Communications Physics

https://doi.org/10.1038/s42005-025-02468-5

Article in Press

Maximum entropy modelling of sub-optimal transport

Received: 14 April 2025

Accepted: 12 December 2025

Cite this article as: Buffa, L., Mazzilli, D., Piombo, R. *et al.* Maximum entropy modelling of sub-optimal transport. *Commun Phys* (2025). https://doi.org/10.1038/s42005-025-02468-5

Lorenzo Buffa, Dario Mazzilli, Riccardo Piombo, Fabio Saracco, Giulio Cimini & Aurelio Patelli

Maximum Entropy modelling of sub-Optimal Transport

Lorenzo Buffa[1,2], Dario Mazzilli[1,*], Riccardo Piombo[1], Fabio Saracco[1], Giulio Cimini[2,1], Aurelio Patelli[1]

[1]Enrico Fermi Research Center, 00184 Rome, Italy
[2]Department of Physics and INFN, University of Rome Tor Vergata, 00133 Rome, Italy
[*]dario.mazzilli@cref.it

December 9, 2025

Abstract

Many natural systems involve structures shaped by competing forces: efficiency, randomness, and incomplete information. To the best of our knowledge, there is currently no robust method to assess the presence of optimization processes in real networks. Here, we introduce a class of bipartite random graphs that bridges two foundational approaches: maximum entropy models and optimal transport theory. By tuning a single parameter, our model generates a continuous family of network configurations, ranging from fully random to cost-minimizing structures. This transition is governed by a variational principle analogous to free energy in statistical physics, where entropy and transport cost play competing roles. We analytically and numerically characterize how dense, entropic graphs evolve into sparse, efficient structures, revealing the most probable network configurations under partial optimization. Beyond clarifying the conceptual link between entropy-based and cost-based methods, our framework offers a generative model for systems where the structure emerges from random and constrained environments.

Introduction

Optimal Transport (OT) theory offers a powerful mathematical framework to describe the transformation of one distribution of mass into another at minimal total cost [1, 2]. Beyond its mathematical elegance, OT has demonstrated wide practical relevance in modeling constrained optimization problems across logistics, economics, and computer science, where limited resources—such as time, energy, or budget—necessitate efficient allocation strategies [3, 4, 5, 6, 7].

Many real-world systems, where agents aim to maximize gain or minimize cost in resource allocation, exhibit behavior consistent with some form of optimization [8, 9]. However, the specific nature of this optimization is often opaque. In particular, it is seldom evident whether the underlying process conforms to the formal structure of an OT problem. Ideally, we would like to infer the presence of an OT mechanism from observational data alone, without requiring detailed access to the governing dynamics or constraints. Yet this task is complicated by various sources of sub-optimality, noise, incomplete information, competing objectives—that naturally arise in empirical settings. As a result, the question of whether a given system is shaped by an OT process remains largely unanswered. To date, no rigorous statistical test has been proposed to detect the presence of OT in real-world data.

In discrete settings, OT problems are naturally represented as bipartite graphs, whose links connect, for instance, sources to sinks or producers to consumers. Interestingly, the exact solution to a discrete OT problem corresponds to a tree spanning the bipartite network [10]. This structural constraint implies that if an OT-like optimization is indeed shaping the system, it would leave distinctive network signatures. Yet, this connection has been largely overlooked in Network Theory, despite its potential implications for foundational tasks in statistical physics of networks [11].

The framework introduced in [12] focuses on the limit where the model collapses to the classical Optimal Transport (OT) solution, and restricts admissible transportation plans by an *a priori* upper bound on edge weights (= 1 by default). A comprehensive extension to generic transportation plans, together with an analysis covering the full range of suboptimal solutions, remains an open problem. Relatedly, Stock et al. [13] study a model not based on the maximum-entropy construction, thus defining possible biased ensembles. Consequently, some of the approximations and regularization choices in [13] limit the generality of their conclusions relative to the present formula-

tion. Further, a complementary line of work in physics explains sparse–dense transitions through nonlinear transport costs or growth-and-selection mechanisms, which typically generate tree-like networks without explicit OT structure [14, 15, 16]; more generally, topologically diversified structures may emerge from both growth processes and dynamical selection [17].

Motivated by this gap, we propose a class of OT-inspired random graph models, derived from maximum-entropy arguments, which interpolate between the strict optimality of tree-like structures and the more diffuse, redundant connectivity observed in empirical data. By blending the OT and the maximum entropy approaches, we position our framework within the broader theory of random network ensembles, where it generalizes known constructions. Our formulation connects these viewpoints: it yields a tunable ensemble of bipartite graphs where a single parameter governs a continuous transition between redundant (dense) and near-optimal (sparse, tree-like) connectivity, while preserving a transparent probabilistic interpretation and explicit links to OT. Practically, our model provides a statistical null for testing the presence of OT-like mechanisms and measuring the level of optimality in real data. Indeed, even though an OT optimization is characterized by the peculiar tree-like structure, this simple observation is not enough to test the presence of this process in a system. Other network models or processes might have a similar structure and a rigorous null model is required to perform model selection and statistical tests. Additionally, by allowing to capture suboptimal yet realistic network configurations, our approach can be relevant not only for solving noisy constrained optimization problems, but also for inference problems on networks such as link prediction, patterns and community detection, network reconstruction and graph combinatorics.

Results

Mathematical framework

In the discrete case, OT is well-suited to be described within a bipartite network framework. This setup involves two sets of elements, e.g. N coal mines and M factories, with assigned physical constraints, such as mining capacity and coal necessity respectively, represented by two vectors $\mathbf{s}$ and $\boldsymbol{\sigma}$, as shown in Fig. 1.

The transport plan is described by a matrix $\mathbf{w}$. Each element $w_{i\alpha}$ specifies

the amount of mass transported from mine i to factory α. To ensure the feasibility of the solution, the plan $\mathbf{w}$ must satisfy the marginal constraints that enforce the preservation of the total mass in each distribution:

$$\sum_{\alpha=1}^{M} w_{i\alpha} = s_i, \quad \text{and} \quad \sum_{i=1}^{N} w_{i\alpha} = \sigma_\alpha \tag{1}$$

These constraints guarantee that the OT solution respects the initial and final distributions: the total amount of mass leaving a source equals its supply, while the total amount received by a destination matches its demand. The constraints in Eq.(1) define a polytope Γ, which represents the set of all matrices w fulfilling the specified conditions.

In this framework, the objective is to minimize the transport cost using a unit cost matrix C. Each unit of mass transported from i to α costs $C_{i\alpha}$ and the total cost of a given transportation plan $\mathbf{w}$ is simply $\sum_{i\alpha} w_{i\alpha} C_{i\alpha}$. In the classic OT problem, $\mathbf{s}, \boldsymbol{\sigma}, \mathbf{C}$ are fixed and given as input to the problem. The optimization only requires to find the optimal transportation plan w^* such that it minimizes the total cost:

$$\mathbf{w}^* = \underset{\mathbf{w}\in\Gamma}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \sum_{\alpha=1}^{M} w_{i\alpha} C_{i\alpha} \right] \tag{2}$$

To enhance computational efficiency and stability, Peyré and Cuturi [18, 19] introduced an entropic regularization to the OT problem in Eq.(2) that can be solved thanks to the Sinkhorn-Knopp algorithm [20].

To build our random graph model we rely on Maximum Entropy (MaxEnt) principle, that provides the most unbiased estimate of a system's microscopic configuration by maximizing Shannon entropy, subject to a set of constraints that determine the values of the observables of interest [21, 22]. In the context of random graphs, the MaxEnt principle prescribes constructing ensembles of graphs that ensure an unbiased representation of all permissible network configurations that remain maximally uninformative beyond the specified constraints [11].

To formalize this framework, we define $P(G)$ as the probability of observing a particular network G in the ensemble $\mathcal{G}$, and let $\pi(G)$ be a generic observable whose expected value is constrained to a specified target π^*. In our setting, each graph G is a bipartite weighted network uniquely described by its biadjacency matrix $\mathbf{w}(G)$—an $N \times M$ matrix whose entries $w_{i\alpha}(G)$ are

positive real numbers that represent the intensity of interactions (weights) between nodes i and α in the two distinct layers. We also have the constraints on the node strengths $s_i(G)$ and $\sigma_\alpha(G)$, which specify how much "mass" or total weight each node can distribute or receive:

$$\langle s_i \rangle = \sum_{G\in\mathcal{G}} P(G) s_i(G) = s_i^*, \quad s_i(G) = \sum_\alpha w_{i\alpha}(G), \quad \forall i = 1, \ldots, N \tag{3a}$$

$$\langle \sigma_\alpha \rangle = \sum_{G\in\mathcal{G}} P(G) \sigma_\alpha(G) = \sigma_\alpha^*, \quad \sigma_\alpha(G) = \sum_i w_{i\alpha}(G), \quad \forall \alpha = 1, \ldots, M \tag{3b}$$

In Equations (3), the terms s_i^* and σ_α^* denote fixed values obtained from empirical observations, specifying the expected strengths for each node. Incorporating these constraints within the MaxEnt framework ensures that the generated ensemble of graphs aligns with the realistic conditions observed in actual data [23, 24, 25, 11, 26].

Sub-Optimal Transport Random network model

A key innovation introduced in this work is the extension of the random graph approach described in the previous section by including of a cost term, which assigns a relative importance to each link through a cost matrix $C_{i\alpha}$. Drawing an analogy from physics, this addition is akin to coupling a system to an external field. Importantly, the introduction of the cost matrix $C_{i\alpha}$ does not act as a new constraint. It represents a conceptual shift: instead of maximizing the entropy of the system, we maximize the analogue of a Helmholtz Free Energy F, subject to the constraints defined in Equations (3). If $U(G) = \sum_{i,\alpha} w_{i\alpha}(G)\, C_{i\alpha}$ represents the energy of the configuration (the network) G, given the disordered external field $C_{i\alpha}$, the F is defined as

$$\mathrm{F}[P] = \mathrm{S}[P] - \beta \mathrm{U}[P], \tag{4}$$

where $\mathrm{S}[P] = -\sum_{G\in\mathcal{G}} P(G) \log P(G)$ is the entropy, and $\mathrm{U}[P] = \sum_{G\in\mathcal{G}} P(G) U(G)$.

Solving the optimization problem in Eq.(4) provides a Boltzmann-like probability distribution:

$$P_{\mathrm{sub\,OT}}(G|\beta, \{\mathbf{t}, \boldsymbol{\theta}\}) = \frac{1}{Z(\beta, \{\mathbf{t}, \boldsymbol{\theta}\})} e^{-\beta U(G) - \sum_i t_i s_i - \sum_\alpha \theta_\alpha \sigma_\alpha} \tag{5}$$

Here $\{\mathbf{t}, \boldsymbol{\theta}\}$ represent the set of all Lagrange multipliers t_i and θ_α associated with the constraints in Eq.(3) and Z represents the partition function (i.e.

the normalization of the probability distribution). It is important to emphasize that the model considers weighted networks where all links are treated as equally significant, as no additional cost is considered. In the limit of vanishing $\beta \to 0$, this model converges to the Bipartite Weighted Configuration Model (BiWCM), discussed in [27]. When β is finite, we label the model *sub OT*.

The probability distribution in Eq.(5) can be factorized into exponential weight probability distributions for each edge (i, α) (see Methods for further details). Thus, it inherently produces a fully connected network, since every pair of nodes i, α the probability to be disconnected (i.e., have weight 0) is zero. Further, the average value of the weight of each link can be computed as

$$\langle w_{i\alpha} \rangle = \frac{1}{\beta C_{i\alpha} + t_i + \theta_\alpha} \tag{6}$$

allowing for the construction of an *average* network. The values of the Lagrange multipliers of Eq.(6) can be determined by maximizing the associated likelihood function $\mathcal{L}$ for a fixed value of β [28].

To solve the sub-optimal model numerically, we follow two primary approaches. The first is to directly solve the maximum log-likelihood problem through optimization techniques, such as stochastic gradient descent or Adam, available in many libraries for machine learning and neural network models [29]. The second approach involves using the analytical gradient of the log-likelihood to derive a system of nonlinear equations that explicitly include the Lagrange multipliers and the node-strength constraints s_i^* and σ_α^*. These equations can then be solved by methods like iterative maps [30] or conjugate gradient based algorithms [31]. Both pathways ultimately converge to the same numerical solution for the multipliers since the solution of the problem is unique. Additional details on the analytical formulation and the numerical procedures for each approach are provided in the Methods.

The parameter β tunes the importance of the energy term on the network probability distribution in Eq.(5). Its impact on the resulting ensemble is illustrated in Fig. 2, which shows how the expected weights $\{\langle w_{i\alpha} \rangle\}$ evolve as β varies. What we described in Fig. 1 corresponds to the rightmost illustration in the upper part of Fig. 2, representing the final stage of the OT framework. As shown, the system evolves from the BiWCM, through an intermediate structured state (SubOT), and finally reaches the exact OT plan.

To be more quantitative, we show in Fig.s 2a, b, c and d a typical behavior of the average weight matrix during the transition.

For low β, our model converges to the distribution of the corresponding BiWCM, where only the entropic term dominates and the external energy vanishes. In this regime, the weights are distributed homogeneously, according to the constraints, across all links (see Fig. 2a). As expected, their values depend entirely on the prescribed strengths s_i^* and σ_α^*, resulting in a dense network since no topological constraint is considered. The heatmap values are shown on a logarithmic scale, highlighting that in this scenario, the weights cluster tightly around their average values.

As β increases (Fig.s 2b and c), the weight progressively concentrates on the lowest-cost links, while others become increasingly suppressed. This behavior emerges from our ensemble construction, which maintains fully connected networks with a fixed total weight. In this regime, the system occupies an intermediate state of sub-optimality, where cost minimization and structural constraints compete. Consequently, a network corresponding to an intermediate β value, between the OT and BiWCM solutions, can be identified as a sub-optimal network relative to the OT solution.

For very large β, the weights condense in a few links with the lowest value of the cost, and when $\beta \to \infty$, the OT transportation plan is fully recovered (Fig. 2d) as already discussed in [12] (see also Methods in [32]).

In conclusion, at varying β, we can generate networks that are closer to or further from the OT solution. Therefore our model offers explicit control over the degree of sub-optimality of a network. Since real-world systems are unlikely to perfectly conform to the OT solution, with our method we can naturally explore this intermediate regime of sub-optimality.

In the following sections, we test alternative cost matrices and different strength distributions, providing a broader understanding of the interplay between cost and constraints in shaping the ensemble.

Maximum Spanning Tree and dense-to-sparse transition

Uniform cost and gaussian strength distribution

Tracking the evolution of the average weights $\langle w_{i\alpha} \rangle$ with β, we observe the link separate into two sets between those that form the OT solution and the rest, as shown in Fig. 3. At low β, the two sets are indistinguishable, but get perfectly apart at high β values. Using the terminology of thermodynamics, we refer to the structure at low β as the dense phase, since the network is fully connected and any heterogeneity of the weight is driven mainly by the

constraints. Instead, the phase at large β corresponds to the sparse phase, where a small number of links (proportional to the number of nodes) get most of the weights. Since the structure of the OT solution is known to be a spanning tree, we can characterize the degree of sub-optimality at any β by looking at the weight associated with the Maximum Spanning Tree (MST) of the average network, defined as the MST mass share:

$$m_{\mathrm{MST}}(\beta) = \frac{\sum\limits_{(i,\alpha)\in MST} \langle w_{i\alpha} \rangle}{\sum\limits_{i,\alpha} \langle w_{i\alpha} \rangle} \tag{7}$$

This particular choice of the 'order parameter' is given by a composition of important factors. First, we know that the limit for $\beta \to \infty$ must be equal to 1 and exactly match the solution of the classic OT problem. Secondly, it contains information about both the network topology and the weight distribution. Third, it is observable in real data. The dependence of m_{MST} from β is shown in Fig. 4a and onward, where we plot it as a function $\ln \beta$. This choice is consistent with the mapping proposed in [33]. Moreover, because the effective range of the control parameter depends on the total weight of the network, we introduce a rescaled parameter

$$\hat{\beta} = K\beta = \beta \sum_i s_i / L^2 \tag{8}$$

where $K = \sum_i s_i/L^2$ sets the overall scale of the weights and represents their average value.

In our numerical experiments, we consider square (bipartite) matrices of varying linear dimension L (from 64 up to 4096), allowing us to investigate how system size influences the phases of the model. In Fig. 4a we observe that the order parameter m_{MST} transitions from near-zero values to $m_{\mathrm{MST}} = 1$ at increasingly larger $\hat{\beta}$ for bigger system sizes, signaling a clear phase transition. Here, the cost matrix $C_{i\alpha}$ has elements drawn from a uniform probability distribution in the range $[0, 1]$ and the strengths are Gaussian with mean $5 \cdot 10^{-5} L$ and standard deviation $10^{-4}\sqrt{12L}$.

As the system size L increases, the order parameter m_{MST} grows noticeably steeper, hinting at a saturation in the thermodynamic limit. To quantify this behavior, we define the transition value $\hat{\beta}_{\mathrm{t}}(L) := \arg\max_{\hat{\beta}} \left[\frac{\partial\, m_{\mathrm{MST}}(\hat{\beta}, L)}{\partial \ln \hat{\beta}} \right]$, following [34], and plot it in Fig. 4b. We then introduce the asymptotic transition value $\beta_{\mathrm{t}}^* \equiv \lim_{L\to\infty} \hat{\beta}_{\mathrm{t}}(L)$ and examine two possible scenarios: one in

which β_t^* diverges with the system size (unbounded power law), and another in which it converges to a finite limit (bounded power law). While fitting the diverging case with a power-law growth yields a reasonably strong adjusted R^2 of about 0.95, the saturating fit aligns almost perfectly with the data, reaching an adjusted R^2 close to 1. Consequently, our results suggest that $\ln \beta_t^* \simeq 3.15(9)$ is finite.

Fig. 4 highlights additional features of the transition in our numerical simulations, suggesting that it is non-critical in the sense that no divergences appear in other observables. In panel (c), for instance, the order parameter evaluated at the transition point seems to converge to a finite value.

Even more telling is the fact that the first derivative of the order parameter remains finite—no divergence is observed—implying that in the thermodynamic limit, the slope of the transition curve remains bounded.

Despite this non-critical character, the transition can still be usefully characterized by examining the behavior of m_{MST} on either side of $\hat{\beta}_t$. To that aim we call $\hat{\beta}_<$ ($\hat{\beta}_>$) the values of $\hat{\beta}$ before (after) $\hat{\beta}_t$. For $\hat{\beta}_< \ll \hat{\beta}_t$, the expected weights $\langle w_{i\alpha} \rangle$ are nearly uniform, consequently, m_{MST} decreases as a power law, acting roughly as $m_{\mathrm{MST}}(\ln \hat{\beta}_<) \simeq 2/L$. In contrast, for $\hat{\beta}_> \gg \hat{\beta}_t$, m_{MST} approaches a size-independent value that depends on $\hat{\beta}$, ultimately tending toward 1 as $\hat{\beta} \to \infty$. Thus, in the thermodynamic limit, the system remains "decoupled" from the cost function for $\hat{\beta} < \hat{\beta}_t$, only to move into a regime where it systematically converges to an OT–like solution for $\hat{\beta} > \hat{\beta}_t$.

For the sake of completeness, we test our results with scaling of transitions that are not critical. We propose a new control parameter $B = \ln(\hat{\beta}/\ln(L))$ that apparently allows for a curve collapse of the m_{MST} and its derivative, as shown in Fig. 5. The curves collapse is consistent with the observation of a limit curve for the order parameter (dashed in green) that is proportional to $1 - \exp(-\alpha(B - B_C))$ for $B > B_C$. The transition does not present a divergence in the first derivative of the order parameter, but shows a slope of $\alpha = 0.846(7)$ after the "critical" point B_C. Indeed, the exponential fit function is maximum in B_C, as expected by empirical behavior.

Different cost and strength distributions

Most of the preceding analyses assumed a specific setup in which the cost functions were uniformly sampled between 0 and 1, and the node strengths were drawn from a Gaussian distribution. However, to gain a more comprehensive understanding of the dense-to-sparse transition and to test the

robustness of our algorithm, we also examined the model under various cost and strength distributions — showing that our results are not limited to the specific case of uniform costs and Gaussian strengths. This step is essential for assessing whether the key features of the transition, particularly its nature, are robust to changes in these distributions, and thus whether the model can reliably capture real-world scenarios, where costs and constraints can be highly heterogeneous and only partial information about them may be available.

We first examined the case where the cost remains uniformly distributed, but the node strengths follow a power-law distribution. Specifically, we used a probability density function $p(x) \sim x^{-4}$, although similar numerical experiments we tested with other power-law exponents revealed no notable qualitative differences. This setup does demand more computational resources because obtaining solutions across different $\hat{\beta}$ values can sometimes require fine-tuning of the simulation parameters. We also examined the case of uniform distributed costs, but with node strengths following a truncated log-normal distribution, with parameters $\mu = 1$ and $\sigma = 0.5$ and with a lower bound of 1 in the support. This has allowed for a similar effect on the strength distribution as the power-law distribution, but with a minor computational effort.

The results, illustrated in Fig. 6, suggest that the transition is qualitatively similar to the case where the strengths are Gaussian-distributed. Moreover, estimating the transition point in the thermodynamic limit yields $\ln \beta_t^* \sim 3.1(3)$ for power-law distributed strengths and $\ln \beta_t^* \sim 3.4(4)$ for log-normal strengths. Both estimates are consistent with the gaussian-distributed strengths case (see Fig. 4b), where we found $\ln \beta_t^* = 3.15(9)$. Although the behavior of the order parameter and its derivative at the transition point is not sufficiently sensitive to decisively distinguish between divergent or saturating fits, our findings are broadly in line with those from the previous setup.

In the second scenario, we vary the cost distribution while retaining Gaussian-distributed node strengths. Because costs must be bounded from below, we cannot simply use standard distributions like the unbounded Gaussian often employed in random matrix theory. Moreover, as the system size L increases, uniform distributions tend to produce many cost values that are extremely close — with average spacing scaling as L^{-2} — which can cause numerical issues, especially in the lower end of the spectrum. In particular, for large L, small cost gaps may create shallow gradients in the Free Energy

landscape for configurations that are not true minima; this complication impedes a straightforward transition to the sparse phase. Simply rescaling the uniform distribution does not resolve this problem, since it amounts to a trivial shift of the control parameter.

Instead, we transform the uniform distribution via a power-law function that depends on L:

$$C_{i\alpha} \;=\; x^{\frac{1}{\log_2(L)}}, \quad x \in [0,1], \tag{9}$$

so that each new cost variable follows a Beta distribution whose parameters are $\log_2(L)$ and 1. This transformation mainly affects the lower end of the spectrum distribution, preserving a finite gap between minimal cost values as L grows. As expected, the resulting behavior differs from the simple uniform-cost case, owing to the explicit dependence of the cost distribution on L, as illustrated in Fig. 7. Although the infinite-size limit of the m_{MST} curves still vanishes below $\ln \hat{\beta}_{\mathrm{t}} \approx 0.4(7)$ and then rises to 1 above it, the finite-size curves exhibit intersection points that shift the transition to smaller values of $\hat{\beta}$. The critical point $\hat{\beta}_t$ goes to zero as the system's size increases (panel b), as well as the value of the order parameter $m_{\mathrm{MST}}(\hat{\beta}_t)$ (panel c), although the two fit, orange and blue lines, have very similar adjusted R^2. Even in this setting, the asymptotic value of $m'_{MST}(\hat{\beta}_t)$ shows a finite value (panel d).

Discussion

The model presented in this study is grounded in an information-theoretic framework similar to maximally entropic null models, but with an added cost term that acts like an external field, hence modifying the standard Entropy into a Free Energy. Our model offers a systematic way to analyze intermediate stages of optimization in bipartite networks, capturing both the uniform *dense* phase, where interaction weights are distributed broadly, and the *sparse* phase, where weights condense onto a minimal set of links, converging toward the spanning tree typical of OT solutions. This framework is particularly relevant for domains in which mutualistic networks exhibit strong link preferences, concentrating most of the total interaction mass on a select few edges. Within such systems, one can leverage the MST mass share, m_{MST}, as an order parameter that tracks the system's transition between these two extremes.

By tuning the control parameter β, the model smoothly moves from a maximum-entropy description of bipartite weighted configuration models—where costs play no role—to a near-OT state in which low-cost edges dominate. Notably, the transition between dense and sparse regimes does not appear to be *critical* in the statistical physics sense, as we observe no divergences in the conventional thermodynamic indicators. Instead, the transition is signaled by the rise of m_{MST}, which vanishes below a threshold β_{t} and grows steadily above it, eventually saturating as β becomes large. Finite-size analysis confirms the results are robust in the thermodynamic limit of large system size. Numerical studies consistently confirm this picture, demonstrating that it persists across various strength distributions (e.g., Gaussian and power-law) and cost matrices, including those whose values are adjusted to address finite-size effects. These results underscore the robustness of the transition, suggesting that the underlying mechanism of weight condensation onto cost-efficient links holds broadly rather than relying on a specific distributional choice.

Interestingly, although the MST structure itself is central to defining m_{MST}, the observed transition is not primarily driven by a topological rearrangement of links. Instead, it reflects a smooth redistribution of weights that eventually "locks in" a small number of cost-favorable edges. As β decreases, these edges lose their advantage, causing weights to diffuse but not necessarily maintaining the OT spanning tree topology. In many practical systems, such sub-optimal configurations emerge from a combination of constrained optimization and other "noise-like" processes that deviate from a complete OT solution. Hence, our model underscores how a partial optimization mechanism, operating under known or estimated cost functions, can naturally coexist with the myriad factors that hold real networks short of the perfectly optimized state.

From an applied perspective, this framework can be used as a null model for tasks such as link prediction, network reconstruction, and statistical validation. If the cost matrix is known or can be reliably estimated, one can deduce the degree of optimization by measuring β. Conversely, if β can be inferred—e.g., via external data or by calibrating the MST mass share—then the cost structure can be approximated or constrained accordingly. Such capabilities are particularly valuable in fields like ecology (e.g., plant–pollinator interactions – see a first tentative approach in [35]) and economics (e.g., trade networks), where sub-optimal strategies may be adaptive responses to uncertain environments or system-level trade-offs. By providing a lens on how

sub-OT mechanisms play out in realistic settings, this model offers a unifying perspective that connects maximal-entropy bipartite null models with the classical OT framework—ultimately helping us quantify, interpret, and predict the patterns found in complex bipartite networks.

Methods

Derivation of the Maximum Free Energy Ensemble

We analyze a canonical ensemble $\mathcal{G}$ of $N \times M$ undirected, weighted bipartite graphs G. Each one is represented by a biadjacency matrix whose elements are continuous real numbers representing edge weights $w_{i\alpha}(G) \in (0, \infty)$. The ensemble is equipped with a probability measure $P(G)$, which will depend on a set of parameters, the Lagrange multipliers $\{\mathbf{t}, \boldsymbol{\theta}\}$ and the coupling term β, characterizing the model. To quantify how well this ensemble represents a given real-world network, we define the expected value for any observable $\pi(G)$ as:

$$\langle \pi \rangle = \sum_{G \in \mathcal{G}} \pi(G)\, P(G) \tag{10}$$

Therefore, the ensemble's expected value of π weights each graph's value $\pi(G)$ by the probability of observing that graph. Our goal is to determine the parameters $\{\mathbf{t}^*, \boldsymbol{\theta}^*\}$ ensuring that the ensemble's expected statistics match empirical values drawn from an observed network G^* at each value of β.

In this work, we impose local constraints on node strengths on each layer, and we call those fixed empirical values as s_i^* and σ_α^*. However, we could have also focused on the degree distribution of each node or considered both aspects simultaneously [30, 11]. Our choice leads to the following constraints, which are already presented in the main text in Equations. (3):

$$\langle s_i \rangle = \textstyle\sum_{G \in \mathcal{G}} P(G)\, s_i(G) = s_i^* \quad \forall i = 1, \ldots, N \tag{11}$$
$$\langle \sigma_\alpha \rangle = \textstyle\sum_{G \in \mathcal{G}} P(G)\, \sigma_\alpha(G) = \sigma_\alpha^* \quad \forall \alpha = 1, \ldots, M \tag{12}$$

where

$$s_i(G) = \sum_\alpha w_{i\alpha}(G) \qquad \sigma_\alpha(G) = \sum_i w_{i\alpha}(G) \tag{13}$$

are the strength variables of the two layers. Obviously, we are also implying the $\sum_{G\in\mathcal{G}} P(G) = 1$ constraint that ensures the correct probability normalization.

Due to the presence of a cost matrix, we seek a maximum-free energy probability measure consistent with Equations (11) and (12). Therefore we introduce a free energy functional defined as:

$$\mathrm{F}[P] = \mathrm{S}[P] - \beta \mathrm{U}[P], \tag{14}$$

where $\mathrm{S}[P] = -\sum_{G\in\mathcal{G}} P(G)\log P(G)$ is the entropy, and $\mathrm{U}[P] = \sum_{G\in\mathcal{G}} P(G)U(G)$.

We obtain the expression of the probability distribution $P(G)$ by maximizing F subject to the constraints in Equations (11) and (12). In practice we have to perform, and then set to zero, the functional derivative with respect to $P(G)$ of the following expression:

$$-\sum_{G\in\mathcal{G}} P(G)\log P(G) + \gamma\left(1 - \sum_{G\in\mathcal{G}} P(G)\right) - \beta\sum_{G\in\mathcal{G}} P(G)\,U(G) + \\ + \sum_i t_i\left(s_i^* - \sum_{G\in\mathcal{G}} P(G)\,s_i(G)\right) + \sum_\alpha \theta_\alpha\left(\sigma_\alpha^* - \sum_{G\in\mathcal{G}} P(G)\,\sigma_\alpha(G)\right) \tag{15}$$

where t_i and θ_α are the Lagrange multipliers associated with each node in both layers and γ is the Lagrange multiplier linked to the normalization constraint. The derivative of Eq.(15) gives:

$$-\log P(G) - 1 - \gamma - \sum_i t_i\,s_i(G) - \sum_\alpha \theta_\alpha\,\sigma_\alpha(G) - \beta\,U(G) = 0 \tag{16}$$

Therefore, the probability measure that characterize the ensemble is:

$$P_{\text{sub OT}}(G\,|\,\beta, \{\mathbf{t}, \boldsymbol{\theta}\}) = e^{-(\gamma+1)} e^{-\mathcal{H}(G|\{\mathbf{t},\boldsymbol{\theta}\}) - \beta\,U(G)}. \tag{17}$$

We enclose all the contributions relative to the Lagrange multipliers in the following quantity:

$$\mathcal{H}(G\,|\,\{\mathbf{t}, \boldsymbol{\theta}\}) \equiv \sum_i t_i s_i(G) + \sum_\alpha \theta_\alpha \sigma_\alpha(G) = \sum_{i,\alpha} w_{i\alpha}(G)(t_i + \theta_\alpha) \tag{18}$$

The partition function is:

$$Z\,(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}) = e^{1+\gamma} = \sum_{G\in\mathcal{G}} e^{-\mathcal{H}(G|\{\mathbf{t},\boldsymbol{\theta}\}) - \beta\,U(G)}, \tag{19}$$

We can rewrite the expression in a more elegant form, resulting in a Boltzmann-like probability distribution for graphs:

$$
\begin{aligned}
P_{\text{sub OT}}\left(G \,|\, \beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right) \quad &\equiv \frac{e^{-\mathcal{H}(G \,|\, \{\mathbf{t}, \boldsymbol{\theta}\}) \,-\, \beta\, U(G)}}{Z\left(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right)} \\
&= \frac{1}{Z\left(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right)} \, \exp\left[-\beta\, U(G) - \sum_i t_i s_i(G) - \sum_\alpha \theta_\alpha \sigma_\alpha(G)\right] \\
&= \frac{1}{Z\left(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right)} \, \exp\left[-\sum_{i,\alpha} w_{i\alpha}(G) \, \left(\beta\, C_{i\alpha} + t_i + \theta_\alpha\right)\right]
\end{aligned}
\tag{20}
$$

The explicit expression of the partition function

$$
Z\left(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right) = \sum_{G \in \mathcal{G}} \exp\left[-\sum_{i,\alpha} w_{i\alpha}(G) \, \left(\beta\, C_{i\alpha} + t_i + \theta_\alpha\right)\right] \tag{21}
$$

can be analytically computed by considering the symbolical sum over the ensemble of graphs as an integral over the real values of the weights $w_{i\alpha}$:

$$
\begin{aligned}
Z\left(\beta, \{\mathbf{t}, \boldsymbol{\theta}\}\right) &= \int_0^{+\infty} \prod_{i,\alpha} dw_{i\alpha} \, e^{-\mathcal{H}(G|\{\mathbf{t},\boldsymbol{\theta}\}) - \beta\, U(G)} = \\
&= \int_0^{+\infty} \prod_{i,\alpha} dw_{i\alpha} \, e^{-\left(\sum_i t_i s_i + \sum_\alpha \theta_\alpha \sigma_\alpha + \beta \sum_{i,\alpha} w_{i\alpha} C_{i\alpha}\right)} = \\
&= \int_0^{+\infty} \prod_{i,\alpha} dw_{i\alpha} \, e^{-\sum_{i,\alpha} w_{i\alpha} (\beta C_{i\alpha} + t_i + \theta_\alpha)} = \\
&= \prod_{i,\alpha} \int_0^{+\infty} dw_{i\alpha} \, e^{-w_{i\alpha} (\beta C_{i\alpha} + t_i + \theta_\alpha)} = \\
&= \prod_{i,\alpha} \frac{1}{\beta\, C_{i\alpha} + t_i + \theta_\alpha}
\end{aligned}
\tag{22}
$$

This result lets us decompose $P_{\text{sub OT}}\left(G \,|\, \{\mathbf{t}, \boldsymbol{\theta}\}\right)$ in edge by edge terms or as a product over the pairs (i, α):

$$
\begin{aligned}
P_{\text{sub OT}}\left(G\,|\,\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right) &= \prod_{i,\alpha}\left(\beta\,C_{i\alpha}+t_i+\theta_\alpha\right)\;\; e^{\,-w_{i\alpha}(G)\,(\beta\,C_{i\alpha}+t_i+\theta_\alpha)} \\
&\equiv \prod_{i,\alpha}\underbrace{r_{i\alpha}\,e^{-r_{i\alpha}\,w_{i\alpha}(G)}}_{P\left(w_{i\alpha}|r_{i\alpha}\right)}
\end{aligned}
\tag{23}
$$

where we define the rate parameter $r_{i\alpha}=(\beta\,C_{i\alpha}+t_i+\theta_\alpha)$. Thus, the weight of each edge (i,α) in the graph is an independent and exponentially distributed random variable $w_{i\alpha}$. Its probability density function is the conditional probability $P\left(w_{i\alpha}|r_{i\alpha}\right)$, with rate parameter $r_{i\alpha}$. When we compute the expected value $\langle w_{i\alpha}\rangle$ with respect to $P\left(w_{i\alpha}|r_{i\alpha}\right)$ we get $1/r_{i\alpha}$.

To determine the values of the Lagrange multipliers that enforce the strengths to match the empirical ones observed in a real graph G^*, we solve the self-consistent equations obtained by maximizing the log-likelihood with respect to each parameter in the set $\{\mathbf{t},\boldsymbol{\theta}\}$:

$$
\begin{aligned}
&\mathcal{L}(G^*\,|\,\beta,\{\mathbf{t},\boldsymbol{\theta}\}) = \log P(G^*\,|\,\beta,\{\mathbf{t},\boldsymbol{\theta}\}) \\
&= -\mathcal{H}(G^*\,|\,\{\mathbf{t},\boldsymbol{\theta}\}) - \log Z\left(\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right) - \beta\,U(G^*)
\end{aligned}
\tag{24}
$$

Since the energy term does not depend on any Lagrange multiplier, the ML condition states that:

$$
\begin{aligned}
&\left.\frac{\partial\mathcal{L}(G^*\,|\,\beta,\{\mathbf{t},\boldsymbol{\theta}\})}{\partial\theta_\alpha}\right|_{\theta_\alpha=\theta_\alpha^*} = \\
&\left[-\frac{\partial\mathcal{H}(G^*\,|\,\{\mathbf{t},\boldsymbol{\theta}\})}{\partial\theta_\alpha} - \frac{1}{Z\left(\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right)}\frac{\partial Z\left(\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right)}{\partial\theta_\alpha}\right]_{\theta_\alpha=\theta_\alpha^*} = 0
\end{aligned}
\tag{25}
$$

$$
\begin{aligned}
&\left.\frac{\partial\mathcal{L}(G^*\,|\,\beta,\{\mathbf{t},\boldsymbol{\theta}\})}{\partial t_i}\right|_{t_i=t_i^*} = \\
&\left[-\frac{\partial\mathcal{H}(G^*\,|\,\{\mathbf{t},\boldsymbol{\theta}\})}{\partial t_i} - \frac{1}{Z\left(\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right)}\frac{\partial Z\left(\beta,\{\mathbf{t},\boldsymbol{\theta}\}\right)}{\partial t_i}\right]_{t_i=t_i^*} = 0
\end{aligned}
\tag{26}
$$

As shown in Eq.(18), all terms coupled to the Lagrange multipliers that appear in $\mathcal{H}(G^*\,|\,\{\mathbf{t},\boldsymbol{\theta}\})$ take the form of a dot product. So we obtain:

$$\left.\frac{\partial \mathcal{H}(G^* \,|\, \{\mathbf{t}, \boldsymbol{\theta}\})}{\partial \theta_\alpha}\right|_{\theta_\alpha = \theta_\alpha^*} = \sigma_\alpha(G^*) \tag{27}$$

$$= -\frac{1}{Z(\beta, \{\mathbf{t}, \boldsymbol{\theta}\})} \sum_G \sigma_\alpha(G)\, e^{-\mathcal{H}(G|\{\mathbf{t},\boldsymbol{\theta}\}) - \beta\, U(G)} = \langle \sigma_\alpha \rangle$$

$$\left.\frac{\partial \mathcal{H}(G^* \,|\, \{\mathbf{t}, \boldsymbol{\theta}\})}{\partial t_i}\right|_{t_i = t_i^*} = s_i(G^*) \tag{28}$$

$$= -\frac{1}{Z(\beta, \{\mathbf{t}, \boldsymbol{\theta}\})} \sum_G s_i(G)\, e^{-\mathcal{H}(G|\{\mathbf{t},\boldsymbol{\theta}\}) - \beta\, U(G)} = \langle s_i \rangle$$

Focusing on Eq.(28) solely, we replace the sum over G by an integral over all $w_{j\gamma}$ and we get that:

$$\frac{1}{Z} \int_0^\infty \Big[\sum_\alpha w_{i\alpha}\Big] \exp\Big[-\sum_{j,\gamma} w_{j\gamma}\,(\beta\, C_{j\gamma} + t_j + \theta_\gamma)\Big] \prod_{j,\gamma} dw_{j\gamma} \tag{29}$$

Because the exponential factorizes,

$$\sum_\alpha \int_0^\infty w_{i\alpha} \prod_{j,\gamma} \Big[e^{-\,(\beta\, C_{j\gamma} + t_j + \theta_\gamma)\, w_{j\gamma}}\, dw_{j\gamma} \Big] \tag{30}$$

we can rewrite the integral as:

$$\sum_\alpha \int_0^\infty w_{i\alpha}\, e^{-\,[\beta\, C_{i\alpha} + t_i + \theta_\alpha]\, w_{i\alpha}}\, dw_{i\alpha} \prod_{(j,\gamma)\neq(i,\alpha)} \int_0^\infty e^{-\,(\beta\, C_{j\gamma} + t_j + \theta_\gamma)\, w_{j\gamma}}\, dw_{j\gamma} \tag{31}$$

Each α term in the integral over i, α gives as a result $1\,/\,(\beta C_{i\alpha} + t_i + \theta_\alpha)^2$ while the integral over each j, γ term yields:

$$\prod_{(j,\gamma)\neq(i,\alpha)} \frac{1}{\beta\, C_{j\gamma} + t_j + \theta_\gamma} \;=\; \frac{Z}{\frac{1}{\beta\, C_{i\alpha} + t_i + \theta_\alpha}} \tag{32}$$

Putting everything together and repeating such computations also for Eq.(27) we have that:

$$s_i^* = \sum_\alpha \frac{1}{\beta\, C_{i\alpha} + t_i^* + \theta_\alpha^*} \qquad \forall i = 1, \ldots, N \tag{33}$$

$$\sigma_\alpha^* = \sum_i \frac{1}{\beta\, C_{i\alpha} + t_i^* + \theta_\alpha^*} \qquad \forall \alpha = 1, \ldots, M \tag{34}$$

Convexity of the optimization problem

Solving Equations (33) and (34) is equivalent to directly computing the gradient of the log-likelihood function while keeping the constraints fixed. This approach effectively transforms the problem into an optimization task. The convexity of the optimization problem described in Equation (4) can be assessed by evaluating the Hessian matrix of $\mathcal{L}(G^* \mid \beta, \{\mathbf{t}, \boldsymbol{\theta}\})$. For the sake of simplicity, we will omit all dependencies of this function in the following discussion. The Hessian takes the form of a block matrix where each block contains negative second-order partial derivatives of $\mathcal{L}$ with respect to all possible combinations of the Lagrange multipliers:

$$\frac{\partial^2 \mathcal{L}}{\partial t_j \, \partial t_k} = \tag{35}$$

$$\frac{\partial}{\partial t_j}\left(\sum_\alpha \frac{1}{\beta\, C_{k\alpha} + t_k + \theta_\alpha}\right) = \begin{cases} -\sum_\alpha \dfrac{1}{\left(\beta\, C_{k\alpha} + t_k + \theta_\alpha\right)^2} & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

Hence, in compact form,

$$\frac{\partial^2 \mathcal{L}}{\partial t_j \, \partial t_k} = -\,\delta_{jk} \sum_\alpha \frac{1}{\left(\beta\, C_{k\alpha} + t_k + \theta_\alpha\right)^2} \tag{36}$$

In other words, the blocks of the Hessian matrix of $\mathcal{L}$ corresponding to the second derivatives with respect to the Lagrange multipliers within the same layer are diagonal. Similarly, we have that:

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_\beta \, \partial \theta_\gamma} = -\,\delta_{\beta\gamma} \sum_i \frac{1}{\left(\beta\, C_{i\beta} + t_i + \theta_\beta\right)^2} \tag{37}$$

$$\frac{\partial^2 \mathcal{L}}{\partial t_j \, \partial \theta_\gamma} = -\,\frac{1}{\left(\beta\, C_{j\gamma} + t_j + \theta_\gamma\right)^2} \tag{38}$$

Implementation of the numerical solution

Here we show how to implement an optimization procedure to determine the Lagrange multipliers $\mathbf{t}$ and $\boldsymbol{\theta}$ by directly optimizing the log-likelihood function rather than explicitly solving the self-consistent equations. The key idea is to leverage gradient-based optimization to find the optimal values of $\mathbf{t}$ and

$\boldsymbol{\theta}$ that enforce the expected strength constraints. The metrics `perc_error` is defined to track how well the row and column constraints are satisfied: it computes the percentage of "mass" of the graph that misses the strength constraint. Instead of explicitly solving the self-consistent Equations (34) and (33) for $\mathbf{t}$ and $\boldsymbol{\theta}$, the approach minimizes the negative of the log-likelihood

$$\mathcal{L}(G^* \,|\, \beta, \{\mathbf{t}, \boldsymbol{\theta}\}) = \sum_{i,\alpha} w_{i\alpha}(G^*) \; (\beta \, C_{i\alpha} + t_i + \theta_\alpha) - \sum_{i,\alpha} \log \; (\beta \, C_{i\alpha} + t_i + \theta_\alpha)$$

via Stochastic Gradient Descent (SGD). The log-likelihood function serves as the loss function, whose gradients are computed using PyTorch [29] automatic differentiation, and the parameters $\mathbf{t}$ and $\boldsymbol{\theta}$ are updated iteratively. In this procedure, each t_i and θ_α are initialized as $2/s_i^*$ and $2/\sigma_\alpha^*$, respectively, and the learning rate is set high at the outset to allow for faster convergence, then it gradually decreases when the Loss function increases instead. The optimization continues up to a maximum number of iterations, unless the error falls below 10^{-3}, in which case it terminates early. If the solution has not converged after the specified number of steps, the process halts and saves the current values of $\mathbf{t}$ and $\boldsymbol{\theta}$ to start a new optimization. This gradient-based method avoids explicitly inverting equations and provides a straightforward means to enforce the required constraints on the row and column sums by allowing the optimizer to drive the system toward the correct parameters $\mathbf{t}$ and $\boldsymbol{\theta}$.

The choice of using Pytorch SGD to solve the model was reached after considering three different algorithms, each with distinct advantages and limitations.

Stochastic Gradient Descent with momentum offers good consistency and is particularly suitable for complex or highly nonlinear cost functions, due to its adaptability. However, its convergence becomes significantly slower as the problem size increases. The time complexity of each step of the SGD algorithm is governed by the computation of the log-likelihood and of its gradient, both of which are $O(NM) = O(N^2)$. Other steps in the process are $O(N + M) = O(N)$, so they can be ignored. Regardless of the system's dimension, with larger values of β, the problem becomes ill-conditioned and the convergence is slower (i.e. more steps are needed).

Fixed-point iterations, on the other hand, are much faster and highly scalable [30], making them attractive for large-scale problems. Yet, in this case they tend to be less stable and more prone to oscillations, especially in

heterogeneous regimes (when the strengths are not Gaussian, for example). This lack of consistency could potentially be mitigated by incorporating updates with momentum. Moreover, with higher values of β, the algorithm becomes ill-conditioned faster than SGD. The time complexity of each step of the fixed-point algorithm is governed by the computation of the contraints' estimates, that are $O(NM) = O(N^2)$, as in the case of SGD.

Finally, Newton's method with Schur complement (analogous to the method used by Koehl et al. in [36]) exhibits better consistency than fixed-point methods and is generally reliable for moderate-size problems. Nevertheless, it tends to be computationally expensive and seems to struggle with more intricate cost landscapes, where convergence is not always guaranteed. Moreover, with higher values of β, the algorithm becomes ill-conditioned faster than SGD. The time complexity of each step is governed by matrix-vector multiplications, of order $O(NM) = O(N^2)$, as in the case of SGD.

Of course, each of these methods has parameters to be tuned, and their effectiveness is bound to the efficacy in finding the optimal values for these parameters. In this sense, more fine tuning is certainly possible to optimize the solution of these models.

As said above, the best method we found to solve our model was PyTorch's version of SGD, run on a NVIDIA GeForce RTX 3060 GPU with 8 GB VRAM. The memory usage of a single run of the SubOT model for a system with $L = 8192$, Gaussian strengths, and uniform cost is 2322.66 MB. This is far enough from the bottleneck, which means that, in principle, it is possible to solve larger systems on such machines. Nevertheless, we chose to stop at the scale reached due to computation time and because it would have been outside of the scope of this work. A more comprehensive discussion of the complexity of the algorithm can be found in Fig. 8.

Require: A set of matrix dimensions `dims`; coupling costant range `betas`; maximum number of steps `num_steps`; initial learning rate `lr0`; momentum for SGD `m0`; maximum weight `maxweigth`.
Ensure: Optimized parameters $\mathbf{t}$, $\boldsymbol{\theta}$ stored for each `dim` and each β.

for `dim` **in** `dims` **do**
  generate `betas`
  generate cost matrix `cost` as random in $[0, 1)$
  transform `cost` ▷ when required
  generate a random matrix `M` in $[0, \texttt{maxweigth}]$ of shape (`dim`,`dim`) ▷ to compute the strengths
  $r \leftarrow$ row-sums of `M`
  $c \leftarrow$ column-sums of `M`
  $t \leftarrow 2/r$ ▷ Element-wise division
  $\theta \leftarrow 2/c$ ▷ Element-wise division
  for `beta` **in** `betas` **do**
    set up optimizer `SGD` with `lr` $\leftarrow$ `lr0`, momentum $\leftarrow$ `m0`
    initialize `increase`$\leftarrow 0$, `patience`$\leftarrow 10$, `last_loss`$\leftarrow \infty$
    for `step` $\leftarrow 0$ **to** `num_steps` **do**
      optimizer.zero_grad()
      `loss` $\leftarrow \mathcal{L}(t, \theta, r, c,$ `cost`, `beta`$)$
      loss.backward()
      optimizer.step()
      `perc_err` $\leftarrow$ `perc_error`(`beta`, `cost`, t, θ, d, u)
      if `loss` $>$ `last_loss` **then**
        `increase` $\leftarrow$ `increase` $+ 1$
      end if
      `last_loss` $\leftarrow$ `loss`
      if `increase` $\geq$ `patience` **then**
        reduce learning rate by factor 0.9
        `increase` $\leftarrow 0$
      end if
      if `perc_err` $< 10^{-3}$ **then**
        save $(\mathbf{t}, \boldsymbol{\theta})$
        break
      end if
      if `step` $=$ `num_steps` **then**
        save $(\mathbf{t}, \boldsymbol{\theta})$ ▷ to start a new opt from the current t and θ
        break
      end if
    end for
  end for
end for

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon request.

Code availability

The codes developed and used for the simulations and analyses presented in this study are available from the corresponding author upon request.

Acknowledgements

D.M., A.P. and R.P. acknowledge the financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title "WECARE – WEaving Complexity And the gReen Economy" – CUP 20223W2JKJ by the Italian Ministry of University and Research (MUR).

Author contributions

L.B., D.M., and R.P. carried out the numerical simulations. L.B., F.S., and A.P. developed the analytical framework. The analysis was designed by L.B., D.M., A.P., G.C., and F.S. All authors contributed equally to the writing and revision of the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] C. Villani. *Optimal transport: old and new*. Springer Science & Business Media, 2009. DOI: https://doi.org/10.1007/978-3-540-71050-9.

[2] L. V. Kantorovich. "On the Translocation of Masses". In: *Journal of Mathematical Sciences* 133.4 (Mar. 2006), pp. 1381–1382. ISSN: 1573-8795. DOI: 10.1007/s10958-006-0049-2.

[3] A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, Jan. 2017. ISBN: 9781400883592. DOI: 10.1515/9781400883592.

[4] P.-A. Chiappori et al. "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness". In: *Economic Theory* 42.2 (May 2009), pp. 317–354. ISSN: 1432-0479. DOI: 10.1007/s00199-009-0455-z.

[5] Y. Rubner. "The Earth Mover's Distance as a Metric for Image Retrieval". In: *International Journal of Computer Vision* 40.2 (2000), pp. 99–121. ISSN: 0920-5691. DOI: 10.1023/a:1026543900054.

[6] M. Kusner et al. "From Word Embeddings To Document Distances". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach et al. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 957–966. URL: https://proceedings.mlr.press/v37/kusnerb15.html.

[7] M. Arjovsky et al. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup et al. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 214–223. DOI: 1701.07875. URL: https://proceedings.mlr.press/v70/arjovsky17a.html.

[8] M. L. Cody. "Optimization in Ecology: Natural selection produces optimal results unless constrained by history or by competing goals". In: *Science* 183.4130 (Mar. 1974), pp. 1156–1164. ISSN: 1095-9203. DOI: 10.1126/science.183.4130.1156.

[9] A. J. Bloom et al. "Resource Limitation in Plants-An Economic Analogy". In: *Annual Review of Ecology and Systematics* 16.1 (Nov. 1985), pp. 363–392. ISSN: 0066-4162. DOI: 10.1146/annurev.es.16.110185.002051.

[10] R. A. Brualdi. *Encyclopedia of mathematics and its applications: Combinatorial matrix classes series number 108*. en. Encyclopedia of mathematics and its applications. Cambridge, England: Cambridge University Press, Aug. 2006. DOI: https://doi.org/10.1017/CBO9780511721182.

[11] G. Cimini et al. "The statistical physics of real-world networks". In: *Nature Reviews Physics* 1.1 (Jan. 2019), pp. 58–71. ISSN: 2522-5820. DOI: 10.1038/s42254-018-0002-6.

[12] P. Koehl et al. "Optimal transport at finite temperature". In: *Phys. Rev. E* 100.1 (July 2019), p. 013310. DOI: 10.1103/PhysRevE.100.013310.

[13] M. Stock et al. "Optimal transportation theory for species interaction networks". In: *Ecology and Evolution* 11.9 (2021), pp. 3841–3855. DOI: https://doi.org/10.1002/ece3.7254.

[14] E. Katifori et al. "Damage and fluctuations induce loops in optimal transport networks". In: *Physical review letters* 104.4 (2010), p. 048704. DOI: https://doi.org/10.1103/PhysRevLett.104.048704.

[15] F. Corson. "Fluctuations and redundancy in optimal transport networks". In: *Physical Review Letters* 104.4 (2010), p. 048703. DOI: https://doi.org/10.1103/PhysRevLett.104.048703.

[16] A. Rinaldo et al. "Evolution and selection of river networks: Statics, dynamics, and complexity". In: *Proceedings of the National Academy of Sciences* 111.7 (2014), pp. 2417–2424. DOI: https://doi.org/10.1073/pnas.1322700111.

[17] V. Colizza et al. "Network structures from selection principles". In: *Physical review letters* 92.19 (2004), p. 198701. DOI: https://doi.org/10.1103/PhysRevLett.92.198701.

[18] M. Cuturi. "Sinkhorn distances: lightspeed computation of optimal transport". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 2292–2300. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[19] G. Peyré et al. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends® in Machine Learning Series. Now Publishers, 2019. ISBN: 9781680835519. DOI: https://doi.org/10.48550/arXiv.1803.00567.

[20] R. Sinkhorn et al. "Concerning nonnegative matrices and doubly stochastic matrices". In: *Pacific Journal of Mathematics* 21.2 (May 1967), pp. 343–348. ISSN: 0030-8730. DOI: 10.2140/pjm.1967.21.343.

[21] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4 (May 1957), pp. 620–630. ISSN: 0031-899X. DOI: 10.1103/physrev.106.620.

[22] T. M. Cover et al. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

[23] J. Park et al. "Statistical mechanics of networks". In: *Phys. Rev. E* 70.6 (Dec. 2004), p. 066117. DOI: 10.1103/PhysRevE.70.066117.

[24] G. Bianconi. "The entropy of randomized network ensembles". In: *Europhysics Letters* 81.2 (Dec. 2007), p. 28005. DOI: 10.1209/0295-5075/81/28005.

[25] T. Squartini et al. *Maximum-entropy networks: Pattern detection, network reconstruction and graph combinatorics*. Springer, 2017. DOI: https://doi.org/10.1007/978-3-319-69438-2.

[26] G. Cimini et al. *Reconstructing Networks*. Elements in Structure and Dynamics of Complex Networks. Cambridge University Press, 2021. DOI: 10.1017/9781108771030.

[27] M. Bruno et al. "Inferring comparative advantage via entropy maximization". In: *Journal of Physics: Complexity* 4.4 (Dec. 2023), p. 045011. ISSN: 2632-072X. DOI: 10.1088/2632-072x/ad1411.

[28] D. Garlaschelli et al. "Maximum likelihood: Extracting unbiased information from complex networks". In: *Physical Review E* 78.1 (July 2008). ISSN: 1550-2376. DOI: 10.1103/physreve.78.015101. URL: http://dx.doi.org/10.1103/PhysRevE.78.015101.

[29] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: 32 (2019). Ed. by H. Wallach et al. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[30] N. Vallarano et al. "Fast and scalable likelihood maximization for Exponential Random Graph Models with local constraints". In: *Scientific Reports* 11 (July 2021), p. 15227. DOI: 10.1038/s41598-021-93830-4.

[31] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2003. ISBN: 9780898715347. DOI: 10.1137/1.9780898718003.ch4.

[32] R. Flamary et al. "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8. URL: http://jmlr.org/papers/v22/20-451.html.

[33] A. Gabrielli et al. "Grand canonical ensemble of weighted networks". In: *Phys. Rev. E* 99.3 (Mar. 2019), p. 030301. DOI: 10.1103/PhysRevE.99.030301.

[34] D. P. Landau et al. *A Guide to Monte Carlo Simulations in Statistical Physics*. 4th ed. Cambridge University Press, 2014. DOI: https://doi.org/10.1017/CBO9781139696463.

[35] M. Stock et al. "Optimal transportation theory for species interaction networks". In: *Ecology and Evolution* 11.9 (2021), pp. 3841–3855. DOI: https://doi.org/10.1002/ece3.7254.

[36] P. Koehl et al. "Statistical Physics Approach to the Optimal Transport Problem". In: *Phys. Rev. Lett.* 123.4 (July 2019), p. 040603. DOI: 10.1103/PhysRevLett.123.040603.

Figure 1: **Example of Optimal Transport framework.** Optimal Transport solution for a system of 4 mines that produce coal to be sold to 4 factories. The indices i and α denote the nodes in each layer of the bipartite network. **(a)** The optimal transport plan as a weighted network, where the color and width of each link represent its weight. **(b)** The biadjacency matrix $w_{i\alpha}$ of the same bipartite network, with matching colors for cells. On the side of the matrix we show the strengths of each node, s_i and σ_α.

Figure 2: **Transition between entropy-driven and cost-driven transport regimes. (a)** Conceptual illustration of the transition between transport regimes as the control parameter β varies. The parameter β tunes the relative importance of the energy with respect to entropy in the ensemble distribution. The OT limit **(iii)** corresponds to the optimal sparse transport solution for moving coal from mines to storage facilities. Sub-panels **(i)** and **(ii)** show how introducing an entropic term modifies this transport plan. Initially, in **(ii)**, it relaxes the strict cost-driven structure, leading to an intermediate sub-optimal regime, where entropy and cost compete in shaping the transport network. Ultimately, in **(i)**, it reaches the BiWCM regime, where transport is fully entropy-driven. **(b)** Quantitative illustration of the effect of the control parameter β: it depicts the average transport plan $\langle w_{i\alpha} \rangle$ in a square network of size $L = 64$, with uniformly distributed costs $C_i \sim U(0, 1)$, for different values of the parameter β. The logarithmic colorbar indicates the link weights in the average transport plans. **(i)** For low β, the weights $\langle w_{i\alpha} \rangle$ are homogeneously distributed, as mass spreads across all links while satisfying the imposed constraints. **(ii)** As β increases, an underlying structure emerges: some links begin to accumulate larger weights $\langle w_{i\alpha} \rangle$, causing the network to resemble the tree-like configuration typical of OT. **(iii)** For very large β, most of the weight condenses onto the OT tree. **(iv)** the exact OT solution.

Figure 3: **The weights $w_{i\alpha}$ as a function of $\ln(\hat{\beta})$ highlight the separation between OT links and the remaining ones.** The figure displays two vertical axes: the left y-axis shows the weights of all links in a 64×64 network with uniformly distributed random costs and Gaussian-distributed strengths, while the right y-axis corresponds to the mass share of the maximum spanning tree, both with respect to the rescaled control parameter $\ln(\hat{\beta}) = \ln(K\beta)$ (Eq. 8) where β tunes the relative importance of the energy with respect to entropy in the ensemble distribution, and $K = \sum_i s_i/L^2$ is the average weight of the network. Blue markers represent the weights of links that are part of the OT solution in the $\ln(\hat{\beta}) \to \infty$ limit, while red markers correspond to all other links (dashed lines indicate average weights). The green curve, relative to the right y-axis, shows the corresponding mass share of the maximum spanning tree of the network, m_{MST} (Eq. 7). Data available in SD1.1 and SD1.2.

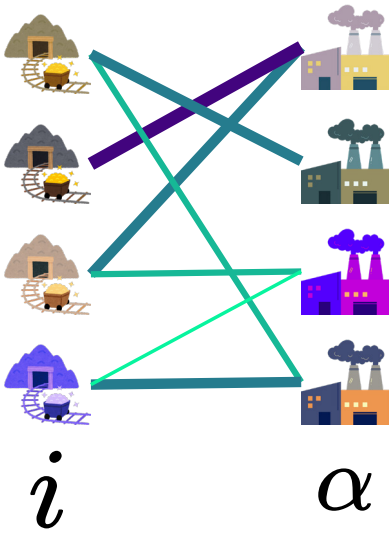
Figure 4: **Sub-optimality transition of the model with uniformly distributed costs and Gaussian-distributed strengths. (a)** Behavior of m_{MST} (Eq. 7) as a function of $\ln \hat{\beta}$ (Eq. 8) for different network sizes L, where L is the number of nodes in each layer of the bipartite network. Shaded areas indicate the standard deviation over at least 100 realizations, except for $L = 4096$, where only 20 realizations were realized due to the computational expense. **(b)** Transition point $\ln \hat{\beta}_{\mathrm{t}}$ as a function of L, defined as the point where the derivative of the m_{MST} is maximum. **(c)** Order parameter m_{MST} evaluated at the transition point $\hat{\beta}_{\mathrm{t}}$, as a function of L. **(d)** Maximum derivative of m_{MST}, evaluated at the transition point $\hat{\beta}_{\mathrm{t}}$, as a function of L. In panels **(b)**, **(c)**, **(d)** error bars denote the standard deviation across realizations, and orange and blue lines correspond to power-law and bounded power-law fits, respectively. Legends include the best-fit parameters along with the corresponding adjusted R-squared values. The control parameter β tunes the balance between transport cost and entropy in the ensemble. m_{MST} (Eq. 7) denotes the mass share of the maximum spanning tree, used as the order parameter in **(a)**, and its value (or the value of its derivative) at the critical point $\hat{\beta}_{\mathrm{t}}$ defines the quantities plotted in **(c)** and **(d)**. Data available in SD1.3.

Figure 5: **The transition properties of the model with a rescaling of the β parameter. (a)** The mass share of the maximum spanning tree m_{MST} (Eq. 7) using the rescaling of the control parameter from $\ln\hat{\beta}$ (Eq. 8) to $B = \ln(\hat{\beta}/\ln(L))$. The curve at different sizes overlap in a region where the value of the order parameter indicates the system is already transitioned to the dense phase. The limit curve for the order parameter (dashed in green) is $1 - \exp(-\alpha(B - B_{\mathrm{C}}))$ for $B > B_{\mathrm{C}}$ and 0 for $B < B_{\mathrm{C}}$ (the fitted values are $B_{\mathrm{C}} = -0.039(7)$ and $\alpha = 0.846(7)$, where α is the exponent of the limit curve and B_{C} is the rescaled crossover point). Shaded areas indicate the standard deviation over at least 100 realizations, except for $L = 4096$, where only 20 realizations were realized due to the computational expense. The fit has been executed using the $L = 4096$ data, which would be our closest approximation of the limit curve. **(b)** The derivatives of m_{MST} with respect to B. The proposed rescaling of the control parameter affects the finite size scaling. Data available in SD1.3.

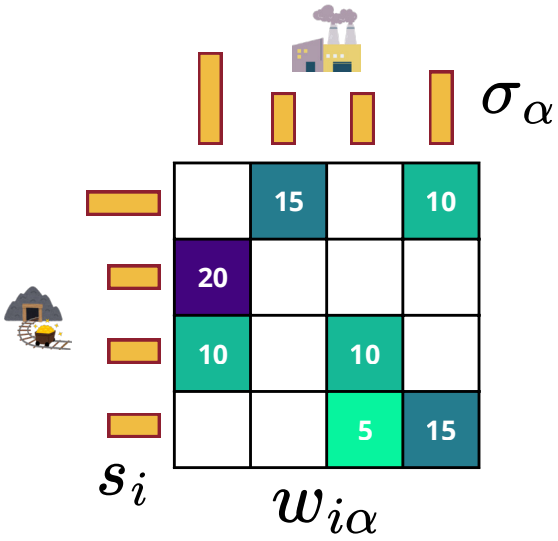
Figure 6: **The transition's properties of the model in the case of Log-Normal-distributed and Power-Law-distributed strengths. (a)** - **(c)** and **(d)** - **(f)** panels refer to the Log-Normal and Power Law distributions, respectively. **(a)** and **(d)** The mass share of the maximum spanning tree m_{MST} (Eq. 7) as a function of $\ln\hat{\beta}$ (Eq. 8), for different sizes L, where L is the number of nodes in each layer of the bipartite network. Each line is the average of at least 10 realizations and the shaded areas indicate the standard deviation over the realizations. **(b)** and **(e)** The transition values $\ln\hat{\beta}_{\mathrm{t}}$ (dots) with respect to L. **(c)** and **(f)**, maximum derivative of m_{MST}, evaluated at $\hat{\beta}_{\mathrm{t}}$. The orange and blue lines in panels **(b)**, **(c)**, **(e)**, **(f)** are a power law and a bounded power law fits. Legends include the best-fit parameters along with the corresponding adjusted R-squared values. Data available in SD1.4 and SD1.5.

Figure 7: **The transition's properties of the model in the case of Beta-distributed cost matrices.** **(a)** The mass share of the maximum spanning tree m_{MST} (Eq. 7) as a function of $\ln \hat{\beta}$ (Eq. 8), for different sizes L, where L is the number of nodes in each layer of the bipartite network. Each line is the average of at least 10 realizations and the shaded areas indicate the standard deviation over the realizations. **(b)** The transition values $\ln \hat{\beta}_{\mathrm{t}}$ (dots) with respect to L. **(c)** The value of the order parameter at the transition point with respect to the layer size L. **(d)** Maximum value of m'_{MST}, evaluated at $\hat{\beta}_{\mathrm{t}}$. The blue line and the red line are a power law and bounded power law fits respectively. Legends include the best-fit parameters along with the corresponding adjusted R-squared values. Data available in SD1.6.

Figure 8: **Performance Analysis across Dimensions and β.** The analysis was performed for Gaussian strengths and uniform costs using the SGD algorithm introduced in the main text. Error bars indicate the standard deviation over at least 10 realizations, except for the largest network sizes, $L = [6144, 8192]$, where only 3 realizations were realized due to the computational expense. **(a)** The cumulative number of SGD steps required to reach convergence at each value of $\log(\hat{\beta})$ (Eq. 8). The colors represent the dimension L of each layer of the bipartite system. For larger dimensions, the scaling of the complexity is more evident. **(b)** The cumulative time (in seconds) required to reach convergence for the SGD algorithm at each value of $\log(\hat{\beta})$. The colors have the same meaning as in **(a)**. **(c)** The average cumulative time (in seconds) to reach convergence at a fixed value of $\log(\hat{\beta})$ with respect to the layer dimension L. Different colors represent different values of $\log(\hat{\beta})$, whereas a quadratic scaling guide have been plotted to better understand the scaling of these curves. The average time is intended as the total time, divided by the number of β values computed up to that point (the bigger the value of $\log(\hat{\beta})$, more values are needed to reach it.). **(d)** The cumulative time (in seconds) to reach convergence at a fixed value of $\log(\hat{\beta})$ with respect to the layer dimension L. The different colors have the same meaning as in (c), as does the scaling guide. Data available in SD1.7.



(a)



(b)