

1 Census TopDown: The Impacts of Differential 2 Privacy on Redistricting

3 Aloni Cohen ✉

4 Hariri Institute for Computing and School of Law, Boston University, USA

5 Moon Duchin ✉

6 Department of Mathematics, Tufts University, USA

7 JN Matthews ✉

8 Tisch College of Civic Life, Tufts University, USA

9 Bhushan Suwal ✉

10 Tisch College of Civic Life, Tufts University, USA

11 — Abstract —

12 The 2020 Decennial Census will be released with a new disclosure avoidance system in place,
13 putting *differential privacy* in the spotlight for a wide range of data users. We consider several key
14 applications of Census data in redistricting, developing tools and demonstrations for practitioners
15 who are concerned about the impacts of this new noising algorithm called TopDown. Based on a
16 close look at reconstructed Texas data, we find reassuring evidence that TopDown will not threaten
17 the ability to produce districts with tolerable population balance or to detect signals of racial
18 polarization for Voting Rights Act enforcement.

19 **2012 ACM Subject Classification** Security and privacy; Applied computing → Law; Applied
20 computing → Voting / election technologies

21 **Keywords and phrases** Census, TopDown, differential privacy, redistricting, Voting Rights Act

22 **Digital Object Identifier** [10.4230/LIPIcs.FORC.2021.5](https://doi.org/10.4230/LIPIcs.FORC.2021.5)

23 **Supplementary Material** <https://megg.org/DP>

24 **Funding** This project was supported on NSF OIA-1937095 (Convergence Accelerator) and by a
25 grant from the Alfred P. Sloan Foundation.

26 *Aloni Cohen*: NSF CNS-1414119 and CNS-1915763; DARPA HR00112020021

27 *Moon Duchin*: NSF DMS-2005512

28 **Acknowledgements** Authors are listed alphabetically. We thank Denis Kazakov, Mark Hansen, and
29 Peter Wayner. Kazakov developed the reconstruction algorithm as a member of Hansen’s research
30 group. Wayner guided our deployment of TopDown in AWS and was an invaluable team member for
31 the technical report. Any opinions, findings, and conclusions or recommendations expressed in this
32 material are those of the authors and do not necessarily reflect the views of our funders.

33 **1 Introduction**

34 A new disclosure avoidance system is coming to the Census: the 2020 Decennial Census
35 releases will use an algorithm called TopDown to protect the data from increasingly feasible
36 *reconstruction attacks* [2]. Census data is structured in a nesting sequence of geographic
37 units covering the whole country, from nation at the top to small *census blocks* at the
38 bottom. TopDown starts by setting a *privacy budget* $\varepsilon > 0$ which is allocated to the levels of
39 a designated hierarchy, then adding noise at each level in a *differentially private* way [12].
40 When $\varepsilon \rightarrow \infty$, the data alterations vanish, while $\varepsilon \rightarrow 0$ yields pure noise with no fidelity to
41 the input data. The algorithm continues with a post-processing step that leaves an output
42 dataset that is designed to be suitable for public use.



© Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal;
licensed under Creative Commons License CC-BY 4.0

2nd Symposium on Foundations of Responsible Computing (FORC 2021).

Editors: Katrina Ligett and Swati Gupta; Article No. 5; pp. 5:1–5:22



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

43 *Redistricting* is the process of dividing a polity into territorially delimited pieces in which
44 elections will be conducted. The Census has a special release—named the PL 94-171 after
45 the law that requires it—that reports the number of residents in every geographic unit in
46 the country by race, ethnicity, and the number of voting-age residents [9]. The 2020 release
47 is slated to occur by September 2021, after which many thousands of district lines will
48 be redrawn: not only U.S. Congressional districts, but those for state legislatures, county
49 commissions, city councils, and many more.

50 Many user groups have expressed concerns about the effects of differential privacy on
51 redistricting. They largely but not exclusively concern two issues. First, “One Person, One
52 Vote” case law calls for balancing population across the electoral districts in a jurisdiction,
53 whether small like city council districts or large like congressional districts. Most states
54 balance congressional districts to within one person based on Census counts. Second, the
55 most reliable legal tool against gerrymandering has been the Voting Rights Act of 1965
56 (VRA), which requires a demonstration of racially polarized voting (RPV). This RPV analysis
57 is typically performed by statistical techniques that infer voting by race from precinct-level
58 returns. Many voting rights advocates worry that noising of Census data will confuse
59 population balancing practices, and others worry that it will attenuate RPV signals, making
60 it harder to press valid claims.

61 The Census Bureau has been commendably transparent about the development of
62 TopDown, making working code publicly available along with documentation and research
63 papers describing the algorithm. The complexity of the algorithm makes it extremely difficult
64 to study analytically, so many people have sought to run it on realistic data. However, since
65 person-level Census data remain confidential for 72 years after collection, detailed input data
66 for TopDown is not public. Data users who would like to understand its impacts are left with
67 two options: decades-old data or a limited demonstration data product.

68 In this paper, we get around the empirical obstacle by use of reconstructed block-level 2010
69 microdata for the state of Texas, and we try to understand the algorithm through theoretical
70 analysis of a much-simplified toy algorithm, ToyDown, that retains the two-stage, top-down
71 structure of TopDown but is much easier to analyze symbolically. We investigate three
72 questions about the count discrepancies created by TopDown in units of census geography
73 and “off-spine” aggregations like districts and precincts.

74 **Hierarchical budget allocation.** We derive easy-to-evaluate expressions for ToyDown errors
75 as a function of the privacy budget allocation. Error at higher levels of the geographic
76 hierarchy impacts lower-level counts with a significant discount, suggesting that bottom-
77 heavy allocations may be optimal for accuracy on small geographies. This is consistent with
78 the small-district errors in our experiments with TopDown. For larger districts, a tract-heavy
79 allocation gives greatest accuracy. Equal allocation over the levels is a strong performer in
80 both cases, making this a good choice from the point of view of multi-scale redistricting.

81 **District construction.** From there, we create further tests to study the impacts of district
82 design. We compare hierarchically greedy to geometrically greedy district-generation schemes,
83 where the former attempt to keep large units whole and the latter attempt to build districts
84 with short boundaries. We find that the ToyDown model gives errors very closely keyed to
85 the fragmentation of the hierarchy, but that spatial factors damp out the primary role of
86 fragmentation in the shift to the TopDown setting.

87 **Robustness of linear regression.** Finally, we consider the unweighted linear regressions
88 commonly used to assess racial polarization in voting rights cases. We find that the noise
89 from both ToyDown and TopDown introduces an attenuation bias that seems alarming at
90 first. However, unweighted linear regression on precincts is already vulnerable to major skews

91 imposed by the inclusion of very small precincts. For any reasonable way of counteracting
 92 that—trimming out the tiny precincts or weighting the regression by the number of votes
 93 cast—the instability introduced by ToyDown and TopDown all but vanishes.

94 Our investigation is set up to answer questions about the status quo workflow in
 95 redistricting. As usual with studies of differential privacy, a finding that DP unsettles the
 96 current practices might lead us to call to refine the way it is applied, but might equally lead
 97 us to interrogate the traditional practices and seek next-generation methods for redistricting.
 98 In particular, it is clear that the practice of *one-person* population deviation across districts
 99 was never reasonably justified by the accuracy of Census data nor required by law, and the
 100 adoption of differential privacy might give redistricters occasion to reconsider that practice.
 101 We make a similar observation about the way that racially polarized voting analysis is
 102 commonly performed in expert reports. On the other hand, by focusing on decisions still to
 103 be announced like the privacy budget and its allocation over the hierarchy, we are able to
 104 make recommendations that can assist the Bureau in protecting privacy while attending to
 105 the important concerns of user groups.

106 **2 Background on Census and redistricting**

107 **2.1 The structure of Census data and the redistricting data products**

108 Every ten years the U.S. Census Bureau attempts a comprehensive collection of person-level
 109 data—called *microdata*—from every household in the country. The microdata are confidential,
 110 and are only published in aggregated tables subject to disclosure avoidance controls. The
 111 Decennial Census records information on the sex, age, race, and ethnicity for each member of
 112 each household, using categories set by the Office of Management and Budget [8]. The 2020
 113 Census used six primary racial categories: White, Black, American Indian, Asian, Native
 114 Hawaiian/Pacific Islander, and Some Other Race. An individual can select these in any
 115 combination but must choose at least one, creating $2^6 - 1 = 63$ possible choices of race.
 116 Separately, *ethnicity* is represented as a binary choice of Hispanic/Latino or not.

117 The 2010 Census divided the nation into over 11 million small units called *census blocks*
 118 which nest in larger geographies in a six-level “central spine”: nation—state—county—
 119 tract—block group—block. Counts of different types are provided with respect to these
 120 geographies. This tabular data is then used in an enormous range of official capacities, from
 121 the apportionment of seats in the U.S. House of Representatives to the allocation of many
 122 streams of federal and state funding. The redistricting (PL 94-171) data includes four such
 123 tables: H1, a table of housing units whose types are occupied/vacant; and four tables of
 124 population, P1 (63 races), P2 (Hispanic, and 63 races of non-Hispanic population), and
 125 P3/P4 (same as P1/P2 but for voting age population). Each table can be thought of as a
 126 *histogram*, with each included type constituting one histogram *bin*. For instance, in table P1
 127 there is 1 person in the $t = \text{White} + \text{Asian}$ bin in the Middlesex County, MA, block numbered
 128 31021002.

129 Treating the 2010 tables as accurate, it is easy to infer information not explicitly presented
 130 in the tables. For instance, the same bin in the P3 table (race for voting age population) also
 131 has a count of 1, implying that there are no White+Asian people under 18 years old in block
 132 31021002. This is the beginning of a *reconstruction* process that would enable an attacker, in
 133 principle, to learn much of the person-level microdata behind the aggregate releases.

134 **2.2 Disclosure avoidance**

135 Title 13 of the U.S. Code requires the Bureau to take measures to protect the privacy of
 136 respondents’ data [1]. In the 2010 Census, this was largely achieved by an ad hoc mechanism
 137 called *data swapping*: a Bureau employee manually swapped data between small census
 138 blocks to thwart re-identification. In 2020, swapping is no longer considered adequate to
 139 protect against more sophisticated (but mathematically straightforward) data attacks that
 140 seek to reconstruct the individual microdata. An internal Census Bureau study concluded
 141 that data swapping was unacceptably vulnerable: Census staff were able to reconstruct the
 142 2010 Census responses of—and correctly reidentify—tens of millions of people.

143 With the reconstruction/reidentification threat in mind, the Bureau has developed an
 144 algorithm called TopDown [2], which begins with a noising step that is *differentially private*,
 145 following a mathematical formalism that provides rigorous guarantees against information
 146 disclosure [12]. Differentially private algorithms obey a quantifiable limit to how much the
 147 output can depend on an individual record in the input. The relationship of output to input
 148 is specified by a tuneable parameter, ϵ , often called the *privacy budget*. When $\epsilon \rightarrow \infty$, the
 149 output approaches equality to the input (high risk of disclosure). When $\epsilon \rightarrow 0$, the output
 150 bears no resemblance to the input whatsoever (no risk of disclosure). Like a fiscal budget,
 151 the privacy budget can be allocated until it is fully spent, in this case by spending parts of
 152 the budget on particular queries and on levels of the hierarchy.

153 TopDown takes an individual-level table of census data and creates a ‘synthetic’ dataset
 154 that will be used in its place to generate the PL 94-171 tables. It can be thought of as
 155 taking as input a histogram with a bin for each person type (i.e., a combination of race, sex,
 156 ethnicity, etc.) and outputting an altered version of the same histogram. It proceeds in two
 157 stages. First, it privatizes the input histogram counts: it adds enough random noise to get
 158 the required level of differential privacy (according to the budget ϵ). At this stage, it also
 159 allocates a portion of the total privacy budget for generating additional noisy histograms of
 160 data of particular importance to the Census Bureau. Second, TopDown does post-processing
 161 on the noisy histograms to satisfy a handful of additional plausibility constraints. Among
 162 other things, post-processing ensures that the resulting histograms contain only non-negative
 163 integers, are self-consistent, and agree with the raw input data on a handful of *invariants*
 164 (e.g., total state population).

165 The overall privacy guarantees of TopDown are poorly understood. In this paper, we
 166 design a simpler cousin of TopDown nicknamed ToyDown and we explore the properties of
 167 both ToyDown and TopDown, primarily focusing on reconstructed Texas data from 2010.

168 **2.3 The use of Census products for redistricting**

169 The PL 94-171 tables are the authoritative source of data for the purposes of apportionment
 170 to the U.S. House of Representatives, and with a very small number of exceptions also for
 171 within-state legislative apportionment. The most famous use of population counts is to
 172 decide how many members of the 435-seat House of Representatives are assigned to each
 173 state. In “One person, one vote” jurisprudence initiated in the *Reynolds v. Sims* case of
 174 1964, balancing Census population is required not only for Congressional districts within
 175 a state but also for districts that elect to a state legislature, a county commission, a city
 176 council or school board, and so on [17, 18, 3].

177 Today, the Congressional districts within a state usually balance total population extremely
 178 tightly: each of Alabama’s seven Congressional districts drawn after the 2010 Census has
 179 a total population of either 682,819 or 682,820 according to official definitions of districts

180 and the Table P1 count, while Massachusetts districts all have a population of 727,514 or
 181 727,515. Astonishingly, though no official rule demands it, more than half of the states
 182 maintain this “zero-balancing” practice (no more than one person deviation) for Congressional
 183 districts [16]. This ingrained habit of zero-balancing districts to protect from the possibility
 184 of a malapportionment challenge is the first source of worry in the redistricting sphere. If
 185 disclosure avoidance practices introduce some systematic bias—say by creating significant
 186 net redistribution towards rural and away from urban areas—then it becomes hard to control
 187 overall malapportionment, which could in principle trigger constitutional scrutiny. In the
 188 end, redistricters may not care very much how many people live in a single census block, but
 189 it could be quite important to have good accuracy at the level of a district.

190 The second major locus of concern for redistricting practitioners is the enforcement of the
 191 Voting Rights Act (VRA). Here, histogram data is used to estimate the share of voting age
 192 population held by members of minority racial and ethnic groups. Voting rights attorneys
 193 must start by satisfying three threshold tests without which no suit can go forward.

- 194 ■ **Gingles 1:** the first “Gingles factor” in VRA liability is satisfied by creating a demonstration
 195 district where the minority group makes up over 50% of the voting age population.
- 196 ■ **Gingles 2-3:** the voting patterns in the disputed area must display *racial polarization*.
 197 The minority population is shown to be cohesive in its candidates of choice, and bloc
 198 voting by the majority prevents these candidates from being elected. In practice, inference
 199 techniques like linear regression or so-called “ecological inference” are used to estimate
 200 voting preferences by race.

201 Since the VRA has been a powerful tool against gerrymandering for over 50 years, many
 202 worry that even where the raw data would clear the Gingles preconditions, the noised data
 203 will tend towards uniformity—blocking deserving plaintiffs from a cause of action.

204 **3** Census TopDown and ToyDown

205 **3.1** Setup and notation

206 For the Census application, the data universe is a set of *types*: for instance, the redistricting
 207 data (the PL 94-171) has the types $T = T_R \times T_E \times T_{VA} \times T_H$, where T_R is the set of 63
 208 races, T_E is binary for ethnicity (Hispanic or not), T_A is binary for age (voting age or not),
 209 and T_H is the set of housing types. (The fuller decennial Census data has more types.)

210 A *hierarchy* H is a rooted tree of some depth d , so that every leaf has distance $\leq d - 1$
 211 from the root. We will usually assume the hierarchy has uniform depth, so that every leaf is
 212 exactly $d - 1$ away from the root. For node $h \in H$, let $n(h) \in \mathbb{N}$ be the number of children
 213 of h in the tree, and let $\ell(h)$ be the level of node h . A hierarchy is called *homogeneous*
 214 if each node at level ℓ has the same number of children, denoted n_ℓ . Let H_ℓ denote the
 215 set of nodes at level ℓ , so that the set of leaves is H_d in the uniform-depth case. Label
 216 the root of the tree $h = 1$. We adopt an indexing of the tree and refer to the i th child of
 217 h as h_i ; the parent of any non-root node h is denoted \hat{h} . In Census data, the hierarchy
 218 represents the large and complicated set of nested geographical units, from the nation at
 219 the root down to the census blocks at the leaves. The standard hierarchy has the six levels
 220 (nation—state—county—tract—block group—block) described above.

221 We associate with hierarchy H and types T a set of *counts* $A_{H,T} = \{a_{h,t} \in \mathbb{N}\}_{h \in H, t \in T}$,
 222 where $a_{h,t}$ is the population of type t in unit h of census geography. We say $A_{H,T}$ is
 223 *hierarchically consistent* if the counts add up correctly: for every non-leaf h and every t , we
 224 require $a_{h,t} = \sum_{i \in [n(h)]} a_{h_i,t}$. For a singleton T , we write $A_H = \{a_h\}$. We set an *allocation*
 225 $(\varepsilon_1, \dots, \varepsilon_d)$ breaking down the privacy budget $\varepsilon = \sum \varepsilon_i$ to the different levels of the hierarchy.

Our *queries* will always be counting queries, so that for instance $q_{F,44}(h)$ returns the number of 44-year-old females in geographic unit h . This particular query is part of a “sex by age” *histogram* $Q_{sex,age} = \{q_{s,a} : s \in T_S, a \in T_A\}$, which partitions T into *bins* by sex and age. In this language, $q_{F,44}$ is a bin of the sex-by-age histogram. By slight abuse of notation, we will use the same terminology for the queries and their outputs, so that the histogram can be thought of as the collection of queries or the collection of counts. Similarly, the “voting age by ethnicity by race” histogram consists of a query for each combination of the $2 \times 2 \times 63$ possible combinations of the three attributes.

3.2 ToyDown and TopDown

The Bureau’s TopDown and our simplified ToyDown are both algorithms for releasing privatized population counts for every $h \in H$. That is, these algorithms protect privacy by noising the data histograms. TopDown releases not just total population counts, but counts by type. We will define *single-attribute* and *multi-attribute* versions of ToyDown that noise A_H and $A_{H,T}$, respectively, where consistency must hold for each type t .

TopDown and ToyDown share the same two-stage structure. Starting with hierarchically consistent raw counts a , the *noising stage* generates differentially private counts \hat{a} . The *post-processing stage* solves a constrained optimization problem to find noisy counts α that are close to the \hat{a} values while satisfying hierarchical consistency and other requirements. TopDown is named after the iterative approach to post-processing: one geographic level at a time, starting at the top (nation) and working down to the leaves (blocks). We sketch the noising and post-processing here, and we describe them in Appendix A in more detail.

The simple ToyDown model can be run in a single-attribute version (only counts A_H), a multi-attribute version (counts by type $A_{H,T}$), or in multi-attribute form enforcing non-negativity. The single-attribute version is easy to describe: level by level, random noise values are selected from a Laplace distribution with scale $1/\epsilon_\ell$ and added to each count, replacing each a_h with $\hat{a}_h = a_h + L_h$. Then, working from top to bottom, the noisy \hat{a}_h are replaced with the closest possible real numbers α_h satisfying hierarchical consistency. Multi-attribute ToyDown is defined analogously, but using $A_{H,T}$ instead of A_H and requiring hierarchical consistency within each type $t \in T$. Non-negative ToyDown adds the inequality requirement that $\alpha_h \geq 0$.

TopDown is structurally similar but much more complex, with more kinds of privatized counts in the noising stage and a great many more constraints in the post-processing stage, including integrality. The privatized counts computed by TopDown are specified by a collection of histograms (or complex queries) called a *workload* W . For each bin of each histogram in the workload and for each node h in the geographic hierarchy, TopDown adds geometric noise to the count. The post-processing step finds the closest integer point that satisfies the requirements given by hierarchical consistency, non-negativity, as well as additional conditions given as invariants and structural inequalities. For example, any block with zero households in the raw counts must have zero households and zero population in the output adjusted counts. Together, the invariants, structural inequalities, integrality, and non-negativity make this optimization problem very hard. The problem is NP-hard in the worst case and TopDown cannot always find a feasible solution. There is a sophisticated secondary algorithm for finding approximate solutions that is beyond the scope of this paper.

ToyDown is simple enough that solutions can often be obtained symbolically. ToyDown simplifies the noising stage by fixing the workload to be the detailed workload partition $Q_{detailed} = \{\{t\}\}_{t \in T}$ consisting of all singleton sets and using the continuous Laplace Mechanism instead of the discrete Geometric Mechanism. It simplifies the post-processing

273 stage by dropping invariants, structural inequalities, integrality, and non-negativity. When
 274 negative answers are permitted, multi-attribute `ToyDown` is equivalent to executing $|T|$
 275 independent instances of single-attribute `ToyDown` on inputs $A_{H,t} = \{a_{h,t}\}_{h \in H}$ for each
 276 $t \in T$. As a result, many of our analytical results for single-attribute `ToyDown` extend
 277 straightforwardly to multi-attribute `ToyDown` (allowing negative answers) by scaling by a
 278 factor of $|T|$ in appropriate places.

279 4 Methods

280 We use both analytical and empirical techniques in this work. This section describes our
 281 high-level empirical approach: what algorithms and raw data we used and how we used
 282 them. See Appendix B for more details. We repeatedly ran `TopDown` and `ToyDown` in
 283 various configurations on a reconstructed person-level Texas dataset created by applying a
 284 reconstruction technique to the block-level data from the 2010 Census, following [15] based on
 285 [11]. The reconstructed microdata records—obtained from collaborators—contain block-level
 286 sex, age, ethnicity, and race information consistent with a collection of tables from 2010
 287 Census Summary File 1.

288 We executed 16 runs of `TopDown` with each of 20 different allocations of the privacy budget
 289 across the five lower levels of the national census geographic hierarchy: $\varepsilon = \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6$.
 290 The 20 allocations consist of five different splits across the levels (Table 1) for each of four
 291 total budgets $\varepsilon \in \{0.25, 0.5, 1.0, 2.0\}$. `TopDown` operates on the six-level Census hierarchy
 292 and requires specifying ε_1 . In our experiments, we ran `TopDown` with a fixed total privacy
 293 budget $\varepsilon_{total} = 10$, with $\varepsilon_1 = 10 - \varepsilon$. Because the nation-level budget is so much higher
 294 than the lower level budgets, we omit further discussion of it. The `TopDown` workload was
 295 modeled after the workload used in the 2018 End-to-End test release, omitting household
 296 invariants and queries.

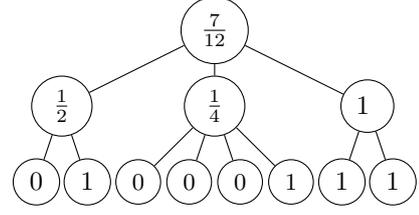
297 We also ran three variants of `ToyDown` (single-attribute, multi-attribute, and non-negative)
 298 on a simplified version of the same data 2010 data. We executed 16 runs of each variant
 299 with each of five different splits of the privacy budget across the five lower levels of the
 300 census geographic hierarchy (Table 1), fixing the total budget for those five levels at $\varepsilon = 1$.
 301 The data was derived from the reconstructed Texas data simplified to include only seven
 302 distinct types: one for the total Hispanic population and one for each of six subgroups of
 303 the non-Hispanic population based on race (White; Black; American Indian; Asian; Native
 304 Hawaiian/Pacific Islander; and Some Other Race or multiple races). Post-processing for single-
 305 attribute `ToyDown` was implemented in NumPy, while post-processing for multi-attribute
 306 and non-negative `ToyDown` used a Gurobi solver.

307 5 Hierarchical budget allocation

308 The relationship of the hierarchical allocation $(\varepsilon_1, \dots, \varepsilon_d)$ to various measures of output
 309 accuracy is not obvious. On one hand, it might seem that higher values of ε_d (the block-level
 310 budget) will best promote accuracy at the block level, for a fixed ε . But on the other
 311 hand, imposing hierarchical consistency forces lower levels to be consistent with the totals at
 312 higher levels, which means that noise at higher levels can trickle down to lower levels. These
 313 competing effects create tradeoffs that are hard to balance without further analysis.

Split name	state ε_2	county ε_3	tract ε_4	BG ε_5	block ε_6
equal	0.2	0.2	0.2	0.2	0.2
state-heavy	0.5	0.25	0.083	0.083	0.083
tract-heavy	0.083	0.167	0.5	0.167	0.083
BG-heavy	0.083	0.083	0.167	0.5	0.167
block-heavy	0.083	0.083	0.083	0.25	0.5

■ **Table 1** Names of designated budget splits used in *ToyDown* and *TopDown* runs below, each with a budget of $\varepsilon_1 = 9$ on the nation and a total of 1 allocated below the national level.



■ **Figure 1** A district in a three-level hierarchy. The 0/1 weight of a leaf indicates its membership in the district; each non-leaf weight is the average of the node’s children.

314 5.1 ToyDown error expressions

315 ► **Definition 1** (District, weights, error). A district $D \subseteq H_d$ is a subset of the leaves (blocks)
316 of the hierarchy H . For hierarchy H , a district D induces weights $w_h \in [0, 1]$ on the hierarchy
317 nodes, defined recursively as follows:

318 ■ For each leaf $h \in H_d$, let $w_h = 1$ if $h \in D$ and $w_h = 0$ otherwise.

319 ■ For $\ell \leq d - 1$ and $h \in H_\ell$, let $w_h = \frac{1}{n(h)} \cdot \sum_{i \in [n(h)]} w_{h_i}$ be the average of the weights of
320 the children.

321 In a homogeneous hierarchy, we can observe that each w_h equals the fraction of the leaves
322 descended from h that belong to D . In particular, the root weight is $w_1 = |D|/|H_d| = 1/k$ if
323 there are k districts of equal population made from nodes of equal population.

324 For node $h \in H$, we record the error $E_h = \alpha_h - a_h$ introduced by *ToyDown* to the count
325 a_h . The total error over district D is $E_D = \sum_{h \in D} E_h$. Let \hat{h} denote the parent of node h .

326 ► **Theorem 2** (Error expressions). $E_1 = L_1$. For $\ell \in \{2, \dots, d\}$ and non-root node $h_i \in H_\ell$,
327 and for every district D with associated weights w_h on the nodes,

$$328 \quad E_{h_i} = L_{h_i} + \frac{1}{n(h)} \left(E_h - \sum_{j \in [n(h)]} L_{h_j} \right), \quad E_D = w_1 L_1 + \sum_{h \in H \setminus \{1\}} (w_h - w_{\hat{h}}) L_h. \quad (1)$$

329 We make several observations. First, our intuition that error at higher levels trickles down
330 to lower levels is correct, but this effect is rather weak. The error at a child h_i is determined
331 by the parent error E_h discounted by the degree $n(h)$, the number of siblings. This suggests
332 that placing more budget at level ℓ is an efficient way to secure accuracy at that level, until
333 a fairly extreme level of error at higher levels overwhelms the degree-based “discount.”

334 Second, because the L_h are all independent random variables with $\mathbb{E}(L_h) = 0$ and
335 $\text{Var}(L_h) = 8/\varepsilon_{\ell(h)}^2$, the theorem provides the following expression for variance that we use
336 repeatedly.

337 ► **Corollary 3** (Error expectation and variance). For all $D \subseteq H_d$ and associated weights w_h ,
338 the expected error and error variance produced by *ToyDown* satisfy $\mathbb{E}(E_D) = 0$ and

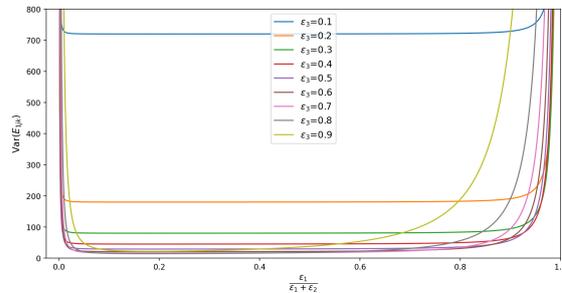
$$339 \quad \text{Var}(E_D) = \frac{8w_1^2}{\varepsilon_1^2} + \sum_{\ell=2}^d \left(\frac{8}{\varepsilon_\ell^2} \cdot \sum_{h \in H_\ell} (w_h - w_{\hat{h}})^2 \right). \quad (2)$$

340 Third, we get a more explicit expression if restricting to homogeneous hierarchies H .
341 Consider the case of a singleton district $\{h\}$ made of a single census block $h \in H_d$.

342 ► **Corollary 4** (Error variance, homogeneous case). *The ToyDown error for a single block*
 343 *$h \in H_d$ satisfies*

$$344 \quad \text{Var}(E_h) = \frac{8}{\varepsilon_1^2(n_1 \cdots n_{d-1})^2} + \sum_{\ell=2}^d \frac{8n_{\ell-1}(n_{\ell-1} - 1)}{\varepsilon_\ell^2(n_{\ell-1} \cdots n_{d-1})^2}. \quad (3)$$

345 Figure 2 plots this expression for various ways of splitting a total privacy budget of
 346 $\varepsilon = 1$ across a three-level hierarchy with $n_1 = n_2 = 10$. The minimum of $f(x_1, \dots, x_d) =$
 347 $\sum_{\ell=1}^d a_\ell/x_\ell^2$ subject to $\sum_\ell x_\ell = \varepsilon$ and $x_\ell \geq 0$ is achieved at $x_\ell = \varepsilon a_\ell^{1/3}/\sum_i a_i^{1/3}$ for all ℓ . For
 348 the example in Figure 2, the minimum-variance split is $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.038, 0.171, 0.791)$ with
 349 variance 14.52. (See accompanying [CoLab notebook](#).) One important note in interpreting
 350 Figure 2 is that these variance numbers are absolute and don't depend on knowing population
 351 counts for the nodes of the hierarchy. They are simply based on sampling Laplace noise with
 352 the given parameters. If a variance of about 15 in the bottom-level counts is too high to be
 353 tolerated in an application, one would have to increase ε to achieve lower variance.



ε	Allocation	L^1 error
1.0	(.16, .16, .16, .16, .16, .2)	0.03
1.0	(.2, .16, .16, .16, .16, .16)	0.03
1.0	(.1, .1, .1, .1, .1, .5)	0.02
1.0	(.02, .02, .02, .02, .02, .9)	0.03
1.0	(.66, .30, .01, .01, .01, .01)	0.09

354 ■ **Figure 2** ToyDown error variance for a leaf node in the three-level hierarchy with $n_1 = n_2 = 10$ and $\varepsilon = 1$. The curves show varying ε_3 (colors) and the relative balance of ε_1 and ε_2 (x-axis).

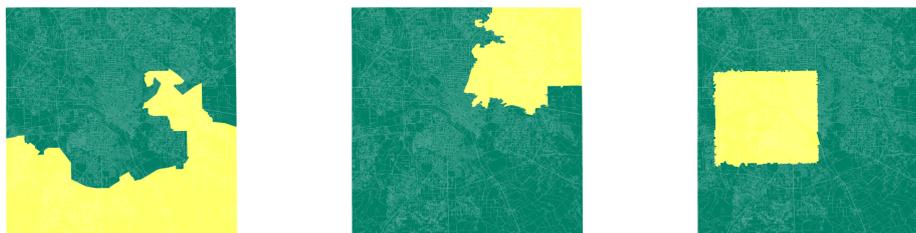
355 ■ **Table 2** L^1 error measurements from selected TopDown runs on reconstructed Texas data. The allocation $(\varepsilon_1, \dots, \varepsilon_6)$ goes from the nation $\ell = 1$ down to census blocks at $\ell = 6$.

354 5.2 Empirical error experiments in TopDown

355 Next, we move to TopDown, which requires the use of input data. First, using reconstructed
 356 2010 Texas data, we varied the relative allocation vector and the total ε , then measured
 357 the effects with an L^1 error metric included in the Census code [5]. This is a measure of
 358 block-level error: it adds the magnitudes of changes in the bins, then divides by twice the
 359 total population in the histogram.

360 Table 2 reports a small selection of the 100+ different scenarios explored. In general, the
 361 lowest error outcomes were observed in a few scenarios: when the budget was distributed
 362 near-equally to the levels of the hierarchy, and when half of the available budget was placed
 363 at the bottom level—beyond $\varepsilon_d = \varepsilon/2$, further bottom-weighting gave diminishing returns in
 364 block-level accuracy.

365 But a budget allocation that produces small block-level errors may not produce small
 366 errors for *districts*, depending on the degree of cancellation or correlation. Next, we use
 367 random district generation to understand the effects of off-spine aggregation. In particular,
 368 we employ the Markov chain sampling algorithm called *recombination* (or ReCom), which runs
 369 an elementary move that fuses two neighboring districts and re-partitions the double-district
 370 by a random balanced cut to a random spanning tree [10].



■ **Figure 3** Three sample districts (yellow) in Dallas County, each within two percent of the ideal population for $k = 4$ districts. These are drawn by tract ReCom, block ReCom, and a square-favoring algorithm, respectively.

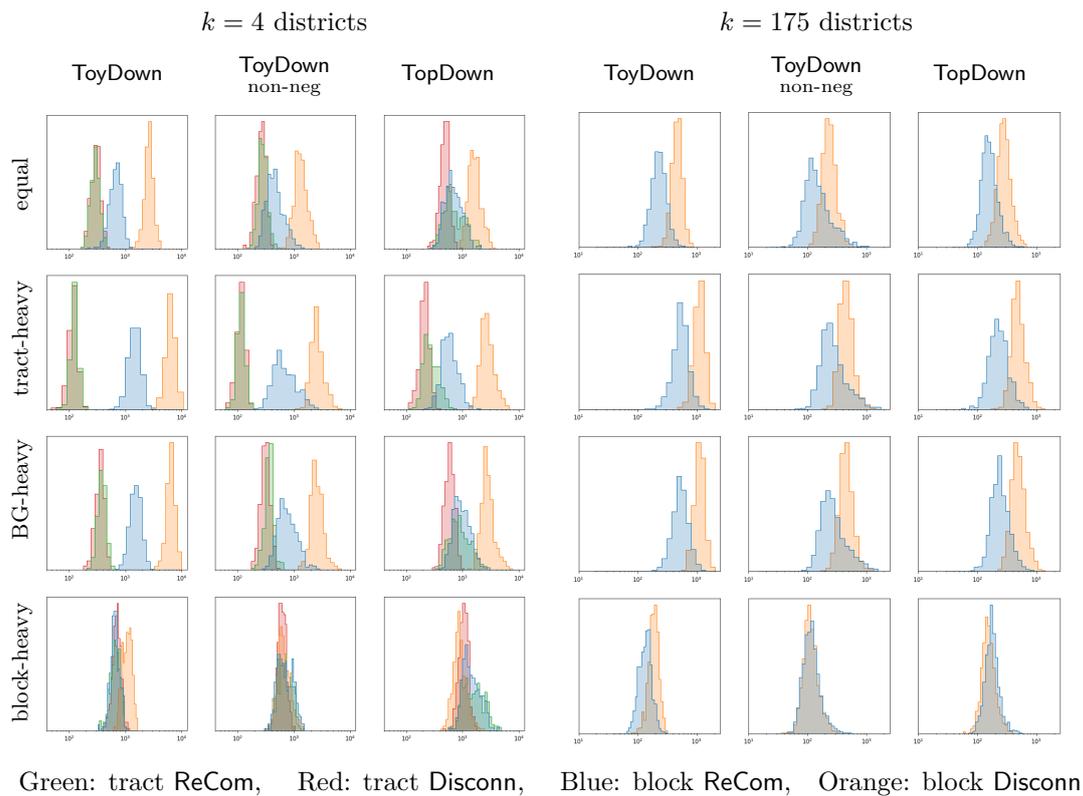
371 We begin with county commission districts in Dallas County, where $k = 4$. Since the 2010
 372 population of Dallas County was roughly 2.4 million, each district will have roughly 600,000
 373 people, making them nearly as big as congressional districts and much larger than tracts.
 374 We also include divisions of the county into $k = 175$ districts of between 13,000 and 14,000
 375 people each for a small-district comparison. Figure 4 plots the data from our experiments on
 376 a logarithmic scale. Each histogram displays 400 values, one for each district drawn by the
 377 specified district-drawing algorithm; each value is the mean observed district-level population
 378 error magnitude over 16 executions of the specified hierarchical noising algorithm using the
 379 specified budget allocation.

380 First, consider two unrealistic forms of district-generation: tract Disconn (red) and block
 381 Disconn (orange), which randomly choose units of the specified type until assembling a
 382 collection with the appropriate population. These are unrealistic because they do not form
 383 connected districts; here, they are used to illustrate the effects of aggregation, neglecting
 384 spatial factors entirely. We see in Figure 4 that block-based methods generate hugely more
 385 error than tract-based methods, except if the budget allocation is concentrated at the bottom
 386 of the hierarchy. The effect is stronger for ToyDown (in keeping with Theorem 2), but is
 387 easily observed for TopDown as well.

388 We compare that with the more realistic district-generation algorithm block ReCom
 389 (blue), which builds compact and connected districts out of block units. This tends to give
 390 error levels in between the extremes set by the other two. Likewise, tract ReCom (green)
 391 builds compact and connected districts from tracts. One reasonable mechanism by which
 392 ReCom has much lower error than Disconn is that ReCom districts will tend to have higher
 393 “hierarchical integrity,” keeping higher-level units whole just by virtue of being connected
 394 and plump. The interior of ReCom districts will thus contain many whole block groups
 395 and tracts. Near the boundary, block groups and tracts are more fragmented, leaving the
 396 corresponding block-level errors uncanceled. These fragmentation ideas are explored more
 397 fully in Section 6 and some sample districts are depicted here.

398 The cancellation effect is significant: in most experiments, the error level for ReCom
 399 districts is much closer to that of tract Disconn than block Disconn (recall the data is plotted
 400 on a logarithmic scale). Overall, drawing districts out of larger pieces (e.g., using tract
 401 Disconn instead of ReCom, or ReCom instead of block Disconn) lowers error magnitude
 402 significantly in the best case and has little or no effect in the worst case.

403 Although tract ReCom and tract Disconn behave very similarly under ToyDown, the
 404 compact districts perform noticeably worse than their disconnected relatives once we pass
 405 to the full complexity of TopDown. At first this seems puzzling, because compact and



■ **Figure 4** These histograms show district-level error on a log scale for various combinations of budget splits (rows), district-drawing algorithms (colors), and noising algorithms (columns). We include both large districts and small districts, dividing the county into $k = 4$ and $k = 175$ equal parts. Each histogram displays 400 values, one for each district drawn by the specified algorithm, plotting the mean observed district-level population error magnitude over 16 executions of the noising algorithm using the specified budget allocation.

406 connected districts are being punished by the geography-aware TopDown. But the reason for
 407 this is apparent on further reflection: *spatial autocorrelation* is causing the post-processing
 408 corrections to move nearby tracts in the same direction, impeding the cancellation that
 409 makes counts usually more accurate on larger geographies.

410 In the end, the story that emerges from these investigations is that, with full TopDown,
 411 the best accuracy that can be observed for large districts occurs when they are made from
 412 whole tracts and the allocation is tract-heavy; an equal split is not much worse. For districts
 413 with population around 13,000, $\varepsilon = 1$ noising creates errors in the low hundreds for compact,
 414 connected districts, with the best performance for block-heavy allocations. Again, an equal
 415 split is not much worse, suggesting that this might be a good policy choice for accuracy in
 416 districts across many scales.

417 6 Geometrically compact vs hierarchically greedy districts

418 The analysis above suggests that the district-level error E_D will depend not only on the
 419 randomness of the noising algorithms, but also on the geometry of D and the structure of H .
 420 This section studies the hypothesis that districts that disrespect the geographical hierarchy
 421 will tend to have higher error magnitude. This section defines the *fragmentation score*,

relates a district’s fragmentation score to its error variance under **ToyDown**, and compares the fragmentation of two simple district-drawing algorithms on homogeneous hierarchies and simple geographies. Ultimately, we find that the explanatory value of the fragmentation score decays as we move to more realistic deployment of **TopDown**. This discrepancy raises important questions for future study: Which of the many additional features of **TopDown** attenuates the fragmentation–variance relationship?

We define a score intended to capture the contribution to $\text{Var}(E_D)$ of the shape of the district with respect to the hierarchy. Recall that \hat{h} denotes the parent of node h .

► **Definition 5** (Fragmentation score). For $D \subseteq H_d$, let $\text{Frag}(D) = \sum_{h \in H} (w_h - w_{\hat{h}})^2$.

Because weights are in $[0, 1]$, the score obeys $0 \leq \text{Frag}(D) < |H|$ for all districts, with higher scores indicating the presence of more units that are only partially included in D .

This fragmentation score is reverse-engineered from the expression for the variance of district-level population errors when using **ToyDown** with privacy divided equally across levels of the hierarchy (Corollary 3): namely, $\text{Var}(E_D) = \frac{8d^2}{\varepsilon^2} (w_1^2 + \text{Frag}(D))$. When the district D itself is a random variable sampled from some distribution, the expected fragmentation $\mathbb{E}(\text{Frag}(D))$ is similarly related to $\text{Var}(E_D)$. Namely, using the law of total variation, when each level gets ε/d privacy budget:

$$\text{Var}(E_D) = \mathbb{E}(\text{Var}(E_D|D)) + \text{Var}(\mathbb{E}(E_D|D)) = \mathbb{E}(\text{Var}(E_D|D)) = \frac{8d^2}{\varepsilon^2} (\mathbb{E}(\text{Frag}(D)) + \mathbb{E}(w_1^2)).$$

When ε is allocated unequally across levels, as for the other splits in Table 1, the simple analytical relationship between the fragmentation score and the error variance breaks down.

Observe that a hierarchy H does not capture all of the geometry relevant to district drawing. In particular, H does not directly encode any information about block adjacency, and therefore we can’t detect from H that a district is contiguous. For algorithms to generate contiguous districts, we need to make use of the plane geometry associated to H . We restrict our attention to the simplest case: homogeneous hierarchies (where every node on level $\ell < d$ has exactly n_ℓ children) and *square tilings*. (where each unit on level ℓ is a square and has n_ℓ children that cover it with a $\sqrt{n_\ell} \times \sqrt{n_\ell}$ grid tiling).

We analyze the fragmentation score for two simple district-drawing algorithms (see Appendix C). The **Greedy** algorithm builds a district from the largest subtrees possible, only subdividing a subtree when necessary. It takes as input H and $k \in \mathbb{N}$ and returns a district of size $N = \lfloor |H_d|/k \rfloor$, assembled by starting with the largest available units at random and adding units that are adjacent in the labeling sequence without passing size N , then allowing one partial unit, and so on recursively at lower levels. Observe that **Greedy** depends only on the hierarchy H . The **Square** algorithm takes as input a square, homogeneous hierarchy H and $k \in \mathbb{N}$ such that the district size is a perfect square, $|D| = |H_d|/k = s_d^2$. It outputs a uniformly random $s_d \times s_d$ square of blocks.

► **Theorem 6.** Let $D_G \sim \text{Greedy}(H, k)$, $D_\square \sim \text{Square}(H, k)$. For $n_1 \cdot n_2 \cdots n_{d-2} \geq k \geq 2$, let $L = \arg \min\{\ell : n_1 \cdot n_2 \cdots n_\ell \geq k\}$.

$$\mathbb{E}(\text{Frag}(D_G)) \leq \frac{k-1}{k^2} \sum_{\ell=1}^L n_\ell + \frac{1}{4} \sum_{\ell=L+1}^{d-1} n_\ell; \quad \mathbb{E}(\text{Frag}(D_\square)) \geq \frac{2}{3} \left(\frac{\sqrt{n_1 \cdots n_{d-1}}}{\sqrt{k}} - \frac{11}{2} \right) \sqrt{n_{d-1}}.$$

Dallas County is nearly a perfect square shape, so it gives us an opportunity to set some roughly realistic parameters to evaluate these bounds. There are 529 tracts in Dallas County,

with an average of 3.2 blocks groups per tract and 26.4 blocks per block group, yielding 44,113 total blocks. We can approximate these parameters by setting $d = 4$, using $k = 4$ as for the county commission districts, and setting $(n_1, n_2, n_3) = (484, 4, 25)$ which has a reasonably similar 48,400 blocks (as a result, $L = 1$). The bounds in the theorem say that $\mathbb{E}(\text{Frag}(D_G)) \leq 98$ and $\mathbb{E}(\text{Frag}(D_\square)) \geq 264$. Note: for homogeneous hierarchies H with equal-population leaves, the score $\text{Frag}(D_G)$ is independent of algorithm randomness and can be computed exactly; for the above parameters $\text{Frag}(D_G) = 90.75$. So the bound in the theorem is fairly tight, at least in this case.

To interpret the theorem, it is helpful to think of **Greedy** as being hierarchically greedy and **Square** as being geometrically greedy. That is, the former is oriented toward using the biggest possible units and keeping them whole, so that spatial considerations are secondary; the latter is oriented towards “compact” geographies with a lot of area relative to perimeter, and unit integrity is secondary. The theorem shows that compactness alone (a function of the plane geometry) does not keep down the fragmentation score (a function of the hierarchy), and indeed the bounds get farther apart as the hierarchy gets larger and more complicated. In Appendix C, we compare these theoretical results to empirical district errors, finding that fragmentation tracks well with errors in **ToyDown**, but that the complexity of the **TopDown** model weakens the relationship, suggesting a need for more sophisticated tools.

7 Ecological regression with noise

7.1 Inference methods for Voting Rights Act enforcement

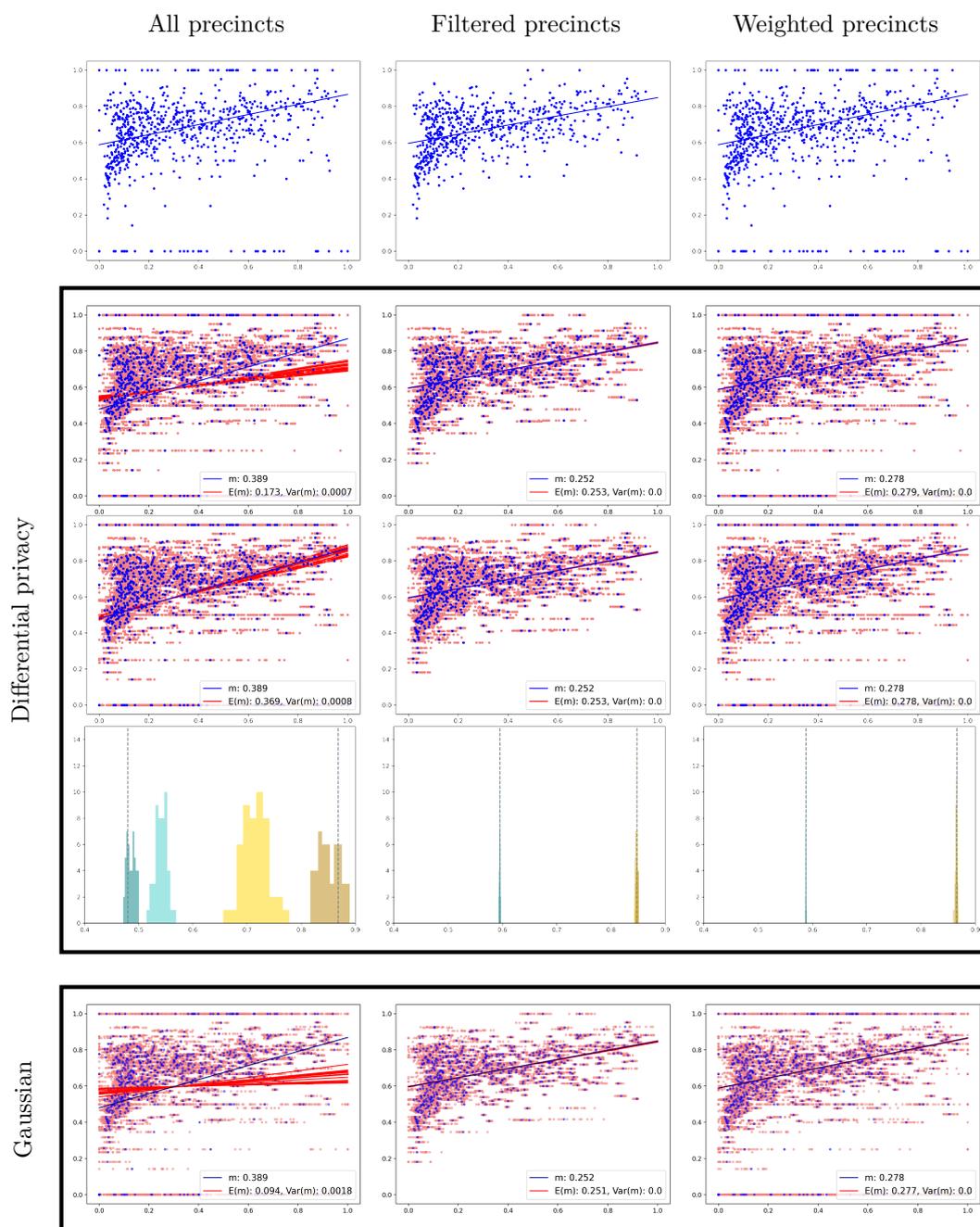
When elections are conducted by secret ballot, it is fundamentally impossible to precisely determine voting patterns by race from the reported outcomes alone. The standard methods for estimating these patterns use the cast votes at the precinct level, combined with the demographics by precinct, to infer racial polarization. Because the general aggregate-to-individual inference problem is called “ecological” (cf. ecological paradox, ecological fallacy), the leading techniques are called *ecological regression* (ER) and *ecological inference* (EI). It is rare that EI and ER do not substantively agree, and we focus on ER here because it lends itself to easily interpretable pictures.

ER is a simple linear regression, fitting a line to the data points determined by the precincts on a demographics-vs-votes plot. A high slope (positive or negative) indicates a likely strong difference in voting preferences, which is necessary to demonstrate the Gingles 2-3 tests for a VRA lawsuit.

The top row of Figure 5 shows standard ER run on the precincts of Dallas County, with each precinct plotted according to its percentage of Hispanic voting age population or HVAP (x -axis) and the share of cast votes that went to Lupe Valdez (y -axis). Strong racial polarization would show up as a fit line of high slope. This process produces a point estimate of Hispanic support for Valdez, found by intersecting the fit line with the $x = 1$ line, which represents the scenario of 100% Hispanic population. The point estimate of non-Hispanic support for Valdez is at the intersection of the fit line with $x = 0$.

7.2 Summary of Experiments

ToyDown and **TopDown** were both run on the full Texas reconstruction from 2010. We plotted Dallas County votes from three contests: votes for Obama for president in 2012 general election, votes for Valdez for governor in the 2018 Democratic Party primary runoff, and votes for Chevalier for comptroller in the 2018 general election. We chose these contests



■ **Figure 5** Comparing ecological regression on un-noised data (top row) with various styles of noising. ER is re-run on data noised by differentially private ToyDown (second row), and data noised by TopDown (third row), both with $\epsilon = 1$, equal split. The blue dots repeat the un-noised data, the pink dots show 16 runs of noised data with pink fit lines re-computed each time. Below that, the histograms show the point estimates of Latino (gold) and non-Latino (teal) support for Valdez estimated from ER on data noised by ToyDown (lighter) and TopDown (darker). The last row contrasts the differentially private algorithms with a naive variant that adds noise to each precinct from a mean-zero Gaussian distribution, set to match the average precinct level L^1 error observed in the ToyDown runs (in this case, this is $\sigma = 20$). Across all of these experiments, the conclusion is striking: TopDown performs better than ToyDown and far better than a naive Gaussian variant, even without filtering precincts; if precincts are filtered or weighted, none of the noising alternatives threatens the ability to detect racially polarized voting.

Race	All precincts (827)		Filtered precincts (626)		Weighted precincts (827)	
	this group	complement	this group	complement	this group	complement
Hispanic	0.869	0.480	0.848	0.596	0.866	0.588
Black	0.917	0.518	0.851	0.620	0.835	0.595
White	0.555	0.623	0.474	0.811	0.478	0.805

Race	Algorithm	statistic	All (827)		Filtered (626)		Weighted (827)	
			group	compl.	group	compl.	group	compl.
Hispanic	ToyDown	mean	0.715	0.541	0.848	0.595	0.867	0.588
Hispanic	ToyDown	variance	36000	7000	250	43	160	19
Black	ToyDown	mean	0.798	0.543	0.851	0.62	0.835	0.595
Black	ToyDown	variance	39000	2100	89	5.9	25	2.1
White	ToyDown	mean	0.476	0.674	0.473	0.811	0.478	0.805
White	ToyDown	variance	17000	8000	64	36	33	17
Hispanic	TopDown	mean	0.853	0.485	0.848	0.595	0.865	0.587
Hispanic	TopDown	variance	45000	6700	480	100	120	16
Black	TopDown	mean	0.91	0.52	0.85	0.62	0.835	0.595
Black	TopDown	variance	30000	1200	250	23	45	2.4
White	TopDown	mean	0.582	0.607	0.472	0.81	0.47	0.804
White	TopDown	variance	10000	3400	92	37	92	10

■ **Table 3** Point estimates from ER for Dallas County in the Valdez/White primary runoff in 2018. In the first table, estimates are made with (un-noised) VAP data from the 2010 Census. In the *filtered precincts* case, precincts with fewer than 10 cast votes are excluded from the initial set of 827 precincts. In the *weighted precincts* case, precincts are weighted by the number of cast votes. The *ToyDown* and *TopDown* estimates are made from VAP data from 16 runs with $\epsilon = 1$ and an ϵ -budget with all levels given equal weighting. Variance is the empirical variance over the repeated runs of the noising algorithm and is in units of 10^{-8} , shown to two significant digits.

500 because in each, ER finds evidence of strong racially polarized voting when using published
501 2010 census data. All three contests gave similar findings; we'll choose the Valdez runoff
502 contest as our focus here.

503 For both *ToyDown* and *TopDown*, we vary how we handle the inclusion of small precincts in
504 the ecological regression. The options are All (every precinct is a data point in the scatterplot,
505 all weighted equally); Filtered (only including precincts with at least 10 votes cast in that
506 election); or Weighted (weighting the terms in the objective function in least-squares fit by
507 number of votes cast). Filtering and weighting are done using the exact number of cast votes,
508 not the differentially private precinct population totals, which is realistic to the use case.

509 For each noising run we have a block- or precinct-level matrix, \hat{M} of noised counts, with
510 height b , the number of geographic units (blocks or precincts), and width c , the number of
511 attributes for which there are counts recorded. We also have a corresponding matrix M of
512 un-noised counts. We can compute the L_1 error by summing over the absolute value of every
513 entry in $M - \hat{M}$. *ToyDown* and *TopDown* were run 16 times for each configuration. Let E_{avg}
514 be the average L_1 error across noising runs.

515 If we add *Gaussian* noise to each count instead, the expected L_1 error is $\sum_{i,j} E[|X_{i,j}|]$,
516 where $X_{i,j} \sim \mathcal{N}(0, \sigma^2)$. This is the half-normal distribution, so $E[|X_{i,j}|] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$. We
517 rearrange to find the standard deviation $\sigma = \frac{E_{avg}\sqrt{\pi}}{bc\sqrt{2}}$ that defines the Gaussian distribution
518 (with $\mu = 0$), so that adding a random variable drawn from it to each unit count will produce
519 an expected L^1 error matching the average E_{avg} observed across the runs.

520 **7.3 The role of small precincts**

521 Practitioners who use ER have raised two questions regarding the effect of differential privacy:
 522 (1) How robust will the estimate be after the noising? (2) Will noising diminish the estimate
 523 of candidate support from a minority population? We analyzed the effects of TopDown and
 524 ToyDown on the 2018 Texas Democratic primary runoff election, where Lupe Valdez was a
 525 clear minority candidate of choice in Dallas county.¹

526 We begin by observing that of the 827 precincts in Dallas County, 201 have fewer than
 527 10 cast votes from that election day—in fact, 99 precincts recorded zero cast votes. These
 528 precincts are a big driver of instability under DP. This is not surprising; percentage swings
 529 are much higher in small numbers even if the noise injected might be low. However, down-
 530 weighting these small precincts makes the estimate almost always agree with the un-noised
 531 estimate. Specifically, we assign weights to the precincts equivalent to the number of total
 532 votes in the precinct. Figure 5 shows how the estimates vary by run type and data treatment.

533 **8 Conclusion**

534 The central goal of this study has been to take the concerns of redistricting practitioners
 535 seriously and to investigate potential destabilizing effects of TopDown on the status quo. A
 536 second major goal is to make recommendations, both to the Disclosure Avoidance team at
 537 the Census Bureau and to the same practitioners—the attorneys, experts, and redistricting
 538 line-drawers in the field. Texas generally, and Dallas County in particular, was selected
 539 because it has been the site of several interesting Voting Rights Act cases in the last 20
 540 years.²

541 Our top-line conclusion is that, at least for the Texas localities and election data we
 542 examined, TopDown performs far better than more naive noising in terms of preserving
 543 accuracy and signal detection for election administration and voting rights law. Perhaps
 544 more importantly, we have created an experimental apparatus to help other groups conduct
 545 independent analyses.

546 This work has led us to isolate several elements of common redistricting practice that lead
 547 to higher-variance outputs and more error under TopDown. The first example is the common
 548 use of a full precinct dataset, with no population weighting, in running racial polarization
 549 inference techniques. The second major example is the use of the smallest available units,
 550 census blocks, for building districts of all sizes, with no particular priority on intactness
 551 for larger units of Census geography. In both cases, we find that these were already likely
 552 sources of silent error. Filtering small precincts (or, better, weighting by population) and
 553 building districts that prioritize preserving whole the largest units that are suited to their
 554 scale are two examples of simple updates to redistricting practice. Besides being sound on
 555 first principles, these adjustments can insulate data users from DP-related distortions and
 556 help safeguard the important work of fair redistricting.

¹ We also examined the general elections for President in 2012 and Comptroller in 2018, with similar findings.

² This is a large county with considerable racial and ethnic diversity. Follow-up work will consider smaller and more racially homogeneous localities.

557 ——— **References** ———

- 558 1 *13 U.S.C. Section 9*. URL: <https://www.law.cornell.edu/uscode/text/13/9>.
- 559 2 John Abowd, Daniel Kifer, Brett Moran, Robert Ashmead, Philip Leclerc, William
560 Sexton, Simson Garfinkel, and Ashwin Machanavajjhala. Census topdown: Differentially
561 private data, incremental schemas, and consistency with public knowledge. 2019. URL:
562 [https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711_](https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf)
563 [0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf](https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf).
- 564 3 Avery v. Midland County, 390 U.S. 474 (1968).
- 565 4 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration*. URL:
566 <https://github.com/uscensusbureau/census2020-das-e2e>.
- 567 5 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration,*
568 *L1 metric*. URL: [https://github.com/uscensusbureau/census2020-das-e2e/blob/](https://github.com/uscensusbureau/census2020-das-e2e/blob/3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20)
569 [3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20](https://github.com/uscensusbureau/census2020-das-e2e/blob/3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20).
- 570 6 U.S. Census Bureau. *TopDown: Adding Geometric Noise to Counts*.
571 URL: [https://github.com/uscensusbureau/census2020-das-e2e/blob/](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678)
572 [d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678)
573 [L678](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678).
- 574 7 U.S. Census Bureau. *2010 Demonstration Data Products*, 2010. URL: [https://www.](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html)
575 [census.gov/programs-surveys/decennial-census/2020-census/planning-management/](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html)
576 [2020-census-data-products/2010-demonstration-data-products.html](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html).
- 577 8 U.S. Census Bureau. *2010 Census Summary File 1*, 2012. URL: [https://www.census.gov/](https://www.census.gov/prod/cen2010/doc/sf1.pdf)
578 [prod/cen2010/doc/sf1.pdf](https://www.census.gov/prod/cen2010/doc/sf1.pdf).
- 579 9 U.S. Census Bureau. *Census P.L. 94-171 Redistricting Data*, 2017. URL: [https://www.census.](https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html)
580 [gov/programs-surveys/decennial-census/about/rdo/summary-files.html](https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html).
- 581 10 Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: A family of markov chains
582 for redistricting. *arXiv preprint arXiv:1911.05725*, 2019.
- 583 11 Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings*
584 *of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database*
585 *systems*, pages 202–210, 2003.
- 586 12 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to
587 sensitivity in private data analysis. Halevi S., Rabin T. (eds) *Theory of Cryptography. TCC*
588 *2006. Lecture Notes in Computer Science*, 3876, 2006.
- 589 13 Peter Wayner JN Matthews, Bhushan Suwal. *Accompanying GitHub repository*. URL: [https:](https://github.com/mggg/census-diff-privacy)
590 [//github.com/mggg/census-diff-privacy](https://github.com/mggg/census-diff-privacy).
- 591 14 Denis Kazakov. *Census Scripts GitHub repository*, 2019. URL: [https://github.com/](https://github.com/94kazakov/census_scripts)
592 [94kazakov/census_scripts](https://github.com/94kazakov/census_scripts).
- 593 15 U.S. Census Bureau Michael Hawes. *Differential Privacy and the 2020 Decennial Census*, 2020.
594 URL: [https://www2.census.gov/about/policies/2020-03-05-differential-privacy.](https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf)
595 [pdf](https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf).
- 596 16 National Conference of State Legislatures. *2010 Redistricting Deviation Table*. URL: [https:](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx)
597 [//www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx)
598 [aspx](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx).
- 599 17 Reynolds v. Sims, 377 U.S. 533 (1964).
- 600 18 Wesberry v. Sanders, 376 U.S. 1 (1964).

601 **A** ToyDown and TopDown

602 ToyDown is described in Algorithm 2. It uses the *Laplace distribution* $\text{Lap}(b)$ with scale
 603 parameter b , i.e., the probability distribution over \mathbb{R} with mean zero and probability density
 604 function $\mathbb{P}[L] = \frac{1}{2b}e^{-|L|/b}$. It has variance $2b^2$. TopDown uses the *geometric* distribution, a
 605 discretized version of the Laplace distribution with integer support.

606 The inputs to TopDown are as follows. $A_{H,T} = \{a_{h,t}\}_{h \in H, t \in T}$, where $a_{h,t}$ is the number
 607 of people in h of type t ; $W = (Q_1, \dots, Q_{|W|})$ is a *workload* consisting of a collection of
 608 histograms Q ; $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$ is a hierarchical allocation of the privacy budget, with $\varepsilon_\ell > 0$
 609 at each level; $B : W \rightarrow [0, 1]$ with $\sum_{Q \in W} B(Q) = 1$ is a probability vector describing the
 610 relative privacy budget on each histogram in the workload; *invariants* V ; and *structural*
 611 *inequalities* S . We write $\mathbf{a}_h = \{a_{h,t}\}_{t \in T}$ (and $\boldsymbol{\alpha}_h$ analogously). For a query q , we write
 612 $q(\mathbf{a}_h) = \sum_{t \in q} a_{h,t}$ (and $q(\boldsymbol{\alpha}_h)$ analogously).

613 In the first stage (lines 2-5), a geometric random variable is added to the raw counts a to
 614 produce noised counts \hat{a} . In the second stage (lines 6-8), the noised counts are adapted to
 615 the nearest integer values that meet a collection of equality and inequality conditions. These
 616 equalities and inequalities, over the real numbers, describe a convex polytope; therefore the
 617 post-processing can be thought of geometrically as a closest-point projection to the integer
 618 points in the convex body under L^2 distance.

619 The noising stages of both ToyDown and TopDown are ε -differentially private for $\varepsilon =$
 620 $\sum_{\ell=1}^d \varepsilon_\ell$. In ToyDown, this stage can be viewed as generating a single histogram at each
 621 level ℓ using budget ε_ℓ . Following the Census Bureau, we use bounded differential privacy,
 622 wherein the global sensitivity of histogram queries is 2. In TopDown, the budget at level
 623 ℓ is further divided among the $|W|$ histograms Q in the workload, each receiving $B(Q)\varepsilon_\ell$
 624 of the budget. Because ToyDown's post-processing is data independent, ToyDown is ε -DP.
 625 TopDown's post-processing is not data independent: the invariants and structural inequalities
 626 may depend on the original data.

■ Algorithm 1 TopDown, based on [2]

```

1: procedure TOPDOWN( $A_{H,T}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d, W, B, V, S$ )
2:   for  $h \in H, Q \in W, q \in Q$  do
3:      $\beta \leftarrow \exp(-B(Q) \cdot \varepsilon_{\ell(h)}/2)$ 
4:      $G_{h,q} \leftarrow \text{Geom}(\beta)$  ▷ See [6]
5:      $\hat{a}_{h,q} \leftarrow q(\mathbf{a}_h) + G_{h,q}$  ▷ Geometric mechanism with
sensitivity 2, budget  $B(Q) \cdot \varepsilon_{\ell(h)}$ 

6:   for  $\ell = 1, \dots, d$  do
7:     Compute hierarchically-consistent ▷ A sophisticated heuristic algorithm
       non-negative integers  $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$  out of scope for this work
       minimizing  $\sum_{h \in H_\ell} \sum_{q \in W_\ell} (q(\boldsymbol{\alpha}_h) - \hat{a}_{h,q})^2$ ,
       subject to the invariants:  $v^*(\boldsymbol{\alpha}_h) = v^*(\mathbf{a}_h)$  for all  $h \in H_\ell, v \in V$ 
       and structural inequalities:  $s(\boldsymbol{\alpha}_h, \mathbf{a}_h) \leq 0$  for all  $h \in H_\ell, s \in S$ 

8:   return  $\{\alpha_{h,t}\}_{h \in H, t \in T}$ 

```

Algorithm 2 ToyDown

```

1: procedure TOYDOWN( $A_H = \{a_h\}_{h \in H}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d$ ) ▷ (Single attribute)
2:   for  $h \in H$  do
3:      $L_h \sim \text{Lap}(2/\varepsilon_{\ell(h)})$ 
4:      $\hat{a}_h \leftarrow a_h + L_h$  ▷ Laplace mechanism with sensitivity 2, budget  $\varepsilon_{\ell(h)}$ 
5:   for  $\ell = 1, \dots, d$  do
6:     Compute hierarchically consistent  $\{\alpha_h\}_{h \in H_\ell}$ 
       minimizing  $\sum_{h \in H_\ell} (\alpha_h - \hat{a}_h)^2$ 
7:   return  $\{\alpha_h\}_{h \in H}$ 

8: procedure MultiAttrTOYDOWN( $A_{H,T} = \{a_{h,t}\}_{h \in H, t \in T}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d$ )
9:   for  $h \in H, t \in T$  do
10:     $L_{h,t} \sim \text{Lap}(2/\varepsilon_{\ell(h)})$ 
11:     $\hat{a}_{h,t} \leftarrow a_{h,t} + L_{h,t}$  ▷ Laplace mechanism with sensitivity 2, budget  $\varepsilon_{\ell(h)}$ 
12:   for  $\ell = 1, \dots, d$  do
13:     Compute hierarchically consistent
       (optionally, non-negative)  $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$ 
       minimizing  $\sum_{h \in H_\ell, t \in T} (\alpha_{h,t} - \hat{a}_{h,t})^2$ 
14:   return  $\{\alpha_{h,t}\}_{h \in H, t \in T}$ 

```

B Detailed materials and methods
B.1 Primary data sources

2010 US Census demographic data was downloaded using the Census API, and the 2010 census block, block group, and tract shapefile for Dallas County were downloaded from the US Census Bureau’s TIGER/Line Shapefiles. For our VRA analysis, we obtained both statewide election results and a statewide precinct shapefile from the Texas Capitol Data Portal, which we then trimmed to the precincts within Dallas County.³

We use a person-level dataset obtained by applying a reconstruction technique to the block-level data from Texas from the 2010 Census.⁴ The reconstructed microdata records contain block-level sex, age, ethnicity, and race information consistent with a collection of tables from 2010 Census Summary File 1. We note that this reconstruction follows the same strategy used by the Census Bureau itself as the first step of its reidentification experiment [15], based on [11].

The reconstructed data is far from perfect. Unlike the Bureau, we do not have access to the ground truth data needed to quantify the errors. The Bureau’s own reconstruction experiment reconstructed 46% of entries exactly, plus an additional 25% within ± 1 year error in age [15]. We note that our reconstructed data contains no household information, because this was not present in the tables used in the constraint system. This is significant because the TopDown configurations for the US Census Bureau’s 2010 Demonstration Data Products [7] include household-based workload queries and invariants.

³ Data comes from data.capitol.texas.gov/topic/elections and data.capitol.texas.gov/topic/geography.

⁴ A team led by data scientist and journalist Mark Hansen at Columbia, including Denis Kazakov, Timothy Donald Jones, and William Reed Palmer, designed an algorithm to solve for the detailed data, which we describe in this section. Code is available upon request [14].

647 **B.2 TopDown configuration**

648 The exact configuration files and code for all the runs are available in this paper’s accompanying
 649 repository [13]. The TopDown code used for this paper was modified from the publicly
 650 available demonstration release of the US Census Bureau’s Disclosure Avoidance System
 651 2018 End-to-End test release [4]. The input data fed to the algorithm was obtained by
 652 restructuring the reconstructed 2010 block-level Texas microdata into the 1940s IPUMs
 653 data format. Most importantly, the reconstructions allowed for 63 distinct combination of
 654 races whereas the End-to-End release only allows for 6 races, so all multi-racial entries were
 655 re-categorized as Other in our TopDown runs.

656 Because TopDown’s post-processing is done level by level, the noisy counts in Dallas
 657 County do not depend on the noisy counts at the tract-level or below in counties other than
 658 Dallas. We modified the census reconstructed data to focus on Dallas county and minimize
 659 the computation time spent processing the other 253 counties in Texas. Specifically, for every
 660 non-Dallas county, we placed all of the population into a single block.

661 We do not enforce certain household invariants that the Census Bureau is planning to
 662 enforce, and our workload omits household queries that are used in Census’s demonstration
 663 data products. Our choice to omit household queries and invariants is result of our use of
 664 reconstructed 2010 census microdata which does not include household information. We
 665 did perform additional runs with household invariants and queries using crude synthetic
 666 household data, the results of which are available in the data repository accompanying this
 667 paper [13]. In those runs, the population in each block was grouped into households of size 5
 668 with at most one group smaller than 5. Ultimately, we focused on the experiments that did
 669 not require synthetic household data.

670 The TopDown runs without the household workload or invariants use a workload consisting
 671 of two histograms: $Q_{detailed}$ and $Q_{va,eth,race}$ with 10% and 90% of the budget respectively.
 672 (The additional runs with households includes an additional households and group quarters
 673 histogram in the workload assigned 22.5% of the budget, leaving 10% and 67.5% for $Q_{detailed}$
 674 and $Q_{va,eth,race}$ respectively.) The End-to-End TopDown code reports a differentially private
 675 estimate of the L^1 error with $\varepsilon = 0.0001$ not included in privacy budget specified elsewhere
 676 in the configuration file and discussed elsewhere in this paper.

677 **C District fragmentation****Algorithm 3 Greedy**

```

1: procedure GREEDY( $H, k$ )
2:   if  $k = 1$  then
3:     Return  $H$ 
4:    $N \leftarrow \lfloor |H_d|/k \rfloor$ ,  $D \leftarrow \emptyset$ ,  $h^* \leftarrow h_1$ 
5:   while  $N > 0$  do
6:     For  $h^*$  and  $D$ , let  $S(h^*, D)$  be the set of
       children  $h$  of  $h^*$  that are disjoint from  $D$ .
7:     while  $\exists h \in S(h^*, D) : |h| \leq N$  do
8:        $D \leftarrow D \cup h$             $\triangleright$  Associating  $h$  with the blocks descendent from it
9:        $N \leftarrow N - |h|$ 
10:    Pick  $h^* \in S(h^*, D)$ 
return  $D$ 

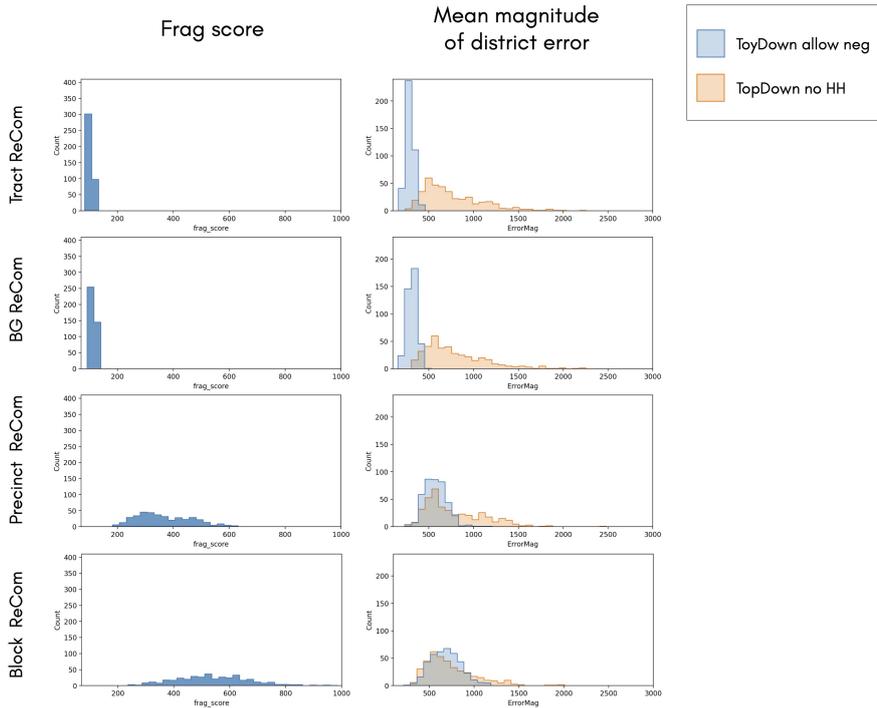
```

■ **Algorithm 4** Square

-
- 1: **procedure** SQUARE(H, k)
 - 2: $s_d \leftarrow \sqrt{|H_d|/k}$ ▷ Side length in blocks of the district
 - 3: $S_d \leftarrow \sqrt{n_1 \cdot n_2 \cdots n_{d-1}}$ ▷ Side length in blocks of the region
 - 4: Sample $i, j \in \{1, \dots, S_d - s_d + 1\}$ uniformly at random
 - 5: **return** $D_{i,j}$, the square district with top left corner at (i, j)
-

678 In Section 6, we defined the fragmentation score and its relationship to error variance for
 679 ToyDown, and analyzed the expected fragmentation score of districts produced by different
 680 district drawing algorithms. Now we apply TopDown to examine the relationship between a
 681 district’s population error and geometry, as captured by the fragmentation score.

682 We fix the a total budget and an equal allocation across levels: $0.2 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \varepsilon_5 =$
 683 ε_6 , as in Table 1. (We do not need to noise the nation because we are focusing on Texas; we
 684 do need to noise Texas even though its total population is invariant, because its population
 685 by race is allowed to vary.) We apply ReCom to build districts out of tracts, block groups,
 686 and blocks—all of which are part of the census hierarchy—and add a realistic variant that
 687 builds from whole *precincts*. These are about the same size as block groups and are more
 688 commonly used in redistricting.



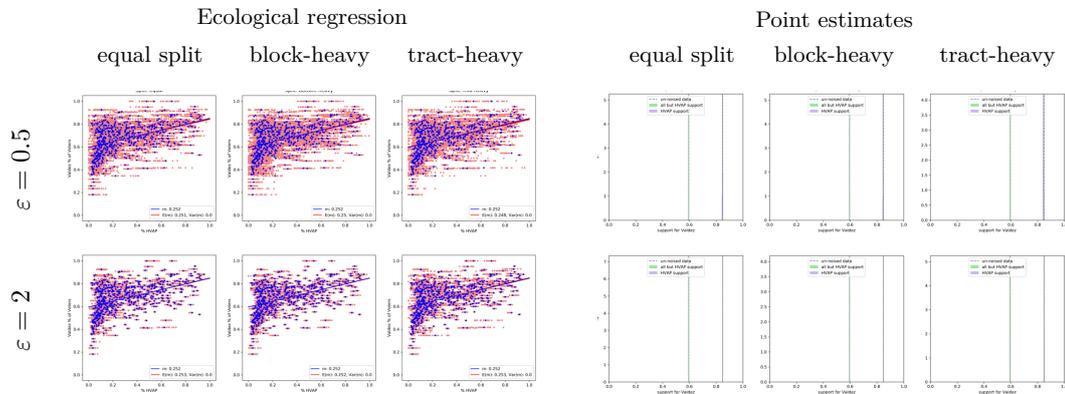
■ **Figure 6** Do the building-block units of districts matter? Histograms of fragmentation score (left column) and mean error magnitude (right column) are shown across four district-drawing algorithms that prioritize compactness. (Dallas County, $k = 4$.) We see that using larger units leads to significantly lower fragmentation and correspondingly low district-level error in ToyDown, but the advantage erodes when we pass to TopDown.

689 Figure 6 plots the data from our experiments. Each of the 12 histograms displays 400
 690 values, one for each district drawn by the specified district-drawing algorithm. The histograms
 691 on the left plot the fragmentation score of each district; the histograms on the right plot the
 692 mean observed district-level population error magnitude over 16 executions of the specified
 693 hierarchical noising algorithm.

694 The size of the constituent units is observed to have a controlling effect on the fragmentation
 695 score, as expected. As we would expect, this carries over to the simplest ToyDown (allowing
 696 negativity). (Note that since the error has zero mean, higher variance drives up the mean
 697 magnitude of error.) But the choice of base units makes far less difference by the time we
 698 move to full TopDown. These observations are consistent, again, with a strong similarity
 699 across spatially nearby units. All four kinds of ReCom will tend to produce compact, squat
 700 districts whose units are more closely geographically proximal than would be observed with
 701 disconnected or elongated shapes. Random noise is uncorrelated, but the post-processing
 702 effects can be highly spatially correlated because of spatial relationships in the underlying
 703 counts by race, ethnicity, and voting age.

704 **D Robustness of noisy ER**

705 Figure 7 extends the findings from Figure 5 with more splits and allocations, showing that
 706 as long as small precincts are filtered out, ecological regression for RPV analysis in Dallas
 707 County is robust to changes in the allocation of the privacy budget across the levels of the
 708 hierarchy and the total privacy budget for TopDown. The corresponding plots for ToyDown
 709 are essentially indistinguishable. (ER with precincts weighted by population is similarly
 710 robust.)



■ **Figure 7** Ecological regression for the Valdez-White runoff election with $\epsilon = .5$ and $\epsilon = 2$ and three different budget allocations, together with corresponding point estimates for Latino and non-Latino support for Valdez, with small precincts filtered out as in Figure 5. Findings stay remarkably stable.