Quantifying Gerrymandering in North Carolina

Gregory Herschlag[a,b] and Han Sung Kang[a,c,d] Justin Luo[a,d] Christy Vaughn Graves[f] Sachet Bangia[e] Robert Ravier[a] Jonathan C. Mattingly*[a,g]

[a]Department of Mathematics, Duke University
[b]Department of Biomedical Engineering, Duke University
[c]Department of Computer Science, Duke University
[d]Department of Electrical and Computer Engineering, Duke University
[e]Department of Economics, Duke University
[f]Program in Applied and Computational Mathematics, Princeton University
[g]Department of Statistical Science, Duke University

October 15, 2019

Abstract

Using an ensemble of redistricting plans, we evaluate whether a given political districting faithfully represents the geo-political landscape. Redistricting plans are gathered from a Monte Carlo algorithm and adhere to realistic and non-partisan criteria. Using the collection of redistricting plans along with historical voting data, we produce an ensemble of elections that reveal geo-political structure within the state. We showcase our methods on the two most recent districtings of NC for the U.S. House of Representatives, as well as a plan drawn by a bipartisan redistricting panel. We find the two state enacted plans are highly atypical outliers whereas the

*A preliminary version of the results presented in this paper were first reported to the arXiv repository [3]. The Duke Math Department, the Information Initiative at Duke (iID), the PRUV and Data+ undergraduate research programs provided financial and material support. Bridget Dou and Sophie Guo were integral members of the Quantifying Gerrymandering Team and have significantly influenced this work. Mark Thomas, and the Duke Library GIS staff helped with data extraction. Robert Calderbank, Galen Reeves, Henry Pfister, Scott de Marchi, and Sayan Mukherjee have provided help throughout this project. John O'Hale procured the 2016 election Data. Tom Ross, Fritz Mayer, Land Douglas Elliott, and B.J.Rudell invited us observe the Beyond Gerrymandering Project and this work relies on what we learned there.

bipartisan plan accurately represents the ensemble both in partisan outcome and in the fine scale structure of district-level results.

An earlier version of this analysis was central to the lower court finding in *Common Cause v. Rucho*; the lower court found that the NC Congressional district maps were a political gerrymander. Analysis which builds on the techniques described here was central to the state court finding in *Common Cause v. Lewis* that the N.C. State Legislative maps violated the state constitution and had to be redrawn.

Keywords: Gerrymandering | Redistricting | Monte Carlo Sampling | Common Cause v. Rucho

In the 2012 NC congressional election, over half the total votes went to Democratic candidates, yet only four of the thirteen congressional representatives were Democrats. Furthermore, the most Democratic district had 29.63% margin of victory, whereas the most Republican district had a 13.11% margin of victory. These results may be due to political gerrymandering or, alternatively, be natural outcomes of NC's geo-political structure as determined by the spatial distribution of partisan votes.

A partisan gerrymandering alters election results away from what would have happened under a 'fair' or neutral redistricting process. To detect a gerrymander we must first understand what this baseline is, along with how it may fluctuate. To understand the how redistricting in a neutral setting would impact elections, we probe the geo-political structure by (i) gathering a representative collection of redistricting plans that adhere to non-partisan redistricting criteria; (ii) next we simulate an election with each of our sampled redistricting plans using the actual partisan votes cast by North Carolinians in the 2012 and 2016 congressional elections and a collection of statewide contests; and (iii) we aggregate election results to construct the distributions of partisan vote balance on each district and of the congressional delegation's partisan composition. Districts that do not respect typical election results are considered gerrymandered. When a districting is gerrymandered, the congressional delegation's partisan composition may not be representative of what is typical. We use a collection of different elections as each election provides a different instrument to probe the structure of a given map. Each election not only has a different statewide partisan vote fraction, but also has different spatial voting patterns. For example, sometimes the suburbs vote with the city core and sometimes they do not.

Having probed the impact of the geo-political structure, we analyze three specific districting plans: the two most recent districting plans of NC for the U.S. House of Representatives and a plan proposed by a bipartisan panel of retired NC judges. By situating the election outcomes of these three districting plans in our sampled ensemble, we determine whether the three districting plans contain unlikely partisan favoritism and thwart the underlying geo-political structure, as expressed by the people's votes, by shifting each district's partisan vote balance significantly away from what is typical.

An earlier version of this analysis played a central in role in *Common Cause v. Rucho* which was the second federal judgement in the past 30 years that declared a districting map was a political gerrymander. The case was brought to the U.S. Supreme court, and the analysis was discussed during oral arguments. The Court ultimately decided that partisan gerrymandering was not a matter for the Federal Courts. Subsequently, a related analysis, adapted to the North Carolina Legislative setting, was central in state court case *Common Cause v. Lewis* which found the N.C. State Legislative maps in violation of the state constitution and resulted in their redrawing.

The qualitative results and subsequent conclusions of this note agree fully with those in earlier iterations of this analysis reported in [3]. They also agree in identifying the current maps as outliers with the much more preliminary and simplistic analysis performed in [24] which began the author's investigations of these questions.

For elaborations of the ideas presented here and applications to other settings, please see the Quantifying Gerrymandering Blog [18].

Methods

To sample from the space of congressional redistricting plans, we construct a family of probability distributions that are concentrated on plans adhering to non-partisan design criteria from proposed legislation. The non-partisan design criteria ensures that

1. the state population is evenly divided between the thirteen congressional districts,
2. the districts are connected and compact,
3. splitting counties is minimized, and
4. African-American voters are sufficiently concentrated in two districts to affect the winner.

The first three criteria come from House Bill 92 (HB92) of the NC General Assembly, which passed the House during the 2015 legislative session. HB92 also states that a districting should comply with the Voting Rights Act (VRA); the fourth criteria ensures the VRA is satisfied and is based on a redistricting plan proposed by the legislature along with recent court rulings. HB92 proposed establishing a bipartisan redistricting commission guided solely by these principles (see Section S1.2 of the supplemental material for the precise criteria).

There is no consensus probability distribution to select compliant redistricting plans. For example, there is no criteria that determines when a plan contains districts that are not compact enough; it is also unclear if a distribution of plans should more heavily weight more compliant plans, or whether it should equally weight all compliant plans. In short, there is no 'correct' choice for this distribution. However, our experience to date is that most resonable distributions identify the same maps as outliers. We define a particular score function and use it to define a Gibbs distribution; for our main results, we find compliant plans through a simulated annealing procedure based on this distribution. We demonstrate that our proceedure converges to a fixed distribution and that our results are remarkably stable when the distribution is changed (see Section S3.3 of the supplemental material). We also demonstrate that our results are remarkably stable when changing annealing parameters and when redefining compliance for redistricting plans (see Section S3 of the supplemental material).

Redistricting plans are sampled with a standard Markov Chain Monte Carlo algorithm combined with a simulate annealing heating and cooling schedule. About 66,500 random redistricting plans were produced. Additionally, we demonstrated the convergence and robustness of our sampling proceedure in Sections S2, and S3 of the supplemental material). For each generated redistricting, we re-tally the actual historic votes from a variety of electoral races, including the 2012 and 2016 US congressional elections, producing ensembles of election outcomes. When re-tallying the votes, we make the assumption that people vote for parties rather than people. We use this ensemble of election outcomes to quantify how representative a particular districting is by observing its place in this collection; we also use the ensemble to quantify gerrymandering by identifying districts which have an atypical partisan concentration of voters.

We analyze and critique the NC U.S. Congressional districting plans used in the 2012 and 2016 elections, as well as the districting developed by a bipartisan group of retired judges as part of the "Beyond Gerrymandering" project spearheaded by Thomas Ross and the Duke POLIS Center [34]. We refer to these districting plans of interest as NC2012, NC2016, and Judges respectively (see Figure S1 for the district maps).

Using a related methodology, we assess to what degree three districting plans of interest (NC2012, NC2016, and Judges) are gerrymander when compared to plans with similar spatial structure. This is done by seeing how close each districting plans' properties are to the collection of nearby redistricting plans. Small changes to district boundaries should not have a significant effect on the character of election results.

Results

Using our ensemble of over 66,500 redistricting plans, we tabulate the observed probability distribution of the congressional delegation's partisan composition for the 2012 and 2016 vote counts. We then situate the NC2012, NC2016 and Judges districting plans on this probability distribution (see Figure 1). The partisan composition of the NC2012 and NC2016 districting plans occur in less than 0.3% and 1.1% of our generated redistricting plans for the 2012 and 2016 congressional elections, respectively, and is heavily biased toward the Republicans. When repeating this analysis for the 2012 presidential race and the 2016 United states senatre race, we find that the partisan composition of the NC2012 and NC2016 districting plans occur in less than 0.2% and 0.8% of our generated redistricting plans, respectively. In contrast, the partisan composition of Judges districting plan occurs in 33.5% and 30% of our generated redistricting plans for the 2012 and 2016 congressional votes, respectively; similarly, the partisan composition of the Judges districting plan occurs in 39% and 27% of our generated redistricting plans for the presidential and senate races, respectively. In all four cases, the Judges plan provides the second most likely outcome.

By keeping the vote counts fixed and changing district boundaries, we have ignored any impact on incumbency. To test the effects of incumbency, we repeat the above analysis over a range of historic elections that include senatorial, presidential, gubernatorial races occurring between 2012 and 2016. The statewide races have the further advantage over the congressional elections in that that they are not influenced by the district boundaries which we are critiquing. We plot the histograms as a function of the statewide Democratic vote fraction in Figure 2. Each election provides a different tool to probe the properties of the maps under discussion as (i) each election leads to a different state wide fraction, but

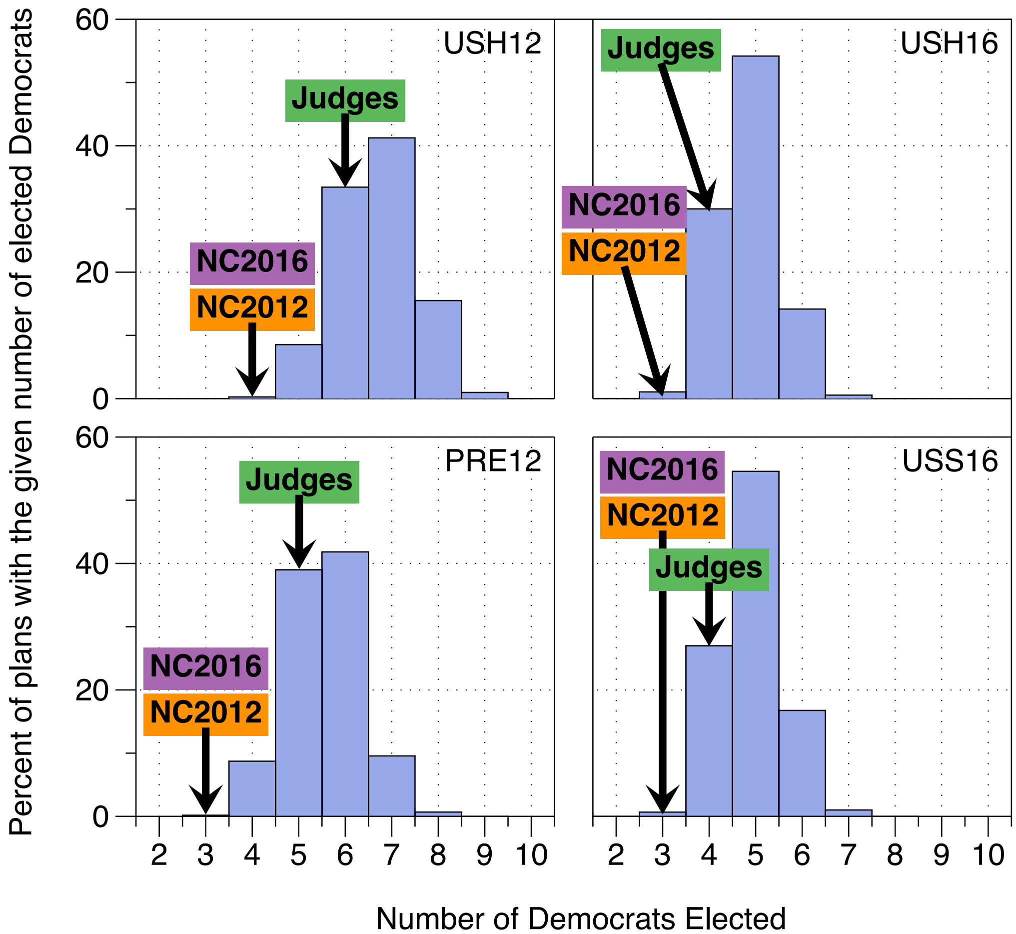

Figure 1: Probability of a given number of Democratics elected among the 13 congressional seats using votes from the 2012 elections (left) and 2016 elections (right), using the congressional elections (USH; top) and the statewide elections (bottom) of the presidential race (PRE12) and the United States Senate race (USS16).

(ii) each election also contains variations in the spatial patterns of the vote. The NC2012 and NC2016 districting plans robustly elect three Democratic candidates over nearly all sets of examined historical voting data, and these results nearly always occur in less than 1% of the ensemble of redistricting plans when examined on the same set of election data. In contrast, the Judges plan gradually shifts from electing four to six Democrats as the statewide Democratic vote fraction changes from 44% to 52% of the vote; when situated within the ensemble of redistricting plans, the results are nearly always one of the two most expected outcomes.

Although the above results are compelling, the partisan balance in election outcomes is not the only signature of gerrymandering and gives little detail of the structures that produce the atypical results. To further probe the geo-political structure, we order the thirteen congressional districts in any given redistricting from the lowest to highest percentage of Democratic votes in each district to construct an ordered thirteen dimensional vector. For each index, we construct a marginal distribution. We summarize the thirteen distributions in a classical box-plot in Figure 3. The full extent of the box plots represent the mininum and maximum ovserved values and the dashes represent the 1% and 10% outliers. On these box-plots, we overlay the percentage of the Democratic vote for the ordered districts in the NC2012, NC2016, and Judges districting plans. We display the resulting box plots when considering the 2012 congressional and presidential votes, along with the 2016 congressional and US senate votes.

The structure of the box plot reveals interesting features in the three examined districting plans. The Judge's districts gradually increase, roughly linearly, from the most Republican district (labeled 1) to the most Democratic (labeled 13); this behavior is identical to the behavior of the marginal distributions. Furthermore, most of the voting percentages from the Judges districts fall inside the boxes on the box-plot which mark the central 50% of the marginal distributions. The NC2012 and NC2016 districting plans have a different structure: Both plans jump in partisan voter percentage between the tenth and eleventh most Republican districts (labeled 10 and 11, respectively). In the NC2012 districting, the fifth through tenth most Republican districts have more Republicans than predicted by the ensemble (labeled 5-10). In the NC2016 districting, the sixth through the tenth

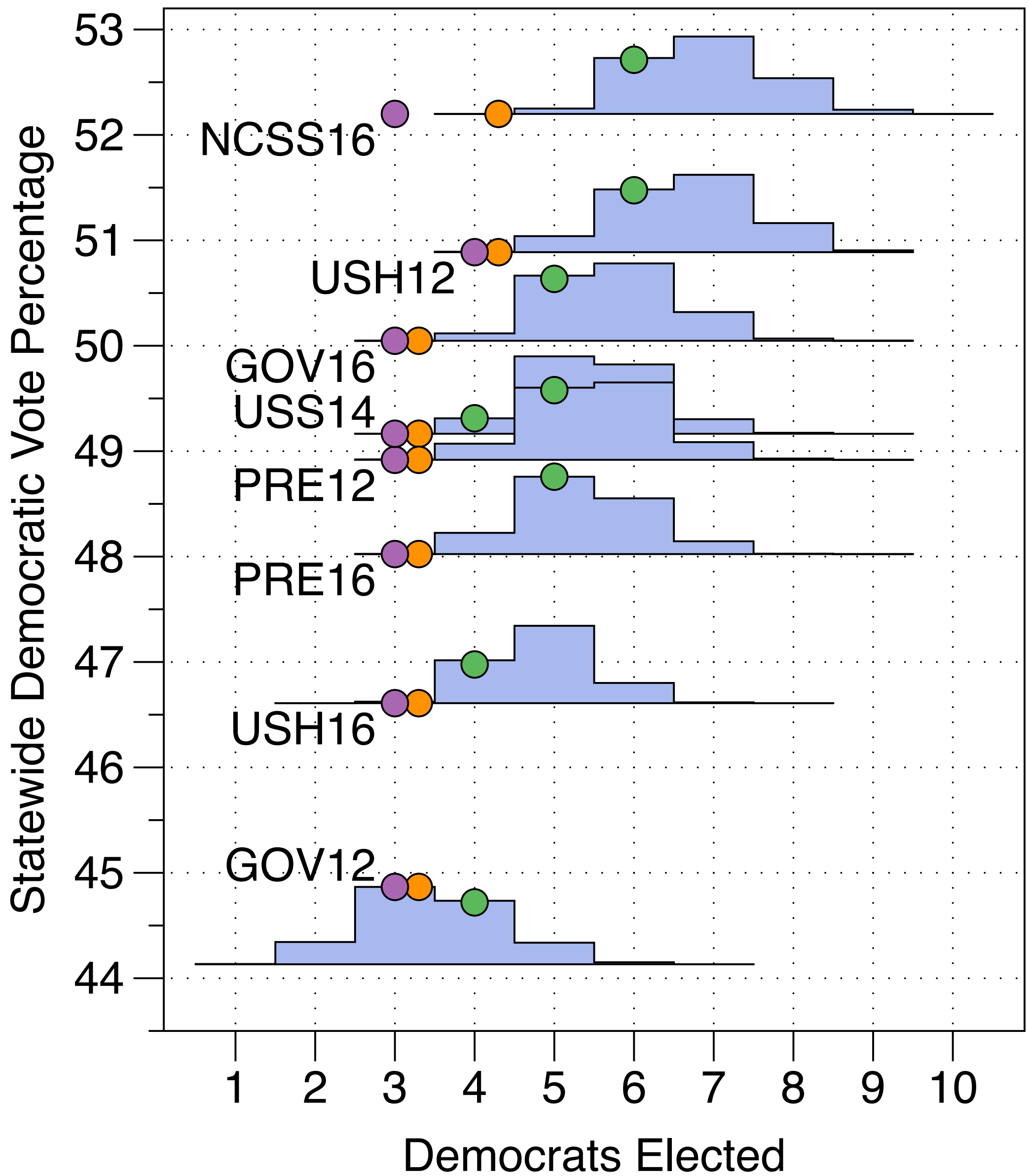

Figure 2: Probability of a given number of Democratic wins among the 13 congressional seats using vote counts from a variety of elections. The y-axis shows the statewide democratic vote fraction. Elections shown are the 2012 and 2016 presidential races (PRE12, PRE16), the 2016 North Carolina secretary of state race (NCSS16), the 2012 and 2016 gubernatorial races (GOV12, GOV16), the 2014 and 2016 US senatorial races (USS14), and the 2012 and 2016 US congressional races (USH12, USH16; also shown in Figure 1).

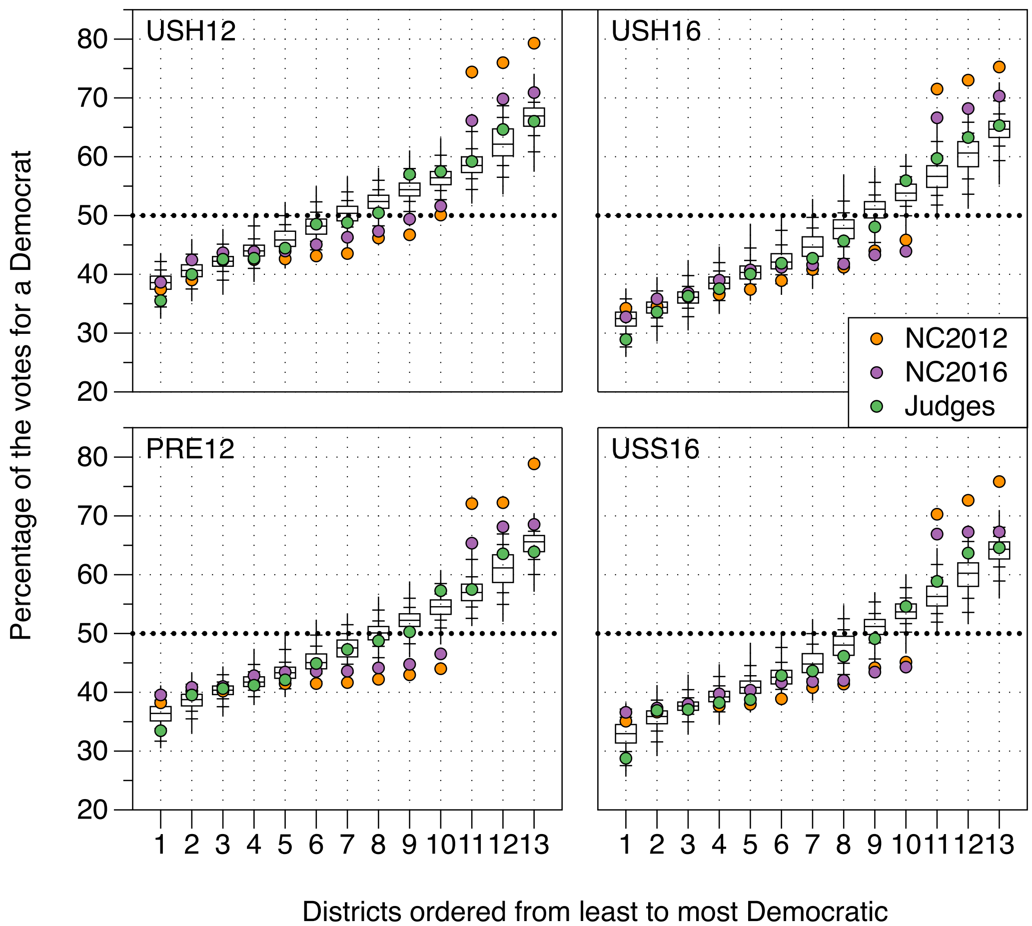

Figure 3: Box-plot summary of districts ordered from most Republican to most Democratic, for the congressional voting data from 2012 (left) and 2016 (right), using We compare our statistical results with the three redistricting plans of interest – the NC2012 plan (orange), NC2016 (purple) and Judges plan (green).

most Republican districts have more Republicans than predicted by the ensemble. In both the NC2012 and NC2016 districting plans, Democratic votes are removed from the central districts and added to the three most Democratic districts (labeled 11-13) – this is strong evidence that the middle districts have been cracked to reduce the Democrats' influence whereas the most Democratic districts have been packed with more Democrats than is expected.

To quantify this observation, there are nearly no eleventh districts within the ensemble that have more Democratic votes than the third most Democratic district (11) in the NC2012 and NC2016 data. Similarly there are nearly no fourth most Democratic districts (10) within the ensemble that have fewer Democratic voters than the NC2012 and NC206 redistricting plans. The only exceptions to this trend are in the NC2016 plan under the 2012 votes: There are 51 fourth most Democratic districts in the ensemble (0.08%) with fewer Democratic voters than the fourth most Democratic district of the NC2016 plan, and there are 59 eleventh districts in the ensemble (0.09%) with more Democratic voters than the third most Democratic district (11) of the NC2016 plan.

Packing and cracking potentially reduce a party's political power. When considering the 2012 votes, the box-plot demonstrates that the 8th-10th least Democratic districts (8-10) are typically above the 50% line, meaning that we expect these districts to elect a Democratic representative; when comparing these districts to the NC2012 and NC2016 districting plans, we see that the 8th-10th districts fall below this line, meaning that these districts elected a Republican. When considering the 2016 votes, the NC2012 and NC2016 districting plans lead to a similar change in the delegation's partisan composition when compared to the ensemble.

In addition to counting number of elected officials for each party, Figure 3 demonstrates partisan safety – districts 6-10 are all more robustly republican than they would otherwise be. The effect is that elections within these districts may become more robustly Republican. This effect maybe seen in both the NC2012 and NC2016 plans for districts labeled six through ten under the 2012 vote counts, and districts labeled eight through ten under the 2016 vote counts.

Under a standard uniform partisan swing hypothesis, a statewide shift in the votes results in the box-plots shifting globally up or down in the direction of the swing (e.g. [19]). Hence, under this assumption, the jump in partisan vote fraction in the NC2012 and NC2016 plans results in a wide range of statewide partisan outcomes that produce an identical partisan composition of the Congressional delegation. This effect is absent from the typical ensemble plans as well as the Judges's plan and is demonstrated clearly in Figure 2.

The Democratic vote fractions of the NC2012 and NC2016 plans show a large jump between the thrid and fourth most Democratic districts (labeled 10 and 9 respectively); in comparison, ensemble and Judges's plan have ranked districts that increase fairly linearly and gradually in Democratic vote fractions. The atypical structure in the NC2012 and NC2016 plans provides a signature of gerrymandering. This structure further reveals the districts which have had votes from one party removed and dispersed to other districts. The cracking and packing dilutes the Democratic party's political power. In the next section, we further quantify this signature by contextualizing the plans of interest within the ensemble of redistricting plans when analyzed with summary metrics.

Summary Metrics

Although the above visualizations provide a clear picture of the signature of gerrymandering, there is a long history of employing summary metrics that seek to encapsulate the above structures with a single number: Such metrics include electoral responsiveness, partisan bias, the efficiency gap, mean-median difference, declination and more (see, for example [14, 32, 35, 9, 36]. Typically these metrics are contextualized with historical data across past elections, districting plans, and states. However it is unclear how meaningful the measures are, as they fail to consider the geo-political makeup of a region (for example, see [5, 8]). For example the geo-political makeup of a state may cause it to have a natural partisan biased [5]. Hence, zero partisan-bias might not be a realistic or even desirable goal. This is particularly at odds with how such measure are often discussed in the popular press.

Based on these observations, we propose two novel metrics that contextualize a redistricting plan within the underlying space of possible plans: The two indices are

- *Gerrymandering Index:* Quantifies how typical the observed level of packing and cracking is for a given redistricting by measuring how the individual districts deviate from the expected percentage of partisan votes.

- *Representativeness Index:* Measures how typical the resulting balance of power obtained by a given districting plan is by contextualizing the number of elected representatives with the ensemble of redistricting plans.

All of the above mentioned summary statistics are based on the ordered districting vote percentages shown as the dots in Figure 3, however the two novel statistics also utilized marginal distributional data (shown as the box plots in Figure 3).

In the present work, we consider two of the established statistics – partisan bias and the efficiency gap. We also consider the two new proposed statistics. We provide detailed descriptions of all utilized indices below. Each summary statistic is computed for each redistricting plan in our ensemble and for each districting plan of interest (NC2012, NC2016, and Judges).

We contextualize the partisan bias and the efficiency gap, using the 2012 and 2016 congressional voting data, with histograms in Figure 4. In both statistics and under both vote counts, the NC2012 is an extreme outlier; it shows extreme bias toward the Republican party and wastes an atypical number of Democratic party. The NC2016 map is not an extreme outlier under either set of vote counts with respect to partisan bias, but is an extreme outlier with respect to the efficiency gap; under both metrics, however, it favors the Republican party. Under the 2012 voting data, only 231 of the generated redistricting plans (0.35% of the plans) are as, or more, biased than the NC2012 districting plan; 3603 redistricting plans (5.4%) are as, or more, biased than the NC2016. For the 2016 voting data, only 161 of the generated redistricting plans (0.24%) are as, or more, biased than the NC2012 districting plan; 4179 plans (6.28%) are as, or more, biased than the NC2016 plan. Under the 2012 and 2016 voting data, the NC2012 districting plan has a higher efficiency gap than all but 2 (0.003%) and 7 (0.01%) of the generated redistricting plans, respectively. The efficiency gap for the NC2016 map is lower than 171 of the redistricting plans in the ensemble (0.26%) for the 2012 voting data, and 489 of the redistricting plans in the ensemble (0.73%) for the 2016 voting data.

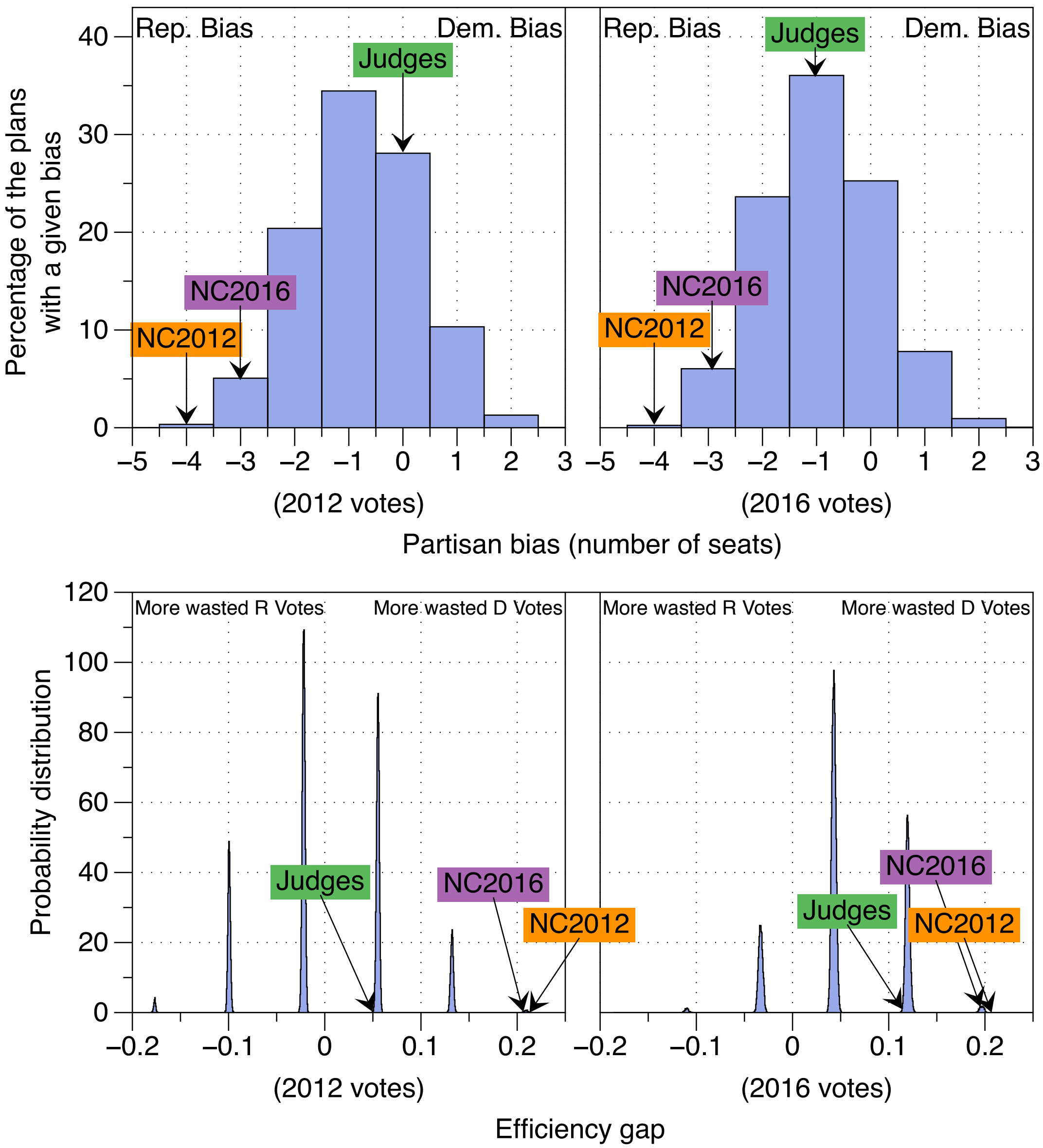

Figure 4: Partisan bias (top) and the efficiency gap (bottom) for the three districts of interest and the ensemble of plans. The data is based on the voting data from 2012 (left) and 2016 (right) congressional races.

In stark contrast, the Judges districting plan has no partisan bias under the 2012 votes and is only slightly biased under the 2016 votes. Under the 2016 votes, however, the bias in the Judges plan takes on the most likely value of the ensemble. This demonstrates the natural Republican bias inherent geography of the votes. In terms of the efficiency gap, the Judges districting plan wastes fewer Democratic votes than 28,125 of the redistricting plans in the ensemble (42.3%) and 20,652 redistricting plans (31.0%) under the 2012 and 2016 votes, respectively.

We also contextualize the Gerrymandering Index and Representativeness index of the three plans of interest within the distribution of the ensemble of plan (see Figure 5). None of the generated redistricting plans constructed have a partisan bias Gerrymandering Index bigger than NC2012 and NC2016 plans, regardless of the voting data used. Similarly, none of the redistricting plans have a Representativeness Index greater than NC2012 plan when the 2012 votes are used and only 620 of the redistricting plans (0.93%) have a greater Representativeness Index when the 2016 votes are used. Only 2 redistricting plans (0.003%) and 468 redistricting plans (or 0.70%) have a Representativeness Index greater than NC2016 under the 2012 and 2016 votes, respectively.

Again, in stark contrast, the Judges districting plan has a lower Gerrymandering index than 25,074 redistricting plans (37.68%) and 36,628 redistricting plans (55.04%) under the 2012 and 2016 votes, respectively. Similarly, the Judges districting plan has a lower Representativeness index 17,558 redistricting plans (26.39%) and 22,547 redistricting plans (33.88%) under the 2012 and 2016 votes, respectively.

All four metrics over both election years indicate that the Judges plan is very typical. The Judges plan shows low partisan bias, has a reasonable efficiency gap, has a comparatively low level of gerrymandering, and is reasonably representative. In contrast, the NC2012 and NC2016 plans have strong partisan bias toward the republicans, have a large efficiency gap, are unrepresentative, and are highly gerrymandered.

While some might like the persuasive simplicity of a single number and arguments which state that the chosen metric is an outlier, we prefer the more descriptive images in the previous section along with the outlier status of the packing and cracking they identify. We find these more explanatory and prefer leading with understanding rather

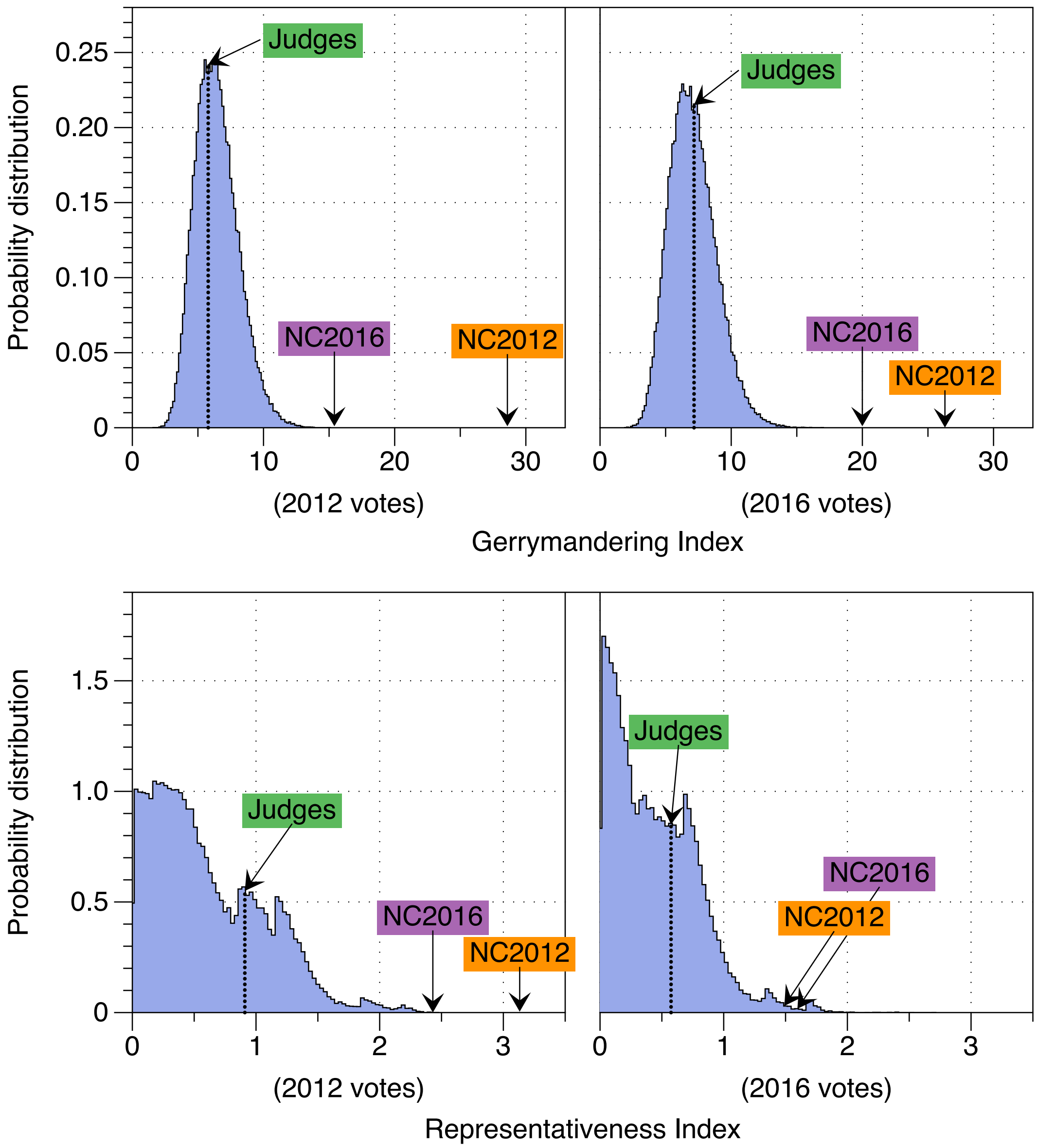

Figure 5: Gerrymandering (top) and Representativeness (bottom) Indices for the three districts of interest, and the ensemble of plans. Figures are based on the voting data from 2012 (left) and 2016 (right).

than generic, one–size–fits–all statistical tests. We find statistical characterizations most compelling when directly supporting the narrative already suggested by the visualizations.

Comparing to locally similar redistricting plans

If relatively small changes in a redistricting dramatically change the partisan vote balance in each district then it raises questions how representative the results generated by the redistricting are, and suggests the redistricting was selected or engineered.[1]

We explore the degree to which the three most Democratic districts of NC2012 and NC2016 are locally typical by examining local ensembles that preserve the core districts on each of the three plans. When generating these new ensembles, we require that each district in the ensemble directly correspond to a district within a plan of interest and that it's population deviates no more than a certain percentage away from the original district. We generate six new ensembles, two for each plan, that deviate by no more than 10% and 30% of the population on each district. Details on the generative procedure and on the ensembles are presented in Section S4 of the supplemental material.

We display our results in Figure 6 under the 2012 congressional votes. We find that in the NC2012 and NC2016 districts, the top three democratic districts of the two local ensembles precipitously lose Democratic voters and this effect is continues as the districts are allowed to deviate more. In contrast, in the Judges plan, there is no appreciable trend of how the structure of the local districts deviate, save some spreading of the marginal distributions. In other words, small changes to district boundaries make the NC2012 and NC2016 redistricting plans less partisan but do not change the characteristics of the Judges redistricting. This suggests that NC2012 and NC2016 plans, in contrast to the Judge's plan, were precisely engineered to achieve a partisan goal.

[1]As the initial working paper reporting these results of this paper [3] was being completed an the work in [7] appeared which provides an interesting set of ideas to assess if samples being drawn are typical or outliers exactly in our context. We hope to explore these ideas in the future.

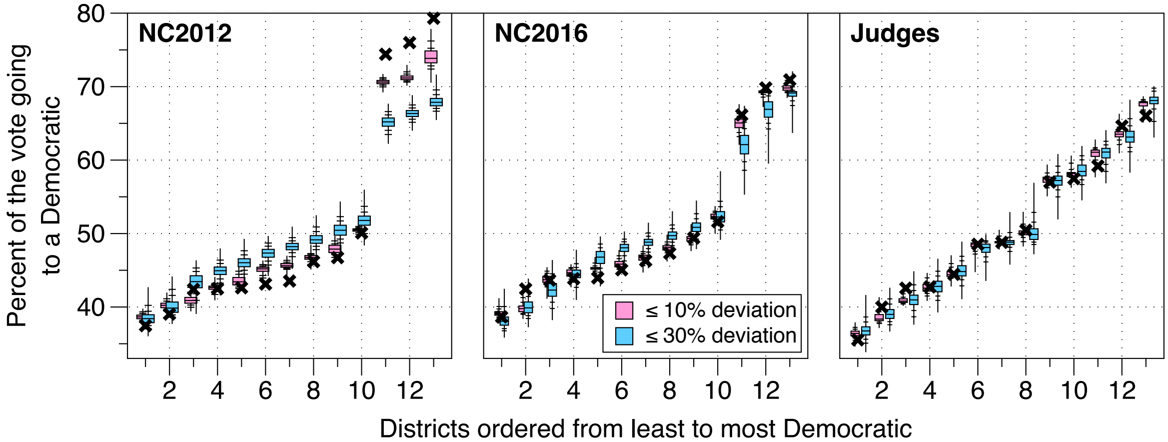

Figure 6: We display the ranked marginal distributions based on random samples drawn from nearby the three redistricting plans of interest: NC2012 (left), NC2016 (center), and Judges (right). Black X's mark the plans ranked district margins, whereas as light purple and turquoise box plots mark the 10% and 30% deviations, respectively. All plots use the 2012 congressional votes.

Generating redistricting plans

Based on the criteria for a reasonable redistricting outlined above, we define a family of probability distributions on the space of redistricting plans by first defining a score function on any given redistricting. This score function will return lower score to those redistricting which better adhere to the outlined design criteria. We will then use the score function to define a measure on the space of redistricting plans which is used in the simulated annealing process of the Markov Chain Monte Carlo algorithm.

Possible methods for generating the ensemble of redistricting plans

Before detailing the method we use we discuss a number of alternative methods and situate them relative to the one we have employed. Ideas to generate redistricting plans with computational algorithms have been being developed since the 1960's [26, 33, 13]. There are three main classifications of redistricting algorithms: constructive randomized algorithms [10, 5, 6], moving boundary MCMC algorithms [23, 24, 2, 37, 12], and optimization algorithms [25, 22]. Constructive randomized algorithms begin each new redistricting with an initial random seed and either grow a fixed number of districts or combine small districts until the desired number of districts is achieved. Moving boundary MCMC algorithms find new redistricting plans by altering existing district boundaries to sample from a specified distribution on redistricting plans. Optimization algorithms seek optimal solutions on a highly non-convex space; most recently a genetic algorithm was proposed for this task in [22]

In [12], the authors demonstrate that MCMC algorithms are able to recover how the sampling procedure reweights the space of plan, since the distribution is known *a priori*; in contrast, it is not known how constructive and optimization routines favor some redistricting plans over others. MCMC algorithms can theoretically sample the space of redistricting plans with the correct probability distribution,[2] whereas constructive algorithms may construct many similar redistricting plans while not generating other types of equally valid redistricting plans, leading to a skewed distribution. In particular, one is unsure what distribution on redistricting the algorithm is producing.

One advantage of the MCMC approach over all of the other options is that it samples from a explicitly specified and constructed probability distribution on redistricting plans. Hence, the biases and preferences are explicit and open to critique. We define the this family of target distributions first by developing a score function that evaluates the overall "goodness" of a districting plan, and then use the score function to define a probability distribution. The family of distributions focuses around compliant redistricting plans. While we know our basic MCMC algorithm will sample from the desired probability distribution on restricting plans if run long enough, MCMC has been seen to not correctly sampling the known distribution if the chains do not run for a sufficient long amount of time [1]. There are a number of empirical tests that may be used to check if a system has mixed

[2]One can rigorously prove that the Markov Chain given by this algorithm converges to the desired distribution if run long enough. One only needs to establish that the Markov Chain transition matrix is irreducible and aperiodic. Since one can evolve from any connection redistricting to another through steps of the chain, it is irreducible. Aperiodicity follows as there exist redistricting plans which are connect to itself through a loop consisting of two steps and a loop consisting of three steps. Since 2 and 3 are prime and hence have greatest common divisor 1, the chain is aperiodic. See the Perron-Frobenius Theorem for more details.

(e.g. [15, 16, 17, 31]).

We found that a straight froward MCMC algorithm on our chain was not sufficiently exploring the redistricting possibilities in the time we could a lot. This could likely be ameliorated by improving the proposal which we are exploring. While waiting for these sampling improvements to bare fuit, we have taken a more pragmatic approach. We have augmented the basic MCMC algorithm with a periodic heating and cooling schedule. This combination is typically referred to as simulated annealing. Viewed at periodic time intervals, it again produces a Markov Chain which can be shown to converge to a unique statistical steady state. The only caveat is that it not necessarily precisely the measure used in the MCMC step. However, it shares many of its characteristics. The areas of high probability are the same as they are concentrated at the minimum of the score function; the relative waiting of these low score regions might well be different. We will have applied the convergence algoritms mentioned above to the simulated annealing chain and have shown empirically it mixes well, converges according to a Gelman-Rubin test and variants, and produces robust conclusions (see sections S2.1 and S3 in the supplemental material).

Defining the score function

To define the score function, we introduce several mathematical formalisms, the first of which is represents the state of North Carolina as a graph G with edges E and vertices V. Each vertex represents a Voting Tabulation District (VTD); an edge between two vertices exists if the two VTDs share boundaries with non-zero length. In general VTDs may be split into census blocks, however (i) the utilized redistricting criteria requires this splitting to minimized and (ii) we demonstrate that splitting VTDs to achieve zero population deviation in a district has nearly no effect on our results (see the supplemental material section S2.2).

Defining the graph this way allows us to formally define a redistricting plan: Assuming each VTD belongs to a single district, a redistricting plan is defined as a function from the vertices, V, to one of the possible districts, which are represented by sequential integers – there are thirteen congressional districts in North Carolina, so we define a redistricting plan as a function $\xi : V \to \{1, 2, \ldots, 13\}$. The redistricting plan function ξ is interpreted as follows: If a VTD is represented by a vertex $v \in V$, then $\xi(v) = i$ means that the VTD in question belongs to district i; similarly, for any $i \in \{1, 2, \ldots, 13\}$ and plan ξ, the i-th district, denoted $D_i(\xi)$, is given by the set $\{v \in V : \xi(v) = i\}$. We restrict the space of considered redistricting plans ξ such that each district $D_i(\xi)$ is a single connected component; this restriction, along with our edge criteria, ensures that districts are always contiguous. We denote the collection of all redistricting plans with connected districts by $\mathcal{R}$.

A plan ξ is rated with our score function denoted J. J maps each redistricting $\xi \in \mathcal{R}$ to a nonnegative number. Lower scores signify redistricting plans that more closely adhere to the criteria of HB92. To define the score function J, we employ several sub-functions that measure how well a given redistricting satisfies the individual principles outlined in HB92. We will denote these sub-functions as J_p, J_I, J_c, and J_m: the *population score* $J_p(\xi)$ measures how well the redistricting ξ partitions the population of North Carolina into 13 equal population groups; the *isoparametric score* $J_I(\xi)$ measures how compact the districts are; the *county score* $J_c(\xi)$ measures the number of counties split between multiple districts; lastly, the *minority score* $J_m(\xi)$ measures the extent to which a districting plan adheres to the VRA. Once the sub-functions are specified, our score function J is defined as a weighted sum of J_p, J_I, J_c, and J_m; since all of the sub-score functions are not on the same scale, we use a weighted combination to balance the influence of each criteria. Specifically, we define

$$J(\xi) = w_p J_p(\xi) + w_I J_I(\xi) + w_c J_c(\xi) + w_m J_m(\xi), \tag{1}$$

where w_p, w_I, w_c, and w_m are a collection of positive weights.

To describe the individual sub-functions, data is associated to our graph G which allows the recovery of relevant features on each VTD. The positive functions $\text{pop}(v)$, $\text{area}(v)$, and $\text{AA}(v)$, defined on a vertex $v \in V$, represent (respectively) the total population, geographic area, and minority population of the VTD associated with v; the symbol AA represents the minority population because African-Americans are the only minority in North Carolina with a large enough population to gain representation under the VRA. The functions $\text{pop}(v)$, $\text{area}(v)$, and $\text{AA}(v)$ are extended to a collection of vertices $B \subset V$ by

$$\text{pop}(B)=\sum_{v \in B} \text{pop}(v),\ \text{area}(B)=\sum_{v \in B} \text{area}(v),\ \text{AA}(B)=\sum_{v \in B} \text{AA}(v). \tag{2}$$

The boundary of a district $D_i(\xi)$, denoted $\partial D_i(\xi)$, is the subset of the edges E that connect vertices inside of $D_i(\xi)$ to vertices outside of $D_i(\xi)$. According to this definition, VTDs that border another state or the ocean will not have an edge that signifies this fact. To incorporate state boundary information, we add the vertex o to V, which represents the "outside" and connect it with an edge to each vertex representing a VTD which shares perimeter with the boundary of the state. We assume that any redistricting ξ always satisfies $\xi(v) = 0$ if and only if $v = o$; since ξ always satisfies $\xi(o) = 0$, and $o \notin D_i(\xi)$ for $i \geq 1$, it does not matter that we have not defined $\text{pop}(o)$, $\text{area}(o)$ or $\text{AA}(o)$, as o is never included in the districts.

Given an edge $e \in E$ which connects two vertices $v, \tilde{v} \in V$, $\text{boundary}(e)$ will represent the length of common border of the VTDs associated with the vertex v and $\tilde{v}$. As before, the definition is extended to the boundary of a set of edges $B \subset E$ by

$$\text{boundary}(B) = \sum_{e \in B} \text{boundary}(e)\,. \tag{3}$$

With these preliminaries out of the way, we define the score sub-functions used to assess the goodness of a redistricting.

The population score function

The population score, which measures how evenly populated the districts are, is defined by

$$J_p(\xi) = \sqrt{\sum_{i=1}^{13} \Big(\frac{\text{pop}(D_i(\xi))}{\text{pop}_{\text{Ideal}}} - 1\Big)^2}, \quad \text{pop}_{\text{Ideal}} = \frac{N_{pop}}{13},$$

where N_{pop} is the total population of North Carolina, $\text{pop}(D_i(\xi))$ is the population of the district $D_i(\xi)$ as defined in equation (2), and $\text{pop}_{\text{Ideal}}$ is the population that each district should have according to the 'one person one vote' standard: $\text{pop}_{\text{Ideal}}$ is equal to one-thirteenth of the total state population.

The Isoparametric score function

The Isoperimetric score, which measures the overall compactness of a redistricting, is defined by

$$J_I(\xi) = \sum_{i=1}^{13} \frac{\left[\text{boundary}(\partial D_i(\xi))\right]^2}{\text{area}(D_i(\xi))}.$$

It is is the ratio of the square perimeter to the total area of each district. The Isoparametric score is minimized for a circle, which is the most compact shape.

This compactness measure is one of two measures often used in the legal literature, where its reciprocal is proportional to as *the Polsby-Popper score* or *the perimeter score* [29, 30]. The second measure, usually referred to as *the dispersion score*, is more sensitive to overly elongated districts, although the perimeter score also penalizes them. We select the *the perimeter score* because it penalizes undulating boundaries, whereas the the dispersion score does not. There are over 20 measures to evaluate compactness (see, for example, [27]). We chose the *the Polsby-Popper score* both because of its historical significance and because it is consistent with the utilized compactness criteria.

The county score function

The county score function measures how many, and to what degree, counties are split between districts. If two VTDs belong to different districts but the same county, the county is called a split county. The score function is defined as

$$\begin{aligned} J_c(\xi) = &\{\#\text{ counties split between} \geq 2 \text{ districts}\} \cdot W_2(\xi) \\ &+ M_C \cdot \{\#\text{ counties split between} \geq 3 \text{ districts}\} \cdot W_3(\xi), \end{aligned}$$

where M_C is a large constant that heavily penalizes three county splits, and W_2 and W_3 are weight factors that smooth abrupt transitions between split/non-split counties and are defined by

$$W_2(\xi) = \sum_{\substack{\text{counties}\\ \text{split between}\\ \geq 2 \text{ districts}}} \sqrt{1 - F_1}$$

$$W_3(\xi) = \sum_{\substack{\text{counties}\\ \text{split between}\\ \geq 3 \text{ districts}}} \sqrt{1 - F_2},$$

where F_i is the fraction of VTDs within a county that are in the i districts that have the highest VTD fractional overlap with the county.

The Voting Rights Act or minority score function

The VRA mandates that minorities have the ability to elect a number of representatives; the number is determined by the fraction of the population comprised of the minority. In North Carolina, the only minority large enough to warrant consideration under the VRA is the African American population. African-American voters make up approximately 20% of the eligible voters in North Carolina; since 0.2 is between $\frac{2}{13}$ and $\frac{3}{13}$, the current judicial interpretation of the VRA stipulates that at least two districts should have enough African-American voters so that this demographic may elect a candidate of their choice.

African-American voters should not be overly represented in a district either. The NC2012 districting plan was ruled unconstitutional because two districts, each containing over 50% African-Americans, were ruled to have been packed too heavily with African-Americans, diluting their influence in other districts. The NC2016 districting was accepted based on racial considerations of the VRA and contained districts that held 44.48% African-Americans, and 36.20% African-Americans. The amount of deviation constitutionally allowed from these numbers is unclear.

Based on these considerations, we chose a VRA score function which awards lower scores to redistricting plans which had one district close to 44.48% African-Americans and a second district close to 36.20% African-Americans. We write the score function as

$$J_m(\xi) = \sqrt{H(44.48\% - m_1)} + \sqrt{H(36.20\% - m_2)}, \tag{4}$$

where m_1 and m_2 represent the percentage of African-Americans in the districts with the highest and second highest percent of African-Americans, respectively. H is the function defined by $H(x) = 0$ for $x \leq 0$ and $H(x) = x$ for $x > 0$. The use of the square root function steepens the score function as districts near the desired population percentage; when used in conjunction with the Monte Carlo algorithm presented below, the steepening will have the effect that districts close to the set desired minority populations are more likely to move toward achieving these populations. Notice that whenever $m_1 \geq 44.84\%$ and $m_2 \geq 36.20\%$ we have that $J_m = 0$; this feature allows the possibility for high minority populations, but allows such instances to arise naturally and does not target such an outcome.

A Family of Target Probability Distributions on Redistricting Plans

We now use the score function $J(\xi)$ to assign a probability to each redistricting $\xi \in \mathcal{R}$ that makes redistricting plans with lower scores more likely. Fixing a $\beta > 0$, we define the probability of ξ, denoted by $\mathcal{P}_\beta(\xi)$, by

$$\mathcal{P}_\beta(\xi) = \frac{e^{-\beta J(\xi)}}{\mathcal{Z}_\beta} \tag{5}$$

where $\mathcal{Z}_\beta$ is the normalization constant defined so that $\mathcal{P}_\beta(\mathcal{R}) = 1$. Specifically,

$$\mathcal{Z}_\beta = \sum_{\xi \in \mathcal{R}} e^{-\beta J(\xi)} .$$

The positive constant β is often called the "inverse temperature" in analogy with statistical mechanics and gas dynamics. When β is very small (the high temperature regime), different elements of $\mathcal{R}$ have close to equal probability. As β increases ("the temperature decreases"), the measure concentrates the probability around the redistricting plans $\xi \in \mathcal{R}$ which minimize $J(\xi)$. This idea has been previously used in [24] and [12].

Sampling the distribution

We use a standard Metropolis-Hastings algorithm. We define the proposal chain Q used for proposing new redistricting plans in the following way:

1. Uniformly pick a conflicted edge at random. An edge, $e = (u, v)$ is a conflicted edge if $\xi(u) \neq \xi(v)$, $\xi(u) \neq 0$, $\xi(v) \neq 0$.

2. For the chosen edge $e = (u, v)$, with probability $\frac{1}{2}$, either:

$$\xi'(w) = \begin{cases} \xi(w) & w \neq u \\ \xi(v) & u \end{cases} \qquad \text{or} \qquad \xi'(w) = \begin{cases} \xi(w) & w \neq v \\ \xi(u) & v \end{cases}$$

Let conflicted(ξ) be the number of conflicted edges for redistricting ξ. Let crossEdges(ξ, ξ') be the number of conflicted edges in redistricting ξ that may be chosen such that when one of the edge vertices is reassigned to the other edge vertex's district, the resulting districting plan is ξ'. We propose state ξ' from state ξ with probability $Q(\xi, \xi') = \frac{1}{2}\frac{\text{crossEdges}(\xi,\xi')}{\text{conflicted}(\xi)}$, which is the probability of proposing state ξ' from state ξ. The acceptance probability is given by:

$$p = \min\left(1, \frac{Q(\xi', \xi)}{Q(\xi, \xi')} e^{-\beta(J(\xi') - J(\xi))}\right)$$

If a redistricting ξ' has districts that are not connected, then we refuse the step, which is equivalent to setting $J(\xi') = \infty$.

We utilize simulated annealing to sample the space: β starts starts and remains at zero until $10,000$ steps are accepted, which allows the MCMC algorithm to freely explore the space of connected redistricting plans; next β grows linearly to one over the course of $60,000$ accepted steps, growing only on an accepted step, which allows the algorithm to search for a redistricting with a low score without getting caught in local minimum; finally, β is fixed at one for $40,000$ steps, accepted or not, so that the algorithm locally samples the measure $\mathcal{P}_\beta$. This process is repeated for each sampled redistricting. We run 64 chains, each starting from a unique initial redistricting plan. Three of these chains begin from the three plans of interest; the remaining chains start from random plans generated by a constructive algorithm that is described in [5]. The different chains are aggregated for the complete ensemble and used to test for convergence (see Section S2.1 of the Supplemental Material). For other Monte Carlo algorithms see [23, 24, 2, 37, 12]; for other redistricting algorithms see [10, 5, 6, 25, 22].

Determining the weight parameters

As we have mentioned above, we have four independent weights (w_p, w_I, w_c, w_m) used in balancing the effect of the different scores in the total score $J(\xi)$. In addition to these parameters, we also have the low and high temperatures corresponding respectively to the maximum and minimum β values used in simulated annealing. We have set the minimum value of β to be zero which corresponds to infinite temperature. In this regime, no district is favored over any other, which allows the redistricting plan freedom to explore the space of possible redistricting plans. To consider a high β value, we note that β multiplies the four weights in the probability distribution function: This means that one of the five remaining degrees (the four weights and high β) is redundant and can be set arbitrarily. We therefore chose to fix the low temperature (high value of β) to be one.

To select appropriate weights, we employ the following tuning method:

1. Set all weights to zero.

2. Find the smallest w_p such that a fraction of the results are within a desired threshold (for the current work we ensured that at least 10% of the redistricting plans were below 1.5% population deviation, however we typically did much better than this).

3. Using the w_p from the previous step, find the smallest w_I such that a fraction of the redistricting plans have all districts below a given isoparametric ratio (we ensured that at least 10% of the results were below this threshold; we chose a threshold of 80, as we have done above).

4. If above criteria for population is no longer met, repeat steps 2 through 4 until both population and compactness conditions are satisfied.

5. Using the w_p and w_I from the previous steps, find the smallest w_m such that at least 10% of all redistricting plans have at least one district with more than 40% African-Americans and a second district has at least 33.5% African-Americans.

6. If the thresholds for population were overwhelmed by increasing w_m, repeat steps 2 through 6. If the thresholds for compactness were overwhelmed, repeat steps 3 through 6.

7. Using the w_p, w_I, and w_m from the previous steps, find the smallest w_c such that we nearly always only have two county splits.

8. If the thresholds for population are no longer satisfied, repeat steps 2 through 8. If the criteria for the compactness is no longer met, repeat steps 3 through 8. If the criteria for the minority populations is not satisfied, repeat steps 5 through 8. Otherwise, finish with a good set of parameters.

With this process, we settle on parameters $w_p = 4500$, $w_I = 5.0$, $w_c = 0.4$, and $w_m = 800$ and have used these parameters for all of the results presented in the main text. We remark that this choice of parameters allows us to sample the space more quickly. Both the primary and local redistricting plans are available for download[3]. Other choices in parameters lead to similar results as is demonstrated Section S3.3 of the Supplemental Material.

Thresholding the sampled redistricting plans

It is possible for the simulated annealing algorithm to draw a redistricting plan that is far from the compliance criteria. For example, under simulated annealing, it is possible to get stuck in a local minimum that has a high score. Furthermore, the individual score functions take an aggregate score over all districts where we may wish to only accept a districting plan that is below some threshold over all districts.

In ensuring a high degree of compliance with HB92, we only use samples which pass an additional set of thresholds, one for each of the selection criteria. This additional layer of rejection sampling was also used in reference [12], though the authors of this work chose to reweigh the samples to produce the uniform distribution over the set redistricting plans that satisfy the thresholds. We prefer to continue to bias our sampling according to the score function so better redistricting plans are given higher weights; we note that the idea of preferring some plans to others is consistent with the provisions HB92 (for explicit criteria see Section S1.2 of the Supplemental Material).

From our experience from the Beyond Gerrymandering project, redistricting plans which use VTDs as their building blocks and have low population deviation can readily be driven

[3]`https://git.math.duke.edu/gitlab/gjh/nccongressionalensembles.git`; source code is also available

to negligible population deviation by breaking the VTDs into census tracts and performing minimal alterations to the overall redistricting plan; we also demonstrate in Section S2.2 that splitting VTDs to achieve zero population deviation has a negligible effect on our results. We thus only accept redistricting plans that have no districts above 1.5% population deviation. Many of our samples have deviations considerably below this value. We require this threshold of every district in the redistricting.

We have found that districts with isoparametric scores under 80 often look pleasing to the eye; furthermore, the enacted plan in 2016 has a district with an isoparametric score of just over 80. Thus, we choose to accept a redistricting only if each district in the plan has an isoparametric ratio less than 80. The Judges redistricting plan would be accepted under this threshold as its least compact district has an isoparametric score of 53.5. The NC2012 would not be accepted with this thresholding as the least compact district in this plan has an isoparametric score of 434.65. We note that only four of the thirteen districts for the NC2012 plan meet our isoparametric score threshold. Although we examine our principle results over a space of compact redistricting plans, we also demonstrate that our results are insensitive to lifting this restriction in Section S3.1 of the Supplemental Material.

Although redistricting plans that split a single county in three are infrequent, they do occur among our samples. Since these are undesirable, we only accept plans for which no counties are split across three or more districts. In order to satisfy population requirements, some counties must be split into two districts; an example making this clear is that Wake and Mecklenburg Counties each contain a population larger than a single Congressional district's ideal population. We do not explicitly threshold based on number of split counties, though redistricting plans with more split counties have a higher scores, and hence are less favored. We remark that our plans nearly always split fewer counties than the NC2012 redistricting plan, and that the NC2012 plan was never critiqued or challenged based on the number of county splits.

To build a threshold based on minority requirements of the VRA, we note that the NC2016 redistricting was deemed by the courts to satisfy the VRA. The districts in this plan with the two highest proportion of African-Americans to total population are composed of 44.5% and 36.2% African-Americans. With this in mind, we only accept redistricting plans if the districts with the two highest percentages of African-American population have at least 40% and 33.5% African-American voters, respectively.

Thresholding in this way sub-samples to roughly 14% of the samples initially produced by our MCMC runs. Although this leads to many unused samples, it ensures that all of the utilized redistricting plans meet certain minimal standards. This better adheres to the spirit of HB92. The reported 66,544 samples used in our study refer to those left after thresholding. The full data set of samples was in excess of 480,000. That being said, we show in Section S3.1 of the Supplemental material that results without compactness thresholding were very close to those with thresholding.

Details of the Indices

Details on symmetry summary statistics

Partisan bias is defined to be the symmetric difference between the expected number of seats won, as a function of the votes cast. If we define $E(s|v)$ as the expected number of seats, s, given a statewide Democratic vote fraction, v, then the partisan bias is defined to be (for a 13 district state) $(E(s|0.5 + v_s) - E(s|0.5 - v_s))$, where v_s is a uniform shift in the overall votes (e.g. [14]). Often the partisan bias is scaled by the number of districts to show the difference in the fraction of seats, however we have elected to not use this scaling. Historically vote shifts are computed by uniformly shifting the vote fractions in the marginal distributions, and we adopt this convention in the current work; we have set $v_s = 0.05$ (5%) in all of the presented results.

The Efficiency Gap is an index that was used in the decision Whitford Op. and Order, Dkt. 166, Nov. 21, 2016 (see also [32]). It quantifies the difference of how many "wasted votes" each party cast; a larger number means that one party wasted more votes than another. More precisely, the Efficiency Gap measures the difference of the relative efficiency for the Democrats and Republicans. The efficiency for each party is the sum of the fraction of votes in districts the party loses plus the sum over the percentage points above 50% in the districts won. The relative efficiency is the efficiency of a given party divided by the sum of percentage of votes obtained by the party in each district. The efficiency gap is the difference between the relative efficiencies of the two parties.[4]

Details of Gerrymandering Index

To compute the Gerrymandering Index, we begin by extracting the mean percentage of Democratic votes in each of the thirteen districts when the districts are ordered from most to least Republican (see Figure 3). For any given redistricting plan, we take the Democratic votes for each district when the districts are again ordered from most to least Republican and consider the differences between the mean and the observed democratic percentage. This gives us a set of thirteen numbers on which we consider a two-norm.

The Gerrymandering Index is smallest when all of the ordered Democratic vote percentages are precisely the mean values. However, this is likely not possible as the percentages in the different districts are highly correlated. To understand the range of possible values, we have plotted the probability distribution function of the Gerrymandering Index of our ensemble of randomly generated reasonable redistricting plans in the main text (see Figure 5). This gives a context in which to interpret any one score.

To provide an example, we note that for the 2012 votes, the mean percentages for the collection of redistricting plans we generate is

$$\vec{m} = (39, 41, 42, 44, 46, 48, 50, 52, 54, 56, 59, 62, 67)\,.$$

If a given redistricting is associated with the sorted winning Democratic percentages

$$\vec{p}_i(36, 38, 39, 40, 41, 42, 43, 44, 49, 52, 64, 66, 70)\,.$$

then the Gerrymandering Index for the redistricting is of

$$||\vec{p}_i - \vec{m}||_2 = 18.1.$$

[4]The original used actual votes, but when the population of each district is equal then the two measures are exactly equivalent. If the actual votes is close to equal then they are almost the same.

Details of Representativeness Index

To calculate the Representativeness Index, we first construct a modified histogram of election results that captures how close an election was to flipping a congressional seat. To do this for a given redistricting plan, we examine the least Republican district in which a Republican won, and the least Democratic district in which a Democrat won. We then linearly interpolate between these districts and find where the interpolated line intersects with the 50% line. For example, in the 2012 election (NC2012 map with 2012 votes), the 9th most Republican district elected a Republican with 53.3% of the vote, and the fourth most Democratic district won their district with 50.1% of the vote. We would then calculate where these two vote counts cross the 50% line, which will be

$$\frac{50 - (100 - 50.1)}{53.3 - (100 - 50.1)} \approx 0.03, \tag{6}$$

and add this to the number of Democratic seats won to arrive at the continuous value of 4.03. This index allows us to construct a continuous variable that contains information on the number of Democrats and Republicans elected. It also contains information on the relative safety between the most competitive seat for each party.

Fractional parts close to zero suggest that the most competitive Democratic race is less likely to go Democratic than the most competitive Republican race is to go Republican. On the other hand, fractional parts close to one suggest that the most competitive Republican race is less likely to go Republican than the most competitive Democratic race is to go Democratic. Instead of simply creating a histogram of the number of seats won by the Democrats, we construct a histogram of the continuous interpolated number of elected Democrats in Figure 7. We define the representativeness as the distance from the interpolated value to the mean value of this histogram (shown in the dashed line). These are the values we report in Figure 5. For the 2012 congressional voting data, we find that the mean interpolated Democratic seats won is 7.16, and the Judges plan yields a value of 6.28, giving a Representativeness Index of $|7.16 - 6.25| = 0.91$. The NC2012 and NC2016 plans both have Representativeness Indices greater than two.

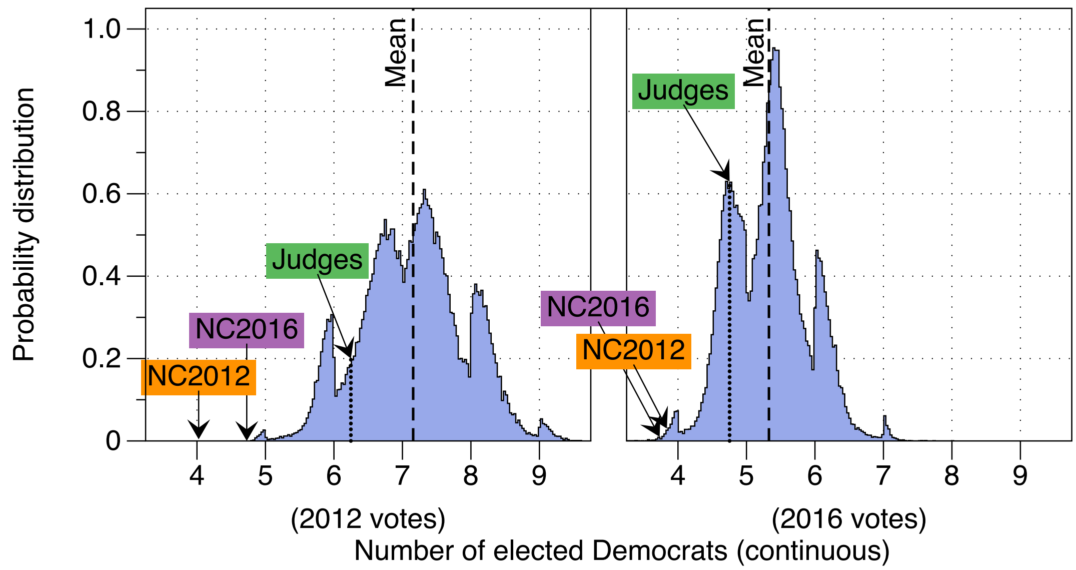

Figure 7: For the 2012 votes (left) and the 2016 votes (right), we plot the interpolated winning margins, which give the number of seats won by the Democrats in finer detail. We determine the mean of this new histogram and display it with the dashed line. The Representativeness Index is defined to be the distance from this mean value. The histograms presented in Figure 1 are overlaid on this plot for reference.

Discussion

We have found evidence that that the NC2012 and NC2016 are heavily gerrymandered: they employ packing and cracking to generate a dramatic jump of partisan vote counts in the ordered district structures. This jump is significantly larger than those found in the ensemble of redistricting plans which is a signature of gerrymandering. When summarizing these observations in single number metrics, we find that the NC2012 and NC2016 redistricting plans are often extreme outliers, suggesting that (i) these districts are heavily gerrymandered and (ii) do not represent the geo-political landscape expressed in a number of elections occurring between 2012 and 2016. Further analysis reveals the NC2012 and NC2016 plans are locally engineered for partisan benefit. All of these results are consistent with the findings of [22].

On the contrary, the districting plan produced by a bipartisan redistricting commission of retired judges from the Beyond Gerrymandering project produced results which are highly typical. The Judges plan is not gerrymandered, was typically representative of the people's will, and displayed consistent statistics with nearby redistricting plans.

The ideas presented in this report are generalizable at both the state and federal level. We note that each state may have different requirements when drawing district boundaries and so care must be taken when considering the criteria to be included in the generative procedure. We hope that the analysis in this report is utilized across different states and levels of government to test the viability of districting plans.

As stated in the introduction, the qualitative results and subsequent conclusions of this note agree fully with those in earlier iterations of this analysis reported in [3]. We have taken the opportunity of this report to improve upon the convergence validations used in generating our ensemble. We have also expanded our analysis to include more statewide elections both because we have grown to prefer statewide elections and generally because each election provides a different tool to probe the properties of the maps under discussion. Each election has different state wide fraction but also different spatial patterns of votes. We have also taken the opportunity to correct a small logical error in [3] which miscalculated the proposals transition probability. In practice this had little effect on the Markov chain as witnessed by the close agreement of the results of this note and [3].

The current notes methodology employs simulated annealing paired with Markov Chain Monte Carlo sampling techniques. We have demonstrated that our methodology converges to a robust result, however we have not provided guarantees that the target probability distribution agrees with the sampled distribution beyond valuing the same regions of high probability. Sampling directly using MCMC in the current context will only become possible under more sophisticated sampling techniques with fixed measure, and these techniques are the subject of our future research. Directly sampling using the MCMC described here has already been used successfully to analysis the North Carolina State Legislative districts and presented to the court in *Common Cause v. Lewis*.

References

[1] William T. Adler and Samuel S.-H. Wang. Response to cho and liu, "sampling from complicated and unknown distributions: Monte carlo and markov chain monte carlo

methods for redistricting". *Physica A*, 516, 2019.

[2] Sachet Bangia, Bridget Dou, Sophie Guo, and Christy Vaughn. Quantifying Gerrymandering Data+ project. https://services.math.duke.edu/projects/gerrymandering/, 2015.

[3] Sachet Bangia, Christy Vaughn Graves, Gregory Herschlag, Han Sung Kang, Justin Luo, Jonathan C. Mattingly, and Robert Ravier. Redistricting: Drawing the line. *arXiv*, 1704.03360, 2017.

[4] U.S. Census Borough. Shapefiles for congressional districts. http://www.census.gov/geo/maps-data/data/tiger-line.html, 2010.

[5] Jowei Chen and Jonathan Rodden. Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*, 8:239–269, 2013.

[6] Jowei Chen and Jonathan Rodden. Cutting through the thicket: Redistricting simulations and the detection of partisan gerrymanders. *Election Law Journal*, 14(4):331–345, 2015.

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

[8] Wendy K. Tam Cho. Measuring partisan fairness: How well does the efficiency gap guard against sophisticated as well as simple-minded modes of partisan discrimination? *University of Pennsylvania Law Review Online*, 166, 2017.

[9] Wendy K. Tam Cho and Yan Y. Liu. Toward a talismanic redistricting tool: A computational method for identifying extreme redistricting plans. *Election Law Journal*, 15(4):351–366, 2016.

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

[11] Harvard's Election Data Archive Dataverse. 2012 house elections votes. https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21919, 2012.

[12] Benjamin Fifield, Michael Higgins, Kosuke Imai, and Alexander Tarr. A new automated redistricting simulator using markov chain monte carlo. *Work. Pap., Princeton Univ., Princeton, NJ*, 2015.

[13] Burton C Gearhart and John M Liittschwager. Legislative districting by computer. *Systems Research and Behavioral Science*, 14(5):404–417, 1969.

[14] Andrew Gelman and Gary King. Enhancing democracy through legislative redistricting. *American Political Science Review*, 88(3), 1994.

[15] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[16] John Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4:641–649, 1992.

[17] Philip Heidelberger and Peter D Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[18] Gregory Herschlag and Jonathan Mattinlgly. Quantifying Gerrymandering blog. https://sites.duke.edu/quantifyinggerrymandering/, 2018 − Present.

[19] Gary King and Andrew Gelman. Systematic consequences of incumbency advantage in the u.s. house elections. *American Journal of Political Science*, 35(1):110–138, 1991.

[20] North Carolina Legislature. 2010 VTD geographic data. http://www.ncleg.net/representation/Content/BaseData/BD2016.aspx, 2010.

[21] North Carolina Legislature. 2010 VTD minority data. http://www.ncleg.net/representation/Content/BaseData/BD2011.aspx, 2010.

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

[23] William Macmillan. Redistricting in a gis environment: An optimisation algorithm using switching-points. *Journal of geographical systems*, 3(2):167–180, 2001.

[24] J. C. Mattingly and C. Vaughn. Redistricting and the Will of the People. *ArXiv e-prints*, October 2014.

[25] Anuj Mehrotra, Ellis L Johnson, and George L Nemhauser. An optimization based heuristic for political districting. *Management Science*, 44(8):1100–1114, 1998.

[26] Stuart S Nagel. Simplified bipartisan computer redistricting. *Stanford Law Review*, pages 863–899, 1965.

[27] Richard G. Niemi, Bernard Grofman, Carl Carlucci, and Thomas Hofeller. Measuring comactness and the role of a compactness standard in a test for partisan and racial gerrymandering. *Journal of Politics*, 52(4):1155–1181, 1990.

[28] North Carolina Board of Elections. Voting data. http://www.ncsbe.gov/ncsbe/Election-Results.

[29] Richard H. Pildes and Richard G. Niemi. Expressive harms, "bizarre districts," and voting rights: Evaluating election-district appearances after Shaw v. Reno. *Michigan Law Review*, 92(3):483–587, Dec 1993.

[30] American Bar Association Section of Administrative Law Practice, Regulatory, and J. Gerald Hebert. *The realist's guide to redistricting: avoiding the legal pitfalls*. American Bar Association, Section of Administrative Law and Regulatory Practice,, 2nd ed. edition, 2010.

[31] AL RAFTERY and S LEWIS. One long run with diagnostics: implementation strategies 474 for markov chain monte carlo [comment]. *Statistical Science, Hayward*, 7(4):475.

[32] Nicholas Stephanopoulos and Eric McGhee. Partisan gerrymandering and the efficiency gap. *82 University of Chicago Law Review*, 831, 2015.

[33] James D Thoreson and John M Liittschwager. Computers in behavioral science. legislative districting by computer simulation. *Systems Research and Behavioral Science*, 12(3):237–247, 1967.

[34] Tom Ross and POLIS center at Duke. Beyond gerrymandering project. See https://sites.duke.edu/polis/projects/beyond-gerrymandering/, 2016.

[35] Samuel S.-H. Wang. Three tests for practical evaluation of partisan gerrymandering. *Stanford Law Review*, 68, 2016.

[36] Gregory S. Warrington. Quantifying gerrymandering using the vote distribution. *arXiv:1705.09393*, 2017.

[37] Lucy Chenyun Wu, Jason Xiaotian Dou, Danny Sleator, Alan Frieze, and David Miller. Impartial redistricting: A markov chain approach. *arXiv:1510.03247v1*, 2015.

SUPPLEMENTARY MATERIAL

Title: Convergence studies, Robustness checks, threshold sensitivity, and further details on the sampling procedure and results

The supplemental material is organized as follows: (1) we begin by detailing the utilized data, including the HB92 criteria, voting data, the enacted NC2012 and NC2016 districting plans, and the Judges districting plan; (2) we then provide evidence that our MCMC algorithm has converged along with an error analysis for how our results might change due to splitting VTDs to achieve zero population; (3) we demonstrate that our main results are insensitive to changing the sampling procedure, the target probability distribution, and to the thresholds; finally, (4) we give some details on the properties of our sampled redistricting plans.

S1 Details on Data (Examined Districting Plans of Interest; Redistricting Criteria; Voting Data)

S1.1 Information on the three considered redistricting plans

Maps for the NC2012, NC2016, and Judges districting plans are shown in Figure S1. The Democratic vote fractions for each district are given in Table S1; the vote fractions are given for the U.S. congressional elections in 2012 and 2016. In the table, vote fractions are adjacent to numbers in parentheses that give the numerical label of the individual districts as identified in the maps in Figure S1. The last two columns contain the mean values of percentage of Democratic votes for the ensemble of redistricting plans.

The three most Democratic districts (labeled 1, 4 and 12 in both the NC2012 and NC2016 plans) have significantly more Democratic votes than the expected average for these districts. Districts 9 and 13 both show evidence of having less Democrats than one would expect from their rankings. These conclusions are consistent across the 2012 and 2016 votes.

The raw data used to produce Figure 1 of the main text is given in Table S2. It underscores how atypical the results produced by the NC2012 and NC2016 redistricting

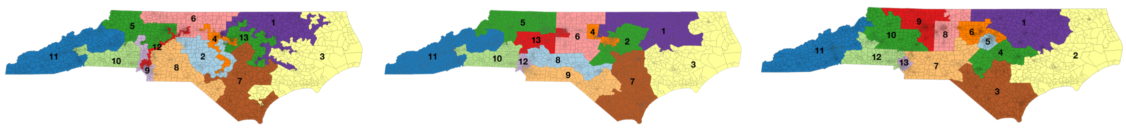

Figure S1: Map for NC2012 (left), NC2016 (middle) and Judges (right). Numbers correspond to labels in Table S1.

plans are. If one is ready to accept four seats for Democrats in the 2012 vote then one should equally accept nine. Similarly in the 2016 votes, if one accepts three seats for Democrats as a legitimate outcome then one should also be willing to accept seven seats. None of these results are particularly representative of the votes cast.

S1.2 Redistricting Criteria

The Judges districting plan was established in the Beyond Gerrymandering project[5] was a collaboration between UNC system President Emeritus and Davidson College President Emeritus Thomas W. Ross, Common Cause, and the POLIS center at the Sanford School at Duke University. The project's goal was to educate the public on how an independent, impartial redistricting process would work. The project formed an independent redistricting commission made up of ten retired jurists – five Democrat and five Republican.

The commission used strong and clear criteria to create a new North Carolina congressional map based on NC House Bill 92 (HB92) from the 2015 legislative season, which states

- §120-4.52(f): Districts must be contiguous; areas that meet only at points are not considered to be contiguous.
- §120-4.52(c): Districts should have close to equal populations, with deviations from the ideal population division within 0.1%.

[5]For more information see `https://sites.duke.edu/polis/projects/beyond-gerrymandering/`

	NC2012		NC2016		Judges		Mean	
Rank	2012	2016	2012	2016	2012	2016	2012	2016
1	37.5 (3)	34.2 (3)	38.7 (3)	32.8 (3)	35.5 (10)	28.9 (10)	38.5	32.3
2	39.0 (6)	34.6 (11)	42.5 (10)	35.8 (11)	40.0 (2)	33.6 (2)	40.6	34.3
3	42.4 (5)	36.2 (7)	43.7 (6)	36.8 (10)	42.6 (12)	36.3 (7)	42.2	36.1
4	42.5 (11)	36.6 (8)	43.9 (11)	39.0 (7)	42.7 (7)	37.6 (12)	44.1	38.6
5	42.6 (2)	37.4 (10)	44.0 (2)	40.7 (6)	44.5 (9)	40.0 (9)	46.1	40.4
6	43.1 (10)	38.9 (5)	45.1 (5)	41.2 (8)	48.5 (8)	41.9 (3)	48.1	42.3
7	43.5 (13)	40.8 (6)	46.3 (13)	41.6 (5)	48.8 (11)	42.7 (11)	50.3	44.7
8	46.2 (8)	41.2 (2)	47.3 (8)	41.8 (9)	50.5 (4)	45.7 (4)	52.3	47.7
9	46.7 (9)	44.0 (9)	49.4 (9)	43.3 (2)	57.0 (3)	48.1 (8)	54.4	50.9
10	50.1 (7)	45.8 (13)	51.6 (7)	43.9 (13)	57.5 (5)	55.9 (1)	56.4	54.0
11	74.4 (4)	71.5 (1)	66.1 (4)	66.6 (12)	59.2 (1)	59.7 (5)	58.7	56.7
12	76.0 (1)	73.0 (4)	69.8 (12)	68.2 (4)	64.6 (6)	63.3 (13)	62.4	60.4
13	79.3 (12)	75.3 (12)	70.9 (1)	70.3 (1)	66.0 (13)	65.3 (6)	66.6	64.6

Table S1: Percentage of Democratic votes in each district when districts are ranked from most Republican to most Democratic under the congressional vote counts. Numbers in parentheses give label of actual district using the numbering convention from maps in Figure S1.

- §120-4.52(g): Districts should be reasonably compact, with (1) the maximum length and width of any given district being as close to equal as possible and (2) the total perimeter of all districts being as small as possible.

- §120-4.52(e): Counties will be split infrequently and into as few districts as possible. The division of Voting Tabulation Districts (VTDs) will also be minimized.

- §120-4.52(d): Redistricting plans should comply with pre-existing federal and North Carolina state law, such as the Voting Rights Act (VRA) of 1965.

- §120-4.52(h): Districts shall not be drawn with the use of (1) political affiliations of

	# of Democratic Winners									
	1	2	3	4	5	6	7	8	9	10
2012 House Votes	0	0	0	182	5689	22258	27445	10322	648	0
2012 Presidential Votes	0	0	124	5808	25954	27837	6363	455	3	0
2016 House Votes	0	7	709	19973	36055	9429	370	1	0	0
2016 Senate Votes	0	2	448	17958	36310	11147	674	5	0	0

Table S2: Among the 66,544 random redistricting plans generated, we present the number of plans that produce the indicated number of Democratic seats under historic congressional vote counts.

registered voters, (2) previous election results, or (3) demographic information other than population. An exception may be made only when adhering to federal law (such as the Voting Rights Act (VRA)).

All federal rules related to the Voting Rights Act were followed but no political data, election results or incumbents addresses were considered when creating new districts. The commission met twice over the summer of 2016 to deliberate and draw maps. The maps resulting from this simulated redistricting commission were released in August 2016. The Judges agreed on a redistricting at the level of Voting Tabulation Districts (VTD). This coarser redistricting was refined at the level of census blocks to achieve districts with less than 0.1% population deviation. The original VTD based maps are used in our study and are denoted Judges; our results are insensitive to splitting VTDs to zero population (for a careful study, see below in Section S2.2).

Although unratified, HB92 was passed in the House, which provides some credence to using it as a guide to draw fair redistricting plans. In contrast, the criteria adopted in the drawing of the 2016 congressional districts contains a "Partisan Advantage" clause which seeks to predetermine the number of elected officials from each party[6]; with the balance of power pre-selected by this criteria, it would be impossible determine any sense of the statistical range in the balance of power. Because we do not wish to use partisan data to

[6]The 2016 Contingent Congressional Plan Committee Adopted Criteria may be found here: https://www.ncleg.gov/Files/GIS/ReferenceDocs/2016/CCP16_Adopted_Criteria.pdf

draw maps, we utilize the criteria from the non-partisan HB92.

S1.3 Data sources and extraction

VTD geographic data was taken from the NCGA website [20] and the United States Census Bureau website [4] , which provide for each VTD its area, population count of the 2010 census, the county in which the VTD lies, its shape and location. Perimeter lengths shared by VTDs were extracted in ArcMap from this data. Minority voting age population was found on the NCGA website using 2010 census data [21].

Data for the vote counts in each VTD in each election was taken from the NCSBE Public data [28]. For the 2016 election, VTD data was reported for precincts rather than VTDs, but rather for each precinct; 2447 of the precincts are VTDs, meaning that we have data for the majority of the 2692 VTDs. However 172 precincts contain multiple VTDs, 66 VTDs were reported with split data, and 7 VTDs were reported with complex relationships. To extrapolate VTD data on those contained in the 172 precincts containing multiple VTDs, we split the votes for a precinct among the VTDs it contained proportional to the population of each VTD. For the 66 split VTDs, VTDs were comprised of multiple precincts all contained with in a certain VTD, so we simply added up the votes among the precincts that were contained in each VTD – there was no extrapolation for these VTDs and these results are precise. For the VTDs with complex relationships, we divided up the votes using estimates based on the geography and population of the VTDs. We note that roughly 10% of the population lies in the VTDs with imperfect data, and that we do not expect significant deviation in our results based on the above approximations.

In using 2012 and 2016 data we have only used presidential election year data. Unfortunately, the 2014 U.S. congressional election in North Carolina contained an unopposed race which prevents the support for both parties being expressed in the VTDs contained in that district. In reference [2], the missing votes were replaced with votes from the Senate race. However, since we had two full elections, namely 2012 and 2016, which needed little to no alterations, we chose not to include the 2014 U.S. congressional votes in our study.

S2 Convergence and error analysis

S2.1 Convergence analysis

To test for convergence in our 64 independent chains we compare each chain to the aggregate ensemble after n steps. For the distribution of elected Democrats, we use the total variation as a comparison. For the ranked marginal ensembles found in Figure 3, we will use the Jeffery-divergence (defined below). We then employ the Gelman-Rubin index to demonstrate strong numerical evidence that the sampling has converged.

For each of the comparisons we consider two observables: the distribution of the number of elected officials, and the marginal distributions of the Democratic vote fractions ordered from least to most Democratic. For distance metrics, we use total variation on the histograms. On each ranked marginal distribution, we use the Jeffery divergence, or J-divergence, defined as

$$\langle p, q\rangle_J = \langle p, q\rangle_{KL} + \langle q, p\rangle_{KL},$$

where $\langle p, q\rangle_{KL}$ is the Kullback–Leibler divergence. In computing the J-divergence, we make a normal approximation for each ranked marginal distribution. We then take the maximum J-divergence across all ranks. For each metric, we examine two elections – the 2012 Presidential election and the 2016 United States Senate election – and then take the maximum distance over both elections.

We plot the maximal error in total variation for the distribution of Democrats elected Figure S2. The error is shown as a function of the total number of cycles in our simulated annealing procedure and decays at an exponential rate with exponent equal to to 2.2×10^{-4} times the number of simulated annealing cycles. After roughly 8000 simulated annealing cycles we have sampled the 66,544 compliant plans across the 64 independent chains. We find that at the end of the sampling the 29th chain under the 2016 US Senate votes yields the maximal total variation. In the ensemble, we find the worst case total variation is just over 10% which is qualitatively shown to be quite close to the aggregate ensemble.

To continue our convergence study we examine the J-divergence across each of the marginal distributions found by ordering the districts by their Democratic vote share. The maximal error over all chains, all ordered districts, and the two examined elections are

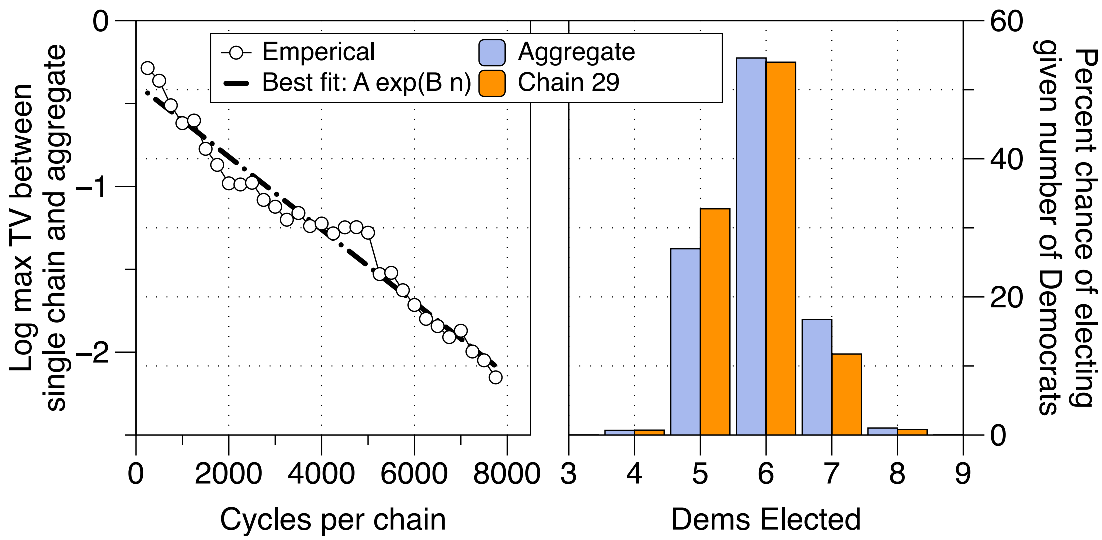

Figure S2: We display the worst case convergence between the total variation on the distribution of Democrats elected (left). At the final step, the 29th chain using the 2016 votes from the US Senate has the worst case error. We compare the histograms between the aggregate from all chains and the the 29th chain (right).

plotted as a function of the number of simulated annealing cycles for each chain (see left side of Figure S3). We find the error decays exponentially with exponent equal to to 4.7×10^{-4} times the number of simulated annealing cycles. At the end of the chains, we find that the maximal J-divergence to the aggregate ensemble occurs in the 13th ranked marginal distribution of the 54th chain under the 2012 Presidential votes (see right side of Figure S3).

In Figure S3 we reveal a plot (right) that is analogous to the box plot found in Figure 3 yet reveals more information. Instead of presenting the location of the quantiles, this plot reveals a vertical histogram over the Democratic vote fraction. Such plots reveal the structure of each marginal distribution and will be used throughout the supplemental section.

As a final test, we employ the Gelman-Rubin index [15]. The original test assumes each chain has an equal number of samples. Because we threshold our samples, this assumption does not hold. When employing the Gelman-Rubin index we therefore use the averaged

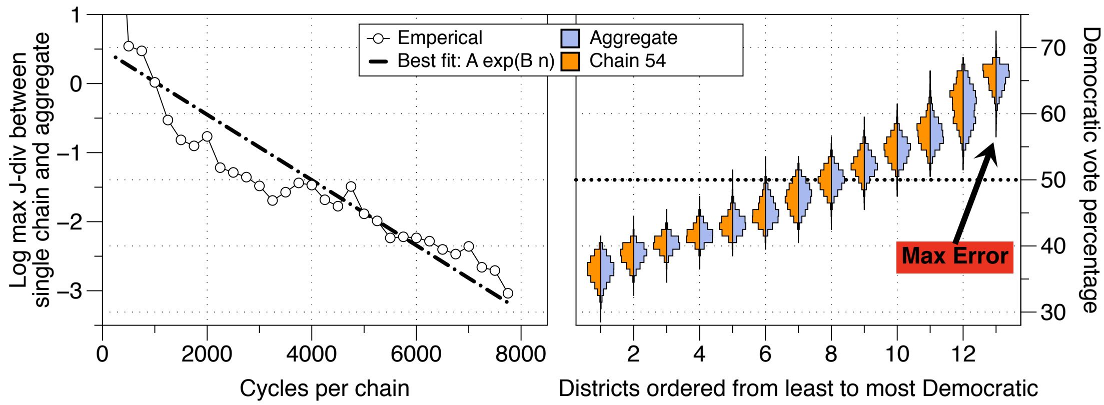

Figure S3: We display the worst case convergence between the J-divergence on the ranked marginal distributions of Democratic vote fractions (left). At the final step, the 54th chain using the 2012 votes from the US Senate has the worst case error in the most democratic marginal distribution. We compare the ranked marginal distributions between the aggregate from all chains and the the 54th chain. The maximal error occurs in the 13th ordered marginal distribution (right).

number of samples per chain within the calculation. We track the index over the two elections for the number of Democrats elected and for the worst case over all ranked marginal distributions (see Figure S4). When employing the Gelman-Rubin index, we must ensure that the ratio of the final variance estimate to the within sequence variance, denoted $\hat{R}$, is close to unity. At the termination of the chains, this ratio, at worst, is equal to $\hat{R} = 1.0036$ when examining the number of Democrats elected on the 2016 US Senate vote counts; similarly, the worst-case ratio is $\hat{R} = 1.0098$ for the twelfth least Republican district using the 2012 Presidential votes. In all cases this number has become empirically very close to one.

S2.2 Error analysis

All of our redistricting plans do not split any VTDs. In practice, VTDs must be split to achieve a population deviation below 0.1%. Splitting VTDs may change the vote count in each district. In this section, we demonstrate that splitting VTDs will have a negligible impact on the partisan vote fractions in each district. To do this, we derive error bounds

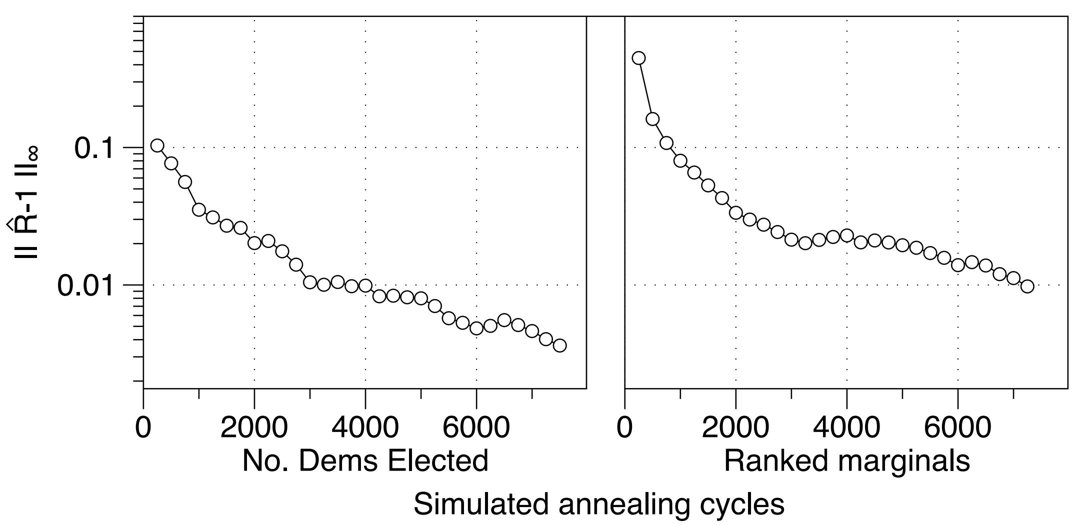

Figure S4: We display the worst case ratio of the current variance estimate to the within sequence variance after a given number of simulated annealing cycles (denoted $\hat{R}$. We subtract one from this value as it is expected to converge to 1. We display this value for the number of elected Democrats (left) and ranked marginal distributions for each district (right).

on the democratic vote fractions by assuming that every person moving into/out of the district will be to the benefit of one of the parties. Under this assumption, we find the minimal and maximal possible Democratic vote fractions of each district. The results are plotted in Figure S5. The small range of possible errors validate the idea that our results are robust when splitting VTDs to achieve a zero population deviation.

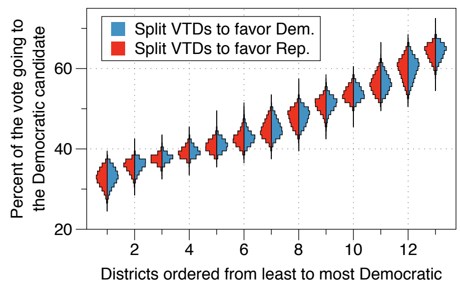

Figure S5: We display the marginal statistics for the ensemble when zeroing population helps the democrats (dark blue) and republicans (red), and compare the two sets of marginal distributions against each other when considering the 2016 US Senate vote counts.

S3 Robustness of results in thresholding, sampling, and distributions

S3.1 Robustness of results in thresholding

We have chosen to ensure that every district within a plan has an isoparametric ratio less than 80. However, there is no corresponding law to dictate a definite value in the choice of threshold. The NC2016 districts have a maximum isoparametric ratio of around 80, and the NC2012 districts have a maximum of over 400. HB92 mentions that when two districting plans are compared, the one with more compact districts should be preferred. The Judges redistricting has a district with maximum isoparametric ratio of around 54. To test the effect of setting different compactness thresholds, we alter the compactness thresholds to be 60, 100 and then remove a compactness threshold. Setting the alternate thresholds results in 5,014, 66,544, 120,886, and 141,232 redistricting plans, respectively. We compare the resulting distributions of the number of Democrats elected and one of the sets of ranked marginal distributions using the US Senate 2016 vote counts in Figure S6. We see almost no change in the results.

We note that having no threshold does not mean that we have arbitrarily large compactness values because of the cooling process in the simulated annealing algorithm penalizes large compactness scores. When considering no thresholding, we have an average maximum isoparametric ratio of around 83 and rarely see redistricting plans with maximal ratio larger than 140.

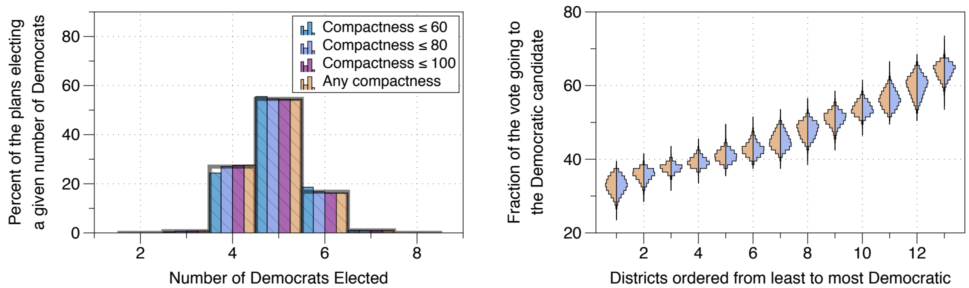

Figure S6: We display changes of the distribution of election results with changes to the compactness threshold (left). The histogram formed with a maximum of 80 for the isoparametric ratio overlays this image with the gray shaded histogram. We display changes to the histogram of the ranked marginal distributions when comparing a maximum of 80 in the isoparametric ratio to a distribution without any thresholding on compactness (right). Both results use the US Senate 2016 voting data.

S3.2 Robustness of results in sampling

It is possible that our method, and the above results, are sensitive to cooling times used under simulated annealing. To test this, we double the simulated annealing time: Instead of remaining hot ($\beta = 0$) for $10,000$ steps, cooling linearly for $60,000$ accepted steps, and remaining cold ($\beta = 1$) for $40,000$ steps, we instead remain hot for $20,000$ steps, cool linearly for $120,000$ accepted steps, and remain cold for $80,000$ steps. We generate a second ensemble of 8,689 compliant plans run on 32 independent chains, each starting from a unique initial condition. We compare the number of elected Democrats and the ranked marginal distributions using the 2016 US Senate voting data in Figure S7. We find nearly identical structure in both cases, which provides evidence that extending the hot, cooling, and sampling times will not impact our results.

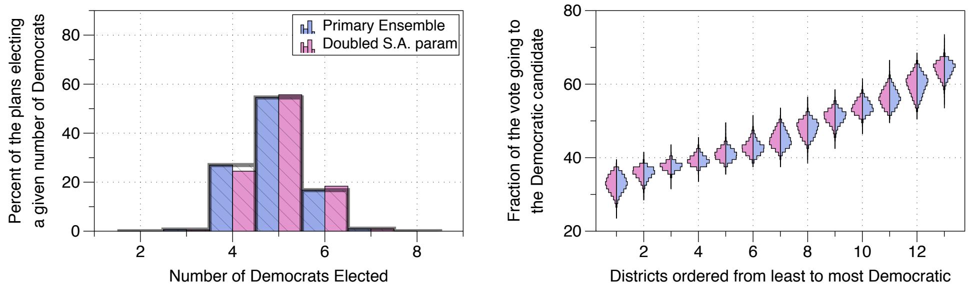

Figure S7: We display the probability distribution of elected Democrats with respect the original versus doubled simulated annealing parameters (left). The histogram presented in the primary text overlays this image with the gray shaded histogram for comparison. We display marginal distributions, rather than box plots, to further display the similarity of the sampled structure (right). The 2016 US Senate votes are used in generating both figures.

S3.3 Robustness of results in the target distribution

We fixed the weights in the target probability distribution function, yet there is not a single 'correct' target distribution. If the results are robust across a range of target distributions, we may conclude that the precise choice of distributions is not important. To investigate this possibility, we change the weights in the score function.

There are four linearly independent dimensions to explore along each weight of the score function: w_p, w_I, w_c and w_m. Exploring this space exhaustively would come at an large computational cost. To reduce this cost, we perform a sensitivity test about the reported weights used in the primary analysis: We first increase and decrease w_p, w_I, and w_m. For the fourth direction, we could simply increase or decrease w_c; we could also increase and decrease the maximum value of β instead, and choose this path instead. Because changing β is equivalent to changing all parameters, this forms a fourth linearly independent search direction, and provides us with information equivalent to changing w_c. This leads to eight different parameter sets, which still require a large number of runs.

To cut down on the computational cost, we take advantage of the result presented above, where we conclude that ignoring the compactness threshold has a minimal effect on our results. The compactness threshold is by far the most restrictive threshold, so we relax it to accept plans with all districts having an isoparametric score of less than 120; this allows us to sample more redistricting plans with fewer simulated annealing cycles.

We run each new set of weights on eight chains with unique initial conditions. In general, we find that increasing the weights tends to slow the computation since it takes more attempts to accept steps in the cooling process. We run each chain for the same amount of wall clock time. When increasing $w_p = 5000$, there are very few completed simulated annealing cycles and so we omit this run. When increasing $\beta = 1.2$, we find 5,660 compliant plans across the eight chains; when decreasing $\beta = 0.8$ we find 11,738 plans. When decreasing $w_m = 900$ and increasing $w_m = 700$, we find 31,566 and 22,210 plans, respectively. When decreasing $w_I = 6$ and decreasing $w_I = 4$, we find 36,172 and 3,194 plans, respectively. When decreasing $w_p = 4000$, we find 32,311 plans across the eight independent chains.

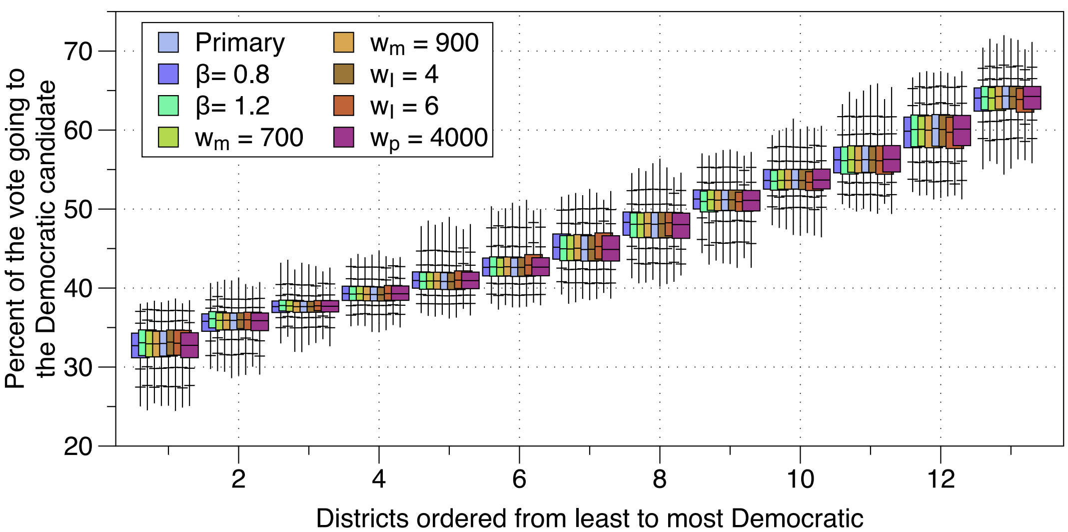

Figure S8: We display box-plots with minimum/maximum values displayed by the extent of the lines, along with 1% and 10% outliers displayed by the dashes outside the boxes. Election results change little with respect to changing the values of the weights. The results for the 2016 US Senate race is shown in this figure.

We plot the ranked marginal distributions in Figure S8 and find almost no change in the 1% and 10% outliers, nor in the region covering the central 50% domain of each marginal distribution. Similarly the medians are also highly insensitive to changes in the target distribution. We conclude that significant changes in the parameters will have little effect on the statistical results of the election data.

S3.3.1 Lowering county splits

In the above analysis compact districts are prioritized over those with low county splits. In this section we determine the sensitivity of our results when we prioritize keeping a low number of county splits. To make this examination, we double the county weight ($w_c = 0.8$) and reduce the compactness weight ($w_I = 4$). We then run 16 independent chains from random unique initial states and threshold compactness on an isoparametric score of 100. We obtain 86,667 redistricting plans when combining the chains. Under these new weights, we find that the median number of split counties drops from 34 to 29. We further threshold the new ensemble to only consider the 1,381 plans that split 22 or fewer counties. Despite the this large change in the specified sample space along with the drastic reduction in samples, the results are remarkably stable. We display the number of elected officials and the ranked marginal distributions of Democratic vote fractions under the 2016 US Senate votes in Figure S9.

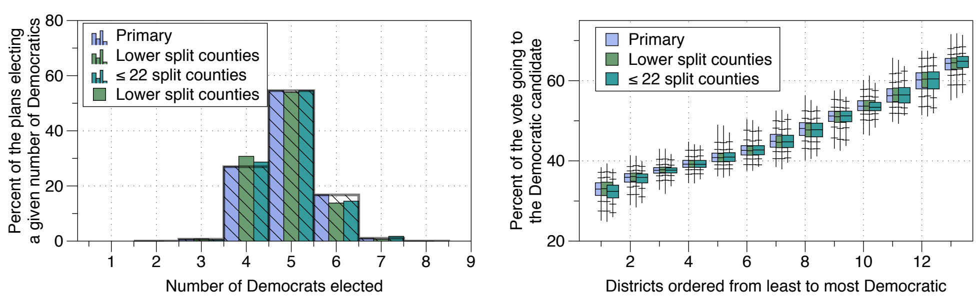

Figure S9: By changing the weights on the energy function we alter the distribution of two county splits, but find qualitatively identical results under the 2016 US Senate voting data. We display the number of Democrats elected over both the primary and reduced-county split ensembles (right) and the ranked marginal distributions of the Democratic vote fractions (left). Despite the changes in the ensemble, the results remain stable.

S4 Details on examining nearby redistricting plans

To randomly sample nearby districts, we run the same MCMC algorithm described above, but add a small modification: If a proposed step ever would increase the population difference between the current state of a district and its initial configuration above a certain fraction of the ideal district population, the step is rejected. Alternatively, one can think of $J(\xi) = \infty$ for any $\xi \in \mathcal{R}$ which has a district that differs from the original state by more than a certain percent of its population. As before, we then threshold the results on the population score and the minority score. We do not threshold on the isoparametric Score as we have demonstrated (see Section S3.1) that thresholding on compactness does not qualitatively affect our results. We also do not threshold on county splitting as the county splitting structure will remain similar due to the constraint on local deviations and with the county score function kept active.

The only exception to this is that we use a 4% population threshold for the NC2012 at 10% population deviation because the redistricting plans close to NC2012 would fail the 1.5% population threshold since the compactness energy dominates here due to the highly non-compact structure; we remark that taking away the population threshold, in this case, does not change our qualitative results.

We run 8 chains, this time each with the same initial condition, for a fixed wall clock time. Because cooling is dependent on accepting steps, we have a differing number of total cycles in each ensemble. Sub-sampling the above ensembles with the above thresholds leads to ensembles at 10% population deviation of 1,060 redistricting plans (10,578 without a population threshold) near the NC2012 plan, 25,969 redistricting plans near the NC2016 plan, and 17,513 redistricting plans near the Judges plan. Similarly, for the ensembles with 30% population deviation, we obtain 15,414 redistricting plans about the NC2012 plan, 26,787 redistricting plans about the NC2016 plan, and 14,980 redistricting plans about the Judges plan.

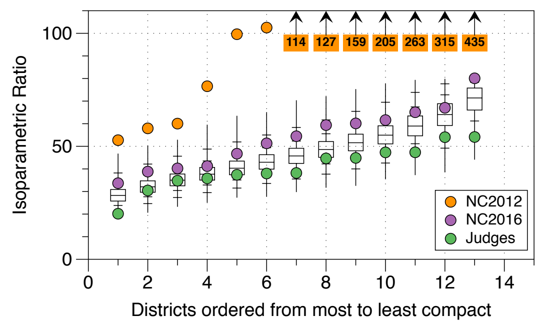

Figure S10: We examine the isoparametric ratios of the ranked districts from most to least compact. We display the ranked marginal distribution of the ensemble as box plots with dashes representing 1% and 10% outliers and lines extending to the maximum and minimum values. We display the ordered isoparametric ratios for the NC2012, NC2016 and Judges plans.

S5 Characteristics of the redistricting plans in the ensemble

We report the properties of the over 66,544 random restricting plans we have generated using the algorithm described in the preceding sections. We begin by examining the ranked marginal distributions on the isoparametric ratios, from most to least compact on our ensemble used in the main text. We contextualize the ranked isoparametric ratios of the three plans of interest in Figure S10. We find that the Judges plan is typically more compact than plans in the ensemble and that the NC2016 plan is generally less compact than plans in the ensemble. The NC2012 plan has plans with extremely non-compact districts.

Next, we examine the fraction of African Americans in the districts with the two highest fractions of this population. As mentioned previously, the NC2016 plan remedied the NC2012 racial gerrymander by reducing the fraction of African Americans in two most representative districts from 52% and 50% to 44.5% and 36.2%. We accept a map so long as there is one district with more than 40% fraction of African Americans, and a second with more than 33.5%. In our ensemble, we find that roughly 43% of the districts with the highest fraction of African Americans have more than 44.5% African Americans; for the collection of districts with the second most fractions of African Americans, we find that

78% of the districts are made up of more than 36.2% African Americans.

Finally, we exam the county splits of the primary ensemble. No plan in this ensemble splits any county into three districts. The median number of counties split across two districts is 34. We remark that the NC2012 plan split 40 counties and was never challenged on this basis. The NC2016, and Judges districting plans split 13, and 12 counties respectively. We have shown above that our results are not sensitive to reducing county splitting (see Section S3.3.1).

We provide all of the ensembles and source code in a git repository. This repository contains the primary ensemble, the ensemble that doubles the simulated annealing parameters, and the ensembles that vary the weights of the score functions. Each redistricting file holds the FID of the VTD, which are consistent with the FIDs in the Harvard's Election Data Archive Dataverse [11]. A read-me is provided. Each district, labeled by the FID, is associated with a district labeled 1-13 in the second column. In addition to the ensembles, we also provide the base data used in our analysis.

The repository may be accessed at
`https://git.math.duke.edu/gitlab/gjh/nccongressionalensembles.git`.