Inference on Ohio Redistricting Maps

Tara Abrishami
Department of Applied Mathematics and Statistics
Johns Hopkins University

Desh Raj
Department of Computer Science
Johns Hopkins University

Noah Scribner
Department of Applied Mathematics and Statistics
Johns Hopkins University

Vasileios Papaioannou
Department of Electrical and Computer Engineering
Johns Hopkins University

Abstract

Gerrymandering is the practice of drawing congressional districts to influence election outcomes in favor of a preferred result. There are few tools currently in place to determine whether a given district map has been gerrymandered in favor of one political party. Expanding on recent work, we use Markov Chain Monte Carlo (MCMC) sampling to estimate the distribution of possible district maps in the state of Ohio, using novel sampling techniques to achieve better candidate district maps. We propose several new statistical tools that use the MCMC posterior distribution to analyze district maps for political gerrymandering.

1 Introduction

In the United States, legislative power is divided between two bodies of Congress: the Senate and the House of Representatives. The Senate is comprised of two elected officials from each state. A Senator from a less populous state represents fewer people than a Senator from a more populous state, thereby giving citizens from less populous states greater political power. To balance this, the number of congresspeople in the House of Representatives from each state depends on the state's population, ensuring that every representative represents approximately the same number of people. For example, the most populous state, California, has fifty-three representatives, whereas the least populous state, Wyoming, has only one.

1.1 Gerrymandering Overview

Each state is assigned a number of representatives n based on their population. The state government divides the state into n congressional districts, where each district corresponds to the electorate for a single representative. A district map is a geographical partition of a state into the n districts. There are three main rules that a state must follow when creating a district map [1]:

1. The districts must be compact.
2. The districts must be contiguous.
3. Each district must contain approximately the same number of people.

There are several other laws at the state and federal levels that govern discrimination against gender, race, ethnicity, and other protected groups. However, these three rules are the main guidelines for

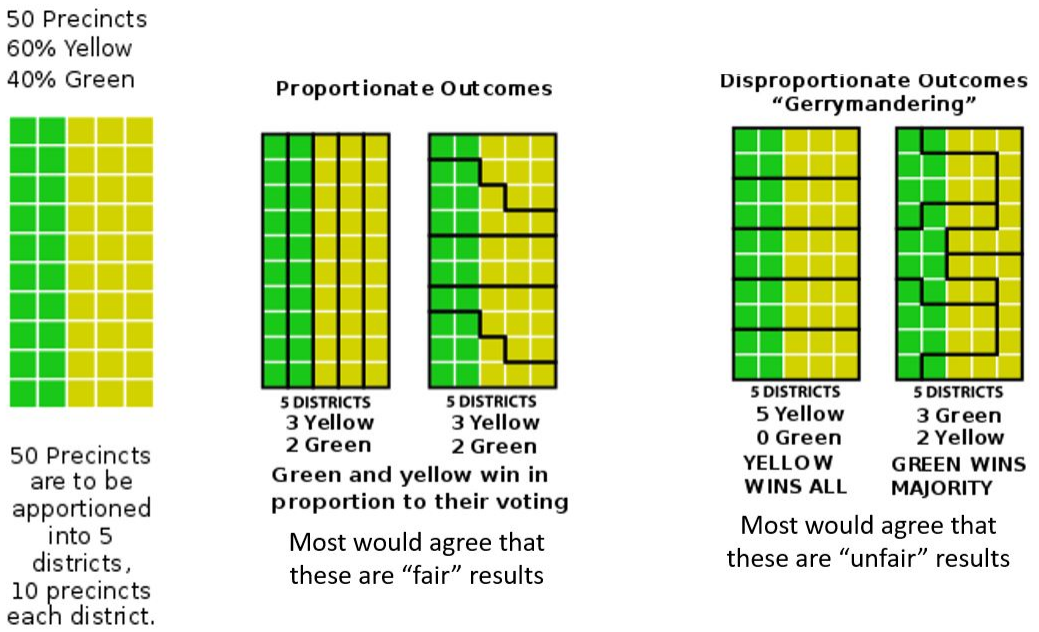


Figure 1: Simple Example of Gerrymandering [2].

states to create a district map. There is no overall standard method or procedure to develop a district map, leaving room for states to gerrymander.

Gerrymandering is the process of drawing district boundaries to disenfranchise one group over another. Figure 1 shows an example of how gerrymandering can be used to change election results.

It is clear that drawing district maps can sway which political party or group has the legislative power for that state. Gerrymandering is of dubious legality, but there are no laws or standards governing what legally constitutes gerrymandering. Though many state district maps have been challenged in court, with several cases making their way to the Supreme Court[1], proving that a map is gerrymandered is difficult without a strong definition of gerrymandering or a definitive procedure to identify a gerrymandered map. Recently, mathematicians have begun to apply statistical methods to analyze district maps for gerrymandering. The development of mathematical tools to study gerrymandering could have big implications for the legality of gerrymandered maps; indeed, testimonies from mathematicians have already been taken into account in court cases considering gerrymandering.

1.2 Related Research

Most of the mathematical modeling to identify gerrymandering involves the application of Markov Chain Monte Carlo sampling. The Metric Geometry and Gerrymandering Group (MGGG) at Tufts University and Massachusetts Institute of Technology and The Quantifying Gerrymandering Research Group (QG) at Duke University used Markov Chain Monte Carlo sampling to analyze the possibility of gerrymandering in the states of North Carolina and Wisconsin [3, 4]. Our methods for generating redistricting samples are similar to the approach described in Bangia et al [5], but we propose some new statistical techniques to quantify gerrymandering.

1.3 Ohio State

We chose Ohio to be our test state for a number of reasons. Ohio is a swing state, which means gerrymandered district maps may greatly affect election outcomes. Furthermore, Ohio's current district map is being challenged in court over accusations of partisan gerrymandering.

[1] Redistricting and the Supreme Court: the most significant cases

Data: A, $\mathcal{E}$, β, (w_p, w_i, w_c), T
Result: $(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_T)$
$\mathcal{C} \leftarrow$ Boundary Conflicts in $\mathcal{E}$;
$samples \leftarrow$ empty list of size T;
$samples[1] \leftarrow \mathcal{E}$;
$accept \leftarrow 0$;
for $t \leftarrow 2$ ***to*** T **do**
  $\mathcal{E}_t, \mathcal{E}_{t-1} \leftarrow samples[t-1]$;
  $c \leftarrow$ pick random edge from $\mathcal{C}$, $(p_1, p_2) \leftarrow c$;
  $u \leftarrow$ Uniform(0,1);
  if $u < \frac{1}{2}$ **then**
    for $q : \mathcal{E}_t[p_1] = \mathcal{E}_t[q]$ **do** $\mathcal{E}_t[q] \leftarrow \mathcal{E}_t[p_2]$;
    $\mathcal{E}_t[p_1] \leftarrow \mathcal{E}_t[p_2]$;
  else
    for $q : \mathcal{E}_t[p_2] = \mathcal{E}_t[q]$ **do** $\mathcal{E}_t[q] \leftarrow \mathcal{E}_t[p_1]$;
    $\mathcal{E}_t[p_2] \leftarrow \mathcal{E}_t[p_1]$;
  end
  $J_t \leftarrow$ score for $\mathcal{E}_t$;
  $J_{t-1} \leftarrow$ score for $\mathcal{E}_{t-1}$;
  $\mathcal{C}' \leftarrow$ new list of Boundary Conflicts;
  accept_prob $\leftarrow \frac{|\mathcal{C}|}{|\mathcal{C}'|} \cdot \exp(-\beta(J_t - J_{t-1}))$;
  if *accept_prob* > *Uniform(0,1)* **then**
    $samples[t] \leftarrow \mathcal{E}_t$;
    $accept \leftarrow accept + 1$;
    $\mathcal{C} \leftarrow \mathcal{C}'$;
  else
    $samples[t] \leftarrow \mathcal{E}_{t-1}$;
  end
end
return $samples$;

Algorithm 1: Metropolis-Hastings sampling for generating redistricting maps.

2 Methodology of Research

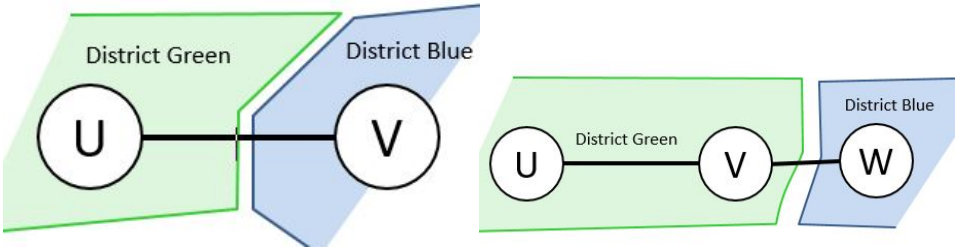
In this section, we discuss the methodology used for this study. Let us suppose that we are given a state S, comprised of precincts $p_i \in P, \forall i \in \{1, \ldots, n\}$ and districts $d_j \in D, \forall j \in \{1, \ldots, m\}$. A redistricting map is a function $\mathcal{E} \in \mathbb{E} : P \to D$. If no constraints are present on a redistricting, we have m^n possibilities, which is unfeasible to enumerate.

The output space $\mathbb{E}$ is constrained by several policies, as outlined earlier in subsection 1.1. This allows us to model the stated objective as a subgraph learning problem. We represent S as a graph where P is the set of vertices. Two vertices p_i and p_j have an edge between them if and only if they are geographically adjacent on the map. Our problem now reduces to dividing this graph into vertex-complete subgraphs which follow the other constraints.

Still, even with this simplification, the problem is known to be NP-hard [6]. Therefore, to solve it approximately, we use Metropolis-Hastings (MH) sampling approach [7]. In the remainder of this section, we describe each component of the MH sampling technique. The entire method is summarized in pseudocode in Algorithm 1.

2.1 Defining a sample

A sample refers to a redistricting map $\mathcal{E}$, which is obtained from a probability distribution $P_{\mathcal{E}}$ over all possible redistrictings. In our case, it is not possible to sample directly from this distribution, so we use an MCMC approach for sampling and then define an acceptance criteria based on score functions discussed on page 4 in subsection 2.4.



(a) (b)

Figure 2: (a) The edge $U - V$ is a boundary conflict since U and V lie in different districts. (b) This can be resolved by moving V to District Green (or U to District Blue), thus generating a new sample.

2.2 Initialization

For an MH sampling, we need to define a starting sample, from which the chain will then be generated. An ideal starting sample would be a random distribution, since we don't want any bias in our redistricting. Nonetheless, in a random redistricting, several adjacent precinct pairs would be allocated to different districts, thereby violating the contiguity principle (see subsection 1.1). Even if we reject such samples and modify them to improve contiguity, it would take several thousand iterations to converge to an acceptable sample.

For this reason, we choose to begin with a known valid redistricting map of the state S. We will now describe how a new redistricting sample is generated from a given sample.

2.3 Chain propagation

Let $\mathcal{E}$ be a given redistricting sample. We define a "boundary conflict" as an edge $e = (p_i, p_j)$ such that $\mathcal{E}(p_i) \neq \mathcal{E}(p_j)$, i.e., all edges such that the adjacent vertices belong to different districts. Simply put, these are all the edges which cross a district boundary (see Figure 2a, for example).

Suppose a boundary conflict $e = (u, v)$ is selected at random from $\mathcal{E}$. We generate a new redistricting $\mathcal{E}'$ as

$$\mathcal{E}'(p) = \begin{cases} \mathcal{E}(p), & \text{if } p \notin \{u, v\} \\ \mathcal{E}(u) \text{ or } \mathcal{E}(v), & \text{otherwise.} \end{cases} \tag{1}$$

The second case says that, with 50% probability, we assign u to $\mathcal{E}(v)$ or v to $\mathcal{E}(u)$. This is shown in Figure 2b.

A limitation with using (1) for chain propagation is that two consecutive redistricting maps differ in only one precinct. As such, it will take many iterations for redistrictings to be sufficiently distinct in electoral outcomes. To remedy this, we propose a second propagation scheme as follows.

$$\mathcal{E}'(p) = \begin{cases} \mathcal{E}(p), & \text{if } p \notin \{ne(u), ne(v)\} \\ \mathcal{E}(u) \text{ or } \mathcal{E}(v), & \text{otherwise,} \end{cases} \tag{2}$$

where $ne(p) = p \cup \{q : (p, q) \in A \mid \mathcal{E}(p) = \mathcal{E}(q)\}$, where A is the adjacency matrix of precincts for the state S. In other words, instead of shifting a single precinct, the precinct is moved along with all its immediate neighbors in the same district. This is represented in Figure 3.

2.4 Acceptance probability

Our initialization and chain propagation schemes ensure that districts remain contiguous throughout the MH sampling process.

Redistricting maps also need to be compact and have similar populations (see Section 1.1). Modeling these criteria in sample generation is difficult, so we model them instead in the acceptance probability. This is done using a scoring function $J(\mathcal{E})$ which assigns a real-valued score to a redistricting map.

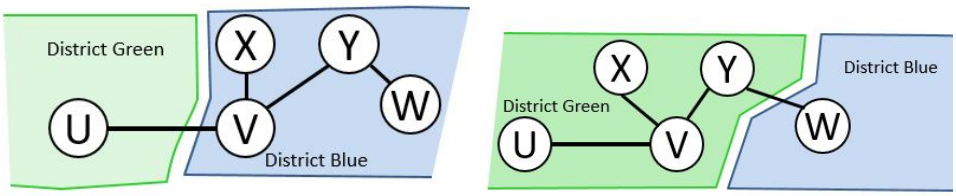


Figure 3: To obtain more different samples in each iteration, we move all the neighbors of a precinct along with it.

This scoring function is a weighted sum of three component , i.e.

$$J(\mathcal{E}) = w_p J_p(\mathcal{E}) + w_i J_i(\mathcal{E}) + w_c J_c(\mathcal{E}),$$

where J_p, J_i, and J_c are the population score, isoperimetric score, and county split score, respectively. The population score and isoperimetric score directly address two requirements of a district map. The county split score is an additional criterion that is used in several legal cases regarding gerrymandering. Combined together, the overall score measures how well the current redistricting conforms to the district guidelines.

2.4.1 Population score

For a given state S, we define the "ideal population" pop_{ideal} as the mean population of all districts in S. The population score quantifies the criterion that each district should contain approximately the same number of people in the sense of deviation of the population of a specific district D_i (denoted by $pop(D_i)$ from pop_{ideal}. Specifically,

$$J_p(\mathcal{E}) = \sqrt{\sum_i \left(\frac{pop(D_i)}{pop_{ideal}} - 1\right)^2},$$

where $pop_{ideal} = \frac{\sum_i pop_{D_i}}{\sum_i 1}$. As $J_p \to 0$, this deviation is minimized.

2.4.2 Isoperimetric Score

The isoperimetric score (also known as the perimeter score) measures the overall redistricting map's compactness. For district D_i, suppose $area(D_i)$ and $per(D_i)$ denote the area and perimeter, respectively, of D_i. To measure how compact D_i is, we use: $cpm(D_i) = \frac{per(D_i)^2}{area(D_i)}$. Then, the isoperimetric score for the district map $\mathcal{E}$ is:

$$J_i(\mathcal{E}) = \sum_i cpm(D_i)$$

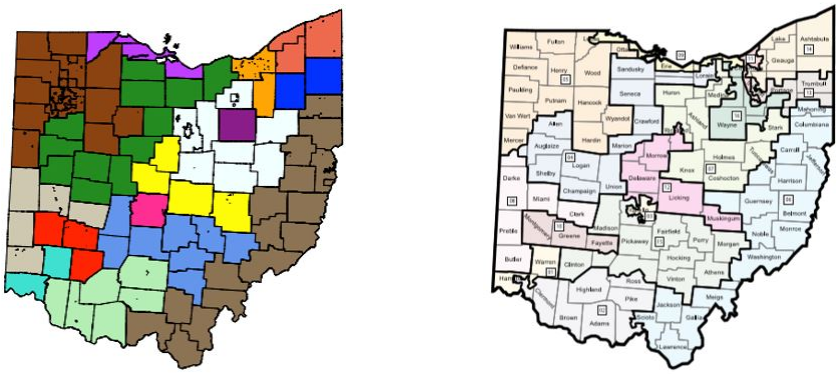
The isoperimetric score penalizes long, winding districts so that regular polygonal shapes are preferred. For this reason, this score has been frequently used in court to either refute or support the legality of district maps. In particular, Ohio's 'Snake on the Lake' 9th district and Illinois' 'Earmuff' 4th district score poorly on this measure [8], because of their unusual shapes which were drawn to avoid certain geographical areas.

2.4.3 County Split Score

The county split score penalizes district maps that divide local (i.e. county) government between representatives. The motivation is that congresspeople work closely with local officials to request funds for building projects, lobby issues, and request tax code changes. Therefore, in an ideal world, the local representative officials should have to work with only one representative and not have their jurisdiction split among several representatives. When this does not happen, a county split score is assigned to the district map.

Formally, define SF_i as the number of counties split in exactly i districts, and $SF_{\geq i}$ as the number of counties split in i or more districts. Then, we have,

$$J_c(\mathcal{E}) = w_1 SF_2 + w_2 SF_{\geq 3},$$



(a) (b)

Figure 4: (a): The starting point for MCMC. The map lines represent counties and the colors represent district assignments. (b): The current Ohio district map [9]. Counties are drawn in gray lines and labeled; districts are drawn in black lines and colored.

where the weights w_1, w_2 are chosen to penalize proposed district maps containing counties that are more heavily split. In practice, we choose $w_1 = 0.2$ and $w_2 = 0.8$, so that districts containing counties with more splits face heavier penalties.

3 Data description

For this project, we use the data compiled by the Metric Geometry and Gerrymandering Group on the state of Ohio. [2] The data contains both descriptive and geographical information on the state of Ohio.

Each of Ohio's 12,947 precincts is associated with 26 descriptive fields, the most relevant of which include population, county, voter turnout, and election results, and geographical details such as coordinates of boundary corners, geographical area, etc.

4 Results & Analysis

4.1 MCMC Convergence

We ran Markov Chain Monte Carlo (MCMC) as described in section 2 to obtain a distribution of potential redistricting maps. The initial district map we used is very similar to the current Ohio map, except that in the initial district map, no county is split across district lines; in other words, the county score for the initial map is 0. The initial district map and the current Ohio district map are shown in Figure 4.

Without access to significant compute power, we were only able to run the chain for 500 iterations [3], getting an acceptance rate of 63.4%. This is not nearly enough to converge to the underlying distribution. Indeed, if $\mathcal{E}_i$ represents the ith district map in the chain,

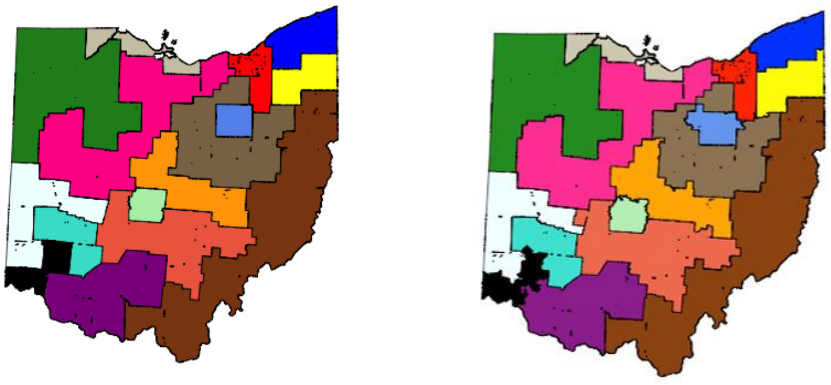
$$\max_{i \in \{1,\ldots,500\}} |\{p : \mathcal{E}_1(p) \neq \mathcal{E}_i(p)\}| = 266.$$

In other words, out of a total of 12,947 precincts, only a maximum of 266 of them ever had a district assignment different from their initial district in any of the maps $\mathcal{E}_i$ sampled by the Markov Chain. The weakness of the sampling is visualized in Figure 5.

We expect that to get reasonable convergence of the chain, we would need to generate on the order of 50,000 samples. Though we were unable to generate a chain with good convergence, our results do provide a proof-of-concept for our methods and suggest that, given better computing resources, the tools we develop here would be compelling advances in the mathematics to analyze gerrymandering.

[2] https://github.com/mggg/ohio-precincts

[3] This required 7 hours on a 2.7 GHz Core i7 (I7-8559U)



(a) (b)

Figure 5: (a): The initial district map $\mathcal{E}_1$. (b): The 500th district map $\mathcal{E}_{500}$. Note that, though there are visible differences between the maps, they are not appreciably different on a global scale.

4.2 Data Analysis

Using the results of the MCMC, we propose a number of statistical tools for data analysis and gerrymandering evaluation.

4.2.1 Election Results Distribution

A distribution of election results can be obtained by calculating the election results for each district in the chain of potential redistricting maps generated by MCMC. This type of analysis has been used in previous work to determine the fairness of election results for a specific district map [3]. In particular, an empirical distribution of election results would enable the calculation of the probability of a specific election result given the posterior distribution obtained via MCMC. If the election result probability is sufficiently low, it could indicate that partisan gerrymandering or other district manipulation is behind the statistically unlikely election outcome.

We attempted to compute an empirical distribution of election results using the posterior distribution of district maps from our MCMC; however, because our chain had such poor convergence, every map in the chain had the same election outcome. Therefore, this method of analysis is reserved for the future, when we have sufficient computational power to run the Markov Chain for more iterations and achieve better convergence.

4.2.2 Precinct Adjacency Probability Matrix

Using the posterior distribution of election maps obtained by MCMC, we can create a matrix of precinct adjacency probabilities. Specifically, we can create a matrix P such that the entry $P[i,j]$ corresponds to the probability that precincts p_i and p_j are in the same district. In other words, if MCMC$(\mathbb{E})$ represents the set of district maps in the posterior distribution,

$$P[i,j] := P\left(\mathcal{E}(p_i) = \mathcal{E}(p_j) \mid \mathcal{E} \in \text{MCMC}(\mathbb{E})\right).$$

An example of a precinct adjacency matrix is shown in Figure 6.

We propose a method of reducing the computational complexity of MCMC sampling through an analysis of the precinct adjacency probability matrix P. Rather than using the precincts as the units of area comprising a district, we seek to create *subdistricts*, or sets of precincts that are minimal units of area in sampling. Specifically, we can select a threshold value t such that if $P[i,j] > t$, then precincts p_i and p_j are always in the same district. Precincts p_i and p_j are then in the same subdistrict. Subdistricts can be computed from the precinct adjacency probability matrix P for a given probability threshold t using a union-find algorithm.

Performing MCMC sampling on a graph with subdistricts, rather than precincts, as the vertex set reduces the number of vertices in the graph, and therefore reduces the computational complexity of the sampling procedure. We conjecture that this will allow for faster convergence in a second

Figure 6: A heat map of the precinct adjacency probabilities for the subset of precincts initially assigned to District 16.

round of MCMC sampling, thus enabling more sophisticated sampling and potentially increasing the accuracy of the resulting posterior distribution.

4.2.3 Likelihood Ratio Test

We can also use the precinct adjacency probability matrix P to compute the likelihood of a given district map $\mathcal{E} \in \mathbb{E}$. We can obtain the set of boundary conflicts of $\mathcal{E}$, $bc(\mathcal{E})$, as follows:

$$bc(\mathcal{E}) = \{(p_i, p_j) \in A \mid \mathcal{E}(p_i) \neq \mathcal{E}(p_j)\}.$$

Each edge $e = (p_i, p_j) \in bc(\mathcal{E})$ has probability $1 - P[i, j]$. Each edge $e = (p_i, p_j) \notin bc(\mathcal{E})$ has probability $P[i, j]$. Therefore, the likelihood of $\mathcal{E}$ is

$$f(\mathcal{E}) = \prod_{i,j} P[i, j]\mathbb{1}_{(p_i,p_j)\notin bc(\mathcal{E})} \cdot (1 - P[i, j])\mathbb{1}_{(p_i,p_j)\in bc(\mathcal{E})}.$$

Since we have now defined a likelihood function $f : \mathbb{E} \to \mathbb{R}$, we can define the likelihood ratio of two district maps $\mathcal{E}_1$ and $\mathcal{E}_2$ as $\frac{f(\mathcal{E}_1)}{f(\mathcal{E}_2)}$. The likelihood ratio can be used to compare two proposal redistrictings $\mathcal{E}_1$ and $\mathcal{E}_2$. Since lawmakers are often faced with the task of selecting a district map out of several possible contenders, using the likelihood ratio to compare district maps has immediate and relevant political applications.

5 Conclusion

In this paper, we implement MCMC sampling to estimate the distribution of possible district maps in the state of Ohio. We encountered significant difficulties running the chain to convergence due to the computational power required to conduct MCMC on this scale. Nonetheless, our results demonstrate the potential of MCMC to simulate the distribution of district maps. We propose several mathematical tools to analyze gerrymandering, including a probability analysis on the precinct level and the development of a likelihood function for district maps. With better convergence of the MCMC, our tools can be applied to analyze partisan gerrymandering. All our code has been made publicly available for academic research [4].

References

[1] J Gerald Hebert, Martina E Vandenberg, and Paul Smith. The realist's guide to redistricting: Avoiding the legal pitfalls. American Bar Association, 2010.

[2] Wikipedia. Gerrymandering.

[3] Gregory Herschlag, Han Sung Kang, Justin Luo, Christy Vaughn Graves, Sachet Bangia, Robert Ravier, and Jonathan C Mattingly. Quantifying gerrymandering in north carolina. *arXiv preprint arXiv:1801.03783*, 2018.

[4] Link to code

[4] Robert Ravier, Jonathan Christopher Mattingly, and Gregory Joseph Herschlag. Evaluating partisan gerrymandering in wisconsin. *arXiv preprint arXiv:1709.01596*, 2017.

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

[6] Andreas Emil Feldmann and Luca Foschini. Balanced partitions of trees and applications. *Algorithmica*, 71(2):354–376, 2015.

[7] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[8] The Roll Call Staff. Top 5 ugliest districts: Partisan gerrymandering 101. *Roll Call*, November 10, 2011.

[9] Office of the Secretary of State for Ohio. Federal congressional districts, February 2018.