Extracting spatial information from networks with low-order eigenvectors

Mihai Cucuringu[*]

Program in Applied and Computational Mathematics (PACM), Princeton University, Fine Hall, Washington Road, Princeton, New Jersey 08544-1000, USA

Vincent D. Blondel[†] and Paul Van Dooren[‡]

Department of Applied Mathematics, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

(Received 15 March 2012; revised manuscript received 20 October 2012; published 8 March 2013)

We consider the problem of inferring meaningful spatial information in networks from incomplete information on the connection intensity between the nodes of the network. We consider two spatially distributed networks: a population migration flow network within the US, and a network of mobile phone calls between cities in Belgium. For both networks we use the eigenvectors of the Laplacian matrix constructed from the link intensities to obtain informative visualizations and capture natural geographical subdivisions. We observe that some low-order eigenvectors localize very well and seem to reveal small geographically cohesive regions that match remarkably well with political and administrative boundaries. We discuss possible explanations for this observation by describing diffusion maps and localized eigenfunctions. In addition, we discuss a possible connection with the weighted graph cut problem, and provide numerical evidence supporting the idea that lower-order eigenvectors point out local cuts in the network. However, we do not provide a formal and rigorous justification for our observations.

DOI: 10.1103/PhysRevE.87.032803 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.Cd, 89.65.Ef

I. INTRODUCTION

Extensive research over the past decades has greatly increased our understanding of the topology and the spatial distribution of many social, biological, and technological networks. This paper considers the problem of inferring meaningful spatial and structural information from incomplete data sets of pairwise interactions between nodes in a network.

The way people interact in many aspects of everyday life often reflect surprisingly well geopolitical boundaries. This inhomogeneity of connections in networks leads to natural divisions, and identifying such divisions can provide valuable insight into how interactions in a network are influenced by its topology. The problem of finding the so-called network communities, i.e., groups of tightly connected nodes, has been extensively studied in recent years and many community detection algorithms exist with different levels of success [1]. In this paper, we consider two particular networks: a county-to-county migration network constructed from 1995 to 2000 US Census data, and a city-to-city communication network built from mobile phone data over a six month period in Belgium. Communities in these networks emerge naturally and are revealed, often at different scales [2], by the eigenvectors of a normalized matrix constructed from the weighted adjacency matrix of the network. We discuss possible explanations for this observation by describing diffusion maps and localized eigenfunctions.

In the remaining part of this Sec. I we report on some related contributions that deal with communities in networks and spectrum of matrices. However, in none of these contributions were we able to find an explanation of why low-order eigenvectors localize so well and seem to identify meaningful geographical boundaries.

One example of a study that is related to our work both in terms of the technique and end goal is a paper by Ratti *et al*. [3]. Starting from measures of the communication intensities between counties in the UK, the authors propose a spectral modularity[1] optimization algorithm that partitions the country into small nonoverlapping geographically cohesive regions that correspond remarkably well with administrative regions.

In [4], Shi and Malik develop a spectral-based algorithm that solves the perceptual grouping problem in computer vision by treating the task of image segmentation as a graph partitioning problem. Their approach is to segment the graph by introducing a global criterion called normalized cut, that measures not just the dissimilarity between different groups but also the total similarity within the groups themselves. They successfully extract global impressions of a scene and provide a hierarchical description of it.

Reades *et al*. [5] connect mobile data from Telecom Italia Mobile to a series of human activities derived from data on commercial premises advertised through the Italian version of "Yellow Pages." The eigendecomposition of a specific correlation matrix provides a top eigenvector which clearly indicates a common underlying pattern to mobile phone usage in Rome, while the second and third eigenvectors indicate spatial variation that is very suggestive of temporally related and activity-related patterns.

Another line of work where lower-order eigenvectors provide useful information comes from the community

[*] mcucurin@math.princeton.edu; www.math.princeton.edu/∼mcucurin/

[†] vincent.blondel@uclouvain.be; www.inma.ucl.ac.be/∼blondel/

[‡] paul.vandooren@uclouvain.be; www.inma.ucl.ac.be/∼vdooren/

[1] Many popular methods for community detection in networks are based on the optimization of the modularity function, a measure of the quality of a network partition into communities.

detection literature. Newman [6] shows that the modularity of a network can be expressed in terms of the top eigenvalues and eigenvectors of a matrix called the modularity matrix, which plays a role in the maximization of the modularity equivalent to that played by the Laplacian in standard spectral partitioning. In related work, Richardson *et al.* [7] extend previously available methods for spectral optimization of modularity by introducing a computationally efficient algorithm for spectral tripartitioning of a network using the top two eigenvectors of the modularity matrix. We also mention here the recent work of Van Mieghem *et al.* [8], who present bounds for modularity, and discuss the influence of the spectrum of the modularity matrix on the maximum modularity. Furthermore, the authors present an analysis of the relationships among the modularity, the assortativity (correlation of the similarities of nodes sharing a link), the largest eigenvalues of the adjacency and modularity matrices, the number of clusters, and the effective graph resistance.

Arenas *et al.* [9] give a geometrical interpretation to modular organization in complex networks. They introduce a mathematical object denoted as the "contribution matrix," which contains information about the partitions of interest, and later use a truncated singular value decomposition to extract the best representation of this matrix in the plane and reveal the skeleton of the associated network. The proposed approach is applied to real networks including the worldwide air transportation network and the AS-P2P Internet network.

Recent work [10], coauthored by one of the authors of this paper, investigates the constraints imposed by space on the network topology, and focuses on community detection by proposing a modularity function adapted to spatial networks. The proposed methods were tested on a large mobile phone network and computer-generated benchmarks, and showed that it is possible to factor out the effect of space in order to reveal more clearly any hidden structural similarities between the nodes. Onnela *et al.* [11] investigate social networks of individuals whose most frequent geographic locations are known. The authors classify the members into groups using community detection algorithms, and explore the relationship between their topological and geographic positions.

On a more general note, we point out that extracting information from the top spectrum of adjacency matrices expands well beyond the detection of spatial proximity, and has also been used to reveal information in other contexts such as semantic analysis [12,13] and navigability of networks. Almost two decades ago, the authors of [12] introduced latent semantic indexing, an algorithm for retrieving textual materials from scientific databases, which relied on the singular value decomposition of large sparse matrices to extract information from the high-order structure in the association of terms with documents.

Motivated by the scalability problems with the Internet routing architecture, Boguñá *et al.* [14] propose an efficient mechanism that explains the connection between network structure and its functions, by relying on the presence of an underlying metric space hidden behind the observable network. In related work [15], a subset of the same authors introduced a method that maps the Internet to a hyperbolic space, allowing them to increase the scalability of network routing algorithms and at the same time to provide a different perspective on community structure in complex networks. Their findings have practical applications that include Internet routing, searching social networks, and the study of information flow in gene regulatory networks.

Finally, we point out the work of Liu *et al.* on the spectral reconstructability of complex networks [16]. They introduce and investigate the reconstructability coefficient θ of a network as the maximum number of eigenvalues that can be set to zero, while still being able to exactly reconstruct the adjacency matrix. Their main finding is that for sufficiently large networks, an apparently universal linear scaling law holds, which allows for a portion of the smallest eigenvalues (in absolute values) to be removed from the spectrum, as long as one still uses the exact eigenvectors. The main difference with respect to the work we present in this paper is that Liu *et al.* use the top eigenvectors and eigenvalues of the spectrum for an (exact) reconstruction of the entire network, while we use the top eigenvectors to discover and highlight the local structure within the network.

This paper is organized as follows. Section II is an introduction to the diffusion map technique and some of its underlying theory. Section III contains the results of numerical simulations in which we applied diffusion maps and eigenvector colorings to the US migration data set. In Sec. IV we present the outcome of similar experiments on the Belgium mobile phone data set. In Sec. V, we explore the connection with localized eigenfunctions, a phenomenon observed before in the mathematics and physics community. Finally, the last section is a summary and a discussion of possible extensions of our approach and its usefulness in other applications.

II. DIFFUSION MAPS AND EIGENVECTOR COLORINGS

This section is a brief introduction to the diffusion maps literature and references therein. We also clarify the notion of eigenvector localizations and eigenvector coloring that we use in subsequent sections. Diffusion maps were introduced in [17] as a dimensionality reduction tool, and connected data analysis and clustering techniques based on eigenvectors of similarity matrices with the geometric structure of nonlinear manifolds. In recent years, diffusion maps have gained a lot of popularity. A nonexhaustive list of references to its underlying theory and applications includes [17–21]. Often called Laplacian eigenmaps, these manifold learning techniques identify significant variables that live in a lower-dimensional space, while preserving the local proximity between data points. Consider a set of N points $V = \{x_1,x_2,\dots,x_N\}$ in an n-dimensional space $\mathbb{R}^n$, where each point (typically) characterizes an image (or an audio stream, text string, etc.). If two images x_i and x_j are similar, then $||x_i - x_j||$ is small. A popular measure of similarity between points in $\mathbb{R}^n$ is defined using the Gaussian kernel $w_{ij} = e^{-||x_i - x_j||^2/\epsilon}$, for some constant ϵ, so that the closer x_i is from x_j, the larger w_{ij}. The matrix $W = (w_{ij})_{1 \leqslant i,j \leqslant N}$ is symmetric and has positive coefficients. To normalize W, we define the diagonal matrix D, with $D_{ii} = \sum_{j=1}^{N} w_{ij}$ and define A by

$$A = D^{-1}W,$$

such that every row of A sums to 1.

Next, one may also define the symmetric matrix $S = D^{-1/2} W D^{-1/2}$, which can also be written as $S = D^{1/2} A D^{-1/2}$ and hence is similar to A. As a symmetric matrix, S has an orthogonal basis of eigenvectors $v_0, v_1, \ldots, v_{N-1}$ associated to the N real ordered eigenvalues $1 = \lambda_0 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_{N-1}$. If we decompose S as $S = V \Lambda V^T$ with $VV^T = V^T V = I$ and $\Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$, then A becomes $A = \Psi \Lambda \Phi^T$ where $\Psi = D^{-1/2} V$ and $\Phi = D^{1/2} V$. Therefore, $A\Psi = \Psi\Lambda$ and the columns of Ψ (denoted by $\psi_0, \psi_1, \ldots, \psi_{N-1}$) form a D- orthogonal basis (i.e., $\langle \psi_i, D\psi_j \rangle = 0, \forall i \neq j$) associated to the N real eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{N-1}$ such that $A\psi_i = \lambda_i \psi_i$, for $i = 0, 1, \ldots, N-1$. Also, $\Phi^T A = \Lambda \Phi^T$ implies that the columns of Φ are left eigenvectors of A, which we denote by $\phi_0, \phi_1, \ldots, \phi_{N-1}$. Since $\Phi^T \Psi = I$, it follows that the vectors ϕ_i and ψ_j are biorthonormal $\langle \phi_i, \psi_j \rangle = \delta_{i,j}$.

Since A is a row-stochastic matrix, $\lambda_0 = 1$ and $\psi_0 = (1, 1, \ldots, 1)^T$, and we disregard this trivial eigenvalue-eigenvector pair as irrelevant. Using the stochasticity of A, we can interpret it as a random-walk matrix on a weighted graph $G = (V, E, W)$, where the set of nodes consists of the points x_i, and there is an edge between nodes i and j if and only if $w_{ij} > 0$. Taking this perspective, A_{ij} denotes the transition probability from point x_i to x_j in one step time $\Delta t = \epsilon$,

$$\Pr\{x(t+\epsilon) = x_j | x(t) = x_i\} = A_{ij}.$$

The parameter ϵ can now be interpreted in two ways. On the one hand, it is the squared radius of the neighborhood used to infer local geometric and density information, in particular w_{ij} is $O(1)$ when x_i and x_j are in a ball of radius $\sqrt{\epsilon}$, but it is exponentially small for points that are more than $\sqrt{\epsilon}$ apart. On the other hand, ϵ represents the discrete time step at which the random walk jumps from one point to another. We refer the reader to [22] for a detailed survey of random walks on graphs, and their applications.

Interpreting the eigenvectors as functions over our data set, the *diffusion map* (also called *Laplacian eigenmap)* maps points from the original space to the first k eigenvectors, $\mathcal{L} : V \mapsto \mathbb{R}^k$, and is defined as

$$\mathcal{L}_t(x_j) = \left[\lambda_1^t \psi_1(j), \lambda_2^t \psi_2(j), \ldots, \lambda_k^t \psi_k(j)\right], \tag{1}$$

where the meaning of the integer exponent t will be made clear in what follows.

Using the left and right eigenvectors denoted earlier, we now write the entries of A as $A_{ij} = \sum_{r=0}^{N-1} \lambda_r \phi_r(i) \psi_r(j)$, and note that $A_{ij}^t = \sum_{r=0}^{N-1} \lambda_r^t \phi_r(i) \psi_r(j)$. However, recall that the probability distribution of a random walk landing at location x_j after exactly t steps, given that it starts at point x_i, is precisely given by the expression $A_{ij}^t = \Pr\{x(t) = x_j | x(0) = x_i\}$. Given the random-walk interpretation, it is natural to quantify the similarity between two points according to the evolution of their probability distributions,

$$D_t^2(i,j) = \sum_{k=1}^{N} \left(A_{ik}^t - A_{jk}^t\right)^2 \frac{1}{d_k},$$

where the weight $\frac{1}{d_k}$ takes into account the empirical local density of the points by giving larger weight to the vertices of lower degree. Since $D_t(i,j)$ naturally depends on the random walk on the graph, it is denoted as the *diffusion distance* at time t. In the diffusion map introduced above, it is a matter of choice to tune the parameter t corresponding to the number of time steps of the random walk. We used $t = 1$ in the diffusion map embeddings throughout our simulations, and that using different values of t corresponds to rescaling the axis. The Euclidean distance between two points in the diffusion map space introduced in (1) is given by

$$||\mathcal{L}(x_i) - \mathcal{L}(x_j)||^2 = \sum_{r=1}^{N-1} \left[\lambda_r^t \psi_r(i) - \lambda_r^t \psi_r(j)\right]^2. \tag{2}$$

The first eigenvalue λ_0 does not enter this expression, since it cancels out. Moreover, as shown in [23], the expression (2) equals the diffusion distance $D_t^2(i,j)$, when $k = N-1$, i.e., when all $N-1$ eigenvectors are considered. For ease of visualization, we used the top $k = 2$ eigenvectors for the projections shown in Figs. 1 and 6.

Finally, we denote by $\mathcal{C}_k$ the coloring of the N data points given by the eigenvector ψ_k, where the color of point $x_i \in V$ is given by the ith entry in ψ_k, i.e.,

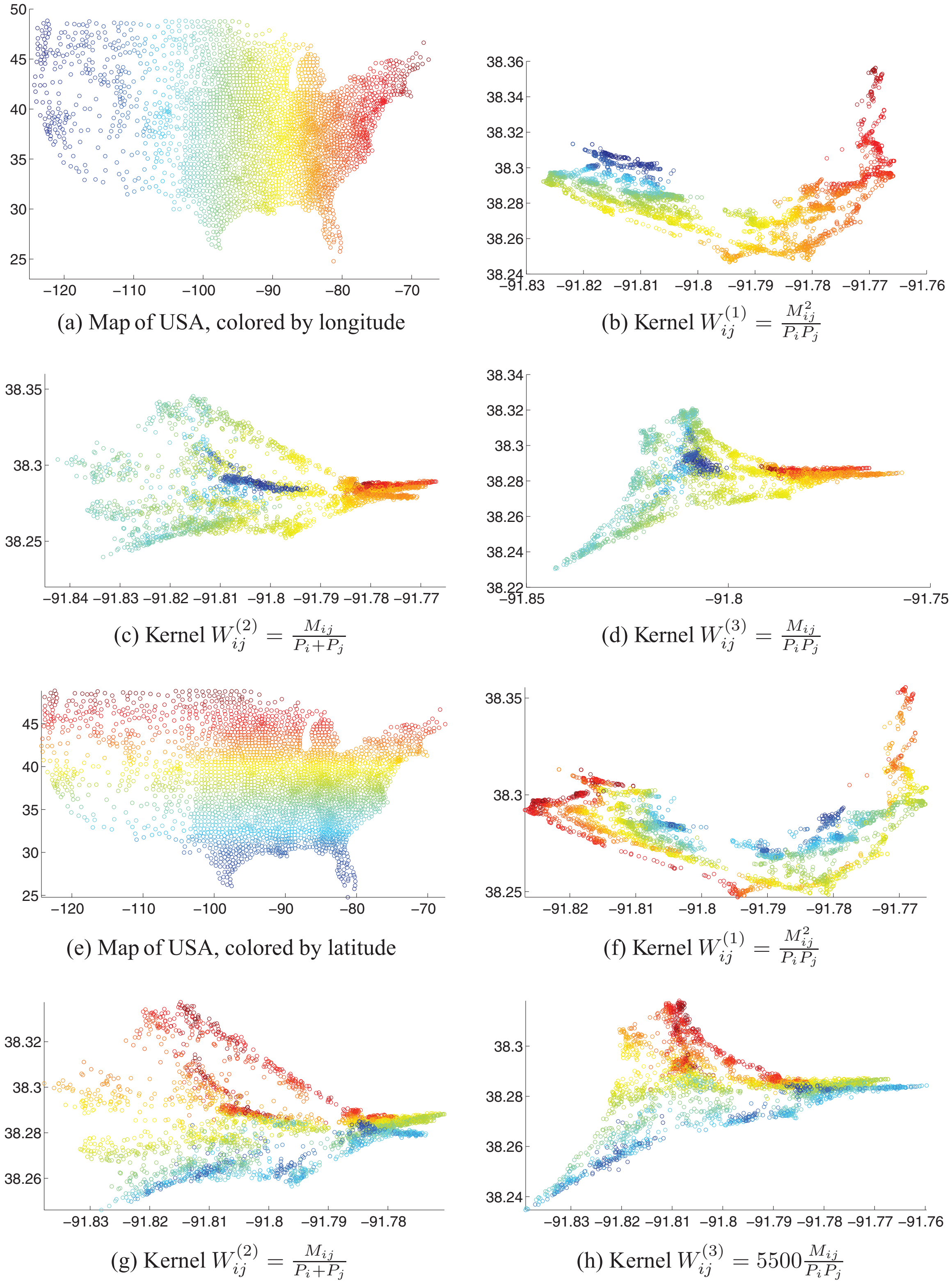
$$\mathcal{C}_k(x_i) = \psi_k(i), \text{ for all } k = 0, \ldots, N-1 \text{ and } i = 1, \ldots, N.$$

We refer to $\mathcal{C}_k$ as an *eigenvector coloring*[2] of order k. The top left plot in Fig. 3 shows the eigenvector coloring of order $k = 1$, together with the associated color bar where red denotes high values and blue denotes low values (consistent for the eigenvector colorings throughout the paper). In practice, only the first k eigenvectors are used in the diffusion map introduced in (1), with $k \ll N-1$ chosen such that $\lambda_1^t \geqslant \lambda_2^t \cdots \geqslant \lambda_k^t > \delta$ but $\lambda_{k+1}^t < \delta$, where δ is a chosen tolerance. Typically, only the top few eigenvectors of A are expected to contain meaningful information, but as illustrated by the eigenvector colorings shown in this paper, one can extract relevant information from eigenvectors of much lower order. The phenomenon of *eigenvector localization* occurs when most of the components of an eigenvector are zero or close to zero, and almost all the mass is localized on a relatively small subset of nodes. On the contrary, delocalized eigenvectors have most of their components small and of roughly the same magnitude. Furthermore, note there is no issue with the fact that the eigenvectors are defined up to a scalar. Since each of them is normalized and real, we can just consider eigenvectors of different sign; however, this can only reverse the color map used, and does not change the localization phenomenon.

III. US CENSUS MIGRATION DATA

We apply the diffusion map technique to the 2000 US Census that reports the number of people that migrated from every county to every other county in the US during the 1995–2000 time frame [24,25]. We denote by $M = (M_{ij})_{1 \leqslant i,j \leqslant N}$ the total number of people that migrated between county i and county j (so $M_{ij} = M_{ji}$) during the five-year period, where $N = 3107$ denotes the number of counties in mainland US We let P_i denote the population of county i. Figure 1 shows the results of the diffusion map technique for

[2] Not to be confused with the "coloring" terminology from graph theory, where the colors are integers.



(a) Map of USA, colored by longitude

(b) Kernel $W^{(1)}_{ij} = \frac{M^2_{ij}}{P_iP_j}$

(c) Kernel $W^{(2)}_{ij} = \frac{M_{ij}}{P_i+P_j}$

(d) Kernel $W^{(3)}_{ij} = \frac{M_{ij}}{P_iP_j}$

(e) Map of USA, colored by latitude

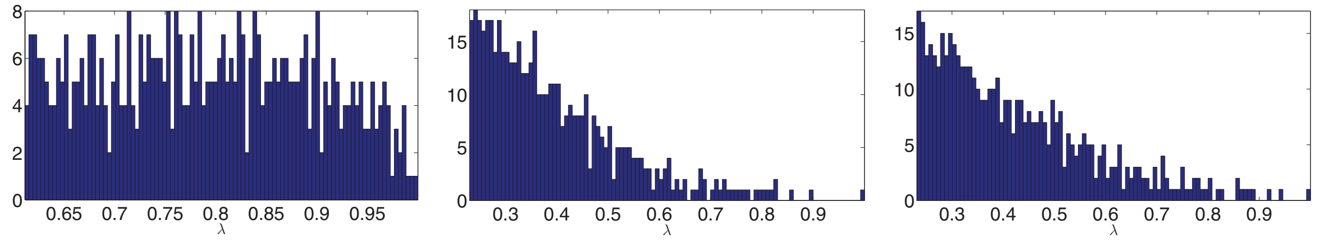
(f) Kernel $W^{(1)}_{ij} = \frac{M^2_{ij}}{P_iP_j}$

(g) Kernel $W^{(2)}_{ij} = \frac{M_{ij}}{P_i+P_j}$

(h) Kernel $W^{(3)}_{ij} = 5500\frac{M_{ij}}{P_iP_j}$

FIG. 1. (Color) Diffusion map reconstructions from the top two eigenvectors, for various kernels, with nodes colored by longitude [(a),(b),(c),(d)] and latitude [(e),(f),(g),(h)].

longitude and latitude colorings when the following kernels are used: $W^{(1)}_{ij} = \frac{M^2_{ij}}{P_iP_j}$, $W^{(2)}_{ij} = \frac{M_{ij}}{P_i+P_j}$, and $W^{(3)}_{ij} = 5500\frac{M_{ij}}{P_iP_j}$. The diffusion map resulting from these kernels places the Midwest closer to the West coast (Fig. 1), but further from the East coast. Similarly, the colorings based on latitude reveal the north-south separation. The kernel $W^{(1)}$ does a better job at separating the East and West coasts, Fig. 1(b), while kernel



(a) $A = D^{-1}W^{(1)}$ (b) $A = D^{-1}W^{(2)}$ (c) $A = D^{-1}W^{(3)}$

FIG. 2. (Color online) Histogram of the top 500 eigenvalues of matrix A for different kernels.

$W^{(2)}$ highlights best the separation between north and south as shown in Fig. 1(g). Figure 2 shows the histogram of the top 500 eigenvalues of the diffusion matrix A, when different kernels are used.

Our kernel of choice for the eigenvector colorings in Fig. 3 was $W^{(1)}$, as it produced more visually appealing results in terms of state boundary detection. For the same reason, we omit the numerical simulations where we used exponential weights to compute the similarity between the nodes. Note also that the spectrum of $A = D^{-1}W^{(1)}$ in the left of Fig. 2 is rather different from the other two spectra, with many more large eigenvalues and without a visible spectral gap. For the rest of this section, we drop the superscript from matrix $W^{(1)}$ and refer to it as W.

In Fig. 4 we plot the histograms of the entries of several eigenvectors of A. The top eigenvector provides a meaningful partitioning that separates the East from the Midwest, and has its entries spread in the interval [−0.03,0.03] with few entries of zero magnitude. On the other hand, the eigenvectors ϕ_7, ϕ_{28}, and ϕ_{83} are *localized* in the sense that they have their larger entries localized on a specific subregion of the US map (highlighted in blue or red in the eigenvector colorings), while taking small values in magnitude on the rest of the domain. We explore in Sec. V the connection with the phenomenon of "localized eigenfunctions" of the Laplace operator.

We use the rest of this section to provide a possible interpretation of the color coded regions that stand out in the eigenvector colorings in Fig. 3. By interpreting the matrix W as a weighted graph, we explore a possible connection of such geographically cohesive colored subgraphs with the graph partitioning problem. In general, the graph partitioning problem seeks to decompose a graph into K disjoint subgraphs (clusters), while minimizing the sum of the weights of the "cut" edges, i.e., edges with end points in different clusters. Given the number of clusters K, the weighted-min-cut problem is an optimization problem that computes a partition $\mathcal{P}_1,\ldots,\mathcal{P}_K$ of the vertex set, by minimizing the weights of the cut edges,

$$\text{Weighted Cut}(\mathcal{P}_1,\ldots,\mathcal{P}_k) = \sum_{i=1}^{k} E_w(\mathcal{P}_i,\overline{\mathcal{P}_i}), \tag{3}$$

where $E_w(X,Y) = \sum_{i\in X, j\in Y} W_{ij}$, and $\overline{X}$ denotes the complement of X. For an extensive literature survey on spectral clustering algorithms we refer the reader to [26], and point out the popular spectral relaxation of (3) introduced by Shi and Malik [4].

When dividing a graph into two smaller subgraphs, one wishes to minimize the sum of the weights on the edges across two different subgraphs and, simultaneously, maximize the sum of the weights on the edges within the subgraphs. Alternatively, one tries to maximize the ratio between the latter quantity and the former, i.e., between the weights of the inside edges and the weights of the outside edges. To that end, we perform the following experiment, where we regard the US states as the clusters, and investigate the possibility that the isolated colored regions that emerge correspond to local cuts in the weighted graph.

We denote by S the matrix of size $N \times N$ ($N = 49$ the number of mainland US states) that aggregates the similarities between counties at the level of states. In particular, if state i has k counties with indices $x_1,\ldots,x_k$, and state j has l counties with indices $y_1,\ldots,y_l$, then we consider the $k \times l$ submatrix,

$$\tilde{W}_{i,j} = W_{\{x_1,\ldots,x_k\},\{y_1,\ldots,y_l\}}, \tag{4}$$

and denote by S_{ij} the sum of the kl entries in $\tilde{W}_{i,j}$. In other words, matrix S is a "state-collapsed" version of the matrix W, and gives a measure of similarity between pairs of states. The heat map in Fig. 5 shows the components of the matrix S on a logarithmic scale, where the intensity of entry (i,j) denotes the aggregated similarity between states i and j.

We refer to the diagonal entry S_{ii} as the "inside degree" of state i, $d_i^{\text{in}} = S_{ii}$, which measures the internal similarity between the counties of state i. We denote by $d_i^{\text{out}} = \sum_{u=1,u\neq i}^{N} S_{i,u}$ (i.e., the sum of the nondiagonal elements in row i) the "outside degree" of node i, which measures the similarity or migration between the counties of state i and all other counties outside of state i. Finally, we denote by $d_i^{\text{ratio}} = \frac{d_i^{\text{in}}}{d_i^{\text{out}}}$ the "ratio degree" of node i which straddles the boundary between intrastate and interstate migration. A large ratio degree is a good indication that a state is very well connected internally, and has little connectivity with the outside world, and thus is a good candidate for a cluster. In other words, a large ratio degree of a cluster (i.e., state) denotes a high measure of separation between that cluster and its environment, which is something discovered by the localization properties of the low-order eigenvectors. Table I ranks the top 14 states within the US in terms of their ratio degree.

Next, we examine the top several eigenvector colorings in Fig. 3, and point out the individual states on which the eigenvectors localize, together with its rank in terms of ratio degree. The entries of large magnitude are colored in red and blue, while the rest of the spectrum denotes values of smaller

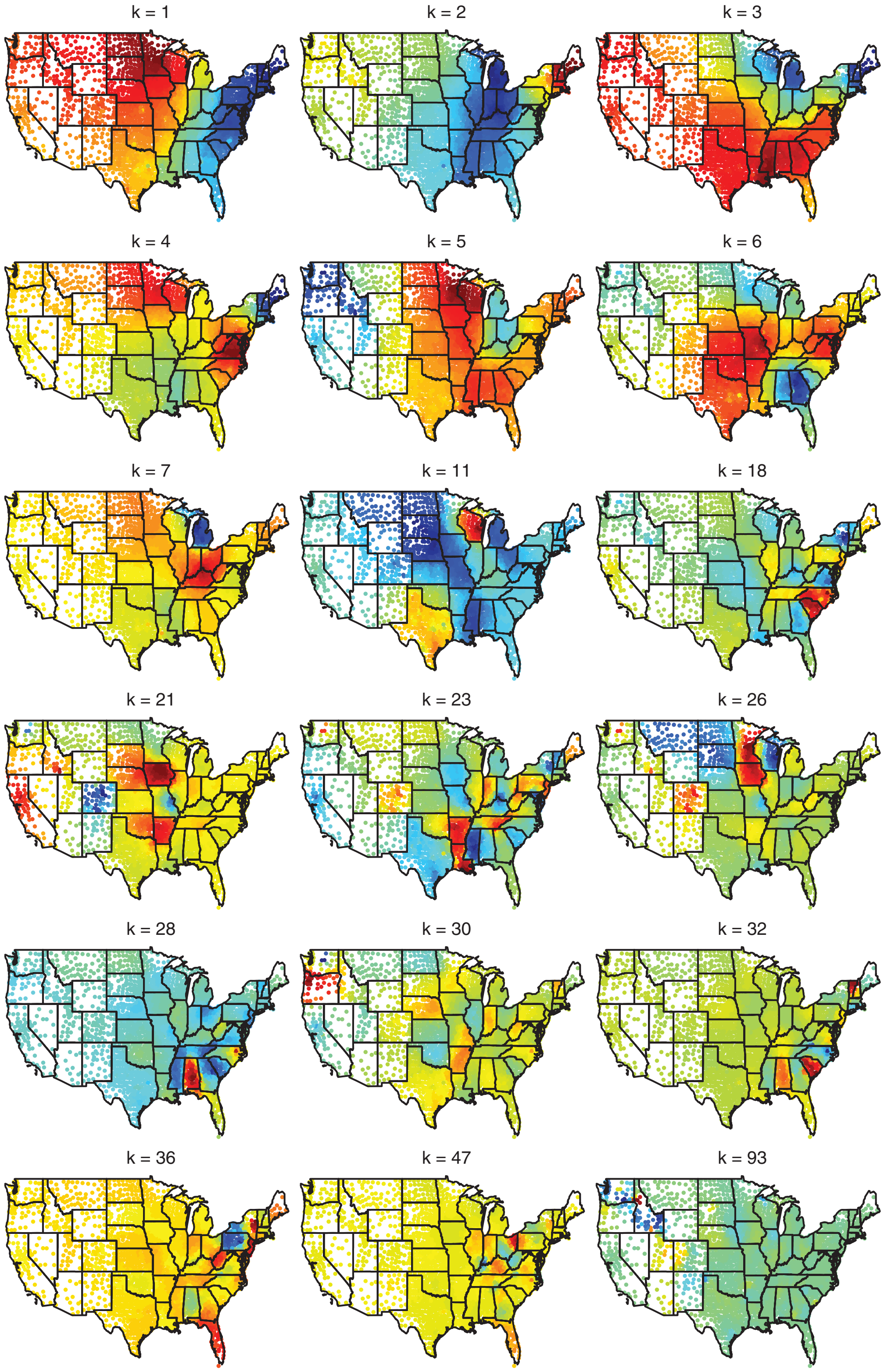
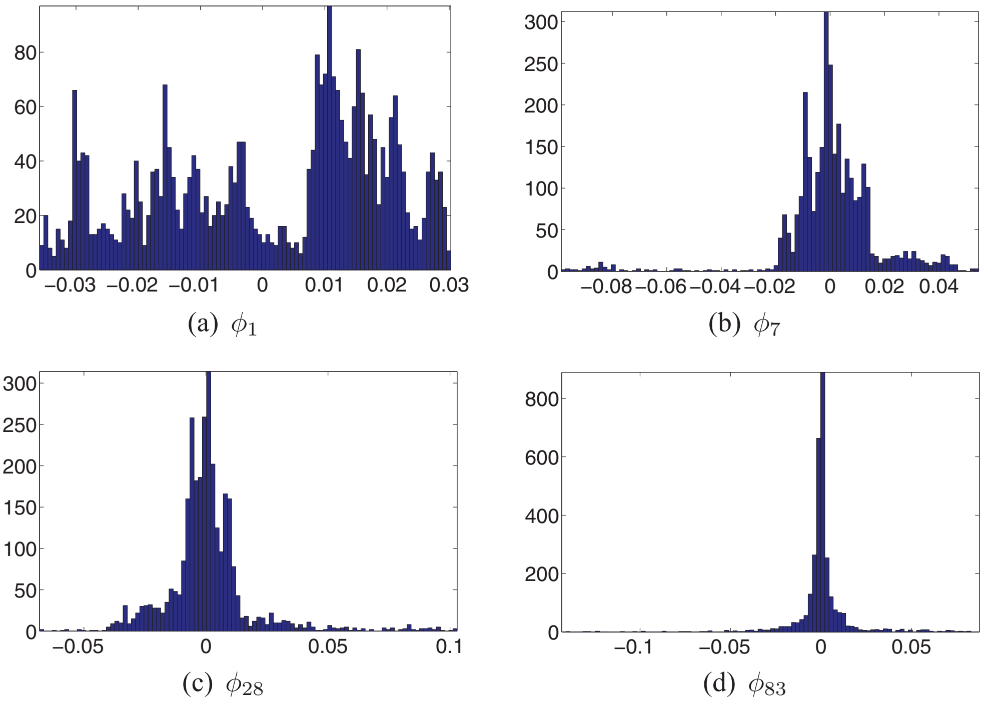


FIG. 3. (Color) Top eigenvector colorings for the similarity matrix $W_{ij} = \frac{M_{ij}^2}{P_i P_j}$.

magnitude or very close to zero. The top three eigenvectors correspond to global cuts between various coasts within the US The only state that stands out individually is Michigan (MI) for $k = 3$, which has rank 2. For $k = 4$, the largest entries correspond to counties in Virginia (VA) which is also ranked 1st, and similarly for Wisconsin (WI) for $k = 5$, ranked 14. For $k = 6$, the states colored in dark red and dark blue are Georgia (GA) with rank 3, and Missouri (MO) of rank 8. When $k = 7$,



(a) ϕ_1 (b) ϕ_7

(c) ϕ_{28} (d) ϕ_{83}

FIG. 4. (Color online) Histogram of the entries in the eigenvectors ϕ_1, ϕ_7, ϕ_{28}, and ϕ_{83} of matrix $A = D^{-1}W^{(1)}$.

Michigan (MI), of rank 2, stands out as the only dark blue colored state. For $k = 8$, we point out Georgia, rank 3, together with Mississippi (MS) of rank 11, and Louisiana (LA) of rank 10. Eigenvector $k = 9$ localizes mostly on Maine (ME) of rank 6, and the New York (NY) area with rank 7. Finally, eigenvector $k = 10$ localizes on a combination of states we already pointed out. We have thus enumerated nine states that stand out in the top ten eigenvector colorings, and all nine of them appear in Table I that ranks the top 14 states in terms of ratio degree. Although this experiment does not provide a formal justification for the eigenvector localization phenomenon, we believe it is a first step in providing evidence that the low-order eigenvectors point out local cuts in the network.

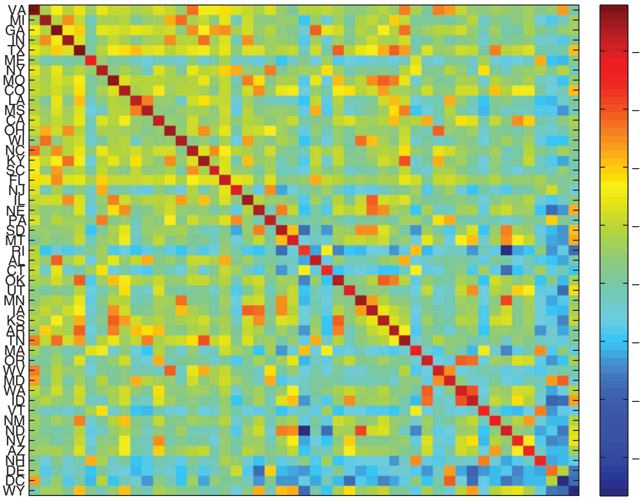


FIG. 5. (Color online) Heat map of the interstate migration flows, where the rows and columns of the matrix are sorted by the ratio degrees of the states. The intensity of entry (i,j) denotes, on a logarithmic scale, the similarity between states i and j, i.e., the sum of all entries in the submatrix $\tilde{W}_{i,j}$ defined in Fig. (4). Table I lists the top 14 states in terms of ratio degree.

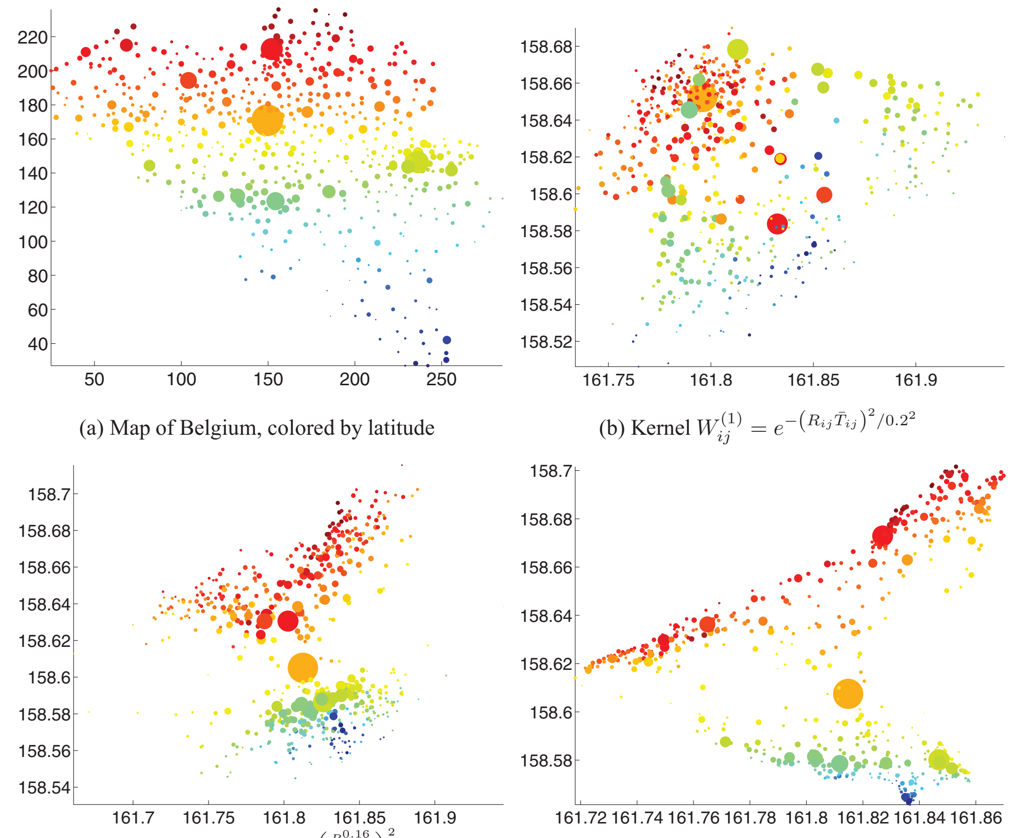
IV. BELGIUM MOBILE NETWORK

In a recent work [27], we studied the anonymized mobile phone communication from a Belgian operator and derived a statistical model of interaction between cities, showing that intercity communication intensity is characterized be a gravity model: the communication intensity between two cities is proportional to the product of their sizes divided by the square of their distance. In this section, we briefly describe the Belgium mobile data set, summarize the results in [27], and apply the diffusion map technique. We refer the reader to [28] for more information on the mobile phone data set.

The data set contains anonymous communication patterns of 2.5 million mobile phone customers, grouped in 571 cities

TABLE I. Top 14 states within the US, ordered by ratio degree.

Rank	State	Ratio degree	Rank	State	Ratio degree
1.	VA	26.7	8.	MO	18.5
2.	MI	20.4	9.	CO	17.1
3.	GA	19.9	10.	LA	16.6
4.	IN	19.7	11.	MS	16.1
5.	TX	19.0	12.	CA	15.7
6.	ME	18.9	13.	OH	15.6
7.	NY	18.7	14.	WI	14.5



(a) Map of Belgium, colored by latitude

(b) Kernel $W^{(1)}_{ij} = e^{-(R_{ij}\bar{T}_{ij})^2/0.2^2}$

(c) Kernel $W^{(2)}_{ij} = e^{-\left(\frac{R_{ij}^{0.16}}{\bar{N}_{ij}^{0.26}}\right)^2}$

(d) Kernel $W^{(3)}_{ij} = \frac{\bar{T}_{ij}}{R_{ij}} = \frac{N_{ij}}{P_i P_j}$

FIG. 6. (Color) Diffusion map reconstructions from the top two eigenvectors, for various kernels. T_{ij} denotes the aggregated communication time in seconds, N_{ij} the number of phone calls between cities i and j, $R_{ij} = \frac{T_{ij}}{N_{ij}}$ the average duration of a call, and P_i the population of city i. We normalize by the population size by defining $\bar{N}_{ij} = \frac{N_{ij}}{P_i P_j}$ and $\bar{T}_{ij} = \frac{T_{ij}}{P_i P_j}$.

in Belgium over a period of six months in 2006 (see also [29] for a description of the data set). Every customer is associated with the ZIP Code of her or his billing address. Calls involving other operators were filtered out, meaning that both the calling and receiving individuals in the data set are customers of the mobile phone company. Also, there is a link between two customers if at least three calls were made in both directions during the six month interval. After this preprocessing, the network has 2.5 million nodes and 38 million links. For every pair of customers we associate a communication intensity by computing the total communication time in seconds. After grouping the customers into their corresponding cities, we compute T_{ij}, the aggregated communication time in seconds between the customers of city i and j, and denote the resulting matrix by $T = (T_{ij})_{1 \leqslant i < j \leqslant n}$. We denote by N_{ij} the number of phone calls between cities i and j, by $R_{ij} = \frac{T_{ij}}{N_{ij}}$ the average duration of a call, and by P_i the number of customers that have the ZIP Code billing address of city i (from now on, we refer to P_i as the population of city i). Furthermore, the normalized number of phone calls with respect to the population of the cities is denoted by $\bar{N}_{ij} = \frac{N_{ij}}{P_i P_j}$, and similarly the normalized communication time by $\bar{T}_{ij} = \frac{T_{ij}}{P_i P_j}$. Finally, $D = (d_{ij})_{1 \leqslant i < j \leqslant n}$ represents the distances between the centroids of the areas of cities i and j. Using these quantities, we now consider the following three kernels: $W^{(1)}_{ij} = e^{-(R_{ij}\bar{T}_{ij})^2/0.2^2}$, $W^{(2)}_{ij} = e^{-(\frac{R_{ij}^{0.16}}{\bar{N}_{ij}^{0.26}})^2}$, and $W^{(3)}_{ij} = \frac{\bar{T}_{ij}}{R_{ij}} = \frac{N_{ij}}{P_i P_j}$.

Figure 6 shows the diffusion map reconstructions for various matrices W that relate cities based on their communication intensities and population sizes. For $W^{(2)}$ and $W^{(3)}$, there is an obvious separation between the north and south parts of Belgium, which stems from the fact that the two regions belong to different linguistic groups. The same separation is emphasized by the colorings associated to the top eigenvector of matrix A, shown in Fig. 7. The remaining eigenvector colorings in Fig. 7 clearly isolate various subregions in Belgium. For example, eigenvectors ψ_1 and ψ_{11} highlight language communities (French, Dutch, and German), while ψ_3 and ψ_5 isolate the regions of Liège and Limburg.

V. LOCALIZED EIGENFUNCTIONS

Let us first make more precise what is meant by a localized eigenfunction. This phenomenon of localization occurs when there exist eigenfunctions supported by small regions of the domain, i.e., they are localized in these regions. An

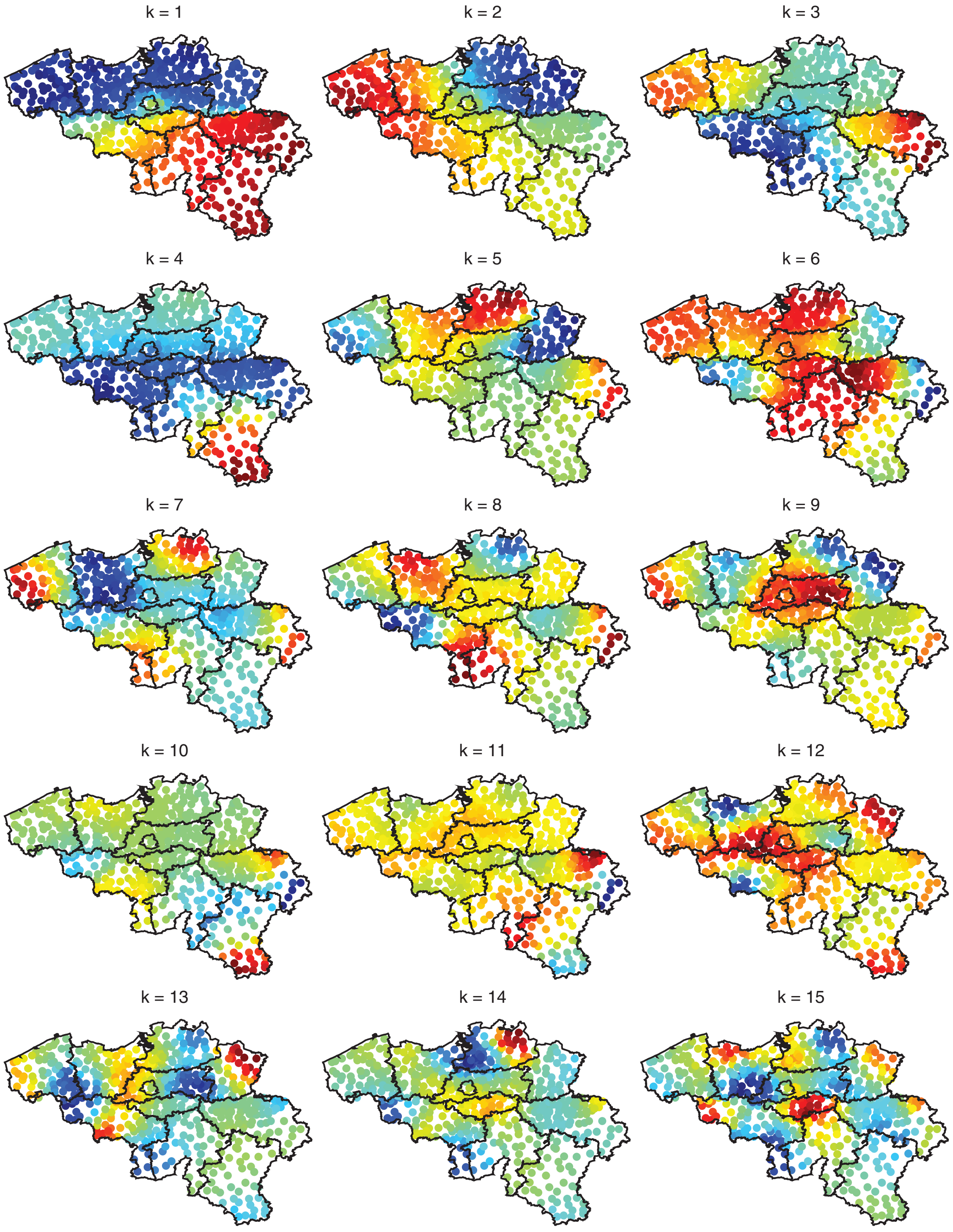


FIG. 7. (Color) Colorings by the top 18 eigenvectors of $A = D^{-1}W^{(3)}$, where $W^{(3)}_{ij} = \frac{\bar{T}_{ij}}{R_{ij}} = \frac{N_{ij}}{P_i P_j}$.

eigenfunction localized on a domain Ω_1 has support on Ω_1 significantly larger than on the complement $\Omega \backslash \Omega_1$, and yet it cannot vanish on $\Omega \backslash \Omega_1$ since eigenfunctions of isolated eigenvalues are real analytic functions and cannot vanish on any open set. This is also observed in the histogram of the entries of the eigenvectors ϕ_7, ϕ_{28}, and ϕ_{83} shown in Fig. 4 and the corresponding colorings in Fig. 3. In contrast, eigenfunctions that do not localize have their support "uniformly" distributed across the domain, similar to the case of eigenvector ϕ_1 from Fig. 4. For example, in the case of the unit interval, the eigenfunctions of the Laplacian are the sine or cosine functions (depending on the boundary conditions) with the larger eigenvalues corresponding to higher oscillations, and they are not localized in the sense that there is no specific subinterval that carries the most (potential) energy of the eigenfunction, and any subinterval supports an amount of energy that is proportional to its length. In other words, the energy of the top eigenfunctions is distributed uniformly across the domain, and similar results are known to hold for the disk and the sphere, where the Laplacian eigenvalues and eigenfunctions are explicitly known.

The behavior observed in the eigenvector colorings from Figs. 3 and 7 is related to the notion of *localized eigenfunctions*, a phenomenon observed before in the mathematics and physics community. The spectrum of the (continuous) Laplace operator has been extensively studied, and there exists a rich literature on the relationship between the spectrum and the geometry of the domain. As more complicated objects, eigenfunctions are more difficult to analyze than the spectrum, and less is known about them. Most of the literature is focused on high-frequency eigenfunctions (associated to larger eigenvalues), such as [30–32], although recent studies such as [33] advocated localized eigenfunctions associated to small eigenvalues. In our experiments, we found the bottom eigenvectors uninteresting as they did not contain any meaningful geometric information. In his work, Sapoval [34] studied localized eigenfunctions in different domains and pointed out their importance for physical applications, such as designing efficient noise-protective walls.

Finally, considering that A is a stochastic matrix, one may further explore ideas from the theory of nearly completely decomposable matrices developed by Simon and Ando to describe and identify the short-, medium-, and long-term behaviors of a dynamical system [35]. Building on earlier work [36], the very recent article of [37] explores this idea in the context of stochastic data clustering and proposes a technique that uses the evolution of the system to infer information on the initial structure.

VI. SUMMARY AND DISCUSSION

We have shown how the diffusion map technique can be used to obtain informative visualizations and capture natural subdivisions within two different real networks. We find surprisingly that some low-order eigenvectors localize very well and seem to reveal small geographically cohesive regions; it is natural to ask for an explanation for our observation.

In looking at Fig. 3 many more questions come to mind. Are the state boundaries a consequence of people migrating within the same state or not? In other words, do states emerge as communities because of people migrating from one county to the other within the state, or because of similar migration patterns directed outside the state? Preliminary analysis on the migration data set in the context of local clustering on graphs supports the idea that the localized low-order eigenvectors highlight local cuts in the network. This is perhaps counterintuitive since such low-order eigenvectors must satisfy the global requirement of exact orthogonality with respect to all of the earlier delocalized eigenvectors, and they must do so while keeping most of their components zero or close to zero. Another question to consider is whether, besides the state boundary detection, the eigenvector colorings reveal any extra information on the intensity of the migration from one region to the other. Furthermore, intercounty migration is most common among young adults and declines as people age, and one may ask how the age composition (or income level) of individual US counties impacts the migration pattern.

In answering these questions, one needs to complement the mathematical description of diffusion maps and clustering by eigenvectors with a sociodemographic behavioral interpretation of migration trends, as considered for example in [38,39]. A more recent paper by Slater [40] is of particular interest since it analyzes migration patterns in the US Census data from 1965 to 1970 and 1995 to 2000. Amongst others, it highlights cosmopolitan or hublike regions, as well as isolated regions that emerge when there is a high measure of separation between a cluster and its environment.

Another interesting direction worth exploring is seeing how the diffusion map reconstructions and colorings change when the matrices used are no longer symmetric. In the case of the US migration data, it may be the case that there are many states for which the most common migration destination is the major city or capital of that state (although there might be other destinations spread across the US that attract people migrating out from that state). It is therefore natural to expect that major cities will stand out in the colorings; however, this is not the case in our simulations since we symmetrize the migration matrix and take into account both the in and out migration from a given state.

ACKNOWLEDGMENTS

We would like to thank Amit Singer for introducing M.C. to diffusion maps and suggesting this as a possible approach for analyzing the Belgium mobile data set, Etienne Huens and Gautier Krings for useful discussions and help with processing the same data set, Michael Mahoney, Mauro Maggioni, Peter Mucha, Mark Newman, and Mason Porter for useful discussions on community detection and references to the literature on localized eigenfunctions, and Thomas Espenshade, Kevin O'Neil, John Palmer, and Ashton Verdery for references and useful discussions on demographic trends in US migration.

[1] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[2] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. **10**, 053039 (2008).

[3] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, PLoS ONE **5**, e14248 (2010).

[4] J. Shi and J. Malik, IEEE Trans. Pattern Anal. Mach. Intel. **22**, 888 (2000).

[5] J. Reades, F. Calabrese, and C. Ratti, Environ. Plan. B **36**, 824 (2009).

[6] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).

[7] T. Richardson, P. J. Mucha, and M. A. Porter, Phys. Rev. E **80**, 036111 (2009).

[8] P. Van Mieghem, X. Ge, P. Schumm, S. Trajanovski, and H. Wang, Phys. Rev. E **82**, 056113 (2010).

[9] A. Arenas, J. Borge-Holthoefer, S. Gomez, and G. Zamora-Lopez, New J. Phys. **12**, 053009 (2010).

[10] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proc. Natl. Acad. Sci. USA **108**, 7663 (2011).

[11] J. P. Onnela, S. Arbesman, M. C. González, A. L. Barabási, and N. A. Christakis, PLoS one **6**, e16939 (2011).

[12] M. W. Berry, S. T. Dumais, and G. W. O'Brien, SIAM Rev. **37**, 573 (1995).

[13] T. K. Landauer and S. T. Dumais, Psychol. Rev. **104**, 211 (1997).

[14] M. Boguñá, D. Krioukov, and K. C. Claffy, Nature Phys. **5**, 74 (2009).

[15] M. Boguñá, F. Papadopoulos, and D. Krioukov, Nat. Commun. **1**, 62 (2010).

[16] D. Liu, H. Wang, and P. Van Mieghem, Phys. Rev. E **81**, 016101 (2010).

[17] S. Lafon, Ph.D. thesis, Yale University, 2004.

[18] M. Belkin and P. Niyogi, Neural Comput. **6**, 1373 (2003).

[19] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, SIAM Multiscale Model. Simul. **7**, 842 (2008).

[20] R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).

[21] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. USA **102**, 7426 (2005).

[22] L. Lovász, Combinatorics, Paul Erdos is Eighty **2**, 1 (1993).

[23] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, in *Advances in Neural Information Processing Systems 18* (MIT Press, Cambridge, MA, 2005), pp. 955–962.

[24] US Census Bureau, 2002, www.census.gov/population/www.cen2000/ctytoctyflow/index.html.

[25] M. J. Perry, Census 2000 Special Reports, 2003.

[26] U. von Luxburg, Stat. Comput., Springer Netherlands **17**, 395 (2007).

[27] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, J. Stat. Mech.: Theory Exp. (2009) L07003.

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.: Theory Exp. (2008) P10008.

[29] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, Physica A **387**, 5317 (2008).

[30] N. Burq and M. Zworski, SIAM Rev. **47**, 43 (2005).

[31] V. M. Babič and V. F. Lazutkin, Problems Math. Phys., Spectral Theory, Diffract. Problems (Russian) **2**, 15 (1967).

[32] P. W. Jones, M. Maggioni, and R. Schul, Proc. Natl. Acad. Sci. USA **6**, 1803 (2008).

[33] S. M. Heilman and R. S. Strichartz, Notices Am. Math. Soc. **57**, 624 (2010).

[34] S. Russ, B. Sapoval, and O. Haeberlé, Phys. Rev. E **55**, 1413 (1997).

[35] H. A. Simon and A. Ando, Econometrica **29**, 111 (1961).

[36] C. D. Meyer, SIAM Rev. **31**, 240 (1989).

[37] C. D. Meyer and C. D. Wessell, arXiv:1008.1758.

[38] E. S. Lee, Pop. Assoc. Am., Demogr. **3**, 47 (1966).

[39] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor, Pop. Dev. Rev. **19**, 431 (1993).

[40] P. B. Slater, ISBER, University of California, Santa Barbara, 2008, arXiv:0809.2768v3.