CLUSTERING PROPENSITY:
A MATHEMATICAL FRAMEWORK FOR MEASURING SEGREGATION

EMILIA ALVAREZ, MOON DUCHIN, EVERETT MEIKE, MARSHALL MUELLER

Abstract. We propose a new family of metrics called CAPY (or *clustering propensity*) scores, designed to measure the clustering level of one or more subgroups within a population. The intended application is to offer new ways of measuring the segregation of demographic subgroups. We discuss two main CAPY scores, Edge and HalfEdge (as well as weighted variants of each) and we compare them to existing segregation scores in the political science, geography, and network science literature. To evaluate the scores, we compute and plot values of minority proportion ρ vs. clustering score C for test distributions on large $n \times n$ grids, and on actual demographic data from U.S. states and cities. We argue that CAPY scores successfully discern qualitatively important differences while providing a stabler baseline for interpretation than classic scores like the Dissimilarity Index and Moran's I.

Keywords: Segregation, network clustering, assortativity, dissimilarity.

Contents

Date: November 2018.

1. Introduction

1.1. **Background and goals.** In this paper we present a family of "clustering propensity" scores that in part unites and in part adds to segregation and assortativity scores that already exist in the geography and network science literature. The goal is to present numerical tools for describing aspects of spatial distribution of populations that can help inform policy considerations.

We have set up the problem as follows: given a region of interest that has been partitioned into geographic units (such as census tracts or precincts), we construct a *dual graph* that records the geographic and demographic information. These dual graphs are flexible network structures that allow for mathematical analysis of spatial population distributions, which in turn leads to a very general framework for measuring segregation.

To analyze performance, we will consider a suite of questions aimed at evaluating whether a proposed score has adequate discernment and stability. That is, scores should offer a stable numerical baseline: similar scores should mean something qualitatively similar across scenarios; in particular, scores should not be heavily or chiefly sensitive to a non-pattern-related variable like city size, minority share, or choice of units. The units issue—in which changing the aggregation level has a drastic impact on output—is well known in the geography literature as a MAUP, or *Modifiable Areal Unit Problem.* Avoiding undue sensitivity to factors that are in some sense orthogonal to clustering or segregation will give us grounds to prefer capy scores to some classical alternatives. And at the same time, we will prefer scores that register meaningful qualitative differences in segregation scenarios.

This direction of investigation was motivated by the study of electoral redistricting. Demographic clustering has a major impact on political representation under the system of single-member districts that dominates the United States electoral scene. This is even made explicit in the checklist of features that must be established to bring a lawsuit under the Voting Rights Act of 1965—litigants must demonstrate that a minority group is "sufficiently large and geographically compact to constitute a majority in a single-member district" in order to press a claim that the group has been denied rightful representation.[1] This phrasing acknowledges legally what is mathematically clear: the size of a minority population alone, without sufficient spatial clustering or "compactness," is not enough to guarantee that the group can secure representation in a districted system. We were motivated by wanting to measure clustering with tools compatible with statistical physics models, like the Ising model, that would allow us to design dynamical systems to intensify and relax the level of clustering and study the representational consequences. The intimate relationship between segregation and district-based representation will be discussed in future work.

2. The theoretical framework of capy scores

2.1. **Geographical units and dual graphs.** We begin by setting up definitions and notation to treat a city, state, or any other jurisdiction as a graph decorated with relevant demographic data. In our examples, we will use geographical units from the census, such as census tracts or census blocks, that partition the jurisdiction into pieces. The *dual graph* of a geographical partition is the graph formed by using a vertex (or node) to represent each unit, then connecting two vertices by an edge if the geographical units are adjacent. We can either adopt edges for *rook adjacency* (in which the shared boundary has to have positive length) or *queen adjacency* (in which we count units as being adjacent even if they just meet at a point). This is illustrated below in Figure 1.

At each node we can record demographic information for the geographic unit, including the total population and racial breakdown, based for instance on census data. The geographical units that make up a jurisdiction have populations of different sizes and compositions. Suppose we have two

[1] In the VRA literature, this is called the Gingles 1 test. See Thornburg v. Gingles, 478 U.S. at 50, 1986.

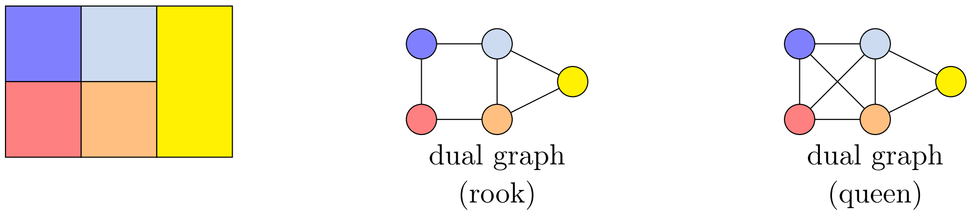

Figure 1. On the left is a partition of a region into five units. The middle and righthand figures represent dual graphs of this partition, where the middle figure has used rook adjacency and the righthand figure uses queen adjacency.

types of population, X and Y, such as Black and White residents.[2] If the nodes of the dual graph are denoted v_i, then we can record integer-valued populations x_i and y_i in each unit, with total population p_i at the ith node. We may have $p_i = x_i + y_i$ if each population member is classified in group X or group Y, or there may be other groups in the population. We will record the X population data as a vector $\mathbf{x} : V \to \mathbb{Z}$, and likewise write $\mathbf{y}$ for the Y population figures. For example, Figure 6 shows the dual graph of the 99 counties in Iowa. The sizes of the nodes in the figure reflect 2010 Census population of the counties, which in fact varies by more than two orders of magnitude, from a minimum of 4029 to a maximum of 430,640.

The total population of a jurisdiction will be denoted $\bar{p} = \sum_i p_i$, and likewise $\bar{x}$ and $\bar{y}$ represent the total number of residents of X or Y type, respectively. We will introduce the notation $\rho = \bar{x}/\bar{p}$ to represent the proportion of population X in the population at large, so that $0 \le \rho \le 1$. Since we typically focus on a population in the numerical minority, most of the plots will have $0 \le \rho < 1/2$.

2.2. **The exploded graph and an inner product expression.** We would like to measure the extent to which people of population X tend to live next to other people of population of X, rather than next to people of population Y. So we will classify within-unit adjacencies as well as adjacencies between neighboring units. There are scores for this in the literature when each node corresponds to a single person, but we have not found existing segregation scores that handle arbitrary percentages at each node of a network.

In the network science and applied mathematics literature, authors sometimes consider constructions that *aggregate* and *disaggregate* nodes in graphs; that is, a graph can be modified by collapsing a subgraph to a node, or by replacing a node with an appropriate subgraph. We will describe a massively disaggregated secondary graph associated to our dual graph which we call the *exploded graph*. We expand each node v_i into a complete graph (or *clique*) K_{p_i} on p_i nodes such that exactly x_i are of X type. If two nodes v_i and v_j are adjacent in the initial dual graph, then the exploded graph contains $p_i \cdot p_j$ edges between the members of the respective cliques. This graph has an enormous number of nodes (one for each person in the jurisdiction) and edges, but it is a theoretical construction that we use to explain the logic of the main definitions; we note that the exploded graph never has to be built or stored.

We can define two expressions as follows:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_i x_i y_i + \sum_{i \sim j} x_i y_j + x_j y_i \ ;$$

$$\langle\!\langle \mathbf{x}, \mathbf{y} \rangle\!\rangle := \frac{1}{2} \left(\sum_i \left(x_i y_i - \frac{x_i + y_i}{2} \right) + \sum_{i \sim j} x_i y_j + x_j y_i \right) .$$

[2]We note that census data includes a count of Black-only population and White non-Hispanic population, among many other racial classifications, including membership in more than one racial group. Census classification allows researchers to treat racial categories as though they are much more stable and clear than the social reality.

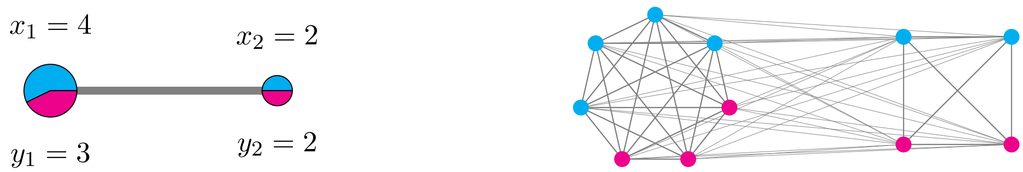

FIGURE 2. This figure shows the *exploded graph* associated to an initial graph with $\mathbf{x} = (4, 2)$, $\mathbf{y} = (3, 2)$, and no other type of population. Here, the exploded graph has $\langle \mathbf{x}, \mathbf{y} \rangle = 30$ edges between different-colored nodes, $\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle = 15$ edges between X nodes, and $\langle\!\langle \mathbf{y}, \mathbf{y} \rangle\!\rangle = 10$ edges between Y nodes, making 55 edges in all. The proportion of X population in the jurisdiction is $\rho = 6/11$.

Here in both expressions the first summation is over all the nodes, and the second is over pairs of adjacent nodes. Note that the number of edges between populations X and Y within the clique associated to vertex i is $x_i y_i$, which means that $\langle \mathbf{x}, \mathbf{y} \rangle$ is a precise count of the edges of XY type when X and Y are disjoint populations. On the other hand, the number of edges between two people of population X is

$$\binom{x_i}{2} = \frac{x_i^2 - x_i}{2},$$

so $\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle$ simplifies to a precise count of the number of edges of XX type.

We note another relationship between these expressions. Since quadratic terms dominate linear terms when the x_i and y_i are large, we get $\langle x, y \rangle \approx 2\langle\!\langle x, y \rangle\!\rangle$ for large populations.

Observe that $\langle x, y \rangle$ is an inner product, so it has a nice representation in terms of matrix multiplication. Letting A be the adjacency matrix of the dual graph, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T (A + I)\mathbf{y}.$$

2.3. **Measuring clustering propensity.** With the information above, we can define clustering propensity scores on the exploded graphs which have a clear probabilistic interpretation.

We can use this to define a one-sided score of the skew via

$$\mathsf{Skew}(\mathbf{x}, \mathbf{y}) := \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle} = \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x},\ \mathbf{x} + 2\mathbf{y} \rangle}.$$

Using the fact that $\langle \mathbf{x}, \mathbf{y} \rangle \approx 2\langle\!\langle \mathbf{x}, \mathbf{y} \rangle\!\rangle$, we see that the skew is approximately $\frac{\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle}{\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle + \langle \mathbf{x}, \mathbf{y} \rangle}$, which is the ratio of the number of XX edges to the number of edges of either XX or XY type. In other words, among the edges that connect X population to either X or Y population, it records the share of XX edges. This measures the prevalence of X living next to X rather than Y, weighted by edges.

Therefore to devise a score of the clustering propensity between populations X and Y from an edge point of view, we can average the X and Y skews, arriving at the edge CAPY score

$$\mathsf{Edge}(\mathbf{x}, \mathbf{y}) := \frac{1}{2} \left(\frac{\langle \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle} + \frac{\langle \mathbf{y}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle} \right) \tag{1}$$

Note that the score can be extended to compare the clustering of multiple disjoint sets, such as with

$$\mathsf{Edge}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{3} \left(\frac{\langle \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{z} \rangle} + \frac{\langle \mathbf{y}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + 2\langle \mathbf{y}, \mathbf{z} \rangle} + \frac{\langle \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{z}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{z} \rangle + 2\langle \mathbf{y}, \mathbf{z} \rangle} \right),$$

and so on to arbitrarily many populations.

However, if we want to reframe this as a propensity in terms of the vertices (the people) rather than the edges (the adjacencies of people), it is more natural to set up the ratio in terms of half-edges rather than edges. A half-edge is a vertex-edge pair (v, e) in which edge e is incident to vertex v. The share of X type half-edges which belong to an XX edge is $\frac{2\langle\langle \mathbf{x},\mathbf{x}\rangle\rangle}{2\langle\langle \mathbf{x},\mathbf{x}\rangle\rangle+\langle \mathbf{x},\mathbf{y}\rangle}$, which is asymptotic to

$$\mathsf{Skew}'(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{y}\rangle}.$$

This has the intuitively appealing interpretation as the probability that a neighbor of an X person is another X person rather than a Y person. Accordingly, we define the half-edge CAPY score to be

$$\mathsf{HalfEdge}(\mathbf{x},\mathbf{y}) := \frac{1}{2}\left(\frac{\langle \mathbf{x},\mathbf{x}\rangle}{\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{y}\rangle} + \frac{\langle \mathbf{y},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle + \langle \mathbf{x},\mathbf{y}\rangle}\right), \tag{2}$$

noting that it can just as easily be extended to more than two populations.

This will be the clustering propensity score that receives our strongest focus in this paper: it averages the average tendency of each subgroup of population to have members of their own subgroup, and not the other, as neighbors.

2.4. **Within-unit and between-unit weighting.** A natural variant on these scores is to weight the connections within geographical units differently than those between neighboring units. To accomplish this, we choose $\lambda \geq 0$ and set

$$\langle \mathbf{x},\mathbf{y}\rangle_\lambda := \lambda\left(\sum_i x_i y_i\right) + \sum_{i\sim j} x_i y_j + x_j y_i.$$

With this, we can simply repeat the formulas for clustering scores using the weighted inner products, such as

$$\mathsf{HalfEdge}_\lambda(\mathbf{x},\mathbf{y}) := \frac{1}{2}\left(\frac{\langle \mathbf{x},\mathbf{x}\rangle_\lambda}{\langle \mathbf{x},\mathbf{x}\rangle_\lambda + \langle \mathbf{x},\mathbf{y}\rangle_\lambda} + \frac{\langle \mathbf{y},\mathbf{y}\rangle_\lambda}{\langle \mathbf{y},\mathbf{y}\rangle_\lambda + \langle \mathbf{x},\mathbf{y}\rangle_\lambda}\right).$$

In this way, any normalization factor one might introduce for $\langle\ ,\ \rangle_\lambda$ cancels out of the numerator and denominator, and we obtain a score that weights the two kinds of neighbors differently.

For instance, if one is working with geographical units that are chosen in part for their social unity, such as census tracts, then it would be reasonable to weigh the within-tract adjacencies more heavily than those between neighboring tracts, such as by taking $\lambda = 2$ or $\lambda = 5$. If the units are counties, then there are some states in which people identify strongly with their county, such as Texas, and other states in which most people don't know what county they live in, such as Massachusetts. Some choice of λ-weighting could then be appropriate for studies of changing segregation over time in Texas.

Note that as $\lambda \to \infty$, the vertex terms dominate the weighted terms, so that in the limit we have $\lim_{\lambda\to\infty}\langle x, y\rangle_\lambda = \sum_i x_i y_i$. This defines the following weighted CAPY scores in the limit, defined by summing over the geographical units.

$$\mathsf{HalfEdge}_\infty(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left(\frac{\sum x_i^2}{\sum x_i(1 + x_i y_i)} + \frac{\sum y_i^2}{\sum y_i(1 + x_i y_i)}\right).$$

Of course, because the interaction between neighboring nodes has been dropped out, this becomes a node-based score (i.e., ignoring edges) like several classical scores discussed in the next section (§3.1).

3. Comparison to existing literature

We will survey some of the numerous existing segregation scores in the social science and applied mathematics literature, translating them into the notation of this paper for ease of comparison. Recall that $\mathbf{p}$ is the vector of population at each node, and $\bar{x}, \bar{y}, \bar{p}$ are the jurisdiction-wide populations of X type, Y type, and all residents, respectively. We also have ρ as the jurisdiction-wide proportion of X population, and $\rho_i = x_i/p_i$ the proportion at node i.

3.1. **Node-based scores: Dissimilarity, Frey, and Gini.** The segregation literature has three major scores that have been described as measuring "evenness," or the consistency of the levels of a sub-population over the units that make up a jurisdiction.

$$D(\mathbf{x}) = \frac{1}{2\bar{x}(\bar{p}-\bar{x})}\sum_i |x_i\bar{p} - p_i\bar{x}| \ ; \qquad F(\mathbf{x},\mathbf{y}) = \frac{1}{2\bar{x}\bar{y}}\sum_i |x_i\bar{y} - y_i\bar{x}| \ ;$$

$$G(\mathbf{x}) = \frac{1}{2\bar{x}(\bar{p}-\bar{x})}\sum_{i,j} |x_i p_j - p_i x_j| \, .$$

These are called the Dissimilarity score, the Frey index, and the segregation Gini index, respectively. We note that all three are based on a similar determinant-like expression: $|vw' - wv'|$ can be interpreted as twice the area of the triangle described by vector (v, w) and vector (v', w'), as in Figure 3.

$$A = \tfrac{1}{2}|vw' - wv'| = \tfrac{1}{2}\left|\begin{smallmatrix} v & v' \\ w & w' \end{smallmatrix}\right|$$

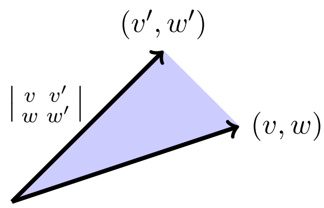

FIGURE 3. This area term is only zero if the vectors point the same direction, which occurs when there is an equality of ratios: $\frac{v}{w} = \frac{v'}{w'}$.

So all three of these formulas, while set up slightly differently from one another, measure how even the distribution of population X is:

- $D(\mathbf{x})$ measures how closely the unit proportions $\rho_i = \frac{x_i}{p_i}$ line up with the citywide proportion $\rho = \frac{\bar{x}}{\bar{p}}$;
- $F(\mathbf{x}, \mathbf{y})$ measures how nearly two groups X and Y have equal proportion of each unit's population;
- $G(\mathbf{x})$ looks over all pairs of units and measures how nearly $\rho_i = \rho_j$.

The determinental interpretation of the scores makes it easy to see that $D(\mathbf{x}) = F(\mathbf{x}, \mathbf{p} - \mathbf{x})$, so Frey's index can be seen as a generalization of dissimilarity to pairs of (not necessarily complementary) populations.[3]

Dissimilarity and this Gini score (which borrows its name from the more famous area-based index of wealth distribution) are among the 20 segregation scores discussed in the classic Massey–Denton survey of segregation indices [8]. This or very similar formulations of Dissimilarity go back to at least the 1950s and have been much used and discussed since then (see [2, 5, 8] and their references).

Note that each of these three scores is given by summing over the nodes without reference to adjacency, none of them can take into account the spatial relationship between geographic units, so they all treat neighboring units no differently than units on opposite sides of a city.

3.2. **Spatial scores in the geography literature, including Moran's I.** Many authors in the geography literature have attempted to modify these scores to take spatial relationships between units into account by "spatial weighting," which can be set up to take into account when units are adjacent, or within a fixed distance, or simply to upweight pairs of units when they are relatively closer or share longer boundary segments. For instance Dawkins in two papers in the 2000s [3, 4] provides spatialized variants of the Gini score from the last section.

[3]In the papers of Frey, the index we call F is referred to as dissimilarity and denoted D, for example in [7].

But the most widely used spatial statistic is very likely Moran's I, introduced in 1950 by a statistician named P.A.P. Moran. Consider a node-wise value $\mathbf{x} = (x_1, \ldots, x_n)$, such as population of group X in our setup. Let $x_0 = \bar{x}/n$ be the average level over the nodes. We might choose to translate $\mathbf{x}$ so that its mean is zero, defining $\mathbf{v} = (x_1 - x_0, \ldots, x_n - x_0)$. Then we can define

$$I = \frac{n}{|E|} \cdot \frac{\displaystyle\sum_{i\sim j}(x_i - x_0)(x_j - x_0)}{\sum_i (x_i - x_0)^2} = \frac{n}{|E|} \cdot \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T \mathbf{v}},$$

in terms of the adjacency matrix A, which in linear algebra terms is just a normalized Rayleigh quotient for the vector $\mathbf{v}$.

To compute this for several test patterns, notice that it can be interpreted as the average of $v_i v_j$ values for adjacent pairs of units divided by the average v_i^2 over the single units. Moran's coefficient for a checkerboard pattern of 0 and 1 on a grid graph would be -1, because every $v_i = x_i - x_0$ would be $\pm 1/2$, but all of the signs in the numerator would be negative because of the alternation. On the other hand, uniformly distributing 0 and 1 values on the vertices of a large graph would give a score near $I = 0$, because of the expected cancellation of positive (like) and negative (unlike) terms. And a heavily clustered 0-1 configuration would tend toward $I = 1$, because nearly all $v_i v_j$ terms would be between like pairs, giving $v_i v_j = v_i^2$, and the two types of adjacency occur in the same proportion as the two types of nodes.

A local version of this score has been proposed, defined in the neighborhood of the jth unit. This can be useful to locate clustering. It can be defined by

$$I_j = n(x_j - x_0) \cdot \frac{\displaystyle\sum_{i\sim j}(x_i - x_0)}{\sum_i (x_i - x_0)^2},$$

which is just like the global I except that the numerator only looks at adjacencies involving node j and we have dropped the normalization by the total number of edges. This has been applied to redistricting in work of Chen–Rodden [1].

One important critique of Moran's I is that it is heavily subject to MAUP, or the *modifiable areal unit problem* discussed in the introduction. This is an important concern in geography: if a score depends too heavily on the choice of geographical units—such as census blocks versus block groups, tracts, etc—that undermines its diagnostic usefulness. To see this problem in Moran's I, consider again the 0-1 checkerboard configuration on a large grid. If the individual units are used, we get $I = -1$, but if we reaggregate mildly so that the 2×2 pieces are used as units, then each unit has an identical composition and we get $I = 0$.

3.3. **Assortativity scores in network science.** In network science, techniques from graph theory, geometry, and data analysis are used to study the structure of networks that come from real-world data. The field largely developed through applications to ecology, epidemiology, and social networks. The term *assortativity* is attached to a range of network scores that are broadly designed to assess whether nodes are more often adjacent to nodes like or unlike themselves, making it precisely aligned with the motivation used to define CAPY scores above. Some of the early focus in the study of assortativity was on graph-theoretic properties, asking for instance whether neighbors are likely to have similar degree or connectivity properties. But demographic sorting has also been considered. For instance, one common example is to study the racial assortativity of social networks; this is clearly relevant to the current application, which is racial assortativity of geographical networks. With an example like this in mind, a recent survey by Mark Newman [9] gives as its main example an assortativity coefficient Q that had been developed to study the spread of HIV. Generally defined with respect to any number of non-overlapping groups that make up a population, it simplifies to something familiar in the case of a group and its complement: it is built from the fraction of XX edges among the XX and XY edges and the corresponding term for YY.

$$Q = \left[\frac{\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle}{\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle + \langle \mathbf{x}, \mathbf{y} \rangle} + \frac{\langle\!\langle \mathbf{y}, \mathbf{y} \rangle\!\rangle}{\langle\!\langle \mathbf{y}, \mathbf{y} \rangle\!\rangle + \langle \mathbf{x}, \mathbf{y} \rangle}\right] - 1.$$

Dropping the linear terms (so that $\langle \mathbf{x}, \mathbf{x} \rangle \approx 2\langle\!\langle \mathbf{x}, \mathbf{x} \rangle\!\rangle$), we have $Q \approx 2\mathsf{Edge} - 1$, which means that it captures just the same information as Edge, but affinely rescaled to vary over $[-1, 1]$ rather than $[0, 1]$.

Thus assortativity is in a sense already in the CAPY family. However, Q only handles nodes whose attributes vary over a finite set, and our exploded graph construction enables us to deal with percentage values, which is a significant generalization. In addition, we think that the $\mathsf{HalfEdge}$ score is a valuable variant on the edge-centered view.

4. Asymptotics on grid graphs

We derive the theoretical behavior of the edge and half-edge CAPY scores in different configurations. Consider an $n \times n$ grid with each node holding a population of M people, so that the total population of the grid is $\bar{p} = Mn^2$. We recall that $\rho = \bar{x}/\bar{p}$ (so that $0 \le \rho < 1/2$) is the parameter representing the (minority) proportion of population X in the grid. In this section we will analyze scores asymptotically as $n \to \infty$.

4.1. **Test configurations on asymptotic grids.**

4.1.1. *Perfect checkerboards.* A perfect checkerboard configuration with density ρ, which we call *Checkerboard* and denote by Ch_ρ, alternates between $x_i = 0$, $y_i = M$ and $x_j = 2\rho M$, $y_j = (1-2\rho)M$ on adjacent nodes. In this way it maintains the global proportion ρ of population X.

That is, the pattern of population X is made up of repeating blocks of the form $\begin{bmatrix} 2\rho & 0 \\ 0 & 2\rho \end{bmatrix}$.

This gives

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{x} \rangle &= \frac{n^2 4\rho^2 M^2}{2}; \\
\langle \mathbf{y}, \mathbf{y} \rangle &= \frac{n^2M^2 + n^2M^2(1-2\rho)^2}{2} + 4n^2M^2(1-2\rho); \text{ and} \\
\langle \mathbf{x}, \mathbf{y} \rangle &= \frac{n^2M^2 2\rho(1-2\rho)}{2} + 2n^2M^2 2\rho.
\end{aligned}$$

The CAPY scores become

$$\mathsf{Edge}(Ch_\rho) = \frac{25 - 50\rho + 20\rho^2 - 4\rho^3}{2(5-\rho)(5-2\rho^2)} \quad \text{and} \quad \mathsf{HalfEdge}(Ch_\rho) = \frac{5 - 8\rho}{2(5-5\rho)}.$$

4.1.2. *Constant/uniform distributions.* Next, consider the *constant* or *uniform* configuration $Const_\rho$, where each node has $x_i = \rho M$ and $y_i = (1-\rho)M$. Then,

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{x} \rangle &= n^2M^2\rho^2 + 2n^2M^2 2\rho^2; \\
\langle \mathbf{y}, \mathbf{y} \rangle &= n^2M^2(1-\rho)^2 + 2n^2M^2 2(1-\rho)^2; \text{ and} \\
\langle \mathbf{x}, \mathbf{y} \rangle &= n^2M^2\rho(1-\rho) + 2n^2M^2 2\rho(1-\rho).
\end{aligned}$$

The CAPY scores are then

$$\mathsf{Edge}(Const_\rho) = \frac{1 - \rho + \rho^2}{2 + \rho - \rho^2} \quad \text{and} \quad \mathsf{HalfEdge}(Const_\rho) = \frac{1}{2}.$$

4.1.3. *Isolated configurations.* Next, consider binary grid configurations in which no two nodes with X population are adjacent. For a given ρ, there must be ρn^2 nodes of X type to get a total X proportion of ρ. Any such configuration is called an *isolated* configuration, and denoted $Isol_\rho$. We compute

$$\begin{aligned}\langle \mathbf{x}, \mathbf{x}\rangle &= n^2 M^2 \rho;\\ \langle \mathbf{y}, \mathbf{y}\rangle &= n^2 M^2 (1-\rho) + 2(2n^2 - 4n^2\rho)M^2; \text{ and}\\ \langle \mathbf{y}, \mathbf{y}\rangle &= 4n^2 M^2 \rho.\end{aligned}$$

We get

$$\mathsf{Edge}(Isol_\rho) = \frac{25-41\rho}{9(5-\rho)} \quad \text{and} \quad \mathsf{HalfEdge}(Isol_\rho) = \frac{3-5\rho}{5-5\rho}.$$

4.1.4. *Clusters.* As in the isolated configuration, the *one-cluster* configurations $OneClust_\rho$ will have $x_i = 0$ or M at each node. But this time the ρn^2 nodes of type X are in a single large cluster. The only contributions to the count of XY edges ($\langle \mathbf{x}, \mathbf{y}\rangle$) will be the *perimeter* of the X cluster. We will choose the cluster to be a asymptotic to the square with side length $\sqrt{\rho}n$, giving $2n^2\rho$ XX edges and $2n^2(1-\rho)$ YY edges to first order, i.e., up to an error term that is linear rather than quadratic in n. We have

$$\begin{aligned}\langle \mathbf{x}, \mathbf{x}\rangle &= n^2 M^2 \rho + 4n^2 M^2 \rho;\\ \langle \mathbf{y}, \mathbf{y}\rangle &= n^2 M^2 (1-\rho) + 4n^2 M^2 (1-\rho); \text{ and}\\ \langle \mathbf{x}, \mathbf{y}\rangle &= 2nM^2\sqrt{\rho},\end{aligned}$$

with CAPY scores

$$\mathsf{Edge}(OneClust_\rho) = \mathsf{HalfEdge}(OneClust_\rho) = 1.$$

In Section 4.3, we will plot configurations with one and multiple clusters to illustrate how, as the perimeter of minority clusters increased, the CAPY scores decrease.

4.2. **Asymptotic comparisons.** We can plot the four test configurations over $0 < \rho < \frac{1}{2}$.

4.3. **Corroboration on finite grids.** To test our analysis of the CAPY scores for clustering, we generated test configurations as described in the last section on a 90×90 grid graph, where each unit has a population of 1000. We plot the following configurations for $\rho = .1, .2, .3, .4, .5$.

- Isolated configurations where some cells are entirely X and no X cell has any rook-adjacent X neighbors;
- One cluster in which cells are entirely X;
- Two to ten clusters of cells that are entirely X;
- Checkerboard where cells alternate between $x_i = 2\rho$ and 0;
- Constant X population of ρ in each cell.

The results, plotted in Figure 4, are nicely consistent with the theoretical behavior derived above, showing that the asymptotic calculations already work at that scale.

5. Observed network example: Counties of Iowa

Many papers in computational social science propose network-based scores but only try them on grid graphs. We next move to the real-data setting that is as close as possible to the grid configuration: the 99 counties of Iowa, whose (rook) dual graph is extremely patterned, with triangle and square structure. Besides being slightly more combinatorially complex than a grid, it also has substantial variation in the population by node, as noted above. We carry out $\mathsf{HalfEdge}$ calculations on the test configurations from above, as follows:

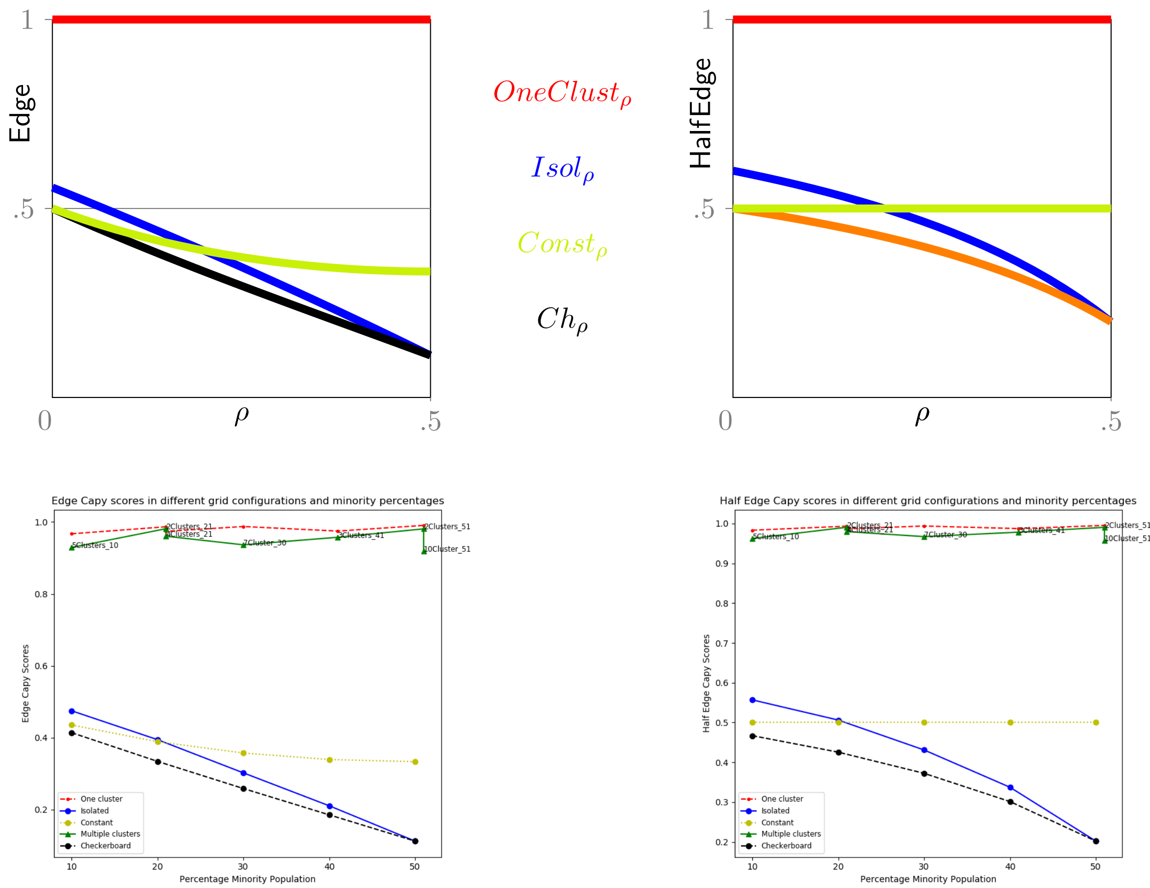

FIGURE 4. CAPY scores for test patterns, on asymptotic grids (top) and 90×90 grid graphs (bottom).

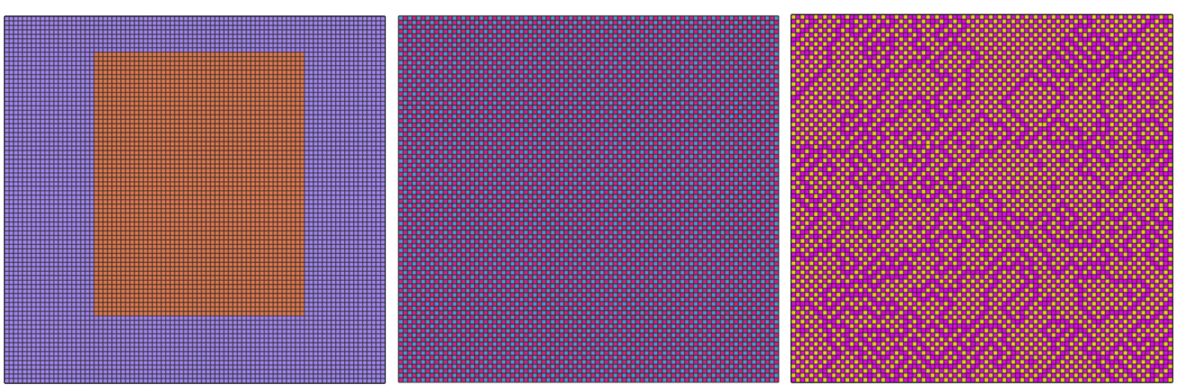

FIGURE 5. These are one-cluster, perfect checkerboard, and isolated configurations, respectively, on a large grid. Each has $\rho = 40\%$ minority population.

- Isolated configurations are produced by randomly choosing nodes to fill with X population (shown in cyan) such that no two X nodes are adjacent;
- Constant configurations are produced by varying ρ from 0 to 1/2 and giving each node that share of X population;

- One-cluster configurations are produced by randomly choosing nodes to be all X and growing the cluster by adding random neighbors.

The results are shown in Figure 6. We do not feature checkerboard configurations, since those are only defined on bipartite graphs.

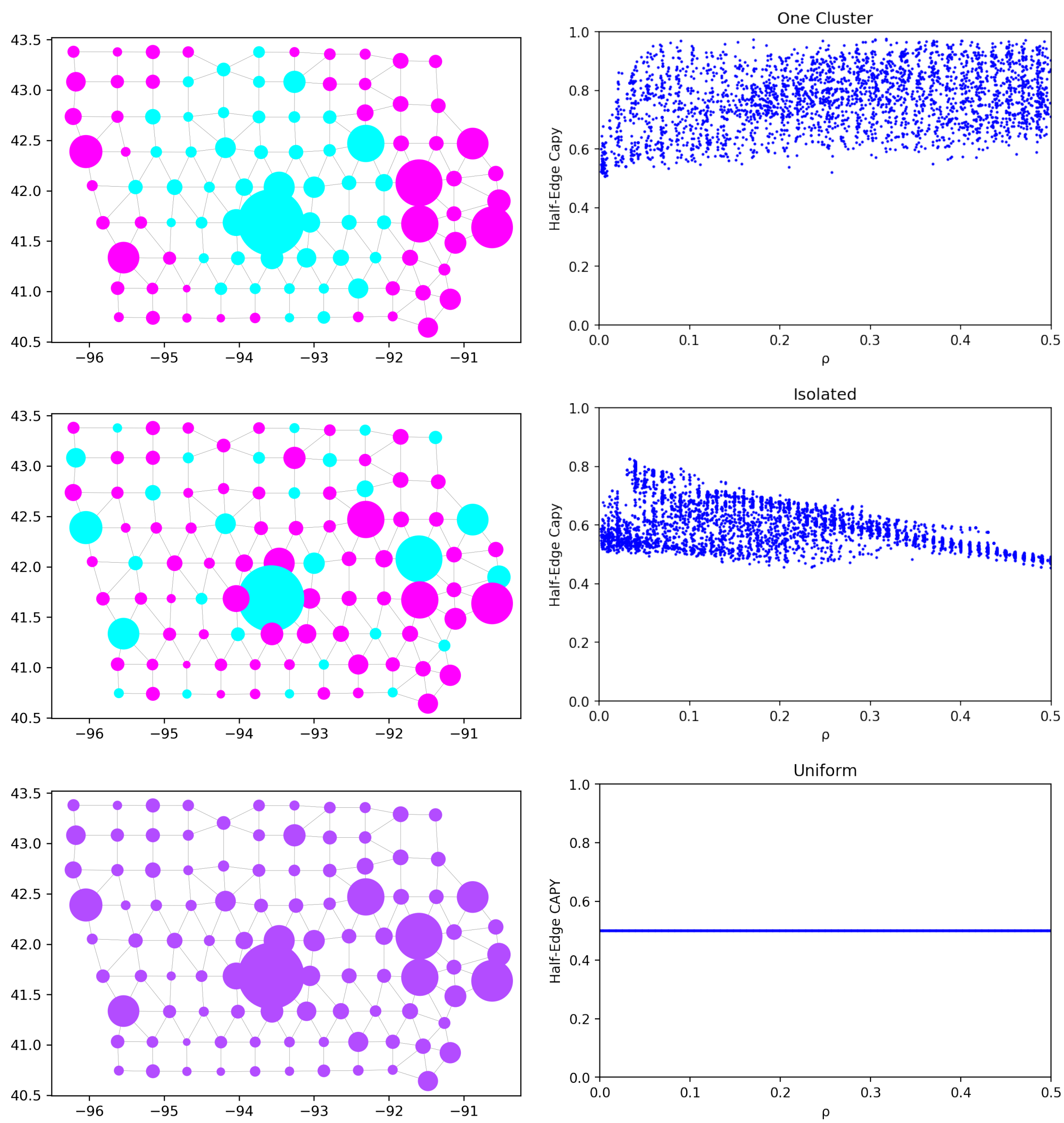

FIGURE 6. The left-hand side shows one example each of an isolated, uniform, and one-cluster configuration of populations X (cyan) and Y (magenta) on the dual graph for Iowa's counties. By generating thousands of test configurations in these patterns at different levels of ρ (the proportion of population of X type), we can observe trends in the HalfEdge score.

References

[1] Chen, J., Rodden, J., (2013), "Unintentional gerrymandering: Political geography and electoral bias in legislatures," *Quarterly Journal of Political Science*, **8(3)**, 239–269.

[2] Cortese, C.F., Falk, R.F., and Cohen, J.K., (1976), "Further Considerations on the Methodological Analysis of Segregation Indices," *American Sociological Review*, **41(4)**, 630–637.

[3] Dawkins, C.J., (2004), "Measuring the Spatial Pattern of Residential Segregation," *Urban Studies*, **41(4)**, 833–851.

[4] Dawkins, C.J., (2006), "The Spatial Pattern of Black-White Segregation in US Metropolitan Areas: An Exploratory Analysis," *Urban Studies*, **43(11)**, 1943–1069.

[5] Duncan, O.D., Duncan, B., (1955), "A Methodological Analysis of Segregation Indexes," *American Sociological Review*, **20(2)**, 210–217.

[6] Fifield, B., Higgins, M., Imai, K., and Tarr, A., (2018), "A New Automated Redistricting Simulator Using Markov Chain Monte Carlo,"

[7] Frey, W.H., Myers, D., (2005), "Racial Segregation in U.S. Metropolitan Areas and Cities, 1990-2000: Patterns, Trends, and Explanations," *Population Studies Center, University of Michigan Institute for Social Research*, Report**05-573**, 1–65.

[8] Massey, D.S., Denton, N.A., (1988), "The Dimensions of Residential Segregation," *Social Forces*, **67(2)**, 281–315.

[9] Newman, M.E.J., *The structure and function of complex networks*. SIAM Rev., 45(2), 167–256.

[10] Metric Geometry and Gerrymandering Group, *Study of voting systems for Santa Clara, CA*. https://mggg.org/MGGG-SantaClara.pdf

Data Appendices

Moon Duchin, Tyler Piazza

Appendix A. Description of data

We have chosen 100 metropolitan areas (Metros) with significant geographical and demographic variation, including all of the top 40 most populous metro areas in the continental United States (by 2010 Census population).

We began with the Core Based Statistical Area (CBSA) shapefiles from the year 2013, fixing these as the definitions of the 100 Metros. We then intersected these with census tracts from 1990, 2000, and 2010, creating three timestamps with a constant geographical extension in which to compare change over time. Demographic data were joined to these tracts and partial tracts from NHGIS data by Python scripts acting on json files. (See GITHUB for all data and code.) Tracts with zero population were removed from the dataset, and the corresponding nodes removed from the dual graphs (though if two vertices were mutually adjacent to a zero-population tract, they were made adjacent to one another after deletion).

In this appendix, we present a range of tables and plots to illustrate features of segregation scores computed on these U.S. Metros. For consistency, we will fix two population subgroups to compare: White and POC. “White" denotes white non-Hispanic Census population, while “POC" (or people of color) represents the complement of this, encompassing all other racial and ethnic groups.

Appendix B. Comparisons and observations

B.1. **Change over time.** We used the three timestamps in our dataset to construct plots to represent the change in segregation over time as reflected in these scores.

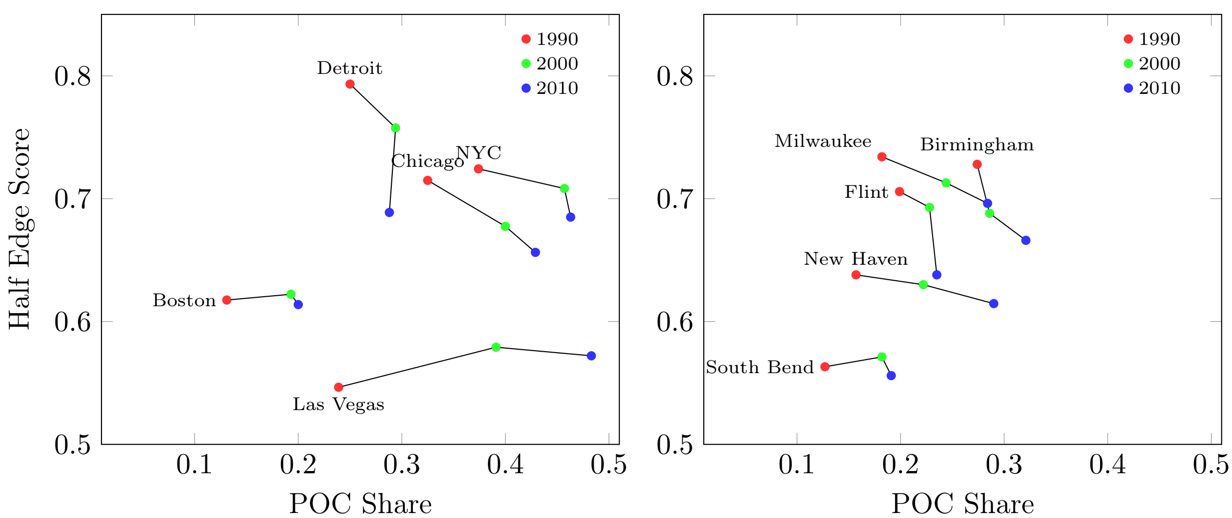

Figure 7. Capy scores for 5 large and 5 medium-sized Metros (population over 1.8 million and 1-1.8 million, respectively) at three timestamps. Most are getting more diverse and less segregated over time.

B.2. **Comparing the scores pairwise.** One notable feature of these scores is that they disagree significantly from one another on how to rank the 100 Metro areas.

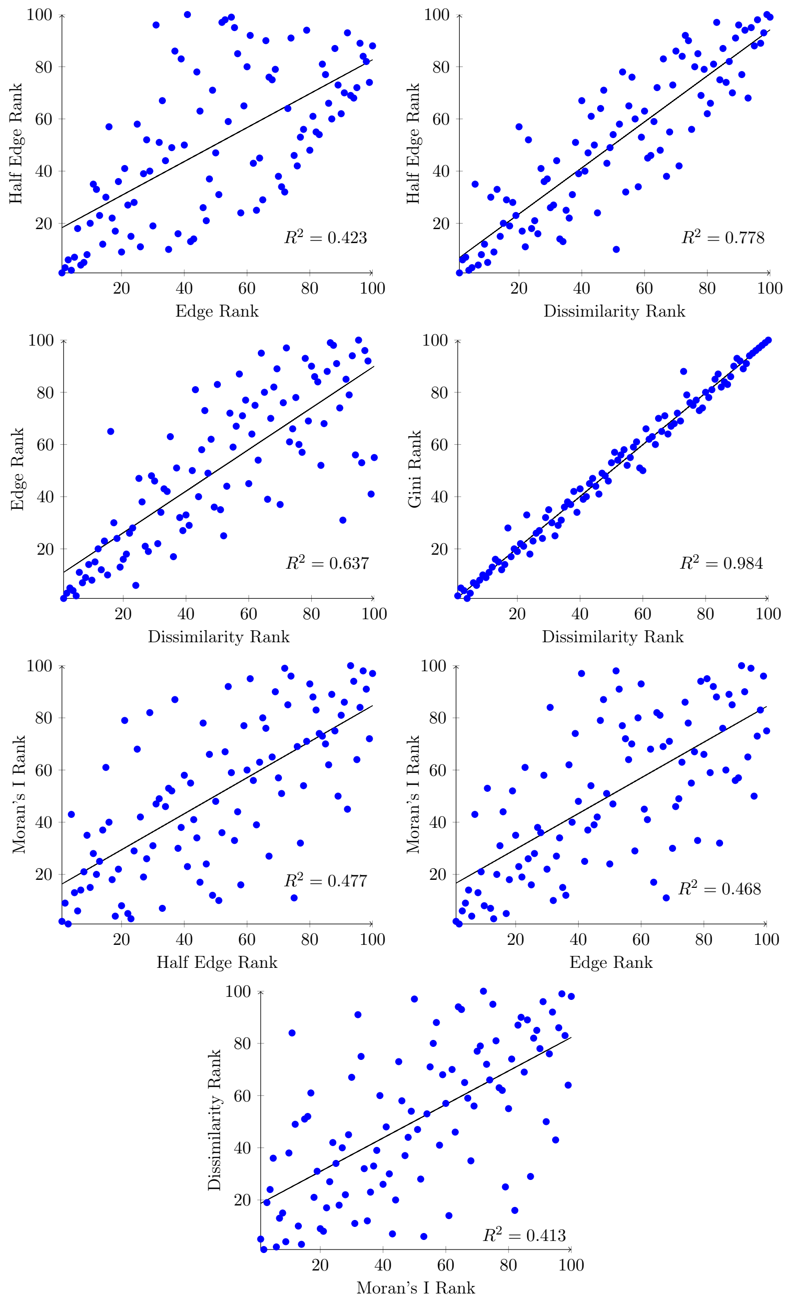

FIGURE 8. Pairwise comparisons of how the segregation scores rank the 100 Metro areas with respect to 2010 data.

B.3. **Within-tract and between-tract measurements.** The next set of plots reports the differences that are imposed by varying the weighting of within-tract comparisons relative to between-tract comparisons. Recall that $\lambda = \infty$ is the node-only variant (i.e., which disregards adjacency of tracts) and at the other extreme $\lambda = 0$ is the edge-only variant.

We find that the Capy scores are making nontrivial use of the tract adjacency patterns, as reflected modest but visible change in rankings as $\lambda \to \infty$. Thus the score is actually and not just theoretically sensitive to the spatial arrangement of tracts. In the other direction, the finding is more surprising: varying $0 \leq \lambda \leq 1$ has virtually no effect at all on the Metro rankings. This means that there is essentially no information loss in practice when discarding the within-tract scoring.

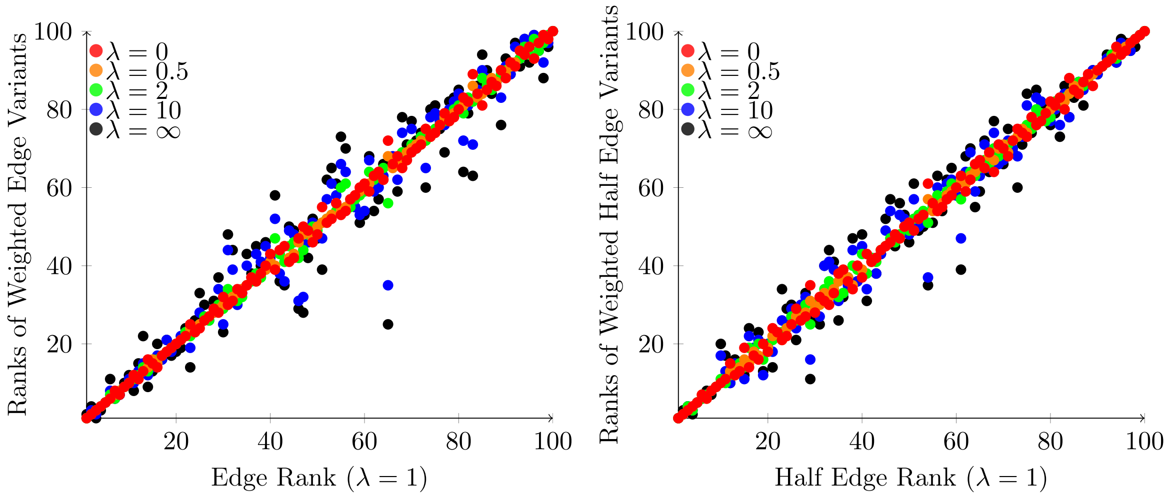

FIGURE 9. These plots compare the Capy scores to their weighted variants, where the edge terms are weighted λ times as heavily as the node terms.

B.4. **Stability.** Ideally, a segregation score should not simply reflect information that is more simply captured by an aggregate Metro statistic, such as the size of the city, the POC share of the population, or the choice of units. We address the choice of units below in §B.6. Here we consider the relationship with city size and POC share.

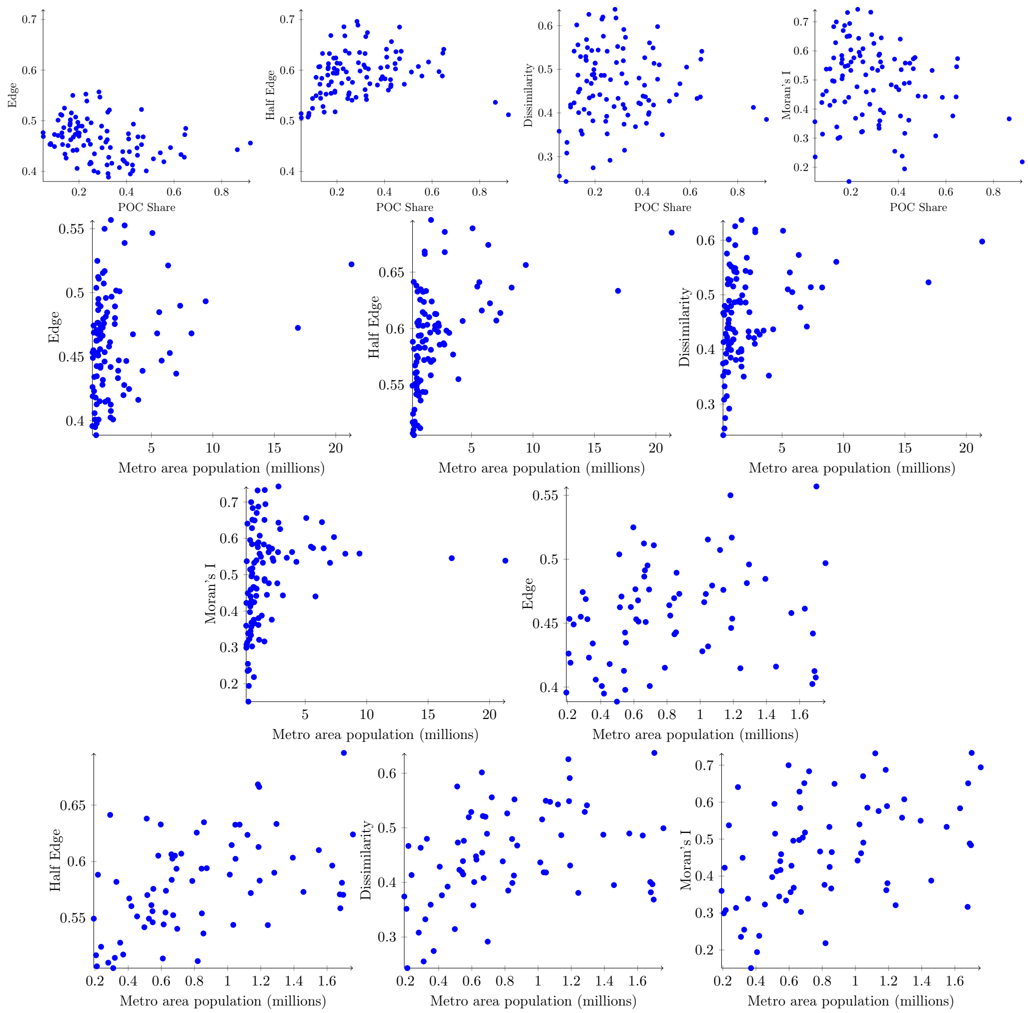

FIGURE 10. These plots compare Capy scores to the POC Share (ρ) in and population of each Metro area. The last row has cities with population less than 1.8 million.

HEdge tends to score the whitest Metros at .5, which is the lowest score ever observed—so it's systematically scoring the whitest metro areas as the least segregated. The Edge score also tends to .5, but that's right in the middle of the observed scores.

B.5. **Edge vs. half-edge discrepancy.** Half-edge is better set up for earth-mover notion of segregation

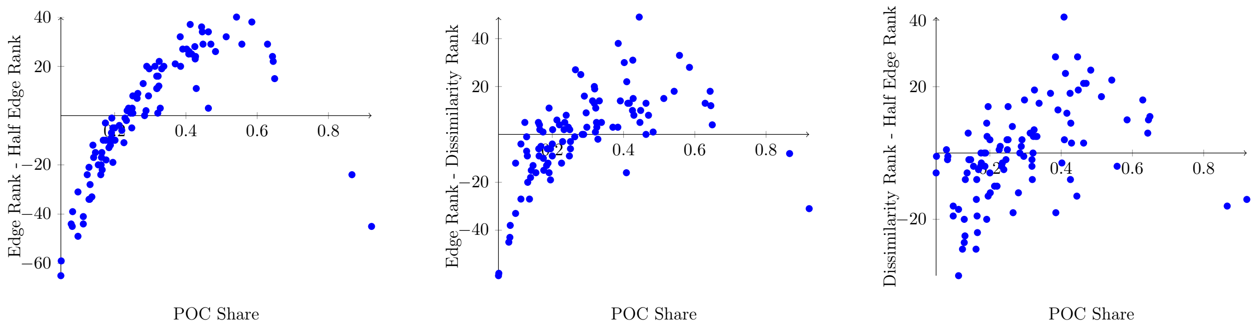

FIGURE 11. Are the differences between scores primarily explained by factors orthogonal to segregation?

B.6. **Modifiable Areal Unit Problem.** The MAUP.

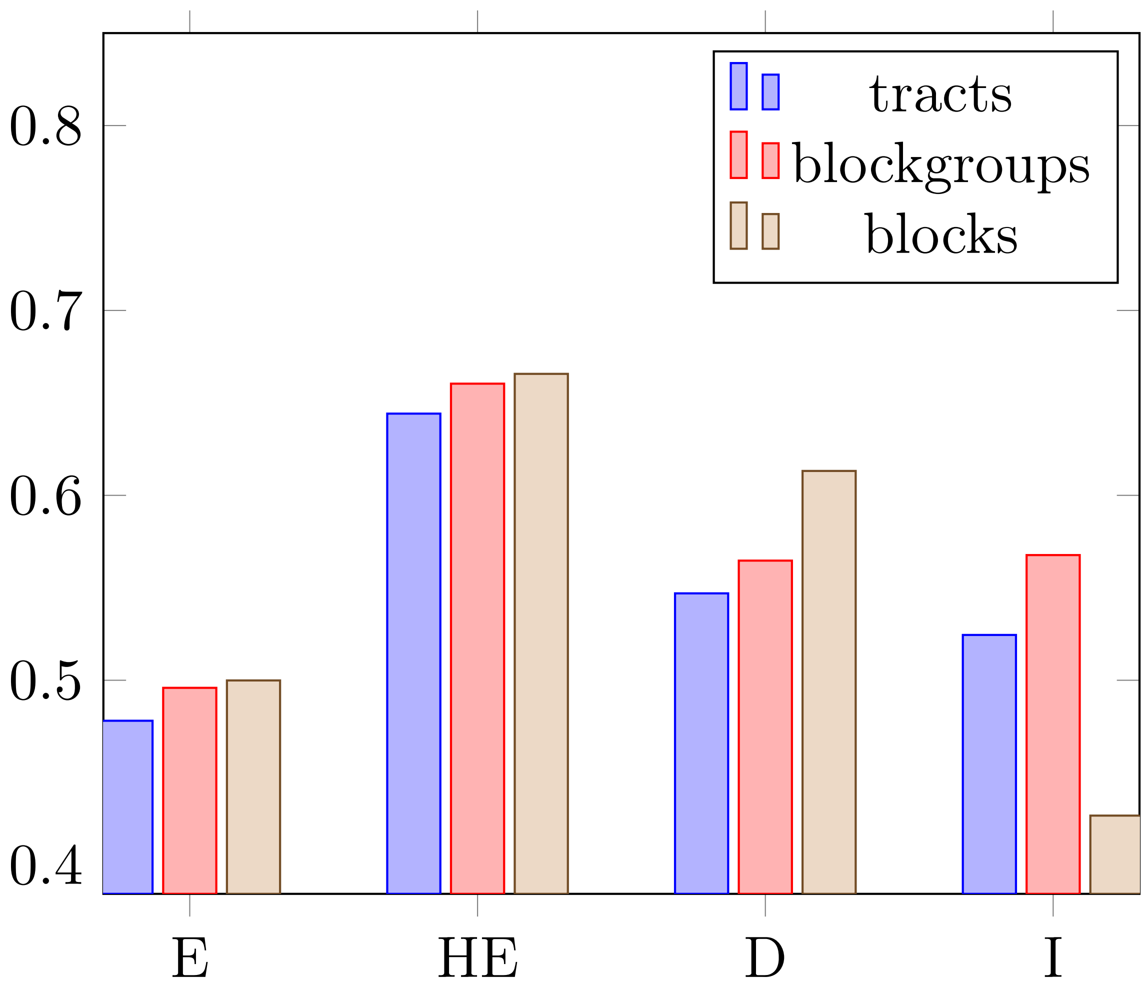

FIGURE 12. There are about 3 times as many block groups as tracts in Chicago, and about 22 times as many blocks as block groups.

B.7. **Tests on manufactured grids.** Tests on generated grids.

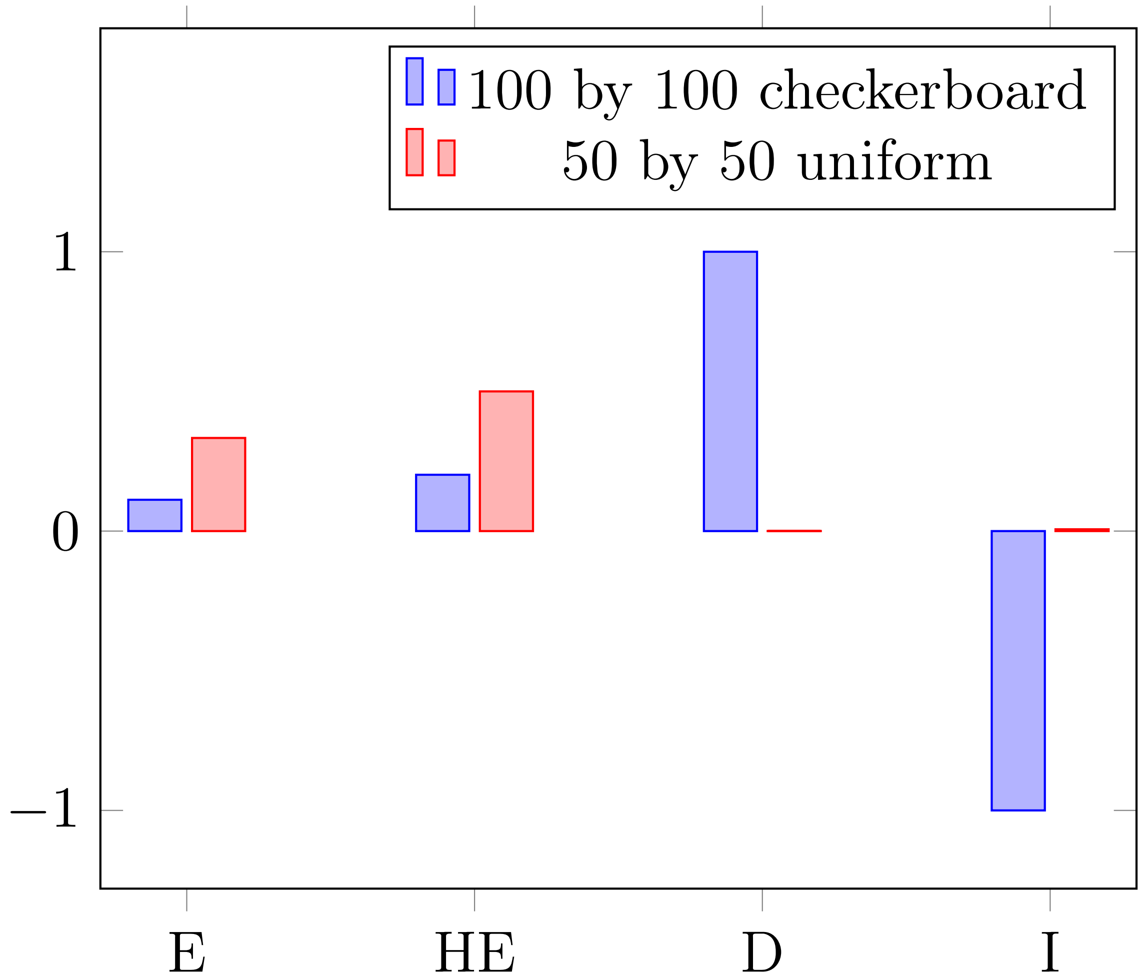

FIGURE 13. Average values from tests where a 100 by 100 grid is filled in a checkerboard with alternating 1 plus epsilon or 0 plus epsilon. Epsilon is a uniform random variable in $[-0.001, 0.001]$. The 50 by 50 grid was made from the 100 by 100 grid by compressing 2 by 2 blocks into a single block.

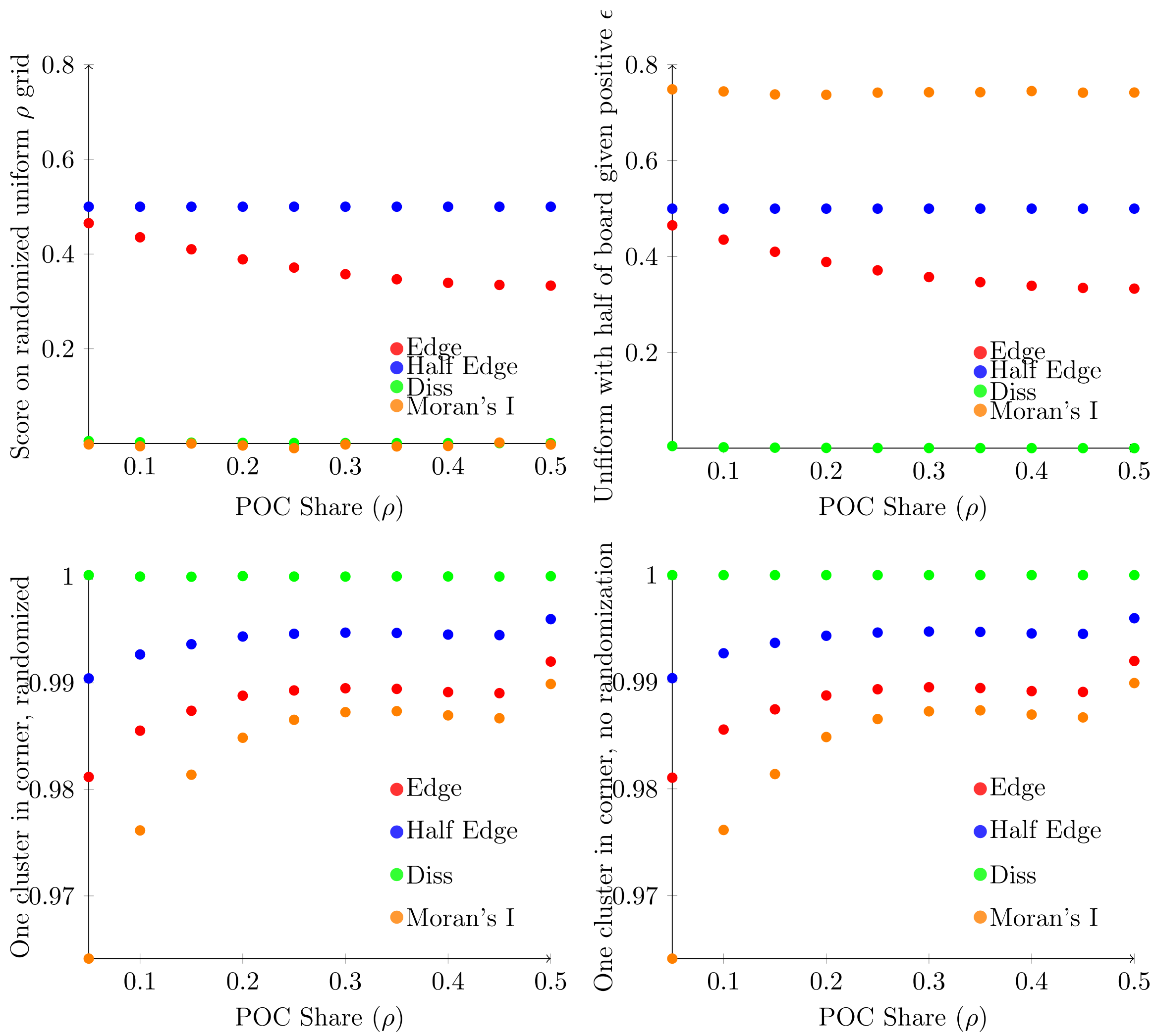

FIGURE 14. All on 100 by 100 grids

Total population of each square is 1, possibly plus epsilon. Epsilon (ϵ) is generally a uniform random variable in $[-0.001, 0.001]$. In the graph with "half of the board given positive ϵ", we specified that the epsilons on the left half of the grid were non negative, and the epsilons on the right half were non positive. For the one cluster tests, we created rectangles which were as close to a square as possible while having area equal to $100 \cdot 100 \cdot \rho$. The width, height pairs of the clusters for the ρ values in increasing order are: $(20, 25), (25, 40), (30, 50), (40, 50), (50, 50), (50, 60), (50, 70), (50, 80), (60, 75), (50, 100)$.

Appendix C. Tabular results

Metro Area	E '10	HE '10	D '10	I '10	E '00	HE '00	D '00	I '00	E '90	HE '90	D '90	I '90
Albany NY	0.476	0.572	0.487	0.576	0.479	0.568	0.462	0.573	0.49	0.556	0.475	0.46
Ann-Arbor MI	0.418	0.551	0.392	0.323	0.429	0.553	0.434	0.349	0.441	0.539	0.435	0.222
Athens GA	0.396	0.549	0.374	0.36	0.413	0.553	0.389	0.286	0.398	0.53	0.387	0.354
Atlanta GA	0.468	0.637	0.51	0.577	0.5	0.659	0.559	0.558	0.545	0.69	0.599	0.655
Austin TX	0.413	0.581	0.397	0.488	0.421	0.584	0.413	0.472	0.428	0.58	0.408	0.536
Baton Rouge LA	0.464	0.626	0.526	0.376	0.474	0.633	0.536	0.456	0.473	0.628	0.54	0.352
Birmingham AL	0.517	0.666	0.591	0.589	0.546	0.688	0.612	0.618	0.588	0.728	0.68	0.666
Bloomington IN	0.449	0.525	0.414	0.537	0.466	0.522	0.424	0.523	0.478	0.519	0.463	0.359
Boston MA	0.49	0.614	0.514	0.604	0.499	0.622	0.534	0.633	0.517	0.618	0.556	0.616
Boulder CO	0.406	0.518	0.274	0.151	0.417	0.516	0.265	0.206	0.437	0.511	0.27	0.224
Bridgeport CT	0.479	0.633	0.548	0.585	0.496	0.634	0.579	0.547	0.506	0.625	0.597	0.483
Buffalo NY	0.55	0.668	0.626	0.687	0.561	0.681	0.656	0.653	0.577	0.689	0.695	0.672
Burlington VT	0.453	0.507	0.243	0.423	0.469	0.504	0.233	0.581	0.484	0.502	0.241	0.336
Cedar-Rapids IA	0.453	0.515	0.333	0.449	0.47	0.514	0.351	0.395	0.484	0.511	0.374	0.338
Charlotte SC	0.447	0.602	0.468	0.545	0.447	0.587	0.466	0.465	0.456	0.578	0.474	0.38
Chattanooga TN	0.495	0.605	0.52	0.504	0.501	0.607	0.551	0.503	0.52	0.619	0.638	0.455
Chicago IL	0.493	0.656	0.56	0.558	0.52	0.678	0.605	0.492	0.572	0.715	0.666	0.508
Cincinnati OH	0.502	0.613	0.568	0.571	0.513	0.62	0.6	0.521	0.539	0.642	0.688	0.487
Cleveland OH	0.539	0.668	0.62	0.643	0.595	0.725	0.67	0.727	0.651	0.771	0.734	0.762
Colorado-Springs CO	0.401	0.54	0.292	0.518	0.412	0.54	0.319	0.49	0.426	0.532	0.333	0.407
Columbia SC	0.415	0.583	0.439	0.466	0.418	0.581	0.421	0.468	0.433	0.59	0.455	0.3
Columbus OH	0.476	0.599	0.514	0.562	0.488	0.6	0.519	0.551	0.524	0.624	0.595	0.528
Dallas TX	0.453	0.622	0.477	0.573	0.46	0.624	0.481	0.522	0.478	0.624	0.496	0.516
Dayton OH	0.515	0.633	0.55	0.67	0.531	0.651	0.594	0.668	0.58	0.697	0.679	0.713
Denver CO	0.447	0.598	0.433	0.626	0.464	0.614	0.44	0.634	0.469	0.594	0.43	0.575
Des Moines IA	0.468	0.555	0.448	0.496	0.478	0.558	0.459	0.573	0.493	0.552	0.498	0.536
Detroit MI	0.547	0.689	0.617	0.656	0.622	0.758	0.707	0.75	0.672	0.793	0.775	0.765
Duluth MN	0.455	0.51	0.308	0.314	0.468	0.51	0.334	0.281	0.48	0.508	0.363	0.232
El-Paso TX	0.443	0.536	0.413	0.366	0.432	0.545	0.434	0.178	0.41	0.554	0.457	0.104
Flint MI	0.504	0.638	0.576	0.595	0.558	0.693	0.65	0.615	0.58	0.706	0.711	0.593
Fort Wayne IN	0.525	0.633	0.529	0.7	0.525	0.635	0.552	0.774	0.538	0.631	0.616	0.752
Fresno CA	0.428	0.588	0.437	0.442	0.416	0.582	0.422	0.314	0.425	0.597	0.442	0.398
Grand Rapids MI	0.481	0.59	0.529	0.558	0.485	0.589	0.51	0.604	0.51	0.593	0.522	0.647
Greensboro NC	0.458	0.61	0.49	0.533	0.454	0.594	0.457	0.436	0.477	0.604	0.488	0.457
Greenville SC	0.415	0.544	0.381	0.321	0.418	0.536	0.378	0.318	0.435	0.543	0.431	0.284
Harrisburg PA	0.491	0.603	0.522	0.584	0.507	0.608	0.576	0.559	0.525	0.605	0.64	0.48
Hartford CT	0.496	0.633	0.541	0.608	0.513	0.635	0.567	0.604	0.551	0.663	0.605	0.665
Houston TX	0.447	0.616	0.505	0.44	0.468	0.637	0.519	0.4	0.45	0.617	0.497	0.386
Huntsville AL	0.435	0.576	0.476	0.459	0.449	0.582	0.484	0.478	0.444	0.567	0.519	0.44
Indianapolis IN	0.489	0.613	0.544	0.576	0.525	0.639	0.604	0.619	0.566	0.676	0.682	0.659
Iowa-City IA	0.426	0.517	0.352	0.299	0.451	0.515	0.389	0.206	0.468	0.513	0.425	0.105
Jacksonville FL	0.416	0.573	0.395	0.388	0.445	0.589	0.438	0.443	0.492	0.628	0.487	0.495
Kansas City MO	0.48	0.603	0.486	0.477	0.506	0.623	0.542	0.52	0.545	0.653	0.605	0.553
Kingsport TN	0.469	0.506	0.255	0.235	0.478	0.505	0.304	0.033	0.486	0.507	0.387	0.063
Knoxville TN	0.473	0.544	0.419	0.462	0.481	0.543	0.415	0.543	0.506	0.57	0.512	0.525
Lafayette LA	0.413	0.561	0.419	0.345	0.414	0.553	0.425	0.39	0.475	0.52	0.464	0.421
Lancaster PA	0.486	0.584	0.455	0.628	0.506	0.586	0.499	0.647	0.529	0.6	0.541	0.634
Lansing MI	0.462	0.57	0.473	0.515	0.461	0.564	0.463	0.51	0.458	0.553	0.46	0.454
Las Vegas NV	0.401	0.572	0.35	0.445	0.411	0.579	0.336	0.562	0.419	0.547	0.334	0.375
Lexington KY	0.451	0.553	0.408	0.303	0.456	0.555	0.416	0.314	0.48	0.573	0.478	0.345
Lincoln NE	0.434	0.528	0.359	0.339	0.451	0.522	0.361	0.367	0.473	0.514	0.367	0.483
Little Rock AR	0.489	0.635	0.552	0.465	0.469	0.605	0.541	0.498	0.483	0.604	0.57	0.522
Los Angeles CA	0.473	0.633	0.523	0.546	0.491	0.65	0.546	0.531	0.497	0.663	0.55	0.604
Louisville KY	0.485	0.603	0.487	0.55	0.511	0.622	0.547	0.507	0.547	0.653	0.645	0.443

Metro Area	E '10	HE '10	D '10	I '10	E '00	HE '00	D '00	I '00	E '90	HE '90	D '90	I '90
Madison WI	0.453	0.544	0.401	0.428	0.466	0.541	0.438	0.47	0.479	0.526	0.471	0.504
McAllen TX	0.456	0.512	0.385	0.218	0.446	0.52	0.391	0.249	0.427	0.518	0.377	0.139
Miami FL	0.485	0.641	0.541	0.574	0.491	0.656	0.566	0.61	0.51	0.675	0.615	0.578
Milwaukee WI	0.557	0.696	0.637	0.733	0.585	0.713	0.684	0.771	0.617	0.734	0.703	0.838
Minneapolis MN	0.468	0.577	0.435	0.547	0.485	0.587	0.469	0.594	0.497	0.564	0.477	0.565
Mobile AL	0.463	0.605	0.519	0.334	0.502	0.652	0.578	0.426	0.555	0.7	0.659	0.475
Nashville TN	0.461	0.596	0.486	0.583	0.474	0.591	0.496	0.451	0.505	0.619	0.546	0.499
New Orleans LA	0.446	0.613	0.549	0.362	0.492	0.657	0.598	0.393	0.479	0.642	0.593	0.392
New York NY	0.522	0.685	0.598	0.539	0.55	0.708	0.631	0.507	0.575	0.724	0.661	0.521
New-Haven CT	0.466	0.615	0.515	0.54	0.497	0.63	0.557	0.595	0.521	0.638	0.587	0.578
Oklahoma City OK	0.454	0.583	0.431	0.381	0.418	0.549	0.345	0.265	0.445	0.557	0.376	0.404
Omaha IA	0.473	0.594	0.468	0.65	0.491	0.604	0.506	0.716	0.509	0.596	0.563	0.676
Orlando FL	0.42	0.587	0.421	0.477	0.417	0.574	0.398	0.434	0.426	0.549	0.384	0.372
Philadelphia PA	0.521	0.674	0.573	0.645	0.557	0.702	0.604	0.663	0.601	0.736	0.667	0.676
Phoenix AZ	0.439	0.607	0.437	0.536	0.45	0.609	0.47	0.456	0.451	0.588	0.462	0.44
Pittsburgh PA	0.501	0.586	0.541	0.538	0.519	0.605	0.573	0.56	0.536	0.616	0.627	0.515
Port St. Lucie FL	0.398	0.546	0.415	0.416	0.452	0.579	0.477	0.382	0.511	0.631	0.589	0.376
Portland ME	0.477	0.514	0.358	0.356	0.478	0.506	0.255	0.337	0.487	0.503	0.261	0.272
Providence RI	0.497	0.624	0.499	0.694	0.504	0.614	0.529	0.649	0.511	0.593	0.531	0.557
Raleigh NC	0.408	0.571	0.369	0.483	0.399	0.54	0.32	0.308	0.423	0.551	0.361	0.309
Reno NV	0.389	0.542	0.315	0.397	0.402	0.54	0.329	0.348	0.418	0.52	0.279	0.341
Rochester NY	0.507	0.624	0.543	0.732	0.513	0.624	0.558	0.743	0.526	0.621	0.583	0.765
Sacramento CA	0.439	0.605	0.423	0.572	0.434	0.596	0.408	0.551	0.428	0.569	0.393	0.496
Salt Lake City UT	0.442	0.559	0.382	0.651	0.436	0.551	0.367	0.445	0.454	0.523	0.311	0.341
San Jose/Francisco CA	0.437	0.607	0.442	0.533	0.418	0.588	0.417	0.426	0.422	0.592	0.418	0.471
San-Antonio TX	0.433	0.597	0.433	0.376	0.454	0.622	0.462	0.378	0.461	0.63	0.487	0.493
San-Diego CA	0.425	0.596	0.427	0.443	0.438	0.608	0.447	0.459	0.444	0.605	0.436	0.508
Santa-Cruz CA	0.474	0.641	0.463	0.641	0.473	0.635	0.46	0.641	0.467	0.62	0.477	0.8
Santa-Fe NM	0.419	0.588	0.467	0.308	0.401	0.572	0.47	0.259	0.386	0.555	0.469	0.297
Sarasota FL	0.441	0.554	0.399	0.425	0.466	0.574	0.496	0.374	0.484	0.557	0.57	0.194
Savannah GA	0.401	0.567	0.429	0.194	0.418	0.579	0.482	0.13	0.434	0.593	0.53	0.116
Seattle WA	0.416	0.555	0.352	0.563	0.415	0.551	0.323	0.489	0.455	0.558	0.354	0.445
South Bend IN	0.443	0.556	0.439	0.44	0.463	0.571	0.483	0.456	0.477	0.563	0.519	0.373
St. Louis MO	0.553	0.686	0.615	0.743	0.559	0.689	0.633	0.669	0.586	0.705	0.696	0.627
Syracuse NY	0.511	0.607	0.556	0.683	0.512	0.606	0.55	0.695	0.521	0.603	0.584	0.615
Tallahassee FL	0.395	0.561	0.376	0.238	0.393	0.554	0.381	0.18	0.411	0.565	0.461	0.127
Tampa FL	0.428	0.586	0.41	0.562	0.462	0.599	0.466	0.527	0.485	0.594	0.517	0.514
Toledo OH	0.476	0.594	0.489	0.651	0.494	0.611	0.526	0.694	0.518	0.627	0.577	0.712
Tucson AZ	0.432	0.602	0.418	0.49	0.435	0.602	0.44	0.461	0.443	0.599	0.462	0.513
Tulsa OK	0.47	0.594	0.48	0.533	0.405	0.542	0.293	0.358	0.439	0.548	0.333	0.347
Tuscaloosa AL	0.423	0.582	0.48	0.255	0.433	0.595	0.497	0.421	0.442	0.589	0.501	0.238
Virginia-Beach VA	0.402	0.571	0.401	0.317	0.4	0.564	0.407	0.264	0.417	0.57	0.432	0.317
Washington DC	0.468	0.636	0.514	0.558	0.458	0.623	0.495	0.485	0.492	0.648	0.522	0.56
Wichita KS	0.451	0.574	0.442	0.368	0.454	0.569	0.447	0.423	0.482	0.575	0.468	0.51
York PA	0.471	0.549	0.423	0.413	0.496	0.559	0.524	0.307	0.509	0.557	0.565	0.43
Youngstown OH	0.512	0.606	0.601	0.497	0.522	0.616	0.626	0.479	0.537	0.626	0.684	0.563

This table of 100 metropolitan areas has the scores for 3 decades (2010, 2000, 1990) for the scores Edge, Half Edge, Dissimilarity, and Moran's I (E,HE,D,I respectively).

Metro Area	E ’10	HE ’10	D ’10	I ’10	E ’00	HE ’00	D ’00	I ’00	E ’90	HE ’90	D ’90	I ’90
Albany NY	33	67	40	27	41	70	57	28	45	74	60	55
Ann-Arbor MI	84	81	82	88	80	79	68	81	80	84	75	93
Athens GA	98	82	87	83	92	78	82	89	99	86	84	76
Atlanta GA	42	13	34	25	27	11	22	32	16	12	24	15
Austin TX	90	62	80	56	81	58	77	54	86	59	81	31
Baton Rouge LA	47	21	25	79	46	23	32	60	63	28	41	77
Birmingham AL	9	8	8	21	9	8	9	19	5	5	10	12
Bloomington IN	61	92	73	45	53	92	71	39	58	91	66	75
Boston MA	22	27	31	19	28	30	33	17	30	38	37	19
Boulder CO	92	93	98	100	88	94	98	96	82	96	98	92
Bridgeport CT	30	19	17	22	31	22	16	35	38	31	25	51
Buffalo NY	3	6	2	6	4	9	4	12	9	13	6	11
Burlington VT	55	99	100	72	49	100	100	26	51	100	100	83
Cedar-Rapids IA	56	95	94	64	48	96	88	73	50	95	88	82
Charlotte SC	62	43	48	41	71	56	54	57	70	60	61	70
Chattanooga TN	19	36	28	52	26	40	26	46	28	37	18	57
Chicago IL	20	9	12	35	14	10	10	48	11	7	13	43
Cincinnati OH	15	30	11	31	17	32	13	41	18	21	7	49
Cleveland OH	5	7	3	14	2	2	3	5	2	2	2	5
Colorado-Springs CO	96	89	97	50	93	88	95	49	90	85	94	65
Columbia SC	87	60	57	60	85	61	73	56	85	56	72	86
Columbus OH	34	44	32	34	37	46	37	34	25	32	26	33
Dallas TX	58	24	45	29	60	26	49	40	59	33	53	37
Dayton OH	10	20	15	8	10	15	15	10	8	11	11	7
Denver CO	64	45	61	17	56	35	64	16	65	50	78	25
Des Moines IA	44	78	53	54	44	74	60	27	42	78	51	32
Detroit MI	4	2	4	9	1	1	1	3	1	1	1	3
Duluth MN	53	98	96	91	51	97	91	90	55	97	90	91
El-Paso TX	66	90	74	81	79	83	69	98	98	76	71	99
Flint MI	14	12	9	20	6	6	5	20	7	8	3	22
Fort Wayne IN	6	18	24	4	11	20	25	1	19	24	20	6
Fresno CA	77	53	59	67	89	60	72	85	91	48	73	67
Grand Rapids MI	28	52	23	36	39	53	39	22	33	53	44	16
Greensboro NC	51	31	37	47	65	50	61	66	61	43	54	56
Greenville SC	88	87	85	89	82	90	85	84	83	83	77	88
Harrisburg PA	21	41	27	23	21	38	18	31	24	42	17	52
Hartford CT	18	17	21	18	18	19	20	23	14	16	22	13
Houston TX	63	25	35	68	52	18	38	72	74	39	52	69
Huntsville AL	73	64	46	63	70	59	46	53	76	66	47	61
Indianapolis IN	24	28	18	26	12	17	11	18	12	14	9	14
Iowa-City IA	79	94	92	94	67	95	83	95	66	94	79	98
Jacksonville FL	86	66	81	76	73	52	67	65	44	27	55	47
Kansas City MO	29	40	41	58	22	28	30	42	17	19	23	30
Kingsport TN	41	100	99	97	42	99	96	100	47	98	83	100
Knoxville TN	37	86	70	62	40	84	76	36	37	64	49	34
Lafayette LA	89	73	69	85	91	77	70	75	62	89	65	64
Lancaster PA	25	58	52	16	23	57	41	14	22	46	40	17
Lansing MI	49	71	47	51	59	72	56	43	69	77	70	58
Las Vegas NV	94	68	93	65	94	62	90	29	94	82	93	72
Lexington KY	60	80	76	93	62	75	75	86	54	62	57	79
Lincoln NE	74	91	89	86	68	91	87	79	64	93	89	50
Little Rock AR	23	15	14	61	50	42	31	47	52	44	33	35
Los Angeles CA	38	16	26	40	36	16	29	37	41	17	38	21
Louisville KY	27	39	39	38	20	31	28	44	15	18	16	60
Madison WI	57	85	77	70	55	86	66	55	56	87	62	44
McAllen TX	52	97	83	98	72	93	81	94	88	92	86	95
Miami FL	26	11	22	28	35	13	21	21	34	15	21	23

Metro Area	E '10	HE '10	D '10	I '10	E '00	HE '00	D '00	I '00	E '90	HE '90	D '90	I '90
Milwaukee WI	1	1	1	2	3	3	2	2	3	4	4	1
Minneapolis MN	45	63	60	39	38	55	53	25	40	68	59	26
Mobile AL	48	37	29	87	25	14	17	68	13	10	15	53
Nashville TN	50	47	42	24	45	51	44	63	39	36	39	45
New Orleans LA	65	29	16	82	33	12	14	74	57	22	27	68
New York NY	7	4	7	43	8	4	7	45	10	6	14	36
New-Haven CT	46	26	30	42	29	24	24	24	27	23	29	24
Oklahoma City OK	54	59	63	77	83	82	89	91	75	73	87	66
Omaha IA	36	49	49	12	34	44	40	6	35	49	36	10
Orlando FL	82	55	68	59	87	66	80	67	89	80	85	74
Philadelphia PA	8	5	10	13	7	5	12	11	4	3	12	9
Phoenix AZ	71	34	58	46	69	37	51	62	73	58	67	62
Pittsburgh PA	16	57	20	44	15	43	19	30	21	40	19	38
Port St. Lucie FL	97	84	72	73	66	64	50	76	32	25	28	71
Portland ME	31	96	90	84	43	98	99	83	46	99	99	89
Providence RI	17	22	36	5	24	34	34	13	31	52	42	29
Raleigh NC	91	70	88	57	99	89	94	87	92	79	91	85
Reno NV	100	88	95	75	96	87	92	82	95	90	97	80
Rochester NY	13	23	19	3	16	25	23	4	23	34	31	4
Sacramento CA	70	38	67	30	77	48	78	33	87	65	82	46
Salt Lake City UT	68	75	84	11	75	81	86	64	72	88	96	81
San Jose/Francisco CA	72	32	54	49	84	54	74	69	93	55	80	54
San-Antonio TX	75	46	62	78	64	29	58	77	68	26	56	48
San-Diego CA	80	48	65	66	74	39	63	59	77	41	74	42
Santa-Cruz CA	35	10	51	15	47	21	59	15	67	35	58	2
Santa-Fe NM	83	54	50	92	97	67	52	93	100	75	63	87
Sarasota FL	69	79	79	71	54	65	43	78	49	72	34	94
Savannah GA	95	72	64	99	86	63	48	99	84	54	43	97
Seattle WA	85	77	91	32	90	80	93	50	71	70	92	59
South Bend IN	67	76	56	69	57	68	47	61	60	69	46	73
St. Louis MO	2	3	5	1	5	7	6	9	6	9	5	18
Syracuse NY	12	33	13	7	19	41	27	7	26	45	30	20
Tallahassee FL	99	74	86	96	100	76	84	97	97	67	69	96
Tampa FL	78	56	75	33	58	47	55	38	48	51	48	39
Toledo OH	32	51	38	10	32	36	35	8	29	29	32	8
Tucson AZ	76	42	71	55	76	45	65	58	78	47	68	40
Tulsa OK	40	50	44	48	95	85	97	80	81	81	95	78
Tuscaloosa AL	81	61	43	95	78	49	42	71	79	57	50	90
Virginia-Beach VA	93	69	78	90	98	71	79	92	96	63	76	84
Washington DC	43	14	33	37	61	27	45	51	43	20	45	28
Wichita KS	59	65	55	80	63	69	62	70	53	61	64	41
York PA	39	83	66	74	30	73	36	88	36	71	35	63
Youngstown OH	11	35	6	53	13	33	8	52	20	30	8	27

This table of 100 metropolitan areas has the ranks of the scores for 3 decades $(2010, 2000, 1990)$ for the scores Edge, Half Edge, Dissimilarity, and Moran's I (E,HE,D,I respectively). A rank of 1 in $E'10$ means that the score was the largest Edge score of the 100 metropolitan area scores for Edge of 2010.

Massachusetts, United States