SIAM J. Matrix Anal. Appl.
Vol. 44, No. 1, pp. 80–105

MEASURING SEGREGATION VIA ANALYSIS ON GRAPHS*

MOON DUCHIN†, JAMES M. MURPHY†, AND THOMAS WEIGHILL‡

Abstract. In this paper, we use analysis on graphs to study quantitative measures of segregation. We focus on a classical statistic from the geography and urban sociology literature known as *Moran's* I, which in our language is a score associated to a real-valued function on a graph, computed with respect to a spatial weight matrix such as the adjacency matrix associated to the geographic units that tile a city. Our results characterizing the extremal behavior of I illustrate the important role of the underlying graph structure, especially the degree distribution, in interpreting the score. In addition to the standard spatial weight matrices encoding unit adjacency, we consider the Laplacian L and a doubly-stochastic approximation M. These alternatives allow us to connect I to ideas from Fourier analysis and random walks. We offer illustrations of our theoretical results with a mix of stylized synthetic examples and real geographic/demographic data.

Key words. spectral graph theory, analysis on graphs, geography, computational social science

MSC codes. 05C50, 65D18, 68T09, 91D30

DOI. 10.1137/21M1466773

1. Introduction.

A central question for geographers, urban sociologists, and demographers is to identify and measure levels of spatial correlation for a social statistic that is associated to geographic units. When the topic is the human geography of a population subgroup, the presence of strong correlation between number and place goes by the general name of *segregation*.

In recent decades, researchers have made increasing use of network structure to model the relationship between the geographical units that make up an area under study. The nodes might stand for individual people or for geographical units like census blocks or counties. Network topology can be given by simple adjacency of units, or by proximity (placing edges between units that are within a threshold distance apart).

Figure 1 shows the basic motivating example: a square lattice graph is first decorated with a checkerboard pattern and then with a clustered pattern. The central question under consideration in this paper is the design of a numerical indicator that detects the intermixing of types on the checkerboard, in contrast to the separation of types on the clustered grid—and that lets us know, on the other hand, when no pattern is present at all.

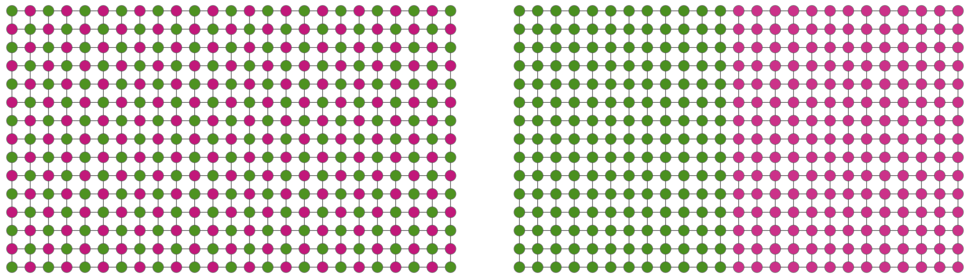
While there is no shortage of proposed metrics to quantify segregation (see section 2.3), the go-to choice in spatial statistics is Moran's I, formally defined below

[*] Received by the editors December 21, 2021; accepted for publication (in revised form) by D. J. Higham August 19, 2022; published electronically February 6, 2023.
https://doi.org/10.1137/21M1466773

Funding: The work of the first author was partially supported by National Science Foundation grant DMS-2005512. The work of the first and third authors was partially supported by National Science Foundation grant OIA-1937095. The work of the second author was partially supported by National Science Foundation grants DMS-1924513 and DMS-1912737.

[†] Department of Mathematics, Tufts University, Medford, MA 02155 USA (moon.duchin@tufts.edu, jm.murphy@tufts.edu).

[‡] Department of Mathematics, UNC Greensboro, Greensboro, NC 27402 USA (t_weighill@uncg.edu).



Checkerboard: $\mathtt{I} = -1$

Clusters: $\mathtt{I} = 0.9576$

FIG. 1. *In these images, green and purple represent two different numerical values, say* 0 *and* 1*. On the left, they are arranged in a checkerboard pattern; on the right, a clustered pattern. Moran's* $\mathtt{I}$ *returns a low value of* -1 *for the checkerboard and a high value of nearly* 1 *for the clusters.*

in Definition 2.1. Introduced in the mid-20th century by P.A.P. Moran [28], this score is so prominent in the study of spatial structure in numerical data that it is almost synonymous with the concept of *spatial autocorrelation*. Over the years, social scientists have developed multiscale generalizations and extensive statistical frameworks that allow for hypothesis tests in which the null hypothesis is of the form "this population is not segregated on this network" [13]. Despite the widespread currency of $\mathtt{I}$ in the field of geography, authors have articulated concerns about the feasibility of reducing a complex social phenomenon such as segregation to a simple score [26]. Basic questions about how to make comparisons using Moran's $\mathtt{I}$—both to compare populations on a common network and to compare across networks—are wide open.

1.1. Summary of contributions.

Broadly, this paper seeks to describe features and properties of Moran's $\mathtt{I}$ as it is commonly used, and to propose related alternatives that have improved properties.

First, after introducing notation and definitions (section 2), we provide a spectral graph theory description of Moran's $\mathtt{I}$ (with respect to a spatial weight matrix W) that we then use to derive basic properties and to consider the standard claims pertaining to its use in spatial statistics (section 3). For various choices of W closely related to a graph, we show that the graph topology (degree distribution, cut lengths) controls the range of attainable values (section 4), which impacts our ability to interpret $\mathtt{I}$ within and especially across localities. Then, in section 5, we consider alternatives to the standard choices of spatial weight matrix (classically, the adjacency matrix A and its row-standardization P). We particularly focus on the Laplacian L and a doubly-stochastic alternative we call M, which offer connections to other rich mathematical concepts. We derive a relationship between $\mathtt{I}(\mathsf{v}; L)$ and Dirichlet energies (section 6) that connects quantitative "smoothness" on a graph, from the point of view of harmonic analysis, to qualitative notions of segregation. Next, we develop a random-walk interpretation of $\mathtt{I}(\mathsf{v}; M)$ in section 7 that offers the version of Moran's $\mathtt{I}$ that seems most promising for bounds and comparisons of any yet proposed.

Finally, supported by a mix of theoretical and empirical work, we make concrete recommendations for the practical use of network-based segregation metrics in geography (section 8).

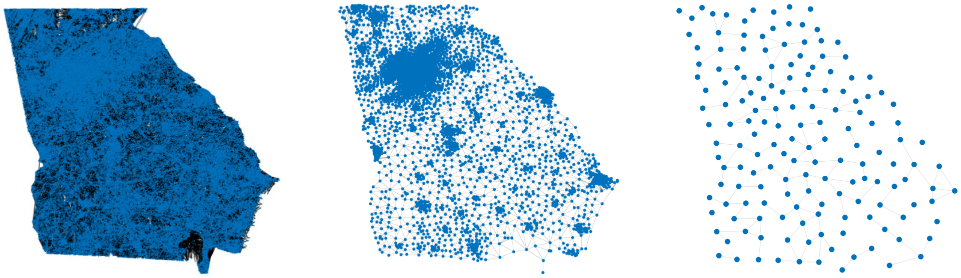


Fig. 2. *The dual graphs of the partition of Georgia into* 291,086 *census blocks (left),* 5,533 *block groups (middle), and* 159 *counties (right). The Census Bureau's geographical hierarchy is nested: each of these can be seen as a quotient of the previous one, collapsing several smaller units into each larger one. All are shown with centroidal embedding. Features of the underlying graphs (like the degree distribution and spectrum) will be shown to provide bounds on the range of* I *values that are achievable.*

2. Background.

2.1. Notation and basic definitions. Suppose we have n geographic units indexed by $1, 2, \ldots, n$ with adjacency data $A \in \mathbb{R}^{n\times n}$. The matrix A has entries $A_{ij} = 1$ if the ith and jth units are adjacent and $A_{ij} = 0$ if not, with the convention that $A_{ii} = 0$ for all i. For example, the units may be census tracts, with $A_{ij} = 1$ if tracts i and j have a shared boundary of positive length (but not if they meet at a corner). Mathematically, A is the (symmetric) adjacency matrix for a graph with nodes corresponding to the geographic units, which we will call the *dual graph* $\mathcal{G}$. We will use $P \in \mathbb{R}^{n\times n}$ to denote the row-standardized adjacency matrix, i.e., the matrix with entries $P_{ij} = A_{ij}/\sum_{k=1}^{n} A_{ik}$. By construction, P is row-stochastic and has real eigenvalues because $P = D^{-1}A$ is conjugate to the symmetric matrix $D^{-1/2}AD^{-1/2}$ where D is the diagonal matrix with $D_{ii} = \sum_{k=1}^{n} A_{ik}$. It achieves a largest eigenvalue of 1 and has all eigenvalues greater than or equal to -1, which is achieved iff the graph is bipartite [11]. Examples using real census data are shown in Figure 2; the Supplementary Materials (SIAM_Supplement.pdf [local/web 1.02MB]) include a histogram of vertex degrees.

Consider a function $\mathsf{v} : V(\mathcal{G}) \to \mathbb{R}$ on the graph nodes, which we will treat as a column vector $\mathsf{v} = (v_1, v_2, \ldots, v_n)^\top \in \mathbb{R}^{n\times 1}$. For example, v_i may be the percentage of residents in tract i who are identified as belonging to a particular demographic (e.g., Hispanic) by the U.S. Census Bureau. Figure 3 shows real examples drawn from Chicago, IL. Let $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ be the average of the entries of v. We will write $\mathsf{0}, \mathsf{1}$ for the vectors (of length n) whose entries are all zero or all one, respectively.

Definition 2.1 (Moran's $\mathtt{I}$). *With notation as above, let $W \in \mathbb{R}^{n\times n}$ be a matrix that is not the zero matrix, and let $w = \sum_{i,j=1}^{n} |W_{ij}|$. Moran's $\mathtt{I}$ with respect to W is a functional $\mathtt{I}(\,\cdot\,; W) : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\mathtt{I}(\mathsf{v}; W) := \left(n \sum_{i,j=1}^{n} W_{ij}(v_i - \bar{v})(v_j - \bar{v}) \right) \Big/ \left(w \sum_{i=1}^{n} (v_i - \bar{v})^2 \right) = \frac{n}{w} \left(\frac{\mathsf{x}^\top W \mathsf{x}}{\mathsf{x}^\top \mathsf{x}} \right),$$

where $\mathsf{x} = \mathsf{v} - \bar{v}\mathsf{1}$.

The most common choices of W in geography are $W = A$ and $W = P$ [5]. We shall refer to any $n \times n$ matrix that is not identically 0 as a *weight matrix*; when it

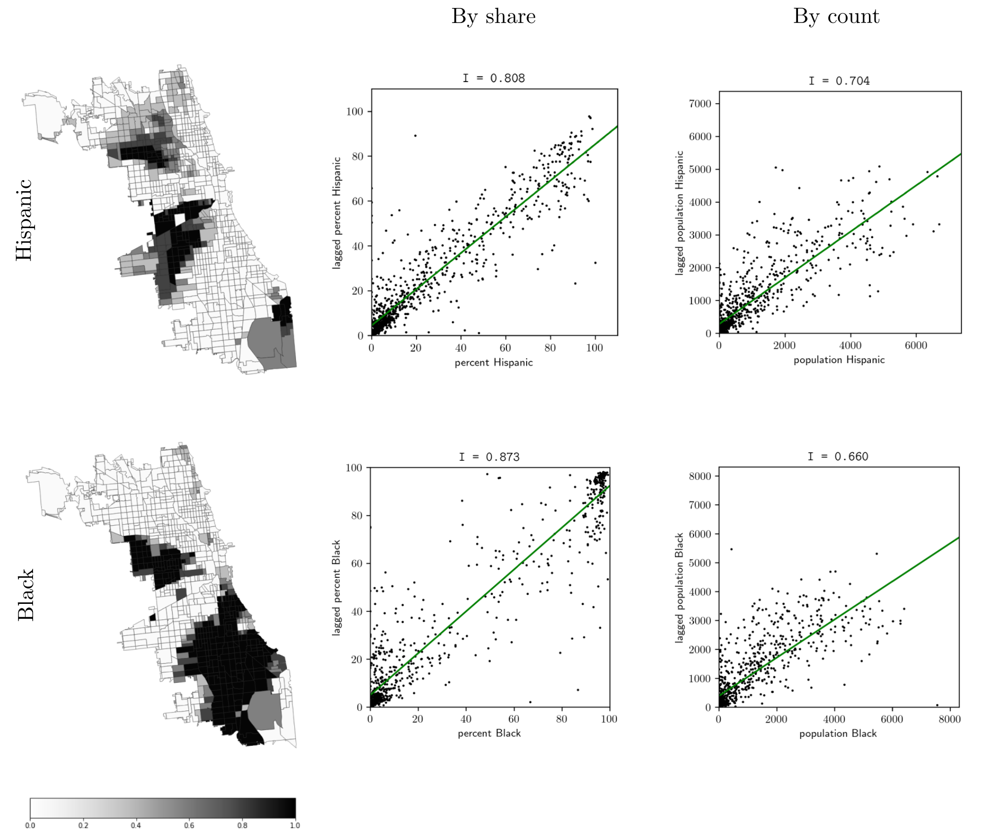


Fig. 3. *Hispanic and Black population data in the* 2010 *census tracts of Chicago, colored by proportion with white corresponding to* 0*% and black corresponding to* 100*%. In the scatterplots, each dot is positioned according to the share (proportion) or count (total number) of the subgroup in that tract (x-axis) and its neighboring tracts (y-axis). Here, Moran's* $\mathtt{I}$ *is calculated with respect to the row-standardized adjacency matrix* P. *(Color available online.)*

is associated to geographic features (as A and P are) we will call it a spatial weight matrix. The usual interpretation in the geography community, dating to Moran's original work, is that for either standard choice of adjacency weights, or for weights based on geographic distance, $\mathtt{I}(\mathsf{v}; W)$ takes higher positive values when the v_i are "spatially correlated" in the sense that neighboring/nearby units tend to have similar values [28]. Conversely, the standard understanding is that $\mathtt{I}$ is negative when neighboring units tend to have very different values, and near zero when there is little or no relationship between the values of neighboring units. The precise notion of spatial correlation is therefore graph-dependent. In short, large values of $\mathtt{I}$ are taken to indicate segregation and small or negative values of $\mathtt{I}$ indicate a lack of segregation. This article clarifies this intuition in a precise mathematical sense through the lens of analysis on graphs. A glossary of notation can be found in Table 1.

Remark 2.2 (Zero-centering and rescaling). It suffices to consider $\mathtt{I}$ on vectors with ℓ^2-norm 1 and mean 0. Let $X := \{\mathsf{x} \in \mathbb{R}^n | \mathsf{x} \cdot \mathbf{1} = 0\} = \mathbf{1}^\perp$ be the subspace of vectors with mean 0. For an arbitrary $\mathsf{v} \in \mathbb{R}^n$ with mean value $\bar{v}$, we let $\mathsf{x} = \mathsf{v} - \bar{v}\mathbf{1}$

TABLE 1
Notation used throughout the paper.

Notation	Definition
$\mathcal{G}$	simple, undirected graph with n nodes
$\mathsf{v} = (v_1, v_2, \ldots, v_n)^\top$	function on graph $\mathcal{G}$, denoted as a column vector
$\bar{v}$	mean of v values
$\\|\mathsf{v}\\|_2$	Euclidean norm of v
$\mathsf{0}, \mathsf{1}$	vector of all 0's and 1's, respectively
$\mathsf{x} = (x_1, x_2, \ldots, x_n)^\top$	arbitrary mean-0 vector
$X = \mathsf{1}^\perp$	space of all mean-0 vectors
Π	orthogonal projection onto X
W	arbitrary weight matrix, not identically zero
Q	arbitrary bistochastic matrix (rows and columns sum to one)
A	graph adjacency matrix associated to $\mathcal{G}$
D	diagonal vertex degree matrix associated to $\mathcal{G}$
P	row-standardized adjacency matrix associated to $\mathcal{G}$
L	Laplacian matrix associated to $\mathcal{G}$
M	bistochastic Metropolis–Hastings matrix associated to $\mathcal{G}$
$\mathtt{I}(\mathsf{v}; W)$	Moran's $\mathtt{I}$ applied to vector v with respect to matrix W
$\mathtt{I}(X; W)$	range of all possible $\mathtt{I}$ values for weight matrix W
$\{(\lambda_i, \Phi_i)\}_{i=1}^n$	eigenvalues and eigenvectors of arbitrary W
$\{(\mu_i, \Psi_i)\}_{i=1}^n$	Laplacian eigenvalues and eigenvectors
$\mathsf{d} = (d_1, d_2, \ldots, d_n)$	vector of W-degrees $d_i = \sum_{j=1}^n \lvert W_{ij}\rvert$
$d_{\min}, d_{\max}, d_{\text{avg}}$	minimum, maximum, and average W-degree over $i = 1, \ldots, n$
$\{\alpha_i\}_{i=1}^n$	coefficients in some orthonormal basis
$\mathcal{E}$	Dirichlet energy functional on the graph
$\mathbb{T}^m = [0, 2\pi]^m$	m-dimensional torus with periodic boundaries

denote its orthogonal projection onto X. Then we immediately see for any v and W that $\mathtt{I}(\mathsf{v}; W) = \mathtt{I}(\mathsf{x}; W)$ and that $\mathtt{I}(\alpha\mathsf{v}; W) = \mathtt{I}(\mathsf{v}; W)$ and $\mathtt{I}(\mathsf{v}; \alpha W) = \mathtt{I}(\mathsf{v}; W)$ for all scalars $\alpha \neq 0$.

We will refer to the ith rowsum $d_i = \sum_{j=1}^n |W_{ij}|$ as the W-*degree* of node i; when $W = A$, this is the standard degree counting the number of edges ending at node i, which we will call either the A-degree or simply the vertex degree, to distinguish it from the W-weighted versions. Define $d_{\text{avg}} := \frac{1}{n}\sum_{i=1}^n d_i = w/n$ to be the average W-degree of the graph; note that if $W = P$, then d_i is identically 1, so $d_{\text{avg}} = 1$. We have

$$\mathtt{I}(\mathsf{v}; W) = \mathtt{I}(\mathsf{x}; W) = \frac{\frac{1}{w}\displaystyle\sum_{i,j=1}^n W_{ij} x_i x_j}{\frac{1}{n}\displaystyle\sum_{i=1}^n x_i^2} = \frac{\displaystyle\sum_{i,j=1}^n W_{ij} x_i x_j}{d_{\text{avg}}\displaystyle\sum_{i=1}^n x_i^2}.$$

Remark 2.3 (Pairs versus singletons). From the function v, the zero-centered x records deviations above and below the mean. The score $\mathtt{I}(\mathsf{x}; W)$ measures the patterns in these deviation values. From the second-to-last expression above, we find an appealing interpretation of $\mathtt{I}$ as *comparing the product of values at related nodes* $(x_i x_j)$ *to the squared values at individual nodes* (x_i^2). The pair average is spatially weighted by the coefficients W_{ij}. In particular, when $W = A$, the score $\mathtt{I}$ is precisely one-half the ratio of the average product across an edge to the average squared value at a vertex. When $W = P$, it is a different ratio: the sum of the products across edges versus the sum of squared node values. These have subtly different properties, as we will see below.

2.2. Moran scatter plots and linear regression. Luc Anselin, a scholar of geography and spatial statistics, is credited with the fundamental observation that when W is row-stochastic, such as for the row-standardized adjacency matrix P, the definition of Moran's I can be rearranged so that it contains a regression coefficient [5]. Let u be defined coordinatewise as $u_i := \sum_{j=1}^{n} W_{ij} v_j$, and denote the mean of u by $\bar{u}$. In the $W = P$ case, this is the average of the function values at the neighbors of i; if W encodes proximity along some dimension, then this is a weighted average. For this reason, u is often called the *(spatially) lagged variable*, by analogy with autocorrelation for time series. Moran's I then reduces to the slope of a regression of the lagged variable (u_i) on the original variable (v_i), as follows:

$$\mathtt{I}(\mathsf{v}; W) = \frac{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v}) \sum_{j=1}^{n} W_{ij} (v_j - \bar{v})}{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v})^2} = \frac{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v}) \cdot (u_i - \bar{v})}{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v})^2} = \frac{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v}) \cdot (u_i - \bar{u})}{\displaystyle\sum_{i=1}^{n} (v_i - \bar{v})^2},$$

where we use the fact that $\sum_{i,j=1}^{n} W_{ij} = 1$ and $\sum_{i=1}^{n} (v_i - \bar{v})\bar{v} = \sum_{i=1}^{n} (v_i - \bar{v})\bar{u} = 0$. We recognize the final expression as the slope of a regression of u on v.

A scatterplot of u (spatially lagged variable) versus v (original variable) has been called a *Moran scatterplot* [5]. Figure 3 shows such a plot for the Hispanic and Black populations of Chicago by census tract. Moran's I is just the slope of the best fit line—shown in green in each plot. The positive correlation, and thus positive value of Moran's I, is easily observable in these cases. Indeed, as the figure shows, the Hispanic and Black populations in Chicago are both very clustered. The connection between segregation, clustering, and graph geometry is developed in section 6.

2.3. Brief summary of prior work. We consider several related methods of quantifying segregation; see the broad surveys [27, 12] for further details and the book chapter [16] for an accessible introduction to I and related measures.

Moran's I, initially introduced by P.A.P. Moran, was brought into geography during the rise of spatial analysis in the late 1940s. In 1969, Cliff and Ord presented an influential conference paper [39] which introduced and argued for the use of alternative spatial weight matrices for Moran's I, in particular W matrices based on more than just contiguity (see [22] for a further discussion on spatial weight matrices). A common use for Moran's I is as part of a significance test to see if data are or are not spatially autocorrelated; in such cases, one compares an observed Moran's I to its distribution under a random function v on the nodes [2, 5, 8, 13]. One can also test for spatial autocorrelation in the residuals of ordinary least squares models [6]. Direct comparison of Moran's I values has sometimes been used in comparing segregation levels between regions [37]; section 4.2 will suggest that this is a dangerous practice. Anselin introduced the Moran scatterplot in [5] as a way to visualize Moran's I (see section 2.2). Moran scatterplots can also be useful in determining the contribution of particular subregions to the overall value of Moran's I. This idea was developed further as part of Anselin's *local indicators of spatial association (LISA)*, a class of methods which forms calculations based on the neighborhood of one node in a network, including local Moran's I and local Geary's c [4]. Outside of geography

Moran's $\mathtt{I}$ has been applied in fields as diverse as epidemiology, urban planning, and environmental studies [21].

Geographers have observed that $\mathtt{I}$ is sensitive to the *modifiable areal unit problem*, or MAUP [8]. Indeed, $\mathtt{I}$ depends quite heavily on the choice of geographic units used (e.g., finer-scale census blocks or coarser-scale census tracts) to construct the associated graph. As a prototype example of this phenomenon, a checkerboard pattern has a Moran's $\mathtt{I}$ of roughly -1 (total anticorrelation) as in Figure 1. But if a simple checkerboard is aggregated so that its 2×2 regions become the new units of analysis, then the distribution becomes roughly uniform and Moran's $\mathtt{I}$ is roughly 0 (no correlation). When choosing units for analysis, the scale and placement of the units will impact all the measurable properties of the region, including the value of Moran's $\mathtt{I}$. On the other hand, there has been considerable development of multiscale graph signal processing tools [41, 34] that allow for multiscale partitions of data-driven graphs using diffusion processes, wavelets, and neural networks; these are potentially interesting tools to capture notions of segregation across spatial scales. We will address the choice of units in real examples throughout the present paper.

The spatial statistics literature contains numerous examples interpreting Moran's $\mathtt{I}$ in linear algebraic terms, as we do here. This has been used to provide a framework for regression modeling [23, 44] and other statistical analysis [13, 43]. Linear algebra is particularly relevant in the context of statistical testing [43] where the goal is to understand, for a given weight matrix W and function v on the nodes, whether v is more segregated in a statistically significant sense than would be expected under a null model. That is, does $\mathtt{I}(\mathsf{v}; W)$ deviate significantly from $\mathbb{E}(\mathtt{I}(\,\cdot\,; W))$, where the expectation is taken over a suitable null model of vectors v? For example, under the model of no spatial correlation in the graph, where node values are randomly sampled independently from each other, Moran himself observed that $\mathbb{E}(\mathtt{I}(\,\cdot\,; W)) = -1/(n-1)$ [28]. Distributions around the mean are given in some cases in terms of spectral properties of the graph [43].

In contrast to the static characterization of segregation captured by $\mathtt{I}$—namely that $\mathtt{I} \gg 0$ indicates segregation—the *Schelling model* provides a dynamical perspective for segregation on graphs [38]. In this model, every node on a network has a label (e.g., the membership in a demographic group), and the network evolves randomly in time as nodes change their label with a propensity towards being similar to their neighbors. The degree of *homophily* is a model parameter and quantifies how much nodes want to have the same labels as their neighbors. One expects a network with high homophily to converge in the limit to a more segregated pattern than one with low homophily. The Schelling model—which bears a family resemblance to models of ferromagnetism in statistical physics such as the Ising model [25]—has been generalized to characterize complex segregation dynamics on extremely regular networks (e.g., hexagonal lattices) [9, 48]. The perspective taken by this literature is to determine the basic properties of the steady-state distribution of the dynamics (e.g., whether large homogeneous regions emerge, depending on the homophily or related parameters) rather than how the underlying network geometry and population distribution impact the dynamics.

Network assortativity [29, 30] was developed in network science to measure the propensity of like nodes to connect to one another, by counting the proportion of edges that link similarly labeled nodes and comparing that to the number expected under a null hypothesis of no special preference. As with $\mathtt{I}$, assortativity scores are influenced by the degree distribution of the network and the underlying sizes of the populations being measured [31]. Previous work by Alvarez et al. [3] has generalized this to the setting that node properties can be real-valued rather than discrete, and the authors construct generalized assortativity scores called clustering propensity (or *capy*) scores that have linear algebra definitions similar to the ones that will be discussed in the present paper. Based on the observation (Remark 2.2) that $\mathtt{I}$ is invariant under translation and rescaling, Alvarez et al. also note that interpretations of $\mathtt{I}$ become risky when comparing datasets with different variances, and that the interpretation is particularly noisy when a population is near uniform. To give a stylized example, consider a city where the east side is 100% Hispanic and the west side is 0% Hispanic and another city where every east side tract is 51% Hispanic while every west side tract is 49% Hispanic. To an observer, it would be obvious that the first city is far more segregated than the second, but Moran's $\mathtt{I}$ sees no difference at all. Indeed, $\mathtt{I}$ can take any value at all when node values are all between $\bar{v}-\epsilon$ and $\bar{v}+\epsilon$, even for very small ϵ. We will return to the question of scale-sensitivity in the discussion of future directions presented in the conclusion.

3. Spectral graph interpretation.

In interpreting the values taken by $\mathtt{I}$, it is essential to understand how the graph itself determines the range of achievable values. In this section we will show that when W is symmetric, $\mathtt{I}(\,\cdot\,;W)$ achieves maximum and minimum values at generalized eigenvectors for the pair $(\Pi W \Pi, \Pi)$, i.e., solutions to the equation $\Pi W \Pi \mathsf{v} = \lambda \Pi \mathsf{v}$, where Π is the orthogonal projection onto X. When the weight matrix is symmetric and has constant rowsum, this reduces to a standard eigenvalue problem.

Definition 3.1 (Rayleigh quotient). *For* $W \in \mathbb{R}^{n\times n}$ *and* $\mathsf{v} \in \mathbb{R}^{n\times 1}$, *the* Rayleigh quotient is $R(\mathsf{v};W) := \frac{\mathsf{v}^\top W \mathsf{v}}{\mathsf{v}^\top \mathsf{v}}$.

It is a standard linear algebra fact that when W is symmetric, the functions v that realize extreme values of $R(\mathsf{v};W)$ are the eigenvectors corresponding to the smallest and largest eigenvalues of W [24]. That is, if W has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $\Phi_1, \ldots, \Phi_n$, then

$$\min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top W \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} = \lambda_n, \quad \max_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top W \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} = \lambda_1,$$

and those extreme values are realized in the eigenspaces corresponding to λ_n and λ_1, respectively. This means that the v realizing extreme values are determined up to scaling when the extreme eigenvalues are simple (multiplicity one). As we will see below, $\mathtt{I}$ is in general not quite a Rayleigh quotient, but close enough to allow us to analyze it in terms of generalized eigenvalues.

3.1. Bounds with adjacency weights.

When $W = A$, the standard adjacency matrix of an undirected, simple (no self-loops) graph, a range of properties of the graph can be inferred from the spectrum of A. Recall that a graph is *d-regular* if $d_i = d$ for all i. Equivalently, a graph $\mathcal{G}$ is d-regular iff $\mathbb{1}$ is an eigenvector of A with eigenvalue d. To emphasize that eigenvalues of A (not necessarily regular) encode connectivity facts about the graph or network $\mathcal{G}$, we record some standard facts from [11]. In these statements we refer to the A-degrees of vertices, which are the standard vertex degrees of the graph. Recall that a graph is called *bipartite* if there are two disjoint sets $A, B \subset V(\mathcal{G})$ such that all edges of $\mathcal{G}$ have one endpoint in each set.

- $\sum_{i=1}^n \lambda_i = 0$, $\sum_{i=1}^n {\lambda_i}^2 = n d_{\mathrm{avg}}$, and $\sum_{i=1}^n {\lambda_i}^3 = 6t$, for t the number of triangles in $\mathcal{G}$.

- $\lambda_1 \leq d_{\max} := \max_i d_i$, the largest degree of any vertex, with equality iff $\mathcal{G}$ is regular.
- $|\lambda_i| \leq \lambda_1$ for all i, and $\mathcal{G}$ is connected iff $\lambda_2 < \lambda_1$.
- If $\mathcal{G}$ is bipartite, then the eigenvalues are symmetric: $\lambda_i = -\lambda_{n+1-i}$ for all i.
- If $\mathcal{G}$ is not bipartite, then $|\lambda_n| < \lambda_1$.

These fundamental facts from spectral graph theory immediately yield bounds on $\mathtt{I}(\cdot\,; A)$ and $\mathtt{I}(\cdot\,; P)$.

Theorem 3.2 ($\mathtt{I}$ bounds for general graphs with adjacency weights). *Let A be the adjacency matrix of an undirected graph $\mathcal{G}$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Let $d_{\min}, d_{\mathrm{avg}}, d_{\max}$ be the minimum, average, and maximum vertex degree (i.e., A-degree) of $\mathcal{G}$, and let P be the row-standardized adjacency matrix described above.*

(a) *The range of possible $\mathtt{I}$ values satisfies $\mathtt{I}(X;A) \subseteq \left[\frac{\lambda_n}{d_{\mathrm{avg}}}, \frac{\lambda_1}{d_{\mathrm{avg}}}\right] \subseteq \left[\frac{-d_{\max}}{d_{\mathrm{avg}}}, \frac{d_{\max}}{d_{\mathrm{avg}}}\right]$ and $\mathtt{I}(X;P) \subseteq \left[\frac{\lambda_n}{d_{\min}}, \frac{\lambda_1}{d_{\min}}\right] \subseteq \left[\frac{-d_{\max}}{d_{\min}}, \frac{d_{\max}}{d_{\min}}\right]$. Note that the bounds for P are still in terms of the eigenvalues and degrees of A.*

(b) *If $\mathcal{G}$ is irregular and not bipartite, $\mathtt{I}(X;A) \subsetneq \left(-\frac{d_{\max}}{d_{\mathrm{avg}}}, \frac{d_{\max}}{d_{\mathrm{avg}}}\right)$.*

Proof. To see (a), note that

$$\frac{\lambda_n}{d_{\mathrm{avg}}} = \min_{\mathsf{v}\neq 0} \frac{1}{d_{\mathrm{avg}}} \frac{\mathsf{v}^\top A \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} \leq \min_{\mathsf{x}\neq 0, \mathsf{x}\perp 1} \frac{1}{d_{\mathrm{avg}}} \frac{\mathsf{x}^\top A \mathsf{x}}{\mathsf{x}^\top \mathsf{x}} = \min \mathtt{I}(X;A).$$

The upper bound is similar. To see the inclusion for $\mathtt{I}(X;P)$, note that

$$\begin{aligned} \min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top P \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} &= \min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top D^{-1} A \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} = \min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top D^{-1/2}(D^{-1/2} A D^{1/2}) D^{-1/2} \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} \\ &= \min_{\mathsf{v}\neq 0} \frac{\left(D^{-1/2}\mathsf{v}\right)^\top (D^{-1/2} A D^{1/2})(D^{-1/2}\mathsf{v})}{\mathsf{v}^\top \mathsf{v}} = \min_{\mathsf{u}\neq 0} \frac{\mathsf{u}^\top (D^{-1/2} A D^{1/2}) \mathsf{u}}{\mathsf{u}^\top D \mathsf{u}}, \end{aligned}$$

where in the last equality we simply make the change of variables $\mathsf{u} = D^{-1/2}\mathsf{v}$. Now, because D is diagonal with smallest entry $d_{\min}$, we have $0 \leq d_{\min} \mathsf{u}^\top \mathsf{u} \leq \mathsf{u}^\top D \mathsf{u}$. Moreover, $D^{-1/2} A D^{1/2}$ is symmetric and similar to A; hence it has the same eigenvalues as A which lie in the range $[-d_{\max}, d_{\max}]$. We conclude that

$$\frac{-d_{\max}}{d_{\min}} \leq \frac{\lambda_n}{d_{\min}} \leq \min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top P \mathsf{v}}{\mathsf{v}^\top \mathsf{v}} \leq \min_{\mathsf{x}\neq 0, \mathsf{x}\perp 1} \frac{\mathsf{x}^\top P \mathsf{x}}{\mathsf{x}^\top \mathsf{x}} = \min \mathtt{I}(X;P).$$

The upper bound is similar.

To see (b), note that the upper bound follows from (a) and observing that $\lambda_1 \leq d_{\max}$, with equality iff $\mathcal{G}$ is regular (in which case $d_{\max} = d_{\mathrm{avg}}$). The lower bound follows similarly, since nonbipartiteness gives $\lambda_n > -\lambda_1 > -d_{\max}$. □

We note that nonbipartiteness is easily observed visually on the graphs of interest in geographic applications by the presence of at least one triangle, and that the real-world graphs are never exactly regular.

While the proof of Theorem 3.2 suggests the eigenvectors Φ_1 and Φ_n as candidate maximizers and minimizers of $\mathtt{I}$, respectively, we shall see that the projection onto X complicates matters when the graph is not regular, and requires us to pass to generalized eigenvectors (Theorem 3.4).

To test the sharpness of the bounds, we can look at three graphs based on real data from Georgia: the graphs dual to blocks, block groups, and counties that are depicted in Figure 2. Results are summarized in Table 2.

We note that non-bipartiteness is easily observed visually on the graphs of interest in geographic applications by the presence of at least one triangle, and that the real-world graphs are never exactly regular.

While the proof of Theorem 3.2 suggests the eigenvectors Φ_1 and Φ_n as candidate maximizers and minimizers of $\mathtt{I}$, respectively, we shall see that the projection onto X complicates matters when the graph is not regular, and requires us to pass to generalized eigenvectors (Theorem 3.4).

TABLE 2

Values of $\mathtt{I}(\cdot;A)$ and $\mathtt{I}(\cdot;P)$ achievable on Georgia dual graphs from Figure 2, to four decimal places. We see that the degree bounds from Theorem 3.2 can be far from the eigenvalue bounds when $d_{\min} \ll d_{\text{avg}} \ll d_{\max}$. Note that there are very high-degree nodes in the block graph; these typically occur when large blocks within bodies of water are adjacent to high numbers of coastal blocks. The true range of achievable $\mathtt{I}$ values is obtained by numerically solving the generalized eigenvalue problem for A, as in Theorem 3.4, or by the Lagrange multiplier method for P described in Remark 3.7. The eigenvalue bounds for P are not tight because the estimate that divides by $d_{\min}$ is far from sharp, and these graphs all have leaves ($d_{\min} = 1$).

TABLE 2

Values of $\mathtt{I}(\cdot;A)$ and $\mathtt{I}(\cdot;P)$ achievable on Georgia dual graphs from Figure 2, to four decimal places. We see that the degree bounds from Theorem 3.2 can be far from the eigenvalue bounds when $d_{\min} \ll d_{\text{avg}} \ll d_{\max}$. Note that there are very high-degree nodes in the block graph; these typically occur when large blocks within bodies of water are adjacent to high numbers of coastal blocks. The true range of achievable $\mathtt{I}$ values is obtained by numerically solving the generalized eigenvalue problem for A, as in Theorem 3.4, or by the Lagrange multiplier method for P described in Remark 3.7. The eigenvalue bounds for P are not tight because the estimate that divides by $d_{\min}$ is far from sharp, and these graphs all have leaves ($d_{\min} = 1$).

	blocks		block groups		counties	
# nodes	291,086		5,533		159	
# edges	1,393,216		15,344		418	
$d_{\min} < d_{\text{avg}} < d_{\max}$	$1 < 9.5725 < 92$		$1 < 5.5464 < 16$		$1 < 5.2579 < 10$	
	A	P	A	P	A	P
deg. bounds	$(-9.6108, 9.6108)$	$(-92, 92)$	$(-2.8848, 2.8848)$	$(-16, 16)$	$(-1.9019, 1.9019)$	$(-10, 10)$
eig. bounds	$(-1.0957, 1.4486)$	$(-10.4886, 13.8669)$	$(-0.6978, 1.1554)$	$(-3.8702, 6.4079)$	$(-0.5849, 1.1151)$	$(-3.0751, 5.8629)$
true $\mathtt{I}$ range	$(-1.0957, 1.4475)$	$(-1.1007, 1.2249)$	$(-0.6978, 1.1526)$	$(-.7510, 1.0326)$	$(-0.5845, 1.0763)$	$(-.7673, 1.0260)$

To test the sharpness of the bounds, we can look at three graphs based on real data from Georgia: the graphs of blocks, block groups, and counties that are depicted in Figure 2. Results are summarized in Table 2.

This shows that it is possible for $\mathtt{I}$ to fall outside of the $[-1, 1]$ range for realistic graphs; later, we will see that in fact the values of $\mathtt{I}$ with respect to A and P can get arbitrarily large (positive or negative) when the degree disparity is exaggerated. Note that these examples contradict what seems to be the prevailing understanding of the behavior of $\mathtt{I}(\cdot, P)$ in the geography community, where it is said that $|\mathtt{I}(\mathsf{v}; P)| \le 1$ for all vectors v.[1] Presumably this belief can be traced to the fact that, for arbitrary graphs $\mathcal{G}$, the vertices have P-degree one and the eigenvalues of P fall in $[-1, 1]$, "fixing" the fact that the A-degrees may be nonconstant and the eigenvalues of A may be large. However, we know of no way to bound $\mathtt{I}(\cdot; P)$ in terms of the eigenvalues of P; our spectral analysis for P depends on the eigenvalues of A.

In contrast, the extremal behavior of $\mathtt{I}$ over X is straightforward when the underlying graph is regular, and a statement similar to the previous theorem becomes sharp.

This shows that it is possible for $\mathtt{I}$ to fall outside of the $[-1, 1]$ range for realistic graphs; later, we will see that in fact the values of $\mathtt{I}$ with respect to A and P can get arbitrarily large (positive or negative) when the degree disparity is exaggerated. Note that these examples contradict what seems to be the prevailing understanding of the behavior of $\mathtt{I}(\cdot, P)$ in the geography community, where it is said that $|\mathtt{I}(\mathsf{v}; P)| \le 1$ for all vectors v.[1] Presumably this belief can be traced to the fact that, for arbitrary graphs $\mathcal{G}$, the vertices have P-degree one and the eigenvalues of P fall in $[-1, 1]$, "fixing" the fact that the A-degrees may be nonconstant and the eigenvalues of A may be large. However, we know of no way to bound $\mathtt{I}(\cdot; P)$ in terms of the eigenvalues of P; our spectral analysis for P depends on the eigenvalues of A.

In contrast, the extremal behavior of $\mathtt{I}$ over X is straightforward when the underlying graph is regular, and a statement similar to the previous theorem becomes sharp.

THEOREM 3.3 ($\mathtt{I}$ bounds for regular graphs with adjacency weights). *Let A be the adjacency matrix of an undirected, d-regular graph $\mathcal{G}$, with row-standardization $P = \frac{1}{d}A$. Let $d = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ be the eigenvalues of A with associated eigenvectors $\Phi_1, \Phi_2, \ldots, \Phi_n$. Then*

THEOREM 3.3 ($\mathtt{I}$ bounds for regular graphs with adjacency weights). *Let A be the adjacency matrix of an undirected, d-regular graph $\mathcal{G}$, with row-standardization $P = \frac{1}{d}A$. Let $d = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ be the eigenvalues of A with associated eigenvectors $\Phi_1, \Phi_2, \ldots, \Phi_n$. Then*

(a) $\mathtt{I}(X; A) = [\mathtt{I}(\Phi_n; A), \mathtt{I}(\Phi_2; A)] \subseteq [-1, 1]$;

(b) $\mathtt{I}(\Phi_2; A) = 1$ *iff $\mathcal{G}$ is disconnected;*

(c) $\mathtt{I}(\Phi_n; A) = -1$ *iff $\mathcal{G}$ is bipartite.*

Since A is just a scalar multiple of P in the regular case, the same bounds and equalities hold for $\mathtt{I}(\cdot; P)$.

Proof. Theorem 3.2 gives $\mathtt{I}(X; A) \subseteq [-1, 1]$ after noting that $\lambda_1 \le d_{\max}$ and $\lambda_n \ge -\lambda_1$. To finish (a), first note that if A is the adjacency matrix of a d-regular graph, then $\Phi_1 = \mathbb{1}$. By the Courant–Fischer–Weyl min-max principle [24],

$$\Phi_2 \in \operatorname*{arg\,max}_{\mathsf{v} \neq 0, \mathsf{v} \perp \Phi_1} \frac{\mathsf{v}^\top A \mathsf{v}}{\|\mathsf{v}\|_2{}^2} = \operatorname*{arg\,max}_{\mathsf{v} \neq 0, \mathsf{v} \perp \mathbb{1}} \frac{\mathsf{v}^\top A \mathsf{v}}{\|\mathsf{v}\|_2{}^2} = \operatorname*{arg\,max}_{\mathsf{v} \neq 0} \mathtt{I}(\mathsf{v}; A).$$

[1] For instance, this is a stated reason to use $W = P$ in the user guide material for ArcGIS [1], the dominant spatial statistics software package, which states that "In general, the Global Moran's Index is bounded by −1.0 and 1.0. This is always the case when your weights are row standardized." And later: "Row standardized weighting is often used with fixed distance neighborhoods and almost always used for neighborhoods based on polygon contiguity. This is to mitigate bias due to features having different numbers of neighbors."

[1] For instance, this is a stated reason to use $W = P$ in the user guide material for ArcGIS [1], the dominant spatial statistics software package, which states that "In general, the Global Moran's Index is bounded by −1.0 and 1.0. This is always the case when your weights are row standardized." And later: "Row standardized weighting is often used with fixed distance neighborhoods and almost always used for neighborhoods based on polygon contiguity. This is to mitigate bias due to features having different numbers of neighbors."

The other bound follows from noting that $\Phi_n \perp \Phi_1$, giving

$$\arg\min_{\mathsf{v}\neq 0, \mathsf{v}\perp \mathsf{1}} \frac{\mathsf{v}^\top A\mathsf{v}}{\|\mathsf{v}\|_2{}^2} = \arg\min_{\mathsf{v}\neq 0} \frac{\mathsf{v}^\top A\mathsf{v}}{\|\mathsf{v}\|_2{}^2} \ni \Phi_n.$$

For (b), note that $\max_{\mathsf{v}\neq 0} \mathtt{I}(\mathsf{v}; A) = \frac{\lambda_2}{d} \leq \frac{\lambda_1}{d}$. The Perron–Frobenius theorem gives $\lambda_2 = \lambda_1 = d$ iff $\mathcal{G}$ is disconnected [24].

To see (c), we use $\min_{\mathsf{v}\neq 0} \mathtt{I}(\mathsf{v}; A) = \frac{\lambda_n}{d}$. Perron–Frobenius gives $-\lambda_n \leq \lambda_1 = d$, with equality iff A is bipartite. □

3.2. Analysis for symmetric weight matrices. In preparation for proposing alternatives for the spatial weight matrix W, we now provide an analysis for arbitrary symmetric matrices. When the graph is regular, its adjacency matrix A has $\mathsf{1}$ as the eigenvector with largest eigenvalue. This makes projection onto $X = \mathsf{1}^\perp$ interact nicely with spectral analysis. For general graphs, we will handle the projection more carefully.

Let $\Pi = I - \frac{1}{n}\mathsf{1}\mathsf{1}^\top$ denote the orthogonal projection onto X, i.e., $\Pi\mathsf{v} = \mathsf{v} - \bar{v}\mathsf{1} = \mathsf{x}$. Noting that $\Pi\Pi^\top = \Pi^2 = \Pi$, we have

$$\mathtt{I}(\mathsf{v}; W) = \frac{\mathsf{v}^\top \Pi W \Pi \mathsf{v}}{\mathsf{v}^\top \Pi \mathsf{v}}.$$

This is no longer a Rayleigh quotient, but rather a *generalized Rayleigh quotient*. Since Π is singular, having $\mathsf{1}$ in its kernel, $\mathtt{I}(\mathsf{v}; W)$ cannot be reduced to a standard Rayleigh quotient via $\Pi^{-\frac{1}{2}}$. Instead, one can use the theory of generalized eigenvalues for the matrix pair (A, B), i.e., solutions to $A\mathsf{v} = \lambda B\mathsf{v}$ [46, 36]. In our case we will consider the generalized spectrum $\{(\lambda_i, \Phi_i)\}_{i=1}^{n-1}$ of nonconstant Φ_i satisfying $\Pi W \Pi \Phi_i = \lambda_i \Pi \Phi_i$. Since W and Π and thus $A = \Pi W \Pi$ are symmetric and $B = \Pi$ is positive semidefinite, the generalized eigenvalues are real, and the eigenvectors can be chosen to satisfy $(\Pi\Phi_i)^\top \Pi \Phi_j = \Phi_i^\top \Pi \Phi_j = 0$, $i \neq j$, and $(\Pi\Phi_i)^\top \Pi\Phi_i = \Phi_i^\top \Pi \Phi_i = 1$. This means that the vectors are orthonormal after projection, so we will say such generalized eigenvectors are Π-*orthonormal*. Using the fact that $\mathsf{1}$ is orthogonal to each $\Pi\Phi_i$, we get the following diagonalization-style statement for $\mathtt{I}$.

Theorem 3.4 (Spectral interpretation of $\mathtt{I}$ for symmetric weight matrices). *Let W be a symmetric $n \times n$ weight matrix, and let $\{(\lambda_i, \Phi_i)\}_{i=1}^{n-1}$ be Π-orthonormal generalized eigenvectors for the pair $(\Pi W \Pi, \Pi)$. Then for all nonzero $\mathsf{v} \in \mathbb{R}^{n\times 1}$,*

(a) $\mathsf{v} = \left(\sum_{i=1}^{n-1} \alpha_i \Pi \Phi_i\right) + \bar{v}\mathsf{1}$ *for some coefficients* $\{\alpha_i\}_{i=1}^{n-1}$;

(b) $\mathtt{I}(\mathsf{v}; W) = \sum_{i=1}^{n-1} \alpha_i^2 \lambda_i \big/ \sum_{i=1}^{n-1} \alpha_i^2$.

Proof. The result in (a) follows immediately from the Π-orthogonality of $\{\Phi_i\}_{i=1}^{n-1}$ and the fact that $\mathsf{1}$ generates the kernel of Π. Note that this could be done either on the left or on the right in the nonsymmetric case, but is unambiguous since W is symmetric.

To see (b), we compute

$$\mathtt{I}(\mathsf{v}; W) = \frac{\mathsf{v}^\top \Pi W \Pi \mathsf{v}}{\mathsf{v}^\top \Pi \mathsf{v}} = \frac{\left(\sum_{i=1}^{n-1} \alpha_i \Pi \Phi_i + \bar{v}\mathsf{1}\right)^\top \Pi W \Pi \left(\sum_{i=1}^{n-1} \alpha_i \Pi \Phi_i + \bar{v}\mathsf{1}\right)}{\left(\sum_{i=1}^{n-1} \alpha_i \Pi \Phi_i + \bar{v}\mathsf{1}\right)^\top \Pi \left(\sum_{i=1}^{n-1} \alpha_i \Pi \Phi_i + \bar{v}\mathsf{1}\right)}$$

$$
=\frac{\left(\sum_{i=1}^{n-1}\alpha_i\Pi^2\Phi_i+\bar{v}\Pi\mathsf{1}\right)^{\top}W\left(\sum_{i=1}^{n-1}\alpha_i\Pi^2\Phi_i+\bar{v}\Pi\mathsf{1}\right)}{\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i+\bar{v}\mathsf{1}\right)^{\top}\left(\sum_{i=1}^{n-1}\alpha_i\Pi^2\Phi_i+\bar{v}\Pi\mathsf{1}\right)}
$$

$$
=\frac{\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i\right)^{\top}W\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i\right)}{\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i+\bar{v}\mathsf{1}\right)^{\top}\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i\right)}=\frac{\left(\sum_{i=1}^{n-1}\alpha_i\Phi_i\right)^{\top}\left(\sum_{i=1}^{n-1}\alpha_i\Pi W\Pi\Phi_i\right)}{\left(\sum_{i=1}^{n-1}\alpha_i\Phi_i\right)^{\top}\left(\sum_{i=1}^{n-1}\alpha_i\Pi\Phi_i\right)}
$$

$$
=\frac{\left(\sum_{i=1}^{n-1}\alpha_i\Phi_i\right)^{\top}\left(\sum_{i=1}^{n-1}\alpha_i\lambda_i\Pi\Phi_i\right)}{\sum_{i,j=1}^{n-1}\alpha_i\alpha_j\Phi_i^{\top}\Pi\Phi_j}=\frac{\sum_{i,j=1}^{n-1}\alpha_i\alpha_j\lambda_i\Phi_i^{\top}\Pi\Phi_j}{\sum_{i,j=1}^{n-1}\alpha_i\alpha_j\Phi_i^{\top}\Pi\Phi_j}=\frac{\sum_{i=1}^{n-1}\alpha_i^2\lambda_i}{\sum_{i=1}^{n-1}\alpha_i^2}. \qquad \square
$$

By the usual scale-invariance, Theorem 3.4 allows us to understand the behavior of $\mathtt{I}$ just considering $\sum_{i=1}^{n-1}\alpha_i^2\lambda_i$ when $\sum_{i=1}^{n-1}\alpha_i^2=1$. Theorem 3.4 says, in other words, that $\mathtt{I}$ values are precisely the convex combinations of the generalized eigenvalues. Since $\{\Pi\Phi_i\}_{i=1}^{n-1}$ is an orthonormal basis for X, this is analogous to the classical spectral analysis of the Rayleigh quotient.

We note that the generalized eigenpairs of $(\Pi W\Pi,\Pi)$ can be put in correspondence with those of $\Pi W\Pi$ as follows.

Lemma 3.5. *Let W be a symmetric matrix.*

(a) *If (λ,Φ) is an eigenpair of $(\Pi W\Pi,\Pi)$, then $(\lambda,\Pi\Phi)$ is an eigenpair of $\Pi W\Pi$.*
(b) *If (λ,Φ) is an eigenpair of $\Pi W\Pi$, then (λ,Φ) is an eigenpair of $(\Pi W\Pi,\Pi)$.*

Proof. To see (a), note that $\Pi^2=\Pi$ and thus $\Pi W\Pi(\Pi\Phi)=\Pi W\Pi\Phi=\lambda\Pi\Phi$. To see (b), note that $\Pi W\Pi\Phi=\lambda\Phi$ implies $\Pi^2W\Pi\Phi=\lambda\Pi\Phi$. Again, $\Pi^2=\Pi$ and the result follows. $\square$

Corollary 3.6 (Extreme values of $\mathtt{I}$). *When W is symmetric, the minimum and maximum values of $\mathtt{I}(\cdot\,;W)$ are the smallest and largest generalized eigenvalues of $(\Pi W\Pi\,,\Pi)$, respectively, and are achieved at the corresponding generalized eigenvectors.*

Suppose additionally that W has constant rowsum k, i.e., the graph is k-regular with respect to W-degree. Then the eigenvectors of W are equal to the generalized eigenvectors of $(\Pi W\Pi,\Pi)$. The eigenvalues agree except possibly for the eigenvalue associated to $\mathsf{1}$, which is k for W and zero for the generalized problem.

Proof. The observation that $\mathtt{I}$ values are convex combinations establishes that the extremes are realized at the largest and smallest generalized eigenvalues. Next, we note that $W\mathsf{1}=k\mathsf{1}$ while $\Pi W\Pi\mathsf{1}=\mathsf{0}$, which establishes the last statement.

Now consider $\mathsf{x}\in X$. From Theorem 3.4, we can express x in the eigenbasis $\{(\lambda_i,\Phi_i)\}_{i=1}^{n-1}$ that spans X, and we note that Π is the identity on X, so it preserves x and $W\mathsf{x}$. This gives us

$$
W\mathsf{x}=\lambda\mathsf{x}\iff W\Pi\mathsf{x}=\lambda\Pi\mathsf{x}\iff\Pi W\Pi\mathsf{x}=\lambda\Pi\mathsf{x},
$$

identifying the eigenvectors with the generalized eigenvectors, as needed. $\square$

Remark 3.7 (Extension to $\mathtt{I}(\cdot\,; P)$). The matrix P is not symmetric, so the above orthogonal decomposition does not apply (in particular, the left and right eigenvectors of P are different). However, the variational problem of minimizing or maximizing $\frac{\mathsf{v}^\top \Pi P \Pi \mathsf{v}}{\mathsf{v}^\top \Pi \mathsf{v}}$ over the space of nonzero vectors reduces, by scale-invariance, to the constrained optimization of $\mathsf{v}^\top \Pi P \Pi \mathsf{v}$ subject to the constraint that $\mathsf{v}^\top \Pi \mathsf{v} = 1$. This has associated Lagrangian function $\mathsf{v}^\top \Pi P \Pi \mathsf{v} + \zeta(\mathsf{v}^\top \Pi \mathsf{v} - 1)$ for some scalar ζ. The v-derivative of the Lagrangian is $\mathsf{v}^\top \Pi (P + P^\top)\Pi + \zeta(2\mathsf{v}^\top \Pi)$. Setting equal to 0, we see $\frac{\mathsf{v}^\top \Pi P \Pi \mathsf{v}}{\mathsf{v}^\top \Pi \mathsf{v}}$ is maximized and minimized at the generalized eigenvectors of $(\frac{1}{2}\Pi(P + P^\top)\Pi, \Pi)$ corresponding to the extreme eigenvalues.

4. Comparing $\mathtt{I}$ within and across graphs.

Theorem 3.4 gives us tools to study the question of what kinds of vectors $\mathsf{v}_1, \mathsf{v}_2$ have $\mathtt{I}(\mathsf{v}_1; W) \approx \mathtt{I}(\mathsf{v}_2; W)$. We consider two cases: $|\mathtt{I}| \gg 0$ and $|\mathtt{I}| \approx 0$. We will continue to suppose that vectors are scaled so that $\mathsf{x} = \mathsf{v} - \bar{v}\mathbb{1}$ has ℓ^2-norm 1, i.e., $\sum_{i=1}^{n-1} \alpha_i^2 = 1$, and we let the generalized eigenvalues of $(\Pi W \Pi, \Pi)$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1}$.

4.1. Analysis when $|\mathtt{I}| \gg 0$.

We first consider $\mathtt{I}(\mathsf{v}; W) \gg 0$. Since we can assume $\sum_{i=1}^{n-1} \alpha_i^2 = 1$, we have $\mathtt{I}(\mathsf{v}; W) = \sum_{i=1}^{n-1} \alpha_i^2 \lambda_i$. So, $\mathtt{I}(\mathsf{v}; W)$ is large iff most of the coefficient energy in $(\alpha_1, \ldots, \alpha_{n-1})$ is localized towards the lowest-indexed values (corresponding to largest eigenvalues).

To study large $\mathtt{I}$, let us write $\beta_i = \alpha_i^2$; then we can express the zero-centered vectors with $\mathtt{I} \geq \lambda$ as

$$X^+(\lambda) = \left\{ \mathsf{x} = \sum_{i=1}^{n-1} \sqrt{\beta}_i \Pi \Phi_i \;\middle|\; \sum_{i=1}^{n-1} \beta_i = 1, \quad \beta_i \geq 0, \quad \sum_{i=1}^{n-1} \beta_i \lambda_i \geq \lambda \right\}.$$

In the generic case that λ_1 is simple, we have $X^+(\lambda_1) = \{\Pi\Phi_1\}$. Clearly $X^+(\lambda) \subseteq X^+(\lambda')$ when $\lambda \geq \lambda'$ and $X^+(\lambda) = \emptyset$ for $\lambda > \lambda_1$.

The expression for $X^+(\lambda)$ can be understood geometrically as a portion of the standard simplex $\sum \beta_i = 1$, $\beta_i > 0$, to one side of the hyperplane $\sum \beta_i \lambda_i = \lambda$. In high dimensions, most of the mass of the simplex concentrates away from the vertices [35], which implies that the volume of $X^+(\lambda)$ is small for $\lambda \approx \lambda_1$. In particular, if λ is large, any $\mathsf{x} \in X^+(\lambda)$ will need to have a large portion of its coefficient energy coming from the $\Pi\Phi_i$ with largest eigenvalues. In the case that there is a significant spectral gap ($\lambda_1 \gg \lambda_2$), then the coefficient energy needs to localize on $\Pi\Phi_1$.

A consequence of the concentration of $X^+(\lambda)$ for λ close to λ_1 is that distinct $\mathsf{x}, \mathsf{x}' \in X^+(\lambda)$ share certain *qualitative* properties when λ is close to λ_1. Indeed, under mild assumptions (see section 6), the largest eigenvectors correspond to *clustered* patterns in the data. If there is a spectral gap, then a single clustered pattern must dominate for λ large enough. See Figure 4 for a visualization, showing conversely that with a small spectral gap, many different clustered patterns can achieve the same high $\mathtt{I}$ scores.

The same arguments apply to sublevel sets $X^-(\lambda)$ when $\lambda \approx \lambda_{n-1}$. In the case of irregular graphs, the interpretation of the eigenvectors of $(\Pi W \Pi, \Pi)$ with smallest eigenvalues is more difficult [15], but one could qualitatively comment that they capture (localized) checkerboard patterns, as in Figure 4.

4.2. Analysis when $|\mathtt{I}| \approx 0$.

On the other hand, the x with $\mathtt{I}(W; \mathsf{x}) \in (-\epsilon, \epsilon)$ need not have any meaningful qualitative properties in common, even as $\epsilon \to 0^+$. The constraint that $\mathtt{I}$ lie in $(-\epsilon, \epsilon)$ does not imply that the energy of the coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_{n-1})$ must localize on specific indices. If $\mathtt{I}(\mathsf{x}; W) \approx 0$, then there must

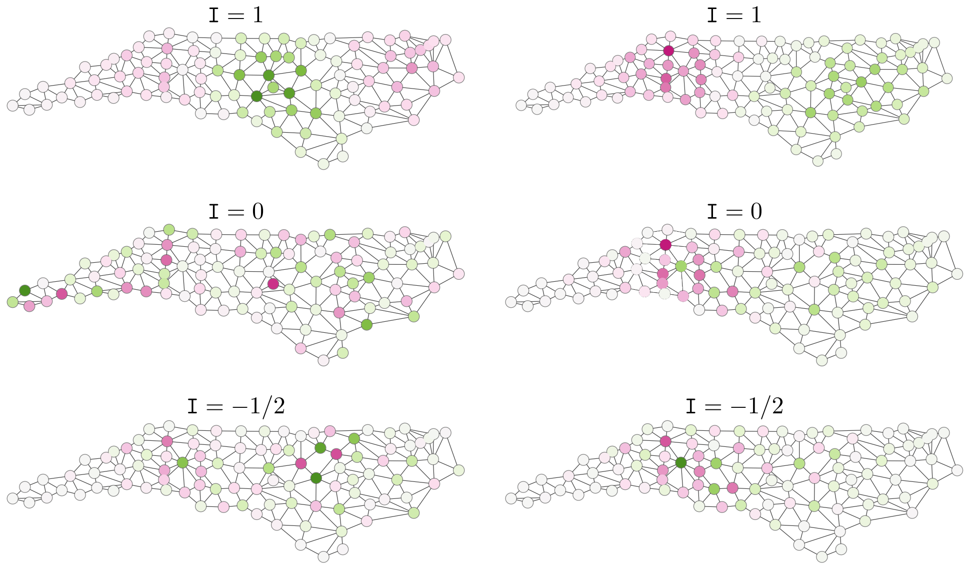


FIG. 4. *The maximum $\mathtt{I}(\mathsf{v}; A)$ value on this graph is ≈ 1.1034, and there are several eigenvectors yielding $\mathtt{I} > 1$ with respect to A; in particular there is not a large spectral gap between λ_1 and λ_2. This means that visibly different cluster patterns can realize the same* I *value, even when it is quite close to the maximum. The two functions in the middle row both realize $\mathtt{I}(\mathsf{v}; A) = 0$, but they are qualitatively rather different: the one on the left is close to spatially uncorrelated, while the one on the right is a linear combination of a cluster pattern with a localized checkerboard. The bottom row shows two functions with $\mathtt{I}(\mathsf{v}; A) = -.5$, reasonably close to the minimum of $-.5983$, both exhibiting something like a "localized checkerboard" pattern.*

be coefficients that place energy on both positive and negative eigenvalues, so that x could show some cluster structure or some localized checkerboarding, or appear spatially uncorrelated (recall that the expected value of $\mathtt{I}(\cdot\ ; A)$ is $-\frac{1}{n-1}$ under a spatially uncorrelated random model, giving values near zero for large graphs). See Figure 4 for two qualitatively very different functions, both with $\mathtt{I} = 0$.

5. Overview of choices of W.

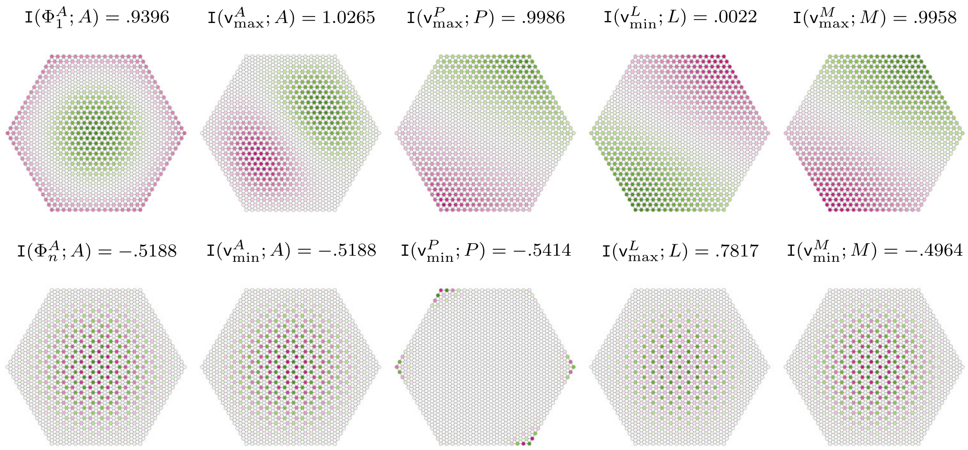
The choice of spatial weight matrix W can have a major impact on the interpretability of I. We now overview and compare four choices of W for use in Moran's I.

DEFINITION 5.1 (Alternative spatial weight matrices). *Given a graph $\mathcal{G}$, let D be the diagonal matrix given by D_{ii} as the ith vertex degree. Then we will consider the following matrices:*

- *$W = A$, the standard adjacency matrix;*
- *$W = P$, the row-standardization $P = D^{-1}A$;*
- *$W = L$, the unnormalized graph Laplacian $L = D - A$; and*
- *$W = M$, the doubly-stochastic matrix defined by*

$$M_{ij} = \begin{cases} A_{ij}/\max(D_{ii}, D_{jj}), & i \neq j, \\ 1 - \sum_{k \neq i} M_{ik}, & i = j. \end{cases}$$

The Laplacian is a ubiquitous choice of matrix associated to a graph that encodes its geometry and topology, so it is a natural choice here. In section 7, we will motivate M as a doubly-stochastic approximation to A, which will provide both nice numerical properties and an appealing random-walk interpretation. To see that M is indeed



$\mathtt{I}(\Phi_1^A; A) = .9396$ $\mathtt{I}(\mathsf{v}_{\max}^A; A) = 1.0265$ $\mathtt{I}(\mathsf{v}_{\max}^P; P) = .9986$ $\mathtt{I}(\mathsf{v}_{\min}^L; L) = .0022$ $\mathtt{I}(\mathsf{v}_{\max}^M; M) = .9958$

$\mathtt{I}(\Phi_n^A; A) = -.5188$ $\mathtt{I}(\mathsf{v}_{\min}^A; A) = -.5188$ $\mathtt{I}(\mathsf{v}_{\min}^P; P) = -.5414$ $\mathtt{I}(\mathsf{v}_{\max}^L; L) = .7817$ $\mathtt{I}(\mathsf{v}_{\min}^M; M) = -.4964$

FIG. 5. *These hexagonal graphs are* 6*-regular except on the boundary. Depending on the spatial weight matrix, the extremizers can differ, particularly in the extent to which the lower-degree vertices along the boundary are reflected in the pattern. (The classical eigenvectors of* A *are shown to the left for comparison.) In particular, the minimizer of* $\mathtt{I}(\cdot\,; L)$ *and maximizer of* $\mathtt{I}(\cdot\,; M)$ *are less impacted by low-degree nodes than the maximizer of* $\mathtt{I}(\cdot\,; A)$*.*

doubly-stochastic, i.e., has rows and columns summing to one, note that the off-diagonals are defined symmetrically in i and j. Since the diagonal entries are defined to make the rows sum to one, the columns must sum to one as well.

While A, L, and M are symmetric, P is not, but it can often be handled by similar techniques, as we saw in Remark 3.7. Extremizing $\mathtt{I}$ amounts to solving a standard eigenvalue problem on L and M, and a generalized eigenvalue problem on A and P. (See Corollary 3.6.)

Remark 5.2 (Vertex-regular case). These spectra, and the $\mathtt{I}$ scores, will differ in general across the choice of weight matrix. However, in the case of d-regular graphs, P is symmetric, and $P = M = \frac{1}{d}A$. The vertex degree matrix is $D = d\cdot I$, so that $L = d\cdot I - A$ and the spectra are related by $\mu_i = d - \lambda_i$ for each i. In particular,

$$\mathtt{I}(\mathsf{x}; A) = \frac{1}{d}\cdot\frac{\mathsf{x}^\top A\mathsf{x}}{\mathsf{x}^\top\mathsf{x}} = \frac{1}{d}\cdot\frac{\mathsf{x}^\top(dI + A - dI)\mathsf{x}}{\mathsf{x}^\top\mathsf{x}} = 1 - \frac{1}{d}\cdot\frac{\mathsf{x}^\top L\mathsf{x}}{\mathsf{x}^\top\mathsf{x}} = 1 - 2\cdot\mathtt{I}(\mathsf{x}; L),$$

so the $\mathtt{I}$ scores for A, P, M are equal and compare to L by a precise affine relationship.

Though this relationship will not be exact for general graphs, it helps translate the conventional wisdom of anticorrelation, noncorrelation, and clustering to $\mathtt{I}(\cdot\,; L)$ values of roughly 1, 1/2, and 0, respectively—in particular, lower values of $\mathtt{I}$ are more segregated when the spatial weight matrix is the Laplacian.

5.1. Comparison on families of graphs. In sections 6 and 7, we explore the different mathematical connections and interpretations made possible by using L or M instead of the more traditional A or P. First, we empirically compare $\mathtt{I}(\cdot\,; W)$ for the different spatial weight matrices. We compare our four spatial weight matrices using a common nearly regular graph: a hexagon with 16 vertices on each side, drawn in the hex lattice. The empirical maximizers and minimizers are shown in Figure 5.

The leftmost pair of plots in Figure 5 shows the extreme eigenvectors of A, and we see that the lower degrees on the boundary have a visible effect on the pattern.

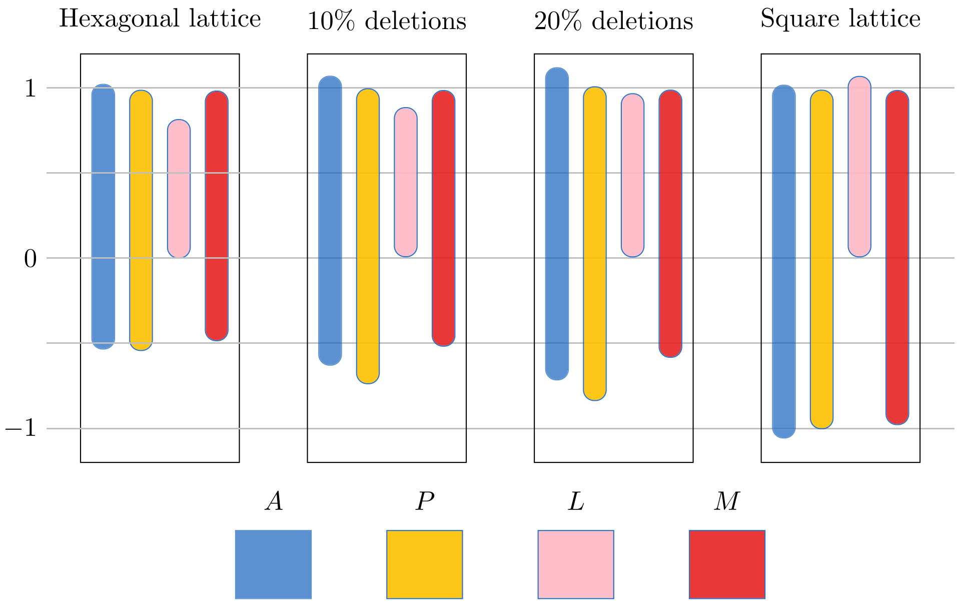


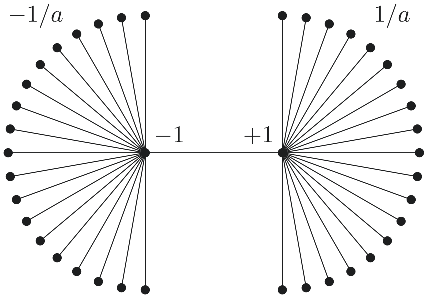
FIG. 6. *Ranges of possible* I$(\cdot\ ; W)$ *values for* $W = A, P, L, M$. *We compare graphs on* 169 *nodes formed within the hexagonal lattice and the square lattice, and we use random edge deletions to interpolate between them and introduce degree variation, as explained in the text. We can observe that* L *gives nonnegative* I *values, and that only* M *gives values always between* −1 *and* 1.

Passing to the true A extremizer (via the generalized eigenvector) or row-normalizing to obtain P might be thought to fix the degree effects, but this particularly fails with the minimizer of P. It is also interesting to note that the maximizer for P takes its strongest values on the second rung—adjacent to the vertices of lowest degree. Note that M and L both give the expected clustered configuration on one end of the I range, but give two interestingly different approximations to checkerboards on the other.

Next, Figure 6 shows the numerically computed ranges of achievable I for $W = A, P, L, M$ on various planar graphs. First, we take a subset of the hexagonal (triangular) lattice, formed as a hexagon with 8 vertices along each side (a smaller version of the graph in Figure 5). From the square lattice, we use a square with 13 vertices along a side, so that it agrees with the hexagon in having 169 vertices. Note that the hexagonal lattice can also be viewed, combinatorially, as the square lattice with diagonals added. Therefore we can produce graphs that in a sense interpolate between the two lattices by deleting edges from the hexagonal lattice. We choose to do so at random in order to also introduce more variation in vertex degree.[2]

This example makes it clear that even quite "reasonable" planar graphs, when they are irregular, can realize I values outside of $[-1, 1]$. In the case when the graph is nearly regular bipartite (such as a large square grid graph), maximal and minimal values of Moran's I can be seen to converge to ±1, respectively, as the number of nodes in the graph increases and the variance in degree converges to 0. On the other hand, one can construct graphs that realize arbitrarily large and small I as the degree disparity gets large.

[2] To be precise, we order the edges, then randomly select 10% (or 20%, respectively) for deletion, rejecting the final product if it is disconnected. Over 1000 successful trials, we then report the average of the numerical minimizer and of the numerical maximizer of I. See the Supplementary Materials (SIAM_Supplement.pdf [local/web 1.02MB]) for more information.



limit as $n \to \infty$			
$\mathtt{I}(\mathsf{v}_a;A)$	$\mathtt{I}(\mathsf{v}_a;P)$	$\mathtt{I}(\mathsf{v}_a;L)$	$\mathtt{I}(\mathsf{v}_a;M)$
a	a	$\frac{1}{4}(a-1)^2$	1
$\mathtt{I}(\mathsf{v}_\infty;A)$	$\mathtt{I}(\mathsf{v}_\infty;P)$	$\mathtt{I}(\mathsf{v}_\infty;L)$	$\mathtt{I}(\mathsf{v}_\infty;M)$
$-\frac{1}{2}$	0	∞	0

FIG. 7. *The family of double-star graphs is depicted here to emphasize the major impact of having very uneven vertex degrees in a graph. Letting v_a take the value $\pm 1/a$ on leaves and ± 1 on hubs, as indicated on the graph above, gives $\mathtt{I}(\mathsf{v}_a;A)$ values approaching any arbitrary a as the number of leaves $n \to \infty$. In particular, this illustrates that all real values of $\mathtt{I}(\mathsf{v};A)$ are achievable for $W = A, P$ and all nonnegative values are achievable for $W = L$. That is, passing to row-normalized P does not mitigate the degree effect, and the problem is even more pronounced for the Laplacian (see section 6). Only the use of the doubly-stochastic approximation M keeps $\mathtt{I}$ bounded (see section 7).*

To see this, consider a double-star graph (Figure 7) where each hub is connected to n leaves. Consider a function v_a that takes the values ± 1 on the hubs and $\pm 1/a$ on the leaves; note v_a has average value 0. For arbitrary fixed $a \neq 0$, as n gets large, the average product across an edge is nearly $1/a$ while the average squared value at a vertex is nearly $1/a^2$. This means that $\mathtt{I}(\mathsf{v}_a;A) \to a$ as $n \to \infty$ (by Remark 2.3). This construction works for both positive and negative values of a. Interestingly, however, putting 0 on the leaves (denoted by v_∞) gives different limiting behavior, with $\mathtt{I}(\mathsf{v}_\infty;A) \to -1/4$ as $n \to \infty$. Normalizing A to be row-stochastic does not solve this problem; if we use P, we get $\mathtt{I}(\mathsf{v}_a;P) \to a$ as $n \to \infty$, while putting 0 on the leaves gives $\mathtt{I}(\mathsf{v}_\infty;P) \to 0$.

This double-star example is designed to exaggerate the phenomenon that causes $\mathtt{I}$ to explode, but degree effects of this kind are reflected in the real-world examples below: when there are adjacent nodes of relatively high vertex degree, extreme values of $\mathtt{I}$ can be obtained by placing positive and negative values on those nodes, and near-zero values everywhere else (see Figure 8, lower right).

5.2. Correlation on realistic examples. Next, we confirm that, despite significant differences observable in theory, the choices of W give outputs that are fairly tightly correlated on real-world dual graphs $\mathcal{G}$ and population functions v. In particular, we consider the spatial weight matrices A, L, and M applied to Black and Hispanic population in the census tracts of all 50 states. Since the underlying graphs are not regular, we know of no theoretical relationship between the $\mathtt{I}$ values when we change the matrix W. Despite this, Figure 9 shows strong correlations. The Supplementary Materials (SIAM_Supplement.pdf [local/web 1.02MB]) contain a more extensive pairwise comparison among all four choices of weight matrix.[3]

It is interesting to look at states for which $\mathtt{I}(\mathsf{v};A)$ and $\mathtt{I}(\mathsf{v};M)$ differ substantially. In order to localize the source of the disparity, we employ Anselin's definition of local

[3] Census tract graphs for all states (based on 2010 census geography) were obtained from [14]. For tracts with zero population, we define v_i using the average population values (Hispanic, Black, and total) of the neighboring tracts.

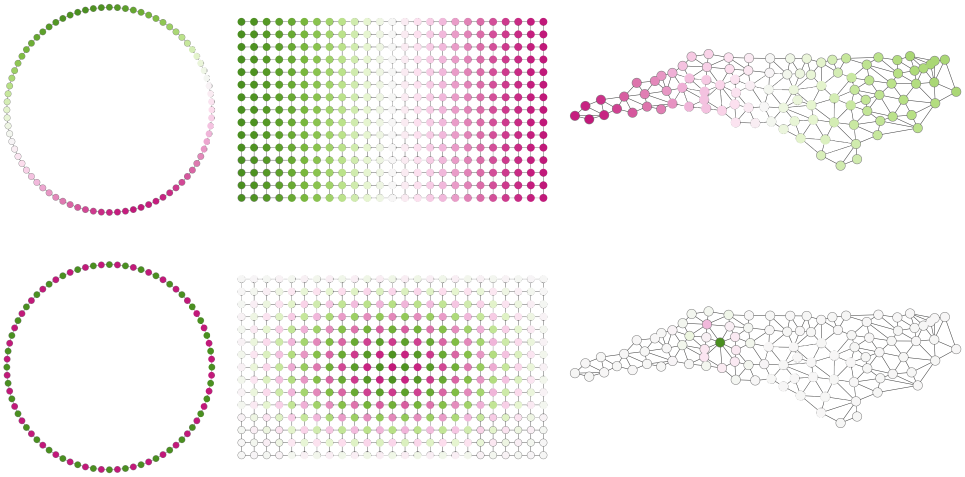


FIG. 8. *For the (even-length) cycle graph, the lowest-frequency nontrivial eigenvector Ψ_2 of L—called the* Fiedler vector*—oscillates slowly and realizes a large $\mathtt{I}$ value, while the highest-frequency eigenvector Ψ_n is a perfect alternating pattern and yields the extremal $\mathtt{I} = -1$. For the grid graph, Ψ_2 oscillates slowly. On the other hand, Ψ_n oscillates very rapidly. Note that the low-frequency eigenvector is highly clustered, while the high-frequency eigenvector is not. The right column shows the county dual graph of North Carolina with its lowest- and highest-frequency eigenvectors. One captures cluster structure, while the other is highly localized at a high-degree vertex.*

Moran's $\mathtt{I}$ at the ith vertex [4], which looks just like the standard definition except the numerator only considers neighbors of i:

$$\mathtt{I}_i(\mathsf{v}; W) := \left(n \sum_{j=1}^{n} W_{ij}(v_i - \bar{v})(v_j - \bar{v}) \right) \Bigg/ \left(w \sum_{j=1}^{n} (v_j - \bar{v})^2 \right).$$

We can then study the contribution of tract i towards the difference by defining $D_i(\mathsf{v}; A, M) := |\mathtt{I}_i(\mathsf{v}; A) - \mathtt{I}_i(\mathsf{v}; M)|$, motivated by the fact that $\mathtt{I}(\mathsf{v}; A)$ and $\mathtt{I}(\mathsf{v}; M)$ agree for regular graphs. Nodes with much higher v values than their neighbors will tend to have high $D_i(\mathsf{v}; A, M)$ values since A does not have diagonal entries, while M does. Also, nodes with low A-degrees tend to have higher diagonal entries in M, thus higher D_i. In Figure 9 a handful of states—North Dakota, Montana, and Mississippi, especially—stand out as having a large discrepancy between $\mathtt{I}(\mathsf{v}; A)$ and $\mathtt{I}(\mathsf{v}; M)$ for one or both v. When we localize to the tracts that contribute most to the disparity (Table 3), we find, as expected, nodes with low A-degree (typically 1 or 2), with concentrations of the minority group that are typically 5–10 times greater than the share in the state overall, and than the share in the neighboring nodes.

The definition of $D_i(\mathsf{v}; A, M)$ was motivated by the fact that $\mathtt{I}(\mathsf{v}; A)$ and $\mathtt{I}(\mathsf{v}; M)$ agree for regular graphs. We could similarly compare $\mathtt{I}(\mathsf{v}; A)$ and $\mathtt{I}(\mathsf{v}; L)$ using the theoretical relationship for regular graphs given in Remark 5.2. We note that ND, MS, and MT remain noticeable outliers in the A versus L comparisons in Figure 9.

6. Laplacian weights. As noted above, the Laplacian L is a very natural choice for matrix-based analysis of a network. The Laplacian has been closely connected with the topic of *community detection* in networks, especially when potential communities are of different sizes [32, 47]. The Laplacian also has a rich theoretical

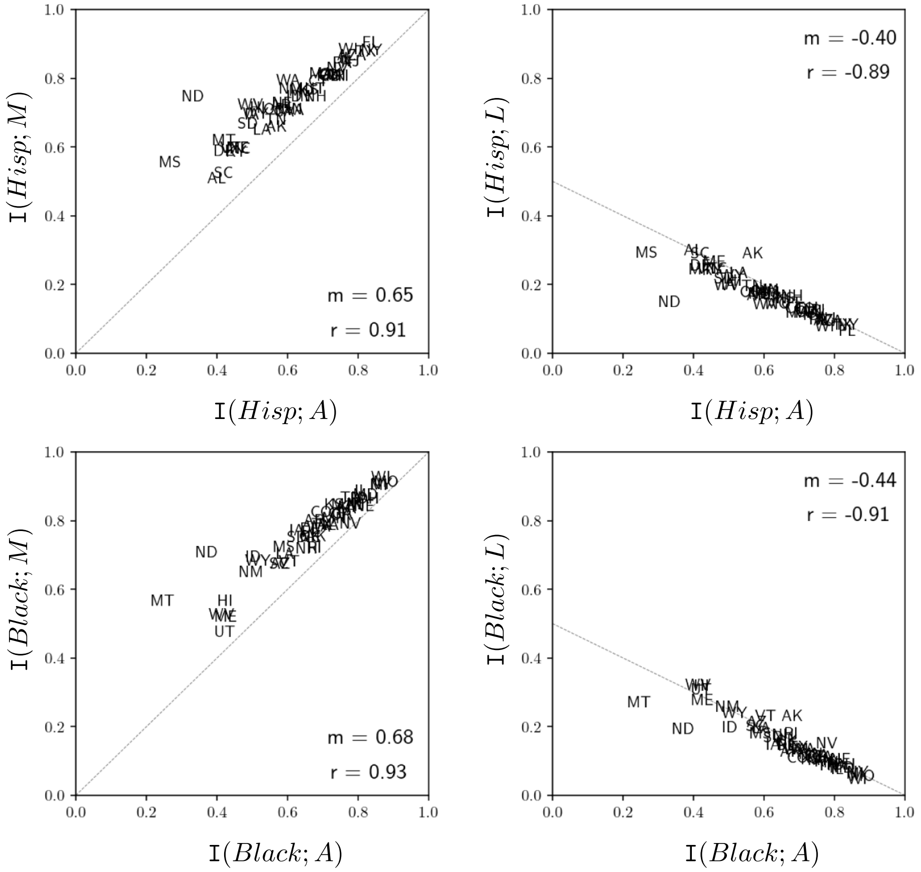


FIG. 9. *Comparing* $\mathtt{I}$ *for the matrices* A, L, M *on* 50*-state census tract data for Hispanic and Black population shares. The line* $y = x$ *is shown for the* A *versus* M *comparison, and the line* $y = (1 - x)/2$ *is shown for comparing* A *to* L*, because the data points would have to fall on these lines if the graphs were regular. The correlation* r *and slope of the best fit line* m *are reported for each plot, but those fit lines are not plotted; their slopes are not equal to* $m = 1$ *and* $m = -1/2$*, indicating different ways of handling degree disparity.*

interpretation in terms of relaxed graph cuts—which lends itself well to measurements of clustering—and in terms of notions of smoothness on graphs via Dirichlet energy functionals. Let $\mathbb{T}^m$ be the m-dimensional torus ($[0, 2\pi]^m$ with opposite boundary faces identified) and recall the L^2 inner product of $f, g \in L^2(\mathbb{T}^m)$, defined as $\langle f, g\rangle := \frac{1}{(2\pi)^m}\int_{\mathbb{T}^m} f(x)\overline{g(x)}dx$.

6.1. High- and low-frequency eigenvectors. By construction, L is symmetric and positive semidefinite, and therefore has a basis of orthonormal eigenvectors $\{\Psi_i\}_{i=1}^n$ with associated real eigenvalues $0 = \mu_1 \leq \cdots \leq \mu_n$. Since $L\mathbb{1} = \mathbb{0}$, we have $\Psi_1 = \frac{1}{\sqrt{n}}\mathbb{1}$, $\mu_1 = 0$. These may be interpreted in the framework of Fourier analysis, in which the eigenvectors $\{\Psi_i\}_{i=1}^n$ of L are the Fourier modes on the graph $\mathcal{G}$ with frequencies $\{\mu_i\}_{i=1}^n \subsetneq [0, \infty)$. In classical Fourier analysis on $\mathbb{T}^1$, the Laplacian operator $\mathcal{L} : f \mapsto -\Delta f = -\nabla \cdot \nabla f = -\frac{d^2}{dx^2}f$ has eigenfunctions $\{\exp(-ikx)\}_{k=-\infty}^{\infty}$ with corresponding eigenvalues $\{k^2\}_{k=-\infty}^{\infty}$, which can be organized from *low frequency* ($|k|$ small) to *high frequency* ($|k|$ large). There is a well-developed literature interpreting

Table 3

Examining the tracts that contribute most to differences between $\mathtt{I}(\mathsf{v}; A)$ and $\mathtt{I}(\mathsf{v}; M)$. We identify the top three $D_i(\mathsf{v}; A, M)$ values, and for those tracts we report the share of Hispanic or Black population (v_i), the average share in neighboring tracts ($(P\mathsf{v})_i$), and the vertex degree d_i. For each state, we also report the average v value and the average value of D_i.

State	v	$\bar{v}$	avg D_i	Tracts with highest D_i	v_i	$(P\mathsf{v})_i$	d_i	D_i
ND	Hisp	0.020	0.0004	Grafton	0.214	0.100	2	0.024
				Minot Air Force Base	0.099	0.012	1	0.009
				Grand Forks Air Force Base	0.100	0.024	2	0.006
MS	Hisp	0.028	0.0006	Morton	0.269	0.049	1	0.061
				Forest	0.268	0.044	2	0.049
				Key Field Air National Guard Base	0.238	0.025	2	0.038
ND	Black	0.009	0.0001	Minot Air Force Base	0.096	0.004	1	0.008
				Grand Forks Air Force Base	0.095	0.006	2	0.006
				Fargo	0.053	0.0013	9	0.001
MT	Black	0.004	1.85e-5	Malmstrom Air Force Base	0.086	0.015	4	0.0025
				Crossroads Correctional Center	0.026	0.003	1	0.0006
				Yellowstone National Park	0.030	0.002	2	0.0005

the graph Laplacian L as a discretization of the continuum differential operator $\mathcal{L}$ [7, 20, 19]. Using analogous language, we can say that Ψ_2 is the lowest-frequency nonconstant eigenvector, while Ψ_n is the highest-frequency eigenvector. (There is such a large literature on Ψ_2 that it has its own name: the *Fiedler vector*.) As noted in Corollary 3.6, these are the extremizers of $\mathtt{I}(\cdot\,; L)$.

We can leverage well-known facts about the Laplacian to rephrase some of the empirical observations in this paper. In particular, if the underlying graph has k internally well-connected components that are weakly connected to each other, then L will have k eigenvalues close to 0 and $\mu_{k+1} \gg 0$ [33]. This gives many qualitatively different functions on the graph that all have a low $\mathtt{I}(\cdot\,; L)$ indicative of clustering.

As we have seen, the case of vertex-regular graphs is one where the eigenvectors of L provide exact solutions to the extremization problem for all four weight matrices. These can be phrased in familiar spectral graph theory language as *nodal decompositions* of the graph into maximal regions where the eigenfunction does not change sign [45]. For the lowest-frequency nonconstant eigenvector Ψ_2, this optimal partition solves a relaxation of the combinatorial normalized cuts functional [40]. Though realistic graphs are not regular, Figure 8 gives an indication that configurations registering maximal segregation are not so far from what a nodal decomposition might predict.

In lattice graphs, which are regular except for along their boundaries, the high-frequency eigenvectors are damped checkerboard patterns, shown in Figure 5. In highly irregular graphs, such as in Figure 8, the checkerboarding may be strongly localized around vertices of high degree. It is clear from both this example and from the double-star graphs in Figure 7 that L is still highly sensitive to degree disparities. In general, the largest eigenvalue may be characterized as a measure of how close the graph is to being bipartite [15], but a corresponding characterization of the highest-frequency eigenvectors is elusive.

6.2. Dirichlet energy functionals.

With the graph Laplacian L, we can quantify a kind of smoothness via a notion of *Dirichlet energy* on a graph.

Suppose that a vector is written as $\mathsf{v} = \sum_{i=1}^{n} \alpha_i \Psi_i$ in the orthonormal basis of eigenvectors of L, so that $\mathsf{x} = \sum_{i=2}^{n} \alpha_i \Psi_i$. Then $\mathsf{v}^\top L \mathsf{v} = \sum_{i=1}^{n} {\alpha_i}^2 \mu_i$, where the right-hand side is large when a large portion of the coefficient energy localizes on the highest-frequency eigenvectors. In analogy with classical Dirichlet energy functionals [17], we may define a *graph Dirichlet energy* for general functions v on the graph, which are not necessarily zero-centered.

DEFINITION 6.1 (Dirichlet energy). *Let* $\mathsf{v}=\sum_{i=1}^{n}\alpha_i\Psi_i$ *be a function on a graph with Laplacian* L *and associated orthonormal eigenvectors* $\{\Psi_i\}_{i=1}^{n}$. *The Dirichlet energy associated to* L *is given by* $\mathcal{E}(\mathsf{v})=\sqrt{\sum_{i=1}^{n}\alpha_i^2\mu_i}=\sqrt{\mathsf{v}^\top L\mathsf{v}}$.

Note that for any scalar $\alpha>0$, $\mathcal{E}(\alpha\mathsf{v})=\alpha\mathcal{E}(\mathsf{v})$. In particular, $\mathcal{E}$ is not scale-invariant, since it lacks the denominator of $\mathsf{v}^T\mathsf{v}$ used when computing $\mathtt{I}$.

Compare this to the classical Dirichlet energy functional on $f:\mathbb{T}^m\to\mathbb{R}$ given by

$$E(f)=\int_{\mathbb{T}^m}\|\nabla f(x)\|_2{}^2dx,$$

where ∇f is interpreted in a weak sense. By Stokes' theorem, $\frac{1}{(2\pi)^m}\int_{\mathbb{T}^m}\|\nabla f(x)\|_2{}^2dx$ $=\langle-\Delta f,f\rangle$ so that $E(f)=\sqrt{\langle\mathcal{L}f,f\rangle}$, in direct analogy to Definition 6.1. To further develop the connection between the graph definition and the classical definition, assume $m=1$ for simplicity. If f has Fourier expansion $f(x)=\sum_{k=-\infty}^{\infty}c_k\exp(-ikx)$, then $\nabla f(x)=-\sum_{k=-\infty}^{\infty}c_kik\exp(-ikx)$. By Parseval's theorem,

$$\|f\|_2{}^2=\sum_{k=-\infty}^{\infty}c_k{}^2,\qquad\|\nabla f\|_2{}^2=\sum_{k=-\infty}^{\infty}k^2c_k{}^2.$$

Noting that $\{k^2\}_{k=0}^{\infty}$ are precisely the eigenvalues of $\mathcal{L}$ defined on $L^2(\mathbb{T}^1)$, we see that the classical and graph definitions agree.

If $\int_{\mathbb{T}^m}\|\nabla f(x)\|_2{}^2dx$ is small, then f is locally smooth in the sense that ∇f has small magnitude in most areas. Noting again that

$$E(f)=\int_{\mathbb{T}^m}\|\nabla f(x)\|_2{}^2dx=-\int_{\mathbb{T}^m}\Delta f(x)f(x)dx=\langle\mathcal{L}f,f\rangle$$

suggests that $\mathsf{x}^\top L\mathsf{x}$, the graph discretization of $\langle\mathcal{L}f,f\rangle$, is a measure of local smoothness of the function x on the graph. This connection is elaborated in [11]. Under this interpretation, $\mathtt{I}(\mathsf{x};L)$ is small (indicating segregation) when x is mostly smooth. Importantly, $\mathtt{I}(\cdot\,;L)$ is scale-invariant, unlike $\mathcal{E}(\cdot)$; a note about developing scale-sensitive measures of segregation will be discussed in section 8.

7. Random walks and I.

We now consider a class of weight matrices for which $\mathtt{I}$ has a random walk interpretation, namely bistochastic matrices (those with rows and columns summing to one).

We first recall some basic facts about Markov chains and random walks. By definition, a *Markov chain* on a finite state space (with states indexed $1,\dots,n$) is a random process encoded by a stochastic $n\times n$ matrix K. The associated random walk steps from the ith to the jth state with probability K_{ij}. This can be visualized as random walk on a graph with n nodes, and an edge (i,j) present when K_{ij} or $K_{ji}>0$. We can encode the walk by matrix multiplication if a probability vector v is interpreted as describing a probabilistic position on the state space. Then $\mathsf{v}^\top K$ is the new position after one step of the walk.

With this, the reader can verify that our spatial weight matrix P discussed above has an interpretation as the simple random walk on the geography units, making all neighboring units equally likely at each stage. As long as the graph is connected and aperiodic (for instance, if it has any triangles), this random walk converges to a unique stationary distribution in which the probability of being at any node in the long term is proportional to its degree. When the graph encodes the tracts of a city, as in many of our examples here, this stationary distribution for P is not very meaningful.

Given an arbitrary Markov chain K, the classical *Metropolis–Hastings* construction allows us to modify the random walk so that it targets a specified stationary distribution π (an arbitrary probability distribution on the states $1, \dots, n$). The Metropolis–Hastings matrix $M = M(K, \pi)$ gives a reversible chain, meaning that $\pi(i)M_{ij} = \pi(j)M_{ji}$ for all i, j. Note that if the Metropolis–Hastings matrix is set to target the uniform distribution, then reversibility means the matrix is symmetric and therefore bistochastic. Next, we establish that for any bistochastic matrix Q, such as for the uniformizing Metropolis matrix $M = M(P, \frac{1}{n}\mathsf{1})$, Moran's $\mathtt{I}$ can be interpreted in terms of variance reduction.

THEOREM 7.1 (Random walk interpretation of $\mathtt{I}$). *For a bistochastic matrix Q and a column vector v, consider $\mathsf{w} = \mathsf{v}^\top Q$, the value of v after one step of the Markov chain given by Q. Let σ_0 and σ_1 be the standard deviation of the values in v and w, respectively, so that the ratio σ_1/σ_0 gives the variance reduction in one step of the walk. Let $\rho(\mathsf{v}, \mathsf{w})$ be the correlation between the values in v and w. Let $\mathsf{x} = \mathsf{v} - \bar{v}\mathsf{1}$ and $\mathsf{y} = \mathsf{w} - \bar{w}\mathsf{1}$ be the zero-centered vectors before and after applying Q. Then*

(a) $\mathtt{I}(\mathsf{v}; QQ^\top) = (\frac{\sigma_1}{\sigma_0})^2$;

(b) $\mathtt{I}(\mathsf{v}; Q) = \frac{\mathsf{y}^\top \mathsf{x}}{\mathsf{x}^\top \mathsf{x}} = \rho(\mathsf{v}, \mathsf{w}) \cdot \frac{\sigma_1}{\sigma_0}$.

Proof. To see (a), note that because Q is bistochastic, the average values satisfy $\overline{\mathsf{v}^\top Q} = \frac{1}{n} \cdot \mathsf{v}^\top Q\mathsf{1} = \frac{1}{n} \cdot \mathsf{v}^\top \mathsf{1} = \overline{\mathsf{v}^\top}$. Thus

$$\frac{{\sigma_1}^2}{{\sigma_0}^2} = \frac{(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top}\mathsf{1}^\top)(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top}\mathsf{1}^\top)^\top}{(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})^\top(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})} = \frac{(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top}\mathsf{1}^\top Q)(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top}\mathsf{1}^\top Q)^\top}{(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})^\top(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})}$$

$$= \frac{(\mathsf{v}^\top - \overline{\mathsf{v}^\top}\mathsf{1}^\top)QQ^\top(\mathsf{v}^\top - \overline{\mathsf{v}^\top}\mathsf{1}^\top)^\top}{(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})^\top(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})} = \mathtt{I}(\mathsf{v}; QQ^\top).$$

A similar calculation yields (b):

$$\mathtt{I}(\mathsf{v}; Q) = \frac{(\mathsf{v}^\top - \overline{\mathsf{v}^\top}\mathsf{1}^\top)Q(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})}{(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})^\top(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})} = \frac{(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top Q}\mathsf{1}^\top)(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})}{(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})^\top(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})} = \frac{\mathsf{y}^\top \mathsf{x}}{\mathsf{x}^\top \mathsf{x}}$$

$$= \frac{(\mathsf{v}^\top Q - \overline{\mathsf{v}^\top Q}\mathsf{1}^\top)(\mathsf{v} - \overline{\mathsf{v}}\mathsf{1})}{||\mathsf{v}^\top Q - \overline{\mathsf{v}^\top Q}\mathsf{1}^\top||_2 \cdot ||\mathsf{v} - \overline{\mathsf{v}}\mathsf{1}||_2} \cdot \frac{||\mathsf{v}^\top Q - \overline{\mathsf{v}^\top Q}\mathsf{1}^\top||_2}{||\mathsf{v} - \overline{\mathsf{v}}\mathsf{1}||_2} = \rho(\mathsf{v}^\top, \mathsf{v}^\top Q) \cdot \frac{\sigma_1}{\sigma_0}. \qquad \square$$

Note $\rho(\mathsf{v}, \mathsf{w})$ is just the one-step autocorrelation (i.e., time lag 1) for the Markov chain Q. Part (a) of Theorem 7.1 states that $\mathtt{I}(\mathsf{v}; QQ^\top)$ can be interpreted as the factor by which the variance is reduced in two steps of evolution under the Markov chain associated to Q. Note that a general weight matrix W admits a decomposition $W = QQ^\top$ for such a matrix Q iff W is bistochastic, positive semidefinite, and symmetric. We also observe that $\mathtt{I}(\mathsf{v}; QQ^\top)$ is always nonnegative. Part (b) of Theorem 7.1 states that $\mathtt{I}(\mathsf{v}; Q)$ decomposes as the product of the one-step autocorrelation for v and the reduction in standard deviation after one step. To see how this plays out in extreme cases, observe that if $|\mathtt{I}(\mathsf{v}; Q)| \approx 1$, then the standard deviation of v must remain roughly the same after one step of Q, with the value of v after one step being either highly correlated ($\mathtt{I}(\mathsf{v}; Q) \approx 1$) or anticorrelated ($\mathtt{I}(\mathsf{v}; Q) \approx -1$) with the initial value of v. The alternative expression $\mathtt{I}(\mathsf{v}; Q) = \frac{\mathsf{y}^\top \mathsf{x}}{\mathsf{x}^\top \mathsf{x}}$ makes it clear that if a step of Q

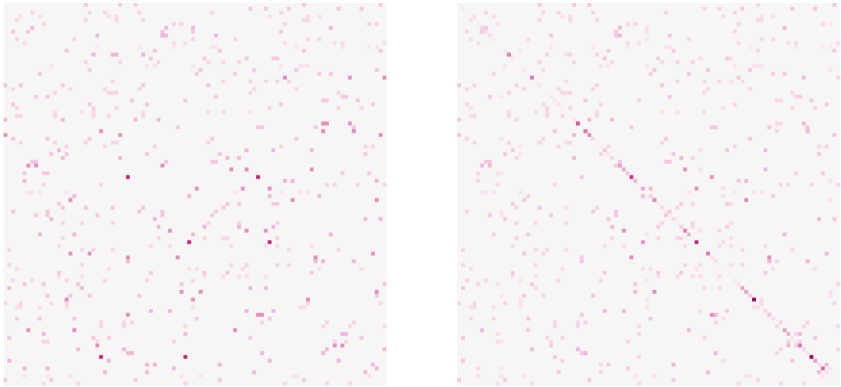


FIG. 10. *The matrices P (left) and M (right) for the North Carolina county dual graph shown in Figure 8. Darker colors indicate higher matrix entries. We see that M has some large diagonal entries, meaning that the associated random walk is very lazy at some nodes.*

changes v to something near-uniform, then $\mathtt{I}$ will have small magnitude. Large $\mathtt{I}$ can only occur when the diffusion process leaves $\mathsf{y}^\top \mathsf{x} \approx \mathsf{x}^\top \mathsf{x}$.

When we pass from the (symmetric) adjacency matrix A to the stochastic normalization P, this need not be bistochastic unless the underlying graph is regular. To take advantage of this theorem, we will use the uniformizing Metropolis–Hastings matrix $M = M(P, \frac{1}{n}\mathbb{1})$, which is both symmetric and bistochastic.

In fact, M is a best symmetric bistochastic approximation to P in the sense that it minimizes the difference $\sum_{i\neq j} |Q_{ij} - P_{ij}|$ among all symmetric bistochastic matrices Q. This follows from the work of Billera and Diaconis in [10], which proves the more general statement that the Metropolis–Hastings matrix $M(K, \pi)$ whose i, j term is

$$M(K,\pi)_{ij} = \min\left(K_{ij}, \quad \frac{\pi(j)}{\pi(i)} K_{ji}\right)$$

minimizes the π-weighted ℓ^1 distance from K to $\{Q\}$. In particular, in our setting, if the graph $\mathcal{G}$ is regular, then P is already symmetric and bistochastic, so $M = P$. In the irregular case, M can be thought of as a modification of the simple random walk; it works by introducing a rejection step that makes the walk extremely lazy when a low-degree vertex is next to a high-degree vertex. (See Figure 10 for an illustration.)

Using M for the spatial weighting in $\mathtt{I}$ succeeds where P does not in mitigating the degree effects discussed throughout this paper. This is because M eliminates W-degree discrepancies in both the rows and columns of A, while P standardizes only the rows. Indeed, because M is a nonnegative stochastic matrix, its largest eigenvalue is 1, realized by its stationary vector $\mathbb{1}$. (This is a well-known Markov chain property following from the Perron–Frobenius theorem.) We recall from Corollary 3.6 that the extreme values of $\mathtt{I}(\cdot\,; M)$ are realized at the eigenvalues λ_2 and λ_n of M.

COROLLARY 7.2 (The range of $\mathtt{I}$ with weights from M). *For any graph $\mathcal{G}$ and any function v, we have $-1 \leq \mathtt{I}(\mathsf{v}; M) \leq 1$.*

This discussion suggests another way in which the random walk interpretation can be fruitful. It is a standard fact in Markov chain theory that the convergence statistics (such as *mixing time*) of a chain have upper and lower bounds in terms of the spectral gap, here $1 - \lambda_2$, of the associated matrices. Random walks that converge more slowly correspond to smaller spectral gaps and smaller Cheeger constants (graphs that have relatively short cuts into large pieces). In the world of geography dual graphs, this says that if the locality itself can be cut in half with a relatively short cut, as in all of the realistic examples here, then $\mathtt{I}$ scores for M can get close to 1. Indeed, planar graphs where the number of vertices is much greater than the largest vertex degree are slow-mixing and have a small spectral gap [42], which ensures that $\mathtt{I}(\cdot\,; M) \approx 1$ is achievable.

In sum, using M for the spatial weight matrix allows us to characterize $\mathtt{I}$ in extremely intuitive language. We imagine a diffusion process that begins with the observed demographic distribution in a locality and conducts a random walk of residents that targets the uniform distribution. Over the long term, this random walk process must reduce the variance, which is initially $\|\mathsf{x}\|$, to zero. Moran's $\mathtt{I}$ now measures *how well a uniformizing diffusion succeeds in a single step*. It is quite reasonable to regard this as a measurement of segregation: a certain group will be considered very far from uniformly dispersed in a population if many steps through neighboring geography are required for the group to approach uniformity.

8. Conclusions, recommendations, and future directions. Moran's $\mathtt{I}$ is a valuable way to detect spatial patterns, or to test for spatial correlation in the residuals of some models, especially when combined with a statistical significance test. However, users must exercise caution when using Moran's $\mathtt{I}$ as a gradated measurement (and not just a qualitative test) for a number of reasons. The underlying graph topology and the choice of spatial weight matrix used in computing $\mathtt{I}$ both strongly impact the range of possible $\mathtt{I}$ values, so using $\mathtt{I}$ to compare across localities remains challenging.

We summarize the findings of this paper with the following practical recommendations.

(1) For a given graph $\mathcal{G}$ and weight matrix W, use the methods here to compute the range of achievable $\mathtt{I}$ values in order to decide whether a particular demographic vector v has an extreme score. However, intermediate $\mathtt{I}$ values (say, $\mathtt{I} = .6$) remain hardest to interpret.

(2) Use circumspect language when comparing $\mathtt{I}$ values for different graphs $\mathcal{G}$ and $\mathcal{G}'$, particularly with standard choices of weight matrix like A and P. The computation $\mathtt{I}(\mathsf{v}; P_{\mathcal{G}}) > \mathtt{I}(\mathsf{v}'; P_{\mathcal{G}'})$ should not be presented as a finding that the first city is more segregated than the second.

(3) Both for within-graph and between-graph comparisons, the best-suited spatial weight matrix is M, which makes $\mathtt{I}$ interpretable in terms of how the subgroup's population diffuses in a random walk. Furthermore, with M weights, $\mathtt{I}$ is actually bounded between -1 and 1 and for large planar graphs can achieve $\mathtt{I} \approx 1$.

(4) The discussion above suggests a novel role for Moran's $\mathtt{I}$ when a demographic function v is fixed on a sufficiently fine graph. Recall that the dependence of a measurement on the choice of units is an important problem in geography called the *modifiable areal unit problem*. Given v, a choice of units for which $\mathtt{I} < 0$ can be interpreted as an aggregation of the underlying fine data that captures regions that are demographically distinct—having similar demographics within the unit and different demographics on neighboring units. That is, when considering alternative choices of geographical units (like census tracts in Chicago versus official neighborhoods maintained in city statistics), a *negative* Moran's score can be interpreted as a signal that the units track with demographic differences. From this point of view, $\mathtt{I}$ can facilitate a kind of demographic community detection.

This study suggests several interesting questions for future exploration. While the connection between high segregation and low-frequency eigenvectors follows from our analysis, the connection between low segregation and high-frequency eigenvectors is more subtle. This is due to the large impact that the underlying graph geometry has on even the local properties of high-frequency eigenvectors (see Figure 8). Understanding the extent to which high-frequency eigenvectors localize (i.e., have concentrated support) on irregular graphs would be interesting in its own right, and has potential connections to Anderson localization in the continuum setting [18].

Another useful direction of inquiry would be to modify the definition of $\mathtt{I}$ by building new metrics that make use of $\mathsf{x}^\top W \mathsf{x}$ without normalizing by the denominator $\mathsf{x}^\top \mathsf{x}$. By creating scale-sensitive scores for the deviation in a population, we could remediate the degeneration in the interpretation of $\mathtt{I}$ for low-variance distributions.

Acknowledgments. We thank Larry Guth and Eugene Henninger-Voss for enlightening conversations. We also thank the two anonymous referees for their helpful comments that significantly improved the paper.

REFERENCES

[1] *How Spatial Autocorrelation (Global Moran's* I*) Works*, https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm (accessed 19 December 2021).

[2] J. A. Agnew and D. N. Livingstone, *The Sage Handbook of Geographical Knowledge*, Sage Publications, London, 2011.

[3] E. Alvarez, M. Duchin, E. Meike, and M. Mueller, *Clustering Propensity: A Mathematical Framework for Measuring Segregation*, preprint, 2018, https://mggg.org/publications/capy.pdf.

[4] L. Anselin, *Local indicators of spatial association—LISA*, Geogr. Anal., 27 (1995), pp. 93–115.

[5] L. Anselin, *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*, in Spatial Analytical Perspectives on GIS, Routledge, 2019, pp. 111–126.

[6] L. Anselin, *Spatial Econometrics: Methods and Models*, Vol. 4, Springer, Dordrecht, 2013.

[7] M. Belkin and P. Niyogi, *Convergence of Laplacian eigenmaps*, in Advances in Neural Information Processing Systems 19 (NIPS 2006), MIT Press, 2006, pp. 129–136.

[8] B. Berry, J. Lobley, and D. Marble, *Spatial Analysis: A Reader in Statistical Geography*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[9] P. Bhakta, S. Miracle, and D. Randall, *Clustering and mixing times for segregation models on $\mathbb{Z}^2$*, in Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2014, pp. 327–340, https://doi.org/10.1137/1.9781611973402.24.

[10] L. Billera and P. Diaconis, *A geometric interpretation of the Metropolis-Hastings algorithm*, Statist. Sci., 16 (2001), pp. 335–339.

[11] F. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, RI, 1997.

[12] C. Dawkins, *Measuring the spatial pattern of residential segregation*, Urban Stud., 41 (2004), pp. 833–851.

[13] P. De Jong, C. Sprenger, and F. V. Veen, *On extreme values of Moran's I and Geary's c*, Geogr. Anal., 16 (1984), pp. 17–24.

[14] D. DeFord, *Census Dual Graphs for* 2010 *Census Units*, https://people.csail.mit.edu/ddeford/dual_graphs.html.

[15] M. Desai and V. Rao, *A characterization of the smallest eigenvalue of a graph*, J. Graph. Theory, 18 (1994), pp. 181–194.

[16] M. Duchin and J. Murphy, *Measuring clustering and segregation*, in Political Geometry, M. Duchin and O. Walch, eds., Birkhäuser, Cham, 2022, pp. 293–302.

[17] L. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

[18] M. Filoche and S. Mayboroda, *Universal mechanism for Anderson and weak localization*, Proc. Natl. Acad. Sci. USA, 109 (2012), pp.14761–14766.

[19] N. Garcia Trillos, M. G. M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator*, Found. Comput. Math., 20 (2020), pp. 827–887.

[20] N. Garcia Trillos and D. Slepčev, *A variational approach to the consistency of spectral clustering*, Appl. Comput. Harmon. Anal., 45 (2018), pp. 239–281.

[21] A. Getis, *A history of the concept of spatial autocorrelation: A geographer's perspective*, Geogr. Anal., 40 (2008), pp. 297–309.

[22] A. Getis, *Spatial weights matrices*, Geogr. Anal., 41 (2009), pp. 404–410.

[23] D. Griffith and P. Peres-Neto, *Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses*, Ecology, 87 (2006), pp. 2603–2613.

[24] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 2012.

[25] E. Ising, *Beitrag zur Theorie des Ferromagnetismus*, Z. Phys., 31 (1925), pp. 253–258.

[26] P. Legendre, *Spatial autocorrelation: Trouble or new paradigm?*, Ecology, 74 (1993), pp. 1659–1673.

[27] D. Massey and N. Denton, *The dimensions of residential segregation*, Soc. Forces, 67 (1988), pp. 281–315.

[28] P. Moran, *Notes on continuous stochastic phenomena*, Biometrika, 37 (1950), pp. 17–23.

[29] M. Newman, *Assortative mixing in networks*, Phys. Rev. Lett., 89 (2002), 208701.

[30] M. Newman, *Mixing patterns in networks*, Phys. Rev. E, 67 (2003), 026126.

[31] M. Newman, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256, https://doi.org/10.1137/S003614450342480.

[32] M. Newman, *Networks*, Oxford University Press, Oxford, 2018.

[33] A. Ng, M. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Advances in Neural Information Processing Systems 14 (NIPS 2001), MIT Press, 2001, pp. 849–856.

[34] A. Ortega, P. Frossard, J. Kovačević, J. Moura, and P. Vandergheynst, *Graph signal processing: Overview, challenges, and applications*, Proc. IEEE, 106 (2018), pp. 808–828.

[35] G. Paouris, *Concentration of mass on convex bodies*, Geom. Funct. Anal., 16 (2006), pp. 1021–1049.

[36] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998, https://epubs.siam.org/doi/pdf/10.1137/1.9781611971163.fm.

[37] B. R. Roberts and R. H. Wilson, *Urban Segregation and Governance in the Americas*, Palgrave Macmillan, New York, 2009.

[38] T. Schelling, *Dynamic models of segregation*, J. Math. Sociol., 1 (1971), pp. 143–186.

[39] A. D. Cliff and J. K. Ord, *The problem of spatial autocorrelation*, in Studies in Regional Science, A. J. Scott ed., Pion, London, 1969, pp. 25–55.

[40] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.

[41] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*, IEEE Signal Process. Mag., 30 (2013), pp. 83–98.

[42] D. A. Spielman and S.-H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, in 37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996), IEEE Comput. Soc. Press, Los Alamitos, CA, 1996, pp. 96–105, https://doi.org/10.1109/SFCS.1996.548468.

[43] M. Tiefelsdorf and B. Boots, *The exact distribution of Moran's I*, Environ. Plann. A, 27 (1995), pp. 985–999.

[44] M. Tiefelsdorf and D. Griffith, *Semiparametric filtering of spatial autocorrelation: The eigenvector approach*, Environ. Plann. A, 39 (2007), pp. 1193–1221.

[45] J. C. Urschel, *Nodal decompositions of graphs*, Linear Algebra Appl., 539 (2018), pp. 60–71, https://doi.org/10.1016/j.laa.2017.11.003.

[46] C. F. Van Loan and G. Golub, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1996.

[47] U. Von Luxburg, *A tutorial on spectral clustering*, Statist. Comput., 17 (2007), pp. 395–416.

[48] Z. Zhao and D. Randall, *A Heterogeneous Schelling Model for Wealth Disparity and Its Effect on Segregation*, preprint, https://arxiv.org/abs/2108.01657v2, 2022.