Information and Inference: A Journal of the IMA (2019) **00**, 1–31
doi:10.1093/imaiai/iaz026

The Gromov–Wasserstein distance between networks and stable network invariants

SAMIR CHOWDHURY
Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA
chowdhury.57@osu.edu

AND

FACUNDO MÉMOLI†
Departments of Mathematics and Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA
†Corresponding author. Email: memoli@math.osu.edu

[Received on 29 May 2019; revised on 24 August 2019; accepted on 29 August 2019]

We define a metric—the network Gromov–Wasserstein distance—on weighted, directed networks that is sensitive to the presence of outliers. In addition to proving its theoretical properties, we supply network invariants based on optimal transport that approximate this distance by means of lower bounds. We test these methods on a range of simulated network datasets and on a dataset of real-world global bilateral migration. For our simulations, we define a network generative model based on the stochastic block model. This may be of independent interest for benchmarking purposes.

Keywords: asymmetric networks; Gromov–Wasserstein distances; network data analysis; metric measure spaces.

1. Introduction

1.1 *Motivation and related literature*

Advances in data mining are beinning to lead to the acquisition of large networks that are directed, weighted and possibly even signed [22]. In light of the ready availability of such data, a natural problem is to devise methods for comparing network datasets. These methods in turn lead to a wide range of applications. An example is the *network retrieval task*: given a database of networks and a query network, return an ordered list of the networks in the database that are most similar to the query. Additionally, because there may be redundant data in the networks that are not relevant to the query, one may wish to impose a notion of significance to certain substructures of the query network. The task then is to retrieve networks that are similar to the query network both globally and also at the scale of relevant substructures.

While there has been some work in devising directed, weighted analogues of conventional network analysis tools such as edge overlap and clustering coefficients, we are more interested in pairwise comparison of individual networks. The intuitive idea behind this comparison is to search for the best possible alignment of edges (according to weights) while simultaneously aligning nodes with similar significance.

Techniques based on optimal transport (OT) provide an elegant solution to this problem by endowing a network with a probability measure. The user adjusts the measure to signify important network

substructures and to smooth out the effect of outliers. This approach was adopted in [15] to compare various real-world network datasets modelled as *metric measure (mm) spaces*—metric spaces equipped with a probability measure. This work was based in turn on the formulation of the *Gromov–Wasserstein (GW) distance* between mm spaces presented in [20,21]. Specifically, this setting considered triples (X, d_X, μ_X) where (X, d_X) is a compact metric space and μ_X is a Borel probability measure.

Exact computation of GW distances amounts to solving a non-convex quadratic program. Towards this end, the computational techniques presented in [20,21] included both readily computable lower bounds and an alternate minimization scheme for reaching a minimum of the GW objection function. This alternate minimization scheme involved solving successive linear optimization problems, and was used for the computations in [15].

An alternative definition of the GW distance due to Sturm (the *transportation* formulation) appeared in [32], although this formulation is less amenable to practical computations than the one in [20] (the *distortion* formulation). Both the transportation and distortion formulations were studied carefully in [20,21,33]. It was further observed by Sturm in [33] that the definition of the (distortion) GW distance can be extended to *gauged measure spaces* of the form $(X, \hat{d}_X, \mu_X)$. Here X is a Polish space, $\hat{d}_X$ is a symmetric L^2 function on $X \times X$ (that does not necessarily satisfy the triangle inequality) and μ_X is a Borel probability measure on X. These results are particularly important in the context of the current paper. From here on, we always refer to the distortion formulation of the GW distance.

Sturm's work in [33] showed that while the collection of isomorphism classes of metric measure spaces is not complete, elements in its completion can be represented by triples $(X, \tilde{d}_X, \mu_X)$ where X, μ_X are as above, and $\tilde{d}_X$ is a symmetric, measurable, square integrable function satisfying the triangle inequality almost everywhere. He further showed that the ambient space of gauged measure spaces, which is interpreted as being 'more linear' due to giving up the triangle inequality, admits explicit descriptions of geometric properties.

In Sturm's work, symmetry is desirable because, for example, it allows an easy definition of open balls, whose volume growth is of theoretical interest (the asymmetric case would require 'forward-open' and 'backward-open' balls). However, practical data are often characterized by lack of symmetry, e.g. inhibitory/excitatory effects in neurons, unidirectional gene regulation in cell signalling pathways and human migration between countries. The asymmetric case is of primary interest in the current work.

From now on, we reserve the term *network* for network datasets that cannot necessarily be represented as metric spaces, unless qualified otherwise. An illustration is provided in Fig. 1. Already in [15], it was observed that numerical computation of GW distances worked well for comparing graph-structured data even when the underlying datasets failed to be metric. This observation was further developed in [24], where the focus from the outset was to compute the GW distance (and related discrepancies) between arbitrary matrices, i.e. what we refer to as finite networks. While the experiments of [24] were on symmetric datasets, their implementations remain valid and theoretically justified even on matrices that do not satisfy symmetry. We emphasize this point in the current work, and extend from matrices to the continuous setting. Thus this work should be viewed as a theoretical complement to [24].

On the computational front, the authors of [24] directly attacked the non-convex optimization problem by considering an *entropy-regularized* form of the GW distance (ERGW) following [31], and using a projected gradient descent algorithm based on results in [3,31]. This approach was also used (for a generalized GW distance) on graph-structured datasets in [35]. It was pointed out in [35] that the ERGW approach occasionally requires a large amount of regularization to obtain convergence, and that this could possibly lead to over-regularized solutions. A different approach, developed in [20,21], considers the use of lower bounds on the GW distance as opposed to solving the full GW optimization

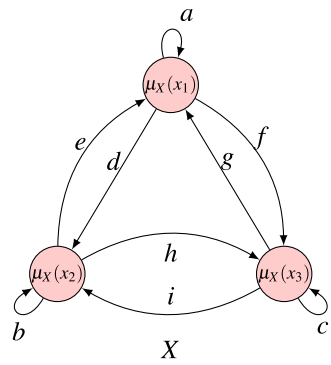


FIG. 1. The networks in this work have asymmetric pairwise weights and a significance value for each node.

problem. This is a practical approach for many use cases, in which it may be sufficient to simply obtain lower bounds for the GW distance. One of the lower bounds in [20] involved linearizing the GW objective by decoupling the alignment term into two separate terms (thus removing the quadratic dependence), and optimizing over each term separately (referred to as the *Third Lower Bound* (TLB)). This approach was also used in [27], with a further relaxation of one of the marginal terms.

As a complement to the alternate minimization scheme of [21] and the ERGW scheme of [24], our numerical experiments are carried out using the lower bound approach, specifically the (TLB). This is certainly faster than alternate minimization (see [15] for computational aspects), but potentially slower than the ERGW scheme of [24]. However, it has the benefit of not needing any parameter tuning, which is an issue with entropic regularization. This makes it useful for exploratory network data analysis.

1.2 *Contributions*

We adopt the setting of networks (X, ω_X, μ_X), where X is a Polish space, μ_X is a Borel probability measure and ω_X is any measurable, integrable function on $X \times X$ (decoupled from the topology of X beyond Borel measurability). Using the GW distance formulation, we define and develop a metric structure on the 'space of networks'. The crux of this construction is that many of the critical theoretical developments in [20,21,33] rely on measure-theoretic properties and not metric properties, hence they extend to the ambient space of networks. Certain interpretations and results cannot carry over: typically these are the statements involving volumes of open balls, which are hard to define in the asymmetric setting. The main algorithms of [24,31] for computing local minima of the ERGW objective do carry over to the network setting.

To complement these algorithms, we adapt ideas from [20,21] to obtain network invariants/features that yield a hierarchy of lower bounds on the network GW distance. The lower bounds arise from satisfying a certain stability property, and are computed by solving (at most) a linear program. In experiments, we focus particularly on the (TLB) from [20].

We strengthen some of the inequalities in the lower bound hierarchy to equalities (Theorem 3.1). As a consequence, we see that the (TLB), which involves solving an ensemble of OT problems over a Polish space $X \times Y$, can be computed by solving OT problems over $\mathbb{R}$ ($\mathbb{R}$-TLB). These can be directly computed via closed-form solutions.

We also define a network Gromov–Prokhorov (GP) distance, propose a new class of invariants (the 'sublevel/superlevel size functions') and use the GP distance to show that these new invariants

satisfy a notion of *interleaving* stability typically arising in the field of applied topology. We exhibit the theoretical utility of these invariants by using them to distinguish between spheres of different dimensions.

Finally, we illustrate our constructions on some highly asymmetric networks (both simulated and real). Our code and datasets are available on https://github.com/samirchowdhury/GWnets.

1.3 *Organization of the paper*

In the following section, we define some notation and terms that will be used throughout the paper. §2 contains details about couplings and the network GW and GP distances. In §3 we present network invariants along with stability results. We conclude with experiments in §4. Appendix A contains additional notes on computations.

1.4 *Notation and basic terminology*

We write $\mathbb{R}_+$ to denote the non-negative reals. The indicator function of a set S is denoted $\mathbf{1}_S$. Given a topological space X (always a Polish space in this paper, and always equipped with the Borel σ-field Borel(X)), we will write Prob(X) to denote the collection of Borel probability measures on X. The *support* of $\mu_X \in \text{Prob}(X)$, denoted supp(μ_X) (or supp(X) when the context is clear), is the set of $x \in X$ such that every open neighbourhood of x has positive measure. Unless specified otherwise, we will always deal with fully supported measures. The Lebesgue measure on the reals will be denoted by $\mathscr{L}$.

The product σ-field on $X \times Y$, denoted Borel($X \times Y$), is defined as the σ-field generated by the measurable rectangles $A \times B$, where $A \in \text{Borel}(X)$ and $B \in \text{Borel}(Y)$. The product measure $\mu_X \otimes \mu_Y$ is defined on the measurable rectangles by writing

$$\mu_X \otimes \mu_Y(A \times B) := \mu_X(A)\mu_X(B), \text{ for all } A \in \text{Borel}(X) \text{ and for all } B \in \text{Borel}(Y).$$

Given a Borel space (X, μ_X) and a Borel measurable function $f : X \to \mathbb{R}$, we write $\|f\|_p := (\int |f|^p \, d\mu_X)^{1/p}$ for $p \in [1, \infty)$, and $\|f\|_\infty := \inf\{M \in [0, \infty] : \mu_X(|f| > M) = 0\}$ for $p = \infty$. For each $p \in [1, \infty]$, $L^p = L^p(\mu_X)$ consists of the Borel measurable functions f with $\|f\|_p < \infty$.

Given a measurable real-valued function $f : X \to \mathbb{R}$ and $t \in \mathbb{R}$, we will occasionally write $\{f \leq t\}$ to denote the set $\{x \in X : f(x) \leq t\}$.

Given (X, μ_X), Y and a Borel-measurable map $f : X \to Y$, the pushforward of μ_X via f is the measure defined by $f_*\mu(A) := \mu(f^{-1}[A])$ for any measurable subset of Y.

2. The structure of measure networks

We will always assume that our measures are fully supported, unless explicitly said otherwise.

2.1 *Networks and isomorphism*

DEFINITION 2.1 A *(measure) network* is a triple (X, ω_X, μ_X) where X is Polish, μ_X is a fully supported Borel probability measure and ω_X is a bounded measurable function on X^2. The naming convention arises from the case when X is finite; in such a case, we can view the pair (X, ω_X) as a complete directed graph with asymmetric real-valued edge weights that is further equipped with node significance values given by μ_X, cf. Fig. 1. Accordingly, the points of X are called *nodes*, pairs of nodes are called *edges* and ω_X is called the *edge weight function* of X. The collection of all measure networks will be denoted $\mathscr{N}$.

REMARK 2.1 (Network data). A large class of objects—including metric spaces, manifolds (Riemannian or Finslerian) and similarity/kernel matrices [24]—can be viewed as networks. Network datasets arising in the sciences typically satisfy the regularity assumptions needed to fit the preceding definition.

We point out one caveat: network datasets in the wild are often incomplete, i.e. ω_X is not fully defined on $X \times X$. In such cases, one needs to preprocess the data (see e.g. [18]) to make it fit within our framework. In many other use cases, however, network datasets are complete by construction. For example, in gene regulatory network inference [25], the only data that can be measured are gene expression levels. In the corresponding network, the nodes are genes and the edge weights are gene dependencies that are *estimated* from the expression levels. The resulting edge weight function is thus completely determined.

REMARK 2.2 Sturm has studied symmetric, L^2 versions of measure networks (called *gauged measure spaces*) in [33], and we point to his work as an excellent reference on the geometry of such spaces. Our motivation comes from studying network datasets, hence the difference in our naming conventions.

When defining any type of distance between networks, as we eventually will, it is necessary to first decide which networks should be viewed as being at 0-distance. We make these choices now. The information contained in a network should be preserved under relabelling. Additionally, if a node is split into multiple nodes, all with the same incoming and outgoing edge weights, the information in the network remains unchanged. Conversely, if multiple nodes have the same incoming/outgoing edge weights, then they can be merged together without information loss. We formalize these ideas via the following notions of *isomorphism*.

DEFINITION 2.2 (Strong isomorphism). To say $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathscr{N}$ are *strongly isomorphic* means that there exists a Borel measurable bijection $\varphi : X \to Y$ (with Borel measurable inverse φ^{-1}) such that

- $\omega_X(x, x') = \omega_Y(\varphi(x), \varphi(x'))$ for all $x, x' \in X$, and
- $\varphi_* \mu_X = \mu_Y$.

We will denote a strong isomorphism between measure networks by $X \cong^s Y$.

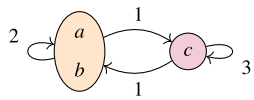
The following definition is a relaxation of strong isomorphism.

DEFINITION 2.3 (Weak isomorphism). $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathscr{N}$ are *weakly isomorphic*, denoted $X \cong^w Y$, if there is a Borel probability space (Z, μ_Z) with measurable maps $f : Z \to X$ and $g : Z \to Y$ such that

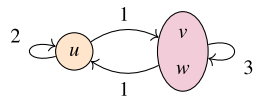
- $f_* \mu_Z = \mu_X$, $g_* \mu_Z = \mu_Y$, and
- $\|f^* \omega_X - g^* \omega_Y\|_\infty = 0$.

Here $f^* \omega_X : Z \times Z \to \mathbb{R}$ is the pullback weight function given by the map $(z, z') \mapsto \omega_X(f(z), f(z'))$. The map $g^* \omega_Y$ is defined analogously. Note that these pullbacks are measurable. Figure 2 provides an illustration.

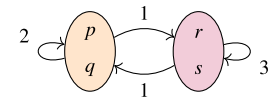
REMARK 2.3 (Interpretation for real data). According to the notion of weak isomorphism, two nodes x, x' are informally the same if they have the same 'internal perception', i.e. $\omega_X(x, x) = \omega_X(x, x') = \omega_X(x', x) = \omega_X(x', x')$, and the same external perception, i.e. all the incoming and outgoing edge weights



$$\omega_X(a,b,c) = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \mu_X = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}$$



$$\omega_Y(u,v,w) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 3 \\ 1 & 3 & 3 \end{pmatrix}, \mu_Y = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$



$$\omega_Z(p,q,r,s) = \begin{pmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 3 \end{pmatrix}, \mu_Z = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

FIG. 2. Weakly isomorphic networks X, Y, Z. Note that Z maps surjectively onto X and Y, and this surjection induces compatible pushforward measures and pullback weight functions, as required by Definition 2.3.

are the same. A relaxation would be to say that x, x' are ε-similar if, for $\varepsilon > 0$,

$$\max(\|\omega_X(x,\cdot) - \omega_X(x',\cdot)\|_\infty, \|\omega_X(\cdot,x) - \omega_X(\cdot,x')\|_\infty) < \varepsilon.$$

The network stochastic block model in §4.2 describes networks that admit partitions into ε-similar blocks.

EXAMPLE 2.1 Networks with one or two nodes will be very instructive in providing examples and counterexamples, so we introduce them now with some special terminology.

- By $N_1(a)$ we will refer to the network with one node $X = \{p\}$, a weight $\omega_X(p,p) = a$ and the Dirac measure $\mu_X = \delta_p$.
- By $N_2\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \alpha, \beta\right)$ we will mean a two-node network with node set $X = \{p,q\}$, and weights and measures given as follows:

$$\begin{aligned} \omega_X(p,p) &= a & \mu_X(\{p\}) &= \alpha \\ \omega_X(p,q) &= b & \mu_X(\{q\}) &= \beta \\ \omega_X(q,p) &= c \\ \omega_X(q,q) &= d. \end{aligned}$$

- Given a k-by-k matrix $\Sigma \in \mathbb{R}^{k\times k}$ and a $k \times 1$ vector $v \in \mathbb{R}^k_+$ with sum 1, we automatically obtain a network on k nodes that we denote as $N_k(\Sigma, v)$. Notice that $N_k(\Sigma, v) \cong^s N_\ell(\Sigma', v')$ if and only if $k = \ell$, and there exists a permutation matrix P of size k such that $\Sigma' = P\,\Sigma\,P^T$ and $Pv = v'$.

Notation. Even though μ_X takes sets as its argument, we will often omit the curly braces and use $\mu_X(p,q,r)$ to mean $\mu_X(\{p,q,r\})$.

We wish to define a notion of distance on $\mathscr{N}$ that is compatible with isomorphism. A natural analogue is the GW distance defined between metric measure spaces [**20**]. To adapt that definition for our needs, we first recall the definition of a measure coupling.

2.2 *Couplings and the distortion functional*

Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ be two measure networks. A *coupling* between these two networks is a probability measure μ on $X \times Y$ with marginals μ_X and μ_Y, respectively. Stated differently, couplings

satisfy the following property:

$$\mu(A \times Y) = \mu_X(A) \text{ and } \mu(X \times B) = \mu_Y(B), \text{ for all } A \in \text{Borel}(X) \text{ and for all } B \in \text{Borel}(Y).$$

The collection of all couplings between (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) will be denoted $\mathscr{C}(\mu_X, \mu_Y)$, abbreviated to $\mathscr{C}$ when the context is clear. Couplings are also referred to as *transport plans*.

EXAMPLE 2.2 (Product coupling). Let (X, ω_X, μ_X), $(Y, \omega_Y, \mu_Y) \in \mathscr{N}$. The set $\mathscr{C}(\mu_X, \mu_Y)$ is always non-empty, because the *product measure* $\mu := \mu_X \otimes \mu_Y$ is always a coupling between μ_X and μ_Y.

EXAMPLE 2.3 (1-point coupling). Let $(X, \omega_X, \mu_X) \in \mathscr{N}$, and let $Y = N_1(a)$ be a network on a single point $\{p\}$. Then there exists a unique coupling $\mu = \mu_X \otimes \delta_p$ between μ_X and δ_p.

EXAMPLE 2.4 (Diagonal coupling). Let $(X, \omega_X, \mu_X) \in \mathscr{N}$. The *diagonal coupling* Δ between μ_X and itself is the transport plan that sends each point to itself, and is defined by writing

$$\Delta(A \times B) := \int_X \mathbf{1}_{A \times B}(x, x)\, \mathrm{d}\mu_X(x) \qquad \text{for all } A, B \in \text{Borel}(X).$$

To see that this is a coupling, let $A \in \text{Borel}(X)$. Then,

$$\Delta(A \times X) = \int_X \mathbf{1}_{A \times X}(x, x)\, \mathrm{d}\mu_X(x) = \int_X \mathbf{1}_A(x)\, \mathrm{d}\mu_X(x) = \mu_X(A),$$

and similarly $\Delta(X \times A) = \mu_X(A)$. Thus $\Delta \in \mathscr{C}(\mu_X, \mu_X)$.

Now we turn to the notion of the *distortion* of a coupling. Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ be two measure networks. Next let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$, and consider the probability space $(X \times Y)^2$ equipped with the product measure $\mu \otimes \mu$. For each $p \in [1, \infty]$ the p-distortion of μ is defined as $\text{dis}_p(\mu) := \left\| \omega_X - \omega_Y \right\|_p$. For $p \in [1, \infty)$, this is written as:

$$\text{dis}_p(\mu) = \left(\int_{X \times Y} \int_{X \times Y} |\omega_X(x, x') - \omega_Y(y, y')|^p \, \mathrm{d}\mu(x, y)\, \mathrm{d}\mu(x', y') \right)^{1/p}.$$

For $p = \infty$, this becomes:

$$\text{dis}_p(\mu) := \text{ess sup} |\omega_X - \omega_Y|.$$

We end by introducing the Wasserstein distance [2, §7], which metrizes the topology of narrow convergence that we introduce below. Let (X, d_X) be a Polish space, let $p \in [1, \infty]$ and let $\mu, \nu \in \text{Prob}(X)$ be such that $\left\| d_X(x_0, \cdot) \right\|_{L^p(\tau)} < \infty$ for $\tau \in \{\mu, \nu\}$ and some $x_0 \in X$. The *pth Wasserstein distance* between μ, ν is defined to be:

$$W_p(\mu, \nu) := \inf_{\tau \in \mathscr{C}(\mu, \nu)} \left\| d_X \right\|_{L^p(\tau)}.$$

2.3 *Optimality of couplings in the network setting*

We now collect some results about probability spaces. Let X be a Polish space. A subset $P \subseteq \text{Prob}(X)$ is said to be *tight* if for all $\varepsilon > 0$, there is a compact subset $K_\varepsilon \subseteq X$ such that $\mu_X(X \setminus K_\varepsilon) \leq \varepsilon$ for all $\mu_X \in P$.

A sequence $(\mu_n)_{n\in\mathbb{N}} \in \mathrm{Prob}(X)^{\mathbb{N}}$ is said to *converge narrowly* to $\mu_X \in \mathrm{Prob}(X)$ if

$$\lim_{n\to\infty} \int_X f \, \mathrm{d}\mu_n = \int_X f \, \mathrm{d}\mu_X \qquad \text{for all } f \in C_b(X),$$

the space of continuous, bounded, real-valued functions on X. Narrow convergence is induced by a distance [2, Remark 5.1.1], in particular by W_p when X is bounded, hence the convergent sequences in $\mathrm{Prob}(X)$ completely determine a topology on $\mathrm{Prob}(X)$. This topology on $\mathrm{Prob}(X)$ is called the *narrow topology*. In some references [33], narrow convergence (resp. narrow topology) is called *weak convergence* (resp. *weak topology*).

A further consequence of having a metric on $\mathrm{Prob}(X)$ [2, Remark 5.1.1] is that singletons are closed. This simple fact will be used below.

THEOREM 2.1 (Prokhorov, [2] Theorem 5.1.3). Let X be a Polish space. Then $P \subseteq \mathrm{Prob}(X)$ is tight if and only if it is relatively compact, i.e. its closure is compact in $\mathrm{Prob}(X)$.

LEMMA 2.1 (Lemma 4.4, [37]). Let X, Y be two Polish spaces, and let $P_X \subseteq \mathrm{Prob}(X)$, $P_Y \subseteq \mathrm{Prob}(Y)$ be tight in their respective spaces. Then the set $\mathscr{C}(P_X, P_Y) \subseteq \mathrm{Prob}(X \times Y)$ of couplings with marginals in P_X and P_Y is tight in $\mathrm{Prob}(X \times Y)$.

LEMMA 2.2 (Compactness of couplings; Lemma 1.2, [33]). Let X, Y be two Polish spaces. Let $\mu_X \in \mathrm{Prob}(X)$, $\mu_Y \in \mathrm{Prob}(Y)$. Then $\mathscr{C}(\mu_X, \mu_Y)$ is compact in $\mathrm{Prob}(X \times Y)$.

Proof. The singletons $\{\mu_X\}$, $\{\mu_Y\}$ are closed and of course compact in $\mathrm{Prob}(X)$, $\mathrm{Prob}(Y)$. Hence by Prokhorov's theorem, they are tight. Now consider $\mathscr{C}(\mu_X, \mu_Y) \subseteq \mathrm{Prob}(X \times Y)$. Since this is obtained by intersecting the preimages of the continuous projections onto the marginals μ_X and μ_Y, we know that it is closed. Furthermore, $\mathscr{C}(\mu_X, \mu_Y)$ is tight by Lemma 2.1. Then by another application of Prokhorov's theorem, it is compact. □

The following lemma appeared for symmetric weight functions in the L^2 case in [33], along with a slightly different proof using parametrizations by the unit interval. The proof is actually simpler in the network setting because we do not need to enforce symmetry of the approximating functions.

LEMMA 2.3 (Continuity of the distortion functional). Let $1 \leq p < \infty$, and let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathscr{N}$. The distortion functional dis_p is continuous on $\mathscr{C}(\mu_X, \mu_Y)$. For $p = \infty$, dis_∞ is lower semicontinuous.

Proof. First suppose $p \in [1, \infty)$. We will construct a sequence of continuous functionals that converges uniformly to dis_p. Since the uniform limit of continuous functions is continuous, this will show that dis_p is continuous.

Bounded continuous functions are dense in L^p (in our setting of Polish spaces with finite measures, see e.g. [11, §7.2]), so for each $n \in \mathbb{N}$, we pick continuous, bounded functions $\omega_X^n \in L^p(\mu_X^{\otimes 2})$ and $\omega_Y^n \in L^p(\mu_Y^{\otimes 2})$ such that

$$\left\|\omega_X - \omega_X^n\right\|_{L^p(\mu_X \otimes \mu_X)} \leq 1/n, \qquad \left\|\omega_Y - \omega_Y^n\right\|_{L^p(\mu_Y \otimes \mu_Y)} \leq 1/n.$$

For each $n \in \mathbb{N}$, define the functional $\mathrm{dis}_p^n : \mathscr{C}(\mu_X, \mu_Y) \to \mathbb{R}_+$ by $\mathrm{dis}_p^n(\nu) := \left\|\omega_X^n - \omega_Y^n\right\|_{L^p(\nu \otimes \nu)}$. Note that $\left|\omega_X^n - \omega_Y^n\right|^p \in C_b((X \times Y)^2)$.

We claim that dis_p^n is continuous. Since the narrow topology on $\mathrm{Prob}(X \times Y)$ is induced by a distance [2, Remark 5.1.1], it suffices to show sequential continuity. Let $\nu \in \mathscr{C}(\mu_X, \mu_Y)$, and let $(\nu_m)_{m\in\mathbb{N}}$ be a

sequence in $\mathscr{C}(\mu_X, \mu_Y)$ converging narrowly to ν. Then in fact $\nu_m \otimes \nu_m$ converges narrowly to $\nu \otimes \nu$ [4, Theorem 2.1]. Thus we have

$$\lim_{m\to\infty} \mathrm{dis}_p^n(\nu_m) = \lim_{m\to\infty} \left(\int_{(X\times Y)^2} \left|\omega_X^n - \omega_Y^n\right|^p \mathrm{d}\nu_m \otimes \mathrm{d}\nu_m \right)^{1/p}$$
$$= \left(\int_{(X\times Y)^2} \left|\omega_X^n - \omega_Y^n\right|^p \mathrm{d}\nu \otimes \mathrm{d}\nu \right)^{1/p} = \mathrm{dis}_p^n(\nu).$$

Here the second equality follows from the definition of convergence in the narrow topology and the fact that the integrand is bounded and continuous. This shows sequential continuity (hence continuity) of dis_p^n.

Finally, we show that $(\mathrm{dis}_p^n)_{n\in\mathbb{N}}$ converges to dis_p uniformly. Let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$. Then,

$$\left|\mathrm{dis}_p(\mu) - \mathrm{dis}_p^n(\mu)\right| = \left| \left\|\omega_X - \omega_Y\right\|_{L^p(\nu\otimes\nu)} - \left\|\omega_X^n - \omega_Y^n\right\|_{L^p(\nu\otimes\nu)} \right|$$
$$\leq \left\|\omega_X - \omega_X^n\right\|_{L^p(\mu_X\otimes\mu_X)} + \left\|\omega_Y - \omega_Y^n\right\|_{L^p(\mu_Y\otimes\mu_Y)} \leq 2/n.$$

But $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ was arbitrary. This shows that dis_p is the uniform limit of continuous functions, hence is continuous. Here the first inequality followed from Minkowski's inequality.

Now suppose $p = \infty$. Let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ be arbitrary. Recall that because we are working over probability spaces, Jensen's inequality can be used to show that for any $1 \leq q \leq r < \infty$, we have $\mathrm{dis}_q(\mu) \leq \mathrm{dis}_r(\mu)$. Moreover, we have $\lim_{q\to\infty} \mathrm{dis}_q(\mu) = \mathrm{dis}_\infty(\mu)$. The supremum of a family of continuous functions is lower semicontinuous. In our case, $\mathrm{dis}_\infty = \sup\{\mathrm{dis}_q : q \in [1,\infty)\}$, and we have shown above that all the functions in this family are continuous. Hence dis_∞ is lower semicontinuous. □

Definition 2.1 (Optimal couplings). Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathscr{N}$, and let $p \in [1,\infty]$. A coupling $\mu \in \mathscr{C}(\mu_X, \mu_Y)$ is *optimal* if $\mathrm{dis}_p(\mu) = \inf_{\nu\in\mathscr{C}(\mu_X,\mu_Y)} \mathrm{dis}_p(\nu)$.

Theorem 2.2 Let (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) be two measure networks, and let $p \in [1,\infty]$. Then there exists an optimal coupling, i.e. a minimizer for $\mathrm{dis}_p(\cdot)$ in $\mathscr{C}(\mu_X, \mu_Y)$.

Proof. The result follows from Lemmas 2.2 and 2.3, because lower semicontinuity and compactness are sufficient to guarantee that dis_p achieves its infimum on $\mathscr{C}(\mu_X, \mu_Y)$, for any $p \in [1,\infty]$. □

2.4 *The network GW distance*

For each $p \in [1,\infty]$, we define:

$$d_{\mathscr{N},p}(X,Y) := \frac{1}{2} \min_{\mu\in\mathscr{C}(\mu_X,\mu_Y)} \mathrm{dis}_p(\mu) \qquad \text{for each } (X,\omega_X,\mu_X), (Y,\omega_Y,\mu_Y) \in \mathscr{N}.$$

Here we implicitly use Theorem 2.2 to write min instead of inf. As we will see below, $d_{\mathscr{N},p}$ is a legitimate pseudometric on $\mathscr{N}$. The structure of $d_{\mathscr{N},p}$ is analogous to a formulation of the *GW distance* between metric measure spaces [21,33].

REMARK 2.4 (Boundedness of $d_{\mathcal{N},p}$). Recall from Example 2.2 that for any $X, Y \in \mathcal{N}$, $\mathscr{C}(\mu_X, \mu_Y)$ always contains the product coupling, and is thus non-empty. A consequence is that $d_{\mathcal{N},p}(X, Y)$ is bounded for any $p \in [1, \infty]$. Indeed, by taking the product coupling $\mu := \mu_X \otimes \mu_Y$ we have $d_{\mathcal{N},p}(X, Y) \leq \frac{1}{2} \operatorname{dis}_p(\mu) < \infty$.

In some simple cases, we obtain explicit formulas for computing $d_{\mathcal{N},p}$.

EXAMPLE 2.5 (Easy examples of $d_{\mathcal{N},p}$). Let $a, b \in \mathbb{R}$ and consider the networks $N_1(a)$ and $N_1(b)$. The unique coupling between the two networks is the product measure $\mu = \delta_x \otimes \delta_y$, where we understand x, y to be the nodes of the two networks. Then for any $p \in [1, \infty]$, we obtain:

$$d_{\mathcal{N},p}(N_1(a), N_1(b)) = \frac{1}{2} \operatorname{dis}_p(\mu) = |\omega_{N_1(a)}(x, x) - \omega_{N_1(b)}(y, y)| = |a - b|.$$

Let $(X, \omega_X, \mu_X) \in \mathcal{N}$ be any network and let $N_1(a) = (\{y\}, a)$ be a network with one node. Once again, there is a unique coupling $\mu = \mu_X \otimes \delta_y$ between the two networks. For any $p \in [1, \infty)$, we obtain:

$$d_{\mathcal{N},p}(X, N_1(a)) = \frac{1}{2} \operatorname{dis}_p(\mu) = \frac{1}{2} \left(\int_X \int_X \left| \omega_X(x, x') - a \right|^p \mathrm{d}\mu_X(x)\, \mathrm{d}\mu_X(x') \right)^{1/p}.$$

For $p = \infty$, we have $d_{\mathcal{N},p}(X, N_1(a)) = \operatorname{ess\,sup} \left(\frac{1}{2} |\omega_X - a| \right)$.

REMARK 2.5 $d_{\mathcal{N},p}$ is not necessarily a metric modulo strong isomorphism. This can be seen from Fig. 2.

The definition of $d_{\mathcal{N},p}$ is sensible in the sense that it captures the notion of a distance:

THEOREM 2.3 For each $p \in [1, \infty]$, $d_{\mathcal{N},p}$ is a pseudometric on $\mathcal{N}$.

Proof of Theorem 2.3 Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y), (Z, \omega_Y, \mu_Y) \in \mathcal{N}$. It is clear that $d_{\mathcal{N},p}(X, Y) \geq 0$. Taking the diagonal coupling (see Example 2.4) shows $d_{\mathcal{N},p}(X, X) = 0$. For symmetry, notice that for any $\mu \in \mathscr{C}(\mu_X, \mu_Y)$, we can define $\widetilde{\mu} := f_* \mu$, where $f : X \times Y \to Y \times X$ is the map $(x, y) \mapsto (y, x)$. Then $\operatorname{dis}_p(\mu) = \operatorname{dis}_p(\widetilde{\mu})$, and this will show $d_{\mathcal{N},p}(X, Y) = d_{\mathcal{N},p}(Y, X)$. Note that we are overloading notation here: there are implicitly two dis_p functions, with different domains, for μ and $\widetilde{\mu}$, respectively.

Finally, we need to check the triangle inequality. Let $\mu_{12} \in \mathscr{C}(\mu_X, \mu_Y)$ and $\mu_{23} \in \mathscr{C}(\mu_Y, \mu_Z)$ be couplings such that $2d_{\mathcal{N},p}(X, Y) = \operatorname{dis}_p(\mu_{12})$ and $2d_{\mathcal{N},p}(Y, Z) = \operatorname{dis}_p(\mu_{23})$ (using Theorem 2.2). By the standard gluing lemma (Lemma 1.4 in [33], also Lemma 7.6 in [36]), we obtain a probability measure $\mu \in \operatorname{Prob}(X \times Y \times Z)$ with marginals μ_{12}, μ_{23}, and a marginal μ_{13} that is a coupling between μ_X and μ_Z. This coupling is not necessarily optimal. Then we have:

$$\begin{aligned}
2d_{\mathcal{N},p}(X, Z) &\leq \operatorname{dis}_p(\mu_{13}) \\
&= \left\| \omega_X - \omega_Y + \omega_Y - \omega_Z \right\|_{L^p(\mu \otimes \mu)} \\
&\leq \left\| \omega_X - \omega_Y \right\|_{L^p(\mu \otimes \mu)} + \left\| \omega_Y - \omega_Z \right\|_{L^p(\mu \otimes \mu)} \\
&= \left\| \omega_X - \omega_Y \right\|_{L^p(\mu_{12} \otimes \mu_{12})} + \left\| \omega_Y - \omega_Z \right\|_{L^p(\mu_{23} \otimes \mu_{23})} = 2d_{\mathcal{N},p}(X, Y) + 2d_{\mathcal{N},p}(Y, Z).
\end{aligned}$$

The second inequality above follows from Minkowski's inequality. This proves the triangle inequality. □

REMARK 2.6 This result and its proof are analogous to the related results for gauged and metric measure spaces [21,33]. The observation here is that the metric structure on $\mathscr{N}$ is not inherited from its elements, but is rather enforced by the structure of $d_{\mathscr{N},p}$. This is in contrast, for example, to the Wasserstein distance W_p, which inherits its metric structure from an underlying metric space.

It remains to discuss the precise pseudometric structure of $d_{\mathscr{N},p}$. The following result is analogous to a statement about *homomorphisms* in [33]; again, the proof is purely measure-theoretic and hence applies to the asymmetric setting.

THEOREM 2.4 (Pseudometric structure of $d_{\mathscr{N},p}$). Let $(X,\omega_X,\mu_X),(Y,\omega_Y,\mu_Y)\in\mathscr{N}$, and let $p\in[1,\infty]$. Then $d_{\mathscr{N},p}(X,Y)=0$ if and only if $X\cong^w Y$.

Proof of Theorem 2.4 Fix $p\in[1,\infty)$. For the backward direction, suppose there exist (Z,μ_Z) and measurable maps $f:Z\to X$ and $g:Z\to Y$ satisfying the conditions of Definition 2.3. Let $\mu:=(f,g)_*\mu_Z$. Then $\mu\in\mathscr{C}(\mu_X,\mu_Y)$, and we have:

$$2d_{\mathscr{N},p}(X,Y)\leq\left(\int_{(X\times Y)^2}\left|\omega_X-\omega_Y\right|^p\,\mathrm{d}\mu\,\mathrm{d}\mu\right)^{1/p}=\left(\int_{Z^2}\left|f^*\omega_X-g^*\omega_Y\right|^p\,\mathrm{d}\mu_Z\,\mathrm{d}\mu_Z\right)^{1/p}=0.$$

Here the first equality is by the change of variables formula. The case $p=\infty$ is similar.

For the forward direction, let $\mu\in\mathscr{C}(\mu_X,\mu_Y)$ be an optimal coupling with $\mathrm{dis}_p(\mu)=0$ (Theorem 2.2). Define $Z:=X\times Y$, $\mu_Z:=\mu$. Then the projection maps $\pi_X:Z\to X$ and $\pi_Y:Z\to Y$ are measurable. We also have $(\pi_X)_*\mu=\mu_X$ and $(\pi_Y)_*\mu=\mu_Y$. Since $\mathrm{dis}_p(\mu)=0$, we also have $\|(\pi_X)^*\omega_X-(\pi_Y)^*\omega_Y\|_\infty=\|\omega_X-\omega_Y\|_\infty=0$.

The $p=\infty$ case is proved analogously. This concludes the proof. □

REMARK 2.7 A result analogous to Theorem 2.4 holds for networks without measure equipped with a Gromov–Hausdorff-type network distance [7]. The 'tripod structure' $X\leftarrow Z\to Y$ described above is much more difficult to obtain in the setting of [7]. This highlights an advantage of the measure-theoretic setting of the current paper.

2.5 *Additional constructions*

We briefly digress to discuss some theoretical connections to the framework presented above. The first of these is the notion of parametrization, which is used in the setting of mm-spaces to define Gromov's box distance [13]. The second is an explicit development of an alternative distance between networks based on the GP distance between mm-spaces [12]. This in turn leads to interesting and novel lower bounds on the $d_{\mathscr{N},\infty}$-distance between spheres (see §3.2.1).

2.5.1 *Interval representation*

We now record a standard result about mm-spaces that remains valid in the network setting. Let $(X,\omega_X,\mu_X)\in\mathscr{N}$. Because X is Polish and $\mu_X(X)=1$, the pair (X,μ_X) admits a *parameter*, i.e. a (not necessarily unique) surjective Borel-measurable map $\rho:I=[0,1]\to X$ such that $\rho_*\mathscr{L}=\mu_X$ [28, Lemma 4.2]. Here $\mathscr{L}$ denotes Lebesgue measure. By pulling back ω_X, we get a triple $(I,\rho^*\omega_X,\mathscr{L})\in\mathscr{N}$. Note that by construction, (X,ω_X,μ_X) is weakly isomorphic to its parameter.

Parametrizations allow one to define a version of Gromov's box distance [13] for networks. Computing the box distance leads to difficult combinatorial problems and is not the focus of this paper, but we point to it as a source of interesting theoretical problems.

In parametrized form, a network is a measurable, integrable function on the unit square. If the edge weight function is normalized and centred to be in $[0,1]$, then a network corresponds to a *graphon* [19].

2.5.2 *The network GP distance* We now formulate a network distance analogous to the *GP* distance between mm-spaces [12]. This will be used to prove subsequent results.

Let $\alpha \in [0,\infty)$. For any (X,ω_X,μ_X), $(Y,\omega_Y,\mu_Y) \in \mathscr{N}$, we write $\mathscr{C} := \mathscr{C}(\mu_X,\mu_Y)$ and define:

$$d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y) := \frac{1}{2}\inf_{\mu\in\mathscr{C}} \inf\left\{\varepsilon > 0 : \mu\otimes\mu\left(\{x,y,x',y' \in (X\times Y)^2 : \left|\omega_X(x,x') - \omega_Y(y,y')\right| \geq \varepsilon\}\right) \leq \alpha\varepsilon\right\}.$$

THEOREM 2.5 For each $\alpha \in [0,\infty)$, $d_{\mathscr{N},\alpha}^{\mathscr{GP}}$ is a pseudometric on $\mathscr{N}$.

Proof. Let $(X,\omega_X,\mu_X),(Y,\omega_Y,\mu_Y),(Z,\omega_Z,\mu_Z) \in \mathscr{N}$. The proofs that $d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y) \geq 0$, $d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,X) = 0$, and that $d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y) = d_{\mathscr{N},\alpha}^{\mathscr{GP}}(Y,X)$ are analogous to those used in Theorem 2.3. Hence we only check the triangle inequality. Let $\varepsilon_{XY} > 2d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y)$, $\varepsilon_{YZ} > 2d_{\mathscr{N},\alpha}^{\mathscr{GP}}(Y,Z)$, and let μ_{XY},μ_{YZ} be couplings such that

$$\mu_{XY}^{\otimes 2}\left(\left\{(x,y,x',y') : \left|\omega_X(x,x') - \omega_Y(y,y')\right| \geq \varepsilon_{XY}\right\}\right) \leq \alpha\varepsilon_{XY},$$
$$\mu_{YZ}^{\otimes 2}\left(\left\{(y,z,y',z') : \left|\omega_Y(y,y') - \omega_Z(z,z')\right| \geq \varepsilon_{YZ}\right\}\right) \leq \alpha\varepsilon_{YZ}.$$

For convenience, define:

$$A := \left\{((x,y,z),(x',y',z')) \in (X\times Y\times Z)^2 : \left|\omega_X(x,x') - \omega_Y(y,y')\right| \geq \varepsilon_{XY}\right\}$$
$$B := \left\{((x,y,z),(x',y',z')) \in (X\times Y\times Z)^2 : \left|\omega_Y(y,y') - \omega_Z(z,z')\right| \geq \varepsilon_{YZ}\right\}$$
$$C := \left\{((x,y,z),(x',y',z')) \in (X\times Y\times Z)^2 : \left|\omega_X(x,x') - \omega_Z(z,z')\right| \geq \varepsilon_{XY}+\varepsilon_{YZ}\right\}.$$

Next let μ denote the probability measure with marginals μ_{XY},μ_{YZ}, and a marginal $\mu_{XZ} \in \mathscr{C}(\mu_X,\mu_Z)$ obtained from gluing μ_{XY} and μ_{YZ} (cf. Lemma 7.6 in [36]). We need to show:

$$\mu_{XZ}^{\otimes 2}\left((\pi_X,\pi_Z)(C)\right) \leq \alpha(\varepsilon_{XY}+\varepsilon_{YZ}).$$

To show this, it suffices to show $C \subseteq A\cup B$, because then we have $\mu^{\otimes 2}(C) \leq \mu^{\otimes 2}(A) + \mu^{\otimes 2}(B)$ and consequently

$$\mu_{XZ}^{\otimes 2}\left((\pi_X,\pi_Z)(C)\right) = \mu^{\otimes 2}(C) \leq \mu^{\otimes 2}(A) + \mu^{\otimes 2}(B) = \mu_{XY}^{\otimes 2}\left((\pi_X,\pi_Y)(A)\right) + \mu_{YZ}^{\otimes 2}\left((\pi_Y,\pi_Z)(B)\right)$$
$$\leq \alpha(\varepsilon_{XY}+\varepsilon_{YZ}).$$

Let $((x,y,z),(x',y',z')) \in (X\times Y\times Z)^2 \setminus (A\cup B)$. Then we have

$$\left|\omega_X(x,x') - \omega_Y(y,y')\right| < \varepsilon_{XY} \text{ and } \left|\omega_Y(y,y') - \omega_Z(z,z')\right| < \varepsilon_{YZ}.$$

By the triangle inequality, we then have:

$$\left|\omega_X(x,x') - \omega_Z(z,z')\right| \leq \left|\omega_X(x,x') - \omega_Y(y,y')\right| + \left|\omega_Y(y,y') - \omega_Z(z,z')\right| < \varepsilon_{XY}+\varepsilon_{YZ}.$$

Thus $((x,y,z),(x',y',z')) \in (X\times Y\times Z)^2 \setminus C$. This shows $C \subseteq A\cup B$.

The preceding work shows that $2d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Z) \leq \varepsilon_{XY}+\varepsilon_{YZ}$. Since $\varepsilon_{XY} > 2d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y)$ and $\varepsilon_{YZ} > 2d_{\mathscr{N},\alpha}^{\mathscr{GP}}(Y,Z)$ were arbitrary, it follows that $d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Z) \leq d_{\mathscr{N},\alpha}^{\mathscr{GP}}(X,Y) + d_{\mathscr{N},\alpha}^{\mathscr{GP}}(Y,Z)$. □

The next lemma follows by unpacking the definitions of the GP and GW distances.

LEMMA 2.4 (Relation between GP and GW). Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathcal{N}$. We always have:

$$d_{\mathcal{N},0}^{\mathcal{GP}}(X, Y) = d_{\mathcal{N},\infty}(X, Y).$$

3. Invariants and lower bounds

As already remarked, finite networks can be regarded as square matrices equipped with a probability measure on the columns (equivalently, the rows). This was the setting of [24]. Theorem 2.3 completes the theoretical justification behind using the GW distance to compare matrices, as carried out (at least for symmetric matrices) in [24].

We now study a variety of network invariants, which can also be thought of as network features. Informally, a network invariant is a compressed representation of the network satisfying the following compatibility property: if two networks are the same in the sense of $d_{\mathcal{N},p}$, then their invariants should also be 'the same'. The invariants we consider are functions $\iota : (\mathcal{N}, d_{\mathcal{N},p}) \to (\mathcal{I}, d_{\mathcal{I}})$, where $(\mathcal{I}, d_{\mathcal{I}})$ is some pseudometric space. Such invariants translate the original problem of computing GW over the 'space of networks' to computing $d_{\mathcal{I}}$ over spaces $\mathcal{I}$ with more regular geometry, e.g. the real line. Translating the problem to a simpler space is done in a controlled manner. One such form of control is *Lipschitz stability*: an invariant ι is *Lipschitz-stable* if there exists a Lipschitz constant L_ι such that

$$d_{\mathcal{I}}(\iota(X), \iota(Y)) \leq L_\iota \cdot d_{\mathcal{N},p}(X, Y). \qquad \text{(for all } X, Y \in \mathcal{N}.)$$

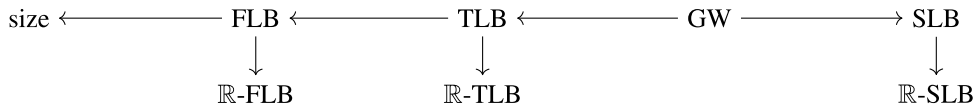
In Section 3.1, we present Lipschitz-stable invariants. In Section 3.2, we present a different notion of control that we refer to as *interleaving stability* as well as associated invariants.

REMARK 3.1 Asymmetry arises in a significant way in this section: for most of our network invariants, we obtain 'outgoing' and 'incoming' versions, based on our choice of functions $\omega_X(x, \cdot)$ or $\omega_X(\cdot, x)$. The network interpretation can be framed in terms of hubs (nodes with high outgoing edge weights) and authorities (nodes with high incoming edge weights) [17].

3.1 *A hierarchy of lower bounds for* $d_{\mathcal{N},p}$

Following [20,21], we now produce a hierarchy of lower bounds for $d_{\mathcal{N},p}$, namely the First Lower Bound (FLB), Second Lower Bound (SLB) and TLB. These are obtained by linearizing the GW objective and/or pushing forward the problem into the real line. Each of these bounds itself has an associated pushforward into the real line, which we denote by adding a prefix $\mathbb{R}$-. Because we are in the asymmetric setting, the FLB, TLB and their $\mathbb{R}$-versions decouple into 'incoming' and 'outgoing' versions. In Remark 3.4, we will show (using Theorem 3.1) that these lower bounds are in fact obtained as Lipschitz stability conditions on certain network invariants.

The hierarchy is illustrated in the following diagram, where the arrows indicate (possibly non-strictly) decreasing complexity.



We now derive these relationships. Let $p \in [1, \infty]$. Start by fixing one pair of coordinates in the $d_{\mathcal{N},p}$ integrand. Then we obtain the *(outgoing) joint eccentricity function* $\mathrm{ecc}^{\mathrm{out}}_{p,X,Y} : X \times Y \to \mathbb{R}_+$ defined by

$$\mathrm{ecc}^{\mathrm{out}}_{p,X,Y}(s,t) := \inf_{\mu \in \mathscr{C}(\mu_X, \mu_Y)} \left\| \omega_X(s,\cdot) - \omega_Y(t,\cdot) \right\|_{L^p(\mu)}, \qquad (s,t) \in X \times Y. \tag{JE}$$

Note that $\omega_X(s,\cdot)$ and $\omega_Y(t,\cdot)$ are both measurable [**11**, Proposition 2.34]. Switching the arguments above produces the *incoming joint eccentricity function*. Taking the norm of an individual term gives the *(outgoing) eccentricity function* $\mathrm{ecc}^{\mathrm{out}}_{p,X} : X \to \mathbb{R}_+$:

$$\mathrm{ecc}^{\mathrm{out}}_{p,X}(s) := \left\| \omega_X(s,\cdot) \right\|_{L^p(\mu_X)}, \qquad s \in X. \tag{E}$$

Flipping the arguments above produces the *incoming eccentricity function*. The norm of the preceding function is the *pth* size *function* $\mathrm{size}_p : \mathcal{N} \to \mathbb{R}_+$:

$$\mathrm{size}_p(X) := \left\| \mathrm{ecc}^{\mathrm{out}}_{p,X} \right\|_{L^p(\mu_X)} = \left\| \omega_X \right\|_{L^p(\mu_X \otimes \mu_X)}. \tag{Sz}$$

The size function is easily seen to be a network invariant: it compresses all the information in a network into a single real number. Theorem 3.1 below shows that this compression occurs in a quantitatively stable manner. Notice that size_p can be computed exactly via a formula, and this computation is extremely cheap. Despite its simplicity, it can be very helpful as a first step in comparing networks. From a procedural perspective, given a network comparison task, one could compute size_p for different networks and compare these values to gain a coarse understanding of the discrepancies between the networks.

A priori, the connections between Equations (E) and (JE) to network invariants are somewhat unclear. We will use Theorem 3.1 to clarify these connections in Remark 3.4, but we present the statements now for convenience. It will turn out that the invariant associated to Equation (E) is the map that takes a network X to the distribution $(\mathrm{ecc}^{\mathrm{out}}_{p,X})_* \mu_X$ over $\mathbb{R}$. The metric between distributions will be taken to be W_p, i.e. the codomain of this invariant is $(\mathrm{Prob}(\mathbb{R}), W_p)$. Next, the invariant associated to Equation (JE) will turn out to be the map that takes X to the distribution over distributions of $\omega_X(x,\cdot)$. Specifically, it will be the pushforward of μ_X under the map $x \mapsto \omega_X(x,\cdot)_* \mu_X$. The codomain of this invariant will be $(\mathrm{Prob}(\mathrm{Prob}(\mathbb{R})), W_p)$, where the ground metric on $\mathrm{Prob}(\mathbb{R})$ is also taken to be W_p.

As a related construction, we note that given any (X, ω_X, μ_X), taking a pushforward of $\mu_X^{\otimes 2}$ via ω_X yields a distribution over $\mathbb{R}$. This produces yet another invariant whose codomain is $(\mathrm{Prob}(\mathbb{R}), W_p)$.

REMARK 3.2 (Local and global invariants). Let $(X, \omega_X, \mu_X) \in \mathcal{N}$. Both the size_p and $(\omega_X)_*(\mu_X \otimes \mu_X)$ invariants are examples of *global* invariants, in the sense that they incorporate data from the network without any reference to particular nodes in the network. In contrast, $(\mathrm{ecc}^{\mathrm{out}}_{p,X})_* \mu_X$ and $(x \mapsto \omega_X(x,\cdot)_* \mu_X)_* \mu_X$ incorporate information at the level of individual nodes within the network, and constitute examples of *local* invariants.

We now state the main theorem of this section, which provides a hierarchy of lower bounds for $d_{\mathcal{N},p}$.

THEOREM 3.1 (Hierarchy of lower bounds). Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathcal{N}$, and let $p \in [1, \infty]$. Let $C : X \times Y \to \mathbb{R}$ denote a cost matrix with entries $C(x,y) := W_p\left(\omega_X(x,\cdot)_* \mu_X, \omega_Y(y,\cdot)_* \mu_Y\right)$. Then we

have the following statements about Lipschitz stability, for $p \in [1, \infty]$:

$$
\begin{aligned}
2d_{\mathcal{N},p}(X,Y) &= \inf_{\mu \in \mathscr{C}(\mu_X,\mu_Y)} \operatorname{dis}_p(\mu) = \inf_{\mu \in \mathscr{C}(\mu_X,\mu_Y)} \left\| \omega_X - \omega_Y \right\|_{L^p(\mu^{\otimes 2})} \\
&\geq \inf_{\mu,\nu \in \mathscr{C}(\mu_X,\mu_Y)} \left\| \omega_X - \omega_Y \right\|_{L^p(\nu \otimes \mu)} \\
&= \inf_{\mu \in \mathscr{C}(\mu_X,\mu_Y)} \left\| \operatorname{ecc}^{\text{out}}_{p,X,Y} \right\|_{L^p(\mu)} && \text{(TLB)} \\
&= \inf_{\mu \in \mathscr{C}(\mu_X,\mu_Y)} \|C\|_{L^p(\mu)} && (\mathbb{R}\text{-TLB}) \\
&\geq \inf_{\mu \in \mathscr{C}(\mu_X,\mu_Y)} \left\| \operatorname{ecc}^{\text{out}}_{p,X} - \operatorname{ecc}^{\text{out}}_{p,Y} \right\|_{L^p(\mu)} && \text{(FLB)} \\
&= W_p\left((\operatorname{ecc}^{\text{out}}_{p,X})_* \mu_X, (\operatorname{ecc}^{\text{out}}_{p,Y})_* \mu_Y \right). && (\mathbb{R}\text{-FLB}) \\
&\geq \left| \operatorname{size}_p(X) - \operatorname{size}_p(Y) \right| && \text{(SzLB)} \\
2d_{\mathcal{N},p}(X,Y) &\geq \inf \left\{ \left\| \omega_X - \omega_Y \right\|_{L^p(\mu)} : \mu \in \mathscr{C}(\mu_X^{\otimes 2}, \mu_Y^{\otimes 2}) \right\} && \text{(SLB)} \\
&= W_p\left((\omega_X)_* \mu_X^{\otimes 2}, (\omega_Y)_* \mu_Y^{\otimes 2} \right). && (\mathbb{R}\text{-SLB})
\end{aligned}
$$

Moreover, analogous bounds hold for the $\operatorname{ecc}^{\text{in}}$ variants as well.

Remark 3.3 The inequalities in Theorem 3.1 appeared in the context of metric measure spaces as the FLB, SLB and TLB and their pushforwards in [**20**). In the asymmetric context of the current paper, we obtain outgoing/incoming versions of the (TLB) and (FLB) inequalities. The main development of the current paper is that we have equalities (FLB)=(ℝ-FLB), (SLB)=(ℝ-SLB) and (TLB)=(ℝ-TLB). The equality (TLB)=(ℝ-TLB) is especially important. *A priori*, each computation of $\operatorname{ecc}^{\text{out}}_{p,X,Y}(x,y)$ involves an OT problem that can be solved via linear programming methods. The equality (TLB)=(ℝ-TLB) shows that this quantity is actually equal to the solution of an OT problem over the real line, which has a closed form solution. Finally, we note that in the discrete case, all of the aforementioned equalities follow from [**27**, Proposition 4.5]. The current theorem proves the equalities in the general setting.

Remark 3.4 (Connecting lower bounds to network invariants). The (TLB) lower bound arises by solving an OT problem with Equation (JE) as a cost matrix, and the (FLB) lower bound arises by solving an OT problem with a difference of terms described by Equation (E) as a cost matrix. By virtue of the equalities (FLB)=(ℝ-FLB), (SLB)=(ℝ-SLB) and (TLB)=(ℝ-TLB), these lower bounds arise precisely as Lipschitz stability conditions on the network invariants described prior to the statement of Theorem 3.1.

Before proving Theorem 3.1, we introduce some terminology from [**16**, §14A]. A subset A of a Polish space X is *analytic* if it is the continuous image of a Polish space Y. Equivalently, A is analytic if there exists a Polish space Y and a Borel subset $B \subseteq X \times Y$ such that $A = \pi_X(B)$, where π_X is the canonical projection. Any Borel measurable map $f : X \to Y$, where Y is Polish, maps analytic sets to analytic sets [**16**, Proposition 14.4].

Lemma 3.1 (Lemma 2.2, [**34**]). Let X, Y be analytic subsets of Polish spaces equipped with the relative Borel σ-fields. Let $f : X \to Y$ be a surjective, Borel-measurable map. Then for any $\nu \in \operatorname{Prob}(Y)$, there exists $\mu \in \operatorname{Prob}(X)$ such that $\nu = f_* \mu$.

The next lemma states that pushforwards of couplings are exactly the couplings between the pushforwards. This was shown in the special case of discrete spaces in [27, Proposition 4.5].

LEMMA 3.2 Let X, Y be Polish, and let $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ be measurable. Let $T : X \times Y \to \mathbb{R}^2$ be the map $(x, y) \mapsto (f(x), g(y))$. Then we have:

$$T_* \mathscr{C}(\mu_X, \mu_Y) = \mathscr{C}(f_* \mu_X, g_* \mu_Y). \tag{3.1}$$

Consequently, we have:

$$W_p(f_* \mu_X, g_* \mu_Y) = \inf_{\mu \in \mathscr{C}(\mu_X, \mu_Y)} \|f - g\|_{L^p(\mu)}. \tag{3.2}$$

Proof of Lemma 3.2 Let $\mu \in \mathscr{C}(\mu_X, \mu_Y)$. It is standard [2, 7.1.6] that $T_* \mu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)$, and hence $W_p(f_* \mu_X, g_* \mu_Y) \leq \|f - g\|_{L^p(\mu)}$.

For the '$\supseteq$' containment of Equation (3.1), let $\nu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)$. The map $T = (f, g)$ is measurable because f, g are measurable. Next note that $X \times Y$ is Polish and hence analytic. Because $X \times Y$ is analytic and $T : X \times Y \to \mathbb{R}^2$ is a measurable map between Polish spaces, the image $T(X \times Y)$ is analytic [16, Proposition 14.4]. The map $T : X \times Y \to T(X \times Y)$ is surjective by construction. Then Lemma 3.1 applies to the map $T : X \times Y \to T(X \times Y)$ and the restriction $\nu|_{T(X \times Y)} \in \text{Prob}(T(X \times Y))$. Thus we obtain $\sigma \in \mathscr{C}(\mu_X, \mu_Y)$ such that $T_* \sigma = \nu|_{T(X \times Y)}$. Finally, note that ν is completely determined by its restriction to $T(X \times Y)$: for any $Z \in \text{Borel}(\mathbb{R}^2)$, we have $\nu(Z) = \nu(Z \cap T(X \times Y))$. Since $\nu|_{T(X \times Y)}$ determines ν, the existence of σ such that $T_* \sigma = \nu|_{T(X \times Y)}$ suffices to show the $\supseteq$ containment. The equality $W_p(f_* \mu_X, g_* \mu_Y) = \|f - g\|_{L^p(\mu)}$ follows immediately.

For Equation (3.2), note that by a change of variables we have ($d_{\mathbb{R}}$ is just the standard distance on $\mathbb{R}$):

$$\|d_{\mathbb{R}}\|_{L^p(T_* \mu)} = \|f - g\|_{L^p(\mu)}.$$

Let $\nu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)$. By the preceding work, $\nu = T_* \sigma$ for some $\sigma \in \mathscr{C}(\mu_X, \mu_Y)$. Hence we have:

$$W_p(f_* \mu_X, g_* \mu_Y) = \inf_{\nu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)} \|d_{\mathbb{R}}\|_{L^p(\nu)} = \inf_{\mu \in \mathscr{C}(\mu_X, \mu_Y)} \|f - g\|_{L^p(\mu)}.$$

□

Proof of Theorem 3.1 Inequality (TLB) holds because ν is allowed to vary and thus we infimize over a larger set. Next fix $(x, y) \in X \times Y$. Applying Lemma 3.7 Equation (3.2), we have

$$C(x, y) = W_p\left(\omega_X(x, \cdot)_* \mu_X, \omega_Y(y, \cdot)_* \mu_Y\right) = \inf_{\nu \in \mathscr{C}(\mu_X, \mu_Y)} \|\omega_X(x, \cdot) - \omega_Y(y, \cdot)\|_{L^p(\nu)} = \text{ecc}^{\text{out}}_{p, X, Y}(x, y).$$

This proves (TLB)=(ℝ-TLB). Next, for any $\nu \in \mathscr{C}(\mu_X, \mu_Y)$, we have by Minkowski's inequality:

$$\|\omega_X(x, \cdot) - \omega_Y(y, \cdot)\|_{L^p(\nu)} \geq \left| \|\omega_X(x, \cdot)\|_{L^p(\nu)} - \|\omega_Y(y, \cdot)\|_{L^p(\nu)} \right| = \left| \text{ecc}^{\text{out}}_{p, X}(x) - \text{ecc}^{\text{out}}_{p, Y}(y) \right|$$

This shows (TLB)≥(FLB). The equality (FLB)=(ℝ-FLB) follows by another application of Lemma 3.7. Next, for any $\mu \in \mathscr{C}(\mu_X, \mu_Y)$, another application of Minkowski's inequality yields:

$$\left\|\mathrm{ecc}^{\mathrm{out}}_{p,X} - \mathrm{ecc}^{\mathrm{out}}_{p,Y}\right\|_{L^p(\mu)} \geq \left|\left\|\mathrm{ecc}^{\mathrm{out}}_{p,X}\right\|_{L^p(\mu)} - \left\|\mathrm{ecc}^{\mathrm{out}}_{p,Y}\right\|_{L^p(\mu)}\right| = \left|\mathrm{size}_p(X) - \mathrm{size}_p(Y)\right|.$$

This shows (FLB)≥(SzLB).

Finally for (SLB), let μ denote the minimizer of dis_p (invoking Theorem 2.2) and define $\sigma := \mu \otimes \mu$. Then $\sigma \in \mathscr{C}(\mu_X^{\otimes 2}, \mu_Y^{\otimes 2})$. Hence

$$2d_{\mathscr{N},p}(X,Y) = \left\|\omega_X - \omega_Y\right\|_{L^p(\mu\otimes\mu)} = \left\|\omega_X - \omega_Y\right\|_{L^p(\sigma)} \geq \inf_{\nu\in\mathscr{C}(\mu_X^{\otimes 2},\mu_Y^{\otimes 2})} \left\|\omega_X - \omega_Y\right\|_{L^p(\nu)}.$$

This shows (SLB). The equality (SLB)=(ℝ-SLB) follows by applying Lemma 3.7 Equation (3.2) with $f = \omega_X$ and $g = \omega_Y$. □

3.2 *Interleaving stable invariants*

We now present a novel family of invariants that satisfies a different type of stability. Let $(X, \omega_X, \mu_X) \in \mathscr{N}$, and let $p \in [1, \infty]$. For each $t \in \mathbb{R}$ and $x \in X$, define the quantity

$$\mathrm{ecc}^{\mathrm{out}}_{p,X}(x,t) := \left\|\omega_X(x,\cdot)\mathbf{1}_{\{\omega_X(x,\cdot)\leq t\}}\right\|_{L^p(\mu_X)}.$$

This is an overload of notation, but the meaning should be clear from the presence of the second parameter. Note that $\{\omega_X(x,\cdot) \leq t\}$ is measurable, and so $\mathbf{1}_{\{\omega_X(x,\cdot)\leq t\}}$ is measurable. Hence the integral is well-defined.

REMARK 3.5 For a metric space (X, d_X, μ_X), the quantity $\left\|\mathbf{1}_{\{d_X(x,\cdot)\leq t\}}\right\|^p_{L^p(\mu_X)}$ is just the measure of the ball of radius t centred at x.

Next, the pth *sublevel* size function is defined for each $(X, \omega_X, \mu_X) \in \mathscr{N}$ and $t \in \mathbb{R}$ by writing

$$\mathrm{subSize}_{p,t}(X) = \left\|\mathrm{ecc}^{\mathrm{out}}_{p,X}(\cdot,t)\right\|_{L^p(\mu_X)}. \tag{subSz}$$

This function is a network invariant. Note that by the Fubini–Tonelli theorem, we can also write $\mathrm{subSize}_{p,t}(X) = \left\|\omega_X\mathbf{1}_{\{\omega_X\leq t\}}\right\|_{L^p(\mu_X\otimes\mu_X)}$. Both formulations are used below.

EXAMPLE 3.1 In [21, Example 5.7], it was shown that the 1-*diameter* invariant (referred to as size_1 in this paper) does not discriminate between spheres of different dimensions. Specifically, it was shown that

$$\mathrm{size}_1(\mathbb{S}^n) = \frac{\pi}{2}. \qquad \text{(for any } n \in \mathbb{N})$$

We now show via explicit computations that the map $t \mapsto \mathrm{subSize}_{1,t}$ does distinguish between spheres. For each $n \in \mathbb{N}$, let $\mathbb{S}^n$ denote the n-sphere with the geodesic metric and normalized volume measure.

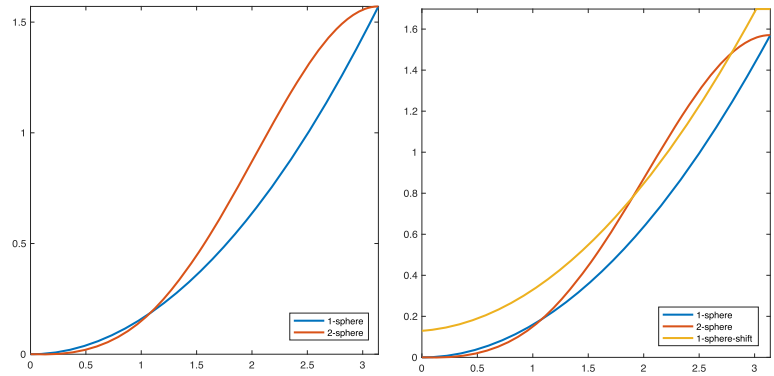


FIG. 3. **Left:** Plots of f_1 and g_1 as described in §3.2.1. **Right:** Plots of f_1, g_1 and a shifted version of f_1.

For each $n \in \mathbb{N}$, let S_n denote the surface area of $\mathbb{S}^n$. We have:

$$S_1 = 2\pi,\ S_2 = 4\pi,\ S_3 = 2\pi^2,\ S_4 = \frac{8}{3}\pi^2.$$

The following formula gives subSize for $\mathbb{S}^n$, $n \in \mathbb{N}$.

PROPOSITION 3.1 Fix $p \in [1, \infty)$. Let $n \in \mathbb{N}$, $n \geq 2$ and $0 \leq t \leq \pi$. Then,

$$\mathrm{subSize}_{p,t}(\mathbb{S}^n)^p = \frac{S_{n-1}}{S_n} \int_0^t \varphi^p \sin^{n-1}(\varphi)\, \mathrm{d}\varphi.$$

For $n = 1$, we have:

$$\mathrm{subSize}_{p,t}(\mathbb{S}^1)^p = \frac{t^{p+1}}{(p+1)\pi}.$$

By applying this result, we obtain $\mathrm{subSize}_{1,t}(\mathbb{S}^1) = \frac{t^2}{2\pi}$ and $\mathrm{subSize}_{1,t}(\mathbb{S}^2) = \frac{\sin(t) - t\cos(t)}{2}$, where $0 \leq t \leq \pi$. Plots of these functions are provided in Fig. 3. Note that by having access to the functions, instead of just the function values at $t = \pi$ (which corresponds to the prior size_1 result of [**21**, Example 5.7]), we are able to distinguish between spheres of different dimensions.

An interesting consequence of the preceding result, along with the result that $\mathrm{size}_1(\mathbb{S}^n) = \frac{\pi}{2}$ for all $n \in \mathbb{N}$, is the following identity for $n \geq 2$:

$$\frac{S_{n-1}}{S_n} \int_0^\pi \varphi \sin^{n-1}(\varphi)\ \mathrm{d}\varphi = \frac{\pi}{2}. \tag{3.3}$$

In particular, this identity and the formula in Proposition 3.1 explain why the 1-diameter (i.e. size_1) cannot distinguish between spheres, and why $\mathrm{subSize}_{1,t}$ is able to do so.

Proof of Proposition 3.1 Let $n = 1$. We obtain the formula as a line integral. Let $r(\theta) = (\cos\theta, \sin\theta)$ be a parametrization of the circle, where $\theta \in [0, 2\pi)$. Using symmetry, we have the following for

$0 \leq t \leq \pi$:

$$\mathrm{subSize}_{p,t}(\mathbb{S}^1)^p = \frac{2}{S_1}\int_0^t \theta^p \|r'(\theta)\|\,\mathrm{d}\theta = \frac{2}{S_1}\left.\frac{\theta^{p+1}}{p+1}\right|_0^t = \frac{t^{p+1}}{(p+1)\pi}.$$

Next let $n \geq 2$. In hyperspherical coordinates, the area element of $\mathbb{S}^n$ is given by

$$\sin^{n-1}(\varphi_1)\sin^{n-2}(\varphi_2)\cdots\sin(\varphi_{n-1})\,\mathrm{d}\varphi_1\mathrm{d}\varphi_2\cdots\mathrm{d}\varphi_n,$$

where the limits of integration are $[0,\pi]$ for $\varphi_1,\ldots,\varphi_{n-1}$, and $[0,2\pi]$ for φ_n. As an example, we have:

$$\mathrm{subSize}_{p,t}(\mathbb{S}^2)^p = \frac{1}{S_2}\int_0^{2\pi}\int_0^t \varphi_1^p \sin\varphi_1\,\mathrm{d}\varphi_1\mathrm{d}\varphi_2.$$

Generalizing to larger values of n, we have:

$$\begin{aligned}\mathrm{subSize}_{p,t}(\mathbb{S}^n)^p &= \frac{1}{S_n}\int_{\varphi_n=0}^{2\pi}\int_{\varphi_{n-1}=0}^{\pi}\cdots\int_{\varphi_1=0}^{t}\varphi_1^p\ \sin^{n-1}(\varphi_1)\sin^{n-2}(\varphi_2)\cdots\sin(\varphi_{n-1})\,\mathrm{d}\varphi_1\mathrm{d}\varphi_2\cdots\mathrm{d}\varphi_n\\ &= \frac{S_{n-1}}{S_n}\int_{\varphi_1=0}^{t}\varphi_1^p\ \sin^{n-1}(\varphi_1)\,\mathrm{d}\varphi_1.\end{aligned}$$

□

Having motivated subSize by at least a theoretical application, we now proceed to its stability.

THEOREM 3.2 (Interleaving stability of subSize). Let $p \in [1,\infty]$, $t \in \mathbb{R}$ and let (X,ω_X,μ_X), $(Y,\omega_Y,\mu_Y) \in \mathscr{N}$. Define $\varepsilon := d_{\mathscr{N},0}^{\mathscr{GP}}(X,Y)$. Then we have the following interleaving stability:

$$\begin{aligned}\mathrm{subSize}_{p,t}(X) &\leq \varepsilon + \mathrm{subSize}_{p,t+\varepsilon}(Y),\\ \mathrm{subSize}_{p,t}(Y) &\leq \varepsilon + \mathrm{subSize}_{p,t+\varepsilon}(X).\end{aligned}$$

Proof. We show the first statement. Invoking Lemma 2.4, we write $\varepsilon = d_{\mathscr{N},\infty}(X,Y)$. Using Theorem 2.2, let $\mu \in \mathscr{C}(\mu_X,\mu_Y)$ be an optimal coupling for which $d_{\mathscr{N},\infty}(X,Y) = \varepsilon$ is achieved. Let $B := \{(x,y,x',y') \in (X\times Y)^2 : \left|\omega_X(x,x') - \omega_Y(y,y')\right| \geq \varepsilon\}$. Let G denote the complement of B, i.e. $G := \left\{(x,y,x',y') \in (X\times Y)^2 : \left|\omega_X(x,x') - \omega_Y(y,y')\right| < \varepsilon\right\}$. By the definition of ε, we have $\mu^{\otimes 2}(B) = 0$, and hence $\mu^{\otimes 2}(G) = 1$. Also define $H := G \cap \left(\{\omega_X \leq t\}\times Y^2\right)$. Then we have:

$$\begin{aligned}\mathrm{subSize}_{p,t}(X) &= \left\|\mathrm{ecc}_{p,X}^{\mathrm{out}}(\cdot,t)\right\|_{L^p(\mu_X)} = \left\|\omega_X\mathbf{1}_{\{\omega_X\leq t\}}\right\|_{L^p(\mu_X^{\otimes 2})} = \left\|\omega_X\mathbf{1}_{\{\omega_X\leq t\}\times Y^2}\right\|_{L^p(\mu^{\otimes 2})}\\ &= \left\|\omega_X\mathbf{1}_H\right\|_{L^p(\mu^{\otimes 2})} = \left\|\left(\omega_X-\omega_Y+\omega_Y\right)\mathbf{1}_H\right\|_{L^p(\mu^{\otimes 2})}\\ &\leq \left\|\left(\omega_X-\omega_Y\right)\mathbf{1}_H\right\|_{L^p(\mu^{\otimes 2})} + \left\|\omega_Y\mathbf{1}_H\right\|_{L^p(\mu^{\otimes 2})}\\ &< \varepsilon + \left\|\omega_Y\mathbf{1}_{\{\omega_Y\leq t+\varepsilon\}}\right\|_{L^p(\mu_Y^{\otimes 2})} = \varepsilon + \mathrm{subSize}_{p,t+\varepsilon}(Y).\end{aligned} \tag{3.4}$$

Here the third equality holds because μ is a coupling measure, and the fourth equality holds because $\mu^{\otimes 2}(G) = 1$. The first inequality holds by Minkowski's inequality. The first part of the second inequality holds because $|\omega_X(x,x') - \omega_Y(y,y)| < \varepsilon$ on H, and the second part holds because $\omega_Y(y,y') \leq \omega_X(x,x') + \varepsilon \leq t + \varepsilon$ on H. Finally, note that repeating the argument with the roles of X and Y switched completes the proof. □

REMARK 3.6 While not applied in the current paper, we may also consider a *superlevel* size function $\mathrm{supSize}_{p,t}(X) := \left\| \omega_X \mathbf{1}_{\{\omega_X \geq t\}} \right\|_{L^p(\mu_X \otimes \mu_X)}$. In the setup of Theorem 3.4, this invariant satisfies the following interleaving stability:

$$\mathrm{supSize}_{p,t}(X) \leq \varepsilon + \mathrm{supSize}_{p,t-\varepsilon}(Y)$$
$$\mathrm{supSize}_{p,t}(Y) \leq \varepsilon + \mathrm{supSize}_{p,t-\varepsilon}(X).$$

To see this, note that the proof of Theorem 3.4 carries through until the step in Inequality (3.4). In this case, for any $(x,y,x',y') \in H$ we have $\omega_Y(y,y') > \omega_X(x,x') - \varepsilon \geq t - \varepsilon$, thus $\mathbf{1}_H$ reduces to $\mathbf{1}_{\{\omega_Y \geq t-\varepsilon\}}$.

3.2.1 *Lower bounds for spheres* Fix $n, m \in \mathbb{N}$. We now invoke Theorem 3.4 to obtain lower bounds on $d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m)$. The explicit value of $d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m)$ is unknown in the existing literature, even for $n = 1$, $m = 2$.

Consider the family $\mathfrak{F} := \{f : [0,\pi] \to \mathbb{R}_+ : f \text{ increasing}\}$. For each $f \in \mathfrak{F}$ and $\varepsilon \in [0,\pi]$, define f^ε by writing, for each $t \in [0,\pi]$,

$$f^\varepsilon(t) := \begin{cases} f(t+\varepsilon) + \varepsilon & : t + \varepsilon \in [0,\pi] \\ f(\pi) + \varepsilon & : \text{ otherwise.} \end{cases}$$

Next define the *interleaving distance* d_{I} on $\mathfrak{F}$ by writing, for each $f, g \in \mathfrak{F}$,

$$d_{\mathrm{I}}(f,g) := \inf\{\varepsilon \geq 0 : f \leq g^\varepsilon \text{ and} g \leq f^\varepsilon\}.$$

This d_{I} is a pseudometric on $\mathfrak{F}$. Next, for $p \in [1,\infty)$, define $f_p, g_p : [0,\pi] \to \mathbb{R}$ by writing:

$$f_p(t) := \mathrm{subSize}_{p,t}(\mathbb{S}^n), \qquad g_p(t) := \mathrm{subSize}_{p,t}(\mathbb{S}^m). \qquad \text{(for all } t \in [0,\pi])$$

Define $\eta := d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m)$. Applying Theorem 3.4, we have $f_p \leq g_p^\eta$ and $g_p \leq f_p^\eta$. Thus $d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m) \geq d_{\mathrm{I}}(f_p, g_p)$. Moreover, by the triangle inequality of d_{I}, we have

$$d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m) \geq d_{\mathrm{I}}(f_p, g_p) \geq \left| d_{\mathrm{I}}(f_p, h) - d_{\mathrm{I}}(h, g_p) \right|,$$

for arbitrary $h \in \mathfrak{F}$. In particular, setting $h \equiv 0$, we have $d_{\mathrm{I}}(f_p, h) = \mathrm{subSize}_{p,\pi}(\mathbb{S}^n) = \mathrm{size}_p(\mathbb{S}^n)$ and $d_{\mathrm{I}}(g_p, h) = \mathrm{subSize}_{p,\pi}(\mathbb{S}^m) = \mathrm{size}_p(\mathbb{S}^m)$. Thus we obtain a size_p bound:

$$d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m) \geq \left| \mathrm{size}_p(\mathbb{S}^n) - \mathrm{size}_p(\mathbb{S}^m) \right|.$$

This bound can be easily improved using different choices of $h \in \mathfrak{F}$.

Using the explicit formula of Proposition 3.1, we are able to computationally obtain lower bounds on $d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^n, \mathbb{S}^m)$. Set $p = 1$, $n = 1$ and $m = 2$. Then $f_1(t) = \frac{t^2}{2\pi}$ and $g_1(t) = \frac{\sin(t) - t\cos(t)}{2}$. Plots of f_1 and g_1 are shown in Fig. 3.

Through Matlab simulations, we find $d_{\mathcal{N},0}^{\mathcal{GP}}(\mathbb{S}^1, \mathbb{S}^2) = d_{\mathcal{N},\infty}(\mathbb{S}^1, \mathbb{S}^2) \geq d_I(f_1, g_1) \geq \mathbf{0.17}$. To contrast this with a previously known lower bound, we refer to [21, Remark 5.16], where the lower bound $d_{\mathcal{N},2}(\mathbb{S}^1, \mathbb{S}^2) \geq 0.0503$ was obtained. Because $d_{\mathcal{N},\infty} \geq d_{\mathcal{N},2}$, this previously known lower bound yields $d_{\mathcal{N},\infty}(\mathbb{S}^1, \mathbb{S}^2) \geq 0.0503$. Our new lower bound of 0.17 improves this threefold.

4. Experiments

4.1 *Computational aspects*

Numerical experiments in [15,20] involved using an alternate optimization procedure to estimate a local minimum of the GW objective. The methods in [24,31] used an ERGW which led to fast algorithms. These methods remain valid in the setting of (possibly asymmetric) networks. To complement the existing literature, in this section we present the use of the (TLB) lower bound to compute dissimilarities between asymmetric networks. By virtue of the equality (TLB)=(ℝ-TLB), this lower bound can be computed by solving a single general OT problem over a cost matrix obtained by solving OT problems over the real line. This is practical because OT problems over $\mathbb{R}$ have closed form solutions, with the caveat that computing all these OT problems is still the main bottleneck in computations. In comparable demonstrations, the ERGW of [24] is orders of magnitude faster, but a standard warning about ERGW is that it is prone to numerical infeasibility issues (see Appendix A). For networks of several hundred nodes, the (ℝ-TLB) can be computed exactly at reasonable speed, i.e. in less than a minute in Matlab on a 2.3 GHz Intel i5 CPU with 8 GB memory. Our experiments show that (ℝ-TLB) works well in discriminating networks.

Next we review the formula for computing OT over $\mathbb{R}$ (see [36, Remark 2.19]). Let networks (X, ω_X, μ_X), (Y, ω_Y, μ_Y) and measurable functions $f : X \to \mathbb{R}$, $g : Y \to \mathbb{R}$ be given. In the ecc$^{\text{out}}$ setting, $f = \omega_X(x, \cdot)_* \mu_X$ and $g = \omega_Y(y, \cdot)_* \mu_Y$. Then let $F, G : \mathbb{R} \to [0, 1]$ denote the cumulative distribution functions of f and g:

$$F(t) := \mu_X(\{x \in X : f(x) \leq t\}), \qquad G(t) := \mu_Y(\{y \in Y : g(y) \leq t\}).$$

The generalized inverses $F^{-1} : [0, 1] \to \mathbb{R}$, $G^{-1} : [0, 1] \to \mathbb{R}$ are given as:

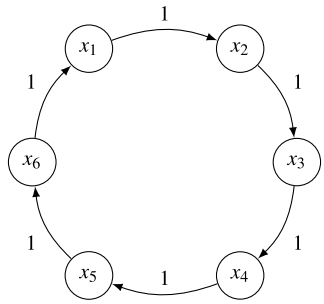
$$F^{-1}(t) := \inf\{u \in \mathbb{R} : F(u) \geq t\}, \qquad G^{-1}(t) := \inf\{u \in \mathbb{R} : G(u) \geq t\}.$$

Then for $p \geq 1$, one has:

$$\inf_{\mu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)} \int_{\mathbb{R} \times \mathbb{R}} |a - b|^p \, \mathrm{d}\mu(a, b) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p \, \mathrm{d}t. \tag{4.1}$$

For $p = 1$, one obtains a reformulation that incurs lower computational cost, at least in a naive implementation:

$$\inf_{\mu \in \mathscr{C}(f_* \mu_X, g_* \mu_Y)} \int_{\mathbb{R} \times \mathbb{R}} |a - b| \, \mathrm{d}\mu(a, b) = \int_{\mathbb{R}} |F(t) - G(t)| \, \mathrm{d}t. \tag{4.2}$$



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	1	2	3	4	5
x_2	5	0	1	2	3	4
x_3	4	5	0	1	2	3
x_4	3	4	5	0	1	2
x_5	2	3	4	5	0	1
x_6	1	2	3	4	5	0

FIG. 4. A cycle network on six nodes corresponding to the weight matrix obtained by right-shifting the vector $[0,1,2,3,4,5]^T$. Note that the weights are highly asymmetric.

In our experiments, we computed both the $\mathrm{ecc}^{\mathrm{out}}$ and $\mathrm{ecc}^{\mathrm{in}}$ versions of ($\mathbb{R}$-TLB) and take their maximum as the lower bound. All computations were done for $p = 2$. For Wasserstein distance computations, we used the `mexEMD` code accompanying [**24**]. Our code and data are available on https://github.com/samirchowdhury/GWnets.

In a prior version of this paper, before the equality (TLB)=($\mathbb{R}$-TLB) was proved in full generality, we were faced with the problem of solving an ensemble of OT problems over the space $X \times Y$. At the time, we resorted to using entropic regularization to compute the (TLB) in a reasonable amount of time. *A priori* this could also have been done by directly solving the associated linear programs, using e.g. `mexEMD`. While entropic regularization is not used in the current paper, we briefly report on these prior approaches in Appendix A.

4.2 *The network stochastic block model*

We now describe a generative model for random networks, based on the popular stochastic block model for sampling random graphs [**1**]. The current network SBM model we describe is a composition of Gaussian distributions. However, the construction can be adjusted easily to work with other distributions.

Fix a number of *communities* $N \in \mathbb{N}$. For $1 \leq i,j \leq N$, fix a mean μ_{ij} and a variance σ_{ij}^2. This collection $\mathscr{G} := \{\mathcal{N}(\mu_{ij}, \sigma_{ij}^2) : 1 \leq i,j \leq N\}$ of N^2 independent Gaussian distributions comprises the network SBM.

To sample a random network (X, ω_X) of n nodes from this SBM, start by fixing $n_i \in \mathbb{N}, 1 \leq i \leq N$ such that $\sum_i n_i = n$. For $1 \leq i \leq N$, let X_i be a set with n_i points. Define $X := \cup_{i=1}^n X_i$. Next sample each node weight as $\omega_X(x,x') \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$, where $x \in X_i$ and $x' \in X_j$. Finally, the pair (X, ω_X) is equipped with the uniform measure μ_X that assigns a mass of $1/n$ to each point.

We now describe the specifics of two experiments on clustering a collection of network SBMs.

4.3 *Experiment: SBMs from cycle networks*

Let $N \in \mathbb{N}$, and let $v = [v_1, \ldots, v_N]$ be an $N \times 1$ vector. Define the right-shift operator ρ by $\rho([v_1, \ldots, v_N]) = [v_N, v_1, \ldots, v_{N-1}]$. The *cycle network* $G_N(v)$ is defined to be the N-node network whose weight matrix is given by $[v^T, \rho(v)^T, (\rho^2(v))^T, \ldots, (\rho^{N-1}(v))^T]$. The cycle network definition appears elsewhere in the literature, see e.g. [**8**]. An illustration is provided in Fig. 4.

In our first experiment on network SBMs, we started with an $N \times 1$ vector of means v and used this to generate $G_N(v)$. We then used $G_N(v)$ as the matrix of means. To keep the experiment simple, we fixed

TABLE 1 ***Left:*** *The five classes of SBM networks corresponding to the experiment in §4.3. N refers to the number of communities, v refers to the vector that was used to compute a table of means via* $G_5(v)$ *and* n_i *is the number of nodes in each community.* ***Right:*** $G_5(v)$ *for* $v = [0, 25, 50, 75, 100]$

Class #	N	v	n_i	Sample cycle network of means				
1	5	[0,25,50,75,100]	10	0	25	50	75	100
2	5	[0,50,100,150, 200]	10	100	0	25	50	75
3	5	[0,25,50,75,100]	20	75	100	0	25	50
4	2	[0,100]	25	50	75	100	0	25
5	5	[−100,−50,0,50,100]	10	25	50	75	100	0

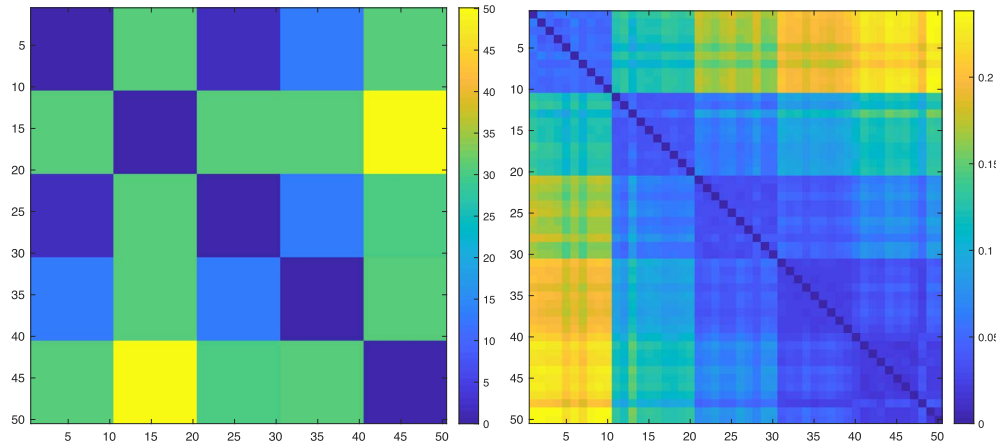


FIG. 5. **Left:** TLB dissimilarity matrix for SBM community networks in §4.3. Classes 1 and 3 are similar, even though networks in Class 3 have twice as many nodes as those in Class 1. Classes 2 and 5 are most dissimilar because of the large difference in their edge weights. Class 4 has a different number of communities than the others, and is dissimilar to Classes 1 and 3 even though all their edge weights are in comparable ranges. **Right:** (ℝ-TLB) dissimilarity matrix for two-community SBM networks in §4.4.

the matrix of variances to be the $N \times N$ matrix whose entries are all 5s. We made 5 choices of v, and sampled 10 networks for each choice. The objective was then to see how well the (ℝ-TLB) could split the collection of 50 networks into 5 classes corresponding to the 5 different community structures. The different parameters used in our experiments are listed in Table 1.

Class 1 is our reference; compared to this reference, class 2 differs in its edge weights, class 3 differs in the number of nodes in each community, class 4 differs in the number of communities and class 5 differs by having a larger proportion of negative edge weights. The (ℝ-TLB) results in Fig. 5 show that classes 1 and 3 are treated as being very similar, whereas the other classes are all mutually well-separated. This is consistent, because $d_{\mathcal{N}}$ is not sensitive to the size of the networks (cf. Theorem 2.4). One interesting suggestion arising from this experiment is that the (ℝ-TLB) can be used for network simplification: given a family of networks which are all at low (ℝ-TLB) distance to each other, it may be reasonable to retain only the smallest network in the family as the 'minimal representative' network.

TABLE 2 *Two-community SBM networks as described in* §4.4

Class #	N	v	n_i
1	2	[0,0]	10
2	2	[0,5]	10
3	2	[0,10]	10
4	2	[0,15]	10
5	2	[0,20]	10

4.4 *Experiment: Two-community SBMs with sliding means*

Having understood the interaction of the ($\mathbb{R}$-TLB) with network community structure, we next investigated how the ($\mathbb{R}$-TLB) behaves with respect to edge weights. In our second experiment, we used a 2×1 means vector v, and varied v as $[0,0], [0,5], \ldots, [0,20]$ (see Table 2). The SBM means were then given by $G_2(v)$ for the various choices of v. The variances were fixed to be the all 5s matrix. The edge weight histograms of the resulting SBM networks then looked like samples from two Gaussian distributions, with one of the Gaussians sliding away from the other. Finally, we normalized each network by its largest weight in absolute value, so that its normalized edge weights were in $[-1, 1]$.

The purpose of this experiment was to test the performance of ($\mathbb{R}$-TLB) on SBMs coming from a mixture of Gaussians. Note that normalization ensures that simpler invariants such as the size invariant would likely fail in this setting. The ($\mathbb{R}$-TLB) still performs reasonably well in this setting, as illustrated by the dissimilarity matrix in Fig. 5. The linear colour gradient is consistent with the 'sliding means' network structure.

4.5 *Experiment: Real migration networks*

For an experiment involving real-world networks, we compared global bilateral migration networks produced by the World Bank [**14**,**23**]. The data consist of 10 networks, each having 225 nodes corresponding to countries/administrative regions. The (i, j)th entry in each network is the number of people living in region i who were born in region j. The 10 networks comprise such data for male and female populations in 1960, 1970, 1980, 1990 and 2000. When extracting the data, we removed the entries corresponding to refugee populations, the Channel Islands, the Isle of Man, Serbia, Montenegro and Kosovo, because the data corresponding to these regions were incomplete/inconsistent across the database. We assigned uniform mass to the nodes.

The result of applying the ($\mathbb{R}$-TLB) to this dataset is illustrated in Fig. 6. To better understand the dissimilarity matrix, we also computed its single linkage dendrogram. The dendrogram suggests that between 1960 and 1970, both male and female populations had quite similar migration patterns. Within these years, however, migration patterns were more closely tied to gender. This effect is also seen between 1980 and 1990, although male migration in 1990 is more divergent. Finally, migration rates are similar for both male and female populations in 2000, and they are different from migration patterns in prior years.

The labels in the dissimilarity matrix are as follows: 1–5 correspond to 'f-1960' through 'f-2000', and 6–10 correspond to 'm-1960' through 'm-2000'. The colour gradient in the dissimilarity matrix suggests that within each gender, migration patterns change in a way that is parametrized by time. This reflects the shifts in global technological and economical forces, which make migration attractive and/or necessary with time.

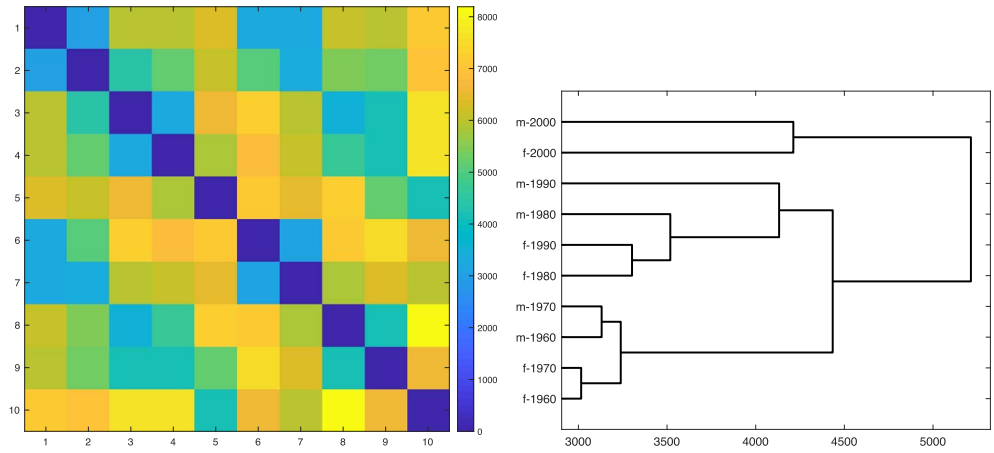


FIG. 6. Result of applying the ($\mathbb{R}$-TLB) to the migration networks in §4.5. **Left:** Dissimilarity matrix. Nodes 1–5 correspond to female migration from 1960–2000, and nodes 6–10 correspond to male migration from 1960–2000. **Right:** Single linkage dendrogram. Notice that overall migration patterns change in time, but within a time period, migration patterns are grouped according to gender.

5. Discussion

We have presented the GW distance as a valid pseudometric on the space of all directed, weighted networks. The crux of this approach is that even though the GW distance was originally formulated for metric measure spaces, the structure of the GW distance automatically forces a metric structure on networks. This yields the insight that the metric structure on the 'space of spaces' is not inherited from the metric on the ground spaces. In particular, while there are several metrics on networks that are combinatorial in nature, and hence hard to compute/sensitive to outliers, this GW metric is considerably more relaxed. The OT-based network invariants that we have presented yield lower bounds on the GW distance, which at most involve linear programming, and hence are readily computable. Finally, we tested our methods on a range of network datasets. The SBM network model that we defined for these tests will likely yield useful benchmarks for such network methods in future applications.

Acknowledgements

We are especially grateful to the anonymous reviewers for their detailed feedback and comments, and also to Justin Solomon for useful insights regarding computation.

Funding

National Science Foundation (grants IIS-1422400, DMS-1723003 and TRIPODS-1740761).

REFERENCES

1. ABBE, E. (2017) Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.*, **18**, 6446–6531.
2. AMBROSIO, L., GIGLI, N., & SAVARÉ, G. (2008) *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich, 2nd ed, Birkhäuser Verlag, Basel, x+334.

3. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L. & Peyré, G. (2015) Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, **37**, A1111–A1138.
4. Billingsley, P. (1999) *Convergence of Probability Measures*. Wiley Series in Probability and Statistics: Probability and Statistics, 2nd ed, New York: John Wiley & Sons pages x+277.
5. Chizat, L. (2017) Transport optimal de mesures positives: modèles, méthodes numériques, applications, *Ph.D. Thesis*.
6. Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. (2018) Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, **87**, 2563–2609.
7. Chowdhury, S. & Mémoli, F. (2017) Distances and isomorphism between networks and the stability of network invariants. *arXiv preprint arXiv:1708.04727*.
8. Chowdhury, S. & Mémoli, F. (2018) Persistent path homology of directed networks. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1152–1169.
9. Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, pp. 2292–2300.
10. Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouve, A. & Peyré, G. (2019) Interpolating between optimal transport and MMD using sinkhorn divergences. *Proceedings of Machine Learning Research*, 2681–2690.
11. Folland, G.B. (1999) *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics (New York), 2nd ed, New York: John Wiley & Sons, xvi+386.
12. Greven, A., Pfaffelhuber, P. & Winter, A. (2009) Convergence in distribution of random metric measure spaces (λ-coalescent measure trees). *Probab. Theory Relat. Fields*, **145**, 285–322.
13. Gromov, M. (1999) Metric structures for Riemannian and non-Riemannian spaces. vol. **152** of *Progress in Mathematics*, Boston, MA: Birkhäuser Boston Inc.
14. World Bank Group (2011) Global bilateral migration database. https://datacatalog.worldbank.org/dataset/global-bilateral-migration-database Accessed: October 3, 2018.
15. Hendrikson, R. (2016) Using Gromov–Wasserstein distance to explore sets of networks. *Master's Thesis*.
16. Kechris, A. (1995) *Classical Descriptive Set Theory*. Graduate Texts in Mathematics, vol. **156**. New York: Springer-Verlag, xviii+402.
17. Kleinberg, J. M. (1999) Authoritative sources in a hyperlinked environment. *J. ACM*, **46**, 604–632.
18. Kumar, S., Spezzano, F., Subrahmanian, V.S. & Faloutsos, C. (2016) Edge weight prediction in weighted signed networks. *Data Mining (ICDM), 2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, pp. 221–230.
19. Lovász, L. (2012) Large networks and graph limits. American Mathematical Society Colloquium Publications, **60**. Providence, RI: American Mathematical Society, xiv+475.
20. Mémoli, F. (2007) *On the use of Gromov–Hausdorff distances for shape comparison*. Prague, Czech Republic: The Eurographics Association.
21. Mémoli, F. (2011) Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, **11**, 417–487. Springer-Verlag.
22. Newman, M. (2010) *Networks: An Introduction*. Oxford: Oxford University Press, xii+772.
23. Özden, Ç., Parsons, C. R., Schiff, M. & Walmsley, T. L. (2011) Where on earth is everybody? The evolution of global bilateral migration 1960–2000. *World Bank Econom. Rev.*, **25**, 12–56.
24. Peyré, G., Cuturi, M. & Solomon, J. (2016) Gromov–Wasserstein averaging of kernel and distance matrices. *International Conference on Machine Learning*, Vol. **48**, New York, NY, USA: JMLR.org, 2664–2672.
25. Sanguinetti, G. & Huynh-Thu, V. A. (2019) Gene regulatory network inference: an introductory survey. *Gene Regulatory Networks*. New York, NY: Humana Press, 1–23.
26. Schmitzer, B. (2019) Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, **41**, A1443–A1481.
27. Schmitzer, B. & Schnörr, C. (2013) Modelling convex shape priors and matching based on the Gromov–Wasserstein distance. *J. Math. Imag. Vis.*, **46**, 143–159.

28. Shioya, T. (2016) *Metric Measure Geometry: Gromov's Theory of Convergence and Concentration of Metrics and Measures*. IRMA Lectures in Mathematics and Theoretical Physics. Zürich: EMS Publishing House, xi+182.
29. Sinkhorn, R. (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Stat.*, **35**, 876–879.
30. Sinkhorn, R. (1967) Diagonal equivalence to matrices with prescribed row and column sums. *Am. Math. Mon.*, **74**, 402–405.
31. Solomon, J., Peyré, G., Kim, V. G. & Sra, S. (2016) Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, **35**, 72.
32. Sturm, K.-T. (2006) On the geometry of metric measure spaces. *Acta Math.*, **196**, 65–131.
33. Sturm, K.-T. (2012) The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*.
34. Varadarajan, V. S. (1963) Groups of automorphisms of Borel spaces. *Trans. Am. Math. Soc.*, **109**, 191–220.
35. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. (2019) Optimal transport for structured data with application on graphs. *Proceedings of the 36th International Conference on Machine Learning,* Chaudhuri, Kamalika and Salakhutdinov, Ruslan. Long Beach, California, USA: PMLR, pp. 6275–6284.
36. Villani, C. (2003) Topics in optimal transportation, vol. **158** of *Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society, xvi+370.
37. Villani, C. (2008) *Optimal Transport: Old and New*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. **338**. Berlin: Springer-Verlag, xxii+973.

A. Computation via entropic regularization

Entropic regularization (ER), as used in [9] and further developed in [3,24,31,35], can be used in an iterative algorithm that approximates a local minimum of the GW objective [24,31]. In this section, we describe some heuristics that we found useful when applying ER-based techniques on network data. The main issue that we deal with is the following: initializing an ER-objective for networks having very different edge weights may create cost matrices with values below machine precision, which causes computations to blow up. A related issue that we found was the problem of 'entropic bias', which can be dealt with using well-understood techniques [10].

We first explain the notion of ER and associated difficulties with numerical stability. Throughout this section, we write M to denote a cost matrix depending on the edge weights of networks (X, ω_X, μ_X), (Y, ω_Y, μ_Y). This M could be the GW objective, as in [24], or one of the lower bound matrices from Theorem 3.1.

A.1 *Numerical stability of ER*

Let (X, ω_X, μ_X), (Y, ω_Y, μ_Y) be networks with $|X| = m$, $|Y| = n$. For a general $m \times n$ cost matrix M, one may consider the entropically regularized OT problem below, where $\lambda \geq 0$ is a regularization parameter and H denotes entropy:

$$\inf_{p \in \mathscr{C}(\mu_X, \mu_Y)} \sum_{i,j} M_{ij} p_{ij} - \frac{1}{\lambda} H(p), \qquad H(m) = -\sum_{i,j} p_{ij} \log p_{ij}.$$

As shown in [9], the solution to this problem has the form $\mathrm{diag}(a) * K * \mathrm{diag}(b)$, where $K := e^{-\lambda M}$ is a kernel matrix and a, b are non-negative scaling vectors in $\mathbb{R}^m, \mathbb{R}^n$, respectively. Here $*$ denotes matrix multiplication, and exponentiation is performed elementwise. An approximation to this solution can be

obtained by iteratively scaling K to have row and column sums equal to μ_X and μ_Y, respectively, and iterating until convergence. This is described in Algorithm 1.

Algorithm 1 Sinkhorn algorithm [9]

procedure SINKHORN(M, λ, mA, mB) ▷M an $m \times n$ cost matrix, mA, mB prob. measures

$a \leftarrow 1_m, b \leftarrow 1_n$ ▷ scaling updates, initialize as all-ones vectors

$K_{ij} \leftarrow \exp(-\lambda M_{ij})$ ▷ initialize kernel

repeat

$b \leftarrow mB./(K'a), a \leftarrow mA./(Kb)$

until convergence

return$\mathrm{diag}(a)K\,\mathrm{diag}(b)$

end procedure

As pointed out in [5,6,26], using a large value of λ (corresponding to a small regularization) leads to numerical instability, where values of K can go below machine precision and entries of the scaling updates a, b can also blow up. For example, Matlab will interpret e^{-1000} as 0, which is a problem with even a moderate choice of $\lambda = 200$ and $M_{ij} = 50$. Theoretically, it is necessary to have K be a positive matrix for the Sinkhorn algorithm to converge to the correct output [29,30]. Practitioners use a range of techniques to deal with the numerical instability, e.g. occasionally 'absorbing' extreme values of a, b into the kernel K (log-domain absorption), or gradually updating λ after starting with a conservative value (see [5] for more details).

Specifically in the network setting, initializing the kernel matrix K can be tricky due to the wide range of edge weights in the dataset: both within a network and between different networks. For example, in the migration network database, the migration into a large country like the USA is separated by several orders of magnitude from that of a smaller country, such as Austria. Furthermore, migration values differ significantly between years, e.g. between 1960 and 2000.

As discussed in [5], many entries of the stabilized kernel obtained as above could be below machine precision, but the entries corresponding to those on which the optimal plan is supported are likely to be above the machine limit. Indeed, this sparsity may even be leveraged for additional computational tricks.

The techniques for stabilizing the entropy regularized OT problem are not the focus of our work, but because these considerations naturally arose in our computational experiments, we describe some strategies we undertook that are complementary to the techniques available in the current literature. In order to provide a perspective complementary to that presented in [5], we impose the requirement that *all* entries of the kernel matrix remain above machine precision.

Initializing in the log domain. A simple adaptation of the 'log domain absorption' step referred to above yields a 'log initialization' method that works well in most cases for initializing K to have values above machine precision. To explain this method, we first present an algorithm (Algorithm 2) for the log domain absorption method. We follow the presentation provided in [5], making notational changes as necessary.

Notice that in Algorithm 2, K might already have values below machine precision at initialization. To circumvent this, we can add a preprocessing step that yields a stable initialization of K. This is

Algorithm 2 Sinkhorn with partial log domain steps

procedure SINKHORNLOG(M, λ, mA, mB) ▷M an $m \times n$ cost matrix, mA, mB prob. measures

  $a \leftarrow 1_m, b \leftarrow 1_n$ ▷ scaling updates

  $u \leftarrow 0_m, v \leftarrow 0_n$ ▷ log domain storage of large a, b

  $K_{ij} \leftarrow \exp(\lambda(-M_{ij} + u_i + v_j))$ ▷ initialize kernel

  while stopping criterion not met **do**

    $b \leftarrow mB./(K'a)$

    $a \leftarrow mA./(Kb)$

    if $\max(\max(a), \max(b)) >$ `threshold` **then**

      $u \leftarrow u + (1/\lambda)\log(a)$ ▷ store a, b in u, v

      $v \leftarrow v + (1/\lambda)\log(b)$

      $K_{ij} \leftarrow \exp(\lambda(-M_{ij} + u_i + v_j))$ ▷ absorb a, b into K

      $a \leftarrow 1_m, b \leftarrow 1_n$ ▷ after absorption, reset a, b

    end if

  end while

  return$\mathrm{diag}(a)K\,\mathrm{diag}(b)$

end procedure

outlined in Algorithm 3. An important point to note about Algorithm 3 is that the user needs to choose a function `decideParam(`α, β`)`, which returns a 'translation factor' γ, where α and β are as stated in the algorithm. This number γ should be such that $\exp(-\lambda\beta + 2\lambda\gamma)$ is above machine precision, but $\exp(-\lambda\alpha + 2\lambda\gamma)$ is not too large. The crux of Algorithm 3 is that by choosing large initial scaling vectors a, b and immediately absorbing them into the log domain, the extreme values of M are cancelled out before exponentiation.

A geometric interpretation in the $p = 2$ case. The preceding initialization method has its limitations: depending on how far $\min(M), \max(M)$ are spread apart, the log initialization step might not be able to yield an initial kernel K that has all entries above machine precision and below the machine limit. In such a case, one recourse is to choose a different value of λ. Thus when given a database of networks $X_1, \ldots, X_n$ and cost matrices arising from comparing these networks, one may need to choose $\lambda_{ij} = \lambda_{ji}$ for each pair $\{X_i, X_j\}$. It turns out that these potentially different λ values can be related to a global λ^* value by rescaling the networks in a geometric manner, using observations from [33]. This is described below. In what follows, we always have $p = 2$.

A.1.1 *Sturm's cosine rule construction* Let $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y) \in \mathscr{N}$. Recall from Example 2.11 that $d_{\mathscr{N},2}(X, N_1(0)) = \frac{1}{2}\,\mathrm{size}_2(X)$. Define $s := \frac{1}{2}\,\mathrm{size}_2(X, \omega_X, \mu_X), t := \frac{1}{2}\,\mathrm{size}_2(Y, \omega_Y, \mu_Y)$.

Algorithm 3 Log domain initialization of K

procedure LOGINITIALIZE(M, λ) ▷M an $m \times n$ cost matrix

$\alpha \leftarrow \min(M)$, $\beta \leftarrow \max(M)$ ▷ scan M for max and min values

$\gamma \leftarrow$ `decideParam`(α, β) ▷`decideParam` is an independent function

$a \leftarrow \exp(-\lambda\gamma)1_m$, $b \leftarrow \exp(-\lambda\gamma)1_n$

$u \leftarrow 0_m$, $v \leftarrow 0_n$

$K_{ij} \leftarrow \exp(\lambda(-M_{ij} + \gamma + \gamma))$ ▷K is stably initialized

perform rest of SINKHORNLOG as usual

end procedure

For an optimal coupling $\mu \in \mathscr{C}(\mu_X, \mu_Y)$, we have:

$$d_{\mathcal{N},2}(X,Y)^2 = \frac{1}{4}\int\int \omega_X(x,x')^2 + \omega_Y(y,y')^2 - 2\omega_X(x,x')\omega_Y(y,y')\,\mathrm{d}\mu(x,y)\,\mathrm{d}\mu(x',y')$$
$$= s^2 + t^2 - \frac{1}{2}\int\int \omega_X(x,x')\omega_Y(y,y')\,\mathrm{d}\mu(x,y)\,\mathrm{d}\mu(x',y'), \tag{A1}$$

where the first equality holds because $|a-b|^2 = \langle a-b, a-b\rangle = |a|^2 + |b|^2 - 2ab$ for all $a, b \in \mathbb{R}$, and the last equality holds because $\omega_X(x,x'), \omega_Y(y,y')$ do not depend on μ_Y, μ_X, respectively. Sturm [**33**, Lemma 4.2] observed the following 'cosine rule' structure. Define

$$\omega_X' := \frac{\omega_X}{2s}, \qquad \omega_Y' := \frac{\omega_Y}{2t}. \tag{A2}$$

Then $\mathrm{size}_2(X, \omega_X') = \frac{1}{2s}\,\mathrm{size}_2(X, \omega_X) = 1 = \frac{1}{2t}\,\mathrm{size}_2(Y, \omega_Y) = \mathrm{size}_2(Y, \omega_Y')$. A geometric fact about this construction is that (X, ω_X, μ_X), (Y, ω_Y, μ_Y) lie on geodesic rays connecting $\mathcal{X} := (X, \omega_X', \mu_X)$ and $\mathcal{Y} := (Y, \omega_Y', \mu_Y)$ respectively to $N_1(0)$. Actually, once (X, ω_X, μ_X) and (Y, ω_Y, μ_Y) are chosen, the geodesic rays are automatically defined to be given by the scalar multiples of ω_X, ω_Y. Then we independently define $\mathcal{X}$ and $\mathcal{Y}$ to be representatives of the weak isomorphism class of networks at $d_{\mathcal{N},2}$ distance $1/2$ from $N_1(0)$ that lie on these geodesics. We illustrate a related situation in Fig. A7, and refer the reader to [**33**] for further details. Implicitly using this geometric fact, we fix $\mathcal{X}, \mathcal{Y}$ as above and treat (X, ω_X, μ_X), (Y, ω_Y, μ_Y) as $2s$ and $2t$-scalings of $\mathcal{X}$ and $\mathcal{Y}$, respectively (i.e. such that Equation A2 is satisfied). Then we have:

$$4d_{\mathcal{N},2}((X, \omega_X, \mu_Y), (Y, \omega_X, \mu_Y))^2 - 4s^2 - 4t^2$$
$$= \int\int 4s^2\omega_X'(x,x')^2 + 4t^2\omega_Y'(y,y')^2 - 8st\omega_X'(x,x')\omega_Y'(y,y')\,\mathrm{d}\mu(x,y)\,\mathrm{d}\mu(x',y') - 4s^2 - 4t^2$$
$$= 4s^2\,\mathrm{size}_2(X,\omega_X')^2 + 4t^2\,\mathrm{size}_2(Y,\omega_Y')^2 - 4s^2 - 4t^2 - 8st\int\int \omega_X'(x,x')\omega_Y'(y,y')\,\mathrm{d}\mu(x,y)\,\mathrm{d}\mu(x',y')$$
$$= -8st\int\int \omega_X'(x,x')\omega_Y'(y,y')\,\mathrm{d}\mu(x,y)\,\mathrm{d}\mu(x',y'), \tag{A3}$$

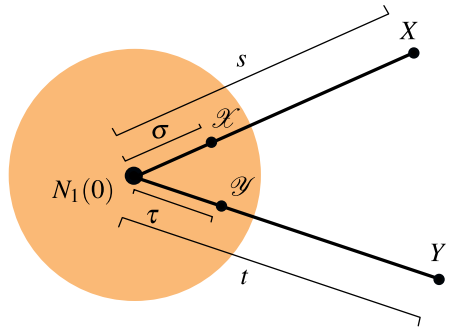


FIG. A7. Interaction between the ER parameter and rescalings, cf. §A.1.2. Choosing a regularization parameter λ_{XY} depending on the edge weights of X and Y is essentially the same as using a fixed parameter λ^* with (edge-weight) rescaled versions $\mathcal{X}, \mathcal{Y}$ of X and Y. Here $\mathcal{X}$, $\mathcal{Y}$ live on geodesic rays connecting $N_1(0)$ to X and Y. The letters s, σ, t, τ represent $d_{\mathcal{N},2}$-distances.

where the last equality holds because $\text{size}_2(X, \omega_X') = 1 = \text{size}_2(Y, \omega_Y')$. Since $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ were $2s, 2t$-scalings of $\mathcal{X}$ and $\mathcal{Y}$ for arbitrary $s, t > 0$, this shows in particular that the quantity

$$(1/2st)\left(d_{\mathcal{M},2}((X, \omega_X, \mu_Y), (Y, \omega_Y, \mu_Y))^2 - s^2 - t^2\right) \qquad \text{(cosine rule)}$$

depends only on the reference networks $\mathcal{X}$ and $\mathcal{Y}$, and is independent of s and t.

A.1.2 *Interpretation of λ and rescaling* Suppose now that we are in a setting where $\lambda^* > 0$, (X, ω_X, μ_X), (Y, ω_Y, μ_Y), and a cost matrix M depending on ω_X, ω_Y are all fixed. Suppose also that $e^{-\lambda^* M}$ contains values below machine precision, and $\lambda_{XY} > 0$ is such that $e^{-\lambda_{XY} M}$ has all values above machine precision. Then one may define $M^* := \frac{\lambda_{XY}}{\lambda^*} M$, so that $e^{-\lambda^* M^*} = e^{-\lambda_{XY} M}$. Here M^* is a rescaled cost matrix, and in typical use cases, it is the cost matrix obtained from rescaled weights ω_X', ω_Y'. For example, if $M = \omega_X \omega_Y$ (as in the integrand of Equation A1, also see [24]), then $M^* = \omega_X' \omega_Y'$, where ω_X', ω_Y' are rescaled from ω_X, ω_Y by $\sqrt{\lambda_{XY}/\lambda^*}$. By the observations from [33] presented above, these rescalings are compatible with the geometry of $(\mathcal{M}, d_{\mathcal{N},2})$, in the sense that the rescaled networks lie on geodesics connecting the original networks to the basepoint $N_1(0)$. This is illustrated in Fig. A7. See [33] for more details about the geodesic structure of gauged measure spaces; the analogous results hold for $(\mathcal{M}, d_{\mathcal{M},2})$.