Random–Walk–Based Segregation Measures*

Coralio Ballester† Marc Vorsatz‡

April 6, 2013

*We are very grateful to Antonio Cabrales, Matthew Jackson, Alfonso Rosa and Yves Zenou for very fruitful discussions. Two anonymous referees helped us to improve the paper significantly. We also thank the seminar/conference participants at ASSET (Alicante), IMEBE (Bilbao), FEDEA, Alicante, Bilbao, Innsbruck, Lisbon, Murcia, U. Pompeu Fabra, Saarbrücken, Valencia, and Vigo for many very helpful comments. It would have been impossible to conduct the empirical analysis without the financial support of FEDEA.

†Corresponding author. Departamento de Fundamentos del Analísis Econónomico, Universidad de Alicante, Campus de San Vicente, 03080 Alicante, Spain. Email: coralio@ua.es. Financial support from Fundación Ramón Areces and the Spanish Ministry of Education and Innovation and FEDER, through the project SEJ–2007–62656, is gratefully acknowledged.

‡Departamento de Análisis Económico II, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain; and Fundación de Estudios de Economía Aplicada (FEDEA), Calle Jorge Juan 46, 28001 Madrid, Spain. Email: mvorsatz@cee.uned.es. Financial support from the Spanish Ministry of Education and Science, through the project ECO2009-07539, and Spanish Ministry of Economy and Competitiveness, through the project ECO2012-31985, is gratefully acknowledged.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Abstract

We propose an intuitive way of how to measure segregation in social and spatial networks. Using random walks, we define the *segregation index* as the probability that an individual meets an individual from the same social group. The segregation index is a generalization of the isolation index and a homophily index introduced in Currarini et al. (2009), and it has a closed–form relation to PageRank that facilitates its computation. We also show that the Spectral Segregation Index proposed by Echenique and Fryer (2007) is not continuous with respect to the network structure. Finally, we apply the measure to Spanish census data and to citations data from Economics, and rationalize the index as the equilibrium outmode of a game.

Keywords: Isolation, Homophily, Network, PageRank, Markov Chain, Segregation.

JEL–Numbers: C0, D85, Z13.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

1 Introduction

The objective of this study is to propose a new measure of residential segregation — the degree to which social groups live apart —, which is important in order to assess both the causes and consequences of social separation. For instance, Cutler et al. (2008) analyze how cultural factors explain the differences in segregation across ethnic groups. Also, residential segregation has a variety of effects on socio-economic variables and its analysis is implicit in the design of integration policies.[1]

The most basic framework that addresses the question of how to measure residential segregation starts with some geographic space (*i.e.*, a city) that is divided into smaller units called census tracts or neighborhoods. Also one takes the distribution of the different social groups (the population can be partitioned using observable characteristics such as ethnicity, religion, or nationality) over census tracts as given. Within this setting, the classical work of Massey and Denton (1988) introduces the following five dimensions of segregation: *evenness*, or the extent to which a group is distributed homogeneously over neighborhoods; *exposure*, or the degree of potential contact to other groups; *concentration*, or the amount of physical space a group occupies; *centralization*, or the degree to which a group resides close to the city center; and *clustering*, or the extent to which individuals from the same group tend to live in neighborhoods next to each other.

In a recent major contribution, Echenique and Fryer (2007) apply graph theory to derive a measure of (residential) segregation that they call the *Spectral Segregation Index* (SSI). The social interaction framework studied is based on the assumption that individuals divide their time among their neighbors/friends, and the more time they spend with individuals from their group, the more segregated the group as a whole is. The authors then apply the SSI to the U.S. Census data from 2000 (replicating the results of Cutler and Glaeser 1997 on ghettos and uncovering a high rank correlation with measures that capture the dimensions of evenness and exposure) and to friendship networks in schools.

[1]See, among others, Benabou (1993), Borjas (1995), Boeri et al. (2012), Card and Rothstein (2007), Cutler and Glaeser (1997), Edin et al. (2003), Kling et al. (2007) and Zenou (2009).

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Our setting departs from Echenique and Fryer (2007) in several crucial aspects. First, we allow multiple individuals to be located in the same node/neighborhood. Second, and most importantly, the segregation of a group depends in our model on the distribution of the other groups in *all* census tracts. The SSI, on the other hand, only considers the information that is contained in the subgraph of the considered group; that is, all individuals not belonging to the group in question are eliminated from the network. Our approach is therefore fundamentally different from the SSI, and, in particular, it is related to the following process: Pick any two individuals from a given group g at random and suppose that the first of the two individuals moves over the network (city) in such a way that in the first period, she advances from her area of residence to some neighboring census tract. In each subsequent period, she either moves from her current position in the network to some adjacent node (this event happens with probability $0 \leq \alpha < 1$) or the process stops (this event happens with probability $1 - \alpha$). Hence, the parameter α can be interpreted as the degree of spatial mobility. The *normalized segregation index* for group g, denoted by $\bar{\sigma}_g$, is then defined as the probability that the two randomly chosen individuals meet when the random–walk of the first individual terminates.[2,3]

After formally introducing the segregation index, we show that the measure directly incorporates three of the five dimensions introduced by Massey and Denton (1988); namely those of exposure, clustering, and centralization. First, the segregation index is a natural spatial generalization of the isolation index (see, Lieberson 1981) to networks. In fact,

[2]People randomly meet friends, friends of friends, or higher–order acquaintances. Also, socio-economic relationships are usually recursive, self–enforcing, and with feedback effects. For example, Jackson and Rogers (2007) show for several situations that a random link formation process matches empirical networks very well.

[3]Since *both* individuals are chosen at random, the measure is controlling for group sizes. It is also robust to the number of census tracts, the definition of the boundaries of census tracts, and the population size. We nevertheless also define the group–size–dependent measure σ_g according to which a randomly chosen individual will meet *any* individual from the same group when the random walk terminates. All our theoretical results hold for both the group–size–dependent and the normalized measure.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

σ_g reduces to the isolation index of group g if the socio–geographical network is empty and individuals only interact in the neighborhoods they reside in (Proposition 1). Note that while σ_g explicitly accounts for the intensity of interactions within group g, it is nevertheless directly related to the level of interaction with other groups (exposure) as well. This is because more intra–group interactions prevents the group from socializing with others.

To see that the segregation index is also related to the dimension of clustering, suppose that $\alpha = 0$. In this case, the only interactions that matter are those with individuals from neighboring census tracts, and the measure becomes equal to the average proportion of all immediate neighbors that belong to the same group. Formally, Proposition 2 establishes that for $\alpha = 0$, the segregation index reduces to the homophily index introduced by Currarini et al. (2009). Since social interactions with individuals who live further away become more and more feasible as α increases, one can regard σ_g as a generalized measure of homophily/clustering that incorporates social relations beyond immediate neighbors.

Third, the segregation index also takes into account the dimension of centralization because groups that are closer to the socio–geographical center of the network are more likely to interact with individuals from the same group everything else equal. The relevance of centralization as a dimension of segregation has been reported in some countries like the U.S., where socio-economic factors can cause minorities to move to the city center. In particular, the spatial mismatch literature emphasizes that blacks who tend to live closer to the city center have worse labor market outcomes (see, Ihlanfeldt and Sjoquist 1998 and Gobillon et al. 2007). Proposition 3 shows that the segregation index of a group g is a weighted sum over the homophily indices of the nodes (the probability that a neighbor belongs to the same group), where the weight assigned to each node is assessed through the PageRank index used by Google to determine the importance of webpages in the World Wide Web.[4] Thanks to this connection to PageRank, we are able to show

[4]We also formally relate the segregation index to the dimension of evenness by showing that the normalized isolation index is minimized if the individuals of a group are equally distributed over all nodes (Proposition 4). However, uneven distributions can lead to

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

how the segregation index can be easily implemented and efficiently calculated in regular computers, even for a large number of census tracts. Implementation files and examples can be found in the web appendix.

In the next step, we analyze a drawback of the SSI. We present an example showing that the SSI is not continuous in the strength of the social ties because in situations when two separated components of a network are merged by a new link, the SSI may change abruptly. We believe, and argue in more detail in Section 4, that continuity is a very important property because both social and geographical networks change over time and are subject to a continuous addition and deletion of links. By considering the network structure we maintain the postulate of Echenique and Fryer (2007) that segregation is formed through social interactions, but we amend their drawback by developing a continuous measure.

In the empirical analysis, we first explore the geographical close–knitness of foreign residents in Spain in 2009. We find that the southern part of the Mediterranean coast of Almería and the autonomous community of Catalonia are the most segregated areas, while Galicia, the Basque Country, and Andalusia are the least segregated ones. Also, the Pakistani and immigrants from Nigeria and Senegal are the most segregated groups, while the Latin Americans turn out to be rather integrated. Given the known detrimental effects of segregation on labor market outcomes (and therefore on income and other correlated variables), this result suggests a proactive housing policy that targets in particular immigrants from poor countries facing a language barrier. One final important feature of the model is that it allows us to uncover network/clustering effects — the change in the measure due to considering the network on top of the isolation index. These effects are strong for the Pakistani and some groups from Africa, but weak for immigrants from Western Europe.

To show that the theoretical model is flexible enough to address a wide range of different questions, we consider afterwards citation data for 140 leading Economics journals (that are divided into 18 different fields) from 2010 to articles published in 2005–2009.

even lower levels of segregation as soon as the network is introduced.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Letting α vary allows us again to uncover clustering in the underlying distribution. It turns out that among the smaller fields without many self citations, Accounting and Sectorial Studies show strong network effects, which suggests the existence of citation circles (journals from these fields cite journals from other fields, ..., that cite journals from the original field). This effect is not found in other fields of similar size (Human Resource Management and Social Economics).

Our final theoretical contribution is the construction of a game–theoretical model that rationalizes the segregation index. We use techniques similar to those developed in Currarini et al. (2009), where individuals are involved in a search process and utilities depend on whether they meet people who belong to the same or a different group. A network of relationships arises as a consequence of these encounters. Ours is an essentially different search model. Individuals search through a fixed spatial network and there is no preference bias towards meetings with people from the same group. To be more concrete, individuals have to decide how much time to invest in each of the groups/cultures and then meet people from a neighboring node. We assume that there are two situations that create benefits for an individual who invests in culture g: (a) her match also invests in culture g and (b) her match belongs to group g. In the unique steady state equilibrium, the average time invested by group g in its own culture is proportional to its segregation index. The fact that the benefits from the meetings are type–independent highlights the unbiased nature of the segregation index. In fact, since the level of interaction of any individual of group g with her neighbors does not depend on her group, the segregation index explains how the spatial setting alone can induce separation.

We proceed as follows. Next, we introduce the segregation index and relate it to the various dimensions of segregation. In Section 4, we discuss the discontinuity of the SSI. Section 5 presents the empirical results. In Section 6, we study the search model. Finally, we conclude.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

2 The Segregation Index

Consider a set $N = \{1, 2, \ldots, n\}$ of n *individuals*. For simplicity, we will call N a *society*. The individuals live in a city that is composed of a finite set $M = \{1, 2, \ldots, m\}$ of m *neighborhoods* or census tracts. A particular neighborhood $i \in M$ will also be referred to as a *node*. It is assumed that every individual lives in exactly one neighborhood, but that each neighborhood possibly inhabits multiple individuals. Also, there is a set G of *groups* that forms a partition of the society. One can think of a group $g \in G$ as a subset of members of the society that share a particular attribute such as religion or ethnicity. Let $n_{g,i}$ be the number of individuals of group $g \in G$ that live in neighborhood $i \in M$. The number of individuals belonging to group $g \in G$ is $n_g = \sum_{i \in M} n_{g,i}$. Similarly, $n_i = \sum_{g \in G} n_{g,i}$ is the number of individuals that reside in neighborhood $i \in M$. The column vectors $\mathbf{c}_g = (n_{g,i}/n_i)_{i \in M}$ and $\mathbf{d}_g = (n_{g,i}/n_g)_{i \in M}$ are referred to as the vectors of *group concentrations* and *group densities*, respectively.

The different neighborhoods in a city are interconnected through links. Formally, $\mathbf{A}$ is an $m \times m$ matrix such that $a_{i,j} = 1$ if there is a connection between i and j, and $a_{i,j} = 0$ otherwise. The intuition is that two neighborhoods are connected if they are geographically adjacent or if individuals from these neighborhoods directly interact with each other.[5] We assume that $a_{i,i} = 1$ so that individuals from the same neighborhood can directly interact with each other. Let $a_i = |\{j \in M : a_{i,j} = 1\}|$ be the number of neighborhoods $i \in M$ is connected to. An $m \times m$ matrix $\mathbf{P}$ is said to be a *row stochastic matrix associated with* $\mathbf{A}$ whenever the following conditions hold: $p_{i,j} = 0$ whenever $a_{i,j} = 0$, $p_{i,j} > 0$ whenever $a_{i,j} = 1$, and $p_i \equiv \sum_{j \in M} p_{i,j} = 1$. With this notation at hand, we can now formally define a *city* as a tuple $C = \langle N, M, G, (n_{g,i})_{g \in G, i \in M}, \mathbf{P} \rangle$.

The advantage of working with a stochastic $\mathbf{P}$ is that it provides a natural interpretation in terms of node–to–node transition probabilities. A *walk* is a sequence

[5] There is no need to assume that the graph is symmetric. Asymmetric graphs occur, for instance, when $\mathbf{A}$ represents a social communication structure. Also, friendship networks are not necessarily symmetric. Finally, the mobility between two different social strata could be more sticky depending on the direction of movement.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

$\omega = (\omega_0, \omega_1, \omega_2, \ldots)$ of at least two, possibly repeated neighborhoods with the restriction that two consecutive neighborhoods must be connected. We assume that in period $t = 0$, a walk passes from the initial neighborhood ω_0 to some adjacent neighborhood ω_1. In every subsequent period $t = 1, 2, 3 \ldots$, a walk continues with probability $\alpha \in [0, 1)$ from the current position ω_t to an adjacent neighborhood ω_{t+1}, and stops with probability $1 - \alpha$. In case of continuation, the walk currently at neighborhood i passes to j with probability $p_{i,j}$. Formally, a *random–walk* is a random variable whose realization is a particular walk ω, where the probability of a walk $\omega = (\omega_0, \omega_1, \ldots, \omega_l)$ of length $|\omega| = l$ conditional on the initial node being ω_0 is

$$\text{Prob}(\omega|\omega_0) = \alpha^{(l-1)} (1 - \alpha) \, p_{\omega_0,\omega_1} \, p_{\omega_1,\omega_2} \, \cdots \, p_{\omega_{l-1},\omega_l}.$$

For example, conditional on that $w_0 = 4$, the walk $\omega = (4, 3, 2)$ occurs with probability $p_{4,3} \cdot \alpha \cdot p_{3,2} \cdot (1 - \alpha)$. Since $\alpha < 1$, any realization ω of a random–walk has finite length with probability 1. Let Ω be the set of all possible realizations of ω. The expected walk length $\mathbb{E}(l)$ is[6]

$$\mathbb{E}(l) = \sum_{l=1}^{\infty} l \sum_{\omega \in \Omega : |\omega| = l} \text{Prob}(\omega|\omega_0) = \frac{1}{1 - \alpha}.$$

This equation shows that α is related to the expected walk length, and can thus be interpreted as the degree of *spatial mobility*. For instance, if $\alpha = 0.85$, then $\mathbb{E}(l) = 6.67$. Similarly, if $\alpha = 0.99$, then $\mathbb{E}(l) = 100$.[7]

[6] To see this, note that since every node has an out–link ($a_i > 0$), the expected length of a walk is independent of its origin. Now, take any initial node ω_0. The random–walk moves to ω_1 with probability one. But the random–walk starting at ω_1 has, due to the independence of the origin, an expected length of α times $\mathbb{E}(l)$. Hence, $\mathbb{E}(l) = 1 + \alpha \, \mathbb{E}(l)$.

[7] It is important to note that the model parameter α gives raise not just to a single measure but to a family of measures, one for each α. This allows us to analyze the phenomenon of segregation as a function of the spatial mobility. In particular, it will become clear in the empirical analysis that a comparison of the degree of the segregation for different α uncovers group clustering (network effects).

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Given the transition matrix $\mathbf{P}$ and the continuation probability $\alpha \in [0, 1)$, let $\mathbf{Q}$ be the $m \times m$ matrix such that $q_{i,j}$ is the probability that a walk ends in neighborhood j, given that it started in i. Our first objectives is to provide a closed–form expression of $\mathbf{Q}$ in terms of the parameters $\mathbf{P}$ and α, yet it is already easy to see that (a) if $\mathbf{P} = \mathbf{I}$, no interaction takes place across neighborhoods, (b) if $\mathbf{P} \neq \mathbf{I}$ and $\alpha = 0$, no interaction takes place beyond immediate neighbors, and (c) if $\mathbf{P} \neq \mathbf{I}$ and α tends to 1, the expected walk length grows arbitrarily large.

Lemma 1. $\mathbf{Q} = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{P}$.

Proof. See the corresponding section. □

The intuition behind this result is as follows. In order to compute $q_{i,j}$, we have to consider all walks of the form $\omega = (i, \omega_1, \omega_2, \ldots, j)$. If $q_{i,j}^{[l]}$ denotes the probability that j is reached in exactly l steps from i, then

$$q_{i,j} = \sum_{l=1}^{\infty} q_{i,j}^{[l]} = (1-\alpha)\, p_{i,j} + \alpha(1-\alpha) \sum_{\omega_1 \in M} p_{i,\omega_1}\, p_{\omega_1,j} + \alpha^2(1-\alpha) \sum_{\omega_1,\omega_2 \in M} p_{i,\omega_1}\, p_{\omega_1,\omega_2}\, p_{\omega_2,j} + \ldots$$

or, in matrix notation,

$$\begin{aligned} \mathbf{Q} &= (1-\alpha)\mathbf{P} + \alpha(1-\alpha)\mathbf{P}^2 + \alpha^2(1-\alpha)\mathbf{P}^3 + \ldots \\ &= (1-\alpha)\left(\mathbf{I} + \alpha\mathbf{P} + \alpha^2\mathbf{P}^2 + \ldots\right)\mathbf{P} \\ &= (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\,\mathbf{P}. \end{aligned}$$

We can now suggest a measure of residential segregation that is related to the level of within–group interactions, taken into account through the matrix $\mathbf{Q}$. The *segregation index* σ_g of group $g \in G$ is defined as the probability that a randomly chosen individual of group g meets another individual from the same group in the neighborhood where her random–walk terminates.

Definition 1. Given city $C = \langle N, M, G, (n_{g,i})_{g \in G, i \in M}, \mathbf{P} \rangle$ and the continuation probabil-

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

ity $\alpha \in [0,1)$, the *segregation index of group* $g \in G$ is

$$\sigma_g(C,\alpha) = \sum_{i\in M} d_{g,i} \sum_{j\in M} q_{i,j}\, c_{g,j} = \mathbf{d}_g^{\mathsf{T}} \mathbf{Q} \mathbf{c}_g.$$

Since σ_g is a probability, it takes values from the interval $[0,1]$. To see that the measure is increasing in the relative size n_g/n of group g, take any city C and consider the city $\hat{C}$ that can be obtained from C by doubling the concentration of individuals of group g in each neighborhood maintaining the total population in each neighborhood constant. Thus, $\hat{\mathbf{d}}_g = \mathbf{d}_g$, $\hat{\mathbf{c}}_g = 2\,\mathbf{c}_g$ and $\hat{\mathbf{Q}} = \mathbf{Q}$, which implies that $\sigma_g(\hat{C},\alpha) = 2\,\sigma_g(C,\alpha)$. Consequently, the segregation of group g has doubled simply because the group has grown but not because its distribution over the network has changed. The reason for this dependence is that it is only required that an individual meets *any* (and not a randomly chosen) individual from the same group. This effect can be accounted for with the following normalization.

Definition 2. Given city $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ and the continuation probability $\alpha \in [0,1)$, the *normalized segregation index of group* $g \in G$ is

$$\bar{\sigma}_g(C,\alpha) = \left(\frac{n_g}{n}\right)^{-1} \sigma_g(C,\alpha) = \left(\frac{n_g}{n}\right)^{-1} \mathbf{d}_g^{\mathsf{T}} \mathbf{Q} \mathbf{c}_g.$$

If group g is distributed homogeneously over all neighborhoods (for all $i \in M$, $c_{g,i} = n_g/n$), then $\sigma_g(C,\alpha) = n_g/n$ and $\bar{\sigma}_g(C,\alpha) = 1$. So if $\bar{\sigma}_g(C,\alpha) > 1$, group g is on average overrepresented in census tracts compared to its overall population share. The size–independence of the normalized measure becomes even more visible if it is rewritten as follows:

$$\bar{\sigma}_g(C,\alpha) = n \sum_{i\in M} d_{g,i} \sum_{j\in M} n_j^{-1}\, q_{ij}\, d_{g,j}.$$

According to this equation, $\bar{\sigma}_g$ is equal to the probability (up to the multiplicative scalar n) that a randomly chosen individual from group g meets another randomly chosen individual from the same group. The fact that *both* individuals are chosen at random prevents this probability from depending on the representativeness of the group in the

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

society. Finally, the *segregation index of city* C is defined as the weighted average over the segregation indices of the groups; that is, $\sigma(C,\alpha) = \sum_{g\in G} n_g/n \cdot \sigma_g(C,\alpha)$. Intuitively, $\sigma(C,\alpha)$ is the probability that a randomly chosen individual (of all groups) meets an individual from the same group when her random–walk terminates. The normalized segregation index $\bar{\sigma}(C,\alpha)$ of city C is defined accordingly.

3 Properties

3.1 Exposure

Isolation, the opposite of exposure, is the degree of potential contact or interaction between members of the same group. This notion is captured through the *isolation index*, which is defined as the likelihood that a typical individual of a group faces a member from the same group in her own neighborhood:

$$I_g(C) = \sum_{i\in M} d_{g,i}\, c_{g,i} = \mathbf{d}_g^{\mathsf{T}} \mathbf{c}_g.$$

The normalized isolation index is then $\bar{I}_g(C) = (n_g/n)^{-1} I_g(C)$. Our first proposition directly relates the segregation to the isolation index.

Proposition 1. *Given city* $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ *and the continuation probability* $\alpha \in [0,1)$*, the segregation index of group* $g \in G$ *is*

$$\sigma_g(C,\alpha) = (1-\alpha)\mathbf{d}_g^{\mathsf{T}}(\mathbf{I}-\alpha\mathbf{P})^{-1}\mathbf{P}\mathbf{c}_g.$$

Proof. This is a direct consequence of Lemma 1 and Definition 1. □

Proposition 1 shows that the segregation index reduces to the isolation index whenever the network is empty ($\mathbf{P} = \mathbf{I}$). Our measure can also be interpreted as a weighted average over the entries of the vector $\mathbf{v}_g \equiv \mathbf{Q}\mathbf{c}_g = (1-\alpha)(\mathbf{I}-\alpha\mathbf{P})^{-1}\mathbf{P}\mathbf{c}_g$, where the weight of each neighborhood $i \in M$ is equal to $d_{g,i}$. Since $\mathbf{v}_g$ contains the contributions of all census tracts to the segregation of group g, we will call it the *vector of local segregations* of group

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

g. In particular, $v_{g,i}$ is the expected concentration of group g in the neighborhood where a random–walk that started in neighborhood i terminates. We conclude with an example.

Example. Consider the city C depicted in Figure 1.

[Include Figure 1 about here]

There are two ethnic groups, blacks and whites. Two whites reside in neighborhoods 1 and 3, three whites in neighborhood 4, and one white in neighborhood 2. Moreover, three blacks live in neighborhood 3 and one black in neighborhoods 1, 2 and 4 each. The transition matrix $\mathbf{P}$ in Figure 1 is such that for all $i, j \in M$, $p_{i,j} = a_{i,j}/a_i$. Since $\mathbf{d}^{\mathsf{T}}_{blacks} = (1/6, 1/6, 1/2, 1/6)$, $\mathbf{d}^{\mathsf{T}}_{whites} = (1/4, 1/8, 1/4, 3/8)$, $\mathbf{c}^{\mathsf{T}}_{blacks} = (1/3, 1/2, 3/5, 1/4)$, and $\mathbf{c}^{\mathsf{T}}_{whites} = (2/3, 1/2, 2/5, 3/4)$, it is easy to verify that $\sigma_{blacks}(C, 0.85) = 0.45$ and $\sigma_{whites}(C, 0.85) = 0.55$. The normalized measures are equal to $\bar{\sigma}_{blacks}(C, 0.85) = 1.05$ and $\bar{\sigma}_{whites}(C, 0.85) = 0.97$, respectively. □

3.2 Clustering

Following Massey and Denton (1988) clustering is the extent to which individuals from the same social groups live in neighborhoods next to each other.[8] It is rather intuitive that the segregation index incorporates clustering in the proper definition: An individual who starts a random–walk in her area of residence is more likely to end up in a neighborhood that is closer to her own census tract than in a neighborhood that is further away as soon $\mathbf{P} \neq \mathbf{I}$. To say it differently, two individuals are more likely to meet the closer they live to each other. Figure 2 visualizes this effect. Since for all $\alpha \in [0, 1)$, $q_{4,5}(A, \alpha) > q_{2,5}(B, \alpha)$ and $q_{5,4}(A, \alpha) > q_{5,2}(B, \alpha)$, the segregation of the blacks is higher in City A than in City B. According to Massey and Denton (1988), this is entirely due to clustering.[9]

[8]In the literature on social networks, clustering refers to the probability that two neighbors of a given individual have a direct link between themselves. The two concepts have to be carefully distinguished.

[9]They write on page 293: *"..., suppose we have two urban areas with the same number of minority members, who comprise the same proportion of the total population. In each*

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

[**Include Figure 2 about here**]

From a formal point of view, we establish that the segregation index captures the dimension of clustering by showing that it naturally extends the homophily measures developed in Currarini et al. (2009). The concept of homophily in social networks reflects the propensity of individuals to be directly linked to other individuals of the same type. In particular, the *homophily index* $h_g(C)$ of group $g \in G$ in city C is the expected fraction of individuals in neighboring census tracts who belong to the same group:

$$h_g(C) = \sum_{i \in M} d_{g,i} \sum_{j \in M} \frac{n_{g,j}\, a_{i,j}}{\sum\limits_{k \in M} n_k\, a_{i,k}}.$$

As usual, we define the normalized homophily index as $\bar{h}_g(C) = (n_g/n)^{-1} h_g(C)$. In the example corresponding to Figure 2, $h_{blacks}(A) = 1/2 \cdot (2/3 + 1) = 5/6$ and $h_{blacks}(B) = 1/2 \cdot (1/3 + 1/2) = 5/12$ — the homophily index for the blacks is higher in City A. Our next proposition shows that if $\alpha = 0$ and the transition probabilities are proportional to the sizes of the neighborhoods, the segregation index coincides with the homophily index. Consequently, the segregation index can be interpreted as a generalized measure of homophily that reaches beyond immediate neighbors for strictly positive α.

Proposition 2. *If* $p_{i,j} = (a_{i,j}\, n_j)/(\sum_{k \in M} a_{i,k}\, n_k)$ *for all* $i, j \in M$*, then* $\sigma_g(C, 0) = h_g(C)$.

Proof. See the corresponding section. □

place, no minority member shares a common residential area with a majority member, all minority areas are located the same average distance from the central business district, and all areas are of the same geographic size. In this case, both urban areas would display identical measures of evenness, exposure, concentration, and centralization. However, if all minority areas in one of the urban areas were contiguous to one another, but in the other area they were separated from one another, then we would probably consider the former urban area to be more segregated, since all minority members live within one single homogeneous ghetto, compared to the latter area, where they reside in minority neighborhoods that are scattered throughout the urban area."

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

3.3 Centralization

Generally speaking the centrality of a node in a network captures its well–connectedness. Depending on the specific context, it can for example be assessed using the notions of degree (the number of connections a node has), betweenness (determine the shortest paths between any two nodes and calculate then, for each node, to how many of these shortest paths it belongs to), or closeness (the mean geodesic distance between a node and all nodes reachable from it). In our case, the segregation index belongs to the class of eigenvector–based centrality measures and, in particular, it is related to the PageRank index of Brin and Page (1998) that underlies Google's search engine.[10]

The main idea of the PageRank vector is that a webpage is more important when it is linked by other relevant pages. This idea is formalized with a finite Markov chain according to which an individual "surfs" randomly over the web. The entries of the PageRank vector are then just the stationary probabilities of this process. More formally, an individual starts at any node (webpage) and surfs the web randomly according to the (column–stochastic) matrix $\mathbf{S}$ that represents the actual links of the web. Additionally, at any point in time, the surfer is "teleported" to a different page (even if the current page i is not directly linked with it) with probability $1-\beta$, while she continues her random–walk by crossing links normally following $\mathbf{S}^{\mathsf{T}}$ with probability β. In case of teleportation, the surfer is taken to page j with probability r_j. The vector $\mathbf{r} = (r_i)_{i\in M}$ is the *personalizaction vector* of the PageRank index summarizing the teleporting mechanism. This process defines a Markov chain whose stationary probabilities are gathered in the PageRank vector, the principal eigenvector of the matrix $(1-\beta)\mathbf{r}\mathbf{1}^{\mathsf{T}} + \beta\mathbf{S}$.

This notion of centrality can be straightforwardly adapted to our framework: An individual from group $g \in G$ randomly walks in the city according to the transition matrix $\mathbf{P}$ (that is, $\mathbf{S} = \mathbf{P}^{\mathsf{T}}$) and is teleported to node j with probability $(1-\alpha)\, d_{g,j}$ (that is, $\beta = \alpha$ and $\mathbf{r} = \mathbf{d}_g$).

Definition 3. Given the city $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ and the continuation probability $\alpha \in [0,1)$, the *PageRank vector* $\mathbf{w}_g$ *of group* $g \in G$ is the principal eigenvector of

[10] See Wasserman and Faust (1994) for an excellent overview of centrality measures.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

the matrix $(1-\alpha)\mathbf{d}_g\mathbf{1}^\mathsf{T} + \alpha\mathbf{P}^\mathsf{T}$.

Note that $\mathbf{w}_g$ captures the centrality of the nodes *for group g*. It is group–dependent because the personalization vector $\mathbf{r}$ is set equal to the density vector $\mathbf{d}_g$ of group g. Thus, in the PageRank model, $w_{g,i}$ is the stationary probability of neighborhood i when teleportation takes place according to $\mathbf{d}_g$, and it accounts for the reachability of neighborhood i by members of group g. Our next proposition is the dual of Proposition 1.

Proposition 3. *The segregation index of group g is equal to* $\sigma_g(C,\alpha) = \mathbf{w}_g^\mathsf{T}\mathbf{P}\mathbf{c}_g$.

Proof. See the corresponding section. □

Proposition 3 shows that individuals who are more central in the sense of PageRank contribute more to the overall segregation of their group because they are easier to reach by other group members. Also, if $p_{i,j} = (a_{i,j}\, n_j)/(\sum_{k\in M} a_{i,k}\, n_k)$ for all $i, j \in M$ (*i.e.,* transition probabilities are proportional to the sizes of the neighborhoods as in Proposition 2), then $\sigma_g(C,\alpha) = \mathbf{w}_g^\mathsf{T}\mathbf{h}_g$, where

$$h_{g,i} = \sum_{j\in M} \frac{n_{g,j}\, a_{i,j}}{\sum_{k\in M} n_k\, a_{i,k}}$$

is the homophily index of neighborhood $i \in M$ for group $g \in G$. This equation highlights the interaction between the position of a neighborhood in the network (centrality) and its clustering (homophily): In the computation of $\sigma_g(C,\alpha)$, the homophily index $h_{g,i}$ of neighborhood $i \in M$ is weighted by its PageRank index $w_{g,i}$. Note also that $\mathbf{w}_g = \mathbf{d}_g$ if the network is empty ($\mathbf{P} = \mathbf{I}$) or if there is no spatial mobility beyond one's own neighborhood ($\alpha = 0$). In the first case, σ_g reduces to the isolation index (because $h_{g,i}$ is then equal to $c_{g,i}$). In the second case, σ_g becomes equal to the homophily index. But if network effects are present ($\mathbf{P} \neq \mathbf{I}$ and $\alpha > 0$), groups that are located more central in the network everything else equal have a higher possibility of intra–group interactions and, as a consequence, a higher segregation index.[11] This is made visible with the help

[11]There is a vast literature that recognizes the importance of centrality in determining socio-economic outcomes. Examples include eduction (Calvo-Armengol et al. 2009),

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

of the following example.

Example. Consider the city C depicted in Figure 3.

[Include Figure 3 about here]

There are eleven individuals belonging to three ethnic groups (blacks, whites, and squares). There is exactly one individual in each neighborhood. The particularity of the network structure is that the blacks and the whites are allocated in a very similar way. Every black is connected to two other blacks plus herself, one white, and one square. Every white is also connected to two other whites plus herself, one black, and one square. The important difference is that all blacks are connected to the same square (individual 11), while two whites are connected to square 9 and two whites to square 10. Square 11 is thus the most central individual of her group, which implies that the blacks are relatively more central.

Since $I_{blacks}(C) = I_{whites}(C) = 1$ and $h_{blacks}(C) = h_{whites}(C) = 3/5$, both groups are equally isolated and clustered. Hence, $\sigma_g(C, \alpha)$ incorporates the difference between the two groups entirely through the PageRank vectors. Straightforward computations yield that $\sigma_{blacks}(C, 0.85) = 0.48 > 0.47 = \sigma_{whites}(C, 0.85)$.[12] □

We complete this part of the analysis with an important comment regarding the implementability of the segregation index in real–life applications. The computation of σ_g is simple and very fast, even for large matrices. One can apply first a simple iterative power method to compute the PageRank vector $\mathbf{w}_g$, and then Proposition 3 to compute the

criminal behavior (Haynie 2001), worker's performance (Mehra et al. 2001), power in organizations (Brass 1984), and the formation and performance of R&D networks (Boje and Whetten 1981, Powell et al. 1996 and Uzzi 1997).

[12]Section 4 shows that the Spectral Segregation Index (SSI) does not relate directly to the dimension of centralization because it only considers the network of within–group interactions. In the example, it misses that the blacks are connected to a more central square than the whites.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

product $\mathbf{w}_g^{\mathsf{T}}\mathbf{P}\,\mathbf{c}_g$. To calculate $\mathbf{w}_g$, one departs from a *randomly* chosen probability vector $\mathbf{w}_g^{(0)}$ (that is, $w_{g,i}^{(0)} \geq 0$ for all $i \in M$ and $\sum_{i\in M} w_{g,i}^{(0)} = 1$) and then iterates by calculating $\mathbf{w}_g^{(t+1)} = (1-\alpha)\mathbf{d}_g + \alpha\mathbf{P}^{\mathsf{T}}\mathbf{w}_g^{(t)}$ for all $t = 0, 1, 2, \ldots$. The asymptotic convergence rate of this iterative power method is equal to α, see Langville and Meyer (2006). For example, if $\alpha = 0.85$, $\alpha^{50} \approx 0.00296$; that is, one expects at least 2-3 places of accuracy after only 50 iterations. [13] We include a description of the algorithm for *Mathematica* and examples in the web appendix.

3.4 Evenness

Evenness refers to how equal the individuals of a given group are spread over the city. Formally, we say that the distribution of group g is *even* if for all $i \in M$, $d_{g,i} = n_i/n$, or equivalently, $c_{g,i} = n_g/n$. In this case, $\bar{\sigma}_g(C,\alpha) = (n_g/n)^{-1}\mathbf{d}_g^{\mathsf{T}}\mathbf{Q}\mathbf{c}_g = \mathbf{d}_g^{\mathsf{T}}\mathbf{Q}\mathbf{1} = \mathbf{d}_g^{\mathsf{T}}\mathbf{1} = 1$ independently of the network structure. This calculation shows that we can use a unit-normalized segregation index as a benchmark level in the sense that uneven distributions will lead to a higher or lower normalized segregation. If $\bar{\sigma}_g(C,\alpha) > 1$, group g is on average overrepresented in neighborhoods. It is underrepresented whenever $\bar{\sigma}_g(C,\alpha) < 1$. We show next that no group can on average be underrepresented if the *isolation index* is applied. Hence, in this case, there is an immediate connection between uneven distributions and a high segregation when the network does not matter.

[13] Note that $\mathbf{P}$ is a sparse matrix meaning that the number of non–zero entries in each row is negligible. Thus, for large datasets (over 5,000 nodes), it is key to work with special data structures representing sparse matrices (*sparse* function in Matlab and *SparseArray* function in Mathematica). Otherwise, one spends a considerable amount of memory space to store zeroes that do not affect matrix multiplications. Also, in order to compute σ_g, it is not necessary to determine all connected components even though the measure is component–separable by definition. Finally, σ_g has an additional computational advantage over the SSI: The convergence speed of the SSI can be very low compared to that of σ_g, which is high if α is well below 1.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Proposition 4. *For all cities* $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$*, all continuation probabilities* $\alpha \in [0,1)$*, and all groups* $g \in G$*,* $\bar{I}_g(C) \geq 1$*.*

Proof. See the corresponding section. □

Since individuals can interact in several neighborhoods as soon as the network matters, it becomes possible that a group is on average underrepresented (low segregation), even if its distribution is uneven. The following example illustrates this effect.

Example. Consider the city depicted in Figure 4.

[Include Figure 4 about here]

Using the node–to-node transition matrix $\mathbf{P}$ one finds that the probabilities that the random–walk terminates in a given neighborhood are equal to

$$\mathbf{d}_{blacks}^{\mathsf{T}}\mathbf{Q} = \frac{1}{12+2\alpha}(4+2\alpha, 4, 4) \quad \text{and} \quad \mathbf{d}_{whites}^{\mathsf{T}}\mathbf{Q} = \frac{1}{12+2\alpha}(6, 3+\alpha, 3+\alpha).$$

We see that the black individual living in the city center is more likely to terminate her walk in her area of residence. Similarly, the black center is also more likely to absorb a walk starting from the white edges. We then obtain that

$$\bar{\sigma}_{blacks}(C,\alpha) = \frac{12+6\alpha}{12+2\alpha} \geq 1 \quad \text{and} \quad \bar{\sigma}_{whites}(C,\alpha) = \frac{18+6\alpha}{24+4\alpha} < 1.$$

We stress that the normalized segregation of the whites is lower than one even though the distribution of the group is uneven. This highlights the idea that an uneven distribution can potentially make individuals hardly accessible by members of the same group. □

4 Discontinuity of the SI

In this part of the paper, it is shown that the SI proposed in Echenique and Fryer (2007) is not continuous in the strength of the social ties. Even though cities do not

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

experience abrupt changes in the short run, people may move within the city or change their pattern of social interactions, and we therefore think that a measure of segregation should behave in a stable way to these changes. Since the only difference in the primitives between Echenique and Fryer (2007) and our model is that they assume a one–to–one mapping between neighborhoods and individuals, it is assumed for the time being that $n_i = 1$ for all $i \in M$.[14]

Given the matrix $\mathbf{P}$, the SSI for group $g \in G$ is calculated by determining first the subgraph $\mathbf{P}^g$ of $\mathbf{P}$ that only contains interactions between members of g. All individuals belonging to a different group are eliminated from the network and, consequently, from the matrix $\mathbf{P}$. Such a subgraph $\mathbf{P}^g$ of $\mathbf{P}$ generally consists of more than one strongly connected component, so the next step is to calculate the SSI for each component separately. In particular, the SSI for group g in the (strongly connected) component $\mathbf{P}^{g,\gamma}$ of $\mathbf{P}^g$ — denoted by $SSI(\mathbf{P}^{g,\gamma})$ — is set equal to the spectral radius of the substochastic matrix $\mathbf{P}^{g,\gamma}$, the largest absolute value of the eigenvalues of $\mathbf{P}^{g,\gamma}$.[15] The segregation of a node (individual) i in this component is denoted by $SSI_i(\mathbf{P}^{g,\gamma})$, and it is the $i-$th entry of the principal eigenvector of the matrix $\mathbf{P}^{g,\gamma}$, normalized so that the vector average is $SSI(\mathbf{P}^{g,\gamma})$. The segregation of group g is a weighted average over the segregation levels of group g in all components of $\mathbf{P}^g$. In particular,

$$SSI(\mathbf{P}^g) = \sum_{\gamma} d_g^{\gamma} \cdot SSI(\mathbf{P}^{g,\gamma}),$$

where $d_g^{\gamma} = 1/n_g \cdot \sum_{i \in \mathbf{P}^{g,\gamma}} n_{g,i}$ is the fraction of individuals from group g that live in $\mathbf{P}^{g,\gamma}$.[16] Alternatively, the SSI of group g can also be expressed as the average over the

[14]The next section presents an extension of the SSI to a setting with multiple inhabitants in the same node.

[15]For the sake of simplicity, given any such matrix $\mathbf{P}^g$, we will also use $\mathbf{P}^g$ to refer to the network corresponding to $\mathbf{P}^g$. Formally, the subgraph $\mathbf{P}^{g,\gamma}$ is said to be a *strongly connected component* of the network represented by $\mathbf{P}^g$ if $\mathbf{P}^{g,\gamma}$ is a network of maximal size such that there is a path between any pair of its nodes.

[16]Proposition 13 in Echenique and Fryer (2007) shows that the SSI is approximately

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

individual segregation levels:

$$SSI(\mathbf{P}^g) = \frac{1}{n_g} \sum_{\gamma} \sum_{i \in \mathbf{P}^{g,\gamma}} SSI_i(\mathbf{P}^{g,\gamma}).$$

To see how the SSI is applied consider the panel on the left–hand side of Figure 5, which corresponds to the motivating example in Echenique and Fryer (2007). The society in City 1 is composed of two groups, blacks and whites. Each dot represents one individual. It is also assumed that individuals only interact with their horizontal and vertical neighbors. So, individual (A,1) spends 50% of her time with each (A,2) and (B,1) each. The subgraph $\mathbf{P}^{blacks}$ in City 1 consists therefore of two connected components $\mathbf{P}^{blacks,1}$ and $\mathbf{P}^{blacks,2}$.

[Include Figure 5 about here]

The SSI for the blacks is determined by taking a weighted average over the spectral radii of the two black connected components. The upper part of Table 1 shows that $SSI(\mathbf{P}^{blacks,1}) = 0.72$ and $SSI(\mathbf{P}^{blacks,2}) = 0.25$. Since 80% of the blacks reside in component 1 and 20% in component 2, the segregation of the blacks is $SSI(\mathbf{P}^{blacks}) = 0.8 \cdot 0.72 + 0.2 \cdot 0.25 = 0.63$.

[Include Table 1 about here]

proportional to the probability that two individuals from the same social group meet *without* leaving the subgroup network. Hence, the SSI is also related to a probability but, at the same time, differs substantially from the segregation index. Indeed, for σ_g it does not matter whether i meets somebody from a different group on his way to j, yet the walks undertaken according to the SSI are restricted to within group interactions. This important distinction can lead to substantial differences in applications; *i.e.*, our empirical analysis in the next section shows that the correlation between the segregation index and the SSI can be rather low.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

To see that the SSI is not continuous in the entries of $\mathbf{P}^{blacks}$ suppose that the black individuals (D,4) and (C,5) start communicating with each other 1% of their respective times. It is irrelevant which neighbor(s) receive now relative less attention from (D,4) and (C,5), for the sake of the example just suppose that this extra time is taken away from the white neighbors. This small change in the network leads to City 2 on the right–hand side of Figure 5, which has now a single component. As it can be seen in the lower part of Table 1, the effect of this change on the segregation of the blacks is substantial, $SSI(\mathbf{P}^{blacks})$ increases from 0.63 to 0.72. In particular, the spectral radius of the integrated component converges to the larger of the two spectral radii. Finally, one also sees that the segregation levels of individuals (B,5) and (C,5) are now very close to 0, whereas they were 0.25 in City 1.

More formally, $SSI(\mathbf{P}^g)$ is discontinuous in $\mathbf{P}^g$, as opposed to what is stated in Proposition 5 in Echenique and Fryer (2007). This becomes evident when two connected components of $\mathbf{P}^g$ are merged or when one component of $\mathbf{P}^g$ is split up.[17] The reason for the discontinuity is the component additivity attached to the measure: The SSI of a connected component $\mathbf{P}^{g,\gamma}$ of $\mathbf{P}^g$ is defined as the spectral radius of the submatrix $\mathbf{P}^{g,\gamma}$, but the spectral radius of $\mathbf{P}^g$ —which is continuous in $\mathbf{P}^g$— can generally not be averaged over its components. In fact,

$$\rho(\mathbf{P}^g) = \max_{\gamma} \rho(\mathbf{P}^{g,\gamma}),$$

where $\rho(\cdot)$ denotes the spectral radius. So, the property of component additivity sacrifices continuity. To restore the continuity of the SSI, one would therefore have to drop this component additivity. However, in this case, the segregation of group g would be equal to the largest (and not the average) segregation of all components of $\mathbf{P}^g$, and the segregation levels in all other components of $\mathbf{P}^g$ would be zero. So, if additivity is dropped, one would still need to explain why this asymmetric outcome is a reasonable measure of segregation.[18]

[17]However, we also note that problems still arise when $\mathbf{P}^g$ is connected but near disconnectedness. In that case, the SSI can still display very unstable behavior.

[18]Note that Proposition 5 on page 475 indicates continuity of $SSI(\mathbf{P}^g)$ in the entire

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

There are several circumstances that can trigger the discontinuity problem of the SSI. The reader should observe first that the disconnectedness of a *group* inside a social network causes the SSI to be discontinuous. Thus, even if a city or social network is connected, the assumption that *all* groups within this network are connected is likely to be violated. Second, if several individuals live in a node, the majority rule employed in Echenique and Fryer (2007) is a sensible choice to determine the representative individual of a census tract. However, one cannot exclude the possibility that the majority group in a node changes (even though the U.S. Census data revealed a considerable margin in most districts) and that this causes two or more smaller components to merge into one big component or vice versa, leading to the same discontinuity problems. Third, even if no particular representativeness rule has to be considered because there is only a single individual in each node, it could happen that one individual changes her location in the spatial network, thus changing the network structure of her group by merging previously separated components or separating previously integrated components. Hence, the movement of a single individual can provoke a significant variation in the segregation of her group. Finally, the definition of adjacent nodes in a spatial network is also likely to become problematic if the measure is not continuous. Echenique and Fryer (2007) reasonably assume that two census tracts are connected whenever their centroids are not more than one kilometer away from each other. Yet, a small change in this distance could potentially merge or separate components of some groups, triggering an abrupt change in the segregation of these groups. Even though it turns out that the application to the U.S. Census in Echenique and Fryer (2007) is robust with respect to this kind of variation in the network, one cannot exclude the possibility that difficulties appear in future applications with different datasets.

To conclude we point out that σ_g and $\bar{\sigma}_g$ are continuous in all their arguments. In particular, continuity with respect to α and $\mathbf{P}$ follows from the fact that $\mathbf{P}$ is stochastic and that $|\alpha| < 1$. Moreover, our measures are quite stable in applications since we are choosing values of α well below 1. Finally, the next proposition provides an upper

network $\mathbf{P}^g$, which is not true. Indeed, $SSI(\mathbf{P}^g)$ is continuous as long as $\mathbf{P}^g$ is connected.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

bound to the change in the segregation index when connections are modified. For that, given two cities $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ and $C' = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}'\rangle$ with a common distribution of individuals over nodes, let U be the subset of nodes whose connection change across the two cities; that is, $U = \{i \in M : p_{i,j} \neq p'_{i,j} \text{ for some } j \in M\}$.

Proposition 5. *The change in segregation index between* $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ *and* $C' = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}'\rangle$ *is bounded above by*

$$|\sigma_g(C', \alpha) - \sigma_g(C, \alpha)| \leq \max_{i\in M} \{c_{g,i}\} \frac{2\alpha}{1-\alpha} \sum_{j\in U} w_{g,j}.$$

Proof. This follows from Theorem 6.5.1 in Langville and Meyer (2006) and Proposition 3. □

5 Applications

5.1 Residential Segregation in Spain

We next apply the segregation index to the Spanish census tract data from January 2009. Among the developed countries Spain is particularly interesting to look at because the country attracted a lot of immigrants from many different parts of the world. During the boom years in the beginning of the century it was easy for the young South Americans to find a job in the construction or the (private) service sector because they were relatively cheaper to hire and had the advantage of speaking the same native language. But also immigrants from the Eastern European countries that recently entered the European Union or Africa were attracted by the vast job opportunities in the labor market. These combined effects led to the situation that more than 10 % of the 45 million residents in Spain in 2009 were foreigners.

[Include Table 2 about here]

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Table 2 presents the population shares of 27 nationalities in Spain.[19] It can be seen that the Romanians form the largest foreign group with 1.76 % of the total population followed by the Moroccans who account for 1.56 % of the residents. The high number Moroccans in Spain is of no surprise given their long tradition in the country. Next are Ecuador (0.93 %) and Colombia (0.65 %). The Nigerians form the smallest foreign group (0.09 %) for which we have data.

The Spanish territory is officially divided into 52 provinces, 47 of those are on the Iberian Peninsula. The remaining ones are the Balearic Islands in the Mediterranean Sea, the Canary Islands in the Atlantic Ocean (2 provinces), and the autonomous cities Ceuta and Melilla in North Africa. We abstain from incorporating Ceuta and Melilla in our analysis because they are too small — each of the two cities has only about 75,000 residents, and very few of them are non–Africans. A list of all provinces can be found in Table 3.

[Include Table 3 about here]

As of January 2009, the National Statistical Institute of Spain (INE) divides the 50 main provinces into a total of 35,757 census tracts. The mean number of residents per census tract is 1,268.98, the corresponding standard deviation is 625.74. As in Echenique and Fryer (2007), we use the geographical location of the centroids of the census tracts to derive the network. In particular, we define two census tracts to be connected if the distance between their centroids is less than 400 meters. A node has then on average 4.84 connections. The corresponding standard deviation is 5.48. Also, there are 16,908

[19] Our analysis excludes immigrants from Britain and Germany who amount to more than 1% of the Spanish population because these groups can be expected to be less of a "social problem" than poor migrants from Africa. Many of them are pensioners who bought second homes in the coastal area of the Mediterranean Sea and it is not clear how much time they actually spend in these residences. We are very grateful to an anonymous referee for pointing out this effect.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

isolated nodes.[20] In order to make consistent comparisons with the homophily index (see Proposition 2), we set the transition probabilities equal to

$$p_{i,j} = \frac{a_{i,j}\, n_j}{\sum_{k \in M} a_{i,k}\, n_k},$$

which means than the transition probability to node j is proportional to the total population in census tract j. Observe that our results remain identical if we choose $p_{i,j} = 1/a_i$.

5.1.1 Results

We first report on the normalized segregation index $\bar{\sigma}_g(C, 0.85)$ for all provinces; that is, C is defined by the borders of the provinces.[21] Recall that only the normalized measure is invariant to the relative group sizes, which is a crucial property if one wants to compare the segregation of different nationalities.

[Include Figure 6 about here]

We can see from Figure 6 that the segregation index is rather low in some parts of the autonomous communities of Galicia (provinces 15, 27, and 32), Asturias (province

[20] Even though the network is constructed from geographical positioning data, we nevertheless proxy true social interactions. This is so because geographical proximity is correlated to close relationships, despite the fact that the new era of information has made weak social links less dependent on physical distance (see, Goldenberg and Levy 2009 and Mok et al. 2010). Also, our qualitative results remain identical for radii of 1,000 and 1,500 meters.

[21] The numerical results for $\alpha \in \{0.00; 0.50; 0.85; 0.99\}$ and the normalized isolation index for all groups and all provinces are detailed in the online appendix. Setting the continuation probability α equal to 0.85 is a prominent choice in applications related to PageRank because in this way, relevant network effects are introduced, and the analysis is not compromised with computation inefficiencies or ill–conditioned problems that can appear if α grows closer to 1.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

33), and the Basque Country (provinces 01, 20, and 48) that are all in northern part of Spain touching the Cantabrian Sea and the Atlantic Ocean. Also, some parts of Andalusia such as Cadiz, Córdoba, Jaén, and Seville (provinces 11, 14, 23, and 41) have a segregation index that lies between 1.00 and 1.10. A substantial portion of the country has a segregation index between 1.11 and 1.20. The central region close to the capital Madrid (province 28), the second biggest city of Barcelona (province 08), the coastal sides of Alicante (province 03) and Murcia (province 30), the Balearic and the Canarian islands (provinces 07, 35, and 38), as well as Salamanca and Badajoz that touch the Portuguese border (provinces 10 and 37) have a moderate segregation index. The segregation index takes high values within other provinces of the autonomous community of Catalonia in northeastern part of the country (for example provinces 17 and 43) and Teruel (province 44). It is highest in Almería (province 04).

Figure 6 uncovers how segregated different parts of the country are, but so far we have not analyzed which groups are causing the results we see. To say it in different words, we still have to investigate which groups are, on average, more segregated. We detail on this by calculating the normalized segregation index for each of the 27 groups in the whole country; that is, C consists now of all 35,757 census tracts. Again, $\alpha = 0.85$.

[Include Figure 7 about here]

It can be seen in Figure 7 that the Spanish are by far the least segregated group. Since the normalized segregation index for group g takes a value of one if $c_{g,i} = n_g/n$ for all nodes, this finding supports the interpretation that the Spanish are very evenly distributed over all census tracts. Also, since the normalized segregation index is considerably larger than one for all other nationalities, the foreign groups show the general tendency to stick together.

The by far most segregated immigrants are the Pakistanis, $\bar{\sigma}_{PK}(\text{Spain}, 0.85) = 23.56$, which implies that interactions within this group occur about 23 times more often than if their members were uniformly spread. Immigrants from Russia and African countries like

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Nigeria and Senegal have a normalized segregation index between 10–15. The immigrants of the South American countries are the most integrated ones, their segregation index ranges from 2.28 for the Colombians to 5.99 for the Uruguayans. Finally, we can also say why Almería is very segregated. In Almería, 4.3 % of the population are Romanians (their segregation index in this province is 2.67) and 5.9 % are Moroccans (their segregation index in this province is 3.94). Consequently, the segregation in Almería is high because two of the largest groups of immigrants are overrepresented in this province.

5.1.2 Network Effects

One important question at this point is whether the incorporation of the network as an additional dimension adds to our understanding of segregation. To investigate this, one has to compare the normalized segregation index when $\alpha = 0.85$ with the situation when $\mathbf{P} = \mathbf{I}$ and the measure reduces to the normalized isolation index. This comparison will in particular reveal which nationalities cluster in neighborhoods close to each other. To see this, imagine that the isolation index of group g is 1 (no individual outside of g is living in a census tract with members of g). There are now two opposite scenarios to be investigated as the network is introduced: The census tracts with inhabitants of group g are spread evenly across the city or they are clustered in one particular area. If the members of g live in evenly spread census tracts, new connections to neighboring census tracts will increase the interactions with members outside of g and the segregation will decrease. If, on the other hand, g forms a clustered group, the segregation of the group will also decrease, but the effect will now be smaller because the new interactions are more likely to be intra–group, as the neighboring nodes are populated by group g as well. Hence, in this example the decrease in segregation due to an increase in α is lower for clustered groups.

[Include Figure 8 about here]

Figure 8 calculates the degree by which the measure decreases when the network is considered; that is, we calculate $1 - \bar{\sigma}_g(C, 0.85)/\bar{I}_s(C)$. One sees that the network effect is

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

substantial for some groups but negligible for others. The clearest effect can be identified for the Pakistanis and the Nigerians. Considering the actual network $\mathbf{P}$ with $\alpha = 0.85$ reduces the normalized isolation index of these groups by 40–60 %. However, it is not only the relative changes that matter. For example, the Russians have the fourth highest normalized isolation index ($\bar{I}_R(C) = 13.26$), while they are the second most segregated group when the network matters and $\alpha = 0.85$. Actually, the network effects for the Russians are rather low because the segregation index does not change much when the network is incorporated. As mentioned before, the differences in the change of the measure is due to clustering. The Russians are overrepresented in the coastal sides of Alicante (province 03), Almería (province 04), Girona (province 17), Málaga (province 29), and Tarragona (province 43), whereas the Pakistanis are spread over many cities, even though they live very concentrated within these cities.[22]

[Include Figure 9 about here]

Finally, Figure 9 analyzes which part of the network effect is due to a change from $\alpha = 0$ (the normalized homophily index) to $\alpha = 0.85$. In fact, the ratio between $\bar{\sigma}_g(C, 0) - \bar{\sigma}_g(C, 0.85)$ and $\bar{I}_g(C) - \bar{\sigma}_g(C, 0.85)$ shows which percentage of the clustering is due to connections beyond immediate neighbors. The data clearly reveals that the normalized homophily index already captures most of the network effects. Nevertheless, there are several group of immigrants for which higher-order relations explain between 20 % and 40 % of the overall network effect (*e.g.*, French, Moroccans, Nigerians, Chinese, and Pakistanis). We also note that almost 40 % of the clustering of Spanish residents is explained by higher-order relations.

[22]The British and the Germans that were excluded from our study because of the before mentioned reasons show a very similar residence pattern to that of the Russians. They live highly segregated in few areas, in particular in Alicante and the Balearic Islands, so that their network effect is rather small. Actually, the British and the Germans are two most segregated groups if $\alpha = 0.85$ (the Pakistanis remain to be the most segregated group if the normalized segregation index is applied).

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

5.1.3 Correlation with the SSI

Next, we calculate the correlation between the SSI and the segregation index. Recall that the basic idea behind the SSI is that individuals interact with their neighbors. If one node corresponds to one individual, the set of neighbors naturally includes all nodes an individual is connected to. We allow for multiple individuals in the same neighborhoods, but the two frameworks become comparable if one defines the set of neighbors as all those individuals that can be reached within one step. The within group substochastic interaction matrix $\mathbf{P}^g$ for group g is thus defined as

$$p^g_{i,j} = \frac{a_{i,j}\, n_{g,j}}{\sum_{k \in M} a_{i,k}\, n_k}.$$

Hence, an individual from group g living in neighborhood i interacts with all $g-$members that can be reached within one step with equal probability. As before, the SSI of group g can then be defined as

$$SSI(\mathbf{P}^g) = \sum_{\gamma} d^{\gamma}_g \cdot SSI(\mathbf{P}^{g,\gamma}).$$

Since the spectral radius for each connected component can be set equal to the weighted average of the corresponding eigenvector, we obtain that

$$SSI(\mathbf{P}^g) = \sum_{\gamma} d^{\gamma}_g \sum_{i \in \mathbf{P}^{g,\gamma}} d^{\gamma}_{g,i} \cdot SSI_i(\mathbf{P}^{g,\gamma}),$$

where $d^{\gamma}_{g,i} = n_{g,i} / \sum_{j \in \mathbf{P}^{g,\gamma}} n_{g,j}$ is the fraction of individuals of group g from component γ who live in neighborhood i. Rewriting this equation as

$$SSI(\mathbf{P}^g) = \sum_{\gamma} \sum_{i \in \mathbf{P}^{g,\gamma}} d^{\gamma}_g \cdot d^{\gamma}_{g,i} \cdot SSI_i(\mathbf{P}^{g,\gamma}) = \sum_{i \in M} d_{g,i} \cdot SSI_i(\mathbf{P}^{g,\gamma(i)})$$

shows that the SSI can be envisioned as a weighted average over all neighborhoods. Since the SSI is not invariant to group sizes, we have to compare it to the size dependent version of the segregation index, which according to Proposition 1 is equal to

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

$\sigma_g(C,\alpha) = \sum_{i\in M} d_{g,i}\, v_{g,i}$. Consequently, we proceed by calculating the correlation between the vectors $\mathbf{v}_g = (v_{g,i})_{i\in M}$ and $(SSI_i(\mathbf{P}^{g,\gamma(i)}))_{i\in M}$ of size 35,757 corresponding to Spain as a whole.

[Include Figure 10 about here]

Figure 10 shows a positive correlation between the SSI and the segregation index for all nationalities. For several groups, the correlation ranges between 0.20 and 0.40. The correlation is rather small for immigrants from Morocco, Chile and Bolivia. Consequently, the data reveals that the SSI and the segregation are rather different in their structure even though they both relate to the probability that two members from the same group meet. This is because the measures use a different underlying network structure to determine this kind of probability (see Footnote 16).

5.2 Segregation in Scientific Publications

5.2.1 Model

Using citation data from scientific publications, we now show that the model of residential segregation can be readily applied to a wide range of socio-economic questions. Let the set of nodes M now be equal to the set of scientific journals. The set of all *fields* G partitions M. The number of articles published by journal $i \in M$ constitutes the population n_i at node i. Since each journal is assumed to belong to exactly one field, $c_{g,i} = 1$ if journal i belongs to field $g \in G$, and $c_{g,i} = 0$ otherwise. The total number of articles published by journals from field g is then equal to $n_g = \sum_{i\in M} c_{g,i}\, n_i$. Similarly, $d_{g,i} = (c_{g,i} \cdot n_i)/n_g$ is the fraction of all articles from journals belonging to field $g \in G$ that are published by journal i. We can see that $d_{g,i} = 0$ whenever journal i does not belong to field g.

Let $l_{i,j}$ be the number of citations from articles in journal i to articles in journal j. The total number of articles cited by articles in journal i is denoted by $l_i = \sum_{j\in M} l_{i,j}$. Setting $p_{i,j} = l_{i,j}/l_i$ induces the stochastic matrix $\mathbf{P}$ with the following interpretation: Imagine a researcher who is currently reading an article from journal $i \in M$. The matrix $\mathbf{P}$ then

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

indicates which journal the researcher is going to read next (following some citation from the current article).[23] If the probability that the random–walk continues after the first step is $0 \leq \alpha < 1$ — the interpretation of α is akin to that in the model of residential segregation —, the matrix $\mathbf{Q} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{P}$ can be derived in exactly the same way as before. In particular, $q_{i,j}$ corresponds to the probability that the researcher stops at journal j given that her random–walk started at journal i.

Given the tuple $C = \langle N, M, G, (n_{g,i})_{g\in G, i\in M}, \mathbf{P}\rangle$ and the continuation probability $\alpha \in [0,1)$, the segregation index $\sigma_g(C,\alpha)$ of field $g \in G$ measures the probability that a researcher who starts reading a randomly chosen article from a journal from field g (that is, each journal i from field g is taken with probability $d_{g,i}$) and follows the citations therein, stops reading an article from a journal that also belongs to field $g \in G$. Formally, by Proposition 3, $\sigma_g(C,\alpha) = \mathbf{w}_g^{\mathsf{T}}\mathbf{P}\,\mathbf{c}_g$, where $\mathbf{w}_g$ is the PageRank vector of field g.

5.2.2 Results

We use citation data from the 2010 Journal Citation Reports for a five year window. It includes citations made in 2010 to articles published in 2005–2009. The set M consists of 140 leading Economics journals according to the measures of Impact Factor™ and Article Influence™. We classify the journals in 18 fields following the CEFAFE–EU ranking from the University of Évora, Portugal (see, Table 4).[24]

[Include Table 4 about here]

For each of the 18 fields, we compute both the homophily index $h_g(C) = \sigma_g(C,0)$, which indicates the percentage of citations that stay within a field, and the segregation index $\sigma_g(C,0.85)$. The results are presented in Table 5.

[23] Modelling citation patterns using transition matrices is standard in the literature. See, for instance, Palacios-Huerta and Volij (2004), and impact measures like Eigenfactor™, Article Influence™ and SCImago Journal Rank.

[24] http://www.cefage.uevora.pt/en/links/revistas_cientificas_rankings.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

[**Include Table 5 about here**]

One can see that the size dependent measures $h_g(C)$ and $\sigma_g(C, 0.85)$ yield very similar results. In fact, the set of the five most segregated fields is identical in both cases, and the corresponding rank correlation is very high (0.942). Also, these two rankings resemble very much the ranking that one obtains if the fields are simply ordered according to their relatives sizes. One should look at the normalized measures to filter out this effect.

The last two columns show that the set of the five most segregated fields is still very much the same for both indices, however there are important changes with respect to the non–normalized measures. For the non–normalized measures, General Economic, which contains all general interest journals and is with 33 journals by far the biggest field, is among the most segregated fields. Yet, it is far less segregated for the normalized measures, which intuitively makes sense. Similarly, Economic History is not much segregated for the size dependent measures but by far the most segregated field if the normalized measures are applied. Also note that the rank correlation between the normalized homophily index and the normalized segregation index for $\alpha = 0.85$ is 0.554. This shows that there is some clustering beyond immediate neighbors that is explained by the segregation index σ: Even if a journal from field g does not cite a journal from the same field, it may cite journals from other fields that usually cite journals from g.

This type of higher order clustering is exactly captured by the difference between $\bar{h}_g(C)$ and $\bar{\sigma}_g(C, 0.85)$, and it is this difference that allows us to uncover some interesting citation patterns. For example, concentrating on the four small fields with a very low normalized homophily index — AccAud, HRM, Sector, and Socio —, one sees that $\bar{\sigma}_g(C, 0.85) - \bar{h}_g(C)$ is large for Sector and AccAud but low for HRM and Socio. Hence, while neither of these four fields has many self citations, there are cycles for Sector and AccAud (these journals cite journals from other fields that then cite back) that do not occur to this large extent for HRM and Socio. Also, these are the only four fields for which the segregation raises as the expected walk length increases, which makes us conclude that these kind of cycles play only a minor/no importance in the other fields.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Among the fields that reduce their segregation when indirect citations are taken into account, General Economics (GenEcon) and Urban, Spatial, Regional, and Transport Economics (UrbSpaReg) deserve some special attention. While GenEcon does not show any network effects (the normalized fraction of field–self–citations is in both cases 1.6 times larger than that obtained from the uniform distribution), UrbSpaReg turns out to have a lot of field–self–citations (roughly 11 times more if compared to the uniform distribution), yet this fraction is reduced by 75% when indirect citations are considered as well. So, journals from other fields that are cited by UrbSpaReg do not have the tendency to cite back. This is to say that UrbSpaReg has a rather high homophily index but a normalized segregation index when $\alpha = 0.85$ that compares to that of many other fields.

6 A Model of Time Investment

We now adapt the model of Currarini et al. (2009) to show that the segregation index σ_g is directly related to the equilibrium outcome of a game in which individuals have to decide how to spend their time.[25] We depart from the spatial network defined through the matrix $\mathbf{A}$ on the set of nodes $M = \{1, 2, \ldots, m\}$. We assume that $\mathbf{A}$ is symmetric and that $a_i > 0$. There is a set of groups/cultures G. At each moment in time, a continuum of individuals of mass $n_{g,i}$ of group $g \in G$ arrives (is born) in neighborhood $i \in M$. We say that an individual is an $i-$*citizen* if she is born in neighborhood i, and a $g-$*member* if she belongs to group g.

Each individual decides how much time to invest in meeting other cultures. We consider an unbiased payoff function that does *not* depend on the type of the individual so that we can identify how the spatial setting alone shapes separation in equilibrium. Thus, let $t_{g,i} \geq 0$ be the time invested by an $i-$citizen in culture $g \in G$, independently of her group. The total time invested by an $i-$citizen is $t_i = \sum_{g \in G} t_{g,i}$. It is the time an

[25] Alternatively, one could also follow an axiomatic approach to rationalize the measure. See, Palacios-Huerta and Volij (2004) and Frankel and Volij (2010) for related examples.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

individual from neighborhood i spends in the matching process, interacting with other citizens who also invest time. After time t_i expires, the $i-$citizen exits the process.

Meetings between individuals are only possible if they live in neighboring census tracts. In particular, an $i-$citizen meets a $j-$citizen with probability $p_{i,j} = a_{i,j}/a_i$; that is, the transition probability is uniform. For all neighborhoods $i, j \in M$ and all groups $g, s, u \in G$, we write $[(i,g) \to (j,s)]^u$ to denote the event in which an $i-$citizen who invests in culture g meets a $j-$citizen who is an $u-$member and invests in culture s. Note that the neighborhoods i and j, or the groups g, s and u may coincide or not. A *mutual–interest–based* meeting of an $i-$citizen with a $j-$citizen (I–meeting from now on) is a meeting of the form $[(i,g) \to (j,g)]^\bullet$; that is, an $i-$citizen meets a $j-$citizen, and they are investing in the same culture. Let $\tau_{i,j}^g$ be the total time an $i-$citizen spends in I–meetings related to culture g with $j-$citizens. A *cultural* meeting of an $i-$citizen with a $j-$citizen (C–meeting from now on) is a meeting of the form $[(i,g) \to (j,\bullet)]^g$; that is, an $i-$citizen who invests in culture g meets a $j-$citizen who is a $g-$member. Let $\bar{\tau}_{i,j}^g$ be the total time an $i-$citizen spends in C–meetings with $j-$citizens who are $g-$members. Finally, note that the meeting $[(i,g) \to (j,g)]^g$ is both I and C.

We assume that individuals only benefit from meetings that are I or C. First, the $i-$citizen gets $A_1 \geq 0$ for each unit of time in the I–meeting $[(i,g) \to (j,g)]^\bullet$. Second, each unit of time in the C–meeting $[(i,g) \to (j,\bullet)]^g$ produces a benefit of $A_0 > 0$ for the $i-$citizen. In case a meeting is both I and C, the $i-$citizen gets $A_0 + A_1$ per unit of time in the meeting. The total benefits of an $i-$citizen are then equal to

$$B_i\left((t_{g,j})_{g\in G, j\in M}\right) = \sum_{g\in G}\sum_{j\in M}\left(A_0\,\bar{\tau}_{i,j}^g + A_1\,\tau_{i,j}^g\right).$$

Costs are supposed to be additively separable with respect to the time spent in the different cultures. Also, they are identical for all cultures and independent of the group:

$$C_i((t_{g,i})_{g\in G}) = \frac{D}{2}\sum_{g\in G} t_{g,i}^2,$$

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

where $D > 0$. The utility of an $i-$citizen is then

$$U_i\left((t_{g,j})_{g\in G, j\in M}\right) = \sum_{g\in G}\left(A_o \sum_{j\in M} \bar{\tau}^g_{i,j} + A_1 \sum_{j\in M} \tau^g_{i,j}\right) - \frac{D}{2}\sum_{g\in G} t^2_{g,i}.$$

One sees that the incentives to express interest in a particular culture g are determined by three factors: The likelihood of meeting an $g-$member (the *compositional effect* that produces a payoff of A_0), the chance of finding other citizens who are also interested in culture g (the *complementarity effect* that produces a payoff of A_1), and the parameter D of the cost function. Next, and in order to be able to calculate the total time spent in different meetings, we have to make some assumptions on the probability that a particular meeting occurs. We follow Currarini et al. (2009) and impose an *unbiased* process. Formally, the probability $\rho^g_{i,j}$ that an $i-$citizen has an $I-$meeting related to culture g with a $j-$citizen is

$$\rho^g_{i,j} \equiv p_{i,j}\,\frac{p_{j,i}\,(n_j\,t_{g,j})}{p_{j,i}\sum\limits_{k\in G} n_j\,t_{k,j}} = p_{i,j}\,\frac{t_{g,j}}{t_j};$$

that is, the probability is determined by the fraction of time $j-$citizens spend in culture g. Similarly, the probability $\bar{\rho}^g_{i,j}$ that an $i-$citizen has a $C-$meeting with an $g-$member from neighborhood j is driven by the size of group g at j:

$$\bar{\rho}^g_{i,j} \equiv p_{i,j}\,\frac{p_{j,i}\,(n_{g,j}\,t_j)}{p_{j,i}\,(n_j\,t_j)} = p_{i,j}\,\frac{n_{g,j}}{n_j} = p_{i,j}\,c_{g,j}.$$

We then find that $\tau^g_{i,j} = \rho^g_{i,j}\cdot t_{g,i} = p_{i,j}\cdot t_{g,i}\cdot t_{g,j}\cdot 1/t_j$ and $\bar{\tau}^g_{i,j} = \bar{\rho}^g_{i,j}\cdot t_{g,i} = p_{i,j}\cdot t_{g,i}\cdot c_{g,j}$, respectively. Using this information, the utility function of an $i-$citizen becomes

$$U_i\left((t_{g,j})_{g\in G, j\in M}\right) = \sum_{g\in G}\left(A_o\,t_{g,i}\sum_{j\in M} p_{i,j}\,c_{g,j} + A_1\,t_{g,i}\sum_{j\in M} p_{i,j}\frac{t_{g,j}}{t_j}\right) - \frac{D}{2}\sum_{g\in G} t^2_{g,i}. \qquad (1)$$

We study the simultaneous move game with utilities given by (1) and where each $i-$citizen chooses the investment times $(t_{g,i})_{g\in G} \geq \mathbf{0}$. Given the investment times from all individuals, the resulting meetings are *balanced* if the following three conditions are satisfied:

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

- *Total balancedness:* The time $i-$citizens meet $j-$citizens is equal to the time $j-$citizens meet $i-$citizens.

- *I–balancedness:* The time $i-$citizens have $I-$meetings related to culture g with $j-$citizens is equal to the time $j-$citizens have $I-$meetings related to culture g with $i-$citizens.

- *C–balancedness:* The time $i-$citizens have $C-$meetings with $g-$members from node j is equal to the time that $g-$members from node j spend in these meetings with $i-$citizens.

Balancedness states that every $i-$citizen who invests t_i units of time has bilateral meetings during this time. Due to the unbiased setting, all three conditions are equivalent and reduce to $n_i \cdot p_{i,j} \cdot t_i = n_j \cdot p_{j,i} \cdot t_j$ (for details, see the proof of Theorem 1). It is also important to note that balancedness can only be ensured in equilibrium if we impose some additional restrictions on the structure of the network. For this reason, we assume from now on that for all $i, j \in M$, $n_i/n_j = a_i/a_j$; that is, in relative terms, the mass of newborns in neighborhood $i \in M$ is proportional to its number of neighbors a_i. Only then meetings with all neighbors are feasible. We nevertheless point out that this condition still gives us the freedom to choose any vectors of group concentrations $(c_g)_{g\in G}$.

We say that $(t^*_{g,i})_{i\in M, g\in G}$ is a *steady state equilibrium* if $(t^*_{g,i})_{i\in M,g\in G}$ is a Nash equilibrium of the simultaneous move game, meetings are balanced, and the inflow and outflow of individuals are equal at any given point in time. The following result states that in the unique steady state equilibrium, the average time invested by $g-$members in their own culture, $\bar{t}^g$, is proportional to the segregation index of group g.

Theorem 1. *For all groups $g \in G$,*

$$\bar{t}^g \equiv \frac{1}{n_g} \sum_{i\in M} n_{g,i}\, t^*_{g,i} = \frac{A_0 + A_1}{D}\, \sigma_g \left(C, \frac{A_1}{A_0 + A_1} \right).$$

Proof. See the corresponding section. □

Observe that the relative magnitude of A_1/A_0 directly relates to the parameter α.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

As α increases in the model of residential segregation, the complementarity effect in the game becomes stronger. We conclude with an example.

Example. Consider the city depicted in Figure 11.

[Include Figure 11 about here]

There are four neighborhoods and two groups. It is assumed that for all $i \in M$, $a_{i,i} = 1$. The vectors of group concentration are equal to $\mathbf{c}_{blacks}^{\mathsf{T}} = (0, 1, 1, 0)$ and $\mathbf{c}_{whites}^{\mathsf{T}} = (1, 0, 0, 1)$. Since the steady state equilibrium is only balanced if the condition $n_i/n_j = a_i/a_j$ is met for all $i, j \in M$, we consider the arrival rates $n_1 = n_2 = 3$, $n_3 = 4$, and $n_4 = 2$. Hence, in Figure 11 each dot indicates a continuum of individuals of mass 1. Also, $\mathbf{d}_{blacks}^{\mathsf{T}} = (0, 3/7, 4/7, 0)$ and $\mathbf{d}_{whites}^{\mathsf{T}} = (3/5, 0, 0, 2/5)$. If $A_0 = 1$, $A_1 = 3$, and $D = 2$, then $\alpha = 0.75$. In this case, the total time invested in the steady state is $t_i^* = 2$ for all $i \in M$, and the average equilibrium investments are twice the segregation indices: $\bar{t}^{blacks} = 2\,\sigma_{blacks}(C, 0.75) = 1.17$ and $\bar{t}^{whites} = 2\,\sigma_{whites}(C, 0.75) = 0.83$.[26]

Theorem 1 allows us to interpret the segregation index as the result of a game where individuals benefit from meetings. In the example, the blacks are more segregated because of the following game–theoretical interpretation. First, a black is more likely to meet blacks since the group is rather clustered ($h_{blacks} = 4/7$). Second, also a white is more likely to meet a black ($h_{whites} = 2/5$). This higher chance of meeting blacks for all individuals increases the compositional incentives to spend time with blacks. Third, complementarities further increase the incentives to spend time in black–related meetings. Finally, we also find that the whites spend more time in black–related than in white–related meetings (1.16 *vs.* 0.83). This is not surprising in a theoretical framework where individuals do not have any bias towards their own culture. □

[26]The latter equations can be verified with the help of the proof of Theorem 1.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

7 Conclusion

We have developed a new measure of residential segregation. In our theoretical model, the nodes of a network represent neighborhoods or census tracts, and links indicate which census tracts are adjacent in the urban space. It has also been assumed that multiple individuals from possibly different social groups can be located in the very same node. Using this information as the only primitive of our analysis, we have studied the following process: In the first period, an individual moves from her area of residence to a neighboring node. In each subsequent period, the individual advances from her current position in the network to an adjacent node with a given probability, or the process stops otherwise. The segregation index has then been defined as the probability that a randomly chosen individual from a given group meets an individual from the same social group in the node where her random–walk terminates.

We have shown in our analysis that the segregation index has several favorable aspects. *First*, the measure reduces to the isolation index in case the network is empty. Consequently, the segregation index can be interpreted as a natural generalization of the isolation index to spatial networks. *Second*, in case the network is not empty and the exogenous probability that the random–walk stops is one, the only interactions that are taken into account are those between nodes that are direct neighbors, and the segregation index reduces to the homophily index introduced by Currarini et al. (2009). *Third*, the segregation index incorporates the idea that social groups that are located closer to the (relative) city center are more segregated everything else equal. In particular, the segregation index turns out to be proportional to the PageRank index applied by Google to determine the importance of webpages in the World Wide Web. *Finally*, the segregation index is a continuous function in the social ties. Indeed, the SSI suggested by Echenique and Fryer (2007), who have been the first to study a graph–theoretical model to develop a measure of segregation, fails to satisfy this important criterion.

In our empirical analysis, we have studied first the Spanish 2009 census tract data. The main result in this regard is that the province of Almeria on the Mediterranean coast is the most segregated area, mainly because two of the more segregated groups, the

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Romanians and the Morrocans, are highly overrepresented in these regions. Immigrants from Latin America, on the other hand, turn out to be very much integrated, an effect that may be related to the fact that these groups speak Spanish natively. We have also seen that clustering effects are important to understand the segregation of some of the some smaller nationalities like the Pakistanis and immigrants from African countries like Nigeria and Senegal, but not for the European communities.

Afterwards, we have indicated that our model residential segregation can be easily adapted to answer a wide range of empirical questions. To illustrate this, we have calculated the segregation of 18 different fields in Economics using citation data from 140 scientific publications. The difference between the normalized homophily index and the normalized segregation index found in the data indicates a substantial amount of higher–order clustering: Journals from field g may not cite journals from the same field but journals from another field that cite journals from g.

Finally, in the last section of the paper, we rationalized the segregation index from a game–theoretical point of view. We considered a model in which individuals arrive continuously at the nodes of an exogenous network and have to decide how much time to invest in the different groups/cultures (stay in the matching process). Using an unbiased utility function, we found that the average amount of time an individual from group g invests in her own culture is directly proportional to segregation index.

References

Benabou, R. (1993). Workings of a city: Location, education, and production. *Quarterly Journal of Economics 108*, 619–652.

Boeri, T., M. De Philippis, E. Patacchini, and M. Pellizzari (2012). Moving to segregation: Evidence from 8 italian cities. *IZA Discussion Papers 6834*.

Boje, D. and D. Whetten (1981). Effects of organizational strategies and contextual constraints on centrality and attributions of influence in interorganizational networks. *Administrative Science Quarterly 26*, 378–395.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Borjas, G. (1995). Ethnicity, neighborhoods, and human-capital externalities. *American Economic Review 85*, 365–389.

Brass, D. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly 29*, 518–539.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 33*, 107–117.

Calvo-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *Review of Economic Studies 76*, 1239–1267.

Card, D. and J. Rothstein (2007). Racial segregation and the blackwhite test score gap. *Journal of Public Economics 91*, 2158–2184.

Currarini, S., M. Jackson, and P. Pin (2009). An economic model of friendship: homophily, minorities, and segregation. *Econometrica 77*, 1003–1043.

Cutler, D. and E. Glaeser (1997). Are ghettos good or bad? *Quarterly Journal of Economics 112*, 827–872.

Cutler, D., E. Glaeser, and J. Vigdor (2008). Is the melting pot still hot? Explaining the resurgence of immigrant segregation. *The Review of Economics and Statistics 90*, 478–497.

Echenique, F. and R. Fryer (2007). A measure of segregation based on social interactions. *Quarterly Journal of Economics 122*, 441–485.

Edin, P., P. Fredriksson, and O. Åslund (2003). Ethnic enclaves and the economic success of immigrants—evidence from a natural experiment. *Quarterly Journal of Economics 118*, 329–357.

Frankel, D. and O. Volij (2010). Measuring school segregation. *Journal of Economic Theory 146*, 1–38.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Gobillon, L., H. Selod, and Y. Zenou (2007). The mechanisms of spatial mismatch. *Urban Studies 44*, 2401–2427.

Goldenberg, J. and M. Levy (2009). Distance is not dead: Social interaction and geographical distance in the internet era. The Hebrew University, Jerusalem.

Haynie, D. (2001). Delinquent peers revisited: Does network structure matter? *American Journal of Sociology 106*, 1013–1057.

Ihlanfeldt, K. and D. Sjoquist (1998). The spatial mismatch hypothesis: A review of recent studies and their omplications for welfare reform. *Housing Policy Debate 9*, 849–892.

Jackson, M. and W. Rogers (2007). Meeting strangers and friends of friends: how random are socially generated networks? *American Economic Review 97*, 890–915.

Kling, J., J. Liebman, and L. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*, 83–119.

Langville, A. and C. Meyer (2006). Google's Pagerank and beyond: The science of search engine rankings. Princeton University Press, Princeton and Oxford.

Lieberson, S. (1981). An asymmetrical approach to segregation. In: Ethnic segregation in cities (eds. Peach, C. and Robinson, V.), 61–82. London, Croom Helm.

Massey, D. and N. Denton (1988). The dimensions of residential segregation. *Social Forces 67*, 281–315.

Mehra, A., M. Kilduff, and D. Brass (2001). The social networks of high and low self-monitors: implications for workplace performance. *Administrative Science Quarterly 46*, 121–146.

Mok, D., B. Wellman, and J. Carrasco (2010). Does distance still matter in connected lives? A pre- and post-internet comparison. *Urban Studies 47*, 2747–2784.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Palacios-Huerta, I. and O. Volij (2004). The measurement of intellectual influence. *Econometrica 72*, 963–977.

Powell, W., K. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly 41*, 116–145.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly 42*, 35–67.

Wasserman, S. and K. Faust (1994). Social network analysis. Methods and applications. Cambridge University Press, Cambridge.

Zenou, Y. (2009). How common is integration policy in europe? In: How unified is the European Union (Eds. Gustavsson, S., Oxelheim, L., and L. Pehrson). Springer, Heidelberg.

Proofs

Proof of Lemma 1

Consider $q_{i,j}$, which is the probability that a walk starting at node i ends at node j. With probability $(1-\alpha)$ the walk has length one, and it will end at node j with probability $p_{i,j}$. With probability α, the walk continues following i's neighbors and it will reach node j with probability $\sum_{k\in M} p_{i,k}\, q_{k,j}$. Thus, $q_{i,j} = (1-\alpha)p_{i,j} + \alpha \sum_{k\in M} p_{i,k}\, q_{k,j}$. Rewriting these equations in matrix form yields $\mathbf{Q} = (1-\alpha)\mathbf{P} + \alpha\mathbf{P}\mathbf{Q}$, which is equivalent to $\mathbf{Q} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{P}$, where the inverse $(\mathbf{I} - \alpha\mathbf{P})^{-1}$ is well-defined and non-negative because $\mathbf{P} \geq \mathbf{0}$ and $\alpha \in [0, 1)$.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Proof of Proposition 2

Note first that $\sigma_g(C,0) = \mathbf{d}_g^\mathsf{T}\mathbf{P}\mathbf{c}_g$. Hence, we show that $h_g(C) = \mathbf{d}_g^\mathsf{T}\mathbf{P}\mathbf{c}_g$. By definition of the homophily index,

$$\begin{aligned} h_g(C) &= \sum_{i\in M} d_{g,i} \sum_{j\in M} \tfrac{a_{i,j}\, n_{g,j}}{\sum_{k\in M} a_{i,k}\, n_k} \\ &= \sum_{i\in M} d_{g,i} \sum_{j\in M} \tfrac{a_{i,j} n_j}{\sum_{k\in M} a_{i,k} n_k}\, c_{g,j} \\ &= \sum_{i\in M} d_{g,i} \sum_{j\in M} p_{i,j}\, c_{g,j} \\ &= \mathbf{d}_g^\mathsf{T}\mathbf{P}\mathbf{c}_g. \end{aligned}$$

This concludes the proof of the proposition.

Proof of Proposition 3

By Lemma 1 and Proposition 1 , $\sigma_g(C,\alpha) = (\mathbf{d}_g^\mathsf{T}\mathbf{M})\mathbf{P}\mathbf{c}_g$, where $\mathbf{M} = (1-\alpha)(\mathbf{I}-\alpha\mathbf{P})^{-1}$. Hence, it only remains to be shown that $\mathbf{w}_g = \mathbf{M}^\mathsf{T}\mathbf{d}_g = (1-\alpha)(\mathbf{I}-\alpha\mathbf{P}^\mathsf{T})^{-1}\mathbf{d}_g$. By definition, $\mathbf{w}_g$ is equal to the principal eigenvector of the matrix $\mathbf{S} = (1-\alpha)\mathbf{d}_g\mathbf{1}^\mathsf{T} + \alpha\mathbf{P}^\mathsf{T}$. Since $\mathbf{S}$ is column stochastic, its spectral radius is 1 and the principal eigenvector $\mathbf{w}_g$ satisfies the equation $\mathbf{S}\mathbf{w}_g = \mathbf{w}_g$. Taking a $\mathbf{w}_g$ whose coordinates sum up to 1, we get that $(1-\alpha)\mathbf{d}_g + \alpha\mathbf{P}^\mathsf{T}\mathbf{w}_g = \mathbf{w}_g$. Solving this equation we get that $\mathbf{w}_g = (1-\alpha)(\mathbf{I}-\alpha\mathbf{P}^\mathsf{T})^{-1}\mathbf{d}_g$.

Proof of Proposition 4

Since $\mathbf{P} = \mathbf{I}$ for the isolation index $I_g(C)$ (see section 3.1), normalized segregation is

$$\bar{\sigma}_g(C,\alpha) = (n_g/n)^{-1} I_g(C) = (n_g/n)^{-1} \sum_{i\in M} d_{g,i}\, c_{g,i} = n \sum_{i\in M} \frac{1}{n_i}\, d_{g,i}^2.$$

The sum of squares is minimized if $d_{g,i} = n_i/n$ for all $i \in M$. Thus,

$$\bar{\sigma}_g(C,\alpha) \geq n \sum_{i\in M} \frac{1}{n_i}\, \frac{n_i^2}{n^2} = \sum_{i\in M} \frac{n_i}{n} = 1.$$

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

This concludes the proof of the proposition.

Proof of Theorem 1

The proof is organized as follows. First, we show existence of a steady state equilibrium and relate it to the segregation index. Afterwards, we establish the uniqueness of the equilibrium. Finally, we prove matching balancedness.

1) Existence

Take any time investments $(t_{g,i})_{i\in M, g\in G}$. Using the utility function (1), one sees that the marginal utility of the time investment $t_{g,i}$ in culture g of an individual born in neighborhood i equals

$$\frac{\partial U_i}{\partial t_{g,i}} = A_0 \sum_{j\in M} p_{i,j} c_{g,j} + A_1 \sum_{j\in M} p_{i,j} \frac{t_{g,j}}{t_j} - D\, t_{g,i}. \tag{2}$$

In order to prove existence, we derive an equilibrium where $\partial U_i/\partial t_{g,i} = 0$. We will show first that the total time efforts are equal for all individuals; that is, in this equilibrium, individuals will only differ in the way they distribute their time over all cultures. Summing over all groups and recalling that $\sum_{g\in G} c_{g,j} = 1$ for all $j \in M$ and $\sum_{j\in M} p_{i,j} = 1$, we have that $A_0 + A_1 - D\, t_i = 0$. It follows then that for all $i \in M$,

$$t_i^* = \frac{A_0 + A_1}{D}. \tag{3}$$

Now, we are going to determine the time spent in $g-$related meetings. Let $\mathbf{t}_g = (t_{g,i})_{i\in M}$ be the vector of times spent in culture $g \in G$. Using equation (3), the first order conditions can be rewritten in matrix form as

$$\frac{A_0}{D}\, \mathbf{P}\, \mathbf{c}_g + \frac{A_1}{A_0 + A_1}\, \mathbf{P}\, \mathbf{t}_g - \mathbf{t}_g = \mathbf{0}.$$

It follows that

$$\mathbf{t}_g^* = \frac{A_0}{D} \left(\mathbf{I} - \frac{A_1}{A_0 + A_1} \mathbf{P} \right)^{-1} \mathbf{P}\, \mathbf{c}_g \geq \mathbf{0}.$$

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

The non–negativity of $\mathbf{t}_g^*$ follows from the fact that the inverse exists and that it is non–negative (remember that $\mathbf{P}$ is stochastic and that $0 \leq A_1/(A_0 + A_1) < 1$). Hence, $\mathbf{t}_g^*$ is the unique equilibrium satisfying $\partial U_i/\partial t_{g,i} = 0$ for all $i \in M$ and $g \in G$.

Next, we will will relate the equilibrium time investments to the segregation index. Recall that the vector of local segregations $\mathbf{v}_g$ for the continuation probability $\alpha = A_1/(A_0 + A_1)$ is defined as

$$\mathbf{v}_g = \frac{A_0}{A_0 + A_1} \left(\mathbf{I} - \frac{A_1}{A_0 + A_1} \mathbf{P} \right)^{-1} \mathbf{P}\, \mathbf{c}_g.$$

Consequently,

$$\mathbf{t}_g^* = \frac{A_0 + A_1}{D}\, \mathbf{v}_g.$$

Finally, the average time investment in culture g by members of group g is

$$\bar{t}^g \equiv \frac{1}{n_g} \sum_{i \in M} n_{g,i}\, t_{g,i}^* = \sum_{i \in M} d_{g,i}\, t_{g,i}^* = \frac{A_0 + A_1}{D} \sum_{i \in M} d_{g,i}\, v_{g,i} = \frac{A_0 + A_1}{D}\, \sigma_g \left(C, \frac{A_1}{A_0 + A_1} \right).$$

This concludes the first part of the proof.

2) Uniqueness

In any equilibrium, $\partial U_i/\partial t_{g,i} \leq 0$ for all $i \in M$ and $g \in G$, with strict inequality only if $t_{g,i}^* = 0$. We show that $\partial U_i/\partial t_{g,i} = 0$. Suppose, by contradiction, that $\partial U_i/\partial t_{g,i} < 0$ for some $i \in M$ and $g \in G$. Then, $t_{g,i}^* = 0$ and it follows from equation (2) that

$$A_0 \sum_{j \in M} p_{i,j} c_{g,j} + A_1 \sum_{j \in M} p_{i,j} \frac{t_{g,j}}{t_j} < 0.$$

However, this cannot hold because all values in the equation are non–negative by definition. Hence, $\partial U_i/\partial t_{g,i} = 0$, and the unique equilibrium is the one calculated in the first part of the proof. This concludes the second part of the proof.

3) Balancedness

The balancedness conditions can be written formally as:

- *Total balancedness.* For all $i, j \in M$, $n_i \cdot t_i \cdot p_{i,j} = n_j \cdot t_j \cdot p_{j,i}$.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

- *I–balancedness.* For all $i, j \in M$ and $g \in G$, $n_i \cdot p_{i,j} \cdot t_{g,i} \cdot t_{g,j} \cdot 1/t_j = n_j \cdot p_{j,i} \cdot t_{g,j} \cdot t_{g,i} \cdot 1/t_i$, which is equivalent to $n_i \cdot t_i \cdot p_{i,j} = n_j \cdot t_j \cdot p_{j,i}$.

- *C–balancedness.* For all $i, j \in M$ and $g \in G$, $n_i \cdot p_{i,j} \cdot t_{g,i} \cdot c_{g,j} = n_{g,j} \cdot t_j \cdot p_{j,i} \cdot t_{g,i} \cdot 1/t_i$, which is equivalent to $n_i \cdot t_i \cdot_{i,j} = n_j \cdot t_j \cdot p_{j,i}$.

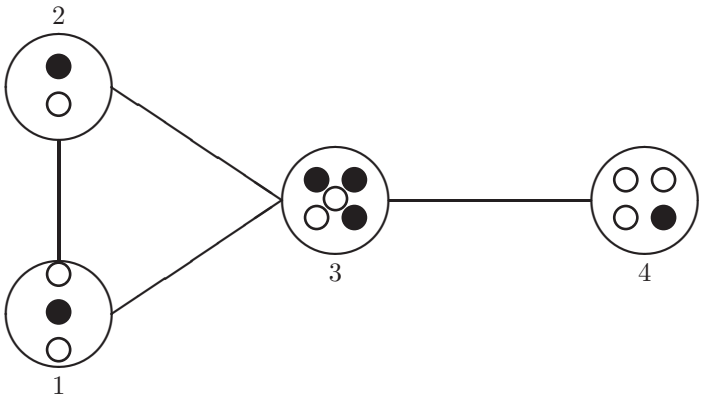
The three conditions are therefore equivalent, and we only have to show that $n_i \, t_i^* \, p_{i,j} = n_j \, t_j^* \, p_{j,i}$. Using that $a_i/a_j = n_i/n_j$ by assumption, $t_i^* = t_j^*$, and $p_{ij} = 1/a_i$, we obtain that

$$\frac{p_{i,j}}{p_{j,i}} = \frac{1/a_i}{1/a_j} = \frac{a_j}{a_i} = \frac{n_j}{n_i} = \frac{n_j}{n_i} \frac{t_j^*}{t_i^*}.$$

This concludes the proof of the Theorem.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Figures and Tables that belong to the main text



$$\mathbf{P} = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

Figure 1: Calculation of the segregation index.

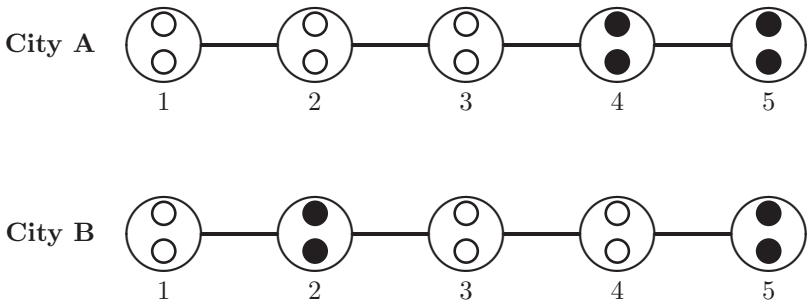


Figure 2: Clustering.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

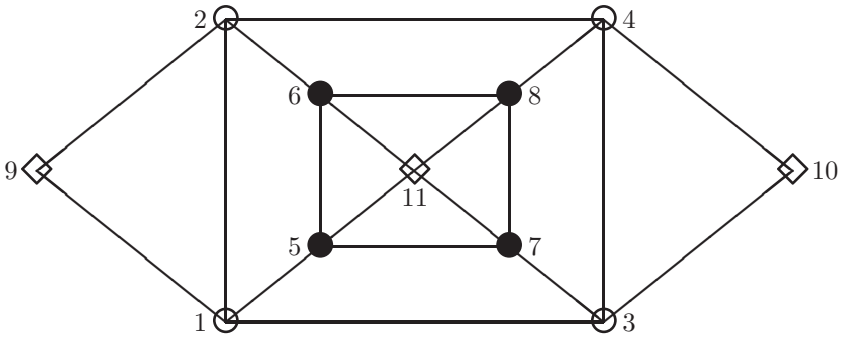
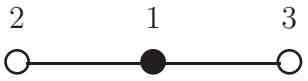


Figure 3: Centrality.



$$\mathbf{P} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}.$$

Figure 4: Evenness.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

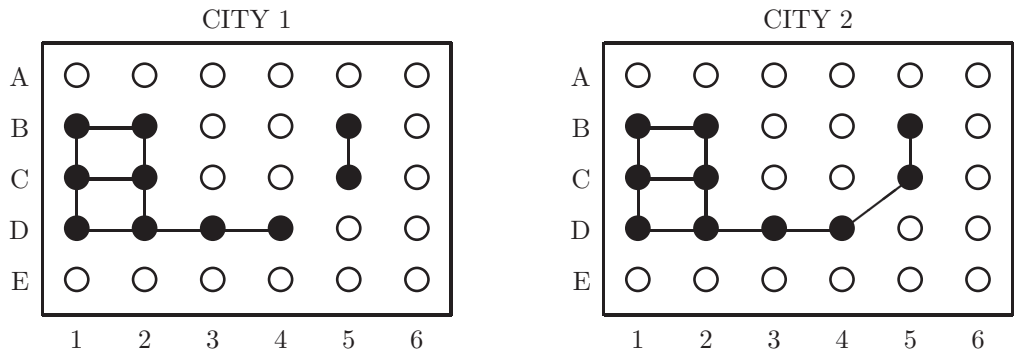


Figure 5: Calculation of the SSI.

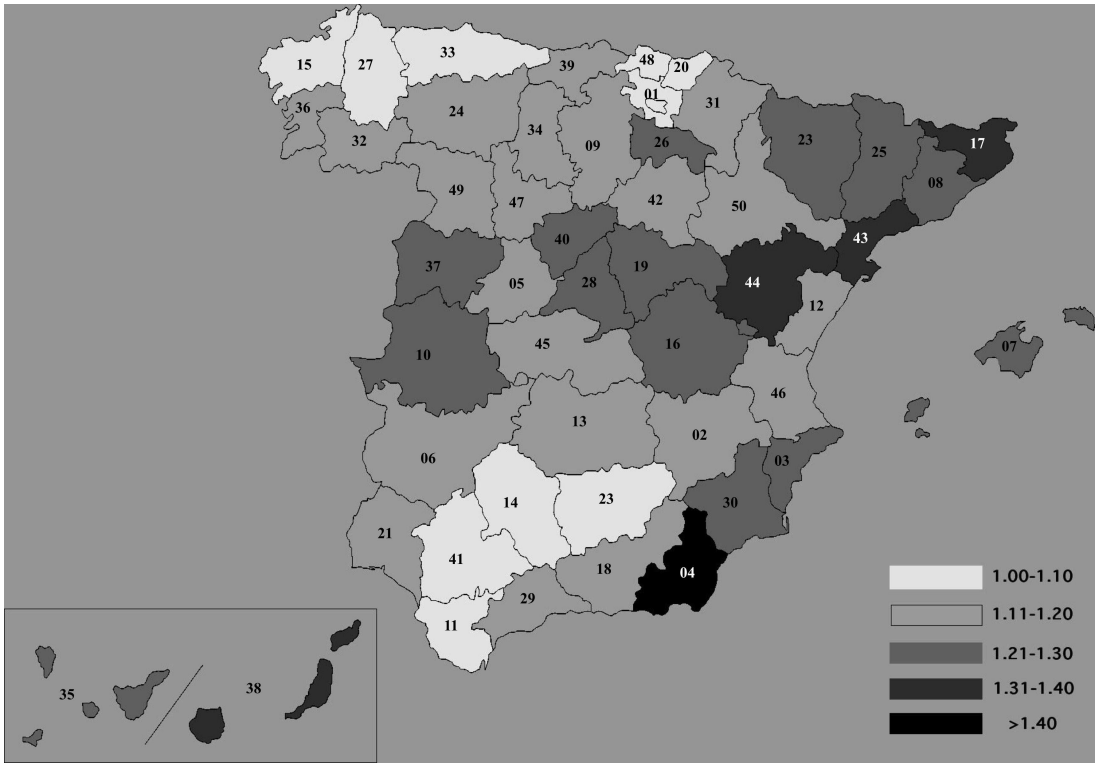


Figure 6: Normalized segregation in Spain by provinces as of January 2009 for $\alpha = 0.85$ and a neighborhood radius of 400 meters. The names of the different provinces can be identified with the help of Table 3.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

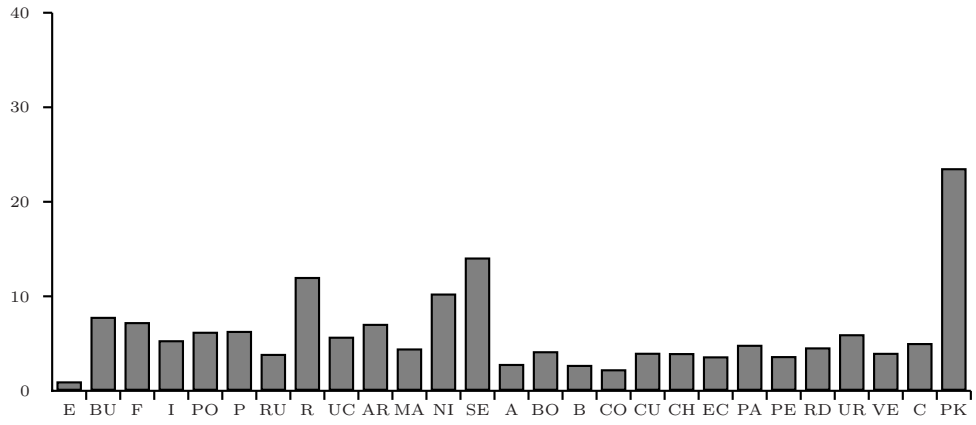


Figure 7: Normalized segregation in Spain by groups as of January 2009 for $\alpha = 0.85$ and a neighborhood radius of 400 meters. The ordering of the different groups corresponds to that in Table 2.

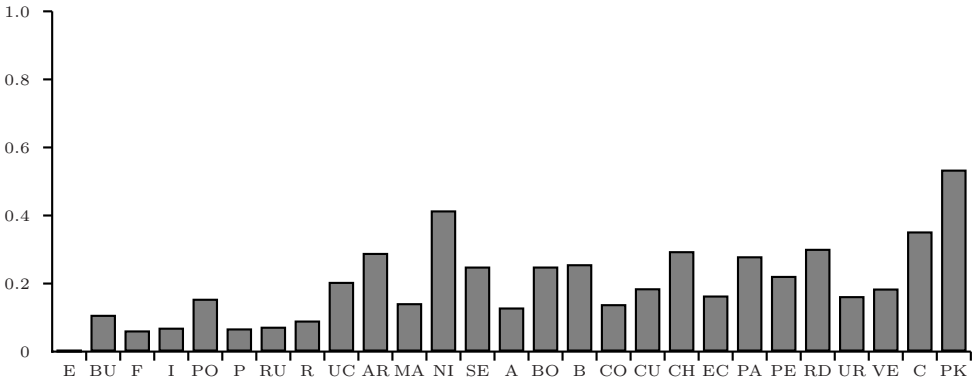


Figure 8: Network effects (in proportions) in Spain by groups as of January 2009 for a neighborhood radius of 400 meters. The ordering of the different groups corresponds to that in Table 2.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

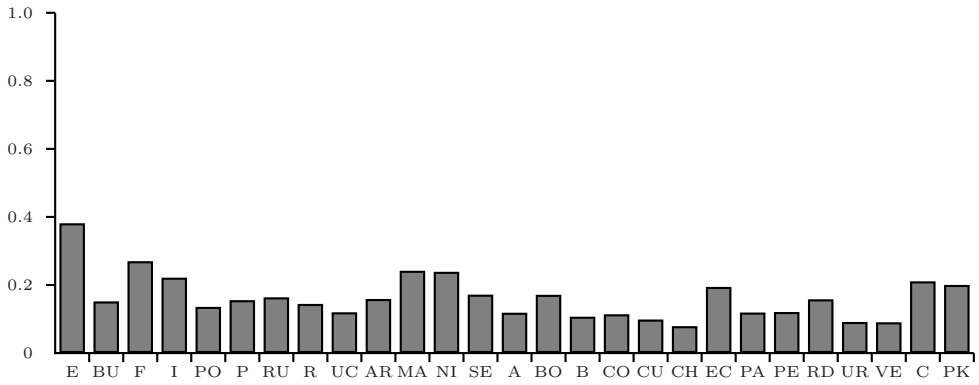


Figure 9: Proportion of the network effect that is due to connections beyond immediate neighbors in Spain by groups as of January 2009 for a neighborhood radius of 400 meters. The ordering of the different groups corresponds to that in Table 2.

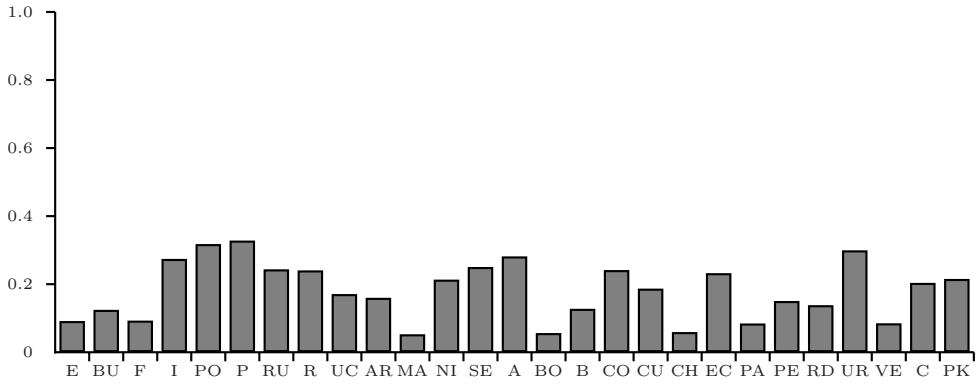


Figure 10: Correlation between $\sigma_g(C, 0.85)$ and the SSI in Spain by groups as of January 2009 for a neighborhood radius of 400 meters. The ordering of the different groups corresponds to that in Table 2.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

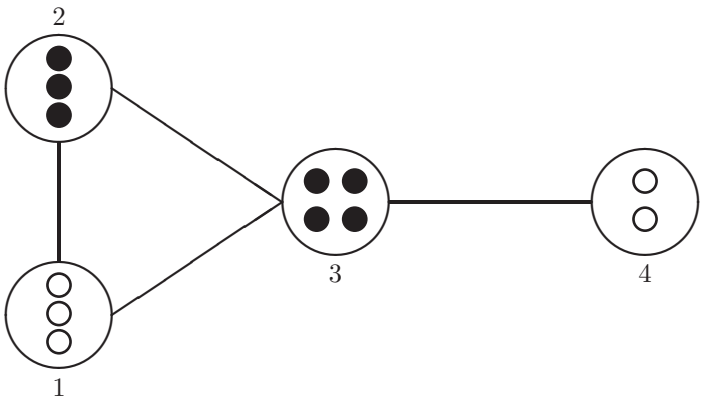


Figure 11: Connections and arrival rates.

Components	Blacks										*SSI*
City 1											
Component 1	(B,1)	(B,2)	(C,1)	(C,2)	(D,1)	(D,2)	(D,3)	(D,4)			
	0.87	0.62	1.26	0.92	0.93	0.75	0.29	0.10			**0.72**
Component 2	(B,5)	(C,5)									
	0.25	0.25									**0.25**
Weighted Average											**0.63**
City 2											
Component	(B,1)	(B,2)	(C,1)	(C,2)	(D,1)	(D,2)	(D,3)	(D,4)	(B,5)	(C,5)	
	1.09	0.78	1.58	1.14	1.16	0.93	0.37	0.13	0.00	0.00	**0.72**

Table 1: Calculation of the SSI.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Nationality	ID	Number of Residents	Share
Spain	E	40,956,149	90.26 %
Europe			
Bulgaria	BU	164,716	0.36 %
France	F	120,262	0.27 %
Italy	I	175,232	0.39 %
Poland	PO	85,007	0.19 %
Portugal	P	140,801	0.31 %
Romania	RU	798,869	1.76 %
Russia	R	47,428	0.10 %
Ukraine	UC	82,263	0.18 %
Africa			
Algeria	AR	56,194	0.12 %
Morocco	MA	708,939	1.56 %
Nigeria	NI	42,322	0.09 %
Senegal	SE	56,589	0.12 %
South America			

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Argentina	A	142,239	0.31 %
Bolivia	BO	230,693	0.51 %
Brazil	B	126,172	0.28 %
Colombia	CO	296,619	0.65 %
Cuba	CU	54,598	0.12 %
Chile	CH	51,032	0.10 %
Ecuador	EC	421,385	0.93 %
Paraguay	PA	81,549	0.18 %
Peru	PE	139,167	0.31 %
Dominican Republic	RD	88,102	0.19 %
Uruguay	UR	50,422	0.11 %
Venezuela	VE	61,448	0.14 %
Asia			
China	C	147,373	0.32%
Pakistan	PK	54,100	0.12%

Table 2: Residents in Spain as of January 2009 according to the country of origin (nationality). The data is made available by the National Statistical Institute (INE) of Spain.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Number	Name	Number	Name	Number	Name
01	Álava	02	Albacete	03	Alicante
04	Almería	05	Ávila	06	Badajoz
07	Baleares	08	Barcelona	09	Burgos
10	Cáceres	11	Cádiz	12	Castellón
13	Ciudad Real	14	Córdoba	15	Cuenca
16	Coruña	17	Girona	18	Granada
19	Guadalajara	20	Guipúzca	21	Huelva
22	Huesca	23	Jaén	24	León
25	Lérida	26	Rioja	27	Lugo
28	Madrid	29	Málaga	30	Murcia
31	Navarra	32	Orense	33	Asturias
34	Palencia	35	Palmas	36	Pontevedra
37	Salamanca	38	Santa Cruz	39	Cantabria
40	Segovia	41	Sevilla	42	Soria
43	Tarragona	44	Teruel	45	Toledo
46	Valencia	47	Valladolid	48	Vizcaya
49	Zamora	50	Zaragoza		

Table 3: Idenfication of provinces.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Field ID	Full Name
AccAud	Accounting and Auditing
AgrEnvEn	Agricultural, Environmental and Energy Economics
DevTrans	Development and Transition Economics
EconStat	Econometrics and Statistics Applied to Economics
FinIns	Finance and Insurance
GenEcon	General Economics
HistPolMet	History of Economic Thought, Political Economy, Methodology
HRM	Human Resource Management and Industrial Relations
InnovEntrSB	Innovation, Entrepreneurship, Small Business
IO	Industrial Organization, Productivity Analysis
LaborHealth	Labor, Education, Population and Health Economics
Law	Law and Economics
Macro	Macroeconomics, International and Monetary Economics
PubEcon	Public Economics, Public Choice
Sector	Sectorial Studies
Socio	Social Economics, Political Science, Philosophy, Sociology, International Relations
Theory	Economic Theory, Game and Decision Theory
UrbSpaReg	Urban, Spatial, Regional and Transport Economics

Table 4: Field abbreviations in the citation model using the CEFAGE–EU ranking from the University of Évora, Portugal.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00399

Field ID	# Journals	n_g/n	$h_g(C)$	$\sigma_g(C, 0.85)$	$\bar{h}_g(C)$	$\bar{\sigma}_g(C, 0.85)$
AccAud	1	0.004 (17)	0.000 (17)	0.011 (14)	0.000 (16)	2.617 (6)
AgrEnvEn	17	**0.120 (2)**	**0.530 (2)**	**0.163 (5)**	4.414 (9)	1.358 (14)
DevTrans	11	0.062 (7)	0.341 (7)	0.118 (7)	5.468 (7)	1.887 (9)
EconStat	9	0.055 (8)	**0.474 (5)**	**0.164 (4)**	**8.636 (3)**	**2.983 (4)**
FinIns	11	**0.094 (3)**	**0.691 (1)**	**0.351 (2)**	**7.356 (4)**	**3.742 (3)**
GenEcon	33	**0.288 (1)**	**0.479 (4)**	**0.458 (1)**	1.662 (14)	1.589 (12)
HistPolMet	3	0.011 (13)	0.365 (6)	0.084 (9)	**31.999 (1)**	**7.343 (1)**
HRM	1	0.004 (17)	0.000 (17)	0.001 (18)	0.000 (16)	0.170 (18)
InnovEntrSB	2	0.009 (14)	0.036 (13)	0.009 (15)	4.097 (11)	1.054 (16)
IO	6	0.032 (10)	0.193 (11)	0.061 (10)	5.799 (6)	1.846 (10)
LaborHealth	11	**0.079 (4)**	**0.528 (3)**	**0.207 (3)**	**6.672 (5)**	**2.622 (5)**
Law	2	0.006 (15)	0.029 (14)	0.012 (13)	4.306 (10)	1.741 (11)
Macro	10	**0.074 (5)**	0.222 (10)	0.111 (8)	3.000 (12)	1.497 (13)
PubEcon	5	0.049 (9)	0.121 (12)	0.051 (12)	2.575 (12)	1.095 (15)
Sector	1	0.001 (18)	0.000 (17)	0.004 (17)	0.000 (16)	**3.927 (2)**
Socio	3	0.016 (12)	0.001 (15)	0.007 (16)	0.001 (15)	0.411 (17)
Theory	9	0.070 (6)	0.331 (8)	0.143 (6)	4.749 (8)	2.057 (8)
UrbSpaReg	5	0.024 (11)	0.264 (9)	0.059 (11)	**10.982 (2)**	2.467 (7)

Table 5: Field, number of journals per field, article shares, homophily index (non–normalized and normalized), and segregation index (non–normalized and normalized) for 18 different areas in Economics using the citation data from 2010 for 140 economics journals. In parentheses, we present the corresponding ordinal rankings.