A Graph-Theoretic Approach for Creating Non-Gerrymandered Congressional Voting Maps

Ethan Rebello
Department of Mathematics
University of North Texas
Denton, Texas, USA
EthanRebello@my.unt.edu

Zachary Li
Department of Mathematics
University of North Texas
Denton, Texas, USA
ZacharyLi@my.unt.edu

Jayanth Pandit
Department of Mathematics
University of North Texas
Denton, Texas, USA
JayanthPandit@my.unt.edu

Lior Fishman
Department of Mathematics
University of North Texas
Denton, Texas, USA
Lior.Fishman@unt.edu

2023 IEEE MIT Undergraduate Research Technology Conference (URTC) | 979-8-3503-0965-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/URTC60662.2023.10534963

***Abstract*—Gerrymandering is one of the biggest problems in American democracy. Political parties use redistricting as a method of maintaining political control. A mathematical solution to political gerrymandering could create a more vibrant, fair democracy. We propose three algorithms - Ideal Partition, Circular Engulfment, and LRPG - that use mathematical methods to create non-gerrymandered congressional maps with substantially less bias than current U.S. House of Representatives districts. Each method is a trade-off between speed, contiguity, and result variance. Ideal Partition creates districts that align with city divisions but with high time complexity. Circular Engulfment is quicker, but it has continuity issues and is better suited for modeling than for true district generation. LRPG is the fastest but only generates one set of districts. Our work showcases a novel mathematical model for non-gerrymandered congressional maps that could significantly improve the quality of democracy.**

***Index Terms*—gerrymander, precinct, district, continuity, underrepresentation**

I. Introduction

A. A History of Redistricting

Redistricting in the United States has been significantly influenced by a few key court cases, such as Baker v. Carr [1], Wesberry v. Sanders [2], and Reynolds v. Sims [3]. These landmark decisions allow judges to intervene and redraw district boundaries to ensure equal representation, By ensuring equal district population. These cases laid the foundation for fair and equitable districting.

Despite the progress made in addressing malapportionment, partisan gerrymandering remains a significant concern. In Rucho v. Common Cause [4], The Supreme Court ruled partisan gerrymandering claims non-justiciable. As a result, political parties often exploit redistricting to their advantage, creating districts that heavily favor one party. This undermines democratic principles of fair representation. To counteract the influence of political interests in redistricting, there is an increasing need for the development and application of mathematical algorithms. This is to ensure fair and unbiased districting that better reflects voters' preferences and diversity.

It remains, however, necessary to develop robust mathematical algorithms for addressing racial vote dilution and promoting fair representation.

979-8-3503-0965-2/23/$31.00 ©2023 IEEE

B. A Redistricting Dilemma

The fundamental problem with redistricting, and thus generating a non-gerrymandered voting map is simple. There is no single metric by which to measure gerrymandering, and even if you build separate metrics that represent vote power [5], geographic compactness [6], and election win margins, it is mathematically impossible to simultaneously optimize all of these metrics [7] with any one redistricting algorithm [8].

Politically, this represents the inability to decide between two major strategies in the world of districting: creating microcosms of a political region that allow for each district to be representative of that region, promoting overall political representation where votes match representatives, or creating homogeneous districts that contain similar people from similar voting blocs, which promotes unity and individual representation. Nurturing diversity makes individual voters feel like they have an impact, and nurturing uniformity makes communities feel that they have an impact.

Of course, no one can speak authoritatively as to which of these approaches is better, but when designing pseudo-random algorithms, there can be an underlying assumption that promotes one strategy over the other.

We seek to promote randomness and avoid a bias toward either method whenever possible. Rather than making an authoritative statement about either approach, we simply seek to facilitate our created algorithms on real maps, to analyze the natural tendencies of given geographic regions. This allows us to adopt a probabilistic view of districting, and determine the likelihood of political outcomes to recognize gerrymandering.

C. Past Studies

In the past, studies have explored graph theoretic approaches to combat gerrymandering. One such work by Liu et al. [9] involves a parallel computation approach to address the combinatoric optimization problem, effectively determining the optimal arrangement of geographical subdivision per federal guidelines, inherent in political redistricting. Another work by Heschlag et al. [10] adopted a Monte Carlo Markov algorithm that generates a series of potential district maps by iteratively adjusting boundaries while respecting geographical and demographic constraints. Various studies by DeFord et al.

[11] introduced a new set of Markov Chains called ReCom (recombination) to facilitate redistricting.

Building upon existing research, this work stands as the first to utilize existing precinct data to create theoretical electrical district maps. This means that districts are based on existing election infrastructure (voting locations, district offices, etc.) and no new arbitrary lines are drawn. Additionally, our study is the first to use undirected map representations in the Ideal Partition and LRPG methods (Circular Engulfment utilizes geographical location data). This allows us to effectively represent district adjacency and create graph theoretical algorithms to group adjacent precincts.

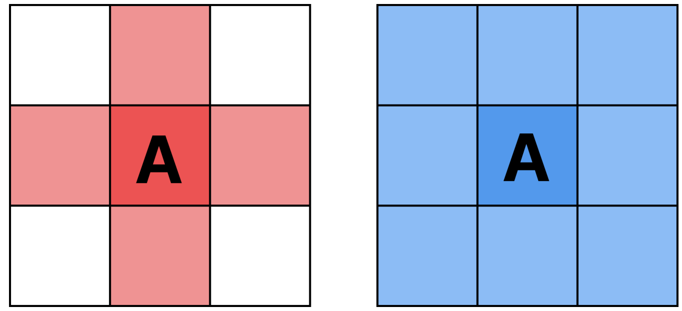

Fig. 1. On the left, we have finite-length adjacency, while on the right, we observe corner adjacency.

II. Data Sourcing and Preprocessing

A. Data Sources

All of our data come from Redistricting Data Hub and Districtr, both tools created by MGGG (Metric Geometry and Gerrymandering Group). We used Redistricting Data Hub [12] for all voting data and to generate simulated elections, as they provide shapefile representations of voting precincts that contain a tabular representation of election results and counts at each polling location in every U.S. state and territory.

We used Districtr [13] population data shapefile that mapped between U.S. Census blocks and the geographical boundaries of those blocks, for each state. Each census block contains both total population numbers, and also a breakdown of the ethnic and age demographics of the residents within each block.

B. Population and Voting Precincts

There are two types of precincts in the data that we reference and process. The first type are population precincts, which are derived from census blocks and contain information about the demographic makeup of a given parcel of land. These precincts may be any size and typically align with neighborhoods or other features, and typically have similar populations with the preponderance being in urban areas. The second type of precincts are voting precincts, which directly correlate with polling locations and vote counting centers in a given geographical area, and these tally the number of votes for each party in a given region. Voting precincts are typically equal in size across a state and vastly outnumber population precincts (e.g. in Georgia, where we had three thousand population precincts, but sixty thousand voting precincts).

For maximal accuracy, our algorithms use a combination of both voting and population precincts. We match the smaller voting precincts to their larger containing population precincts using GeoPandas and Shapely in Python, and break ties where voting precincts fall into two population precincts by the larger geographical overlap.

Whenever we discuss precincts, this refers to their preprocessed and merged forms unless explicitly stated otherwise.

C. Adjacency Calculations

Once the precincts are constructed, the next step involves recalculating their adjacencies to represent the precinct structure as a graph. We adhere to laws specific to each state regarding congressional district contiguity and implemented that in our adjacency calculations. Generally, we utilized rook or finite-length adjacency, rather than corner adjacency, as depicted in Fig. 1. Additionally, for handling islands, we employed the water adjacency method.

Using GeoPandas and Shapely in Python, we look for overlapping borders in between precincts and represent those connections in an edge list, which is then processed by NetworkX into a connected unweighted graph.

The computational aspect of this process presents a challenge due to its time-consuming nature. To address this, we have incorporated several heuristics aimed at excluding unlikely intersections from the calculations. However, in the case of large states such as California, which contains over fifty-five thousand precincts, the calculation can still be resource-intensive in terms of both time and RAM usage.

III. Ideal Partition

Ideal Partition is a graph-based approach that focuses on randomly partitioning distinct chunks of the graph until we find a desirable split. These partitions are then preserved as individual districts once their population totals are optimized to meet the desired criteria.

Random partitions are generated by selecting 2 precincts at random and then grouping all other precincts based on their graph distance to each of the selected precincts, as specified in (1). Here, $\text{SNPD}(a, b)$ represents the Shortest Node Path Distance on the adjacency graph G between any two precincts or nodes, denoted as a and b. SNPD is also depicted in Fig. 2.

$$\text{Group}(p) = \begin{cases} a & \text{if SNPD}(a,p) < \text{SNPD}(b,p) \\ b & \text{if SNPD}(a,p) > \text{SNPD}(b,p) \\ 0 & \text{if SNPD}(a,p) = \text{SNPD}(b,p) \end{cases} \quad (1)$$

Once the assignment of each precinct to a group is determined based on the value of (1), we proceed to evaluate whether either group a or b can correctly form the desired total when group 0 is taken into account. If either group a or b can meet the required total population with the inclusion of precincts from group 0, we select that group and proceed with the split. Subsequently, precincts from group 0 are added to

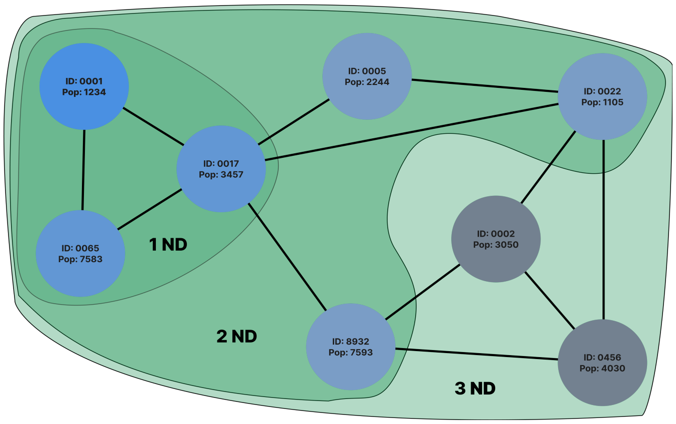

Fig. 2. A visual depiction of Shortest Node Path Distance on a sample graph

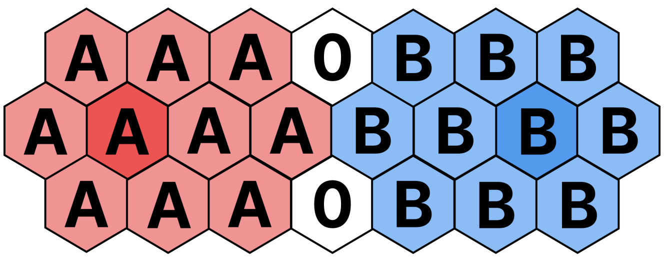

Fig. 3. A visual depiction of Ideal Partition's grouping formula (1)

the selected group to achieve the optimal population total for the final districts.

List 1: Algorithm Steps - Ideal Partition

1) Select 2 random precincts, a and b.
2) Group every other precinct by (1).
3) If the desired total is possible when combining either group a or b and precincts from group 0, save that split.
4) Repeat until $n-1$ substates or districts are formed.
5) If using substates, repeat the whole process until the desired district size is formed.

This method, described in List 1 and pictured in Fig. 3, often needs to be iterated to execute in a reasonable time. For Georgia, which has 14 districts, it was the fastest to split the state into two halves, which we call substates, and then into sevenths to get the resulting districts, but any split, even uneven splits (e.g. a 60-40% split of the population) can be used if faster.

The primary parameters in this method include the set of ratios for the desired splits S, the minimum SNPD between a and b denoted as α, and the maximum SNPD between a and b denoted as β. The parameter α serves to prevent a and b from being excessively close together, thus avoiding oversampling points from within a concentrated urban center where most precincts are located. On the other hand, β ensures that a and b are not too far apart from each other, as this might result in nearly all other precincts falling into group 0.

When tuned properly, α and β create districts that both group and split urban areas, forming compact districts.

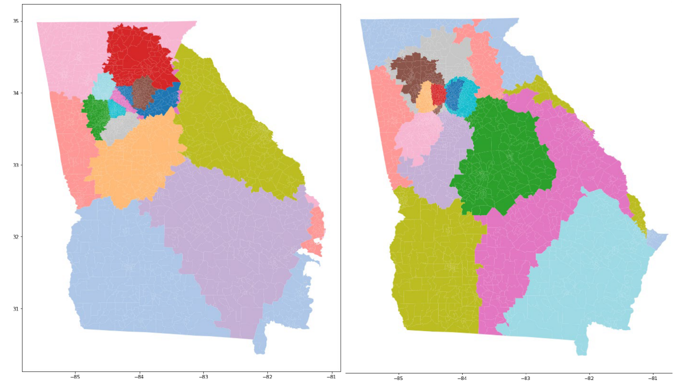

Fig. 4. Three examples of maps generated by Ideal Partition. Each color is one district

Generated regions tend to be densely populated and often extend from urban areas, as depicted in Fig. 4. However, the method exhibits significant edge salience, resulting in small rough spots on the borders between districts due to random precinct selection. Addressing these discrepancies and smoothing out the edges would be necessary before implementing the map in the real world.

IV. Circular Engulfment

Circular engulfment is a geographical method solely reliant on the gradual expansion of spheres to form districts until the desired population for each district is achieved. Unlike other methods, it does not require any adjacency preprocessing. Due to its purely geometric nature, this approach lacks geographical accuracy and often results in the creation of disconnected and non-contiguous districts.

This method is particularly well-suited for data collection and statistical analysis. While it may not be the best option for generating practical district maps, it serves as a valuable tool for measuring the distribution of ethnicity in districts and understanding the implications of population distribution on redistricting.

The method randomly selects n random precincts as starting points called origin precincts, and each of those precincts' spheres of influence, centered on their centroids, is gradually expanded until that sphere of influence has enough population to form a district. This process is described in List 2 and Fig. 5.

List 2: Algorithm Steps - Circular Engulfment

1) Select n random origin precincts.
2) Draw a circle centered at the centroid of each origin precinct.
3) Expand the circle by α and add the population of the subsumed precincts to a running total.
4) Once the running total reaches the required population for a district, save that district.
5) Continue the process until $n-1$ districts have the necessary population.
6) Assign all remaining area to the last district.

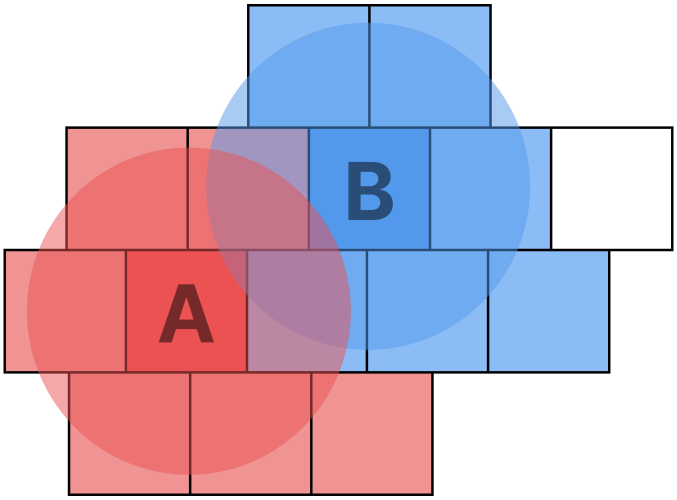

Fig. 5. A visual depiction of Circular Engulfment's Spheres of Influence

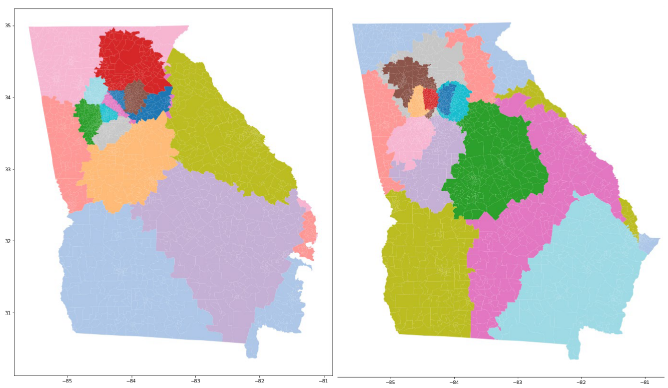

Fig. 6. Two examples of maps generated by Circular Engulfment. Each color is one district

The primary parameters in the circular engulfment method are α and β. The value of α represents the growth rate of the spheres, influencing how quickly the districts are formed. Higher values of α result in faster execution of the program but may introduce more severe geographical features and lead to districts with less equal population distribution.

On the other hand, the parameter β determines the minimum geographical distance between two randomly selected districts. Increasing β helps to prevent oversampling points from cities where a majority of precincts are concentrated. However, this may also push origin precincts to the periphery of geographic regions, possibly leading to stratification.

Most regions are circular and densely pack urban voters into small districts whereas rural voters are taken from all over the state and placed into geographically large districts that engulf urban districts, behavior depicted in the generated maps shown in Fig. 6.

V. Least Required Population Grouping

Least Required Population Grouping, or LRPG, is a joint geographic and graph method, that creates districts by grouping precincts based on their distance from the origin precinct.

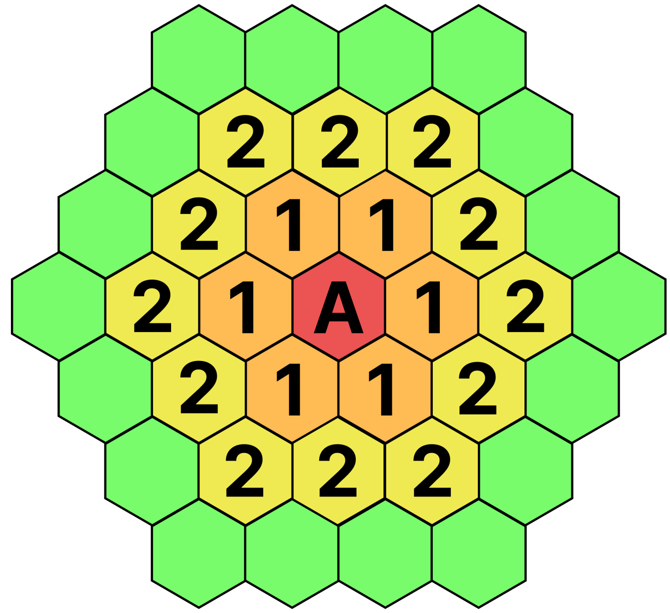

Fig. 7. A visual depiction of LRPG's Joint Distance List Creation

This method creates compact districts with little edge salience, but also very little variance in overall results.

The main mechanism of LRPG revolves around the Joint Distance List. Initially, a random origin node is selected, and all other nodes are ranked based on their SNPD away from the origin node. Within each group of equal SNPD, the nodes are further ranked by their geographical distance between the centroids of the two shapes.

This Joint Distance List is then iterated down as described in List 3 and Fig. 7, and the population of each precinct is added to a running total. Once this total reaches the required population for a district, the corresponding precincts are set aside, and the process is repeated until all n districts are formed.

List 3: Algorithm Steps - LRPG

1) Pick a random origin precinct.
2) Order all nodes first by Shortest Node Path Distance (SNPD), and then by geographic distance between the two centroids.
3) Iterate down the list, keeping a running total of the population and adding precincts one by one.
4) Once the needed population is reached, save the district and start the next one at the next list item.
5) Continue until $n-1$ districts are formed.
6) Save all remaining list items to the nth district.

The primary parameter in LRPG is r, representing the radius from the centroid of the entire geographic region within which an origin precinct can be chosen. The main purpose of r is to prevent stratification that can occur when the chosen origin precinct is situated too close to the edge of the region. When this happens, the equal Shortest Node Path Distance (SNPD) zones emanate as concentric rings away from the

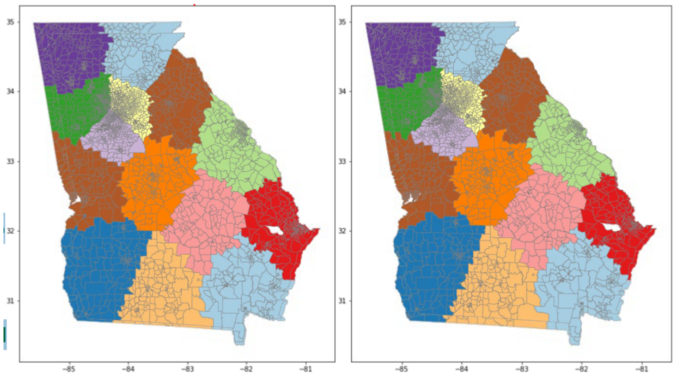

Fig. 8. Two examples of maps generated by LRPG. Each color is one district

origin precinct, leading to the creation of thin and elongated districts.

Increasing r may elevate the risk of stratification, but it also introduces more variability in the generated maps. Safer, smaller values of r result in low variance between maps and stable results.

Generated districts are compact and are typically regular or circular, but they are drawn somewhat irreverently to city lines and demarcations, thoroughly mixing urban and rural precincts, as depicted in the generated maps in Fig. 8.

VI. Discussion

A. Compactness Metrics

One desirable attribute of congressional districts is compactness, as it serves as a safeguard against the egregious use of cracking or packing techniques. To assess compactness, the Polsby-Popper test [6] is widely employed, and it can be calculated for a given district D using (2), where $\mathrm{A_D}$ is the area of a district, and $\mathrm{P_D}$ is the perimeter of a district.

$$\mathrm{PP(D)} = \frac{4\pi \times \mathrm{A_D}}{\mathrm{P_D^2}} \in [0, 1] \tag{2}$$

We saw a significant increase in the Polsby-Popper scores (2) for our generated districts compared to current districts.

B. Representation Metrics

In evaluating representation, we utilize a modified version of the Weighted Efficiency Gap formula [8] proposed by Kristopher Tapp. Let V be the total number of voters, V_p be the number of voters of a given party p, R be the total number of representatives in a given region, and R_p be the number of elected representatives for a given party p, then our formula for underrepresentation can be represented as (3).

$$\mathrm{UR} = \sum_{p}^{parties} \frac{V_p}{V} \times |\frac{V_p}{V} - \frac{R_p}{R}| \tag{3}$$

Equation (3) represents the weighted sum of the discrepancy between the percentage of voters affiliated with a specific party and the percentage of seats occupied by that party's representatives in comparison to the total number of available seats.

In many instances, particularly in smaller states or regions, the formula (3) may simplify due to the limited availability of candidates.

Circular Engulfment was executed for all 45 states with over 1 representative, and the level of underrepresentation was measured. Subsequently, a comparison was drawn between the results obtained and the existing maps. Our analysis revealed a significant decrease of 17% in underrepresentation when using the simplest algorithm. Moreover, preliminary testing indicates that Ideal Partition and LRPG exhibit even more promising outcomes.

VII. Future Work

Our future endeavors encompass two tasks. Firstly, we aim to devise an algorithmic approach for smoothing districts at the county level. By leveraging county-level data whenever possible, we seek to ensure that individuals within the same county are represented by the same representatives.

The second task on our agenda revolves around automating the Ideal Partition split method. Currently, this method relies on human decision-making and produces equal fractions of the whole. However, we envision a fully autonomous process, streamlining the Ideal Partition algorithm through automation, and we seek to enhance its efficiency.

References

[1] U. S. Court, "Baker v. carr," 369 U.S. 186, 1962, case on legislative apportionment and justiciability.

[2] ——, "Wesberry v. sanders," 376 U.S. 1, 1964, case on equal representation for congressional districts.

[3] ——, "Reynolds v. sims," 377 U.S. 533, 1964, case on apportionment of state legislative districts.

[4] ——, "Rucho v. common cause," 588 U.S. 557, 2019, case on partisan gerrymandering and justiciability.

[5] J. Kalcsics and R. Z. Ríos-Mercado, *Districting Problems*. Cham: Springer International Publishing, 2019, pp. 705–743. [Online]. Available: https://doi.org/10.1007/978-3-030-32177-2_25

[6] D. D. Polsby and R. Popper, "The third criterion: Compactness as a procedural safeguard against partisan gerrymandering," *Yale Law & Policy Review*, vol. 9, no. 2, 1991. [Online]. Available: https://ssrn.com/abstract=2936284

[7] B. Alexeev and D. G. Mixon, "An impossibility theorem for gerrymandering," *The American Mathematical Monthly*, vol. 125, no. 10, pp. 878–884, 2018. [Online]. Available: https://doi.org/10.1080/00029890.2018.1517571

[8] K. Tapp, "Measuring political gerrymandering," *The American Mathematical Monthly*, vol. 126, no. 7, pp. 593–609, 2019. [Online]. Available: https://doi.org/10.1080/00029890.2019.1609324

[9] Y. Y. Liu, W. K. T. Cho, and S. Wang, "Pear: a massively parallel evolutionary computation approach for political redistricting optimization and analysis," *Swarm and Evolutionary Computation*, vol. 30, pp. 78–92, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210650216300220

[10] G. Herschlag, H. S. Kang, J. Luo, C. V. Graves, S. Bangia, R. Ravier, and J. C. Mattingly, "Quantifying gerrymandering in north carolina," *Statistics and Public Policy*, vol. 7, no. 1, pp. 30–38, 2020. [Online]. Available: https://doi.org/10.1080/2330443X.2020.1796400

[11] D. DeFord, M. Duchin, and J. Solomon, "Recombination: A Family of Markov Chains for Redistricting," *Harvard Data Science Review*, vol. 3, no. 1, mar 31 2021, https://hdsr.mitpress.mit.edu/pub/1ds8ptxu.

[12] "Metric geometry and gerrymandering group," Website, MGGG Redistricting Lab, 2023, https://mggg.org.

[13] "Districtr," Website, MGGG Redistricting Lab, 2023, https://mggg.org/districtr/.