Published in the Quarterly Journal of Economics
Volume 122(2), May 2007, pages 441-485.

A MEASURE OF SEGREGATION BASED ON SOCIAL INTERACTIONS *

Federico Echenique Roland G. Fryer, Jr.

June 2006

Abstract

We develop an index of segregation based on two premises: (1) a measure of segregation should disaggregate to the level of individuals, and (2) an individual is more segregated the more segregated are the agents with whom she interacts. We present an index which satisfies (1) and (2), and that is based on agents' social interactions: the extent to which Blacks interact with Blacks, Whites with Whites, etc. We use the index to measure school and residential segregation. Using detailed data on friendship networks, we calculate levels of within-school racial segregation in a sample of US schools. We also calculate residential segregation across major US cities, using block-level data from the 2000 US Census.

*We are grateful to Gary Becker, Kim Border, Fernando Borraz, Toni Calvo-Armengol, David Card, Joan Esteban, Drew Fudenberg, Edward Glaeser, Jerry Green, Faruk Gul, Oliver Hart, James Heckman, Matthew Jackson, Kevin Lang, Edward Lazear, Erzo Luttmer, Derek Neal, Jesse Shapiro, and two anonymous referees for useful comments and suggestions. We are especially grateful to Lawrence Katz, Glenn Loury, and Kevin Murphy for advice, encouragement, and detailed comments on previous drafts. We thank seminar participants at Berkeley, Boston, Brown, Caltech, Carnegie Mellon, Chicago, Harvard, McGill, NBER, Pompeu Fabra, Torcuato Di Tella, Universitat Autonoma de Barcelona, and Vanderbilt. Katherine Barghaus, Patricia Foo, and Alex Kaufman provided exceptional research assistance. A portion of this paper was written while Fryer was a visitor at Institut d'Anàlisi Económica at Universitat Autonoma in Barcelona, Spain. Fryer gratefully acknowledges financial support from the Alphonse Fletcher Sr. Fellowship.

I. Introduction

Ethnic and racial segregation is an important and well-studied social phenomenon. For over 50 years, social scientists have been concerned with measuring the extent, and estimating the impact of, segregation in education, housing, and the labor market. The result of this scholarship has been nearly 20 different indices of segregation, and a consensus that the spatial separation of many minorities from jobs, role models, health care, and quality local public goods is a leading cause of racial and ethnic differences on many economic, social, and health related outcomes [Almond, Chay, and Greenstone 2003; Borjas 1995; Case and Katz 1991; Kain 1968; Cutler and Glaeser 1997; Massey and Denton 1993; Collins and Williams 1999].

We propose a new approach to measuring segregation based on two premises: (1) a measure of segregation should disaggregate to the level of individuals, and (2) an individual is more segregated the more segregated are the agents with whom she interacts. Having a measure of segregation with the flexibility to disaggregate to the level of individuals opens up windows of opportunity for empirical work, and a better understanding of the mechanisms by which social interactions affect economic and social outcomes. We also desire a measure that gives a larger level of segregation for individuals whose contacts are more segregated. Consider Figure I, which depicts the distribution of blacks across metropolitan Detroit, Michigan. There is a large oval in the center of the city containing almost exclusively black households. Any measure of segregation should report that the household in the epicenter is more segregated than a household close to the edge, even when each household has all black neighbors.

insert figure I

We use social networks – individuals and their connections – as our mathematical framework. In this framework, we propose three specific properties that any measure of segregation in a network should satisfy. We prove that one and only one index satisfies these properties and the two broad principles above; which we label the "Spectral Segregation Index" (SSI).

The properties require that: (a) [Monotonicity] if all individuals in Network A have a larger share of their interactions with agents of the same group than in Network B, then Network A is more segregated than B; (b) [Linearity] an individual is more segregated the more segregated are the agents with whom she interacts, and this relationship takes on a linear form; and (c) [Homogeneity] if all individuals in a network have half of their interactions with members of the same group, the index of segregation is one-half. The latter condition normalizes the index.

We defer a formal definition of the SSI to Section IV.. Informally, the SSI measures the connectedness of individuals of the same group.[1] Consider the following recursion. Define "first-order segregation" as the share of one's social interactions that are with individuals of their own group. Let "second-order segregation" be the average over all own-group social interactions of their first order segregation. Following this line, an agent's n^{th} order segregation is the average over own group connections of their $n-1$ order segregation, and so on. The SSI of an individual is the limit, as $n \to \infty$, of that individual's n^{th} order segregation.

The SSI has important advantages over existing measures of segregation. First, as a gauge of residential segregation, it is invariant to arbitrary partitions of a city; existing measures are not.[2] Second, it allows one to investigate how segregated multiple minority groups are, permitting comparisons of Asians, Blacks, Hispanics, Native Americans, and so on, within and across cities.[3] The SSI makes it possible to compare Hispanic segregation across cities, compare the Hispanics of east Los Angeles from the Hispanics in south Los Angeles, or compare them to Blacks in Chicago. Third, our index allows one to analyze the full distribution of segregation, allowing researches to move beyond aggregate statistics, which can be misleading. The typical Black household is more segregated than the typical Hispanic household, yet the most segregated Hispanics are orders of magnitude more segregated than any

1. Groups can be defined in terms of gender, political affilation, educational attainment, race/ethnicity, and so on. Our empirical applications are to race/ethnicity.

2. As a practical matter, we use the most disaggregated data publicly available: census blocks.

3. Another way to analyze multiple groups with existing indices is to calculate the weighted average of several dichotomous indices (see Reardon and Firebaugh [2002]). It is not clear how to interpret the findings from such an exercise.

Blacks. Fourth, there are inherent multiplicative effects captured by SSI which other indices omit. An individual's susceptibility to group-transmitted influences depends on how many contacts the individual has with members of the group, the susceptibility of her contacts, the susceptibility of their contacts, and so on.

The SSI has some disadvantages as well. It depends on the quality of the information one can obtain about social interactions. In the case of residential segregation, for example, the information is restricted to where individuals live within a city and not how they interact. Unlike other indices, however, as better information on the nature of social interactions is obtained, the SSI becomes a sharpened proxy of those interactions. Second, it is sensitive to the fraction of individuals in a network who have the race/ethnicity under study. We address this issue by calculating a "baseline,"and adjusting actual SSI taking this into account. Finally, implementing the SSI can be computationally demanding, though our applications demonstrate that the computational tasks are often feasible.[4]

After formally deriving the SSI, we apply the index to two well-known social phenomena: measuring the extent of school and residential segregation. We begin by measuring within-school segregation patterns, by race, using data on friendship networks available in the National Adolescent Study of Health (Addhealth). Our analysis unearths a rich set of new facts. First, the relationship between the share of black students in a school and their segregation is non-linear: When black students are relatively scarce in a school, their friendship networks tend to be integrated. As their share of the student population increases, segregation increases dramatically, plateauing when blacks comprise roughly twenty-five percent of the student population. Schools that have twenty-five percent or more black students exhibit severe within school racial segregation of social interactions. This phenomenon undermines the intuition that a school that has equal shares of black and white students is well integrated. A similar, though less pronounced, pattern exists among Asians and Hispanics, and is weaker still for Whites. The common practice of using the percentage of a racial

4. We have posted results from some of the more computationally intense calculations on the authors' webpages: `http://www.hss.caltech.edu/~fede/` (Echenique);
`http://post.economics.harvard.edu/faculty/fryer/projects.html` (Fryer)

group in a school as a proxy for within school segregation measures for that group is deeply problematic.

We also calculate the extent of segregation across major cities in the US, using block-level data from the 2000 Census. We find that, on average, Blacks are more segregated than any other racial group, but the most segregated Hispanics are more segregated than the most segregated Blacks. A virtue of the SSI is the ability to measure segregation at disaggregated levels, allowing one to measure the intensity of same-race clusters or uncover the most segregated city blocks in America. For example, we find that the largest minority ghetto in the US consists of Hispanics in Los Angeles, CA – 17,909 blocks are connected to each other. It is important to emphasize that these dissagregated results cannot be obtained with any of the existing measures of segregation. We also use SSI to correlate segregation with several MSA-level variables, and replicate Cutler and Glaeser's [1997] classic work on gettoes.

We compare our results to existing calculations applying commonly-used measures. The rank correlation between the SSI and the popular dissimilarity index is .42. The rank correlation with the index of isolation is .93. Our index can be interpreted as a measure of segregation as isolation that is rooted in a social-interactions framework.

The organization of the paper is as follows. Section II. provides a brief discussion of existing segregation indices. Section III. provides an example that previews our general results. Section IV. derives the SSI. Section VI. uses the SSI to estimate the prevalence of within-school and residential segregation. Section VII. concludes. There are two appendices. Appendix A contains the technical proofs of all formal results and additional theoretical results omitted from the text. Appendix B presents a guide to the programs we used to compute our index.

II. Background and Previous Literature

At an abstract level, segregation is the degree to which two or more groups are separated from each other. However, practical definitions can be quite distinct from one another,

conceptually and empirically. Massey and Denton [1988] group existing indices into five classes: evenness, exposure, concentration, centralization, and clustering, which they take to resemble the totality of what is usually meant by "segregation." Evenness refers to the differential distribution of two groups across areas in a city. Measures of exposure are designed to approximate the amount of potential contact and interaction between members of different groups. Concentration indices measure the relative amount of physical space occupied by a minority group. Centralization is the extent to which a group is located near the center of an urban area, and clustering measures the degree to which geographic units inhabited by minority members abut one another, or cluster spatially. Of the five dimensions of segregation, only two are used in the vast majority of applied work in the social sciences: evenness and exposure. Economists ultimately care about the degree to which segregation affects social interactions. For this purpose, concentration and centralization are inadequate, and measures of clustering are largely avoided due to their sensitivity to the number and population of census regions.

The most popular measure of segregation is the "dissimilarity" index (developed by Jahn, Schmid, and Schrag [1947]), a measure of evenness.[5] Suppose a city is divided into N sections. The dissimilarity index measures the percentage of a group's population that would have to change sections for each section to have the same percentage of that group as the whole city. In symbols:

(1)
$$\text{index of dissimilarity} = \frac{1}{2}\sum_{i=1}^{N}\left|\frac{black_i}{black_{total}} - \frac{nonblack_i}{nonblack_{total}}\right|,$$

where $black_i$ is the number of blacks in area i, $black_{total}$ is the total number of blacks in the city as a whole, $nonblack_i$ is the number of non-blacks in area i, and $nonblack_{total}$ is the number of non-blacks in the city. The dissimilarity index has the appealing feature that it

5. Other measures of evenness include the Gini coefficient (the mean absolute difference between minority proportions weighted across all pairs of geographic units, expressed as a proportion of the maximum weighted mean difference), the Atkinson index (similar to Gini coefficient, but allows researchers to decide how to weight geographic units which are over or under the city-wide distribution), and Entropy (the weighted average of each geographic units deviation from the racial entropy of the city as a whole).

is invariant to the size of a minority group.

A second commonly-used measure of segregation is "isolation," a measure of exposure. As Blau [1977] recognized, Blacks can be evenly distributed among residential areas in a city, but experience little exposure to non-Blacks if they are a relatively large proportion of the city. Isolation measures the extent to which Blacks are exposed only to one other, rather than to non-Blacks. The index is computed as the minority-weighted average of each section's minority population:

$$\text{index of isolation} = \sum_i \left(\frac{black_i}{black_{total}} \cdot \frac{black_i}{person_i} \right),$$

where $person_i$ refers to the total population of area i.[6]

insert figure II

Dissimilarity and isolation possess at least two undesirable properties. First, they explicitly depend on the arbitrary ways in which cities are partitioned into sections (e.g. census tracts).[7] That is, fixing the location of minorities and non-minorities in a city and re-drawing the sections can drastically change the measure of segregation. An exaggerated example is depicted in Figure II. The city depicted in the figure has a dissimilarity index of 0 – perfect integration – when sections are drawn vertically and has a dissimilarity index of 1 – extreme segregation – when sections are drawn horizontally; no household has moved. Similarly, vertical partitions yield an isolation index of .5 whereas horizontal partitions produce an index of 1. This is a highly undesirable property of any segregation index, as it may artificially indicate that a city is more or less segregated as a function of how the tracts are drawn. The key flaw is that there is no theory of how the city should be partitioned. Intuition suggests

6. Another commonly used measure of exposure is the interaction index, which is the inverse of the isolation index presented above.

7. We are not the first to draw attention to this flaw in measures of segregation, see Cowgill and Cowgill [1951], Appendix A in Taeuber and Taeuber [1965], and Massey and Denton [1988]. While this property is problematic for measures of residential segregation, it is less likely to effect measures of occupational or school segregation – where there is a natural clustering of individuals.

that the more disaggregated the better, but complete disaggregation results in all sections having only one race: maximum segregation, regardless of the city.

Second, existing measures are not defined when trying to measure segregation at the level of individuals. It is difficult to correctly identify the relationship between segregation and outcomes without individual-level variation in segregation. As a descriptive matter, individual segregation may be more useful than city-wide segregation. Rather than correlate individual economic outcomes with city-wide segregation, one can correlate individual outcomes with individual measures of segregation. On the other hand, the right level of aggregation depends on the problem at hand; group-level, neighborhood, or city-level segregation may be the appropriate level of aggregation in many applications. It is an open empirical question, one that cannot be answered without a measure that disaggregates to the individual level.[8]

The literature in economics involving the measurement of segregation is small (Phillipson [1993], Hutchens [2001], Frankel and Volij [2004]). Similar to our exercise, their approach is axiomatic – identifying desirable properties that an index should possess. The literature takes an arbitrary partition of a city as given, and uses the partition to identify indices axiomatically. There is little in common with our approach.

III. A Motivating Example

insert figure III

Before moving to a full description of the model, we present a stark example which previews the Spectral Index and discusses (informally) some of its properties.

Consider City 1, depicted in Figure III. The nodes in City 1 represent households. Each household can be one of two races: black or white. In the figure, household $(A, 1)$ is white, $(B, 1)$ is black, and so on.

8. This critique is conceptual – not purely data driven. Existing measures are not equipped to measure segregation at the level of individuals, irrespective of the available data.

Our measure of segregation is based on the social network of the members of a race. Consider the black households in City 1. For the purposes of this example, we use the information on where an individual lives to infer whom she interacts with, and trace out a network of social interactions based on residential patterns. Suppose that each individual interacts only with her immediate neighbors; $(A, 1)$ interacts with $(B, 1)$ and $(A, 2)$; $(D, 4)$ interacts with $(C, 4)$, $(E, 4)$, $(D, 3)$, and $(D, 5)$, and so on. The resulting network of black households is shown on the right in Figure III. The thickness of a line connecting two individuals reflects the intensity of their relationships; thicker lines imply a node is at least one-third of an individual's social interactions. Here, $(B, 2)$ has four neighbors, so she has a less intense relation to each one of them than $(B, 1)$, who has only three neighbors.

Black households are partitioned in two separate networks. We call each of these subnetworks a *connected component*. The fact that social networks are often partitioned in such connected components is of practical importance; components often correspond to ghettos or other natural clusterings of individuals. Let the connected component on the left, comprising eight households, be denoted Component 1, and the component on the right, with three households, Component 2.

We envision segregation as the degree of connectivity of the race's social network. The potential effects of segregation arise because Blacks tend to interact with Blacks, and Whites with Whites. The idea that segregation is synonymous with same-race interactions has—once a network of social interactions is constructed—a formal expression in network connectivity.

The SSI is one measure of network connectivity. It arises as the unique measure that satisfies certain properties, the most important of which is a requirement that an individual be more segregated the more segregated are his direct neighbors. Concretely, that an individual's segregation is the weighted sum of her neighbors' segregation, weighted by how much she interacts with each one of them. We discuss the properties in detail in the next section.

insert table I

The SSI for blacks in City 1 is in Table I. Note that Component 1 is more segregated than Component 2, which reflects that the network in Component 1 is more connected than that in Component 2. The SSI also lets us disaggregate the component-wide SSI into individual household SSI: the component-wide SSI is the average of the individual SSI. Note that $(C, 1)$ is the most segregated household in this example, which captures that this is an individual who only interacts with blacks. On the other hand, $(D, 4)$ is the most integrated household in Component 1.

Individual SSI should be interpreted as the distribution of component-wide SSI within a network. So a particular individual's SSI is relative to the SSI of the component she is in. Note how $(D, 4)$'s share in Component 1's segregation is small, while the distribution of segregation in Component 2 is quite even. So $(C, 4)$s SSI is smaller than $(C, 5)$s. The component's SSI is the average of the individual SSIs; hence, an individual's SSI may be much larger than the SSI of her connected component.

Finally, we remark that the SSI is invariant to the size of the population of blacks. If we double the size of City 1 by adjoining a copy of the city to itself, SSI will not change. We would have two new components and their respective SSIs, and the city SSI would be the weighted average of the four components.

IV. Measuring Segregation Based on Social Interactions

IV.A. The Social Interactions Framework

The basic building blocks for our measure of segregation is a set of individuals V and information on whether (and, possibly, how much) any two individuals interact. Hence, the measure depends on the network of social interactions among the individuals in V. Our measure identifies segregation of the members of a group with the intensity of the social interactions among the members of that group.

Given any two individuals, suppose we know whether they interact with each other and

the intensity of their interaction. For any two individuals v and v' in V, let the number $r_{vv'} \geq 0$ represent the nature of their relationship. If $r_{vv'} = 0$, then there is no relation between v and v'; if $r_{vv'} > 0$ then v and v' have a relationship. Abusing notation, we use V to refer to the number of elements in the set V. The information on interactions is then summarized in a $V \times V$ matrix R, with typical element $r_{vv'}$.

We make two important assumptions about the numbers $r_{vv'}$ in R. First, we assume that individuals face a budget constraint for their social interactions:

$$\sum_{v' \in V} r_{vv'} = 1$$

for all v in V. Think of $r_{vv'}$ as the fraction of time that v spends with v'. Second, we assume that if $r_{vv'} = 0$, then $r_{v'v} = 0$, though we allow $r_{vv'}$ and $r_{v'v}$ to be different when they are not zero. We allow for $r_{vv'} \neq r_{v'v}$ because a relationship can have a different level of importance or intensity to v and to v'. In fact, this comes up in empirical applications of SSI: v may interact only with v', in which case $r_{vv'} = 1$, while v' may split his time equally among n other relationships, so $r_{v'v} = 1/n$.

Now, suppose that we know the race of each individual $v \in V$. For the rest of the section, fix one race, called race h, and drop from the set V all individuals from races other than h. Form the matrix B from the matrix R by retaining only those $r_{vv'}$ for which both v and v' belong to race h. The matrix B (a submatrix of R) reflects the network of same-race social interactions among the members of race h.

Let us briefly discuss two examples, which preview our empirical applications in Section 6. First, suppose we construct B using information on residential patterns (and only information on residential patters). We would need to set a criterion for who is a neighbor of whom, and set $r_{v,v'} = 0$ when v and v' are not neighbors. The criterion could be that v and v' are neighbors if they live sufficiently close to each other. We can then suppose, in the absence of additional information on social interactions, that the relation with each of his neighbors is equally important to v, and set $r_{vv'}$ to be the inverse of the number of v's neighbors. Finally, we keep only those agents that belong to the race under analysis (race h). Second, suppose

we construct B from a survey on social interactions where individuals are asked to name their 10 closest friends. We would then set $r_{vv'} = 0$ if v and v' do not name each other as friends, and set $r_{vv'}$ to be the inverse of the number of v's friends, supposing the survey does not let us infer the relative importance of each friendship. The two examples are developed in detail empirically in Section 6.

It is important to note that, while we focus on the network of same-race interactions, the intensity of those interactions is affected by cross-race connections through $r_{vv'}$. For example, let v be a member of race h. If v interacts only with v', and v' is in race h, then $r_{vv'} = 1$ and 1 will be the only non-zero element of v's row of $r_{vv'}$s in B. On the other hand, if v interacts with 9 members of another race, besides v', then $r_{vv'} = 1/10$ and $1/10$ will be the only non-zero element of v's row of $r_{vv'}$s in B. This difference implies that v is more integrated when he has relations with individuals of other races. We discuss this feature of our measure in Section 5.C.

A *segregation index* for race h is a function that assigns a real number $S^h(B)$ to each matrix B of same-race interactions, along with functions assigning a real number $s_v^h(B)$ for each individual member v of race h, such that $S^h(B)$ is the average of the individual $s_v^h(B)$.

Our definition of a segregation index reflects our desire that segregation be measured at the individual level. Individual segregation is measured in the same units as racial segregation; race-h segregation is the average of the segregation of all individuals of race h.

IV.B. Three Properties Which Define The Spectral Segregation Index

We present three properties that jointly define our measure of segregation.

The first property requires that an increase in the intensity of same-race interactions imply an increase in segregation. Concretely, say that a matrix B' has *more intense interactions* than matrix B if all the entries of the matrix B' are at least as large as those of B. Then, if $B = (r_{vv'})$ and $B = (r'_{vv'})$ we have $r_{vv'} \leq r'_{vv'}$ for all v and v'. A segregation index satisfies the property of *monotonicity* if, whenever B' has more intense interactions than B, $S^h(B) \leq S^h(B')$.

The second property is a normalization of the index. Let $d > 0$ be a real number. A matrix B is *homogeneous of degree* d if, for all v in race h, $\sum_{v'} r_{vv'} = d$. An example of a homogeneous of degree 3/4 matrix is

$$\begin{pmatrix} 0 & 1/4 & 1/2 \\ 1/4 & 0 & 1/2 \\ 1/2 & 1/4 & 0 \end{pmatrix}$$

A segregation index is *homogeneous* if, whenever B is homogeneous of degree d, $S^h(B) = d$.

Homogeneous networks rarely occur in practice, but the property gives an interpretation to the segregation of networks one encounters in applications. For example, a measure of 0.8 can be read as the segregation race-h individuals would have if they spent 80 percent of their time with individuals of the same race. Homogeneity also provides a "scale free" property: If City A has more households than City B, but each household in both cities has the same fraction of same-race neighbors, the index will report the same level of segregation for both cities.

Our third property is the most substantial and potentially controversial. We want the segregation of an individual i to depend on the segregation of the individuals with whom she interacts. We require that this dependence takes a linear form. We need some auxiliary concepts to present the third property.

Let N_v be the set of individuals of race h that v interacts with: the set of v' in race h with $r_{vv'} > 0$. In a similar vein, consider the set of individuals who interacts with the members of N_v, and those that interact with those that interact with the members of N_v, and so on. The resulting set of individuals, with direct or indirect interactions with v, is called the *connected component* of B that v belongs to; denote this set of individuals by C_v.

The third property requires that $s^h_v(B)$ be the average of $s^h_{v'}(B)$ among v's race-h social interactions, relative to the average segregation of the individuals in v's connected component. If S^{C_v} is the average segregation of individuals in C_v, say that a segregation index

satisfies *linearity* if

$$s_v^h(B) = \frac{1}{S^{C_v}} \sum_{v' \in N_v} r_{vv'} s_{v'}^h(B).$$

There are two qualitative assumptions behind the linearity property. The first is that v's segregation depends on his neighbors' segregation. As described in the Introduction, if one considers Figure I, which depicts the distribution of blacks across metropolitan Detroit, it seems evident that individuals in the center of the city's black ghetto should be measured as more segregated than those closer to the edge. Linearity is one embodiment of this requirement. In subsection V.D we discuss the implications of relaxing this assumption. Note that, while the weights $r_{v,v'}$ must add to one, an individual's SSI is not bounded by 1.

The second qualitative property is that the dependence is modulated by the connected component's segregation. That is, a decrease in the segregation of one of v's neighbors will affect v less if v lives in a highly segregated component. The key idea is that v receives the effects of segregation from her different neighbors, and any one neighbor is less important when the component is highly segregated.

It is not possible to relax linearity, while retaining the linear influence of neighbors' segregation. Suppose that v's segregation depends directly on her neighbor's segregation, but that it does not take the form assumed in the linearity property. Suppose that the component's segregation does not play a role, and that v's segregation depends directly on the sum of neighbor's segregation. Then, an increase in a neighbors' segregation gives a one-for-one increase in v's segregation, and this in turn directly impacts v's neighbor. The result does not necessarily (in fact, generally will not) converge to new levels of segregation. Our use of the components' segregation guarantees that the effect of an increase in segregation for a neighbor does not impact fully on v, at least not for large values of segregation, ensuring that there is a solution to the problem of determining all individuals' segregation measures.[9]

9. The SSI is the weighted average of the SSI by connected component (S^{C_v}), weighting each component by how many individuals it has. One may be interested in identifying highly segregated components, even where the overall population is not highly segregated. In residential segregation, components can be interpreted as ghettos, and in school segregation as same-race cliques.

The three properties described above jointly define our index. The *spectral segregation index* (SSI) is the (unique) segregation index that satisfies the properties of monotonicity, homogeneity, and linearity (Theorem 1, Appendix A).

On a connected component, SSI is the largest eigenvalue of the corresponding irreducible submatrix of B. The individual SSI are obtained by distributing the component's SSI among individuals using the eigenvector corresponding to the largest eigenvalue. Thus, SSI results from familiar matrix operations and is easy to compute using standard software, such as MATLAB. The irreducible submatrices of B are often very *sparse*, meaning that many of its entries are zeroes. There are efficient algorithms for computing the largest eigenvalues of sparse matrices, and MATLAB comes with one such algorithm incorporated in its `eigs` command.

V. Analysis of the Spectral Segregation Index

The previous section described three properties which provide the precise assumptions underlying the SSI. In this section, we provide further properties and features of SSI, illuminate an alternative interpretation for the index, discuss other ways to incorporate cross-race interactions, and describe the implications of relaxing the linearity property.

V.A. An Alternative Interpretation of SSI.

An alternative way to interpret the SSI is through a model of group-specific capital transmission. SSI is a measure of how fast same-group influences are disseminated purely as a result of social contacts.[10]

Suppose that the matrix of same group social interactions, B, has only one connected component (without this assumption, the result will hold in each connected component of B.) Let x_v be a measure of how much group-specific capital an individual v has. We think of this capital as the depth of one's group identity; something that arises from repeated social interaction with people of one's own group. There is an inherent difference between

10. We thank Erzo Luttmer for suggesting this interpretation.

visiting a church once to listen to their gospel choir and interacting constantly with people who are involved with gospel music. The intensity with which one experiences the same social phenomenon is the key to this difference. Segregation is related to this intensity, and one can show how SSI captures the intensity of same-group social phenomena.

Suppose that, in each period t, individual i's h-capital grows depending on how much h-specific capital her contacts have, and on how much v interacts with them. Specifically, suppose that

(2)
$$x_{vt} = x_{vt-1} + \sum_{v' \in B} r_{vv'} x_{v't-1},$$

and that x_{v0} is given, for all v.

The law of motion in (2) is our assumption that capital reflects the intensity of v's own-race identity. Similar models have been used to capture cultural transmission in networks; see Brueckner and Smirnov [2004].[11]

PROPOSITION 1. For all vectors $(x_{v'0})_{v'}$ of initial stocks of capital, and all v,

$$\lim_{t \to \infty} \frac{x_{vt}}{x_{vt-1}} = 1 + S^h(B).$$

Proposition 1 shows that we can interpret SSI as the rate of growth of group-specific influences. It follows from a familiar calculation in Perron-Froebenius theory; recall that SSI is the largest eigenvalue of B in the case where we have only one connected component. In economics the result is reminiscent of the balanced growth result in the theory of Leontief systems (see, e.g., Samuelson and Solow [1953]).

Examples of this type of group-specific capital transmission may include language (Lazear [1999]) and the choice of first names [Fryer and Levitt 2004]. In a simple model of culture and language, Lazear [1999] shows that incentives to assimilate by learning to speak the native language are decreasing in the size of an ethnic enclave. Fryer and Levitt [2004] argue that

11. The model in Brueckner and Smirnov [2004] is slightly different, as they allow x_{vt} to be a weighted average of x_{vt-1} and $x_{v't-1}$. The statement in Proposition 1 holds for their model with $\theta + S^h(B)$ instead of $1 + S^h(B)$, where θ is the inverse of the number of neighbors each agent has.

the choice of distinctive first names is a cultural investment, and show that this practice is more common in highly segregated areas. Both of these papers are consistent with the basic model of group-specific capital transmission described above and, *ipso facto*, our measure of segregation.

V.B. General Properties

We discuss here some important and more subtle properties of SSI.

First, SSI identifies isolated individuals by marking them as perfectly integrated. If v has no connections ($r_{vv'} = 0$) to individuals of his group, then $s_v^h(B) = 0$. If v has relations with at least one individual of his same group, $s_v^h(B) > 0$ (Proposition 3, Appendix 1). Perfectly-integrated groups are rare, but we do observe perfectly integrated individuals in our applications. These are individuals who only interact with others of different races. SSI singles them out by assigning them a measure of zero.

Second, small changes in the structure of social interactions will entail small changes in SSI. SSI is a continuous function of the elements of B (Proposition 5, Appendix 1).

Third, SSI is related to a calculation of connections between individuals. If v has a relation to v', and v' has one to v'', then information can travel from v to v'' by the path $v - v' - v''$. It is intuitive to think of the number of such paths as a measure of how connected v is to v''. Segregation, on the other hand, is the extent to which individuals of the same group are connected, so counting paths between individuals gives rise to a natural measure of segregation. It turns out that SSI has a close connection to the number of paths that exist between individuals. Counting paths gives another interpretation of SSI.

We flesh out this connection in Appendix A. Here we give some simple calculations suggesting the nature of the relationship between counting paths between individuals within the same group and SSI.

Consider the following special case: each non-zero $r_{vv'}$ takes the same value, so $r_{vv'}$ is either 0 or $r \in (0, 1)$. Let N_v^k be the set of individuals for which there is a path to v with at

most k individuals. Then,

$$s_v^h(B) = \sum_{v' \in N_v^k} \alpha_{vv'} s_{v'}^h(B),$$

where $\alpha_{vv'}$ is proportional to the number of paths between v and v'. Note how all the v' in the same component as v affect v's segregation. The weight of each v' is affected by the number of paths between v and v'. Concretely, $\alpha_{vv'}$ is obtained as the number of paths of length k (with k individuals) from v to v' multipled by $r^k/(S^h(B))^k$. The number of paths from v to v', in turn, is the vv' entry of the matrix $\frac{1}{r^k}B^k$.

Fourth, and related to the previous property, SSI captures certain multiplier effects in the social interactions network. An individual's susceptibility to own-group influences (patterns of speech, names, and other group-specific behavior) depends on how many contacts the individual has with his or her own-group and the susceptibility of those contacts.

insert figure IV

Consider the following thought experiment, depicted in Figure IV. We show the effect of changing the race of one individual in a network, the resulting changes in SSI capture the essence of the multiplier effects. Network A has 3 black individuals who are connected to each other, and all of which are also connected to one white individual. To illustrate the multiplier effects captured in SSI, Network B changes the race of Individual 4 so she is also black now. To keep the calculations transparent, we assume that 4 also has three neighbors in total. Table II shows the levels of segregation before and after Individual 4 changes race.

insert table II

V.C. More on Cross-race Interactions

We argued that SSI captures cross-race interactions by their effect on the intensity of same-race interactions. We expand on this point here using a simple example, and then discuss alternative ways of incorporating cross-race interactions.

We have argued that, if v interacts only with v', and v' is in race h, then v would be more segregated than if she interacts with 9 other individuals who are not in race h. We make the same point here with a concrete example. Consider Figure V. The blacks in the city on the left have a SSI of 0.83. If we add white neighbors, to obtain the city on the right, the blacks have a much lower SSI of 0.5. The change is purely the result of the lower intensity of same-race interactions due to a decrease in $r_{vv'}$s. Note that the SSI for the city on the right follows immediately because all black agents spend exactly 1/2 their time with other blacks.

insert figure V

An alternative way to incorporate cross-race interactions would be to explicitly let the segregation of individual v depend on the segregation of the neighbors that are not the same race as her. There are two potential problems with this. First, we would need to decide whether a more segregated white neighbor makes a black agent more or less segregated. There are simple arguments for both effects: a black agent may be expected to interact less with a highly segregated white, and thus be more isolated from whites, or she may get more white specific capital from a segregated white, and become less isolated from whites. Our approach is agnostic with respect to the effect of one race's segregation on another, and allows for the possibility of deciding the matter empirically.

The second objection is practical. The computational complexity of calculating SSI depends critically on the dimensions of the matrices B. If we need to allow explicitly for the interactions that each v has with all her neighbors, we would tend to get much more connected networks, and thus much larger matrices B. As a result, the already slow task of calculating SSI would become extremely time consuming and likely infeasible in many applications.

V.D. Relaxing Linearity

Without assuming linearity, we would be unable to derive a unique numerical index. If, for example, the linearity assumption is replaced with a monotonicity condition – higher

segregation among i's same-race neighbors imply higher $s_i^h(\beta)$ – one cannot pin down a specific numerical index. The situation is analogous to that of income distribution measures, where general properties lead to orderings of Lorenz curves, that do not allow one to compare any two distributions. In our framework a Lorenz-curve-type ordering is readily obtained: group h is more segregated in β than in β' if the distribution of $(\sum_j r'_{ij})$ dominates that of $(\sum_j r_{ij})$. Something similar arises in the measurement of income distribution. Atkinson [1970] presents a partial order on income distributions, in which two distributions may not be comparable in terms of income inequality. When Lorenz curves cross, one has to decide how much weight to assign to each side of the intersection. Rather than choose adhoc weights which could differ for each application (which, some have argued, is the main reason researchers do not use the Atkinson index as a measure of segregation, Massey and Denton [1988]), we get implicit weights through the Linearity property.

VI. Two Applications of SSI: Measuring School and Residential Segregation

Here we develop two illustrative applications of SSI: estimating racial segregation of friendship networks in schools and residential segregation.[12]

VI.A. School Segregation

There is an impressive literature on the effects of segregation across schools on achievement. Jonathan Guryan [2004] estimates that half of the decline in black dropout rates between 1970 and 1980 is attributable to desegregation plans. Robert Crain and Jack Strauss [1985] find that students randomly offered the chance to be bussed to a suburban school were more likely to work in professional jobs nearly 20 years after the experiment. Christopher Jencks [1972] estimates that desegregation raises black achievement by 2-3 percent. Based on a meta-analysis of ninety-three studies, Robert Crain and Rita Mahard [1981] conclude

12. Fryer and Torelli [2005] provide another natural application of SSI: measuring social popularity in schools.

that desegregation has a significant effect on black achievement, especially younger children, though other meta-analyses are less conclusive [St. John 1975].

Yet, in the spirit of Martin Luther King, who dreamed that one day "little black boys and black girls will be able to join hands with little white boys and white girls and walk together as sisters and brothers," some argue that society should strive for integration *within* schools not just *across* them (Lucas [1999], Mickelson [2001]. Within school segregation, commonly referred to as "second-generation segregation," is thought to be as important as segregation across schools in inhibiting the educational opportunities of racial and ethnic minorities (Mickelson [2001]). Previous studies use traditional measures of segregation (such as exposure and dissimilarity) to measure segregation across schools. These measures do not disaggregate to the individual level and cannot use information on students' actual social contacts – limiting our ability to understand the relationship between within-school segregation and outcomes.

We first describe the data used to estimate SSI; we then present the analysis.

The National Longitudinal Study of Adolescent Health (Addhealth) database is a nationally representative sample of 90,118 students entering grades 7 through 12 in the 1994-1995 school year. A stratified random sample of 20,745 students was given an additional in-home interview; 17,700 parents of these children were also interviewed. Thus far, information has been collected on these students at 3 separate points in time: 1995, 1996, and 2002. There are 175 schools from 80 communities included in the sample, with an average of more than 490 students per school, allowing within school analysis. Students who are missing data on race, grade level, or friendships are dropped from the sample.

A wide range of data are gathered on the students, as described in detail on the Addhealth website (http://www.cpc.unc.edu/projects/addhealth). Our primary outcome variables are divided between measures of academic achievement and those that are more associated with social behaviors. The social variables include smoking, skipping school (without a valid excuse), interracial dating, and whether or not a student is happy at their school. Smoking and skipping school are answers to the question, "During the past 12 months, how often

did you..." Answer choices range from never to nearly everyday. Interracial dating is a dichotomous variable equal to 1 if the student reports ever dating interracially and zero otherwise. Happiness measures whether or not students report being happy at their school. The academic variables include: Peabody Vocabulary Test (PVT) scores, whether or not a student plans to attend college, grades in the previous grading period, and a measure of how much effort the student exerts. All responses (including grades) are self-reported. For each student, grades were calculated by aggregating grades in 4 subjects: math, history, science, and English.

To measure school segregation, we make use of the information on friendship networks within schools available in the Addhealth. All students contained in the in-school survey were asked, "List your closest male/female friends. List your best male/female friend first, then your next best friend, and so on." Students were allowed to list as many as 5 friends from each sex. Each friend can be linked in the data and the full range of covariates in the in-school survey (race, gender, grade point average, etc) can be gleaned from each friend. Friendship links are defined as unions: student A is considered to be "friends" with student B if A lists B as a friend, B lists A as a friend, or both.

The school-level spectral segregation index is calculated by taking, for each racial group, the average SSI of each connected component (CC) in the school that consists of students from that group, weighted by the size of those connected components. In other words, to calculate the black group SSI for school 1, assuming there are two black connected components in that school 1, we find: [(SSI of CC1)(size of CC1) + (SSI of CC2)(size of CC2)]/[size of CC1 + size of CC2]. Students who are singletons (who do not have any friends from their racial group) are considered to be connected components of size 1 with SSI equal to 0 – completely integrated.

In order to make individual SSI comparable across connected components each individual SSI is multiplied by the size of the connected component of which it is a part.

insert figure VI

Figure VI depicts the relationship between the percentage of a racial group in a school and the level of segregation for that racial group in that school, using the Addhealth database. Each observation is a school. Grade levels 7-12 are combined. School level segregation ranges from .014 to .848 across the 175 schools in AddHealth. The mean level of segregation is .618; the standard deviation is .146.

Many researchers assume the relationship between the segregation of a racial group within a school and the percentage of that group in the school is linear (see, for example, Orfield [1983]). This approximation is a good first pass for Whites (though we find nearly all White data points above the 45° line), but less true for Hispanics and Asians. For Blacks, the relationship between percent own-race in a school and own-race segregation is even more non-linear. As the percentage of black students increases from zero to twenty-five percent, black segregation rises sharply. Above twenty-five percent, Blacks are near complete segregation.

It is important to emphasize that our data do not allow one to disentangle why these patterns exist. The segregation observed in Figure VI could be a result of own-race preferences for social interactions or the response to external discrimination or racism. Understanding the causal model underlying these observations is of great importance to our understanding of social interactions, bussing programs, and the optimal organization of schools, among other things.

insert table III

Table III presents estimates of the relationship between individual-level measures of segregation and individual outcomes. Individual level segregation ranges from 0 to 174.973 with a mean of .618 and standard deviation of 2.48.

We estimate models of the form:

$$\begin{aligned} outcome_{i,j} &= \alpha_j + X_i\beta + \gamma segregation_i + \xi_1 black \cdot segregation_i \\ &\quad + \xi_2 asian \cdot segregation_i + \xi_3 hispanic \cdot segregation_i + \varepsilon_{i,j}, \end{aligned} \tag{3}$$

where i indexes individuals, j indexes schools, X_i represents a set of individual level controls, and α_j denotes school fixed-effects. The coefficient γ measures the relationship between the segregation of individual i and a given outcome for i. We concentrate on ξ_i, which measures the differential effect of individual segregation for group i relative to whites, and $\gamma+\xi_i$ which captures the overall relationship between segregation and outcomes for group i.

For Blacks, individuals who are more segregated are less likely to smoke (a behavior predominant among white teens) and have lower test scores. Segregated Asians are less likely to skip school, more likely to have high test scores, put in more effort, and report being happier. Segregated Hispanics are less likely to smoke, more likely to have low test scores, low grades, and low probability of attending college. Not surprisingly, students of all races are less likely to date interracially when schools are more segregated. Similar results are obtained when one excludes school fixed-effects.

VI.B. Residential Segregation

The ideal data to estimate residential segregation would contain information on the nature of each household's interactions with other households. In lieu of this, we proceed like we did for the imaginary city of the example in Section III.: we use geographical distance to infer social interactions. In addition, since we lack individual-level data we work with block-level data from the 2000 US Census. We restrict our sample to the 313 Metropolitan Statistical Areas (MSAs). The data are available from Geolytics Inc. (see http://www.geolytics.com/).

Census blocks contain, on average, 300 households, and are approximately 100 meters in radius. We identify a block with the race/ethnicity of the majority of its inhabitants. This assumption is not too problematic, as blocks are strikingly homogeneous: 94.3 percent of Iowans live in a homogeneous census block and so do 77 percent of Texans. Save Washington DC, more than 60 percent of the blocks in all states contain households of only one race (for half the states, 80 percent or more of the blocks contain only one race).

We assume that two blocks are neighbors if they are within one kilometer of each other.[13]

13. We have used one kilometer radii because one kilometer is the median radius of a census tract

From this, we know when r_{ij} should be non-zero. The next step is to calculate the intensities of social interactions; the values of r_{ij}. We obtain the total number, d_i, of neighbors of block i, i.e. the number of blocks that are within one kilometer of i, independent of race. Absent further information on the structure of social interactions in neighborhoods and consistent with the budget constraint described in Section 4, let $r_{ij} = 1/d_i$. With the resulting matrix B, we are in a position to calculate SSI using the characterization we present in the appendix.[14]

An important caveat to our application of SSI to residential segregation is that it ignores block density.[15] To correct for this, one could assign all individuals in a census block to the centroid of that block, and run the resulting individual-level estimation. This method, however, is computationally very costly.

We first discuss a baseline for comparing residential segregation measures; we then present our results.

Since SSI for race h is a measure of the connectivity of the race-h network it will tend to be larger in cities with larger fractions of race-h individuals, even if individuals located at random in the city.

We refer to the SSI one would expect to see in a city when individuals locate at random as *Baseline SSI*. We provide estimates of both SSI, and of the SSI in excess of Baseline SSI.

We have obtained measures of Baseline SSI by simulating random assignment of races to large regular (in a graph-theoretic sense) cities with the corresponding fraction of race-h inhabitants. Concretely, for each fraction $p = 0.01, 0.02, \ldots 0.99$ we simulated $1,000$ cities of 100 households each, where each household is of race h with probability p.[16]

(1.03), and tracts are the traditional notion of a neighborhood in the literature. Our results alter little when we change criterion to 0.5 or 1.5 kilometers.

14. We need to calculate the largest eigenvalue of (each connected component of) B. The Matlab programs to calculate all indices reported in the paper are available at
`http://post.economics.harvard.edu/faculty/fryer/fryer.html`

15. This likely induces little error in the estimates of segregation, given our definition of neighbor usually encompasses several blocks. In areas such as New York, however, this limitation may be quite restrictive.

16. For a few values of p we ran simulations of much larger cities, with $2,500$ nodes, and we obtain the same results. For the simulation of the full range of p we chose size 100 because the larger simulations are very time intensive. All simulations were done in Matlab; the code is available from the authors.

insert figure VII

Figure VII shows the results of our simulations. On the horizontal axis is the fraction of race-h inhabitants, while the vertical axis shows the average SSI. When the share of race-h inhabitants in a city is relatively small, SSI mirrors the percent race-h in a city closely. This is to be expected. When race-h inhabitants are relatively few and assigned to a city at random, linearity has little power to alter SSI from percent black. As the fraction of race-h individuals increases, however, SSI significantly departs from the percentage of race-h in a city. We have used only large cities, as we can prove (See Appendix B) that baseline SSI converges as a city grows. In fact the simulations show the convergence to be quite fast.

Detroit is the most segregated city for Blacks; Lowell, MA for whites; McAllen, TX for Hispanics and Honolulu, HI for Asians.[17] The list seems quite intuitive. It also confirms that SSI is correlated with the size of a minority group. The latter point begs for a distinction between SSI and "adjusted" SSI: the segregation in excess of baseline SSI. It is unclear which is most closely related to economic outcomes. Adjusted SSI tells us more about preferences, while the original SSI is a better measure of the pure connectedness in a network. The most segregated cities using adjusted SSI for Asians, Blacks, Hispanics, and Whites are: Los Angeles, CA; Milwaukee, WI; Flagstaff, AZ; and Pine Bluff, AR, respectively. Approximately 11 percent of households in Milwaukee are black, implying an expected SSI of .1145 if blocks were allocated at random. The actual measure of segregation is a factor of 9 larger. To generate the level of segregation in Milwaukee, assuming blocks were assigned a race at random, Blacks need to comprise 80 percent of the population.

We have emphasized how the SSI allows one to consider more disaggregated units than the city. One of the most interesting units is the agglomeration of same-race blocks: racially homogenous ghettos, which SSI identifies endogenously as connected components (see Section 4). This is related to city-wide SSI, but SSI weights the ghetto's SSI against members of the same race in other parts of the city, who are more integrated. For Blacks and Whites,

17. For a complete list of the most and least segregated cities, see `http://post.economics.harvard.edu/faculty/fryer/fryer.html`.

the largest ghetto is Detroit – implying an enormous amount of city-wide segregation. Remarkably, 87 percent of black blocks in Detroit comprise one large ghetto. The largest connected component is San Francisco for Asians, and Los Angeles for Hispanics. Hispanics in Los Angeles comprise the largest minority ghetto in America; 17,909 Hispanic blocks are connected.

Along with the variation across cities in SSI, there are several MSA level characteristics which are associated with higher levels of racial segregation. For instance, cities which exhibit higher segregation for blacks tend to be larger cities, have a high percentage of female-headed households, and are less likely to be in the West.

insert table IV

Table IV presents a correlation matrix of popular measures of segregation. These measures include dissimilarity, isolation, Gini coefficient, exposure, entropy, and interaction. Also included in the matrix are SSI, SSI minus the baseline, and the ranking of cities based solely on the their fraction of Blacks. All measures were calculated using data at the census block level for 326 MSAs. The Spectral index has surprisingly little correlation with dissimilarity, gini, entropy, and interaction – averaging less than .5 – and high correlation with isolation and exposure; averaging more than .90. Given the nature of the isolation and exposure indexes, it is not surprising that SSI is more correlated with the measures relative to the others. As a measure of residential segregation, our measure is very similar to existing measures of exposure with the added ability to disaggregate to the level of individuals, and a well-understood theoretical foundation. Adjusted SSI becomes even less correlated with dissimilarity and isolation. The fraction black in a city is highly correlated with SSI, but the linearity property assures that this correlation is less than perfect.

We end with a discussion of the relationship between residential segregation and outcomes.

The economic literature on the effects of segregation on outcomes is impressive. Case and Katz [1991] show that youths in a central city are affected by the characteristics of their

neighbors. Almond, Chay, and Greenstone [2003] show that segregation of hospitals in the Jim Crow era had a significant negative effect on infant mortality. Using evidence from the Moving to Opportunity experiment, Katz, Kling, and Liebman [2001] and Kling, Liebman, and Katz [2005] provide evidence that moving individuals to lower poverty neighborhoods has substantial effects on mental and physical health of parents and children.

Cutler and Glaeser [1997] is one of the most influential papers in economics on the impact of segregation. They use the dissimilarity index as a measure of segregation. We re-estimate the impact of black segregation on economic outcomes with Cutler and Glaeser's specification. Econometrically, we estimate models of the form:

$$\begin{aligned} outcome_i &= X_i'\beta + \beta_1 segregation_j \\ &\quad + \beta_2 segregation_j * black_i + \varepsilon_i, \end{aligned} \tag{4}$$

where $outcome_i$ is measured at the individual level and $segregation_j$ is measured at the MSA level, and compare the results obtained with SSI and the dissimilarity index.

Identical to Cutler and Glaeser [1997], we correlate measures of segregation with various economic and social outcomes for young people aged 20-30. We choose to focus on younger individuals for three reasons. First, they are most susceptible to group level influences as a result of social interactions. Second, the problems of mobility across metropolitan areas is more easily avoided. Third, and most importantly, it mirrors the specifications in Cutler and Glaeser [1997]. For identical reasons, we drop individuals born in a foreign country. Data from the 1990 1 percent Census Public Micro Use Sample are used. Our sample contains 97, 976 individuals aged 20-24 and 139, 715 individuals between the ages of 25 and 30 residing in the 204 MSAs with at least 100,000 people and 10,000 blacks in 1990. This sample is identical to Cutler and Glaeser [1997].

Outcome measures are divided into 3 categories: educational attainment, labor market, and social outcomes. Educational attainment is measured as the probability an individual graduates from high school or college. There are two measures of labor market outcomes. The first is whether or not an individual is idle (not working and not employed). The second

is earnings (sum of wages, salary, and self-employment income). In all specifications, we use the natural logarithm of earnings, conditional on the individual not being in school and reporting positive earnings.[18] The final outcome variable is a social outcome – whether a woman is an unmarried mother.

insert table V

Tables V presents a series of ordinary least squares estimates of the relationship between segregation and outcomes for persons aged 20-24 and 25-30, using the dissimilarity index and the SSI – controlling for the standard set of individual and MSA-level covariates used by Cutler and Glaeser [1997]. Each measure of segregation has been normalized such that they have a mean of zero and a standard deviation of one.

The top panel of Table V replicates Cutler and Glaeser's [1997] results using the dissimilarity index. The bottom panel estimates the same specification using SSI. Results differ slightly between SSI and dissimilarity. On each outcome, cities with higher dissimilarity indices have inferior outcomes: less likely to graduate from high school or college, more likely to be unemployed and not in school, earn less money, and more likely to be a single mother. SSI paints a similar portrait, though the magnitudes are slightly weaker. No qualitative conclusions are unchanged. In all cases, the R-squared from regressions using the dissimilarity index and those using the Spectral index are remarkably similar.

VII. Conclusion

For decades, social scientists have used measures of evenness and exposure to estimate the prevalence and impact of segregation in housing, firms, and schools. These measures have many limitations, which we have discussed throughout. This paper develops a new measure of segregation based on two key ideas: a measure of segregation should disaggregate to the level of individuals, and an individual is more segregated the more segregated are the agents

18. Following Cutler and Glaeser (1997), we omit people in school from the earnings regression, since these individuals are expected to have low income.

with whom they interact. Developing three properties that any segregation measure should satisfy, our main result shows that one and only one segregation index satisfies our three properties and the two aims mentioned above—the Spectral Segregation Index. To illustrate the potential of the index, it is applied to two well-known social problems: measuring within-school and residential segregation and several new facts and insights are gleaned. We hope the Spectral index will be a useful tool for applied researchers interested in the agglomeration of individuals in networks.

Appendix 1: Technical Proofs

We present formally the results stated in Sections IV. and V..

Fix a race h. Let C_k, $k = 1, 2, \ldots K$, be the connected components of B. Abusing notation, let C_k also denote the submatrix of B with columns (and rows) indexed by the elements of C_k. Let λ_k be the largest eigenvalue of C_k, and x_k be its associated eigenvector, normalized so its entries add to one. [19]

The *Spectral Segregation Index (SSI)* is the index

$$B \mapsto \left(\hat{S}^h(B), (\hat{s}_i(B))_{i \in h}\right),$$

where $\hat{S}^h(B) = \sum_{i \in h} \frac{\hat{s}_i(B)}{V}$ and $\hat{s}_i(B) = \lambda_k x_{ki} |C_k|$.

Theorem 2. A segregation index satisfies Monotonicity, Homogeneity and Linearity if and only if it is the Spectral Segregation Index.

We note that the properties of Monotonicity, Homogeneity and Linearity are independent, in the sense that no pair of properties imply the third.

We state two additional properties of SSI. Proposition 3 was stated informally in Section IV.. Proposition 4 is informative about SSI, and used in the proofs below.

Proposition 3. If v has at least one same-race neighbor, $\hat{s}^h_v(B) > 0$. If v has no same-race neighbors, $\hat{s}^h_v(B) = 0$.

Proof. If $i \in h$ has at least one same-race neighbor, then i is in C_k, for some irreducible submatrix C_k. Let λ_k be the largest eigenvalue of C_k, and x_k be its associated eigenvector. By Lemma 6, x_k is strictly positive, so $x_{ki} > 0$. Since $\lambda_k > 0$ (Lemma 6), the definition of $\hat{s}^h_i(B)$ implies that $\hat{s}^h_i(B) > 0$. QED

19. Note that λ_k and x_k must exist by the Perron-Froebenius Theorem.

PROPOSITION 4. If C_k, $k = 1, \ldots, K$ are the connected components (the irreducible submatrices) of B, then

$$\hat{S}^h(B) = \sum_{k=1}^{K} \left(\frac{|C_k|}{V} \right) \hat{S}^{C_k},$$

and S^{C_k} is the largest eigenvalue of C_k. So $\hat{S}^h(B)$ is the weighted average of the components' largest eigenvalues.

Proof. We show that S^{C_k} is the largest eigenvalue of C_k. $S^{C_k} = \sum_{i \in C_k} s_i(B) / |C_k| = \lambda_k \sum_{i \in C_k} x_i$. Since x was normalized so that $\sum_{i \in C_k} x_i = 1$, it follows that $S^{C_k} = \lambda_k$. That $S^h(B)$ is the weighted average of the S^{C_k} follows immediately by definition of $S^h(B)$ and S^{C_k}. QED

PROPOSITION 5. $\hat{S}^h(B)$ is a continuous function of the entries of B

Proof. This is a direct consequence of Theorem 2 and the result in Appendix D of Horn and Johnson [1985]. QED

VII.A. Proof of Theorem 2

The proof of Theorem 2 proceeds by stating and proving 5 lemmas that together establish the theorem.

The first lemma unifies some standard results about irreducible matrices.

LEMMA 6. Let C be a real, non-negative, irreducible matrix. Then A has a real, positive, eigenvalue λ with associated eigenvector y. Such that

1. y is strictly positive, so $y_i > 0$ for all i, and y is the unique, up to a scalar multiple, strictly positive eigenvector of C;

2. λ is larger than $|\sigma|$, for any other eigenvalue σ of C; in particular, λ is larger than any other real eigenvalue.

Proof. By the Perron-Froebenius Theorem (Theorem 8.4.4 in Horn and Johnson [1985]), C has a real, strictly positive, eigenvalue, λ, with associated strictly positive eigenvector y. The multiplicity of λ is one and λ is larger than $|\sigma|$, for any other eigenvalue σ of C (λ is the spectral radius of C).

Let z be any strictly positive eigenvector, by Corollary 8.1.30 in Horn and Johnson, z is associated to eigenvalue λ. The z is a scalar multiple of y, as λ has multiplicity one. QED

Now we verify that the spectral segregation index satisfies our three axioms.

LEMMA 7. The Spectral Segregation Index satisfies Montonicity.

Proof. Let B' have more intense interactions than B. Let $C' = (c'_{ij})$ be an irreducible submatrix of B' Then the set of rows in C' is the union of the rows in some collection $C_1, C_2, \ldots, C_L$ of irreducible submatrices of B. Let $C = (c_{ij})$ be the block-diagonal matrix with $C_1, C_2, \ldots, C_L$ in its diagonal. Let x' be an eigenvector associated to the largest eigenvalue λ' of C'. Then $C'x' = \lambda' x'$, $x_i > 0$ for all i (Lemma 6), and B' having more intense interactions than B imply that

(5)
$$\lambda' = \frac{1}{x'_i} \sum_{j \in C'} c'_{ij} x'_j \geq \frac{1}{x'_i} \sum_{j \in C'} c_{ij} x'_j$$

Let $\lambda = \max\{|\sigma| : \sigma \text{ is an eigenvalue of } C\}$ be the spectral radius of C. Then, by Horn and Johnson's Theorem 8.1.26,

(6)
$$\lambda \leq \max_{i \in C} \frac{1}{x'_i} \sum_{j \in C} c_{ij} x'_j.$$

Statements (5) and (6) imply that $\lambda \leq \lambda'$. But λ' is $S^{C'}$ (Proposition 4); so $\lambda \leq \hat{S}^{C'}$.

Now we prove that $\hat{S}^{C_l} \leq \lambda$, for $l = 1, \ldots, L$. Let λ_l be the largest real eigenvalue of C_l. Let x_l be an eigenvector of C_l, associated to λ_l; Let $y = (y_i)_{i \in C}$ be the vector obtained from x_l by letting $y_i = x_{li}$ if $i \in C_l$ and 0 otherwise. Then, since C is block-diagonal, λ_l is an eigenvalue of C, with associated eigenvector y. By definition of λ, since λ_l is real, $\lambda_l \leq \lambda$. But Proposition 4 implies that $\lambda_l = \hat{S}^{C_l}$, so $\hat{S}^{C_l} \leq \lambda$, for $l = 1, \ldots, L$.

Let C'_k, $k = 1, \ldots, K$ be the irreducible submatrices of $B^{h\prime}$, and let each C'_k be the union of L_k irreducible submatrices of B^h, C'_{kl} with $l = 1, \ldots, L_k$. By Proposition 4

$$\begin{aligned}
\hat{S}^h(B) &= \sum_{k=1}^{K} \sum_{l=1}^{L_k} \frac{|C_k|}{V} \hat{S}^{C_{kl}} \\
&\leq \sum_{k=1}^{K} \hat{S}^{C'_k} \sum_{l=1}^{L_k} \frac{|C_k|}{V} \\
&= \sum_{k=1}^{K} \hat{S}^{C'_k}(B') \frac{|C_k|}{V(B')} = \hat{S}^h(B') \qquad \text{QED}
\end{aligned}$$

LEMMA 8. The Spectral Segregation Index satisfies homogeneity.

Proof. Let $a \in A$ be h-homogeneous of degree d. Let $y = \mathbf{1}$, then homogeneity says that $Ay = d\mathbf{1}$, so d is an eigenvalue with eigenvector y. By Lemma 6 d must coincide with λ, the largest eigenvalue of B, and the rescaled eigenvector must coincide with x. So $\hat{S}^h(B) = d$. QED

LEMMA 9. The Spectral Segregation Index satisfies linearity.

Proof. By Proposition 4, $\hat{S}^{C_k}$ is an eigenvalue with eigenvector (x_i), the eigenvector in the definition of the spectral index. The, for any i, $s_i(B) = S^{C_k} x_i \, |C_k| = |C_k| \, (C_k \cdot x|_i)$. So

$$\begin{aligned}
s_i(B) &= \sum_{j \in C_k} |C_k| \, r_{ij} x_j \\
&= \frac{1}{\lambda_k} \sum_{j \in C_k} |C_k| \, r_{ij} x_j \lambda_k \\
&= \frac{1}{S^{C_k}} \sum_{j \in N_i^a} s_j(B) \qquad \text{QED}
\end{aligned}$$

Second, we prove that any index that satisfies the three axioms must coincide with the spectral index. Let $\left(S^h(B), (s_i(B))_{i \in h}\right)$ be a segregation index that satisfies the three axioms.

LEMMA 10. If B has $b_{ij} = 0$ for all i and j, then $s_i(B) = \hat{s}_i(B)$ for all i.

Proof. By Homogeneity, $S^h(B) = 0$, so we must have and $s_i(B) = 0$ for all i, as $s_i(B) \geq 0$ and $S^h(B)$ is the average $s_i(B)$. Thus the index coincides with the Spectral Segregation Index. QED

LEMMA 11. For any B, $s_i(B) = \hat{s}_i(B)$ for all i.

Proof. If B is such that $b_{ij} = 0$ for all i and j, we are done by Lemma 10. Suppose that $b_{ij} > 0$ for at least one i and j.

Let $\gamma = \min \{b_{ij} : b_{ij} > 0\}$. Let $D = (d_{ij})$ be the matrix defined by $d_{ij} = 0$ if $b_{ij} = 0$, and

$$d_{ij} = \frac{\gamma}{|\{j : b_{ij} > 0\}|}$$

if $b_{ij} > 0$.

Note that $\sum_j d_{ij} = \gamma$ for all i, so D is homogeneous of degree γ. Then Homogeneity implies that $S^h(D) = \gamma$. Now, by definition of D, D has more intense interactions than B. So Monotonicity implies that $S^h(B) \geq S^h(D) = \gamma$. Hence, $S^h(B) > 0$.

Fix a component C_k such that $S^{C_k} > 0$; since $S^h(B) > 0$ there must exist at least one such component. For $i \in C_k$, let $x_i = \frac{s_i^h(B)}{|C_k| S^h(B)}$. Note that, by definition of $S^{C_k} x_i$, $\sum_{i \in C_k} x_i = 1$.

Then $S^{C_k} x_i = s_i(B)/|C_k| = \frac{1}{|C_k|} \sum_{j \in N_i^a} r_{ij} s_j / S^{C_k}$, by Linearity. Then $S^{C_k} x_i = \sum_{j \in N_i^a} r_{ij} x_j$. So $S^{C_k} x = C_k x$; S^{C_k} is an eigenvalue of C_k with eigenvector x.

Now, $s_i(B) > 0$ for all i. Since $s_i(B) = 0$ for some i would imply, by Linearity, that all $j \in N_i$ have $s_j(B) = 0$. Then, by recursion, $s_j(B) = 0$ for all $j \in C_k$, which would contradict that $S^{C_k} > 0$. Hence x is a strictly positive eigenvector.

By Proposition 4 and Lemma 6 now $S^{C_k} = \hat{S}^{C_k}$, and by the rescaling $\sum_{i \in C_k} x_i = 1$, x must coincide with the defining eigenvector in the definition of the spectral segregation index. Then, $s_i(B) = \hat{s}_i(B)$ for all i.

Finally, take a component with $S^{C_k} = 0$. Then Monotonicity and Lemma 10 imply that $b_{ij} = 0$ for all i and j in C_k. QED

Lemmas (7) through (11) establish the theorem.

VII.B. Results in Section V.

We first prove Proposition 1, we then state and prove additional results that were informally announced in Section V.. The results are formalizations of the discussion of network connectivity in Section V..

Proof of Proposition 1. Let I denote the $V \times V$ identity matrix. Let $D = I + B$. Then equation 2 implies that the vector $x_t = (x_{it})_i$ satisfies $x_t = Dx_{t-1}$, for all t. So $x_t = D^t x_0$. By Lemma 8.4.2 in Horn and Johnson [1985], $1 + \hat{S}^h(B)$ is the largest eigenvalue of D. By Lemma 8.2.7 in Horn and Johnson, there is a matrix L such that

$$\lim_{t\to\infty}(1 + \hat{S}^h(B))^{-t} D^t = L$$

Then,

$$\frac{x_{it}}{x_{it-1}} = (1 + \hat{S}^h(B))\frac{((1 + \hat{S}^h(B))^{-t} D^t x_0)_i}{((1 + \hat{S}^h(B))^{-t+1} D^{t-1} x_0)_i} \to (1 + \hat{S}^h(B))$$

We provide two results that help interpret the SSI. The first relates SSI to how many neighbors individuals have. The second result shows how SSI measures the connectivity of the h-race network. Both results hold in the neighborhood model, where r_{ij} is either 0 or $r > 0$.

Here we interpret B as graph, denoted G, for which the vertexes are the individuals and there is an edge (link) between two indexes i and j if $r_{ij} > 0$ The degree of a vertex i, $d(i)$, is the number of edges at i. Let $d_{\min} = \min\{d(v) \,|v \in V\}$ denote the minimum degree of G, $d_{\max} = \max\{d(v) \,|v \in V\}$ represents its maximum degree, and $\overline{d} = \frac{1}{|V|}\sum_{v\in V} d(v)$ the average degree of G. [20]

PROPOSITION 12. Let $d_{\min}$, $\overline{d}$ and $d_{\max}$ be the minimum, average, and maximum degrees

20. We use the most basic notions in Graph Theory. A reader can consult any graph-theory textbook, for example Diestel [1997]. Some of the ideas we use are from the field of Spectral Graph Theory; see e.g. Cvetković, D., Rowlinson, P., and Simić, S. [1997] for a comprehensive treatment.

of B^h, respectively. Then

$$d_{\min} \leq \overline{d} \leq \hat{S}^h \leq d_{\max}$$

Proof. See Cvetkovic and Rowlinson [1990]. QED

Let d_i be the number of same-race neighbors of household i. Proposition 12 proves that, Homogeneity notwithstanding, $\hat{S}^h(B)$ is larger than the average d_i over the individuals with $a(i) = h$.

Now we use walks in a graph to bring out the relation between SSI and network connectivity. A *walk* of length k is a sequence of (not necessarily different) vertexes v_1, v_2, ..., v_k, v_{k+1} such that for each $i = 1, 2, ..., k$ there is an edge from v_i to v_{i+1}. A walk is closed if $v_{k+1} = v_1$. Let W_i^θ be the number of walks of length θ that individual $i \in V$ can take in B, and define $W^\theta = \sum_i W_i^\theta$. Let W_{ij}^θ be the number of walks of length θ between individual $i \in V$ and $j \in V$. A graph is bi-partite if its vertex-set admits a partition into 2 classes such that every edge has its ends in different classes. The graphs one encounters in applications of SSI are never bi-partite.

Proposition 13. For θ sufficiently large: (1) $\dfrac{W_i^\theta}{(\hat{S}^h(B))^{\theta-1}}$ is approximately proportional to $\hat{s}_i^h(B)$, and the constant of proportionality is independent of i; (2) $\sqrt[\theta]{W^\theta/n^h}$ approximates $\hat{S}^h(B)$; and (3) if B is non-bipartite, W_{ij}^θ is approximately proportional to $(\hat{S}^h(B))^{\theta-2}\hat{s}_i^h(B)\hat{s}_j^h(B)$.

Proof. Let $U = (u_i)$ be the eigenvectors of B, normalized to form an orthonormal basis, so $U^T U = I$. Let D be the matrix with the eigenvalues of B on the diagonal, and 0 everywhere else. So $A = UDU^T$.

If 1 is the vector with 1 in all its entries, the vector of θ-long walks (W_i^θ) is defined by $(W_i^\theta) = A^\theta 1$. So $(W_i^\theta) = UD^\theta U^T 1$. The (u_i) vectors form a basis, so there are scalars (ξ_i) such that $1 = \sum_i \xi_i u_i$.

Then $(W_i^\theta) = \sum_i \xi_i U D^\theta U^T u_i$. But $U^T u_i = e_i$, the vector with i-th entry 1, and 0 elsewhere. So $(W_i^\theta) = \sum_i \xi_i \lambda_i^\theta U e_i = \sum_i \xi_{i\lambda i}^\theta u_i$. Let $\lambda_1 = S^h$; λ_1 has multiplicity 1, as B has

a unique non-trivial eigenvector (Theorem 2.1.3 in Cvetkovic, Rowlinson and Simic [1997]). So $S^h(\beta) > \lambda_i$, $i = 2, 3, \ldots, |h|$.

Then

$$\frac{1}{(S^h(B))^{\theta-1}}(W_i^\theta) = S^h(\beta)\sum_i \xi_i \frac{\lambda_i^\theta}{\lambda_1^\theta} u_i \tag{7}$$

$$\rightarrow S^h(B)\xi_1 u_i, \tag{8}$$

as $\lambda_i^\theta/\lambda_1^\theta \rightarrow 0$ for all $i \neq 1$. Since u_1 is a scalar multiple of the (x_i) vector in the definition of the spectral index, $S^h(B)\xi_1 u_1$ is a scalar multiple of s_i^h.

The second statement is a theorem of Cvetkovic, stated in the survey by Cvetkovic and Rowlinson [1990]. The third statement is essentially Theorem 2.2.5 in Cvetkovic, Rowlinson and Simic. QED

Proposition 13 (1) says that, as θ grows, $W_i^\theta(\hat{S}^h(B))^{\theta-1}$ converges. Thus $\hat{S}^h$ measures the growth in the number of walks that i can take. Further, it converges to something proportional to $\hat{s}_i$, thus individual SSI measures explain the differences, among individuals, in how many walks they can take relative to $\hat{S}$. Statement (2) in Proposition 13 says that $W^\theta \sim V\left(\hat{S}^h(\beta)\right)^\theta$. The total number of walks will grow at rate $\hat{S}^h(B)$ (a statement which is similar, and has a similar proof, to that of Proposition 1). Finally, (3) says that two individuals' measures are related to how many walks there are between the two individuals, relative to the total number of walks (given by $\hat{S}^h(B)$, in light of Statement (2)).

VII.C. Baseline Segregation

Here we present a theoretical justification for our "baseline" simulations. SSI converges as a city's size grows, so we can estimate SSI for relatively large cities (the size of 6400 is enough in our simulations).

Let $H = \{0, 1\}$ be the set of races. We are interested in only one race here, so working with $H = \{0, 1\}$ is without loss of generality. Let V_n be set of households, such that if $n \leq m$ then $V_n \subseteq V_m$.

Let $\Omega_n = H^{V_n}$ be the set of possible assignments of households to races. Abusing notation, let $\omega \in \Omega_n$ represent the resulting $V_n \times V_n$ matrix of social interactions. Endow the power set of Ω_n with the probability measure p_k obtained by letting each household be race 1 with probability $\pi \in (0,1)$, independently of the races of other households.

Let

$$E_n \hat{S}^h = \sum_{\omega \in \Omega_n} \hat{S}^h(\omega) p_n(\omega)$$

be the expected value of the SSI.

PROPOSITION 14. There is $\overline{S}$ such that $E_n \uparrow \overline{S}$ as $n \to \infty$.

Proof. We shall prove that, if $n \leq m$, then

$$\sum_{\omega \in \Omega_n} \hat{S}^h(\omega) p_n(\omega) \leq \sum_{\omega \in \Omega_m} \hat{S}^h(\omega) p_m(\omega).$$

Since the $E_n \hat{S}^h$ are bounded above by 1, the result follows.

Let $q_{n,m}$ be the probability distribution on $H^{V_m \setminus V_n}$ induced by letting each household be race 1 with probability $\pi \in (0,1)$, independently of the races of other households. Abusing notation, we shall use $q_{n,m}$ for the probability distribution induced by $q_{n,m}$ on $\left\{\omega \in \Omega_m : \omega|_{V_n} = \{0\}^{V_n}\right\}$. Then,

$$\begin{aligned}
\sum_{\omega \in \Omega_m} \hat{S}^h(\omega) p_m(\omega) &= \sum_{\omega' \in \Omega_n} p_n(\omega') \left[\sum_{\{\omega \in \Omega_m : \omega|_{V_n} = \omega'\}} q_{n,m}(\omega - \omega') \hat{S}^h(\omega) \right] \\
&\geq \sum_{\omega' \in \Omega_n} p_n(\omega') \left[\sum_{\{\omega \in \Omega_m : \omega|_{V_n} = \omega'\}} q_{n,m}(\omega - \omega') \hat{S}^h(\omega') \right] \\
&= \sum_{\omega' \in \Omega_n} p_n(\omega') \hat{S}^h(\omega') \sum_{\{\omega \in \Omega_m : \omega|_{V_n} = \omega'\}} q_{n,m}(\omega - \omega') \\
&= \sum_{\omega' \in \Omega_n} p_n(\omega') \hat{S}^h(\omega') \qquad \text{QED}
\end{aligned}$$

Appendix 2: A Brief Guide to Programs Calculating the Spectral Index

All programs to calculate the Spectral Index are in Matlab. There are three files which are used: callspec.m, neighbors.m, and blockspectral.m. We briefly describe each below. The version of the programs described is for geographic analysis of census blocks at the MSA level. Programs can be easily adapted for use in myriad applications.

callspec.m is the shell program that calls the other programs. It allows you to run the SSI algorithm on a list of cities. The list should be in a text file called list#.txt, where # is an identification string (does not necessarily need to be a number). For instance, you might want to create a list of five cities, and denote it list1.txt. The contents of list1.txt might be:

"001"

"002"

"003"

This list, when supplied as an input to callspec.m, would tell the program to calculate the SSI for cities whose identification numbers are 001, 002, 003, 100, and 369. Identification numbers should be in double quotes, and each should be on a new line. The file list1.txt should be placed in the same folder as callspec.m and the other m-files.

To run the program, simply type 'callspec' at the Matlab prompt. You will receive a prompt for list number. In this case, you would type '1' to call the above list.

Next you will receive a prompt to specify which race you wish to calculate SSI for. As the program stands, you can choose any of four races (or they could be non-race groups, depending on your application), or you can choose to calculate all four at once.

Finally, you are prompted to supply a neighbor radius, in kilometers. When constructing the neighbor matrix, neighbors will be considered anyone within this radius.

callspec.m will call blockspectral.m sequentially on each of the identification numbers in list#.txt, which in turn calls neighbors.m in order to construct the matrix. To construct this matrix, it must reference a set of files named msa_#.txt, where # stands in for the city identifiers. In the case of list1.txt, you would need files msa_001.txt, msa_002.txt,

msa_003.txt, msa_100.txt, and msa_369.txt. All files should again be in the same folder. These files should have the following structure: each line is a census block (or whatever your geographic unit of reference is) and four comma-separated columns. The first column is an identifier and should be in double quotes. The second is latitude. The third is longitude. The fourth is the group identifier for that block. For example, msa_369.txt might be:

"360150102006073",42.24114,-76.81282,1

"360150108003016",42.13062,-76.82308,1

"360150102003009",42.20382,-76.88979,2

This would correspond to city 369 having 8 census blocks, of which 5 are majority group 1, 2 are majority group 2, and 1 is majority group 4. neighbors.m uses this information to make the neighbor matrix needed to calculate the SSI.

The program generates two main types of output. Summary data appears in matrix called sipartial.mat. Information about individual blocks appears in output files called si_#.txt, where again # is the city identifier. The sipartial.mat matrix has 12 columns:

Column 1: city identifier

Column 2: group identifier

Column 3: SSI for group for city

Column 4: number of connected components for group

Column 5: number of singletons for group

Column 6: median connected component size for group

Column 7: largest connected component size for group

Column 8: smallest connected component size for group

Column 9: total number of blocks of group

Column 10: percent of blocks belonging to group

Column 11: average number of neighbors for group

Column 12: average number of same-group neighbors for group

As you can see, columns 1 and 2 identify the unique city/group combination; column 3 gives the SSI; and columns 4-12 give supporting statistics.

If you wish to find the SSI for each individual block you must look at the si_#.txt output files. These files have five columns each:

Column 1: city identifier

Column 2: connected component identifier

Column 3: block identifier

Column 4: SSI for block

Column 5: SSI for connected component

For example, to find the individual SSI for block 360150102006073 in city 369 you would look in the file si_369.txt for the row that has 360150102006073 in the third column. The individual SSI is the value in the fourth column.

If you wish to adapt these files for use in a non-geographic application, the main point of modification would be at line 38 of neighbors.m, which is the linking rule. If you wished to study the segregation of, for instance, a social network, this line of code (which currently calculates geographic distance and compares it with the "neighbor radius" solicited earlier) would be replaced by code that checks whether two people have a link in the social network. Other code would have to change too of course (for instance, latitude and longitude might be replaced by a list of friends' IDs), but the essential thing that determines the type of application is the linking rule.

References

Almond, Douglas, Kenneth Chay, and Michael Greenstone. "Civil Rights, the War on Poverty, and Black-White Convergence in Infant Mortality in Mississippi" mimeo Massachussets Institute of Technology, Department of Economics, 2003.

Atkinson, Anthony , "On the measurement of inequality," *Journal of Economic Theory*, II, (1970), 244-263.

Blau, Peter, "Inequality and Heterogeneity: A Primitive Theory of Social Structure." (New York, NY: Free Press, 1977)

Borjas, George, "Ethnicity, Neighborhoods, and Human-Capital Externalities," *American Economic Review*, LXXXV (1995), 365-390.

Brueckner, Jan and Oleg Smirnov, "Workings of the Melting Pot: Social Networks and the Evolution of Population Attributes." , mimeo, The University of Illinois at Urbana Champaign, 2004.

Case, Anne, and Lawrence Katz, "The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths." NBER Working Paper No. 3705, 1991.

Collins, Chiquita, and David R. Williams, "Segregation and Mortality: The Deadly Effects of Racism?" *Sociological Forum*, XIV (1999), 495-523.

Cowgill, Donald, and Mary Cowgill, "An Index of Segregation Based on Block Statistics." *American Sociological Review*, XVI (1951), 825-831.

Crain, Robert, and Jack Strauss, "School Desegregation and Black Occupational Attainments: Results from a Long-Term Experiment," Center for Social Organization of Schools, Johns Hopkins University, 1985.

Crain, Robert, and Rita Mahard, "Minority Achievement: Policy Implications of Research," in *Effective School Desegregation: Equity, Quality, and Feasibility*, Beverly Hills, CA: Willis D. Hawley Ed. Sage Publications, (1981) 55-84

Cutler, David, and Edward Glaeser, "Are Ghettos Good or Bad?" *Quarterly Journal of Economics*, CXII (1997), 827-872.

Cvetković, Dragos, and Peter Rowlinson, "The Largest Eigenvalue of a Graph: A Survey." *Linear and Multilinear Algebra*, XXVIII (1990), 3-33.

Cvetković, Dragos, Peter Rowlinson, and Slobodan Simić, *Eigenspaces of Graphs.*, (Cambridge, United Kingdom: Cambridge University Press, 1997).

Diestel, Reinhard, *Graph Theory.* (New York, NY: Springer Verlag, 1997).

Frankel, David and Oscar Volij, "Measuring Segregation." mimeo. Iowa State University, 2004.

Fryer, Roland and Steve Levitt, "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics*, CXIX (2004), 767-805.

Fryer, Roland and Paul Torelli. . "An Empirical Analysis of 'Acting White.' NBER Working Paper No. 11334, 2005.

Guryan, Jonathan, "Desegregation and Black Drop-Out Rates," *American Economic Review*, XCIV (2004), 919–944.

Horn, Roger A. and Charles R. Johnson, *Matrix Analysis*. (Cambridge, United Kingdom: Cambridge University Press, 1985).

Hutchens, Robert. . "Numerical Measures of Segregation: Desirable Properties and Their Implications," *Mathematical Social Sciences*, XLII (2001), 13-29.

Jahn, Julius A., Calvin F. Schmid, and Clarence Schrag, "The Measurement of Ecological Segregation" *American Sociological Review*, CIII (1947), 293-303.

Jencks, Christopher *Inequality: A Reassessment of the Effect of Family and Schooling in America*, (New York, NY: Basic Books, 1972).

Kain, John, . "Housing Segregation, Negro Employment, and Metroplitan Decentralization," *Quarterly Journal of Economics*, LXXXII (1968), 175-197.

Katz, Lawrence, Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics*, CXVI (2001), 607-54

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence Katz, "Experimental Analysis of Neighborhood Effects." Working Paper. Princeton University, 2005.

Lazear, Edward. . "Culture and Language," *Journal of Political Economy*, CVII (1999), S95-S129.

Lucas, Samuel R.,*Tracking Inequality: Stratification and Mobility in American High Schools.* (New York, NY: Teachers College Press, 1999).

Massey, Douglas and Nancy Denton, "The Dimensions of Residential Segregation." *Social Forces*, LXVII (1988), 281-315

Massey, Douglas and Nancy Denton, *American Apartheid: Segregation and the Making of the Underclass.* (Cambridge, MA: Harvard University Press, 1993).

Mickelson, Roslyn. A., "Subverting Swann: The Effects of First- and Second-Generation Segregation in the Charlotte-Mecklenburg Schools," *American Educational Research Journal*, XXXVIII (2001), pp. 215–52.

Orfield, Gary, *Public School Desegregation in the United States, 1968-1980.* (Washington, DC: Joint Center for Political Studies, 1983).

Philipson, Tomas, "Social Welfare and Measurement of Segregation." *Journal of Economic Theory*, LX (1993), 322-334

Reardon, Sean F., and Glenn Firebaugh, "Measures of Multigroup Segregation." *Sociological Methodology* XXXII (2002), 33-67.

Samuelson, Paul A. and Robert Solow, "Balanced Growth Under Constant Returns to Scale," *Econometrica* XXI (1953), 412-424.

St. John, Nancy H. . *School Desegregation Outcomes for Children*, New York, NY: John Wiley and Sons (1975).

Taeuber, Karl. and Alma Taeuber, *Negroes in Cities: Residential Segregation and Neighborhood Change.* (Chicago, IL: Chicago Aldine Publishing Co., 1965)

Division of the Humanities and Social Sciences, 228-77, California Institute of Technology, Pasadena, California 91125, USA.

email: fede@hss.caltech.edu

Harvard University Society of Fellows and National Bureau of Economic Research, Cambridge, MA 02138, USA.

email: rfryer@fas.harvard.edu

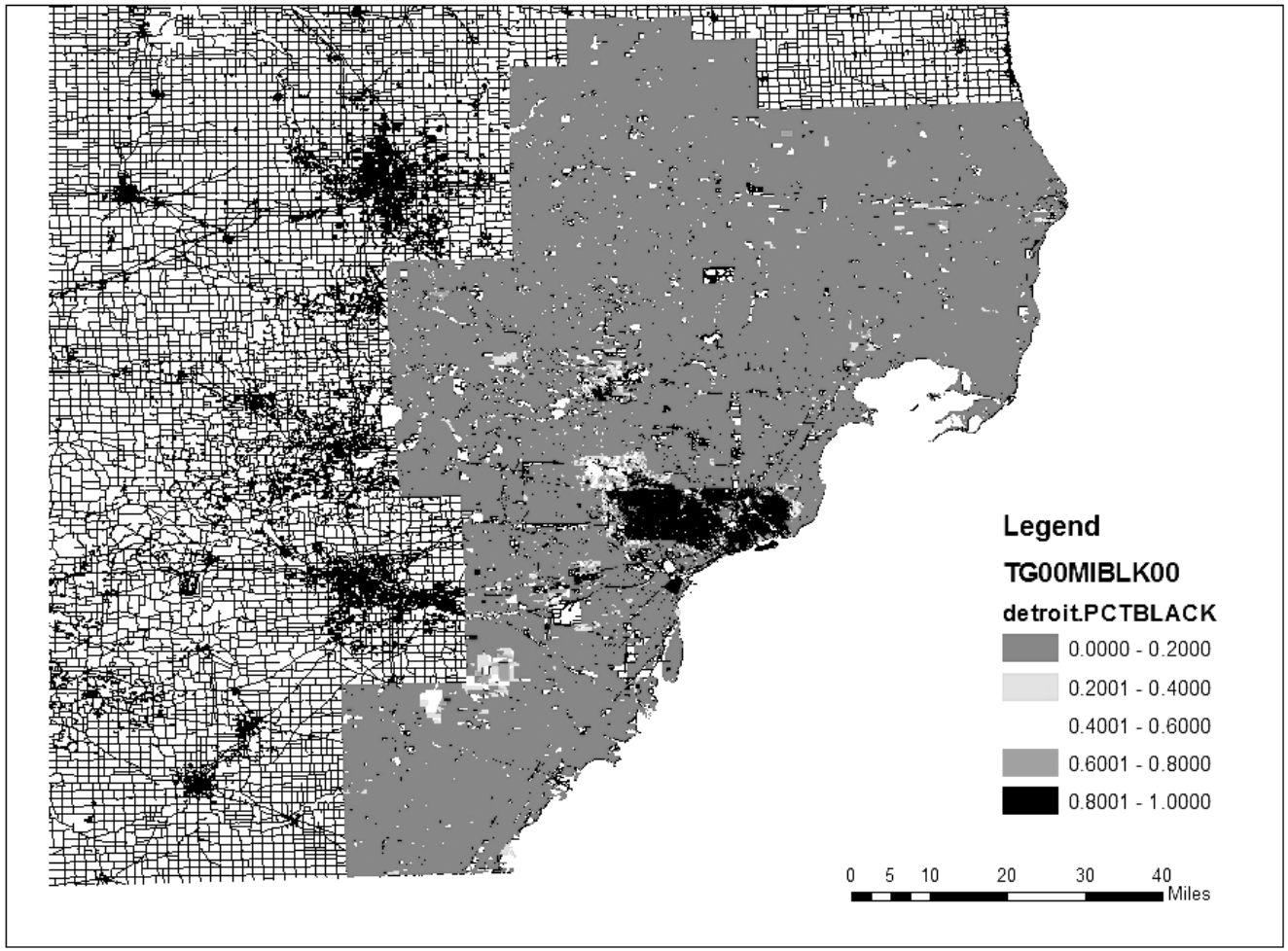

Figure I: Segregation in Metropolitan Detroit

Notes: Figure I is based on block-level data from the 2000 U.S. Census.

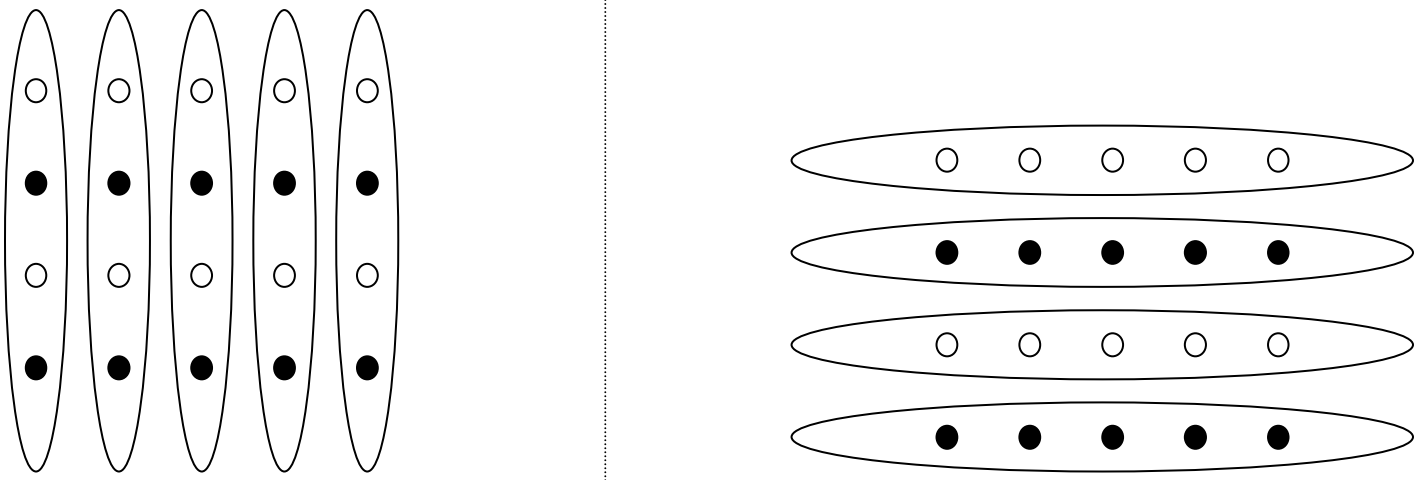

Figure II: A hypothetical city

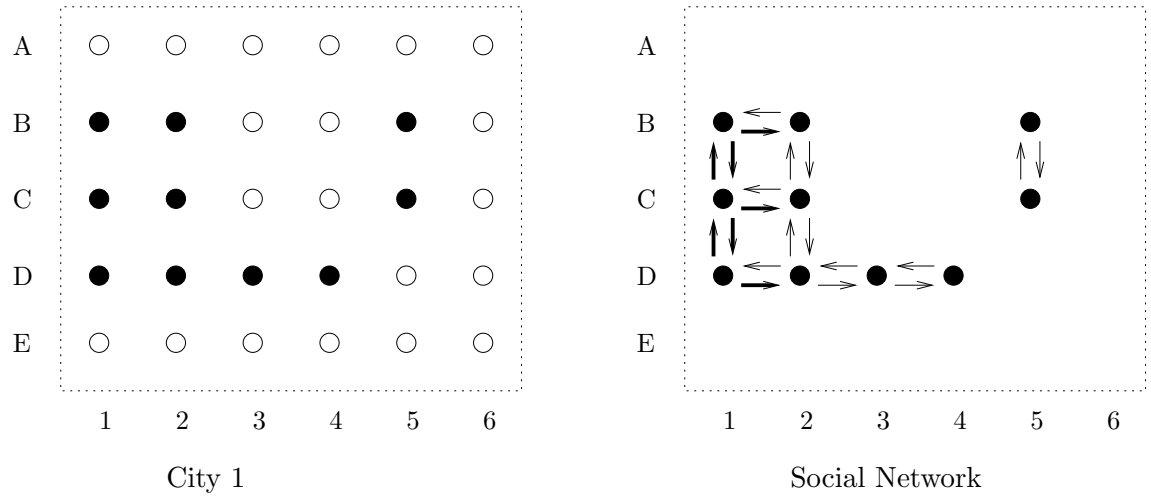

Figure III: A Simple Example.

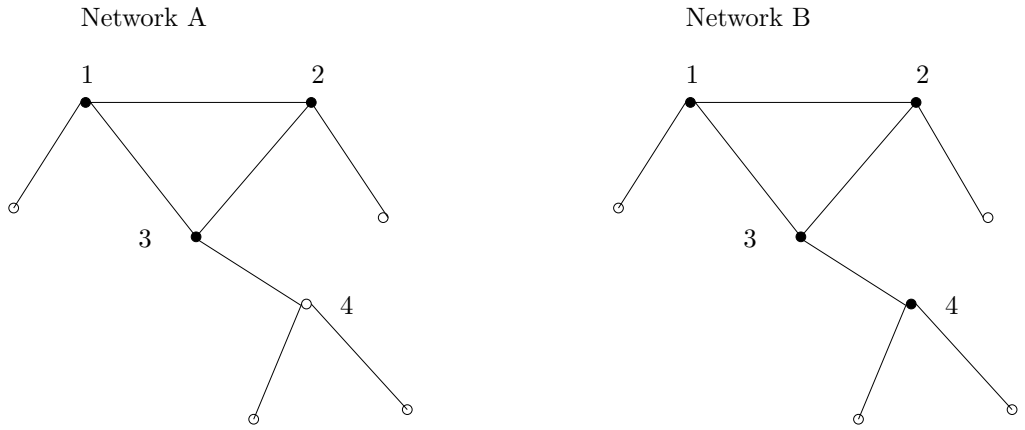

Figure IV: Individual 4 Changes Race.

Figure V: A Change in the Number of White Neighbors.

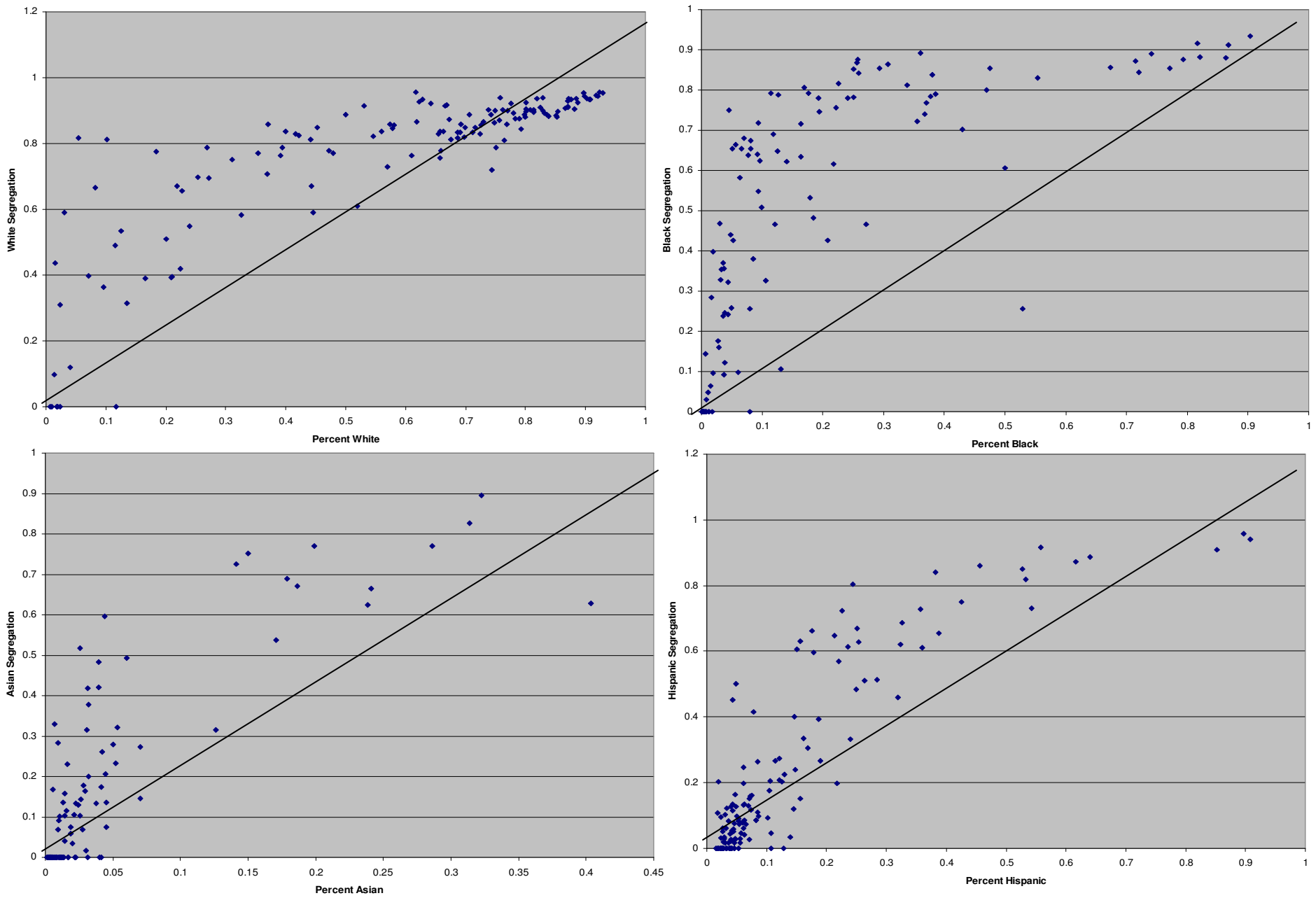

Figure VI: The Relationship Between Group Size and Group Segregation, By Race

Notes: Figure VI is based on data from the National Study of Adolescent Health. Each data point represents segregation calculated at the school level based on students' responses about who their friends are.

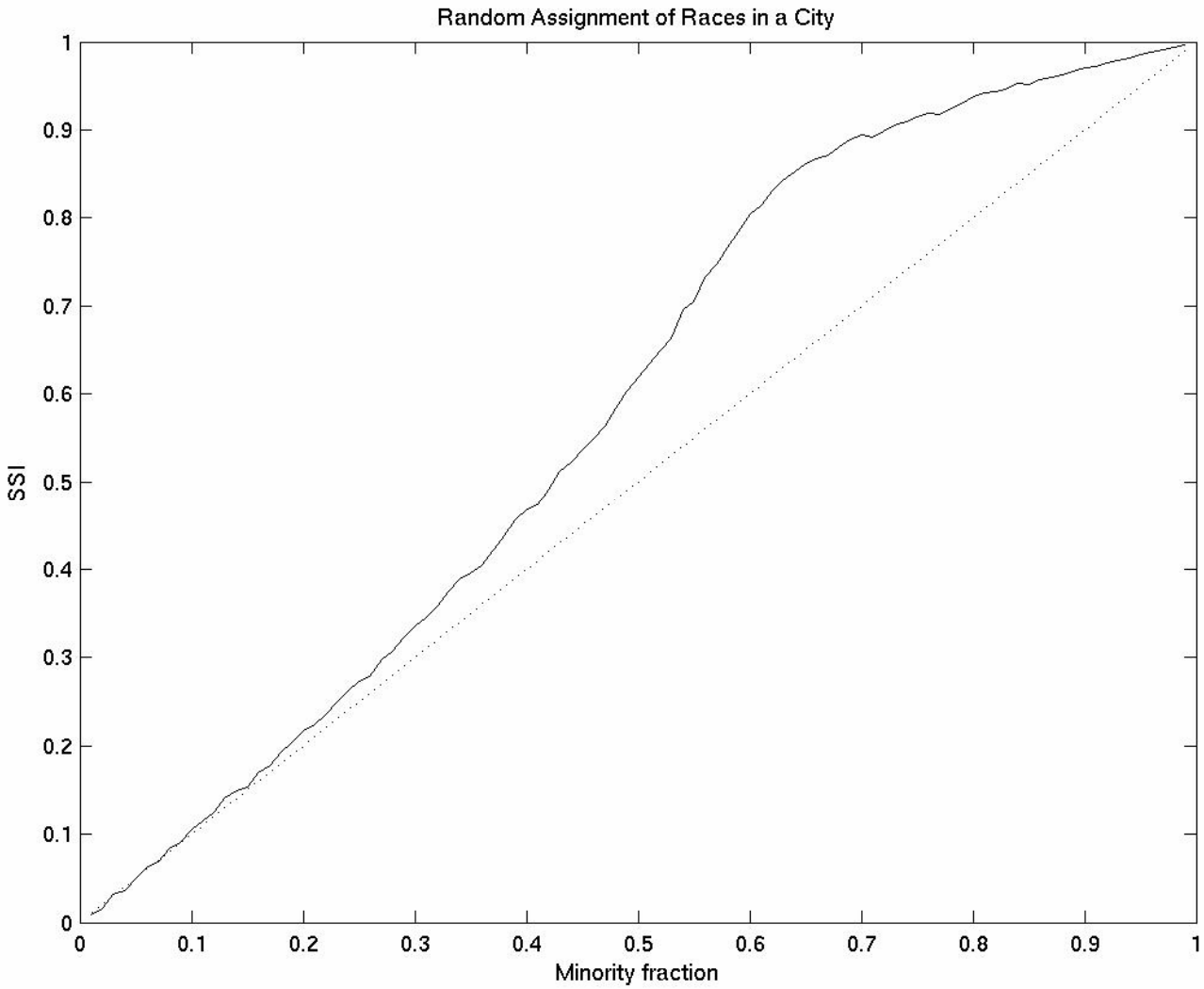

Figure VII: Simulating the Baseline Spectral Segregation Index

Notes: We have obtained measures of Baseline SSI by simulating random assignment of races to large regular (in a graph-theoretic sense) cities with the corresponding fraction of race-h inhabitants. For each fraction p=0.01,0.02,…0.99 we simulated 1,000 cities of 100 households each, where each household is of race h with probability p.

TABLE I

Spectral Segregation of Blacks in City 1

Component 1	$(B,1)$	$(B,2)$	$(C,1)$	$(C,2)$	$(D,1)$	$(D,2)$	$(D,3)$	$(D,4)$	SSI
	0.87	0.62	1.26	0.92	0.93	0.75	0.29	0.10	0.72
Component 2	$(C,5)$	$(B,5)$							SSI
	0.25	0.25							0.25
City 1									SSI
									0.63

TABLE II

SSI Before and After the Change

	1	2	3	4	$\hat{S}$
Before	0.67	0.67	0.67	0	0.67
After	0.78	0.78	0.91	0.42	0.72

TABLE III

The Relationship Between Individual Level Segregation And Outcomes

	Social				**Academic**			
	Smoking	Skip School	Interracial Dating	Happiness	PVT Scores	No College	Grades	Effort
Black	-0.143**	-0.010**	0.085**	-0.091**	-0.424**	-0.013*	-0.216**	0.027**
	(.004)	(0.003)	(0.017)	(0.007)	(0.026)	(0.006)	(0.011)	(0.002)
Asian	-0.081**	-0.008*	0.372**	-0.025*	-0.303**	-0.047**	0.258**	0.036**
	(0.006)	(0.004)	(0.029)	(0.01)	(0.042)	(0.007)	(0.015)	(0.003)
Hispanic	-0.040**	0.026**	0.460**	-0.016*	-0.426**	0.065**	-0.182**	0
	(0.005)	(0.003)	(0.017)	(0.007)	(0.026)	(0.006)	(0.01)	(0.002)
Individual SSI (*1000)	0.007	-0.355	-6.765**	0.11	1.32	-0.02	0.739	0.3
	(0.611)	(0.214)	(2.379)	(0.814)	(3.1877)	(0.679)	(1.038)	(0.2452)
Black*Individual SSI (*1000)	-2.914*	-1.133	-4.311	3.21	-25.453*	0.19	0.212	0.07
	(1.167)	(0.664)	(9.126)	(2.9947)	(12.8544)	(2.6717)	(4.729)	(0.9232)
Asian*Individual SSI (*1000)	-6.377	-8.038**	-66.776**	21.111**	-101.626**	0.12	17.450	4.374*
	(5.173)	(1.834)	(16.963)	(5.791)	(30.061)	(6.619)	(14.144)	(2.081)
Hispanic*Individual SSI (*1000)	-6.355*	-1.608	-14.584	4.04	-47.409**	12.410**	-13.103**	3.867**
	(2.958)	(2.202)	(14.042)	(4.243)	(13.754)	(3.126)	(3.552)	(1.401)
Age	0.029**	0.009**	0	-0.037**	-0.034**	0.021**	-0.024**	-0.011**
	(0.001)	(0.001)	(0.003)	(0.001)	(0.006)	(0.001)	(0.002)	(0.001)
Male	0	0.019**	0	0.047**	0.124**	0.085**	-0.184**	-0.047**
	(0.003)	(0.002)	(0.008)	(0.004)	(0.014)	(0.003)	(0.006)	(0.001)
Mother College Educated	-0.024**	0	0	0.031**	0.099**	-0.080**	0.154**	0.006**
	(0.004)	(0.002)	(0.01)	(0.005)	(0.019)	(0.004)	(0.008)	(0.002)
Father College Educated	-0.032**	-0.010**	0.01	0.021**	0.078**	-0.075**	0.163**	0.013**
	(0.004)	(0.002)	(0.012)	(0.005)	(0.021)	(0.004)	(0.008)	(0.002)
Mother Professional	0	0	0.01	-0.01	0.067**	-0.024**	0.062**	0.003*
	(0.004)	(0.002)	(0.01)	(0.005)	(0.019)	(0.004)	(0.007)	(0.002)
Father Professional	-0.008*	0	0.02	0.022**	0.127**	-0.048**	0.114**	0
	(0.004)	(0.002)	(0.011)	(0.005)	(0.02)	(0.004)	(0.008)	(0.002)
Constant	-0.220**	-0.108**	0.117*	1.134**	0.725**	-0.125**	3.216**	0.989**
	(0.017)	(0.011)	(0.049)	(0.022)	(0.088)	(0.019)	(0.035)	(0.008)
Observations	78075	77903	9553	73837	14387	69257	72744	79599
R^2	0.07	0.04	0.37	0.05	0.28	0.1	0.18	0.08
Mean of Dependent Variable	.166	.049	.303	.580	.055	.211	2.798	.810
SD of Dependent Variable	.372	.215	.460	.494	.975	.408	.809	.171

All regression use data from the National Longitudinal Survey of Adolescent health. Dependent variables vary by column. Smoking and Skip School are binary variables taking the value 1 if the student does the activity once a week or more. Interracial Dating is a binary variable equal to one if a student reports ever dating someone of a different race. Happiness is a binary value taking the value of one if the student agrees or strongly agrees that they are happy to be at their school. No college is a binary variable that equals one if the student reports a probability of .5 or greater that she will attend college. Effort is an ordered categorical variable that takes values .25 if student never tries at all, .50 if they don't try very hard, .75 if the student reports they try hard enough, but not as hard as they could, and 1 if the student reports they try very hard to do their best. Test scores are adjusted to be standard normal. Grade composites are constructed from 4 reported grades: English/languages arts, mathematics, history/social studies, and science. Grades are first converted to their equivalent on a 4-point scale: A=4, B=3, C=2, D=1.

In all cases, dummy variables for missing values and school fixed effects are included. Robust standard errors are beneath the coefficients. * significant at 5%; ** significant at 1%.

TABLE IV

Correlation Between Existing Measures of Segregation and the Spectral Index

	SSI	Dissimilarity	Isolation	Exposure	Entropy	Gini	% Black	Interaction	SSI-Baseline
SSI	1								
Dissimilarity	**0.42**	1							
Isolation	**0.93**	0.56	1						
Exposure	**0.91**	0.59	0.95	1					
Entropy	**0.47**	-0.38	0.36	0.34	1				
Gini	**0.46**	1	0.6	0.63	-0.37	1			
Percent Black	**0.9**	0.31	0.92	0.84	0.56	0.35	1		
Interaction	**0.47**	-0.35	0.37	0.34	0.98	-0.33	0.57	1	
SSI-Baseline	**0.89**	0.39	0.74	0.8	0.2	0.41	0.61	0.18	1

All calculations performed using block-level data from from all 313 MSAs in the 2000 US Census. The sample includes all census blocks in all MSAs. Baseline SSI calculated from simulations described in Section 5.1.B.

TABLE V

The Relationship Between Segregation and Outcomes

	Age 20-24					Age 25-30				
	Education		Income		Social	Education		Income		Social
	High School Graduate	College Graduate	Idle	Earnings	Single Mother	High School Graduate	College Graduate	Idle	Earnings	Single Mother
Dissimilarity Index										
Segregation	.002	.008	-.001	-.008	.001	.003	-.002	.000	-.008	-.003
	(.004)	(.005)	(.002)	(.009)	(.004)	(.003)	(.008)	(.003)	(.008)	(.003)
Segregation*Black	-.041	-.010	.041	-.093	.045	-.032	-.006	.035	-.064	.059
	(.006)	(.004)	(.006)	(.019)	(.008)	(.006)	(.007)	(.005)	(.015)	(.007)
Spectral Segregation Index										
Segregation	-.002	.007	-.002	.002	-.001	-.001	.000	-.004	.012	-.001
	(.005)	(.007)	(.003)	(.009)	(.004)	(.004)	(.011)	(.004)	(.007)	(.004)
Segregation*Black	-.036	-.008	.024	-.091	.070	-.041	-.008	.025	-.051	.073
	(.010)	(.005)	(.009)	(.027)	(.013)	(.008)	(.009)	(.008)	(.023)	(.013)
Summary Statistics										
N	97,976	97, 976	97, 976	56, 627	49, 038	139,715	139,715	139,715	105, 997	71,531
R^2	.034 / .034	.093 / .093	.050 / .048	.090 / .089	.108 / .108	.031 / .031	.040 / .040	.050 / .048	.092 / .091	.109 / .108

All regressions are estimated using the 1990 1% Census Pums. Dependent variables vary by column. Idleness is defined as not working and not enrolled in school. Earnings are the sum of wage, salary, and self-employment income in 1989. The sample for earnings consists of individuals who are not working, not enrolled in school, and have non-negative earnings. All regressions include the following covariates: an exhaustive set of racial dummy variables, gender, single year age dummy variables, log of population, percent black, log median household income, and manufacturing share. The latter four covariates are also interacted with a black dummy. Standard errors, reported in parentheses, are corrected for heteroskedasticity and intra-MSA clustering of the residuals.