Implications of Distance over Redistricting Maps: Central and Outlier Maps*

Seyed A. Esmaeili [1], **Darshan Chakrabarti**[2], **Hayley Grape**[3], **Brian Brubach**[3]

[1]University of Chicago Data Science Institute
[2]Columbia University
[3]Wellesley College
esmaeili@uchicago.edu, darshan.chakrabarti@columbia.edu, hg3@wellesley.edu, bb100@wellesley.edu

Abstract

In representative democracy, a redistricting map is chosen to partition an electorate into districts which each elects a representative. A valid redistricting map must satisfy a collection of constraints such as being compact, contiguous, and of almost-equal population. However, these constraints are loose enough to enable an enormous ensemble of valid redistricting maps. This enables a partisan legislature to gerrymander by choosing a map which unfairly favors it. In this paper, we introduce an interpretable and tractable distance measure over redistricting maps which does not use election results and study its implications over the ensemble of redistricting maps. Specifically, we define a central map which may be considered "most typical" and give a rigorous justification for it by showing that it mirrors the Kemeny ranking in a scenario where we have a committee voting over a collection of redistricting maps to be drawn. We include running time and sample complexity analysis for our algorithms, including some negative results which hold using any algorithm. We further study outlier detection based on this distance measure and show that our framework can detect some gerrymandered maps. More precisely, we show some maps that are widely considered to be gerrymandered that lie very far away from our central maps in comparison to a large ensemble of valid redistricting maps. Since our distance measure does not rely on election results, this gives a significant advantage in gerrymandering detection which is lacking in all previous methods.

1 Introduction

Redistricting is the process of dividing an electorate into districts which each elect a representative. In the United States, this process is used for both federal and state-level representation, and we will use the U.S. House of Representatives as a running example. Subject to both state and federal law, the division of states into congressional districts is not arbitrary and must satisfy a collection of properties such as contiguity and near-equal population. Despite these regulations, redistricting is vulnerable to strategic manipulation in the form of gerrymandering. The body in charge can easily create a map within the legal constraints that leads to election results which favor a particular outcome (e.g., more representatives elected from one political party in the case of *partisan gerrymandering*). In addition, the ability to draw gerrymandered districts has improved greatly with the aid of computers since the historic salamander-shaped district approved by Massachusetts Governor Elbridge Gerry in 1812. For example, assuming voting consistent with the 2016 election, the state of North Carolina with 13 representatives can be redistricted to elect either 3 Democrats and 10 republicans or 10 Democrats and 3 Republicans.

However, despite this obvious threat to functioning democracy, partisan gerrymandering has often eluded regulation partly because it has been difficult to measure. In response, a recent line of impactful research introduced sampling techniques to randomly generate a large collection of redistricting maps[1] (Chikina, Frieze, and Pegden 2017; DeFord, Duchin, and Solomon 2019; Herschlag et al. 2020) and calculate statistics such as a histogram of the number of seats won by each party using this collection. This can show that a proposed or enacted map is an outlier in terms of its election outcome with respect to the sample. In fact, these techniques were used as a key argument in a recent U.S. Supreme Court case on partisan gerrymandering (Rucho v. Common Cause No. 18-422, 588 U.S. 2019) and have supported successful efforts to change redistricting maps in state supreme court cases (LWV vs Commonwealth of Pennsylvania No. 159 MM 2018). More importantly for the present work, at least two states, Michigan and Wisconsin, have used such a sampling tool (DeFord, Duchin, and Solomon 2019) in the recent redistricting process in response to the 2020 U.S. Census (Chen 2021). Recent papers have even extended these methods to locate which regions in a state are most unfairly impacted by a given map (Lin et al. 2022; Ko et al. 2022).

Despite this progress, all of these methods rely on election outcomes to detect any possible gerrymandering. This is a problem in instances where citizens, courts, and/or legislative bodies request methods which do not take partisanship into account. However, there has not been an effective method for detecting gerrymandering by identifying outliers according to a non-partisan metric. Abrishami et al. (2020) make progress toward this goal, but has limitations such a

*Readers are encouraged to see the arxiv version with more results: https://arxiv.org/abs/2203.00872

[1]These are not the truly uniform random samples from the immense and ill-defined space of all possible maps that we ideally want, but they are generally treated as such in courts.

using a small number of samples and not having a clear outlier score (See Section 2 and Appendix C.2 for more details). In this paper, we introduce a framework that resolves these issues and demonstrates the first effective method for detecting gerrymandering based on a distance (dissimilarity) measure over redistricting maps.

Furthermore, while progress has been made on the problem of detecting/labeling gerrymandering through the use of these sampling techniques, the question of drawing a redistricting map in a way that is "fair" and resists strategic manipulation remains largely unclear. We survey some existing proposals to automate redistricting in more detail in Section 2, but none of them have been adopted in practice thus far. Indeed, the issue of finding the "ideal" redistricting map is elusive and one of the main problems in redistricting and gerrymandering. We make progress in this direction, by introducing a novel method which selects the most "central" map. More specifically, given a collection of maps which are voted on by committee members we select the map with the minimum vote-weighted distance from the collection.

1.1 Our Contributions

Our paper introduces a number of contributions:

1. **Distance over Redistricting Maps**: We introduce a tractable family of distance measures over redistricting maps which have a simple edit distance interpretation. This family of distance measures can be adjusted to accommodate considerations such as a population or path length between voting blocks (Subsection 3.1).
2. **Medoid Map**: We introduce the *medoid map* and show that it mirrors the Kemeny ranking in a setting where committee members vote over a collection of maps. We further characterize the complexity of finding this map and introduce heuristics for finding it (Sections 4, 5, 6).
3. **Centroid Map**: We introduce the *centroid map* which is not a valid map, but has interesting properties and implications. We provide algorithms for finding this map and characterize its sample complexity (Sections 5,6).
4. **Gerrymandering Detection and Empirical Validation**: We show empirically that our framework can be used to detect gerrymandering in some instances. Further, we carry out extensive experimental validation of our pipeline where we ensure convergence and reproducibility by repeating the same experiment using different seeds. Remarkably, we reach the same conclusion across all runs: well-known gerrymandered maps of North Carolina and Pennsylvania have a distance in the 99th percentile away from the centroid map in comparison to an ensemble (Section 6 and Appendix B).

Furthermore, in Appendix C we include more details about gerrymandering detection such as the interpretation of having a large distance from the centroid map and a more detailed comparison to the work of (Abrishami et al. 2020). Finally, we note that our framework can adapt to various considerations such the Voting Rights Act (VRA) (Bickel 1966) and other state-specific redistricting rules by simply modifying the sampling method to adapt to these considerations. Due to the space limits, we delay all proofs to Appendix A.

2 Related Work

Less than a decade ago, several early works ushered in the current era of Markov Chain Monte Carlo (MCMC) sampling techniques for gerrymandering detection (Mattingly and Vaughn 2014; Wu et al. 2015; Fifield et al. 2015). Followup work has both refined these techniques and further analyzed their ability to approximate the target distribution. Authors of these works have been involved in court cases in Pennsylvania (Chikina, Frieze, and Pegden 2017) and North Carolina (Herschlag et al. 2020) with sampling approaches being used to demonstrate that existing maps were outliers as evidence of partisan gerrymandering. One of the most recent works in this area introduces the **ReCom** tool (DeFord, Duchin, and Solomon 2019) which was used by the Wisconsin Peoples Maps Commission and the Michigan Independent Citizens Redistricting Commission in the current redistricting cycle following the 2020 U.S. census (Chen 2021). More recently, the works of (Lin et al. 2022; Ko et al. 2022) have made an interesting extension of the previous methods by identifying the voting blocks unfairly impacted in a gerrymandered map. Generally, these techniques have primarily been used to analyze and sometimes reject existing maps rather than draw new maps. However, we may view them as narrowing the search space of maps that can be drawn. Furthermore, it has been shown that even the regulation of gerrymandering via outlier detection is subject to strategic manipulation (Brubach, Srinivasan, and Zhao 2020).

On the automated redistricting side, many map drawing algorithms have favored optimization approaches and in particular, optimizing some notion of compactness while avoiding explicit use of partisan information. Approaches emphasizing compactness include balanced power diagrams (Cohen-Addad, Klein, and Young 2018), a k-median-based objective (Bycoffe et al. 2018), and minimizing the number of cut edges (Hettle et al. 2021). Some works include partisan information for the sake of creating competitive districts (districts with narrow margins between the two main parties). The PEAR tool (Liu, Cho, and Wang 2016) balances nonpartisan criteria like compactness (defined by the Polsby-Popper score (Polsby and Popper 1991)) with other criteria such as competitiveness and uses an evolutionary algorithm with some similarity to the random walks taken by MCMC sampling approaches. Other works go further in the explicitly partisan direction such as the game theoretic approach of Pegden, Procaccia, and Yu (2017) which seeks a map that is fair to two parties. Finally, there are methods which prefer simplicity such as the Splitline (Ryan and Smith 2022) algorithm which iteratively splits a state until the desired number of districts is reached.

In all of these approaches, the aim is to automate redistricting, but it is difficult to determine whether the choices made are the "right" or "fairest" decisions. The question of whether optimizing properties such as compactness while ignoring partisan factors could result in partisan bias is a concern. Cho (2019) notes a comment by Justice Scalia suggesting that such a process could be biased against Democratic voters clustered in cities in Vieth v. Jubelirer (Vieth v. Jubelirer No. 02-1580, 541 U.S. 267 (2004). For those that do take partisan bias into account, there are questions of

whether purposely drawing competitive districts or giving a fair allocation to two parties are really beneficial to voters.

Finally, the work of Abrishami et al. (2020) introduces a distance measure over redistricting maps. However, our distance is easy to compute and does not require solving a linear program. Further, our focus is on the implications of having a distance measure, i.e. the medoid and centroid maps that will be introduced. Moreover, unlike Abrishami et al. (2020) we can detect gerrymandered maps rigorously by specifying where they lie on a distance histogram without using an embedding method and using 200,000 samples instead of only 100. Finally, we give a clear outlier score for a given map (its percentile distance from the central map) whereas (Abrishami et al. 2020) cannot do that. Therefore, it is difficult to see how their methods would be applied in a real practical setting. See Appendix C.2 for a more detailed discussion.

3 Problem Set-Up

A given state is modelled by a graph $G = (V, E)$ where each vertex $v \in V$ represents a voting block (*unit*). Each unit v has a weight $w(v) > 0$ which represents its population. Further, $\forall u, v \in V$ there is an edge $e = (u, v) \in E$ if and only if the two vertices are *connected* (geographically this means that units u and v share a boundary). The number of units is $|V| = n$. A *redistricting* (*redistricting map* or simply *map*) M is a partition of V into $k > 0$ many districts, i.e., $V = V_1 \cup V_2 \cdots \cup V_k$ where each V_i represents a district and $\forall i \in [k], |V_i| \neq 0$ and $\forall i, j \in [k], V_i \cap V_j = \emptyset$ if $i \neq j$. The redistricting map M is decided by the induced partition, i.e., $M = \{V_1, \ldots, V_k\}$. For a redistricting M to be considered valid, it must satisfy a collection of properties, some of which are specific to the given state. We use the most common properties as stated in (DeFord, Duchin, and Solomon 2019; Hettle et al. 2021): **(1)** *Compactness*: The given partitioning should have "compact" districts. Although there is no definitive mathematical criterion which decides compactness for districts, some have used common definitions such as Polsby-Popper or Reock Score (Alexeev and Mixon 2018). Others have used a clustering criterion like the k-median objective (Cohen-Addad, Klein, and Young 2018) or considered the total number of cuts (number of edges between vertices in different districts) (DeFord, Duchin, and Solomon 2019). **(2)** *Equal Population*: To satisfy the desideratum of "one person one vote" each district should have approximately the same number of individuals. I.e., a given district V_i should satisfy $\sum_{v \in V_i} w(v) \in [(1-\epsilon)\frac{\sum_{v \in V} w(v)}{k}, (1+\epsilon)\frac{\sum_{v \in V} w(v)}{k}]$ where ϵ is a non-negative parameter relaxing the equal population constraint. **(3)** *Contiguity*: Each district (partition) V_i should be a connected component, i.e., $\forall i \in [k]$ and $\forall u, v \in V_i$, v should be reachable from u through vertices which only belong to V_i. Our proofs do not rely on these properties and therefore can accommodate additional properties if desired.

Let $\mathcal{M}$ be the set of all valid maps. Let $\mathcal{D}(\mathcal{M})$ be a distribution over these maps. Furthermore, define a distance function over the maps $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$. Then the *population medoid map* is M^* which is a solution to the following:

$$M^* = \arg\min_{M \in \mathcal{M}} \mathop{\mathbb{E}}_{M' \sim \mathcal{D}(\mathcal{M})} [d(M, M')] \tag{1}$$

In words, the population medoid map is a valid map minimizing the expected sum of distances away from all valid maps according to the distribution $\mathcal{D}(\mathcal{M})$. This serves as a natural way to define a central or most typical map with respect to a given distance metric of interest.

Since we clearly operate over a sample (a finite collection) from $\mathcal{D}(\mathcal{M})$; therefore, we assume that the following condition holds:

Condition 3.1. *We can sample maps from the distribution $\mathcal{D}(\mathcal{M})$ in an independent and identically distributed (**iid**) manner in polynomial time.*

We note that although independence certainly does not hold over the sampling methods of (DeFord, Duchin, and Solomon 2019; Mattingly and Vaughn 2014) since they use MCMC methods, it makes the derivations significantly more tractable. Further, the specific choice of the sampling technique is somewhat immaterial to our objective.

Based on the above condition, we can sample from the distribution $\mathcal{M}$ efficiently and obtain a finite set of maps $\mathcal{M}_T$ having T many maps, i.e., $|\,\mathcal{M}_T\,| = T$.

Now, we define the *sample medoid*, which is simply the extension of the population medoid, but restricted to the given sample. This leads to the following definition:

$$\bar{M}^* = \arg\min_{M \in \mathcal{M}_T} \sum_{M' \in \mathcal{M}_T} d(M, M') \tag{2}$$

3.1 Distance over Redistricting Maps

Before we introduce our distance measure, we note that a given map (partition) M can be represented using an "adjacency" matrix A in which $A(i, j) = 1$ if and only if $\exists V_\ell \in M : i, j \in V_\ell$ otherwise $A(i, j) = 0$. We note that this adjacency matrix can be seen as drawing an edge between every two vertices i, j that are in the same district, i.e., where $A(i, j) = 1$. It is clear that we can refer to a map by the partition M or the induced adjacency matrix A. Accordingly, we refer to the population medoid as M^* or A^* and the sample medoid as $\bar{M}^*$ or $\bar{A}^*$

We now introduce our distance family which is parametrized by a weight matrix Θ and have the following form:

$$d_\Theta(A_1, A_2) = \frac{1}{2} \sum_{i,j \in V} \theta(i, j) |A_1(i, j) - A_2(i, j)| \tag{3}$$

where we only require that $\theta(i, j) > 0, \forall i, j \in V$ where $\theta(i, j)$ is the (i, j) entry of Θ. For the simple case where $\theta(i, j) = 1, \forall i, j \in V$, our distance $d_{\mathbf{1}}(A_1, A_2)$ is equivalent to a Hamming distance over adjacency matrices. When $\theta(i, j) = 1, \forall i, j \in V$, we refer to the metric as the *unweighted distance*. We note that such a distance measure was used in previous work that considered adversarial attacks on clustering (Chhabra, Roy, and Mohapatra 2020; Cinà, Torcinovich, and Pelillo 2022).

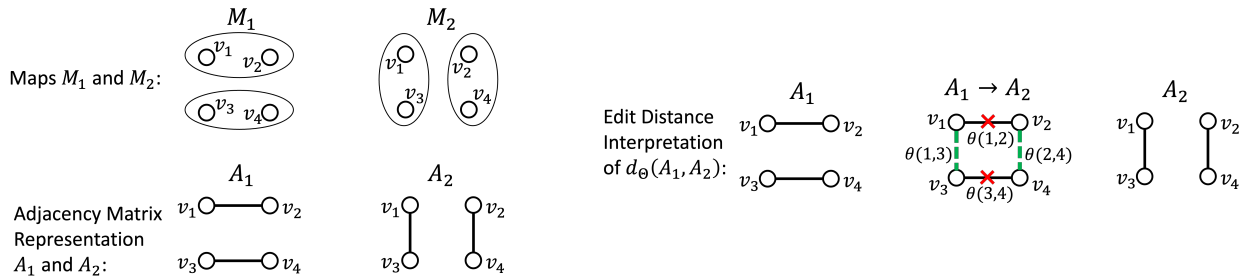

Figure 1: We are given a hypothetical state consisting of 4 vertices $V = \{v_1, v_2, v_3, v_4\}$ with M_1 and M_2 being two valid redistricting maps. The adjacency matrices A_1,A_2, and edit distance interpretation of $d_\Theta(A_1, A_2)$ are demonstrated. Note that $d_\Theta(A_1, A_2) = \theta(1,2) + \theta(3,4) + \theta(1,3) + \theta(2,4)$ which is exactly the minimum sum of edge weights that need to be deleted and added to obtain A_2 from A_1.

Another choice of Θ that leads to a meaningful metric is the *population-weighted distance* where $\theta(i,j) = w(i)w(j)$. This leads to $d_W(A_1, A_2) = \frac{1}{2}\sum_{i,j \in V} w(i)w(j)|A_1(i,j) - A_2(i,j)|$. The population-weighted distance takes into account the number of individuals being separated from one another when vertices i and j are separated from one another[2] by assigning a cost of $w(i)w(j)$. By contrast, the unweighted distance assigns the same cost regardless of the population values and thus has a uniform weight over the separation of units immaterial of the populations which they include.

Another choice of metric which is meaningful, could be of the form $\theta(i,j) = f(l(i,j))$ where $l(i,j)$ is the length of a shortest path between i and j and $f(.)$ is a positive decreasing function such as $f(l(i,j)) = e^{-l(i,j)}$. Such a metric would place a smaller penalty for separating vertices that are far away from each other. In general, our distance has an edit distance interpretation. Specifically, if we were to draw edges between vertices according to the entries with 1 in the adjacency matrix, then given A_1 and A_2, the distance $d_\Theta(A_1, A_2)$ simply equals the minimum total weight (according to $\theta(i,j)$) of the edges that must be added and deleted to obtain A_2 from A_1. In the case of the unweighted distance, it is precisely equal to the minimum number of edges that have to be deleted from and added to A_1 to obtain A_2. See Figure 1 for an illustration.

4 Justification for Choosing a Central Map

Connection to the Kemeny Rule: We note that the Kemeny rule (Kemeny 1959; Brandt et al. 2016) is the main inspiration behind our proposed framework. More specifically, given a set of alternatives and individuals voting by ranking the alternatives, the Kemeny rule provides a method for aggregating the resulting collection of rankings. This is done by introducing a distance measure over rankings (the Kendall tau distance (Kendall 1938)) and choosing the ranking which minimizes the sum of distances away from the other rankings in the collection as the aggregate ranking.

Although we do not deal with rankings here, we follow a similar approach to the Kemeny rule as we introduce a distance measure over redistricting maps and choose the map which minimizes the sum of the distances as the aggregate map. In fact, recently there has been significant citizen engagement in drawing redistricting maps. For example, in the state of Maryland an executive order from the governor has established a web page to collect citizen submissions of redistricting maps (Commission 2021). If each member of a committee was to vote for exactly one map in the given submitted maps, then if we interpret the probability $p_{M'}$ for a map $M' \in \mathcal{M}$ to be the number of votes it received from the total set of votes, then the medoid map M^* (similar to the Kemeny ranking) would be the map which minimizes the weighted sum of distances from the set of maps voted on. We include this result as a proposition and its proof follows directly from the definition we gave above:

Proposition 1. *Suppose we have a committee of $\mathcal{T}$ many voters and that each voter votes for one map from a subset of all possible valid maps $\mathcal{M}$, then given a map M', if we assign it a probability $p_{M'} = \frac{\sum_{\tau=1}^{\mathcal{T}} \nu_{\tau,M'}}{\mathcal{T}}$ where $\nu_{\tau,M'} \in \{0,1\}$ is the vote of member τ for map M', then the medoid map $M^* = \arg\min_{M \in \mathcal{M}} \mathbb{E}_{M' \sim p_{M'}}[d(M, M')]$ is the map that minimizes the sum of distances from the set of valid maps where the distance to each map is weighted by the total votes it receives.*

Connection to Distance and Clustering Based Outlier Detection: The medoid map, by virtue of minimizing the sum of distances, can be considered a central map. Accordingly, one may consider using the medoid map to test for gerrymandering in a manner similar to distance and clustering based outlier detection (He, Xu, and Deng 2003; Knox and Ng 1998). More specifically, given a large ensemble of maps, if the enacted map is far from the medoid[3] in comparison to the ensemble then this suggests possible gerrymandering. In fact, we carry experiments on the states of North Carolina and Pennsylvania (both of which have had enacted maps which were considered gerrymandered) and we indeed find the gerrymandered maps to be faraway whereas the remedial maps are much closer in terms of distance.

[2]Recall that each vertex (unit) is a voting block (AKA voter tabulation district) and units may contain different numbers of voters.

[3]Our experiments use the centroid instead of the medoid map.

5 Algorithms

We show our linear time algorithm for obtaining the sample medoid in Subsection 5.1. In Subsection 5.2, we define the population centroid, derive sample complexity guarantees for obtaining it, and show that its (i, j) entry equals the probability of having i and j in the same district. Finally, in Subsection 5.3, we discuss obtaining the population medoid and show that in general an arbitrarily large sample is not sufficient to approximate it.

Before we introduce our algorithms, we show that our distance family is a metric (satisfies the properties of a metric):

Proposition 2. *For all Θ such that $\forall i, j, \theta(i, j) > 0$, the following distance function is a metric.*

$$d_\Theta(A_1, A_2) = \frac{1}{2} \sum_{i,j \in V} \theta(i, j)|A_1(i, j) - A_2(i, j)|$$

5.1 Obtaining the Sample Medoid

We note that in general obtaining the sample medoid is not scalable since it usually takes quadratic time (Newling and Fleuret 2017) in the number of samples, i.e. $\Omega(T^2)$. An $O(T^2)$ run time can be easily obtained through a brute-force algorithm which for every map calculates the sum of the distances from other maps and then selects the map with the minimum sum. However, for our family of distances $d_\Theta(.,.)$ we show that the medoid map is the closest map to the centroid map (defined below) and show a simple algorithm that runs in $O(T)$ time for obtaining the sample medoid. The fundamental cause behind this speed up is an equivalence between the Hamming distance over binary vectors and the square of the Euclidean distance which is still maintained with our generalized distance. Before introducing the theorem we define $d_{2,\Theta}(A_1, A_2) = \frac{1}{2} \sum_{i,j \in V} \theta(i, j)(A_1(i, j) - A_2(i, j))^2$ where the absolute has been replaced by a square. Now we state the decomposition theorem:

Theorem 5.1. *Given a collection of redistricting maps $A_1, \ldots, A_T$, the sum of distances of the maps from a fixed redistricting map A' equals the following:*

$$\sum_{t=1}^{T} d_\Theta(A_t, A') = \sum_{t=1}^{T} d_{2,\Theta}(A_t, \bar{A}_c) + T\, d_{2,\Theta}(\bar{A}_c, A') \tag{4}$$

where $\bar{A}_c = \frac{1}{T} \sum_{t=1} A_t$.

Notice that the above theorem introduces the centroid map $\bar{A}_c$ which is simply equal to the empirical mean of the adjacency maps. It should be clear that with the exception of trivial cases the centroid map $\bar{A}_c$ is not a valid adjacency matrix, since despite being symmetric it would have fractional entries between 0 and 1. Hence, the centroid map also does not lead to a valid partition or districting. Moreover, we note that it is more accurate to call $\bar{A}_c$ the sample centroid, as opposed to the population centroid A_c (see Subsection 5.2) which we would obtain with an infinite number of samples.

The above theorem leads to Algorithm 1 with the following remark:

Algorithm 1: Finding the Sample Medoid

Input: $\mathcal{M}_T = \{A_1, \ldots, A_T\}$, $\Theta = \{\theta(i, j) > 0, \forall i, j \in V\}$.

1: Calculate the centroid map $\bar{A}_c = \frac{1}{T} \sum_{t=1}^{T} A_t$.

2: Pick the map $\bar{A}^* \in \mathcal{M}_T$ which minimizes the $d_{2,\Theta}$ distance from the centroid $\bar{A}_c$, i.e. $\bar{A}^* = \arg\min_{A \in \mathcal{M}_T} d_{2,\Theta}(A, \bar{A}_c)$.

return $\bar{A}^*$

Remark 1. *Algorithm 1 returns the correct sample medoid and runs in $O(T)$ time.*

We note that calculating the sample medoid in algorithm 1 has no dependence on the generating method. Therefore, if a set of maps are produced through any mechanism and are considered to be representative and sufficiently diverse, then algorithm 1 can be used to obtain the sample medoid in time that is linear in the number of samples.

5.2 Sample Complexity for Obtaining the Population Centroid

In the prior section, we introduced the sample centroid $\bar{A}_c$ which is equal to the empirical mean from taking the average of the adjacency matrices, i.e., $\bar{A}_c = \frac{1}{T} \sum_{t=1}^{T} A_t$. We now consider the population centroid $A_c = \lim_{T \to \infty} \sum_{t=1}^{T} A_t$. Clearly, by the law of large numbers (Zubrzycki 1972), we have $A_c(i, j) = \mathbb{E}[A(i, j)]$. It is also clear that A_c has an interesting property, specifically the (i, j)-entry equals the probability that i and j are in the same district:

Proposition 3. $A_c(i, j) = \Pr[i \textit{ and } j \textit{ in the same district}]$.

Now we show that with a sufficient number of samples, the sample centroid converges to the population centroid entry-wise and in terms of the $d_{2,\Theta}$ value. Specifically, we have the following proposition:

Proposition 4. *If we sample $T \geq \frac{1}{\epsilon^2} \ln \frac{n}{\delta}$ **iid** samples, then with probability at least $1 - \delta$, we have that $\forall i, j \in V : |\bar{A}_c(i, j) - A_c(i, j)| \leq \epsilon$. Further, let $\kappa = \max_{i,j \in V} \sqrt{\theta(i, j)}$, if we have $T \geq \frac{\kappa n^2}{\epsilon} \ln \frac{n}{\delta}$ **iid** samples, then $d_{2,\Theta}(\bar{A}_c, A_c) \leq \epsilon$ with probability at least $1 - \delta$.*

5.3 Obtaining the Population Medoid

Having found the sample centroid $\bar{A}_c$ and shown that it is a good estimate of the population centroid A_c, we now show that we can obtain a good estimate of the population medoid by solving an optimization problem. Assuming that we have the population centroid A_c, then the population medoid is simply a valid redistricting map A which has a minimum $d_{2,\Theta}(A, A_c)$ value. This follows immediately from Theorem 5.1. More interestingly, we show that this optimization problem is a constrained instance of the min k-cut problem:

Theorem 5.2. *Given the population centroid A_c, the population medoid A^* can be obtained by solving a constrained min k-cut problem.*

If we have a good estimate $\bar{A}_c$ of the population centroid A_c, then we can solve the above optimization using $\bar{A}_c$ instead of A_c and obtain an estimate of the population medoid $\bar{A}^*$ instead of the true population medoid A^* and bound the error of that estimate. The issue is that the min k-cut problem is NP-hard (Goldschmidt and Hochbaum 1994; Saran and Vazirani 1995) [4]. Further, the existing approximation algorithms assume non-negativity of the weights. Even if these approximation algorithms can be tailored to this setting, the additional constraints on the partition being a valid redistricting (each partition being contiguous, of equal population, and compact) make it quite difficult to approximate the objective. In fact, excluding the objective and focusing on the constraint alone, only the work of (Hettle et al. 2021) has produced approximation algorithm for redistricting maps but has done that for the restricted case of grid graphs. Further, while there exists heuristics for solving min k-cut for redistricting maps they only scale to at most around 500 vertices (Validi and Buchanan 2022).

Having shown the difficulty in obtaining the population medoid by solving an optimization problem, it is reasonable to wonder whether we can gain any guarantees about the population medoid by sampling. We show the negative result that we cannot guarantee that we can estimate the sample medoid of a distribution with high probability by choosing a sampled map even if we sample an arbitrarily large number of maps. This implies as a corollary that the sample medoid does not converge to the population medoid in contrast to the centroid (see Proposition 4).

Theorem 5.3. *For any arbitrary T many **iid** samples $\{A_1, \ldots, A_T\}$ there exists a distribution over a set of redistricting maps such that: (1) $\Pr[\min_{A \in \{A_1,\ldots,A_T\}} d(A, A^*) \geq 0.331] \geq \frac{2}{3}$ and (2) $\Pr[\min_{A \in \{A_1,\ldots,A_T\}} f(A) \geq 1.1 f(A^*)] \geq \frac{2}{3}$ where $f(.)$ is the medoid cost function.*

We therefore, use a heuristic to find the medoid as mentioned in the next section.

6 Experiments

We conduct experiments on three states, North Carolina (NC), Maryland (MD), and Pennsylvania (PA), which have featured in major court cases on partisan gerrymandering (LWV vs Commonwealth of Pennsylvania No. 159 MM 2018; Rucho v. Common Cause No. 18-422, 588 U.S. 2019). The number of voting units (vertices) are around 2,700, 1,800, and 8,900, and the number of districts are 13, 8, and 18 for NC, MD, and PA, respectively[5]. We focus on the results for NC here and discuss the qualitatively-similar results for PA and MD in Appendix B. NC is especially valuable because we have good examples of both gerrymandered and not gerrymandered maps (enacted maps which are widely considered gerrymandered and a remedial map which is not).

To generate a collection of maps, we use the Recombination algorithm, **ReCom**, from DeFord, Duchin, and Solomon (2019) whose implementation is available online. **ReCom** is a Markov Chain Monte Carlo (MCMC) sampling method, and hence, the generated samples are not actually **iid**. While this means Condition 3.1 does not hold, we believe our theorems still have utility and future work can address more realistic sampling conditions. Moreover, we always exclude the first 2,000 samples from any calculation as these are considered to be "burn-in" samples[6]. Throughout this section, when we say distance, we mean $d_{2,\Theta}(.,.)$ instead of $d_{\Theta}(.,.)$. Further experimental results and figures are included in Appendix B.

Convergence of the Centroid: Previous work (DeFord, Duchin, and Solomon 2019; DeFord and Duchin 2019) has used the **ReCom** algorithm for estimating statistics such as the histogram of election seats won by a party and determined that using 50,000 samples is sufficient for accurate results. However, our setting is more challenging. Specifically, the centroid includes $\Omega(n)$ entries where n is the total number of voting units (vertices) whereas an election histogram includes only k entries where k is the number of districts—usually orders of magnitude smaller than the number of voting units. Thus, we sample 200,000 maps instead to estimate the centroid. Here, we emphasize the importance of our linear-time algorithm since using a quadratic-time algorithm on samples of the order of even 50,000 could be computationally difficult. Following similar practice to Herschlag et al. (2020) for verifying convergence, we repeat the procedure (sampling using **ReCom** and estimating the centroid) for a total of three times for each state, starting from a different seed map each time and confirming that all three runs result in essentially the same centroid estimate.

To verify the closeness of the different centroid estimates, we calculate the distances between them and compare them to their distances from sampled redistricting maps using **ReCom**. We find that the centroids are orders of magnitude closer to each other than to any other sampled map. For example, the maximum unweighted distance between any two centroids is less than 130 whereas the minimum unweighted distance between any of the three centroids and any sampled map is more than 100,000. Similarly, the maximum weighted distance between any two centroids is less than 1.6×10^9 whereas the minimum weighted distance between a sampled map and a centroid is at least 1.3×10^{12} which is again three orders of magnitude higher.

Distance Histogram and Detecting Gerrymandering: For each state, we plot the distance histogram from its centroid. More specifically, having estimated the centroid $\bar{A}_c$, we sample 200,000 maps and calculate $d_{2,\Theta}(\bar{A}_c, A_t)$ where

[4] Note that in our case the min k-cut problem can have negative edge weights while the min k-cut problem is generally stated with non-negative weights. Nevertheless, we still minimize a cut objective and the non-negative weight min k-cut instance is trivially reducible to a min k-cut instance with negative and non-negative weights.

[5] For PA, the number of districts was reduced to 17 after the 2020 census, but we are using past election results with 18 districts.

[6] In MCMC, the chain is supposed to converge to a stationary distribution after some number of steps, called the mixing time. Although the mixing time has not been theoretically calculated for **ReCom**, empirically it seems that 2,000 steps are sufficient.

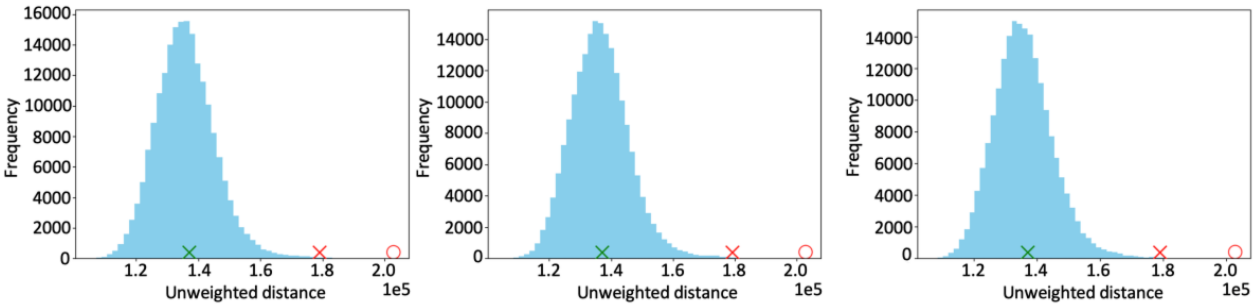

Figure 2: Distance histograms for NC using the unweighted distance measure. Different plots correspond to different seeds. For NC the distances of gerrymandered maps are indicated with red markers whereas the distances of the remedial maps are indicated with green markers (the circle and the X are for 2011 and 2016 enacted maps, respectively).

A_t is the t^{th} sampled map. Figure 2 shows the unweighted distance histogram for NC[7]. The histogram appears like a normal distribution, peaking at the middle (around the mean) and falling almost symmetrically away. This shows that the maps do not concentrate near the centroid even though it minimizes the sum of $d_{2,\Theta}(\bar{A}_c, A_t)$ distances. Interestingly, the histogram has a similar shape for both distances (unweighted and weighted), and this shape remains unchanged across the different seeds.

Furthermore, previous work used similar sampling methods to detect gerrymandered maps (Chikina, Frieze, and Pegden 2017; Mattingly and Vaughn 2014; Herschlag et al. 2020). In essence, these papers demonstrated that the election outcome achieved by the enacted map was rare in comparison to a large sampled ensemble of redistricting maps. Using similar logic, we can also detect gerrymandered maps. Specifically, the 2011 and 2016 enacted maps of NC were widely considered to be gerrymandered, and both maps are at the right tail of the histogram, far from the centroid. By contrast, a remedial NC map drawn by a set of retired judges (Herschlag et al. 2020) is much closer to the centroid (see Figure 2 red and green marked symbols). Interestingly, all gerrymandered maps are in the 99th percentile in terms of distance (for both distance measures and across three seeds).

This suggests that our method can detect gerrymandered maps with two advantages over previous methods: it does not use election results or partisan outcomes and it is very interpretable. Thus, a guideline or rule that maps should not be far away from the centroid can exist alongside reforms that prohibit explicit partisan consideration.

Finding the medoid: Since we have shown in Subsection 5.3 that the medoid cannot be obtained by sampling, we follow a heuristic that consists of these steps: (1) Sample 200,000 maps and pick the one closest to the centroid A_{closest}. (2) Start the **ReCom** chain from A_{closest} but given a specific state (redistricting map) we only allow transitions to new states (maps) that are closer to the centroid, and we do this for 200,000 steps to obtain the final estimated medoid $\hat{A}^*$. We follow this procedure three times one for each centroid [8]. Figure 3 (top row) shows the A_{closest} medoids from two different runs (each comparing to a different centroid), and it is easy to see they are different. The bottom row shows the final medoids $\hat{A}^*$ which are visually more similar to each other and also close in distance.

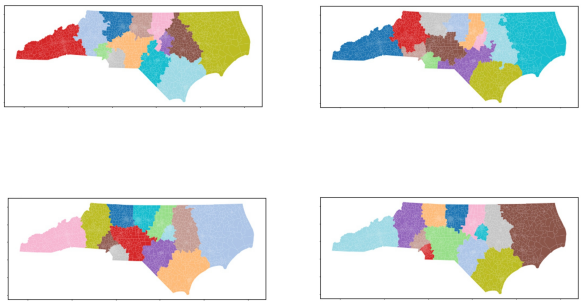

Figure 3: NC medoids, each column is for a specific seed. Top row: A_{closest}, Bottom row: $\hat{A}^*$.

7 Conclusion

In this paper we introduced a framework which accounts for the challenge of choosing from a large space of valid maps in the redistricting process. Specifically, we introduced a well-motivated family of distance measures and showed how we can obtain the medoid map according to this measure. Additionally, we showed experimentally that our framework can be used to find outlier (gerrymandered) maps based on their distance from the centroid map. Finally, we believe that there are further applications of having a distance measure over redistricting maps that are interesting to investigate such as using high dimensional visualization methods to gain further insight into the map ensemble.

[7] In Appendix B, we show the histogram for PA and MD as well.

[8] As mentioned before we get three centroids each from sampling a chain that starts with a different seed.

References

Abrishami, T.; Guillen, N.; Rule, P.; Schutzman, Z.; Solomon, J.; Weighill, T.; and Wu, S. 2020. Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, 42(5): A3340–A3366.

Alexeev, B.; and Mixon, D. G. 2018. An impossibility theorem for gerrymandering. *The American Mathematical Monthly*, 125: 878–884.

Ashtiani, H.; Kushagra, S.; and Ben-David, S. 2016. Clustering with same-cluster queries. *Advances in neural information processing systems*, 29.

Bickel, A. M. 1966. The Voting Rights Cases. *The Supreme Court Review*, 1966: 79–102.

Borg, I.; Groenen, P. J.; Mair, P.; Borg, I.; Groenen, P. J.; and Mair, P. 2013. Mds algorithms. *Applied Multidimensional Scaling*, 81–86.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of computational social choice*. Cambridge University Press.

Brubach, B.; Srinivasan, A.; and Zhao, S. 2020. Meddling metrics: the effects of measuring and constraining partisan gerrymandering on voter incentives. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 815–833.

Bycoffe, A.; Koeze, E.; Wasserman, D.; and Wolfe, J. 2018. The Atlas Of Redistricting. https://projects.fivethirtyeight.com/redistricting-maps/. [Online; published 25-January-2018; accessed 15-August-2019].

Chen, M. 2021. Tufts research lab aids states with redistricting process. *The Tufts Daily*.

Chhabra, A.; Roy, A.; and Mohapatra, P. 2020. Suspicion-free adversarial attacks on clustering algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3625–3632.

Chikina, M.; Frieze, A.; and Pegden, W. 2017. Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences*, 114(11): 2860–2864.

Cho, W. K. T. 2019. Technology-Enabled Coin Flips for Judging Partisan Gerrymandering. *Southern California law review*, 93.

Cinà, A. E.; Torcinovich, A.; and Pelillo, M. 2022. A black-box adversarial attack for poisoning clustering. *Pattern Recognition*, 122: 108306.

Cohen-Addad, V.; Klein, P. N.; and Young, N. E. 2018. Balanced centroidal power diagrams for redistricting. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 389–396.

Commission, M. C. R. 2021. Redistricting Map Submission Process. https://redistricting.maryland.gov/Pages/plan-proposals.aspx. [Online; accessed 20-October-2021].

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

DeFord, D.; Duchin, M.; and Solomon, J. 2019. Recombination: A family of Markov chains for redistricting. *arXiv preprint arXiv:1911.05725*.

Duchin, M.; and Walch, O. 2021. Political Geometry.

Fifield, B.; Higgins, M.; Imai, K.; and Tarr, A. 2015. A new automated redistricting simulator using markov chain monte carlo. *Work. Pap., Princeton Univ., Princeton, NJ*.

Goldschmidt, O.; and Hochbaum, D. S. 1994. A polynomial algorithm for the k-cut problem for fixed k. *Mathematics of operations research*, 19(1): 24–37.

He, Z.; Xu, X.; and Deng, S. 2003. Discovering cluster-based local outliers. *Pattern recognition letters*, 24(9-10): 1641–1650.

Herschlag, G.; Kang, H. S.; Luo, J.; Graves, C. V.; Bangia, S.; Ravier, R.; and Mattingly, J. C. 2020. Quantifying gerrymandering in north carolina. *Statistics and Public Policy*, 7(1): 30–38.

Hettle, C.; Zhu, S.; Gupta, S.; and Xie, Y. 2021. Balanced Districting on Grid Graphs with Provable Compactness and Contiguity. *arXiv preprint arXiv:2102.05028*.

Kemeny, J. G. 1959. Mathematics without numbers. *Daedalus*, 88(4): 577–591.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93.

Knox, E. M.; and Ng, R. T. 1998. Algorithms for mining distancebased outliers in large datasets. In *Proceedings of the international conference on very large data bases*, 392–403. Citeseer.

Ko, S.-H.; Taylor, E.; Agarwal, P.; and Munagala, K. 2022. All Politics is Local: Redistricting via Local Fairness. *Advances in Neural Information Processing Systems*, 35: 17443–17455.

Lin, J.; Chen, C.; Chmielewski, M.; Zaman, S.; and Fain, B. 2022. Auditing for gerrymandering by identifying disenfranchised individuals. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1125–1135.

Liu, Y. Y.; Cho, W. K. T.; and Wang, S. 2016. PEAR: a massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm and Evolutionary Computation*, 30: 78 – 92.

LWV vs Commonwealth of Pennsylvania. No. 159 MM 2018.

Mattingly, J. C.; and Vaughn, C. 2014. Redistricting and the Will of the People. *arXiv preprint arXiv:1410.8796*.

Mead, A. 1992. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1): 27–39.

Newling, J.; and Fleuret, F. 2017. A sub-quadratic exact medoid algorithm. In *Artificial Intelligence and Statistics*, 185–193. PMLR.

Pegden, W.; Procaccia, A. D.; and Yu, D. 2017. A partisan districting protocol with provably nonpartisan outcomes. *arXiv preprint arXiv:1710.08781*.

Polsby, D. D.; and Popper, R. D. 1991. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & Policy Review*, 9(2): 301–353.

Rucho v. Common Cause. No. 18-422, 588 U.S. 2019.

Ryan, I.; and Smith, W. D. 2022. Splitline districtings of all 50 states + DC + PR. https://rangevoting.org/SplitLR.html. [Online; accessed 15-August-2019].

Saran, H.; and Vazirani, V. V. 1995. Finding k cuts within twice the optimal. *SIAM Journal on Computing*, 24(1): 101–108.

Validi, H.; and Buchanan, A. 2022. Political districting to minimize cut edges. *Mathematical Programming Computation*, 1–50.

Vieth v. Jubelirer. No. 02-1580, 541 U.S. 267 (2004).

Wu, L. C.; Dou, J. X.; Sleator, D.; Frieze, A.; and Miller, D. 2015. Impartial redistricting: A markov chain approach. *arXiv preprint arXiv:1510.03247*.

Zubrzycki, S. 1972. *Lectures in probability theory and mathematical statistics*, volume 38. Elsevier Publishing Company.