give unstable results—with respect to starting point, run length, and summary statistics—at lengths in the many millions. Practitioners must opt for fast implementations and very large subsampling time; even then, the Flip approach requires dozens of tuning decisions, which undermines any sense in which the associated stationary distribution is canonical.

The second major design feature of recombination, alongside its natural relationship to compactness, is its efficiency. The ReCom chain is designed so that each step completely destroys the boundary between two districts, in the sense that the previous pairwise boundary has no impact on the next step. As there are at most $\binom{k}{2}$ boundaries in a given k-partition, this observation suggests that we can lose *most* memory of our starting point in a number of steps that is polynomial in

k and does not depend on n at all. The Markov chain literature has examples of processes on grids with constant scaling behavior, such as the Square Lattice Shuffle (Håstad, 2006). That chain has arrangements of n^2 different objects in an $n \times n$ grid as its set of states; a move consists of randomly permuting the elements of each row, then of each column—or just one of those, then transposing. Its mixing time is constant, that is, independent of n. Chains with logarithmic mixing time are common in statistical mechanics: a typical fast-mixing model, like the discrete hard-core model at high temperature, mixes in time $n \log n$ with local moves (because it essentially reduces to the classic coupon collector problem), but just $\log n$ with global moves. See Section 8.1 for more discussion of research questions in this direction for exploring recombination.

Our experiments here suggest that the time needed for effective sampling has moderate growth in the problem size: tens of thousands of recombination steps give stable results on practical-scale problems whether we work with the roughly 9,000 precincts of Pennsylvania or the roughly 100,000 census blocks in our Virginia experiments. We turn to these experiments now.

6. Experimental Comparison

In this section, we run experiments on the standard toy examples for graph problems, including $n \times n$ grids as well as empirical dual graphs generated from census data. All of our experiments were carried out using the GerryChain Python package (Voting Rights Data Institute, 2018), with additional source code available for inspection (DeFord, Duchin, & Solomon, 2019). The state geographic and demographic data was obtained from the Census TIGER/Line geography program accessed through the National Historical Geographic Information System (NHGIS) (Manson et al., 2018). The real-world graphs can be large, but they share key properties with lattices—they tend to admit planar embeddings, and most faces are triangles or squares. Figure 7 shows the state of Missouri at four different levels of census geography, providing good examples of the characteristic structures we see in our applications.

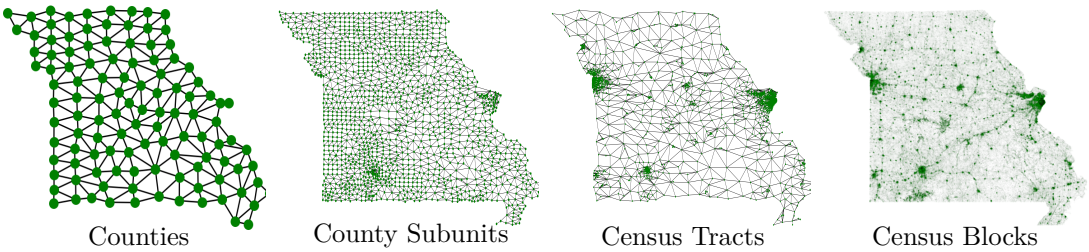
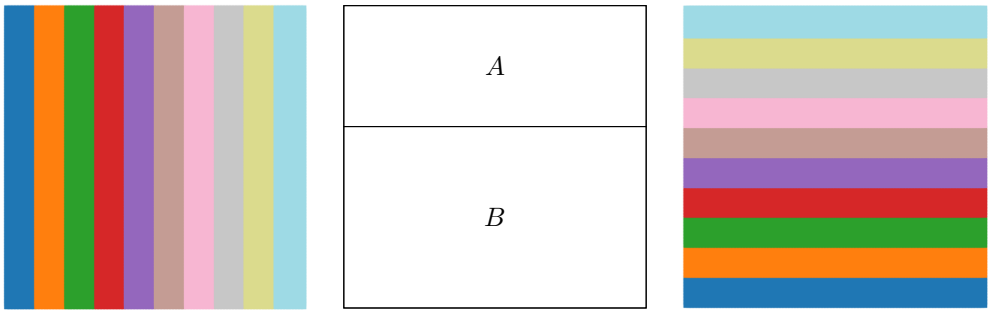
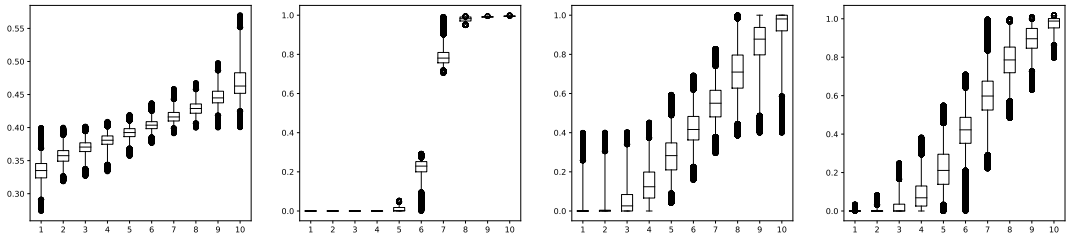


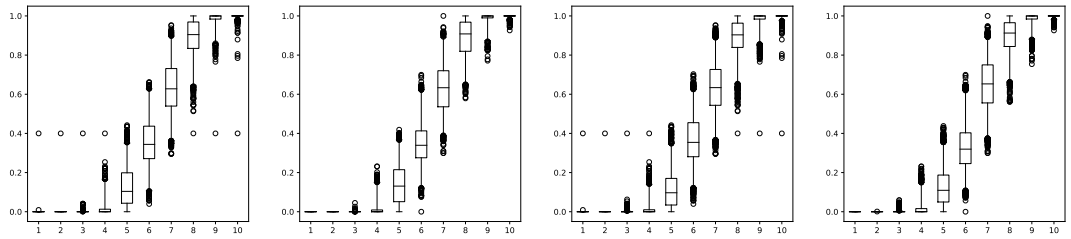
Figure 7. Four dual graphs for Missouri at different levels of geography in the census hierarchy.

6.1. **Projection to Summary Statistics.** The space of districting plans is wildly complicated and high-dimensional. For the redistricting application, we seek to understand the measurable properties of plans that have political or legal relevance, such as their partisan and racial statistics; this amounts to projection to a much lower-dimensional space. In the language of Geyer (2011), these push-forward statistics are called *functionals*.





Flip from vertical 1 million steps | Flip from horizontal 1 million steps | Flip from vertical 1 billion steps | Flip from horizontal 1 billion steps



ReCom from vertical 10,000 steps | ReCom from horizontal 10,000 steps | ReCom from vertical 20,000 steps | ReCom from horizontal 20,000 steps

	Flip (1M)		Flip (1B)		ReCom (10K)		ReCom (20K)	
# A Seats	V seed	H seed	V seed	H seed	V seed	H seed	V seed	H seed
0	878,400	0	1,430,511	0	1	0	1	0
1	121,600	0	13,223,704	0	2	0	2	0
2	0	0	38,333,268	61,711	1	0	1	0
3	0	0	262,315,135	183,597,693	1,656	1,626	2,751	2,978
4	0	1,000,000	480,049,699	605,371,790	7,022	7,364	14,462	15,309
5	0	0	197,367,357	208,772,091	1,318	1,010	2,783	1,713
≥ 6	0	0	7,280,326	2,196,715	0	0	0	0

FIGURE 8. Boxplots and summary statistics for a synthetic election on a 100×100 grid with $k = 10$ districts, comparing Flip runs to ReCom runs from two different starting positions. The boxplots show the proportion of A votes by district, where the districts are ordered from smallest A share to largest A share. Though this multistart heuristic does not rigorously guarantee the convergence of ReCom, we can be certain that one billion steps is not enough for Flip.

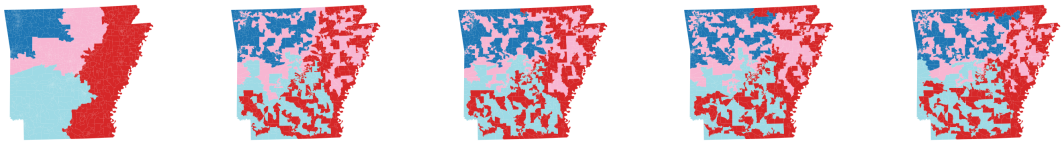
Many of the metrics of interest on districting plans are formed by summing some value at each node of each district. For example, the winner of an election is determined by summing the votes for each party in each geographic unit that is assigned to a given district, and so 'Democratic seats won' is a summary statistic that is real- (in fact integer-) valued. It is plausible that chains that mix slowly in the space of partitions will converge much more quickly in their projection to some summary statistics.

To investigate this possibility, we begin with a toy example with synthetic vote data on a grid, comparing the behavior of the Flip and ReCom proposals (Figure 8). For each Markov chain, we evaluate statistics using a vote distribution on a 100×100 grid—that is, each node is assigned a voting outcome. In this example, each node is assigned to vote for a single party and the votes for Party A are placed in the top 40 rows of the grid. We use two initial districting plans: the familiar vertical-stripes partition and the counterpart horizontal-stripes partition. We collect every state visited by each Markov chain into an ensemble; as we extend the chain, the sample statistics over that ensemble will converge to the push-forward of the stationary distribution, irrespective of starting point. This appeals generally to the family of results called Markov Chain Central Limit Theorems. See Geyer (2011, Section 1.8) for an introduction and some literature pointers. In the figure, the boxes show the 25th–75th percentile statistics over the ensemble and the whiskers span from 1st to 99th. The table records the number of districts with an A majority for each plan; if A were a political party, an A majority in three districts would mean that the party won 3 seats out of 10.

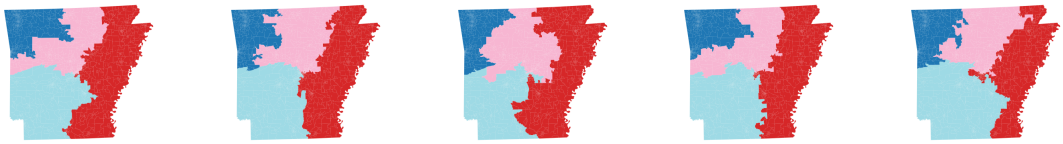
Our results confirm that in this example the Flip chain is unable to produce diverse election outcomes from either starting point after 1,000,000 steps; the Flip ensemble primarily reported one seat outcome in each scenario, giving four seats in the first setup and zero seats in the second. Matters have changed after 1,000,000,000 steps, where the ensemble seeded at the vertical partition has diffused to many possible seat outcomes, but still does not match the summary statistics gathered from the corresponding horizontal-seeded run. The ReCom ensemble nearly exclusively records outcomes of three, four, or five seats, and the histograms from the two seeds are in qualitative agreement after only 10,000 steps. The corresponding boxplots show a more detailed version of this story, highlighting the ways in which each ensemble captures the spatial distribution of voters. The recombination walk takes just a few steps to forget its initial position and then returns consonant answers from the two initial positions. We note that this Flip ensemble is far from convergence after a billion steps, so the evidence here does not offer a conclusive comparison of its stationary distribution to that of ReCom, though it suggests a marked difference.

6.2. **Imposing Constraints.** We begin by noting that the tendency of Flip chains to draw non-compact plans is not limited to grid graphs but occurs on geographic dual graphs just as clearly. The first run in Figure 9 shows that Arkansas's block groups admit the same behavior, with upwards of 90% of nodes on the boundary of a district, and roughly 45% of edges cut, for essentially the entire length of the run. The initial plan has under 20% boundary nodes, and around 5% of edges cut; the basic recombination chain (Run 3) stays well within range of those statistics.

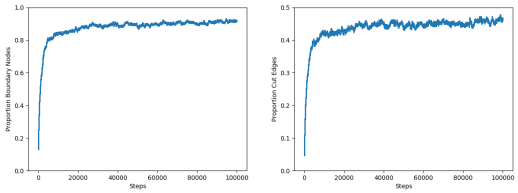
Using thresholds or constraints to ensure that the Flip proposals remain reasonably criteria-compliant requires a major tradeoff. While this enforces validity, it is difficult for Flip Markov chains to generate substantively different partitions under tight constraints. Instead the chain can flip the same set of boundary nodes back and forth and remain in a small neighborhood around the initial plan. See the second run in Figure 9 for an example. Sometimes, this is because an overly



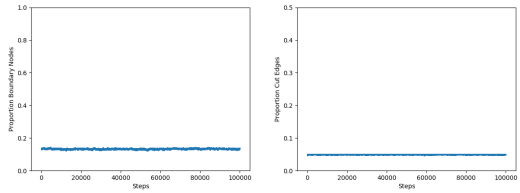
Run 1: 100K Flip steps, shown every 25K, no compactness constraint



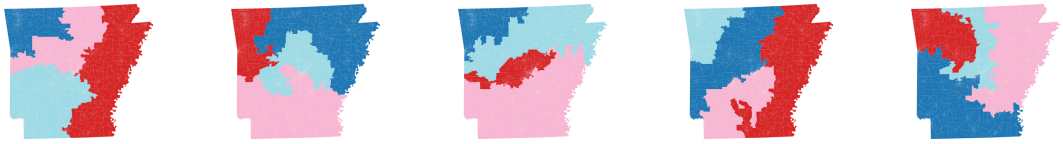
Run 2: 100K Flip steps, shown every 25K, limited to 5% total cut edges



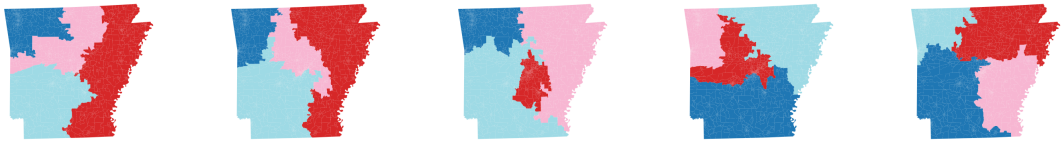
Run 1 boundary statistics



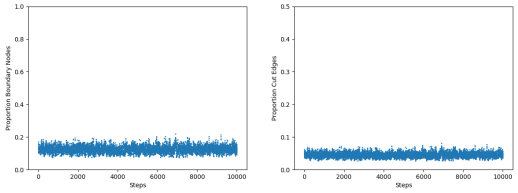
Run 2 boundary statistics



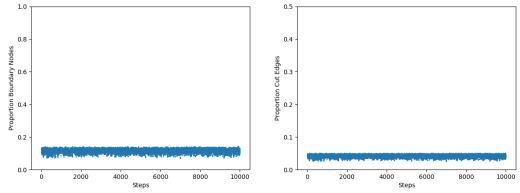
Run 3: 10K ReCom steps, shown every 2500, no compactness constraint



Run 4: 10K ReCom steps, shown every 2500, limited to 5% total cut edges



Run 3 boundary statistics



Run 4 boundary statistics

FIGURE 9. Arkansas block groups partitioned into $k = 4$ districts, with population deviation limited to 5% from ideal. Imposing a compactness constraint makes the Flip chain unable to move very far.

tight constraint disconnects the state space entirely and leaves the chain exploring a small connected component. (An example of this behavior was presented in W. K. T. Cho and Rubinstein-Salzedo (2019, figure 2), though its significance was misinterpreted by the authors with respect to the test in Chikina et al. (2017).) Recombination often responds better to sharp constraints, because large changes at each step mean that ReCom chains do not tend to run at the limit values when constrained. Still, the interactions between various choices of constraints and priorities are so far vastly underexplored. In Section 6.3, we will consider the use of preferential acceptance functions rather than hard rejection constraints on the chains.

6.3. **Temperature Variation.** As we have shown, the Flip proposal tends to create districts with extremely long boundaries, which does not produce a comparison ensemble that is practical for our application. To overcome this issue, we could attempt to modify the proposal to favor districting plans with shorter boundaries. As noted already, this is often done with a standard technique in MCMC called the Metropolis–Hastings algorithm: fix a compactness score, such as a notion of boundary length $|\partial P|$, prescribe a distribution proportional to $x^{|\partial P|}$ on the state space, and use the Metropolis–Hastings rule to preferentially accept more compact plans. Even if we are unable to achieve a sample that approximates this distribution, it could be the case that targeting short boundaries with an accelerated Flip strategy generates a suitably diverse ensemble in reasonable time for our applications.

The Flip distribution was already slow to mix, and Metropolis–Hastings adds an additional score computation and accept/reject decision at every step to determine whether to keep a sample; this typically implies that this variant runs more slowly than the unweighted proposal distribution. To aid in getting reliable results from slow-mixing systems, it is common practice to employ another technique from the statistical physics literature called *simulated annealing*, which iteratively tightens the prescribed distribution toward the desired target—effectively taking larger and wilder steps initially to promote randomness, then becoming gradually more restrictive.

To test the properties of a simulated annealing run based on a Metropolis-style weighting, we run chains to partition Tennessee and Kentucky block groups into nine and six Congressional districts, respectively. We run the Flip walk for 500,000 steps beginning at a random seed drawn by the recursive tree method. The first 100,000 steps use an unmodified Flip proposal; Figure 10 shows that after this many steps, the perimeter statistics are comparable to the Arkansas outputs shown here, with over 90% boundary nodes and nearly 50% cut edges. This initial phase is equivalent to using an acceptance function proportional to $2^{\beta|\partial P|}$ with $\beta = 0$. The remainder of the chain linearly interpolates β from 0 to 3 along the steps of the run.

Figure 10 shows how these Tennessee and Kentucky chains evolved. Ultimately, there is a relatively small difference between the initial and final states in both examples: the simulated annealing has caused the random walk to return to very near its start point. This is due to the properties of the Flip proposal. The districts grow tendrils into each other, but the boundary segments rarely change assignment. Thus, when the annealing forces the tendrils to retract, they collapse near the original districts, and this modified Flip walk fails to move effectively through the space of partitions. These examples do not imply that no annealing-based method can work in practice, but rather that care must be taken to verify that a diverse collection of plans is being created, as the choice of energy function, length of annealing schedule, and choice of state space constraints can all have a major impact on the success and effectiveness of this sampling approach.

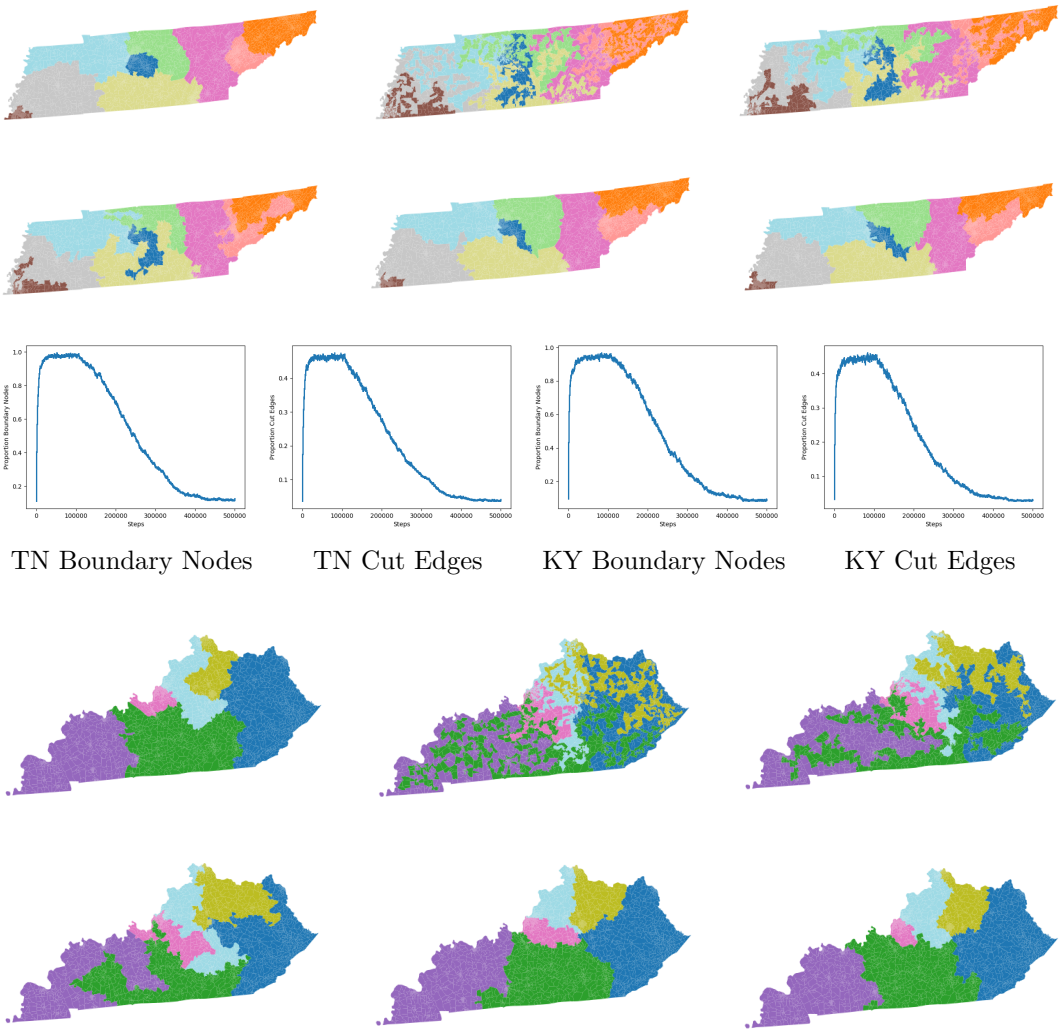


FIGURE 10. Snapshots of the TN and KY annealing ensembles after each 100,000 steps. Comparing the starting and ending states shows only slight changes to the plans as a result of the boundary segments mostly remaining fixed throughout the chain.

Other ensemble generation approaches such as Fifield, Higgins, et al. (2020) use *parallel tempering* (also known as *replica exchange*), a related technique in MCMC also aimed at accelerating its dynamics. In this algorithm, chains are run in parallel from different start points at different temperatures, then the states are occasionally exchanged between temperatures. Exactly the issues highlighted already apply to the individual chains in a parallel tempering run, so this strategy may also struggle to introduce meaningful new diversity.

These experiments suggest that the tendency of Flip chains to produce fractal-esque shapes is extremely difficult to remediate and that direct attempts to do so end up impeding the progress of the chain through the state space. On moderate-sized problems, this can conceivably be countered with careful tuning and extremely long runs. By contrast, ReCom generates plans with relatively few cut edges (usually comparable to human-made plans) by default, and our experiments indicate that its samples are approximately uncorrelated after far fewer steps of the chain—hundreds rather than billions. Weighted variants of ReCom can then be tailored to meet other principles by modifying the proposal or the acceptance probabilities to favor higher or lower compactness scores, or to favor the preservation of larger units like counties and communities of interest. With the use of constraints and weights, one can effectively use ReCom to impose and compare operational versions of the redistricting rules and priorities described in Section 3.2 (Becker et al., 2021; Caldera et al., 2020; DeFord & Duchin, 2019; DeFord et al., 2018; Duchin & Spencer, 2021).

7. Case Study: Virginia House of Delegates

Finally, we offer a brisk demonstration of what a high-quality comparator ensemble can do in a redistricting problem of current legal interest. We look at the Virginia House of Delegates plans that were recently debated in the *Bethune-Hill* cases that went to the Supreme Court in 2017 and 2019. We include a brief discussion here, with supporting materials in Appendix B. For full details, see DeFord et al. (2018).

The districting plan for Virginia's 100-member House of Delegates was commissioned and enacted by its Republican legislative caucus in 2011, following the 2010 census. That plan was challenged in complicated litigation that went before multiple federal courts, with the ultimate finding that the plan was an unconstitutional racial gerrymander. The core of the courts' reasoning was that it is impermissible for the state to have expressly elevated the Black voting age population (BVAP) in 12 districts to the 55% mark. (See Figure 11 to see this conspicuous feature.) Defending the enacted plan, the state variously claimed that the high BVAP was necessary for compliance with the Voting Rights Act and that it was a natural consequence of the state's geography and demographics. The courts disagreed, finding that 55% BVAP was unnecessary for VRA compliance in 11 of 12 districts, and that it caused dilution of the Black vote elsewhere in the state.

We can use a ReCom ensemble to investigate whether the BVAP > 0.55 property might happen by chance, as a mere consequence of human geography. (It does not.) What's more, we can home in on this last question, investigating how and where the 'packing' in the high-BVAP districts leads to 'cracking' in others. In the figure, we can locate the costs to electoral opportunity across the remaining districts: it is the next four districts and even the nine after that that exhibit depressed BVAP, supporting claims of vote dilution.

We emphasize that ensemble analysis cannot stand alone in the study of gerrymandering, but it provides a unique ability to identify outliers against alternative valid plans, holding political and physical geography constant. It is also important to note that this proposed use of ensembles is strictly for *assessment*, and we in no way endorse the use of randomly sampled plans for enactment. The use of modeling to assess human judgment does not demand the excision of human judgment.

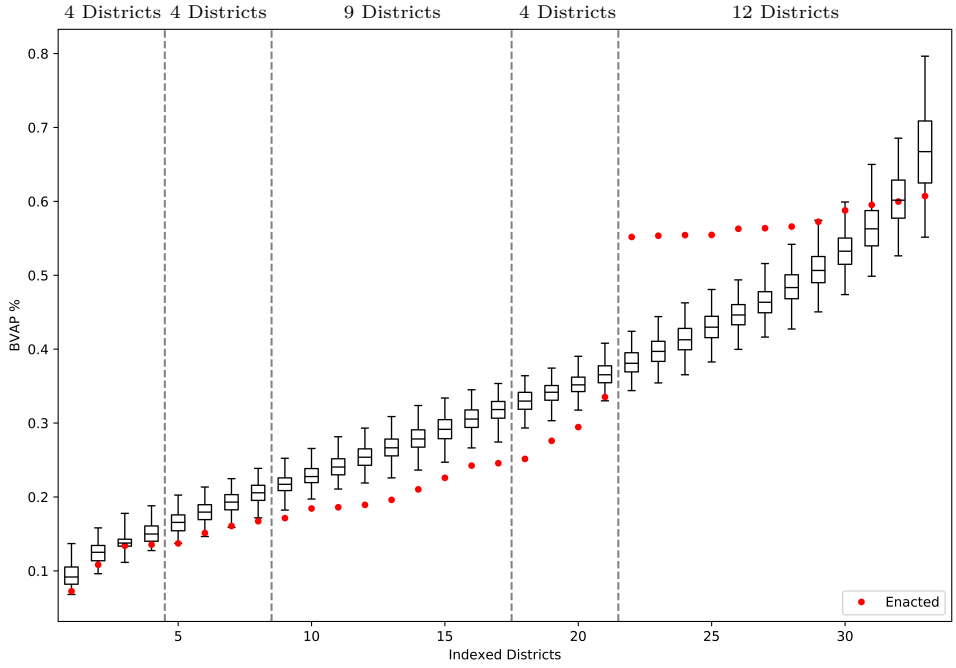


FIGURE 11. Black voting age population (BVAP) in the 33 districts affected by the court challenge. To facilitate comparison, districts are sorted from lowest BVAP to highest in each plan. Dots show the BVAP levels in the challenged plan, and the box-and-whiskers show the ensemble of alternative plans. The 'packing' of the top 12 districts has an unmistakable cost in the following four districts, and even the next nine, which diminishes the opportunity of Black voters to elect candidates of choice.

8. Discussion and Conclusion

Ensemble-based analysis provides much-needed machinery for understanding districting plans in the context of viable alternatives. By assembling a diverse and representative collection of plans, we can learn about the range of possible district properties along several axes, from partisan balance to shape to demographics. When a proposed plan is shown to be an extreme outlier relative to a population of alternatives, we might infer that the plan is better explained by goals and principles that were not stated (and so were not incorporated in the model design).

Due to the extremely large space of possible plans for most realistic redistricting problems, we can come nowhere close to complete enumeration of alternatives. For this reason, we turn to algorithmic sampling, but the design of an ensemble-generation algorithm is a subtle task with major mathematical, statistical, and computational challenges. Comparator plans must be legally viable and pragmatically plausible to draw power from the conclusion that a proposed plan has

very different properties. Moreover, to promote consistent and reliable analysis, we should strive to connect the sampling method to a well-defined distribution on the space of plans that not only has favorable qualitative properties but also can be sampled tractably. This consideration leads us to study design and diagnostics for Markov chains.

In this article we have introduced and surveyed ReCom, focusing on a spanning tree bipartition implementation, discussing it in theoretical and empirical terms. Across a range of small and large experiments with synthetic and observed data, we find that a run assembled in hours to days on a standard laptop produces large, diverse ensembles of plausible districting plans. As we write in early 2021, the new redistricting cycle has begun, and we expect these methods to have a practical impact as line-drawers—and the watchdogs that hold them to high standards—gear up for a districting reboot around the country. At the same time, the associated research questions are very much alive.

8.1. **Open Questions.** We end with a sampling of the many interesting questions and research directions that remain to be explored.

Mathematics.

- Explore the mathematical properties of spanning tree bipartitioning. For instance, what proportion of spanning trees in a grid have a *balanced cut*—an edge whose complementary components have the same number of nodes? How about in a triangular lattice?
- Following Akitaya et al. (2019) and Akitaya et al. (2020), study the ergodicity of the chain, that is, the connectivity of the state space by Flip and ReCom moves, under various constraints on district population balance. In particular, estimate the ReCom diameter of the state space of exactly balanced k-partitions of $n \times n$ grids—the most steps that might be required to connect any two partitions. We conjecture that the diameter is sublinear (in fact, logarithmic) in n for fixed k, despite the super-exponential growth of the state space itself.
- Building on the same two papers and on Najt et al. (2019) and Cohen-Addad et al. (2020), develop our understanding of computational complexity of sampling from various distributions on balanced partitions. Those authors collectively prove results for planar graphs and for bounded tree width; it would be valuable to continue to propose smaller classes of graphs that come closer to the lattice-like structure observed in geographic dual graphs.
- Prove rapid mixing of ReCom for the $n \times n$ grid case—that is, show that the chain approaches stationarity in a number of steps that is polynomial in n.
- Experiments show that the number of cut edges appears to be normally distributed in a ReCom ensemble (see Appendix Figure 14(B)). Prove a central limit theorem for boundary length in ReCom sampling of $n \times n$ grids into k districts, with parameters depending on n and k. Clelland et al. (2020) find a nearly linear relationship between cut edges and sp in an empirical analysis, but with a somewhat different slope for Flip and for ReCom samples. There are many questions to be explored in that direction.

Computation.

- Propose other balanced bipartitioning methods to replace spanning trees, supported by fast algorithms. Subject these methods to similar tests of quality, like adaptability to districting principles and heuristic convergence in projection to summary statistics.
- Find effective parallelizations to multiple CPUs while retaining control of the sampling distribution.

Applied Modeling.

- Study the stability of ReCom summary statistics to perturbations of the underlying graph. This ensures that ensemble analysis is robust to some of the implementation decisions made when converting geographical data to a dual graph.
- Stationarity can be reached more quickly for certain summary statistics than for others. Find conditions on summary statistics that suffice for faster convergence in projection. For the summary statistics most relevant to redistricting, compare the outputs across ensemble generation techniques.
- Identify sources of voting pattern data (e.g., recent past elections) and summary statistics (e.g., metrics in the political science literature) that best capture the signatures of racial and partisan gerrymandering.
- Consider whether these analyses can be gamed: Could an adversary with knowledge of a Markov proposal create plans that are extreme in a way that is hidden, avoiding an outlier finding?

ReCom is available for use as an open-source software package, accompanied by a suite of tools to process maps and facilitate MCMC-based analysis of plans (Voting Rights Data Institute, 2018, 2020). Beyond promoting adoption of this methodology for ensemble generation, we aim to use this release as a model for open and reproducible development of tools for redistricting. By making code and data public—and making a sustained effort to thoughtfully engage with the problem in its full political and legal complexity—we can promote public trust in expert analysis and facilitate broader engagement among the many interested parties in the redistricting process.

Disclosure Statement. The authors acknowledge the generous support of the Prof. Amar G. Bose Research Grant, the Jonathan M. Tisch College of Civic Life, and NSF Convergence Accelerator Grant No. OIA-1937095.

Acknowledgments. We are grateful to the many people whose discussion and input informed our approach to this work. We thank Sarah Cannon, Sebastian Claici, Jeanne Clelland, Lorenzo Najt, Wes Pegden, Dana Randall, Zach Schutzman, Matt Staib, Thomas Weighill, and Pete Winkler for wide-ranging conversations about spanning trees, Markov chain theory, MCMC dynamics, and the interpretation of ensemble results. We are grateful to Brian Cannon for his help and encouragement in making our Virginia analysis relevant to the practical reform effort. The GerryChain software accompanying this article was initiated by participants in the Voting Rights Data Institute (VRDI) at Tufts and MIT, and we are deeply grateful for their hard work, careful software development, and ongoing involvement. We particularly thank Parker Rule for improvements that make our chain code more powerful and efficient and for experimental work to support this article, and Max Hully and Ruth Buck, whose data curation and software engineering were instrumental in this research program.

Contributions. The authors contributed equally to the long-term research program overviewed in this article, and the author order follows the alphabetical convention in mathematics.

References

Abrishami, T., Guillen, N., Rule, P., Schutzman, Z., Solomon, J., Weighill, T., & Wu, S. (2020). Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, *42*(5), A3340–A3366. https://doi.org/10.1137/19m1295258

Akitaya, H. A., Jones, M. D., Korman, M., Meierfrankenfeld, C., Munje, M. J., Souvaine, D. L., Thramann, M., & Tóth, C. D. (2019). Reconfiguration of connected graph partitions.

Akitaya, H. A., Korman, M., Korten, O., Souvaine, D. L., & Tóth, C. D. (2020). Reconfiguration of connected graph partitions via recombination.

Altman, M., & McDonald, M. (2010). The promise and perils of computers in redistricting. *Duke J. Const. L. & Pub. Policy*, *5*, 69.

Angulu, H., Buck, R., DeFord, D., Duchin, M., Fain, H., Hully, M., Khan, M., Schutzman, Z., & York, O. (2020). Study of reform proposals for Chicago city council. *MGGG Technical Report*, 1–31. https://mggg.org/chicago.pdf

Autry, E. A., Carter, D., Herschlag, G., Hunter, Z., & Mattingly, J. C. (2020). Multi-scale merge-split Markov chain Monte Carlo for redistricting.

Bar-Natan, A., Najt, L., & Schutzman, Z. (2020). The gerrymandering jumble: Map projections permute districts compactness scores. *Cartography and Geographic Information Science*, *47*(4), 321–335. https://doi.org/10.1080/15230406.2020.1737575

Barnes, R., & Solomon, J. (2020). Gerrymandering and compactness: Implementation flexibility and abuse. *Political Analysis, First View*. https://doi.org/10.1017/pan.2020.36

Becker, A., Duchin, M., Gold, D., & Hirsch, S. (2021). Computational redistricting and the voting rights act. *preprint*.

Benade, G., & Procaccia, A. (2020). Abating gerrymandering by mandating fairness. *Preprint*. http://procaccia.info/papers/fcf.pdf

Besag, J., & Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, *76*(4), 633–642. https://doi.org/10.1093/biomet/76.4.633

Buchanan, A., Lykhovyd, E., & Validi, H. (2019). Imposing contiguity constraints in political districting models. *Preprint*. http://www.optimization-online.org/DB_HTML/2020/01/7582.html

Caldera, S., DeFord, D., Duchin, M., Gutekunst, S. C., & Nix, C. (2020). Mathematics of nested districts: The case of Alaska. *Statistics and Public Policy*, *7*(1), 39–51. https://doi.org/10.1080/2330443x.2020.1774452

Cannon, S., Duchin, M., Randall, D., & Rule, P. (2020). A reversible recombination chain for graph partitions. *Preprint*. https://mggg.org/ReCom

Carter, D., Herschlag, G., Hunter, Z., & Mattingly, J. (2019). A merge-split proposal for reversible Monte Carlo Markov chain sampling of redistricting plans. *arxiv:1911.01503*.

Chen, J., & Rodden, J. (2013). Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*, *8*(3), 239–269. https://doi.org/10.1561/100.00012033

Chen, J., & Rodden, J. (2016). The losers bonus: Political geography and minority party representation.

Chikina, M., Frieze, A., Mattingly, J. C., & Pegden, W. (2020). Separating effect from significance in markov chain tests. *Statistics and Public Policy*, *7*(1), 101–114. https://doi.org/10.1080/2330443X.2020.1806763

Chikina, M., Frieze, A., & Pegden, W. (2017). Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences*, *114*(11), 2860–2864. https://doi.org/10.1073/pnas.1617540114

Chinta, G., Jorgenson, J., & Karlsson, A. (2010). Zeta functions, heat kernels, and spectral asymptotics on degenerating families of discrete tori. *Nagoya Mathematical Journal*, *198*, 121–172. https://doi.org/10.1215/00277630-2009-009

Cho, W., & Liu, Y. (2016). Toward a talismanic redistricting tool: A computational method for identifying extreme redistricting plans. *Election Law Journal: Rules, Politics, and Policy*, *15*. https://doi.org/10.1089/elj.2016.0384

Cho, W. K. T., & Rubinstein-Salzedo, S. (2019). Understanding significance tests from a non-mixing Markov chain for partisan gerrymandering claims. *Statistics and Public Policy*, *6*(1), 44–49. https://doi.org/10.1080/2330443X.2019.1574687

Cirincione, C., Darling, T. A., & O'Rourke, T. G. (2000). Assessing South Carolina's 1990s congressional districting. *Political Geography*, *19*(2), 189–211. https://doi.org/10.1016/S0962-6298(99)00047-5

Clelland, J. N., Bossenbroek, N., Heckmaster, T., Nelson, A., Rock, P., & VanAusdall, J. (2020). Compactness statistics for spanning tree recombination. https://arxiv.org/abs/2103.02699

Cohen-Addad, V., Klein, P. N., & Marx, D. (2020). On the computational tractability of a geographic clustering problem arising in redistricting. *arXiv:2009.00188*. https://arxiv.org/abs/2009.00188

Cohen-Addad, V., Klein, P. N., & Young, N. E. (2017). Balanced power diagrams for redistricting. *CoRR*, *abs/1710.03358*. http://arxiv.org/abs/1710.03358

Cohen-Addad, V., Klein, P. N., & Young, N. E. (2018). Balanced centroidal power diagrams for redistricting. *SIGSPATIAL/GIS*.

DeFord, D., & Duchin, M. (2019). Redistricting reform in Virginia: Districting criteria in context. *Virginia Policy Review*, *12(2)*, 120–146.

DeFord, D., Duchin, M., & Solomon, J. (2018). Comparison of districting plans for the Virginia House of Delegates. *MGGG Technical Report*, 1–26. https://mggg.org/VA-report.pdf

DeFord, D., Duchin, M., & Solomon, J. (2019). Replication code. *GitHub repository*. Retrieved June 20, 2019, from https://github.com/drdeford/recom-VA

DeFord, D., Duchin, M., & Solomon, J. (2020). A computational approach to measuring vote elasticity and competitiveness. *Statistics and Public Policy*, *7*(1), 69–86. https://doi.org/10.1080/2330443x.2020.1777915

DeFord, D., Lavenant, H., Schutzman, Z., & Solomon, J. (2019). Total variation isoperimetric profiles. *SIAM Journal on Applied Algebra and Geometry*, *3*, 585–613. https://doi.org/10.1137/18m1215943

Diaconis, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc. (N.S.)*, *46*(2), 179–205. https://doi.org/10.1090/S0273-0979-08-01238-X

Duchin, M., & Spencer, D. (2021). Models, race, and the law. *Yale Law Journal Forum*, *130*, 744–797.

Duchin, M., & Tenner, B. (2018). Discrete geometry for electoral geography. *arXiv:1808.05860*.

Fifield, B., Higgins, M., Imai, K., & Tarr, A. (2020). A new automated redistricting simulator using Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, *29*, 715–728. https://doi.org/10618600.2020.1739532

Fifield, B., Imai, K., Kawahara, J., & Kenny, C. T. (2020). The essential role of empirical validation in legislative redistricting simulation. *Statistics and Public Policy*, *7*(1), 52–68. https://doi.org/10.1080/2330443x.2020.1791773

Fryer Jr, R. G., & Holden, R. (2011). Measuring the compactness of political districting plans. *The Journal of Law and Economics*, *54*(3), 493–535. https://doi.org/10.1086/661511

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo* (pp. 3–48). CRC Press, Boca Raton, FL. https://doi.org/10.1201/b10905-2

Haas, C., Hachadoorian, L., Kimbrough, S., Miller, P., & Murphy, F. (2020). Seed-fill-shift-repair: A redistricting heuristic for civic deliberation. *PLoS ONE*, *15*, e0237935. https://doi.org/10.1371/journal.pone.0237935

Håstad, J. (2006). The square lattice shuffle. *Random Structures & Algorithms*, *29*(4), 466–474. https://doi.org/10.1002/rsa.20131

Hebert, J. G., Smith, P. M., Vandenburg, M. E., & DeSanctis, M. B. (2010). *The realists' guide to redistricting: Avoiding the legal pitfalls, 2nd edition*. American Bar Association.

Herschlag, G., Kang, H. S., Luo, J., Graves, C. V., Bangia, S., Ravier, R., & Mattingly, J. C. (2020). Quantifying gerrymandering in North Carolina. *Statistics and Public Policy*, *7*(1), 30–38. https://doi.org/10.1080/2330443x.2020.1796400

Herschlag, G., Ravier, R., & Mattingly, J. C. (2017). Evaluating partisan gerrymandering in Wisconsin [arXiv: 1709.01596]. *arXiv:1709.01596 [physics, stat]*.

Jacobs, M., & Walch, O. (2018). A partial differential equations approach to defeating partisan gerrymandering. *arXiv:1806.07725*.

Jin, H. (2017). *Spatial optimization methods and system for redistricting problems* (Doctoral dissertation). University of South Carolina.

Kenyon, R. (2000). The asymptotic determinant of the discrete Laplacian. *Acta Mathematica*, *185*(2), 239–286. https://doi.org/10.1007/bf02392811

Kim, M. J. (2011). *Optimization approaches to political redistricting problems* (Doctoral dissertation). The Ohio State University.

Kleinberg, J., & Tardos, É. (2006). *Algorithm design*. Addison Wesley.

Levin, H. A., & Friedler, S. A. (2019). Automated congressional redistricting. *Journal of Experimental Algorithmics (JEA)*, *24*, 1–24. https://doi.org/10.1145/3316513

Liu, Y. Y., Cho, W. K. T., & Wang, S. (2016). PEAR: A massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm and Evolutionary Computation*, *30*, 78–92. https://doi.org/10.1016/j.swevo.2016.04.004

Magleby, D. B., & Mosesson, D. B. (2018). A new approach for developing neutral redistricting plans. *Political Analysis*, *26*(2), 147–167. https://doi.org/10.1017/pan.2017.37

Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2018). IPUMS national historical geographic information system: Version 13.0 [database]. https://doi.org/10.18128/D050.V13.0

Mattingly, J. (2017). Expert report of Jonathan Mattingly. https://s10294.pcdn.co/wp-content/uploads/2016/05/Expert-Report-of-Jonathan-Mattingly.pdf

McCartan, C., & Imai, K. (2020). Sequential Monte Carlo for sampling balanced and compact redistricting plans.

Nagel, S. S. (1965). Simplified bipartisan computer redistricting. *Stanford Law Review*, *17*(5), 863–899. https://doi.org/10.2307/1226994

Najt, L., DeFord, D., & Solomon, J. (2019). Complexity of sampling connected graph partitions. *arXiv:1908.08881*.

Nascimento, M. C., & de Carvalho, A. C. (2011). Spectral methods for graph clustering: A survey. *European Journal of Operational Research*, *211*(2), 221–231. https://doi.org/10.1016/j.ejor.2010.08.012

Pegden, W. (2017). Expert report of Wesley Pegden in League of Women Voters of Pennsylvania v. Commonwealth of Pennsylvania. https://www.pubintlaw.org/wp-content/uploads/2017/06/Expert-Report-Wesley-Pegden.pdf

Ricca, F., Scozzari, A., & Simeone, B. (2013). Political districting: From classical models to recent approaches. *Annals of Operations Research*, *204*(1), 271–299. https://doi.org/10.1007/s10479-012-1267-2

Schaeffer, S. E. (2007). Survey: Graph clustering. *Comput. Sci. Rev.*, *1*(1), 27–64. https://doi.org/10.1016/j.cosrev.2007.05.001

Tasnádi, A. (2011). The political districting problem: A survey. *Society and Economy*, *33*(3), 543–554. https://doi.org/10.1556/socec.2011.0001

Temperley, H. N. V. (1972). The enumeration of graphs on large periodic lattices. *Combinatorics (Proc. Conf. Combinatorial Math., Math. Inst., Oxford, 1972)*, 285–294. https://doi.org/10.1017/cbo9780511662072.024

Voting Rights Data Institute. (2018). GerryChain. *GitHub repository*. Retrieved May 17, 2019, from https://github.com/mggg/gerrychain

Voting Rights Data Institute. (2020). GerryChainJulia. *GitHub repository*. Retrieved March 4, 2020, from https://github.com/mggg/gerrychain

Weaver, J. B., & Hess, S. W. (1963). A procedure for nonpartisan districting: Development of computer techniques. *Yale Law Journal*, *73*, 288. https://doi.org/10.2307/794769

Weighill, T., & Rodden, J. (2021). Political geography and representation: A case study of districting in Pennsylvania. In M. Duchin & O. Walch (Eds.), *Political geometry*. Birkhäuser.

Wilson, D. B. (1996). Generating random spanning trees more quickly than the cover time. *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, 296–303. https://doi.org/10.1145/237814.237880

Zhang, P., DeFord, D., & Solomon, J. (2020). Medial axis isoperimetric profiles. *Computer Graphics Forum*, *39*(5), 1–13. https://doi.org/10.1111/cgf.14064

Appendix A. Algorithmic Variants

A.1. **Uniformizing the Flip Walk.** Although Algorithm 2 on its own does not have a uniform steady-state distribution, it is possible to adjust the transition probabilities to *target* a uniform distribution, as in the work of Chikina–Frieze–Pegden (Chikina et al., 2017). This can be done by adding self-loops to each plan in the state space to equalize the degree; the resulting technique is given in Algorithm 4. To see that this has a uniform steady-state distribution over the permissible partitions of $\mathcal{P}_k(G)$, we note that with M set to the maximum degree in the state space and

$$p = \frac{|\{(u, P(v)) : (u, v) \in \partial P\}|}{M \cdot |V|},$$

we have

$$X_P(Q) = \begin{cases} 1 - p & Q = P \\ p & |\{P(u) \neq Q(u) : u \in V\}| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Continuing to follow Chikina et al., we can accelerate the `Uniform Flip` algorithm without changing its proposal distribution by employing a function that returns an appropriate number of steps to wait at the current state before transitioning, so as to simulate the expected self-loops traversed before a non-loop edge is chosen. This variant is in Algorithm 5. Since the geometric variable computes the expected waiting time before selecting a node from $\partial_V P$, this recovers the same walk and distribution with many fewer calls to the proposal function.

Algorithm 4: `Uniform Flip`

Input: Dual graph $G = (V, E)$ and current partition P
Output: New partition Q

Initialize:
$p = \frac{|\{(u, P(v)) : (u, v) \in \partial P\}|}{M \cdot |V|}$

if *Bernoulli(1-p) = 1* **then**
  Return: P
else
  Allowed = `False`
  while *Allowed* = *`False`* **do**
    $Q = \mathsf{Node\ Choice}(G, p)$
    Allowed $= C(Q)$
  end
  Return: Q
end

Algorithm 5: `Uniform (Fast)`

Input: Dual graph $G = (V, E)$ and current partition P
Output: Number of steps to wait in the current state (σ) and next partition (Q)

Initialize:
$p = \frac{|\{(u, P(v)) : (u, v) \in \partial P\}|}{M \cdot |V|}$
$\sigma \sim \text{Geometric}(1 - p)$

$Q = \mathsf{Node\ Choice}(G, p)$
if $C(Q) =$ *`False`* **then**
  Return: (σ, P)
else
  Return: (σ, Q)
end

A.2. **The General Framework for Recombination.** Algorithm 6 offers an extremely general family of related Markov chains.

Algorithm 6: `Recombination (General)`

Input: Dual graph $G = (V, E)$, the current partition P, the number of districts to merge ℓ
Output: The next partition Q

Select $\ell \geq 2$ districts $W_1, W_2, \ldots, W_\ell$ from P.
Form the induced subgraph H of G on the nodes of $W = \bigcup_{i=1}^{\ell} W_i$.
Create a partition $R = \{U_1, U_2, \ldots, U_\ell\}$ of H
Define $Q(v) = \begin{cases} R(v) & \text{if } v \in H \\ P(v) & \text{otherwise} \end{cases}$
Return: Q

There are two algorithmic design decisions that are required to specify the details of a ReCom chain:

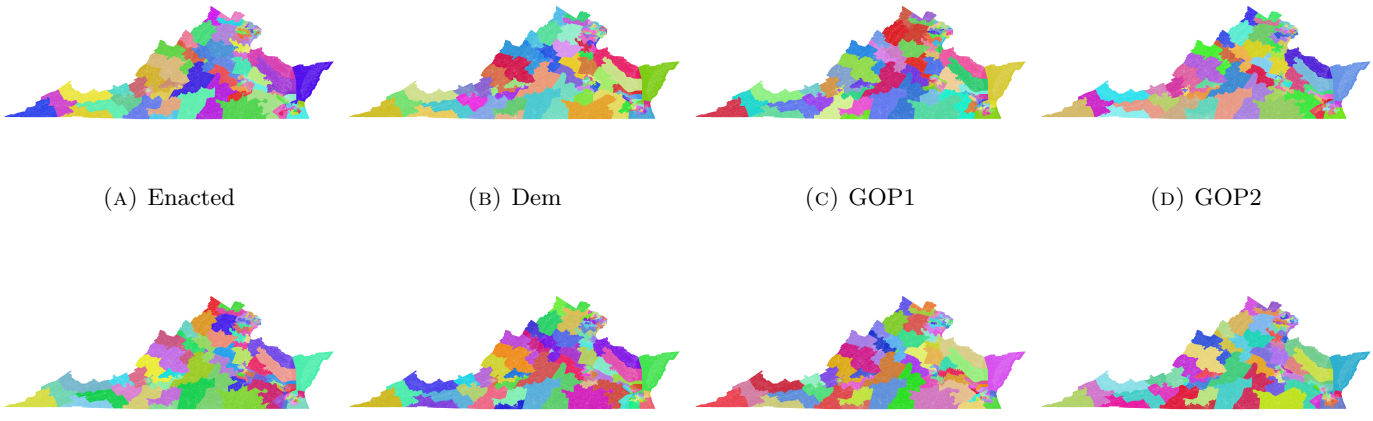
- The first parameter in the ReCom method is how to **choose which districts are merged** at each step. By fixing the partitioning method, we can create entirely new plans as in Section 3.1 by merging all of the districts at each step ($\ell = k$). For most of our use cases, we work at the other extreme, taking two districts at a time ($\ell = 2$), and we select our pair of adjacent districts to be merged proportionally to the length of the boundary between them, which improves compactness quickly. Bipartitioning is usually easier to study than ℓ-partitioning for $\ell > 2$. More importantly for this work, the slow step in a recombination chain is the selection of a spanning tree. Drawing spanning trees for the $\ell = k$ case (the full graph) is many times slower than for $\ell = 2$ when k is large, making bipartitioning a better choice for chain efficiency. This approach also generalizes in a second way: We can take a (maximal) matching on the dual graph of districts and bipartition each merged pair independently, taking advantage of the well-developed and effective theory of matchings.
- The choice of **(re)partitioning method** offers more freedom. Desirable features include full support over contiguous partitions, ergodicity of the underlying chain, ability to control the distribution with respect to legal features (particularly population balance), computational efficiency, and ease of explanation in non-academic contexts like court cases and reform efforts. Potential examples include standard graph algorithms, like the spanning tree partitioning method we will introduce in Section 4.3, as well as methods based on minimum cuts, spectral clustering, or shortest paths between boundary points.

Setting these two choices gives a well-defined Markov chain.

Appendix B. Plots for Virginia Case Study

Here, we present supporting evidence for the Virginia House of Delegates ensemble analysis described in Section 7, focusing on the portion of the state covered by the invalidated districts and their neighbors. Figure 12 shows two possible attempts to assess whether the Black voting age population (BVAP) is excessively elevated in the top 12 districts without the benefit of ensemble analysis. One approach is to use other human-made plans for comparison. Besides the original enacted plan, Figure 12 features some replacement proposals introduced in the legislature—a Democratic caucus plan (Dem) and a sequence of Republican counterproposals (GOP1, GOP2, GOP3), organizing the 33 districts from lowest BVAP to highest for each plan. The figure also shows statistics for reform plans proposed by the civil rights group NAACP and by the Princeton Gerrymandering Project, and finally the plan drawn by a court-appointed expert (or 'special master'). Interpreting these comparisons is difficult because the alternative plans are sharply limited in number, and their designers may have had their own agendas. A second approach is to forego the comparison with other plans and simply make the observation that the enacted plan's BVAP values conspicuously jump the 37-55% BVAP range, the same range that expert reports indicate might be plausibly necessary for Black residents to elect candidates of choice. But neither of these adequately controls for the effects of the actual clustering of Black population across the state geography—maybe the enacted plan just shows how the population would fall across districts formed without undue attention to race. To address that, the ensemble method generates a large, diverse collection of alternative plans made without consideration of racial statistics, holding the state's human and physical geography constant.

Figures 13–15 demonstrate nonconvergence for Flip chains and are suggestive of convergence for ReCom, though it is always difficult to rule out pseudo-convergence. In particular, Figure 15 makes use of a metric called the mean-median score; it is a signed measure of party advantage that is one of the leading partisan metrics in the political science literature. In Figure 16, we apply the ReCom outputs, studying the full ensemble (top plot) and the winnowed subset of the ensemble containing only plans in which no district exceeds 60% BVAP (bottom). This finally allows us to answer questions about whether structural constraints explain the BVAP pattern in the enacted plan. The full ensemble suggests that the pattern is not explained by the human geography of Virginia or by the districting rules of compactness, contiguity, and population balance. The winnowed ensemble helps us determine whether disqualifying plans with extremely elevated Black population will change the overall properties of the sample, mitigating the indicators of packing and cracking. (It does not.)



(A) Enacted (B) Dem (C) GOP1 (D) GOP2

(E) GOP3 (F) NAACP (G) Princeton (H) Remedial

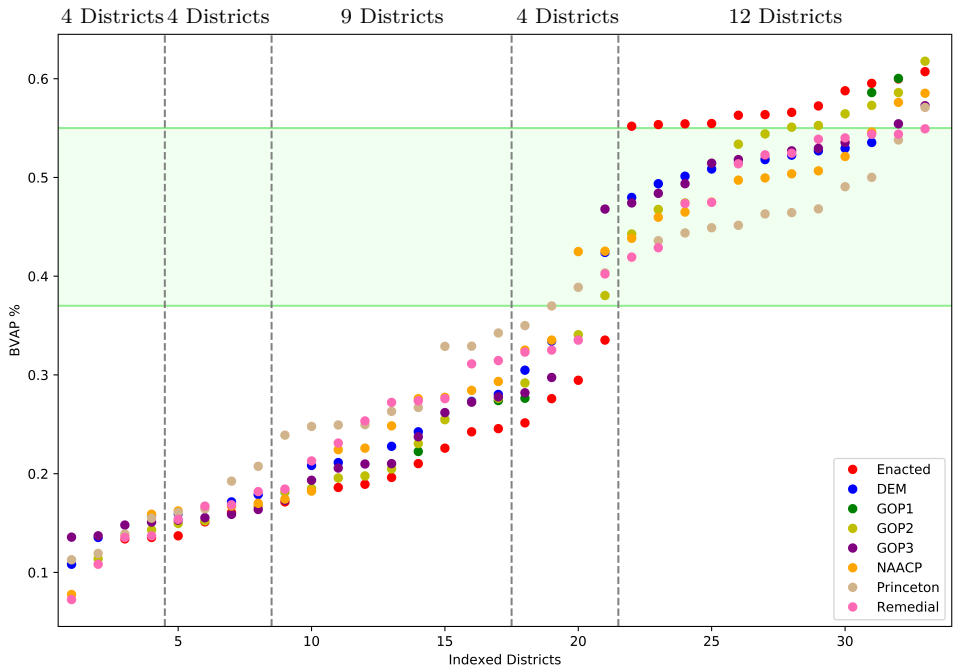


FIGURE 12. Eight proposed House of Delegates plans as described in the text. The boxplot shows the Black voting age population (BVAP) in the 33 districts affected by the court ruling, ordered from lowest to highest BVAP in each plan. The 2011 enacted plan jumps the key 37-55% BVAP range entirely, but the collection of other plans makes it difficult to tell how many more 37-55% BVAP plans might be expected or possible.

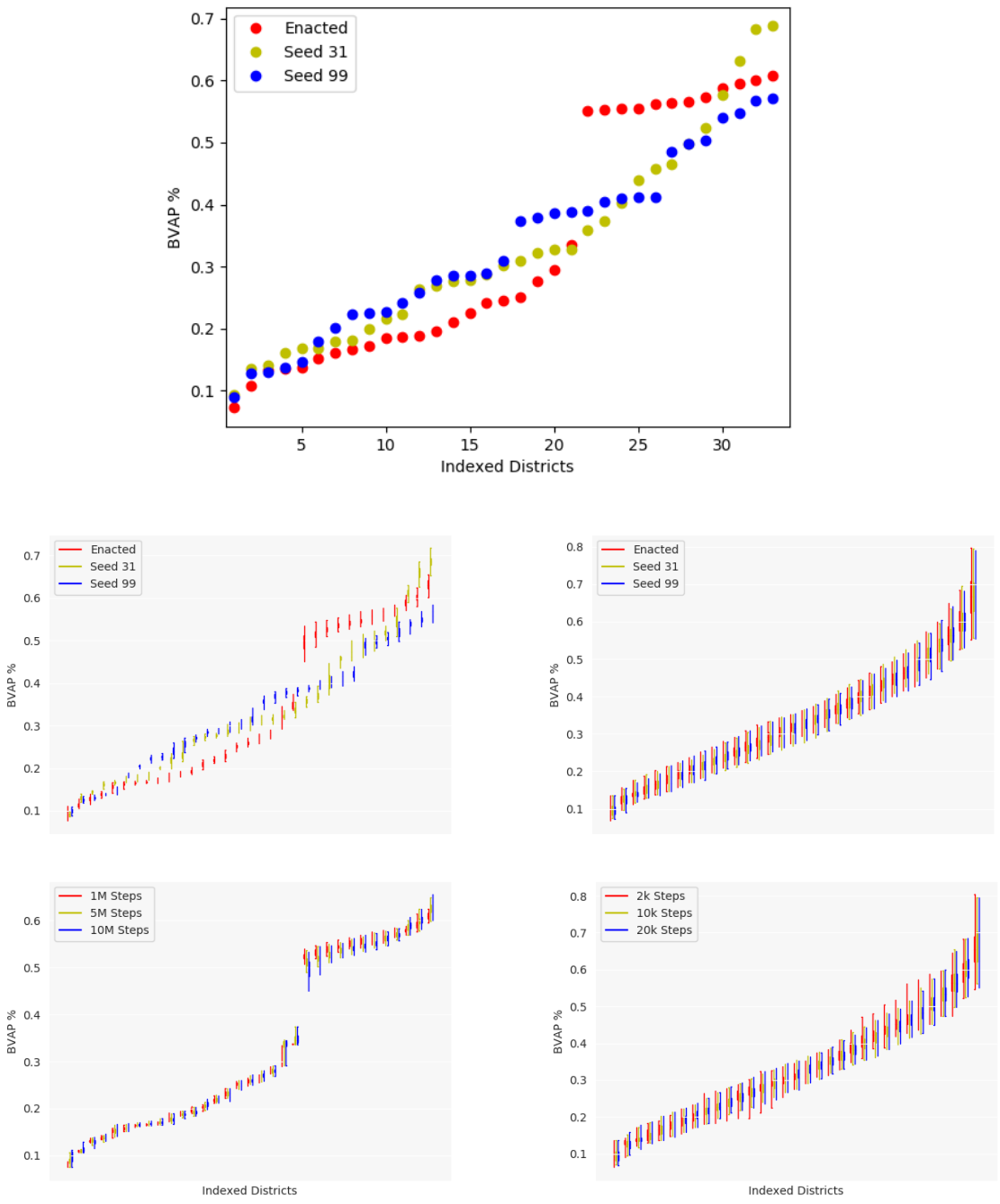
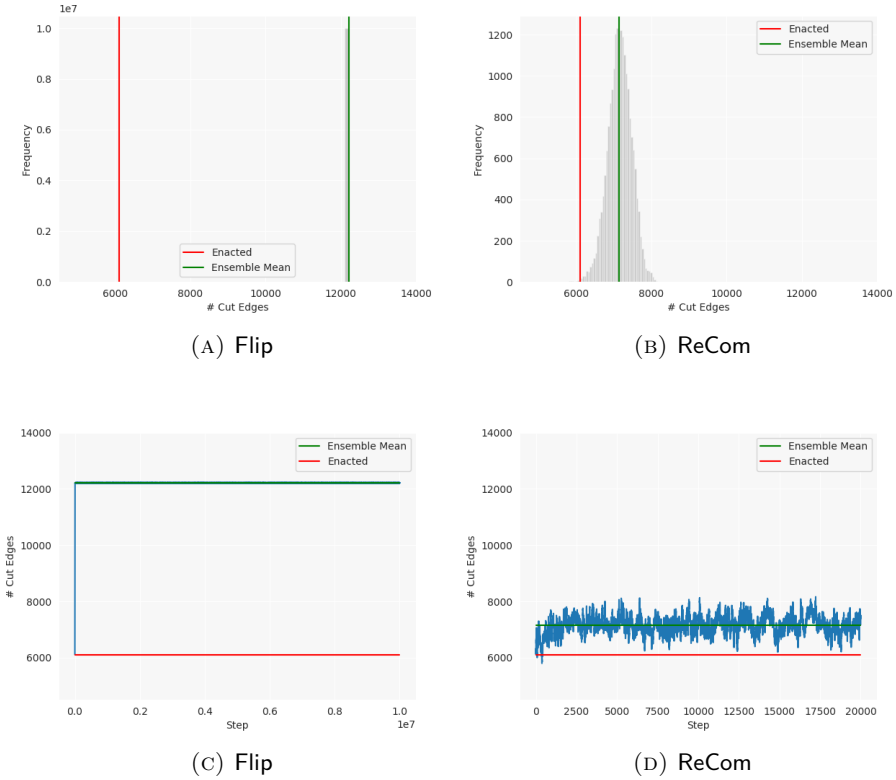


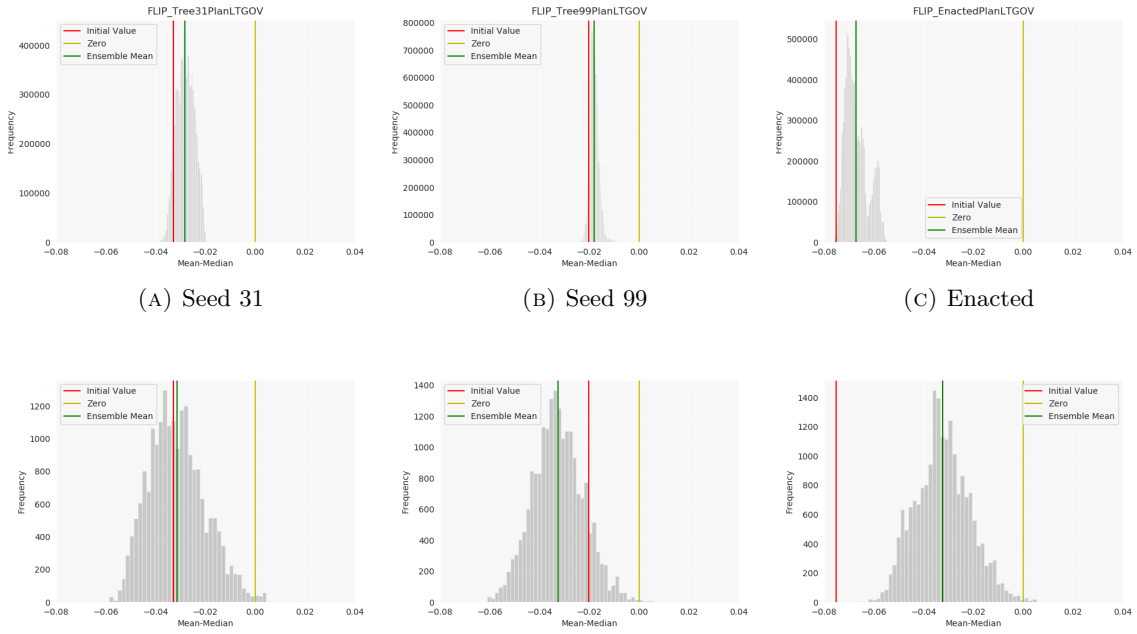
FIGURE 13. Convergence heuristics. Black voting age population (BVAP) levels in the enacted plan are compared to two synthetically generated seed plans. Ten million steps is not enough to mitigate the dependence on the starting point in a Flip run. By contrast, 20,000 steps overcomes the dependence on starting point for a recombination run, with most of the progress in the first 10,000 steps. Top row: levels at starting points. Middle row: Flip (left) and ReCom (right) ensembles from three starting points. Bottom row: runs of varying lengths starting from enacted plan.



(A) Flip (B) ReCom

(C) Flip (D) ReCom

FIGURE 14. Compactness comparison. Histograms (a,b) and traces (c,d) of the boundary length. Flip ensembles saturate the worst allowable compactness score (here, set to twice the value of the enacted plan).



(A) Seed 31 (B) Seed 99 (C) Enacted

(D) Seed 31 (E) Seed 99 (F) Enacted

FIGURE 15. Projection to partisan statistics. Mean-median (partisan symmetry) scores, illustrating dependence of Flip ensembles on starting point after one million steps.

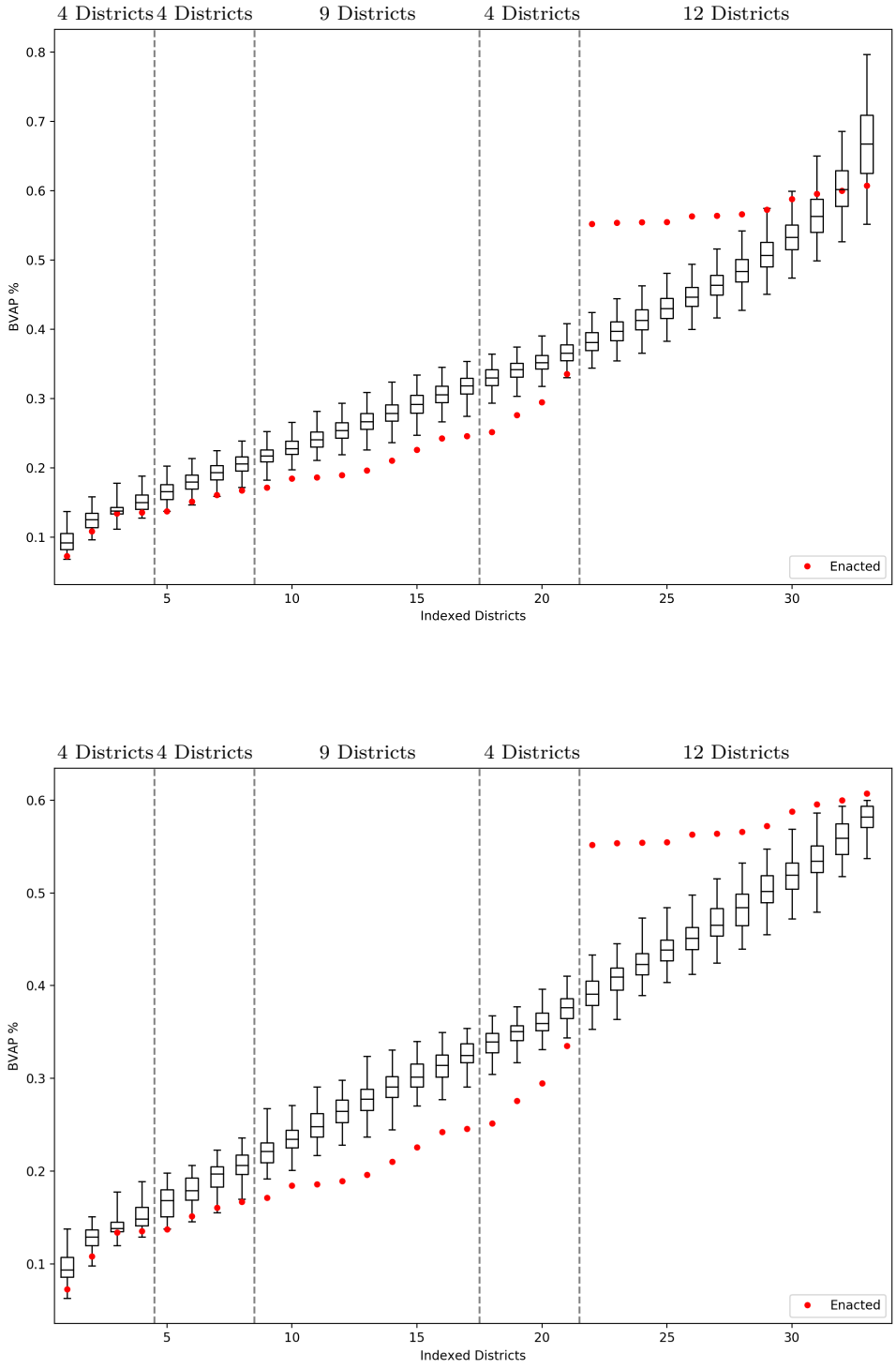


FIGURE 16. Ensemble analysis. Black voting age population (BVAP) levels in the Enacted plan can now be compared to an ensemble of population-balanced, compact plans that hold the state's demographics and geography constant. Top: full ReCom ensemble, repeated from Section 7 for comparison. Bottom: same ensemble, winnowed to $\leq$60% BVAP.