

A topological interpretation of the walk distances

Pavel Chebotarev*

Michel Deza†

Abstract

The walk distances in graphs have no direct interpretation in terms of walk weights, since they are introduced via the *logarithms* of walk weights. Only in the limiting cases where the logarithms vanish such representations follow straightforwardly. The interpretation proposed in this paper rests on the identity $\ln \det B = \text{tr} \ln B$ applied to the cofactors of the matrix $I - tA$, where A is the weighted adjacency matrix of a weighted multigraph and t is a sufficiently small positive parameter. In addition, this interpretation is based on the power series expansion of the logarithm of a matrix. Kasteleyn [13] was probably the first to apply the foregoing approach to expanding the determinant of $I - A$. We show that using a certain linear transformation the same approach can be extended to the cofactors of $I - tA$, which provides a topological interpretation of the walk distances.

Keywords: Graph distances; Walk distances; Transitional measure; Network

MSC: 05C12, 05C50, 51K05, 15A09, 15A15

1 Introduction

The walk distances for graph vertices were proposed in [4] and studied in [5]. Along with their modifications they generalize [5] the logarithmic forest distances [3], resistance distance, shortest path distance, and the weighted shortest path distance. The walk distances are graph-geodetic: for a distance¹ $d(i, j)$ in a graph G this means that $d(i, j) + d(j, k) = d(i, k)$ if and only if every path in G connecting i and k visits j .

It is well known that the resistance distance between two adjacent vertices in a tree is equal to 1. In contrast to this, the walk distances take into account the centrality of vertices. For example, any walk distance between two central adjacent vertices in a path turns out [5] to be less than that between two peripheral adjacent vertices. This property may be desirable in some applications including machine learning, mathematical chemistry, the analysis of social and biological networks, etc.

In the present paper, we obtain a topological interpretation of the simplest walk distances. Such an interpretation is not immediate from the definition, since the walk distances are introduced via the *logarithms* of walk weights. Only in the limiting cases where the logarithms

*Institute of Control Sciences of the Russian Academy of Sciences, 65 Profsoyuznaya Street, Moscow 117997, Russia, E-mail: chv@member.ams.org.

†Laboratoire de Geometrie Appliquee, LIGA, Ecole Normale Supérieure, 45, rue d'Ulm, F-75230, Paris, Cedex 05, France, E-mail: Michel.Deza@ens.fr.

¹In this paper, a *distance* is assumed to satisfy the axioms of metric.

vanish such representations follow straightforwardly [5]. The interpretation we propose rests on the identity $\ln \det B = \text{tr} \ln B$ applied to the cofactors of the matrix $I - tA$, where A is the weighted adjacency matrix of a weighted multigraph and t is a sufficiently small positive parameter. In addition, it is based on the power series expansion of the logarithm of a matrix. We do not employ these identities explicitly; instead, we make use of a remarkable result by Kasteleyn [13] based on them. More specifically, Kasteleyn obtained an expansion of the determinant of $I - A$ and the logarithm of this determinant. We show that using a certain linear transformation the same approach can be extended to the cofactors of $I - tA$, which provides a topological interpretation of the walk distances.

2 Notation

In the graph definitions we mainly follow [10]. Let G be a weighted multigraph (a weighted graph where multiple edges are allowed) with vertex set $V(G) = V$, $|V| = n > 2$, and edge set $E(G)$. Loops are allowed; we assume that G is connected. For brevity, we will call G a *graph*. For $i, j \in V(G)$, let $n_{ij} \in \{0, 1, \dots\}$ be the number of edges incident to both i and j in G ; for every $q \in \{1, \dots, n_{ij}\}$, $w_{ij}^q > 0$ is the weight of the q th edge of this type. Let

$$a_{ij} = \sum_{q=1}^{n_{ij}} w_{ij}^q \quad (1)$$

(if $n_{ij} = 0$, we set $a_{ij} = 0$) and $A = (a_{ij})_{n \times n}$; A is the symmetric *weighted adjacency matrix* of G . In what follows, all matrix entries are indexed by the vertices of G . This remark is essential when submatrices are considered: say, “the i th column” of a submatrix of A means “the column corresponding to the vertex i of G ” rather than just the “column number i .”

By the *weight* of a graph G , $w(G)$, we mean the product of the weights of all its edges. If G has no edges, then $w(G) = 1$. The weight of a set \mathcal{S} of graphs, $w(\mathcal{S})$, is the total weight (the sum of the weights) of its elements; $w(\emptyset) = 0$.

For $v_0, v_m \in V(G)$, a $v_0 \rightarrow v_m$ *walk* in G is an arbitrary alternating sequence of vertices and edges $v_0, e_1, v_1, \dots, e_m, v_m$ where each e_i is a (v_{i-1}, v_i) edge. The *length* of a walk is the number m of its edges (including loops and repeated edges). The *weight* of a walk is the product of the m weights of its edges. The weight of a set of walks is the total weight of its elements. By definition, for any vertex v_0 , there is one $v_0 \rightarrow v_0$ walk v_0 with length 0 and weight 1.

We will need some special types of walks. A *hitting* $v_0 \rightarrow v_m$ *walk* is a $v_0 \rightarrow v_m$ walk containing only one occurrence of v_m . A $v_0 \rightarrow v_m$ walk is called *closed* if $v_m = v_0$ and *open* otherwise. The *multiplicity* of a closed walk is the maximum μ such that the walk is a μ -fold repetition of some walk.

We say that two closed walks of non-zero length are *phase twins* if the edge sequence e_1, e_2, \dots, e_m of the first walk can be obtained from the edge sequence e'_1, e'_2, \dots, e'_m of the second one by a cyclic shift. For example, the walks $v_0, e_1, v_1, e_2, v_2, e_3, v_0$ and $v_2, e_3, v_0, e_1, v_1, e_2, v_2$ are phase twins. A *circuit* [11, 13] in G is any equivalence class of phase twins. The *multiplicity* of a circuit is the multiplicity of any closed walk it contains (all such walks obviously have the same multiplicity). A walk (circuit) whose multiplicity exceeds 1 is *periodic*.

Let r_{ij} be the weight of the set \mathcal{R}^{ij} of all $i \rightarrow j$ walks in G provided that this weight is finite. $R = R(G) = (r_{ij})_{n \times n} \in \mathbb{R}^{n \times n}$ will be referred to as the *matrix of the walk weights* of G .

It was shown in [4] that if R exists then it *determines a transitional measure in G* , that is, (i) it satisfies the transition inequality

$$r_{ij} r_{jk} \leq r_{ik} r_{jj}, \quad i, j, k = 1, \dots, n \quad (2)$$

and (ii) $r_{ij} r_{jk} = r_{ik} r_{jj}$ if and only if every path from i to k visits j .

3 The walk distances

For any $t > 0$, consider the graph tG obtained from G by multiplying all edge weights by t . If the matrix of the walk weights of tG , $R_t = R(tG) = (r_{ij}(t))_{n \times n}$, exists, then²

$$R_t = \sum_{k=0}^{\infty} (tA)^k = (I - tA)^{-1}, \quad (3)$$

where I denotes the identity matrix of appropriate dimension.

By assumption, G is connected, while its edge weights are positive, so R_t is also positive. Apply the logarithmic transformation to the entries of R_t , namely, consider the matrix

$$H_t = \overrightarrow{\ln R_t}, \quad (4)$$

where $\overrightarrow{\varphi(S)}$ stands for elementwise operations, i.e., operations applied to each entry of a matrix S separately. Finally, consider the matrix

$$D_t = \frac{1}{2} (h_t \mathbf{1}^T + \mathbf{1} h_t^T - H_t - H_t^T), \quad (5)$$

where h_t is the column vector containing the diagonal entries of H_t , $\mathbf{1}$ is the vector of ones of appropriate dimension, and h_t^T and $\mathbf{1}^T$ are the transposes of h_t and $\mathbf{1}$. An alternative form of (5) is $D_t = (U_t + U_t^T)/2$, where $U_t = h_t \mathbf{1}^T - H_t$, and its elementwise form is

$$d_{ij}(t) = \frac{1}{2} (h_{ii}(t) + h_{jj}(t) - h_{ij}(t) - h_{ji}(t)), \quad i, j \in V(G), \quad (6)$$

where $H_t = (h_{ij}(t))$ and $D_t = (d_{ij}(t))$. This is a standard transformation used to obtain a distance from a proximity measure (cf. the inverse covariance mapping in [7, Section 5.2] and the cosine law in [8]).

In the rest of this section, we present several known facts (lemmas) which will be of use in what follows, one simple example, and two remarks.

Lemma 1 ([4]). *For any connected G , if $R_t = (r_{ij}(t))$ exists, then the matrix $D_t = (d_{ij}(t))$ defined by (3)–(5) determines a graph-geodesic distance $d_t(i, j) = d_{ij}(t)$ on $V(G)$.*

²In the more general case of weighted *digraphs*, the ij -entry of the matrix $R_t - I$ is called the *Katz similarity* between vertices i and j . Katz [14] proposed it to evaluate the social status taking into account all $i \rightarrow j$ paths. Among many other papers, this index was studied in [13, 23].

This enables one to give the following definition.

Definition 1. For a connected graph G , the *walk distances* on $V(G)$ are the functions $d_t(i, j) : V(G) \times V(G) \rightarrow \mathbb{R}$ and the functions, $d_t^w(i, j)$, positively proportional to them, where $d_t(i, j) = d_{ij}(t)$ and $D_t = (d_{ij}(t))$ is defined by (3)–(5).

Example 1. For the multigraph G shown in Fig. 1,

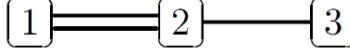


Figure 1: A multigraph G on 3 vertices.

the weighted adjacency matrix is

$$A = \begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

the matrix $R_{\frac{1}{3}}$ of the walk weights of $\frac{1}{3}G$ exists and has the form

$$R_{\frac{1}{3}} = R(\frac{1}{3}G) = (r_{ij}(\frac{1}{3})) = \frac{1}{4} \begin{bmatrix} 8 & 6 & 2 \\ 6 & 9 & 3 \\ 2 & 3 & 5 \end{bmatrix},$$

and the computation (5) of the walk distances $d_t(i, j)$ with parameter $t = \frac{1}{3}$ yields

$$D_{\frac{1}{3}} = (d_{ij}(\frac{1}{3})) = \frac{1}{2} \begin{bmatrix} 0 & \ln 2 & \ln 10 \\ \ln 2 & 0 & \ln 5 \\ \ln 10 & \ln 5 & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 0.35 & 1.15 \\ 0.35 & 0 & 0.80 \\ 1.15 & 0.80 & 0 \end{bmatrix}.$$

Since the walk distances are graph-geodetic (Lemma 1) and all paths from 1 to 3 visit 2, $d_{\frac{1}{3}}(1, 2) + d_{\frac{1}{3}}(2, 3) = d_{\frac{1}{3}}(1, 3)$ holds.

Regarding the existence of R_t , since for a connected graph A is irreducible, the Perron-Frobenius theory of nonnegative matrices provides the following result (cf. [23, Theorem 4]).

Lemma 2. For any weighted adjacency matrix A of a connected graph G , the series $R_t = \sum_{k=0}^{\infty} (tA)^k$ with $t > 0$ converges to $(I - tA)^{-1}$ if and only if $t < \rho^{-1}$, where $\rho = \rho(A)$ is the spectral radius of A . Moreover, ρ is an eigenvalue of A ; as such ρ has multiplicity 1 and a positive eigenvector.

Observe that for the graph G of Example 1, $\rho = \sqrt{5}$, so $\frac{1}{3} = t < \rho^{-1}$ is satisfied.

Lemma 3. For any vertices $i, j \in V(G)$ and $0 < t < \rho^{-1}$,

$$d_t(i, j) = -\ln \left(\frac{r_{ij}(t)}{\sqrt{r_{ii}(t) r_{jj}(t)}} \right). \quad (7)$$

Lemma 3 is a corollary of (4), (5), and Lemma 2.

On the basis of Lemma 3, the walk distances can be given the following short definition: $d_t(i, j) = -\ln r'_{ij}(t)$, where $r'_{ij}(t) = \frac{r_{ij}(t)}{\sqrt{r_{ii}(t) r_{jj}(t)}}$ and $R_t = (r_{ij}(t))_{n \times n}$ is defined by (3).

Remark 1. Consider another transformation of the correlation-like index $r'_{ij}(t) = \frac{r_{ij}(t)}{\sqrt{r_{ii}(t) r_{jj}(t)}}$:

$$d'_t(i, j) = 1 - \frac{r_{ij}(t)}{\sqrt{r_{ii}(t) r_{jj}(t)}}. \quad (8)$$

Is $d'_t(i, j)$ a metric? It follows from Definition 1, (7), and (8) that for any walk distance $d_t^w(i, j)$, there exists $\lambda > 0$ such that

$$d'_t(i, j) = 1 - e^{-\lambda d_t^w(i, j)}. \quad (9)$$

Eq. (9) is the Schoenberg transform [21, 22] (see also [7, Section 9.1] and [1, 15]). As mentioned in [6], an arbitrary function $\tilde{d}(i, j)$ is the result of the Schoenberg transform of some metric if and only if $\tilde{d}(i, j)$ is a *P-metric*, i.e., a metric with values in $[0, 1]$ that satisfies the *correlation triangle inequality*

$$1 - \tilde{d}(i, k) \geq (1 - \tilde{d}(i, j))(1 - \tilde{d}(j, k)), \quad (10)$$

which can be rewritten as $\tilde{d}(i, k) \leq \tilde{d}(i, j) + \tilde{d}(j, k) - \tilde{d}(i, j) \tilde{d}(j, k)$.

This fact implies that (8) defines a P-metric. It is easily seen that the correlation triangle inequality for $d'_t(i, j)$ reduces to the transition inequality (2); obviously, it can be given a probabilistic interpretation.

For the graph G of Example 1, the P-metric $d'_t(i, j)$ with $t = \frac{1}{3}$ is given by the matrix

$$D'_{\frac{1}{3}} = (d'_{ij}(\frac{1}{3})) = \begin{bmatrix} 0 & 1 - \sqrt{0.5} & 1 - \sqrt{0.1} \\ 1 - \sqrt{0.5} & 0 & 1 - \sqrt{0.2} \\ 1 - \sqrt{0.1} & 1 - \sqrt{0.2} & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 0.29 & 0.68 \\ 0.29 & 0 & 0.55 \\ 0.68 & 0.55 & 0 \end{bmatrix}.$$

Remark 2. It can be noted that the Nei standard genetic distance [17] and the Jiang-Conrath semantic distance [12] have a form similar to (7). Moreover, the transformation $-\ln(r(i, j))$ where $r(i, j)$ is a similarity measure between objects i and j was used in the construction of the Bhattacharyya distance between probability distributions [2] and the Tomiuk-Loeschke genetic distance [24] (see also the Leacock-Chodorow similarity [16] and the Resnik similarity [19]). These and other distances and similarities are surveyed in [6].

4 An interpretation of the walk distances

For a fixed t : $0 < t < \rho^{-1}$, where $\rho = \rho(A)$ let us use the notation

$$B = I - tA. \quad (11)$$

Assume that i and $j \neq i$ are also fixed and that $i+j$ is even; otherwise this can be achieved by renumbering the vertices. Hence, using (3)–(6), the positivity of $R_t = (I - tA)^{-1}$, and the determinant representation of the inverse matrix we obtain

$$d_t(i, j) = 0.5(\ln \det B_{\bar{i}\bar{i}} + \ln \det B_{\bar{j}\bar{j}} - \ln \det B_{\bar{i}\bar{j}} - \ln \det B_{\bar{j}\bar{i}}), \quad (12)$$

where $B_{\bar{i}\bar{j}}$ is B with row i and column j removed.

4.1 Logarithms of the cofactors: expressions in terms of circuits

To obtain an interpretation of the right-hand side of (12), we need the following remarkable result due to Kasteleyn.

Lemma 4 (Kasteleyn [13]). *For a digraph Γ with a weighted adjacency matrix \tilde{A} ,*

$$\det(I - \tilde{A}) = \exp \left(- \sum_{c \in \mathcal{C}} \frac{w(c)}{\mu(c)} \right) \quad (13)$$

$$= \prod_{c \in \mathcal{C}_1} (1 - w(c)), \quad (14)$$

where \mathcal{C} and \mathcal{C}_1 are the sets of all circuits and of all non-periodic circuits in Γ , $w(c)$ and $\mu(c)$ being the weight and the multiplicity of the circuit c .

The representation (13) was obtained by considering the generating function of walks in Γ . Basically, the sum $\sum_{c \in \mathcal{C}} \frac{w(c)}{\mu(c)}$ is a formal counting series in abstract *weight variables* (cf. [20, p. 19]). However, as soon as the weights are real and thus the generating function is a function in real counting variables, the issue of convergence arises. Since (13) is based on the power expansion $-\ln(I - \tilde{A}) = \sum_{k=1}^{\infty} k^{-1} \tilde{A}^k$, a necessary condition of its validity in the real-valued setting is $\rho(\tilde{A}) < 1$.

When the arc weights are nonnegative, the same condition is sufficient. However, if some vertices i and j are connected by parallel $i \rightarrow j$ arcs carrying weights of different signs, then the problem of conditional convergence arises. Namely, if the absolute values of such weights are large enough, then, even though $\rho(\tilde{A}) < 1$, by choosing the order of summands in the right-hand side of (13), the sum can be made divergent or equal to any given number.

To preserve (13) in the latter case, the order of summands must be adjusted with an arbitrary order of items in $\sum_{k=1}^{\infty} k^{-1} \tilde{A}^k$. Hence it suffices to rewrite (13) in the form

$$\det(I - \tilde{A}) = \exp \left(- \sum_{k=1}^{\infty} \sum_{c \in \mathcal{C}_k} \frac{w(c)}{\mu(c)} \right), \quad (15)$$

where \mathcal{C}_k is the set of all circuits that involve k arcs in Γ .

Lemma 4 is also applicable to undirected graph. To verify this, it is sufficient to replace an arbitrary undirected graph G with its *directed version*, i.e., the digraph obtained from G by replacing every edge by two opposite arcs carrying the weight of that edge.

Since by (11), $B_{\bar{i}} = I - (tA)_{\bar{i}\bar{i}}$, Lemma 4 can be used to evaluate $\ln \det B_{\bar{i}}$. Let $G_{\bar{i}}$ ($G_{\bar{i}\bar{j}}$) be G with vertex i (vertices i and j) and all edges incident to i (i and j) removed.

Corollary 1.

$$-\ln \det B_{\bar{i}} = \sum_{c \in \mathcal{C}^{\bar{i}}} \frac{w(c)}{\mu(c)} = \sum_{c \in \mathcal{C}^{\bar{i}\bar{j}} \cup \mathcal{C}^{j\bar{i}}} \frac{w(c)}{\mu(c)},$$

where

- $\mathcal{C}^{\bar{i}}$ is the set of circuits in $tG_{\bar{i}}$,
- $\mathcal{C}^{\bar{i}\bar{j}}$ is the set of circuits in $tG_{\bar{i}\bar{j}}$,

• $\mathcal{C}^{j\bar{i}}$ is the set of circuits visiting j , but not i in tG , $w(c)$ and $\mu(c)$ being the weight and the multiplicity of c .

Proof. By assumption, $0 < t < \rho^{-1}(A)$; $B_{\bar{i}\bar{i}} = I - tA_{\bar{i}\bar{i}}$. Since A is irreducible, $\rho(tA_{\bar{i}\bar{i}}) < \rho(tA) < 1$ [9, Ch. III, § 3.4]. Moreover, the edge weights in G are positive by assumption. Therefore, the expansion (13) holds for $B_{\bar{i}\bar{i}}$, which yields the desired statement. \square

To interpret (12), we also need an expansion of $\ln \det B_{\bar{i}\bar{j}}$ ($j \neq i$). Convergence in such an expansion provided by Lemma 4 can be achieved by applying a suitable linear transformation of $B_{\bar{i}\bar{j}}$.

For the fixed i and $j \neq i$, consider the matrix

$$T_{ij} = I(j, i)_{\bar{i}\bar{j}}, \quad (16)$$

where $I(j, i)$ differs from $I_{n \times n}$ by the ji -entry: $I(j, i)_{ji} = -1$.

The reader can easily construct examples of T_{ij} and verify the following properties.

Lemma 5.

1. The columns of T_{ij} form an orthonormal set, i.e., T_{ij} is orthogonal: $T_{ij}^T T_{ij} = I$.
2. If $i + j$ is even (as assumed), then $\det T_{ij} = 1$.
3. $T_{ij}^T = T_{ji}$.
4. For any $M_{n \times n}$, $M_{\bar{i}\bar{j}} T_{ij}^{-1}$ is obtained from M by: (i) deleting row i , (ii) multiplying column i by -1 , and (iii) moving it into the position of column j .

The proof of Lemma 5 is straightforward.

Corollary 2. 1. $I_{\bar{i}\bar{j}} T_{ij}^{-1}$ is obtained from $I_{(n-1) \times (n-1)}$ by replacing the kk -entry with 0, where

$$k = \begin{cases} j, & j < i, \\ j-1, & j > i. \end{cases} \quad (17)$$

2. $I_{\bar{i}\bar{j}} T_{ij}^{-1} I_{\bar{i}\bar{j}} = I_{\bar{i}\bar{j}}$, i.e., T_{ij}^{-1} is a g -inverse [18] of $I_{\bar{i}\bar{j}}$.

Since $\det T_{ji} = 1$ (Lemma 5), we have

$$\det B_{\bar{i}\bar{j}} = \det(B_{\bar{i}\bar{j}} T_{ji}). \quad (18)$$

Now we apply Kasteleyn's Lemma 4 to $B_{\bar{i}\bar{j}} T_{ji}$ by considering a (multi)digraph Γ whose weighted adjacency matrix is

$$\mathcal{A} = I - B_{\bar{i}\bar{j}} T_{ji}, \quad (19)$$

where B is defined by (11). Namely, Lemma 4 in the form (15) along with (18) yield

Lemma 6.

$$-\ln \det B_{\bar{i}\bar{j}} = \sum_{k=1}^{\infty} \sum_{c \in \mathcal{C}'_k} \frac{w(c)}{\mu(c)}, \quad (20)$$

where \mathcal{C}'_k is the set of all circuits that involve k arcs in a digraph Γ whose weighted adjacency matrix is \mathcal{A} , while $w(c)$ and $\mu(c)$ are the weight and the multiplicity of the circuit c .

As well as (13), (20) is applicable to the case of formal counting series. However, in (11), t is a real weight variable. In this case, a necessary and sufficient condition of the convergence in (20) is $\rho(\mathcal{A}) < 1$.

Let us clarify the relation of Γ and its circuits with G and its topology. This is done in the following section.

4.2 The walk distances: An expression in terms of walks

To elucidate the structure of the digraph Γ introduced in Lemma 6, an algorithmic description of the matrix \mathcal{A} is useful.

Lemma 7. \mathcal{A} can be obtained from tA by: replacing ta_{ji} with $ta_{ji} - 1$, deleting row i , multiplying column i by -1 , and moving it into the position of column j .

Proof. By (19), items 1 and 3 of Lemma 5, (16), and (11) we have

$$\mathcal{A} = (T_{ij} - B_{\overline{ij}})T_{ij}^{-1} = (I(j, i) - I + tA)_{\overline{ij}}T_{ij}^{-1}.$$

Now the result follows from item 4 of Lemma 5. \square

Let us reformulate Lemma 7 in terms of G and Γ . Recall that a digraph is the *directed version* of a graph if it is obtained by replacing every edge in the graph by two opposite arcs carrying the weight of that edge.

Corollary 3. A digraph Γ with weighted adjacency matrix \mathcal{A} can be obtained from tG by:

- taking the directed version of the restriction of tG to $V(G) \setminus \{i, j\}$ and
- adding a vertex ij with: two loops of weights 1 and $-ta_{ji}$ (negative³), weights ta_{jm} of outgoing arcs, and weights $-ta_{mi}$ of incoming arcs, where $m \in V(G) \setminus \{i, j\}$.

Vertex ij is represented in \mathcal{A} by row and column k , where k is given by (17).

In what follows, Γ denotes the digraph defined in Corollary 3. The *jump* in Γ is the loop of weight 1 at ij . The walk in Γ that consists of one jump is called the *jump walk* (at ij).

To interpret $\ln \det B_{\overline{ij}}$ in terms of G , we need the following notation.

Definition 2. A walk with i, j jumps in G is any walk in the graph G' obtained from G by attaching two additional loops of weight 1: one adjacent to vertex i and one adjacent to j . These loops are called *jumps*. A walk with i, j jumps (in G) only consisting of one jump is called a *jump walk* (at i or j).

Definition 3. A $j \rightarrow i$ alternating walk with jumps is any $j \rightarrow i$ walk w with j, i jumps such that (a) any $j \dots j$ subwalk of w either visits i or contains no edges except for jumps and (b) any $i \dots i$ subwalk of w either visits j or contains no edges except for jumps.

A $j \rightarrow i \rightarrow j$ alternating walk with jumps is defined similarly: the only difference is that the endpoint of such a walk is j .

To introduce some additional notation, observe that any $j \rightarrow i$ alternating walk w with jumps can be uniquely partitioned into a sequence of subwalks (w_1, \dots, w_t) such that every two neighboring subwalks share one terminal vertex and each w_k is a jump walk or is a $j \rightarrow i$ or an $i \rightarrow j$ hitting walk without jumps. For every $k \in \{1, \dots, t\}$, consider the set $p_k = \{w_k, \tilde{w}_k\}$, where \tilde{w}_k is either w_k written from end to beginning (reversed⁴) when w_k is a hitting walk without jumps, or a jump walk at i (j) when w_k is a jump walk at j (resp., i). The sequence $p(w) = (p_1, \dots, p_t)$ will be called the *route partition* of w . We say that two

³If $a_{ji} = 0$, then this loop is omitted.

⁴Cf. “dihedral equivalence” in [11].

$j \rightarrow i$ alternating walks with jumps, w and w' , are *equipartite* if the route partition of w' can be obtained from that of w by a cyclic shift. Finally, any equivalence class of equipartite $j \rightarrow i$ alternating walks with jumps will be called an *alternating $j \rightarrow i$ route with jumps*. If r is such a route, then its *length* and *weight* are defined as the common length and weight of all walks with jumps it includes, respectively. If a route partition $p(w) = (p_1, \dots, p_t)$ has period (the length of the elementary repeating part) y , then the *multiplicity* of the alternating $j \rightarrow i$ route with jumps that corresponds to $p(w)$ is defined to be t/y .

Completely the same construction can be applied to define *alternating $j \rightarrow i \rightarrow j$ route with jumps* (starting with the above definition of a $j \rightarrow i \rightarrow j$ alternating walk with jumps). A notable difference is that there are alternating $j \rightarrow i \rightarrow j$ routes with jumps that do not visit i : these consist of jumps at j . The weight of such a route with jumps is 1 and its multiplicity is the number of jumps.

Lemma 8. *There is a one-to-one correspondence between the set of circuits in Γ that contain vertex ij and have odd (even) numbers of negatively weighted arcs and the set of alternating $j \rightarrow i$ routes (alternating $j \rightarrow i \rightarrow j$ routes) with jumps in G . The circuit in Γ and route with jumps in G that correspond to each other have the same length, weight, and multiplicity.*

Proof. Every circuit containing vertex ij in Γ can be uniquely represented by a cyclic sequence⁵ of walks each of which either is an $ij \rightarrow ij$ walk including exactly one negatively weighted arc, or is the jump walk at ij . Such a cyclic sequence uniquely determines an alternating $j \rightarrow i$ or $j \rightarrow i \rightarrow j$ route with jumps in G (if the number of negatively weighted arcs involved in the circuit is odd or even, respectively).

On the other hand, every set $p_k = \{w_k, \tilde{w}_k\}$ involved in an alternating $j \rightarrow i$ or $j \rightarrow i \rightarrow j$ route with jumps in G uniquely determines either an $ij \rightarrow ij$ walk containing exactly one negatively weighted arc, or the jump walk at ij in Γ . Thereby, every alternating route with jumps under consideration uniquely determines a circuit in Γ . Furthermore, the two correspondences described above are inverse to each other. Thus, these reduce to a one-to-one correspondence.

Finally, it is easily seen that the corresponding circuits and alternating routes with jumps share the same length, weight, and multiplicity. \square

Remark 3. It can be noted that the multiplicity of an alternating $j \rightarrow i$ route with jumps in G can only be odd.

Both circuits and alternating routes will be called *figures*. Lemmas 6 and 8 enable one to express $\ln \det B_{ij}$ in terms of figures in tG and tG_{ij} .

Lemma 9.

$$-\ln \det B_{ij} = \sum_{k=1}^{\infty} \sum_{c \in (\mathcal{C}_{ij} \cup \mathcal{C}^{j \rightarrow i \rightarrow j} \cup \mathcal{C}^{j \rightarrow i}) \cap \mathcal{C}_k} (-1)^{\zeta(c)} \frac{w(c)}{\mu(c)},$$

where

⁵A *cyclic sequence* is a set $X = \{x_1, \dots, x_N\}$ with the relation “next” $\eta = \{(x_2, x_1), \dots, (x_N, x_{N-1}), (x_1, x_N)\}$.

- $\mathcal{C}^{\overline{ij}}$ is the set of circuits in $tG_{\overline{ij}}$,
- $\mathcal{C}^{j \rightarrow i \rightarrow j}$ is the set of alternating $j \rightarrow i \rightarrow j$ routes with jumps in tG ,
- $\mathcal{C}^{j \rightarrow i}$ is the set of alternating $j \rightarrow i$ routes with jumps in tG ,
- \mathcal{C}_k is the set of figures (in tG or $tG_{\overline{ij}}$) that involve k arcs,

$$\zeta(c) = \begin{cases} 0, & c \in \mathcal{C}^{\overline{ij}} \cup \mathcal{C}^{j \rightarrow i \rightarrow j}, \\ 1, & c \in \mathcal{C}^{j \rightarrow i}, \end{cases}$$

while $w(c)$ and $\mu(c)$ are the weight and the multiplicity of c .

Similarly, we can express $\ln \det B_{\overline{ji}}$ in terms of the sets $\mathcal{C}^{\overline{ij}}$, $\mathcal{C}^{i \rightarrow j \rightarrow i}$, and $\mathcal{C}^{i \rightarrow j}$. There exist natural bijections between $\mathcal{C}^{j \rightarrow i \rightarrow j}$ and $\mathcal{C}^{i \rightarrow j \rightarrow i}$ and between $\mathcal{C}^{j \rightarrow i}$ and $\mathcal{C}^{i \rightarrow j}$. Namely, to obtain an element of $\mathcal{C}^{i \rightarrow j \rightarrow i}$ from $c \in \mathcal{C}^{j \rightarrow i \rightarrow j}$ (or an element of $\mathcal{C}^{i \rightarrow j}$ from $c \in \mathcal{C}^{j \rightarrow i}$), it suffices to reverse all $j \rightarrow i$ and $i \rightarrow j$ hitting walks without jumps in c and to replace every jump walk at j with the jump walk at i and vice versa.

On the other hand, the sets $\mathcal{C}^{i \rightleftharpoons j} \stackrel{\text{def}}{=} \mathcal{C}^{j \rightarrow i \rightarrow j} \cup \mathcal{C}^{i \rightarrow j \rightarrow i}$ and $\mathcal{C}^{i-j} \stackrel{\text{def}}{=} \mathcal{C}^{j \rightarrow i} \cup \mathcal{C}^{i \rightarrow j}$ also make sense. Specifically, they are useful for expressing $d_t(i, j)$. Such an expression is the main result of this paper. It follows by combining (12), Corollary 1, and Lemma 9.

Theorem 1.

$$d_t(i, j) = \frac{1}{2} \sum_{k=1}^{\infty} \sum_{c \in (\mathcal{C}^{i\overline{j}} \cup \mathcal{C}^{\overline{ij}} \cup \mathcal{C}^{i \rightleftharpoons j} \cup \mathcal{C}^{i-j}) \cap \mathcal{C}_k} (-1)^{\zeta(c)} \frac{w(c)}{\mu(c)},$$

where the sets of figures in tG are denoted by:

- $\mathcal{C}^{i\overline{j}}$: of circuits visiting i , but not j ,
- $\mathcal{C}^{\overline{ij}}$: of circuits visiting j , but not i ,
- $\mathcal{C}^{i \rightleftharpoons j}$: of alternating $j \rightarrow i \rightarrow j$ and $i \rightarrow j \rightarrow i$ routes with jumps,
- \mathcal{C}^{i-j} : of alternating $j \rightarrow i$ and $i \rightarrow j$ routes with jumps,
- \mathcal{C}_k : of figures that involve k arcs;

$$\zeta(c) = \begin{cases} 0, & c \in \mathcal{C}^{i \rightleftharpoons j}, \\ 1, & c \in \mathcal{C}^{i\overline{j}} \cup \mathcal{C}^{\overline{ij}} \cup \mathcal{C}^{i-j}, \end{cases}$$

while $w(c)$ and $\mu(c)$ are the weight and the multiplicity of c .

In more general terms, Theorem 1 can be interpreted as follows. The walk distance between i and j is reduced by $j \rightarrow i$ and $i \rightarrow j$ walks (see \mathcal{C}^{i-j}), connections of i with other vertices avoiding j ($\mathcal{C}^{i\overline{j}}$), and connections of j avoiding i ($\mathcal{C}^{\overline{ij}}$). The set $\mathcal{C}^{i \rightleftharpoons j}$ supplies all positive terms in the expansion of $d_t(i, j)$. It comprises constantly jumping walks along with closed walks involving i and j whose positive weights compensate the negative overweight of $j \rightarrow i$ and $i \rightarrow j$ routes with extra jumps.

Note that Theorem 1 supports the observation in the Introduction that the high centrality of i and j reduces, ceteris paribus, the walk distance between them. Indeed, the elements of $\mathcal{C}^{i\overline{j}} \cup \mathcal{C}^{\overline{ij}}$ which account for the centrality of i and j make a negative contribution to the distance.

The following example may provide some additional insight into Theorem 1.

Example 2. For the graph G of Example 1, let us approximate $d_{\frac{1}{3}}(1, 3) = \frac{1}{2} \ln 10 \approx 1.15$ using Theorem 1. Due to (19), $\mathcal{A} = \frac{1}{3} \begin{bmatrix} 0 & -2 \\ 1 & 3 \end{bmatrix}$. As $\rho(\mathcal{A}) = 2/3 < 1$, convergence holds in (20) and thus in Theorem 1. The leading terms of the expansion Theorem 1 provides for $d_{\frac{1}{3}}(1, 3)$ are presented in Table 1. In this table, $\frac{k}{\mu}(v_0 \cdots v_m)$ is the denotation of a collection of figures where each figure has multiplicity μ and contains some walk (or walk with jumps) whose sequence of vertices is v_0, \dots, v_m ; k is the cardinality of the collection. If $\mu = 1$, then μ is omitted; if $\mu = k = 1$, then μ and k are omitted.

\cap	$\mathcal{C}^{1\bar{3}} \cup \mathcal{C}^{\bar{1}3}$	$\mathcal{C}^{1\rightleftharpoons 3}$	\mathcal{C}^{1-3}
\mathcal{C}_1	\emptyset	$(11), (33)$	\emptyset
\mathcal{C}_2	$4(121), (323)$	$\frac{1}{2}(111), \frac{1}{2}(333)$	$2(123), 2(321)$
\mathcal{C}_3	\emptyset	$\frac{1}{3}(1111), \frac{1}{3}(3333)$	$2(1123), 2(3321)$
\mathcal{C}_4	$\frac{4}{2}(12121), 6(12121),$ $\frac{1}{2}(32323)$	$\frac{1}{4}(11111), \frac{1}{4}(33333),$ $\frac{2}{2}(12321), (12321), \frac{2}{2}(32123), (32123)$	$2(11123), 2(33321)$
\mathcal{C}_5	\emptyset	$\frac{1}{5}(111111), \frac{1}{5}(333333), 4(112321), 4(332123)$	$2(111123), 2(333321)$

Table 1: The figures forming the leading terms in the expansion of $d_{\frac{1}{3}}(1, 3)$ in Example 2.

The first terms of the series Theorem 1 provides are:

$$\begin{aligned}
d_{\frac{1}{3}}(1, 3) &= \frac{1}{2} \left[(2 \cdot 1) + \left(-\frac{4}{9} - \frac{1}{9} + 2 \cdot \frac{1}{2} - 2 \cdot \frac{2}{9} \right) + \left(2 \cdot \frac{1}{3} - 2 \cdot \frac{2}{9} \right) \right. \\
&\quad \left. + \left(-\frac{2+6}{81} - \frac{1}{2} \cdot \frac{1}{81} + 2 \left(\frac{1}{4} + \frac{1+1}{81} \right) - 2 \cdot \frac{2}{9} \right) + \left(2 \left(\frac{1}{5} + \frac{4}{81} \right) - 2 \cdot \frac{2}{9} \right) + \dots \right] \\
&= \frac{461}{405} + \dots,
\end{aligned}$$

where $\frac{461}{405} \approx 1.1383$.

In the above expression, the sum (with signs) of the weights of figures that involve k edges is 0 whenever k is even. Thus, the above expansion reduces to

$$d_{\frac{1}{3}}(1, 3) = \frac{1}{2} \left[(2 \cdot 1) + \left(2 \cdot \frac{1}{3} - 2 \cdot \frac{2}{9} \right) + \left(2 \left(\frac{1}{5} + \frac{4}{81} \right) - 2 \cdot \frac{2}{9} \right) + \dots \right].$$

The relative error of this approximation is 1.1%.

In some cases, the convergence of such expansions is extremely slow. On the other hand, the meaning of Theorem 1 is to clarify the concept of walk distance by representing it as the sum of route/circuit weights rather than to provide an effective algorithm for computing it.

References

- [1] F. Bavaud, On the Schoenberg transformations in data analysis: Theory and illustrations, *Journal of Classification* 28 (3) (2011) 297–314.
- [2] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society* 35 (1943) 99–109.
- [3] P. Chebotarev, A class of graph-geodetic distances generalizing the shortest-path and the resistance distances, *Discrete Applied Mathematics* 159 (5) (2011) 295–302.
- [4] P. Chebotarev, The graph bottleneck identity, *Advances in Applied Mathematics* 47 (3) (2011) 403–413.
- [5] P. Chebotarev, The walk distances in graphs, *Discrete Applied Mathematics*, In press. URL <http://dx.doi.org/10.1016/j.dam.2012.02.015>
- [6] M. M. Deza, E. Deza, *Encyclopedia of Distances*, Springer, Berlin–Heidelberg, 2009.
- [7] M. M. Deza, M. Laurent, *Geometry of Cuts and Metrics*, volume 15 of *Algorithms and Combinatorics*, Springer, Berlin, 1997.
- [8] F. Critchley, On certain linear mappings between inner-product and squared-distance matrices, *Linear Algebra and its Applications* 105 (1988) 91–107.
- [9] F. R. Gantmacher, *Applications of the Theory of Matrices*, Interscience, New York, 1959.
- [10] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.
- [11] F. Harary, A. Schwenk, The spectral approach to determining the number of walks in a graph, *Pacific Journal of Mathematics* 80 (2) (1979) 443–449.
- [12] J. J. Jiang, D. W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, 1997, 15 pp.
- [13] P. W. Kasteleyn, Graph theory and crystal physics, in: F. Harary (ed.), *Graph Theory and Theoretical Physics*, Academic Press, London, 1967, pp. 43–110.
- [14] L. Katz, A new status index derived from sociometric analysis, *Psychometrika* 18 (1) (1953) 39–43.
- [15] M. Laurent, A connection between positive semidefinite and Euclidean distance matrix completion problems, *Linear Algebra and its Applications* 273 (1-3) (1998) 9–22.
- [16] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (ed.), *WordNet. An electronic lexical database*, chap. 11, MIT Press, Cambridge, MA, 1998, pp. 265–283.

- [17] M. Nei, Genetic distance between populations, *The American Naturalist* 106 (949) (1972) 283–292.
- [18] C. R. Rao, S. K. Mitra, *Generalized Inverse of Matrices and its Applications*, Wiley, New York, 1971.
- [19] P. Resnik, Using information content to evaluate semantic similarity, in: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, vol. 1, Morgan Kaufmann Publishers, San Francisco, CA, 1995.
- [20] J. Riordan, *An Introduction to Combinatorial Analysis*, Wiley, New York, 1958.
- [21] I. J. Schoenberg, Remarks to M. Fréchet's article "Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert", *Annals of Mathematics* 36 (1935) 724–732.
- [22] I. J. Schoenberg, Metric spaces and positive definite functions, *Transactions of the American Mathematical Society* 44 (1938) 522–536.
- [23] G. L. Thompson, *Lectures on Game Theory, Markov Chains and Related Topics*, Monograph SCR-11, Sandia Corporation, Albuquerque, NM, 1958.
- [24] J. Tomiuk, V. Loeschcke, A new measure of genetic identity between populations of sexual and asexual species, *Evolution* 45 (1991) 1685–1694.