Escuela de Gobierno y
Transformación Pública
Tecnológico de Monterrey

Working papers are preliminary documents that have not been peer-reviewed. They should not be considered conclusive or disseminated as scientifically validated information.

Using Dual Graphs to Model Territorial Communities of Interest

James Syme
jsyme@tec.mx
School of Government and Public Transformation
Tecnológico de Monterrey

School of Government and Public Transformation

Working Paper No. 6

DOI: https://doi.org/10.21203/rs.3.rs-7040797/v1

Publication date: july, 2025

Using Dual Graphs to Model Territorial Communities of Interest

James Syme, Escuela de Gobierno y Transformación Publica
Tecnologico de Monterrey

July 17, 2025

Abstract

Communities of interest are foundational to democratic representation in territorial constituencies yet are often broadly- or un-defined. The ambiguity of definition and process of community identification leaves practitioners, legal experts, and academics with relatively little clear empirical evidence from which to draw reliable conclusions. This reveals a problem with governance in territorial districting, where a foundational theoretical concept–the community as the centerpiece of representation–collides with practical limitations, including the characterization and identification of a community. To help mitigate and explore this problem, I introduce a graph-based model of territorial communities based on assumptions derived from available statutory definitions and use Census data to explore how communities can be represented, though the method is broadly extendable to a range of empirical characterizations of interest. This novel approach to community identification builds on existing graph-based methods for computational redistricting to facilitate new theoretical research into the interactions between communities and representation while providing a new tool to support practitioners' needs in identifying communities.

Introduction

Thirty-four U.S. states require the preservation of communities of interest in federal and/or state legislative districts [Chen et al., 2022]. The theory underlying this criterion is that the preservation of a community within a legislative district gives that community undivided political power, allowing members to organize around salient issues and ideas, communicate concerns and preferences with political representatives, and hold those representatives accountable. However, the identification of a community of interest–which is generally defined as a territorially contiguous community of people with distinct, shared interests worthy of representation in a legislature–is challenging. Communities are subject to a range of competing perceptions about their boundaries from those who live within the community, if those members even hold perceptions on the topic. The ambiguity in most definitions provides constituents and citizens wide latitude to define communities of interest, obfuscating the landscape for those who are responsible for drawing and adopting legislative districts. Compounding the effect, districting officials are tasked with sifting through what information is available on tight budgets and timelines.

The inclusion of communities of interest as a legislative criteria is likely the consequence of conceptualizations of representation in the early United States, which could be traced back to Britain and even feudal systems [Guinier, 1992]. When the House of Commons was established in England, the basic unit of representation was the town, county, or shire [Stephanopoulos, 2012]. Before the Revolutionary War, American colonies maintained the English system of community-based representation; in New England, townships were the basis for representation in local legislatures, while counties and parishes were often the basis in the middle and southern colonies [Zagarri, 1987]. By the mid-19th century, states gaining new admission to the Union largely switched to using districts as the basis of representation rather than towns, though the administrative community continued to influence criteria as many states adopted provisions to ensure that districts did not cut across political subdivisions such as cities or countries [Stephanopoulos, 2012]. Unsurprisingly, this led to many states malapportioning federal representation. Cox and Katz [2002] found that "[i]n the 88th Congress (elected in 1962), 234 congressional districts deviated by at least 10% from the average district population in their respective states, with the maximum deviation being 118% [Cox and Katz, 2002]" . Apportionment for state legislatures were perhaps even worse. A particularly egregious 1962 Alabama districting plan partitioned the state so that the "district in Jefferson County, which is near Birmingham, contained 41 times as many eligible voters as those in another district of the state. [Oyez, 2022b]".

A series of Supreme Court cases–*Baker v. Carr* (1962), *Wesberry v. Sanders* (1964), and *Reynolds v. Simms* (1964), collectively referred to as the *reapportionment revolution*–subsequently transformed legislative redistricting in the United States at both the state and federal levels. At the federal level, the Court in *Wesberry v. Sanders* standardized the "as nearly as practicable" population requirement for congressional districts, meaning that states have very little leeway in differences in population between federal legislative districts–a deviation of over 1% from the ideal population could be justiciable [Stephanopoulos, 2013]. In contrast, the Court found in *Reynolds v. Simms* (377 U.S. 533, 1964) that "mechanical exactness is not required" [Oyez, 2022b] when drawing districts at the state level.

While these cases rightly put an end to egregious malapportionment within states, they altered the statutory role of the community in the constituency. Administrative bodies–such as towns, cities, or counties–were no longer the basis of districts and became secondary considerations instead, where many states strive to preserve them within districts where possible. The preservation of communities of interest emerged as a criterion to help maintain the important role communities play in representation, though the strict equal population requirement often requires the division of communities of interest [Guinier, 1992, Webster, 2013, Gimpel and Harbridge-Yong, 2020]. Indeed, partisan gerrymandering is often facilitated by splitting communities under the guise of this rule in a practice aptly referred to as "cracking". However, problems with community identification make it difficult to fully analyze and characterize the extent of community splitting.

Practical Problems with Identification

Map drawers face a number of practical challenges facing in identifying community boundaries to preserve. First, self-identification–generally in the form of public input sessions or map and descriptive text submissions sent directly to map drawing authorities–is the primary means for identifying communities across state and local governments. While this wisely accounts for what McCartan et al. [2023] call the "subjective" nature of the community, it also brings selection bias. For example, in a retrospective study, Rambo [2020] found that well-organized citizens were far more likely to use the term "communities of interest" in public feedback than citizens from the general public, including groups advocating for communities not traditionally associated with districting protections. A randomized sample of 967 comments (around 13.5% of total comments made from January 1, 2011 to August 15, 2011 to the CRCC) showed no feedback from large swaths of Los Angeles County.

Where information is submitted, input is often conflicting [Rambo, 2020]. Individuals and groups that submit perceived boundaries are–without any outside coordination–likely to disagree on final boundaries. Figure 1 shows 6 different submissions to the 2020 CRC [We Draw the Lines CA, 2022] delineating Boyle Heights, a predominantly Hispanic and Latino neighborhood in Los Angeles, CA. Note that the general area is the same, but the specific boundaries vary significantly.

Figure 1: Six submissions to the 2020 California Citizen's Redistricting Commission delineating Boyle Heights as a COI. Composite generated from public submissions to the CRC [We Draw the Lines CA, 2022].

Furthermore, state definitions of communities of interest–some example of which are presented below in Section –are often broad or vague. Indeed, Kim and Chen [2021] note that "the term [communities of interest] is so open-ended that it can be used to justify abusive practices such as partisan advantage or incumbent protection after the fact. [Kim and Chen, 2021]" Furthermore, the ambiguity in state definitions can reward well-organized groups over groups that reflect more traditional conceptualizations of a community of interest. For example, Rambo [2020] found that environmental advocates and suburban residents were able to effectively leverage this ambiguity in the 2011 California redistricting cycle into preservation of communities.

Finally, redistricting occurs under a number of time and budgetary constraints, which combine with a flood of information to create a frantic environment that can overwhelm map drawers. In 2011, the California Citizens Redistricting Commission (CRC)–the fourteen-person independent commission with partisan balance that draws and adopts state and federal legislative districts in California–received "written submissions from more than 2,000 organizations and 20,000 individuals. [Sonenshein, 2013]" The 2020 CRC website shows over 35,000 public submissions [We Draw the Lines CA, 2022].

The combination leaves map drawers–in addition to researchers, NGOs, and community advocates–with little clear information available to identify the boundaries of communities of interest. Returning to California, public feedback to the California Citizen's Redistricting Commission (CRC) from the 2021 redistricting cycle show 1,820 COIs PDFs and/or GIS shapefiles (around 5.14% of public submissions shown) submitted to the CRC [We Draw the Lines CA, 2022], including those submitted using `https://www.drawmycacommunity.org`[1]. These data on community boundaries were used to draw the 52 Congressional, 40 State Senate, 80 State Assembly districts, and over 39 million people in California. These conditions can lead to the adoptions of communities of interest that are reflective of the loudest voice in the room rather than the "true" community.

Systemic Community Identification can Mitigate Problems

Systematic evidence for the boundaries of communities of interest can mitigate these practical problems by restricting the size of the community of interest search space and highlighting communities that are otherwise not visible in public submissions while simultaneously and providing new evidence for academic and theoretical investigation. Additionally, systematic evidence is reproducible and can be made transparent, which can increase confidence in the approach used to delineate the boundaries. This additional information, when presented in an accessible format, can help compensate for a lack of assessment capacity, reduce uncertainty of boundaries due to conflicting public perceptions and a near-infinite uncertainty space outside of feedback, and mitigate the impact of low resources for self-identification.

Furthermore, map-drawing algorithms play an increasing role in legal challenges to districting plans (e.g., see *Rucho v. Common Cause* [Oyez, 2022a], *Harkenrider v. Hochul* [New York Court of Appeals, 2022], and *South Carolina State Conference of the NAACP vs. Alexander* [The American Redistricting Project, 2023]). Indeed, map-drawing algorithms have even been referred to as the "gold standard in partisan gerrymandering cases. [Chen and Stephanopoulos, 2021]" These algorithms–which can also be used to evaluate tradeoffs in representational outcomes due to criteria and in gerrymandering research–could better reflect redistricting criteria in many states by explicitly incorporating communities of interest or rejecting maps that perform too much splitting. Systematic identification of communities of interest can support improvements in legislative map-drawing algorithms by providing a robust means to integrate community preservation. Systematic evidence for communities could also better facilitate theoretical investigations into the relationship between community preservation and representational outcomes of interest. For example, Gimpel and Harbridge-Yong [2020] investigated the relationship between community preservation and competitiveness in Pennsylvania and North Carolina, relying on "potential communities of interest" derived from economic hot spots and regional identifiers. Computational communities would be well-suited to support these kinds of analyses.

There are a number of projects that aim to elicit community of interest boundaries from the public writ large and aggregate them in databases (e.g., Redistricting Data Hub [2024], rep [2023]). For example, The website Representable (`https://www.representable.org`) is an open-source collection project that allows any user to draw community boundaries for their community, but this project also faces a small sample size; each state only has a small number communities when compared to the population at large (on the order of a tens to a few hundred). McCartan et al. [2023] have effectively used surveys in select cities to identify communities of interest using kernel density estimators, though this reflects a highly specialized approach that may not be generally feasible under the timelines or budgetary environment in which redistricting is conducted.

It is important to emphasize that empirical and computational methods should not supplant public input [Mac Donald and Cain, 2013]. Instead, they should supplement the tightly-constrained process of COI identification [Sonenshein, 2013]. Providing systematic evidence about COIs reduces uncertainty by (1) filling in information gaps on the boundaries of COIs when self-identification is absent, (2) providing a sort-of validation test for single COI submissions, and (3) helping to sort through conflicting information when multiple COIs are present. Providing information and implementation capacity in an accessible format–such as an online tool and a script repository–helps provide policymakers and supporting analysts with the tools to actually use the method.

[1]The vast majority of public feedback is classified by the CRC as "COI Input", though many of these are textual descriptions of community interests themselves and suggestions for other communities that should be included in the same district rather than actionable boundaries.

A Systemic Approach to Identify Communities of Interest

While a clear need for a systematic approach to community identification exists, there is no single, authoritative definition from which an approach can be derived. Morrill [1987] defines communities of interest as "territories within which people share a common sense of identity and value across a range of interests. [Morrill, 1987]" Statutory and legal definitions vary significantly between states. For example, the Supreme Court of North Carolina declared in 2002 that "communities of interest should be considered in the formation of compact and contiguous electoral districts [Supreme Court of North Carolina, 2002, 562 S.E.2d 377]" without providing a definition. Further north, the West Virginia state code requires that communities of interest are taken into account when drawing state senatorial districts but provides no definition: "[t]he Legislature finds and declares that... in dividing the state into senatorial districts... [it has] taken into account in crossing county lines, to the extent feasible, the community of interests of the people involved. (West Virginia §01-2-1(c)(5))"

Other states attempt to define communities of interest in terms of what interests may be considered, though these definitions are varied and somewhat vague. Consider the following examples:

- **Alabama** "(a) A community of interest is a defined area of the state that may be characterized by, among other commonalities, shared economic interests, geographic features, transportation infrastructure, broadcast and print media, educational institutions, and historical or cultural factors. (b) The discernment, weighing, and balancing of the varied factors that contribute to communities of interest is an intensely political process best carried out by elected representatives of the people. (Ala. Code 1975 §17-14-70.1) "[2]
- **California** "[A community of interest is] a contiguous population which shares common social and economic interests that should be included relatively a single district for purposes of its effective and fair representation. (Art. XXI, §2)"
- **Colorado** "[C]ommunities of interest, including ethnic, cultural, economic, trade area, geographic, and demographic factors, shall be preserved within a single district wherever possible. (Art.V §44(3)(b)"
- **Montana** "Communities of interest can be based on Indian reservations, urban interests, suburban interests, rural interests, tribal interests, neighborhoods, trade areas, geographic location, demographics, communication and transportation networks, social, cultural, historic, and economic interests and connections, or occupations and lifestyles. [Montana Districting and Apportionment Commission, 2021]"
- **New Mexico** "[A community of interest is] a contiguous population that shares common economic, social or cultural interests. (NMSA 1978, §79-3-B (2021))"

There are two clear common threads that can be used to identify modeling principles to which a systematic approach should adhere. First, communities of interest should reflect territorial or geospatial **contiguity**. Second, interest is often generally defined to include an array of social and/or economic interests. Given the relative ambiguity of interest, any approach to community identification should reflect the principle of **interest elasticity**–i.e., it should be versatile and adaptable to a range of different of empirical characteristics of socioeconomic interest, including substantive policy preferences, demographics, political boundaries, and even person-level interactions.

However, some key normative components of community formation are missing from statutory definitions and the literature. While communities are contiguous, there is little to no discussion of how those contiguous boundaries should be set in the presence of interest gradients. I propose a principle of **local equilibrium**–if a community is identified along a single substantive or symbolic dimension (such as race or policy preference), it should not be possible to extend its boundaries and preserve that interest as the community's discerning feature. Additionally, **completeness**–the assumption that all citizens live within a community of interest–is untreated in the literature. Presumably all citizens can be said to live in *some* community that has interests associated with it. In the early United States, some instances of corporate representation implicitly assumed

[2] Note: this definition was revised in 2023 (Alabama SB5) to remove the references to "ethnic", "racial", and "tribal" (see Legislature [2021]).

this to be true, as counties, parishes, or townships, which can practically partition states, were the basis of representation ([Zagarri, 1987]). Finally, **uniqueness** specifies that communities are mutually exclusive under a single parameterization. The key normative implication of this assumption is that citizens, under a single dimension of interest, cannot live in multiple communities of interest at the same time. This is both reasonable–a spatial model of preferences would disallow individuals from simultaneously occupying multiple points in a spectrum–and reflective of practice: individuals that submit maps of their communities of interest only submit one.

A Graph-Based Method to Identify Communities of Interest

Graphs, which are collections of vertices V and edges E that connect those vertices referred to as $G = (V, E)$, are particularly well-suited for identifying communities of interest in the context of the descriptive and prescriptive assumptions identified above for two reasons. First, there is a robust literature dedicated to analyzing and identifying community structure on graph structures (e.g., see [Newman, 2006a, Fortunato, 2010, Traag et al., 2011, Riolo and Newman, 2020], etc.). Second, geospatial partitions can easily be transformed into graph representation. Under the method described herein, a legislative unit, such as a state, is treated as a tessellation (or tiling), which induces a dual graph on the tiles. Then, some metric of likeness or strength of connection is assigned to edges along that border each other, and the strengths of connection between adjacent units are used to group units together using a community detection algorithm. Finally, grouped elements are combined to identify communities of interest relevant to legislative redistricting.

These four steps are expanded upon below.

Step 1: Define a Tessellation

Let $S \subset \mathbb{R}^2$ be a set representing the territorial extent of a state, and let $T = \{T_i\}_{i \in I}$ be a tessellation of the state T (so that $\bigcup_{i \in I} T_i = S$). Following the standard approach to drawing districts, where Census blocks or precincts are grouped to form districts, the tessellation forms the basis of new communities of interest. Therefore, the tessellation should be chosen so that each tile T_i is a cohesive entity worthy of community representation–i.e., the tessellation should be of fairly high resolution. Counties and parishes, for example, form tessellations of states yet are too large to constitute a community in and of themself. Furthermore, each element of the tessellation should be associable with both population counts and population characteristics relevant to community formation and legislative interests, and the population variance of the tiling should be acceptable.

The U.S. Census Bureau maintains hierarchical tessellations of each U.S. state for the purposes of counting and estimating populations and households every 10 years that largely satisfy these needs. The three levels are blocks, block groups, and tracts (groups of block groups), and these units never cross county boundaries. Census blocks are the lowest-level tiles for which population counts are available in the decennial Census. However, the Census includes other important characteristics relevant to community formation when drawing these tessellations. As the Census notes, blocks are "statistical areas bounded by visible features, such as streets, roads, streams, and railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county limits and short line-of-sight extensions of streets and roads. [United States Census Bureau, 2022]" Figure 2 demonstrates the key difference between blocks (panel B) and precincts (which also act as a tessellation, panel A) in an area of Ohio centered near Painesville.

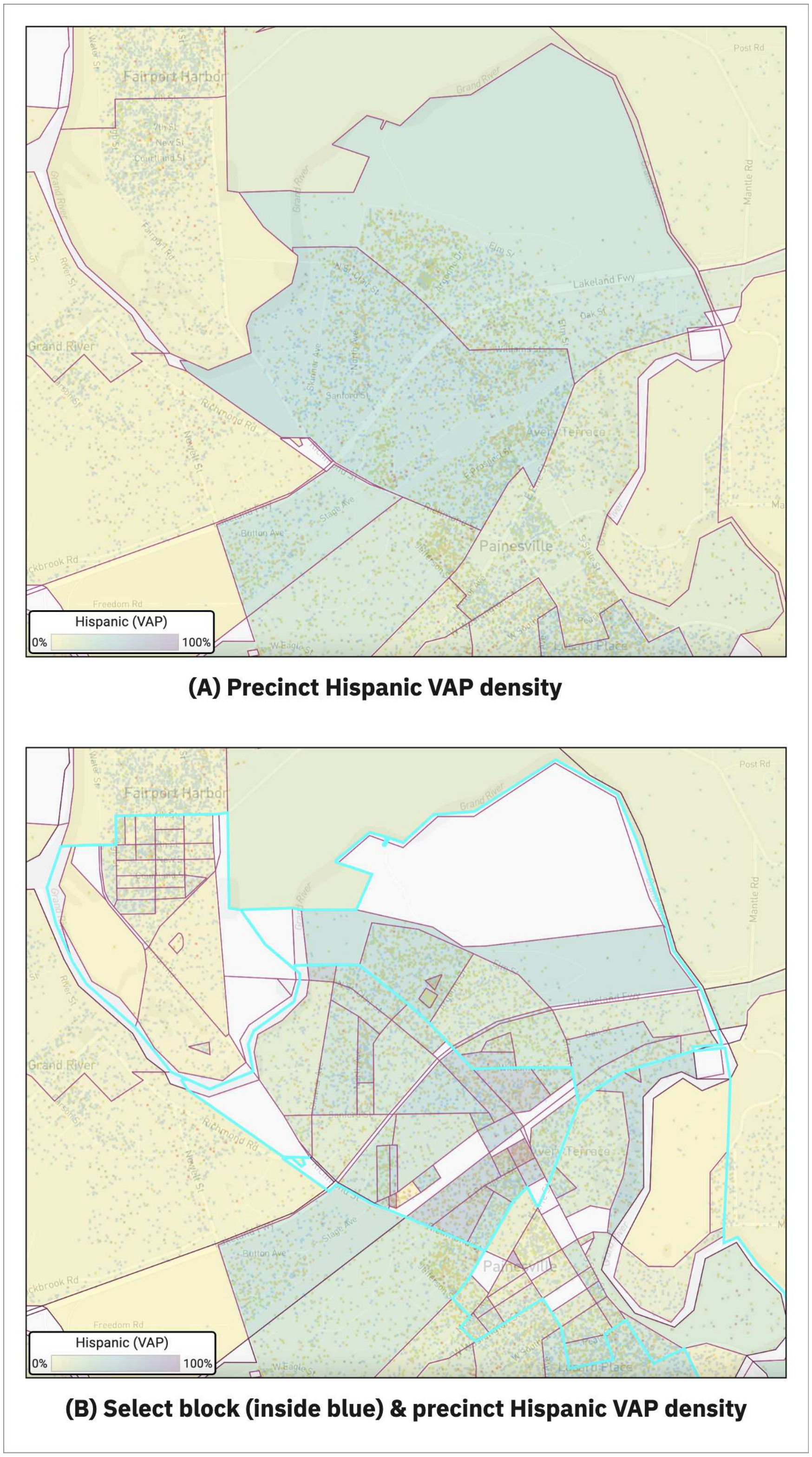

Figure 2: Choropleth maps of Hispanic VAP at the precinct level (A, top) and precinct with select blocks shown (within blue boundaries, B, bottom) near Painesville, OH.

Ohio State Route 2 is clearly visible in both panels, but only the block-level tessellation is composed of individual blocks associated with the freeway, which can divide communities. This phenomena reflects an important concept in transportation geography known as *community severance*, which is the "variable and cumulative negative impact of the presence of transport infrastructure or motorised traffic on the perceptions, behaviour, and well-being of people who use the surrounding areas or need to make trips along or across that infrastructure or traffic. Mindell et al. [2017]"

Depending on the data used to represent interest, different tessellation resolutions come with different risks. The use of high resolution data, such as individual census responses or geolocation data, could elevate the risk of re-identification of participants. This is less of a problem when using Census data, as the Census takes care to add statistical noise to high-resolution population estimates to protect against reidentification while preserving the aggregate population characteristics of a state[3], but individuals should take care if using datasets associated with potentially personally identifiable information.

Step 2: Derive the Dual Graph

A **dual graph** of a tessellation is a graph where each $T_i \in T$ is associated with a vertex v_i on a graph [West, 2001], and adjacency between v_i and v_j obtains when T_i and T_j share a border. To ensure contiguous communities, it is important that the border contain at least a line segment, or so called "Rook" adjacency. If tiles could intersect at only a point–i.e., "Queen" adjacency–then communities could contain impassable corners, which would violate common statutory definitions of communities of interest. Units E and I in Figure 3 give an example of Queen adjacent tiles; if they were considered adjacent, it would be possible to draw a community that contains only those two nodes, which would violate practical contiguity.

Note that it may be necessary or best practice to prune the graph to remove edges that are known to represent invalid community connections. A common example where this may be necessary occurs in Census blocks that contain islands; in many cases, the block may border a wide range of blocks, though the land mass is centered closely to only a few of these. Pruning these ties can significantly improve the ability of this method to reflect territorially communities.

Step 3: Set Edge Weights Using Empirical Data

Edge weights are numerical values assigned to edges that represent a strength of connection between adjacent vertices with respect to some sociopolitical interest. This strength of connection can be determined using various empirical data, and different interests may be better represented by different datasets. Let $\varphi(v_i, v_j) \geq 0$ be used to represent some function that defines the edge weight between units i and j. The more closely that i and j are related–which can be guided by data and modified to reflect a range of relationships, including similarities, probabilities of interactions, and binary relationships (e.g., if i and j located in the same administrative region)–the higher the value of φ should be.

Step 4: Apply Community Detection Algorithms and Dissolve Community Polygons

Community detection algorithms are then used to identify community structures that arise from a combination of graph topology and edge weights. Communities on a graph are generally collections of vertices that "cluster into tightly-knit groups with a high density of within-group edges and a lower density of between-group edges. [Newman, 2004]" These algorithms create a partition $\hat{X}(\beta) = \{\hat{X}_k(\beta)\}$ of the graph G given some parameterization β. A partition follows the uniqueness and completeness assumptions delineated above.

There are several approaches and algorithms available to model community structure. A comprehensive review of community detection algorithms conducted by Yang et al. [2016] compared the computational performance of 8 algorithms–Edge Betweenness [Girvan and Newman, 2002], Fastgreedy [Clauset et al., 2004], Louvain/Multilevel [Blondel et al., 2008], Label Propagation [Raghavan et al., 2007], Infomap [Rosvall and Bergstrom, 2007], Leading Eigenvector [Newman, 2006b], Spinglass [Reichardt and Bornholdt, 2006], and

[3] See the U.S. Census United States Census Bureau [2021a] for more information on Census statistical safeguards.

Walktrap [Pons and Latapy, 2005]–and found that the Louvain/Multilevel algorithm outperformed all others consistently on larger graphs. Community detection at the block level can include hundreds of thousands of nodes and millions of edges, so high-performing algorithms are desirable. However, the Louvain algorithm uses modularity, which suffers from the resolution limit [Fortunato and Barthélemy, 2007]. Furthermore, the Louvain algorithm does not guarantee contiguity.

Instead, the later Leiden algorithm [Traag et al., 2019], which guarantees contiguous community partitions and performs as well or better than the Louvain algorithm computationally, was selected. The Leiden algorithm minimizes the *Constant Potts Model* [Traag et al., 2011], given as

$$\mathcal{H} = -\sum_k (e_k - \gamma n_k^2), \tag{1}$$

where $e_k = \sum_{ij} A_{ij} w_{ij} \delta(\sigma_i, k)\delta(\sigma_j, k)$ is the total edge weight contained in community k, σ_i is the community to which vertex i is assigned, δ is the Kronicker Delta function, and $n_k = \sum_i \delta(\sigma_i, k)$ is the number of nodes in community k.

The CPM (1) also contains a resolution parameter, γ, that reflects the important principle of *scalability* identified by Chen et al. [2022]–in short, communities of interest may differ with the different scale of government, as different levels of government have different authority over different relevant interests. Smaller values of γ reduce the penalty for including more nodes in a community c, while continuing to reward edges contained within the community, leading to larger (lower resolution) communities. As γ increases, the CPM penalizes the number of nodes, leading to smaller (higher resolution) communities.

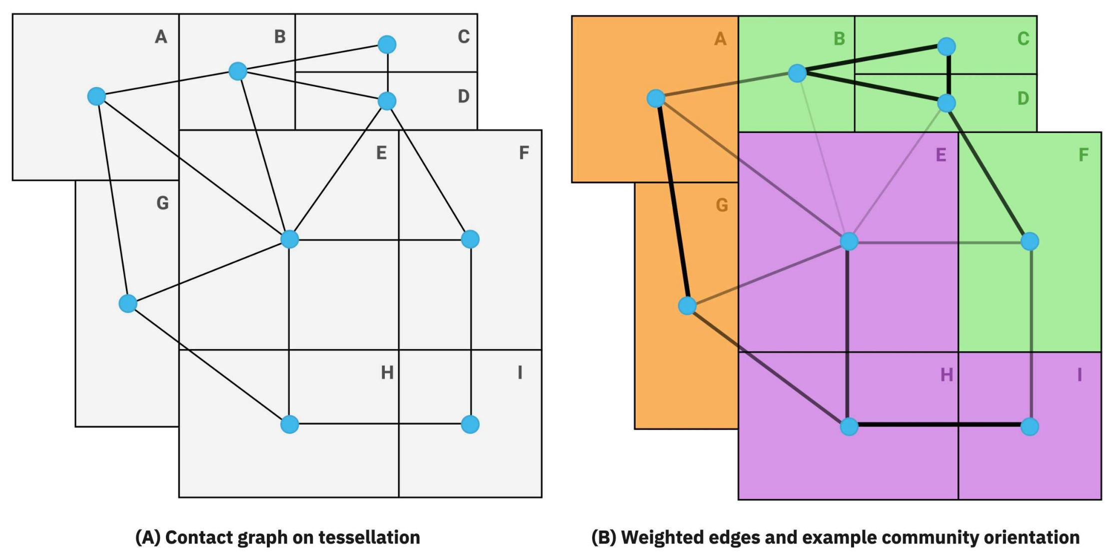

Figure 3: Example of tessellation with contact graph (A, left) and subsequent community identification using community detection algorithm (B, right). Edge thickness represents hypothetical strength of connection while colors indicate community cohabitation.

Community detection is a combinatorics problem, and most algorithms converge to local optima instead of a true global optimum. The Leiden algorithm in particular takes a greedy approach, achieving a local optimum by iterating until the value of the objective function does not improve by more than some small tolerance ε in successive iterations. In larger graphs, this can lead to relative instability in the solution given a fixed parameterization β, a problem identified by Riolo and Newman [2020].

To improve the stability and reproducibility of community orientations, I use a frequentist approach, which I call *cohabitation frequency*, to identify edges that are found to be within the same communities with some probability $\alpha \in [0,1]$ (I set $\alpha = 0.98$ based on tests in Ohio and California). Cohabitation frequency leverages different random seeds to obtain different local optima under the same parameterization β–the community detection algorithm is run to convergence R times with the same parameterization. Edges that are not contained within the same community at least $\lceil \alpha R \rceil$ times are dropped from E to form E'; the resulting edges E' are combined with vertices V to defined the cohabitation frequency subgraph $G' \subseteq G$. The connected components $\{C_k\}$ of G' are then the collections of tiles making up communities of interest. A more detailed description of the cohabitation frequency approach is available in the supplementary materials. The use of cohabitation frequency communities to identify communities of interest ensures that the resulting communities are "strong", ensuring that the edges and vertices that comprise each community are reliable cohabitants.

Once the graph partition $C(\beta; \alpha)$ is identified, the geospatial extent of the k^{th} community $\tilde{C}_k \subset \mathbb{R}^2$ is found as the union of all tiles T_i associated with vertices in the community. This process is referred to as *disolving the tiles*–i.e., $\tilde{C}_k = \bigcup_{v_i \in C_k} T_i$, so that $\tilde{C}_k \subset S$ are polygons. Collectively, I refer to the set of communities $\tilde{C}_k$ as a *community orientation map*. The community orientation maps are the primary product of interest for policymakers and community advocates (e.g., in the form of GIS products such as shapefiles, GEOJson, etc), though weighted graphs and vertex assignment tables (VATs), which represent $C(\beta)$ in long form are also useful in analysis and algorithm integration.

Other methods for identifying COIs quantitatively have been proposed in the literature. For example, Rossiter et al. [2018] generated communities of interest using Thessian polygons and, somewhat similarly to the approach discussed herein, geospatial clustering of block groups based on the minimum spanning tree method of AssunÇão et al. [2006]. However, the method described herein takes a more graph-centric approach, allowing for characterizations of interest using novel data and scalability on tessellations, while also clustering based on well-established notions of community structure on graphs. Other approaches to identifying communities of interest have looked beyond the constraint of geospatial contiguity; for example, Vaghefi and Nazareth [2016] used Twitter data to identify communities of interest in social media, while Makse [2012] combined factor analysis with initiative voting records.

One key advantage of the graph-centric approach taken herein is that a representation of a state as a graph is common in many redistricting algorithms (e.g., Duchin and Tenner [2018], Cannon et al. [2022]). This connection allows for the relatively seamless integration of communities identified using this method with redistricting algorithms, helping to fill an important void in these ever-critical algorithms through a number of potential means, including providing weight information for algorithms to include; adding community vertex membership as information to search algorithms (which can be used to inform heuristic decision-making); or characterizing community splits in ensembles of districting plans in the context of exploratory modeling.

Data

Tessellations of each state that are delineated in Census TIGER shape files. The Census 2020 TIGER shape files include block, block group, and tract level, form the basis of all communities of interest identified in this paper[4]. These data were used to infer the contact graph for each state using Julia scripts included in online supplementary materials.

I used three datasets to quantify one characterization of sociopolitical interests, referred to as *similarity*. First, two socioeconomic datasets from the Census Bureau–the Public Law 94 (PL94) and 2016-2020 5-year American Community Survey (ACS5)–were used to calculate edge weights representative of shared socioeconomic interests under the similarity approach. The Census PL94 dataset includes population counts by race, (2) ethnicity, and (3) citizen voting age population at the Census block level, the highest publicly-available resolution. Second, the 2016-2020 ACS data contain estimates of populations by language spoken at home, household income bracket, education attainment level, industry, and more at tract and block group levels. These datasets are aggregated to Census tessellation levels, meaning that no secondary estimation is needed to map socioeconomic data to each tile.

Additionally, an exploration of calibration and validation relies on empirical community of interest map submissions taken from the Representable project [rep, 2023], a free and open repository where anyone can submit a community map, The Representable project stores user submissions of polygons delineating boundaries and associated descriptions of communities of interest. These data are stored in the form of GeoJSON files that can be downloaded directly from the site. For the calibration, data were obtained for the state of Ohio in August 2022.

[4] 2020 Census geometry shapefiles are available for each state from the Census FTP repository [United States Census Bureau, 2021b].

Characterizing Interest Using Socioeconomic Similarity

There is no unique measurement that can quantify the mosaic of potential sociopolitical interests, though socio-demographic similarity is a reasonable first-step. This approach is rooted in the assumption that people who signal similar backgrounds and lived experiences are likely to share interests pertinent to those shared realities. There are some examples that support this assumption; for one, race has been used in the United States to restrict and prevent political participation of voters. Individuals of the same race likely share some interests in the face of issues like redlining and environmental racism, which have lasting effects on the wealth and health of minority communities [Aaronson et al., 2021]. Language and culture are also important to interest; Spanish and Vietnamese speaking citizens, for example, are entitled to access to ballots in Spanish and Vietnamese, and cultural norms or experiences relevant to representational interest can sometimes be captured using language as a proxy for culture. Language may even shape political preferences and opinion by affecting the structural mechanisms through which we perceive the world [Pérez and Tavits, 2022]. Beyond these, educational attainment, education status, and income are just some measures that act as proxies for other interests relevant to representation.

Socioeconomic characteristics are paired with the graph by associating each vertex with characteristic vectors. Each socioeconomic class–such as race, ethnicity or language–is associated with a characteristic vector. Similarity between adjacent vertices i and j along any of these socioeconomic spaces are then represented geometrically by determining the angle between i and j's vectors. A characteristic vector $\chi_i \in \mathbb{R}^M$ associated with vertex i has elements χ_{im} associated with dimension m of the socioeconomic space. Elements of a characteristic vector can be given as a population or as a proportion, though all vertices must use the same convention. Proportions are useful when downscaling values from lower resolution surveys–e.g., when using block group or tract data at the block level. For adjacent vertices i and j, the cosine similarity

$$s_{ij} = \cos\theta_{ij} = \frac{\chi_i \cdot \chi_j}{||\chi_i||||\chi_j||}, \tag{2}$$

is the cosine of the angle between these characteristic vectors, θ_{ij}.

Cosine similarity is desirable for a few reasons; first, it is easily comparable across tiles with disparate populations, such that $s_{ij} \in [0, 1]$ for $0 \leq \theta \leq \pi/2$ ($\chi_i \geq 0$). Furthermore, when characteristic vectors have the same distribution across a socioeconomic space, they are colinear, or entirely similar, and $\cos(0) = 1$. Orthogonal vectors, which share no non-zero elements, take on the value $\cos\left(\frac{\pi}{2}\right) = 0$ and hence are entirely dissimilar. However, in some graphs, cosine similarity can lead to relatively low variance in edge weights, potentially reducing the value of edge weighting in community detection algorithms. In so, it may be preferable to highlight differences between adjacent vertices using a linear relationship between the angles, and so the normalized angle

$$g(\chi_i, \chi_j) = 1 - \frac{2}{\pi}\theta_{ij} = 1 - \frac{2}{\pi}\arccos\frac{\chi_i \cdot \chi_j}{||\chi_i||||\chi_j||} \tag{3}$$

can be used. The function g has the same desirable properties as cosine similarity–but with a higher degree of variation when θ_{ij} is close to 0 or $\pi/2$ since g is linear in θ. The normalized angle score was used to generate illustrative results discussed herein.

Of course, communities are often dependent on multiple interests. Multiple types of interest can be integrated into a single edge weight using different aggregations; for example, a simple approach would return a weighted mean of similarities across all dimensions d, i.e.,

$$\varphi_{ij} = \frac{1}{W}\sum_d w_d s_{ij}^{(d)}, \tag{4}$$

where $s_{ij}^{(d)}$ is a similarity function applied to each dimension d, $w_d \geq 0$ is the weight applied to dimension d, and $W = \sum_d w_d$ is the total weight. Of course, other aggregation functions can be used to generate the edge weights φ_{ij}, and each function might carry a different interpretation.

The choice of socioeconomic spaces and dimensions within it are important to consider when using similarity. The spaces of sociopolitical interest should be conceptually independent–e.g. age, ethnicity, race, sex, language spoken at home, income, education, jobs and industries, etc. are all components of interest relevant to representation. Individuals who identify with the same race may be more likely to hold similar views about certain interests, though if they are in very different income brackets, they may hold different views on taxation and government spending. Furthermore, within each interest space, categories should be divided in a way that appropriately captures differences in interest. For example, within age, dividing children by pre-school, elementary, junior high, and high school (¡18) age is likely more relevant to representational interests than arbitrarily grouping them into pairs of years. This becomes less obvious with adults, as there are a myriad of potential interests that could be represented by different aggregations–for example, 18-20 year olds are adults, but many states permit new behaviors when these adults reach 21. At the age of 65, benefits from a number of entitlement programs, such as Medicare and Social Security, become available for citizens and taxpayers. The Census includes some reasonable aggregations in the ACS5 that reflect some of these logical groupings (e.g.,18-20, then 21-24, 25-34, 35-44, etc).

Similarity Communities

To demonstrate the behavior of the similarity score approach, I will explore how similarity communities behave in Ohio under different assumptions about socioeconomic similarity. These communities were generated using a Julia codebase that leverages the iGraph C package [Csardi et al., 2006] for community detection and is freely available at `https://www.github.com/**REPOHERE**`. The codebase facilitates analysis at block, block group, and tract levels for any of the 50 states. Additionally, a public facing interactive tool that can be used to explore communities derived from similarity scores in different states is under development. Similarity scores were calculated using a combination of socioeconomic population counts and estimates taken from both Census PL-94 population counts and 2016-2020 American Community Survey 5-year estimates.

Figures 4 and 5 demonstrate how tiles data are used to generate similarity scores and assign edge weights; they show the race similarity graph near Valley View, Ohio overlaid against Census block population densities for black and white racial identities, respectively. Other racial categories (with the exception of two or more races) do not have significant populations in the region. Representative vertices are shown as dots[5], and their position is only used to illustrate how they represent an associated tile on the map. Edges are also shown; gold edges on the graph are associated with at least one vertex with no population (meaning similarity was set to 0.5), while dark gray edges were calculated using similarity between the adjacent tiles. The thickness of the edges is set using the similarity edge weight.

[5]Note that some vertices may be shown outside the boundaries of the tile they represent since they are associated with the crude tile centroid.

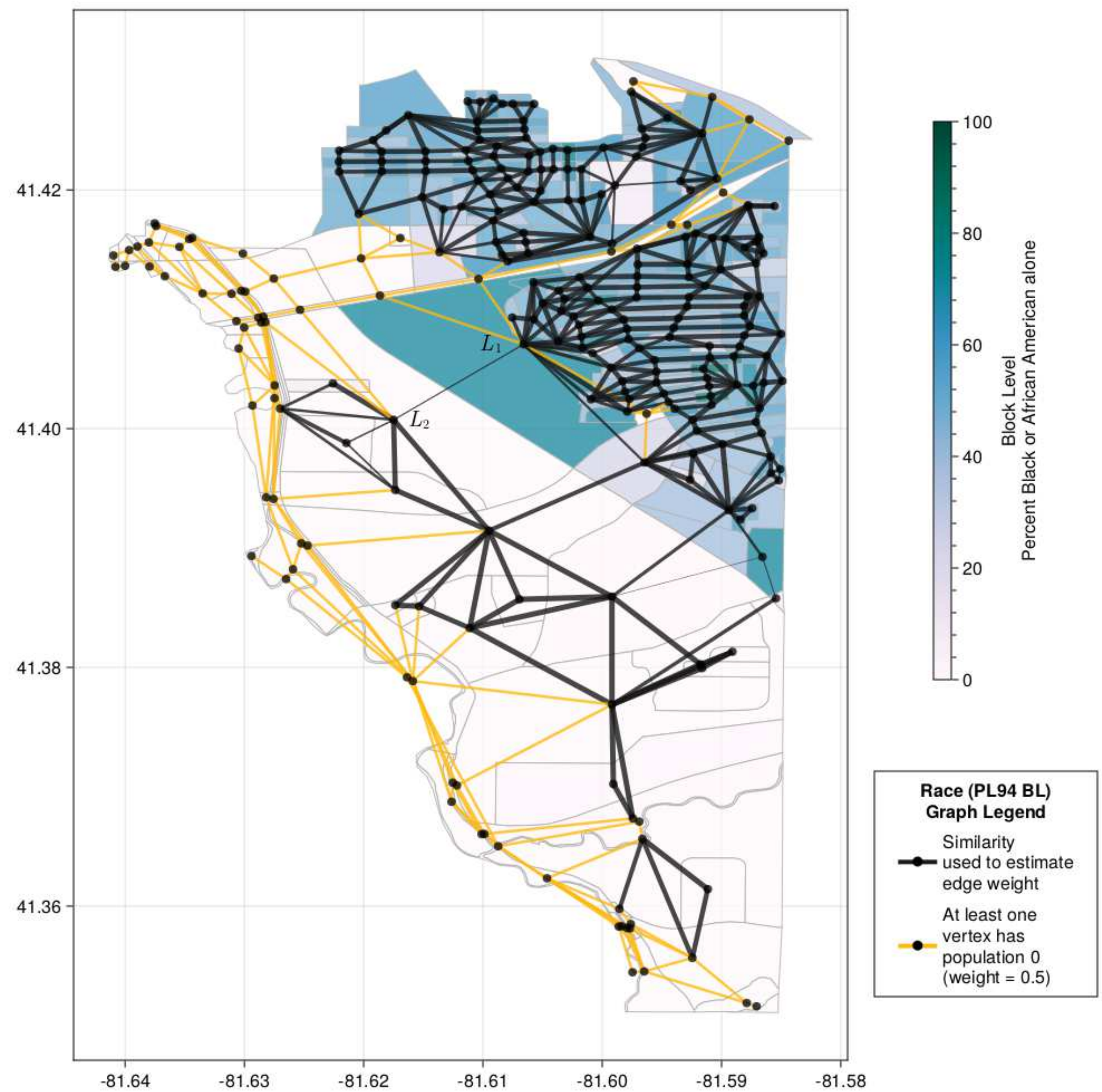

Figure 4: Dual graph near Valley View, Ohio with Race (PL-94) similarity edge weighting and the fraction of each block that identified as Black alone.

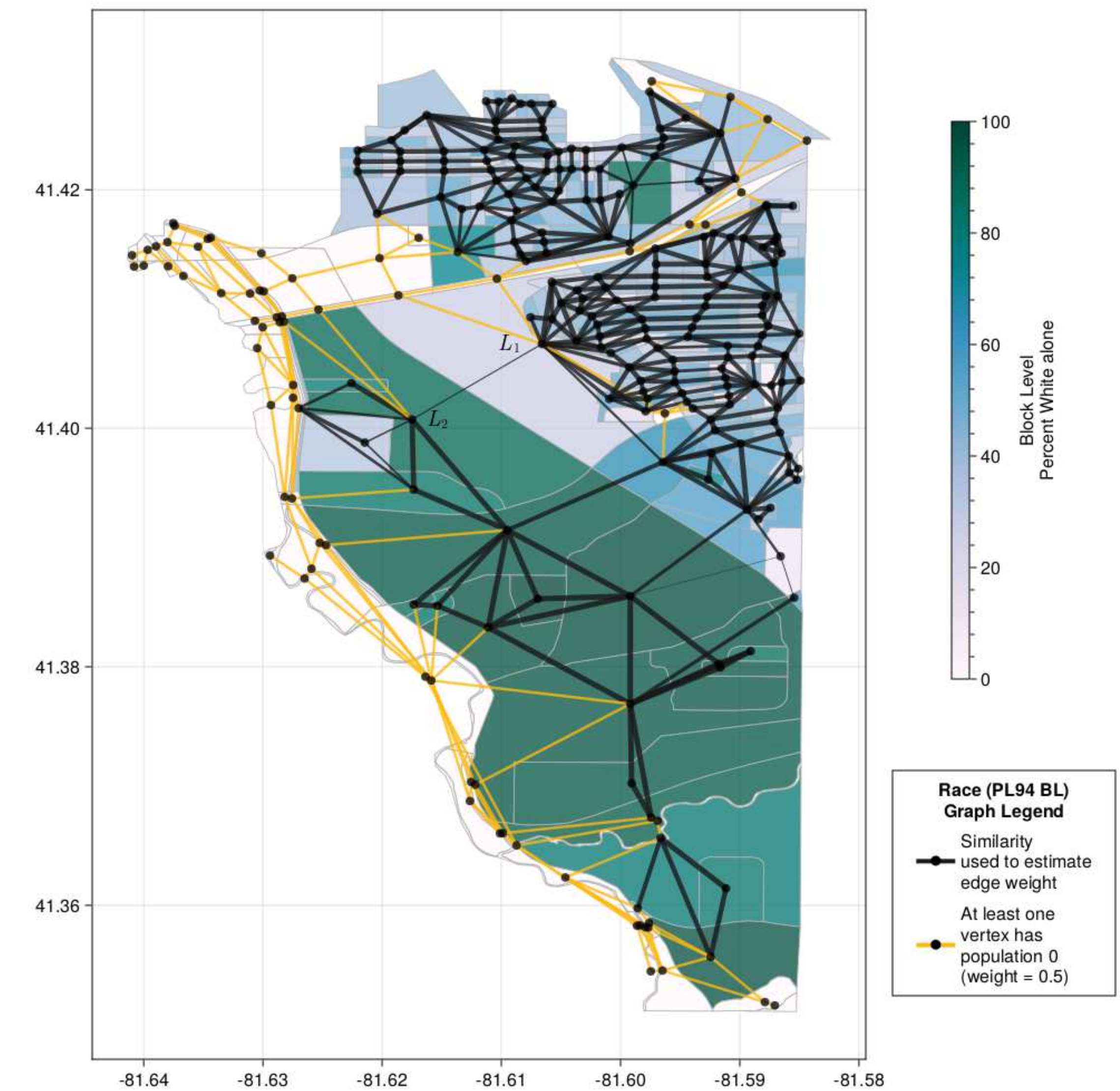

Figure 5: Dual graph near Valley View, Ohio with Race (PL-94) similarity edge weighting and the fraction of each block that identified as White alone.

The region near Valley View, Ohio is notably segregated, making it useful for illustrating graphs and similarity scores due to dissimilarity between adjacent blocks. The population living in the western half of the figure (tract 390351929)–has a high density of persons identifying as White (1,792 out of1,897 people, $\approx 93.94\%$). In contrast, the population living in the eastern tracts is less White: tracts 39035154501 ($\approx 43.46\%$ White, $\approx 48.10\%$ Black), 39035154502 ($\approx 39.11\%, 51.51\%$), 39035154601 ($\approx 33.85\%, 47.54\%$), 39035154603 ($\approx 42.11\%, 50.49\%$), and 39035154604 ($\approx 42.21\%, 49.89\%$) are minority-majority, and several are Black-majority.

Two blocks–$L_1 = 390351545021001$ and $L_2 = 390351929001020$ (2020 FIPS block codes) reveal the similarity edge-weighting calculation intuitively. In the diagrams, populated blocks that are next to each other with highly-contrasted coloring for each of the racial categories should show thinner edges between them due to dissimilarity. Blocks L_1 and L_2 are clearly dissimilar across racial categories of Black and White, as L_1 has a high proportion of residents that identify as black ($\approx 72.04\%$), while L_2 has a very high proportion of White-identifying residents ($\approx 90.12\%$); this compositional difference leads to a low edge weight, shown as a relatively thin edge on each figure.

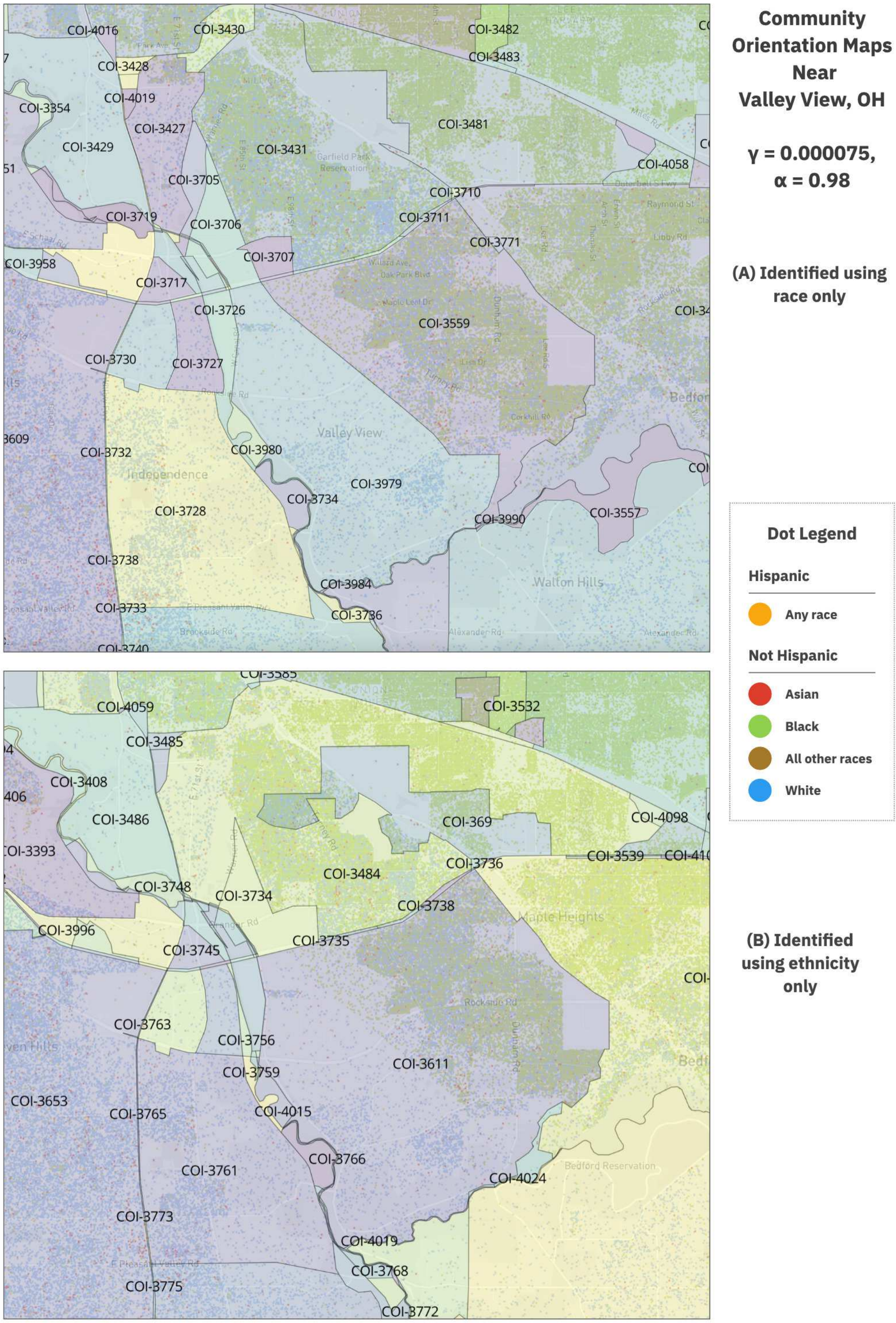

Figure 6: Community boundaries for $\gamma = 0.000075$ and $\alpha = 0.98$ drawn using race only (panel A) and ethnicity only (panel B) near Valley View, Ohio. Maps generated using Dave's Redistricting App [dra, 2022].

Figure 6 shows how the graphs shown in Figures 4 - 5 translate into communities of interest for a resolution of $\gamma = 0.000075$ and a cohabitation frequency threshold of $\alpha = 0.98$ on the block-level graph in Ohio. Each map shows different communities of interest (colored arbitrarily to distinguish between them) along with dots, each of which represents a few hundred residents of a different race/ethnicity combination [dra, 2022]. The communities are oriented quite differently in each case, reflecting the effects of the underlying data on community boundaries.

Note that the race-only community detection reflects the clear racial dissimilarity evident in Figures 4 and 5, as COI-3559 and COI-3979 have two distinct compositions, with COI-3559 having a much higher Black population (13,444/21,901$\approx$ 61.4% Black, 7,333/21,901$\approx$ 33.5% White) than COI-3979 (37/1,898$\approx$ 1.9% Black, 1,763/1,898$\approx$ 92.9% White), which is primarily White. Similarly, COI-3431 (12,259/20,665$\approx$ 59.3% Black, 7,274/20,665$\approx$ 35.1% White), and COI-3481 (9,976/10,421$\approx$ 95.7% Black, 223/10,421$\approx$ 2.1% White), are split, reflecting the greater diversity of COI-3431.

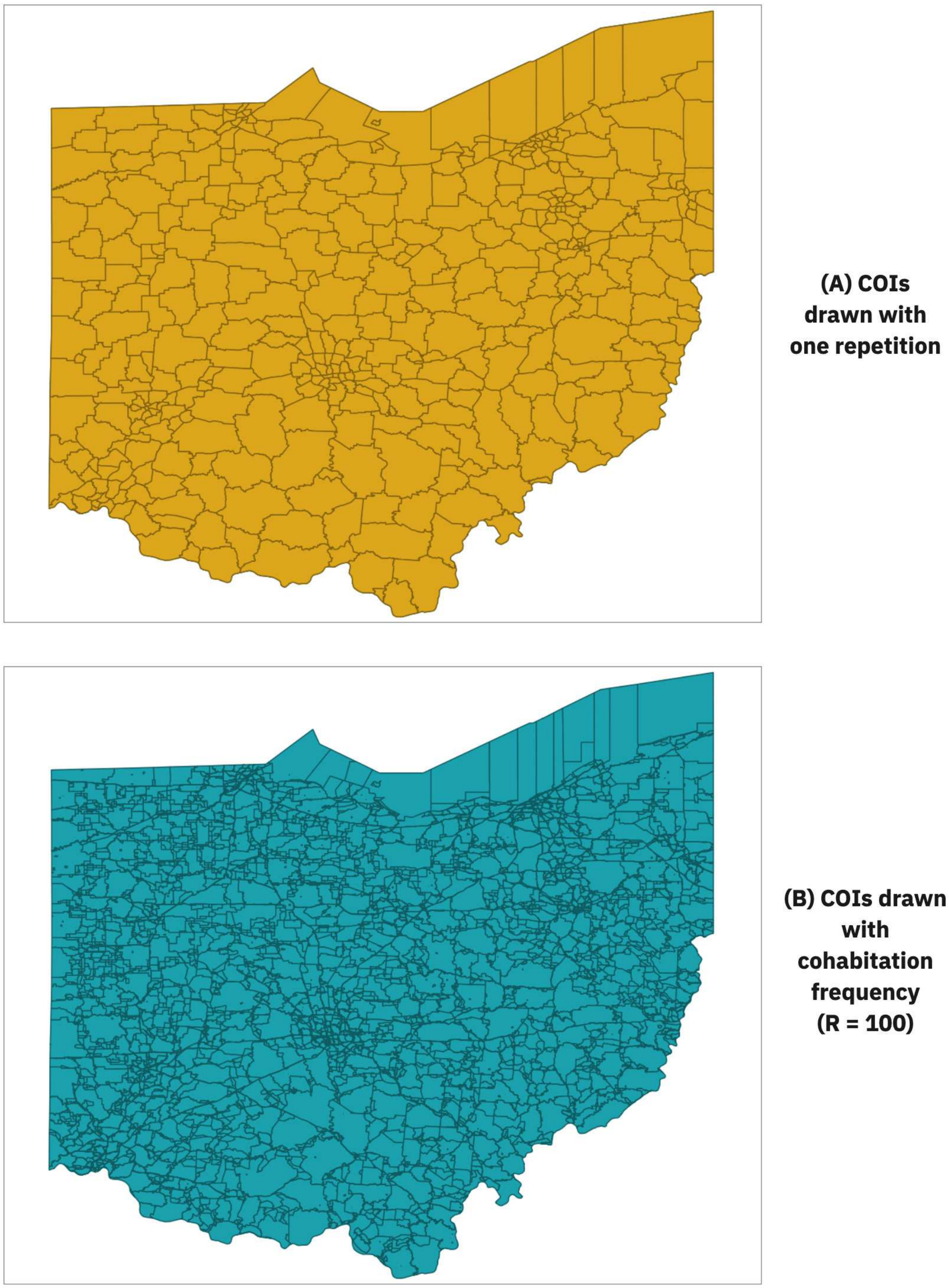

Figure 7: Comparison of communities drawn using a single solution from the Leiden algorithm in Ohio (top) versus communities drawn using $R = 100$ and $\alpha = 0.98$ (bottom).

Finally, the impact of cohabitation frequency is illustrated at the state level in Figure 7. Cohabitation frequency approach can lead to high variation in community sizes for a single resolution parameter; in practice, denser, well-defined community structures often emerge from those associated with cohabitation, while a number of “floating communities”–smaller collections of vertices that are often rural, unpopulated, and/or associated with infrastructure like U.S. Interstate highways (see 7)–surround them. These floating communities–which are only consistently reflected in the cohabitation frequency communities–often reflect well known barriers to community formation.

Validating Computational Communities

The approach to identifying communities of interest taken herein is fundamentally a model of community formation, and it is critical to understand the conditions under which model outcomes can be trusted to inform valid use cases–does the model represent what it claims to represent? The validation of a model of community formation in particular presents numerous challenges. There is no single ground truth to which computationally generated communities of interest can be compared to consistently establish validity if any ground truth is available (e.g., conflicting submissions in Figure 1). Even in the presence of empirical communities, the quality of information may differ considerably between available boundaries. For example, community boundaries drawn by well-organized community advocacy groups that incorporate feedback, comments, and perspectives from community members are likely to be of higher quality than an arbitrary test map.

However, in some cases, self-perceptions of community boundaries align well from different sources, creating an opportunity to evaluate what Adcock and Collier [2001] refer to as *convergent* or *discriminant* validity. In doing so, we ask the question, "[a]re the scores... produced by alternative indicators... of a given systematized concept... empirically associated and thus convergent? [Adcock and Collier, 2001]"

Alignment with Self-Reported Communities

An obvious approach to validation of any model involves modifying parameters to align model estimates with empirical observations in a process known as calibration. This method for validation is somewhat weakened in the absence of an authoritative set of communities of interest, though it helps establish both nomological validity and convergent validity by answering the question: under calibrated conditions, does the systematic approach to community identification correlate with other information about where a community of interest is located?

The construct of community alignment can be conceptualized using different characteristics of a community. A method-drawn map can be aligned with a self-reported community on the basis of area alone, reflecting that a systematically-constructed community of interest reflects the territorial extent of the self-reported community. Alternatively, a method-drawn community might be considered to align well if it preserves the same population. The selection of one one measure over another has normative implications. For example, a self-reported community may intentionally include certain points of interest or landmarks, such as a park, stadium, or airport, as part of the basis of the community that population alignment might miss, while others may include boundaries that extend beyond key population centers to include space that isn't critical to the interest that the community seeks to coalesce.

To help establish validity, I demonstrate a calibration of computational communities of interest to an empirical baseline using two quantifiable requirements to characterize alignment. First, any map that aligns well with a self-reported target community of interest should minimize splitting of that community. Second, overlapping communities that align well should be associated with similar measures (such as area or population).

To quantify splitting, Chen et al. [2022] introduced the *effective split index* (ESI) and *uncertainty of district membership* (UDM) in support of a quantifiable standard of community splitting by districting plans. Here, method-drawn maps are treated as districts, while a self-reported community of interest is treated as a target. The metrics introduced by Chen et al. [2022] are generalizable to arbitrary measures μ[6] defined on T and, therefore, any method-drawn communities X that are collections of tiles. For the following definitions, let C be a method-drawn community orientation map (the overlay), and let $X \subset \mathbb{R}^2$ be target empirical community to which C will be compared (the target). If $K(X,C) = \{k \in K : C_k \bigcap X \neq \emptyset\}$ is the index set of all overlay communities C_k that intersect community X, then

$$p_k(X,C;\mu) = \frac{\mu(C_k \bigcap X)}{\mu(X)} \tag{5}$$

is, for an arbitrary measure μ, the fraction of the measure of C_k and X out of the measure for empirical community X. The effective split index $s(X)$ of a community X is defined by Chen et al. [2022] as

$$s(X,C;\mu) = \frac{1}{\sum_{k \in K(X,C)} p_k(X,C;\mu)^2} - 1, \tag{6}$$

[6] See supplementary materials for a brief discussion of measures.

while the generalized uncertainty of district membership–which is the two-bit Shannon entropy of the measure space implied by the overlay C–is given by Chen et al. [2022] as

$$u(X, C; \mu) = -\sum_{k \in K(X,C)} p_k(X, C; \mu) \log_2 p_k(X, C; \mu). \tag{7}$$

However, community splitting alone is inadequate to identify community alignment–a very large community, for example, might avoid splitting a target if it wholly contains that community. To compensate, I use what I call the *average overlay tile ratio* (AOTR)

$$a(X, C; \mu) = \sum_{k \in K(X,C)} p_k(X, C; \mu) \frac{\mu(C_k)}{\mu(X)}. \tag{8}$$

The average overlay tile ratio is used to characterize measure equivalence between overlaying communities and a target empirical community. The final calibration metric–Manhattan AOTR, or MAOTR [7])–m compares the distance of a from 1, i.e.,

$$m(C, X; \mu) = |1 - a(C, X; \mu)|. \tag{9}$$

The three metrics s, u, and m were applied to two measures associated with each method-drawn community X_k in an orientation map and each target empirical community C: **area** (from Mollweide projection) and **total population** (estimated using block-level Census data in empirical communities).

To demonstrate how to identify parameterizations of algorithms that align well with empirical communities, I drew 30,000 community orientation maps in two stages using similarity scores with 13 parameters, including γ, β, and weights for the 11 socioeconomic factors shown in Table **??**. The first 20,000 maps were used to narrow parameter ranges and identify collections of parameters important to good alignment, while the second set of 10,000 was used to explore convergent validity and factors associated with good fits.

[7] See the supplementary materials for a detailed discussion of AOTR and M-AOTR

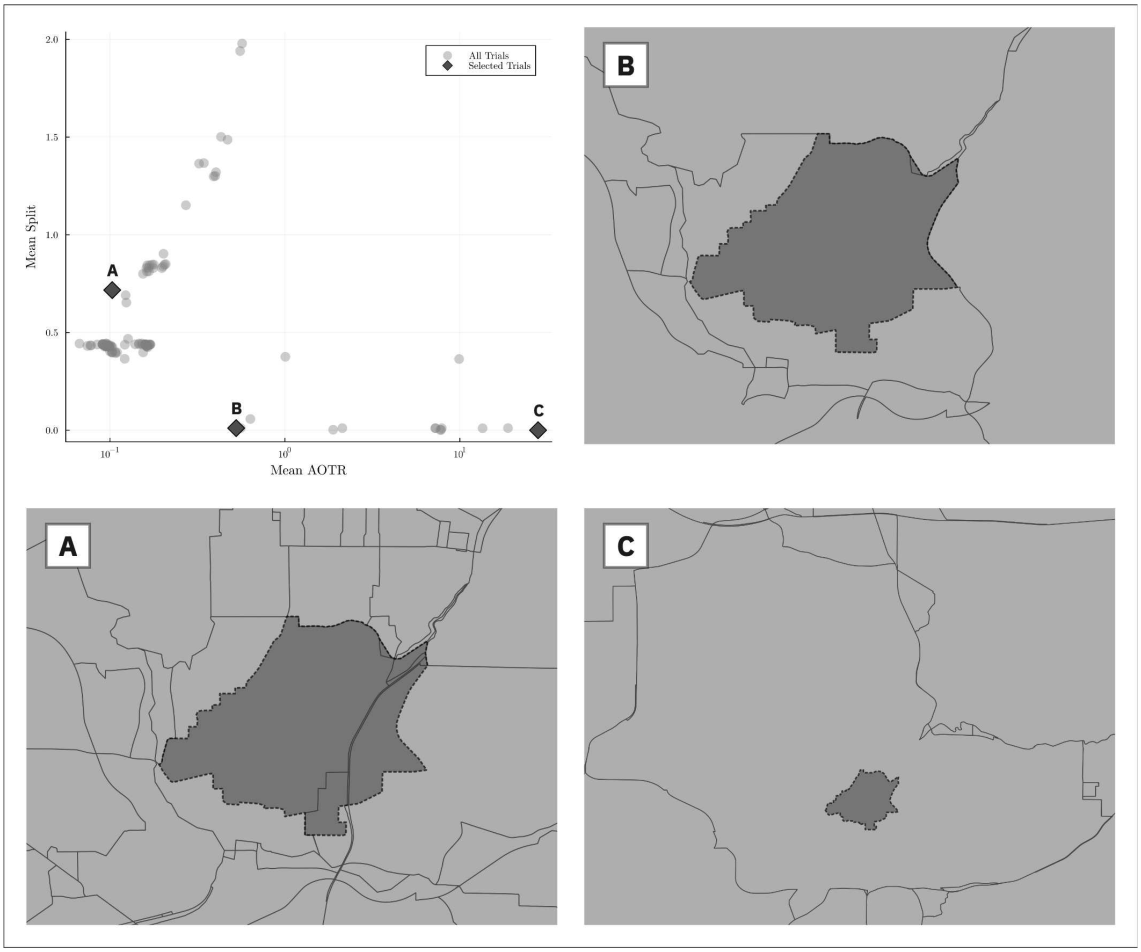

Figure 8: Selected points on calibration Pareto frontiers and associated maps with comparison to Representable "Akron North Hill Community" target [rep, 2023].

Figure 8 shows a subset of 112 layered Pareto fronts taken the larger set of 20,000 runs along with three community orientations taken from the front, each demonstrating how the calibration metrics characterize alignment. The empirical Akron North Hill Community (the target) is shown in darker gray in each of panels A, B, and C, while method-drawn community orientation maps are shown in lighter gray. Panel A illustrates a low-MAOTR, high-split ($(\bar{m}, \bar{s}) = (0.0184, 0.717)$) community orientation. Notice that the communities intersecting the target are of roughly the same size, but there are two clear splits within the community. Panel C, on the other hand, demonstrates a high-MAOTR, low-split ($(\bar{m}, \bar{s}) = (28.008, 0.0)$) community orientation. In this panel, the empirical community is totally preserved, but it is wholly contained in another community that differs vastly in scale. Finally, panel B reflects the calibrated region: low-split, low-MAOTR ($(\bar{m}, \bar{s}) = (0.526, 0.0106)$). Here, the empirical community is contained within a larger community, but there is very little splitting. Closer inspection of the component MAOTR values reveals that close alignment in population drives the community fitting the calibration $m_{AREA} = 0.787, m_{POPULATION} = 0.266$ reflect a fairly close population fit, even if the area is somewhat larger than that of the empirical Akron North Hill Community. Furthermore, among $s_{AREA}, s_{POPULATION}, u_{AREA}$ and $u_{POPULATION}$–only $u_{POPULATION}$ exceeded 0.0011 (≈ 0.0345), reflecting the minute splitting that occurred when overlaying the calibrated community orientation map against the target community.

The calibration dataset also allowed for the exploration of another question related to convergent validity: what factors, if any, lead to a good fit of the community orientation map and the empirical community, and how do those factors compare to the description of the community? To explore this question, I reduced the space of component similarity dimensions to six socioeconomic factors–Age, Educational Attainment, Household Income, Language Spoken at Home, Race, and Ethnicity–and generated 10,000 community orientation maps using random weighting of socioeconomic similarities. Weights of the individuals similarity dimensions were log-normalized on the probability simplex to improve the uniformity of the sampling space, i.e., the weights w_d were calculated as $w_d = \frac{\log x_d}{\sum_{d \in D'} \log x_d}$. Furthermore, I refined the definition of good fit to place bounds on individual measures so that $m_{AREA} < 1.433$, $m_{POPULATION} \leq 0.488$, and $s_{AREA}, s_{POPULATION}, u_{AREA}, u_{POPULATION} \leq 0.01$. The bounds for m were chosen based on author judgement and visual inspection of the component distributions of MAOTR. The threshold of 0.01 used for component expected split index and uncertainty of membership metrics are fairly restrictive, as "[f]or any given community, uncertainty of greater than 0.5 bit may be considered substantial. [Chen et al., 2022]" The bounds were more restrictive than the 25th percentile of each metric's distribution.

I then used the Patient Rule Induction Method (PRIM) [Friedman and Fisher, 1999] and Classification And Regression Trees (CART) [Breiman, 1984] supervised learning algorithms to characterize ranges of socioeconomic similarity weights that correspond with a good fit and compare those to the description of the community. These algorithms work by generating decision trees and input regions that explain a labeled space, generating tradeoffs in density (the fraction of cases within the region that match the condition) and coverage (what fraction of all cases are explained by the region). The Representable community "Akron North Hill Community" is described by the submitter as an "Immigrant and refugee Asian community in Northeast Ohio" [rep, 2023]. This description highlights three key quantifiable components of identity relevant to community formation–immigrant status, refugee status, and race, specifically the Census grouping of Asian. The data used in this exploratory analysis did not include data on immigration status or refugee status–though the Census does estimate population totals by immigrant status in the ACS at the tract level–leaving race as the only quantifiable, comparable component.

The exploratory analysis revealed that race is an important component of community orientation maps that reflect the population and boundaries associated with the individual submission. CART (restricted to three dimensions) and PRIM both determined that the resolution γ (CART: $\gamma > 3.626\text{e-}5$, PRIM: $\gamma > 3.5\text{e-}5$) and the weights of ethnicity (CART: $w_{\text{ethnicity}} < 0.2815$, PRIM: $w_{\text{ethnicity}} < 0.29$) and race (CART: $w_{\text{race}} > 0.133$, PRIM: $w_{\text{race}} > 0.136$) were important to explaining good fits. Notably, over 70% cases require the normalized weight of race to be greater than 0.133.

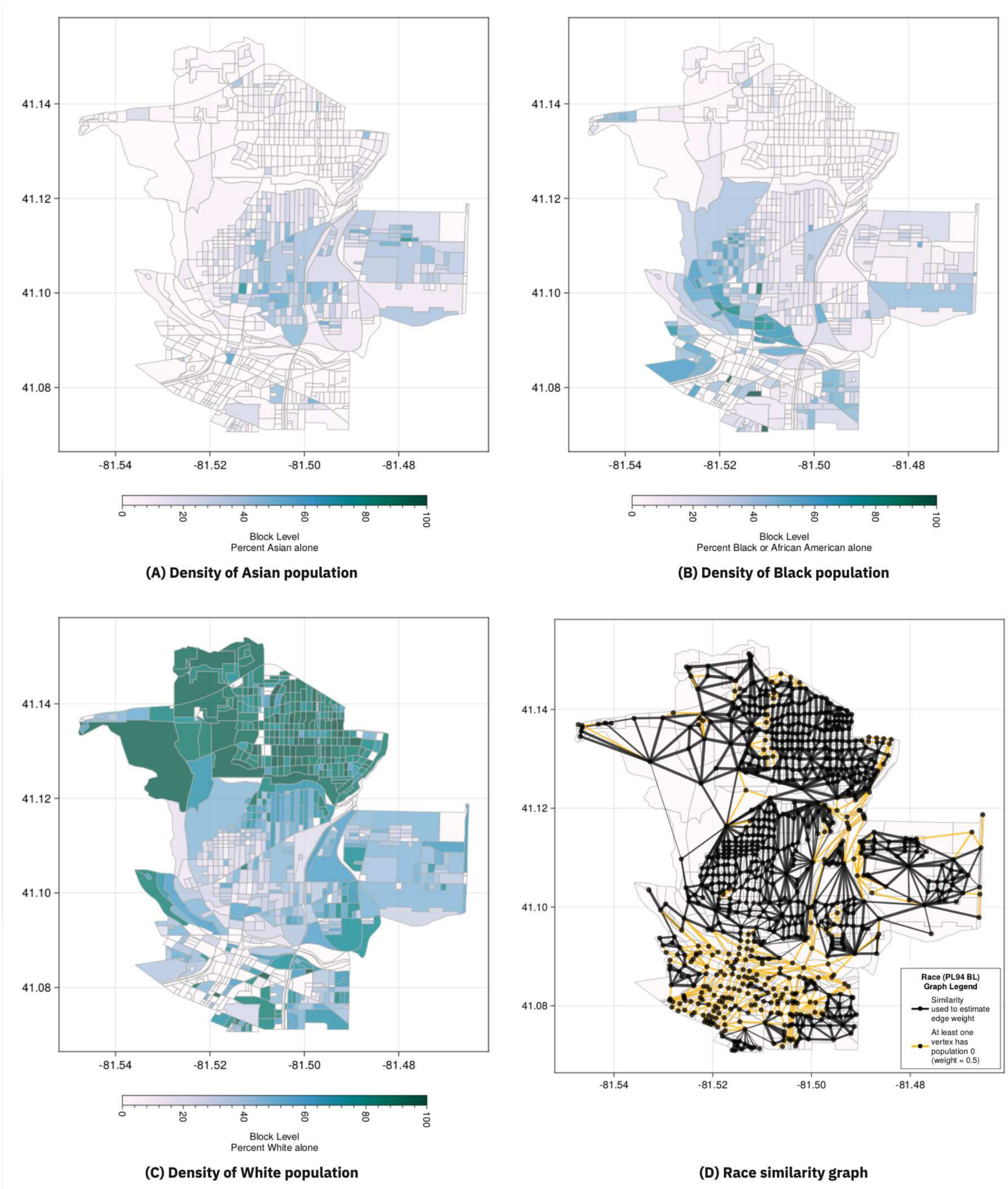

Figure 9: Density of Asian (panel A), Black (panel B), and White (panel C) populations along with race similarity graph (panel D) in the area surrounding "Akron North Hill Community" [rep, 2023].

Figure 9 shows the density of Asian, Black, and White–the three most prevalent racial categories in the area north of central Akron, OH–populations by Census block along with the block-level race similarity graph. It is unsurprising that high weights for race similarity are not associated with higher density and coverage; even though the "Akron North Hill Community" includes "Asian" as a defining component of the community, Figure 9 shows that the distribution of individuals identified as Asian is not dissimilar to the east of the community area. The upper bound on ethnicity is also not surprising; the area is characterized by ethnic homogeneity, reducing the value of ethnic similarity in defining this particular community. However, relative significance of racial similarity as a dimension aligns partially with the individual's description of the community and an encouraging component of convergent validity.

There are some important limitations to this exploratory validation. First, the region determined by CART and PRIM covers 73.3% of all acceptable fits with 24.2% density. The background density of acceptable fits is 8.93%, meaning that this region has about a 2.7x higher rate of acceptable fits than the larger dataset. However, this also means that 75.8% of cases within the box are not acceptable fits, and so it's inadequate to explain what constitutes a good fit. Furthermore, the empirical baseline community reflects one individual's perceptions about the boundaries of a community, and the quality of that submission or how truly it reflects the individual's perceptions or the perceptions of the community itself is impossible to verify. Additionally, this validation was not intended to systematically evaluate how well method-drawn communities reflect empirical submissions; instead, it is intended to support validation by showing it can closely reproduce some reasonable available empirical baselines. A more thorough analysis could help identify factors important to community formation on a broader scale.

Valid Use Cases

Calibration to an empirical community boundary provides some information on measurable factors and covariates that influence how that individual perceives their community. However, a single empirical baseline should not be used to parameterize a map for an entire state. In practice, community self-reporting is highly specialized, informed by experiences, preferences, mobility, and firsthand knowledge of place. One individual's perceptions do not extend to the population at large and may not even extend to the same individual when living in a different location or set of socioeconomic circumstances. Social norms and other key factors–such as language spoken at home–in one community may shape how the boundaries of a community are perceived or what components of interest are most salient.

Geospatially uniform weighting and aggregation of socioeconomic similarity, mobility or other edge weight components, which reflects an assumption of spatially homogenous preferences, is unlikely to generate a map that appropriately captures spatial heterogeneity. Instead, weighting may be dependent on the community itself. In spite of these hurdles, calibration helps to achieve several things. First, it provides a validation test for the method itself. The method's ability to mimic some empirical communities–both in terms of reflecting key calibration metrics and supporting qualitative descriptions of the community–provides a measure of internal validation, demonstrating that it can reasonably reproduce empirically observed targets. It also shows the pliability of the method and its responsiveness to a range of input factors, identifying different communities on the basis of different interests.

Importantly, calibration can help bound algorithmic parameters–such as the resolution parameter or socioeconomic weights in the similarity method–that affect community orientation maps. Indeed, multiple empirical targets can be evaluated simultaneously using the latin hypercube method method described herein, which can help quantify uncertainty surrounding community boundaries. Then, collections of feasible maps can remain informative for characterizing districting plans in terms of community splits. For example, leveraging distributed and high-powered computing, it is possible to evaluate how a bounded range of viable community orientation maps are split by a single districting plan, then use statistical analyses to characterize which plans tend to split which community orientations more. Furthermore, bounded batch map-drawing could be combined with statistical aggregation methods–such as those proposed by McCartan et al. [2023]–to characterize how often tessellation tiles are grouped with other tiles. This information could, for example, be used to set transition probabilities in map drawing algorithms or evaluate whether a map drawing authority's decisions to assign related tiles into different districts is defensible.

Individual maps are still useful. A single map could be generated by members of the public or community organizations to agree upon boundaries of their community. Furthermore, parameterizations associated with multiple maps, which could be elicited through a well-structured tool or workshop environment, can be used to reproduce maps and aggregate preferences across a number of persons or organization, potentially facilitating agreement across a number of individuals. Of course, individual maps can still be qualitatively informative if analyzed contextually, though map drawers and researchers should take care to avoid basing map boundaries on an individual map unless supported by analyses that quantify parametric uncertainty.

Map drawers must be prepared to accept and interpret computational maps of communities of interest. If presented by individuals, computational community orientation maps can be considered as no different from existing maps that reflect constituents' perceptions about where boundaries should lie, even if those perceptions are shaped by algorithmic or systematic approaches to setting community boundaries, though care should be taken to avoid overweighing residents' perceptions about the boundaries of communities in which they do not live. Additionally, analyses derived using these maps are informative, providing a replicable, systematic way to understand which tessellation elements might be grouped with others. However, legal stakeholders, including legislatures, map drawers, and judges, will need to develop standards and practices for interpreting and integrating computational communities into the process.

Importantly, map drawers will face the challenge of identifying rules for transparency in community (and district) maps drawn using algorithms. The adoption of or debate around community maps without accountability increases the risk of systemic abuses by actors that present "black-box" analyses, obfuscating data, motivations, and conclusions, propagating and perhaps even worsening a current concern with the process [Kim and Chen, 2021]. This is especially true with analyses derived from private and novel datasets, which are necessarily less transparent than comparable analyses conducted with publicly available data. This necessity opens the door to further misuse or even tampering without mechanisms for verifying analytical source data.

To mitigate these harms, processes must be in place to allow for verification and accountability–in accordance with data use agreements and privacy laws–by courts and redistricting officials. Traditional principles for accountability in "bureaucratic mechanisms" include transparency and consistency [Trelles et al., 2023]. However, algorithms expose limitations of transparency as an ideal, as complete algorithmic transparency is limited and can "intentionally occlude", privilege "seeing over understanding" , and can "invoke neoliberal models of agency" [Ananny and Crawford, 2018]. However, other regulatory mechanisms can foster accountability. Krafft et al. [2022] evaluated accountability mechanisms in algorithmic decision making systems; in particular, they found that "superficial information disclosure requirements" and "comprehensive testing and auditing... [can] make decisions traceable." Though the computational methodology introduced herein is not an algorithmic decision making system, the system operates similarly, with "learning" in this case corresponding with finding an optimal community orientation based on empirical data.

Finally, the encroachment of computational methods into legal processes and statutes, including redistricting processes, increases the risk of entrenching what Diver [2022] has called *computational legalism*, or "an extreme species of unreflective rule-following that code can so easily impose upon citizens [Diver, 2022]". Adopting computational maps of communities as a standard would become a form of what Hildebrandt [2018] identified as *algorithmic regulation*, or "standard-setting, monitoring and behaviour modification by means of computational algorithms. [Hildebrandt, 2018]". This reinforces the finding of Mac Donald and Cain [2013] that "objective COI approaches... are at best supplements to public testimony, certainly not a substitute. [Mac Donald and Cain, 2013]"

Discussion

This paper has used statutory and academic definitions of communities of interest to develop a mathematical approach to identifying feasible community orientation maps using empirical data and contact graphs. Furthermore, I have demonstrated how similarity data can be used to identify community boundaries, while finishing with an exploration of the validity of these maps alongside some recommendations for redistricting stakeholders. The novel method introduced herein addresses a number of problems with community identification, including selection bias in public commentary, an often untenable community uncertainty space, and a lack of communities of interests in map drawing algorithms.

The model discussed herein, as with any model, is limited by its assumptions that simplify the complexity of the world it attempts to represent. Structural biases–including sampling bias, algorithmic bias, and the unconscious biases of researchers and analysts–persist. The choice of tessellation, community detection algorithm, and data used to characterize interest reflect these biases and all shape community orientations that are produced. Even the use of Census data comes with caveats; use of Census tilings cements Census perspectives on what territorial regions belong together, and Census survey counts and estimates are imperfect. The use of similarity scores and probabilities of neighbor colocation are obvious cognitive simplifications of socio-political interests, which are driven heavily by individual experiences and perceptions.

The combination of computational methodology and "big data" can convey a false sense objectivity and foster overconfidence in the derived products, and this method contributes to a growing body of approaches that quantify and systematize legal entities, increasing the risk of devolution into computational legalism. While the author expects this is unlikely to occur, these maps should be seen as a decision-making and research aid rather than an unassailable ground truth that supplants the public input process.

These limitations, however, should not preclude the use of the approach in identifying communities of interest given its numerous potential practical and academic benefits. It carries the potential to increase the visibility of many communities that would otherwise be overlooked in the community process two fold. It illuminates the boundaries of communities to map drawers in the absence of self-identification. On the other hand, it provides members of the public easily accessible baselines that readily illustrate the normative community of interest and can facilitate discussion, organization, and further self identification.

The use of contact graphs and community detection algorithms is versatile and can, using different empirical datasets, represent numerous important constructs relevant to representation beyond socioeconomic similarity and colocation. Political preferences, which are important signals for substantive interests, can be imperfectly captured by incorporating results from direct ballot initiatives in states where they are carried. A range of data are available that could characterize other policy-relevant substantive interests, such as information about healthcare and insurance access, social and historical trends (including redlining), environmental quality, or labor and industry. The restriction that communities do not cross certain boundaries, such as county lines, can be incorporated using a binary co-membership along an edge. Geographic features such as cultural and economic landmarks–including religious sites, major industrial sites, ports, commerce and trade centers, and more–can be tied to representations of political interest and can even be combined with cellular device data to characterize interactions themselves. Furthermore, more complex edge weighting functions φ_{ij} could be developed to capture how certain dimensions of interest can become more salient in the face of other similarities or differences–for example, race and ethnicity alone misses key nuances that are important to discerning communities within those groups, but are important between groups. There is a frontier to explore on how interests might be best characterized quantitatively.

This method also aligns neatly with current computational approaches to evaluating legislative redistricting and can facilitate theoretical advancements into redistricting research. Notably, the approach taken in this paper also integrates well with Markov Chain Monte Carlo methods for drawing random maps (e.g., Cannon et al. [2022]) given that both model states as contact graphs–next steps might integrate computational communities of interest into batch-drawn maps and use splitting standards as an evaluation metric, or explore how districts based on transition matrices derived from weighted edges compare to those based on unweighted edges. Chen et al. [2022] have argued for a communities of interest as a standard in redistricting, and the introduction of systematic identification of community maps can help facilitate such a standard. In particular, the development of a standardized approach to communities that is comparable across states could facilitate the ability to score state maps in the context of communities, producing community maps and scores across states that are directly comparable. The combination of exploratory communities could further complement emerging statistical methods, such as the boundary aggregation methods proposed by

McCartan et al. [2023], or be used to bound uncertainty in map drawers' decisionmaking.

Maps drawn using this method are not limited to algorithmic or computational experiments and analysis alone. The systematic graph method discussed herein could also be used to facilitate participatory or iterative engagements with community members to evaluate where community boundaries lie; simply providing a reasonable modeled community map can provide community members a basis from which to diverge, converse, and evaluate where a community's boundaries lie.

The primal question driving this research is whether the preservation of communities of interest may provide a pathway to better representation by reducing the degree to which communities of interest and the localized coalitions they form are divided among multiple constituencies. The approach discussed herein facilitates a research agenda to answer the following key questions pertinent to legislative districting: can community preservation mitigate gerrymanders by ensuring community coalitions have the opportunity for fair representation, and, if so, under what conditions? How do map drawing algorithms–a key feature in the future of redistricting and gerrymandering case law–respond when using edge weights representative of shared interest? What are the tradeoffs between representational outcomes and population equality in single member districts with large populations? Systematic community identification can help answer these questions while providing practical support to determining territorial legislative constituencies.

References

Dave's redistricting. https://www.davesredistricting.org, 2022. URL https://www.davesredistricting.org.

Representable project. https://www.representable.org, 2023.

Daniel Aaronson, Daniel Hartley, and Bhashkar Mazumder. The effects of the 1930s holc "redlining" maps. *American Economic Journal: Economic Policy*, 13(4):355–92, November 2021. doi: 10.1257/pol.20190414. URL https://www.aeaweb.org/articles?id=10.1257/pol.20190414.

Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546, 2001. doi: DOI:10.1017/S0003055401003100. URL https://www.cambridge.org/core/article/measurement-validity-a-shared-standard-for-qualitative-and-quantitative-research/91C7A9800DB26A76EBBABC5889A50C8B.

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018. doi: 10.1177/1461444816676645. URL https://doi.org/10.1177/1461444816676645.

R. M. AssunÇão, M. C. Neves, G. Câmara, and C. Da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006. ISSN 1365-8816. doi: 10.1080/13658810600665111. URL https://doi.org/10.1080/13658810600665111.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/p10008. URL http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Leo Breiman. *Classification and Regression Trees*. Number 9781315139470. Routledge, https://doi.org/10.1201/9781315139470, 1st edition, 1984.

Sarah Cannon, Moon Duchin, Dana Randall, and Parker Rule. Spanning tree methods for sampling graph partitions, 2022. URL https://arxiv.org/abs/2210.01401.

Jowei Chen and Nicholas Stephanopoulos. The race-blind future of voting rights. *Yale Law Journal*, 130(4):862–946, February 2021. doi: {https://www.yalelawjournal.org/article/the-race-blind-future-of-voting-rights}.

Sandra J. Chen, Samuel Wang, Bernard Grofman, Richard Ober, Kyle Barnes, and Jonathan Cervas. Turning communities of interest into a rigorous standard for fair districting. *Stanford Journal of Civil Rights and Civil Liberties*, 18(101), February 2022. URL https://law.stanford.edu/publications/turning-communities-of-interest-into-a-rigorous-standard-for-fair-districting/.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004. doi: 10.1103/PhysRevE.70.066111. URL https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

Gary W. Cox and Jonathan N. Katz. *Elbridge Gerry's Salamander: The Electoral Consequences of the Reapportionment Revolution*. Political Economy of Institutions and Decisions. Cambridge University Press, United Kingdom, 2002. ISBN 0521001544.

Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

Laurence E Diver. *A Legal Philosophy Perspective: Code is Less than Law*, pages 69–108. Edinburgh University Press, 2022. ISBN 9781474485326. URL http://www.jstor.org/stable/10.3366/jj.7358696.8.

Moon Duchin and Bridfget Eileen Tenner. Discrete geometry for electroal geography. *arXiv physics.soc-ph*, (1808.05860), 2018. doi: arXiv:1808.05860[physics.soc-ph].

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2009.11.002. URL https://www.sciencedirect.com/science/article/pii/S0370157309002841.

Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0605965104. URL https://www.pnas.org/content/104/1/36.

Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999. doi: 10.1023/A:1008894516817. URL https://doi.org/10.1023/A:1008894516817.

James G. Gimpel and Laurel Harbridge-Yong. Conflicting goals of redistricting: Do districts that maximize competition reckon with communities of interest? *Election Law Journal: Rules, Politics, and Policy*, 19 (4):451–471, 2020. doi: 10.1089/elj.2019.0576. URL https://doi.org/10.1089/elj.2019.0576.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799. URL https://www.pnas.org/doi/abs/10.1073/pnas.122653799.

Lani Guinier. Groups, representation, and race-conscious districting: A case of the emperor's clothes. *Texas Law Review*, 71(7):1589–1642, 1992. URL https://heinonline.org/HOL/P?h=hein.journals/tlr71&i=1609https://heinonline.org/HOL/PrintRequest?handle=hein.journals/tlr71&collection=journals&div=58&id=1609&print=section&sction=58.

Mireille Hildebrandt. Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170355, 2018. doi: 10.1098/rsta.2017.0355. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0355.

Yunsieg Kim and Jowei Chen. Gerrymandered by definition: The distortion of 'traditional' districting principles and a proposal for an empirical redefinition. *Wisconsin Law Review*, (101), 2021. doi: http://dx.doi.org/10.2139/ssrn.3543820.

Tobias D. Krafft, Katharina A. Zweig, and Pascal D. König. How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, 16(1):119–136, 2022. doi: https://doi.org/10.1111/rego.12369. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12369.

Alabma Legislature. *Reapportionment Committee Redistricting Guidelines*, May 2021. URL https://www.legislature.state.al.us/aliswww/reapportionment/Reapportionment%20Guidelines%20for%20Redistricting.pdf.

Karin Mac Donald and Bruce E. Cain. Community of interest methodology and public testimony symposium issue: Foxes, henhouses, and commissions: Assessing the nonpartisan model in election administration, redistricting, and campaign finance. *UC Irvine Law Review*, 3(3):609–636, 2013. URL https://heinonline.org/HOL/P?h=hein.journals/ucirvlre3&i=618.

Todd Makse. Defining communities of interest in redistricting through initiative voting. *Election Law Journal: Rules, Politics, and Policy*, 11(4):503–517, 2012. doi: 10.1089/elj.2011.0144. URL https://www.liebertpub.com/doi/abs/10.1089/elj.2011.0144.

Cory McCartan, Jacob R. Brown, and Kosuke Imai. Measuring and modeling neighborhoods, 2023.

Jennifer S. Mindell, Paulo R. Anciaes, Ashley Dhanani, Jemima Stockton, Peter Jones, Muki Haklay, Nora Groce, Shaun Scholes, and Laura Vaughan. Using triangulation to assess a suite of tools to measure community severance. *Journal of Transport Geography*, 60:119–129, 2017. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2017.02.013. URL https://www.sciencedirect.com/science/article/pii/S0966692316305026.

Montana Districting and Apportionment Commission. *Mandatory Criteria for Congressional Districts*, July 2021. URL https://leg.mt.gov/content/Districting/2020/Topics/Criteria/adopted-criteria-congressional-dac-july-2021.pdf.

Richard L. Morrill. Redistricting, region and representation. *Political Geography Quarterly*, 6(3):241–260, 1987. ISSN 0260-9827. doi: https://doi.org/10.1016/S0260-9827(87)80003-4. URL https://www.sciencedirect.com/science/article/pii/S0260982787800034.

New York Court of Appeals. Harkenrider v. hocul. https://law.justia.com/cases/new-york/court-of-appeals/2022/60.html, April 2022.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, Jun 2004. doi: 10.1103/PhysRevE.69.066133. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066133.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006a. doi: 10.1073/pnas.0601602103. URL http://www.pnas.org/content/103/23/8577.abstract.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006b. doi: 10.1103/PhysRevE.74.036104. URL https://link.aps.org/doi/10.1103/PhysRevE.74.036104.

Oyez. Rucho v. common cause. https://www.oyez.org/cases/2018/18-422, 2022a. URL https://www.oyez.org/cases/2018/18-422.

Oyez. Reynolds v. sims. https://www.oyez.org/cases/1963/23, 2022b. URL https://www.oyez.org/cases/1963/23.

Efrén Pérez and Margit Tavits. *Introduction: LOST IN TRANSLATION?*, volume 45, pages 1–18. Princeton University Press, 2022. ISBN 9780691215136. URL http://www.jstor.org/stable/j.ctv2nnv4x3.4.

Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In pInar Yolum, Tunga Güngör, Fikret Gürgen, and Can Özturan, editors, *Computer and Information Sciences - ISCIS 2005*, pages 284–293, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32085-2.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007. doi: 10.1103/PhysRevE.76.036106. URL https://link.aps.org/doi/10.1103/PhysRevE.76.036106.

Olivia Rambo. Defining communities of interest: political self-definition in los angeles county during the first california citizen's redistricting commission. *GeoJournal*, 2020. ISSN 1572-9893. doi: 10.1007/s10708-020-10141-3. URL https://doi.org/10.1007/s10708-020-10141-3.

Redistricting Data Hub. Community of interest map collection project. https://redistrictingdatahub.org/resources/community-of-interest-map-collection-project/, 2024.

Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul 2006. doi: 10.1103/PhysRevE.74.016110. URL https://link.aps.org/doi/10.1103/PhysRevE.74.016110.

Maria A. Riolo and M. E. J. Newman. Consistency of community structure in complex networks. *Phys. Rev. E*, 101:052306, May 2020. doi: 10.1103/PhysRevE.101.052306. URL https://link.aps.org/doi/10.1103/PhysRevE.101.052306.

Kalyn M. Rossiter, David W. S. Wong, and Paul L. Delamater. Congressional redistricting: Keeping communities together? *The Professional Geographer*, 70(4):609–623, 2018. ISSN 0033-0124. doi: 10.1080/00330124.2018.1443477. URL https://doi.org/10.1080/00330124.2018.1443477.

Martin Rosvall and Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007. doi: 10.1073/pnas.0611034104. URL `https://www.pnas.org/doi/abs/10.1073/pnas.0611034104`.

Raphael J. Sonenshein. When the people draw the lines: an examination of the california citizens redistricting commission. Report, The League of Women Voters, 2013. URL `https://cavotes.org/sites/default/files/jobs/RedistrictingCommission%20Report6122013.pdf`.

Nicholas Stephanopoulos. Redistricting and the territorial community. *University of Pennsylvania Law Review*, 160:1379, 2012. URL `https://ssrn.com/abstract=1777276`.

Nicholas O. Stephanopoulos. Our electoral exceptionalism. *University of Chicago Law Review*, 80(2):769–858, 2013. URL `https://heinonline.org/HOL/P?h=hein.journals/uclr80&i=781`.

Supreme Court of North Carolina. *Stephenson v. Bartlett*. 2002.

The American Redistricting Project. S.c. state conf. of the naacp v. alexander. `https://thearp.org/litigation/sc-state-conf-naacp-v-mcmaster/`, January 2023.

V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E*, 84:016114, Jul 2011. doi: 10.1103/PhysRevE.84.016114. URL `https://link.aps.org/doi/10.1103/PhysRevE.84.016114`.

V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z. URL `https://doi.org/10.1038/s41598-019-41695-z`.

Alejandro Trelles, Micah Altman, Eric Magar, and Michael McDonald. No accountability without transparency and consistency: Evaluating mexico's redistricting-by-formula. *Election Law Journal: Rules, Politics, and Policy*, 22(1):80–99, 2023. doi: 10.1089/elj.2021.0061. URL `https://doi.org/10.1089/elj.2021.0061`.

United States Census Bureau. Statistical safeguards. `https://www.census.gov/about/policies/privacy/statistical_safeguards.html`, November 2021a.

United States Census Bureau. Census ftp site. `https://www2.census.gov/geo/tiger/TIGER2020PL/STATE`, 2021b.

United States Census Bureau. Glossary. `https://www.census.gov/programs-surveys/geography/about/glossary.html`, April 2022.

Mahyar Sharif Vaghefi and Derek L. Nazareth. Detecting communities of interests in social media platforms using genetic algorithms. In *AMCIS*, 2016.

We Draw the Lines CA. Data. `https://wedrawthelines.ca.gov/data-dashboard/`, 2022.

Gerald R. Webster. Reflections on current criteria to evaluate redistricting plans. *Political Geography*, 32:3–14, 2013. ISSN 0962-6298. doi: https://doi.org/10.1016/j.polgeo.2012.10.004. URL `https://www.sciencedirect.com/science/article/pii/S0962629812001151`.

Douglas B. West. *Introduction to Graph Theory*. Number 8178088304. Pearson Education, Inc., 2 edition, 2001.

Zhao Yang, René Algesheimer, and Claudio J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(1):30750, 2016. ISSN 2045-2322. doi: 10.1038/srep30750. URL `https://doi.org/10.1038/srep30750`.

Rosemarie Zagarri. *Apportionment Divides the States*, book section 2, pages 36–60. Cornell University Press, 1987. URL `http://www.jstor.com/stable/10.7591/j.ctv1nhmwm.6`.