Recombination:
A family of Markov chains for redistricting

Daryl DeFord, Moon Duchin, and Justin Solomon*

March 27, 2020

Contents

*Following the convention in mathematics, author order is alphabetical.

Abstract

Redistricting is the problem of partitioning a set of geographical units into a fixed number of districts, subject to a list of often-vague rules and priorities. In recent years, the use of randomized methods to sample from the vast space of districting plans has been gaining traction in courts of law for identifying partisan gerrymanders, and it is now emerging as a promising analytical tool for legislatures and independent commissions. In this paper, we set up redistricting as a graph partition problem and introduce a new family of Markov chains called Recombination (or ReCom). These are large-step random walks on the space of graph partitions, in contrast with commonly used Flip walks, which randomly 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 demonstrate advantages of ReCom on real-world data and explain both the challenges of the Markov chain approach and the analytical tools that it enables. We close with a short case study involving the Virginia House of Delegates.

1 Introduction

In many countries, geographic regions are divided into districts that elect political representatives, 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, and political 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 and in public opinion. Further complicating matters, agenda-driven redistricting makes it common for line-drawers to *gerrymander*, or to design plans specifically skewing elections toward a preferred outcome, such as favoring or disfavoring a political party, demographic group, or collection of incumbents.

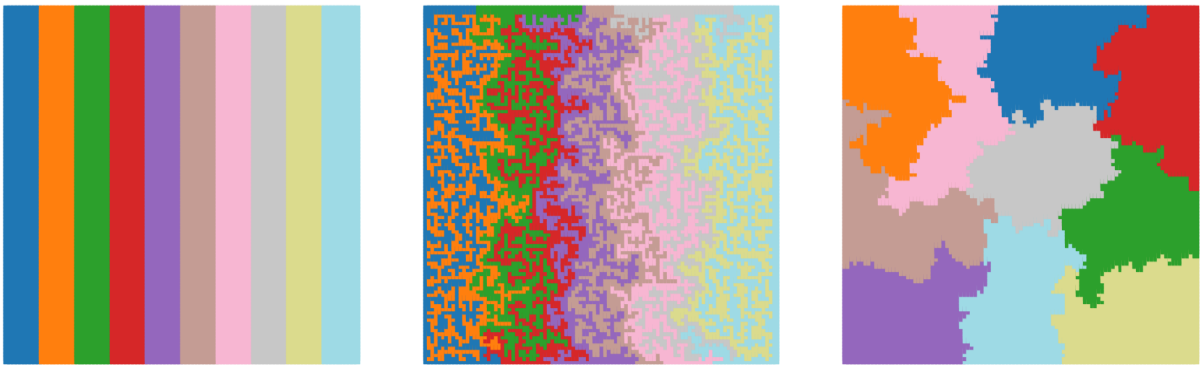
The fundamental technical challenge in the study of redistricting is to contend with the sheer number of possible ways to construct districting plans. State geographies admit enormous numbers of divisions into contiguous districts; even when winnowing down to districting plans that satisfy criteria set forth by legislatures or commissions, 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 below.

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. For them to provide a proper counterfactual, however, we need some assurance of representative sampling, i.e., drawing from a probability distribution that successfully reflects the rules and priorities articulated by redistricters.

In one powerful application, ensembles have been used to conduct *outlier analysis*, arguing that a proposed plan has properties that are extreme outliers relative to the comparison statistics of an ensemble of alternative plans. Such methods have been used in a string of recent legal challenges to partisan gerrymanders (Pennsylvania, North Carolina, Michigan, Wisconsin, Ohio), which were all successful at the district court or state supreme court level. Outliers also received a significant amount of discussion by the U.S. Supreme Court, 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 in 2021 when new plans are enacted around the country. These methods can help clarify the influence of each individual state's political geography as well as the tradeoffs between possible rules and criteria.

The inferences that can be drawn from ensembles rely heavily on the distributions from which the ensembles are sampled. To facilitate sampling, *Markov chain Monte Carlo*, or MCMC, methods offer strong underlying theory and heuristics, in the form of mixing theorems and convergence diagnostics. Drawing from this literature, the new Markov chains described here pass several quality tests, even though (as is nearly always the case in applications) rigorous mixing time bounds are still out of reach.

The first and most natural approach to forming a random walk on graph partitions is to simply re-assign one node at a time through a random process—we will call this the Flip walk and will introduce several variations on this theme. Doing Flip in a districting context amounts to changing the labeling of individual geographic units along district borders. As an alternative, we define a new family of random walks called *recombination* (or ReCom) Markov chains on the space of partitions, based on a step that *merges* two or more districts and randomly *re-partitions* them to form a new plan. To make this method concrete, we will focus on the case of merging two districts at a time, subsequently re-partitioning using a spanning tree-based method. We will argue that spanning tree ReCom has favorable properties that make it well-suited to the study of redistricting. Critically for reliability of MCMC-based analysis, we present evidence that ReCom converges efficiently to a distribution that comports with traditional districting criteria, with little or no parameter-tuning by the user.



(a) Initial Partition (b) 1,000,000 Flip steps (c) 100 ReCom steps

Figure 1: Comparison of the basic Flip proposal versus the spanning tree ReCom proposal to be described below. Each Markov chain was run from the initial partition of a 100×100 grid into 10 parts shown at left. A typical plan drawn from the uniform distribution will have winding, "fractal" districts of the type suggested by (b), simply because there are far more of these than there are geometrically "tame" partitions.

1.1 Contributions

We introduce new proposal distributions called ReCom for MCMC on districting plans and argue that they provide an alternative to the previous Flip-based approaches for ensemble-based analysis with significant advantages in efficiency and replicability. In particular, we:

- discuss practical setup for implementing Markov chains for redistricting;
- describe the ReCom and Flip random walks on the space of graph partitions;
- provide evidence for efficient convergence and stable results with ReCom;
- study qualitative features of sampling distributions through summary statistics from real data, addressing common methodological variants like simulated annealing and parallel tempering; and
- provide a model analysis of racial gerrymandering in the Virginia House of Delegates.

To aid reproducibility of our work, an open-source implementation of ReCom is available online [Ins18]. A 2018 report on the Virginia House of Delegates case study that was written for reform advocates, legislators, and the general public is also available [DDS18]. The present authors and our collaborators have applied recombination chains in numerous theoretical and applied projects to date [CDD+19, DDS19a, ABD+20, BDGM20, DD19, NDS19, WR20], and several other teams have now adopted this methodology [CHHM19, BP20].

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 below in §3.1.) Although this sounds similar to previous successful settings for Markov chain methods, some essential features of the redistricting problem combine to present great challenges that can cause standard techniques to fail.

1.2.1 Non-uniform sampling

It is crucial to understand that sampling *uniformly* from all valid partitions is not an appropriate goal in this application, nor has uniform sampling been attempted in any legal application.[1] Although the uniform distribution over graph partitions might seem to be the canonical choice, non-uniform sampling is needed for two fundamental reasons: an ensemble drawn from the uniform distribution would be regarded as undesirable in the application domain, and there are serious complexity obstructions to uniform sampling at the practical scale of redistricting problems. The theoretical complexity obstructions are corroborated by full-scale experiments. Put simply, there is no hope to accomplish near-uniform sampling 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.

As we will explain in §5.1, a uniform sample will be dominated by wildly-shaped districts (illustrated above in Figure 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 Figures 7 and 13. 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 [NDS19]. The experiments here (and the experimental evidence in a range of other papers discussed below) corroborate the slow convergence of Flip walks—and their uniform variants—in practice.

Since we cannot reasonably target the uniform distribution, we must specify an alternative target. We argue that the natural stationary distribution of ReCom is an attractive choice. We give a closed-form expression that approximates the ReCom distribution in §5.1, and we describe a variant of ReCom whose steady state precisely matches that weighting on plans.

1.2.2 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 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 means that 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 make large changes encounter combinatorial obstructions: The

[1] For example, Wesley Pegden's expert work [Peg17a] 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 [Mat17] targets a prescribed non-uniform distribution.

sequence of changes that would be needed to travel between qualitatively distinct plans becomes exponentially unlikely at scale.

MCMC practitioners have also had great success with a suite of techniques that work by varying a "temperature" parameter, alternating between (1) heating: transiting the state space quickly, having loosened or set aside other constraints, and (2) cooling: tightening the constraints to improve to states with specified features. But in our setting, 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 you would expect, and can even cause temperature variation to produce near-loops returning to near their starting position rather than exploring the state space effectively.[2]

1.3 Review of computational approaches to redistricting

Computational methods for generating districting plans have appeared since at least the work of Weaver, Hess, and Nagel in the 1960s [WH63, Nag65]. Like much modern software for redistricting, early techniques like [Nag65] incrementally improve districting plans in some metric while taking criteria like population balance, compactness, and partisan balance into account. Many basic elements still important for modern computational redistricting approaches were already in place in that work. Quantitative criteria are extracted from redistricting practice (see our §3.2.1); contiguity is captured using a graph structure or "touchlist" (see our §3.1); 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 ("the trading part") even appears in their 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. Below, we highlight a few incomplete but representative examples; see [CDO00, Tas11, AM10, Sax18, RSS13] for broader surveys; only selected recent work is cited below.

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 [Kim11, Jin17]. Many authors have used Voronoi or power diagrams with some variant of k-means [FJH11, CKY17, CAKY18, LF19], and there has been a lineage of approaches through integer programming [BLV19] and even a partial differential equations approach with a volume-preserving curvature flow [JW18].

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 authors 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 a single best example.

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 outwards 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 with a tolerable balance. Typically, these algorithms abandon a plan and re-start if they reach a configuration that cannot be completed into a valid plan, which can happen often. Examples include [CR13, CR16, MM18]. We are not aware of any

[2] We will illustrate this below in Figure 9. Another view of this phenomenon can be found in [AGR+19, Fig. 12] where the annealing procedure does not allow the Markov chain to move a large distance through the state space.

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 [HKL$^+$18, HRM17, FHIT18, CFP17, CFMP19]. An evolutionary-style variant with the same basic step can be found in [CL16, LCW16]. The use of random walks for sampling is well developed across scientific domains in the form of *Markov chain Monte Carlo*, or 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 as well as crossover steps. The experiments below are not intended to reproduce the precise setup of any of these implementations. (in part because the detailed specifications and code are often not made 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. In addition to discussing these aspects of Flip chains, we present an alternative chain that gives us an occasion to debate the mathematical and modeling needs of the application to redistricting.

There have been some attempts to provide benchmarks to compare the various approaches to each other, but this is difficult. For instance, [FIKK19] has 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 future work could re-weight by other target distributions. Re-weighting for benchmark purposes is attempted in [CHHM19]. However, it is unclear if complete enumerations will be possible on a large enough geography to provide an opportunity for all the relevant phenomena of realistic redistricting problems to become apparent.

2 Markov chains

A Markov chain is simply a process for moving between positions in a *state space* according to a transition rule in 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 huge and fast-growing literature and a long track record of modeling success, including in a range of social science applications. See the classic survey [Dia09] 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.

This paper 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. For this application, a gold standard might be (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.

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 [BC89], Chikina et al. [CFP17, CFMP19]) 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.

Most previous MCMC methods for redistricting (for both local and global sampling) are built on variations of the same proposal distribution that we call a "flip step," for which each move reassigns a single geographic unit from one district to a neighboring district. This kind of proposal, for which we record several versions collectively denoted as Flip, is relatively straightforward to implement and in its simplest form satisfies the properties needed for Markov chain theory to apply. We will elaborate serious disadvantages of basic Flip chains, however, in an attempt to catch the literature up with practitioner knowledge: demonstrably slow mixing; stationary distributions with undesirable qualitative properties; and additional complications in response to standard MCMC variations like constraining, re-weighting, annealing, and tempering. We will argue that an alternative Markov chain design we call *recombination*, implemented with a spanning tree step, avoids these problems. We denote this alternative chain by ReCom. Both Flip and ReCom are discussed in detail in §4 below.

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 below can be interpreted as exploring the choice of a suitable design for a Flip chain.[3] 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.[4]

Some of the performance obstructions described below 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.

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 contours with a discrete model [DT18], and in practice the vast majority of algorithmic approaches discussed above adopt the discrete model. 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 ten years) or precincts (defined by state, county, or

[3]For instance, Figures 11 and 13 show that the subsampling parameter would have to be much larger than ten million for a Flip chain in Virginia.

[4]For a discussion of burn time, pseudo-convergence, and the applicability of the Markov Chain Central Limit Theorem to the $m = 0, r = 1$ case, see [Gey11].

local authorities). District boundaries extremely rarely cut through census blocks and typically preserve precincts,[5] 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 five to ten thousand precincts and several hundred thousand census blocks.

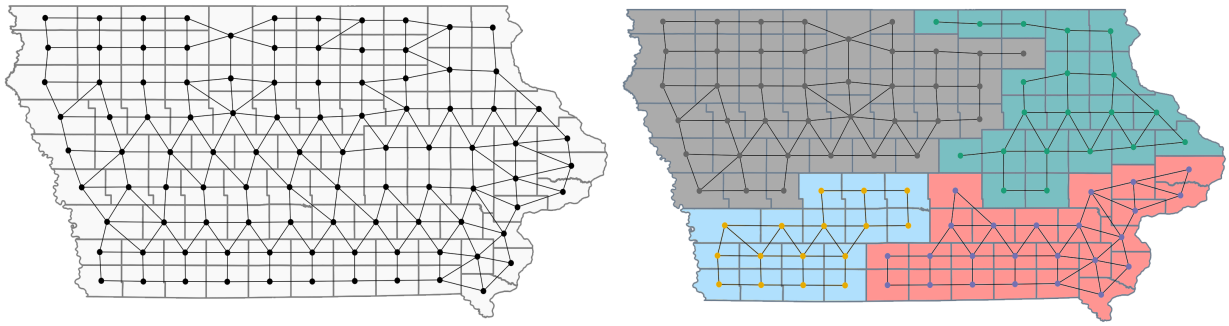


Figure 2: Iowa is currently the only state whose congressional districts are made of whole counties. The dual graph of Iowa's counties is shown here together with the current Iowa congressional districts.

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 2 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 [NdC11, Sch07]. 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.

3.1.1 Generating seed plans

To actually run our Markov chains, we need a valid initial state—or *seed*—in addition to the proposal method. Although in some situations we may want to start chains from the currently enacted plan, we will need other seed plans if we want 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 §1.3 above) or spanning tree methods (see §4.3 below) can be used for plan generation, 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 ten states that dictates that state House districts should nest perfectly inside state Senate districts, either two-to-one (AK, IA, IL, MN, MT, NV,

[5]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 there is no precinct shapefile publicly available.

OR, WY) or three-to-one (OH, WI). A constraining way to interpret 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. Even under this narrow interpretation, 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 in Minnesota [CDD$^+$19]. The actual choice left to redistricters, who in reality control House and Senate lines simultaneously, is far more open, and 10^{100} would not be an unreasonable guess.

3.2.1 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).[6,7]

- **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 disconnected pieces. Unfortunately, contiguity must be handled by building and cleaning dual graphs for each state on a case-by-case basis.

- **Compactness:** Most states have a "compactness" rule preferring regular district shapes, but few attempt a definition, and several of the attempted definitions are unusable.[8] We will handle it in a mathematically natural manner for a discrete model: we count the number of *cut edges* in a plan, i.e., the number of edges in the dual graph whose endpoints belong to different districts (see §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).

- **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 (also 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 [BGH$^+$17, DD19]

- **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. [HSVD10, Ch3-5]. Here, a modeler might reasonably choose to create a comparator ensemble made up of new plans that provide at least as many districts with a substantial minority share of voting age population as the previous plan.[9]

[6]The case law around tolerated population deviation is thorny and still evolving [HSVD10, Chapter 1]. 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.

[7]Excessively tight requirements for population balance can spike the rejection rate of the Markov chain and impede its efficiency. 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.

[8]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 underdefined, unstable, and inconsistent [DT18, BNNS19, BS19, DLSS19]. In practice, compactness is almost everywhere ruled by the eyeball test.

[9]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. However, the percentage of a minority group in the voting age population is frequently used as a loose proxy. For instance, in Virginia, there are two current Congressional districts with over 40% Black Voting Age Population, and a plausible comparator ensemble should contain many plans that preserve that property. This does not ensure that every such plan is fully compliant with the VRA.

- **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 about

- **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 to make plans mutually comparable.[10] A modeler with multiple objective functions must also decide whether to try 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 §6 provides details of how we approached some of the decisions above 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 described above 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.

We next turn to setting out proposal methods for comparison. A proposal method is a procedure for transitioning between states of $\mathcal{P}_k(G)$ according to a proposal distribution. 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 far 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 stochastic algorithm for modifying a partition.

In this section, we introduce the main Flip and ReCom proposals analyzed in the paper and describe some of their qualitative properties. We also devote some attention to the spanning tree method that we employ in our empirical analysis.

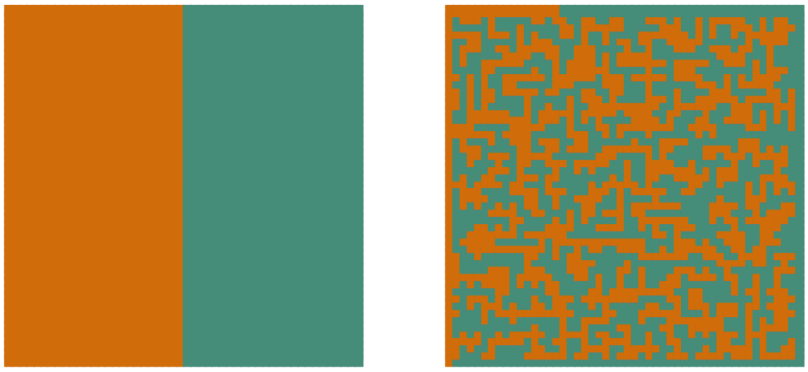
[10] If for instance we use an L^∞ or supremum norm to summarize the compactness scores of the individual districts, then all but the worst district can be altered with no penalty. Choosing L^1 or L^2 aggregation takes all scores into account, but to some extent allows better districts to cover for worse ones. Pegden has argued for L^{-1} to heavily penalize the worst abuses for scores measured on a $[0, 1]$ scale [CFP17, Peg17b].

4.2 Flip proposals

4.2.1 Node choice, contiguity, rejection sampling

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 3 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.

(a) Sequence of four flip steps



(b) Outcome of 500,000 flip steps

Figure 3: At each flip step, a single node on the boundary changes assignment, 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.

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 non-uniformity to the process. To construct a *reversible* Markov chain we follow [CFP17] 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 non-uniform, 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 1,3 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 discuss evidence below that one billion steps is enough to improve matters significantly, but not to the point of approximate convergence to stationarity. 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 under this proposal.[11] Attempts have been made to address this mixing issue in practice, including using simulated annealing or parallel tempering in [HRM17, HKL$^+$18, FHIT18] and a Swendsen-Wang variant in [FHIT18] that changes the assignments of several nodes at a time. However, we will show in §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.

4.2.2 Uniformizing

For practical interpretation of a sample, it can be useful to have a simple description of the distribution of probabilities associated to the Markov chain after a large number of steps. Although Algorithm 2 does not have a uniform steady state distribution, it is possible to re-weight the transition probabilities to target a *uniform* distribution, as in the work of Chikina–Frieze–Pegden [CFP17]. 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 3. To see that this has a uniform steady-state distribution over the permissible partitions of $\mathcal{P}_k(G)$, we note

[11] 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.

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 4. 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 3: `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 4: `Uniform Flip (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

On the other hand, attempting to sample from the uniform distribution causes problems for at least two reasons. First, sampling from uniform distributions over partitions runs into complexity obstructions (see §5.2 below), implying as a corollary that we should not expect these chains to converge efficiently. Even though abstract complexity results do not always dictate practical performance, in practice we will observe that extremely long runs are needed to produce substantial change in the map. We will demonstrate slow convergence on problems approximating real scale by showing that even the projection to summary statistics remains strongly correlated with the initial state (Figure 8). Secondly, even if we had a uniform sampling oracle, the distribution is massively concentrated in non-compact plans: generic connected partitions are remarkably snaky, with long tendrils and complex boundaries (see §5.1). The erraticness of typical shapes in the flip ensembles is undesirable from the perspective of districting, which places a premium on well-behaved boundaries. This also means that it is difficult for these chains to move effectively in the state space when compactness constraints are enforced, since generic steps increase the boundary length, leading to high rejection probabilities or disconnected state spaces. We evaluate some standard techniques for ameliorating this issue in our experiments below. Correcting the shape problem is not straightforward and introduces a collection of parameters that interact in complicated ways with the other districting rules and criteria.

4.3 ReCom proposals

The slow convergence and poor qualitative behavior of the Flip chain leads us to introduce 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 might be considered net favorable thanks to superior convergence and distributional design.

At each step of our new chain, we select a number of districts of the current plan and form the induced subgraph of the dual graph on the nodes of those districts. We then partition this new region according to an algorithm that preserves contiguity of the districts. We call this procedure *recombination* (ReCom), motivated by the biological metaphor of recombining genetic information. A general version of this approach is summarized in Algorithm 5; Figure 4 shows a schematic of a single step with this proposal.

Algorithm 5: `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

The ReCom procedure in Algorithm 5 is extremely general and varying the parameters described in more detail generates a family of related Markov chains. 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 §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, as we will discuss in §5.1.[12]
- 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 §4.3.1, as well as methods based on minimum cuts, spectral clustering, or shortest paths between boundary points.

With these two choices, we have a well-defined Markov chain to study. The experiments shown in the present paper are conducted with a spanning tree method of bipartitioning, which we now describe.

4.3.1 Spanning tree recombination

In all experiments below, we focus on a particular method of bipartitioning that creates a recombination chain whose behavior is well-aligned to redistricting. This method merges two adjacent districts (i.e., $\ell = 2$), selects a spanning tree of the merged subgraph, and cuts it to form two new districts. (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. In the figure below, 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. Our implementation uses the loop-erased random walk method of Wilson's algorithm [Wil96].[13]

[12] 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.

[13] Wilson's algorithm is notable in that it samples uniformly from all possible spanning trees in polynomial time.

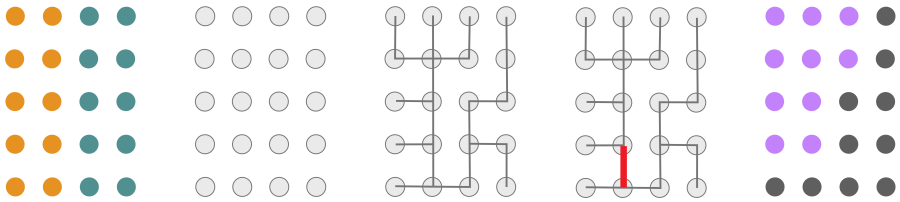


Figure 4: A schematic of a ReCom spanning tree step for a small grid with $k = 2$ districts that are merged ($\ell = 2$) and re-split. Deleting the indicated edge from the spanning tree leaves two connected components with an equal number of nodes.

- 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 example of rejection sampling in our implementation. In practice, the rejection rate is low enough that this step runs efficiently. If there are multiple edges that could be cut to generate partitions with the desired tolerance, we sample uniformly from among them.

Pseudocode for this technique is provided in Algorithm 6.

Algorithm 6: `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 Theoretical comparison

In §6, we will conduct experiments that provide intuition for qualitative behavior of the simple (unweighted) Flip and ReCom chains. However, precise mathematical characterization of their stationary distributions appears to be extremely challenging and is the subject of active research.[14] In this section, we provide high-level explanations of the two main phenomena that can be gleaned from experiments: ReCom samples preferentially from fairly compact districting plans while simple Flip ensembles are composed of plans with long and winding boundaries; and ReCom draws quickly-converging sample statistics while Flip statistics converge slowly. We also describe a reversible variant of ReCom with a closed-form stationary distribution.

5.1 Distributional design: compactness

"Compactness" is a vague but important term 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 this treatment, we focus on the discrete perimeter as a way to measure compactness. Recall from §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). There is a great deal of mathematical literature connected to 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 below—that Flip must contend with serious compactness problems—would apply to any reasonable score, as the figures illustrate.

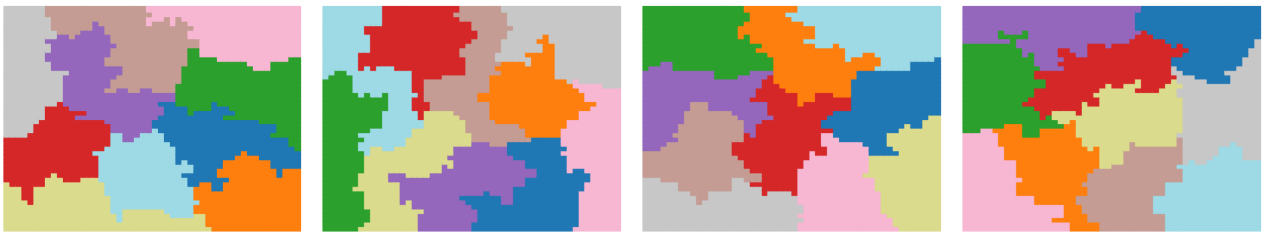


Figure 5: The ReCom proposal tends to produce compact or geometrically tame districts, with favorable isoperimetric ratios. Each of these plans was selected after 100 ReCom steps starting from the same vertical-stripes partition. Unlike the Flip samples, these partitions have relatively short boundaries in addition to displaying low correlation with the initial state.

The 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 is a polynomial number of 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.

Spanning trees provide a useful mechanism to produce contiguous partitions, since the deletion of any single edge from a tree leaves two connected subgraphs. Furthermore, the tendency of the spanning tree process will be to produce districts without skinny necks or tentacles. To see this, consider the $k = 2$ case first. The number of ways for the spanning tree 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.[15] Let us define $\mathsf{sp}(G)$ to be the number of spanning trees of

[14] This is true for Flip despite the ability to target various distributions: reweighting (whether uniform or Gibbs) is possible because the state space can be understood *locally*. There is no usable global description, which would be needed to write down the steady state.

[15] The idea that one can cut spanning trees to create partitions, and that the resulting distribution will have fac-

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.

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 $\mathsf{sp}(P)$. This helps us understand why more compact districts are up-weighted by recombination, as follows: Kirchhoff's counting formula for spanning trees 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 [Tem72]. There are difficult mathematical theorems that suggest that square subgrids have more spanning trees than any other subgraphs of grids with the same number of nodes [Ken00, CJK10]. But this means that if a district has a simple "neck" or "tentacle" with just two or three nodes, it can reduce the number of possible spanning trees by a factor of C^2 or C^3, making the district ten or thirty times less likely to be selected by a spanning tree process. The long snaky districts that are observed in the Flip ensembles are nearly trees themselves, and are therefore dramatically down-weighted by ReCom because they admit far fewer spanning trees than their plumper cousins. For example, the initial partition of the 50×50 grid in Figure 3 has roughly 10^{1210} spanning trees that project to it while the final partition has roughly 10^{282}. That means that the tame partition is over 10^{900} times more likely to be selected by a spanning tree ReCom step than the snaky partition, while uniform Flip weights them exactly the same.

New work of Cannon–Duchin–Randall–Rule [CDRR20] introduces a variant of ReCom that targets precisely this stationary distribution. By adding a correction term to the acceptance probability, the authors show that the 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 than the unweighted version, but that it obtains extremely similar summary statistics on grids and on real data at scale. Their paper also considers the effects of burn time and subsampling.

5.2 Complexity and mixing

Flip distributions and uniform distributions have another marked disadvantage for sampling: computational intractability. In the study of computational complexity, P $\subseteq$ RP $\subseteq$ NP are complexity classes (polynomial time, randomized polynomial time, and nondeterministic polynomial time), and it is widely believed that P = RP, and RP $\neq$ NP. Recent theoretical work of DeFord–Najt–Solomon [NDS19] 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 sampling proportionally to $x^{|\partial P|}$ for any $0 < x \leq 1$ is intractable, in the sense that an efficient solution would imply RP = NP. Note that the $x = 1$ case covers uniform sampling. This analysis implies that methods targeting the uniform distribution and natural variants weighted to favor shorter boundary lengths are likely to face complexity obstructions, particularly with respect to worst-case scenarios, which should require increased scrutiny of the quality of sampling.[16] Our experiments in §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.

Unlike Flip, 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 size of the full state space of balanced k-partitions of an $n \times n$ grid is larger than exponential in n. Based on a mixture of

tors proportional to the number of trees in a block, is a very natural one and appears for instance in the ArXiv note `https://arxiv.org/pdf/1808.00050.pdf`.

[16] In [FIKK19], Fifield et al. attempt to approximate uniform sampling by re-weighting a Gibbs sample. For this to succeed, the Gibbs chain would need to be run for long enough to accept significant numbers of exponentially unlikely proposals. Because the sample size needed would to get good estimates would therefore explode, this scheme does not circumvent the complexity obstructions to uniform sampling.

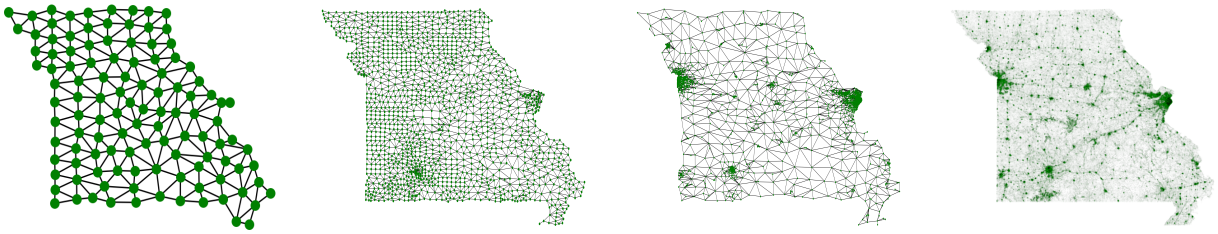
experiments and theoretical exploration, we conjecture that the full ReCom diameter of the state space—the most steps that might be required to connect any two partitions—is sublinear (in fact, logarithmic) in n. We further conjecture that ReCom is rapidly mixing (in the technical sense) on this family of examples, with mixing time at worst polynomial in n. By contrast, we expect that the Flip diameter of the state space for grids, and the mixing time, grow exponentially or faster in n. Even proving the ergodicity of these chains can be difficult in practice, depending on the topology of the underlying graph and the constraints enforced. In [AJK+19], it was shown that the state space is connected when the dual graph is 2–connected and the population can grow to two times ideal size. Generalizing and sharpening this result is an active area of current research; determining what types of population bounds disconnect the state space for a given graph is a difficult question.

The Markov chain literature has examples of processes on grids with constant scaling behavior, such as the Square Lattice Shuffle [Hås06]. 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, i.e., 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. The global nature of ReCom moves leaves open the possibility of this level of efficiency.

Our experiments below support the intuition that the time needed for effective sampling has moderate growth; tens of thousands of recombination steps give stable results on practical-scale problems whether we work with the roughly 9000 precincts of Pennsylvania or the roughly 100,000 census blocks in our Virginia experiments. Note that these observations do not contradict the theoretical obstructions in [NDS19], since ReCom is not designed to target the uniform distribution or any other distribution known to be intractable. While ReCom is decidedly nonuniform, the arguments in §5.1 indicate that this nonuniformity is desirable, as the chain preferentially samples from compact plans, thus comporting with traditional districting principles.

6 Experimental comparison

In this section, we will run experiments on the standard toy examples for graph problems, $n \times n$ grids, as well as on empirical dual graphs generated from census data. The real-world graphs can be large but they share key properties with lattices—they tend to admit planar embeddings, with most faces triangles or squares. Figure 6 shows the state of Missouri at four different levels of census geography, providing good examples of the characteristic structures we see in our applications.



(a) County (b) County Subunit (c) Census Tract (d) Census Block

Figure 6: Four dual graphs for Missouri at different levels of geography in the Census hierarchy. The graphs have 115, 1,395, 1,393, and 343,565 nodes respectively.

All of our experiments were carried out using the GerryChain software [Ins18], with additional source code available for inspection [DDS19b]. The state geographic and demographic data was obtained from the census TigerLine geography program accessed through NHGIS [MSVRR18].

6.1 Sampling distributions, with and without tight 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 7 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 right in 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 distinct 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 7 for an example. Sometimes, this is because an overly tight constraint disconnects the state space entirely and leaves the chain exploring a small connected component.[17] Recombination responds better to sharp constraints, and ReCom chains do not tend to run at the limit values when constrained. The interactions between various choices of constraints and priorities are so far vastly under-explored. In §6.3, we will consider the use of preferentially weighting steps rather than constraining the chains.

6.2 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 [Gey11], these push-forward statistics are called *functionals*.

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 which may mix slowly in the space of partitions will converge much more quickly in their projection to some summary statistics.

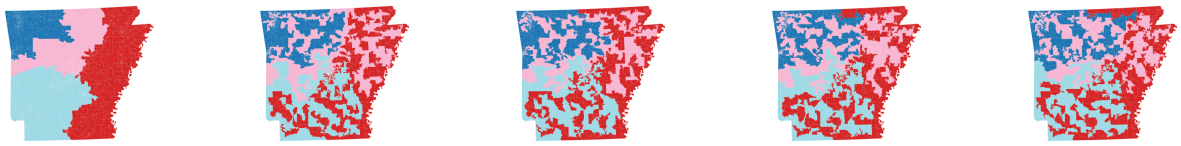
To investigate this possibility, we begin with a toy example with synthetic vote data on a grid, comparing the behavior of the Flip and ReCom proposals (Figure 8). For each Markov chain, we evaluate statistics using a vote distributions on a 100×100 grid, where each node is assigned to vote for a single 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; by the Markov Chain Central Limit Theorem, the sample statistics over that ensemble will converge to the push-forward of the stationary distribution, irrespective of starting point [Gey11, §1.8].

The 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 essentially only saw 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.

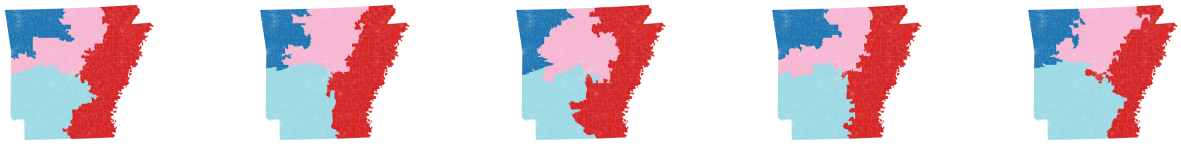
6.3 Weighting, simulated annealing, and parallel tempering

As we have shown above, 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

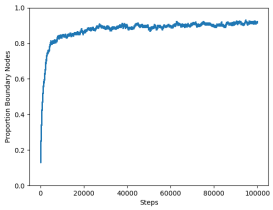
[17] An example of this behavior was presented in [CRS19, Fig 2], though its significance was misinterpreted by the authors with respect to the test in [CFP17].

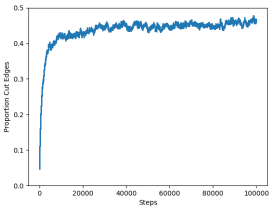


Run 1: 100K Flip steps, shown every 25K, no compactness constraint

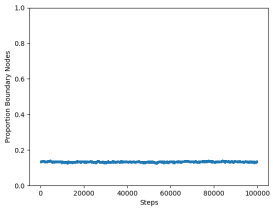


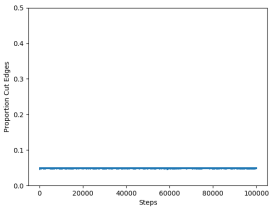
Run 2: 100K Flip steps, shown every 25K, limited to 5% total cut edges



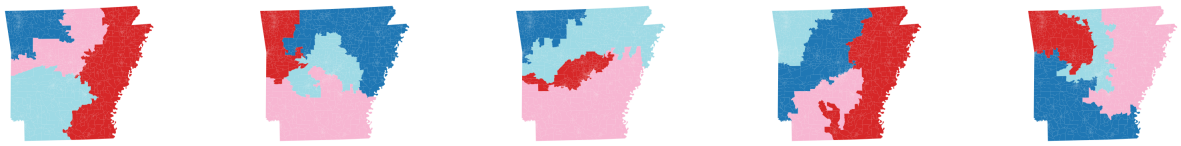


Run 1 boundary statistics

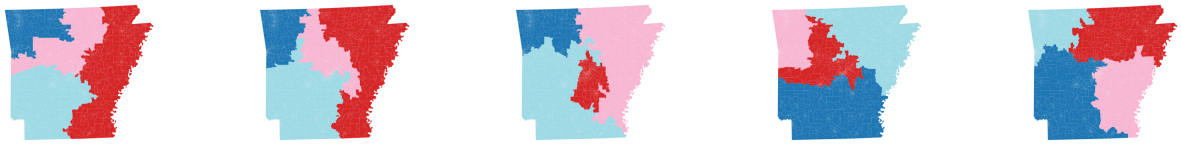




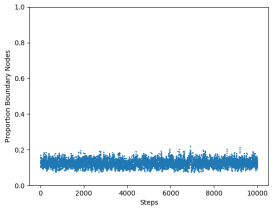
Run 2 boundary statistics

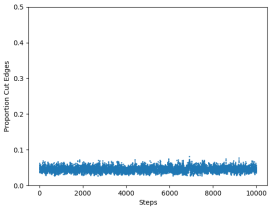


Run 3: 10K ReCom steps, shown every 2500, no compactness constraint

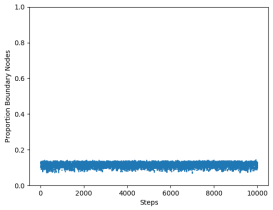


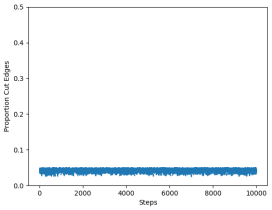
Run 4: 10K ReCom steps, shown every 2500, limited to 5% total cut edges





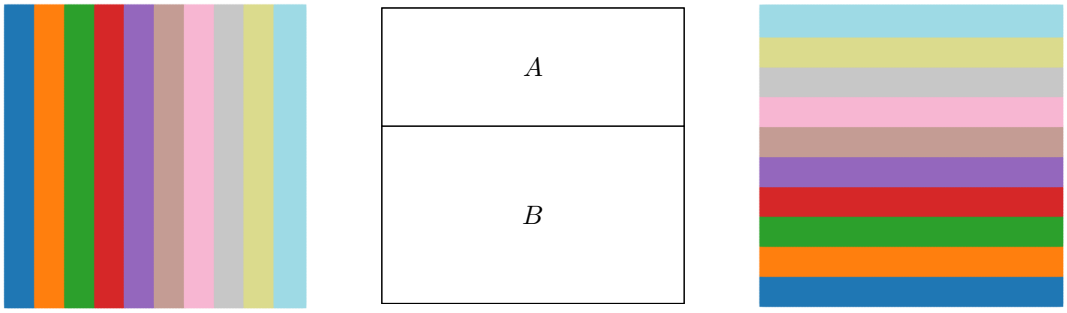
Run 3 boundary statistics

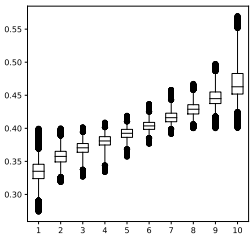




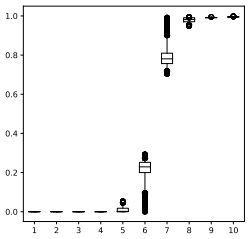
Run 4 boundary statistics

Figure 7: 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.

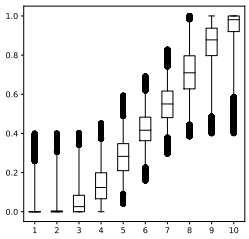




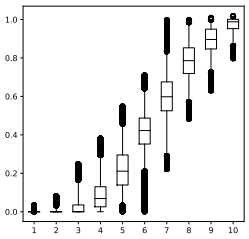
Flip from vertical
1 million steps



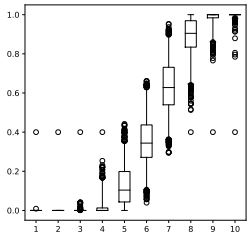
Flip from horizontal
1 million steps



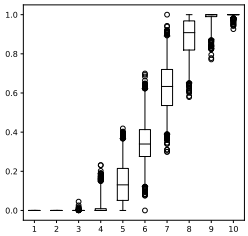
Flip from vertical
1 billion steps



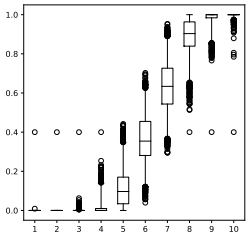
Flip from horizontal
1 billion steps



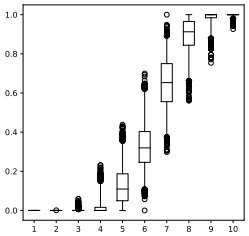
ReCom from vertical
10,000 steps



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 a table of push-forward statistics for a synthetic elections 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 the district made up of A votes across the ten districts of the plan, where the districts are ordered from smallest A share to largest A share. The boxes show the 25th-75th percentile statistics over the ensemble and the whiskers span from 1st-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. Though this multistart heuristic does not rigorously guarantee the convergence of ReCom, we can be sure that 1 billion steps is not enough for Flip.

could attempt to modify the proposal to favor districting plans with shorter boundaries. As noted above, 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. As discussed above in §5.2, there are computational obstructions to sampling proportionally to $x^{|\partial P|}$ [NDS19]. Even if we are unable to achieve a perfect sample from this distribution, however, it could be the case that this 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 9 shows that after this many steps, the perimeter statistics are comparable to the Arkansas outputs above, 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 9 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 has failed to move effectively through the space of partitions.

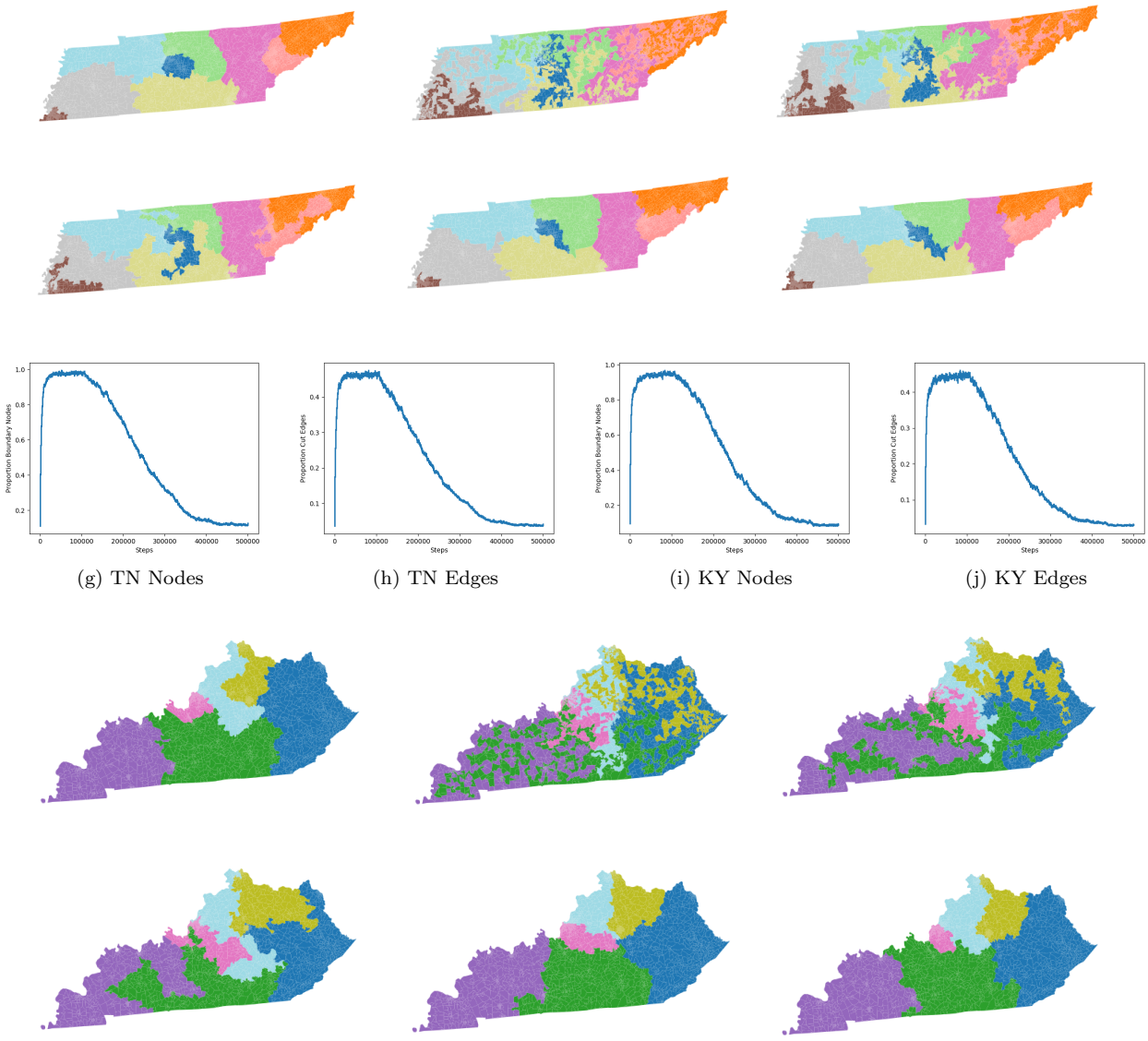
Other ensemble generation approaches such as [FHIT18] use *parallel tempering* (also known as *replica exchange*), a related technique in MCMC also aimed at accelerating its dynamics. In this algorithm, chains are run in parallel from different start points at different temperatures, then the states are occasionally exchanged between temperatures. Exactly the issues highlighted above apply to the individual chains in a parallel tempering run, making this strategy struggle to introduce meaningful new diversity.

These experiments suggest that the tendency of Flip chains to produce fractal shapes is extremely difficult to remediate and that direct attempts to do so end up impeding any 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 acceptance probabilities to favor higher or lower compactness scores, or 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 all of the redistricting rules and priorities described above [CDD+19, DDS18, DD19].[18]

7 Case study: Virginia House of Delegates

Finally, we demonstrate the assessment of convergence diagnostics and the analysis enabled by a high-quality comparator ensemble in a redistricting problem of current legal interest. For details, see [DDS18]; we include a brief discussion here, with accompanying figures in Appendix A.

[18]The weighting of a spanning tree ReCom chain is not implemented with a full (reversible) Metropolis-Hastings algorithm for the same reason that the chain is not reversible in the first place: it is not practical to compute all of the transition probabilities from a given state in this implementation. Nevertheless a weighting scheme preserves the Markov property and passes the same heuristic convergence tests as before. Several teams are now producing Recombination-like algorithms that are reversible and still fairly efficient (references to be added when available).



(g) TN Nodes (h) TN Edges (i) KY Nodes (j) KY Edges

Figure 9: 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.

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 before reaching the Supreme Court earlier this year, 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 constructed the districts in such a way that Black Voting Age Percentage (or BVAP) hit the 55% mark in twelve of the districts. 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 unnecessarily elevated in 11 of 12 districts, and that it caused dilution of the Black vote elsewhere in the state.

In Appendix A, we present various kinds of evidence, focusing on the portion of the state covered by

the invalidated districts and their neighbors. Figure 10 shows two possible attempts to assess whether the 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 10 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 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 11–13 demonstrate that for all the reasons shown in the simpler experiments above—poor compactness and failure of convergence in projection to racial or partisan statistics—individual Flip chains do not produce diverse ensembles, while ReCom chains pass tests of quality.[19] In Figure 14, 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 suggests that it is still not explained by a concern not to allow any districts to have very high BVAP.

We can go further with this analysis, looking for the classic gerrymandering pattern of "packing and cracking"—overly concentrated population in some districts and dispersed population in other districts that were near to critical mass. The top 12 districts have elevated BVAP compared to the neutral plans, and now we can locate the costs 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 does not stand alone in the study of gerrymandering, but it provides a unique ability to identify outliers against a suitable counterfactual of 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 should not be interpreted as an endorsement of using randomly sampled plans for enactment. The use of modeling to assess human judgment does not demand the excision of human judgment.

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 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, 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 any power from the conclusion that a proposed plan has very different properties. Moreover, to promote consistent and reliable analysis, it is valuable to connect the sampling method to a well-defined distribution over plans that not only has

[19]The metric used in Figure 13 is 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.

favorable qualitative properties but also can be sampled tractably. This consideration leads us to study convergence and mixing for Markov chains.

Across a range of small and large experiments with synthetic and observed data, we find that a run assembled in several days on a standard laptop produces ReCom ensembles whose measurements do not vary substantially between trials, whether re-running to vary the sample path through the state space or re-seeding at a new starting point.

Many interesting questions remain to be explored. Here is a selection of open questions and research directions.

Mathematics

- Explore the mathematical properties of spanning tree bipartitioning. For instance, what proportion of spanning trees in a grid have an edge whose complementary components have the same number of nodes?
- Experiments show that the number of cut edges appears to be normally distributed in a ReCom ensemble (see Fig 12(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.
- Prove rapid mixing of ReCom for the grid case. Even finding the edge values (in the balance constraints) for ergodicity of the chain and proving diameter bounds for the state space (when ergodicity is known) are difficult open questions.
- 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.

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 convergence in projection to summary statistics independent of seed.
- 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.
- 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. 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, we can promote public trust in expert analysis and facilitate broader engagement among the many interested parties in the redistricting process.

Acknowledgments

We are grateful to the many individuals and organizations 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 paper 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. Max Hully and Ruth Buck were deeply involved in the data preparation and software development that made the experiments possible. Finally, we acknowledge the generous support of the Prof. Amar G. Bose Research Grant and the Jonathan M. Tisch College of Civic Life.

References

[ABD+20] Hakeem Angulu, Ruth Buck, Daryl DeFord, Moon Duchin, Howard Fain, Max Hully, Maira Khan, Zach Schutzman, and Oliver York. Study of Reform Proposals for Chicago City Council. *MGGG Technical Report*, pages 1–31, 2020.

[AGR+19] Tara Abrishami, Nestor Guillen, Parker Rule, Zachary Schutzman, Justin Solomon, Thomas Weighill, and Si Wu. Geometry of graph partitions via optimal transport, 2019.

[AJK+19] Hugo A. Akitaya, Matthew D. Jones, Matias Korman, Christopher Meierfrankenfeld, Michael J. Munje, Diane L. Souvaine, Michael Thramann, and Csaba D. Tóth. Reconfiguration of connected graph partitions, 2019.

[AM10] Micah Altman and Michael McDonald. The promise and perils of computers in redistricting. *Duke J. Const. L. & Pub. Policy*, 5:69, 2010.

[BC89] Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642, 1989.

[BDGM20] Ruth Buck, Moon Duchin, Dara Gold, and JN Matthews. Community-Centered Redistricting in Lowell, Massachusetts. *MGGG Technical Report*, pages 1–8, 2020.

[BGH+17] Sachet Bangia, Christy Vaughn Graves, Gregory Herschlag, Han Sung Kang, Justin Luo, Jonathan C. Mattingly, and Robert Ravier. Redistricting: Drawing the Line. *arXiv:1704.03360 [stat]*, April 2017. arXiv: 1704.03360.

[BLV19] Austin Buchanan, Eugene Lykhovyd, and Hamidreza Validi. Imposing contiguity constraints in political districting models. *In progress*, 2019.

[BNNS19] Assaf Bar-Natan, Lorenzo Najt, and Zachary Schutzman. The gerrymandering jumble: Map projections permute districts' compactness scores. *arXiv:1905.03173*, 2019.

[BP20] Gerdus Benade and Ariel Procaccia. Abating Gerrymandering by Mandating Fairness. *Preprint*, 2020.

[BS19] Richard Barnes and Justin Solomon. Gerrymandering and compactness: Implementation flexibility and abuse. *Political Analysis (to appear)*, 2019.

[CAKY18] Vincent Cohen-Addad, Philip N. Klein, and Neal E. Young. Balanced centroidal power diagrams for redistricting. In *SIGSPATIAL/GIS*, 2018.

[CDD+19] Sophia Caldera, Daryl DeFord, Moon Duchin, Sam Gutekunst, and Cara Nix. Mathematics of nested districts: The case of Alaska. *arXiv:*, 2019.

[CDO00] Carmen Cirincione, Thomas A Darling, and Timothy G O'Rourke. Assessing South Carolina's 1990s congressional districting. *Political Geography*, 19(2):189–211, February 2000.

[CDRR20] Sarah Cannon, Moon Duchin, Dana Randall, and Parker Rule. A reversible recombination chain for graph partitions. *preprint*, 2020.

[CFMP19] Maria Chikina, Alan Frieze, Jonathan Mattingly, and Wesley Pegden. Practical tests for significance in Markov chains. *arXiv:1904.04052*, 2019.

[CFP17] Maria Chikina, Alan Frieze, and Wesley Pegden. Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences*, 114(11):2860–2864, March 2017.

[CHHM19] Daniel Carter, Gregory Herschlag, Zach Hunter, and Jonathan Mattingly. A merge-split proposal for reversible Monte Carlo Markov chain sampling of redistricting plans. *arxiv:1911.01503*, 2019.

[CJK10] Gautam Chinta, Jay Jorgenson, and Anders Karlsson. Zeta functions, heat kernels, and spectral asymptotics on degenerating families of discrete tori. *Nagoya mathematical journal*, 198:121–172, 2010.

[CKY17] Vincent Cohen-Addad, Philip N. Klein, and Neal E. Young. Balanced power diagrams for redistricting. *CoRR*, abs/1710.03358, 2017.

[CL16] Wendy Cho and Yan Liu. Toward a Talismanic Redistricting Tool: A Computational Method for Identifying Extreme Redistricting Plans. *Election Law Journal: Rules, Politics, and Policy*, 15, November 2016.

[CR13] Jowei Chen and Jonathan Rodden. Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures. *Quarterly Journal of Political Science*, 8(3):239–269, June 2013.

[CR16] Jowei Chen and Jonathan Rodden. The loser's bonus: Political geography and minority party representation. 2016.

[CRS19] Wendy K. Tam Cho and Simon Rubinstein-Salzedo. Understanding significance tests from a non-mixing Markov chain for partisan gerrymandering claims. *Statistics and Public Policy*, 6(1):44–49, 2019.

[DD19] Daryl DeFord and Moon Duchin. Redistricting reform in Virginia: Districting criteria in context. *Virginia Policy Review*, 12(2):120–146, 2019.

[DDS18] Daryl DeFord, Moon Duchin, and Justin Solomon. Comparison of districting plans for the Virginia House of Delegates. *MGGG Technical Report*, pages 1–26, 2018.

[DDS19a] Daryl DeFord, Moon Duchin, and Justin Solomon. A Computational Approach to Measuring Vote Elasticity and Competitiveness. *Preprint*, pages 1–30, 2019.

[DDS19b] Daryl DeFord, Moon Duchin, and Justin Solomon. Replication code. *GitHub repository*, 2019.

[Dia09] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):179–205, 2009.

[DLSS19] Daryl DeFord, Hugo Lavenant, Zachary Schutzman, and Justin Solomon. Total variation isoperimetric profiles. *SIAM Journal on Applied of Algebra and Geometry*, 3:585–613, 2019.

[DT18] Moon Duchin and Bridget Tenner. Discrete geometry for electoral geography. *arXiv:1808.05860*, 2018.

[FHIT18] Benjamin Fifield, Michael Higgins, Kosuke Imai, and Alexander Tarr. A new automated redistricting simulator using Markov Chain Monte Carlo. *Princeton University Working Paper*, pages 1–55, May 2018.

[FIKK19] Benjamin Fifield, Kosuke Imai, Jun Kawahara, and Christophe Kenny. The essential role of empirical validation in legislative redistricting simulation. *preprint*, 2019.

[FJH11] Roland G Fryer Jr and Richard Holden. Measuring the compactness of political districting plans. *The Journal of Law and Economics*, 54(3):493–535, 2011.

[Gey11] Charles J. Geyer. Introduction to Markov chain Monte Carlo. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 3–48. CRC Press, Boca Raton, FL, 2011.

[Hås06] Johan Håstad. The square lattice shuffle. *Random Structures & Algorithms*, 29(4):466–474, 2006.

[HKL+18] Gregory Herschlag, Han Sung Kang, Justin Luo, Christy Vaughn Graves, Sachet Bangia, Robert Ravier, and Jonathan C. Mattingly. Quantifying Gerrymandering in North Carolina. *arXiv:1801.03783 [physics, stat]*, January 2018. arXiv: 1801.03783.

[HRM17] Gregory Herschlag, Robert Ravier, and Jonathan C. Mattingly. Evaluating Partisan Gerrymandering in Wisconsin. *arXiv:1709.01596 [physics, stat]*, September 2017. arXiv: 1709.01596.

[HSVD10] J. Gerald Hebert, Paul M. Smith, Martina E. Vandenburg, and Michael B. DeSanctis. *The realists' guide to redistricting: avoiding the legal pitfalls, 2nd edition*. American Bar Association, 2010.

[Ins18] Voting Rights Data Institute. GerryChain. *GitHub repository*, 2018.

[Jin17] Hai Jin. Spatial optimization methods and system for redistricting problems. 2017.

[JW18] Matt Jacobs and Olivia Walch. A partial differential equations approach to defeating partisan gerrymandering. *arXiv:1806.07725*, 2018.

[Ken00] Richard Kenyon. The asymptotic determinant of the discrete Laplacian. *Acta Mathematica*, 185(2):239–286, 2000.

[Kim11] Myung Jin Kim. *Optimization approaches to political redistricting problems*. PhD thesis, The Ohio State University, 2011.

[LCW16] Yan Y. Liu, Wendy K. Tam Cho, and Shaowen Wang. PEAR: A massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm and Evolutionary Computation*, 30:78–92, October 2016.

[LF19] Harry A. Levin and Sorelle A. Friedler. Automated congressional redistricting. In *JEAL*, 2019.

[Mat17] Jonathan Mattingly. Expert report of jonathan mattingly,, 2017.

[MM18] Daniel B. Magleby and Daniel B. Mosesson. A New Approach for Developing Neutral Redistricting Plans. *Political Analysis*, 26(2):147–167, April 2018.

[MSVRR18] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS national historical geographic information system: Version 13.0 [database], 2018.

[Nag65] Stuart S. Nagel. Simplified Bipartisan Computer Redistricting. *Stanford Law Review*, 17(5):863–899, 1965.

[NdC11] Mariá C.V. Nascimento and André C.P.L.F. de Carvalho. Spectral methods for graph clustering – a survey. *European Journal of Operational Research*, 211(2):221 – 231, 2011.

[NDS19] Lorenzo Najt, Daryl DeFord, and Justin Solomon. Complexity of sampling connected graph partitions. *arXiv:1908.08881*, 2019.

[Peg17a] Wesley Pegden. Expert report of wesley pegden in league of women voters of pennsylvania v. commonwealth of pennsylvania, 2017.

[Peg17b] Wesley Pegden. Pennsylvania's congressional districting is an outlier: Expert report. *League of Women Voters vs. Pennsylvania General Assembly*, November 2017.

[RSS13] Federica Ricca, Andrea Scozzari, and Bruno Simeone. Political districting: from classical models to recent approaches. *Annals of Operations Research*, 204(1):271–299, 2013.

[Sax18] James Saxon. Spatial constraints on gerrymandering: A practical comparison of methods. 2018.

[Sch07] Satu Elisa Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, August 2007.

[Tas11] Attila Tasnádi. The political districting problem: A survey. *Society and Economy*, 33(3):543–554, 2011.

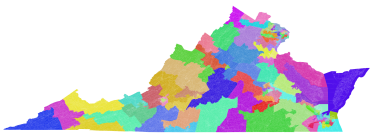
[Tem72] H. N. V. Temperley. The enumeration of graphs on large periodic lattices. In *Combinatorics (Proc. Conf. Combinatorial Math., Math. Inst., Oxford, 1972)*, pages 285–294, 1972.

[WH63] James B Weaver and Sidney W Hess. A procedure for nonpartisan districting: Development of computer techiques. *Yale LJ*, 73:288, 1963.

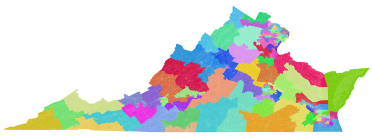
[Wil96] David Bruce Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. ACM.

[WR20] Thomas Weighill and Jonathan Rodden. title. In Duchin et al., editor, *Political Geography*, chapter 12, page pages. publisher, 2020.

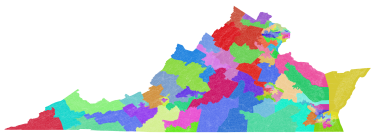
A Plots for Virginia Case Study



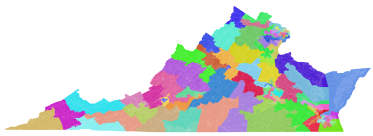
(a) Enacted



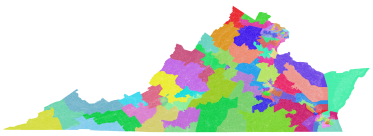
(b) Dem



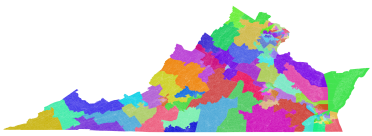
(c) GOP1



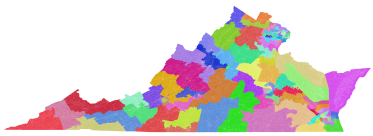
(d) GOP2



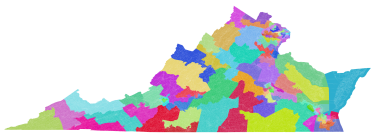
(e) GOP3



(f) NAACP



(g) Princeton



(h) Remedial

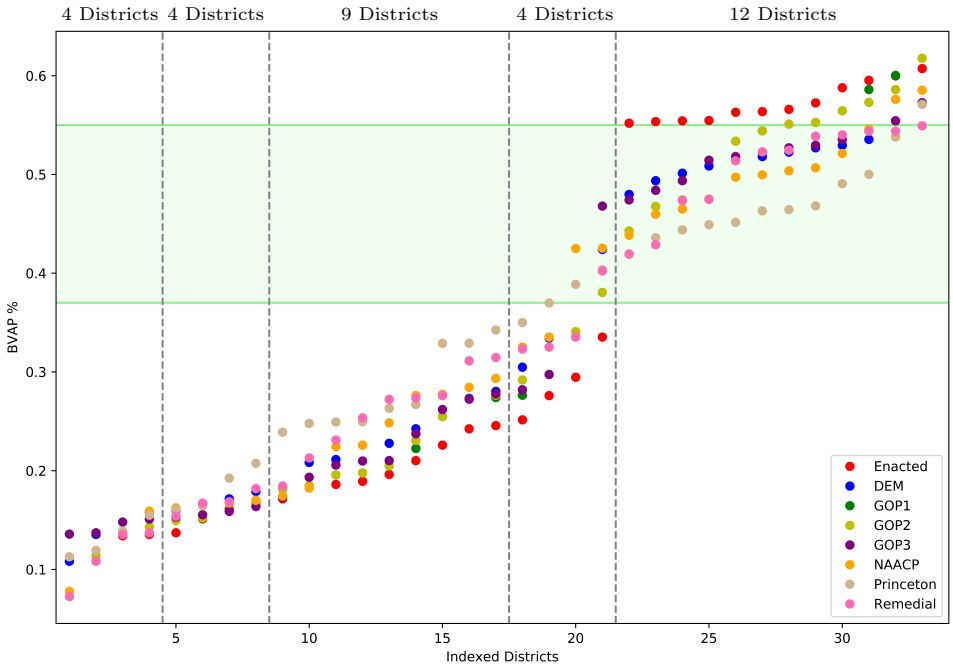


Figure 10: Eight proposed House of Delegates plans as described in the text. The boxplot shows the Black Voting Age Population (BVAP) in the 33 districts affected by the court ruling, ordered from lowest to highest BVAP in each plan. The 2011 enacted plan jumps the key 37-55% BVAP range entirely, but the collection of other plans makes it difficult to tell how many more 37-55% BVAP plans might be expected or possible.

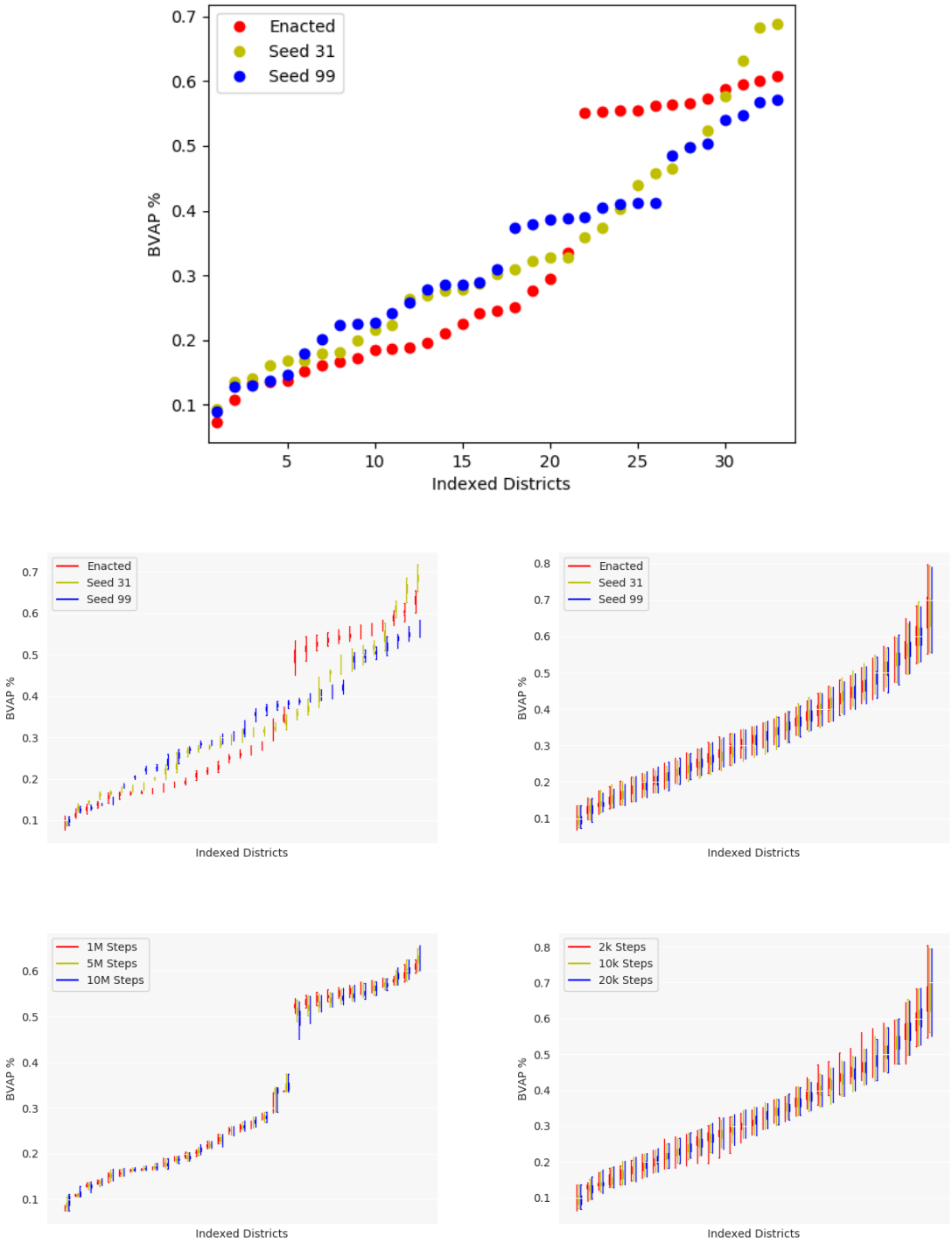
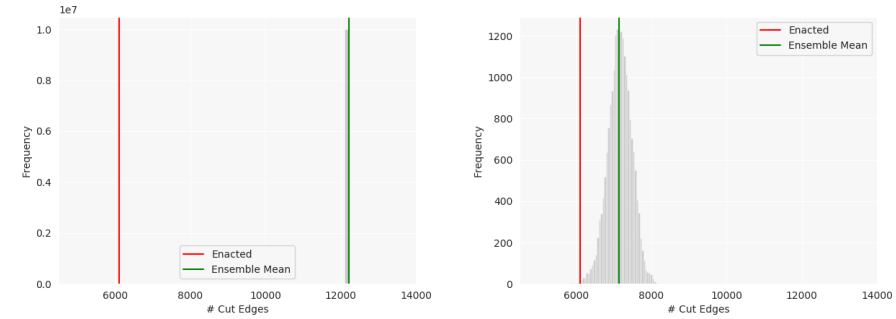
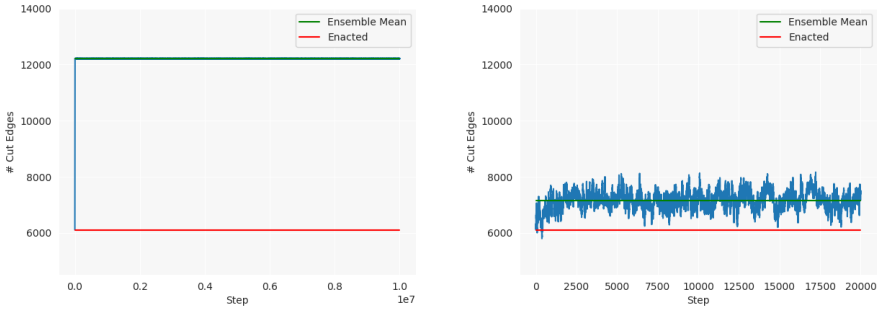


Figure 11: Convergence heuristics. The BVAP levels in the enacted plan are compared to two synthetically generated seed plans. 10 million steps are not enough to mitigate the dependence on the starting point in a Flip run. By contrast, 20,000 steps overcomes the dependence on starting point for a recombination 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.

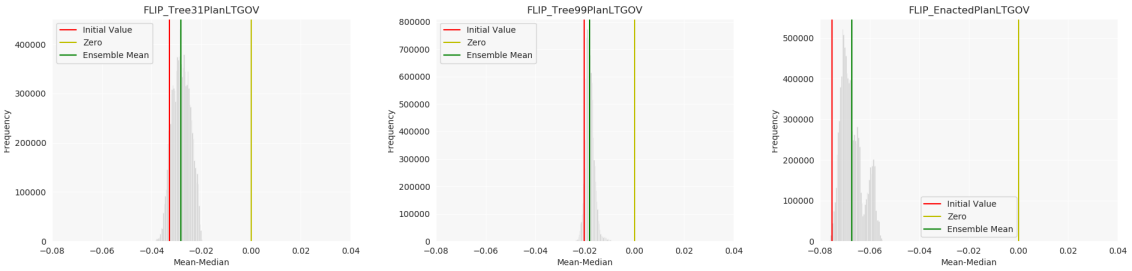


(a) Flip (b) ReCom

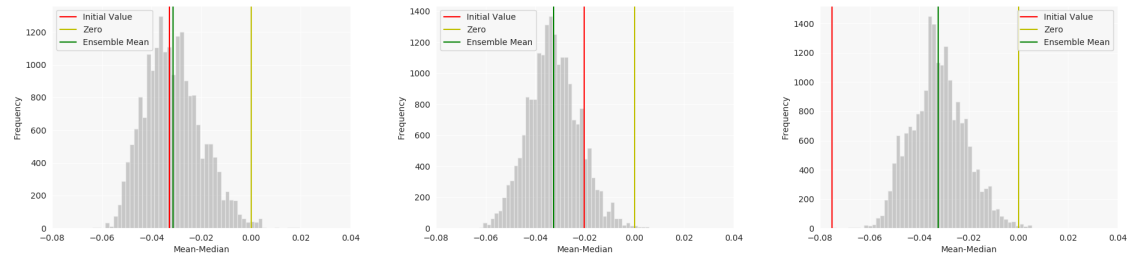


(c) Flip (d) ReCom

Figure 12: 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 13: Projection to partisan statistics. Mean-median (partisan symmetry) scores, illustrating dependence of Flip ensembles on starting point after one million steps.

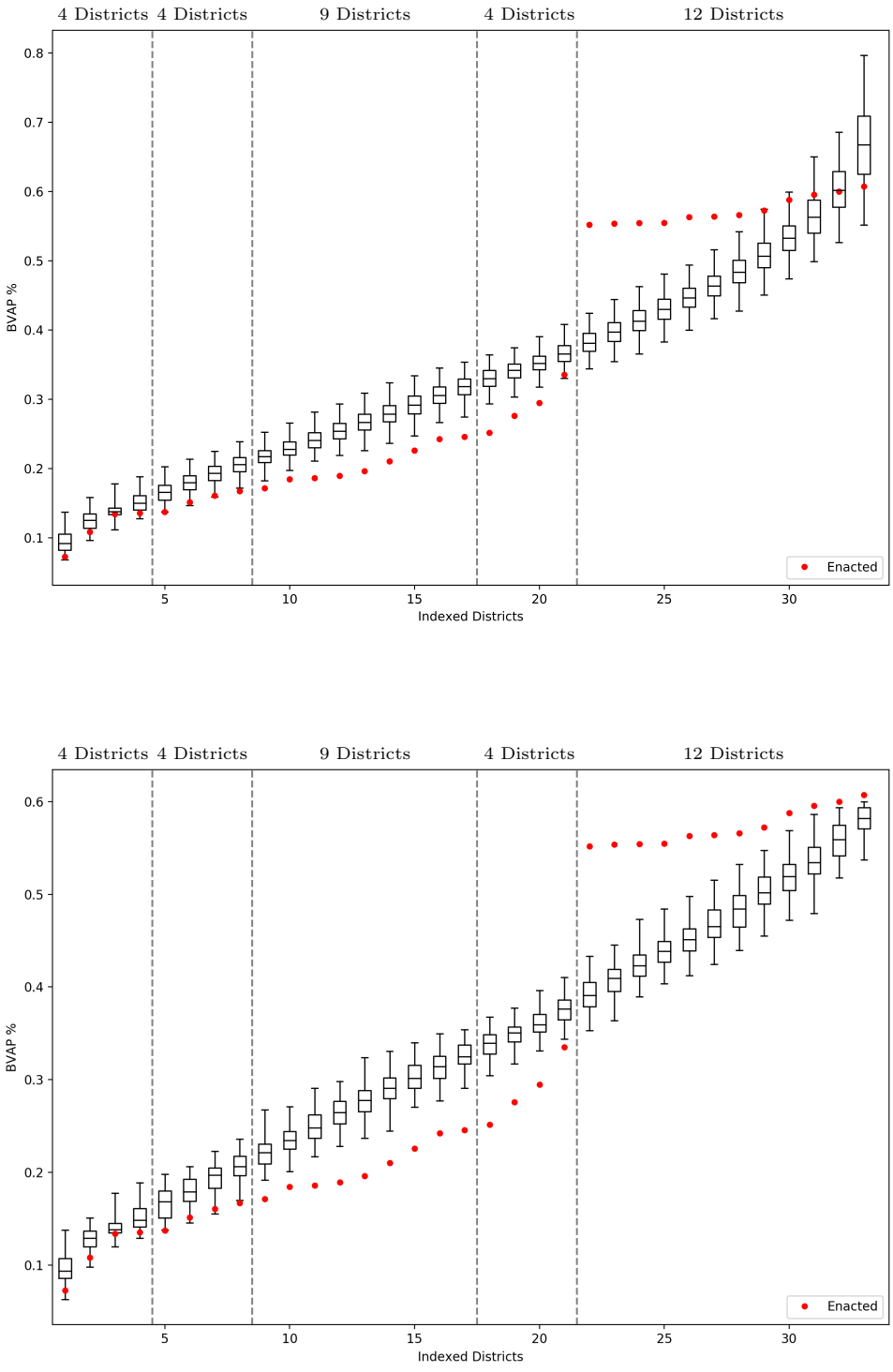


Figure 14: Ensemble analysis. The BVAP levels in the Enacted plan can now be compared to an ensemble of population-balanced, compact plans that hold the state's demographics and geography constant. Top: full ReCom ensemble. Bottom: same ensemble, winnowed to $\leq$60% BVAP.