Political districting to minimize county splits

Maral Shahmizad, Austin Buchanan
School of Industrial Engineering & Management, Oklahoma State University, Stillwater, OK 74078,
maral.shahmizad@okstate.edu buchanan@okstate.edu

When partitioning a state into political districts, a common criterion is that political subdivisions like counties should not be split across multiple districts. This criterion is encoded into most state constitutions and is sometimes enforced quite strictly by the courts. However, map drawers, courts, and the public typically do not know what amount of splitting is truly necessary, even to satisfy basic criteria like contiguity and population balance. In this paper, we provide answers for all congressional, state senate, and state house districts in the USA using 2020 census data. Our approach is based on integer programming. The associated codes and experimental results are publicly available on GitHub.

1. Introduction

When partitioning a state into political districts, traditional redistricting principles dictate that districts should have nearly equal populations, be contiguous on the map, and have reasonably compact shapes. Another important criterion is that political subdivisions such as counties, cities, and towns should not be split across multiple districts. This criterion is encoded into most state constitutions (NCSL 2021) and is sometimes enforced quite strictly. Prominent examples include Texas's *County Line Rule* and North Carolina's *Whole County Provision* which apply to legislative districting (i.e., for state house and/or state senate), as well as Iowa's and West Virginia's insistence on keeping all counties whole in their congressional districting plans.

When redistricting laws are violated, courts can intervene. For example, in 2018 the Pennsylvania Supreme Court ordered new maps to be drawn after finding their existing congressional districting plan to be an unconstitutional partisan gerrymander that favored Republicans (*League of Women*

Voters of Pennsylvania v. the Commonwealth of Pennsylvania). The court remarked that the enacted plan divided 28 of the 67 counties and further specified how many times each individual county was split (with 37 county splits in total). The court then ordered the state legislature to adopt a new map that does "not divide any county, city, incorporated town, borough, township, or ward, except where necessary to ensure equality of population." When the legislature failed to do so, the court adopted its own map, with assistance from Stanford Law Professor Nate Persily. The court pointed out that its remedial plan "splits only 13 counties," with four counties split twice and nine counties split once, giving a total of 17 county splits. The overturned and court-mandated maps are shown in Figure 1. Observe the drastically different splitting patterns around Philadelphia and Pittsburgh in the southeastern and southwestern portions of the state, respectively.

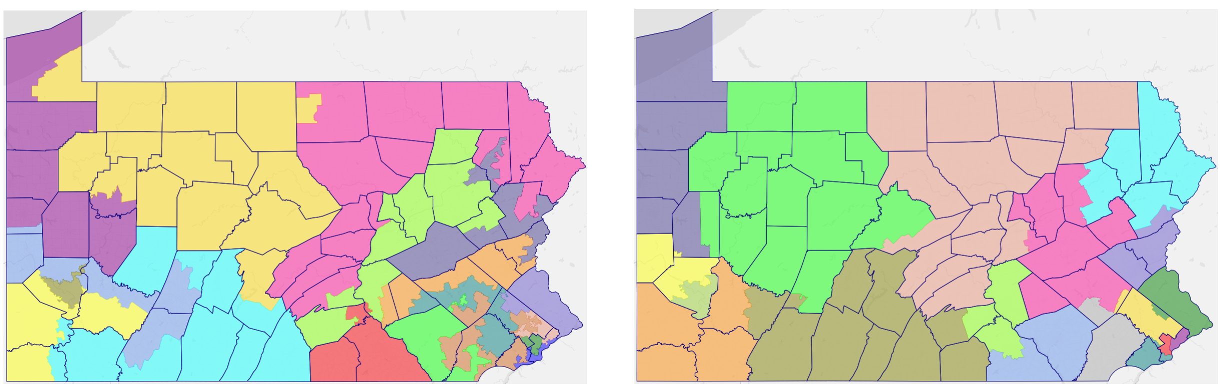

Figure 1 Pennsylvania's overturned and court-mandated congressional districts.

In another example, New York's congressional districts and state senate districts were overturned in 2022, with the state's highest court ruling that they were Democratic partisan gerrymanders in violation of the New York Constitution (*Harkenrider v. Hochul*). The court appointed Jonathan Cervas, a political scientist and Post-Doctoral Fellow at Carnegie Mellon University, as Special Master to redraw the districts. Cervas paid special attention to preserving political subdivisions, reporting that the old congressional map split 34 counties a total of 56 times, while the new map split 16 counties a total of 26 times (Cervas 2022). In fact, Cervas went on to write that "while I was quite successful in limiting the number of counties and cities that were split, some splits are simply inevitable. . . I can assure you that if yours was split it was not because of any kind of animus but was essentially due to the *mathematical necessity of splitting some units*." (Emphasis added.) The two maps are shown in Figure 2.

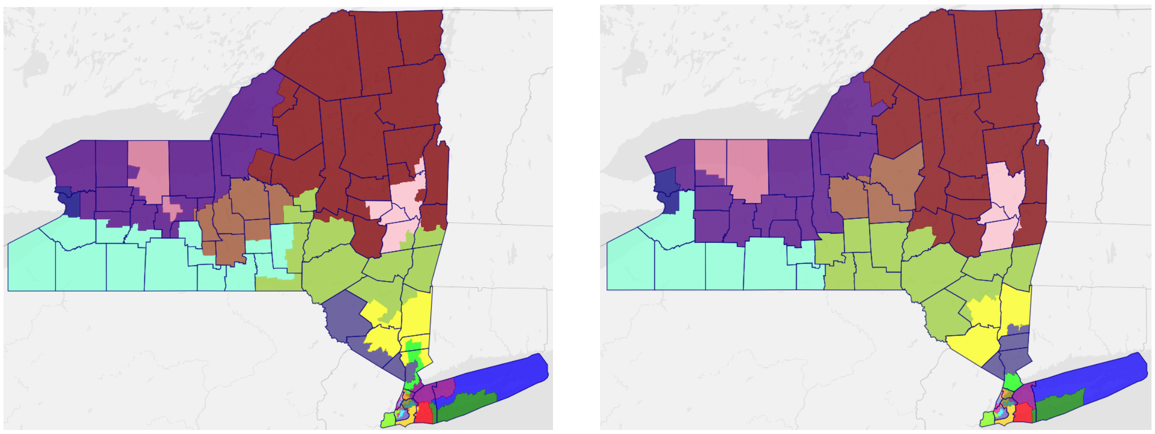

Figure 2 New York's overturned and court-mandated congressional districts.

When is a county split truly necessary? Sometimes an explanation is straightforward. For example, each of New York's congressional districts must have a population near 776,971. Meanwhile, New York County (Manhattan) is much more populous and cannot be kept whole. Further, it must be divided across *at least three* districts (i.e., split at least twice), because its population sufficiently exceeds $2 \times 776,971$. Indeed, the court-mandated plan splits it twice.

However, this simple reasoning cannot be used to justify all necessary county splits. For example, the court-mandated plan splits Orleans County in the northwest of the state, despite it having a small population of roughly 40,000. It is not necessary to split Orleans County *per se*, but a split is necessary somewhere in its vicinity. The reason is that neighboring Monroe County has a population sufficiently below 776,971, and adding any one of Monroe's neighbors to it would make its district's population too large. In a synthetic, but illustrative example, consider a hypothetical state with three counties, each having 100 people, arranged in a triangle. When dividing this state into two 150-person districts, no *specific* county must be split, nevertheless we cannot keep all counties whole. Thus, it may be hopeless to justify any *particular* county split. A more reasonable goal is to justify the *total number of county splits*. So, we may ask, what is the minimum number of county splits possible in a contiguous and population-balanced plan? While a few generic answers to this optimization problem have been proposed in the literature, we find none of them to be wholly satisfactory, as discussed below.

Redistricting folklore states that when drawing k districts, it suffices to have just $k-1$ county splits, i.e., $k-1$ is an upper bound on the minimum number of county splits. Some researchers have further said that $k-1$ county splits is the *right* number, i.e., that $k-1$ is "highly probable" to be a lower bound, particularly if district populations are not permitted to differ by more than one person (Nagle 2022). This $k-1$ figure is stated on popular websites like Dave's Redistricting App (DRA 2023) and can be found in research papers (Cervas and Grofman 2020, McCartan and Imai 2023) and expert testimony (Cervas 2022). In fact, in the 2020 round of redistricting, 14 states drew their congressional districts to have $k-1$ county splits (Nagle 2022). However, we will see that $k-1$ is neither a lower nor upper bound on the minimum number of county splits.

Another solution to the minimum county splits problem was recently proposed by Carter et al. (2020). They state that the minimum number of county splits equals the number of districts minus the *maximum number of county clusters*, "with the exception of rare circumstances which impact the optimal districting." To understand this claim, let us introduce some notation. Denote by C the set of counties, and let G_C be the county-level contiguity graph, which has a vertex for each county and two vertices are connected by an edge if the associated counties are adjacent on the map. For population balance, we suppose that each district should have a population of at least L and at most U. The population of county c is denoted by p_c, and the population of a subset of counties S is indicated by the shorthand $p(S) := \sum_{c \in S} p_c$.

A county clustering decomposes a districting instance into a collection of miniature districting instances. Each miniature districting instance is defined by a subset of counties and a positive integer *size* indicating how many districts to build from. A formal definition follows.

DEFINITION 1. A *county clustering* is a partition $(C_1, C_2, \ldots, C_q)$ of the counties along with associated cluster sizes $(k_1, k_2, \ldots, k_q)$ such that:

1. the cluster sizes $(k_1, k_2, \ldots, k_q)$ are positive integers that sum to k;
2. each cluster C_j is contiguous, i.e., induces a connected subgraph of G_C;
3. each cluster C_j satisfies population balance, i.e., $Lk_j \leq p(C_j) \leq Uk_j$.

The *maximum county clustering problem* is an optimization problem that seeks a county clustering with a maximum number of clusters (i.e., a *maximum county clustering*).

To illustrate, consider the fictional state of Splitigan in Figure 3. It has five rectangular counties (Alpha, Beta, Gamma, Delta, Epsilon) which are to be divided into four districts of equal population. One possible county clustering is the trivial clustering in which all counties are placed in one cluster C_1 of size $k_1 = 4$ and population $p(C_1) = 16$. Alternatively, we may place Gamma County in one cluster C_1 of size 1 and population $p(C_1) = 4$, and the remaining counties in another cluster C_2 of size 3 and population $p(C_2) = 12$; this is *maximum*, as no county clustering has more clusters.

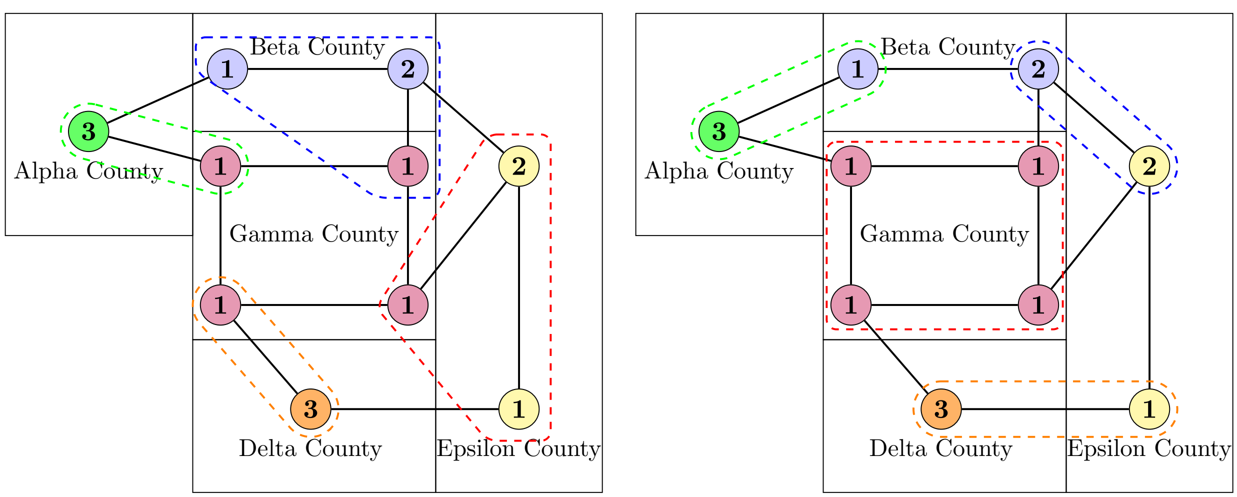

Figure 3 Two possible districting plans for the fictional state of Splitigan. Rectangles represent counties, nodes represent tracts, and their populations are given inside the nodes.

Figure 3 also shows two possible districting plans for Splitigan in dashed lines. The plan on the left pairs each of the exterior counties (Alpha, Beta, Delta, and Epsilon) with one tract from Gamma County in the center. Thus, Gamma County is divided across four districts (i.e., it is split three times), and the others are split zero times, for a total of three county splits. Meanwhile, the plan on the right keeps Alpha, Gamma, and Delta Counties whole, and splits Beta and Epsilon Counties once each, for a total of two county splits. This is the minimum number of county splits possible. So, in this case, Carter et al.'s theorem *does* hold: the minimum number of county splits (2) indeed equals the number of districts (4) minus the maximum number of county clusters (2).

Intuitively, why should Carter et al.'s theorem hold? The idea is as follows. Identify a maximum number of county clusters $(C_1, C_2, \dots, C_q)$ with associated cluster sizes $(k_1, k_2, \dots, k_q)$, and consider each county cluster C_j as a miniature districting instance. If redistricting folklore holds, then we can divide the cluster C_j into k_j districts using $k_j - 1$ county splits. Summing over q clusters gives $(k_1 - 1) + (k_2 - 1) + \cdots + (k_q - 1) = k - q$ county splits. In particular, this decomposition works for Splitigan; the first cluster $C_1 = \{\gamma\}$ already has size $k_1 = 1$ giving a district with $k_1 - 1 = 0$ county splits, while the second cluster $C_2 = \{\alpha, \beta, \delta, \varepsilon\}$ can be divided into $k_2 = 3$ districts using $k_2 - 1 = 2$ county splits.

Unfortunately, the folklore $k - 1$ result does not always hold. In a footnote of their appendix, Carter et al. (2020) give a counterexample in which one of the counties has zero population. In a more realistic example, consider the county-level graph in Figure 4 with the same claw topology. Here, the population of each county is given inside its node, and the task is to build $k = 2$ districts, each with population between $L = 95$ and $U = 105$ in order to satisfy a $\pm 5\%$ deviation. Observe that at least one of the leaf counties must be split, since otherwise all three would be kept whole, forcing two of them to be in the same (overpopulated) district. Each district can take at most 55 people from this leaf county, so both districts will need to extend into the hub county, splitting it as well. So, at least two counties must be split, and this is more than $k - 1$. *This is true irrespective of how counties are made up of census tracts or blocks.* Later, we will extend this example to show that arbitrarily many county splits might be required, many more than $k - 1$.

Fortunately, half of Carter et al.'s theorem holds without the "except in rare circumstances" caveat. That is, the minimum number of county splits is always *at least* the number of districts minus the maximum number of county clusters. For example, in the claw instance, the maximum

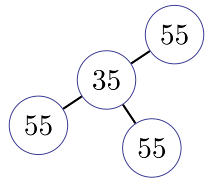

Figure 4 A *claw* instance that requires more than $k-1$ county splits.

number of county clusters is one, and indeed the minimum number of county splits is at least $k-1=1$. This relationship between a minimization problem and a maximization problem is analogous to optimization duality and motivates us to define *weak split duality* and *strong split duality*.

Definition 2. A districting instance exhibits *weak split duality* if the minimum number of county splits is at least the number of districts minus the maximum number of county clusters. It exhibits *strong split duality* if equality also holds.

For example, Splitigan satisfies both weak and strong split duality, while the claw instance only satisfies weak split duality.

In light of this discussion, we seek to answer the following questions. For 2020 census data, what is the minimum number of county splits required in a contiguous and population-balanced plan for each state and for each type of districting (congressional, state senate, state house)? How does this compare to the enacted plans? How often does strong split duality hold in practice? How should we solve the maximum county clustering problem and the minimum county splits problem? Is it true, as Nagle (2022) states, that forcing districts to satisfy a 1-person deviation makes it "highly probable that the minimum number of county splits is uniquely given as the number of districts minus one"? Or, as Autry et al. (2021) posit that "it is reasonable to assume that there is no subset of counties that perfectly can accommodate a subset of the congressional districts... [which] may be used to demonstrate that $k-1$ county splits is optimal"? Further, can we answer these questions in ways that are transparent and understandable by the public and the courts?

Outline. Section 2 provides background on political districting criteria, algorithms, terminology, and notation. Section 3 proves weak split duality and shows that the split duality gap can be arbitrarily large. However, we will later see that every one of our test instances, across all states and district types, satisfies strong split duality. Section 4 proposes an approach for solving the minimum county splits problem. The first step, *Cluster*, finds a maximum number of county clusters $(C_1, C_2, \dots, C_q)$ with associated sizes $(k_1, k_2, \dots, k_q)$ using integer programming techniques. By weak split duality, this provides a lower bound $k - q$ on the minimum number of county splits. The second and third steps, *Sketch* and *Detail*, use integer programming techniques to find a detailed districting plan for each cluster C_j with $k_j - 1$ county splits. Ultimately, this provides a districting plan for the entire state with $(k_1 - 1) + (k_2 - 1) + \cdots + (k_q - 1) = k - q$ county splits, matching the bound from weak split duality and thus proving optimality. Section 5 applies the approach to every state and district type. We find that strong split duality holds on these real-life instances, empirically supporting a claim of Carter et al. (2020). However, contrary to Nagle (2022) and Autry et al. (2021), most real-life districting instances, including *all* of our instances with more than 30 counties, admit nontrivial county clusterings under 1-person deviation. We conclude in Section 6.

2. Background and Literature Review

In the USA, political districts are redrawn every ten years, after the census has been conducted. The resulting populations determine how the 435 seats in the US House of Representatives will be divided amongst the states, a process called reapportionment (Balinski and Young 2010). Then, each state must be divided into the appropriate number of *congressional* districts. After the 2020 census, the least-populous states received one seat, while the most-populous state (California) received 52 seats. Additionally, 49 state governments have two legislative bodies (upper and lower), often referred to as their state senate and state house (or general assembly), that require the drawing of *legislative* districts for their elections. The remaining state, Nebraska, only has a state senate. In these state legislatures, the number of seats and districts varies, with between 13 and 67 state senate districts and between 30 and 204 state house districts (Ballotpedia 2023). Unlike congressional districts, which elect one person, some legislative districts are multi-member (e.g., each of Arizona's 30 House districts elects two representatives). So, from now on, we will typically refer to the number of districts, and not the number of seats.

Congressional and legislative districts must satisfy a number of state and federal laws. For example, after the "one-person, one-vote" revolution of the 1960s, districts must have roughly equal populations, generally exhibiting less than a 1% population deviation for congressional districts and less than 10% for legislative districts (Hebert et al. 2010, Davis et al. 2019). These *population balance* constraints are not numerically specified in federal law, but instead have emerged from federal court cases as a consequence of Article I, Section 2 of the US Constitution and the Equal Protection Clause of the 14th Amendment, beginning with landmark Supreme Court cases like *Baker v. Carr* (1962), *Wesberry v. Sanders* (1964), and *Reynolds v. Sims* (1964). The Voting Rights Act of 1965 and the Equal Protection Clause of the 14th Amendment place federal restrictions on racial gerrymandering (Hebert et al. 2010, Davis et al. 2019), see the landmark Supreme Court cases *Thornburg v. Gingles* (1986) and *Shaw v. Reno* (1993). The US Supreme Court decided in *Rucho v. Common Cause* (2019) and *Lamone v. Benisek* (2019) that partisan gerrymandering, while undemocratic, lies outside the scope of the federal courts.

Essentially all other districting laws are state laws and thus vary across the country. They often codify traditional redistricting principles like contiguity, compactness, and the preservation of counties and other political subdivisions. In recent years, some states have passed redistricting reforms, including the use of independent redistricting commissions to draw maps (instead of state legislatures) and the requirement to abide by additional criteria like promoting competitiveness, proportionality, or partisan fairness (NCSL 2021).

State law sometimes conflicts with federal law. Notably, many states have strict rules about the preservation of political subdivisions that, if enforced, would violate federal case law regarding population balance. For example, Wake County in North Carolina had a population in 2020 that exceeded one million people, making it significantly larger than that of an ideal congressional district (745,671), state senate district (208,788), or state house district (86,995). Meanwhile, the *Whole County Provision* of North Carolina's constitution (Article II, Sections 3 & 5) states that "No county shall be divided in the formation of a [legislative] district." These legal contradictions have led to court battles on how to interpret these state laws when they conflict with federal law, see *Stephenson v. Bartlett* (2002) which went before North Carolina's Supreme Court.

Many reasons have been proposed to keep political subdivisions whole, or to minimize the extent to which they are split. We have already seen that many states require the preservation of political subdivisions by law. Motivations for these laws include: simplifying election administration, making it easier for voters to know their representative, empowering voters to elect candidates that will represent their community's interests, and obstructing the most extreme partisan gerrymandering efforts (Gladkova et al. 2019, Wachspress and Adler 2021).

After deciding that political subdivisions should be preserved, one of the next questions is how to *quantify* splitting, prompting several different splitting scores (Becker and Gold 2022). Few state laws actually specify which ones to use, but the most common scores are the number of split counties and the number of county splits. They are used by Dave's Redistricting (DRA 2023), the Redistricting Report Card of the Princeton Gerrymandering Project (Princeton Gerrymandering Project 2023), and Harvard's ALARM Project (McCartan et al. 2022a,b). Several more complicated scores exist in the literature (Gladkova et al. 2019, Wachspress and Adler 2021, Becker and Gold 2022), like pieces or traversals (Carter et al. 2020), naked boundary length, Pennsylvania fouls, effective splits, split pairs, and various entropy-based and population-weighted scores. These scores can guide redistricting officials when drawing maps, can be used to challenge bad maps in court, and can be used to mobilize members of a political subdivision to speak up when they are being divided unnecessarily (Wachspress and Adler 2021). For more information about redistricting, we refer the reader to the books and surveys of Grofman (1985), Hebert et al. (2010), Bullock III (2010), Levitt (2010), Davis et al. (2019), Duchin and Walch (2022).

2.1. Computational Techniques for Redistricting

Districting problems are generally NP-hard (Altman 1997), in part because they can be used to express instances of the NP-hard Partition problem via the population balance constraints. In the presence of contiguity constraints, districting remains hard in planar bipartite graphs, even when each node has one person and each district should have three people (Dyer and Frieze 1985). This has led researchers to develop greedy construction heuristics (Vickrey 1961, Kim 2019), local search heuristics (King et al. 2012, 2015, 2018), and those based on metaheuristic frameworks like tabu search, genetic algorithms, and simulated annealing (Bozkaya et al. 2003, Ricca and Simeone 2008, Altman et al. 2011, Guo and Jin 2011, Liu et al. 2016, Olson 2022, Gutiérrez-Andrade et al. 2019). There are also many different generalizations of Voronoi diagrams (Cohen-Addad et al. 2018, Levin and Friedler 2019, Miller 2007, Ricca et al. 2008, Svec et al. 2007). For more on heuristic approaches to districting, we refer the reader to the surveys of Ricca et al. (2013), Goderbauer and Winandy (2018), Becker and Solomon (2022).

In the past decade, a new area of computational districting has arisen called *ensemble analysis*, where the idea is to generate a huge collection of redistricting plans so that proposed or enacted plans can be placed in the context of the distribution of possible plans. Many early approaches were based on a Markov chain Monte Carlo framework, where an algorithm repeatedly and randomly moved from a districting plan to a neighboring plan (Adler and Wang 2019, Cho and Liu 2018, DeFord and Duchin 2019, Fifield et al. 2015), often based on search neighborhoods like "flip" where a single node on the boundary of two districts is flipped to the other district. One criticism of this approach is that the flip neighborhood makes small changes and may require a huge number of iterations to adequately explore the distribution of possible plans. This led to larger search neighborhoods like recombination (DeFord et al. 2021) which repeatedly merges two (or more) districts, draws a random spanning tree on their nodes, and splits the tree into districts by removing one (or more) edges. More recent approaches (McCartan and Imai 2023, Autry et al. 2021) aim to generate collections of districting plans drawn randomly from an explicit target distribution, making the approaches more credible than previous approaches when labeling proposed or enacted plans as (gerrymandered) outliers. Many ensemble approaches have had trouble dealing with (hard) constraints on county splitting and (if at all) try to capture them with "soft" constraints (penalties) with poor results. For example, Herschlag et al. (2020) find that their ensemble had a median of

34 split counties when applied to North Carolina congressional districting, much larger than the 2016 remedial plan (13 split counties) and 2019 plan (12 split counties).

The desire to limit county splits has prompted some of the most recent work in ensemble approaches (Autry et al. 2021, McCartan and Imai 2023). For example, the sequential Monte Carlo algorithm of McCartan and Imai (2023) is designed to produce at most $k-1$ county splits, where k is the number of districts to draw. It works by repeatedly drawing a random spanning tree in (a subgraph of) graph G and then carving off a district by deleting one of the tree's edges. If there is no suitable edge to delete that would carve off a population-balanced district, a new random spanning tree is generated. Each iteration of this procedure introduces at most one county split. So, after $k-1$ iterations, we have carved off $k-1$ districts and introduced $k-1$ county splits, with what remains constituting the final district. Presumably, this procedure would fail on instances that require k or more county splits, as it would eventually get stuck in an infinite loop drawing new spanning trees over and over in search of an edge to delete to carve off the next district.

In the integer programming literature, there are many optimization models for political districting and graph partitioning (Ricca et al. 2013). Many of them are inspired by the model of Hess et al. (1965), which is a constrained k-median model with variables of the form x_{ij} that equal one when geographic unit i is assigned to (the district centered at) geographic unit j; for example, see Oehrlein and Haunert (2017), Alès and Knippel (2020), Swamy et al. (2022). Thus, the number of variables in these models grows (at least) quadratically in the number n of geographic units. When the geographic units are counties, these models can often be solved directly, but may require sophisticated variable fixing procedures when dealing with more granular units (Validi et al. 2022). Some other models are based on labeling or assignment variables of the form x_{ij} that equal one when geographic unit i is assigned to district number $j \in [k]$; for example, see Ferreira et al. (1996), Borndörfer et al. (1998), Shirabe (2009), Kim and Xiao (2017), Becker and Solomon (2022), Validi and Buchanan (2022). Thus, the number of variables in these models grows as kn, which is typically much smaller than n^2. However, these models come with a great deal of symmetry which must be dealt with carefully (Validi and Buchanan 2022). The contiguity constraints must also be imposed with care to maintain tractability, often through flow-based (Shirabe 2005, 2009) or cut-based constraints (Carvajal et al. 2013, Wang et al. 2017, Oehrlein and Haunert 2017, Validi et al. 2022). There are also set partitioning models that have a binary variable for each possible district and have constraints requiring that each geographic unit is covered exactly once (Garfinkel and Nemhauser 1970). These set partitioning models generally have an exponential number of variables, and may require the use of column generation to solve their linear programming relaxations (Mehrotra et al. 1998), and branch-and-price to solve the integer programs themselves. Many existing models from the literature seek compactness as their objective and impose population balance and contiguity in the constraints (Hess et al. 1965, Mehrotra et al. 1998, Validi et al. 2022, Validi and Buchanan 2022, Belotti et al. 2023). Recent models consider partisan fairness (Swamy et al. 2022, Gurnee and Shmoys 2021) and minority representation (Arredondo et al. 2021, Belotti et al. 2023).

To our knowledge, there is only one published optimization model from the literature that explicitly promotes the preservation of political subdivisions, which is due to Birge (1983). The proposed subdivision-preserving objective function is quadratic and does not correspond to an established splitting score. Birge found the optimization model too large and unwieldy to solve and instead used a heuristic. There are also two unpublished technical reports. Motivated by a court case in Kentucky, Norman and Camm (2003) propose an optimization model for a problem equivalent to ours. They propose a county-level mixed integer program (MIP), including binary variables x_{ij} indicating whether a portion of county i is assigned to a district seated at county j and continuous variables p_{ij} indicating how many people from county i are assigned to the district seated at county j. To impose contiguity, their first model requires that assignments can only be made to adjacent counties ("one-ring adjacency"), which they then relax to counties at most two hops away ("two-ring adjacency") in a second model. After observing that their MIPs sometimes generate non-compact districts, they fix select x_{ij} variables to zero. They report computational results for a handful of states, with a 46-county 15-district instance taking 610,800 seconds to solve using CPLEX 7.5 on an 866-MHz PC.

A more recent, comprehensive study was undertaken by Önal and Patrick (2016). They start off with a Hess-style MIP where census tracts as taken as the basic geographic unit, and the binary variable x_{ij} indicates whether tract i is assigned to (the district rooted at) tract j. They apply their model to Illinois for 2000 and 2010 census data. To deal with the large number of census tracts ($n \approx 3,000$), they propose heuristic techniques to reduce the number of possible district centers (i.e., fix some x_{jj} variables to zero). They also remove some x_{ij} variables by defining a zone around the possible district centers and disallowing assignments that reach beyond the zone. To impose contiguity, they use existing distance-based constraints stating that if tract i is assigned to tract j, then a neighbor of i that is closer to j is also assigned to j, see Zoltners and Sinha (1983), Mehrotra et al. (1998), Cova and Church (2000). For the objective function, they first minimize the sum of population-weighted distances between tracts and their district centers. They add additional variables to capture the number of county splits and add their sum to the objective, with a user-chosen weight of α. In a similar way, they add variables to capture the number of majority-minority districts and subtract this from the population-weighted-distance objective, with a user-chosen weight of β. The MIPs, which have roughly 70,000 variables and constraints after the (inexact) variable-fixing routines, are applied to generate legislative districts for Illinois that fare better than the enacted maps with respect to minority representation and/or county splitting. In a footnote, they remark that they can reduce the number of county splits for a 2000 Kentucky instance from 16 to 14. Due to the variable-fixing routines, distance-based contiguity constraints, and multiple objectives, it is not known whether the generated maps have a minimum number of county splits.

Last, we revisit the highly relevant work of Carter et al. (2020). They consider North Carolina's *Whole County Provision*, as interpreted by the North Carolina Supreme Court in *Stephenson v. Barlett* (2002) and later clarified in *Dickson v. Rucho* (2015). To limit county splitting, the court required map drawers to follow a clustering procedure that refers to *q-county clusters*, which are subsets of q contiguous counties whose combined population lies within an integer multiple of the population balance window $[L, U]$. Among all possible county clusterings, one must first maximize the number of 1-county clusters, then maximize the number of 2-county clusters, etc, and then divide each cluster into the appropriate number of districts. Carter et al. (2020) provide a combinatorial algorithm for this county clustering problem and apply it to North Carolina legislative districting. They observe that if the real goal is to minimize the number of county splits, then a better objective is to maximize the number of county clusters. They argue that the minimum number of county splits equals the number of districts k minus the maximum number of county clusters, where k is the number of districts to draw. Observe that any contiguous state admits the trivial clustering in which all counties belong to the same cluster, in which case the number of county splits would be $k-1$. However, the claw example in Figure 4 shows that the minimum number of county splits can exceed $k-1$. To understand this discrepancy, let us consider both the "basic" and "enlarged" versions of the county splits theorem of Carter et al. (2020). Their basic theorem states that "A clustering that maximizes the number of county clusters also minimizes the number of county splits." In the enlarged version of this theorem, they add the caveat that this statement holds "except in rare circumstances which impact the optimal districting." The theorem statement does not specify what this means, but their proof refers to "bad combines," which are cases where their algorithmic proof fails. They add that "we think it is rare to require [bad combines] in most real-world scenarios," but no theoretical or empirical evidence is provided. In this paper, we provide empirical evidence that supports this claim.

2.2. Terminology and Notation

Consider a simple graph $G=(V,E)$ that represents geographic units and their adjacencies. Typically, we take G to be a tract-level graph, meaning that the vertex set V consists of a state's census tracts, and the edges in E show which pairs of tracts share a boundary of positive length; it is not enough to meet at a point. In some cases, we take G to be a block-level graph whose vertex set is the set of census blocks. This is sometimes necessary because a tract may have more population than is permitted in a district (e.g., for New Hampshire State House districts). The set of counties is denoted by C, and the set of vertices in V that belong to county $c \in C$ is denoted by V_c.

We also use a county-level graph G_C whose vertices represent counties. The county-level graph can be obtained from G by taking each county $c \in C$, identifying the vertices V_c that belong it, and merging them into a single node. We assume that each county is contiguous, i.e., that the subgraph $G[V_c]$ induced by V_c is connected. Strictly speaking this is not true, and indeed some input graphs are disconnected; however, in these cases, we connect them by adding a subset of "least cost" edges, as in Validi et al. (2022).

By design, each block belongs to precisely one tract, and each tract belongs to precisely one county. Each geographic unit i, which could be a block, tract, or county, has an associated population p_i that is nonnegative (and sometimes zero). The populations across levels are consistent, e.g., the population of a county equals the sum of its tracts' populations. When working with geographic units across the census hierarchy, it is helpful to understand GEOIDs which the Census Bureau uses to uniquely identify them and show their nested relationships. For example, the authors' offices are located in the census block whose GEOID is '401190104001002'. The first two digits '40' correspond to the state of Oklahoma, the next three digits '119' correspond to Payne County, the next five digits '01040' correspond to the tract, and the last five digits '01002' correspond to the block. So, for example, the blocks in Payne County are those whose GEOIDs begin with '40119'.

The desired number of districts is denoted by k. This number is known and set by apportionment for congressional districting and by state law for legislative districting. The ideal district population $\bar{p} := p(V)/k$ equals the state's total population $p(V) := \sum_{i \in V} p_i$ divided by k. We impose population deviations of 1% ($\pm 0.5\%$) for congressional districting and 10% ($\pm 5\%$) for legislative districting. Specifically, we require each congressional district to have a population between $L = \lceil 0.995\bar{p} \rceil$ and $U = \lfloor 1.005\bar{p} \rfloor$. Meanwhile, we use $L = \lceil 0.95\bar{p} \rceil$ and $U = \lfloor 1.05\bar{p} \rfloor$ for legislative districting.

A districting plan $f : V \to [k]$ assigns each vertex to a district from the set $[k] := \{1, 2, \ldots, k\}$. In districting plan f, county c's vertices are assigned to the districts $f[V_c] = \{f(v) \mid v \in V_c\}$, where $f[\cdot]$ denotes the image of a subset. A county c is *whole*, *intact*, or *preserved* if its vertices are assigned to only one district, i.e., if $|f[V_c]| = 1$; otherwise, it is *split*. The number of times that county c

is split is given by $|f[V_c]| - 1$. The minus-one adjustment is done so that whole counties are split zero times. District $j \in [k]$ is contiguous if its vertices $f^{-1}[\{j\}]$ induce a connected subgraph, where $f^{-1}[\cdot]$ denotes the preimage. A districting plan is contiguous if each of its districts is contiguous. A districting plan is population-balanced if each district's population lies between L and U. From now on, we require all districting plans to be contiguous and population-balanced. The total number of county splits is given by $\sum_{c\in C}(|f[V_c]| - 1)$. We seek to minimize this quantity.

PROPOSITION 1. *In a districting plan, vertices from county c must belong to $\lceil p_c/U \rceil$ or more districts, so the total number of county splits is at least*

$$\text{(obvious lower bound)} \quad \sum_{c\in C} \left(\lceil p_c/U \rceil - 1\right).$$

Finally, we remark that *block-level* districting plans are common in practice, but we will try as much as possible to use tracts as our most granular geographic units. The number of census blocks in some states approaches one million (see Texas) which can make computations difficult. Also, the US Census Bureau has indicated that blocks should first be aggregated into larger units before districting to mitigate any undesirable effects that their new disclosure avoidance system may have (US Census Bureau 2021b). Census tracts are also more cohesive, designed to be homogeneous in terms of "population characteristics, economic status, and living conditions" (US Census Bureau 1994).

3. Split Duality

In this section, we prove weak split duality using the county-district incidence graph. Then, we show that *strong* split duality does not always hold. In fact, we generate synthetic districting instances with arbitrarily large *split duality gap*, which we define as the difference between the minimum number of county splits and the lower bound coming from weak split duality.

LEMMA 1. *Denote by k the number of districts. If there is a districting plan with s county splits, then there is a county clustering with at least $k - s$ clusters.*

Proof. Suppose there is a districting plan with s county splits. Construct the *county-district incidence graph*, which has a vertex for each county $c \in C$ and a vertex for each district number $j \in [k]$, as illustrated in Figure 5. An edge connects county c to district j if some portion of county c is assigned to district j. This graph has $n = |C| + k$ vertices and $m = |C| + s$ edges, and thus has at least $n - m = (|C| + k) - (|C| + s) = k - s$ connected components. For each of these components, construct a cluster from its county nodes. $\square$

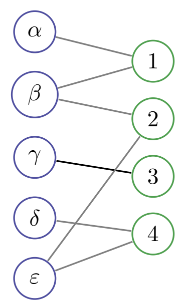

Figure 5 The county-district incidence graph for the plan in Figure 3 (right).

By straightforwardly applying Lemma 1, we get weak split duality.

Theorem 1 **(Weak split duality, Thm. 2 of Carter et al. (2020))**. *The minimum number of county splits is at least the number of districts minus the maximum number of county clusters.*

Proof. Let s be the minimum number of county splits. By Lemma 1, we can construct at least $\hat{c} := k - s$ county clusters. Moreover, the *maximum* number of county clusters c must be at least $\hat{c}$. So, $c \geq \hat{c} \geq k - s$, and thus $s \geq k - c$. $\square$

Proposition 2 **(Arbitrarily large split duality gap)**. *For all nonnegative integers q and h, there exists a districting instance with $k = 2$ districts and a $\pm h$-person deviation that requires at least $k + q$ county splits.*

Proof. Consider the county-level graph depicted in Figure 6. The county populations are given inside the nodes. The total population is $24q + 6h + 18$, to be divided over $k = 2$ districts, so the ideal district population is $p(C)/k = 12q + 3h + 9$, while the lower and upper population bounds are $L = 12q + 2h + 9$ and $U = 12q + 4h + 9$. See that no two leaves can be kept whole in the same district, because their combined population $12q + 4h + 10$ is larger than U. So, consider a districting plan in which one of the leaves l is split, and is thus divided between districts 1 and 2. We claim that all vertices along the path to the population-3 hub (including the hub itself) must be split. If not, then one of them, say v, is kept whole. Without loss, suppose that v is fully assigned to district 1. Then, since district 2 is contiguous and contains part of the leaf county l, it cannot contain portions of counties that lie beyond the hub, and thus has population at most $8q + 2h + 8$ which is less than L, a contradiction. So, the number of split counties (and county splits) is at least $k + q$, irrespective of how counties are made up of census tracts or blocks. $\square$

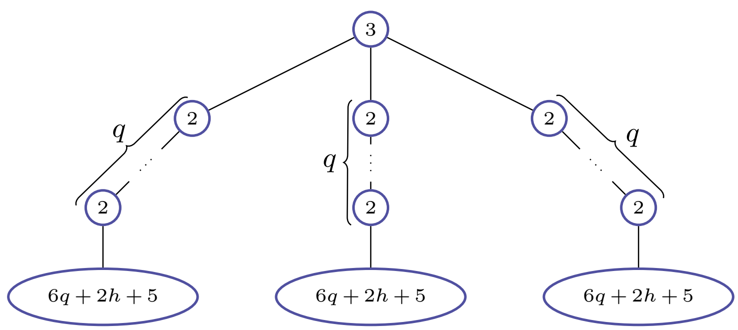

Figure 6 A districting instance that requires at least $k + q$ county splits.

Despite the negative result given in Proposition 2, we will see that the split duality gap is typically zero in practice. For this reason, our exact approach for solving this problem heavily exploits weak split duality.

4. Solving the Minimum County Splits Problem

We now propose to solve the minimum county splits problem, which is the task of finding a (contiguous and population-balanced) districting plan that has a minimum number of county splits. Our overall approach has three steps:

1. *Cluster*. Partition the counties into a maximum number of county clusters $(C_1, C_2, \dots, C_q)$ with associated cluster sizes $(k_1, k_2, \dots, k_q)$. If there are multiple such clusterings, pick one that is compact, i.e., few cut edges (Validi and Buchanan 2022).

2. *Sketch.* For each cluster C_j, sketch a districting plan for it that has k_j districts and $k_j - 1$ county splits. A *sketch* indicates what *proportion* of each county is assigned to each district.

3. *Detail.* For each cluster C_j, find a detailed districting plan that *abides by* the sketch. That is, a tract (or block) is permitted in district j only if its county is (partially) assigned to district j in the sketch.

We solve each step using integer programming techniques. If the *Cluster* step is successful, then weak split duality implies that at least $k - q$ county splits are required (Theorem 1). If the *Sketch* and *Detail* steps are successful, then we obtain a districting plan with a number of county splits equal to

$$(k_1 - 1) + (k_2 - 1) + \cdots + (k_q - 1) = k - q.$$

So, if all three steps are successful (as will be true in our experiments), we conclude that $k - q$ is the minimum number of county splits. These three steps are illustrated for the Tennessee State Senate in Figure 7.

At a high level, our approach follows the proof strategy of Carter et al. (2020). However, they proposed a combinatorial algorithm for the *Clustering* step and were unable to declare optimality or provide an optimality gap for instances from North Carolina. Their approach to the *Sketch* step is not described in enough detail to be implemented, and no procedure is given for the *Detail* step. So, the *Cluster-Sketch-Detail* approach described here constitutes the first exact approach for the problem and generates the first provably optimal solutions for the minimum county splits problem.

4.1. Cluster

In the maximum county clustering problem, the task is to find a county clustering with a maximum number of clusters. To solve this problem, we propose the following mixed integer programming (MIP) model. It is inspired by the classic districting model of Hess et al. (1965), but with tweaks to permit cluster sizes larger than one. For every pair of counties $i, j \in C$, we create a binary variable x_{ij} that equals one when county i is assigned to (the county cluster rooted at) county j.

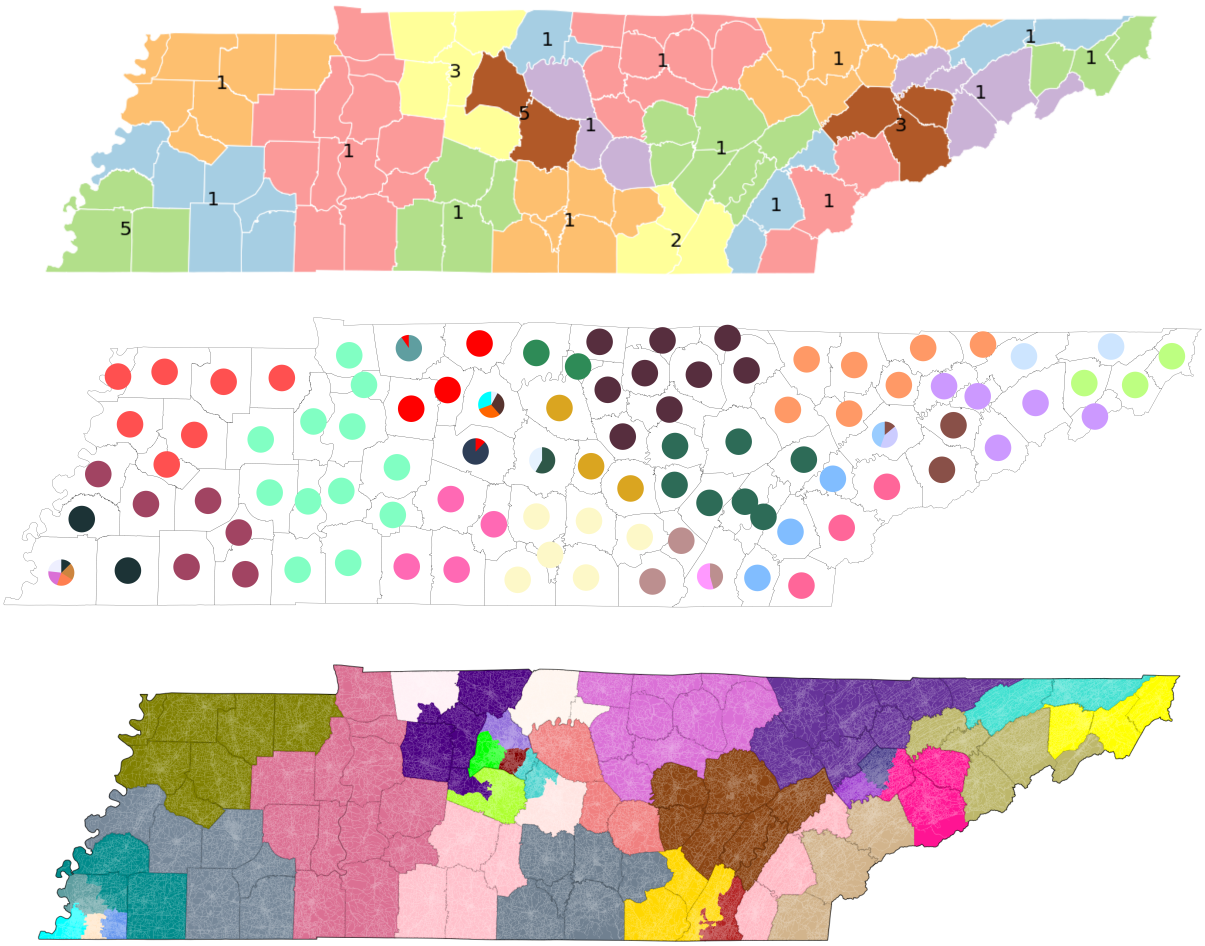

Figure 7 Application of *Cluster-Sketch-Detail* to the Tennessee State Senate. In the first image (*Cluster*), each cluster is given a color. In the second and third images (*Sketch* & *Detail*), each district is given a color.

In particular, the variable x_{jj} equals one if county j roots a county cluster, and the variable y_j represents the size of this cluster. The maximum county clustering model is as follows.

$$\max \sum_{j \in C} x_{jj} \tag{1a}$$

$$\text{s.t.} \sum_{j \in C} x_{ij} = 1 \qquad \forall i \in C \tag{1b}$$

$$\sum_{j \in C} y_j = k \tag{1c}$$

$$C_j = \{i \in C \mid x_{ij} = 1\} \text{ is connected} \qquad \forall j \in C \tag{1d}$$

$$L y_j \leq \sum_{i \in C} p_i x_{ij} \leq U y_j \qquad \forall j \in C \tag{1e}$$

$$x_{ij} \leq x_{jj} \qquad \forall i, j \in C \tag{1f}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in C \tag{1g}$$

$$y_j \in \mathbb{Z}_+ \qquad \forall j \in C. \tag{1h}$$

The objective (1a) maximizes the number of county clusters. The assignment constraints (1b) ensure that each county is assigned to one cluster. The size constraint (1c) ensures that the cluster sizes sum to k. The contiguity constraints (1d) state that each cluster should be connected. For this, we use the flow-based constraints proposed by Shirabe (2005, 2009), cf. Oehrlein and Haunert (2017) and Validi et al. (2022). The population balance constraints (1e) ensure that the cluster rooted at county j has a population between Ly_j and Uy_j. The coupling constraints (1f) ensure that if a county is not selected as a root, then no other counties can be assigned to it.

In our experience, model (1) is not particularly useful "out of the box." For example, one obstacle is model symmetry: a county clustering $(C_1, C_2, \dots, C_q)$ can be represented in the model in $|C_1| \times |C_2| \times \cdots \times |C_q|$ different ways by changing which vertex is selected as the root of each cluster. A popular remedy to this symmetry is the asymmetric representatives trick, which was first introduced for graph coloring problems (Campêlo et al. 2008) but has also been used for districting (Validi et al. 2022, Validi and Buchanan 2022). In our case, it amounts to choosing an ordering of the counties $(c_1, c_2, \dots, c_n)$ and imposing $x_{ij} = 0$ when county i comes before county j in the ordering. In this way, among the counties in a cluster, only the earliest one in the ordering can be its root, eliminating the model symmetry. This diagonal fixing removes nearly half of the x_{ij} variables; see Validi and Buchanan (2022) for other fixing tricks that exploit the population balance constraints.

Even after these variable fixing tricks, model (1) frequently takes more than one hour to identify a maximum county clustering and prove its optimality. To illustrate this behavior, Table 1 provides results for the ten largest US states (by population) after the 2020 census, i.e., those with at least 13 congressional districts. (Details on our computational setup are given with the final set of experiments in Section 5.) The situation did not change substantially when using alternative contiguity constraints besides Shirabe's, such as a,b-separator inequalities (Carvajal et al. 2013, Oehrlein and Haunert 2017, Fischetti et al. 2017, Wang et al. 2017, Validi et al. 2022) or variants of length-U a,b-separator inequalities (Validi et al. 2022, Validi and Buchanan 2022), whether using integer or fractional separation. This prompted us to develop our own heuristics to warm start the MIP. We also propose some valid inequalities to strengthen its LP relaxation.

Table 1 Initial results for maximum county clustering model (1) on the ten largest states. We report the maximum number of county clusters (or best lower and upper bounds $[LB, UB]$) and solve time in a 3600s time limit (TL).

		Congressional		State Senate		State House	
state	$\|C\|$	obj	time	obj	time	obj	time
CA	58	$[11, 12]$	TL	14	303.32	20	106.39
FL	67	$[8, 10]$	TL	16	739.05	26	81.10
GA	159	$[7, 12]$	TL	$[1, 33]$	TL	$[46, 61]$	TL
IL	102	8	461.72	20	1,251.34	$[29, 32]$	TL
MI	83	$[7, 9]$	TL	18	580.82	32	2,054.85
NC	100	$[7, 12]$	TL	$[26, 30]$	TL	$[40, 41]$	TL
NY	62	8	560.00	20	228.36	26	344.60
OH	88	$[9, 12]$	TL	20	948.20	$[35, 36]$	TL
PA	67	$[8, 10]$	TL	23	202.14	39	17.58
TX	254	$[12, 19]$	TL	$[1, 19]$	TL	$[29, 51]$	TL

4.1.1. Upper Bound and Valid Inequalities First, we propose an upper bound on the number of county clusters. A motivating insight is that an overpopulated county i (with population $p_i > U$) must belong to a cluster C_j whose size k_j is more than one. This cluster consumes at least

$\lceil p_i/U \rceil$ units of the size budget k, which reduces the number of possible clusters by $\lceil p_i/U \rceil - 1 = \lfloor p_i/(U+1) \rfloor$. Summing over all counties shows that the number of clusters is at most

$$k - \sum_{i \in C} \left\lfloor \frac{p_i}{U+1} \right\rfloor .$$

This upper bound on the number of county clusters, together with weak split duality, imply the obvious lower bound on the number of county splits that was given in Proposition 1. The cluster upper bound can be generalized to a family of bounds with parameter t as follows.

PROPOSITION 3. *For positive integers t, the number of county clusters is at most*

$$tk - \sum_{i \in C} \left\lfloor \frac{tp_i}{U+1} \right\rfloor .$$

Proof. Let $C_1, C_2, \ldots, C_q$ be a county clustering with sizes $k_1, k_2, \ldots, k_q$. By definition, each cluster C_j satisfies $p(C_j) \leq k_j U$. Multiply by $t/(U+1)$ to get:

$$\sum_{i \in C_j} \frac{tp_i}{U+1} \leq \frac{tk_j U}{U+1}.$$

A weaker version of this inequality also holds:

$$\sum_{i \in C_j} \left\lfloor \frac{tp_i}{U+1} \right\rfloor \leq \frac{tk_j U}{U+1}.$$

Since the left-hand side is an integer, we can round the right-hand side to get:

$$\sum_{i \in C_j} \left\lfloor \frac{tp_i}{U+1} \right\rfloor \leq \left\lfloor \frac{tk_j U}{U+1} \right\rfloor \leq tk_j - 1. \tag{2}$$

Then, we have

$$q + \sum_{i \in C} \left\lfloor \frac{tp_i}{U+1} \right\rfloor = \sum_{j=1}^{q} \left(1 + \sum_{i \in C_j} \left\lfloor \frac{tp_i}{U+1} \right\rfloor \right) \leq \sum_{j=1}^{q} tk_j = tk,$$

where the inequality holds by (2). $\square$

Example. Consider six counties, each with a population of 200, that are to be divided into $k=4$ equal-population districts, i.e., $L=U=300$. If we apply Proposition 3 for $t=1$, we get that the number of clusters is at most

$$tk-\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor=1(4)-6\left\lfloor\frac{1(200)}{300+1}\right\rfloor=4,$$

which is not helpful. However, if we use $t=2$, we get the tight bound

$$tk-\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor=2(4)-6\left\lfloor\frac{2(200)}{300+1}\right\rfloor=2.\quad\square$$

Sometimes a bound coming from Proposition 3 is tight, in which case it may not be necessary to solve a MIP model to prove optimality of a given county clustering. Otherwise, we can still generate valid inequalities using similar ideas.

THEOREM 2 **(Rounding Inequalities)**. *Let t be a positive integer, and let j be a county. The following rounding inequality is valid for model* (1)*.*

$$\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor x_{ij}\le ty_j-x_{jj}.$$

Proof. Let (x^*,y^*) be a feasible solution to model (1). If $y_j^*=0$, then we have

$$\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor x_{ij}^*=0\le 0=ty_j^*-x_{jj}^*.$$

In the other case, where $y_j^*\ge 1$, we have $\sum_{i\in C}p_ix_{ij}^*\le Uy_j^*$ by population balance. Multiply both sides by $t/(U+1)$ to get:

$$\sum_{i\in C}\left(\frac{tp_i}{U+1}\right)x_{ij}^*\le\frac{tUy_j^*}{U+1}.$$

A weaker version of this inequality also holds:

$$\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor x_{ij}^*\le\frac{tUy_j^*}{U+1}.$$

Since the left-hand-side is integer, we can round the right-hand side to get:

$$\sum_{i\in C}\left\lfloor\frac{tp_i}{U+1}\right\rfloor x_{ij}^*\le\left\lfloor\frac{tUy_j^*}{U+1}\right\rfloor\le ty_j^*-1\le ty_j^*-x_{jj}^*.$$

The middle inequality requires $ty_j^*>0$, which is why the case $y_j^*=0$ is considered separately. $\square$

Next, we generalize the a, b-separator inequalities, which have previously been used to impose contiguity, to *cluster-separator inequalities*. In their original form, these inequalities state that if vertices a and b are disconnected after removing vertex subset S from the graph, then the inequality $x_{ab} \leq \sum_{i \in S} x_{ib}$ is valid. In words, if vertex a is assigned to b, then at least one vertex from S must join them. We observe that the same inequality can sometimes be applied even when a and b are connected in $G_C - S$ by exploiting population balance.

For example, recall the Splitigan example from Figure 3. If Gamma and Epsilon County are removed, this leaves a component consisting of Alpha and Beta County, each having a population of three. Their combined population of six does not lie within an integer multiple of the population bounds $[L, U] = [4, 4]$, so they cannot form a county cluster by themselves. We conclude that if Alpha County is assigned to Beta County, then Gamma and/or Epsilon County must join them, i.e., $x_{\alpha\beta} \leq x_{\gamma\beta} + x_{\varepsilon\beta}$. Generally, we have the following theorem, where a *county cluster* is a subset of contiguous counties whose population lies within an integer multiple of the population balance window $[L, U]$.

THEOREM 3 **(Cluster-Separator Inequalities)**. *Let S be a subset of counties, and let $a, b \notin S$ be two other counties (possibly $a = b$). If there is no county cluster in $G_C - S$ that contains a and b, then the following cluster-separator inequality is valid for model* (1).

$$x_{ab} \leq \sum_{i \in S} x_{ib}.$$

Proof. Suppose that $(\hat{x}, \hat{y})$ is feasible to model (1) and that there is no county cluster in $G_C - S$ that contains a and b. We are to show that $\hat{x}_{ab} \leq \sum_{i \in S} \hat{x}_{ib}$. If $\hat{x}_{ab} = 0$ then the inequality is easily satisfied, so suppose that $\hat{x}_{ab} = 1$. Consider the subset of counties assigned to b, i.e., $C_b = \{i \in C \mid \hat{x}_{ib} = 1\}$. By constraints (1f), we have $1 = \hat{x}_{ab} \leq \hat{x}_{bb}$ so b also belongs to C_b. By constraints (1d), C_b is connected. By constraints (1e), C_b has a population satisfying $L\hat{y}_b \leq p(C_b) \leq U\hat{y}_b$, where $\hat{y}_b$ is a nonnegative integer by constraints (1h) (and positive by constraints (1e)). This is all to say that C_b qualifies as a county cluster of size $\hat{y}_b$ that contains a and b. Now, if there is an element c common to both S and C_b, then $\hat{x}_{ab} = 1 = \hat{x}_{cb} \leq \sum_{i \in S} \hat{x}_{ib}$, as desired. Otherwise, S and C_b are disjoint, meaning that C_b is a county cluster in $G_C - S$ that contains a and b, a contradiction. $\square$

The particular case of these inequalities where $a = b$ was essentially proposed by Oehrlein and Haunert (2017), see their inequalities (14).

We evaluate the effectiveness of the rounding inequalities from Theorem 2 and the cluster-separator inequalities from Theorem 3 by testing them on the same ten states as in Table 1. We apply the rounding inequalities for the parameter value $t = 1$, as this seemed to work best in initial testing. We apply the cluster-separator inequalities for sets S that are near[1] to b. We also attempted to separate the cluster-separator inequalities on-the-fly, but adding the subset of them "near" to b a priori seemed to work better.

Results are given in Table 2. We find the valid inequalities to be very helpful for the state senate and state house instances, solving most of them within a one-hour time limit or leaving an absolute optimality gap of one cluster. The exception is Texas, which has the most counties of any state (254). However, most of the congressional districting instances remain troublesome, which motivates the warm start heuristics that are developed next.

4.1.2. MIP-Based Construction Heuristic First, we propose a MIP-based construction heuristic. It is inspired by a procedure of McCartan and Imai (2023). At a high-level, each step of their approach is to draw a random spanning tree and delete one of its edges to carve off a district. Our approach is different in that we carve off *clusters* and use a MIP for the carving step. In a greedy attempt to maximize the number of clusters, we minimize the *size* k_j of the cluster C_j that is carved off. To promote compactness, we impose a secondary objective to minimize the weight of the cut edges emanating from the cluster. The weight of each edge is chosen uniformly at random between zero and one. In this way, we can run the construction heuristic multiple times and get different starting points for local search. Given inputs p, L, U, k, and county-level graph G_C, the construction heuristic generates a county clustering as follows.

1. let $G' \leftarrow G_C$ and $k' \leftarrow k$ and $j \leftarrow 0$
2. choose edge weights uniformly at random from $[0, 1]$
3. while $k' > 0$ do

Table 2 Initial results when using valid inequalities from Theorems 2 and 3.

		Congressional		State Senate		State House	
state	$\lvert C\rvert$	obj	time	obj	time	obj	time
CA	58	$[10,12]$	TL	14	108.24	20	11.69
FL	67	$[8,10]$	TL	16	104.10	26	4.82
GA	159	$[6,12]$	TL	$[31,32]$	TL	$[57,58]$	TL
IL	102	8	1,663.47	20	90.73	$[30,31]$	TL
MI	83	9	2,886.09	18	87.65	32	186.04
NC	100	$[10,12]$	TL	28	709.68	40	802.24
NY	62	8	2,628.22	20	27.14	26	75.91
OH	88	$[10,12]$	TL	20	294.78	35	2,206.51
PA	67	10	457.92	23	30.34	39	5.96
TX	254	$[1,19]$	TL	$[13,19]$	TL	$[1,51]$	TL

- $j \leftarrow j+1$
- find a nonempty cluster $C_j \subseteq V(G')$ and integer $k_j \geq 1$ such that

  (a) the cluster C_j and its complement $V(G') \setminus C_j$ are connected,

  (b) the cluster C_j and its complement are population-balanced, i.e.,

$$k_j L \leq p(C_j) \leq k_j U \quad \text{and} \quad (k'-k_j)L \leq p(V(G') \setminus C_j) \leq (k'-k_j)U,$$

  (c) the size k_j is minimum, with a secondary objective to minimize the weight of the edges between the cluster and its complement

- update $G' \leftarrow G' - C_j$ and $k' \leftarrow k' - k_j$

4. return county clusters $(C_1, C_2, \ldots, C_j)$ with sizes $(k_1, k_2, \ldots, k_j)$

Each iteration selects a cluster C_j and associated size k_j. For this task, we use a two-cluster labeling MIP (Validi and Buchanan 2022) in which the size of the first cluster is k_j, the size of the second cluster is $k' - k_j$, and contiguity is imposed by adding violated a,b-separator inequalities in a cut callback. To handle the two objectives, we initially fix the size to the smallest possible value

$k_j = 1$ and minimize the weight of the cut edges. If this MIP is infeasible, then we increase the size k_j by one and re-optimize, repeating until feasibility. Typically, each MIP solve takes less than one second, and at most k iterations are required, so the total time is reasonable. The solution quality is usually good but suboptimal often enough to motivate local search. Also, even when a county clustering is known to be maximum, local search can help make the clusters more compact.

4.1.3. MIP-Based Local Search We improve the feasible solutions coming from the construction heuristic using MIP-based local search. For the local search neighborhood, we use recombination, which was originally proposed to generate large ensembles of districting plans (DeFord et al. 2021). In an ordinary recombination move, two districts are merged into a *double district*, a random spanning tree is drawn over their vertices, and one of its edges is deleted to split the spanning tree into two new districts. The authors also proposed a more general t-opt recombination move in which t districts are merged together and re-partitioned into t new districts, say, by deleting $t-1$ edges from a random spanning tree. Our approach is different in that we are working with *clusters* and use a MIP for the re-partitioning step. Again, our primary objective is to maximize the number of clusters, with a secondary objective to minimize the number of cut edges between clusters. Given inputs p, L, U, k, county-level graph G_C, and initial county clustering $(C_1, C_2, \dots, C_q)$ with sizes $(k_1, k_2, \dots, k_q)$, the t-opt recombination heuristic works as follows.

1. Select t clusters from the county clustering, say, $(C'_1, C'_2, \dots, C'_t)$ with sizes $(k'_1, k'_2, \dots, k'_t)$.
2. Merge the clusters $C' = C'_1 \cup C'_2 \cup \dots \cup C'_t$ and sizes $k' = k'_1 + k'_2 + \dots + k'_t$.
3. If $G[C']$ is connected, find a maximum county clustering of the counties C' with total size k', with a secondary objective to minimize the number of cut edges.
4. If there is an improvement, then update the county clustering.
5. Repeat, trying different subsets of t clusters, until no more improvements are possible.

To find a maximum county clustering of (C', k') in step 3, we simply solve the maximum county clustering model (1). Denote by $t' \geq t$ the resulting number of clusters. We then minimize the number of cut edges by solving a t'-cluster labeling MIP in which contiguity is imposed by adding violated a,b-separator inequalities in a cut callback (Validi and Buchanan 2022).

The results of our heuristics and their effectiveness as a MIP warm start are illustrated in Table 3. For the purposes of these initial tests, we apply them just to congressional instances, as this is where they are needed most. We run the carving construction heuristic, followed by t-opt recombination local search for $t=2$, then $t=3$, and lastly $t=4$. We iterate this MIP-based carve-and-recombination heuristic three times. Afterwards, we use the best county clustering to warm start model (1). When possible, we use the upper bound from Proposition 3 for $t=1$ to avoid a MIP solve (see GA, MI, and TX). Otherwise, we apply the inequalities from Theorems 2 and 3.

Table 3 Initial heuristic results for ten congressional districting instances.

		Carve and Recom		MIP Solve	
state	$\lvert C\rvert$	LB	time	obj	time
CA	58	11	1,671.48	11	2,269.01
FL	67	8	2,063.04	$[8, 10]$	TL
GA	159	12	8,457.05	12	0.00
IL	102	8	2,165.48	8	8.78
MI	83	9	5,272.17	9	0.00
NC	100	11	1,255.11	$[11, 12]$	TL
NY	62	8	2,905.24	8	31.37
OH	88	11	3,849.23	11	161.38
PA	67	10	4,031.30	10	64.01
TX	254	19	14,747.39	19	0.00

4.2. Sketch

The next step after *Cluster* is *Sketch*. A sketch indicates what proportion of each county is assigned to each district. We begin with a maximum county clustering $(C_1, C_2, \dots, C_q)$ with sizes $(k_1, k_2, \dots, k_q)$. For each cluster C' with size k', we sketch a districting plan with k' districts and

$k'-1$ county splits, eventually leading to a districting plan with $(k_1-1)+(k_2-1)+\cdots+(k_q-1)=k-q$ county splits, matching the bound from weak split duality.

For this task, we propose a MIP. For notational simplicity, we write the formulation for the entire set of counties C and total size k, but in actuality it should be applied to each cluster separately. The binary variable x_{ij} indicates whether some portion of county $i\in C$ is assigned to district $j\in[k]:=\{1,2,\dots,k\}$. The variable z_{ij} indicates what proportion of county $i\in C$ is assigned to district $j\in[k]$. The integer variable s_i counts the number of times that county $i\in C$ is split. We begin with some *basic* constraints.

$$\sum_{i\in C} s_i = k-1 \tag{3a}$$

$$\sum_{j=1}^{k} x_{ij} = s_i + 1 \qquad \forall i\in C \tag{3b}$$

$$\sum_{j=1}^{k} z_{ij} = 1 \qquad \forall i\in C \tag{3c}$$

$$L\le \sum_{i\in C} p_i z_{ij}\le U \qquad \forall j\in[k] \tag{3d}$$

$$0\le z_{ij}\le x_{ij} \qquad \forall i\in C,\ \forall j\in[k] \tag{3e}$$

$$x_{ij}\in\{0,1\} \qquad \forall i\in C,\ \forall j\in[k]. \tag{3f}$$

Constraint (3a) imposes that there are $k-1$ county splits. Constraints (3b) count the number of splits for each county. Constraints (3c) impose that 100% of each county is assigned. Constraints (3d) ensure population balance. Constraints (3e) relate the continuous assignment variables z_{ij} and their binary counterparts x_{ij}.

So far, the model captures the population balance constraints and splitting constraints. Other properties that we would like in our sketch include compactness and contiguity. Accordingly, we introduce binary variables y_{ej} indicating whether edge e *belongs* to district $j\in[k]$, i.e., if its endpoints are assigned to district j. We permit each edge to belong to *at most one* district; this removes the incentive for multiple districts to cross the border between two counties, yielding computational speedups and better sketches. This leads to the following *edge consistency* constraints over the edges $E(C)$ of the cluster.

$$y_{ej} \leq x_{ij} \qquad \forall i \in e \in E(C),\ \forall j \in [k] \tag{4a}$$

$$y_{ej} \geq \sum_{i \in e} x_{ij} - 1 \qquad \forall e \in E(C),\ \forall j \in [k] \tag{4b}$$

$$\sum_{j=1}^{k} y_{ej} \leq 1 \qquad \forall e \in E(C) \tag{4c}$$

$$y_{ej} \in \{0,1\} \qquad \forall e \in E(C),\ \forall j \in [k]. \tag{4d}$$

If an edge belongs to a district, then it is *preserved*; otherwise, it is *cut*. In a contiguous district, its number of preserved edges is at least its number of nodes minus one.

$$\sum_{e \in E(C)} y_{ej} \geq \sum_{i \in C} x_{ij} - 1 \qquad \forall j \in [k]. \tag{5}$$

When the input graph is a tree (which is often true for small county clusters), these constraints (4) and (5) suffice for contiguity (Chopra et al. 2017, Wang et al. 2017). Generally, however, our implementation adds violated a, b-separator inequalities in a callback. For compactness, we maximize the number of preserved edges which is equivalent to minimizing the number of cut edges:

$$\max \sum_{e \in E(C)} \sum_{j=1}^{k} y_{ej}.$$

As a secondary objective, we minimize the sum of squares of counties touched by the districts

$$\min \sum_{j=1}^{k} \left(\sum_{i \in C} x_{ij} \right)^2,$$

which can be linearized by introducing a binary variable b_{jt} for each district j and each possible number of counties touched $t = 1, 2, \ldots, |C|$ and writing

$$\begin{aligned}
\min \quad & \sum_{j=1}^{k} \sum_{t=1}^{|C|} t^2 b_{jt} & \\
\text{s.t.} \quad & \sum_{i \in C} x_{ij} = \sum_{t=1}^{|C|} t b_{jt} & \forall j \in [k] \\
& \sum_{t=1}^{|C|} b_{jt} = 1 & \forall j \in [k] \\
& b_{jt} \in \{0,1\} & \forall j \in [k],\ \forall t = 1, 2, \ldots, |C|.
\end{aligned}$$

To motivate the secondary objective, consider an instance with three counties (c_1, c_2, c_3) arranged in a line with populations $(140, 120, 140)$ that are to be divided into four districts of 100 people. In one possible sketch, a 100-person district is created inside each county, with the fourth district snaking through all counties to pick up the leftover populations $(40, 20, 40)$. This sketch has three county splits and two preserved edges. In another possible sketch, we build the districts from left-to-right, with populations taken from counties (c_1, c_2, c_3) as $(100, 0, 0)$, $(40, 60, 0)$, $(0, 60, 40)$, and $(0, 0, 100)$. This sketch also has three county splits and two preserved edges. However, the latter sketch scores better with respect to some splitting scores as well as to our secondary objective, $1^2 + 2^2 + 2^2 + 1^2 = 10$ versus $1^2 + 1^2 + 1^2 + 3^2 = 12$, and we found it easier to extend the latter sketch into a detailed districting plan. Indeed, imagine an example with 100 counties arranged in a line, each with a population of 101. The 100-inside-each-county plus snaking-leftovers plan would have one very strange district with one person from each county, which would give it an awful secondary score of $1^2 + 1^2 + \cdots + 1^2 + 100^2 = 10100$, while the left-to-right plan would have a secondary score of $1^2 + 2^2 + 2^2 + \cdots + 2^2 + 1^2 = 398$. Further, the left-to-right sketch is easier to detail.

To speed up the MIP solve, we eliminate some model symmetry by identifying a county $i \in C$ with largest population and imposing $z_{i1} \geq z_{i2} \geq \cdots \geq z_{ik}$. This in turn forces $x_{ij} = 1$ for $j = 1, 2, \ldots, \lceil p_i/U \rceil$. Ultimately, the time needed for *Sketch* is negligible, less than one second in 99% of cases. This is explainable by the fact that the clusters usually have small size and few counties.

4.3. Detail

The last step is *Detail*, which extends the sketches into a detailed districting plan. At a high-level, we use a capacitated k-means heuristic, but with additional constraints for contiguity and county-splitting. Where possible, we speed up the computation by building districts from tracts, but sometimes resort to blocks. We also apply simplified (and speedy) contiguity constraints at first, only resorting to more expensive a, b-separator inequalities when needed. Below, we motivate and explain these implementation decisions, which do not sacrifice optimality.

Detail finds a detailed districting plan for each county cluster that *abides by* the sketch. For example, suppose the sketch has a given county being 50% assigned to district 3, 30% assigned to district 5, and 20% assigned to district 9. Then, in the *Detail* step, we permit the tracts (or blocks) of this county to be assigned *only* to districts 3, 5, and 9. Even though the percentages from the sketch are ignored, the number of county splits will remain the same, $(k_1 - 1) + (k_2 - 1) + \cdots + (k_q - 1) = k - q$, again matching the bound from weak split duality.

Districts can usually be built from tracts, but this already poses a computational challenge as the number of tracts in many states exceeds one thousand. Worse, some instances require more granular units such as blocks (see New Hampshire's State House), and the number of blocks in a state can approach one million. Accordingly, we must be very careful in our modeling and approach. A direct application of standard integer programming models like the Hess model or the labeling model is simply out of the question. Working in our favor is the fact that clusters usually have few counties, particularly for state house instances. Additionally, the sketches keep many counties whole, and for such counties there is nothing for us to decide. Nevertheless, we will still encounter challenges in big cities. For example, consider the county clusters that contain Los Angeles County, Cook County (Chicago), Harris County (Houston), or Maricopa County (Phoenix). These counties must be split, so it is inevitable that we will have to deal with large tract-level or block-level instances. This motivates us to develop MIP-based heuristics for *Detail* that can handle large instances.

First, we apply a capacitated k-means heuristic, similar to Hess et al. (1965), Bradley et al. (2000), and Validi et al. (2022). With a suitable map projection, we identify (x, y)-coordinates of each tract's centroid. Then, after obtaining an initial assignment of tracts to k districts, we find the (population-weighted) mean of each district. The cost to assign tract i to district j is taken as $(p_i + 1)d_{ij}^2$ where p_i is the population of i and d_{ij} is the Euclidean distance between the centroid of i and the mean of district j, where the $+1$ term ensures that zero-population tracts still prefer nearby districts. We reassign tracts to districts by minimizing the reassignment cost subject to the population balance constraints. We repeat this procedure until convergence. At termination, we have a partition of the tracts into k reasonably compact—although typically not contiguous—districts. Throughout this procedure, we enforce the sketch constraints. That is, tract i is permitted in district j only if its county is partially assigned to district j. By our sketch, this means that there will be at most $k-1$ county splits.

If we are lucky, this procedure returns connected districts; otherwise, we add explicit contiguity constraints. For computational efficiency, we first try constraints inspired by the tree-based contiguity constraints of Zoltners and Sinha (1983), see also Mehrotra et al. (1998), Cova and Church (2000), and Gurnee and Shmoys (2021). After applying the capacitated k-means heuristic, find the node (tract) that is closest to each district's mean. This gives a set of k roots. From each root, find the (graph-based) distance to all other vertices, where intra-county edges have weight one and inter-county edges have large weight (e.g., equal to the number of nodes in the graph). In this way, the distances within a county are shorter than distances across county boundaries, cf. Clelland et al. (2022). For each district j, order the vertices by increasing distance from its root, and let $\text{pos}_j(i)$ be the position of vertex i in this ordering. Then, for each non-root vertex i and for each district j, we impose that if i is assigned to j, then at least one of its neighbors that appears earlier in the ordering must be assigned to j, i.e.,

$$x_{ij} \leq \sum_{\substack{v \in N(i) \\ \text{pos}_j(v) < \text{pos}_j(i)}} x_{vj}. \tag{6}$$

Any x that satisfies these DAG constraints gives connected districts, although the converse is not true. We call them DAG constraints because they can equivalently be expressed in terms of a directed acyclic graph in which all m edges are oriented based on the ordering, whereas the tree-based constraints of Zoltners and Sinha are based on the $n-1$ edges of a shortest-path tree. In this way, the DAG constraints permit more solutions than the tree-based constraints, at the cost of having more nonzeros—$O(km)$ versus $O(kn)$—although planar instances satisfy $O(km) = O(kn)$. The DAG constraints also permit more solutions than distance-based contiguity constraints if distances are not unique. If we are unsuccessful with the DAG constraints, then we resort to using the a,b-separator constraints in callback.

The above steps are first attempted on the tract-level graph. If that is unsuccessful, then we resort to using the block-level graph. When applying the DAG constraints, intra-tract edges have weight one, other intra-county edges have weight $|V|$, and inter-county edges have weight $|V|^2$. Given the large size of the block-level instances and the desire to keep tracts whole, we add additional constraints to the a,b-separator model requiring that each tract is divided across at most two districts and that only $k-1$ tracts are split where k is the size of the cluster. For the vast majority of instances, these tract-level or block-level procedures suffice for the *Detail* step. For a handful of tricky clusters, we use other ad-hoc methods to find a suitable districting plan.

5. Final Results

In our computational experiments, we use a Dell Precision Tower 7000 Series (7810) machine running Windows 10 enterprise, x64, with an Intel Xeon Processor E52630 v4 (10 cores, 2.2GHz, 3.1GHz Turbo, 2133MHz, 25MB, 85W) and 32 GB memory. Our MIP solver is Gurobi v10.0. Our code is written in Python and is available at `https://github.com/maralshahmizad/Political-Districting-to-Minimize-County-Splits`.

The raw districting data comes from the US Census Bureau (2021a). Initial data processing was conducted by Redistricting Data Hub (2021a). Daryl DeFord finished the data processing, including creating the graphs and storing them as `json` files, and kindly shared the files with us. We are also happy to share the 10+ GB of data with others after they agree to the terms set by RDH. The GerryChain package (MGGG 2023) is used to read the `json` files and convert them to NetworkX graphs. The number of legislative districts for each state was taken from Ballotpedia (2023). Adjustments were made in select cases to handle multi-member districts[2]. We do not provide results for Hawaii given that it is a collection of islands, and contiguity is far from attainable. The number of county splits in enacted plans was calculated from block equivalency files for the 118th Congress (US Census Bureau 2023) and 2022 state legislative districts (US Census Bureau 2022).

To draw maps, we use the TIGER/Line Shapefiles from the US Census Bureau (2021c). We do not consider any laws that vary by state. For example, some states reallocate incarcerated individuals so that they are counted at their previous residence (Redistricting Data Hub 2021b), but we directly use the P.L. 94-171 data in our experiments (US Census Bureau 2021a). Unless noted otherwise, we impose a 1% (±0.5%) population deviation for congressional districts and a 10% (±5%) population deviation for legislative districts, which are chosen to follow legal norms.

For the maximum county clustering problem, we apply the valid inequalities (rounding and cluster-separator inequalities) to all instances, but apply the warm start heuristics (carving and recombination) only to Texas and the congressional instances. We impose a 24-hour time limit on the final MIP solve. If the final MIP solve identifies a larger county clustering than the warm start, then we re-apply local search for $t = 2, 3, 4$, but only to reduce the number of cut edges, a process that we call *cleanup*. All maximum county clustering instances are solved by our standard implementation, with the exception of NC/CD and GA/SS, which terminated with bounds of $[11, 12]$ and $[31, 32]$, respectively. Using ad-hoc analyses, we show that 11 and 31 are optimal, as documented in the GitHub repository.

The sketch and detail codes were successful for all but seven instances: PA/CD, IN/SS, NC/SS, FL/SH, ME/SH, NH/SH, WY/SH. To solve them, we make ad-hoc tweaks to the code (e.g., longer time limit, tweaks to sketch, tweaks to detail). Final results for congressional, state senate, and state house instances are provided in Tables 5, 6, and 7 of Appendix B, respectively.

5.1. Strong Split Duality Holds in Practice

First, we observe that strong split duality holds for all 140 instances. That is, the `split LB` column (which reports the bound from weak split duality) always equals the minimum number of county splits in the `min splits` column. This provides the first empirical evidence to support claims of Carter et al. (2020) who wrote that "we think it is rare [to have strong split duality violations] in most real-word scenarios." So, while strong split duality should not be considered as a mathematical theorem (see Proposition 2), our experiments have found it to be a safe working hypothesis.

5.2. Strength of the Obvious Lower Bound

Next, we evaluate the strength of the obvious lower bound from Proposition 1. For each districting instance, we calculate the absolute optimality gap between the obvious lower bound and the minimum number of county splits. For each district type, we construct a histogram showing the frequency of each absolute optimality gap, see Figure 8.

For many congressional instances, the obvious lower bound is often strong, being sharp (i.e., with zero gap) for 25/43 instances and yielding an absolute optimality gap of at most three for 40/43 instances. The exceptions are Florida (10 vs. 19), New Jersey (3 vs. 9), and New York (13 vs. 18). For state senate instances, the obvious lower bound is again reasonably strong, yielding an absolute optimality of at most three for 46/49 instances. The exceptions are Florida (18 vs. 24), Mississippi (19 vs. 23), and South Carolina (26 vs. 30). However, for state house instances, the obvious lower bound is less powerful, yielding an absolute optimality gap of at most three for only 20/48 instances. Not once is it sharp, not even for states like Delaware or Rhode Island that have just a handful of counties. Moreover, the obvious lower bound is off by five or more on 22 state house instances. For these instances, a different bound, such as the one coming from weak split duality, is crucial to prove optimality.

5.3. Comparison with Enacted Plans

Now, we compare the minimum number of county splits to that of the enacted plans. For congressional instances, we see a noticeable difference for states such as California (41 vs. 72), Georgia (2 vs. 21), Illinois (9 vs. 53), and Texas (19 vs. 59). Interestingly, the congressional plans enacted by Illinois and Texas after the 2020 Census have been declared the worst Democratic and Republican gerrymanders, respectively (Kenny et al. 2023).

However, partisan gerrymandering is not the only explanation for these differences. One could rightly speculate that compliance with federal law (e.g., the Voting Rights Act of 1965), state law, or other criteria may force more county splits than Table 5 suggests. Recall that we enforce a 1% ($\pm$0.5%) population deviation in our congressional experiments, as the US Supreme Court

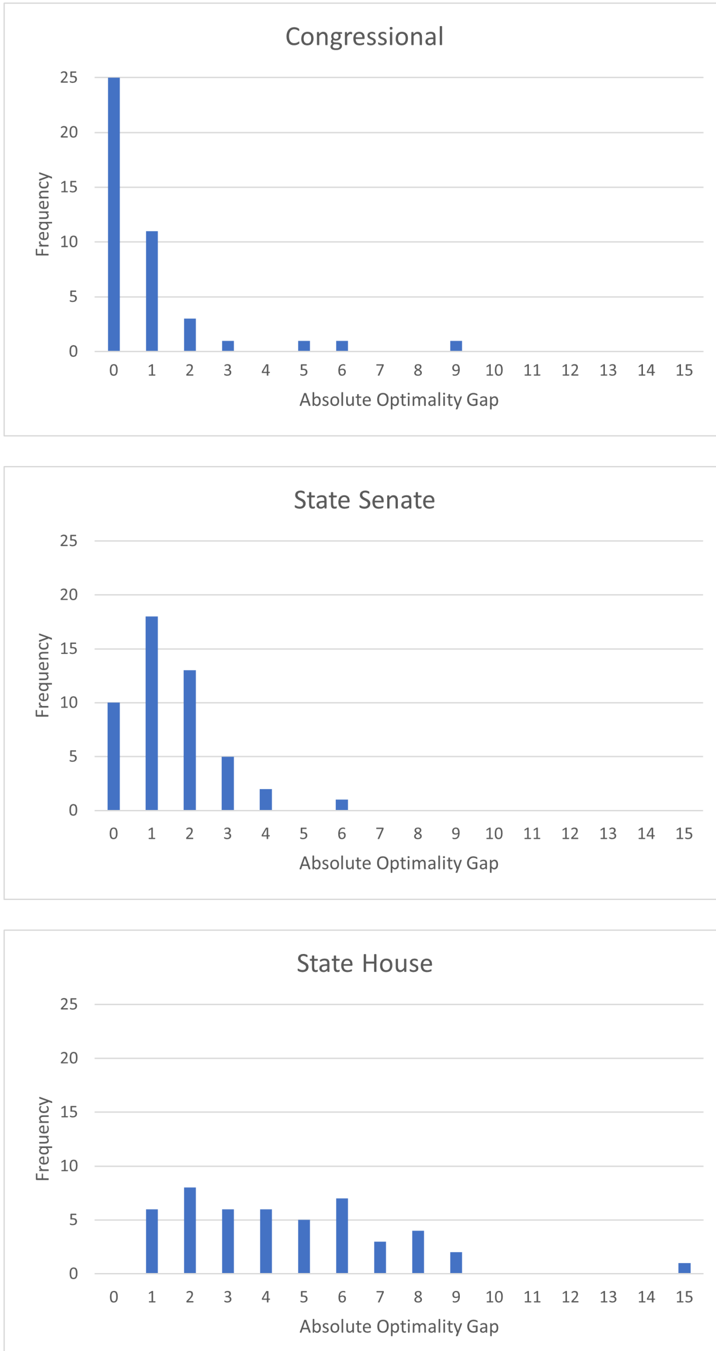

Figure 8 By how much does the obvious lower bound underestimate the minimum number of county splits?

has permitted population deviations approaching 1% in the recent past, see *Tennant v. Jefferson County* (2012). However, most states adopt congressional plans with 1-person deviation. The reason is that any deviation from "precise mathematical equality" risks a lawsuit, at which point the state must show that the deviation was "necessary to achieve some legitimate state objective" including but not limited to "making districts compact, respecting municipal boundaries, preserving the cores of prior districts, and avoiding contests between incumbent Representatives," see *Karcher v. Daggett* (1983) and Hebert et al. (2010). Rather than deal with these issues, states often opt for 1-person deviation, which may require more county splits than Table 5 suggests. (In Section 5.4, we dive deeper into 1-person deviation.) However, we suspect that none of these justifications can explain the large number of county splits in the congressional plans for Illinois and Texas.

For the state senate and state house instances, many enacted plans have an enormous number of county splits compared to the minimum possible. This is despite the fact that many states impose the same 10% ($\pm$5%) population deviation that we do. For state senate, consider Illinois (39 vs. 135), Louisiana (21 vs. 77), Minnesota (41 vs. 100), and Wisconsin (13 vs. 73). For state house, consider Illinois (87 vs. 220), Indiana (61 vs. 129), Mississippi (86 vs. 181), and Wisconsin (69 vs. 159). Meanwhile, for New Hampshire and Vermont, the enacted plans actually have *fewer* county splits, but only because of the particular way we handled their floterial and multi-member districts of different sizes (see endnote 2).

Which state legislative districts perform the best/worst in terms of county splitting? To answer this question, we calculate the *county splits ratio* for each instance, which we define as the number of county splits in the enacted plan divided by the minimum number of county splits possible (in a contiguous and population-balanced plan). A ratio of 1 indicates that the enacted plan achieved a minimum number of county splits, while a ratio of 2 indicates that the enacted plan has twice as many county splits than the contiguity and population-balance constraints require. The county splits ratios for each state senate and state house are plotted in Figure 9. The plot excludes MD, NH, SD, and VT for multi-member district reasons, as well as NE because it has no house.

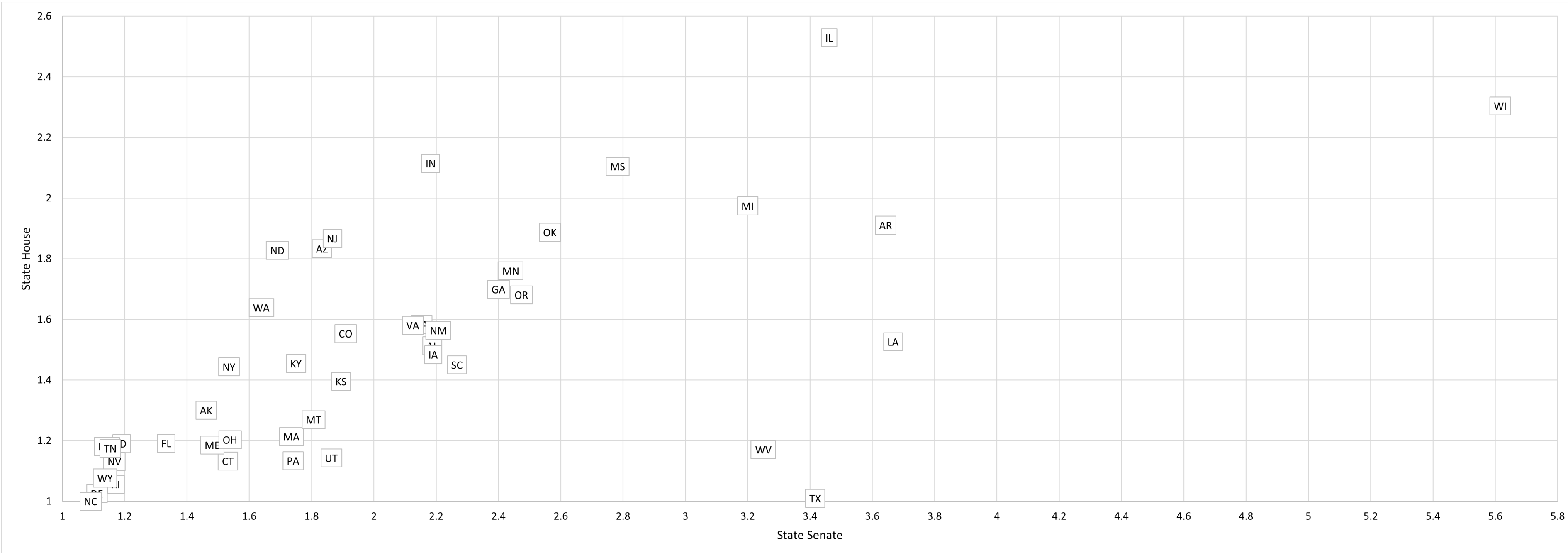

Figure 9 County Splits Ratios for State Senate and State House

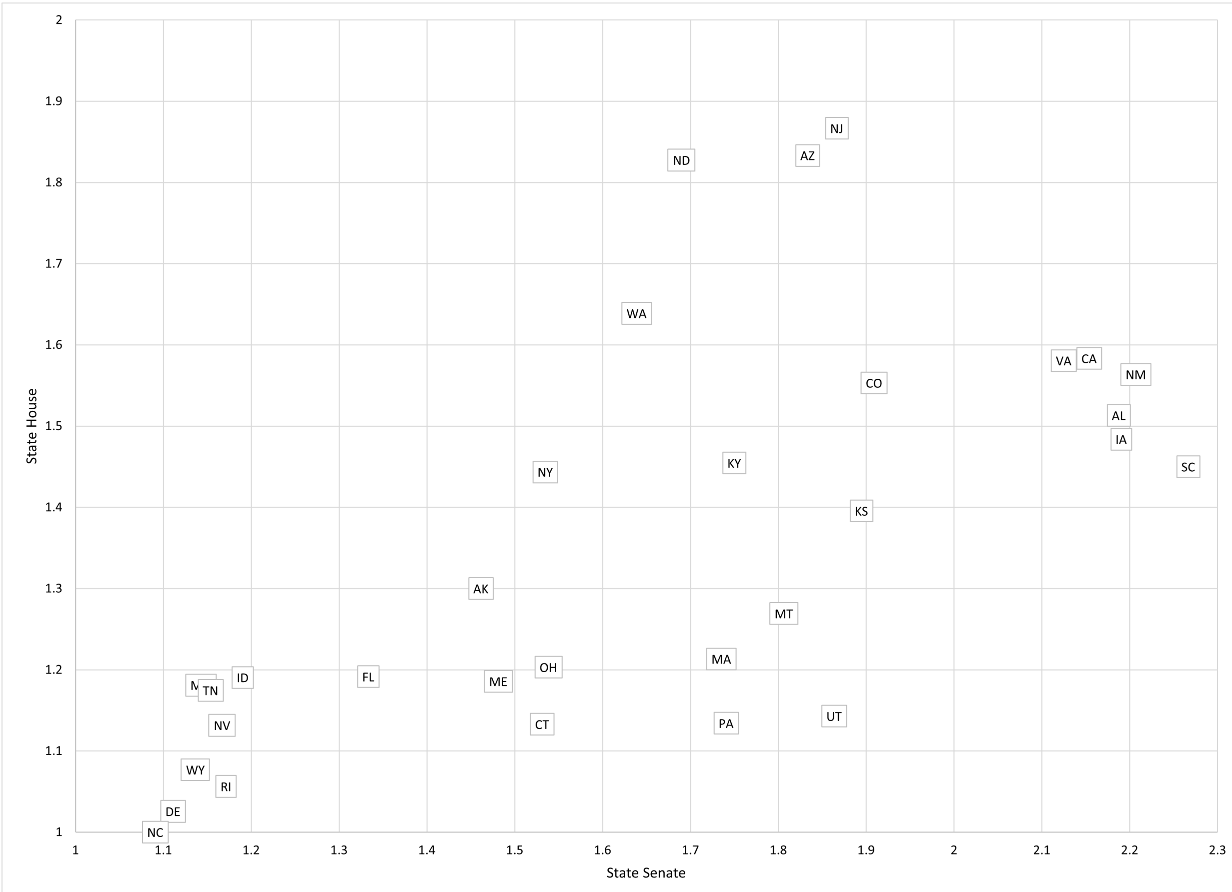

Figure 10 County Splits Ratios, Zoom-in to the State Senate window $[1, 2.3]$ and State House window $[1, 2]$.

We see that Texas's State House plan has a tiny county splits ratio of $101/100 = 1.01$, while its State Senate plan has a much larger ratio of $41/12 \approx 3.42$. To explain this, recall the *County Line Rule* from its constitution, which puts strict limits on county splitting in their House map, but not in their Senate map. Likewise, the *Whole County Provision* in North Carolina tightly constrains county splits, leading to ratios of $24/22 \approx 1.09$ and $80/80 = 1.00$ in their plan. Meanwhile, Illinois's constitution only specifies that legislative districts "shall be compact, contiguous and substantially equal in population," allowing their plans to be an outlier in Figure 9, with county splits ratios of $135/39 \approx 3.46$ and $220/87 \approx 2.53$, respectively. Arguably the worst offender is Wisconsin, which has county splits ratios of $73/13 \approx 5.62$ and $159/69 \approx 2.30$. Wisconsin is such an outlier in Figure 9 that many other states cannot be seen, so Figure 10 provides a zoomed-in version of this plot.

5.4. County Clusterings with 1-Person Deviation

In recent years, enacted congressional districts have often exhibited a 1-person deviation, i.e., all districts have a population between $L = \lfloor p(C)/k) \rfloor$ and $U = \lceil p(C)/k \rceil$. Researchers have speculated that, with such tight population balance constraints, the only possible county clustering is the trivial one where all counties belong to the same cluster (Autry et al. 2021, Nagle 2022). Is this actually true? To answer this question, it suffices to check the feasibility of a labeling MIP model (Validi and Buchanan 2022) with two clusters. Results are given in Table 4. Contrary to conventional wisdom, 79% of real-life districting instances admit a nontrivial county clustering. This includes *all* instances with more than 30 counties. One explanation for this surprising result is that the number of ways to partition $|C|$ counties into two clusters is $2^{|C|-1}$, which grows so rapidly that some of them are bound to be contiguous and satisfy a 1-person deviation. Indeed, when the number of counties exceeds 30, we have more than one billion options to choose from.

Table 4 Which congressional, state senate, and state house instances (CD, SS, SH) admit nontrivial county clusterings for 1-person deviation (excluding HI)? Black squares indicate uninteresting or nonexistent instances.

state	$\|C\|$	CD	SS	SH	state	$\|C\|$	CD	SS	SH	state	$\|C\|$	CD	SS	SH	state	$\|C\|$	CD	SS	SH
AK	30	■	✗	✗	IN	92	✓	✓	✓	ND	53	■	✓	✓	SD	66	■	✓	✓
AL	67	✓	✓	✓	KS	105	✓	✓	✓	NE	93	✓	✓	■	TN	95	✓	✓	✓
AR	75	✓	✓	✓	KY	120	✓	✓	✓	NH	10	✗	✗	✓	TX	254	✓	✓	✓
AZ	15	✗	✗	✗	LA	64	✓	✓	✓	NJ	21	✗	✗	✗	UT	29	✗	✓	✓
CA	58	✓	✓	✓	MA	14	✗	✗	✗	NM	33	✓	✓	✓	VA	133	✓	✓	✓
CO	64	✓	✓	✓	MD	24	✗	✗	✗	NV	17	✗	✗	✗	VT	14	■	✗	✓
CT	8	✗	✗	✓	ME	16	✗	✗	✓	NY	62	✓	✓	✓	WA	39	✓	✓	✓
DE	3	■	✗	✗	MI	83	✓	✓	✓	OH	88	✓	✓	✓	WI	72	✓	✓	✓
FL	67	✓	✓	✓	MN	87	✓	✓	✓	OK	77	✓	✓	✓	WV	55	✓	✓	✓
GA	159	✓	✓	✓	MO	115	✓	✓	✓	OR	36	✓	✓	✓	WY	23	■	✓	✓
IA	99	✓	✓	✓	MS	82	✓	✓	✓	PA	67	✓	✓	✓					
ID	44	✓	✓	✓	MT	56	✓	✓	✓	RI	5	✗	✗	✗					
IL	102	✓	✓	✓	NC	100	✓	✓	✓	SC	46	✓	✓	✓					

6. Conclusion and Future Work

In this paper, we consider the task of partitioning a state into k contiguous and population-balanced districts using a minimum number of county splits. Contrary to conventional wisdom, the minimum number of county splits is generally not equal to $k-1$. However, as observed by Carter et al. (2020), there is a close connection between this problem and the maximum county clustering problem. Indeed, we have a powerful tool in weak split duality, which says that the minimum number of county splits s is at least $k-c$, where c is the maximum number of county clusters. Carter et al. posited that strong split duality holds as well, which we have empirically confirmed over a large set of real-life districting instances. This is despite the fact that strong split duality generally does not hold, and the split duality gap $g = s - (k-c)$ can be arbitrarily large.

To solve the minimum county splits problem, we propose a three-step *Cluster-Sketch-Detail* approach, which heavily exploits integer programming techniques. Ultimately, it provides the first exact approach for the minimum county splits problem and the first answers to the question "how many county splits are mathematically necessary to satisfy basic criteria like contiguity and population balance?" across a large set of real-life districting instances. The answers could be used to challenge "bad" maps in court for states that place strict legal requirements on the number of county splits.

We do *not* claim that the computer-generated maps from our experiments are "good" or that they should be enacted in practice. They do not consider the Voting Rights Act (VRA) or any laws that vary by state. We also do not evaluate the partisan performance of the computer-generated maps nor of county splitting restrictions in general; for more on these subjects, see Chen et al. (2013), Gurnee and Shmoys (2021), Kenny et al. (2023), Duchin and Schoenbach (2023). However, as also suggested Carter et al. (2020), it is possible for map drawers to use the maximum county clusterings as starting points and incorporate the VRA, state-specific laws, or other criteria in the *Sketch* and *Detail* steps. To achieve a minimum number of county splits, all that is required is to limit each cluster to $k'-1$ county splits, where k' is the cluster's size. The maximum county clusterings generated in our experiments are typically not unique, and using other maximum county clusterings adds another layer of flexibility to the map-drawing process. Enumerating them is an interesting question for future research. In some cases, legal or political considerations not considered in this paper may make it impossible to achieve the minimum number of county splits reported here; in these cases, the results provided here may still give insightful baselines for comparison. In future research, it would be interesting to investigate to what extent the VRA or other constraints force a larger number of county splits. Understanding the precise tradeoffs between the number of county splits and other legitimate goals is an interesting task for future work.

Endnotes

1. First, for every edge $\{a,b\}$, we apply the inequality for $S = N(\{a,b\})$ (if this S satisfies the theorem's conditions), where $N(\cdot)$ denotes the open neighborhood. Second, for every vertex b and every subset $N' \subseteq N[b]$ that contains b, we apply the inequality for $S = N(N')$ (if this S satisfies the theorem's conditions), where $N[\cdot]$ denotes the closed neighborhood. Empirically, we have found these sets S to be most helpful, cf. Oehrlein and Haunert (2017). Also, for these sets S, it is relatively straightforward to check the theorem's conditions.

2. In an easy case, each of Arizona's 30 State House districts elects two members, so we simply use $k = 30$. We apply the same approach to the state house districts of Idaho, New Jersey, North Dakota, and Washington. However, a few other states have multi-member districts of varying size, which we handle as follows. New Hampshire's State House has multi-member districts with sizes varying between 1 to 11 members, with some of the districts being *floterial* (meaning that they "float" above other districts), so we simplify things by setting $k = 400$ equal to the number of seats. We apply the same approach to Vermont's State Senate and State House. Maryland's State House has 47 multi-member districts with three members in each, some of which are subdivided into three single-member districts or into a two-member district and a single-member district; we simply set $k = 47$. Similarly, South Dakota's State House has 35 multi-member districts with two members in each, two of which are subdivided into single-member districts; we simply set $k = 35$.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1942065. This analysis was conducted using data from the Redistricting Data Hub. We thank Daryl DeFord for sharing the districting instances as `json` files. We thank Hamidreza Validi for helpful comments. We thank two anonymous referees whose comments helped to improve the presentation.

References

Adler WT, Wang SSH (2019) Response to Cho and Liu, "Sampling from complicated and unknown distributions: Monte Carlo and Markov chain Monte Carlo methods for redistricting". *Physica A: Statistical Mechanics and its Applications* 516:591–593.

Alès Z, Knippel A (2020) The k-partitioning problem: Formulations and branch-and-cut. *Networks* 76(3):323–349.

Altman M (1997) The computational complexity of automated redistricting: Is automation the answer? *Rutgers Computer & Tech. LJ* 23:81.

Altman M, McDonald MP, et al. (2011) BARD: Better automated redistricting. *Journal of Statistical Software* 42(4):1–28.

Arredondo V, Martínez-Panero M, Peña T, Ricca F (2021) Mathematical political districting taking care of minority groups. *Annals of Operations Research* 305(1):375–402.

Autry EA, Carter D, Herschlag GJ, Hunter Z, Mattingly JC (2021) Metropolized multiscale forest recombination for redistricting. *Multiscale Modeling & Simulation* 19(4):1885–1914.

Balinski ML, Young HP (2010) *Fair representation: meeting the ideal of one man, one vote* (Brookings Institution Press), second edition.

Ballotpedia (2023) State legislative districts. `https://ballotpedia.org/State_Legislative_Districts`, accessed: 2023-02-17.

Becker A, Gold D (2022) The gameability of redistricting criteria. *Journal of Computational Social Science* 5:1735–1777.

Becker A, Solomon J (2022) Redistricting algorithms. Duchin M, Walch O, eds., *Political Geometry: Rethinking Redistricting in the US with Math, Law, and Everything In Between* (Birkhauser).

Belotti P, Buchanan A, Ezazipour S (2023) Political districting to optimize the Polsby-Popper compactness score. Available on Optimization-Online.

Birge JR (1983) Redistricting to maximize the preservation of political boundaries. *Social Science Research* 12(3):205–214.

Borndörfer R, Ferreira CE, Martin A (1998) Decomposing matrices into blocks. *SIAM Journal on Optimization* 9(1):236–269.

Bozkaya B, Erkut E, Laporte G (2003) A tabu search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research* 144(1):12–26.

Bradley PS, Bennett KP, Demiriz A (2000) Constrained k-means clustering. Technical Report MSR-TR-2000-65, Microsoft Research, Redmond, WA.

Bullock III CS (2010) *Redistricting: The most political activity in America* (Rowman & Littlefield Publishers).

Campêlo M, Campos VA, Corrêa RC (2008) On the asymmetric representatives formulation for the vertex coloring problem. *Discrete Applied Mathematics* 156(7):1097–1111.

Carter D, Hunter Z, Teague D, Herschlag G, Mattingly J (2020) Optimal legislative county clustering in North Carolina. *Statistics and Public Policy* 7(1):19–29.

Carvajal R, Constantino M, Goycoolea M, Vielma JP, Weintraub A (2013) Imposing connectivity constraints in forest planning models. *Operations Research* 61(4):824–836.

Cervas J (2022) Report of the Special Master, *Harkenrider v. Hochul*. `http://jonathancervas.com/2022/NY/CERVAS-SM-NY-2022.pdf`.

Cervas JR, Grofman B (2020) Tools for identifying partisan gerrymandering with an application to congressional districting in Pennsylvania. *Political Geography* 76:102069.

Chen J, Rodden J, et al. (2013) Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science* 8(3):239–269.

Cho WKT, Liu YY (2018) Sampling from complicated and unknown distributions: Monte Carlo and Markov chain Monte Carlo methods for redistricting. *Physica A: Statistical Mechanics and its Applications* 506:170–178.

Chopra S, Filipecki B, Lee K, Ryu M, Shim S, Van Vyve M (2017) An extended formulation of the convex recoloring problem on a tree. *Mathematical Programming* 165(2):529–548.

Clelland J, Colgate H, DeFord D, Malmskog B, Sancier-Barbosa F (2022) Colorado in context: Congressional redistricting and competing fairness criteria in Colorado. *Journal of Computational Social Science* 5:189–226.

Cohen-Addad V, Klein PN, Young NE (2018) Balanced centroidal power diagrams for redistricting. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 389–396 (ACM).

Cova TJ, Church RL (2000) Contiguity constraints for single-region site search problems. *Geographical Analysis* 32(4):306–329.

Davis M, Strigari F, Underhill W, Wice JM, Zamarripa C (2019) *Redistricting Law 2020* (National Conference of State Legislatures).

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

DeFord D, Duchin M, Solomon J (2021) Recombination: A family of Markov chains for redistricting. *Harvard Data Science Review* 3(1).

DRA (2023) Dave's redistricting. `https://davesredistricting.org/`, accessed: 2023-02-17.

Duchin M, Schoenbach G (2023) Redistricting for proportionality. *The Forum* 20(3-4):371–393.

Duchin M, Walch O, eds. (2022) *Political Geometry: Rethinking Redistricting in the US with Math, Law, and Everything In Between* (Birkhauser).

Dyer ME, Frieze AM (1985) On the complexity of partitioning graphs into connected subgraphs. *Discrete Applied Mathematics* 10(2):139–153.

Ferreira CE, Martin A, de Souza CC, Weismantel R, Wolsey LA (1996) Formulations and valid inequalities for the node capacitated graph partitioning problem. *Mathematical Programming* 74(3):247–266.

Fifield B, Higgins M, Imai K, Tarr A (2015) A new automated redistricting simulator using Markov chain Monte Carlo. *Work. Pap., Princeton Univ., Princeton, NJ* .

Fischetti M, Leitner M, Ljubić I, Luipersbeck M, Monaci M, Resch M, Salvagnin D, Sinnl M (2017) Thinning out Steiner trees: a node-based model for uniform edge costs. *Mathematical Programming Computation* 9(2):203–229.

Garfinkel RS, Nemhauser GL (1970) Optimal political districting by implicit enumeration techniques. *Management Science* 16(8):B–495.

Gladkova T, Goldbloom-Helzner A, Khan M, Kolstoe B, Noory J, Schutzman Z, Stucky E, Weighill T (2019) Discussion of locality splitting measures. `https://github.com/vrdi/splitting/blob/master/SplittingReport.pdf`.

Goderbauer S, Winandy J (2018) Political districting problem: Literature review and discussion with regard to federal elections in Germany. URL `https://www.or.rwth-aachen.de/files/research/repORt/LitSurvey_PoliticalDistricting__Goderbauer_Winandy_20181024.pdf`.

Grofman B (1985) Criteria for districting: A social science perspective. *UCLA L. Rev.* 33:77.

Guo D, Jin H (2011) iRedistrict: Geovisual analytics for redistricting optimization. *Journal of Visual Languages & Computing* 22(4):279–289.

Gurnee W, Shmoys DB (2021) Fairmandering: A column generation heuristic for fairness-optimized political districting. *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA21)*, 88–99 (SIAM).

Gutiérrez-Andrade MÁ, Rincón-García EA, de-los Cobos-Silva SG, Lara-Velázquez P, Mora-Gutiérrez RA, Ponsich A (2019) Simulated annealing and artificial bee colony for the redistricting process in Mexico. *INFORMS Journal on Applied Analytics* 49(3):189–200.

Hebert JG, Vandenberg ME, Smith P (2010) *The Realist's Guide to Redistricting: Avoiding the Legal Pitfalls* (American Bar Association).

Herschlag G, Kang HS, Luo J, Graves CV, Bangia S, Ravier R, Mattingly JC (2020) Quantifying gerrymandering in North Carolina. *Statistics and Public Policy* 7(1):30–38.

Hess S, Weaver J, Siegfeldt H, Whelan J, Zitlau P (1965) Nonpartisan political redistricting by computer. *Operations Research* 13(6):998–1006.

Kenny CT, McCartan C, Simko T, Kuriwaki S, Imai K (2023) Widespread partisan gerrymandering mostly cancels nationally, but reduces electoral competition. *Proceedings of the National Academy of Sciences* 120(25):e2217322120.

Kim M, Xiao N (2017) Contiguity-based optimization models for political redistricting problems. *International Journal of Applied Geospatial Research (IJAGR)* 8(4):1–18.

Kim MJ (2019) Give-and-take heuristic model to political redistricting problems. *Spatial Information Research* 27:539–552.

King DM, Jacobson SH, Sewell EC (2015) Efficient geo-graph contiguity and hole algorithms for geographic zoning and dynamic plane graph partitioning. *Mathematical Programming* 149(1-2):425–457.

King DM, Jacobson SH, Sewell EC (2018) The geo-graph in practice: creating United States congressional districts from census blocks. *Computational Optimization and Applications* 69(1):25–49.

King DM, Jacobson SH, Sewell EC, Cho WKT (2012) Geo-graphs: an efficient model for enforcing contiguity and hole constraints in planar graph partitioning. *Operations Research* 60(5):1213–1228.

Levin HA, Friedler SA (2019) Automated congressional redistricting. *Journal of Experimental Algorithmics (JEA)* 24(1):1–10.

Levitt J (2010) *A citizen's guide to redistricting* (Brennan Center for Justice at New York University School of Law).

Liu YY, Cho WKT, Wang S (2016) PEAR: a massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm and Evolutionary Computation* 30:78–92.

McCartan C, Imai K (2023) Sequential Monte Carlo for sampling balanced and compact redistricting plans. *Annals of Applied Statistics* To appear.

McCartan C, Kenny C, Simko T, Kuriwaki S, Garcia III G, Wang K, Wu M, Imai K (2022a) ALARM project. `https://alarm-redist.github.io/fifty-states/`, accessed: 2022-02-10.

McCartan C, Kenny CT, Simko T, Garcia III G, Wang K, Wu M, Kuriwaki S, Imai K (2022b) Simulated redistricting plans for the analysis and evaluation of redistricting in the United States. *Scientific Data* 9(1):689.

Mehrotra A, Johnson EL, Nemhauser GL (1998) An optimization based heuristic for political districting. *Management Science* 44(8):1100–1114.

MGGG (2023) GerryChain 0.2.20. `https://gerrychain.readthedocs.io/en/latest/`.

Miller S (2007) The problem of redistricting: the use of centroidal Voronoi diagrams to build unbiased congressional districts. *Senior project, Whitman College* .

Nagle J (2022) Euler's formula determines the minimum number of splits in maps of election districts. *Available at SSRN 4115039* .

NCSL (2021) Redistricting criteria. `http://www.ncsl.org/research/redistricting/redistricting-criteria.aspx`, accessed: 2023-02-17.

Norman SK, Camm JD (2003) The Kentucky redistricting problem: Mixed-integer programming model. Technical Report 03-04, College of Business Administration, Northern Arizona University.

Oehrlein J, Haunert JH (2017) A cutting-plane method for contiguity-constrained spatial aggregation. *Journal of Spatial Information Science* 2017(15):89–120.

Olson B (2022) Impartial automatic redistricting. `https://bdistricting.com/2020/`, accessed: 2023-02-17.

Önal H, Patrick KT (2016) A mathematical programming approach to political redistricting with compactness and community integrity considerations. Technical report, University of Illinois at Urbana-Champaign.

Princeton Gerrymandering Project (2023) Redistricting report card. `https://gerrymander.princeton.edu/redistricting-report-card/`, accessed: 2023-02-17.

Redistricting Data Hub (2021a) Download Data. `https://redistrictingdatahub.org/data/download-data/`.

Redistricting Data Hub (2021b) States that adjust the census data for redistricting. `https://redistrictingdatahub.org/data/ongoing-data-projects/states-that-adjust-the-census-data-for-redistricting/`.

Ricca F, Scozzari A, Simeone B (2008) Weighted Voronoi region algorithms for political districting. *Mathematical and Computer Modelling* 48(9-10):1468–1477.

Ricca F, Scozzari A, Simeone B (2013) Political districting: from classical models to recent approaches. *Annals of Operations Research* 204(1):271–299.

Ricca F, Simeone B (2008) Local search algorithms for political districting. *European Journal of Operational Research* 189(3):1409–1426.

Shirabe T (2005) A model of contiguity for spatial unit allocation. *Geographical Analysis* 37(1):2–16.

Shirabe T (2009) Districting modeling with exact contiguity constraints. *Environment and Planning B: Planning and Design* 36(6):1053–1066.

Svec L, Burden S, Dilley A (2007) Applying Voronoi diagrams to the redistricting problem. *The UMAP Journal* 28(3):313–329.

Swamy R, King DM, Jacobson SH (2022) Multiobjective optimization for politically fair districting: A scalable multilevel approach. *Operations Research* To appear.

US Census Bureau (1994) Census tracts and block numbering areas. *Geographic Areas Reference Manual* (US Department of Commerce Washington, DC).

US Census Bureau (2021a) Decennial Census P.L. 94-171 Redistricting Data. `https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html`.

US Census Bureau (2021b) Disclosure avoidance for the 2020 Census: An introduction.

US Census Bureau (2021c) TIGER/Line Shapefiles. `https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html`.

US Census Bureau (2022) State legislative districts. `https://www.census.gov/programs-surveys/decennial-census/about/rdo/state-legislative-district.html`.

US Census Bureau (2023) Congressional districts. `https://www.census.gov/programs-surveys/decennial-census/about/rdo/congressional-districts.html`.

Validi H, Buchanan A (2022) Political districting to minimize cut edges. *Mathematical Programming Computation* 14:623–672.

Validi H, Buchanan A, Lykhovyd E (2022) Imposing contiguity constraints in political districting models. *Operations Research* 70(2):867–892.

Vickrey W (1961) On the prevention of gerrymandering. *Political Science Quarterly* 76(1):105–110.

Wachspress J, Adler WT (2021) Split decisions: Guidance for measuring locality preservation in district maps. Center for Democracy & Technology, URL `https://cdt.org/wp-content/uploads/2021/11/2021-11-04-Locality-splitting-report-final.pdf`.

Wang Y, Buchanan A, Butenko S (2017) On imposing connectivity constraints in integer programs. *Mathematical Programming* 166(1-2):241–271.

Zoltners AA, Sinha P (1983) Sales territory alignment: A review and model. *Management Science* 29(11):1237–1256.

Appendix A: Full MIP for Sketch

For the *Sketch* step, the full MIP is as follows, where the objective is to maximize the number of preserved edges. Again, for notational simplicity, this MIP is written across the entire set of counties C with size k, but in practice it should be applied to each county cluster separately.

$$
\begin{aligned}
\max \quad & \sum_{e \in E(C)} \sum_{j=1}^{k} y_{ej} & \\
\text{s.t.} \quad & y_{ej} \geq \sum_{i \in e} x_{ij} - 1 & \forall e \in E(C),\ \forall j \in [k] \\
& \sum_{j=1}^{k} y_{ej} \leq 1 & \forall e \in E(C) \\
& y_{ej} \leq x_{ij} & \forall i \in e \in E(C),\ \forall j \in [k] \\
& \sum_{i \in C} s_i = k - 1 & \\
& \sum_{j=1}^{k} x_{ij} = s_i + 1 & \forall i \in C \\
& \sum_{j=1}^{k} z_{ij} = 1 & \forall i \in C \\
& L \leq \sum_{i \in C} p_i z_{ij} \leq U & \forall j \in [k] \\
& x_{ij} \in \{0,1\} & \forall i \in C,\ \forall j \in [k] \\
& y_{ej} \in \{0,1\} & \forall e \in E(C),\ \forall j \in [k] \\
& 0 \leq z_{ij} \leq x_{ij} & \forall i \in C,\ \forall j \in [k].
\end{aligned}
$$

After solving this MIP, we set the number of preserved edges equal to its optimal objective value and apply a secondary objective to minimize the sum of squares of counties touched by the districts

$$\min \ \sum_{j=1}^{k} \sum_{t=1}^{|C|} t^2 b_{jt},$$

which requires the following additional constraints:

$$
\begin{aligned}
& \sum_{i \in C} x_{ij} = \sum_{t=1}^{|C|} t b_{jt} & \forall j \in [k] \\
& \sum_{t=1}^{|C|} b_{jt} = 1 & \forall j \in [k] \\
& b_{jt} \in \{0,1\} & \forall j \in [k],\ \forall t = 1, 2, \ldots, |C|.
\end{aligned}
$$

Appendix B: Final Results for Congress, State Senate, and State House

Table 5: Final Congressional Results (excludes HI and trivial states).

state	$\lvert C\rvert$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
AL	67	7	714,166	721,342	0	7	0	0	6
AR	75	4	749,117	756,645	0	4	0	0	3
AZ	15	9	790,639	798,584	6	2	7	7	15
CA	58	52	756,549	764,152	39	11	41	41	72
CO	64	8	718,106	725,322	1	6	2	2	20
CT	8	5	717,583	724,794	3	1	4	4	10
FL	67	28	765,375	773,067	10	9	19	19	31
GA	159	14	761,311	768,961	2	12	2	2	21
IA	99	4	793,605	801,580	0	4	0	0	0
ID	44	2	914,956	924,150	0	2	0	0	1
IL	102	17	749,909	757,445	7	8	9	9	53
IN	92	9	750,178	757,717	1	8	1	1	8
KS	105	4	730,798	738,142	0	4	0	0	4
KY	120	6	747,218	754,727	1	5	1	1	6
LA	64	6	772,412	780,174	0	6	0	0	15
MA	14	9	777,197	785,007	5	2	7	7	22
MD	24	8	768,293	776,013	3	4	4	4	9
ME	16	2	677,774	684,585	0	2	0	0	1
MI	83	13	771,304	779,055	4	9	4	4	21
MN	87	8	709,746	716,878	1	6	2	2	12
MO	115	8	765,518	773,210	1	7	1	1	10
MS	82	4	736,619	744,021	0	4	0	0	4
MT	56	2	539,402	544,823	0	2	0	0	1
NC	100	14	741,943	749,398	2	11	3	3	13
NE	93	3	650,566	657,103	0	3	0	0	2

Continued on next page

Table 5 – continued from previous page

state	$\lvert C \rvert$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
NH	10	2	685,321	692,208	0	1	1	1	5
NJ	21	12	770,213	777,953	3	3	9	9	20
NM	33	3	702,312	709,369	0	3	0	0	10
NV	17	4	772,273	780,034	2	2	2	2	5
NY	62	26	773,087	780,855	13	8	18	18	26
OH	88	15	782,697	790,563	3	11	4	4	14
OK	77	5	787,912	795,829	1	4	1	1	7
OR	36	6	702,679	709,740	1	5	1	1	16
PA	67	17	761,041	768,689	4	10	7	7	17
RI	5	2	545,947	551,432	1	1	1	1	1
SC	46	7	727,548	734,859	0	6	1	1	10
TN	95	9	764,032	771,710	1	7	2	2	11
TX	254	38	763,153	770,821	19	19	19	19	59
UT	29	4	813,815	821,993	1	3	1	1	7
VA	133	11	780,749	788,595	1	9	2	2	11
WA	39	10	766,676	774,380	4	6	4	4	11
WI	72	8	733,032	740,398	1	7	1	1	13
WV	55	2	892,374	901,342	0	2	0	0	0

Table 6: Final State Senate Results (excludes HI).

state	$\|C\|$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
AK	30	20	34,837	38,503	12	7	13	13	19
AL	67	35	136,374	150,728	13	19	16	16	35
AR	75	35	81,742	90,345	14	21	14	14	51
AZ	15	30	226,465	250,302	22	6	24	24	44
CA	58	40	939,033	1,037,878	23	14	26	26	56
CO	64	35	156,716	173,211	22	13	22	22	42
CT	8	36	95,157	105,173	31	4	32	32	49
DE	3	21	44,784	49,497	18	3	18	18	20
FL	67	40	511,532	565,377	18	16	24	24	32
GA	159	56	181,720	200,848	23	31	25	25	60
IA	99	50	60,618	66,997	20	29	21	21	46
ID	44	35	49,919	55,173	19	14	21	21	25
IL	102	59	206,304	228,019	39	20	39	39	135
IN	92	50	128,926	142,496	20	28	22	22	48
KS	105	40	69,775	77,119	19	21	19	19	36
KY	120	38	112,646	124,503	11	26	12	12	21
LA	64	39	113,459	125,401	20	18	21	21	77
MA	14	40	166,961	184,535	31	6	34	34	59
MD	24	47	124,859	138,001	35	10	37	37	45
ME	16	35	36,979	40,870	24	8	27	27	40
MI	83	38	251,934	278,452	19	18	20	20	64
MN	87	67	80,913	89,430	38	26	41	41	100
MO	115	34	171,976	190,078	14	20	14	14	16
MS	82	52	54,101	59,795	19	29	23	23	64
MT	56	50	20,601	22,768	30	19	31	31	56
NC	100	50	198,349	219,227	20	28	22	22	24
ND	53	47	15,748	17,405	28	18	29	29	49

Continued on next page

Table 6 – continued from previous page

state	$\|C\|$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
NE	93	49	38,030	42,032	26	21	28	28	37
NH	10	24	54,528	60,266	19	4	20	20	40
NJ	21	40	220,614	243,836	28	10	30	30	56
NM	33	42	47,897	52,938	28	13	29	29	64
NV	17	21	140,447	155,230	17	3	18	18	21
NY	62	63	304,623	336,687	42	20	43	43	66
OH	88	33	339,682	375,436	12	20	13	13	20
OK	77	48	78,363	86,610	22	25	23	23	59
OR	36	30	134,180	148,303	17	11	19	19	47
PA	67	50	247,052	273,056	26	23	27	27	47
RI	5	38	27,435	30,322	33	3	35	35	41
SC	46	46	105,707	116,833	26	16	30	30	68
SD	66	35	24,067	26,600	17	16	19	19	29
TN	95	33	198,949	219,890	13	20	13	13	15
TX	254	31	893,169	987,186	12	19	12	12	41
UT	29	29	107,174	118,455	22	7	22	22	41
VA	133	40	204,996	226,574	16	24	16	16	34
VT	14	30	20,365	22,507	23	6	24	24	18
WA	39	49	149,389	165,113	34	13	36	36	59
WI	72	33	169,668	187,527	12	20	13	13	73
WV	55	17	100,238	110,788	2	13	4	4	13
WY	23	31	17,678	19,538	21	9	22	22	25

Table 7: Final State House Results (excludes HI).

state	$\|C\|$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
AK	30	40	17,419	19,251	28	10	30	30	39
AL	67	105	45,458	50,242	71	29	76	76	115
AR	75	100	28,610	31,621	61	33	67	67	128
AZ	15	30	226,465	250,302	22	6	24	24	44
CA	58	80	469,517	518,939	57	20	60	60	95
CO	64	65	84,386	93,267	46	18	47	47	73
CT	8	151	22,687	25,074	139	8	143	143	162
DE	3	41	22,938	25,352	38	2	39	39	40
FL	67	120	170,511	188,459	89	26	94	94	112
GA	159	180	56,536	62,486	118	57	123	123	209
IA	99	100	30,309	33,498	54	38	62	62	92
ID	44	35	49,919	55,173	19	14	21	21	25
IL	102	118	103,152	114,009	85	31	87	87	220
IN	92	100	64,463	71,248	55	39	61	61	129
KS	105	125	22,328	24,678	87	34	91	91	127
KY	120	100	42,806	47,311	47	45	55	55	80
LA	64	105	42,142	46,577	72	29	76	76	116
MA	14	160	41,741	46,133	145	10	150	150	182
MD	24	47	124,859	138,001	35	10	37	37	67
ME	16	151	8,572	9,473	137	11	140	140	166
MI	83	110	87,032	96,192	73	32	78	78	154
MN	87	134	40,457	44,715	92	34	100	100	176
MO	115	163	35,873	39,648	110	47	116	116	137
MS	82	122	23,060	25,486	79	36	86	86	181
MT	56	100	10,301	11,384	74	22	78	78	99
NC	100	120	82,646	91,344	71	40	80	80	80
ND	53	47	15,748	17,405	28	18	29	29	53

Continued on next page

Table 7 – continued from previous page

state	$\lvert C\rvert$	k	L	U	obvious LB	max clusters	split LB	min splits	enacted splits
NH	10	400	3,272	3,616	375	10	390	390	154
NJ	21	40	220,614	243,836	28	10	30	30	56
NM	33	70	28,738	31,762	52	15	55	55	86
NV	17	42	70,224	77,615	35	4	38	38	43
NY	62	150	127,942	141,408	115	26	124	124	179
OH	88	99	113,228	125,145	57	35	64	64	77
OK	77	101	37,242	41,161	67	30	71	71	134
OR	36	60	67,090	74,151	44	13	47	47	79
PA	67	203	60,851	67,255	158	39	164	164	186
RI	5	75	13,901	15,363	70	4	71	71	75
SC	46	124	39,214	43,341	94	24	100	100	145
SD	66	35	24,067	26,600	17	16	19	19	31
TN	95	99	66,317	73,296	55	36	63	63	74
TX	254	150	184,589	204,018	99	50	100	100	101
UT	29	75	41,441	45,802	59	12	63	63	72
VA	133	100	81,999	90,629	55	38	62	62	98
VT	14	150	4,073	4,501	137	12	138	138	118
WA	39	49	149,389	165,113	34	13	36	36	59
WI	72	99	56,556	62,509	63	30	69	69	159
WV	55	100	17,041	18,834	70	24	76	76	89
WY	23	62	8,839	9,769	49	10	52	52	56