Bilevel Optimization for Traffic Mitigation in Optimal Transport Networks

Alessandro Lonardi[1,*] and Caterina De Bacco[1,†]

[1]*Max Planck Institute for Intelligent Systems, Cyber Valley, Tübingen 72076, Germany*

Global infrastructure robustness and local transport efficiency are critical requirements for transportation networks. However, since passengers often travel greedily to maximize their own benefit and trigger traffic jams, overall transportation performance can be heavily disrupted. We develop adaptation rules that leverage Optimal Transport theory to effectively route passengers along their shortest paths while also strategically tuning edge weights to optimize traffic. As a result, we enforce both global and local optimality of transport. We prove the efficacy of our approach on synthetic networks and on real data. Our findings on the International European highways reveal that our method results in an effective strategy to lower car-produced carbon emissions.

Introduction.—Transport networks are ubiquitous in nature and engineering, spanning from living organisms to cities and telecommunications. Many of these systems can be modeled by adaptation rules that follow the principle of minimum energy, regulating edge flows to optimize transportation costs. Examples in biology are plants, whose profiles emerge from a trade-off between minimization of hydraulic resistance and carbon cost [1], and leaves, shaped by the interplay of nutrients' transport efficiency and robustness to damage [2–4].

Similarly, adaptation rules have been employed to model traffic flows in urban transportation by jointly minimizing the energy dissipated by the passengers and the construction cost of the infrastructure [5–9]. While these models set forth a first approach to simulate traffic flows using adaptation, they crucially neglect that passengers in a transportation network do not move cohesively to minimize a unique cost. Instead, they choose their routes greedily to maximize their benefit (Wardrop's first principle) [10–12]. As a consequence, transport networks may be very inefficient. It is sufficient to think of a city where two points of interest are joined by a fast highway and by longer secondary roads. If all passengers neglected each other's routes, the highway would get congested, slowing the passengers down. In this case, the system would be globally more inefficient than if, unrealistically, a fraction of the passengers were to take user-suboptimal solutions, and be rerouted onto longer secondary connections to increase the overall network efficiency.

In this work, we propose a set of adaptation equations to find traffic flows that mitigate traffic congestion, considered as a proxy for global efficiency, while still accounting for the greedy nature of passengers. In other words, we design a scheme to trade off the shortest routes for the passengers against longer ones that prevent heavy loads on the edges.

We frame the problem in a bilevel optimization setup, which poses a competition between greedy passengers and a network manager. The passengers minimize their origin-destination path cost (lower-level problem), whereas the network manager guarantees global efficiency by mitigating traffic bottlenecks on edges (upper-level problem), while implicitly accounting for passengers' shortest path. We tackle the optimization problem by alternating Optimal Transport (OT)-inspired adaptation rules for the lower-level optimization, and a Projected Gradient Descent (PGD) scheme for the upper-level optimization.

More in detail, greedy passenger flows are found by solving a dynamical system that governs the evolution in time of edge capacities, which can be thought of as widths of roads, so that they optimally allocate passengers on their shortest paths. Adaptation rules are a well-established mechanism for route assignment on networks [3, 5–9, 13–19] and in continuous domains [20–24]. Here, we propose a model that exploits OT theory to prove that, at convergence, passengers move along the shortest path. Particularly, our dynamical system admits a Lyapunov functional [22] which asymptotically converges to the shortest path (Wasserstein) distance between entry and exit distributions of passengers [14, 15, 25].

Traffic mitigation is performed by minimizing a quadratic loss function that penalizes passengers traveling along edges whose traffic exceeds a prefixed threshold. The minimization problem can be treated analytically by assuming that the network edges are endowed with capacities and weights (resistances) and their flows are the gradient of a scalar potential, as for electrical networks. In detail, we derive a closed-form Gradient Descent (GD) update for the weights, which can be interpreted as the cost that passengers have to pay to travel on the network. In practice, network managers would implement these weights by strategically designing incentives or disincentives, e.g., assigning road tolls, to encourage passengers to relocate from jammed edges. The task of traffic mitigation has been addressed using a variety of methods. These include belief propagation [26–28], adaptive dynamical networks [29], MCMC schemes [30], cellar automata [31, 32], and heuristic routing models [33].

Recently, a bilevel optimization problem similar to the one studied here was solved using message-passing [34]. This method yields an approximate computationally efficient algorithm by imposing a quadratic ansatz on the messages (auxiliary functions that determine passenger flows) and with a distributed update of the weights. While the problem's setting is similar to ours, the methodologies radically differ since we alternate adaptation rules for the capacities with global GD for the weights, whereas message-passing updates flows asynchronously and probabilistically.

We find that our method effectively trades off between traffic mitigation and the shortest routes taken by passengers. Namely, both on synthetic topologies and real roads, it returns optimal transport networks where congestion is heavily reduced. We

argue that this result is highly beneficial for reducing the carbon footprint of roads. Remarkably, we also show that the uncoordinated actions of network manager and passengers can be counterproductive, i.e., they may increase traffic, hence producing an outcome opposite to that initially intended.

Problem.—We take a network $G(V, E)$ where $M \geq 1$ groups of greedy passengers i can travel from origin nodes O^i (one per group), to possibly multiple destination nodes D^i. Numbers of entry and exit passengers are stored in a mass matrix with entries $\tilde{S}^i_v > 0$ for each $v = O^i$, $\tilde{S}^i_v < 0$ for $v \in D^i$, and $S^i_v = 0$ otherwise. We assume that the system is isolated, i.e., that passengers entering the network must also exit. This condition is $\sum_v \tilde{S}^i_v = 0$ for all i. When traveling along an edge, passengers pay a cost $\tilde{w}_e > 0$, and lastly, each edge is equipped with a capacity—width of a road to allocate passengers i—$\tilde{c}^i_e \geq 0$. All the main problem variables have been introduced in their respective units to highlight the physical nature of the problem. However, these can be nondimensionalized as $S = \tilde{S}/S_c$, where S_c are characteristic units (respectively for w and c) to derive scale-independent adaptation rules [35].

Lower-level optimization.—The lower-level problem allows us to find the cheapest routes from O^i to D^i. In order to model traffic flows, we introduce the fluxes F^i_e, specifying the displacement of S^i along and edge e. In analogy with electrical networks, we assume that there exists an auxiliary pressure potential p^i_v on each node v due to index i. With this, we define the potential-based fluxes for all $e = (u, v)$ and i, i.e., Poiseuille's Law, as

$$F^i_e = \frac{c^i_e}{w_e}(p^i_u - p^i_v)\,. \tag{1}$$

Fluxes must obey Kirchhoff's law. We can write it as $\sum_e B_{ve} F^i_e = S^i_v$, where B is a conventionally oriented incidence matrix of the network. Substituting Eq. (1) in Kirchhoff's law, the potential becomes a function of c and w, namely $p^i_v = \sum_u (L^{i\dagger})_{vu} S^i_u$, where $\dagger$ denotes the Moore-Penrose inverse and $L^i_{uv} = \sum_e (c^i_e/w_e) B_{ue} B_{ve}$ are entries of the network weighted Laplacian. With this substitution, also F is a function of only c and w, with an explicit dependence as in Eq. (1), and an implicit one in the potential.

For any fixed set of weights, we write the lower-level problem as

$$J = \sum_e w_e ||F_e||_1 \tag{2}$$

$$\min_{c \geq 0} J\,. \tag{3}$$

The convex OT cost J in Eq. (2) is the sum over M indexes of the w-shortest path costs $J^i = \sum_e w_e |F^i_e|$ [14, 15]. Hence its only minimizer is the overlap of M shortest paths between all O^i and D^i, which can be computed with c using Eq. (1) and Kirchhoff's law.

Upper-level optimization.—The upper-level problem formalizes the task of the network manager of tuning w to mitigate traffic jams triggered by the passengers. We measure traffic by penalizing congested links where $||F_e||_1$ exceeds a threshold $\theta \geq 0$. Conveniently, we introduce $\Delta_e = ||F_e||_1 - \theta$.

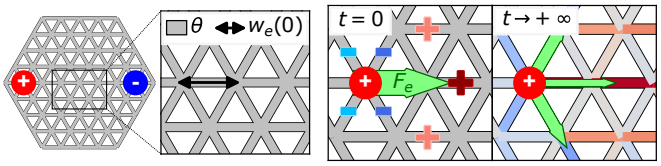

FIG. 1. Bilevel optimization scheme on a lattice. Entry and exit inflows are the red and blue nodes, respectively. Initially, (green) fluxes distribute minimizing the travel cost $w_e(t = 0) = \ell_e$, being the length of an edge. If they exceed θ they get penalized, hence, the network manager tunes the weights to encourage rerouting over more expensive (red), or cheaper (blue) edges (for a companion Fig. [35]).

Analogously to Eqs. (2)-(3), for any set of capacities, the upper-level optimization is

$$\Omega = \frac{1}{2} \sum_e \Delta^2_e\, H(\Delta_e) \tag{4}$$

$$\min_{w \geq \epsilon} \Omega\,, \tag{5}$$

where H is the Heaviside step function. In Eq. (4), all edges with $\Delta_e \geq 0$ are penalized with the square of the traffic in excess, Δ_e. Other objective functions, e.g., the hinge loss can be utilized [34, 36], we do not explore this here. Furthermore, the weights are constrained to be larger than a small $\epsilon > 0$. This means that passengers cannot profit ($w < 0$), or travel for free ($w = 0$). Practically, this ensures that the Laplacian L is well-defined.

Bilevel Optimization.—We combine the two optimization setups into a unique problem. Suppose that the network manager is regularly informed of the passengers' routes, and using such information they tune the weights to mitigate traffic. Moreover, after each update also the passengers reroute accordingly to the updated weights.

Formally, this translates into searching for c and w so that the upper-level objective is minimized, while implicitly accounting for the lower-level, i.e.,

$$\min_{w \geq \epsilon} \Omega(w; \hat{c}) \tag{6}$$

$$\text{s.t. } \hat{c} = \arg\min_{c \geq 0} J(c; w)\,, \tag{7}$$

where the equality in Eq. (7) comes from the convexity of J [14, 15]. In Eq. (6) we make explicit the dependence on w as a variable and on c as a fixed parameter (conversely for Eq. (7)).

Optimal Transport dynamics.—To find the shortest paths required for the lower-level problem, we couple fluxes and capacities with the ODEs

$$\frac{dc^i_e}{dt} = \frac{F^{i\,2}_e}{c^i_e} - c^i_e\,, \tag{8}$$

where fluxes obey Kirchhoff's law. In Eq. (8), edges with high flux enlarge, whereas those where the negative decaying term

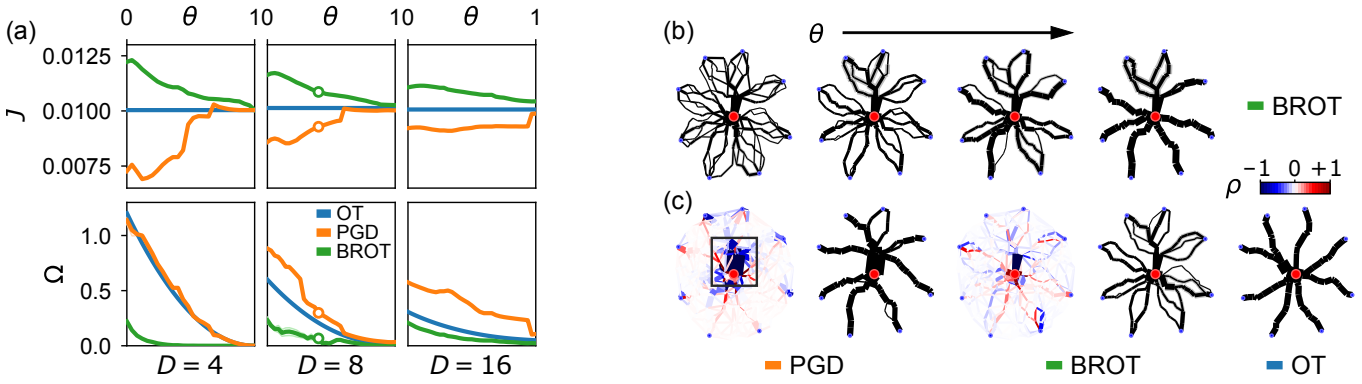

FIG. 2. Overview of the routing schemes. (a) OT cost J and congestion cost Ω against θ. Columns refer to networks with different numbers of destinations $D = 4, 8, 16$. Marked points are the networks in (c). (b) Networks extracted from BROT at different θ for $D = 8$. Edge widths are proportional to the average fluxes in 50 runs of the algorithm. Shaded gray edge contours are fluxes' standard deviations. (c) Cost (left) and flux (right) networks for all methods and $\theta/\theta^\star = 0.4$. Flux networks are as in (b), whereas edges in the cost networks are colored with $\rho = w^\star_X - \ell$ ($X =$ BROT, PGD) and their widths are proportional to the fluxes. The black rectangle frames a region where, in PGD, the network manager naively decreases edge weights, triggering congestion.We conveniently normalize $\theta^\star$ to 1 and ρ in $[-1, 1]$. Figs. for $D = 4, 16$ and detailed network visualizations are in Supp. Mat. [35].

prevails shrink. Crucially, asymptotic solutions converge to the minimum OT cost J in Eq. (2), that is the Wasserstein distance between passengers' entry and exit distributions [35], and whose minimizers are origin-destination shortest paths.

Projected Gradient Descent (PGD).—Minimization of Eq. (6) is performed using vanilla GD, with an additional projection step to enforce $w \geq \epsilon$. Importantly, we can derive a closed-form expression for the gradients $\Psi_e = \partial\Omega/\partial w_e$ [35].

Bilevel optimization scheme.—In order to find the optimal c and w, and hence F, we iterate between Eq. (8) and PGD recursively. The alternating scheme is repeated until J and Ω converge. A diagram outlining the optimization scheme is in Fig. 1, we also provide an open-source implementation (BROT, Bilevel Routing on networks with Optimal Transport) [37].

One could adopt several approaches to solve Eqs. (6)-(7) [38]. However, our model admits a straightforward physical interpretation, i.e., the analogy with electrical or hydraulic networks, which carries significant benefits. We obtain a principled algorithm to solve the lower-level optimization guided by physical laws, and also an efficient numerical implementation that scales robustly with both the system's size and M [5]. Furthermore, our OT-based approach admits a convenient closed-form alternating update of c and w. This is not a straightforward task in standard best-first search algorithms, e.g. Dijkstra's [39], which do not require potential-based fluxes. Finally, contrary to other bilevel optimization schemes [34], our formulation allows to distinguish the paths of different passengers as it outputs fluxes—and thus trajectories—for any individual $i = 1, \ldots, M$.

Experimental setup.—We analyze BROT's optimal networks against two other baseline methods. The first, referred to as OT, consists of finding passengers' shortest paths without any intervention from the network manager. Practically, we assume a unitary cost per unit of length fare, i.e., we set $w = \ell$ with ℓ the Euclidean lengths of the edges, and numerically integrate Eq. (8). The second, referred to as PGD, reflects the scenario of a network manager that tunes w only relying on the shortest paths taken when $w = \ell$, and that neglects how fluxes redistribute while updating w. In practice, this corresponds to running the Projected Gradient Descent only, with initial conditions being $w(0) = \ell + \xi$ and $c^i_e \simeq |F^i_{\text{Dij},e}|$ [35], and then, to integrating Eq. (8) with $w = w^\star_{\text{PGD}}$ being the optimal weights returned by the network manager. Here, ξ is a small zero-sum uniform noise, F_{Dij} are the shortest path fluxes computed with Dijkstra's algorithm, and the approximation arises because, to avoid numerical instabilities, a small non-zero c^i_e is allocated to all edges. We fix BROT's initial conditions to $w(0) = \ell + \xi$ and $c^i_e(0) = S^i_{O^i}$, i.e., we assign uniform cost fares to the edges, and we assume that the network manager is agnostic about passengers' routes at time $t = 0$.

Synthetic experiments.—First, we study a network of size $|V| = 300$, $|E| = 864$, with nodes that are placed uniformly at random in the unitary disk, and edges that are extracted from their Delaunay triangulation. Entry and exit inflows are $S^i_{O^i} = +1$ on an origin node placed at the center of the network, and $S^i_{D^i} = -1/D$, on $D = 4, 8, 16$ destination nodes D^i that are distributed on the edge of the disk. Here $M = 1$, i.e., there is only a single index i. A second experimental setup where $M = 5, 10, 15$ origin-destination pairs are extracted at random from the nodes is discussed in Supp. Mat. [35].

We evaluate J and Ω at convergence for all methods, and for θ ranging from $\theta = 0$, where congestion is reached easily, to $\theta = \theta^\star > 0$, a large value where only a small fraction of edges is congested. Results are in Fig. 2(a).

Since for OT the network manager does not intervene, i.e., w remains fixed at ℓ, J is also constant for all values of θ, and its value is the origin-destination shortest length. The profile of J changes when the network manager influences passengers' routes by tuning the weights. Specifically, for PGD J drops when reducing θ, making it cheaper for the passengers

to move on the network. On the contrary, lower θ corresponds to a larger J for BROT. This behavior seemingly favors an uninformed network manager (PGD) over an informed one (BROT). However, the profile of Ω shows that, even though the traveling cost of PGD is cheaper, all transport networks at convergence are highly congested (large values of Ω). In order to reduce congestion, BROT successfully trades off the cost of traveling against traffic, outputting low values of Ω for all θ, with only a mild increase at θ approaches zero. In Fig. 2(b) we make this clear by showing optimal networks extracted with BROT at different values of θ. For low θ, fluxes heavily distribute giving rise to ramified loopy networks. This structure gets lost while increasing θ, since fluxes progressively concentrate in tree-like structures similar to the OT tree in Fig. 2(c).

Remarkably, when increasing the number of destinations D ($D = 8, 16$) the Price of Anarchy [40] becomes greater for PGD than for OT, i.e., the network manager's intervention increases traffic congestion, having the opposite effect to that intended. In order to better elucidate this phenomenon, we illustrate exemplary networks at convergence in Fig. 2(c). The parameter $\rho = w^\star_X - \ell$ (X = BROT, PGD), expressing the variation of cost for each method, indicates that the uninformed network manager naively—and significantly—decreases the cost of a small fraction of edges (squared in Fig. 2(c)). This encourages the fluxes to largely concentrate on them, thus creating congestion. Conversely, the informed network manager of BROT tunes w by distributing cheaper and more costly edges over the whole network, hence fluxes spread out, leading to a trade-off between J and Ω.

In order to further discern the nature of traffic congestion, we propose two additional metrics. First, the Gini coefficient of the fluxes, $\text{Gini} = \sum_{mn} |x_m - x_n|/2|E|^2\bar{x}$, where $\bar{x} = \sum_e x_e/|E|$ and $x_e = ||F_e||_1$; this measures the statistical dispersion of passengers over the network. $\text{Gini} = 0$ corresponds to uniformly distributed fluxes, and larger Gini to high congestion. Second, we compute the total travel time $T_\theta(s) = \sum_e t_{\theta,e}(s)||F_e||_1$ using an affine latency function for over-trafficked edges [34, 41], namely $t_{\theta,e}(s) = \ell_e(1 + \Delta_e/\theta)/v_\infty$ if $||F_e|| \geq \theta$, and $t_{\theta,e}(s) = \ell_e/v_\infty$ otherwise. Here $v_\infty = 1$ is a (conventionally fixed) free-flow velocity, and s is a sensitivity coefficient to measure traffic congestion. Results for $D = 8$ are in Fig. 3.

The Gini coefficient of PGD fluctuates slightly around the high values attained by the congested shortest path network extracted with OT. For BROT, as θ decreases—hence more flux gets penalized—Gini sharply drops, signaling that more distributed networks are outputted. The total travel time reveals once again that the uncoordinated action of passengers and network manager may be detrimental compared to having no tuning of w. In fact, times for PGD are higher than those for OT. Remarkably, BROT manages to keep $T_\theta(s)$ relatively small for any value of θ and for both low ($s = 1$) and high ($s = 50$) sensitivity. Finally, as θ increases, traffic gradually mitigates, with $\lim_{\theta\to+\infty} T_\theta(s) = T_\infty$ being the travel time for infinite capacities, when all passengers flow freely. From

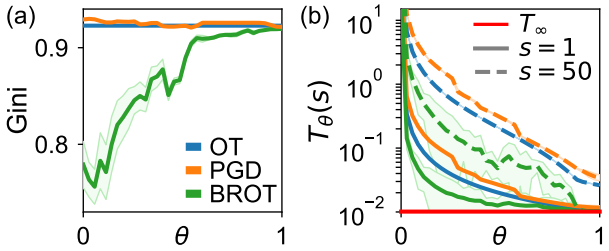

FIG. 3. Measuring traffic congestion, $D = 8$. (a) Gini coefficient against θ. (b) $T_\theta(s)$ against θ. Solid lines correspond to low sensitivity $s = 1$ and dashed ones to $s = 50$, in red we draw T_∞ (free flow). For both plots, shaded areas are standard deviations over 50 realizations of the algorithms. Figs. for $D = 4, 16$ are in Supp. Mat. [35].

the definition $T_\theta(s)$, one can see that $T_\infty = J_{\text{OT}}$ (solution of Eq. (3) with $w = \ell$).

The E-road network.—We study the methods on a graph extracted from the International European highways (E-road= [42, 43], of size $|V| = 541$ and $|E| = 712$. Entry inflows of passengers are populations of 15 large cities. We assume that all passengers travel from one city to another. Therefore, we set for O^i and $v \in D^i$ (being also an origin node O^j), $\tilde{S}_v = r_v S_{O^i}$, with $r_v = \tilde{S}_{O_j}/\sum_k \tilde{S}_{O^k}$. In this way, cities with high inflows also have large outflows. The total number of passengers to be routed is $\sum_i \tilde{S}_{O^i} \simeq 3 \cdot 10^7$. We choose to fix $\tilde{\theta}$ (opportunely dimensionalized by S_c) so that approximately 43% of the passengers get rerouted from their congested shortest path, found with Dijkstra's and $w = \ell$.

Results are in Fig. 4. We observe that in the shortest path configuration of OT, a large volume of passengers travels between the two most populous cities, Madrid and Berlin, on the southernmost region of the network. The action of the uninformed network manager in PGD is that of heavily increasing the price of the connections to Milan (see Supp. Mat. [35] for a companion Figure displaying the cost variation ρ on the networks). This causes a heavy rerouting of passengers from Madrid to the north, and congests the roads connecting Madrid to Paris, and then from Paris to Berlin. In contrast, on the network outputted by BROT fluxes are more distributed, and traffic mitigation gives a largely ramified road network.

We study the effect that different routing schemes have on the average travel time per passenger. We define it as $\langle \tilde{T}_{\tilde{\theta}}(s) \rangle = \sum_e \tilde{t}_{e,\tilde{\theta}}(s)||\tilde{F}_e||_1 / \sum_e ||\tilde{F}_e||_1$, where $\tilde{t}_{\tilde{\theta}}$ is a dimensionalized latency function computed using $\tilde{\ell}$, the Euclidean distance between cities, and $v_\infty = 100$ (km/hours).

Results for sensitivities $s = 1$ and $s = 5$ (Fig. 4) show that the average travel time found with BROT is substantially lower than that of OT and PGD. Particularly, for low sensitivity BROT's $\langle \tilde{T}_{\tilde{\theta}}(s) \rangle$ is approximately 2 (hours), while PGD's and OT's are 5.7 (hours) and 3 (hours). Here, BROT leads to a reduction in traveled time of approximately 33% and 64% when compared to OT and PGD. This result becomes starker if the sensitivity to congestion increases, in this case BROT reduces $\langle \tilde{T}_{\tilde{\theta}}(s) \rangle$ of 50% compared to OT—from 6.3 (hours)

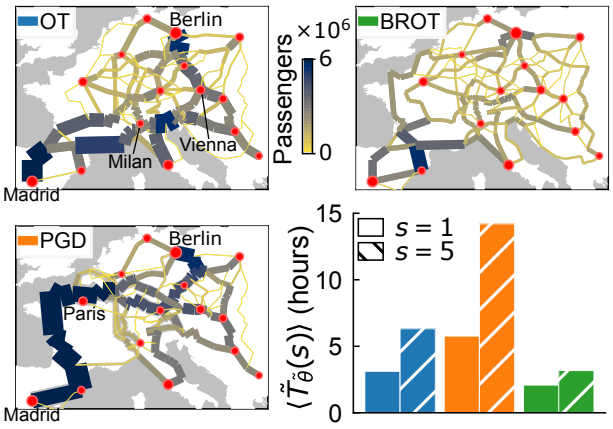

FIG. 4. E-road transport networks. Nodes in red are 15 main cities taken as passenger inflows, their size is proportional to the entry inflows. Edge widths are the total number of passengers $||\tilde{F}_e||_1$, gray shaded areas correspond to standard deviations over 50 realizations of the algorithms where the small noise ξ in $w(0)$ is varied. Colors of $\langle \tilde{T}_{\tilde{\theta}}(s) \rangle$ correspond to those in the networks' legends. Hatch styles are different sensitivities.

to 3.1 (hours)—and of 78% compared to PGD—whose heavy congestion leads to $\langle \tilde{T}_{\tilde{\theta}}(s) \rangle \simeq 15$ (hours).

This analysis elucidates again that the Price of Anarchy (here quantified by the increase in average travel time) is higher when the network manager's intervention is uncoordinated with the passengers (PGD), as opposed to when there is no intervention (OT). Our results also show that BROT provides an effective solution to tune road pricing so that travel times can be drastically lowered.

Conclusion and outlook.—We developed BROT, an algorithm that can trade off traffic mitigation against the shortest path routes. This is done by numerically integrating a set of adaptation rules that solve a bilevel optimization problem where traffic congestion is minimized, while implicitly accounting for passengers' minimizing their travel costs. Our model proved its effectiveness on both synthetic and real data. Particularly, experiments on the E-road network demonstrate how an informed tuning of road tolls—where the network manager factors in passengers' rerouting—is profoundly beneficial to reduce the carbon footprint of roads, since traffic jams, and hence longer travels, critically impact greenhouse gas emissions of vehicles [44–46].

To facilitate practitioners using our algorithms, we provide an open-source code [37].

Acknowledgments.—The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Alessandro Lonardi.

* alessandro.lonardi@tuebingen.mpg.de
† caterina.debacco@tuebingen.mpg.de

[1] L. Koçillari, M. E. Olson, S. Suweis, R. P. Rocha, A. Lovison, F. Cardin, T. E. Dawson, A. Echeverría, A. Fajardo, S. Lechthaler, C. Martínez-Pérez, C. R. Marcati, K.-F. Chung, J. A. Rosell, A. Segovia-Rivas, C. B. Williams, E. Petrone-Mendoza, A. Rinaldo, T. Anfodillo, J. R. Banavar, and A. Maritan, The widened pipe model of plant hydraulic evolution, Proceedings of the National Academy of Sciences **118**, 10.1073/pnas.2100314118 (2021).
[2] E. Katifori, G. J. Szöllősi, and M. O. Magnasco, Damage and Fluctuations Induce Loops in Optimal Transport Networks, Phys. Rev. Lett. **104**, 048704 (2010).
[3] H. Ronellenfitsch and E. Katifori, Global Optimization, Local Adaptation, and the Role of Growth in Distribution Networks, Phys. Rev. Lett. **117**, 138301 (2016).
[4] H. Ronellenfitsch and E. Katifori, Phenotypes of vascular flow networks, Phys. Rev. Lett. **123**, 248101 (2019).
[5] A. Lonardi, E. Facca, M. Putti, and C. De Bacco, Designing optimal networks for multicommodity transport problem, Phys. Rev. Research **3**, 043010 (2021).
[6] A. Lonardi, M. Putti, and C. De Bacco, Multicommodity routing optimization for engineering networks, Scientific Reports **12**, 7474 (2022).
[7] A. Lonardi, E. Facca, M. Putti, and C. De Bacco, Infrastructure adaptation and emergence of loops in network routing with time-dependent loads, Phys. Rev. E **107**, 024302 (2023).
[8] A. A. Ibrahim, A. Lonardi, and C. De Bacco, Optimal transport in multilayer networks for traffic flow optimization, Algorithms **14**, 10.3390/a14070189 (2021).
[9] A. A. Ibrahim, D. Leite, and C. De Bacco, Sustainable optimal transport in multilayer networks, Phys. Rev. E **105**, 064302 (2022).
[10] H. Youn, M. T. Gastner, and H. Jeong, Price of anarchy in transportation networks: Efficiency and optimality control, Phys. Rev. Lett. **101**, 128701 (2008).
[11] M. T. Gastner and M. E. J. Newman, Optimal design of spatial distribution networks, Phys. Rev. E **74**, 016117 (2006).
[12] R. Selten, T. Chmura, T. Pitz, S. Kube, and M. Schreckenberg, Commuters route choice behaviour, Games and Economic Behavior **58**, 394 (2007).
[13] A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki, Rules for Biologically Inspired Adaptive Network Design, Science **327**, 439 (2010).
[14] V. Bonifaci, K. Mehlhorn, and G. Varma, Physarum can compute shortest paths, Journal of Theoretical Biology **309**, 121 (2012).
[15] V. Bonifaci, Physarum can compute shortest paths: A short proof, Information Processing Letters **113**, 4 (2013).
[16] A. Tero, R. Kobayashi, and T. Nakagaki, A mathematical model for adaptive transport network in path finding by true slime mold, Journal of Theoretical Biology **244**, 553 (2007).
[17] D. Hu and D. Cai, Adaptation and Optimization of Biological Transport Networks, Phys. Rev. Lett. **111**, 138701 (2013).
[18] J. B. Kirkegaard and K. Sneppen, Optimal Transport Flows for Distributed Production Networks, Phys. Rev. Lett. **124**, 208101 (2020).
[19] V. Bonifaci, E. Facca, F. Folz, A. Karrenbauer, P. Kolev, K. Mehlhorn, G. Morigi, G. Shahkarami, and Q. Vermande, Physarum-inspired multi-commodity flow dynamics, Theoretical Computer Science **920**, 1 (2022).
[20] D. Baptista, D. Leite, E. Facca, M. Putti, and C. De Bacco, Network extraction by routing optimization, Scientific Reports **10**, 088702 (2020).
[21] E. Facca, F. Cardin, and M. Putti, Towards a Stationary Monge-Kantorovich Dynamics: The Physarum Polycephalum Experience, SIAM Journal on Applied Mathematics **78**, 651 (2016).

[22] E. Facca, S. Daneri, F. Cardin, and M. Putti, Numerical Solution of Monge-Kantorovich Equations via a Dynamic Formulation, Journal of Scientific Computing **82**, 68 (2020).

[23] E. Facca, F. Cardin, and M. Putti, Branching structures emerging from a continuous optimal transport model, Journal of Computational Physics **447**, 110700 (2021).

[24] D. Leite and C. D. Bacco, Revealing the similarity between urban transportation networks and optimal transport-based infrastructures (2022), arXiv:2209.06751 [physics.soc-ph].

[25] A. Lonardi, D. Baptista, and C. De Bacco, Immiscible color flows in optimal transport networks for image classification, Frontiers in Physics **11**, 10.3389/fphy.2023.1089114 (2023).

[26] H. F. Po, C. H. Yeung, and D. Saad, Futility of being selfish in optimized traffic, Phys. Rev. E **103**, 022306 (2021).

[27] Y.-Z. Xu, H. F. Po, C. H. Yeung, and D. Saad, Scalable node-disjoint and edge-disjoint multiwavelength routing, Phys. Rev. E **105**, 044316 (2022).

[28] C. H. Yeung, Coordinating dynamical routes with statistical physics on space-time networks, Phys. Rev. E **99**, 042123 (2019).

[29] J. Jiang, X. Wang, and Y.-C. Lai, Optimizing biologically inspired transport networks by control, Phys. Rev. E **100**, 032309 (2019).

[30] V. Colizza, J. R. Banavar, A. Maritan, and A. Rinaldo, Network structures from selection principles, Phys. Rev. Lett. **92**, 198701 (2004).

[31] T. S. Tai and C. H. Yeung, Global benefit of randomness in individual routing on transportation networks, Phys. Rev. E **100**, 012311 (2019).

[32] T. S. Tai and C. H. Yeung, Adaptive strategies for route selection en-route in transportation networks, Chinese Journal of Physics **77**, 712 (2022).

[33] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, Efficient routing on complex networks, Phys. Rev. E **73**, 046108 (2006).

[34] B. Li, D. Saad, and C. H. Yeung, Bilevel optimization in flow networks: A message-passing approach, Phys. Rev. E **106**, L042301 (2022).

[35] See Supplemental Material at [URL will be inserted by publisher], includes [47–50].

[36] J. W. Rocks, H. Ronellenfitsch, A. J. Liu, S. R. Nagel, and E. Katifori, Limits of multifunctionality in tunable networks, Proceedings of the National Academy of Sciences **116**, 2506 (2019).

[37] Bilevel Routing on network with Optimal Transport: BROT.

[38] A. Sinha, P. Malo, and K. Deb, A Review on Bilevel Optimization: From Classical to Evolutionary Approaches and Applications, IEEE Transactions on Evolutionary Computation **22**, 276 (2018).

[39] E. Dijkstra, A note on two problems in connexion with graphs, Numerische Mathematik **1**, 269 (1959).

[40] The price of anarchy measures how much a system degrades due to greedy behavior of its agents. In our case, the global efficiency of the system is measured by Ω, which increases as the network gets more congested by greedy passengers.

[41] T. Roughgarden and E. Tardos, How Bad is Selfish Routing?, J. ACM **49**, 236–259 (2002).

[42] J. Kunegis, KONECT: The Koblenz Network Collection, in *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion (Association for Computing Machinery, New York, NY, USA, 2013) p. 1343–1350.

[43] L. Šubelj and M. Bajec, Robust network community detection using balanced propagation, The European Physical Journal B **81**, 353 (2011).

[44] T. Van Woensel, R. Creten, and N. Vandaele, Managing the environmental externalities of traffic logistics: The issue of emissions, Production and Operations Management **10**, 207 (2001).

[45] F. Kellner, Exploring the impact of traffic congestion on CO2 emissions in freight distribution networks, Logistics Research **9**, 21 (2016).

[46] M. Barth and K. Boriboonsomsin, Traffic Congestion and Greenhouse Gases, ACCESS Magazine **1**, 2 (2009).

[47] M. Muehlebach and M. I. Jordan, On Constraints in First-Order Optimization: A View from Non-Smooth Dynamical Systems, Journal of Machine Learning Research **23**, 1 (2022).

[48] G. H. Golub and V. Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, SIAM Journal on Numerical Analysis **10**, 413 (1973).

[49] L. Kantorovich, Mathematical Methods of Organizing and Planning Production, Management Science **6**, 366 (1960).

[50] C. Villani, *Optimal Transport: Old and New*, Vol. 338 (Springer Berlin, Heidelberg, 2009).

Bilevel Optimization for Traffic Mitigation in Optimal Transport Networks: Supplementary Material (SM)

Alessandro Lonardi[1,*] and Caterina De Bacco[1,†]

[1]*Max Planck Institute for Intelligent Systems, Cyber Valley, Tübingen 72076, Germany*

PROBLEM FORMULATION

When building the model setup, we assume that passengers move greedily according to Wardrop's first principle. This means that, for a given set of weights w, they travel from their origins O^i to their destinations D^i minimizing the OT cost $J = \sum_e w_e ||F_e||_1$. As a consequence, a meaningful comparison of BROT is proposed against two methods. The first, PGD, consists of a scheme where initially only Ω is minimized by a network manager that tunes w while ignoring how passengers reroute while the weights update. We suppose that, at $t = 0$, the network manager knows which would be the paths that the passengers were to take if they moved minimizing J with $w = \ell$—we compute these fluxes with Dijstra's algorithm. In PGD, only at convergence of w the passengers choose on which path to travel. The second scheme is OT, where the passengers find their shortest path for $w = \ell$. Here, the network manager does not intervene on w.

We give an overview of the schemes in Fig. S1, which is a companion Figure of Fig. 1 (main text). Here we consider an entry and an exit node where unitary mass flows in (red plus) and flows out (blue minus) the network. Additionally to BROT, PGD, and OT, we also extract the paths the passengers would take if they did not act greedily. This simulates the unrealistic—for our case of study—situation where the agents on a road network follow the instruction of an oracle, and accept to reroute in order to achieve a social optimum rather than maximizing their own benefit.

The costs J and Ω in Fig. S1 show that if passengers unrealistically moved in a coordinated way, then traffic congestion would be greatly minimized. However, this would cause J to explode. The OT cost J can be minimized by OT and BROT, which in turn trigger traffic congestion. As discussed in the main text, it may happen that congestion is larger PGD than in OT, showing that the uncoordinated actions of passengers and network manager can be detrimental to the global efficiency of the transport network. Only BROT is able to trade off between J and Ω.

The networks in Fig. S1 reflect the costs profiles. Particularly, in the coordinated transport network (purple), passengers distribute over the whole area of the network and keep congestion low. Here, $\rho = w^\star_{\text{PGD}} - \ell$ ($w^\star_{\text{PGD}}$ are the weights tuned by the network manager) shows that the straight line connecting the origin and destination nodes is heavily penalized, while other parallel connections and oblique links branching from the origin and destination become cheaper. In the OT network (blue), passengers simply travel on the shortest straight path. In PGD (orange), passengers split in two separate branches, which however are congested. Only BROT (green) is able to trade off between congestion and total travel cost by outputting a network where fluxes distribute hierarchically to prevent over-trafficking and to minimize the travel cost. Noticeably, the action of the network manager (here quantified by $\rho = w^\star_{\text{BROT}} - \ell$) is less invasive in this scenario, i.e., edge costs need to be varied less to find an optimal configuration.

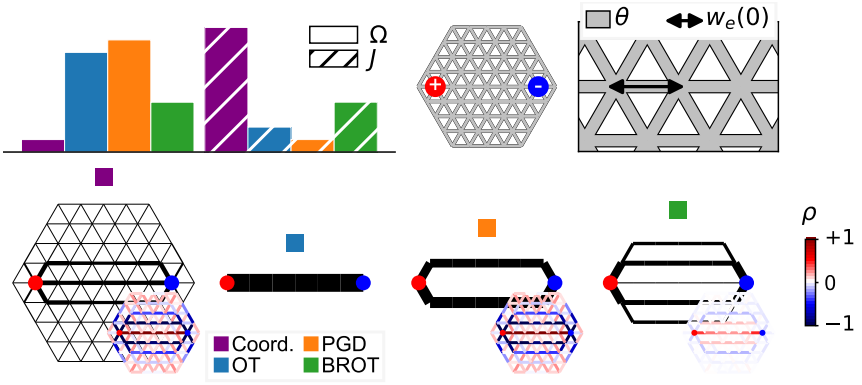

FIG. S1. Bilevel optimization scheme in detail. We plot costs and network topology for different adaptation rules. Histograms and networks labeled with purple, blue, orange, and green squares correspond to coordinated traffic, OT, PGD, and BROT, respectively. Hatch styles are J and Ω. Edges in black are proportional to $||F_e||_1$, while those colored with ρ express the change in w (normalized in $[-1, 1]$).

NONDIMENSIONALIZATION OF THE MODEL

The scale-independent adaptation equations for the evolution of weights and capacities can be derived by rescaling dimension-dependent quantities. We start with the constant coefficients ODEs

$$\frac{d\tilde{c}_e^i}{d\tilde{t}} = \alpha \frac{\tilde{F}_e^{i2}}{\tilde{c}_e^i} - \beta \tilde{c}_e^i \tag{S1}$$

$$\frac{d\tilde{w}_e}{d\tilde{t}} = -\gamma \frac{\partial \tilde{\Omega}(\tilde{w}(t), \tilde{c})}{\partial \tilde{w}_e} \,. \tag{S2}$$

In Eq. (S1) we write a dimension-dependent version of Eq. (8, main text). In Eq. (S2) the gradient flow equation associated with the GD update used to tune the weights. Here, α, β, and γ are constant coefficients with appropriate dimensions. We then choose the following nondimensionalization:

$$t = \tilde{t}/t_c \tag{S3}$$

$$c_e = \tilde{c}_e/c_c \tag{S4}$$

$$w_e = \tilde{w}_e/w_c \tag{S5}$$

$$S_v = \tilde{S}_v/S_c \tag{S6}$$

where t_c, c_c, w_c and S_c are characteristic units. As mentioned in the main text, in order to fix ideas, one could think of such units as time, length, price of tolls applied to roads, and number of passengers, for t_c, c_c, w_c, and S_c, respectively. Substituting Eqs. (S3)-(S6) in Kirchhoff's law yields $F_e = \tilde{F}_e/S_c$, where F are nondimensional fluxes. By recasting all nondimensional variables in Eq. (S1) and Eq. (S2) we get

$$\frac{dc_e^i}{dt} = \alpha \left(\frac{t_c S_c^2}{c_c^2} \right) \frac{F_e^2}{c_e^i} - \beta t_c \, c_e^i \tag{S7}$$

$$\frac{dw_e}{dt} = -\gamma \left(\frac{t_c S_c^2}{w_c^2} \right) \frac{\partial \Omega(w(t), c)}{\partial w_e} \tag{S8}$$

showing that, to recover the adimensional model, we can conventionally set

$$t_c = 1/b \tag{S9}$$

$$c_c/S_c = \sqrt{a/b} \tag{S10}$$

$$w_c/S_c = \sqrt{c/b} \,. \tag{S11}$$

We also notice that $\tilde{J} = \sum_e \tilde{w}_e ||\tilde{F}_e||_1 = (w_c S_c) \sum_e w_e ||F_e||_1$, which is $\tilde{J} = (w_c S_c) J$. Moreover, $\Omega = \tilde{\Omega}/S_c^2$, with $\theta = \tilde{\theta}/S_c$ for an opportunely dimension-dependent capacity threshold $\tilde{\theta}$, and $\tilde{T}(s) = (w_{\mathrm{OT},c} S_c) T(s)$, where $w_{\mathrm{OT},c}$ is the nondimensionalization coefficient used for the constant weights of OT. As a consequence, when comparing the costs Ω and J, and the total traveled time $T_\theta(s)$ between different methods we perform the following nondimensionalization. We fix $w_{X,c} = \sum_e \tilde{w}^\star_{X,e}$ where X is one between OT, BROT, or PGD, and starred weights are those at convergence. Additionally, we set S_c to be the total inflowing number of passengers, i.e., $S_c = \sum_i \tilde{S}^i_{O^i}$. This yields, for any algorithm X,

$$J_X = \frac{\tilde{J}_X}{\left(\sum_e \tilde{w}^\star_{X,e}\right)\left(\sum_i \tilde{S}^i_{O^i}\right)} \quad \Omega_X = \frac{\tilde{\Omega}_X}{\left(\sum_i \tilde{S}^i_{O^i}\right)^2} \quad T_{\theta,X}(s) = \frac{\tilde{T}_{\tilde{\theta},X}(s)}{\left(\sum_e \tilde{w}^\star_{\mathrm{OT},e}\right)\left(\sum_i \tilde{S}^i_{O^i}\right)} \,. \tag{S12}$$

CONNECTION WITH OPTIMAL TRANSPORT

The adaptation rules governing the evolution of the capacities (Eq. (8, main text)) is tightly connected with Optimal Transport theory. In detail, consider the problem formulation proposed in the main text. By fixing $i = 1$ we obtain that the passengers inflows $\mu = S_{O^i}$ and outflows $\nu = -S_{D^i}$ are two atomic probability distributions supported on the origin and destination nodes, respectively, that need to be mapped one into the other by minimizing the transportation cost. This problem, using a standard OT formulation (primal Kantorovich Problem) is the linear program [49]

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{u \in V, v \in V} w_{uv} \pi_{uv} \,. \tag{S13}$$

Here $\Pi(\mu,\nu)$ is the set of transport paths π (expressing the probability of an assignment of passengers on uv) that satisfy the conservation constraints $\sum_v \pi_{uv} = \mu_u$ and $\sum_v \pi_{uv} = \nu_v$. The cost w_{uv} corresponds to the price one needs to pay to move passengers from u to v. Since the transport of passengers is supported on a network, we fix $w_{uv} = +\infty$ if two nodes u, v are not connected. It can be proved [25] that the problem in Eq. (S13) and the minimization problem in Eq. (2, main text) correspond, i.e., the minimization objectives and the search space defined by $\Pi(\mu,\nu)$ and Kirchhoff's law are the same, therefore shortest path fluxes are exactly Optimal Transport paths. Moreover, Optimal Transport fluxes are asymptotic solutions of Eq. (8, main text) [25].

Noticeably, the optimal cost J is exactly the Wasserstein distance between μ and ν, when w_{uv} satisfies the properties of a metric, i.e., (i) symmetry, (ii) vanishing along the diagonal, (iii) triangular inequality [50]. These requirements may not hold in our setup, nevertheless, J can still be interpreted as the cost that passengers pay to move on the network. We also remark that even if disconnected nodes yield an infinite edge cost, the finiteness of Eq. (S13) is guaranteed by assuming the existence of at least one path joining O^i and D^i where passengers can travel.

Optimal Transport paths for $i > 1$ are the overlap of M independent solutions of Eq. (S13), with passengers that move from $\mu^i = S_{O^i}$ to $\nu^i = -S_{D^i}$ for all i.

CLOSED-FORM EXPRESSION OF THE UPPER-LEVEL PROBLEM GRADIENTS

We derive the gradients of the upper-level problem in Eq. (6, main text) in closed-form. Let us start by applying the chain rule to Ω, which is defined as

$$\Omega = \frac{1}{2}\sum_e \Delta_e^2 H(\Delta_e)\,, \tag{S14}$$

with $\Delta_e = ||F_e||_1 - \theta$ and H Heaviside step function. In order to ease notations, for any feasible set of capacities c, we denote all edges for which $\Delta_e \geq 0$ as $w_e \in W$. Conversely, if $\Delta_e < 0$ (hence $H(\Delta_e) = 0$), then $w_e \notin W$. This allows us to write

$$\Psi_e = \frac{\partial \Omega}{\partial w_e} \tag{S15}$$

$$= \sum_{e' \in W} \frac{\partial \Omega}{\partial \Delta_{e'}} \frac{\partial \Delta_{e'}}{\partial w_e} \tag{S16}$$

$$= \sum_{e' \in W} \Delta_{e'} \frac{\partial ||F_{e'}||_1}{\partial w_e}\,. \tag{S17}$$

To manipulate Eq. (S17), we introduce the auxiliary variables $r_e^i = w_e / c_e^i$, that we can use to write compactly all problem's main variables. In particular, we denote fluxes, as variables of r, as $\mathcal{F}_e^i = (\mathcal{P}_u^i - \mathcal{P}_v^i)/r_e^i$. Least-squares potentials are $\mathcal{P}_v^i = \sum_u (\mathcal{L}^{i\dagger})_{vu} S_u^i$, and entries of the Laplacian are $\mathcal{L}_{vu}^i = \sum_e (1/r_e^i) B_{ve} B_{ue}$. With this change of variable, derivatives of the fluxes in Eq. (S17) become

$$\frac{\partial ||F_{e'}||_1}{\partial w_e} = \sum_i \frac{\partial ||\mathcal{F}_{e'}||_1}{\partial r_e^i} \frac{\partial r_e^i}{\partial w_e}\,. \tag{S18}$$

Now, write differences of pressure along the network edges as $\Delta\mathcal{P}_e^i = \mathcal{P}_u^i - \mathcal{P}_v^i$, and when computing the gradients with respect to an edge e, we separate contributions for $e' \neq e$, and for $e' = e$. In particular, we write

$$\frac{\partial ||\mathcal{F}_{e'}||_1}{\partial r_e^i} = \frac{\partial}{\partial r_e^i}\left(\sum_j \frac{1}{r_{e'}^j} |\Delta\mathcal{P}_{e'}^j|\right) \tag{S19}$$

$$= \begin{cases} \dfrac{1}{r_{e'}^i} \mathrm{sgn}(\mathcal{P}_{e'}^i) \dfrac{\partial \Delta\mathcal{P}_{e'}^i}{\partial r_e^i} & \forall e' \neq e \\ -\dfrac{1}{r_{e'}^{i2}} |\Delta\mathcal{P}_{e'}^i| + \dfrac{1}{r_{e'}^i} \mathrm{sgn}(\mathcal{P}_{e'}^i) \dfrac{\partial \Delta\mathcal{P}_{e'}^i}{\partial r_{e'}^i} & \forall e' = e\,. \end{cases} \tag{S20}$$

To simplify notations, we substitute Eq. (S20) into Eq. (S18), and group all terms in one unique expression. This reads

$$\frac{\partial ||F_{e'}||_1}{\partial w_e} = \sum_i \left(\frac{1}{r_{e'}^i} \mathrm{sgn}(\Delta\mathcal{P}_{e'}^i) \frac{\partial \Delta\mathcal{P}_{e'}^i}{\partial r_e^i} - \frac{1}{r_{e'}^{i2}} |\Delta\mathcal{P}_{e'}^i| \delta_{e'e} \right) \frac{\partial r_e^i}{\partial w_e} \tag{S21}$$

with $\delta_{e'e}$ Kronecker delta. Moreover, derivatives of pressure differences can be written making explicit the definition of the least-squares potential, that is,

$$\frac{\partial \Delta \mathcal{P}^i_{e'}}{\partial r^i_e} = \sum_{vu} B_{ve'} \frac{\partial (\mathcal{L}^{i\dagger})_{vu}}{\partial r^i_e} S_u \,. \tag{S22}$$

Now, in order to conclude the derivations, we need to calculate the derivatives of the Laplacian Moore-Penrose inverse in Eq. (S22). This can be done by following the detailed calculation of **??** . We denote their closed-form expression with $\mathcal{Q}^i_{uve} = \partial(\mathcal{L}^{i\dagger})_{vu}/\partial r^i_e$ as defined in Eq. (S37).

Substituting Eq. (S37) into Eq. (S22), and then Eq. (S22) back into Eq. (S21), we obtain

$$\frac{\partial ||F_{e'}||_1}{\partial w_e} = \sum_i \left(\frac{1}{r^i_{e'}} \text{sgn}(\Delta \mathcal{P}^i_{e'}) \sum_{vu} B_{ve'} \mathcal{Q}^i_{vue} S^i_u - \frac{1}{r^{i2}_{e'}} |\Delta \mathcal{P}^i_{e'}| \delta_{e'e} \right) \frac{\partial r^i_e}{\partial w_e} \,. \tag{S23}$$

Making explicit capacities and weights in all terms of Eq. (S23) yields

$$\frac{\partial ||F_{e'}||_1}{\partial w_e} = \sum_i \left(\frac{c^i_{e'} c^i_e}{w_{e'} w^2_e} \text{sgn}(\Delta P^i_e) \sum_{vu} B_{ve'} \Lambda^i_{vue} S^i_u - \frac{c^i_{e'}}{w^2_{e'}} |\Delta P^i_e| \delta_{e'e} \right) \quad \forall e \in E,\, e' \in W \tag{S24}$$

$$\Lambda^i_{vue} = L^{i\dagger}_{ux} L^{i\dagger}_{xv} + L^{i\dagger}_{uy} L^{i\dagger}_{yv} - L^{i\dagger}_{uy} L^{i\dagger}_{xv} - L^{i\dagger}_{ux} L^{i\dagger}_{yv} \qquad \forall e = (x,y) \in W,\, u,v \in V,\, i \in M \tag{S25}$$

$$\Delta P^i_e = p^i_x - p^i_y \qquad \forall e = (x,y) \in W,\, i \in M \,, \tag{S26}$$

which we simplify further by writing

$$\frac{\partial ||F_{e'}||_1}{\partial w_e} = \sum_i \left(\frac{c^i_{e'} c^i_e}{w_{e'} w^2_e} \text{sgn}(\Delta P^i_{e'}) \sum_{vu} B_{ve'} \Lambda^i_{vue} S^i_u - \frac{|F^i_{e'}|}{w_{e'}} \delta_{e'e} \right) \tag{S27}$$

$$= \sum_i \left(\frac{c^i_{e'} c^i_e}{w_{e'} w^2_e} \text{sgn}(\Delta P^i_{e'}) \sum_{vu} B_{ve'} (L^{i\dagger}_{ux} L^{i\dagger}_{xv} + L^{i\dagger}_{uy} L^{i\dagger}_{yv} - L^{i\dagger}_{uy} L^{i\dagger}_{xv} - L^{i\dagger}_{ux} L^{i\dagger}_{yv}) S_u - \frac{|F^i_{e'}|}{w_{e'}} \delta_{e'e} \right) \tag{S28}$$

$$= \sum_i \left(\frac{c^i_{e'} c^i_e}{w_{e'} w^2_e} \text{sgn}(\Delta P^i_{e'}) \sum_{vu} B_{ve'} (L^{i\dagger}_{vx} p_x + L^{i\dagger}_{vy} p_y - L^{i\dagger}_{vx} p_y - L^{i\dagger}_{vy} p_x) - \frac{|F^i_{e'}|}{w_{e'}} \delta_{e'e} \right) \tag{S29}$$

$$= \sum_i \left(\frac{c^i_{e'} c^i_e}{w_{e'} w^2_e} \text{sgn}(\Delta P^i_{e'}) \sum_{vu} B_{ve'} B_{ue} L^{i\dagger}_{vu} \Delta P^i_e - \frac{|F^i_{e'}|}{w_{e'}} \delta_{e'e} \right) \tag{S30}$$

$$= \sum_i \left(\frac{c^i_{e'}}{w_{e'}} \frac{F^i_e}{w_e} \text{sgn}(F^i_{e'}) \sum_{vu} B_{ve'} B_{ue} L^{i\dagger}_{vu} - \frac{|F^i_{e'}|}{w_{e'}} \delta_{e'e} \right) \tag{S31}$$

$$= \sum_i \frac{F^i_e}{w_e} \left(\frac{c^i_{e'}}{w_{e'}} \text{sgn}(F^i_{e'}) G^i_{e'e} - \text{sgn}(F^i_e) \delta_{e'e} \right) , \tag{S32}$$

where we introduced $G^i_{e'e} = \sum_{vu} B_{ve'} B_{ue} L^{i\dagger}_{vu}$. Notice that G yields non-zero terms for all edges in W and not, hence, congestion on an edge e can affect the weight of any other edge of the network. We conclude by substituting Eq. (S32) in Eq. (S17), and finally obtain the gradients of Ω in closed-form. These read

$$\Psi_e = \frac{\partial \Omega}{\partial w_e} = \sum_{e' \in W} \Delta_{e'} \sum_i \frac{F^i_e}{w_e} \left(\frac{c^i_{e'}}{w_{e'}} \text{sgn}(F^i_{e'}) G^i_{e'e} - \text{sgn}(F^i_e) \delta_{e'e} \right) . \tag{S33}$$

A similar result can be found in [34] (Appendix IIIC).

Differentiation of the Laplacian Moore-Penrose inverse

The following calculations can be carried out identically for any index $i \in M$, thus we omit it. We also assume that $\mathcal{L}$ has constant rank, i.e., the network does not disconnect in multiple connected components. With this assumption, we can write ([48], Theorem 4.3) the Laplacian Moore-Penrose inverse derivatives as

$$\frac{\partial \mathcal{L}^\dagger}{\partial r_e} = -\mathcal{L}^\dagger \frac{\partial \mathcal{L}}{\partial r_e} \mathcal{L}^\dagger + \mathcal{L}^\dagger \mathcal{L}^{\dagger\top} \frac{\partial \mathcal{L}^\top}{\partial r_e} \left(I - \mathcal{L}\mathcal{L}^\dagger \right) + \left(I - \mathcal{L}^\dagger \mathcal{L} \right) \frac{\partial \mathcal{L}^\top}{\partial r_e} \mathcal{L}^{\dagger\top} \mathcal{L}^\dagger \,, \tag{S34}$$

where we introduced I, identity matrix of size $|V|$, and $\top$ denotes the transposed of a matrix. The expression in Eq. (S34) can be combined with the network Laplacian properties $\mathcal{L}\mathcal{L}^\dagger = \mathcal{L}^\dagger\mathcal{L} = I - \vec{1}\otimes\vec{1}/|V|$, and $\sum_u \partial\mathcal{L}_{uv}/\partial r_e = \sum_v \partial\mathcal{L}_{uv}/\partial r_e = 0$, with $\vec{1}$ being a $|V|$-dimensional array of ones and $\otimes$ the Kronecker product. This yields

$$\frac{\partial\mathcal{L}^\dagger}{\partial r_e} = -\mathcal{L}^\dagger\frac{\partial\mathcal{L}}{\partial r_e}\mathcal{L}^\dagger\,, \tag{S35}$$

which we further simplify, for all $u, v \in V$ and $e = (x, y) \in W$, as follows:

$$\frac{\partial\mathcal{L}^\dagger_{uv}}{\partial r_e} = -\sum_{u'v'e'}\mathcal{L}^\dagger_{u'u}\frac{\partial}{\partial r_e}\left(\frac{B_{u'e'}B_{v'e'}}{r_{e'}}\right)\mathcal{L}^\dagger_{v'v} \tag{S36}$$

$$= \frac{1}{r_e^2}\left(\mathcal{L}^\dagger_{ux}\mathcal{L}^\dagger_{xv} + \mathcal{L}^\dagger_{uy}\mathcal{L}^\dagger_{yv} - \mathcal{L}^\dagger_{uy}\mathcal{L}^\dagger_{xv} - \mathcal{L}^\dagger_{ux}\mathcal{L}^\dagger_{yv}\right) \tag{S37}$$

$$= \mathcal{Q}_{uve}\,. \tag{S38}$$

NUMERICAL IMPLEMENTATION

Projected Gradient Descent

In order to constrain the weights onto their feasibility set $C = \{w \in \mathbb{R}^{|E|} : w_e \geq \epsilon > 0\}$, we perform a projection step at every GD iteration. Thus, for every discrete time step $n \in \mathbb{N}_0$, we modify the update of the weights as

$$\text{GD: } w_e(n+1) = w_e(n) - \eta\Psi_e(n) \tag{S39}$$

$$\text{PGD: } w_e(n+1) = \underset{x_e\geq\epsilon}{\arg\min}\,|(w_e(n) - \eta\Psi_e(n)) - x_e|\,. \tag{S40}$$

Other projection methods could be used to constrain the weights. Particularly, in our numerical code [37] we also implement the method of Muehlebach and Jordan [47]. This consists of adding a "cleverly-designed" momentum term to vanilla GD, which ensures that, given that $w(0) \in C$, then w will fall in the feasibility region at convergence. Remarkably, this approach has the advantage of being easily adaptable to highly non-linear constraints. However, since the structure of C in our case is simple, we observe that it does not give any numerical benefit compared to PGD. Hence, we opt for the latter. The threshold ϵ has been set to $\epsilon = 0.01 \cdot \min w$.

Initial conditions

Initial conditions for OT, PGD, and BROT are set as follows:

$$\text{OT: } c_e^i(0) = S^i_{O^i} \qquad w_e = \ell_e \tag{S41}$$

$$\text{PGD: } c_e^i = (1-\lambda)|F^i_{\text{Dij},e}| + \lambda S^i_{O^i} \qquad w_e(0) = \ell_e + \xi_e \tag{S42}$$

$$\text{BROT: } c_e^i(0) = S^i_{O^i} \qquad w_e(0) = \ell_e + \xi_e\,, \tag{S43}$$

where $\lambda = 0.95$, F_{Dij} are the shortest path fluxes computed with Dijkstra's algorithm, and ξ is zero-sum small noise, defined as $\xi = 0.1 \cdot \min w(0)(\zeta_0/\max|\zeta_0|)$ where ζ_0 the correspondent mean-centered vector of ζ, $\zeta_e \sim U(0,1)$.

Conditions in Eq. (S41) allow us to compute the shortest path fluxes without the intervention of the network manager, hence with $w = \ell$. In this case, the width of roads at $t = 0$ is set to be uniform and equal to the inflowing passengers $S^i_{O^i}$, since passengers could potentially travel on any edge of the network. In PGD, we suppose that the network manager is—*only initially*—informed about passengers' shortest paths, hence we should initialize the capacities as $c_e^i = |F_e^i|$. This condition comes from the fact that at convergence of Eq. (8, main text) and for optimal solutions of Eq. (2, main text) the scaling $c_e^i \sim |F_e^i|$ holds [5]. In order to prevent numerical instabilities, in Eq. (S42) we assign $c_e^i \simeq |F^i_{\text{Dij},e}|$ to edges that are traversed by shortest path fluxes, and $c_e^i = 0.05 \cdot S^i_{O^i}$ to all the others. The weights are initialized as equal to the lengths, with a small zero-sum noise used to explore the cost landscape of Ω. In PGD, c gets updated only after the full update of the weights by the network manager is performed, then the OT paths are computed for $w = w^\star_{\text{PGD}}$. For BROT, the network manager is initially uninformed about passengers' routes, hence $c_e^i(0) = S^i_{O^i}$. Similarly to PGD, $w(0) = \ell + \xi$.

Implementation Details

In Algorithm 1 we write a pseudocode for the implementation of OT, PGD, and BROT. In Table S1, we provide a detailed list of all parameters used for our experiments. Random seeds for ξ were ranged in $0, 1, \ldots, 49$.

Algorithm 1: BROT, PGD, and OT

Input: network G, inflows S, θ, additional parameters as in Table S1.
Initialize: OT: w, $c(0)$ with Eq. (S41); PGD: $w(0)$, c with Eq. (S42); BROT: $w(0)$, $c(0)$ with Eq. (S43)
while convergence is False **do**
  OT: update c with Eq. (8, main text), $w^\star_{\text{OT}} = \ell$ // The weights remain fixed until convergence
  PGD: update w with Eq. (S40), at convergence fix $w = w^\star_{\text{PGD}}$ // The capacities remain fixed until convergence
    update c with Eq. (8, main text) // The weights are now fixed
  BROT: alternate Eq. (8, main text) and Eq. (S40)
end
Output: optimal weights, capacities, and fluxes $w^\star$, $c^\star$, $F^\star(c^\star, w^\star)$

Network configuration	Parameters			
	θ	T (tot. number of iterations)	ε_J (conv. threshold of J)	ε_Ω (conv. threshold of Ω)
Lattice exps	range(0.0, 0.505, 0.005)	$5 \cdot 10^3$	10^{-6}	10^{-6}
Synth. exps main text				
$D = 4$	range(0.0, 0.13, 0.005)	$5 \cdot 10^3$	10^{-6}	10^{-6}
$D = 8$	range(0.0, 0.072, 0.002)	$5 \cdot 10^3$	10^{-6}	10^{-6}
$D = 16$	range(0.0, 0.033, 0.001)	$3 \cdot 10^3$	10^{-4}	10^{-4}
Synth. exps Supp. Mat.				
$M = 5$	range(0.0, 0.155, 0.005)	$3 \cdot 10^3$	10^{-6}	10^{-6}
$M = 10$	range(0.0, 0.155, 0.005)	$3 \cdot 10^3$	10^{-6}	10^{-6}
$M = 15$	range(0.0, 0.155, 0.005)	$2 \cdot 10^3$	10^{-6}	10^{-6}
E-roads exps	range(0.0, 0.155, 0.005)	$2 \cdot 10^3$	10^{-5}	10^{-5}

TABLE S1. Experimental parameters. With range(x, y, z) we denote evenly spaced arrays that range from x to $y - z$, with steps of size z.

SYNTHETIC EXPERIMENTS

We show additional results that complement those of the synthetic experiments discussed in the main text. The interpretation of these results is fundamentally identical to those already discussed in the manuscript, therefore, here we only present them briefly.

First, Fig. S2, Fig. S3, and Fig. S4 are companion figures to Fig. 2 (main text). These consist of a detailed visualization of the transport networks that we extract with OT, PGD, and BROT for D (number of destination nodes) being $D = 4$, $D = 8$, and $D = 16$. Fig. S5 is a companion Figure to Fig 3 (main text). Here we display the Gini coefficient and the total travel time $T_\theta(s)$ for $D = 4$ and $D = 16$.

All plots from Fig. S6 to Fig. S10 refer to experiments performed on a different origin-destination configuration than that in the main text. Particularly, here we fix $M = 5, 10, 15$ origin-destination pairs that are extracted at random among the network nodes. We build S in such a way that groups of greedy passengers move from each origin to each destination node, potentially having to interact with other types of passengers—indexed by a different i—that travel on the network edges. In other words, traffic congestion on edges can be triggered by the interaction of greedy passengers, e.g., i and j, which may cause $|F^i_e| + |F^j_e| > \theta$. In Fig. S6 (companion of Fig. 2 (main text)), we give an overview of the results produced by the routing schemes. In Fig. S7 we plot the Gini coefficient and the total traveled times $T_\theta(s)$ for all configurations $M = 5, 10, 15$. Finally, in Fig. S8, Fig. S9, and Fig. S10 we plot detailed visualizations of the transport networks extracted with OT, PGD, and BROT.

E-ROAD NETWORK

In Fig. S11 we show companion plots to those in Fig. 4 (main text). Particularly, we color the E-road network with the nondimensionalized Euclidean lengths ℓ, which are used for the initial conditions Eqs. (S41)-(S43). Then, we show the change of cost at convergence of the numerical schemes, i.e., $\rho_X = w_X^\star - \ell$, with $X =$ PGD, BROT. We also display the distribution of passengers on edges at convergence of Dijkstra's, i.e., the configuration based on which the uninformed network manager of PGD tunes the weight. The highlighted value of $\tilde{\theta}$ is that fixed for all numerical experiments on the E-road network.

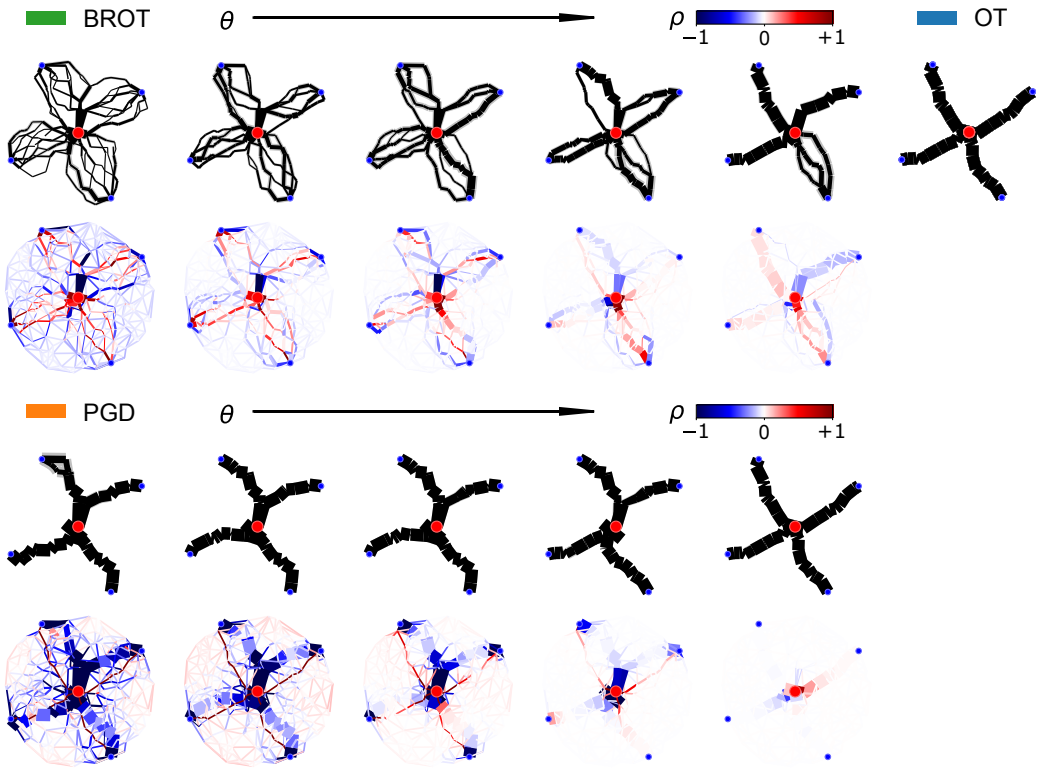

FIG. S2. Detailed visualization of transport networks for different methods. For an extensive caption one can refer to Fig. 2 (main text).

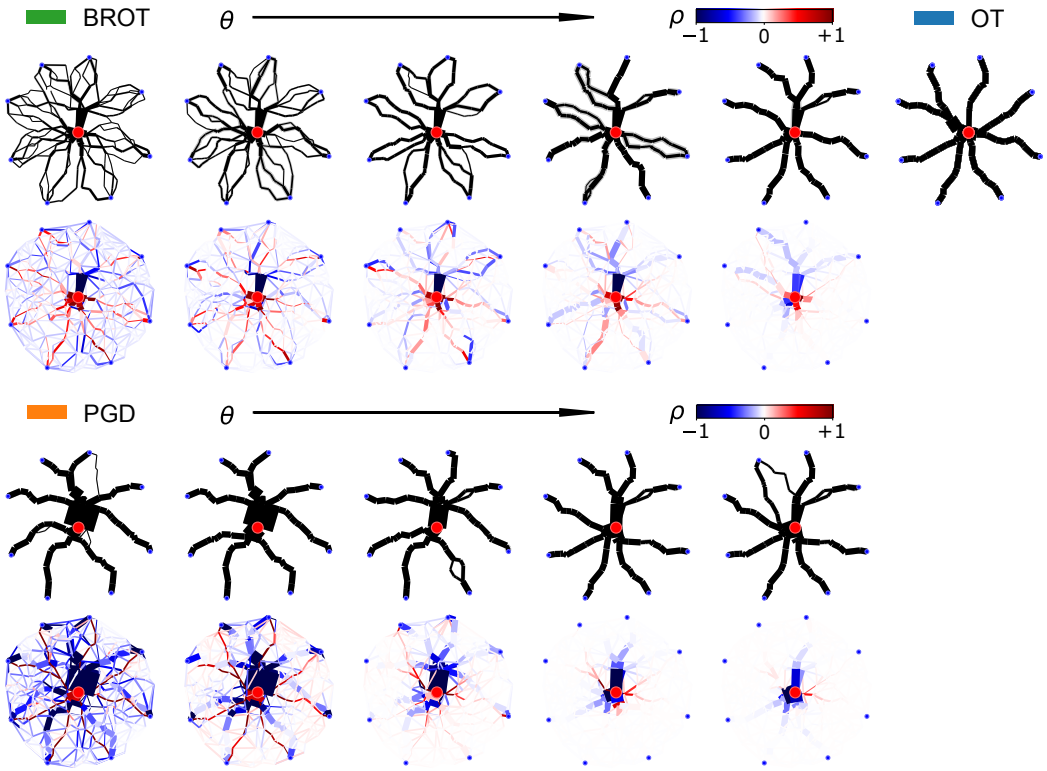

FIG. S3. Detailed visualization of transport networks for different methods. For an extensive caption one can refer to Fig. 2 (main text).

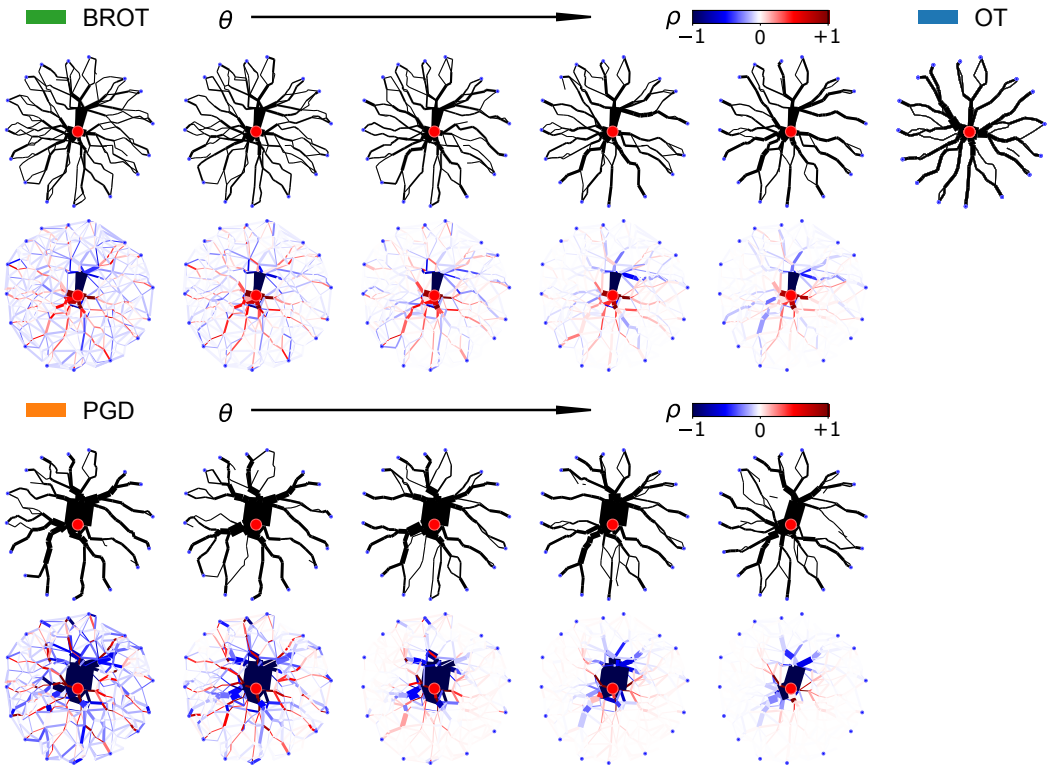

FIG. S4. Detailed visualization of transport networks for different methods. Spurious isolated edges that are not connected to the origin and destination nodes are caused by the adopted coloring scheme, i.e., drawing in black all over a small threshold. For an extensive caption one can refer to Fig. 2 (main text).

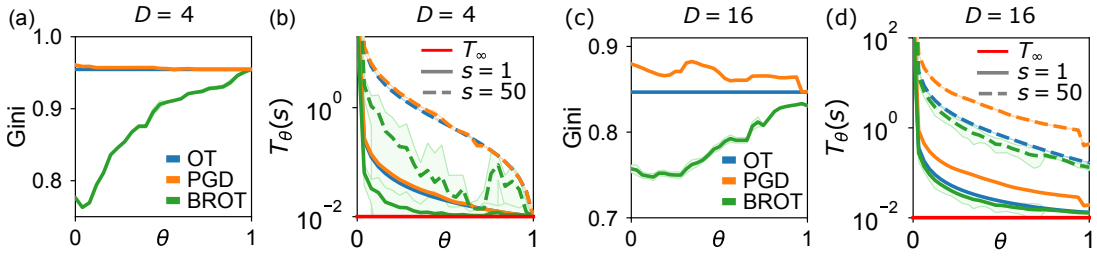

FIG. S5. Measuring traffic congestion, $D = 4, D = 16$. (a,c) Gini coefficient against θ. (b,d) $T_\theta(s)$ against θ. For an extensive caption one can refer to Fig. 3 (main text).

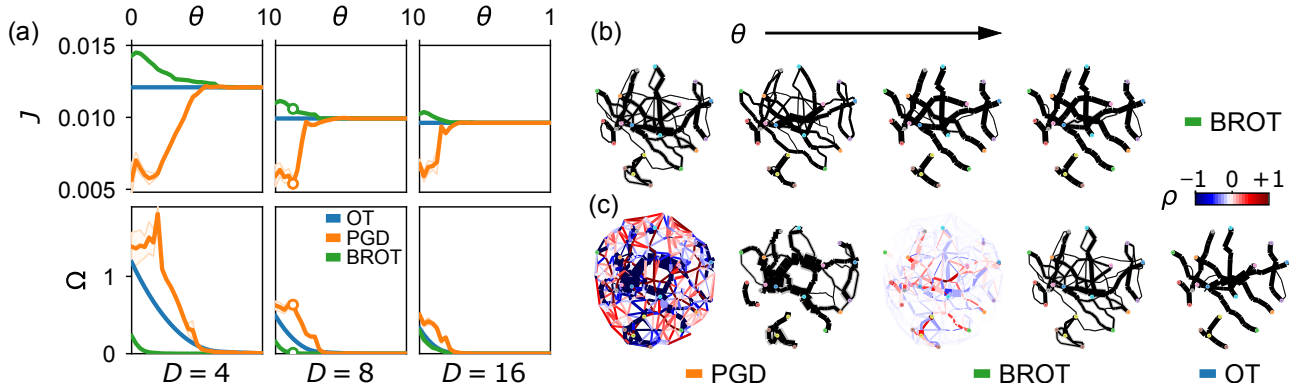

FIG. S6. Overview of the routing schemes for origin-destination pairs. Node colors refer to different indexes i. For an extensive caption of all subplots (a)-(c) one can refer to Fig. 2 (main text).

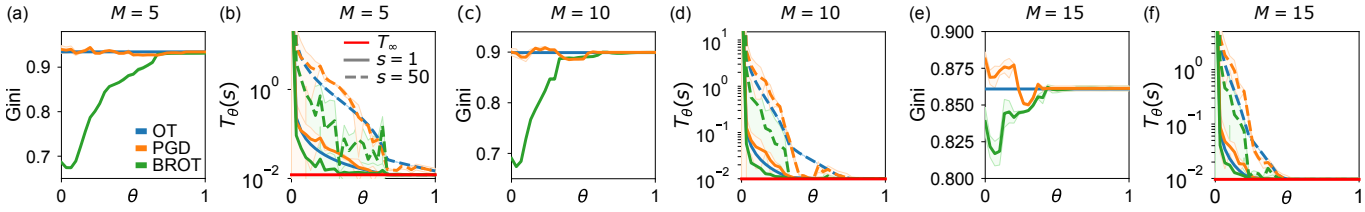

FIG. S7. Measuring traffic congestion. (a,c,e) Gini coefficient against θ. (b,d,f) $T_\theta(s)$ against θ. For an extensive caption one can refer to Fig. 3 (main text).

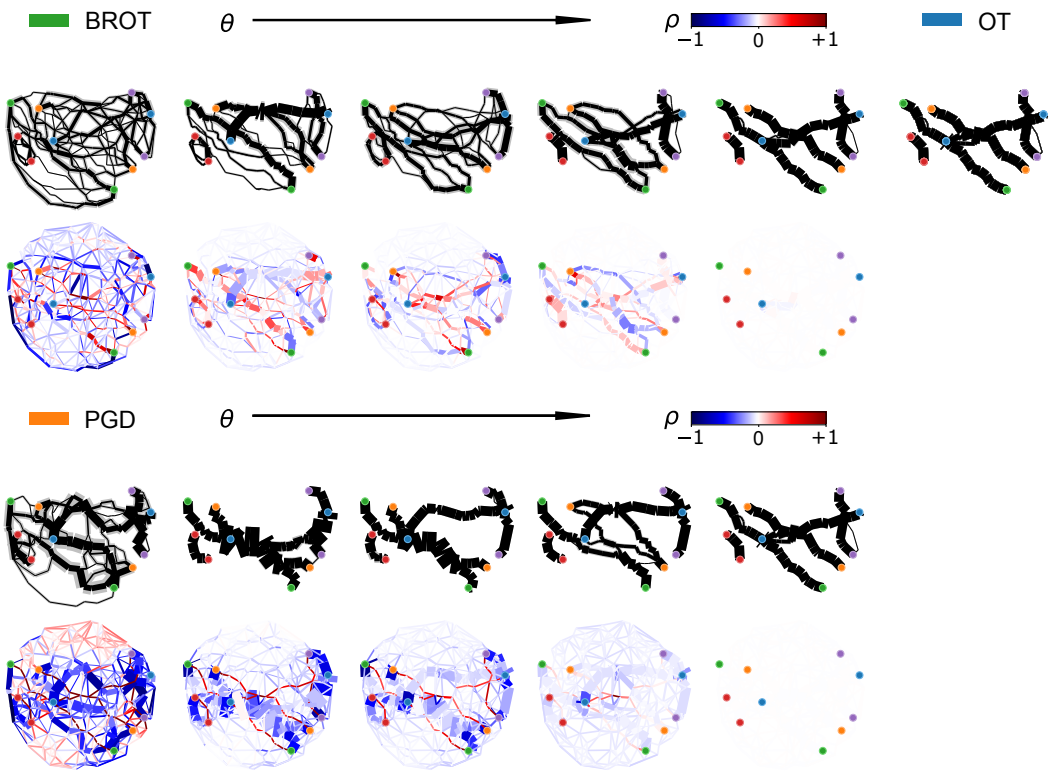

FIG. S8. Detailed visualization of transport networks for different methods. For an extensive caption one can refer to Fig. 2 (main text).

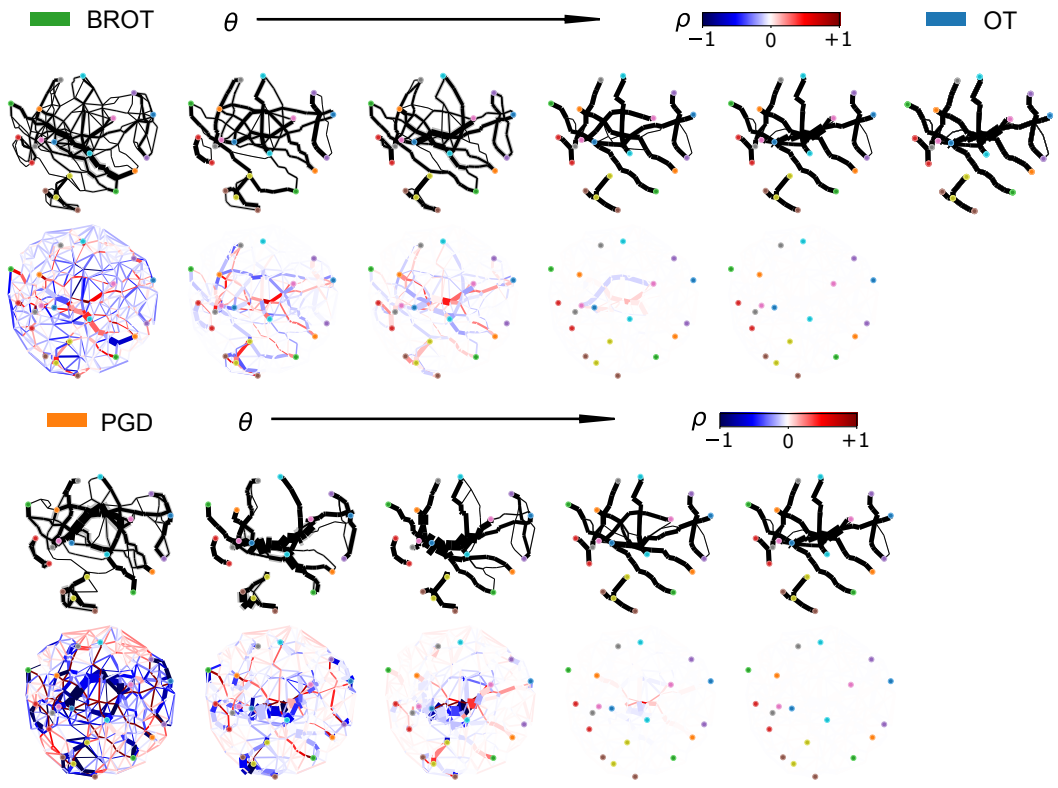

FIG. S9. Detailed visualization of transport networks for different methods. For an extensive caption one can refer to Fig. 2 (main text).

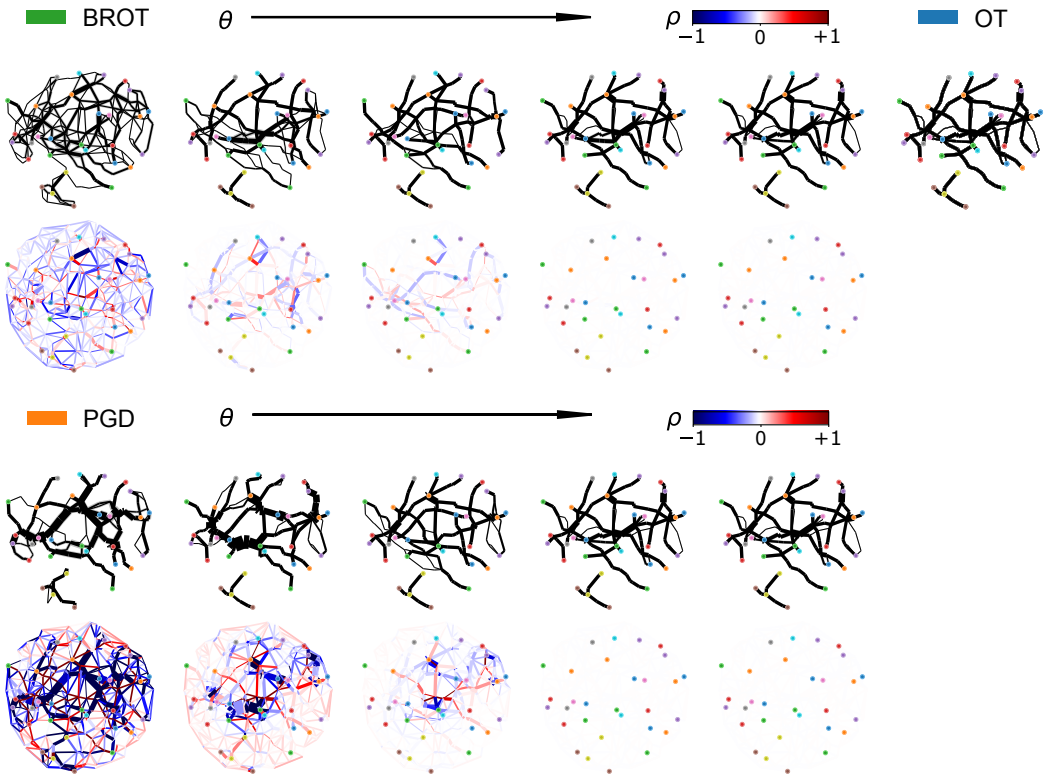

FIG. S10. Detailed visualization of transport networks for different methods. For an extensive caption one can refer to Fig. 2 (main text).

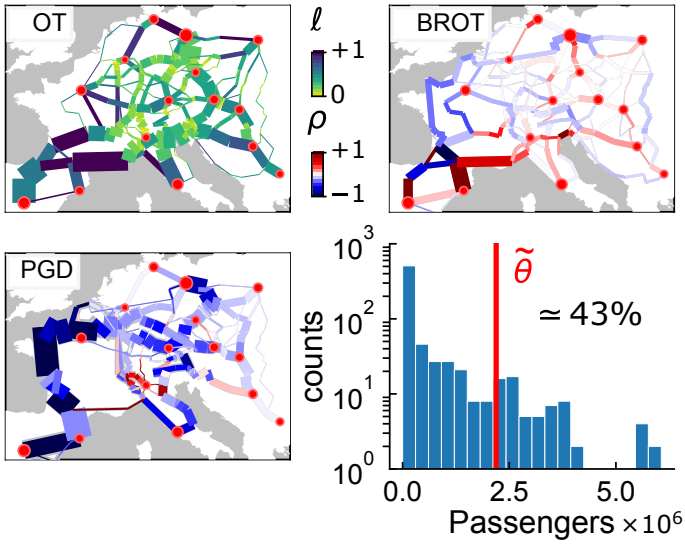

FIG. S11. E-road transport networks, companion Figure. Colors of edges follow the colorbars for ℓ (the Euclidian length between nodes) and for ρ, being the difference in cost between the weights at convergence and their initial configuration. The bottom right histogram is the shortest path distribution, computed with Dijkstra, of the fluxes. In red we mark the value of $\tilde{\theta}$ that has been used for all experiments on the E-road network, which penalizes approximately 43% of the total number of passengers traveling along their shortest path.