Noname manuscript No.
(will be inserted by the editor)

Linear-size formulations for connected planar graph partitioning and political districting

Jack Zhang · Hamidreza Validi · Austin Buchanan · Illya V. Hicks

Received: date / Accepted: date

Abstract Motivated by applications in political districting, we consider the task of partitioning the n vertices of a planar graph into k connected components. We propose an extended formulation for this task that has two desirable properties: (i) it uses just $O(n)$ variables, constraints, and nonzeros, and (ii) it is *perfect*. To explore its ability to solve real-world problems, we apply it to a political districting problem in which contiguity and population balance are imposed as hard constraints and compactness is optimized. Computational experiments show that, despite the model's small size and integrality for connected partitioning, the population balance constraints are more troublesome to effectively impose. Nevertheless, we share our findings in hopes that others may find better ways to impose them.

1 Introduction

In political districting, the task is to partition a state into a given number of districts for elections. Two basic criteria are that districts have roughly equal populations and be contiguous on the map. There are many other traditional and emerging redistricting principles that can guide the redistricting process.

Jack Zhang
Operations Research Center, Massachusetts Institute of Technology
E-mail: jackz111@mit.edu

Hamidreza Validi
Industrial, Manufacturing & Systems Engineering, Texas Tech University
E-mail: hvalidi@ttu.edu

Austin Buchanan
Industrial Engineering & Management, Oklahoma State University
E-mail: buchanan@okstate.edu

Illya V. Hicks
Computational Applied Mathematics and Operations Research, Rice University
E-mail: ivhicks@rice.edu

In the operations research literature, districting problems are usually cast in terms of a graph $G = (V, E)$ whose vertices represent geographic units (e.g., counties, census tracts, or census blocks) and whose edges indicate which pairs of geographic units are adjacent on the map. Each geographic unit $i \in V$ has an associated population p_i. The task is to partition the vertices into k districts $(V_1, V_2, \ldots, V_k)$ so that each district V_j induces a connected subgraph $G[V_j]$ and has a population $p(V_j) := \sum_{i \in V_j} p_i$ that is near to the ideal $p(V)/k$.

For an integer programmer, the question is how to best capture these constraints in a mathematical model. A common refrain among researchers is that the contiguity constraints pose a "major challenge" [28], are "particularly difficult to deal with" [23], and make districting "much more difficult than other partitioning problems" [24]. Another huge challenge is the sheer size of redistricting instances [31]. In many states, the number of census tracts exceeds one thousand, and the number of census blocks approaches one million, thus n can be quite large. Classic integer programming models for districting include a capacitated k-median model that uses n^2 variables [15], a set partitioning model that uses exponentially many variables [10], and a labeling model that uses kn variables and suffers from symmetry [23, 13, 30].

These challenges with contiguity constraints and huge scale prompted us to develop something completely different. We propose an extended formulation for partitioning the n vertices of a planar graph into k connected components. The model has linear size, using just $O(n)$ variables, constraints, and nonzeros. Further, it is *perfect*, i.e., it projects down to the convex hull of feasible solutions. Key ingredients to the approach include the characterization of branching polyhedra due to Edmonds [6] and the linear-size formulation for spanning trees in planar graphs due to Williams [33], cf. [22, 29].

At a high level, the model works by selecting a spanning tree and picking $n-k$ edges from the tree's edges. This gives a forest with k components. In each component, a vertex is selected as its root, and the edges of the component are oriented away from it. Figure 1 illustrates the idea for a 5-by-5 grid instance that requires $k = 5$ districts.

Then, we propose a flow-based formulation to impose the population balance constraints. To evaluate its performance, we apply it to a stylized districting problem in which contiguity and population balance are imposed as hard constraints, and compactness is sought by minimizing the flow variables. Computational experiments show that, despite the model's small size and perfectness for connected partitioning, the population balance constraints become the new bottleneck. Nevertheless, we hope that, by sharing our experience here, other researchers may find better ways to impose the population balance constraints and make the approach more tractable.

2 Background and Literature Review

We consider a simple (undirected) graph $G = (V, E)$ with $n = |V|$ vertices and $m = |E|$ edges. Denote by $E(S)$ the subset of edges that have both endpoints

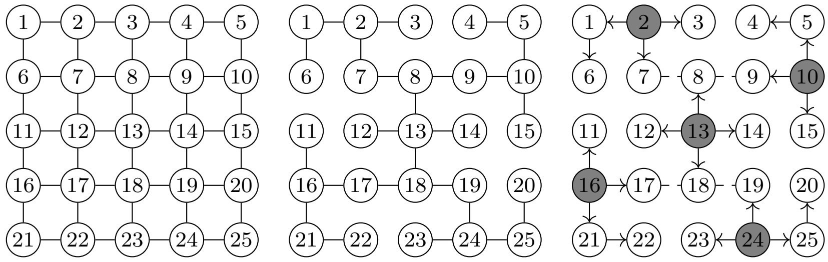

Fig. 1: 5×5 grid graph (left); a spanning tree (center); and 5 out-trees (right).

in $S \subseteq V$. We call $G = (V, E)$ a tree if it is connected and satisfies $|E| = |V| - 1$. A forest is a disjoint union of trees. A spanning tree is a tree subgraph over the same vertex set. Often, G will be planar, in which case its number of edges is linear with respect to its number of vertices, namely $m \leq 3n - 6$ if $n \geq 3$.

We also consider directed graphs $D = (V, A)$. Often, they are obtained from G by replacing each undirected edge $\{i, j\} \in E$ by two oppositely directed arcs (i, j) and (j, i). The subsets of arcs that point towards vertex i and away from vertex i are denoted by $\delta^-(i)$ and $\delta^+(i)$, respectively. Their sizes are called the indegree and outdegree. Denote by $A(S)$ the subset of arcs that have both endpoints in $S \subseteq V$. Associated with a directed graph $D = (V, A)$ is its *underlying undirected graph* which has the same vertex set and has the edge $\{i, j\}$ if at least one of the arcs (i, j) or (j, i) belongs to A.

An *out-tree* is a directed graph in which each vertex has indegree at most one and whose underlying graph is a tree. An out-tree has one vertex with indegree zero that is often called its *root*. An *out-forest* is a disjoint union of out-trees. A *spanning out-tree* is an out-tree subgraph that has the same vertex set. Terms like out-tree and out-forest are used by [11,1,4]. Elsewhere, spanning out-trees are often called arborescences, and out-forests are often called branchings, see [25, Ch. 52] and [18, Ch. 6].

Forests, spanning trees, and their directed counterparts are well-studied in the literature from the polyhedral perspective. For example, the forest polytope of undirected graph $G = (V, E)$, which is the convex hull of characteristic vectors of its forests, admits the polyhedral representation:

$$x(E(S)) \leq |S| - 1 \qquad \forall S \subseteq V,\ S \neq \varnothing \tag{1a}$$

$$x_e \geq 0 \qquad \forall e \in E, \tag{1b}$$

where $x(E(S))$ is shorthand for $\sum_{e \in E(S)} x_e$. Further, adding the constraint $x(E) = n - 1$ gives the spanning tree polytope, see [7] and [25, Ch. 50].

Similarly, the out-forest polytope of $D = (V, A)$, which is the convex hull of its out-forests, admits the following polyhedral representation [25, Ch. 52]:

$$x(A(S)) \leq |S| - 1 \qquad \forall S \subseteq V,\ S \neq \varnothing \tag{2a}$$

$$x(\delta^-(i)) \leq 1 \qquad \forall i \in V \tag{2b}$$

$$x_a \geq 0 \qquad \forall a \in A. \tag{2c}$$

By adding the constraint $x(A) = n - 1$ to system (2), we get the spanning out-tree polytope, i.e., the convex hull of out-forests with $n-1$ arcs. Generally, the convex hull of out-forests with $n - k$ arcs is obtained from system (2) by adding the cardinality constraint $x(A) = n - k$, see Corollary 52.6b of [25].

These polyhedral characterizations of (out-)forests and spanning (out-)trees are quite large as they use exponentially many cycle-elimination constraints (1a) and (2a). However, they admit much smaller representations. For example, there are extended formulations for spanning (out-)tree polytopes of arbitrary graphs that use $O(nm)$ variables, constraints, and nonzeros [34,19]. Meanwhile, when G is planar, there are extended formulations of size $O(n)$, as shown by Williams [33]. While Williams' original formulation had a flaw in the sense that it permitted (disconnected) solutions other than spanning trees, there is a simple fix [22,29]. Building on Williams' formulation, Fiorini et al. [8] provide extended formulations of size $O(g^{1/2}n^{3/2} + g^{3/2}n^{1/2})$ for spanning tree polytopes of genus g graphs. Later this was generalized by Aprile et al. [2] who give size $O(n^{1+\beta})$ extended formulations for spanning tree polytopes of graphs that admit 1/2-balanced separators of size $O(n^{\beta})$ with $0 < \beta < 1$.

3 Extended Formulation for Connected Partitioning

Before giving the linear-size extended formulation for connected partitioning in planar graphs, we first give an exponentially large formulation for selecting k disjoint out-trees. It is obtained from the out-forest polytope (2) by adding a cardinality constraint. Further, to capture which vertices are roots (which will be important in the districting context), we introduce slack variables s_i for the indegree constraints (2b) which gives:

$$x(A(S)) \leq |S| - 1 \qquad \forall S \subseteq V,\ S \neq \varnothing \tag{3a}$$

$$x(\delta^-(i)) + s_i = 1 \qquad \forall i \in V \tag{3b}$$

$$x(A) = n - k \tag{3c}$$

$$x_a \geq 0 \qquad \forall a \in A \tag{3d}$$

$$s_i \geq 0 \qquad \forall i \in V. \tag{3e}$$

This formulation selects an out-forest with $n-k$ arcs, thus giving k disjoint out-trees. Further, system (3) defines an integral polyhedron, as introducing slack variables in (3b) preserves integrality by the following folklore lemma.

Lemma 1 *Suppose the polytope $P = \{x \mid Ax \leq b, Cx = d, x \geq 0\}$ is integral. If A and b are integral, then introducing slack variables preserves integrality, i.e., $Q = \{(x, s) \mid Ax + s = b, Cx = d, x \geq 0, s \geq 0\}$ is integral.*

Proof Let (x^*, s^*) be an extreme point of Q. We are to show that (x^*, s^*) is integral. First, we claim that x^* is an extreme point of P. If not, then x^* can be written as a strict convex combination of points in P, i.e., $x^* = \sum_{i=1}^{q} \lambda_i x^i$ and $\sum_{i=1}^{q} \lambda_i = 1$ for some λ strictly positive and $q \geq 2$. For each $i = 1, 2, \ldots, q$, let $s^i = b - Ax^i$. Then, we can write (x^*, s^*) as a strict convex combination of points (x^i, s^i) from Q as follows:

$$\begin{aligned}(x^*, s^*) = (x^*, b - Ax^*) &= (\sum_{i=1}^{q} \lambda_i x^i, b - A(\sum_{i=1}^{q} \lambda_i x^i)) \\ &= \sum_{i=1}^{q} \lambda_i (x^i, b - Ax^i) = \sum_{i=1}^{q} \lambda_i (x^i, s^i).\end{aligned}$$

This contradicts that (x^*, s^*) is an extreme point of Q, thus showing that x^* is an extreme point of P. Then, by assumption that P is integral, we have that x^* is integral, and so is $s^* = b - Ax^*$ as A and b are integral. We have thus shown that an arbitrary extreme point (x^*, s^*) of Q is integral, as desired. □

Proposition 1 summarizes our observations about the linear system (3).

Proposition 1 *The linear system* (3) *gives a correct formulation for selecting k disjoint out-trees of a directed graph $D = (V, A)$ and defines an integral polyhedron.*

Aside from the cycle-elimination constraints (3a), formulation (3) has just $O(n)$ variables, constraints, and nonzeros. To get the full formulation's size down to $O(n)$, we use Williams' spanning tree formulation $\mathscr{P}_{\text{Williams}}(G)$. Specifically, we impose that the spanning tree edge variables y satisfy the constraints of Williams' formulation, which we simply write as $y \in \mathscr{P}_{\text{Williams}}(G)$, and that the out-forest arc variables satisfy $x_{ij} + x_{ji} \leq y_e$. That is, the proposed formulation is as follows, where $G = (V, E)$ is a connected simple planar graph, and $D = (V, A)$ is obtained from G by bidirecting its edges, i.e., replacing each undirected edge $\{i, j\}$ by oppositely directed arcs (i, j) and (j, i).

$$\begin{aligned}
& y \in \mathscr{P}_{\text{Williams}}(G) && && \text{(4a)}\\
& x_{ij} + x_{ji} \leq y_e && \forall e = \{i, j\} \in E && \text{(4b)}\\
& x(\delta^-(i)) + s_i = 1 && \forall i \in V && \text{(4c)}\\
& x(A) = n - k && && \text{(4d)}\\
& s_i \geq 0 && \forall i \in V && \text{(4e)}\\
& x_a \geq 0 && \forall a \in A. && \text{(4f)}
\end{aligned}$$

Theorem 1 *The linear system* (4) *gives a correct formulation of size $O(n)$ for selecting k disjoint out-trees of the bidirected graph $D = (V, A)$ and projects to an integral polyhedron in the (s, x)-space of variables.*

Proof Denote by P as the set of all (s, x) satisfying the exponentially large formulation (3), and denote by Q as the set of all (s, x, y) satisfying the Williams-based extended formulation (4). By Proposition 1, P is a correct formulation and is integral. So, to prove the claim, we show that $P = \text{proj}_{s,x} Q$.

To prove $P \subseteq \operatorname{proj}_{s,x} Q$, it suffices to show that the extreme points of P belong to $\operatorname{proj}_{s,x} Q$. So, let $(\hat{s}, \hat{x})$ be an extreme point of P. By Proposition 1, $(\hat{s}, \hat{x})$ is integral and thus binary. The arcs $A' = \{(i,j) \in A \mid \hat{x}_{ij} = 1\}$ selected by $\hat{x}$ thus form an out-forest consisting of k out-trees. The underlying edge set

$$E' = \{\{i,j\} \in E \mid (i,j) \in A' \text{ or } (j,i) \in A'\}$$

induces a subgraph of G that is a forest with k components. Since G is connected, it admits a spanning tree (V, E'') such that $E' \subseteq E''$. Let $\hat{y}$ be the characteristic vector of E''. Since $\hat{y}$ represents the edge set of a spanning tree, we have $\hat{y} \in \mathscr{P}_{\text{Williams}}(G)$, and by construction of E' and E'' we have $\hat{x}_{ij} + \hat{x}_{ji} \leq \hat{y}_e$ for each edge $\{i,j\} \in E$. So, $(\hat{s}, \hat{x}, \hat{y})$ satisfies constraints (4a) and (4b). The other constraints from (4) also define P and are thus satisfied by the assumption that $(\hat{s}, \hat{x}) \in P$. So, $(\hat{s}, \hat{x}, \hat{y}) \in Q$ and thus $(\hat{s}, \hat{x}) \in \operatorname{proj}_{s,x} Q$.

To prove $\operatorname{proj}_{s,x} Q \subseteq P$, let $(\hat{s}, \hat{x}, \hat{y})$ be a point from Q. It suffices to show that it satisfies constraints (3a), so consider $S \subseteq V$ with $S \neq \varnothing$. We have

$$\hat{x}(A(S)) = \sum_{(i,j) \in A(S)} \hat{x}_{ij} \leq \sum_{e \in E(S)} \hat{y}_e = \hat{y}(E(S)) \leq |S| - 1,$$

where the first inequality holds by constraints (4b), and the last inequality holds by the fact that $\hat{y}$ belongs to $\mathscr{P}_{\text{Williams}}(G)$ which is equivalent to the spanning tree polytope, and any point $\hat{y}$ in the spanning tree polytope satisfies the cycle-elimination constraints $\hat{y}(E(S)) \leq |S| - 1$, see constraints (1a). So, $(\hat{s}, \hat{x})$ satisfies the constraints of (3) and thus belongs to P. $\square$

4 Application in Political Districting

Three of the most basic political districting criteria are population balance, contiguity, and compactness. The formulation (4) captures the contiguity constraints, which researchers have often suggested is the bottleneck for approaches based on integer programming. In this section, we consider how to incorporate population balance constraints and a compactness objective in the context of model (4). One goal is to preserve the model's linear size, $O(n)$. This task has eluded previous research. Even those who adopt Williams' spanning tree formulation arrive at models of size $\Theta(kn)$ or $\Theta(n^2)$, see [12, 16, 17].

In the USA, the population balance constraints are enforced quite strictly by the courts, see [14]. A common rule-of-thumb is that congressional districts should satisfy a 1% population deviation ($\pm 0.5\%$), meaning that each district should have a population between $L = \lceil 0.995 p(V)/k \rceil$ and $U = \lfloor 1.005 p(V)/k \rfloor$. However, to avoid litigation, many states draw their congressional districts to meet a 1-person deviation, $L = \lfloor p(V)/k \rfloor$ and $U = \lceil p(V)/k \rceil$. Other states like Iowa and West Virginia strive to keep all counties whole and thus may be able to justify the need for a larger deviation. For example, the US Supreme Court upheld West Virginia's 2010 districts which exhibited a 0.79% deviation [20].

To ensure population balance, we propose a flow-based formulation. The root of each out-tree serves as a source vertex that sends out flow to the other vertices in the same district. The flow generated at vertex i is represented by the variable g_i. Flow is only permitted to be sent across arcs of the out-forest, written using a flow variable f_{ij} for each arc $(i,j) \in A$ of the bidirected graph $D = (V, A)$. The population balance constraints are then written as follows.

$$Ls_i \leq g_i \leq Us_i \qquad \forall i \in V \tag{5a}$$
$$g_i + f(\delta^-(i)) = f(\delta^+(i)) + p_i \qquad \forall i \in V \tag{5b}$$
$$0 \leq f_{ij} \leq (U - p_i)x_{ij} \qquad \forall (i,j) \in A. \tag{5c}$$

Constraints (5a) state that if vertex $i \in V$ is selected as a root, then it generates a flow of between L and U for its district; otherwise, it generates zero flow. The flow-balance constraints (5b) ensure that each vertex $i \in V$ consumes a flow equal to its population p_i. By constraints (5c), flows are not sent across unselected arcs, with a big-M coefficient of $U - p_i$. Under these population balance constraints (5), the integrality of the s and x variables is certainly not guaranteed, so we will place explicit integrality restrictions on them. Technically, it suffices to impose integrality solely on x, as the integrality of s would follow by the indegree constraints (4c). However, our implementation defines the root variables s as binary, as they represent key decisions on which a MIP solver may find it advantageous to branch.

To favor districts that are compact in shape, we choose to minimize the total population flow, subject to contiguity and population balance constraints:

$$\min \left\{ f(A) \;\middle|\; (f, g, s, x, y) \text{ satisfies (4) and (5)},\ (s, x) \text{ binary} \right\}.$$

The population flow objective has the following interpretation. For each district, place its total population at its root. Have the people walk home along the edges of their district. The objective value is then the total number of edges crossed, aggregated over all people. Observe that once a district $S \subseteq V$ and root r have been chosen, an optimal out-tree for the district can be constructed as a shortest path tree of $G[S]$ rooted at r. To find the best out-tree for a district S, one can try all possible roots $r \in S$ and compute the resulting objectives. This is very similar to a k-median objective and the moment-of-inertia objective that are discussed next.

4.1 An existing MIP for comparison

To evaluate the computational performance of the proposed districting model, we compare it with the classic integer programming model of Hess et al. [15]. The Hess model has a binary variable z_{ij} for each pair of vertices $i, j \in V$ indicating whether vertex i is assigned to (the district centered at) vertex j. Setting $z_{jj} = 1$ represents vertex j being selected as a district center (root). Originally, Hess et al. choose to minimize the moment-of-inertia:

$$\min \sum_{i \in V} \sum_{j \in V} p_i d_{ij}^2 z_{ij},$$

where distances d_{ij} are Euclidean distances between centroids. However, we will instead use hop-based distances in graph G and will not square them, in an attempt to mimic the flow objective. The (revised) Hess model is then:

$$\begin{aligned}
\min \quad & \sum_{i\in V}\sum_{j\in V} p_i d_{ij} z_{ij} && && \text{(6a)}\\
\text{s.t.} \quad & \sum_{j\in V} z_{ij} = 1 && \forall i \in V && \text{(6b)}\\
& L z_{jj} \leq \sum_{i\in V} p_i z_{ij} \leq U z_{jj} && \forall j \in V && \text{(6c)}\\
& \sum_{j\in V} z_{jj} = k && && \text{(6d)}\\
& z_{ij} \leq z_{jj} && \forall i,j \in V && \text{(6e)}\\
& z_{ij} \in \{0,1\} && \forall i,j \in V. && \text{(6f)}
\end{aligned}$$

The assignment constraints (6b) ensure that each vertex is assigned to one district. Constraints (6c) are the population balance constraints. Constraints (6d) ensure that k district centers (and thus k districts) are selected. The coupling constraints (6e) ensure that assignments are not made to unselected district centers. Observe that the Hess model (6) has n^2 binary variables and a proportional number of constraints and nonzeros.

To guarantee contiguity, we add the popular constraints of Shirabe [26, 27] which use a flow variable f^j_{uv} for each vertex $j \in V$ and arc $(u,v) \in A$ of the bidirected graph D. Several other formulations for imposing contiguity in the Hess model have been proposed, including a cut-based model that uses a,b-separator inequalities [21], cf. [3,9,32], as well as the length-bounded cut model proposed by Validi et al. [31]. In our experience, these cut-based models perform slightly better than flow-based models on county-level instances (see Table 3 of Validi et al. [31]) and significantly better on tract-level instances (see Table 4 of Validi et al. [31]). Additional computational tricks like Lagrangian reduced-cost fixing are needed to handle larger tract-level instances [31]. However, for the purposes of this paper, the Hess-Shirabe model suffices to show the limitations of the proposed linear-size formulation.

5 Computational Experiments

For our experiments, we apply the linear-size formulation and the Hess-Shirabe formulation to real districting instances. The data arises with the 2020 U.S. Census, which was processed by Daryl DeFord and kindly shared with us. This includes the county-level graphs $G = (V, E)$, county populations p_i, and number of congressional districts k. We permit a 1% population deviation ($\pm 0.5\%$). Experiments are conducted on a PC running Windows 10 Pro with an Intel Core i7-10700T processor (2 GHz base, 4.5 GHz turbo) and 16 GB RAM. Our code is written in Python and solves MIPs using Gurobi 9.5.2 with 16 threads.

Default settings are used, with the exception of a zero `MIPGap` to ensure truly optimal solutions and a `branchPriority` of one for the s_i and z_{jj} variables as these key decisions are convenient to branch on, following Validi et al. [31]. For Williams' spanning tree formulation, we must find a planar embedding of G and construct the planar dual graph, which we obtain using [5]. Our code and data are available at https://github.com/JackDaihanZhang/Linear-size-formulations-for-connected-planar-graph-partitioning-and-political-districting.

Most county-level instances are trivial ($k = 1$) or infeasible (e.g., because of a county whose population exceeds U). Such states are excluded, and computational results for the remaining feasible states are given in Table 1.

Table 1: Computational results on county-level instances, without MIP start.

			Linear-size model			Hess-Shirabe model		
state	n	k	B&B	objective	time(s)	B&B	objective	time(s)
ME	16	2	97	1,200,838	0.07	1	1,200,838	0.03
NM	33	3	286	1,753,079	0.45	1	1,743,350	0.70
ID	44	2	1	3,237,073	0.40	1	3,237,073	0.27
WV	55	2	2,072	3,900,450	0.79	1	3,900,450	0.52
MT	56	2	805	1,750,471	2.64	1	1,750,471	0.46
AL	67	7	36,729	4,844,289	126.13	3,834	4,844,289	53.46
AR	75	4	11,315	3,933,012	83.20	1	3,930,146	1.67
MS	82	4	12,572	4,723,142	148.47	1	4,721,804	3.23
NE	93	3	945	2,525,974	7.18	1	2,525,974	2.68
IA	99	4	23,062	5,288,753	333.63	1	5,286,772	5.23
KS	105	4	55,429	4,015,151	367.65	1	4,007,553	10.34

The objective values of the two models are similar across the states, with the linear-size model sometimes having a slightly larger objective. This is because it effectively measures distances *inside* the districts, while the Hess-Shirabe model is based on distances in G, which can be less, cf. [12]. Figure 2 illustrates solutions for New Mexico, showing that the two models sometimes generate different solutions, not just different objective values.

Fig. 2: Solutions to the linear-size model (left) and Hess-Shirabe model (right).

In terms of computational performance, both models easily handle smaller instances. However, as the number of nodes and districts increases, the linear-size model performs worse than the Hess-Shirabe model, with running times often larger by one or two orders of magnitude. The linear-size model visits many more branch-and-bound nodes, by up to four orders of magnitude.

Inspecting the computational logs, we observe that our MIP solver (Gurobi) sometimes takes a while to find a good feasible solution for the linear-size model. So, we ask—maybe its performance could be significantly improved if we provide a MIP start? Table 2 shows results when an *optimal* MIP start is given. Still, the linear-size model underperforms the Hess-Shirabe model. It is no better at proving the optimality of a given solution.

Table 2: Computational results on county-level instances, when given an optimal MIP start.

			Linear-size model			Hess-Shirabe model		
state	n	k	B&B	objective	time(s)	B&B	objective	time(s)
ME	16	2	1	1,200,838	0.05	1	1,200,838	0.03
NM	33	3	199	1,753,079	0.20	1	1,743,350	0.21
ID	44	2	197	3,237,073	0.32	1	3,237,073	0.19
WV	55	2	2,140	3,900,450	0.65	1	3,900,450	0.40
MT	56	2	802	1,750,471	2.18	1	1,750,471	0.22
AL	67	7	18,808	4,844,289	64.34	921	4,844,289	13.40
AR	75	4	7,267	3,933,012	22.95	1	3,930,146	1.17
MS	82	4	2,3015	4,723,142	210.33	1	4,721,804	2.78
NE	93	3	1,145	2,525,974	6.14	1	2,525,974	1.66
IA	99	4	15,808	5,288,753	261.04	1	5,286,772	4.71
KS	105	4	9,226	4,015,151	220.79	1	4,007,553	4.39

6 Conclusion and Future Work

In this paper, our motivation was twofold. First, redistricting instances can be huge, involving potentially thousands of census tracts or nearly one million census blocks. This poses a challenge for existing integer programming models that use $\Theta(kn)$, $\Theta(n^2)$, or exponentially many variables. Second, many researchers believe that contiguity constraints make districting "much more difficult than other partitioning problems." This motivated us to develop a size $\Theta(n)$ integer programming model for political districting that exhibits desirable theoretical properties, at least in terms of contiguity.

Two key ingredients to our approach are the polyhedral characterization of cardinality-constrained out-forests due to Edmonds [6] and the linear-size extended formulation for spanning trees due to Williams [33]. They permit us to write a linear-size extended formulation for partitioning the n vertices of a planar graph into k districts; moreover, the formulation is perfect, projecting down to an integral polytope. Using additional flow variables, we impose the population balance constraints and a compactness objective. Computational experiments on 2020 US Census data show that the resulting integer programming model, despite its small size, is outperformed by existing models, e.g., the model of Hess et al. [15] with contiguity constraints of Shirabe [26, 27], cf. even faster approaches [21, 31].

So, we finish this paper in a districting whack-a-mole situation; after striking the "contiguity mole" with our MIP mallet, a new "population balance mole" has popped up. While this is an unfortunate outcome, we hope that sharing our experiences here may lead others to find better ways to impose the population balance constraints. Aside from the flow-based population balance constraints discussed in this paper, we have also (unsuccessfully) attempted to use analogues of the rounded capacity inequalities originally designed for capacitated vehicle routing. That rabbit hole has many moles.

Acknowledgments

This material is partially based upon work supported by the National Science Foundation under Grant No. 1942065 and by Rice University's Building Research on Inequality and Diversity to Grow Equity (BRIDGE) seed grant.

Conflict of interest

The authors declare that they have no conflict of interest.

Data Availability

The datasets generated during and/or analysed during the current study are available at: https://github.com/JackDaihanZhang/Linear-size-formulations-for-connected-planar-graph-partitioning-and-political-districting.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows: theory, algorithms, and applications. Prentice-Hall, Inc. (1993)
2. Aprile, M., Fiorini, S., Huynh, T., Joret, G., Wood, D.R.: Smaller extended formulations for spanning tree polytopes in minor-closed classes and beyond. The Electronic Journal of Combinatorics pp. P4–47 (2021)
3. Carvajal, R., Constantino, M., Goycoolea, M., Vielma, J.P., Weintraub, A.: Imposing connectivity constraints in forest planning models. Operations Research **61**(4), 824–836 (2013)
4. Cygan, M., Fomin, F.V., Kowalik, Ł., Lokshtanov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., Saurabh, S.: Parameterized Algorithms. Springer (2015)
5. DeFord, D., Najt, L., Stucky, E.: PlanarEmbeddings (2019). `https://github.com/vrdi/NetworksWeek5/tree/master/PlanarEmbeddings`
6. Edmonds, J.: Optimum branchings. Journal of Research of the National Bureau of Standards B **71**(4), 233–240 (1967)

7. Edmonds, J.: Matroids and the greedy algorithm. Mathematical Programming **1**(1), 127–136 (1971)
8. Fiorini, S., Huynh, T., Joret, G., Pashkovich, K.: Smaller extended formulations for the spanning tree polytope of bounded-genus graphs. Discrete & Computational Geometry **57**(3), 757–761 (2017)
9. Fischetti, M., Leitner, M., Ljubić, I., Luipersbeck, M., Monaci, M., Resch, M., Salvagnin, D., Sinnl, M.: Thinning out Steiner trees: a node-based model for uniform edge costs. Mathematical Programming Computation **9**(2), 203–229 (2017)
10. Garfinkel, R.S., Nemhauser, G.L.: Optimal political districting by implicit enumeration techniques. Management Science **16**(8), B–495 (1970)
11. Gibbons, A.: Algorithmic Graph Theory. Cambridge University Press (1985)
12. Haunert, J.H.: Aggregation in map generalization by combinatorial optimization. Ph.D. thesis, Fachrichtung Geodäsie und Geoinformatik der Leibniz-Univ. (2008)
13. Haunert, J.H., Wolff, A.: Generalization of land cover maps by mixed integer programming. In: Proceedings of the 14th annual ACM international symposium on advances in geographic information systems, pp. 75–82 (2006)
14. Hebert, J.G., Vandenberg, M.E., Smith, P.: The Realist's Guide to Redistricting: Avoiding the Legal Pitfalls. American Bar Association (2010)
15. Hess, S., Weaver, J., Siegfeldt, H., Whelan, J., Zitlau, P.: Nonpartisan political redistricting by computer. Operations Research **13**(6), 998–1006 (1965)
16. Kim, M., Xiao, N.: Contiguity-based optimization models for political redistricting problems. International Journal of Applied Geospatial Research (IJAGR) **8**(4), 1–18 (2017)
17. Kim, M.J.: Multiobjective spanning tree based optimization model to political redistricting. Spatial Information Research **26**(3), 317–325 (2018)
18. Korte, B., Vygen, J.: Combinatorial Optimization: Theory and Algorithms, *Algorithms and Combinatorics*, vol. 21, 6th edn. Springer (2018)
19. Martin, R.K.: Using separation algorithms to generate mixed integer model reformulations. Operations Research Letters **10**(3), 119–128 (1991)
20. NCSL: 2010 redistricting deviation table. `https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx` (2020). Accessed: 2022-07-21
21. Oehrlein, J., Haunert, J.H.: A cutting-plane method for contiguity-constrained spatial aggregation. Journal of Spatial Information Science **15**, 89–120 (2017)
22. Pashkovich, K.: Extended formulations for combinatorial polytopes. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg (2012)
23. Ricca, F., Scozzari, A., Simeone, B.: Political districting: from classical models to recent approaches. Annals of Operations Research **204**(1), 271–299 (2013)
24. Ricca, F., Simeone, B.: Local search algorithms for political districting. European Journal of Operational Research **189**(3), 1409–1426 (2008)
25. Schrijver, A.: Combinatorial Optimization: Polyhedra and Efficiency, *Algorithms and Combinatorics*, vol. 24. Springer (2003)
26. Shirabe, T.: A model of contiguity for spatial unit allocation. Geographical Analysis **37**(1), 2–16 (2005)
27. Shirabe, T.: Districting modeling with exact contiguity constraints. Environment and Planning B: Planning and Design **36**(6), 1053–1066 (2009)
28. Swamy, R., King, D.M., Jacobson, S.H.: Multiobjective optimization for politically fair districting: A scalable multilevel approach. Operations Research **71**(2), 536–562 (2023)
29. Validi, H., Buchanan, A.: A note on "a linear-size zero-one programming model for the minimum spanning tree problem in planar graphs". Networks **73**(1), 135–142 (2019)
30. Validi, H., Buchanan, A.: Political districting to minimize cut edges. Mathematical Programming Computation **14**(4), 623–672 (2022)
31. Validi, H., Buchanan, A., Lykhovyd, E.: Imposing contiguity constraints in political districting models. Operations Research **70**(2), 867–892 (2022)
32. Wang, Y., Buchanan, A., Butenko, S.: On imposing connectivity constraints in integer programs. Mathematical Programming **166**(1), 241–271 (2017)
33. Williams, J.C.: A linear-size zero-one programming model for the minimum spanning tree problem in planar graphs. Networks **39**(1), 53–60 (2002)
34. Wong, R.T.: A dual ascent approach for Steiner tree problems on a directed graph. Mathematical Programming **28**(3), 271–287 (1984)