A Family of Tractable Graph Distances

Jose Bento∗
Stratis Ioannidis†

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 and the CKS 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 CKS distance, 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.

1 Introduction

Graph similarity and the related problem of graph isomorphism have a long history in data mining, machine learning, and pattern recognition [21, 13, 38]. 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 [20, 19, 11], clustering [3], outlier detection [7], and diameter computation [31] 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 [40]. Formally, given graphs $G_A$ and $G_B$, represented by their adjacency matrices $A, B \in \{0, 1\}^{n \times n}$, the chemical distance is $d_{\text{chem}}(A, B)$ is defined in terms of a mapping between the two graphs that minimizes their edge discrepancies, i.e.:

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

where $\mathbb{P}^n$ is the set of permutation matrices of size $n$ and $\| \cdot \|_F$ is the Frobenius norm (see Sec. 2 for definitions). The Chartrand-Kubiki-Shultz (CKS) [18] distance is an alternative: CKS is again given by (1.1) 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; 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 $P$ s.t. $AP = PB$ (for both adjacency and path matrices), is famously a problem that is neither known to be in $P$ nor shown to be NP-hard [8].

Despite their computational advantages, unfortunately, using them to approximate $d_{\text{chem}}(A, B)$ 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 [3]. We also demonstrate this empirically in Section 5 (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.

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

Copyright © 2018 by SIAM
Unauthorized reproduction of this article is prohibited
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 generalization 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 (1.1) of the form:

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

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 (i) adjacency, path distances, weights, etc.). We make the following contributions:

- We prove sufficient conditions on $S$ and norm $\| \cdot \|$ for which (1.2) is a metric. In particular, we show that $d_S$ is a so-called *pseudo-metric* (see Sec. 2) when:
  1. $S = \mathbb{P}^n$ and $\| \cdot \|$ is any entry-wise or operator norm;
  2. $S = \mathbb{W}^n$, the set of doubly stochastic matrices, $\| \cdot \|$ is any entry-wise or operator 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
  3. $S = \mathbb{O}^n$, the set of orthogonal matrices, and $\| \cdot \|$ is the operator or entry-wise 2-norm.

Relaxations (ii) and (iii) are very important from a practical standpoint. For all matrix norms, computing (1.2) with $S = \mathbb{W}^n$ is tractable, as it is a convex optimization. For $S = \mathbb{O}^n$, (1.2) is non-convex but is still tractable, as it reduces to a spectral decomposition. This was known for the Frobenius norm [56]; we prove this is the case for the operator 2-norm also.

- 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. (1.2). Crucially, we do this without affecting the metric 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 method of multipliers [15], which we implement over a compute cluster using Apache Spark [62].

**Related Work.** Graph distance (or similarity) scores find applications in varied fields such as in image processing [21], chemistry [60][20], and social network analysis [13][38]. Graph distances are easy to define when, contrary to our setting, the correspondence between graph nodes is known, i.e., graphs are labeled [46][38][55]. Beyond the chemical distance, classic examples of distances between unlabeled graphs are the edit distance [28][51] and the maximum common subgraph distance [17][10], both of which also have versions for labeled graphs. Both are metrics and are hard to compute, while existing heuristics [48][36] are not metrics. The reaction distance [36] is also a metric directly related to the chemical distance [10] when edits are restricted to edge additions and deletions. Jain [32] 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 metrics [32][36] to attain tractability.

A metric can also be induced by embedding graphs in a metric space and measuring the distance of these embeddings [50][27][39]. Several works follow such an approach, mapping graphs, e.g., to spaces determined by their spectral decomposition [63][60][24]. In general, in contrast to our metrics, such approaches are not as discriminative, as embeddings summarize graph structure. Continuous relaxations of graph isomorphism, both convex and non-convex [42][41][56], have found applications in a variety of contexts, including social networks [57], computer vision [52], shape detection [63][30], and neuroscience [57]. None of the above works focus on metric properties of resulting relaxations, which several fail to satisfy [57][57][63][30].

Metrics naturally arise in data mining tasks, including clustering [61][29], NN search [20][19][11], and outlier detection [7]. Some of these tasks become tractable or admit formal guarantees precisely when performed over a metric space. For example, finding the nearest neighbor [20][19][11] or the diameter of a dataset [31] become polylogarithimic under metric assumptions; similarly, approximation algorithms for clustering (which is NP-hard) rely on metric assumptions, whose absence leads to a deterioration on known bounds [3]. Our search for metrics is motivated by these considerations.

### 2 Notation and Preliminaries

**Graphs.** We represent an undirected graph $G(V, E)$ with node set $V = [n] = \{1, \ldots, n\}$ and edge set $E \subseteq [n] \times [n]$ by its adjacency matrix, i.e., $A = [a_{i,j}]_{i,j \in [n]} \in \{0,1\}^{n \times n}$ s.t. $a_{ij} = a_{ji} = 1$ 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 graphs are represented by (possibly non-
Matrix Norms. Given a matrix $A = [a_{ij}]_{i,j \in [n]} \in \mathbb{R}^{n \times n}$ and $p \in \mathbb{N} \cup \{\infty\}$, 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 $\|A\|_\infty = \max_{i,j} |a_{ij}|$. We denote the entry-wise 2-norm (i.e., the Frobenius norm) as $\|\cdot\|_F$.

Permutation, Doubly Stochastic, and Orthogonal Matrices. We denote the set of permutation matrices as $\mathbb{P}^n = \{P \in \{0,1\}^{n \times n} : P1 = 1, P^T 1 = 1\}$, the set of doubly-stochastic matrices (i.e., the Birkhoff polytope) as $\mathbb{W}^n = \{W \in [0,1]^{n \times n} : W1 = 1, W^T 1 = 1\}$, and the set of orthogonal matrices (i.e., the Stiefel manifold) as $\mathbb{O}^n = \{U \in \mathbb{R}^{n \times n} : U^TU = I\}$. Note that $\mathbb{P}^n = \mathbb{W}^n \cap \mathbb{O}^n$. Moreover, the Birkoff-von Neumann Theorem \cite{12} states that $\mathbb{W}^n = \text{conv}(\mathbb{P}^n)$, i.e., the Birkhoff polytope is the convex hull of $\mathbb{P}^n$.

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

\begin{align*}
(2.3a) & \quad d(x, y) \geq 0 \quad \text{(non-negativity)} \\
(2.3b) & \quad d(x, y) = 0 \iff x = y \quad \text{(pos. definiteness)} \\
(2.3c) & \quad d(x, y) = d(y, x) \quad \text{(symmetry)} \\
(2.3d) & \quad d(x, y) \leq d(x, z) + d(z, y) \quad \text{(triangle inequality)}
\end{align*}

A function $d$ is called a pseudometric if it satisfies \(2.3a\), \(2.3c\), and \(2.3d\), but not the positive definiteness property \(2.3b\) is replaced by the (weaker) property:

\begin{align*}
(2.3e) & \quad d(x, x) = 0 \quad \text{for all } x \in \Omega.
\end{align*}

If $d$ is a pseudometric, then $d(x, y) = 0$ defines an equivalence relation $\sim_d$ over $\Omega$. A pseudometric is then a metric over $\Omega/\sim_d$, the quotient space of $\sim_d$. A $d$ that satisfies \(2.3a\), \(2.3b\), and \(2.3d\) but not the symmetry property \(2.3c\) is called a quasimetric. If $d$ is a quasimetric, then its symmetric extension $\bar{d} : \Omega \times \Omega \to \mathbb{R}$, defined as $\bar{d}(x, y) = d(x, y) + d(y, x)$, is a metric over $\Omega$.

Graph Isomorphism, Chemical, 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. $P^T AP = B$ or, equivalently, $AP = PB$. The chemical distance, given by \(1.1\), extends the latter relationship to capture distances between graphs. 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:

\begin{align*}
(2.4) & \quad d_S(A, B) = \min_{P \in S} \|AP - PB\|,
\end{align*}

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 \(1.1\) when $\Omega = \mathbb{R}^{n \times n}$, $S = \mathbb{P}^n$ and $\|\cdot\| = \|\cdot\|_F$.

In CKS distance \cite{13}, matrices $A, B$ contain pairwise path distances between any two nodes; equivalently, CKS is the chemical distance of two weighted complete graphs with path distances as edge weights. Our main contribution is determining general conditions on $\Omega$ and $\|\cdot\|$ under which $d_S$ is a metric over $\Omega$, for arbitrary weighted graphs, thereby including both the chemical and CKS distances as special cases.

For concreteness, we focus on distances between graphs of equal size. Extensions to graphs of unequal size are described in \cite{10}.

3 A Family of Graph Metrics

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

**Theorem 3.1.** If $S = \mathbb{P}^n$ and $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then $d_S$ given by \(2.4\) is a pseudometric over $\Omega = \mathbb{R}^{n \times n}$.

Hence, $d_{P^n}$ is a pseudometric under any entry-wise or operator norm over arbitrary directed, weighted graphs. Our second result states that this property extends to the relaxed version of the chemical distance, in which permutations are replaced by doubly stochastic matrices.

**Theorem 3.2.** If $S = \mathbb{W}^n$ and $\|\cdot\|$ is an arbitrary entry-wise or operator norm, then $d_S$ given by \(2.4\) 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 \(2.4\) defines a pseudometric over undirected graphs. The symmetry property \(2.3c\) breaks if $\|\cdot\|$ is an operator norm or graphs are directed. In either case, $d_S$ is a quasimetric over the quotient space $\Omega/\sim_d$, and symmetry is attained via the symmetric extension $\bar{d}_S$.

**Theorem 3.3.** If $S = \mathbb{O}^n$ and $\|\cdot\|$ is either the operator or the entry-wise (i.e., Frobenius) 2-norm, then $d_S$ given by \(2.4\) is a pseudometric over $\Omega = \mathbb{R}^{n \times n}$.

Though \(2.4\) is not a convex problem when $S = \mathbb{O}^n$, it is also tractable. Umeayama \cite{58} 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 [10]). In the general case of directed graphs, (2.4) is a classic example of a problem that can be solved through optimization on manifolds [2].

**Equivalence Classes.** The equivalence of matrix norms implies that all pseudometrics $d_S$ defined through (2.4) for a given $S$ have the same quotient space $\Omega/\sim_{d_S}$: if $d_S(A, B) = 0$ for one matrix norm $\| \cdot \|$ in (2.4), it will be so for all. When $S = \mathbb{P}^n$, $\Omega/\sim_{d_{\mathbb{P}}} = \Omega$ is the quotient space defined by graph isomorphism: any two adjacency matrices $A, B \in \mathbb{R}^{n \times n}$ satisfy $d_{\mathbb{P}}(A, B) = 0$ if and only if their (possibly weighted) graphs are isomorphic. When $S = \mathbb{W}^n$, the quotient space $\Omega/\sim_{d_{\mathbb{W}}}$ has a connection to the Weisfeiler-Lehman (WL) algorithm [59] described in [10]; Ramana et al. [17] show that $d_{\mathbb{W}}(A, B) = 0$ if and only if $G_A$ and $G_B$ receive identical colors by the WL algorithm. If $S = \mathbb{O}^n$ and $S = \mathbb{S}^n$, i.e., graphs are undirected, then $\Omega/\sim_{d_{\mathbb{O}}}$ is determined by co-spectrality: $d_{\mathbb{O}}(A, B) = 0$ if and only if $A, B$ have the same spectrum. When $\Omega = \mathbb{R}^{n \times n}$, $d_O(A, B) = 0$ implies that $A, B$ are co-spectral, but co-spectral matrices $A, B$ do not necessarily satisfy $d_O(A, B) = 0$.

### 3.1 Proof of Theorems 3.1–3.3

We define several properties that play a crucial role in our proofs. We say that a set $S \subseteq \mathbb{R}^{n \times n}$ is closed under multiplication if $P, P' \in S$ implies that $P \cdot P' \in S$. We say that $S$ is closed under transposition if $P \in S$ implies that $P^\top \in S$, and closed under inversion if $P \in S$ implies that $P^{-1} \in S$. Finally, given a matrix norm $\| \cdot \|$, we say that set $S$ is contractive w.r.t. $\| \cdot \|$ if $\| A P \| \leq \| A \|$ and $\| P A \| \leq \| A \|$, for all $P \in S$ and $A \in \mathbb{R}^{n \times n}$. But differently, $S$ is contractive if and only if every $P \in S$ is a contraction w.r.t. $\| \cdot \|$. We rely on several lemmas, whose proofs can be found in [10]. The first three establish conditions under which (2.4) satisfies the triangle inequality (2.3d), symmetry (2.3c), and weak property (2.3e), respectively:

**Lemma 3.1.** Given a matrix norm $\| \cdot \|$, suppose that set $S$ 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 (2.4) satisfies $d_S(A, C) \leq d_S(A, B) + d_S(B, C)$.

**Lemma 3.2.** Given a matrix norm $\| \cdot \|$, suppose that $S \subseteq \mathbb{R}^{n \times n}$ is (a) contractive w.r.t. $\| \cdot \|$, and (b) closed under inversion. Then, for all $A, B \in \mathbb{R}^{n \times n}$, $d_S(A, B) = d_S(B, A)$.

**Lemma 3.3.** If $I \in S$, then $d_S(A, A) = 0$ for all $A \in \mathbb{R}^{n \times n}$.

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 $\| \cdot \|$, $d_{\mathbb{P}}$ and $d_{\mathbb{O}}$ defined in terms of this norm satisfy all assumptions of Lemmas 3.1 and 3.3. We therefore turn our attention to this property.

**Lemma 3.4.** Let $\| \cdot \|$ be any operator or entry-wise norm. Then, $S = \mathbb{P}^n$ is contractive w.r.t. $\| \cdot \|$.

Hence, Theorem 3.1 follows as a direct corollary of Lemmas 3.1 and 3.4. Indeed, $d_{\mathbb{P}}$ is non-negative, symmetric by Lemmas 3.2 and 3.4, satisfies the triangle inequality by Lemmas 3.1 and 3.4, as well as property (2.3e) by Lemma 3.3; hence $d_{\mathbb{P}}$ 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 3.5.** Let $\| \cdot \|$ be the operator 2-norm or the Frobenius norm. Then, $S = \mathbb{O}^n$ is contractive w.r.t. $\| \cdot \|$.

Theorem 3.3 follows from Lemmas 3.1 and 3.5 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 $\| \cdot \|_{\infty}$. Lemma 3.5 along with the Birkhoff-von Neumann theorem imply that $\mathbb{W}^n$ is also contractive:

**Lemma 3.6.** Let $\| \cdot \|$ be any operator or entry-wise norm. Then, $\mathbb{W}^n$ is contractive w.r.t. $\| \cdot \|$. 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 $d_{\mathbb{W}}$, we leverage the following lemma:

**Lemma 3.7.** Suppose that $\| \cdot \|$ is transpose invariant, and $S$ is closed under transposition. Then, $d_S(A, B) = d_S(B, A)$ for all $A, B \in \mathbb{S}^n$.

The first part of Theorem 3.2 therefore follows from Lemmas 3.1 and 3.6, as $\mathbb{W}^n$ is closed under transposition, contains the identity $I$, and is closed under multiplication, while all entry-wise norms are transpose invariant. Operator norms are not transpose invariant. However, if $\| \cdot \|$ is an operator norm, or $\Omega = \mathbb{R}^{n \times n}$, then Lemmas 3.6 and 3.1 imply that $d_{\mathbb{W}}$ satisfies non-negativity (2.3a) and the triangle inequality (2.3d), while Lemma 3.3 implies that it satisfies (2.3e) by (2.3e) These properties are inherited by extension $d_S$, which also satisfies symmetry (2.3c), and Theorem 3.2 follows. □

### 4 Incorporating Metric Embeddings

We have seen that the chemical distance $d_{\mathbb{P}}$ can be relaxed to $d_{\mathbb{W}}$ or $d_{\mathbb{O}}$, 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 $\psi_A : [n] \rightarrow \Omega$ from the nodes of the graph to a metric space $(\Omega, \bar{d})$. That is, $\psi_A$ maps nodes of the graph to $\Omega$, where $\Omega$ is endowed with a metric $\bar{d}$. We refer to a graph endowed with an embedding $\psi_A$ as an embedded graph, and denote this by $(A, \psi_A)$, 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 $\Omega$ be an arbitrary finite set endowed with the Kronecker delta as a metric, that is, for $s, s' \in \Omega$, $\bar{d}(s, s') = 0$ if $s = s'$, while $\bar{d}(s, s') = \infty$ if $s \neq s'$. Given a graph $G_A$, a mapping $\psi_A : [n] \rightarrow \Omega$ is then a metric embedding. The values of $\Omega$ are invariably called colors or labels, and a graph embedded in $\Omega$ 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 [59] after $k$ iterations.

As discussed below, node attributes translate to soft constraints in metric (2.4), while node colors correspond to hard constraints. The unified view through embeddings allows us to establish metric properties for both simultaneously (c.f. Thm. 4.1 and 1.2).

**Embedding Distance.** Consider two embedded graphs $(A, \psi_A), (B, \psi_B)$ of size $n$ that are embedded in the same metric space $(\Omega, \bar{d})$. 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 $\bar{d}(\psi_A(u), \psi_B(v))$. Let $D_{\psi_A, \psi_B} = [\bar{d}(\psi_A(u), \psi_B(v))]_{u \in V, v \in V} \in \mathbb{R}_{+}^{n \times n}$ be the corresponding matrix of embedded distances. After mapping nodes to the same metric space, it is natural to seek $P \in \mathbb{R}^n$ that preserve the embedding distance. This amounts to finding a $P \in \mathbb{R}^n$ that maps nodes in $A$ nodes in $B$ of equal color. It is not hard to verify\footnote{This follows from Thm. 4.1 for $A = B = 0$, i.e., for distances between embedded graphs with no edges.} that $\min_{P \in \mathbb{R}^n} \text{tr} (P^T D_{\psi_A, \psi_B})$ induces a metric between graphs embedded in $(\Omega, \bar{d})$. Despite the combinatorial nature of $\mathbb{R}^n$, (4.5) is a maximum weighted matching problem, which can be solved through, e.g., the Hungarian algorithm [39] in polynomial time in $n$.

We note that this metric is not as expressive as (2.4): depending on the definition of the embeddings $\psi_A, \psi_B$, attributes may only capture “local” similarities between nodes, as opposed to the “global” view of a mapping attained by (2.4).

**A Unified, Tractable Metric.** Motivated by the above considerations, we focus on unifying the “global” metric (2.4) with the “local” metrics induced by arbitrary graph embeddings. Proofs for the two theorems below are provided in the supplement. Given a metric space $(\Omega, \bar{d})$, let $\Psi^\Omega_\bar{d} = \{ \psi : [n] \rightarrow \Omega \}$ be the set of all mappings from $[n]$ to $\Omega$. Then, given two embedded graphs $(A, \psi_A), (B, \psi_B) \in \mathbb{R}^{n \times n} \times \Psi^\Omega_\bar{d}$, we define:

$$d_S((A, \psi_A), (B, \psi_B)) = \min_{P \in S} \|AP - PB\| + \ldots + \text{tr}(P^T D_{\psi_A, \psi_B}) \tag{4.6}$$

for some compact set $S \subset \mathbb{R}^{n \times n}$ and matrix norm $\| \cdot \|$.

**Theorem 4.1.** If $S = \mathbb{R}^n$ and $\| \cdot \|$ is an arbitrary entry-wise or operator norm, then $d_S$ given by (4.6) is a pseudometric over the set of embedded graphs $\Omega = \mathbb{R}^{n \times n} \times \Psi^\Omega_\bar{d}$.

We stress here that this result is non-obvious: is not true that adding any linear term to $d_S$ leads to a quantity that satisfies the triangle inequality. It is precisely because $D_{\psi_A, \psi_B}$ contains pairwise distances that Theorem 4.1 holds. We can similarly extend Theorem 3.2.

**Theorem 4.2.** If $S = \mathbb{R}^n$ and $\| \cdot \|$ is an arbitrary entry-wise or operator norm, then $d_S$ given by (4.6) is a pseudometric over $\Omega = \mathbb{R}^{n \times n} \times \Psi^\Omega_\bar{d}$. Furthermore, if $\| \cdot \|$ is an arbitrary entry-wise or operator norm, then the symmetric extension $d_S$ of (4.6) is a pseudometric over $\Omega = \mathbb{R}^{n \times n} \times \Psi^\Omega_\bar{d}$.

Adding the linear term (15) 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.
Figure 1: A clustering experiment using metrics and non-metrics (y-axis) for different clustering parameters (x-axis) is shown in (a), left. We sample graphs with $n = 50$ nodes from the six classes, shown in the adjacent table in (d), bottom-center. We compute distances between them using nine different algorithms from Table 1. Only the distances in our family (DSL1, DSL2, ORTHOP, and ORTHFR) are metrics. The resulting graphs are clustered using hierarchical agglomerative clustering [20] 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 $\approx 0.8$. Subfigures (b) and (c), top center and right, shows that non-metric distances produce triangle inequality violations (TIVs) which contribute to poor clustering results; the figure shows the fraction of TIVs within different 10-node and 50 node graph families under these algorithms. Finally, subfigure (e), bottom right, shows the fraction of triangle inequality violations for different algorithms on the small graphs dataset of all 7-node graphs.

Theorem 4.2 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 (2.4) is $n^2$, incorporating such hard constraints can significantly reduce the problem’s computation time; we illustrate this in the next section. Note that adding (4.5) to $d_C^n$ does not preserve the metric property.

5 Experiments

Graphs. We use synthetic graphs from six classes summarized in the table in Fig. 1(d). In addition, we use a dataset of small graphs, comprising all 853 connected graphs of 7 nodes [15]. Finally, we use a collaboration graph with 5242 nodes and 14496 edges representing author collaborations [11].

Algorithms. We compare our metrics to several competitors outlined in Table 1 (see also 10 for additional details). All receive only two unlabeled undirected simple graphs $A$ and $B$ and output a matching a matrix $\hat{P}$ either in $W^n$ or in $F^n$ estimating $P^\star$. If $\hat{P} \in F^n$, we compute $\|A\hat{P} - \hat{P}B\|_1$. If $\hat{P} \in W^n$, then we compute both $\|A\hat{P} - \hat{P}B\|_1$ and $\|A\hat{P} - \hat{P}B\|_F$; all norms are entry-wise. We also implement our two relaxations $d_W$ and $d_C^n$, for two different matrix norm combinations.

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(a). Graphs of size $n = 50$ from the 6 classes in Fig. 1(d) 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.
Effect of TIVs

(a) Cosine Similarity to EXACT

(b) Convergence of ADMM

(c) NN Graph vs. NN Graph of EXACT

Figure 2: (a) Effect of introducing TIVs on the performance of different algorithms on the clustering experiment of Figure 1(a) when using the Ward method. (b) Cosine similarity between the Laplacian of distances produced by each algorithm and the one by EXACT. (c) Distance between nearest neighbor (NN) graphs induced by different algorithms and NN graph induced by EXACT.

Triangle Inequality Violations (TIV). Given graphs $A$, $B$, and $C$ and a distance $d$, a 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 1. Figure 1(b) and (c) shows the TIV fraction across the synthetic graphs of Figure 1(d), with Figure 1(c) showing the fraction of TIVs found on the 835 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 Figure 1(a). Figure 2(a) 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. We also see that, even after introducing TIVs, clustering based on metrics outperforms clustering based on non-metrics.

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 Figure 2(b), we compare the distances between small graphs with 7 nodes produced by the different algorithms and EXACT using the DISTATIS method of [1]. Let $D \in \mathbb{R}^{835 \times 835}$ 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 $U = I - \frac{1}{n} 11^T$. 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 graphs. 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 2(c) 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 [4.6]. To show this, we produce a copy of the 5242 node collaboration graph with permuted node labels. We then run the WL algorithm [59] to produce structural colors, which induce coloring constraints on $P \in \mathbb{W}^n$. The support of $P$ (i.e., the number of variables in the optimization (2.4)), the support of $AP - PA$...
(i.e., the number of non-zero summation terms in the objective of (2.4)), as well as the execution time $\tau$ of the WL algorithm, are summarized in Fig. 3(b). The original unconstrained problem involves $2^{2422}$ ≈ 27.4M variables. However, after using WL and induced constraints, the effective dimension of the optimization problem (2.4) reduces considerably. This, in turn, speeds up convergence time, shown in Fig. 3(b): including the time to compute constraints, a solution is found 110 times faster after the introduction of the constraints.

6 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, especially in parallel implementations like ADMM. Determining the necessity of the conditions used in proving that $d_S$ is a metric is also an open problem.

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).

References


Copyright © 2018 by SIAM
Unauthorized reproduction of this article is prohibited


