SIAM J. MATH. DATA SCI.
Vol. 3, No. 2, pp. 467–493

A Metric on Directed Graphs and Markov Chains Based on Hitting Probabilities*

Zachary M. Boyd†, Nicolas Fraiman‡, Jeremy Marzuola†, Peter J. Mucha§, Braxton Osting¶, and Jonathan Weare‖

Abstract. The shortest-path, commute time, and diffusion distances on undirected graphs have been widely employed in applications such as dimensionality reduction, link prediction, and trip planning. Increasingly, there is interest in using asymmetric structure of data derived from Markov chains and directed graphs, but few metrics are specifically adapted to this task. We introduce a metric on the state space of any ergodic, finite-state, time-homogeneous Markov chain and, in particular, on any Markov chain derived from a directed graph. Our construction is based on hitting probabilities, with nearness in the metric space related to the transfer of random walkers from one node to another at stationarity. Notably, our metric is insensitive to shortest and average walk distances, thus giving new information compared to existing metrics. We use possible degeneracies in the metric to develop an interesting structural theory of directed graphs and explore a related quotienting procedure. Our metric can be computed in $O(n^3)$ time, where n is the number of states, and in examples we scale up to $n = 10,000$ nodes and $\approx 38M$ edges on a desktop computer. In several examples, we explore the nature of the metric, compare it to alternative methods, and demonstrate its utility for weak recovery of community structure in dense graphs, visualization, structure recovering, dynamics exploration, and multiscale cluster detection.

Key words. directed graph, metric space, Markov chain

AMS subject classifications. 05C20, 05C50, 05C62, 05C82, 05C85, 60J10, 60J22, 62H11, 62H30, 62M15, 68W40

DOI. 10.1137/20M1348315

*Received by the editors June 26, 2020; accepted for publication (in revised form) January 14, 2021; published electronically April 19, 2021.

https://doi.org/10.1137/20M1348315

Funding: The first and fourth authors were supported by the James S. McDonnell Foundation 21st Century Science Initiative-Complex Systems Scholar Award grant 220020315, with additional support from the Army Research Office (MURI award W911NF-18-1-0244). The third author was supported in part by NSF CAREER grant DMS 13-52353 and NSF grant DMS 19-09035 and thanks Duke University and MSRI for hosting him during part of the completion of this work. The fifth author acknowledges partial support from NSF DMS 16-19755 and 17-52202. The sixth author was supported by the Advanced Scientific Computing Research Program within the DOE Office of Science through award DE-SC0020427. The content is solely the responsibility of the authors and does not necessarily reflect the views of any of the funding agencies.

†Department of Mathematics, University of North Carolina–Chapel Hill, Chapel Hill, NC 27514 USA (zachboyd@email.unc.edu, marzuola@email.unc.edu).

‡Department of Statistics and Operations Research, University of North Carolina–Chapel Hill, Chapel Hill, NC 27514 USA (fraiman@email.unc.edu).

§Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics and Department of Applied Physical Sciences, University of North Carolina–Chapel Hill, Chapel Hill, NC 27514 USA (mucha@unc.edu).

¶Department of Mathematics, University of Utah, Salt Lake City, UT 84112 USA (osting@math.utah.edu).

‖Courant Institute of Mathematical Sciences, New York University, New York, NY 10012-1185 USA (weare@nyu.edu).

1. Introduction.

1.1. Motivation. Many finite spaces can be endowed with meaningful metrics. For undirected graphs, the geodesic (shortest-path), commute time (effective resistance), and diffusion distance [28, 16, 15] metrics are widely applied [16, 31, 2]. The first two can be naively generalized to directed graphs by summing shortest/average walk length in each direction, whereas the third is specifically undirected. We know of only one graph metric specifically designed for directed graphs, namely, the generalized effective resistance distance developed in [54, 55]. Overlaying a metric onto a directed structure is a challenge since, by definition, the metric is symmetric.

A related problem is finding metrics on the state space of a finite-state, discrete-time Markov chain. In this case, there is also limited prior work, consisting of mean commute time [42, 7, 11] and a constant-curvature metric [52].

Metrics fit into the broader category of dissimilarity measures, with the decision whether to impose all metric axioms being application dependent. When a metric is used, this additional structure can enable various algorithmic accelerations, improved guarantees, and useful inductive biases [19, 36, 23, 4, 41]. Furthermore, the metric structure is a key ingredient in proofs of convergence, consistency, and stability. While mostly settled for undirected graphs [38, 45, 46, 48, 49], the development of such theories for directed graphs (digraphs) and Markov chains is an open research problem. The first positive result for digraphs appeared recently [56].

In the present work, we introduce and analyze a new metric for digraphs and Markov chains based on the *hitting probability* from one node to another, by which we mean the probability that a random walker starting at one node will reach the other node before returning to its starting node. By correctly combining these probabilities with the invariant distribution of an irreducible Markov chain, a metric can be constructed. This metric differs from other metrics by being insensitive to walk length, thus measuring information that is, in a sense, orthogonal to commute time, as illustrated in examples. In the special case of undirected graphs and with the scale parameter $\beta = 1$ (defined below), the hitting probabilities metric is actually the logarithm of effective resistance/commute time (plus a constant), a striking fact proven in [18, section 1.3.4]. For other values of the scale parameter, the hitting probabilities metric is a new addition to the limited catalogue of undirected graph metrics. We illustrate the utility of our metric in several examples, both analytical and numerical, related to graph symmetrization, clustering, structure detection, data exploration, and geometry detection.

1.2. Our contributions. Let $(X_t)_{t\geq 0}$ be a discrete-time Markov chain on the state space $[n] = \{1, \dots, n\}$ with initial distribution λ and irreducible transition matrix P, i.e.,

$$\mathbf{P}(X_0 = i) = \lambda_i \qquad \text{and} \qquad \mathbf{P}(X_{t+1} = j \mid X_t = i) = P_{i,j}\,.$$

We emphasize that X is not required to be aperiodic.

Let $\phi \in \mathbb{R}^n_+$ be the invariant distribution for P, i.e., $P^T\phi = \phi$. The *hitting time* (starting from a random state distributed like λ) for a state $i \in [n]$ is the random variable given by

$$\tau_i := \inf\{t \geq 1 \colon X_t = i\}\,.$$

For $i, j \in [n]$, let us define

$$Q_{i,j} := \mathbf{P}_i[\tau_j < \tau_i]\,, \tag{1.1}$$

which denotes the probability that starting from site i (i.e., the subscript on $\mathbf{P}_i$ is used to indicate that $\lambda = \delta_i$) the hitting time of j is less than the time it takes to return to i. We emphasize that we consider $\tau_j < \tau_i$ here for a single walk and take the probability of such an event over all walks starting at i when computing $Q_{i,j}$. An expression for the *hitting probability matrix*, Q, in terms of the transition matrix will be given in (3.1); see section 3 on computational methods.

Lemma 1.1. *The following relationship holds*[1] *for* $i \neq j$*:*

$$Q_{i,j}\phi_i = Q_{j,i}\phi_j\,. \tag{1.2}$$

The weighting by the invariant measure is motivated by connections between the invariant measure and random walks as found in [37, section 1.7]. A proof of Lemma 1.1 is given in section 2.

Remark 1.2. Lemma 1.1 implies that, with appropriate choice of Q_{ii}, $\frac{1}{n}Q$ is a reversible Markov chain with invariant distribution ϕ.

We define the *normalized hitting probabilities matrix*, $A^{(\mathrm{hp},\beta)} \in \mathbb{R}^{n\times n}$, by

$$A^{(\mathrm{hp},\beta)}_{i,j} := \begin{cases} \dfrac{\phi_i^\beta}{\phi_j^{1-\beta}} Q_{i,j}, & i \neq j, \\ 1, & i = j, \end{cases} \tag{1.3}$$

where $\beta \in [1/2, \infty)$. In contexts where the choice of β is not important, we simply write $A^{(\mathrm{hp})} = A^{(\mathrm{hp},\beta)}$. Two useful choices for β are 1 and 1/2. The $Q_{i,j}$ matrix has recently been shown to play a key role in determining the error of a family of stratified Markov chain Monte Carlo methods [17, 47].

From Lemma 1.1, we immediately have the following corollary.

Corollary 1.3. *The matrix* $A^{(\mathrm{hp},\beta)}$ *defined in* (1.3) *is symmetric. In particular,*

$$A^{(\mathrm{hp},1/2)}_{i,j} = \sqrt{Q_{i,j}Q_{j,i}}. \tag{1.4}$$

Proof. We observe

$$\begin{aligned} A^{(\mathrm{hp},\beta)}_{i,j} &= \frac{\phi_i^\beta}{\phi_j^{1-\beta}} Q_{i,j} = \frac{\phi_i^{\beta-1}}{\phi_j^{1-\beta}} \phi_i Q_{i,j} \\ &= \frac{\phi_i^{\beta-1}}{\phi_j^{1-\beta}} \phi_j Q_{j,i} = \frac{\phi_j^\beta}{\phi_i^{1-\beta}} Q_{j,i} = A^{(\mathrm{hp},\beta)}_{j,i}. \end{aligned}$$

Hence, $A^{(\mathrm{hp},\beta)}$ is symmetric.

To prove (1.4), we observe that $\left(A^{(\mathrm{hp},1/2)}_{i,j}\right)^2 = \frac{\phi_i}{\phi_j} Q_{i,j}^2 = Q_{i,j}Q_{j,i}$ by (1.2). ∎

[1] Lemma 1.1 was previously (and independently) proven in [10], in the context of Markov chain perturbation theory applied to the internet. It was possibly known even earlier.

In some applications, information about relatedness of vertices in a graph will be most immediately encoded in the form of a nonstochastic adjacency matrix A. In this case the input adjacency matrix can be transformed into a stochastic matrix P either by a similarity transformation involving the dominant right eigenvector of A or by normalization of the rows of A so that they sum to 1. The resulting stochastic matrix P can then be used as in (1.3) to construct $A^{(\mathrm{hp},\beta)}$, itself a symmetric adjacency matrix on the vertices of the network. In this article we do not address the relative merits of methods to transform an adjacency matrix into a stochastic matrix. We use row normalization unless otherwise stated.

Given an irreducible stochastic matrix P, we can thus define a distance $d^\beta\colon [n]\times[n]\to\mathbb{R}$, which we refer to as the *hitting probability metric*, by

$$d^\beta(i,j) = -\log\left(A^{(\mathrm{hp},\beta)}_{i,j}\right). \tag{1.5}$$

Theorem 1.4. *The hitting probability metric, $d^\beta\colon [n]\times[n]\to\mathbb{R}$, defined in (1.5) is a metric for $\beta\in(1/2,1]$. For $\beta = 1/2$, d^β is a pseudometric,*[2] *and there exists a quotient graph on which the distance function becomes a metric.*

In Theorem 2.15, we show that there exists a quotient graph on which $d^{1/2}$ is a metric and which preserves many of the metric properties of the original graph.[3] The key observation for the $d^{1/2}$ pseudometric is that in order for two vertices to be distance 0 from each other, the probability of hitting the other vertex before returning must be 1 for both. Hence we provide (in subsections 2.3 and 2.4) a means of effectively collapsing these vertices to a single vertex, carefully preserving the overall probabilities relative to the remaining vertices.

Remark 1.5. In light of Lemma 1.1 and Theorem 1.4, $A^{(\mathrm{hp})}$ has two interpretations, first as a symmetrization of A and second as a weighted similarity graph corresponding to d, since $A(i,j) = e^{-d(i,j)}$. The practice of associating a finite (subset of a) metric space with a similarity graph in this way is widespread, especially in the manifold learning and graph-based machine-learning communities.[4] Thus, in our experiments, we favored the use of $A^{(\mathrm{hp})}$ for certain applications where it seemed more natural.

Finally, we show how advances from [47] enable us to compute the distance matrix in $O(n^3)$ operations, allowing us to scale up to $\approx 38M$ edges in examples on a Lenovo ThinkStation P410 desktop with Xeon E5–1620V4 3.5 GHz CPU and 16 GB RAM using MATLAB R2019a Update 4 (9.6.0.1150989) 64-bit (glnxa64). We also provide various synthetic examples to help develop an intuition for the metric and its differences from other measures. We conclude with an example using New York City taxi data to illustrate how our metric can aid in data exploration.

1.3. Relationship to other notions of similarity and metrics. In this section, we discuss some related notions of similarity and metrics on finite state spaces with asymmetric

[2] Recall that a pseudometric on $[n]$ is a nonnegative real function $f\colon [n]\times[n]\to\mathbb{R}_{\geq 0}$ satisfying $d(i,i)=0$, symmetry $d(i,j)=d(j,i)$, and the triangle inequality $d(i,j)\leq d(i,k)+d(k,j)$. A pseudometric is a metric if we can identify indiscernible values, i.e., $d(i,j)=0 \iff i=j$.

[3] While the usual pseudometric quotienting procedure could apply here, there is no guarantee that there would be a corresponding subgraph, which is why Theorem 2.15 is needed.

[4] [57] cites [25, 44] as this similarity function's first use specifically for graph-based clustering.

(directional) relationships. Our focus is on symmetric notions of dissimilarity, with an emphasis on metrics. While, in some applications, asymmetric similarity scores may be the right choice (see, e.g., Tversky's seminal work on features of similarity [50]), we restrict our scope to symmetric notions. We do, however, wish to mention directed metrics (also called quasi-metrics), which are a natural analogue to metric spaces for relaxations of digraph cut problems [6].

From [7, 27, 42], we know that commute time is a metric on ergodic Markov chains. In [11, 54, 55], generalizations of effective resistance are developed for ergodic Markov chains and directed graphs. Commute-time- and resistance-based metrics are popular and more robust than shortest-path distances, although they are not informative in certain large-graph limits [53]. In subsection 4.2.1, we compare the effective resistance of [54, 55] to the hitting probability metric on a particular example.

In [52], a metric is developed on Markov chains. This metric gives the chain constant curvature in an appropriate generalized sense. Distance in this metric is then related to the distinguishability after one step of random walks beginning at the two distinct nodes. The metric is constructed jointly with the curvature using a fixed point argument. It is expected to be useful in proving, for example, concentration inequalities for Markov chains.

Notions of diffusion distance to a set B on undirected graphs have been explored recently for the connection Laplacian [45] and for the graph Laplacian [9]. The notion of distance is determined by taking ℓ steps using the random walk generated by the symmetric graph adjacency matrix A with degree matrix D, i.e., it counts the number of walks of length 2ℓ from i to j. Diffusion distances from a vertex i to a subgraph B in [9] is defined as the smallest number of steps for all random walks started at i to reach B. The work [45] established that diffusion distances converge to geodesic distances in the high density limit of random graphs on manifolds, and [9] explored how eigenvectors relate to this notion of distance. Directed graphs have been represented as magnetic connection Laplacians on undirected graphs through a notion of polarization (see [20]), after which a version of diffusion distance can be applied.

A variety of methods exist in machine learning to compute "graph representations," which are learned embeddings of nodes, subgraphs, or entire (possibly directed) graphs into Euclidean space so that they can be fed into standard machine learning tools [24]. These can be seen as imposing a metric on directed graphs, with the main drawbacks relative to the hitting probability metric being model complexity, difficulty of interpretation, and difficulty of analysis.

In [33], existing symmetrization techniques for directed graphs are surveyed. In particular, we mention [43, 58, 29, 8]. In each of these articles, clustering, community detection, and/or semi-supervised learning techniques are considered on directed graphs using various symmetrizations, such as that of Fan Chung (e.g., [43]) or using commute times similar to those in the effective resistance metric (e.g., [8]). Our results use $A^{(\mathrm{hp})}$ as a symmetrization, and we will see that this enables us to perform the tasks just mentioned, although with different and sometimes more helpful results.

In [21], the metric of [55, 54] is used as the basis for a digraph symmetrization technique. It is guaranteed to preserve effective resistances, possibly relying on negative entries. Rigorous applications to directed cut and graph sparsification are given.

Outline. We prove Lemma 1.1, Corollary 1.3, and Theorem 1.4 in section 2. In section 3, we describe computational methods to compute the normalized hitting probabilities matrix, $A^{(\mathrm{hp},\beta)}$. In section 4, we give some examples of the computed metric. We conclude in section 5.

2. Proofs and discussion of structural properties.

2.1. Structure of the normalized hitting probabilities matrix.

Proof of Lemma 1.1. The probability that X_t starts from i and hits j at least $k+1$ times before returning to i can be expressed as

$$\mathbf{P}_i[\tau_j < \tau_i]\mathbf{P}_j[\tau_j < \tau_i]^k .$$

We let V_i^j be the number of times X_t hits j before returning to i, $V_i^j = \sum_{t=1}^{\tau_i} 1_{X_t=j}$. Then, we have

$$\mathbf{P}_i[\tau_j < \tau_i]\mathbf{P}_j[\tau_j < \tau_i]^k = \mathbf{P}_i[V_i^j \geq k+1] .$$

Now observe that

$$\sum_{k=0}^{\infty} \mathbf{P}_i[\tau_j < \tau_i]\mathbf{P}_j[\tau_j < \tau_i]^k = \sum_{k=0}^{\infty} \mathbf{P}_i[V_i^j \geq k+1] = \sum_{k=0}^{\infty} \mathbf{E}_i\left[1_{V_i^j \geq k+1}\right] \tag{2.1}$$

$$= \mathbf{E}_i\left[\sum_{k=0}^{\infty} 1_{V_i^j \geq k+1}\right] = \mathbf{E}_i\left[\sum_{k=0}^{\infty} k 1_{V_i^j = k}\right] = \sum_{k=0}^{\infty} k\mathbf{P}[V_i^j = k]$$

$$= \mathbf{E}_i[V_i^j] . \tag{2.2}$$

The expectation in (2.2) is known to satisfy

$$\mathbf{E}_i[V_i^j] = \frac{\phi_j}{\phi_i} , \tag{2.3}$$

which is proved in, for example, [37, Theorem 1.7.6].

However, we recognize the expression (2.1) as a geometric series and hence have

$$\sum_{k=0}^{\infty} \mathbf{P}_i[\tau_j < \tau_i]\mathbf{P}_j[\tau_j < \tau_i]^k = \mathbf{P}_i[\tau_j < \tau_i](1 - \mathbf{P}_j[\tau_j < \tau_i])^{-1}$$

$$= \mathbf{P}_i[\tau_j < \tau_i]\mathbf{P}_j[\tau_i < \tau_j]^{-1} = Q_{i,j}Q_{j,i}^{-1} .$$

Combining this with (2.3) we arrive at $Q_{i,j}Q_{j,i}^{-1} = \frac{\phi_j}{\phi_i}$. ∎

To prove Theorem 1.4, we will need one more lemma.

Lemma 2.1. *The following inequality holds:*

$$Q_{i,j} \geq Q_{i,k}Q_{k,j} . \tag{2.4}$$

Proof. Consider the corresponding auxiliary Markov process restricted to nodes i, j, and k with 3×3 transition matrix F, the elements of which we denote by, e.g., $F_{i,j} = \mathbf{P}_i[\tau_j < \min\{\tau_i, \tau_k\}]$ and $F_{i,i} = \mathbf{P}_i[\tau_i < \min\{\tau_j, \tau_k\}]$. That is, $F_{i,j}$ gives the probability of a random walker starting at i eventually reaching j before either reaching k or returning to i, while $F_{i,i}$ gives the probability of a random walker starting at i returning to i before reaching either j or k. Since $F_{i,i} + F_{i,j} + F_{i,k} = 1$, we have

$$Q_{i,j} = F_{i,j} + \frac{F_{i,k}F_{k,j}}{1 - F_{k,k}} , \quad Q_{i,k} = F_{i,k} + \frac{F_{i,j}F_{j,k}}{1 - F_{j,j}} , \quad Q_{k,j} = F_{k,j} + \frac{F_{k,i}F_{i,j}}{1 - F_{i,i}} .$$

Hence, we observe

$$Q_{i,k}Q_{k,j} = \left(F_{i,k} + \frac{F_{i,j}F_{j,k}}{1-F_{jj}}\right)\left(F_{k,j} + \frac{F_{k,i}F_{i,j}}{1-F_{ii}}\right)$$
$$= F_{i,k}F_{k,j} + F_{i,j}\left(\frac{F_{j,k}F_{k,j}}{1-F_{jj}} + \frac{F_{i,k}F_{k,i}}{1-F_{ii}} + \frac{F_{j,k}F_{k,i}F_{i,j}}{(1-F_{ii})(1-F_{jj})}\right).$$

Using

$$\frac{F_{j,k}}{1-F_{j,j}} = \frac{F_{j,k}}{F_{j,i}+F_{j,k}} < 1,$$

we then observe

$$\begin{aligned} Q_{i,k}Q_{k,j} &\le F_{i,k}F_{k,j} + F_{i,j}\left(F_{k,j} + \frac{F_{i,k}F_{k,i}}{1-F_{ii}} + \frac{F_{k,i}F_{i,j}}{1-F_{ii}}\right) \\ &= F_{i,k}F_{k,j} + F_{i,j}\left(F_{k,j} + F_{k,i}\frac{F_{i,k}+F_{i,j}}{1-F_{ii}}\right) \\ &= F_{i,k}F_{k,j} + F_{i,j}\left(F_{k,j}+F_{k,i}\right) \\ &= \left(\frac{F_{i,k}F_{k,j}}{1-F_{k,k}} + F_{i,j}\right)(1-F_{k,k}) \\ &= Q_{i,j}(1-F_{k,k}) \le Q_{i,j}. \end{aligned}$$

■

2.2. Hitting probability metric. In this section we establish Theorem 1.4. In particular, we explore the notion that, much like effective resistance, the normalized hitting probabilities matrix provides a natural notion of distance on the digraph (or between states of a Markov chain).

To begin, we recall the definition of d^β from (1.5) and note that we have already established the symmetry $d^\beta(i,j) = d^\beta(j,i)$ for all i,j. As seen from the statement of Theorem 1.4, we will observe that the triangle inequality holds for all $\beta \ge 1/2$ and that positivity holds for all $\beta > 1/2$. In the case $\beta = 1/2$, $d^{1/2}$ gives a pseudometric structure, as there can indeed exist structures in a directed graph or Markov chain on which $d^{1/2}(i,j) = 0$ and $i \ne j$. As an example, consider the nodes on a cycle with in-degree and out-degree 1. (See section 4.)

When $d^{1/2}(i,j) = 0$ there are specific structures that restrict all random walks leaving i so that they must hit j before returning to i. We show in Theorem 2.15 that for any graph, there exists a canonical quotient graph on which $d^{1/2}$ is indeed a metric that is closely related to $d^{1/2}$ on the original graph. Let us now proceed to the proofs.

Proof of Theorem 1.4. First, we show positivity for $i \ne j$. Note that $d^\beta(i,j) > 0$ if and only if $A^{(\mathrm{hp},\beta)}_{i,j} < 1$. Consider the converse:

$$1 \le A^{(\mathrm{hp},\beta)}_{i,j} = \frac{\phi_i^\beta}{\phi_j^{1-\beta}}Q_{i,j} = \frac{\phi_j^\beta}{\phi_i^{1-\beta}}Q_{j,i},$$

that is,

$$\frac{\phi_j^{1-\beta}}{\phi_i^\beta} \le Q_{i,j} \quad \text{and} \quad \frac{\phi_i^{1-\beta}}{\phi_j^\beta} \le Q_{j,i}.$$

Then if $\beta > 1/2$, we have

$$1 \geq Q_{i,j}Q_{j,i} \geq \phi_j^{1-2\beta}\phi_i^{1-2\beta} > 1\,,$$

a contradiction. For $\beta = 1/2$, the last inequality above becomes an equality, so the corresponding argument by contradiction requires only that $A^{(\mathrm{hp},1/2)} \leq 1$ and thus $d^{1/2}(i,j) \geq 0$.

Symmetry follows from Corollary 1.3, and $d^\beta(i,i) = 0$ is immediate, so all that remains is the triangle inequality.

To prove the triangle inequality, we observe for $i \neq j \neq k$ that

$$\begin{aligned} d^\beta(i,j) &= -\log\left(A_{i,j}^{(\mathrm{hp},\beta)}\right) = -\beta\log\phi_i - (\beta-1)\log\phi_j - \log Q_{i,j} \\ &= d^\beta(i,k) + d^\beta(k,j) + (2\beta-1)\log\phi_k + [\log Q_{i,k} + \log Q_{k,j} - \log Q_{i,j}]\,, \end{aligned} \tag{2.5}$$

which, applying Lemma 2.1, proves that the triangle inequality holds for all $\beta \geq 1/2$. ∎

We observe that the $2\beta - 1$ coefficient of $\log\phi_k$ in (2.5) vanishes when $\beta = 1/2$; hence the triangle inequality is as tight as possible (since $Q_{i,k}Q_{k,j}/Q_{i,j} = 1$, e.g., for a directed cycle graph). From the above argument, we can see that the only obstruction to $d^{\frac{1}{2}}$ being a metric is if there is a pair i,j such that

$$A^{(\mathrm{hp},1/2)} = \sqrt{\frac{\phi_j}{\phi_i}}Q_{j,i} = \sqrt{\frac{\phi_i}{\phi_j}}Q_{i,j} = 1\,.$$

In this case,

$$Q_{j,i} = \sqrt{\frac{\phi_i}{\phi_j}} \quad \text{and} \quad Q_{i,j} = \sqrt{\frac{\phi_j}{\phi_i}}\,.$$

Thus, $Q_{i,j}Q_{j,i} = 1$, which, as they are both probabilities, means in fact $Q_{i,j} = Q_{j,i} = 1$. Hence, also $\phi_i = \phi_j$.

Observation 2.2. The condition $\phi_i = \phi_j$ is not an extra restriction beyond $Q_{i,j} = Q_{j,i} = 1$: if $Q_{i,j} = Q_{j,i} = 1$, then a random walker must visit i every time it visits j (and vice versa), and hence the invariant probabilities of sites i and j must be equivalent.

2.3. Structure theory of digraphs where $d^{1/2}$ is not a metric. In this subsection, we investigate the structure of graphs where $d^{1/2}$ is not a metric, which we refer to as ($d^{1/2}$-) degenerate. This is useful for understanding our metric embedding and is foundational for subsection 2.4, where we derive the quotienting procedure to repair graph degeneracies. In this section, we first give a general construction to produce degenerate graphs and show that all degenerate graphs can be constructed in this way. Next, we give a general decomposition of degenerate graphs into equivalence classes and their segments.

2.3.1. A general construction for degenerate graphs. A simple example of a graph with $Q_{i,j} = Q_{j,i} = 1$ is a closed, directed cycle. However, we also have the following much more general construction: Take any two directed acyclic graphs, G_1 and G_2. Connect all the leaves (sinks/nodes of out-degree zero) of G_1 to i and all the leaves of G_2 to j. Connect j to all the roots (sources/nodes of in-degree 0) of G_1 and i to all the roots of G_2. Possibly add edges between i and j. Then, for each node k except i and j, replace it with an arbitrary

strongly connected graph H_k (corresponding to an irreducible Markov chain), replacing each edge to (from) k with at least one edge to (from) a node in H_k. The resulting graph is strongly connected and has i and j only reachable through each other.

In fact, all graphs with $Q_{i,j} = Q_{j,i} = 1$ can be constructed this way. To see this, note that i and j must have positive in- and out-degree by strong connectedness. Let C_i be those nodes reachable from i without passing through j, and define C_j similarly. Consider the following claim.

Claim 2.3. $C_i \cup C_j \cup \{i, j\}$ includes all nodes of the graph, and $C_i \cap C_j$ is empty.

Proof. For the first part, consider a fixed node k which is not i or j, together with a shortest path C_{ik} from i to k. If C_{ik} does not pass through j, then $k \in C_i$; otherwise, k is reachable from j without passing through i since C_{ik} is a shortest path.

For the second part, assume otherwise; that is, pick $k \in C_i \cap C_j$. Then there exist (1) a path C_{ik} from i to k not passing through j, (2) a path C_{jk} from j to k not passing through i, and (3) a path C_{kj} from k to j (by strong connectedness). Assume without loss of generality (WLOG) that C_{kj} passes through j only at the end. C_{kj} cannot pass through i since otherwise $C_{ik} + C_{kj}$ contains a walk from i to i without passing through j, violating $Q_{i,j} = 1$. But then $C_{jk} + C_{kj}$ would be a walk from j to j that does not pass through i, contradicting $Q_{j,i} = 1$. ∎

Since C_i and C_j can only be connected through i and j, removing i and j disconnects these two sets. Now consider the subgraphs induced by C_i and C_j, respectively. As can be done with any directed graph, we reduce each of these subgraphs to their quotients under the mutual reachability equivalence relation, yielding a pair of directed acyclic graphs. The next subsection generalizes this decomposition to account for all nodes for which $d^{1/2}$ vanishes rather than a single pair.

2.3.2. Decomposition into equivalence classes and segments. Consider an equivalence class $\alpha = \{a_1, a_2, \ldots, a_K\}$ of nodes under the equivalence relation $i \sim j \Leftrightarrow d_{i,j}^{1/2} = 0$. We refer to a node in a nonsingleton equivalence class as *($d^{1/2}$-) degenerate*. A graph is $d^{1/2}$-degenerate if it has a degenerate node.

Definition 2.4. • *A* walk *is a sequence of nodes* $\{i_1, i_2, \ldots, i_K\}$ *such that* $P_{i_k, i_{k+1}} > 0$ *for* $1 \le k < K$.
• *A walk is* closed *if* $i_K = i_1$.
• *A closed walk is a* commute *from* i_1 *if* $i_k \ne i_1$, *for* $1 < k < K$.
• *A walk is a* path *if* $i_k \ne i_{k'}$ *when* $k' \ne k$. *A commute is a* cycle *if it is a path when the last element is removed.*

Lemma 2.5. *A commute from* $a_k \in \alpha$ *must include each of the other members of* α *exactly once in an order that depends only on the graph.*

Proof. The proof is in three assertions. We assume WLOG that commutes are from a_1. First, each a_k is visited. This is the same as claiming that $Q_{a_1, a_k} = 1$, which was shown in the proof of Theorem 1.4. Second, each a_k is visited at most once, since $Q_{a_k, a_1} = 1$. Lastly, each a_k is visited in a fixed order: Let J_1 and J_2 be commutes from a_1 that visit, respectively, a_2 before a_3 and vice versa. Then let J_1' and J_2' be the subwalks from a_1 to a_2 and from a_2 to a_1 in J_1 and J_2, respectively. The concatenation of J_1' and J_2' is thus a commute from a_1 that does not visit a_3, a contradiction. ∎

In the rest of subsection 2.3, we assume that equivalence classes under $\sim$ are sorted so that they must be visited in the order by commutes from their first element. Similarly, if the K elements of an equivalence class are numbered $a_1, \ldots, a_K$, we naturally identify $a_x = a_{x \bmod K}$.

Lemma 2.6. *Given an equivalence class α under $\sim$, for each $j \in G - \alpha$ there is a unique k such that all walks J containing j with $\alpha \cap J \neq \varnothing$ include either a_{k-1} before j or a_k after j.*

Proof. It is enough to consider paths. Suppose J_1 and J_2 are two paths from j which reach a_k and $a_{k'}$, respectively, before reaching any other elements of α, with $k \neq k'$. By strong connectivity, we can select a shortest path from a_k to j to extend J_1 to a cycle from j, which we call J_3. That is, if we select a shortest (in number of distinct steps) path, γ, from a_k to j, then $J_1 \cup \gamma = J_3$ is the required extension of J_1.

Now, J_3 can be cyclically reordered to be a commute from a_k. Thus, J_3 includes $a_{k'}$, and since γ was shortest possible, it includes $a_{k'}$ exactly once. Let $J_4 \subset J_3$ be the subwalk from $a_{k'}$ to j. Then concatenating J_4 and J_2 gives a commute from $a_{k'}$ that does not include a_k, a contradiction. Thus, j has a unique successor a_k in α.

The conclusion that there is a unique predecessor of j in α follows by reversing the direction of all edges and reapplying the above argument. It must be a_{k-1} since a_k is the first member of the equivalence class encountered in any commute from j. ∎

Definition 2.7. *We will here refer to the equivalence classes on $G - \alpha$ induced by Lemma 2.6 as (α-) segments of G.*[5]

Lemma 2.8. *Given distinct equivalence classes α and β under $\sim$, every element of α must lie within a single segment induced by β.*

Proof. Let $\alpha = \{a_1, a_2, \ldots, a_{K_\alpha}\}$ and $\beta = \{b_1, b_2, \ldots, b_{K_\beta}\}$. Suppose, by way of contradiction, that a_1 lies between b_{k_1} and b_{k_1+1} and a_2 lies between b_{k_2} and b_{k_2+1} for $k_1 \neq k_2$. By strong connectedness, there exists a (shortest) path from b_{k_1} to a_1 to b_{k_1+1}. If b_{k_1+1} and b_{k_2} are distinct nodes, there also exists a shortest path from b_{k_1+1} to b_{k_2}. Since $Q_{b_{k_2}, a_2} < 1$, there exists a shortest path from b_{k_2} to b_{k_2+1} not passing though a_2. Finally, if b_{k_2+1} and b_{k_1} are distinct nodes, there exists a shortest path from b_{k_2+1} to b_{k_1}. Concatenating all these paths gives a commute from a_1 to itself not passing though a_2, a contradiction. ∎

The foregoing lemmata show that the nontrivial equivalence classes in a $d^{1/2}$-degenerate digraph induce a structure of equivalence cycles and their segments, with distinct equivalence cycles restricted to lie within the segments of each other. This has potential application in segmentation of directed graphs and will be an important technical tool in the proofs in the next subsection.

2.4. Quotients of $d^{1/2}$-degenerate Markov chains. Next, we develop a way to transform a Markov chain X for which $d^{1/2}$ is not a metric into a quotient Markov chain X', for which $d^{1/2}$ is a metric.

[5] Alternatively, we could define segments more generally with respect to any node set α. Then the segment corresponding to $i \in \alpha$ is the set of nodes reachable from i without passing through any other elements of α. From this perspective, the absolute segments described later are simply the intersection of the segments with respect to all the equivalence classes.

Remark 2.9. In subsection 2.4, we identify singleton classes with their member. Additionally, we append a prime to any symbol when it is meant to refer to X' rather than X.

The quotient graph is given by the following construction, which has appeared in [35] as well as in [32, 34, 5] and possibly other places.

Definition 2.10. *Given a Markov chain X and an equivalence relation on the states of X, the* quotient Markov chain *has one state for each equivalence class, and the transition probabilities are given by*

$$P'_{U,V} = \frac{1}{\phi_U} \sum_{i \in U} \phi_i P_{i,V} = \frac{1}{\phi_U} \sum_{i \in U} \sum_{j \in V} \phi_i P_{i,j} \,,$$

where $\phi_U = \sum_{i \in U} \phi_i$.

The map that sends X to X' is denoted ι. It can be shown [35] that the invariant measure on X' evaluated at state U is $\phi'_U = \phi_U$. Furthermore, P carries information about the equilibration rate in ergodic chains [34, 32, 5], although we do not use this fact in this paper. When applying Definition 2.10 to $\sim$, the definition reduces to

$$P'_{U,V} = \frac{1}{|U|} \sum_{i \in U} \sum_{j \in V} P_{i,j} \,,$$

since ϕ is constant within equivalence classes (see proof of Theorem 1.4).

Lemma 2.11 (quotienting one class at a time). *Let $\sim$ induce the nonsingleton classes $\{\alpha_1, \alpha_2, \ldots, \alpha_L\}$. For a node set S, let $\sim_S$ be the relation with nonsingleton class S, keeping all other nodes in individual (singleton) classes. One can then produce a graph with the same nodes as P' by performing a series of quotienting operations $P \underset{\sim_{\alpha_1}}{\to} P_1 \underset{\sim_{\alpha_2}}{\to} \cdots \underset{\sim_{\alpha_L}}{\to} P_L$. Then $P_L = P'$, after identifying nested classes with the nodes in them,* e.g., $\{\{a\}, \{b\}\} \to \{a, b\}$.

Proof. The proof is by induction on L. If $L = 1$, the result is vacuously true. So assume the result is true for graphs having L nonsingleton equivalence classes, and we proceed to establish the result for $L + 1$ nonsingleton classes. Let G have classes $\{\alpha_1, \ldots, \alpha_{L+1}\}$ under $\sim$. Then we apply $\sim_{\alpha_1}$ to get P_1 and then use the inductive assumption to conclude that $P_{L+1} = P'_1$. So we need to prove that $P'_1 = P'$. We have

$$P'_{1\alpha,\beta} = \begin{cases} P_{\alpha,\beta}, & \alpha \neq \alpha_1, \beta \neq \alpha_1, \\ \sum_{j \in \beta} P_{1\alpha_1,j}, & \alpha = \alpha_1, \beta \neq \alpha_1, \\ \frac{1}{|\alpha|} \sum_{i \in \alpha} P_{1i,\alpha_1}, & \alpha \neq \alpha_1, \beta = \alpha_1, \\ P_{1\alpha_1,\alpha_1}, & \alpha = \alpha_1 = \beta \,, \end{cases}$$

where we have implicitly used the fact that ϕ_{P_1} has the form given in Lemma 2.11. Expanding further gives

$$P'_{1\alpha,\beta} = \begin{cases} P_{\alpha,\beta}, & \alpha \neq \alpha_1, \beta \neq \alpha_1, \\ \sum_{j\in\beta} \frac{1}{|\alpha_1|} \sum_{i\in\alpha_1} P_{i,j}, & \alpha = \alpha_1, \beta \neq \alpha_1, \\ \frac{1}{|\alpha|} \sum_{i\in\alpha} \sum_{j\in\alpha_1} P_{i,j}, & \alpha \neq \alpha_1, \beta = \alpha_1, \\ \frac{1}{|\alpha_1|} \sum_{i\in\alpha_1, j\in\alpha_1} P_{i,j}, & \alpha = \alpha_1 = \beta\,. \end{cases}$$

Rearranging sums gives

$$P'_{1\alpha,\beta} = \frac{1}{|\alpha|} \sum_{i\in\alpha, j\in\beta} P_{i,j} = P_{\alpha,\beta}\,,$$

as expected. ∎

Lemma 2.12. *Collapsing a single equivalence class α respects Q in the following sense. Let i and j be two nonequivalent nodes.*

- *If i and j lie in the same α-segment, then $Q_{i,j} = Q'_{i,j}$.*
- *If i and j lie in different α-segments, then $\frac{1}{2}Q_{i,j} < Q'_{i,j} < Q_{i,j}$.*
- *If $i \in \alpha$, then $Q_{i,j} = |\alpha| Q'_{\alpha,j}$.*
- *If $j \in \alpha$, then $Q_{i,j} = Q'_{i,\alpha}$.*

Proof. Let $\alpha = \{a_1, \ldots, a_K\}$, where $K = |\alpha|$. It is clear that $Q_{i,j}$ is unaffected by taking quotients if i and j lie in the same α-segment or if $j \in \alpha$.

For $i = a_k \in \alpha$, we know that $Q_{a_k,j} = Q_{a_\ell,j}$ for all ℓ, so WLOG assume that j lies in the segment between a_k and a_{k+1}. Now, let us denote by Q_{i_1,i_2,i_3} the probability of a random walker starting at i_1 and reaching i_2 before reaching i_3 (in particular, $Q_{i,j} = Q_{i,j,i}$). Then,

$$\begin{aligned} Q'_{\alpha,j} &= P'_{\alpha,j} + \sum_{i'\neq j,\alpha} P'_{\alpha,i'} Q'_{i',j,\alpha} = \frac{1}{K} P_{i,j} + \frac{1}{K} \sum_{\ell=1}^{K} \sum_{i'\neq j, i'\notin\alpha} P_{a_\ell,i'} Q_{i',j,\alpha} \\ &= \frac{1}{K} P_{i,j} + \frac{1}{K} \sum_{i'\neq j, i'\notin\alpha} P_{a_k,i'} Q_{i',j,\alpha} = \frac{1}{K} P_{i,j} + \frac{1}{K} \sum_{i'\neq j,i} P_{i,i'} Q_{i',j,i} = \frac{1}{K} Q_{i,j}\,. \end{aligned}$$

Finally let i and j be such that any path from i to j must pass through $a_1, \ldots, a_k$ before encountering j. Then the following reasoning applies. Let $a = a_1, b = a_K$, $x = Q_{a,b,j}$ and $y = Q_{b,i,a}$. Then we have

$$\begin{aligned} Q_{i,j} &= Q_{i,a,i} Q_{a,j,i}\,, \\ Q_{a,j,i} &= (1-x) + x Q_{b,j,i}\,, \\ Q_{b,j,i} &= (1-y) Q_{a,j,i}\,. \end{aligned}$$

Solving for $Q_{i,j}$ yields

$$Q_{i,j} = Q_{i,a} \frac{1-x}{1-x+xy}\,.$$

On the quotiented graph, we also have

$$Q'_{i,j} = Q_{i,a} Q'_{\alpha,j,i}, \quad Q'_{\alpha,j,i} = \frac{1}{2}\left[1 - x + x Q'_{\alpha,j,i}\right] + \frac{1}{2}\left[(1-y) Q'_{\alpha,j,i}\right]\,.$$

Hence, $Q'_{i,j} = Q_{i,a} \frac{1-x}{1-x+y}$ and thus $Q'_{i,j} < Q_{i,j}$.

Furthermore, we can bound the ratio

$$\frac{Q_{i,j}}{Q'_{i,j}} = \frac{1}{1 - \frac{(1-x)y}{1-x+y}}. \tag{2.6}$$

Since the function $g(x_1, x_2) = \frac{x_1 x_2}{x_1+x_2}$ is bounded above by $\frac{1}{2}$ on $(0,1)^2$, (2.6) cannot exceed 2, which gives the bound. The bound is tight because all values of x and y can be attained when considering arbitrary weighted graphs. (A graph with only the four nodes a, b, i, j mentioned in the proof and edges $a \to j$, $a \to b$, $j \to b$, $b \to i$, $b \to a$, and $i \to a$ suffices to attain all possible values of x, y.) ∎

Definition 2.13. *An* absolute segment *is a maximal set of nodes which lie in the same segment with respect to all nonsingleton equivalence classes.*

Lemma 2.14. *ι respects Q in the following sense for nodes i and j in distinct equivalence classes α and β:*

- *If i and j lie in the same absolute segment, then $Q_{i,j} = Q'_{i,j}$.*
- *Otherwise, $\frac{1}{2^c|\alpha|} Q_{i,j} \le Q'_{\alpha,\beta} < Q_{i,j}$, where c is the number of equivalence classes with respect to which i and j lie in different segments. (In particular, $c < L$.) Equality holds only when $c = 0$.*

Proof. If i is degenerate, first collapse α, scaling $Q_{i,j}$ by $|\alpha|$. Next, collapse all other equivalence classes one at a time, further scaling $Q_{i,j}$ by the appropriate factor in $(\frac{1}{2}, 1)$ whenever i and j lie in different segments with respect to the collapsing class. ∎

From this lemma we immediately get the following theorem.

Theorem 2.15. *X' is a metric space with metric $(d')^{1/2}$. In particular, for $i \in \alpha$ and $j \in \beta$, with $\alpha \neq \beta$,*

- *if i and j lie in the same absolute segment, then $d_{i,j}^{1/2} = (d')_{i,j}^{1/2}$;*
- *otherwise $d_{i,j} < d'_{\alpha,\beta} \le d_{i,j} + \frac{1}{2}\log|\alpha||\beta| + c\log 2$, where c is the number of equivalence classes with respect to which i and j lie in different segments. Equality holds only when $c = 0$.*

Thus, ι pushes apart the different absolute segments. All other distances are unaffected.

Remark 2.16. ι is analogous to a rigid motion on each absolute segment, in that none of the in-absolute-segment distances are distorted.

3. Computational methods.

To compute the normalized hitting probabilities matrix and metric structure on a Markov chain (or network) consisting of n nodes/states with probability transition matrix P, we require only the computation of the invariant measure and the Q matrix. The invariant measure can be computed using iterative eigenvector methods, which need $O(m)$ operations per iteration for m edges.

We briefly recall the work in [47, Theorem 5] that shows the Q matrix can be computed in $O(n^3)$ time. The key idea from [47, Lemma 5] is that one can compute

$$Q_{i,j}(P) = \frac{e_i^T (I - P_j)^{-1} P_j e_j}{e_i^T (I - P_j)^{-1} e_i} = \frac{M(j)_{i,j}^{-1}}{M(j)_{i,i}^{-1}}, \tag{3.1}$$

where $e_j \in \mathbb{R}^n$ is the vector with a 1 in the jth entry and zeros elsewhere, $P_j = (I - e_j e_j^T)P \in \mathbb{R}^{n\times n}$, and the invertible matrix $M(j) = I - P + e_j e_j^T P \in \mathbb{R}^{n\times n}$. See Theorem 5 of [47] for full details, but this identity follows from realizing that as defined $M(j)$ is invertible with inverse

$$M(j)^{-1} = \begin{pmatrix} (I-P_j)^{-1} & (I-P_j)^{-1}P_j e_j \\ 0 & 1 \end{pmatrix}$$

given in block form on the $e_j^{\perp}$, e_j basis.

If we then compute $M(1)^{-1}$ on the way to obtaining the first column $Q_{i,1} = M(1)^{-1}_{i1}/M(1)^{-1}_{ii}$, then $M(j)$ is a rank-2 perturbation of $M(1)$ and we can apply the Sherman–Morrison–Woodbury identity to compute $M(j)^{-1}$. Since we only access $2n-2$ elements of $M(j)^{-1}$, the full $O(n^2)$-time Sherman–Morrison–Woodbury update is not needed, and we can get the jth column $Q_{i,j}$ in $O(n)$ computations from $M(1)^{-1}$. A MATLAB implementation of this procedure, along with code for all of the numerical experiments described in the paper, is available at https://github.com/zboyd2/hitting_probabilities_metric.

The matrix Q encodes the hitting probabilities of a random walk on the nodes of a graph, and the order of the method we present here is very well documented in [47]. However, there are several results that consider the computational complexity of the related problem of commute times; see, for instance, the works [30, 3]. The computational cost of computing hitting probabilities through inversion of the Laplacian has been explored further in [22, 14, 13], resulting in some cases in which the method may be improved to better than $O(n^3)$. As we are mostly interested in the construction of the metric here, we will not further explore the question of optimal order of the computation.

4. Examples.

We consider examples of Markov chains and directed graphs to illustrate the proposed metric. We start with simple graphs for which the calculations can be performed exactly. We then numerically explore a variety of synthetic graphs and a real-world example defined from New York City taxi cab data.

4.1. Exact formulations.

Here we consider some simple graphs on which the invariant measure and hitting probabilities can be computed exactly to help us understand $A^{(\mathrm{hp},\beta)}$ and d^{β}.

1. *Directed cycles:* Consider a directed cycle on n nodes. Then $\phi_i = 1/n$ for all i, and $Q_{i,j} = 1$ for all $i \neq j$. Therefore, $A^{(\mathrm{hp},\beta)}$ is a weighted clique, and d^{β} has all points equidistant. For $\beta = 1/2$, the weights equal to 1, and all nodes are identified with each other in the metric topology.
2. *Complete graphs:* Consider a complete graph on $n > 2$ nodes. Then $\phi_i = 1/n$ for all i, and $Q_{i,j} = \mathrm{const} < 1$ for all $i \neq j$. Therefore, $A^{(\mathrm{hp},\beta)}$ is a weighted clique. Unlike the directed cycle case, the weights in the clique are < 1 for all $\beta \geq 1/2$.
3. *Glued cycles:* Consider graphs of the type depicted in Figure 4.1, namely, graphs composed of n_b "backbone nodes" forming a directed chain, which then branches into C chains of length n_c, each of which finally connects back to the beginning of the backbone chain. Intuitively, a random walker on this graph transitions between $C+1$ groups of nodes, namely, each of the C branches and the backbone. As illustrated

Table 4.1

Values of Q, $A^{(\mathrm{hp},\beta)}$, and d evaluated at distinct nodes i and j for the glued cycles example from subsection 4.1. We include extra columns for the case $\beta = 1/2$, which is particularly interpretable. Observe that neither $A^{(\mathrm{hp},\beta)}$ nor d^{β} depends on n_b or n_t (except up to scaling), which is a manifestation of their blindness to walk length. Also, the nodes that are closest together are those which lie on common chains. Note that we scaled $A^{(\mathrm{hp},\beta)}$ for visual clarity. The invariant measure is easily verified to be $(n_b + n_c)^{-1}$ on the backbone and $(C(n_b + n_c))^{-1}$ elsewhere.

i	j	Q	$\frac{A^{(\mathrm{hp},\beta)}}{(n_b+n_c)^{1-2\beta}}$	$A^{(\mathrm{hp},1/2)}$	$d^{\frac{1}{2}}$
Branch	Same branch	1	$1/C^{2\beta-1}$	1	0
Branch	Different branch	$1/2$	$1/2C^{2\beta-1}$	$1/2$	$\log 2$
Backbone	Branch	$1/C$	$1/C^{\beta}$	$C^{-1/2}$	$1/2 \log C$
Branch	Backbone	1	$1/C^{\beta}$	$C^{-1/2}$	$1/2 \log C$
Backbone	Backbone	1	1	1	0

in Table 4.1, our metric captures this intuition by placing each node very close to the others on its chain. This is in contrast to commute-time-based metrics, where the length of the chain must be taken into account. (See Figure 4.2.) In subsection 4.2, we consider some numerical results based on this example.

4.2. Synthetic numerical examples. We consider four examples. The first two demonstrate that the spectrum of $A^{(\mathrm{hp},\beta)}$ (for $\beta = 1/2$ or $\beta = 1$) identifies cyclic and clique-like sets in a useful manner. We compare to two alternative symmetrizations and another metric. The second example additionally shows the scalability of our approach. In the third example, we explore when it is advantageous to use d for visualization and clustering purposes, using a directed planted partition model for ground truth comparisons. In dense, difficult-to-detect regimes, our method is more accurate than clustering using the input adjacency matrix directly. Finally, in the fourth example, we compare $d^{1/2}$, d^1, and spatial distance for geometric graphs, finding that our distance captures comparable information to the spatial distance, with the similarity being especially tight when $\beta = 1/2$.

4.2.1. Glued-cycles networks. For the two-glued-cycles networks illustrated in Figure 4.1, we construct a probability transition matrix P by taking a uniform edge weight for all connected vertices and performing a row normalization. We then compute the Fiedler eigenvector corresponding to the second smallest eigenvalue of the graph Laplacian (sometimes called the Fiedler vector) for different symmetrized adjacency matrices. For the adjacency matrices A constructed below, we calculate the graph Laplacian $L = D - A$ with D the diagonal matrix of node degrees (row or column sums of A). The examples here are two directly glued cycles, as well as two glued cycles with a bidirectional edge between the cycles. In the first case, the results are all very similar regardless of the symmetrization, but for the second case the results differ significantly. In each case, we group the nodes based on whether the corresponding vector element is positive, negative, or zero.

For the two glued cycles without the bidirectional edge, the naive symmetrizations of the directed adjacency matrix, either $A = (P + P^T)/2$ or $A = \max\left(P, P^T\right)$, have a Fiedler vector that is 0 on the spine and splits each cycle into signed components; see the top right plot

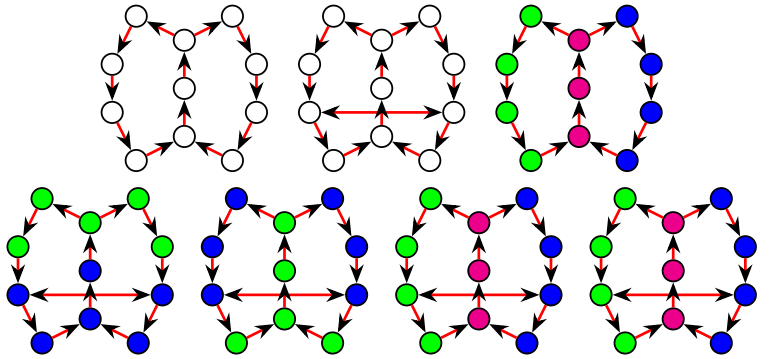


Figure 4.1. *(Top left) Two-glued-cycles example from subsection 4.1 with $n_b = 3$, $n_c = 4$, and $C = 2$. The "backbone" nodes run along the center, and the two partial cycles split off from and then return to it. (Top middle) Similar two-glued-cycles network with a bidirectional edge. (Top right) Sign of the Fiedler vector of the Laplacian for several different symmetrizations. (Bottom) The sign of the Fiedler vector of the Laplacian for several different symmetrizations. The sign of the Fiedler vectors is encoded as (−, green), (0, magenta), and (+, blue).*

in Figure 4.1. However, in the bottom left component Figure 4.1, for the two-glued-cycles network with the bidirectional edge, the naive symmetrization splits the network horizontally, which is reasonable, since the resulting graph cut is small, although this (by construction) does not reflect the coherent, directed structure of the original graph.

One way to account for directed structure in a way that minimizes equilibrium flux across the cut was suggested by Fan Chung [12] (cited in subsection 1.3), defining the Laplacian by $L = I - \frac{1}{2}\left[\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}\right]$, where $\Phi = \mathrm{diag}(\phi) \in \mathbb{R}^{n \times n}$. Chung uses L to establish a Cheeger-type inequality for digraphs, which is used to study the rate of convergence for Markov chains. Using Chung's Laplacian again gives a comparable outcome for the two-glued-cycles example (Figure 4.1), but in the example with the bidirectional edge, this symmetrization places most of the nonbackbone nodes in one class and all backbone nodes in the other (second plot in Figure 4.1).

The normalized hitting probabilities matrices $A^{(\mathrm{hp},1)}$ and $A^{(\mathrm{hp},1/2)}$ each distinguish between the two branches, with the backbone set equal to zero in both the cases of the glued cycles and the glued cycles with a bidirectional edge as seen in Figure 4.1. Thus, all three approaches uncover different structure in the two-glued-cycles graph with the bidirectional edge, with the naive symmetrization yielding small undirected cuts, Chung's approach yielding (perhaps) two different dynamical states, and $A^{(\mathrm{hp},\beta)}$ showing all three chains in a natural way for both $\beta = 1/2, 1$.

Finally, we compare the total effective resistance metric of [54] to our metric on the example of the two-glued-cycles network (with no bidirectional edge). As one might expect given the relationship between effective resistance and commute times in the undirected case, the total effective resistance of [54] is sensitive to cycle length. Figure 4.2 demonstrates that the commute time approach views the distances on each cycle quite differently and that the relative distances from the total effective resistance metric are more difficult to interpret in the second loop.

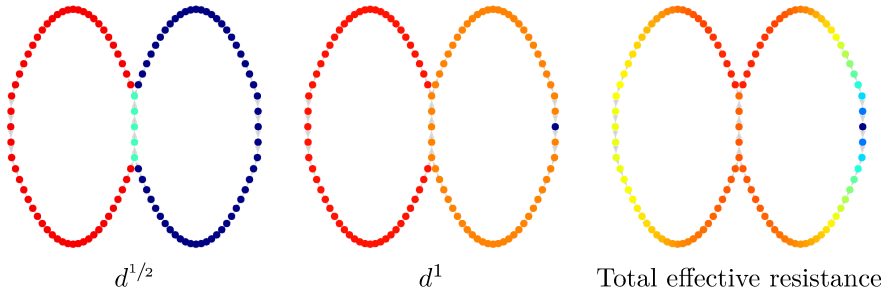


Figure 4.2. *Two glued cycles with $n_c = 55$ and $n_b = 5$, with nodes colored by distance from a node on the far right. Blue denotes small distances. The metric $d^{1/2}$ has three levels of distance corresponding to nodes on the same branch, backbone, and opposite branch, respectively. The metric d^1 is similar, except nodes on the same branch are not distinguished from backbone nodes. Finally, the total effective resistance metric from [54] gives a smoother notion of distance on the right branch and backbone, but on the left branch, proceeding counterclockwise, one finds the distance decreasing and then increasing again, which is somewhat difficult to interpret. This example shows how different resistance/commute time are from hitting-probability distance.*

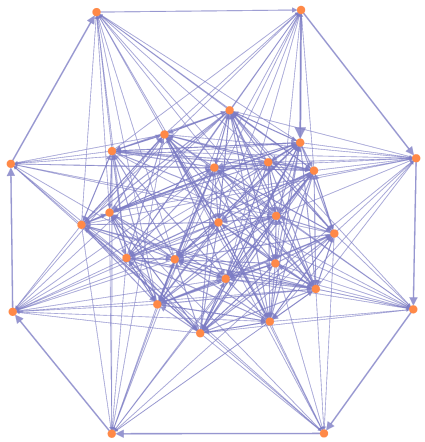


Figure 4.3. *Erdős–Rényi plus cycle example.*

4.2.2. Cycle adjoined to directed Erdős–Rényi (ER) graph. Consider the following construction, illustrated in Figure 4.3. Let $n = n_{\text{er}} + n_{\text{cycle}}$, and let the first n_{er} nodes form an unweighted, directed ER graph with connection probability p. The remaining n_{cycle} nodes form an unweighted, directed cycle. The ER graph and cycle are connected by adding $2\,\text{round}(np) - 1$ edges of weight w to each cycle node from randomly selected nodes in the ER graph.[6] Finally, a single, bidirectional edge of weight 1 is added from one cycle node to one ER node. Normalizing the rows to form a probability transition matrix, a random walker on this graph would transition between the ER and cycle subgraphs, where the cycle subgraph is difficult to escape quickly because of the single exit. For the particular choice of $n_{\text{er}} = 20$, $n_{\text{cycle}} = 8$, $p = .5$, and $w = 3$, we find that the Fiedler vector of (the Laplacian associated with) $A^{(\text{hp},1/2)}$ is

[6] These edges are drawn with replacement with multiedges merged to a single edge of weight w. Results were similar when we added the weights instead.

positive on the cycle nodes and negative elsewhere. In contrast, the Fiedler vector of the naive symmetrization $A = (P + P^T)/2$ or Chung's L [12] does not separate the cycle and ER nodes. Scaling up to $n = n_{\text{er}} + n_{\text{cycle}} = 7,200 + 2,800 = 10,000$ nodes, keeping the other parameters the same (≈ 38.7 million edges), gives similar eigenvector results. The computation takes 31 seconds on a Lenovo ThinkStation P410 desktop with Xeon E5–1620V4 3.5 GHz CPU and 16 GB RAM using MATLAB R2019a Update 4 (9.6.0.1150989) 64-bit (glnxa64): 18 seconds to compute Q, 6 seconds to compute ϕ, 2 seconds to form $A^{(\text{hp},1/2)}$, and 5 seconds to compute the Fiedler vector.

4.2.3. Cluster detection and visualization for digraphs. We next use d for clustering and dimension reduction. We consider directed graphs generated by a planted partition model with nodes grouped into three ground truth communities and form a uniformly weighted adjacency matrix by connecting an edge from i to j with probability p_{in} if i and j are in the same community and p_{out} ($< p_{\text{in}}$) otherwise. A probability transition matrix can then be formed using row normalization. We then attempt to recover the ground truth node assignments. The difficulty of this problem is generally understood in terms of $\Delta = p_{\text{in}} - p_{\text{out}}$ and $\rho = \frac{p_{\text{in}}+2p_{\text{out}}}{3}$. Small values of Δ correspond to more difficult clustering problems that may be solved less accurately (relative to the ground truth). In this example we attempt to cluster the nodes into $k = 3$ clusters using several approaches: (1) principal component analysis[7] (PCA) [40] on the adjacency matrix, A, followed by k-means clustering on the first $k-1$ PCA vectors; (2) PCA on $d^{1/2}$ followed by k-means; and (3) k-medoids on $d^{1/2}$. (The k-medoids algorithm is similar in spirit to the k-means unsupervised clustering algorithm but applies in arbitrary metric spaces; see, for instance, [26, 39].) Results are shown in Figures 4.4 and 4.5. We find that method (1) works best on sparse or well-separated clusters, method (2) works best with dense, difficult-to-detect clusters, and method (3) has no clear advantage. More specifically, using $d^{1/2}$ in method (2) enhances our ability to get a better-than-chance clustering in dense networks.[8] (We note that spectral methods in undirected graphs give asymptotically optimal almost-exact recovery but are not optimal for harder cases where only better-than-chance recoverability is possible [1]. This is consistent with Figures 4.4 and 4.5.) Finally, we can also use PCA on $d^{1/2}$ to visualize the directed network. The first and second principal components, generated using the built-in routine in MATLAB, are plotted in Figure 4.6, clearly showing the separation into three clusters, which are in accordance with the three ground-truth communities.

4.2.4. Distances on geometric graphs. Given known convergence properties of various graph models to continuum problems (e.g., [49, 48, 45, 46, 38]), we are motivated by the question of how our distance metric compares to a standard notion of distance when the

[7] Specifically, we used the PCA routine from MATLAB R2019a Update 4 (9.6.0.1150989) 64-bit (glnxa64). As expected, this gives different results in general when applied to a matrix versus its transpose. In this case, the matrix is stochastically equivalent with its transpose, and in the New York taxi example below, the PCA-based plots are similar regardless of whether the transpose is used.

[8] We also tried using the shortest commute and generalized effective resistance metrics [54, 55] as substitute for the hitting time metric in this example and found similar improvements over using the raw adjacency matrix. In particular, the shortest commute was the most effective metric for this task (although this metric is not robust, so the real-life performance may be different).

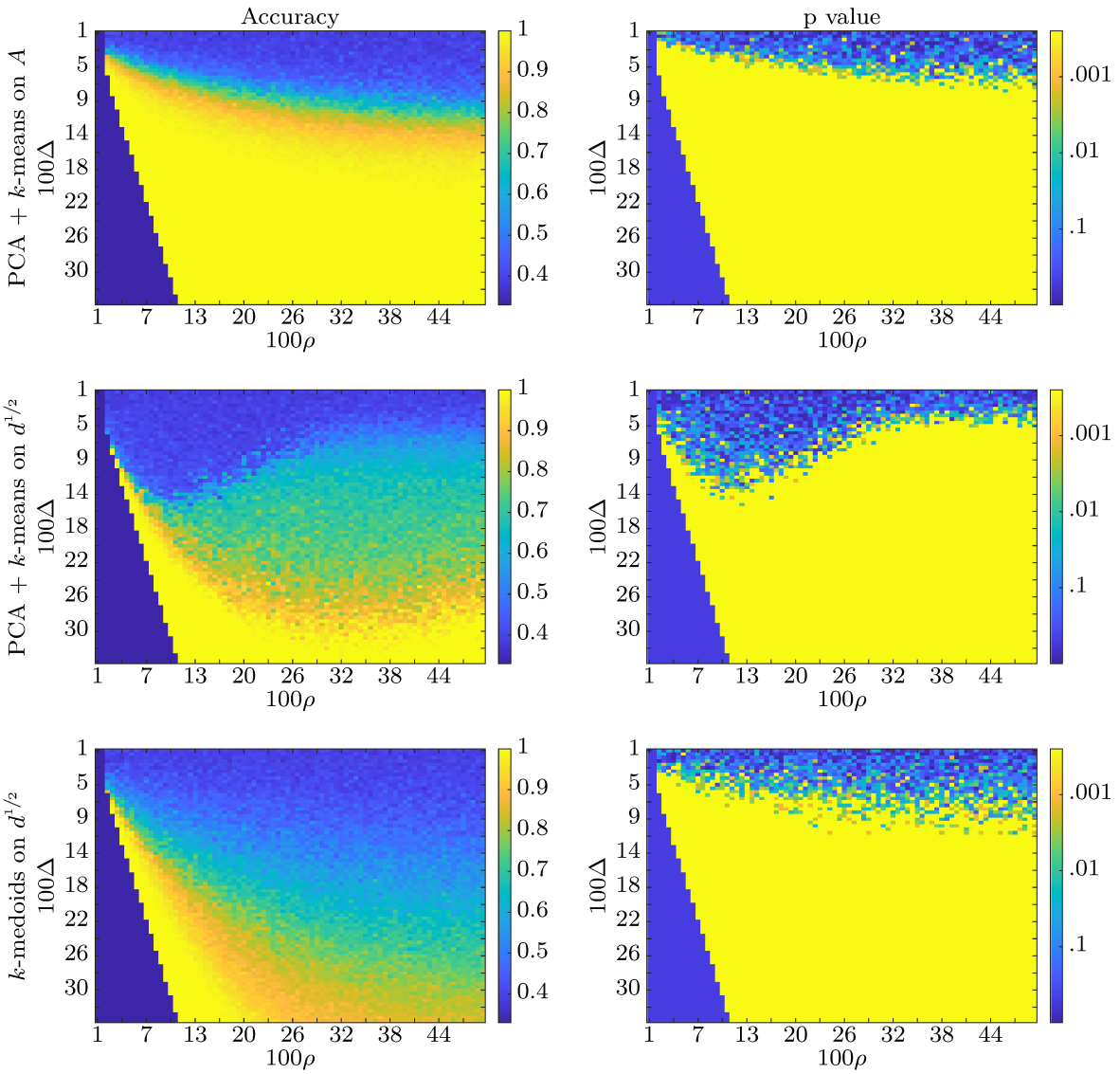


Figure 4.4. *Results of (top row) PCA on A followed by k-means, (middle) PCA on $d^{1/2}$ followed by k-means, and (bottom) k-medoids on $d^{1/2}$ on 300-node graphs generated using the directed planted partition model with three clusters, as described in subsection 4.2.3. We varied the mean edge density, ρ, and cluster quality, $\Delta = p_{\text{in}} - p_{\text{out}}$. Since results depend on the random initialization, we report best of 5 runs for each entry. If any generated graph was not strongly connected, we did not try to cluster it. The left column is the accuracy (purity) of the recovered partition, and the right value is the empirical p value of the accuracy relative to 4,000 random partitions obtained by drawing each community label uniformly at random. Notably, method (2) has the best performance for dense, weakly clustered graphs. (Note that the triangular blue region on the lower left of each plot represents a (ρ, Δ) parameter combination that cannot exist.)*

network arises from a natural geometric setting. For instance, as mentioned in the introduction, [45] proves that the notion of diffusion distance converges to that of geodesic distance as a point cloud samples a closed manifold at higher and higher densities.

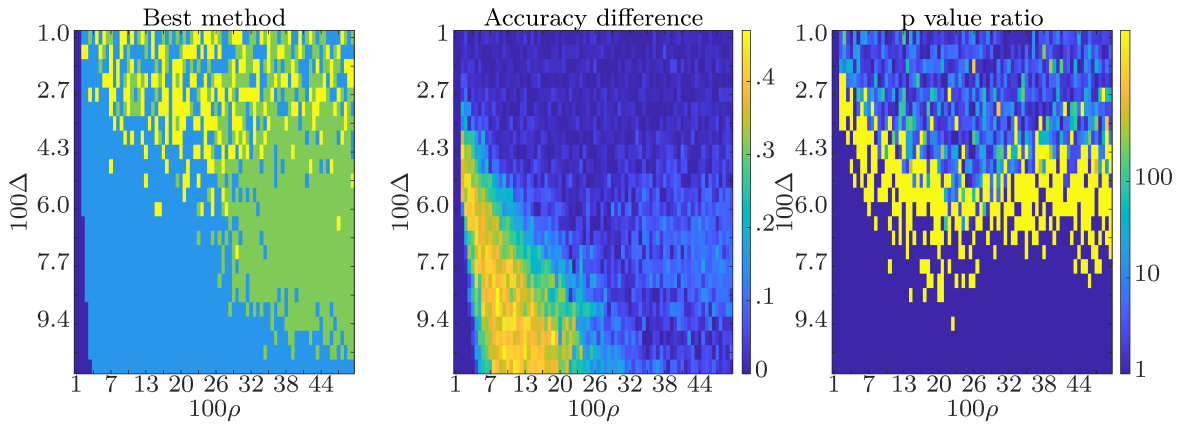


Figure 4.5. *(Left) Regions where the methods from Figure 4.4 perform best. Here, light blue is method* (1)*, green is method* (2)*, and yellow is method* (3)*. (Middle) Difference in accuracy between the best and second-best methods. (Right) Ratio of p value of the best and second best methods from Figure 4.4. Combining these plots, we see that there is a significant parameter regime consisting of dense, difficult to detect structure, where using* $d^{1/2}$ *instead of A enhances the spectral detection of structure by* 5–20% *for graphs where method* (1) *is recovering essentially no structure in A. Note that the y axis is different from Figure 4.4.*

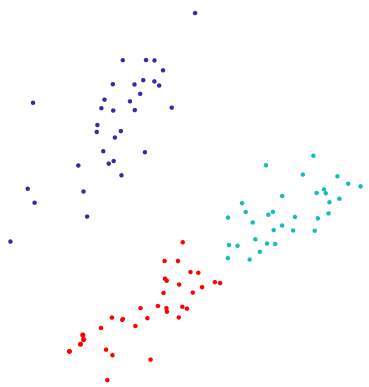
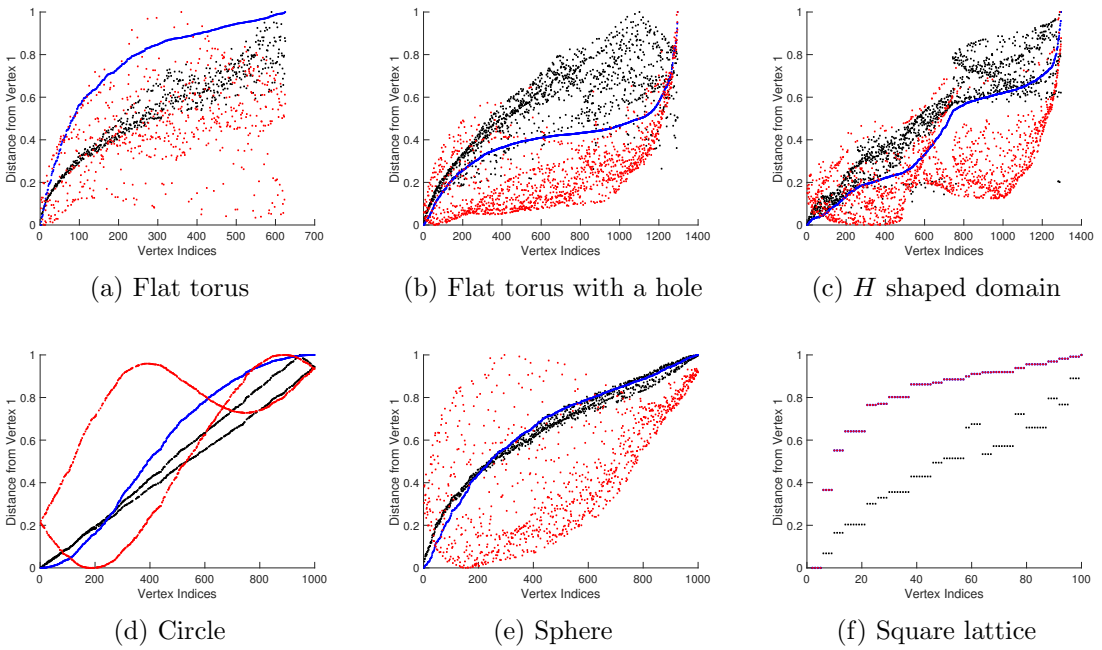


Figure 4.6. *PCA embedding of* $d^{1/2}$*, colored by ground truth community.*

In Figure 4.7, we consider distances computed using our metric structure in a family of geometric graphs constructed using Euclidean distances to determine edge weights. The following geometric graphs are considered:

(a) A random point cloud on a flat torus $[0, 2\pi]^2$ with 36^2 points.
(b) A random point cloud on a flat torus with a hole $[0, 2\pi]^2 \setminus B((\pi, \pi), \pi/2)$ with 36^2 points (distances relative to a point in the bottom left of the torus).
(c) An H shaped domain $([0, 2\pi]^2 \cap \{|x_1 - \pi| \geq \pi/2\}) \cup ([0, 2\pi]^2 \cap \{|x_2 - \pi| \leq \pi/4\})$ with 36^2 points (distances relative to a point in the bottom right of the H).
(d) A random point cloud on the circle of length 2π with 1000 points.
(e) A random point cloud on a sphere of radius 1 in $\mathbb{R}^3$ with 1000 points.
(f) A square 10×10 lattice on the flat torus $[0, 2\pi]^2$.



(a) Flat torus (b) Flat torus with a hole (c) H shaped domain

(d) Circle (e) Sphere (f) Square lattice

Figure 4.7. *Normalized distance plots comparing the scaled distances from one node in a geometric graph computed using Euclidean distances (black), $d^{1/2}$ (blue), and d^1 (red). The geometric graphs from top left to bottom right are* (a) *a random point cloud on a flat torus,* (b) *a random point cloud on a flat torus with a hole,* (c) *a random point cloud on an H shaped domain,* (d) *a random point cloud on the circle,* (e) *a random point cloud on a sphere, and* (f) *a square lattice on the flat torus. Note, in all subplots, we have ordered the vertices from closest to farthest from a reference node given by the first vertex generated relative to the $d^{1/2}$ metric.*

For the regular lattice example, the edge weights are only carried on nearest neighbor vertices. In all other cases, we consider the edge weights to be of the form $e^{-\gamma d_{\text{Euc}}(x_i,x_j)^2}$, where d_{Euc} is just the Euclidean distance metric (determined with periodicity if the domain is periodic, i.e., we take shortest-path distance in the flat torus). We have chosen the scale factor $\gamma = 1$ uniformly throughout.

Once the geometric graph is constructed, we computed the pairwise Euclidean distances, as well as the pairwise distances $d^{1/2}$ and d^1 for comparison. To assist with interpretation and comparison, we have ordered the vertices in Figure 4.7 from closest to farthest relative to the $d^{1/2}$ metric and plotted for each distance function the rescaled distances $(d - d_{\min})/(d_{\max} - d_{\min})$ to normalize all of them to the same scale.

Throughout, we note that $d^{1/2}$ is a reasonable fit to the measured Euclidean distances, while d^1 seems to do well only when the geometry is such that the invariant measure normalization (that is, the choice of β) does not matter as much. Note that the distance $d^{1/2}$ and d^1 are identical on the square lattice, up to scaling. In this case, we are really studying the structure of the Q hitting probability matrix. Our results give some preliminary indication that in the consistency limit the $d^{1/2}$ metric may converge to the Euclidean distance while the d^1 metric converges to something else entirely. However, we leave this pursuit for future analytical studies.

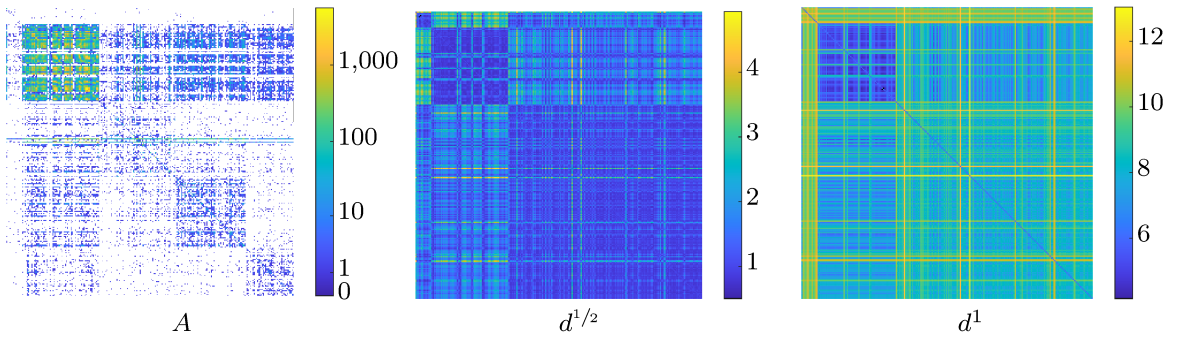


Figure 4.8. *An example based on New York City taxi transit data, where nodes are* 250 *neighborhoods and* $A_{i,j}$ *is the number of trips from* i *to* j*. (Left) Heatmap of* A *with the nodes sorted by borough in this order: EWR airport (one node), Staten Island, Manhattan, Queens, Brooklyn, and the Bronx. The taxi traffic is dominated by the Manhattan block in the upper left, with Queens, Brooklyn, and the Bronx forming three blocks further down and to the right. (Middle) A similarly arranged heatmap of* $d^{1/2}$*. The Manhattan neighborhoods are close together, and the smaller upper left block corresponding to Staten Island is distinguished as a coherent submodule, despite having only* 8 *interior edges. Queens, Brooklyn, and the Bronx form a large block in the lower right. See subsection* 4.3 *for an explanation of these differences. (Right) A heatmap of* d^1*. We observe that in this normalization Staten Island is quite far from everything, including itself. Manhattan is relatively close to almost everything, especially itself. This is exactly what we should expect because* $d^1_{i,j}$ *is small only when both* ϕ_i *and the* $i \to j$ *hitting probability are large. (In the right two heatmaps, the diagonal is set to a nonzero value to improve contrast.)*

4.3. Real-world example: The New York City taxi network. Consider the movement over time of a New York City taxi, which we interpret as a Markov chain where the states are neighborhoods and $P_{i,j}$ is the probability that a trip begun in neighborhood i ends in neighborhood j. Using publicly available data from the New York City Taxi and Limousine Commission,[9] we computed an adjacency matrix where $A_{i,j}$ is the number of Yellow Taxi trips in January 2019 that started at i and ended at j, where i and j are chosen from 262 neighborhoods[10] spread across the city's five boroughs (Manhattan, Staten Island, Queens, Brooklyn, and the Bronx). We also included trips to and from Newark Liberty International Airport (EWR) in New Jersey. We restricted our analysis to the 250-neighborhood strongly connected component.

The data is dominated by degree, as shown in Figure 4.8, with the busier Manhattan neighborhoods having tens of thousands of trips, and the Staten Island neighborhoods having median out-degree of 4. Traffic is also organized by borough, although the distinctions between the spatially adjacent Brooklyn, Queens, and Bronx boroughs are perhaps less apparent, and they might be properly considered as peripheral to the Manhattan core. Staten Island is notable for its remoteness, which is reflected in the sparsity of A in that block.

In Figure 4.8, we compare A with $d^{1/2}$ and d^1. While $d^{1/2}$ highlights Manhattan in a manner similar to A, it does not distinguish much between Queens, Brooklyn, and the Bronx, showing them instead as a single interconnected group. In contrast, Staten Island is very

[9] Accessed at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page in April 2020.

[10] Two additional neighborhoods are marked "unknown" and appear to designate out-of-city or out-of-state endpoints. We excluded these from our analysis.

clearly highlighted as its own, close group, which is reasonable given the geographic proximity of these neighborhoods and the fact that a disproportionately large number of trips involving Staten Island both started and ended there. Although the purpose of this example is not to provide an optimal clustering of the data, we note that Staten Island does represent a difficult cluster to detect, and arguably is not even a cluster, since there are only eight interior edges (counting multiplicity but excluding self-edges) and 309 incoming or outgoing edges, all of which are hidden in over 1 million edges (again, counting multiplicity).

Note that the fact that Staten Island is highlighted by $d^{1/2}$ is not simply because of degree scaling, as a heat map of P does not highlight Staten Island as a block. The true explanation seems to involve two factors: (1) Staten Island has eight nondiagonal in-edges 13 neighborhoods,[11] and the median out-degree is 4. Thus, a taxi that does enter Staten Island has a relatively large likelihood of visiting another Staten Island location next, relative to taxis starting at other neighborhoods. (2) The average frequency of visiting Staten Island at all is so low that the pattern of visiting is almost memoryless, with taxis leaving Staten Island having plenty of time to mix in other areas before visiting Staten Island again, so that the probability of leaving Staten Island and then reaching another Staten Island location before returning to the first one is about $\frac{1}{2}$, despite the low degree of Staten Island neighborhoods. In contrast, Staten Island is far from other locations, especially Manhattan, since by (1.4), mutually high hitting probabilities are required for closeness, but the probability of starting in a Manhattan neighborhood and reaching Staten Island before returning is very low.

The distance d^1 places the Manhattan nodes close to most other nodes, especially each other, while the Staten Island nodes are far from everything, especially each other. Since $d^1_{i,j} = -\log(\phi_i) - \log(Q_{i,j})$, this distance is small only when (1) ϕ_i is large and (2) $Q_{i,j}$ is far from zero. Thus, the Staten Island nodes, which have small values of ϕ, cannot be close to anything, and the Manhattan nodes, which have the largest values of ϕ, can be close to other nodes, depending on $Q_{i,j}$. Empirically, $Q_{i,j}$ is usually not very small, with 77% of the entries in Q being at least 0.1, which explains Manhattan's overall closeness to other nodes. The fact that the Manhattan nodes are closer to each other than to other nodes is accounted for by the fact that $Q_{i,j}$ for i in Manhattan is generally larger if j is also in Manhattan, which might be expected. (The medians differ by a factor of 5.4.) A similar observation explains why the Staten Island nodes are considered farther from each other than they are from nodes in the other boroughs.

Finally, we used $d^{1/2}$ to perform PCA, with the first two principal components (PCs) visualized in Figure 4.9. These two PCs explained 64% and 34% of the variation, respectively, with the first PC being closely related to out-degree (Pearson correlation with $\log k_{\text{out}}$ is .96) and the second PC being well correlated (Pearson correlation .9978) with the column means of $d^{1/2}$. So over 98% of the variance is explained by these two PCs. Interestingly, both the highest- and lowest-degree nodes were on average far from other nodes. Recalling that $d^{1/2}_{i,j}$ is the negative log of the geometric mean of $Q_{i,j}$ and $Q_{j,i}$ (see (1.4)), closeness requires that both of these factors be high. If i is a high-degree core node, then $Q_{i,j}$ is small for most j. In contrast, a mid-degree peripheral node in Queens, the Bronx, or Brooklyn enjoys reasonable

[11] Staten Island has 20 neighborhoods, but 7 have 0 out-degree and are thus excluded from the strongly connected component.

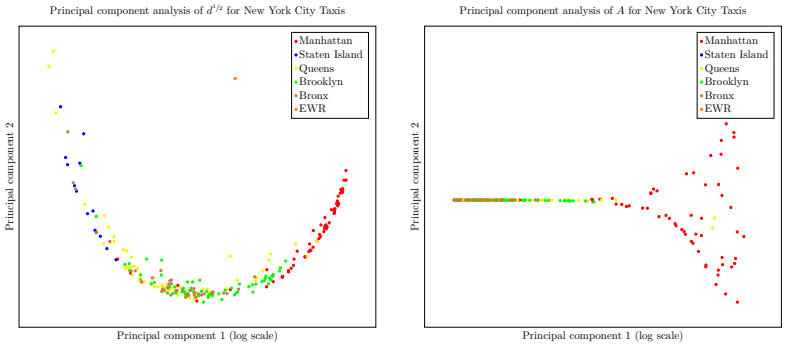


Figure 4.9. *(Left) These two PCs explain* 98.3% *of the variance. The first PC has a* .96 *correlation coefficient with* $\log k_{\text{out}}$*, and the second PC as a correlation coefficient of* .9978 *with the column means of* $d^{1/2}$*, which we interpret as the average distance to other nodes. Notably, the highest-degree nodes also have high average distance to other nodes. This is also true of the lowest-degree nodes, while the mid-degree nodes in Queens, Brooklyn, and the Bronx are closer to other nodes on average. We interpret this by noting that, while high-degree nodes are common endpoints for trips (so* $Q_{i,j}$ *might be high when* j *is a high-degree node), they have a lot of self-loops, and the taxis that leave them tend to return relatively quickly (so* $Q_{j,i}$ *is low for most* i*). Using* (1.4)*, we see that* $d_{i,j}^{1/2}$ *will then not be very small for high-degree* i*. The mid-degree nodes, in contrast, send a lot of taxis into the Manhattan core, which are likely to mix through the city for a long time before returning (so* $Q_{i,j}$ *is not very small for almost all destinations* j*). (Right) PCA on* A *gives a similar first PC. The second PC is nearly constant except on Manhattan, where it is correlated with the East-West coordinate (Pearson* .42*,* $p = .0004$*). The second PC explains about half as much variance for* A *as for* $d^{1/2}$*.*

values of $Q_{i,j}$ for other peripheral nodes j, since once a taxi enters the Manhattan core, it is likely to visit a significant portion of the other nodes before returning to i. Finally, if a node's degree is too small, the probability of a taxi reaching it at all is too small for the hitting probabilities to be high. For comparison, performing PCA directly on A gives a similar first PC, with a different second PC that explains about half as much variance as the second PC of $d^{1/2}$. The second PC is nearly constant, except on Manhattan, where it correlates with the East-West coordinate.

5. Conclusion. Given a probability transition matrix for an ergodic, finite-state, time-homogeneous Markov chain, we have constructed a family of (possibly pseudo-)metrics on the state space, which we refer to as hitting probability distances. Alternatively, this construction gives a metric on the nodes of a strongly connected, directed graph. In the cases where we do not obtain a proper metric, the degeneracies give global structural information, and we can quotient them away. Our metrics can be computed in $O(n^3)$ time and $O(n^2)$ space, in one example scaling up $10,000$ nodes and $\approx 38M$ edges on a desktop computer. Our metric captures different information compared to other directed graph metrics and captures multiscale structure in the taxi example. We have considered the utility of this metric for structure detection, dimension reduction, and visualization, finding in each case advantages of our method compared to existing techniques.

Some other possible applications include efficient nearest-neighbor search, new notions of graph curvature [51], Cheeger inequalities, and provable optimality of weak recovery for dense, directed communities. Additionally, in our experiments, we observed that several eigenvalues

of the symmetrized adjacency matrix contained useful information about structure such as cycles, and it would be good to understand better which structures get encoded in leading eigenspaces. Empirically, it is important to know how commonly $d^{1/2}$ is degenerate and what useful structure is revealed in practice. A natural theoretical question is consistency of the distances in the large graph limit as we approach a natural geometric object embedded in a standard Euclidean space [38, 45, 46, 48, 49, 56].

In terms of possible improvements to our method, an effective means of thresholding the symmetrized hitting probability matrix could improve scalability. A natural question to pursue in a variety of settings would be the sparsification of $A^{(\mathrm{hp},\beta)}$ and its implications for spectral analysis and clustering applications. In particular, the potentially sparse P will map into a full (but symmetric) matrix $A^{(\mathrm{hp},\beta)}$. In large systems the $O(n^2)$ storage requirement may become a burden. Hence, it is natural to ask, If we sparsify the $A^{(\mathrm{hp},\beta)}$ matrix to have a comparable number of edges to that of the original P, how much information can be stably preserved in the spectrum? This will be a topic of future work on the hitting probability matrices we have constructed.

REFERENCES

[1] E. Abbe, *Community Detection and Stochastic Block Models*, http://www.princeton.edu/~eabbe/publications/abbe_FNT_final4_plain.pdf (27 March 2020).

[2] I. Abraham, A. Fiat, A. Goldberg, and R. Werneck, *Highway dimension, shortest paths, and provably efficient algorithms*, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2010.

[3] D. Boley, G. Ranjan, and Z.-L. Zhang, *Commute times for a directed graph using an asymmetric Laplacian*, Linear Algebra Appl., 435 (2011), pp. 224–242.

[4] L. Boytsov and B. Naidan, *Learning to prune in metric and non-metric spaces*, in Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 1574–1582.

[5] S. Caracciolo, A. Pelissetto, and A. Sokal, *Two Remarks on Simulated Tempering*, manuscript, 1992.

[6] M. Charikar, K. Makarychev, and Y. Makarychev, *Directed metrics and directed graph partitioning problems*, in Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2006, pp. 51–60.

[7] P. Chebotarev and E. Deza, *Hitting time quasi-metric and its forest representation*, Optim. Lett., 14 (2020), 291307.

[8] M. Chen, J. Liu, and X. Tang, *Clustering via random walk hitting time on directed graphs*, in Proceedings of the 23rd National Conference on Artificial Intelligence, AAAI, 2008, pp. 616–621.

[9] X. Cheng, M. Rachh, and S. Steinerberger, *On the diffusion geometry of graph Laplacians and applications*, Appl. Comput. Harmon. Anal., 46 (2019), pp. 674–688.

[10] S. Chien, C. Dwork, R. Kumar, D. R. Simon, and D. Sivakumar, *Link evolution: Analysis and algorithms*, Internet Math., 1 (2004), pp. 277–304.

[11] M. C. H. Choi, *On resistance distance of Markov chain and its sum rules*, Linear Algebra Appl., 571 (2019), pp. 14–25.

[12] F. Chung, *Laplacians and the Cheeger inequality for directed graphs*, Ann. Comb., 9 (2005), pp. 1–19.

[13] M. B. Cohen, J. Kelner, R. Kyng, J. Peebles, R. Peng, A. B. Rao, and A. Sidford, *Solving directed Laplacian systems in nearly-linear time through sparse LU factorizations*, in Proceedings of the 59th Annual Symposium on Foundations of Computer Science, IEEE, 2018, pp. 898–909.

[14] M. B. Cohen, J. Kelner, J. Peebles, R. Peng, A. Sidford, and A. Vladu, *Faster algorithms for computing the stationary distribution, simulating random walks, and more*, in Proceedings of the 57th Annual Symposium on Foundations of Computer Science, IEEE, 2016, pp. 583–592.

[15] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
[16] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 7426–7431.
[17] A. R. Dinner, E. H. Thiede, B. Van Koten, and J. Weare, *Stratification as a general variance reduction method for Markov chain Monte Carlo*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 1139–1188.
[18] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*, https://arxiv.org/abs/math/0001057, 2000.
[19] C. Elkan, *Using the triangle inequality to accelerate k-means*, in Proceedings of the Twentieth International Conference on International Conference on Machine Learning, AAAI, 2003, pp. 147–153.
[20] M. Fanuel, C. M. Alaíz, Á. Fernández, and J. A. Suykens, *Magnetic eigenmaps for the visualization of directed networks*, Appl. Comput. Harmon. Anal., 44 (2018), pp. 189–199.
[21] K. Fitch, *Effective resistance preserving directed graph symmetrization*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 49–65.
[22] G. Golnari, Z.-L. Zhang, and D. Boley, *Markov fundamental tensor and its applications to network analysis*, Linear Algebra Appl., 564 (2019), pp. 126–158.
[23] G. Hamerly, *Making k-means even faster*, in Proceedings of the 2010 SIAM International Conference on Data Mining, 2010, pp. 130–140.
[24] W. L. Hamilton, R. Ying, and J. Leskovec, *Representation learning on graphs: Methods and applications*, IEEE Data Eng. Bull., 40 (2017), pp. 42–51.
[25] Jianbo Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.
[26] L. Kaufmann, *Clustering by means of medoids*, in Proceedings of the Conference on Statistical Data Analysis Based on the L^1 Norm, Neuchatel, 1987, 1987, pp. 405–416.
[27] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, Grad. Texts in Math. 40, Springer, New York, 1976.
[28] S. S. Lafon, *Diffusion Maps and Geometric Harmonics*, Ph.D. thesis, Yale University, New Haven, CT, 2004.
[29] D. Lai, H. Lu, and C. Nardini, *Extracting weights from edge directions to find communities in directed networks*, J. Stat. Mech. Theory Exp., 2010 (2010), P06003.
[30] Y. Li and Z.-L. Zhang, *Random walks on digraphs, the generalized digraph Laplacian and the degree of asymmetry*, in Proceedings of the International Workshop on Algorithms and Models for the Web-Graph, 2010, pp. 74–85.
[31] D. Liben-Nowell and J. Kleinberg, *The link prediction problem for social networks*, in Proceedings of the Twelfth International Conference on Information and Knowledge Management, New York, 2003.
[32] N. Madras and D. Randall, *Markov chain decomposition for convergence rate analysis*, Ann. Appl. Probab., 12 (2001), pp. 581–606.
[33] F. D. Malliaros and M. Vazirgiannis, *Clustering and community detection in directed networks: A survey*, Phys. Rep., 533 (2013), pp. 95–142.
[34] R. A. Martin and D. Randall, *Sampling adsorbing staircase walks using a new Markov chain decomposition method*, in Proceedings of the 41st Annual Symposium on Foundations of Computer Science, New York, 2000, IEEE, pp. 492–502.
[35] B. Mitavskiy, J. Rowe, A. Wright, and L. M. Schmitt, *Quotients of Markov chains and asymptotic properties of the stationary distribution of the Markov chain associated to an evolutionary algorithm*, Genet. Program. Evolvable Mach., 9 (2008), pp. 109–123.
[36] A. Moore, *The anchors hierarchy: Using the triangle inequality to survive high dimensional data*, in Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence, AAAI, 2000, pp. 397–405.
[37] J. R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.
[38] B. Osting and T. H. Reeb, *Consistency of Dirichlet partitions*, SIAM J. Math. Anal., 49 (2017), pp. 4251–4274.
[39] H.-S. Park and C.-H. Jun, *A simple and fast algorithm for k-medoids clustering*, Expert Systems Appl., 36 (2009), pp. 3336–3341.

[40] K. Pearson, *On lines and planes of closest fit to systems of points in space*, London Edinburgh Dublin Philos. Mag. J. Sci., 2 (1901), pp. 559–572.
[41] S. Pitis, H. Chan, K. Jamali, and J. Ba, *An inductive bias for distances: Neural nets that respect the triangle inequality*, in Proceedings of the Eighth International Conference on Learning Representations, 2020.
[42] M. R. Rozinas, *Metric on State Space of Markov Chain*, preprint, arXiv:1004.4264 [math. PR], 2010, https://arxiv.org/abs/1004.4264.
[43] V. Satuluri and S. Parthasarathy, *Symmetrizations for clustering directed graphs*, in Proceedings of the 14th International Conference on Extending Database Technology, 2011, pp. 343–354.
[44] G. L. Scott and H. C. Longuet-higgins, *Feature grouping by relocalisation of eigenvectors of proximity matrix*, in Proceedings of the British Machine Vision Conference, 1990, pp. 103–108.
[45] A. Singer and H.-T. Wu, *Vector diffusion maps and the connection Laplacian*, Comm. Pure Appl. Math., 65 (2012), pp. 1067–1144.
[46] A. Singer and H.-T. Wu, *Spectral convergence of the connection Laplacian from random samples*, Inf. Inference, 6 (2017), pp. 58–123.
[47] E. Thiede, B. V. Koten, and J. Weare, *Sharp entrywise perturbation bounds for Markov chains*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 917–941.
[48] N. G. Trillos and D. Slepčev, *A variational approach to the consistency of spectral clustering*, Appl. Comput. Harmon. Anal., 45 (2018), pp. 239–281.
[49] N. G. Trillos, D. Slepčev, J. Von Brecht, T. Laurent, and X. Bresson, *Consistency of Cheeger and ratio graph cuts*, J. Mach. Learn. Res., 17 (2016), pp. 6268–6313.
[50] A. Tversky, *Features of similarity*, Psychol. Rev., 84 (1977), pp. 327–352.
[51] Y. van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi, *Mean curvature, threshold dynamics, and phase field theory on finite graphs*, Milan J. Math., 82 (2014), pp. 3–65.
[52] F. Völlering, *Constant Curvature Metrics for Markov Chains*, preprint, arXiv:1712.02762 [math. PR], 2017, https://arxiv.org/abs/1712.02762.
[53] U. Von Luxburg, A. Radl, and M. Hein, *Hitting and commute times in large random neighborhood graphs*, J. Mach. Learn. Res., 15 (2014), pp. 1751–1798.
[54] G. F. Young, L. Scardovi, and N. E. Leonard, *A new notion of effective resistance for directed graphs—Part* I*: Definition and properties*, IEEE Trans. Automat. Control, 61 (2016), pp. 1727–1736.
[55] G. F. Young, L. Scardovi, and N. E. Leonard, *A new notion of effective resistance for directed graphs—Part* II*: Computing resistances*, IEEE Trans. Automat. Control, 61 (2016), pp. 1737–1752.
[56] A. Yuan, J. Calder, and B. Osting, *A Continuum Limit for the PageRank Algorithm*, preprint, arXiv:2001.08973 [math. AP], 2020, https://arxiv.org/abs/2001.08973.
[57] L. Zelnik-Manor and P. Perona, *Self-tuning spectral clustering*, in Neural Information Processing Systems, Cambridge, MA, 2004, MIT Press, pp. 1601–1608.
[58] D. Zhou, T. Hofmann, and B. Schölkopf, *Semi-supervised learning on directed graphs*, in Proceedings of the 17th International Conference on Neural Information Processing Systems, pp. 1633–1640.