Bento and Ioannidis *Applied Network Science* (2019) 4:107
https://doi.org/10.1007/s41109-019-0219-z

Applied Network Science

RESEARCH **Open Access**

A family of tractable graph metrics

José Bento[1] and Stratis Ioannidis[2*]

*Correspondence:
ioannidis@ece.neu.edu
[2]Department of Electrical and Computer Engineering, Northeastern University, 360 Huntington Avenue, 02115 Boston, MA, USA
Full list of author information is available at the end of the article

Abstract
Important data mining problems such as nearest-neighbor search and clustering admit theoretical guarantees when restricted to objects embedded in a metric space. Graphs are ubiquitous, and clustering and classification over graphs arise in diverse areas, including, e.g., image processing and social networks. Unfortunately, popular distance scores used in these applications, that scale over large graphs, are not metrics and thus come with no guarantees. Classic graph distances such as, e.g., the chemical distance and the Chartrand-Kubiki-Shultz distance are arguably natural and intuitive, and are indeed also metrics, but they are intractable: as such, their computation does not scale to large graphs. We define a broad family of graph distances, that includes both the chemical and the Chartrand-Kubiki-Shultz distances, and prove that these are all metrics. Crucially, we show that our family includes metrics that are tractable. Moreover, we extend these distances by incorporating auxiliary node attributes, which is important in practice, while maintaining both the metric property and tractability.

Keywords: Metric spaces, Graph distances, Graph matching, Graph isomorphism, Convex optimization, Spectral algorithms

Introduction

Graph similarity and the related problem of graph isomorphism have a long history in data mining, machine learning, and pattern recognition (Conte et al. 2004; Macindoe and Richards 2010; Koutra et al. 2013). *Graph distances* naturally arise in this literature: intuitively, given two (unlabeled) graphs, their distance is a score quantifying their structural differences. A highly desirable property for such a score is that it is a *metric*, i.e., it is non-negative, symmetric, positive-definite, and, crucially, satisfies the triangle inequality. Metrics exhibit significant computational advantages over non-metrics. For example, operations such as nearest-neighbor search (Clarkson 2006; 1999; Beygelzimer et al. 2006), clustering (Ackermann et al. 2010), outlier detection (Angiulli and Pizzuti 2002), and diameter computation (Indyk 1999) admit fast algorithms precisely when performed over objects embedded in a metric space. To this end, proposing *tractable* graph metrics is of paramount importance in applying such algorithms to graphs.

Unfortunately, graph metrics of interest are often computationally expensive. A well-known example is the *chemical distance* (Kvasnička et al. 1991). Formally, given graphs G_A and G_B, represented by their adjacency matrices $A, B \in \{0,1\}^{n\times n}$, the chemical distance $d_{\mathbb{P}^n}(A,B)$ is defined in terms of a mapping between the two graphs that minimizes their edge discrepancies, i.e.:

$$d_{\mathbb{P}^n}(A,B) = \min_{P\in\mathbb{P}^n} \|AP - PB\|_F, \tag{1}$$

where $\mathbb{P}^n$ is the set of permutation matrices of size n and $\|\cdot\|_F$ is the Frobenius norm (see "Notation and preliminaries" section for definitions). The *Chartrand-Kubiki-Shultz (CKS)* (Chartrand et al. 1998) distance is an alternative: CKS is again given by (1) but, instead of edges, matrices A and B contain the pairwise shortest path distances between any two nodes.

The chemical and CKS distances have important properties. First, they are zero if and only if the graphs are isomorphic, which appeals to both intuition and practice; second, as desired, they are metrics over the quotient space defined by graph isomorphism (see "Notation and preliminaries" section); third, they have a natural interpretation, capturing global structural similarities between graphs. However, finding an optimal permutation P is notoriously hard; graph isomorphism, which is equivalent to deciding if there exists a permutation P such that $AP = PB$ (for both adjacency and path matrices), is famously a problem that is neither known to be in P nor shown to be NP-hard (Babai 2016). There is a large and expanding literature on scalable heuristics to estimate the optimal permutation P (Klau 2009; Bayati et al. 2009; Lyzinski et al. 2016; El-Kebir et al. 2015). Despite their computational advantages, unfortunately, using them to approximate $d_{\mathbb{P}^n}(A, B)$ breaks the metric property.

This significantly degrades the performance of many important tasks that rely on computing distances between graphs. For example, there is a clear separation on the approximability of clustering over metric and non-metric spaces (Ackermann et al. 2010). We also demonstrate this empirically in "Experiments" section (c.f. Fig. 1): attempting to cluster graphs sampled from well-known families based on non-metric distances significantly increases the misclassification rate, compared to clustering using metrics.

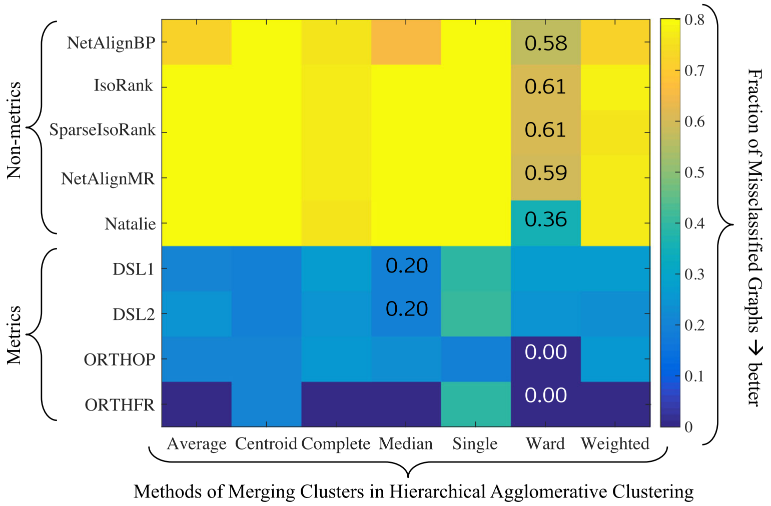


Fig. 1 Clustering Misclassification Error. A clustering experiment using metrics and non-metrics (y-axis) for different clustering parameters (x-axis). We sample graphs with $n = 50$ nodes from the six classes, shown in Table 3. We compute distances between them using nine different algorithms from Table 2. Only the distances in our family (DSL1, DSL2, ORTHOP, and ORTHFR) are metrics. The resulting graphs are clustered using hierarchical agglomerative clustering (Hartigan 1975) using *Average*, *Centroid*, *Complete*, *Median*, *Single*, *Ward*, *Weighted* as a means of merging clusters. Colors represent the fraction of misclassified graphs, with the minimal misclassification rate per distance labeled explicitly. Metrics outperform other distance scores across all clustering methods. The error rate of a random guess is ≈ 0.8.

An additional issue that arises in practice is that nodes often have attributes not associated with adjacency. For example, in social networks, nodes may contain profiles with a user's age or gender; similarly, nodes in molecules may be labeled by atomic numbers. Such attributes are not captured by the chemical or CKS distances. However, in such cases, only *label-preserving* permutations P may make sense (e.g., mapping females to females, oxygens to oxygens, etc.). Incorporating attributes while preserving the metric property is thus important from a practical perspective.

Contributions

We seek generalizations of the chemical and CKS distances that (a) *satisfy the metric property* and (b) are *tractable*: by this, we mean that they can be computed either by solving a convex optimization problem, or by a polynomial time algorithm. Specifically, we study generalizations of (1) of the form:

$$d_S(A,B) = \min_{P \in S} \|AP - PB\| \tag{2}$$

where $S \subset \mathbb{R}^{n \times n}$ is closed and bounded, $\|\cdot\|$ is a matrix norm, and $A, B \in \mathbb{R}^{n \times n}$ are arbitrary real matrices (representing adjacency, path distances, weights, etc.). We make the following contributions:

- We prove sufficient conditions on S and norm $\|\cdot\|$ under which (2) is a *pseudometric*, i.e., a metric over a quotient space defined by equivalence relation $d_S(A,B) = 0$. In particular, we show that d_S is a pseudometric when:

  (i) $S = \mathbb{P}^n$ and $\|\cdot\|$ is any entry-wise or operator norm;
  (ii) $S = \mathbb{W}^n$, the set of *doubly stochastic* matrices, $\|\cdot\|$ is an arbitrary entry-wise norm, and A, B are symmetric; a modification on d_S extends this result to both operator norms as well as arbitrary matrices (capturing, e.g., directed graphs); and
  (iii) $S = \mathbb{O}^n$, the set of orthogonal matrices, and $\|\cdot\|$ is the operator or entry-wise 2-norm.

  We also characterize the corresponding equivalence classes (see "Main results" section). Relaxations *(ii)* and *(iii)* are very important from a practical standpoint. For all matrix norms, computing (2) with $S = \mathbb{W}^n$ is tractable, as it is a convex optimization. For $S = \mathbb{O}^n$, (2) is non-convex but is still tractable, as it reduces to a spectral decomposition. This was known for the Frobenius norm (Umeyama 1988); we prove this is also the case for the operator 2-norm.
- We include node attributes in a natural way in the definition of d_S as both *soft* (i.e., penalties in the objective) or *hard* constraints in Eq. (2). Crucially, we do this *without affecting the pseudometric property and tractability*. This allows us to explore label or feature preserving permutations, that incorporate both (a) exogenous node attributes, such as, e.g., user age or gender in a social network, as well as (b) endogenous, structural features of each node, such as its degree or the number of triangles that pass through it. We numerically show that adding these constraints can speed up the computation of d_S.

From an experimental standpoint, we extensively compare our tractable metrics to several existing heuristic approximations. We also demonstrate the tractability of our metrics by parallelizing their execution using the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011), which we implement over a compute cluster using Apache Spark (Zaharia et al. 2010).

Related Work

Graph distance (or similarity) scores find applications in varied fields such as in image processing (Conte et al. 2004), chemistry (Allen 2002; Kvasnička et al. 1991), and social network analysis (Macindoe and Richards 2010; Koutra et al. 2013). Graph distances are easy to define when, contrary to our setting, the correspondence between graph nodes is known, i.e., graphs are *labeled* (Papadimitriou et al. 2010; Koutra et al. 2013; Soundarajan et al. 2014). Beyond the chemical distance, classic examples of distances between unlabeled graphs are the *edit distance* (Garey and Johnson 2002; Sanfeliu and Fu 1983) and the *maximum common subgraph distance* (Bunke and Shearer 1998; Bunke 1997), both of which also have versions for labeled graphs. Both are pseudometrics and are hard to compute, while existing heuristics (Riesen and Bunke 2009; Fankhauser et al. 2011) do not satisfy the triangle inequality. The *reaction distance* (Koca et al. 2012) is also a pseudometric, and is directly related to the chemical distance (Kvasnička et al. 1991) when edits are restricted to edge additions and deletions. Jain (Jain 2016) also considers an extension of the chemical distance, limited to the Frobenius norm, that incorporates edge attributes. However, it is not immediately clear how to relax the above pseudometrics (Jain 2016; Koca et al. 2012) to attain tractability, while keeping the pseudometric property.

A pseudometric can also be induced by embedding graphs in a metric space and measuring the distance between embeddings (Riesen et al. 2007; Ferrer et al. 2010; Riesen and Bunke 2010). Several works follow such an approach, mapping graphs, e.g., to spaces determined by their spectral decomposition (Zhu and Wilson 2005; Wilson and Zhu 2008; Elghawalby and Hancock 2008). In general, in contrast to our pseudometrics, such approaches are not as discriminative, as embeddings summarize graph structure. Continuous relaxations of graph isomorphism, both convex and non-convex (Lyzinski et al. 2016; Aflalo et al. 2015; Umeyama 1988), have found applications in a variety of contexts, including social networks (Koutra et al. 2013), computer vision (Schellewald et al. 2001), shape detection (Sebastian et al. 2004; He et al. 2006), and neuroscience (Vogelstein et al. 2011). Lyzinski et al. (Lyzinski et al. 2016) in particular show (both theoretically and experimentally) that a non-convex relaxation is advantageous over one of the relaxations we consider here (namely, d_S with $S = \mathbb{W}^n$, $\|\cdot\| = \|\cdot\|_F$) in recovering the optimal permutation P. They also incorporate features via a trace penalty as we do in "Incorporating metric embeddings" section (c.f. Eq. (17)). None of the above works however focus on the metric properties of the resulting relaxations, which several fail to satisfy (Vogelstein et al. 2011; Koutra et al. 2013; Sebastian et al. 2004; He et al. 2006; Lyzinski et al. 2016).

Metrics naturally arise in data mining tasks, including clustering (Xing et al. 2002; Hartigan 1975), Nearest Neighbour (NN) search (Clarkson 2006; 1999; Beygelzimer et al. 2006), and outlier detection (Angiulli and Pizzuti 2002). Some of these tasks become tractable, or admit formal guarantees, precisely when performed over a metric space. For example, finding the nearest neighbor (Clarkson 2006; 1999; Beygelzimer et al. 2006) or the diameter of a data-set (Indyk 1999) become polylogarithimic under metric assumptions; similarly, approximation algorithms for clustering (which is NP-hard) rely on metric assumptions, whose absence leads to a deterioration of known bounds (Ackermann et al. 2010). Our search for metrics is motivated by these considerations.

The present paper is an extended version of a paper by the same authors that appeared in the 2018 SIAM International Conference on Data Mining (Bento and Ioannidis 2018), which did not contain any proofs. In addition to the material included in the conference version, the present paper contains (a) proofs of all main theorems, establishing sufficient conditions under which a solution to (2) yields a pseudo-metric, (b) a polynomial-time spectral algorithm for computing (2) over the Stiefler manifold, (c) extensions of our metrics to graphs of unequal sizes, and (d) an extended experiment section.

Notation and preliminaries

We begin by introducing some terminology that we use throughout the paper. A summary of our notation can be found in Table 1.

Graphs We represent an undirected unweighted graph $G(V,E)$ with node set $V = [n] \equiv \{1,\dots,n\}$ and edge set $E \subseteq [n] \times [n]$ by its *adjacency matrix*, i.e. $A = [a_{i,j}]_{i,j\in[n]} \in \{0,1\}^{n\times n}$ such that $a_{ij} = a_{ji} = 1$ if and only if $(i,j) \in E$. In particular, A is symmetric, i.e. $A = A^\top$. We denote the set of all real, symmetric matrices by $\mathbb{S}^n$. *Directed* unweighted graphs are

Table 1 Notation Summary

$[n]$	Set $\{1,\dots,n\}$
$\mathbb{R}^{n\times n}$	The set of real $n \times n$ matrices.
$\mathbb{S}^n$	The set of real, symmetric matrices.
I	The identity matrix of size $n \times n$.
$\mathbf{1}$	The n-dimensional vector whose entries are all equal to 1.
$\sigma_{\max}(\cdot)$	Largest singular value of a matrix.
$\mathrm{tr}(\cdot)$	The trace of a matrix.
$\mathtt{conv}(\cdot)$	The convex hull of a set.
$G(V,E)$	Graph with vertex set V and edge set E.
A, B	Matrices $[a_{i,j}]_{i,j\in[n]}$, $[b_{i,j}]_{i,j\in[n]}$.
$\Vert\cdot\Vert_p$	Operator or entry-wise p-norm.
$\Vert\cdot\Vert_F$	Frobenius norm.
$\mathbb{P}^n$	Set of permutation matrices of size $n \times n$, c.f. (4)
$\mathbb{W}^n$	Set of doubly stochastic matrices (a.k.a. the Birkhoff polytope) of size $n \times n$, c.f. (5)
$\mathbb{O}^n$	Set of orthofonal matrices (a.k.a. the Stiefel manifold) of size $n \times n$, c.f. (6)
$\Omega, \tilde{\Omega}$	Sets over which a metric is defined.
$d(x,y)$	A metric over space Ω.
$\bar{d}(x,y)$	The symmetric extension of $d(x,y)$.
(Ω, d)	A metric space.
G_A, G_B	Graphs with adjacency matrices A, B.
P, W, O	$n \times n$ matrices.
S	A closed and bounded subset of $\mathbb{R}^{n\times n}$.
$d_S(A,B)$	A class of distance scores defined by minimization (12) over set S.
$d_{\mathbb{P}^n}$	Pseudometric d_S, where S is the set of permutation matrices.
$d_{\mathbb{W}^n}$	Pseudometric d_S, where S is the set of doubly stochastic matrices.
$d_{\mathbb{O}^n}$	Pseudometric d_S, where S is the set of orthogonal matrices.
$\Psi^n_{\tilde{\Omega}}$	Set of all embeddings from $[n] \to \tilde{\Omega}$, where $(\tilde{\Omega}, \tilde{d})$ is a metric space.
ψ_A, ψ_B	Embeddings in $\Psi^n_{\tilde{\Omega}}$ of nodes in graphs G_A and G_B, respectively.
D_{ψ_A,ψ_B}	$n \times n$ matrix of all pairwise distances between images of nodes in G_A and G_B, under embeddings ψ_A and ψ_B.

represented by (possibly non-symmetric) binary matrices $A \in \{0,1\}^{n\times n}$, and *weighted* graphs by real matrices $A \in \mathbb{R}^{n\times n}$.

Matrix Norms. Given a matrix $A = [a_{ij}]_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ and a $p \in \mathbb{N}_+ \cup \{\infty\}$, its *induced* or *operator p-norm* is defined in terms of the vector p-norm through $\|A\|_p = \sup_{x\in\mathbb{R}^n:\|x\|_p=1} \|Ax\|_p$, while its *entry-wise p-norm* is given by $\|A\|_p = (\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^p)^{1/p}$, for $p \in \mathbb{N}_+$, and $\|A\|_\infty = \max_{i,j} |a_{i,j}|$. We denote the entry-wise 2-norm (i.e., the *Frobenius* norm) as $\|\cdot\|_F$. Recall that all matrix norms on $\mathbb{R}^{n\times n}$ are *equivalent*: given two norms $\|\cdot\|, \|\cdot\|'$, there exist constants $c_1, c_2 > 0$ such that

$$c_1\|A\|' \leq \|A\| \leq c_2\|A\|' \quad \text{for all } A \in \mathbb{R}^{n\times n}. \tag{3}$$

Permutation, Doubly Stochastic, and Orthogonal Matrices. We denote the set of *permutation* matrices as

$$\mathbb{P}^n \equiv \{P \in \{0,1\}^{n\times n} : P\mathbf{1} = \mathbf{1}, P^\top\mathbf{1} = \mathbf{1}\}, \tag{4}$$

the set of *doubly-stochastic* matrices (i.e., the *Birkhoff polytope*) as

$$\mathbb{W}^n \equiv \{W \in [0,1]^{n\times n} : W\mathbf{1} = \mathbf{1}, W^\top\mathbf{1} = \mathbf{1}\}, \tag{5}$$

and the set of *orthogonal matrices* (i.e., the *Stiefel manifold*) as

$$\mathbb{O}^n \equiv \{U \in \mathbb{R}^{n\times n} : UU^\top = U^\top U = I\}. \tag{6}$$

Note that

$$\mathbb{P}^n = \mathbb{W}^n \cap \mathbb{O}^n, \tag{7}$$

i.e., permutation matrices are precisely the matrices that are both doubly stochastic and orthogonal. Moreover, the Birkoff-von Neumann Theorem (Birkhoff 1946) states that

$$\begin{aligned}\mathbb{W}^n &= \mathrm{conv}(\mathbb{P}^n)\\ &= \left\{W \in [0,1]^{n\times n} : W = \textstyle\sum_{P\in\mathbb{P}^n} \theta_P P, \text{ for some } \theta \in \mathbb{R}_+^{|\mathbb{P}^n|} \text{ s.t. } \theta^\top\mathbf{1} = 1\right\},\end{aligned} \tag{8}$$

i.e., the Birkoff polytope is the convex hull of $\mathbb{P}^n$. Hence, every doubly stochastic matrix can be written as a convex combination of permutation matrices.

Metrics Given a set Ω, a function $d : \Omega \times \Omega \to \mathbb{R}$ is called a *metric*, and the pair (Ω, d) is called a *metric space*, if for all $x, y, z \in \Omega$:

$$d(x,y) \geq 0 \qquad \text{(non-negativity)} \tag{9a}$$

$$d(x,y) = 0 \text{ iff} x = y \qquad \text{(pos. definiteness)} \tag{9b}$$

$$d(x,y) = d(y,x) \qquad \text{(symmetry)} \tag{9c}$$

$$d(x,y) \leq d(x,z) + d(z,y) \qquad \text{(triangle inequality)} \tag{9d}$$

A function d is called a *pseudometric* if it satisfies (9a), (9c), and (9d), but the positive definiteness property (9b), also known as the *identity of indiscernibles*, is replaced by the (weaker) property:

$$d(x,x) = 0 \text{ for all } x \in \Omega. \tag{9e}$$

If d is a pseudometric, then $d(x,y) = 0$ defines an equivalence relation $x \sim_d y$ over Ω. A pseudometric is then a metric over $\Omega/\sim_d$, the quotient space of $\sim_d$.

A function d that satisfies (9a), (9b), and (9d) *but not* the symmetry property (9c) is called a *quasimetric*. If d is a quasimetric, then its *symmetric extension* $\bar{d} : \Omega \times \Omega \to \mathbb{R}$, defined as

$$\bar{d}(x, y) = d(x, y) + d(y, x), \tag{10}$$

is a metric over Ω.

Graph Isomorphism, Chemical Distance, and CKS Distance. Let $A, B \in \mathbb{R}^{n\times n}$ be the adjacency matrices of two graphs G_A and G_B. Then, G_A and G_B are *isomorphic* if and only if there exists $P \in \mathbb{P}^n$ s.t.

$$P^\top AP = B \qquad \text{or, equivalently,} \qquad AP = PB. \tag{11}$$

The *chemical distance*, given by (1), extends the second relationship in (11) to capture distances between graphs. More generally, let $\| \cdot \|$ be a matrix norm in $\mathbb{R}^{n\times n}$. For some $\Omega \subseteq \mathbb{R}^{n\times n}$, define $d_S : \Omega \times \Omega \to \mathbb{R}_+$ as:

$$d_S(A, B) = \min_{P\in S} \|AP - PB\|, \tag{12}$$

where $S \subset \mathbb{R}^{n\times n}$ is a closed and bounded set, so that the infimum is indeed attained.

Note that d_S is the chemical distance (1) when $\Omega = \mathbb{R}^{n\times n}$, $S = \mathbb{P}^n$ and $\| \cdot \| = \| \cdot \|_F$. The Chartrant-Kibiki-Shultz (CKS) distance (Chartrand et al. 1998) can also be defined in terms of (12), with matrices A, B containing pairwise path distances between any two nodes; equivalently, CKS is the chemical distance of two weighted complete graphs with path distances as edge weights.

The Weisfeiler-Lehman (WL) Algorithm. The WL algorithm (Weisfeiler and Lehman 1968) is a heuristic for solving the graph isomorphism problem. We use this algorithm to (a) describe the quotient space over which (12) is a metric when $S = \mathbb{W}^n$ (see "Main results" section), and (b) to generate node embeddings in our experiments (see "Experiments" section).

To gain some intuition on the algorithm, note that two isomorphic graphs must have the same degree distribution. More broadly, the distributions of k-hop neighborhoods in the two graphs must also be identical. Building on this observation, to test if two undirected, unweighted graphs are isomorphic, WL *colors* the nodes of a graph $G(V, E)$ iteratively. At iteration 0, each node $v \in V$ receives the same *color* $c^0(v) := 1$. Colors at iteration $k + 1 \in \mathbb{N}$ are defined recursively via

$$c^{k+1}(v) := \mathsf{hash}\left(\mathsf{sort}\left(\mathsf{clist}_v^k\right)\right) \tag{13}$$

where hash is a perfect hash function, and

$$\mathsf{clist}_v^k = [\, c^k(u) : (u, v) \in E)] \tag{14}$$

is a list containing the colors of all of v's neighbors at iteration k.

Intuitively, two nodes in V share the same color after k iterations if their k-hop neighborhoods are isomorphic. WL terminates when the partition of V induced by colors is stable from one iteration to the next. This coloring extends to weighted directed graphs by appending weights and directions to colors in clist_v^k.

After coloring two graphs G_A, G_B, WL declares a non-isomorphism if their color distributions differ. If not, then they *may* be isomorphic and WL gives a set of *constraints* on candidate isomorphisms: a permutation P under which $AP = PB$ must map nodes in G_A to nodes in G_B of the same color.

Main results

Motivated by the chemical and CKS distances, we establish general conditions on S and $\|\cdot\|$ under which d_S is a metric over Ω, for arbitrary weighted graphs. For concreteness, we focus here on distances between graphs of equal size. Extensions to graphs of unequal size are described in "Graphs of different sizes" section.

A family of graph metrics

Optimization over Permutation Matrices. Our first result establishes that $d_{\mathbb{P}^n}$ is a pseudometric over *all* weighted graphs when $\|\cdot\|$ is an *arbitrary* entry-wise or operator norm.

Theorem 1 *If $S = \mathbb{P}^n$ and $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then d_S given by (12) is a pseudometric over $\Omega = \mathbb{R}^{n\times n}$.*

Hence, $d_{\mathbb{P}^n}$ is a pseudometric under any entry-wise or operator norm over arbitrary directed, weighted graphs.

Optimization over the Birkhoff Polytope. Our second result states that the pseudometric property extends to the *relaxed* version of the chemical distance, in which permutations are replaced by doubly stochastic matrices.

Theorem 2 *If $S = \mathbb{W}^n$ and $\|\cdot\|$ is an arbitrary entry-wise norm, then d_S given by (12) is a pseudometric over $\Omega = \mathbb{S}^{n\times n}$. If $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then its symmetric extension $\bar{d}_S(A,B) = d_S(A,B) + d_S(B,A)$ is a pseudometric over $\Omega = \mathbb{R}^{n\times n}$.*

Hence, if $S = \mathbb{W}^n$ and $\|\cdot\|$ is an arbitrary entry-wise norm, then (12) defines a pseudometric over *undirected* graphs. The symmetry property (9c) breaks if $\|\cdot\|$ is an operator norm or graphs are directed. In both of these two cases d_S is a quasimetric over the quotient space $\Omega/\sim_d$, and symmetry is attained via the symmetric extension $\bar{d}_S$.

Theorem 2 has significant practical implications. In contrast to $d_{\mathbb{P}^n}$ and its extensions implied by Theorem 1, computing $d_{\mathbb{W}^n}$ under any operator or entry-wise norm is tractable, in the sense that involves minimizing a convex function subject to linear constraints (Boyd and Vandenberghe 2004; Nesterov and Nemirovskii 1994; Bertsekas 1997).

Optimization over the Stiefler Manifold. A more limited result applies to the case when S is the Stiefel manifold $\mathbb{O}^n$:

Theorem 3 *If $S = \mathbb{O}^n$ and $\|\cdot\|$ is either the operator (i.e., spectral) or the entry-wise (i.e., Frobenius) 2-norm, then d_S given by (12) is a pseudometric over $\Omega = \mathbb{R}^{n\times n}$.*

Though (12) is not a convex problem when $S = \mathbb{O}^n$, it is also tractable. Umeyama (Umeyama 1988) shows that the optimization can be solved exactly when $\|\cdot\| = \|\cdot\|_F$

and $\Omega = \mathbb{S}^n$ (i.e., for undirected graphs) by performing a spectral decomposition on A and B. We extend this result, showing that the same procedure also applies when $\|\cdot\|$ is the operator 2-norm (see Thm. 7 in "Metric computation over the Stiefler manifold" section). In the general case of directed graphs, (12) is a classic example of a problem that can be solved through optimization on manifolds (Absil et al. 2009).

Equivalence Classes. Observe that the equivalence of matrix norms, as stated by Eq. (3), implies that if $d_S(A,B) = 0$ for one matrix norm $\|\cdot\|$ in (12), it will be so for all. As a result, pseudometrics d_S defined through (12) for a given S have the same quotient space $\Omega/\sim_{d_S}$, irrespectively of norm $\|\cdot\|$. We therefore turn our attention to characterizing this quotient space in the three cases when S is the set of permutation, doubly stochastic, and orthononal matrices.

When $S = \mathbb{P}^n$, $\Omega/\sim_{d_{\mathbb{P}^n}}$ is the quotient space defined by graph isomorphism: any two adjacency matrices $A, B \in \mathbb{R}^{n\times n}$ satisfy $d_{\mathbb{P}^n}(A,B) = 0$ if and only if their (possibly weighted) graphs are isomorphic.

When $S = \mathbb{W}^n$, the quotient space $\Omega/\sim_{d_{\mathbb{W}^n}}$ has a connection to the Weisfeiler-Lehman (WL) algorithm (Weisfeiler and Lehman 1968) described in "The Weisfeiler-Lehman (WL) Algorithm" section: Ramana et al. (Ramana et al. 1994) show that $d_{\mathbb{W}^n}(A,B) = 0$ if and only if G_A and G_B receive identical colors by the WL algorithm (see also (Tinhofer 1986) for another characterization of this quotient space). This equivalence relation is sometimes called called fractional linear isomorphism (Ramana et al. 1994).

Finally, if $S = \mathbb{O}^n$ and $\Omega = \mathbb{S}^n$, i.e., graphs are undirected, then $\Omega/\sim_{d_{\mathbb{O}^n}}$ is determined by *co-spectrality*: $d_{\mathbb{O}^n}(A,B) = 0$ if and only if A, B have the same spectrum. When $\Omega = \mathbb{R}^{n\times n}$, $d_{\mathbb{O}^n}(A,B) = 0$ implies that A, B are co-spectral, but co-spectral matrices A, B do not necessarily satisfy $d_{\mathbb{O}^n}(A,B) = 0$. Put differently, the quotient space $\Omega/\sim_{d_{\mathbb{O}^n}}$ in this case is a refinement of the quotient space of co-spectrality.

Incorporating metric embeddings

We have seen that the chemical distance $d_{\mathbb{P}^n}$ can be *relaxed* to $d_{\mathbb{W}^n}$ or $d_{\mathbb{O}^n}$, gaining tractability while still maintaining the metric property. In practice, nodes in a graph often contain additional attributes that one might wish to leverage when computing distances. In this section, we show that such attributes can be seamlessly incorporated in d_S either as soft or hard constraints, *without violating the metric property*.

Metric Embeddings. Given a graph G_A of size n, a *metric embedding* of G_A is a mapping $\psi_A : [n] \to \tilde{\Omega}$ from the nodes of the graph to a metric space $(\tilde{\Omega}, \tilde{d})$. That is, ψ_A maps nodes of the graph to $\tilde{\Omega}$, where $\tilde{\Omega}$ is endowed with a metric $\tilde{d}$. We refer to a graph endowed with an embedding ψ_A as an *embedded graph*, and denote this by (A, ψ_A), where $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix of G_A. We list two examples:

Example 1: Node Attributes. Consider an embedding of a graph to $(\mathbb{R}^k, \|\cdot\|_2)$ in which every node $v \in V$ is mapped to a k-dimensional vector describing "local" attributes. These can be *exogenous*: e.g., features extracted from a user's profile (age, binarized gender, etc.) in a social network. Alternatively, attributes may be *endogenous* or *structural*, extracted from the adjacency matrix A, e.g., the node's degree, the size of its k-hop neighborhood, its page-rank, etc.

Example 2: Node Colors. Let $\tilde{\Omega}$ be an arbitrary finite set endowed with the Kronecker delta as a metric, that is, for $s, s' \in \tilde{\Omega}$,

$$\tilde{d}(s,s') = \begin{cases} 0, & \text{if } s = s' \\ \infty, & \text{o.w.} \end{cases} \tag{15}$$

Given a graph G_A, a mapping $\psi_A : [n] \to \tilde{\Omega}$ is then a metric embedding. The values of $\tilde{\Omega}$ are invariably called *colors* or *labels*, and a graph embedded in $\tilde{\Omega}$ is a *colored* or *labeled* graph. Colors can again be *exogenous* or *structural*: e.g., if the graph represents an organic molecule, colors can correspond to atoms, while structural colors can be, e.g., the output of the WL algorithm (see "The Weisfeiler-Lehman (WL) Algorithm" section) after k iterations.

As discussed below, node attributes translate to *soft* constraints in metric (12), while node colors correspond to *hard* constraints. The unified view through embeddings allows us to establish metric properties for both simultaneously (c.f. Theorems 4 and 5) .

Embedding Distance. Consider two embedded graphs (A, ψ_A), (B, ψ_B) of size n that are embedded *in the same metric space* $(\tilde{\Omega}, \tilde{d})$. For $u \in [n]$, a node in the first graph, and $v \in [n]$, a node in the second graph, the embedded distance between the two nodes is given by $\tilde{d}(\psi_A(u), \psi_B(v))$. Let $D_{\psi_A,\psi_B} = [\tilde{d}(\psi_A(u), \psi_B(v))]_{u\in V, v\in V} \in \mathbb{R}_+^{n\times n}$ be the corresponding matrix of embedded distances. After mapping nodes to the same metric space, it is natural to seek $P \in \mathbb{P}^n$ that preserve the *embedding distance*. This amounts to finding a $P \in \mathbb{P}^n$ that minimizes:

$$\text{tr}\left(P^\top D_{\psi_A,\psi_B}\right) = \textstyle\sum_{u,v\in[n]} P_{u,v} \tilde{d}(\psi_A(u), \psi_B(v)). \tag{16}$$

Note that, in the case of colored graphs and the Kronecker delta distance, minimizing (16) finds a $P \in \mathbb{P}^n$ that maps nodes in A to nodes in B of equal color. It is not hard to verify that $\min_{P\in\mathbb{P}^n} \text{tr}\left(P^\top D_{\psi_A,\psi_B}\right)$ induces a metric between graphs embedded in $(\tilde{\Omega}, \tilde{d})$. In fact, this follows from the more general theorem we prove below (Theorem. 4) for $A = B = 0$, i.e., for distances between embedded graphs with no edges.

Despite the combinatorial nature of $\mathbb{P}^n$, the problem of minimizing (16) over $\mathbb{P}^n$ is a maximum weighted matching problem, which can be solved through, e.g., the Hungarian algorithm (Kuhn 1955), in $O(n^3)$ time (Jonker and Volgenant 1987). We note that this metric is not as expressive as (12): depending on the definition of the embeddings ψ_A, ψ_B, attributes may only capture "local" similarities between nodes, as opposed to the "global" view of a mapping attained by (12).

A Unified, Tractable Metric. Motivated by the above considerations, we focus on unifying the "global" metric (12) with the "local" metrics induced by arbitrary graph embeddings. Given a metric space $(\tilde{\Omega}, \tilde{d})$, let $\Psi^n_{\tilde{\Omega}} = \{\psi : [n] \to \tilde{\Omega}\}$ be the set of all mappings from $[n]$ to $\tilde{\Omega}$. Then, given two embedded graphs $(A, \psi_A), (B, \psi_B) \in \mathbb{R}^{n\times n} \times \Psi^n_{\tilde{\Omega}}$, we define:

$$d_S\left((A, \psi_A), (B, \psi_B)\right) = \min_{P\in S} \left[\|AP - PB\| + \text{tr}(P^\top D_{\psi_A,\psi_B})\right] \tag{17}$$

for some compact set $S \subset \mathbb{R}^{n\times n}$ and matrix norm $\|\cdot\|$. Our next result states that incorporating this linear term does not affect the pseudometric property of d_S.

Theorem 4 *If $S = \mathbb{P}^n$ and $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then d_S given by (17) is a pseudometric over the set of embedded graphs $\Omega = \mathbb{R}^{n\times n} \times \Psi^n_{\tilde{\Omega}}$.*

We stress here that this result is non-obvious: it is not true that adding *any* linear term to d_S leads to a quantity that satisfies the triangle inequality. It is precisely because D_{ψ_A,ψ_B} contains pairwise distances that Theorem 4 holds. We can similarly extend Theorem 2:

Theorem 5 *If $S = \mathbb{W}^n$ and $\|\cdot\|$ is an arbitrary entry-wise norm, then d_S given by (17) is a pseudometric over $\Omega = \mathbb{S}^n \times \Psi^n_{\tilde{\Omega}}$, the set of symmetric graphs embedded in $(\tilde{\Omega}, \tilde{d})$. Moreover, if $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then the symmetric extension $\bar{d}_S$ of (17) is a pseudometric over $\Omega = \mathbb{R}^{n\times n} \times \Psi^n_{\tilde{\Omega}}$.*

Adding the linear term (16) in d_S has significant practical advantages. Beyond expressing exogenous attributes, a linear term involving colors, combined with a Kronecker distance, translates into *hard* constraints: any permutation attaining a finite objective value *must* map nodes in one graph to nodes of the same color. Theorem 5 therefore implies that such constraints can thus be added to the optimization problem, while maintaining the metric property. In practice, as the number of variables in optimization problem (12) is n^2, incorporating such hard constraints can significantly reduce the problem's computation time; we illustrate this in "Experiments" section. Note that adding (16) to $d_{\mathbb{O}^n}$ does *not* preserve the metric property.

Proofs of Main results

Proof of Theorems 1–3

We define several properties that play a crucial role in our proofs.

Definition 1 *We say that a set $S \subseteq \mathbb{R}^{n\times n}$ is* closed under multiplication *if $P, P' \in S$ implies that $P \cdot P' \in S$.*

Definition 2 *We say that $S \subseteq \mathbb{R}^{n\times n}$ is* closed under transposition *if $P \in S$ implies that $P^\top \in S$, and* closed under inversion *if $P \in S$ implies that $P^{-1} \in S$.*

Definition 3 *Given a matrix norm $\|\cdot\|$, we say that set $S \subseteq \mathbb{R}^{n\times n}$ is* contractive *w.r.t. $\|\cdot\|$ if $\|AP\| \leq \|A\|$ and $\|PA\| \leq \|A\|$, for all $P \in S$ and $A \in \mathbb{R}^{n\times n}$. Put differently, S is contractive if and only if every linear transform $P \in S$ is a contraction w.r.t. $\|\cdot\|$.*

The proofs of Theorems 1–3 rely on several common lemmas. The first three establish conditions under which (12) satisfies the triangle inequality (9d), symmetry (9c), and weak property (9e), respectively:

Lemma 1 *Given a matrix norm $\|\cdot\|$, suppose that set $S \subseteq \mathbb{R}^{n\times n}$ is (a) contractive w.r.t. $\|\cdot\|$, and (b) closed under multiplication. Then, for any $A, B, C \in \mathbb{R}^{n\times n}$, d_S given by (12) satisfies $d_S(A, C) \leq d_S(A, B) + d_S(B, C)$.*

Proof Consider $P' \in \arg\min_{P\in S} \|AP - PB\|$, and $P'' \in \arg\min_{P\in S} \|BP - PC\|$. Then, from closure under multiplication, $P'P'' \in S$. Hence,

$$\begin{aligned} d_S(A,C) &\leq \|AP'P'' - P'P''C\| = \|AP'P'' - P'BP'' + P'BP'' - P'P''C\| \\ &\leq \|AP'P'' - P'BP''\| + \|P'BP'' - P'P''C\| \\ &= \|(AP' - P'B)P''\| + \|P'(BP'' - P''C)\| \\ &\leq \|AP' - P'B\| + \|BP'' - P''C\| = d_S(A,B) + d_S(B,C), \end{aligned}$$

where the last inequality follows from the fact that P', P'' are contractions. □

Lemma 2 *Given a matrix norm $\|\cdot\|$, suppose that $S \subset \mathbb{R}^{n\times n}$ is (a) contractive w.r.t. $\|\cdot\|$, and (b) closed under inversion. Then, for all $A, B \in \mathbb{R}^{n\times n}$, $d_S(A,B) = d_S(B,A)$.*

Proof Observe that property (b) implies that, for all $P \in S$, P is invertible and $P^{-1} \in S$. Hence,

$$\|AP - PB\| = \|PP^{-1}AP - PBP^{-1}P\| = \|P(P^{-1}A - BP^{-1})P\| \leq \|BP^{-1} - P^{-1}A\|,$$

as P is a contraction w.r.t $\|\cdot\|$. We can similarly show that $\|BP^{-1} - P^{-1}A\| \leq \|AP - PB\|$, hence $\|AP - PB\| = \|BP^{-1} - P^{-1}A\|$. As S is closed under inversion,

$$\min_{P\in S} f(P) = \min_{P:P^{-1}\in S} f(P),$$

for every $f : S \to \mathbb{R}$. Hence

$$\begin{aligned} d_S(A,B) &= \min_{P\in S} \|BP^{-1} - P^{-1}A\| = \min_{P:P^{-1}\in S} \|BP^{-1} - P^{-1}A\| \\ &= \min_{P\in S} \|BP - PA\| = d_S(B,A). \end{aligned}$$

□

Lemma 3 *If $I \in S$, then $d_S(A,A) = 0$ for all $A \in \mathbb{R}^{n\times n}$.*

Proof Indeed, if $I \in S$, then $0 \leq d_S(A,A) \leq \|AI - IA\| = 0$. □

Both the set of permutation matrices $\mathbb{P}^n$ *and* the Stiefel manifold $\mathbb{O}^n$ are *groups* w.r.t. matrix multiplication: they are closed under multiplication, contain the identity I, and are closed under inversion. Hence, if they are also contractive w.r.t. a matrix norm $\|\cdot\|$, $d_{\mathbb{P}^n}$ and $d_{\mathbb{O}^n}$ defined in terms of this norm satisfy all assumptions of Lemmas 1–3. We therefore turn our attention to this property.

Lemma 4 *Let $\|\cdot\|$ be any operator or entry-wise norm. Then, $S = \mathbb{P}^n$ is contractive w.r.t. $\|\cdot\|$.*

Proof Observe first that all vector p-norms are invariant to permutations of a vector's entries; hence, for any vector $x \in \mathbb{R}^d$, if $P \in \mathbb{P}^n$, $\|Px\|_p = \|x\|_p$. Hence, if $\|\cdot\|$ is an operator p-norm, $\|P\| = 1$, for all $P \in S$. Every operator norm is *submultiplicative*; as a result $\|PA\| \leq \|P\|\|A\| = \|A\|$ and, similarly, $\|AP\| \leq \|A\|$, so the lemma follows for operator norms. On the other hand, if $\|\cdot\|$ is an entry-wise norm, then $\|A\|$ is invariant to permutations of either A's rows or columns. Matrices PA and AP precisely amount to such permutations, so $\|PA\| = \|AP\| = \|A\|$ and the lemma follows also for entrywise norms. □

Hence, Theorem 1 follows as a direct corollary of Lemmas 1–4. Indeed, $d_{\mathbb{P}^n}$ is non-negative, symmetric by Lemmas 2 and 4, satifies the triangle inequality by Lemmas 1 and 4, as well as property (9e) by Lemma 3; hence $d_{\mathbb{P}^n}$ is a pseudometric over $\mathbb{R}^{n\times n}$. Our next lemma shows that the Stiefel manifold $\mathbb{O}^n$ is contractive for 2-norms:

Lemma 5 *Let $\|\cdot\|$ be the operator (i.e., spectral) or the entry-wise (i.e., Frobenius) 2-norm. Then, $S = \mathbb{O}^n$ is contractive w.r.t. $\|\cdot\|$.*

Proof Any $U \in \mathbb{O}^n$ is an orthogonal matrix; hence, $\|U\|_2 = \|U\|_F = 1$. Both norms are submultiplicative: the first as an operator norm, the second from the Cauchy-Schwartz inequality. Hence, for $U \in \mathbb{O}^n$, we have $\|UA\| \leq \|U\|\|A\| = \|A\|$.

Note that an alternative proof can be obtained by the fact that both norms are unitarily invariant (see Lemma 12). □

Theorem 3 therefore follows from Lemmas 1–3 and Lemma 5, along with the fact that $\mathbb{O}^n$ is a group. Note that $\mathbb{O}^n$ is *not* contractive w.r.t. other norms, e.g., $\|\cdot\|_1$ or $\|\cdot\|_\infty$.

To prove Theorem 2, we first show that Lemma 4 along with the Birkoff-von Neumann theorem imply that $\mathbb{W}^n$ is also contractive:

Lemma 6 *Let $\|\cdot\|$ be any operator or entry-wise norm. Then, $\mathbb{W}^n$ is contractive w.r.t. $\|\cdot\|$.*

Proof By the Birkoff-con Neumann theorem (Birkhoff 1946),

$$\mathbb{W}^n = \mathtt{conv}(\mathbb{P}^n).$$

Hence, for any $W \in \mathbb{W}^n$ there exist $P_i \in \mathbb{P}^n$, $\theta_i > 0$, $i = 1, \ldots, k$, such that $W = \sum_{i=1}^k \theta_i P_i$ and $\sum_{i=1}^k \theta_i = 1$. Both operator and entrywise p-norms are convex functions; hence, for any $A \in \mathbb{R}^{n\times N}$:

$$\begin{aligned} \|WA\| &= \left\|\sum_{i=1}^k \theta_i P_i A\right\| \leq \sum_{i=1}^k \theta_i \|P_i A\|, && \text{by Jensen's ineqality,} \\ &\leq \sum_{i=1}^k \theta_i \|A\|, && \text{by Lemma 4,} \\ &= \|A\| \end{aligned}$$

The statement $\|AW\| \leq \|A\|$ follows similarly. □

Unfortunately, the Birkhoff polytope $\mathbb{W}^n$ is *not* a group, as it is not closed under inversion. Nevertheless, it is closed under transposition; in establishing (partial) symmetry of $d_{\mathbb{W}^n}$, we leverage the following lemma:

Lemma 7 *Suppose that $\|\cdot\|$ is transpose-invariant, and $S \subseteq \mathbb{R}^{n\times n}$ is closed under transposition. Then, $d_S(A,B) = d_S(B,A)$ for all $A, B \in \mathbb{S}^n$.*

Proof By transpose invariance and the symmetry of A and B, we have that:

$$\|AP - PB\| = \|BP^\top - P^\top A\|.$$

Moreover, as S is closed under transposition, for every $f : S \to \mathbb{R}$,

$$\min_{P \in S} f(P) = \min_{P : P^\top \in S} f(P).$$

Hence, $d_S(A, B) = \min_{P \in S} \|BP^\top - P^\top A\| = \min_{P:P^\top \in S} \|BP^\top - P^\top A\| = d_S(B, A)$. □

The first part of Theorem 2 therefore follows from Lemmas 1, 3, 6, and 7: this is because $\mathbb{W}^n$ contains the identity I and is closed under both transposition and multiplication, while all entry-wise norms are transpose invariant.

To prove the second part, observe that operator norms are not transpose invariant. However, if $\|\cdot\|$ is an operator norm, or $\Omega = \mathbb{R}^{n\times n}$, then Lemma 6 and Lemma 1 imply that $d_{\mathbb{W}^n}$ satisfies non-negativity (9a) and the triangle inequality (9d), while Lemma 3 implies that it satisfies (9e). These properties are inherited by extension $\bar{d}_S$, given by (10), which also satisfies symmetry (9c), and the second part of Theorem 2 follows.

Proof of Theorems 4 and 5

We begin by establishing conditions under which d_S satisfies the triangle inequality (9d). We note that, in contrast to Lemma 1, we require the additional condition that $S \subseteq \mathbb{W}^n$, which is not satisfied by $\mathbb{O}^n$.

Lemma 8 *Given a norm $\|\cdot\|$, suppose that $S \subseteq \mathbb{R}^{n\times n}$ is (a) contractive w.r.t. $\|\cdot\|$, (b) closed under multiplication, and (c) is a subset of $\mathbb{W}^n$, i.e., contains only doubly stochastic matrices. Then, for any $(A, \psi_A), (B, \psi_B), (C, \psi_C)$ in $\mathbb{R}^{n\times n} \times \Psi_{\tilde{\Omega}}$,*

$$d_S((A, \psi_A), (C, \psi_B)) \leq d_S((A, \psi_A), (B, \psi_B)) + d_S((B, \psi_B), (C, \psi_C)).$$

Proof Consider

$$P' \in \arg\min_{P \in S} \left(\|AP - PB\| + \operatorname{tr}\left(P^\top D_{\psi_A, \psi_B}\right)\right),$$

and

$$P'' \in \arg\min_{P \in S} \left(\|BP - PC\| + \operatorname{tr}\left(P^\top D_{\psi_B, \psi_C}\right)\right).$$

Then, from closure under multiplication, $P'P'' \in S$. We have that

$$d_S((A, \psi_A), (C, \psi_C)) \leq \|AP'P'' - P'P''C\| + \operatorname{tr}\left[(P'P'')^\top D_{\psi_A \psi_C}\right]$$

As in the proof of Lemma 1, we can show that

$$\begin{aligned}
\|AP'P'' - P'P''C\| &= \|AP'P'' - P'BP'' + P'BP'' - P'P''C\| \\
&\leq \|AP'P'' - P'BP''\| + \|P'BP'' - P'P''C\| \\
&= \|(AP' - P'B)P''\| + \|P'(BP'' - P''C)\| \\
&\leq \|(AP' - P'B)\|\|P''\| + \|P'\|\|(BP'' - P''C)\| \\
&\leq \|AP' - P'B\| + \|BP'' - P''C\|
\end{aligned}$$

using the fact that both P' and P'' are contractions. On the other hand,

$$\operatorname{tr}\left[(P'P'')^\top D_{\psi_A \psi_C}\right] = \sum_{u,v \in [n]} \sum_{k \in [n]} \left(P'_{uk} P''_{kv} \tilde{d}(\psi_A(u), \psi_C(v)))\right)$$

$$\leq \sum_{u,v\in[n]}\sum_{k\in[n]} \left[P'_{uk}P''_{kv}\big(\tilde{d}(\psi_A(u),\psi_B(k)) + \tilde{d}(\psi_B(k),\psi_C(v))\big)\right]$$

(as $\tilde{d}$ is a metric, and P', P'' are non-negative)

$$= \sum_{u,k\in[n]} P'_{uk}\;\tilde{d}(\psi_A(u),\psi_B(k)) \sum_{v\in[n]} P''_{kv} + \sum_{k,v\in[n]} P''_{kv}\tilde{d}(\psi_B(k),\psi_C(v)) \sum_{u\in[n]} P'_{uk}$$

$$\leq \mathrm{tr}\left((P')^\top D_{\psi_A,\psi_B}\right) + \mathrm{tr}\left((P'')^\top D_{\psi_B,\psi_C}\right),$$

where the last inequality follows as both $P, P^\top$ are $\|\cdot\|_1$-norm bounded by 1 for every $P \in S$. □

The weak property (9e) is again satisfied provided the identity is included in S.

Lemma 9 *If* $I \in S$, *then* $d_S((A,\psi_A),(A,\psi_A)) = 0$ *for all* $A \in \mathbb{R}^{n\times n}$.

Proof Indeed, $0 \leq d_S((A,\psi_A,(A,\psi_A)) \leq \|AI - IA\| + \sum_{u\in[n]} \tilde{d}(\psi_A(u),\psi_A(u)) = 0$. □

To attain symmetry over $\Omega = \mathbb{R}^{n\times n} \times \Psi^n_{\tilde{\Omega}}$, we again rely on closure under inversion, as in Lemma 2; nonetheless, in contrast to Lemma 2, due to the linear term, we also need to assume the orthogonality of elements of S.

Lemma 10 *Given a norm* $\|\cdot\|$, *suppose that* S *(a) is contractive w.r.t.* $\|\cdot\|$, *(b) is closed under inversion, and (c) is a subset of* $\mathbb{O}^n$, *i.e., contains only orthogonal matrices. Then,* $d_S((A,\psi_A),(B,\psi_B)) = d_S((B,\psi_B),(A,\psi_A))$ *for all* $(A,\psi_A),(B,\psi_B) \in \mathbb{R}^{n\times n} \times \Psi_{\tilde{\Omega}}$.

Proof As in the proof of Lemma 2, we can show that contractiveness w.r.t. $\|\cdot\|$ along with closure under inversion imply that: $\|AP - PB\| = \|BP^{-1} - P^{-1}A\|$. As S is closed under inversion, $\min_{P\in S} f(P) = \min_{P:P^{-1}\in S} f(P)$ for all $f : S \to \mathbb{R}$, while orthogonality implies $P^{-1} = P^\top$ for all $P \in S$. Hence, $d_S((A,\psi_A),(B,\psi_B))$ equals

$$\begin{aligned}
\min_{P\in S}\left[\|AP - PB\| + \mathrm{tr}\left(P^\top D_{\psi_A,\psi_B}\right)\right] &= \min_{P\in S}\left[\|BP^{-1} - P^{-1}A\| + \mathrm{tr}\left(P^{-1}D_{\psi_A,\psi_B}\right)\right]\\
&= \min_{P\in S}\left[\|BP^{-1} - P^{-1}A\| + \mathrm{tr}\left(\left(P^{-1}\right)^\top D^\top_{\psi_A,\psi_B}\right)\right]\\
&= \min_{P\in S}\left[\|BP^{-1} - P^{-1}A\| + \mathrm{tr}\left(\left(P^{-1}\right)^\top D_{\psi_B,\psi_A}\right)\right]\\
&= \min_{P:P^{-1}\in S}\left[\|BP^{-1} - P^{-1}A\| + \mathrm{tr}\left(\left(P^{-1}\right)^\top D_{\psi_B,\psi_A}\right)\right]\\
&= d_S((B,\psi_B),(A,\psi_A)).
\end{aligned}$$

□

Theorem 4 therefore follows from the above lemmas, as $S = \mathbb{P}^n$ contains I, it is closed under multiplication and inversion, is a subset of $\mathbb{W}^n \cap \mathbb{O}^n$ by (7), and is contractive w.r.t. all operator and entrywise norms. Theorem 5 also follows by using the following lemma, along with Lemmas 8 and 9.

Lemma 11 *Suppose that $\|\cdot\|$ is transpose invariant, and S is closed under transposition. Then, $d_S((A,\psi_A),(B,\psi_B)) = d_S((B,\psi_B),(A,\psi_A))$ for all $(A,\psi_A),(B,\psi_B) \in \mathbb{S}^n \times \Psi_{\tilde{\Omega}}$.*

Proof By the transpose invariance of $\|\cdot\|$ and the symmetry of A and B, we have that: $\|AP - PB\| = \|BP^\top - P^\top A\|$. Moreover, as S is closed under transposition, $\min_{P\in S} f(P) = \min_{P:P^\top \in S} f(P)$ for any $f: S \to \mathbb{R}$. Hence, $d_S((A,\psi_A),(B,\psi_B))$ equals

$$\begin{aligned}\min_{P\in S}\left[\|AP-PB\| + \operatorname{tr}\left(P^\top D_{\psi_A,\psi_B}\right)\right] &= \min_{P\in S}\left[\|BP^\top - P^\top A\| + \operatorname{tr}\left(PD^\top_{\psi_A,\psi_B}\right)\right]\\ &= \min_{P:P^\top\in S}\|BP^\top - P^\top A\| + \operatorname{tr}\left((P^\top)^\top D_{\psi_B,\psi_A}\right)\\ &= d_S((B,\psi_B),(A,\psi_A))\end{aligned}$$

□

Metric computation over the Stiefler manifold.

In this section, we describe how to compute the metric d_S in polynomial time when $S = \mathbb{O}^n$ and $\|\cdot\|$ is the Frobenius norm or the operator 2-norm. The algorithm for the Frobenius norm, and the proof of its correctness, is due to Umeyama (Umeyama 1988); we reprove it for completeness, along with its extension to the operator norm.

Both cases make use of the following lemma:

Lemma 12 *For any matrix $M \in \mathbb{R}^{n\times n}$ and any matrix $P \in \mathbb{O}^n$ we have that $\|PM\| = \|MP\| = \|M\|$, where $\|\cdot\|$ is either the Frobenius or operator 2-norm.*

Proof Recall that the operator 2-norm $\|\cdot\|_2$ is $\|M\|_2 = \sup_{x\neq 0}\|Mx\|_2/\|x\|_2 = \sqrt{\sigma_{\max}(M^\top M)} = \sqrt{\sigma_{\max}(MM^\top)} = \|M^\top\|_2$. where $\sigma_{\max}$ denotes the largest singular value. Hence, $\|PM\|_2 = \sup_{x\neq 0}\|PMx\|_2/\|x\|_2 = \sqrt{\sigma_{\max}(M^\top P^\top PM)} = \sqrt{\sigma_{\max}(M^\top M)} = \|M\|_2$. as $P^\top P = I$. Using the fact that $\|M\|_2 = \|M^\top\|_2$ for all $M \in \mathbb{R}^{n\times n}$, as well as that $PP^\top = I$, we can show that $\|MP\|_2 = \|P^\top M^\top\|_2 = \|M^\top\|_2 = \|M\|_2$.

The Frobenius norm is $\|M\|_F = \sqrt{\operatorname{tr}(M^\top M)} = \sqrt{\operatorname{tr}(MM^\top)} = \|M^\top\|_F$, hence $\|PM\|_F = \sqrt{\operatorname{tr}(M^\top P^\top PM)} = \sqrt{\operatorname{tr}(M^\top M)} = \|M\|_F$ and, as in the case of the operator norm, we can similarly show $\|MP\|_F = \|P^\top M^\top\|_F = \|M^\top\|_F = \|M\|_F$. □

In both norm cases, for $A, B \in \mathbb{S}^n$, we can compute d_S using a simple spectral decomposition, which dominates computations and can be performed in $O(n^3)$ time. Let $A = U\Sigma_A U^T$ and $B = V\Sigma_B V^T$ be the spectral decomposition of A and B. As A and B are real and symmetric, we can assume $U, V \in \mathbb{O}^n$. Recall that $U^{-1} = U^\top$ and $V^{-1} = V^\top$, while Σ_A and Σ_B are diagonal and contain the eigenvalues of A and B sorted in increasing order; this ordering matters for computations below.

The following theorem establishes that this decomposition readily yields the distance d_S, as well as the optimal orthogonal matrix P^*, when $\|\cdot\| = \|\cdot\|_F$:

Theorem 6 *(Umeyama 1988) $d_S(A,B) \triangleq \min_{P\in S}\|AP - PB\|_F = \|\Sigma_A - \Sigma_B\|_F$ and the minimum is attained by $P^* = UV^\top$.*

Proof The proof makes use of the following lemma by Hoffman and Wielandt (Hoffman and Wielandt 1953): □

Lemma 13 ((Hoffman and Wielandt 1953)) *If A and B are Hermitian matrices with eigenvalues $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \leq b_2 \leq \ldots \leq b_n$ then $\|A - B\|_F^2 \geq \sum_{i=1}^{n}(a_i - b_i)^2$.*

Note that if Σ_A and Σ_B are diagonal matrices with the ordered eigenvalues of A and B in the diagonal, then the conclusion of Lemma 13 can be written as $\|A - B\|_F \geq \|\Sigma_A - \Sigma_B\|_F$. For any $P \in \mathbb{O}^n$ and $\|\cdot\| = \|\cdot\|_F$ we have

$$\begin{aligned}\|AP - PB\| &= \|(A - PBP^{-1})P\| \overset{\text{Lemma 12}}{=} \|A - PBP^\top\| = \|U\Sigma_A U^\top - PV\Sigma_B V^\top P^\top\| \\ &= \|U(\Sigma_A - U^\top PV\Sigma_B V^\top P^\top U)U^\top\| \overset{\text{Lemma 12}}{=} \|\Sigma_A - U^\top PV\Sigma_B V^\top P^\top U\| \\ &= \|\Sigma_A - \Delta\Sigma_B\Delta^\top\|\end{aligned}$$

where we define $\Delta \equiv U^\top PV$. As a product of orthogonal matrices, $\Delta \in \mathbb{O}^n$. Notice that

$$\begin{aligned}\|\Sigma_A - \Delta\Sigma_B\Delta^\top\| &= \|\Sigma_A - \Delta\Sigma_A\Delta^\top + \Delta(\Sigma_B - \Sigma_A)\Delta^\top\| \leq \|\Sigma_A - \Delta\Sigma_A\Delta^\top\| + \|\Delta(\Sigma_B - \Sigma_A)\Delta^\top\| \\ &\overset{\text{Lemma 12}}{=} \|\Sigma_A - \Delta\Sigma_A\Delta^\top\| + \|\Sigma_B - \Sigma_A\|.\end{aligned}$$

Therefore, for any $P \in \mathbb{O}^n$, $\|\Sigma_A - \Sigma_B\| \leq d_S(A,B) \leq \|\Sigma_A - \Delta\Sigma_A\Delta^\top\| + \|\Sigma_B - \Sigma_A\|$, where the first inequality follows by Lemma 13 if we notice that $\|AP - PB\| = \|A - PBP^{-1}\|$ and that PBP^{-1} and B have the same spectrum for any P. If we choose $P = UV^\top$ then $\Delta = I$ and the result follows.

We can compute d_S when $S = \mathbb{O}^n$ and $\|\cdot\|$ is the operator norm in the exact same way.

Theorem 7 *Let $\|\cdot\| = \|\cdot\|_2$ be the operator 2-norm. Then, $d_S(A,B) \triangleq \min_{P\in S}\|AP - PB\|_2 = \|\Sigma_A - \Sigma_B\|_2$ and the minimum is attained by $P^* = UV^\top$.*

Proof The proof follows the same steps as the proof of Theorem 6, using Lemma 14 below instead of Lemma 13. □

Lemma 14 *If A and B are Hermitian matrices with eigenvalues $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \leq b_2 \leq \ldots \leq b_n$ then $\|A - B\|_2 \geq \max_i |a_i - b_i|$.*

Proof This is the second exercise following Corollary 6.3.4 in Horn and Johnson (Horn and Johnson 2012). We reprove this here for completeness. □

Let $\tilde{B} = -B$ have eigenvalues $\tilde{b}_1 \leq \tilde{b}_2 \leq \ldots \leq \tilde{b}_n$ and let $C = A + \tilde{B}$ have eigenvalues $c_1 \leq c_2 \leq \ldots \leq c_n$. We make use of the following lemma by Weyl (see Theorem 4.3.1 (Weyl), page 239, in (Horn and Johnson 2012)) to lower-bound c_n.

Lemma 15 (Weyl) *If X and Y are Hermitian with eigenvalues $x_1 \leq \ldots \leq x_n$ and $y_1 \leq \ldots \leq y_n$ and if $X + Y$ has eigenvalues $w_1 \leq \ldots \leq w_n$ then $x_{i-j+1} + y_j \leq w_i$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, i$.*

If we choose $X = \tilde{B}$, $Y = A$ and $i = n$ we get $a_j + \tilde{b}_{n+1-j} \leq c_n$ for all $j = 1, \ldots, n$. Since $\tilde{b}_{n+1-j} = -b_j$ we get that $a_j - b_j \leq c_n$, for any j. Similarly, by exchanging the role of A and B, we can lower bound the largest eigenvalue of $B - A$, say d_n, by $b_j - a_j$ for any j. Notice that, by definition of the operator norm and the fact that $A - B$ is Hermitian, $\|A - B\|_2 \geq |c_n|$ and $\|B - A\|_2 \geq |d_n|$. Since $\|B - A\|_2 = \|A - B\|_2$ we have that $\|A - B\|_2 \geq$

$\max\{|c_n|, |d_n|\} \geq \max\{c_n, d_n\} \geq \max\{a_j - b_j, b_j - a_j\} = |a_j - b_j|$ for all j. Taking the maximum over j we get that $\|A - B\|_2 \geq \max_j |a_j - b_j|$, and the lemma follows.

Note again that if Σ_A and Σ_B are diagonal matrices with the ordered eigenvalues of A and B in the diagonal, then the conclusion of Lemma 14 can be written as $\|A - B\|_2 \geq \|\Sigma_A - \Sigma_B\|_2$. The proof of Thm. 7 proceeds along the same steps as the above proof, using again the fact that, by Lemma 12, $\|M\|_2 = \|MP\|_2 = \|PM\|_2$ for any $P \in \mathbb{O}^n$ and any matrix M, along with Lemma 15.

Graphs of different sizes

For simplicity, we have described distances over graphs of equal sizes. There are several applications (Hu et al. 2016; Shen et al. 2015; Lyzinski et al. 2016; Pachauri et al. 2013) where by design we want to compare (and align the nodes of) equal-sized graphs. E.g., in computer vision, one might want to establish a correspondence among the nodes of two graphs, each representing a geometrical relation among m special points in two images of objects of the same type. When poses of objects do not differ significantly, the same number, m, of special points will be extracted from each image, and hence the graphs being compared will have the same size.

We can nevertheless extend our approach to graphs of different sizes. We can do so by extending two graphs, G_A and G_B, with dummy nodes such that the new graphs G'_A and G'_B have the same number of nodes. Many papers follow this approach, e.g. (Zaslavskiy et al. 2009b; 2009a; Narayanan et al. 2011; Zaslavskiy et al. 2010; Zhou and De la Torre 2012; Gold and Rangarajan 1996; Yan et al. 2015; Solé-Ribalta and Serratosa 2010; Yan et al. 2015). If G_A has n_A nodes and G_B has n_B nodes we can, for example, add n_B dummy nodes to G_A and n_A dummy nodes to G_A. Once we have G'_A and G'_B of equal size, we can use the methods we already described to compute a distance between G'_A and G'_B and return this distance as the distance between G_A and G_B.

Possible graph extensions differ in how the dummy nodes connect to existing graph nodes, how dummy nodes connect to themselves, and what kind of penalty we introduce for associating dummy nodes with existing graph nodes.

Method 1. One way of extending the graphs is to add dummy nodes and leave them isolated, i.e., with no edges to either existing nodes or other dummy nodes. Although this might work when both graphs are dense, it might lead to non desirable results when one of the graphs is sparse. For example, let G_A be 3 isolated nodes and G_B be the complete graph on 4 nodes minus the edges forming triangle $\{(1, 2), (2, 3), (3, 1)\}$. Let us assume that $S = \mathbb{P}^n$, such that, when we compute the distance between G_A and G_B, we produce an alignment between the graphs. One desirable outcome would be for G_A to be aligned with the three nodes in G_B that have no edges among them. This is basically solving the problem of finding a sparse subgraph inside a dense graph. However, computing $d_S(A', B')$, where A' and B' are the extended adjacency matrices, could equally well align G_A with the 3 dummy nodes of G'_B.

Method 2. Alternatively, one could add dummy nodes and connect each dummy node to all existing nodes and all other dummy nodes. This avoids the issue described for method 1. However, this creates a similar non-desirable situation: since the dummy nodes in each extended graph form a clique, we might align G_A, or G_B, with just dummy nodes, instead of producing an alignment between existing nodes in G_A and existing nodes in G_B.

Method 3. If both G_A and G_B are unweighted graphs, a method that avoids both issues above (aligning a sparse graph with isolated dummy nodes or aligning a dense graphs with cliques of dummy nodes) is to connect each dummy node to all existing nodes and all other dummy nodes with edges of weight 1/2. This method works because, when $S = \mathbb{P}^n$, it discourages alignments of edges between existing nodes in G_A to dummy-dummy edges or dummy-existing node edges in G_B, and vice versa.

Method 4. One can also discourage aligning existing nodes with dummy nodes by introducing a soft linear term as in (17), penalizing mappings between dummy and existing nodes.

Method 5. Finally, a method of ensuring that the graphs have equal size is repeating them, i.e., creating "super" graphs that consist of multiple replicas of the same graph as connected components, resulting in two graphs of size equal to the least common multiple (LCM) of the sizes of the two original graphs. This is most appropriate when a spectral approach is used, like the ones used to optimize over $\mathbb{O}^n$: this is because repetition, in effect, only changes the multiplicity of each value in the spectrum, which can be done (a) without affecting the spectrum structure, and (b) efficiently, once the LCM is computed.

Experiments

We experimentally study the properties of different graph distance measures, including metrics from our family, over several graph classes. Our main observation is that computing a heuristic estimate $\hat{P}$ of $P^* = \arg\min_{P\in\mathbb{P}^n} \|AP - PB\|$, and using $\hat{P}$ to estimate $d_{\mathbb{P}^n}(A, B)$ leads to violations of the metric property. In contrast, our proposed approach of computing $d_S(A, B)$ for some S for which d a metric, and for which its computation is tractable, yields significantly improved performance in tasks such as clustering graphs (see Fig. 1).

Experimental setup

Graphs We use *synthetic graphs* from six classes summarized in Table 3: Barabasi Albert with degree d (Bd), Erdos Renyi with probablity p (Ep), Power Law Tree (P), Regular with degree d (Rd), Small World (S), Watts Strogatz with degree d (Ws). In addition, we use a dataset of *small graphs*, comprising all 853 connected graphs of 7 nodes. Finally, we use a *collaboration graph* with 5242 nodes and 14496 edges representing author collaborations.

Algorithms We compare our metrics to several competitors outlined in Table 2. All receive only two unlabeled undirected simple graphs A and B and output a matching a matrix $\hat{P}$ either in $\mathbb{W}^n$ or in $\mathbb{P}^n$ estimating P^*. If $\hat{P} \in \mathbb{P}^n$, we compute $\|A\hat{P} - \hat{P}B\|_1$. If $\hat{P} \in \mathbb{W}^n$, then we compute both $\|A\hat{P} - \hat{P}B\|_1$ and $\|A\hat{P} - \hat{P}B\|_F$; all norms are entrywise. We also implement our two relaxations $d_{\mathbb{W}}$ and $d_{\mathbb{O}^n}$, for two different matrix norm combinations.

We briefly review here additional impementation details about the algorithms summarized in Table 2.

- **NetAlignBP**, **IsoRank**, **SparseIsoRank** and **NetAlignMR** are described by (Bayati et al. 2009). **Natalie** is described in (El-Kebir et al. 2015). All five algorithms output $P \in \mathbb{P}^n$.
- The algorithm in (Lyzinski et al. 2016) outputs one $P \in \mathbb{P}^n$ and one $P' \in \mathbb{W}^n$. We use $P \in \mathbb{P}^n$ to compute $\|AP - PB\|_1$ and call this **InnerPerm**. We use $P' \in \mathbb{W}^n$ to compute $\|AP' - P'B\|_1$ and $\|AP' - P'B\|_2$ and call these algorithms **InnerDSL1** and **InnerDSL2** respectively. We use our own CVX-based projected gradient descent solver for the non-convex optimization problem the authors propose.
- **DSL1** and **DSL2** denote $d_S(A, B)$ when $S \in \mathbb{W}^n$ and $\|\cdot\|$ is $\|\cdot\|_1$ (element-wise) and $\|\cdot\|_F$, respectively. We implement them in Matlab (using CVX) as well as in C, aimed for medium size graphs and multi-core use. We also implemented a distributed version in Apache Spark (Zaharia et al. 2010) that scales to very large graphs over multiple machines based on the Alternating Directions Method of Multipliers (Boyd et al. 2011).
- **ORTHOP** and **ORTHFR** denote $d_S(A, B)$ when $S \in \mathbb{O}^n$ and $\|\cdot\|$ is $\|\cdot\|_2$ (operator norm) and $\|\cdot\|_F$ respectively. We compute them using an eigendecomposition.
- For small graphs, we compute $d_{\mathbb{P}^n}(A, B)$ using our brute-force GPU-based code. For a single pair of graphs with $n \geq 15$ nodes, **EXACT** already takes several days to finish. For $\|\cdot\| = \|\cdot\|_1$ in d_S (element-wise or matrix norm), we have implemented the chemical distance as an integer value LP and solved it using branch-and-cut. It did not scale well for $n \geq 15$.
- We implemented the WL algorithm over Spark to run, multithreaded, on a machine with 40 CPUs.

We use all public algorithms as black boxes with their default parameters, as provided by the authors.

Table 2 Competitor Distance Scores & Our Metrics

(Non-metric) Distance Score Algorithms	
NetAlignBP	Network Alignment using Belief Propagation (Bayati et al. 2009)
IsoRank	Neighborhood Topology Isomorphism using Page Rank (Singh et al. 2007)
SparseIsoRank	Neighborhood Topology Sparse Isomorphism using Page Rank (Bayati et al. 2009)
InnerPerm	Inner Product Matching with Permutations (Lyzinski et al. 2016)
InnerDSL1	Inner Product Matching with Matrices in $\mathbb{W}^n$ and entry-wise 1-norm (Lyzinski et al. 2016)
InnerDSL2	Inner Product Matching with Matrices in $\mathbb{W}^n$ and Frobenius norm (Lyzinski et al. 2016)
NetAlignMR	Iterative Matching Relaxation (Klau 2009)
Natalie (V2.0)	Improved Iterative Matching Relaxation (El-Kebir et al. 2015)
Metrics from our Family (2)	
EXACT	Chemical Distance via brute force search over GPU
DSL1	Doubly Stochastic Chemical Distance $d_{\mathbb{W}^n}$ with entry-wise 1-norm
DSL2	Doubly Stochastic Chemical Distance $d_{\mathbb{W}^n}$ with Frobenius norm
ORTHOP	Orthogonal Relaxation of Chemical Distance $d_{\mathbb{O}^n}$ with operator 2-norm
ORTHFR	Orthogonal Relaxation of Chemical Distance $d_{\mathbb{O}^n}$ with Frobenius norm

Experimental results

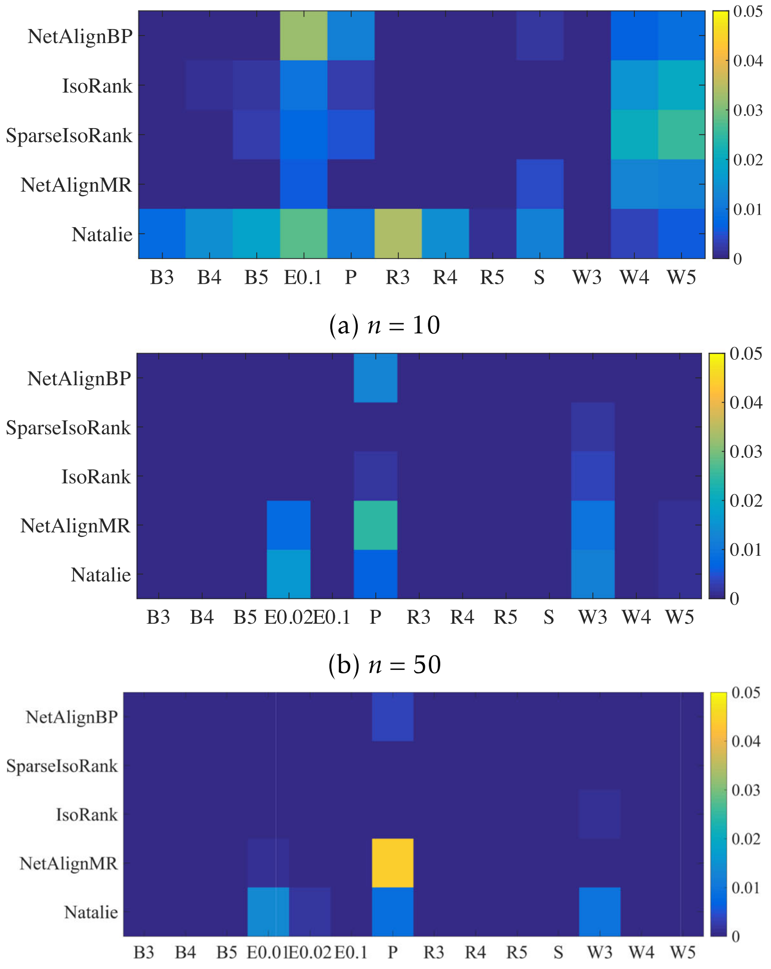
Clustering Graphs. The difference between our metrics and non-metrics is striking when clustering graphs. This is illustrated by the clustering experiment shown in Fig. 1. Graphs of size $n = 50$ from the 6 classes in Table 3 are clustered together through hierarchical agglomerative clustering. We compute distances between them using nine different algorithms; only the distances in our family (DSL1, DSL2, ORTHOP, and ORTHFR) are metrics. The quality of clusters induced by our metrics are far superior than clusters induced by non-metrics; in fact, **ORTHOP** and **ORTHFR** can lead to no misclassifications. This experiment strongly suggests our produced metrics correctly capture the topology of the metric space between these larger graphs.

Triangle Inequality Violations. Given graphs A, B and C and a distance d, a Triangle Inequality Violation (TIV) occurs when $d(A, C) > d(A, B) + d(B, C)$. Being metrics, none of our distances induce TIVs; this is not the case for the remaining algorithms in Table 2. Figure 2 shows the TIV fraction across the synthetic graphs of Table 3 while Fig. 3 shows the fraction of TIVs found on the 853 small graphs ($n = 7$). **NetAlignMR** also produces no TIVs on the small graphs, but it does induce TIVs in synthetic graphs. We observe that it is easier to find TIVs when graphs are close: in synthetic graphs, TIVs abound for $n = 10$. No algorithm performs well across all categories of graphs.

Effect of TIVs on Clustering. Next, to investigate the effect of TIVs on clustering, we artificially introduced triangle inequality violations into the pairs of distances between graphs. We then re-evaluated clustering performance for hierarchical agglomerative clustering using the *Ward* method, which performed best in Fig. 1. Figure 4 shows the fraction of misclassified graphs as the fraction of TIVs introduced increases. To incur as small a perturbation on distances as possible, we introduce TIVs as follows: For every three graphs, A, B, C, with probability p, we set $d(A, C) = d(A, B) + d(B, C)$. Although this does not introduce a TIV w.r.t. A, B, and C, this distortion does introduce TIVs w.r.t. other triplets involving A and C. We repeat this 20 times for each algorithm and each value of p, and compute the average fraction of TIVs, shown in the x-axis, and the average fraction of misclassified graphs, shown in the y-axis. As little as 1% TIVs significantly deteriorate clustering performance. Note that the fraction of TIVs is computed over the total number of TIVs possible, which grows cubicly with the number of graphs being clustered. We also see that, even after introducing TIVs, clustering based on metrics outperforms clustering based on non-metrics.

Table 3 Synthetic Graph Classes

	Description
B*d*	Barabasi Albert of degree *d* (Albert and Barabási 2002)
E*p*	Erdős-Rényi with probability *p* (Erdös and Rényi 1959)
P	Power Law Tree (Mahmoud et al. 1993)
R*d*	Regular Graph of degree *d* (Bollobás 1998)
S	Small World (Kleinberg 2000)
W*d*	Watts Strogatz of degree *d* (Watts and Strogatz 1998)



(a) $n = 10$

(b) $n = 50$

(c) $n = 100$

Fig. 2 Fraction of triangle inequality violations (TIVs) for different algorithms on random graphs of different types and sizes (a n = 10, b n = 50, c n =100): **DSL1**, **DSL2**, **ORTHOP**, and **ORTHOP**, are not shown since, as pseudometrics, they have zero TIVs. **a** $n = 10$, **b** $n = 50$, **c** $n = 100$

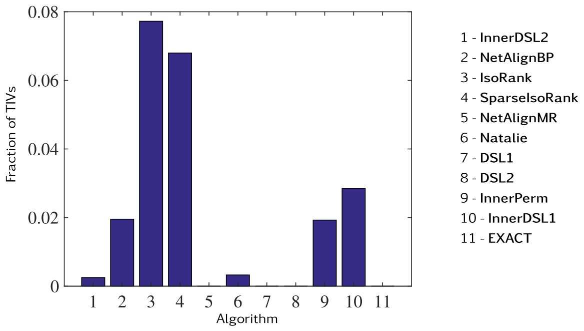


Fig. 3 Triangle Inequality Violations (TIVs) over the *small graphs* dataset

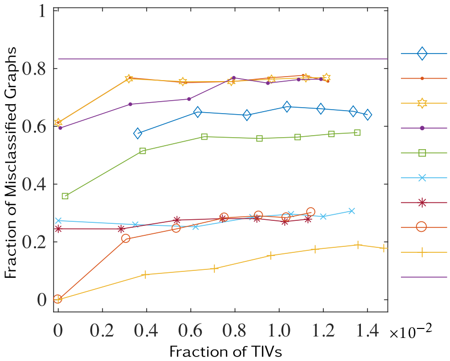


Fig. 4 Effect of introducing triangle inequality violations on the performance of different algorithms on the clustering experiment of Fig. 1 when using the Ward method

Comparison to Chemical Distance. We compare how different distance scores relate to the chemical distance **EXACT** through two experiments on the small graphs (computation on larger graphs is prohibitive). In Fig. 5a), we compare the distances between small graphs with 7 nodes produced by the different algorithms and **EXACT** using the DISTATIS method of (Abdi et al. 2005). Let $D \in \mathbb{R}_{+}^{835\times835}$ be the matrix of distances between graphs under an algorithm. DISTATIS computes the normalized Laplacian of this matrix, given by $L = -UDU/\|UDU\|_2$ where $U = I - \frac{\mathbf{1}\mathbf{1}^\top}{n}$. The DISTATIS score is the cosine similarity of such Laplacians (vectorized). We see that our metrics produce distances attaining high similarity with **EXACT**, though **NetAlignBP** has the highest similarity. We measure proximity to **EXACT** with an additional test. Given D, we compute the nearest neighbor (NN) meta-graph by connecting a graph in D to every graph at distance less than its average distance to other graps. This results in a (labeled) meta-graph, which we can compare to the NN meta-graph induced by other algorithms, measuring the fraction of distinct edges. Figure 5b shows that our algorithms perform quite well, though **Natalie** yields the smallest distance to **EXACT**.

Incorporating Constraints. Computation costs can be reduced through metric embeddings, as in (17). To show this, we produce a copy of the 5242 node collaboration graph

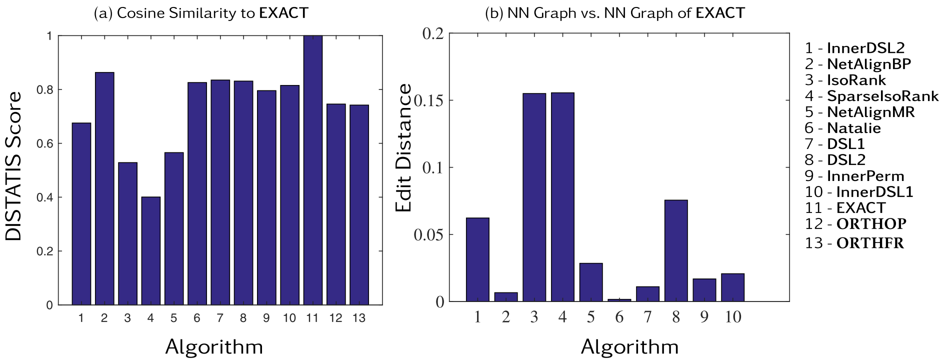


Fig. 5 **a** Cosine similarity between the Laplacian of distances produced by each algorithm and the one by **EXACT**, measured via DISTATIS (Abdi et al. 2005). **b** Edit distance between nearest neighbor (NN) meta-graphs induced by different algorithms and NN meta-graph induced by **EXACT**

Table 4 Effect of coloring/hard constraints

k	$\|P\|_0$	$\|AP—PA\|_0$	τ
1	3,747,960	100.569	133s
2	239,048	3004	104s
3	182,474	2036	136s
4	182,016	2030	169s
5	182,006	2030	200s

Effect of coloring/hard constraints on the numbers of variables ($\|P\|_0$) and terms of objective ($\|AP - PA\|_0$) using k iterations of the WL coloring algorithm. The last column shows the execution time of WL on a 40 CPU machine using Apache Spark (Zaharia et al. 2010)

with permuted node labels. We then run the WL algorithm (Weisfeiler and Lehman 1968) to produce structural colors, which induce coloring constraints on $P \in \mathbb{W}^n$. The WL algorithm reaches a fixed point after $k = 5$ iterations. The support of P (i.e., the number of variables in the optimization (12)), the support of $AP - PA$ (i.e., the number of non-zero summation terms in the objective of (12)), as well as the execution time τ of the WL algorithm, are summarized in Table 4. The original unconstrained problem involves $5242^2 \approx 27.4$M variables. However, after using WL and induced costraints, the effective dimension of the optimization problem (12) reduces considerably. This, in turn, speeds up convergence time, shown in Fig. 6: including the time to compute constraints, a solution is found 110 times faster after the introduction of the constraints.

Conclusion

Our work suggests that incorporating soft and hard constraints has a great potential to further improve the efficiency of our metrics. In future work, we intend to investigate and characterize the resulting equivalence classes under different soft and hard constraints, and to quantify these gains in efficiency. We also plan to develop scalable distributed solvers for our family of metrics. A good starting point is the Alternating Direction Method of Multipliers (Gabay and Mercier 1976; Glowinski and Marroco 1975), which enjoys several useful properties. Specifically, under proper tuning and mild convexity assumptions, it achieves the convergence rate of the fastest-possible first-order method (França and Bento 2016; Nesterov 2013), it can be less affected by the topology

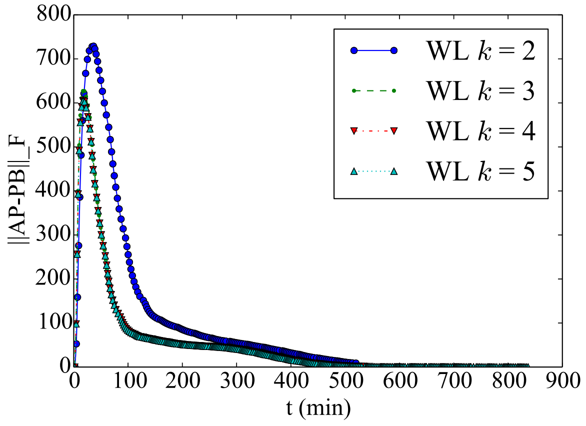


Fig. 6 Convergence of ADMM algorithm (Boyd et al. 2011) computing **DSL2** on two copies of the collaboration graph as a function of time, implemented using Apache Spark (Zaharia et al. 2010) on a 40 CPU machine

of the communication network in a cluster than, e.g. gradient descent (França and Bento 2017a; 2017b), and it parallelizes well both on share-memory multiprocessor systems, GPUs and computer clusters (Boyd et al. 2011; Parikh and Boyd 2014; Hao et al. 2016). Determining the necessity of the conditions used in proving that d_S is a metric is also an open problem. Finally, we are investigating generalizations of our family of metrics to multi-metrics, i.e. we want to define a tractable closeness score for a set of $n > 2$ graphs that satisfies a generalization of the properties of metrics for more than two elements (Safavi and Bento 2018).

Abbreviations

ADMM: Alternating Directions Method of Multipliers; CKS: Chartrand-Kubiki-Shultz; GPU: Graph Processing Unit; LCM: Least Common Multiple; WL: Weisfeiler-Lehman

Acknowledgements

The authors gratefully acknowledge the support of the National Science Foundation (grants IIS-1741197,IIS-1741129) and of the National Institutes of Health (grant 1U01AI124302).

Authors' contributions

The authors proved the main theorems collaboratively. All experimental results were implemented and executed by JB, with the exception of the experiments in Fig. 6, which was implement and executed by SI. Both authors read and approved the final manuscript.

Availability of data and materials

McKay, B.: List of 7 node connected graphs. http://users.cecs.anu.edu.au/ bdm/data/graphs.html Leskovec, J., Kleinberg, J., Faloutsos, C.: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/ca-GrQc.html Khan, A., Gleich, D., Halappanavar, M., Pothen, A.: Multicore codes for network alignment. https://www.cs.purdue.edu/homes/dgleich/codes/netalignmc/ El-Kebir, M., Heringa, J., Klau, G.: Natalie, a tool for pairwise global network alignment. http://www.mi.fu-berlin.de/w/LiSA/Natalie Bento, J.: Graph distance via brute force GPU computation. https://github.com/bentoayr/exact_graph_match

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, Boston College, St. Mary's Hall, 2nd floor, 02467 Chestnut Hill, MA, USA. [2]Department of Electrical and Computer Engineering, Northeastern University, 360 Huntington Avenue, 02115 Boston, MA, USA.

Received: 1 March 2019 Accepted: 2 October 2019
Published online: 15 November 2019

References

Abdi H, O'Toole AJ, Valentin D, Edelman B (2005) DISTATIS: The analysis of multiple distance matrices. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops. https://doi.org/10.1109/cvpr.2005.445

Absil P-A, Mahony R, Sepulchre R (2009) Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton

Ackermann MR, Blömer J, Sohler C (2010) Clustering for metric and nonmetric distance measures. ACM Trans Algoritm (TALG) 6(4):59

Aflalo Y, Bronstein A, Kimmel R (2015) On convex relaxation of graph isomorphism. PNAS 112(10):2942–2947

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr B Struct Sci 58(3):380–388

Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: Principles of Data Mining and Knowledge Discovery. https://doi.org/10.1007/3-540-45681-3_2

Babai L (2016) Graph isomorphism in quasipolynomial time [extended abstract]. In: Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing - STOC 2016. https://doi.org/10.1145/2897518.2897542

Bayati M, Gerritsen M, Gleich DF, Saberi A, Wang Y (2009) Algorithms for large, sparse network alignment problems. In: Ninth IEEE International Conference on Data Mining. https://doi.org/10.1109/icdm.2009.135

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends® Mach Learn 3(1):1–122

Bento J, Ioannidis S (2018) A family of tractable graph distances. In: Proceedings of the 2018 SIAM International Conference on Data Mining. SIAM. pp 333–341. https://doi.org/10.1137/1.9781611975321.38

Bento, J, Ioannidis S Graph distance via brute force GPU computation. https://github.com/bentoayr/exact_graph_match

Bertsekas DP (1997) Nonlinear programming. J Oper Res Soc 48(3):334–334

Beygelzimer A, Kakade S, Langford J (2006) Cover trees for nearest neighbor. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. https://doi.org/10.1145/1143844.1143857

Birkhoff G (1946) Three observations on linear algebra. Univ Nac Tucumán Revista A 5:147–151
Bollobás B (1998) Random graphs. In: Modern Graph Theory. Springer, Berlin/Heidelberg. pp 215–252
Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge university press, Cambridge, UK
Bunke H (1997) On a relation between graph edit distance and maximum common subgraph. Pattern Recog Lett 18(8):689–694
Bunke H, Shearer K (1998) A graph distance metric based on the maximal common subgraph. Pattern Recog Lett 19(3):255–259
Chartrand G, Kubicki G, Schultz M (1998) Graph similarity and distance in graphs. Aequationes Math 55(1-2):129–145
Clarkson KL (1999) Nearest neighbor queries in metric spaces. Discret Comput Geom 22(1):63–93
Clarkson, KL (2006) Nearest-neighbor searching and metric space dimensions. In: Nearest-Neighbor Methods for Learning and Vision: Theory and Practice. MIT Press, Cambridge. pp 15–59
Conte D, Foggia P, Sansone C, Vento M (2004) Thirty years of graph matching in pattern recognition. Int J Pattern Recog Artif Intell 18(03):265–298
El-Kebir M, Heringa J, Klau GW (2015) Natalie 20: Sparse global network alignment as a special case of quadratic assignment. Algorithms 8(4):1035–51
Elghawalby H, Hancock ER (2008) Measuring graph similarity using spectral geometry. In: Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-540-69812-8_51
Erdös P, Rényi A (1959) On random graphs, i. Publ Math (Debrecen) 6:290–297
Fankhauser S, Riesen K, Bunke H (2011) Speeding up graph edit distance computation through fast bipartite matching. In: Graph-Based Representations in Pattern Recognition. https://doi.org/10.1007/978-3-642-20844-7_11
Ferrer M, Valveny E, Serratosa F, Riesen K, Bunke H (2010) Generalized median graph computation by means of graph embedding in vector spaces. Pattern Recog 43(4):1642–1655
França G, Bento J (2016) An explicit rate bound for over-relaxed admm. In: Information Theory (ISIT), 2016 IEEE International Symposium On. IEEE. pp 2104–2108. https://doi.org/10.1109/isit.2016.7541670
França G., Bento J. (2017) Markov chain lifting and distributed ADMM. IEEE Sig Process Lett 24(3):294–298
França G, Bento J. (2017) How is distributed admm affected by network topology?. arXiv preprint arXiv:1710.00889
Gabay D, Mercier B (1976) A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput Math Appl 2(1):17–40
Garey MR, Johnson DS (2002) Computers and Intractability vol. 29. W. H. Freeman and Company, New York
Glowinski R, Marroco A (1975) Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. ESAIM: Math Model Numer Anal Modélisation Math Anal Numérique 9(R2):41–76
Gold S, Rangarajan A (1996) Softmax to softassign: Neural network algorithms for combinatorial optimization. J Artif Neural Netw 2(4):381–399
Hao N., Oghbaee A., Rostami M., Derbinsky N., Bento J. (2016) Testing fine-grained parallelism for the admm on a factor-graph. In: Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International. IEEE. pp 835–844
Hartigan JA (1975) Clustering Algorithms. Wiley, New York
He L, Han CY, Wee WG (2006) Object recognition and recovery by skeleton graph matching. In: 2006 IEEE International Conference on Multimedia and Expo. https://doi.org/10.1109/icme.2006.262700
Hoffman AJ, Wielandt HW (1953) The variation of the spectrum of a normal matrix. Duke Math J 20(1):37–39
Horn RA, Johnson CR (2012) Matrix Analysis. Cambridge University Press, Cambridge, UK
Hu N, Thibert B, Guibas L (2016) Distributable consistent multi-graph matching. arXiv preprint arXiv:1611.07191
Indyk P. (1999) Sublinear time algorithms for metric space problems. In: Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing. ACM. pp 428–434. https://doi.org/10.1145/301250.301366
Jain BJ (2016) On the geometry of graph spaces. Discret Appl Math 214:126–144
Jonker R, Volgenant A (1987) A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38(4):325–340
Klau GW (2009) A new graph-based method for pairwise global network alignment. BMC Bioinformatics 10(1):59
Kleinberg J (2000) The small-world phenomenon: An algorithmic perspective. In: STOC. ACM, New York
Koca J, Kratochvil M, Kvasnicka V, Matyska L, Pospichal J (2012) Synthon Model of Organic Chemistry and Synthesis Design vol. 51. Springer, Berlin/Heidelberg
Koutra D, Tong H, Lubensky D (2013) Big-align: Fast bipartite graph alignment. In: ICDM. https://doi.org/10.1109/icdm.2013.152
Koutra D, Vogelstein JT, Faloutsos C (2013) Deltacon: A principled massive-graph similarity function. In: SDM. https://doi.org/10.1137/1.9781611972832.18
Kuhn HW (1955) The hungarian method for the assignment problem. Nav Res Logist Q 2(1-2):83–97
Kvasnička V, Pospíchal J, Baláž V (1991) Reaction and chemical distances and reaction graphs. Theor Chem Acc Theory Comput Model (Theoretica Chimica Acta) 79(1):65–79
Lyzinski V, Fishkind DE, Fiori M, Vogelstein JT, Priebe CE, Sapiro G (2016) Graph matching: Relax at your own risk. IEEE Trans Pattern Anal Mach Intell 38(1):60–73
Macindoe O., Richards W. (2010) Graph comparison using fine structure analysis. In: SocialCom. https://doi.org/10.1109/socialcom.2010.35
Mahmoud HM, Smythe RT, Szymański J (1993) On the structure of random plane-oriented recursive trees and their branches. Random Struct Algoritm 4(2):151–176
Narayanan A, Shi E, Rubinstein BI (2011) Link prediction by de-anonymization: How we won the kaggle social network challenge. In: Neural Networks (IJCNN), The 2011 International Joint Conference On. IEEE, New York. pp 1825–1834
Nesterov Y (2013) Introductory Lectures on Convex Optimization: A Basic Course. vol. 87. Springer, Berlin/Heidelberg
Nesterov Y, Nemirovskii A (1994) Interior-point Polynomial Algorithms in Convex Programming vol. 13. Siam, Philadelphia
Pachauri D, Kondor R, Singh V (2013) Solving the multi-way matching problem by permutation synchronization. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., Red Hook. pp 1860–1868

Papadimitriou P, Dasdan A, Garcia-Molina H (2010) Web graph similarity for anomaly detection. J Internet Serv Appl 1(1):19–30
Parikh N, Boyd S (2014) Block splitting for distributed optimization. Math Program Comput 6(1):77–102
Ramana M. V., Scheinerman E. R., Ullman D. (1994) Fractional isomorphism of graphs. Discret Math 132(1-3):247–265
Riesen K, Bunke H (2009) Approximate graph edit distance computation by means of bipartite graph matching. Image Vis Comput 27(7):950–959
Riesen, K, Bunke H (2010) Graph Classification and Clustering Based on Vector Space Embedding vol. 77. World Scientific, Singapore
Riesen K, Neuhaus M, Bunke H (2007) Graph embedding in vector spaces by means of prototype selection. In: Graph-Based Representations in Pattern Recognition. https://doi.org/10.1007/978-3-540-72903-7_35
Safavi S, Bento J (2018) n-metrics for multiple graph alignment. arXiv preprint arXiv:1807.03368
Sanfeliu A, Fu K (1983) A distance measure between attributed relational graphs for pattern recognition. Trans Syst IEEE Man Cybern SMC-13(3):353–362
Schellewald C, Roth S, Schnörr C (2001) Evaluation of convex optimization techniques for the weighted graph-matching problem in computer vision. In: Lecture Notes in Computer Science. https://doi.org/10.1007/3-540-45404-7_48
Sebastian TB, Klein PN, Kimia BB (2004) Recognition of shapes by editing their shock graphs. IEEE Trans Pattern Anal Mach Intell 26(5):550–571
Shen Y, Lin W, Yan J, Xu M, Wu J, Wang J (2015) Person re-identification with correspondence structure learning. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE Computer Society, Washington. pp 3200–3208
Singh R., Xu J., Berger B. (2007) Pairwise global alignment of protein interaction networks by matching neighborhood topology. In: Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-540-71681-5_2
Solé-Ribalta A, Serratosa F (2010) Graduated assignment algorithm for finding the common labelling of a set of graphs. In: Lecture Notes in Computer Science. Springer. pp 180–190. https://doi.org/10.1007/978-3-642-14980-1_17
Soundarajan S, Eliassi-Rad T, Gallagher B (2014) A guide to selecting a network similarity method. In: Proceedings of the 2014 SIAM International Conference on Data Mining. https://doi.org/10.1137/1.9781611973440.118
Tinhofer G (1986) Graph isomorphism and theorems of Birkhoff type. Computing 36(4):285–300
Umeyama S (1988) An eigendecomposition approach to weighted graph matching problems. IEEE Trans Pattern Anal Mach Intell 10(5):695–703
Vogelstein JT, Conroy JM, Podrazik LJ, Kratzer SG, Harley ET, Fishkind DE, Vogelstein RJ, Priebe CE (2011) Large (brain) graph matching via fast approximate quadratic programming. arXiv preprint arXiv:1112.5507
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
Weisfeiler B, Lehman AA (1968) A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Technicheskaya Informatsia 2(9):12–16
Wilson RC, Zhu P (2008) A study of graph spectra for comparing graphs and trees. Pattern Recog 41(9):2833–2841
Xing EP, Ng AY, Jordan MI, Russell S (2002) Distance metric learning with application to clustering with side-information. In: NIPS. MIT Press, Cambridge Vol. 15. p 12
Yan J, Cho M, Zha H, Yang X, Chu S (2015) A general multi-graph matching approach via graduated consistency-regularized boosting. arXiv preprint arXiv:1502.05840
Yan J, Wang J, Zha H, Yang X, Chu S (2015) Consistency-driven alternating optimization for multigraph matching: A unified approach. IEEE Trans Image Process 24(3):994–1009
Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: Cluster computing with working sets. HotCloud 10(10-10):95
Zaslavskiy M, Bach F, Vert J-P (2009a) A path following algorithm for the graph matching problem. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(12):2227–2242
Zaslavskiy M, Bach F, Vert JP (2009b) Global alignment of protein–protein interaction networks by graph matching methods. Bioinformatics 25(12):259–1267
Zaslavskiy M, Bach F, Vert J-P (2010) Many-to-many graph matching: a continuous relaxation approach. In: Machine Learning and Knowledge Discovery in Databases. Springer. pp 515–530. https://doi.org/10.1007/978-3-642-15939-8_33
Zhou F, De la Torre F (2012) Factorized graph matching. In: IEEE Conference on Computer Vision & Pattern Recognition (CVPR). IEEE. pp 127–134. https://doi.org/10.1109/cvpr.2012.6247667
Zhu P, Wilson RC (2005) A study of graph spectra for comparing graphs. In: Procedings of the British Machine Vision Conference. https://doi.org/10.5244/c.19.69

Publisher's Note