Optimal Transport Networks in Spatial Equilibrium*

[Preliminary and Incomplete]

Pablo D. Fajgelbaum
UCLA and NBER

Edouard Schaal
CREI and NYU

First Draft: October 2016

Abstract

We develop a framework to study optimal transport networks in general equilibrium spatial models. We embed an optimal transport problem into a general neoclassical environment with arbitrary many locations arranged on a graph. Goods must be shipped through linked locations subject to congestion in transport. In addition, resources can be invested to lower trade costs in any link. The framework nests the neoclassical models used in international trade and allows for factor mobility. We define the globally optimal transport network as the solution to a social planner's problem of simultaneously choosing the allocation, the gross trade flows across the network, and the investment in each link. If the congestion in transport is strong relative to the returns to infrastructure investments, the planner's is a standard convex optimization problem, guaranteeing convergence of efficient gradient-descent based algorithms. We use the model to characterize the globally optimal transport network in different spatial equilibrium environments and to contrast it with sub-optimal networks. We also implement optimal networks when the planner's problem is non-convex due to increasing returns to network investments, and contrast their properties with convex cases.

*This paper was presented at the Toulouse SED Conference 2016 under the title "Endogenous Transport Networks."

1 Introduction

Trade costs are a ubiquitous force in international trade and economic geography, as they rationalize spatial distributions prices, real incomes, and trade flows. Some central questions involve counterfactuals with respect to trade costs. For example, the gains-from-trade counterfactual comparing trade and autarky scenarios is a canonical theoretical question, as well as the focus of quantitative research summarized by Costinot and Rodríguez-Clare (2013). The shape and quality of transport networks within and between countries are important components of trade costs, and particularly so in developing countries (Limao and Venables, 2001; Atkin and Donaldson, 2015).[1] Partly motivated by the observation that every year the world economy invests massive amounts of resources in transport infrastructure (WorldBank, 2009), recent studies assess the economic impact of infrastructure investments. These studies include both the ex post empirical analyses of actual changes in transport technologies or infrastructure, summarized by Donaldson (2015) and Redding and Turner (2015), and the simulation of changes in trade costs in quantitative spatial setups using estimated models, as reviewed for example by Redding and Rossi-Hansberg (2016).

A current limitation of this burgeoning body of research is the ability to study optimal transport networks. What would be the gains from replacing actual, potentially inefficient networks with the optimal transport network, and how do these gains vary across countries? How does the optimal transport network interact with the sources of comparative advantages, such as differences in relative productivity and factor endowments, vis-a-vis natural geographic features? How does the optimal response of the transport network diffuse local shocks and impact the spatial distribution of economic activity?

In this paper, we develop a framework to study optimal transport networks in standard general-equilibrium trade and economic-geography models. We model a general neoclassical environment with multiple goods and factors in which arbitrary many locations are arranged on a graph. Goods can only be shipped through connected locations; for example, connected locations may correspond to bordering locations in the geographical space. Shipping is subject to decreasing returns, the source of which may be congestion or specific factors in transport technologies.[2] In addition, resources can be invested to increase the capacity of the transport infrastructure in any link (e.g., the number of lanes or the quality of the road). The transport network is defined as the set of capacities across links. The framework nests commonly used neoclassical trade models (e.g., the Ricardian, Armington, and factor-endowment models), and it allows for either a fixed spatial distribution of the primary factors (as in international trade models) or for labor and potentially other factors to be mobile (as in economic geography models).

Solving for the globally optimal transport network is challenging because of dimensionality –the space of all networks is large– and because of potentially increasing returns due to the complementarity between network investments and shipping. Our approach deals with both hurdles.

[1] See and WorldBank (2011) and IADB (2013) for assessments of transport costs in Africa and Latin-America, respectively.

[2] For recent evidence on road congestion in the U.S. see Duranton and Turner (2011).

First, rather than optimizing over the network in the competitive equilibrium, we tackle the planner's problem of simultaneously choosing the allocation, the gross trade flows, and the capacity investments in every link.[3] Second, we convexify the social planner's problem through continuous infrastructure investments (instead of a binary choice of whether to build or not). Third, to deal with increasing returns we introduce curvature in the planner's problem through decreasing returns in transport activities: the more is shipped between connected locations, the higher is the marginal cost of shipping an extra unit of any commodity. The welfare theorems hold, so that the planner's optimal allocation and gross trade flows given the network capacities correspond to a competitive equilibrium.

The first main implication of these assumptions is a reduction in dimensionality. Using the first-order conditions with respect to capacity investments avoids a search in the space of all networks. Because the investment is continuous, the optimal infrastructure investment between connected locations i and j is determined as function of goods' shadow values –the equilibrium prices in the decentralized allocation– in locations i and j alone. Instead of searching in the space of all networks, these properties allow us to search in the considerably smaller space of equilibrium prices. The globally optimal network then results from the combination of the optimization conditions in every link. The second main implication is convexity: if congestion in transport is strong relative to the returns to network-capacity investments, the planner's problem is a standard convex optimization problem. Therefore, besides being a realistic force, congestion in transport ensures the sufficiency of the first-order conditions from the planner's problem and guarantees the convergence of efficient gradient-descent based algorithms to find the solution.

These properties hold regardless of the number of goods, sectors, and factors, and regardless of whether labor is fixed or mobile, as long as the model lies within the neoclassical realm. In the absence of strong enough congestion relative to the returns to network building, the planner's problem is not convex and there may be multiple local maxima. We implement these non-convex cases combining the necessary first-order conditions from the planner's problem with standard simulated annealing methods. The ensuing networks in these cases are sparser, and the capacity distribution is more skewed towards fewer but wider "highways". In some specific cases, we show that the optimal network is a tree, so that every pair of locations is connected through only one route.

In this preliminary version of our paper, we first develop the framework, characterize its key main properties, and discuss the computational implementation. Then, we illustrate its uses by computing the optimal transport network in different spatial equilibrium environments. Our applications start from the simplest case nested within our model, an endowment economy without labor mobility and only one traded and one non-traded good in a symmetric graph. Then, we progressively move to more complex cases with randomly located cities, multiple sectors, labor mobility, geographic accidents, and increasing returns to network building. Our examples illustrate

[3] The problem of choosing the gross trade flows is related to the minimum cost flow problem and to the optimal transport problem on a network studied in the optimal transport literature, as we discuss in the literature review.

the differences between optimal and suboptimal networks in terms of regional effects and aggregate welfare, as well as the impact of the optimal network on the spatial distribution of economic activity. We also illustrate the contrast between the globally optimal networks in convex cases, where the decreasing returns to shipping offset the increasing returns due to network building, and the locally optimal networks in non-convex cases.

Our paper is related to a recent quantitative literature in international trade and spatial economics. Eaton and Kortum (2002) and Anderson and Van Wincoop (2003) developed quantitative versions of the Ricardian and Armington trade models, respectively, allowing counterfactuals with respect to trade costs in multi-country competitive equilibrium. Recently, these frameworks have been applied to spatial setups allowing for factor mobility and trade frictions within countries, among other forces; e.g., see Allen and Arkolakis (2014), Redding (2016a), Fajgelbaum et al. (2015), and Caliendo et al. (2014), among others.

Recent studies build upon these gravity frameworks by introducing a least-cost route optimization problem of traders, and then undertake counterfactuals with respect to infrastructure. Along these lines, Allen and Arkolakis (2014) simulate the aggregate welfare effect of the U.S. highway system, Redding (2016a) compares the impact of infrastructure changes in models with varying degrees of increasing returns, Alder (2016) simulates counterfactual transport networks in India, Nagy (2016) measures the historical impact of railroad growth in the U.S. on the spatial distribution of economic, and Sotelo (2016) simulates the impact of highway investments on agricultural productivity in Peru. In an urban setup, Redding (2016b) develops a framework to study innovations to urban transport systems and applies it to Berlin. More recently, Allen and Arkolakis (2016) adds Fréchet shocks to the costs of traders choosing least-cost routes, leading to analytic expressions for the welfare effect of local infrastructure improvements, and evaluate innovations to the U.S. highway system. Together with Allen and Arkolakis (2016), their model allows to compute the maximal welfare gradient with respect to local changes in infrastructure around the initially observed data. These studies rest upon methodological insights from Dekle et al. (2008), who showed how gravity trade models can be used to undertake counterfactuals relative to observed trade data.[4]

Relative to this literature, the distinctive aspect of our framework is that it allows to compute the globally optimal transport network in a general neoclassical environment. The framework nests the Ricardian and Armington models commonly used in the recent quantitative economic geography literature, canonical factor-endowment models such as the Hecksher-Ohlin and specific-factors models, and standard urban economics models such as Rosen-Roback (Roback, 1982). We establish conditions under which the planner's problem is a convex problem, ensuring that the globally optimal network can be computed using standard gradient-descent methods. We also discuss the properties and the numerical implementation of the optimal network when that condition does not

[4]Felbermayr and Tarasov (2015) studies optimal investments in infrastructure by competing planners in an Armington model where locations are arranged on the line. Some recent studies also allow for endogenous transport costs in different historical contexts. Swisher IV (2015) model U.S. transport investments as the result of a Nash Equilibrium across competing companies in the context. Trew (2016) endogeneizes trade costs in the spatial-development framework of Desmet and Rossi-Hansberg (2014) by making them depend on the amount of activity in a location, and studies the role of transport infrastructure in shaping structural change in England and Wales.

hold, due to for example increasing returns to network building or weak transport congestion.[5]

There is also a qualitative difference between the optimal transport problem embedded in our model and the one in gravity models. Gravity models assume that each commodity is produced in only one location (as in Armington) or that all the technologies to produce traded goods are linear (as in the Ricardian model), and that transport technologies have constant returns to scale. These assumptions imply that solving the least-cost route optimization across pairs of locations is sufficient to fully characterize the optimal transport problem independently from any general-equilibrium outcome. In our case, the same good may be produced in many regions, markets may source the same good from different suppliers, and there may be decreasing returns in transport. Therefore, endogenous variation in a commodity's spatial supply and demand impacts the optimal transport and vice-versa, linking the optimal-transport problem with the neoclassical allocation problem.[6]

The optimal-transport problem embedded within our framework is related to problems studied early on by Monge (1781) and Kantorovich (1942), and more recently surveyed by Villani (2003) and Galichon (2016). However, our problem differs from this literature in three important aspects. First, optimal transport problems in this strand of literature usually study the direct assignment of sources to destinations, sidestepping the optimal route problem. In that regard, our approach is more closely related to optimal flow and transport problems on a network as reviewed by Bertsekas (1998) and in Chapter 8 of Galichon (2016). Second, in our model, consumption and production in every location are endogenous because they respond to standard general-equilibrium forces. Instead, the aforementioned optimal transport problems are typically concerned with mapping sources with fixed supply to sinks with fixed demand. Third, our focus is on the optimization over network investments, whereas this literature takes the transport costs between links as a primitive. Despite these differences, we are able to follow the optimal transport literature in adopting the convenient duality approach to solve for the global optimization problem. As a special case of a convex optimization problem, our model shares the strong duality property that makes optimal transport problems tractable. Therefore, an important insight of our approach is that embedding an optimal transport problem into a general neoclassical spatial equilibrium extended with a network design problem does not preclude the application of such resolution methods.[7]

[5] We focus on transport networks and their impact on goods' trade. Chaney (2014a) studies endogenous networks of traders in contexts with imperfect information. For a review of recent literature on the role of various types of networks in international trade see Chaney (2014b).

[6] A useful case nested within our model that illustrates this property is the pure endowment economy, where the supply of each good is exogenously given in each location. Even in the absence of transport congestion, least-cost routes are no longer sufficient to solve for the optimal transport, for it is necessary to determine in what quantities each source supplies each market. The optimal-transport problem embedded within our model yields the least-cost route as the solution under the Armington or Ricardian assumptions in the absence of congestion in transport.

[7] Our paper also relates to the network-design literature in operations research, which studies related network-design problems without embedding them in general-equilibrium spatial models. That literature is often focused on heuristic approaches to finding local and global maxima in non-convex cases, with few general results. In the trade literature, Alder (2016) applies an heuristic algorithm based on Gastner and Newman (2006) that progressively eliminates links according to their impact on market-access measures to determine which cities should be connected if a Chinese-style transport network was imposed in India.

Finally, our framework could be combined with empirical research that estimates how transport costs impact economic activity. For instance, Chandra and Thompson (2000), Baum-Snow (2007) and Duranton et al. (2014) estimate the impact of the U.S. highways on various local economic outcomes; Donaldson (2010) and Donaldson and Hornbeck (2016) estimate the impact of access to railways in India and the U.S., respectively; and Faber (2014) estimates the impact of connecting regions to the expressway system in China. Since our model determines the optimal location of transport investments in a general geography, it may serve as a basis to construct instruments for the location of transport infrastructure as function of observed economic fundamentals. Feyrer (2009) and Pascali (2014) assess how the arrival of air transport and steam shipping, respectively, impacted countries or cities whose geographic position made them more likely to use the new transport mode, while studies such as Davis and Weinstein (2002) and Michaels and Rauch (2013) study the degree of adjustment and persistence in networks of cities in response to large shocks. Our model could be used to determine the impact of new transport technologies operating through the optimal investments reshaping the network, and to study inefficient network lock-in due to existing investments corresponding to dated economic fundamentals, two potential uses that we illustrate in our examples.

2 Model

2.1 Environment

Preferences The economy consists of a discrete set of locations $\mathcal{J} = \{1, .., J\}$. We let L_j be the number of workers located in $j \in \mathcal{J}$, and L be the total number of workers. We will entertain cases with labor mobility, where L_j is determined endogenously, and cases without mobility in which L_j is given. Workers consume traded goods aggregating the services of tradable commodities and a non-traded good in fixed supply, such as land or housing. Utility of an individual worker who consumes c units of traded goods h units of non-traded goods is

$$U\left(c, h\right)$$

where U is a homothetic function. This formulation will encompass cases where locations vary in how attractive they are, e.g., because of amenities. In location j, per-capita consumption of traded services is

$$c_j = \frac{C_j}{L_j},$$

where C_j is the aggregate supply of traded services in location j. There is a discrete set of tradable sectors $n = 1, .., N$, combined into C_j through a CRS and concave aggregator,

$$C_j\left(C_j^1, .., C_j^N\right)$$

where C_j^n is the quantity of sector n's output consumed in j. A convenient, but not necessary, functional form corresponding to what is typically assumed in the literature is the CES technology,

$$C_j = \left(\sum_{n=1}^{N} \left(C_j^n \right)^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \tag{1}$$

where $\sigma > 1$ is the elasticity of substitution.[8]

Production The supply-side of the economy corresponds to a general neoclassical economy. In addition to labor, there is a fixed supply $\bar{V}_j = \left\{ V_j^m \right\}$ of primary factors $m = 1, .., M$ in location j. These factors are immobile across regions, and may be either mobile or immobile across sectors. Output of sector n in location j is:

$$Y_j^n = F_j^n \left(L_j^n, \bar{V}_j^n \right), \tag{2}$$

where L_j^n is the number of workers and $\bar{V}_j^n = \left\{ V_j^{mn} \right\}_m$ is the quantity of other primary factors allocated to the production of sector n in location j. The production function F_j^n either has constant returns to scale or is a constant (in which case the supply is an endowment). Therefore, the production structure encompasses the neoclassical trade models. The Armington model (Anderson and Van Wincoop, 2003) corresponds to $N = J$ (as many sectors as regions) and $F_j^n = 0$ for $n \neq j$ $z_j^n = 0$, so that Y_j^j is region j's output in the differentiated commodity that (only) region j provides. The Ricardian model corresponds to labor as the only factor of production and linear technologies, $Y_j^n = z_j^n L_j^n$. The specific-factors and Hecksher-Ohlin models are also special cases of this production structure.

Underlying Geography The locations $\mathcal{J}$ are arranged on an undirected graph $(\mathcal{J}, \mathcal{E})$, where $\mathcal{E}$ denotes the set of edges (i.e., unordered pairs of $\mathcal{J}$). For each location j there is a set $\mathcal{N}(j)$ of connected locations, or neighbors in short. Goods can be shipped only through connected locations; i.e., goods shipped from j can be sent to any $k \in \mathcal{N}(j)$, but to reach any $k' \notin \mathcal{N}(j)$ they must transit through a sequence of connected locations. The transport-network design problem will consist in determining the quality of the infrastructure linking each pair of connected locations. A natural case encompassed by this setup corresponds to j being a geographic unit such as county, $\mathcal{N}(j)$ being its bordering counties, and shipments being done by land. More generally, neighbors in our theory do not need to be geographically contiguous; e.g., it could be possible to ship directly between distant locations by air or sea. Locations may engage in entrepôt trade, i.e., exporting imported goods. We let Q_{jk}^n be the quantify of goods in sector n shipped from j to $k \in \mathcal{N}(j)$.

[8]For simplicity we assume that traded commodities are only used for final consumption. The framework can be generalized to allow traded goods as intermediates in production.

Transport Technology Because of dimensionality, optimizing over networks is usually intractable. We approach this problem by convexifying the network. Specifically, we assume that transporting goods entails resource costs. Transporting Q^n_{jk} units of commodity n from j to k requires $\tau^n_{jk}Q^n_{jk}$ units of the good n itself, where τ^n_{jk} denotes the per-unit cost of transporting good n from j to k. This cost may depend on the quantity shipped, Q^n_{jk}, and on a shifter I_{jk} that captures investments in infrastructure along link jk through a congestion function $\tau_{jk}(Q, I)$:

$$\tau^n_{jk} = \tau_{jk}\left(Q^n_{jk}, I_{jk}\right)$$

Note that $1 + \tau^n_{jk}$ is the iceberg cost typically considered in the literature, except that here it may depend on how much is shipped, Q^n_{jk}, and on the shifter I_{jk}.[9] We assume:

$$\frac{\partial \tau_{jk}(Q, I)}{\partial Q} \geq 0.$$

Assuming $\frac{\partial \tau_{jk}}{\partial Q} > 0$ implies decreasing returns in shipping, due to for example road congestion. More generally, these decreasing returns may originate in the presence of specific factors in transportation.[10] In short, the more is shipped, the higher the per-unit shipping cost. When $\frac{\partial \tau_{jk}}{\partial Q} = 0$, the marginal cost of shipping is invariant to the quantity shipped.

In addition, the shipping cost between j and k depends on the investments undertaken on that link, I_{jk}. A higher I_{jk} determines a higher capacity in the link jk, representing for example a better road quality or more lanes. Hence, we assume:

$$\frac{\partial \tau_{jk}(Q, I)}{\partial I} < 0.$$

The mapping from the infrastructure investment I to the congestion τ compounds the rate at which investments translate into road capacity, and the rate at which road capacity translates into unit costs. In the absence of investment, transport along jk could be either feasible, $\tau_{jk}\left(Q^n_{jk}, 0\right) > 0$, or prohibitively costly, $\tau_{jk}\left(Q^n_{jk}, 0\right) = \infty$. Only when investments go to infinity there is free transport, $\tau_{jk}\left(Q^n_{jk}, \infty\right) = 0$. Finally, we note that the congestion function τ_{jk} is allowed to vary by jk, denoting that, due to geographic accidents, shipping along some links may be more costly than along others given the same quantity shipped and infrastructure investment.[11]

Flow Constraint In every location there are tradable commodities being produced, coming in, and coming out. The balance of these flows requires that, for all locations $j = 1, .., J$ and

[9] In the standard formulation in the trade literature, the iceberg trade cost is defined as a coefficient greater than one such that one unit arrives if that many units are shipped. Here, 1 unit arrives if $1 + \tau$ units are shipped.

[10] Our framework and main results can be extended to encompass cases where the cost of transporting goods is defined per unit and nominated in terms of primary factors, the final traded good, or the final non-traded good. We skip the presentation of these cases to save notation, but their incorporation in the planner's problem is straightforward.

[11] The locations $k \notin \mathcal{N}(j)$ unconnected to j can be equivalently modeled as connected locations for which $\tau_{jk}(Q, I) = \infty$ for all Q and I.

commodities $n = 1, .., J$:

$$C_j^n + \sum_{k \in N(j)} \left(Q_{jk}^n + \tau \left(Q_{jk}^n, \kappa_{jk}^n \right) Q_{jk}^n \right) \leqslant Y_j^n + \sum_{i \in N(j)} Q_{ij}^n \tag{3}$$

The left-hand side of this inequality is location j's consumption of good n, exports to neighbors and quantities lost in transit. These flows must be less than domestic production and imports from neighbors.[12]

We let P_j^n be the multiplier of this constraint. This multiplier reflects society's valuation of a marginal unit of good n in location j. In the decentralized allocation, this multiplier will equal the price of good n in location j; therefore, we simply refer to P_j^n as the price of good j in location n.

Network-Building Constraint The network-design problem will consist in choosing the investments $\{I_{jk}\}_{j,k \in \mathcal{N}(j)}$ for any two connected nodes. For simplicity, we assume that the investment is nominated in units of a freely-mobile resource in fixed supply K which cannot be used for any purpose other than building infrastructure. This assumption leads to the intuitive feature that the opportunity cost of building road capacity in any location corresponds to not building optimally somewhere else. Hence this approach is consistent with assuming that society has sunk a fraction of its resources into network-building, but must still decide how to allocate these resources. The network-building constraint is

$$\sum_j \sum_{k \in \mathcal{N}(j)} I_{jk} = K. \tag{4}$$

We assume $I_{jk} = I_{kj}$, so that the investment on one link applies to shipments in either direction. The framework can easily accommodate more general cases where I is an aggregator of local primary factors, in which case the opportunity cost of building roads would vary across locations.[13]

A Convenient Parametrization For illustrative purposes and to implement the model, a convenient parametrization is the following constant-elasticity transport technology,

$$\tau_{jk}(Q, I) = \frac{Q^\beta}{\kappa_{jk}(I)} \tag{5}$$

where $\kappa_{jk}(I)$ is the road's capacity, itself a function of the investment,

$$\kappa_{jk}(I) = \left(\frac{I}{\delta_{jk}} \right)^\gamma. \tag{6}$$

If $\beta > 0$, this formulation implies congestion in shipping: the more is shipped, the higher the per-unit shipping cost; when $\beta = 0$, the marginal cost of shipping is invariant to the quantity shipped,

[12] In standard minimum-cost flow problems this restriction is referred to as "conservation of flows constraint". E.g., see Bertsekas (1998) and Chapter 8 of Galichon (2016).

[13] E.g., letting $I = F_j^I \left(L_j^I, \bar{V}_j^I \right)$ where L_j^I and $\bar{V}_j^I$ are local labor and factors used in network building and F_j^I is a CRS production function.

as in the standard iceberg formulation. In turn, γ captures the returns to scale in road-building. If $\gamma < 1$ then there are decreasing returns in the technology to build roads. The shifter δ_{jk} denotes how costly it is to build a lane from j to k, e.g. because of geographic accidents.

2.2 Planner's Problem

We solve the problem of a utilitarian social planner who maximizes worker's welfare. Letting ω_j be the planner's weight per worker located in region j, we define this problem as follows.

Definition 1. *The planner's problem without labor mobility is*

$$W = \max_{c_j, h_j, C_j, \left\{C_j^n, L_j^n, \bar{V}_j^n, \left\{Q_{jk}^n\right\}_{k \in \mathcal{N}(j)}\right\}_n, \{Ijk\}_{k \in \mathcal{N}(j)}} \sum_j \omega_j L_j U\left(c_j, h_j\right)$$

subject to:

(i) availability of traded commodities,

$$c_j L_j \leqslant C_j^T\left(C_j^1, .., C_j^N\right) \text{ for all } j;$$

and availability of non-traded commodities,

$$h_j L_j \leqslant H_j \text{ for all } j;$$

(ii) the balanced-flows constraint,

$$C_j^n + \sum_{k \in N(j)} \left(Q_{jk}^n + \tau_{jk}\left(Q_{jk}^n, I_{jk}\right) Q_{jk}^n\right) \leqslant z_j^n F\left(L_j^n, \bar{V}_j^n\right) + \sum_{i \in N(j)} Q_{ij}^n \text{ for all } j,\ n;$$

(iii) the network-building constraint,

$$\sum_j \sum_{k \in \mathcal{N}(j)} I_{jk} \leq K,$$

with $I_{jk} = I_{kj}$*;*

(iv) local labor-market clearing,

$$\sum_n L_j^n \leq L_j \text{ for all } j;$$

and local factor market clearing for the remaining factors,

$$\sum_n V_j^{ln} \leq V_j^l \text{ for all } j \text{ and } l;$$

(v) the non-negativity constraints on flows, trade costs and factors

$$\begin{aligned} Q_{jk}^n &\geq 0 \text{ for all } j, k \in N\left(j\right), n \\ I_{jk} &\geq 0 \text{ for all } j, k \in N\left(j\right), n \\ L_j^n, V_j^n &\geq 0 \text{ for all } j, n. \end{aligned}$$

If labor is mobile then the problem is defined as follows.

Definition 2. *The planner's problem with labor mobility is*

$$W = \max_{c_j, h_j, C_j, \left\{C_j^n, L_j^n, \bar{V}_j^n, \left\{Q_{jk}^n, \kappa_{jk}^n\right\}_{k \in \mathcal{N}(j)}\right\}, L_j, u} u$$

subject to restrictions (i)-(v) above; as well as:

(vi) free labor mobility,

$$uL_j \leq U\left(c_j, h_j\right) L_j \text{ for all } j;$$

(vii) aggregate labor-market clearing,

$$\sum_j L_j = L; \text{ and}$$

This formulation restricts the planner's problem to allocations satisfying utility equalization across locations. Since U is strictly increasing in its arguments, restriction (vi) implies that the planner will allocate $u = U\left(c_j, h_j\right)$ across all populated locations, and $c_j = 0$ otherwise. Since per-capita utility equalization across locations holds in the competitive allocation, we restrict the planner's problem to allocations that can be implemented by the market.

The planner problem from Definition 1 can be expressed as nesting three problems:

$$W = \max_{I_{jk}} \max_{Q_{jk}^n} \max_{C_j^n, Y_j^n} \sum_j \omega_j L_j U\left(c_j, h_j\right)$$

subject to constraints. A similar nesting can be expressed in the case with labor mobility from Definition 2. We now discuss some intuitive properties of the planner's solution.

Neoclassical Allocation Problem The innermost maximization problem over C_j^n, Y_j^n can be cast as a standard neoclassical allocation problem of choosing consumption and output given goods prices subject to the production possibility frontier. This problem does not depend on the flows Q_{jk}^n nor on the network investments I_{jk} other than through the prices P_j^n. Since, as we show below, the second welfare theorem holds, the neoclassical allocation and gross flows from the planner's problem correspond to a competitive market equilibrium with transfers set in relation to the planner's weights, ω_j.

Optimal Transport Problem The problem over Q_{jk}^n given the neoclassical allocation is the optimal-transport problem. Given quantities supplied and demanded, this problem determines the efficient shipping throughout the network accounting for congestion through the balanced-flows constraint (3). Given production Y_j^n and consumption C_j^n, in the absence of congestion this is a standard problem considered in the optimal transport literature, e.g., chapter 8 of Galichon (2016). To understand the solution to this problem, remember that P_j^n is the multiplier of the flows constraint (ii), equal to the price of good n in location j in the market allocation. The solution to the planner's problem gives the following equilibrium price differential for commodity

n between j and $k \in \mathcal{N}(j)$:

$$\frac{P_k^n}{P_j^n} - 1 \leq \left(\varepsilon_{Q,jk}^n + 1\right) \tau_{jk}^n, = \text{if } Q_{jk}^n > 0, \tag{7}$$

where we have used the notation $\tau_{jk}^n \equiv \tau_{jk}\left(Q_{jk}^n, I_{jk}\right)$, and where

$$\varepsilon_{Q,jk}^n \equiv \frac{\partial \tau_{jk}^n}{\partial Q_{jk}^n} \frac{Q_{jk}^n}{\tau_{jk}^n}$$

is the elasticity of the per-unit transport cost with respect to the quantity shipped. Condition (7) is a standard no-arbitrage condition: the price differential between location j and any of its connected locations k must be less than or equal than the marginal cost. From the planner's perspective, this cost includes both the per-unit shipping cost τ_{jk}^n and the marginal congestion $\varepsilon_{Q,jk}^n$. In the absence of decreasing returns in transport, $\varepsilon_{Q,jk}^n = 0$, the price differential only reflects transport cost, as in the standard constant iceberg cost formulation.

This expression has a number of intuitive properties that we exploit throughout our analysis. First, given the network investment, (7) identifies the trade flow Q_{jk}^n as function of the price differential as long as the right-hand side is an invertible function of Q_{jk}^n. This is the case whenever the total shipping cost $\tau_{jk}^n Q_{jk}^n$ is convex in Q_{jk}^n, as established in Proposition (1). Second, whenever that condition holds, the gross trade flow Q_{jk}^n is increasing in the price differential $\frac{P_k^n}{P_j^n}$: the larger the difference in marginal valuations, the higher the flow to the location where the product is more scarce. It also implies that goods in each sector flow in only one direction; i.e. $Q_{jk}^n > 0 \Rightarrow Q_{kj}^n = 0$. However, there may be flows in opposite direction along the same link corresponding to different goods. Third, given the relative prices, the flows are decreasing with the marginal congestion, $\varepsilon_{Q,jk}^n$.

For example, assuming constant-elasticity congestion as in (5), we have $\varepsilon_{Q,jk}^n = \beta$. Then, we obtain the following gross trade flow of good n from j to k as function of prices:

$$Q_{jk}^n = \begin{cases} \left[\frac{\kappa_{jk}}{\beta+1}\left(\frac{P_k^n}{P_j^n} - 1\right)\right]^{\frac{1}{\beta}} & \text{if } P_k^n \geq P_j^n \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

This solution naturally implies that increases in capacity are associated with higher flows given the price differentials.

The first-order condition (8) reduces the dimension of the optimal-transport problem. Substituting the solution for Q_{jk}^n as function of the price differentials P_k^n / P_j^n into the balanced-flows constraints (ii) yields a system of nonlinear equations with as many prices as regions for each good. Therefore, for each commodity, the optimal transport problem can be solved finding the J prices P_k^n, one for each location. Rather than solving globally for least-cost routes, this characterization optimizes on the local flows. The global paths and quantities delivered to each destination are defined once the prices are solved for. The least-cost route typically present in the applications of the gravity literature is the solution in the absence of congestion ($\beta = 0$) if, in equilibrium, each market sources each product from only one source, as in the Armington model where products are differentiated by origin.

Network Investment Problem Consider now the outer problem of choosing the network I_{jk} given the optimal transport and the neoclassical allocation. Letting μ be the multiplier of the network-building constraint (iv), the planner's choice of I_{jk} gives

$$I_{jk}\mu = \sum_n P_j^n Q_{jk}^n \tau_{jk}^n \varepsilon_{I,jk}^n + \sum_n P_k^n Q_{kj}^n \tau_{kj}^n \varepsilon_{I,kj}^n, \tag{9}$$

where

$$\varepsilon_{I,jk}^n \equiv \left| \frac{\partial \tau_{jk}^n}{\partial I_{jk}} \frac{I_{jk}}{\tau} \right|$$

is the elasticity of the per-unit transport cost with respect to the network investment.[14] To interpret this condition, note that its left-hand side is the opportunity cost of investing along jk, defined as the marginal valuation for the scarce resource used to build transport infrastructure (μ) times the quantity invested I_{jk} along jk. This marginal cost of improving the transport capacity of the link jk must equal the marginal savings in shipping in the right-hand side of 9. The first sum includes the savings in shipments from j to k. The investment reduces the per-unit cost by $\tau_{jk}^n \varepsilon_{I,jk}^n$, saving on the Q_{jk}^n units shipped of each good n with social value P_j^n. The second sum includes the savings in shipments in the opposite direction.

Assuming constant-elasticity technologies (5) and (6), we have $\varepsilon_{I,jk} = \gamma$. Further imposing $\delta_{jk} = \delta_{kj}$, so that given the infrastructure investment the congestion function is symmetric with respect to the direction of the flows, we obtain the following road capacities:[15]

$$\kappa_{jk} = \kappa_{kj} = \left(\frac{\gamma}{\mu \delta_{jk}} \sum_n \left(P_j^n \left(Q_{jk}^n\right)^{1+\beta} + P_k^n \left(Q_{kj}^n\right)^{1+\beta} \right) \right)^{\frac{\gamma}{1+\gamma}}. \tag{10}$$

The optimal road capacity is higher along links with more gross flows in either direction ($Q_{jk}^n > 0$ or $Q_{kj}^n > 0$). Given these flows, the optimal capacity also increases with prices at origin: because shipping requires the good being shipped as an input, a higher price at origin implies a higher marginal saving from investing.[16] Conditioning on these outcomes, the optimal network capacity

[14] Note that, for simplicity, we have constrained the planner's problem to symmetric investments in both directions, i.e. $I_{jk} = I_{kj}$. Still, we allow for $\delta_{jk} \neq \delta_{kj}$, so that, given the amount invested I_{jk}, the shipping cost function $\tau_{jk}(Q, I_{jk})$ depends on the direction of the flow. E.g., it can be cheaper drive downhill or to sail with the wind's direction.

[15] More generally, the solution with asymmetric terrain or shipping costs is $\kappa_{jk} = \left(\frac{I_{jk}}{\delta_{jk}}\right)^{\gamma}$, where:

$$I_{jk} = I_{kj} = \left[\frac{\gamma}{\mu} \sum_n \left(P_j^n \delta_{jk}^{\gamma} \left(Q_{jk}^n\right)^{1+\beta} + P_k^n \delta_{kj}^{\gamma} \left(Q_{kj}^n\right)^{1+\beta} \right) \right]^{\frac{1}{1+\gamma}}.$$

[16] In cases where shipping requires local resources such as labor in j, a similar logic implies that a higher shadow

is lower when building infrastructure is more costly (higher δ_{jk}).

Importantly, using the solution for Q^n_{jk} from (7), the network investment problem inherits from the optimal transport problem the useful property that prices in locations j and k are sufficient to determine the optimal investment in the link jk.

Convexity We next establish the convexity of the planner's problem.

Proposition 1. *(Convexity of the Planner's Problem) Given the network investments* $\{I_{jk}\}$*, the optimal transport+neoclassical allocation problem is a convex optimization problem if* $Q\tau_{jk}(Q, I_{jk})$ *is convex in* $Q \in \mathbb{R}_+$ *for all* j *and* $k \in \mathcal{N}(j)$*. The full planner's problem from Definition* (1) *(resp. Definition* (2)*) is a convex (resp. quasiconvex) optimization problem if* $Q\tau_{jk}(Q, I)$ *is convex in* $Q \in \mathbb{R}_+$ *and* $I \in \mathbb{R}_+$ *for all* j *and* $k \in \mathcal{N}(j)$*. When the transport technology is given by* (5) *and* (6)*, this property is ensured by* $\beta \geq \gamma$*.*

The proof is straightforward. Given the neoclassical assumptions, all the constraints are convex, except potentially the balanced-flows constraint. Adding decreasing returns in transport ensures convexity of that constraint as well. Without labor mobility, the objective is concave due to decreasing marginal utility from consumption of the traded good. In the case with labor mobility, the objective function is only quasiconcave, but the Arrow-Enthoven theorem for sufficiency of the Kuhn-Tucker conditions under quasiconcavity is satisfied (Arrow and Enthoven, 1961).[17] This result has the important implication of establishing conditions under which the Kuhn-Tucker conditions are both necessary and sufficient to identify the optimal solutions. Therefore, well-known numerical algorithms can be applied to find numerically the solution (Boyd and Vandenberghe, 2004).

The condition ensuring convexity restricts how congestion in shipping Q and the network investment I combine in the transport technology in each link through $\tau_{jk}(Q, I)$. Intuitively, the condition requires the increase in marginal transport costs resulting from higher traffic along a link to offset the reduction in those costs resulting from investments in that link. This interpretation can be clearly grasped when the transport technology is given by (5) and (6). In that case $\beta \geq \gamma$ denotes stronger congestion in network-building than in transport. In the absence of decreasing returns in transport (i.e., if $\frac{\partial \tau_{jk}}{\partial Q} = 0$), convexity necessarily fails.

Non-Convex Cases When the condition guaranteeing global convexity in Proposition 1 fails we cannot ensure the sufficiency of the first-order conditions. We discuss below how to deal with these cases computationally. Non-convex cases are interesting, for they give rise to networks with distinctive properties.

Proposition 2. *When the transport technology is given by* (5) *and* (6) *and* $\gamma > \beta$ *and there is a single commodity, the optimal transport network is a disjoint union of trees.*

value of labor in j, or wages in the market allocation, translate into more investments in the link jk to economize on the relatively scarce resource.

[17] This theorem requires that the the gradient of the objective function is different from zero at the optimal point. This property holds in our context.

A tree is a connected graph with no loops. Intuitively, cycles cannot be optimal, because it is strictly better, from the planner's perspective, to remove a link from the loop and concentrate resources in other links. As a result, only one path links any two locations in the network. Note that, despite this property, the optimal-transport problem is still not trivial because it is necessary to determine the quantity of output from each location that is sourced by every other location.

2.3 Decentralized Allocation

We establish that the planner's allocation corresponds to a decentralized equilibrium. We focus on the decentralization of the neoclassical allocation ($\max_{C_j^n, Y_j^n}$) and optimal-transport ($\max_{Q_{jk}^n}$) problems with and without labor mobility given the network investments $\{I_{jk}\}$. In the decentralization, the network is taken as given and assumed to be constructed by a benevolent government who owns the stock K of the input used to build the network.[18]

Given the network, the decentralized economy corresponds to the perfectly competitive equilibrium of a standard neoclassical economy where consumers maximize utility given their budget, producers maximize profits subject to their production possibilities, and goods and factor markets clear.

Relative to standard international trade and economic geography models, there only one slightly less standard feature: because of aggregate congestion in transport, we must account for externalities and rents in the transport sector. We assume perfect competition and free entry into transport. In each location, atomistic traders purchase goods and ship them to connected locations. As long as there is congestion, atomistic traders who take the iceberg trade cost as given engage in too much shipping. To correct this externality, we allow for Pigouvian sales taxes t_{jk}^n on companies shipping good n on leg $j \to k$. The profits made by an individual company shipping good n from j to k are:

$$\pi_{jk}^n = \max_{q_{jk}^n} \left[p_k^n \left(1 - t_{jk}^n\right) - p_j^n \left(1 + \tau_{jk}^n\right) \right] q_{jk}^n,$$

where p_j^n is the price of good n in location j in the market allocation and q_{jk}^n is the quantity shipped. Each individual shipping company takes the iceberg trade cost τ_{jk} as given, although this trade cost is determined endogenously through (5) as function of the aggregate quantity shipped. Free entry ensures that $\pi_{jk}^n \leq 0$, with equality if there are actual shipments, $Q_{jk}^n > 0$. As a result, the rents from the transport sector equal the tax revenue collected by the government.

We must also allocate the returns to factors other than labor. Under no labor mobility we assume that, in addition to the wage, each worker in j receives a transfer b_j such that $\sum_j b_j L_j = \Pi$, where Π is an aggregate portfolio including all the sources of income, including tax revenues, except for labor. Hence, no other agent except for workers own the primary factors or the non-traded goods, or are rebated the tax revenues. This formulation allows for trade imbalances, which are needed to implement the planner's allocation under arbitrary weights. Under perfect labor mobility, we

[18] This structure can be generalized to more general formulations, where resources used to produce traded and non-traded goods can also be used to build the network, and where the benevolent planner taxes the private sector to obtain the resources to build the network.

assume that all workers own an equal fraction of all the sources of income other than labor regardless of location, so that $b_j = \frac{\Pi}{L}$ for all j.

Since it is standard, we relegate the Definition 3 of the competitive allocation with and without labor mobility to the appendix. Using that definition, we establish that the welfare theorems given the transport network hold.

Proposition 3. *(First and Second Welfare Theorems) If the sales tax on shipments of product n from j to k is*

$$1 - t_{jk}^n = \frac{1+\tau_{jk}^n}{1+\left(\varepsilon_{Q,jk}^n + 1\right)\tau_{jk}^n},$$

then:

(i) if labor is immobile, the competitive allocation coincides with the planner's problem under specific planner's weights ω_j. Conversely, the planner's allocation can be implemented by a market allocation with specific transfers b_j; and

(ii) if labor is mobile, the competitive allocation coincides with the planner's problem.

In either case, the price of good n in location j, p_j^n, equals the multiplier on the balanced-flows constraint in the planner's allocation, P_j^n.

2.4 Numerical Implementation

Convex Case We use solvers that rely on the dual approach typically pursued in optimal transport problems. Specifically, letting $\mathcal{L}$ be the Lagrangian of the planner's problem as function of the variables controlled by the planner, $\mathbf{x}$, and the multipliers λ, the strong duality implies that the solution to the dual problem,

$$\inf_{\lambda} \sup_{\mathbf{x}} \mathcal{L}\left(\mathbf{x}, \lambda\right),$$

is equivalent to the solution to the primal problem,

$$\sup_{\mathbf{x}} \inf_{\lambda} \mathcal{L}\left(\mathbf{x}, \lambda\right).$$

The advantage of the dual approach is that we are able to use the first-order conditions from the optimal transport problem and the optimal investment problem, (7) and (10), as well as those from the neoclassical allocation problem, to express the control variables as function of the multipliers, $\mathbf{x}\left(\lambda\right)$. The remaining minimization problem,

$$\inf_{\lambda} \mathcal{L}\left(\mathbf{x}\left(\lambda\right), \lambda\right)$$

is a globally convex minimization problem without constraints over NJ variables corresponding to the goods prices across locations.[19] Instead, solving the problem using the primal approach would

[19] There is an additional multiplier, corresponding to the the network-building constraint (4), but its value can be easily computed to satisfy that constraint.

require solving a much higher dimensional system with additional constraints. The convexity of the dual problem ensures the convergence of gradient-descent methods. For practical purposes, we use a convex solver based on an interior-point method, which converges in polynomial time and can easily handle thousands of variables at a time, when the sparsity of the problem is taken into account.[20]

Nonconvex case To compute non-convex cases in which the condition stated in Proposition (1) fails we exploit the property, stated in the first of Proposition 1, that the neoclassical allocation+optimal flows problem nested within the planner problem is convex as long as $Q\tau_{jk}(Q, I_{jk})$ is convex in Q. This property holds for weak assumptions on τ, including the standard formulation where iceberg costs are independent from the quantity shipped. We follow an iterative procedure, where we solve for the global optimum over C_j^n, Y_j^n, Q_{jk}^n given the network investments I_{jk}, and then use the optimal network investment condition (9) to characterize I_{jk} as function of the solution to the neoclassical-allocation and optimal-transport problems. In practice, this algorithm always converges to a local optimum. We then refine that solution using a simulated annealing method that perturbs the local optimum and searches for a better network in its vicinity. Our current work in progress explores the implementation of branch-and-bound methods which provide arbitrarily close approximations to the global optimum.

3 Examples

We implement examples that illustrate some economic forces captured by the framework. We start with simple, low-dimensional environments, and build up to more complex cases. Preferences between traded and non-traded goods are CRRA over consumption bundles and Cobb-Douglas between the traded and non-traded good:

$$U = \frac{\left(c^\alpha h^{1-\alpha}\right)^{1-\rho}}{1-\rho},$$

with $\alpha = \frac{1}{2}$ and $\rho = 2$. There is a single factor of production, labor, and all technologies are linear.

We adopt the constant-elasticity functional forms (5) and (6) for the congestion and network-building technologies. Except otherwise noted, we impose $\beta = \gamma = 1$. I.e., both the technology to build capacity and the congestion function are linear. This parametrization is at the boundary of the parameter space guaranteeing global convexity.

We start by inspecting cases with a single good, no labor mobility, and convexity of the planner solution. Then, we introduce multiple sectors, labor mobility, and cases with increasing returns to building the network where the convexity of the planner's problem fails.

[20] We use the open-source large-scale optimization package IPOPT (https://projects.coin-or.org/Ipopt), available for C/C++, Fortran, MATLAB and other languages.

3.1 One Good on a Regular Geometry

Comparative Statics over K in a Symmetric Network Figure A.1 presents a network with 13^2 locations uniformly distributed in a square, each connected to 8 neighbors. All fundamentals except for productivity are symmetric: $(L_j, H_j, \delta_{ij}) = (1, 1, 1)$. Labor productivity is $z_j = 1$ at the center and 10 times smaller elsewhere.

Figure A.2 shows the globally optimal network when $K = 1$ (panel (a)) and when $K = 100$ (panel (b)). The upper-left figure in each panel displays the optimal network capacities κ_{ij} corresponding to (6). The optimal network investments radiate from the center, and so do shipments. The bottom figures in each panel display the multipliers of the flows constraint (3) –the prices in the market allocation– and consumption. Because tradable goods are less abundant in the outskirts, marginal utility is higher and so are prices. As the aggregate investment grows from $K = 1$ to $K = 100$, the network grows into the outskirts and differences in the marginal utility shrink. Panel (a) of Figure A.3 displays the spatial distribution of prices (upper panel) and consumption (bottom). The left panels display outcomes across locations ordered by euclidean distance to the center. As the network grows, relative prices and consumption converge to the center, and spatial inequalities are reduced.

Panel (b) of Figure A.3 illustrates the difference between the welfare gains from uniform and optimal network expansion. For K close to zero, the capacities κ_{jk} are small everywhere and all locations are close to autarky We simulate an increase in K in two cases: uniformly allocating I across all links (a "rescaled" network) and optimally doing so. The figure reports the welfare increase associated with each network. Broadly speaking, the uniform network expansion corresponds to the standard counterfactual implemented in international trade, in which trade costs are reduced uniformly from autarky to trade. As K grows, the economy converges to the free-trade level of welfare regardless of whether the network is optimal. Moving from close to autarky to close to free trade increases welfare by 2.5%. However, investing optimally leads to faster convergence to the free-trade welfare level. In the example, the welfare level attained in the uniform network when $K = 10^5$ is attained in the optimal network when $K = 10^2$.

Lock-In versus Optimal Network The top panel of Figure A.4 illustrates a case with losses from lock-in in a suboptimal network. Starting from the same symmetric geography as our previous example, Panel (a) shows the optimal size-K network when the high-productivity city is located at the southwest corner instead of in the middle. In Panel (b), productivity increases in the city at the northeast, but decreases at the southwest. At the same time, an additional quantity ΔK of resources is invested. Optimal network investments are diverted to the northeast, and the southwest network shrinks. Panel (c) shows a spatial distribution of productivities identical to (b), but now the network is constrained to to be built upon the existing network.

Hence, the networks in (b) and (c) entail the exact same amount of investments, $K + \Delta K$. However, network (c) is a suboptimal network corresponding to lock-in in an older network targeting a different spatial productivity distribution, while network (b) targets the final distribution of productivity. In the lock-in network, close-by regions do not develop despite the large productivity increase in the northeast. The bottom panel in Figure A.4 shows the welfare loss from lock-in, defined as the difference in welfare between panels (c) and (b) in the top panel, as function of the new city's productivity. The larger is the new city's productivity, the largest is the loss from lock-in.

Random Cities and Non-Convex Cases We now explore more complex networks and non-convex cases. Figure A.5 shows 20 cities randomly located in a space where each location has six neighbors. Population is $L_j = 1$ in the cities and 0 otherwise. Productivity is again ten times larger at the center. The top panel shows the capacity and goods flows in the optimal network. The optimal network radiates from the center to reach all destinations. Due to congestion, multiple sub-routes are built. As a result, some destinations are reached through multiple routes. However, to reach some faraway locations only one route is built.

The middle panel inspects the same spatial configuration but assumes $\gamma = 2$. Now, the sufficient condition for global convexity from Proposition 1 fails. We see a qualitative change in the shape of the network. Due to increasing returns in network building, less roads with higher capacity are built. In particular, there is now only one route linking any two destinations, consistent with the no-loops result in Proposition 2.

Because in the non-convex network we can only guarantee convergence to a local optimum, we refine the solution applying simulated annealing. The bottom panel compares the non-convex network before and after the annealing refinement. The refined network economizes on the number of links, leading to a welfare increase but preserving the no-loops property.

3.2 Multiple Sectors, Labor Mobility, and Non-Convexity

One Homogeneous Good and 10 Differentiated Goods All the applications so far included only one traded sector. We now show a case with multiple traded goods. We allow for 11 traded commodities, one "agricultural" good that may be produced everywhere ($z_j = 1$ everywhere) and ten "industrial" goods each produced in one random city only ($z_j = 1$ in only one city and $z_j = 0$ otherwise). These goods are combined into via the CES aggregator (1) with elasticity $\sigma = 2$. The first panel in Figure A.6 shows the optimal network. In the figure, each circle's size denotes the exogenous population, equal to 1 everywhere. The remaining figures show the shipments of each good, with the circle sizes representing production. Figure A.7 shows the optimal network with annealing when $\gamma > 2$. In these examples, we observe complex shipping patterns. There are bilateral flows over each link, now involving several commodities. The homogeneous good travels short distances to fulfill the needs to the industrial cities. Overall, the optimal network in the first panel reflects the spatial distribution of comparative advantages. Even though population is the same everywhere, the network has high capacity close to places specialized in industrial goods, and branches out into the agricultural hinterland. Figures A.8 and A.9 replicate figures A.6 and A.7 allowing for labor mobility. Workers now endogenously concentrate in the locations with capacity to produce the industrial goods, but the network shape is preserved otherwise.

A Ricardian Economy In Figure A.10 we now consider the Ricardian economy with 2 goods. Productivity in good 1 is higher at the center, as shown in the upper-left panel of panel (a), and equal to 1 everywhere else. Panel (a) shows the allocation of labor in each sector, the optimal network and the pattern of shipping under a uniform population distribution, whereas Panel (b) shows the same outcomes when workers are freely mobile. Both figures correspond to the convex network. We allow for aggregate decreasing returns in each sector, so that some locations may be incompletely specialized. Locations at the center with comparative advantages in good 1 specialize in that good and ship it out to the periphery. As labor becomes mobile, workers sort based on absolute advantages, and population density increases at the center but the pattern of specialization remains unchanged.

3.3 Geographic Features and New Transport Technologies

We now show how the framework can accommodate geographic accidents. For simplicity we focus again on the case with a single good and no factor mobility. Panel (a) of Figure A.11 shows 20 cities randomly allocated in a space where each location is connected to 8 other locations. Population equals 1 in all cities. Productivity is the same everywhere (equal to 0.1), except in the central city where it is 10 times larger. Aggregate consumption in the optimal allocation varies in proportion to each city's size in the figure.

The function mapping infrastructure investments to a link's capacity depends on the link-specific coefficient δ_{ij} in 6. In panel (a) we show the optimal network under the assumption that the cost of adding capacity is directly proportional to geographic distance:

$$\delta_{ij} = \delta_0 + \delta_1 \text{Euclidean Distance}_{ij}. \tag{11}$$

As in our first set of examples, the optimal network radiates from the highest-productivity cities in order to reduce differences in marginal utility across the populated locations.

In panel (b), we add a "mountain" by stretching the euclidean distance between links in the northeast. The optimal network now circles the mountain to reach small cities in the northwest, although it remains optimal to build roads on the hillside. Because more resources need to be invested in that region, the network shrinks elsewhere.

In the subsequent figures, we either increase or reduce the cost of building the network in specific links. Specifically, we allow for the more general function:

$$\delta_{ij} = \delta_0 + \delta_1 \text{Euclidean Distance}_{ij} + \delta_2 \text{CrossingRiver}_{ij} + \delta_3 \text{AlongRiver}_{ij}. \tag{12}$$

In panel (c) we include a river and assume that $\delta_2 = \delta_3 = \infty$, so that either crossing or navigating along the river is prohibitively costly. The optimal network linking cities on either side of the river can only be built through the only patch of dry land. In panel (d) we assume that no dry patch exists, but building bridges is feasible, $\delta_2 < \infty$. Now, the planner builds two bridges, directly connecting the two pairs of cities on either side of the river. In turn, panel (e) allows for water transport by allowing building network capacity along the river ($\delta_3 < \infty$). The planner retains the bridges, but it now reaches faraway locations in the southeast via water instead of ground transport.

Finally, panel (f) moves to the non-convex case, $\gamma = 2 > \beta$, implemented through the FOC + simulated annealing approach we have described. Now, a unique route links any two cities, water transport is not used, and a single bridge is built.

We show how the arrival of new transport technologies can lead to a reconfiguration of city sizes based on their initial geographic position. Both panels of Figure A.12 corresponds to an economy with random cities, all with same population, where productivity is 10 times larger in the city represented with the bigger circle than in every other city. The circle sizes represent consumption per capita. Panel (a) shows an economy with strong dependence on water transport (a low δ_3 in (12)). The optimal network leads to high consumption per capita at the river crossing. In panel (b) we assume that ground transport becomes cheap (e.g., due to the arrival of railways) (a lower δ_1 in (12)). As a result, water transport is no longer used. Instead, the economy relies only ground transport and bridges. Because of the new transport technologies, the spatial distribution of consumption per worker is reconfigured. In particular, the city originally located at the river's crossing shrinks and and consumption in the city at the river intersection shrinks, while all the other cities in the hinterland grow.

4 Conclusion

In this paper we developed a framework to study optimal transport networks in spatial equilibrium models. The framework combines three components: a general neoclassical model where each location is a node in a graph, an optimal transport problem subject to congestion in shipping, and an optimal network investment problem. The framework nests the neoclassical models used in international trade and allows for factor mobility. If congestion in transport is strong relative to the returns to infrastructure investments, the planner's is a standard convex optimization problem, guaranteeing convergence of efficient gradient-descent based algorithms. In the absence of this property, we are still able to implement the framework combining the first order conditions of the planner's problem with standard global-search numerical methods.

In this preliminary version of the paper we have used the model to characterize the globally optimal transport network in different spatial equilibrium environments and to contrast it with sub-optimal networks. We implemented the optimal network in simple regular geometries with a single good and no mobility, and in more complex environments with randomly located cities, labor mobility, and many sectors. Our examples have illustrated the differences between optimal and suboptimal networks in terms of regional effects and aggregate welfare, as well as the impact of the optimal network on the spatial distribution of economic activity. We have also illustrated the contrast between globally optimal networks in convex cases, where congestion in shipping offsets increasing returns due to network building, and locally optimal networks in non-convex cases, in which case the optimal network is sparser, more concentrated in fewer links, and tree-shaped.

Our work in progress explores applications of the framework to some of the specific questions we have initially posed: What would be the gains from replacing actual, potentially inefficient networks with the optimal transport network, and how do these gains vary across countries? How does the optimal transport network interact with the sources of comparative advantages, such as differences in relative productivity and factor endowments, vis-a-vis natural geographic features? How does the optimal response of the transport network diffuse local shocks and impact the spatial distribution of economic activity? Our framework could also be combined with empirical research that estimates how transport costs impact economic activity. For instance, it may serves as basis to construct instruments for the location of transport infrastructure as function of observed economic fundamentals, to determine the aggregate and regional impacts of new transport technologies operating through the optimal investments reshaping the network, and to study cases with inefficient lock-in due to existing investments corresponding to past fundamentals.

References

Alder, S. (2016). Chinese roads in india: The effect of transport infrastructure on economic development. *Manuscript, Univ. North Carolina, Chapel Hill*.

Allen, T. and C. Arkolakis (2014). Trade and the topography of the spatial economy. *Quarterly Journal of Economics 1085*, 1139.

Allen, T. and C. Arkolakis (2016). The welfare effects of transportation infrastructure improvements. *Manuscript, Dartmouth and Yale*.

Anderson, J. E. and E. Van Wincoop (2003). Gravity with gravitas: a solution to the border puzzle. *the american economic review 93*(1), 170–192.

Arrow, K. J. and A. C. Enthoven (1961). Quasi-concave programming. *Econometrica: Journal of the Econometric Society*, 779–800.

Atkin, D. and D. Donaldson (2015). Who's getting globalized? the size and implications of intra-national trade costs. Technical report, National Bureau of Economic Research.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, 775–805.

Bertsekas, D. P. (1998). *Network optimization: continuous and discrete models.* Citeseer.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization.* Cambridge university press.

Caliendo, L., F. Parro, E. Rossi-Hansberg, and P.-D. Sarte (2014). The impact of regional and sectoral productivity changes on the U.S. economy. Technical Report 20168, National Bureau of Economic Research.

Chandra, A. and E. Thompson (2000). Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system. *Regional Science and Urban Economics 30*(4), 457–490.

Chaney, T. (2014a). The network structure of international trade. *The American Economic Review 104*(11), 3600–3634.

Chaney, T. (2014b). Networks in international trade. In *The Oxford Handbook of the Economics of Networks*.

Costinot, A. and A. Rodríguez-Clare (2013). Trade theory with numbers: Quantifying the consequences of globalization, vol. 4 of handbook of international economics.

Davis, D. R. and D. E. Weinstein (2002). Bones, bombs, and break points: the geography of economic activity. *The American Economic Review 92*(5), 1269–1289.

Dekle, R., J. Eaton, and S. Kortum (2008). Global rebalancing with gravity: Measuring the burden of adjustment. Technical Report 3, International Monetary Fund.

Desmet, K. and E. Rossi-Hansberg (2014). Spatial development. *The American Economic Review 104*(4), 1211–1243.

Donaldson, D. (2010). Railroads of the raj: Estimating the impact of transportation infrastructure. Technical report, National Bureau of Economic Research.

Donaldson, D. (2015). The gains from market integration. *Annual Review of Economics 7*(1), 619–647.

Donaldson, D. and R. Hornbeck (2016). Railroads and american economic growth: A market access approach. *The Quarterly Journal of Economics*, qjw002.

Duranton, G., P. M. Morrow, and M. A. Turner (2014). Roads and trade: Evidence from the us. *The Review of Economic Studies 81*(2), 681–724.

Duranton, G. and M. A. Turner (2011). The fundamental law of road congestion: Evidence from us cities. *The American Economic Review 101*(6), 2616–2652.

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from china's national trunk highway system. *The Review of Economic Studies*, rdu010.

Fajgelbaum, P. D., E. Morales, J. C. S. Serrato, and O. M. Zidar (2015). State taxes and spatial misallocation. Technical report, National Bureau of Economic Research.

Felbermayr, G. J. and A. Tarasov (2015). Trade and the spatial distribution of transport infrastructure.

Feyrer, J. (2009). Trade and income–exploiting time series in geography. Technical report, National Bureau of Economic Research.

Galichon, A. (2016). *Optimal Transport Methods in Economics.* Princeton University Press.

Gastner, M. T. and M. Newman (2006). Optimal design of spatial distribution networks. *Physical Review E 74*(1), 016117.

IADB (2013). Too far to export: Domestic transport costs and regional export disparities in latin america and the caribbean.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk SSSR*, Volume 37, pp. 199–201.

Limao, N. and A. J. Venables (2001). Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review 15*(3), 451–479.

Michaels, G. and F. Rauch (2013). Resetting the urban network: 117-2012.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale.

Nagy, D. (2016). City location and economic development. *Manuscript, CREI*.

Pascali, L. (2014). The wind of change: Maritime technology, trade and economic development.

Redding, S. J. (2016a). Goods trade, factor mobility and welfare. *Journal of International Economics 101*, 148–167.

Redding, S. J. (2016b). A quantitative framework for evaluating the impact of urban transport improvements. *Manuscript, Princeton University*.

Redding, S. J. and E. Rossi-Hansberg (2016). Quantitative spatial economics. Technical report, National Bureau of Economic Research.

Redding, S. J. and M. A. Turner (2015). Transportation costs and the spatial organization of economic activity, vol. 5 of handbook of regional and urban economics.

Roback, J. (1982). Wages, rents, and the quality of life. *The Journal of Political Economy*, 1257–1278.

Sotelo, S. (2016). Domestic trade frictions and agriculture. *Manuscript, University of Michigan*.

Swisher IV, S. (2015). Reassessing railroads and growth: Accounting for transport network endogeneity. Technical report, mimeo.

Trew, A. W. (2016). Endogenous infrastructure development and spatial takeoff.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

WorldBank (2009). World development report 2009: Reshaping economic geography.

WorldBank (2011). *Africa's Transport Infrastructure: Mainstreaming Maintenance and Management*. World Bank Publications.

A Figures

Figure A.1: A Simple Underlying Geography

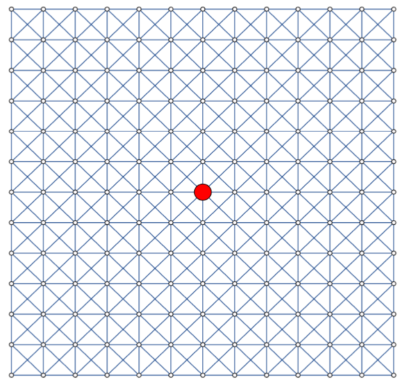
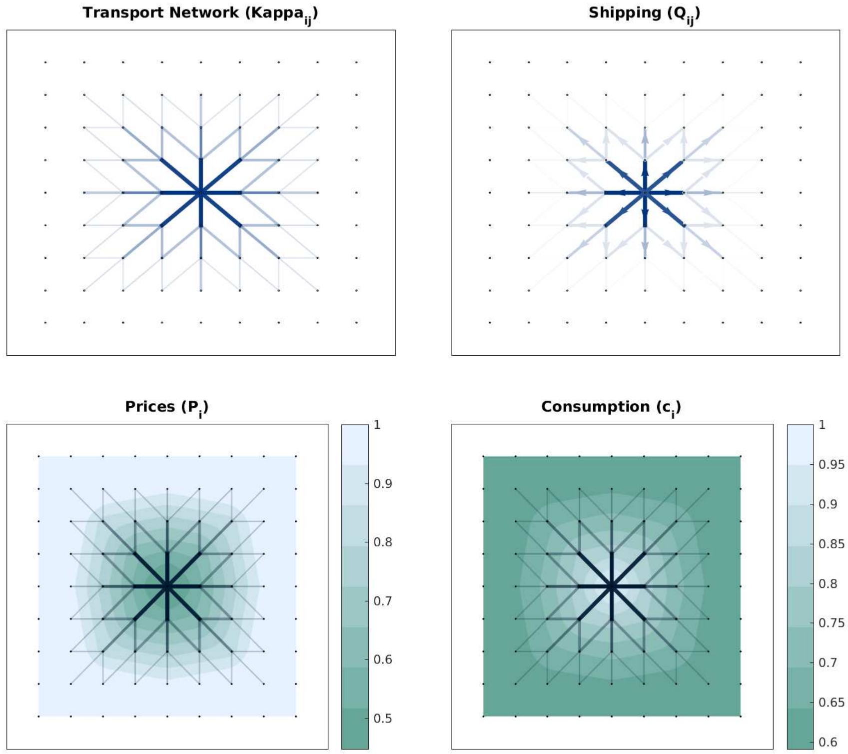


Figure A.2: The Optimal Network for $K = 1$ and $K = 100$

(a) K=1



(b) K=100

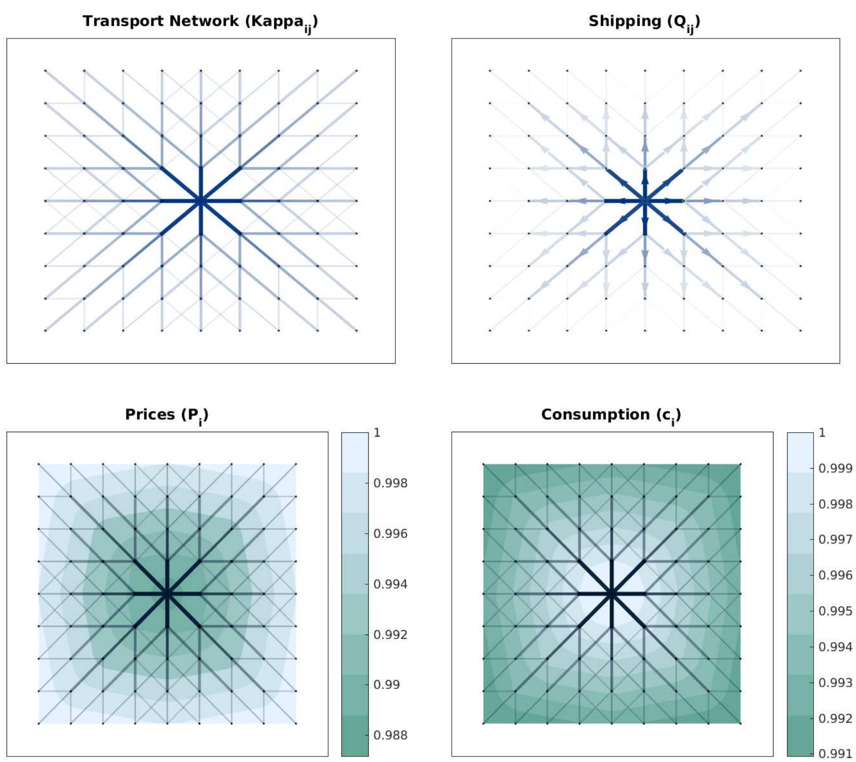
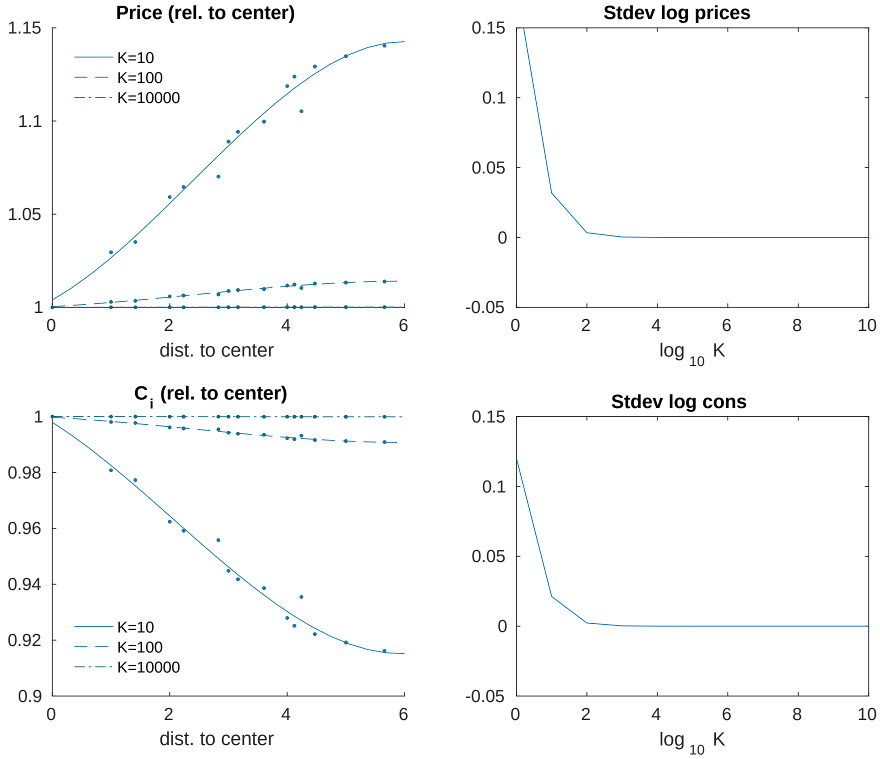


Figure A.3: Optimal Network Growth

(a) Spatial Inequalities



(b) Uniform versus Optimal Network

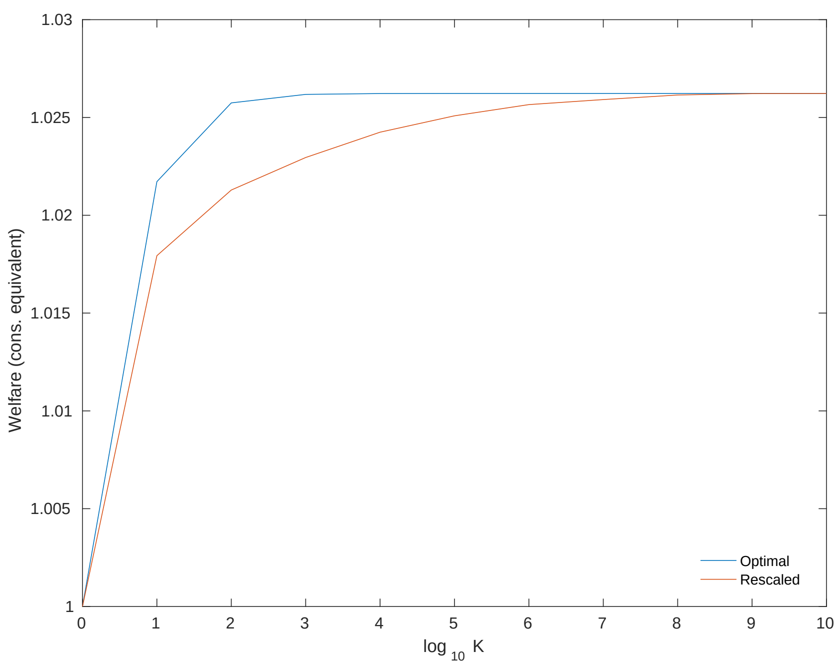
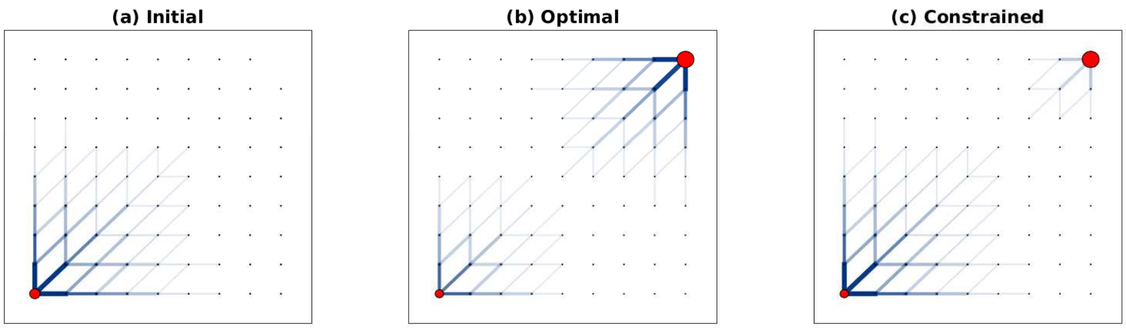


Figure A.4: Inefficient Network Lock-in

(a) Lock-in and Optimal Network



(b) Welfare Differences between Optimal and Lock-in Network

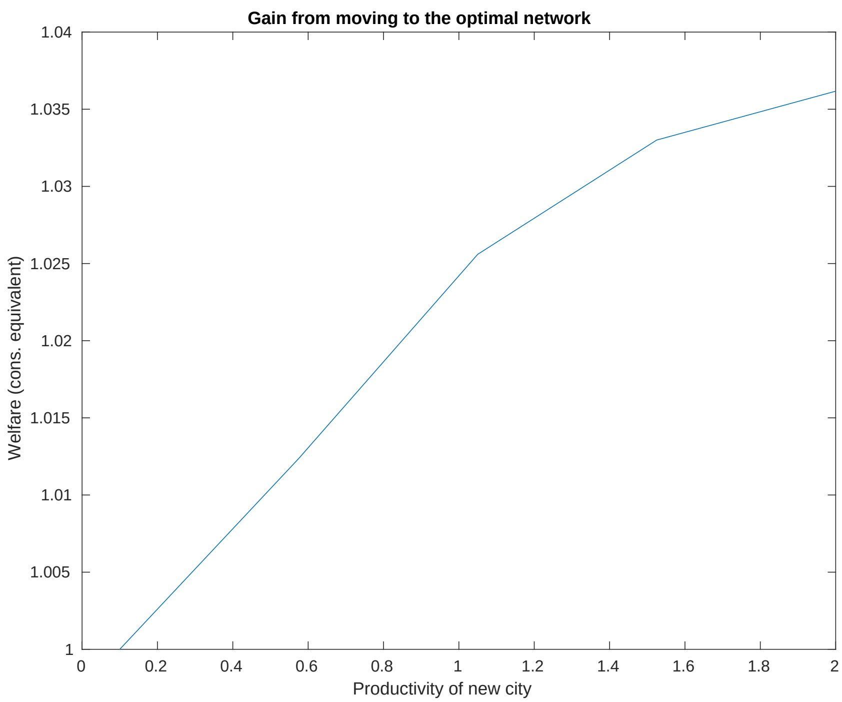
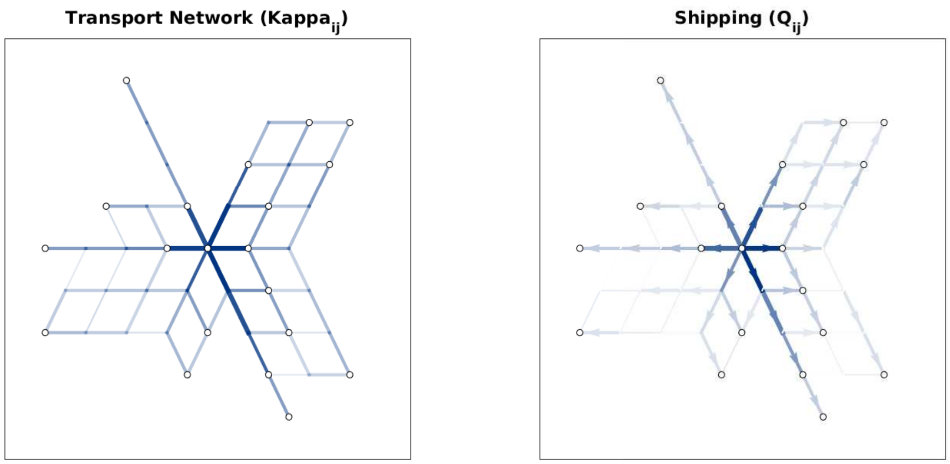
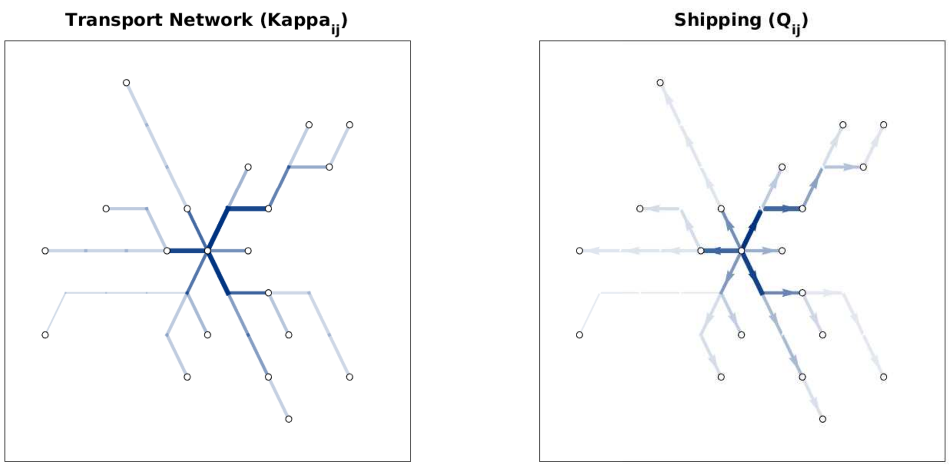


Figure A.5: Optimal Network with Random Cities, Convex and Non-Convex Cases

(a) Convex Case: $\gamma = \beta = 1$



(b) Non-Convex Case: $\gamma = 2 > \beta = 1$



(c) Optimal Capacities Before and After Annealing Refinement in Non-Convex Case

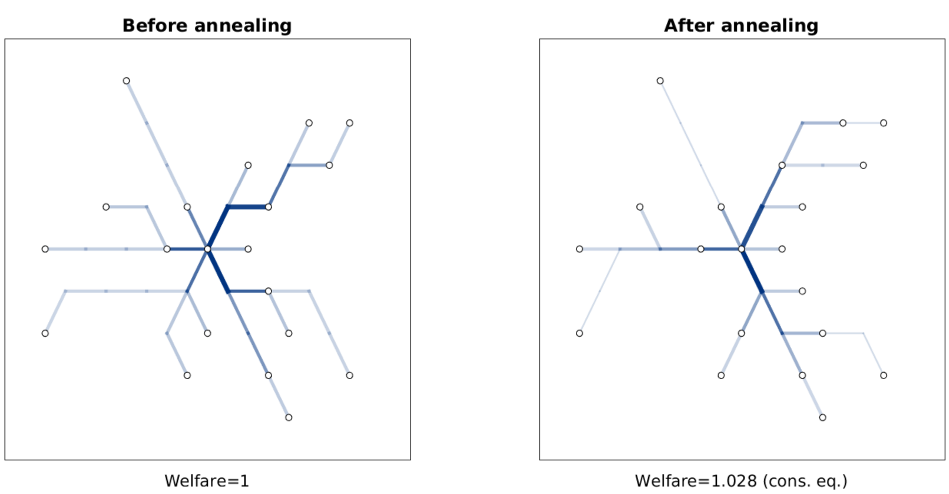


Figure A.6: Optimal Network with 11 Goods, Convex Case, No Labor Mobility

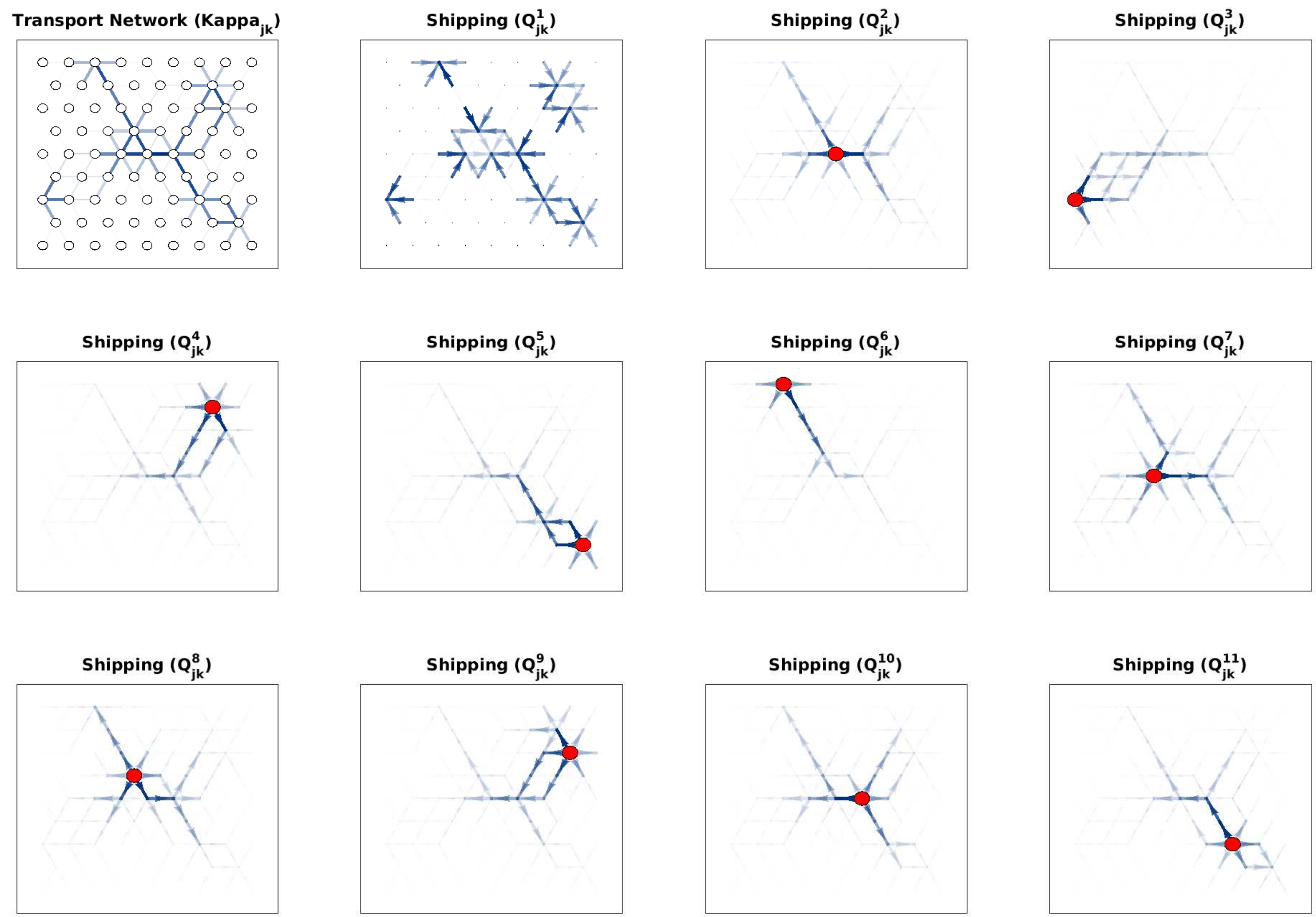


Figure A.7: Optimal Network with 11 Goods, Nonconvex Case, No Labor Mobility

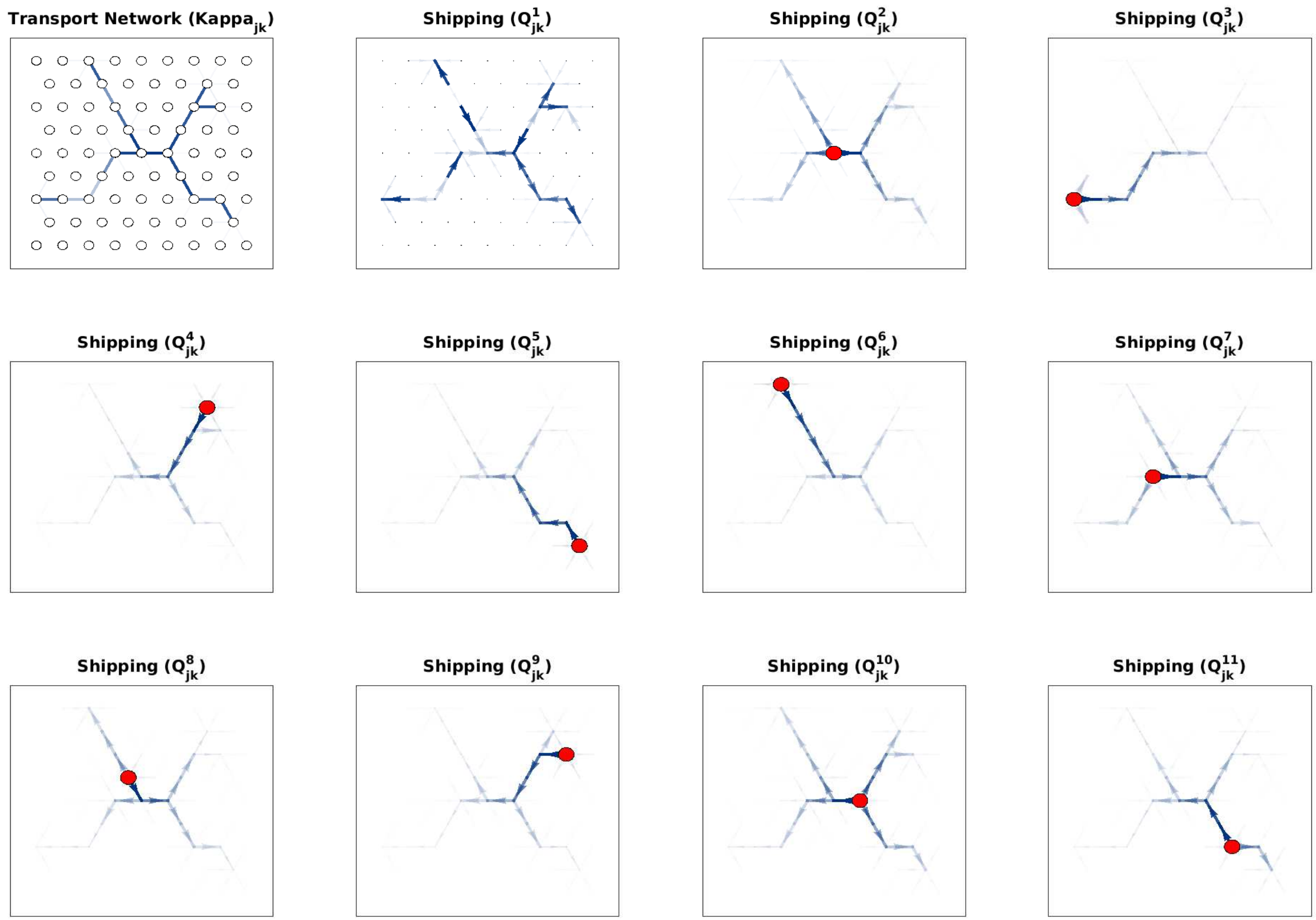


Figure A.8: Optimal Network with 11 Goods, Convex Case, Labor Mobility

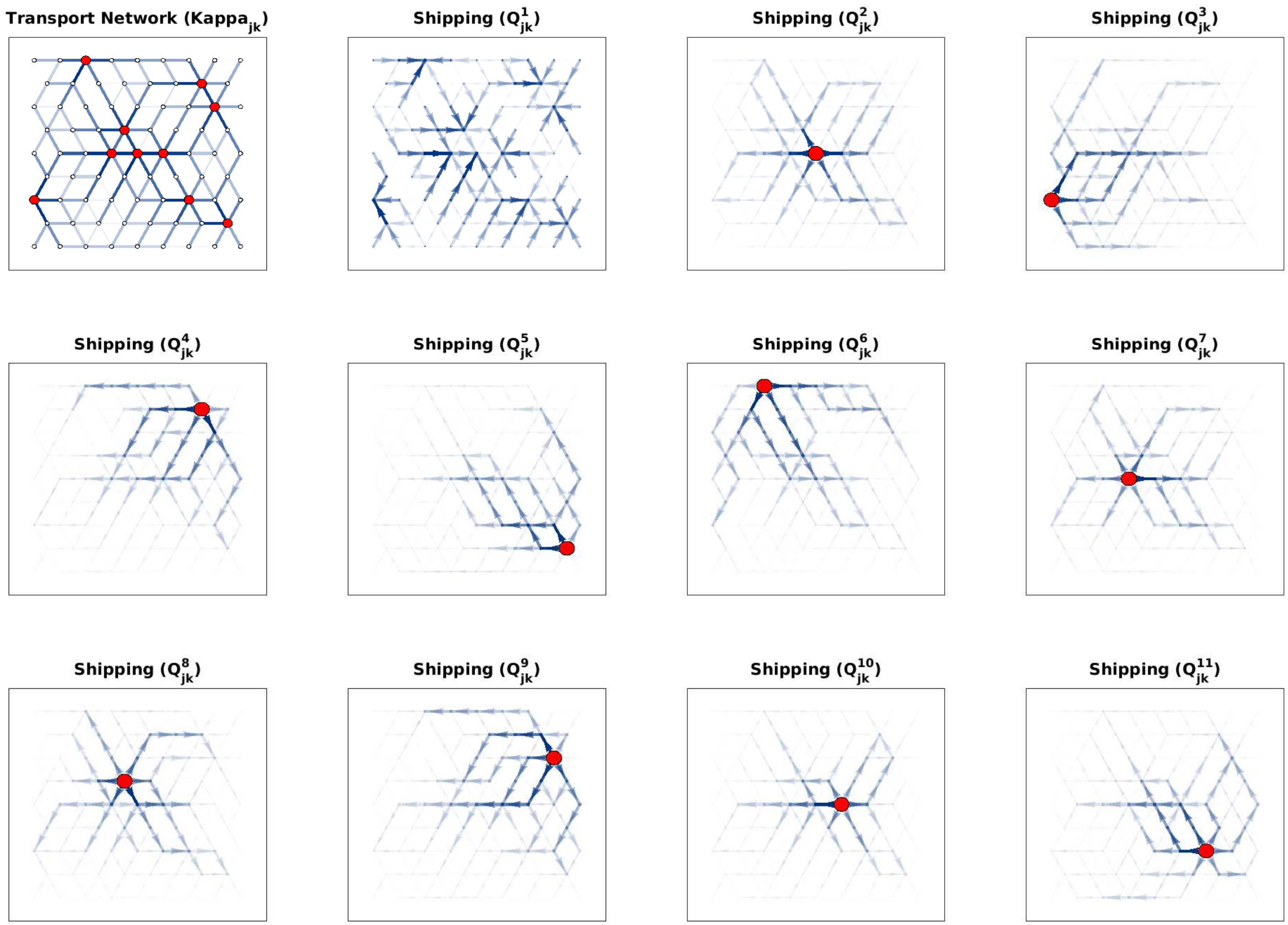


Figure A.9: Optimal Network with 11 Goods, Nonconvex Case, Labor Mobility

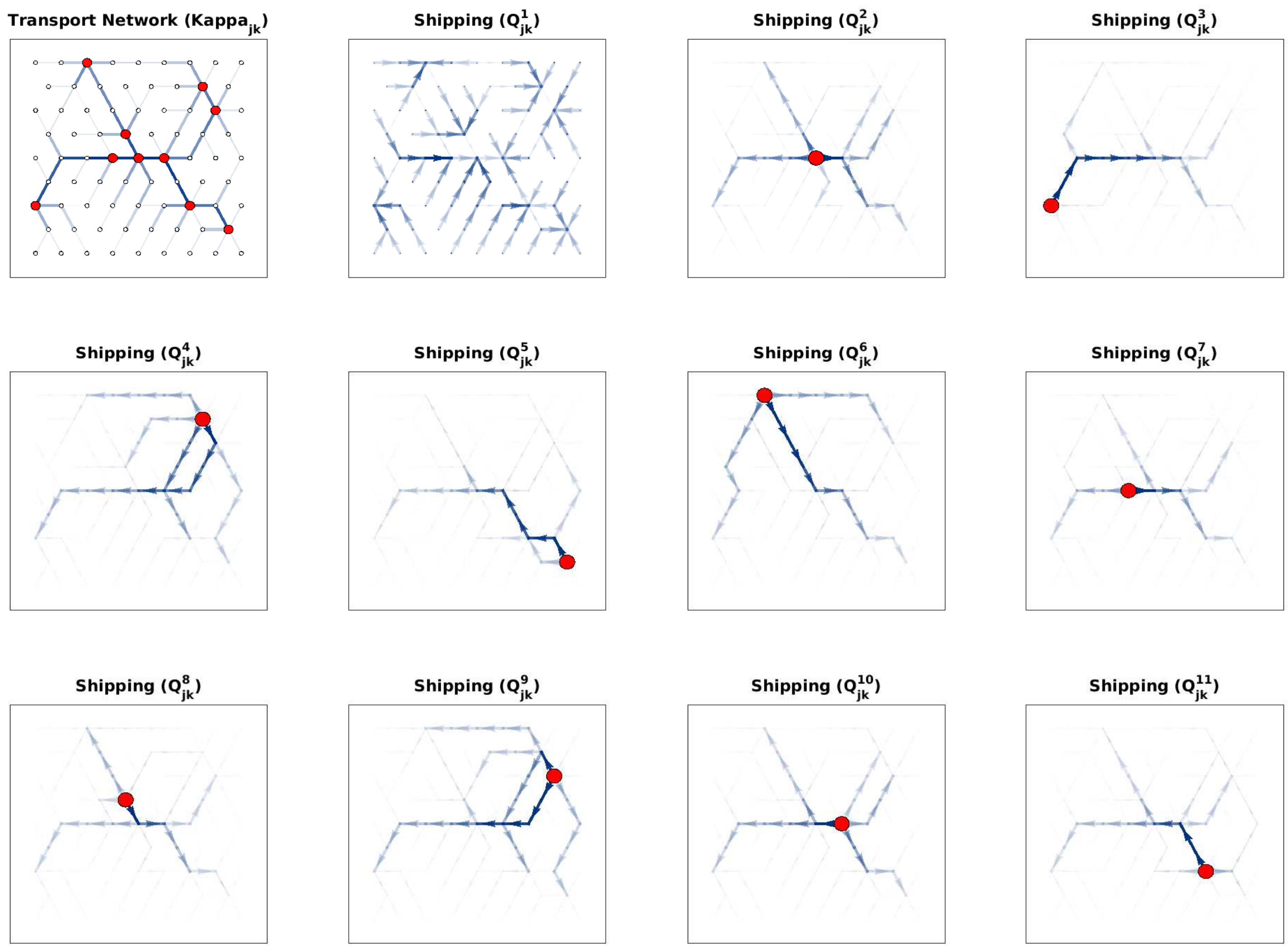
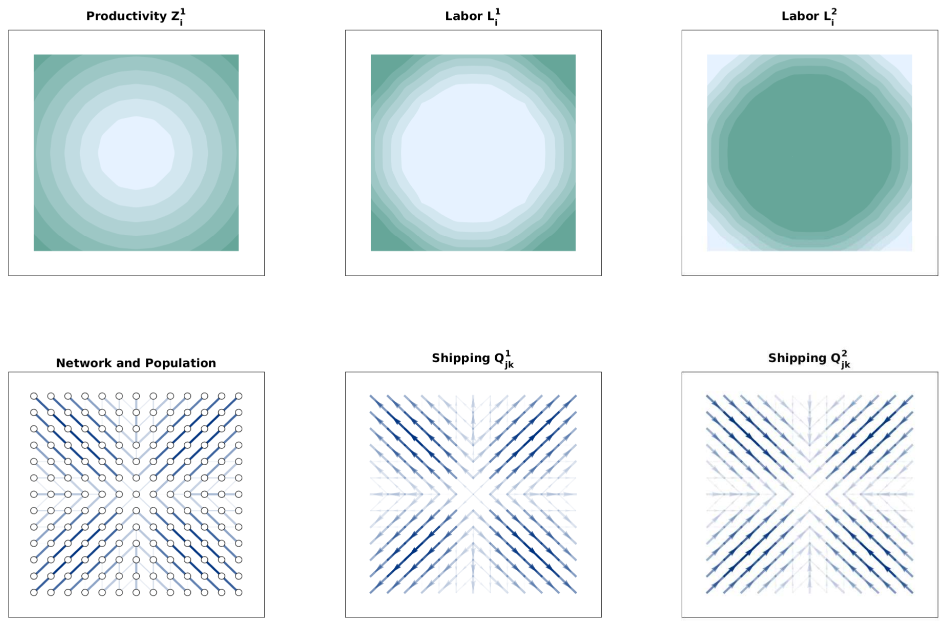


Figure A.10: A 2-Goods Ricardian Economy

(a) No Labor Mobility



(b) Labor Mobility

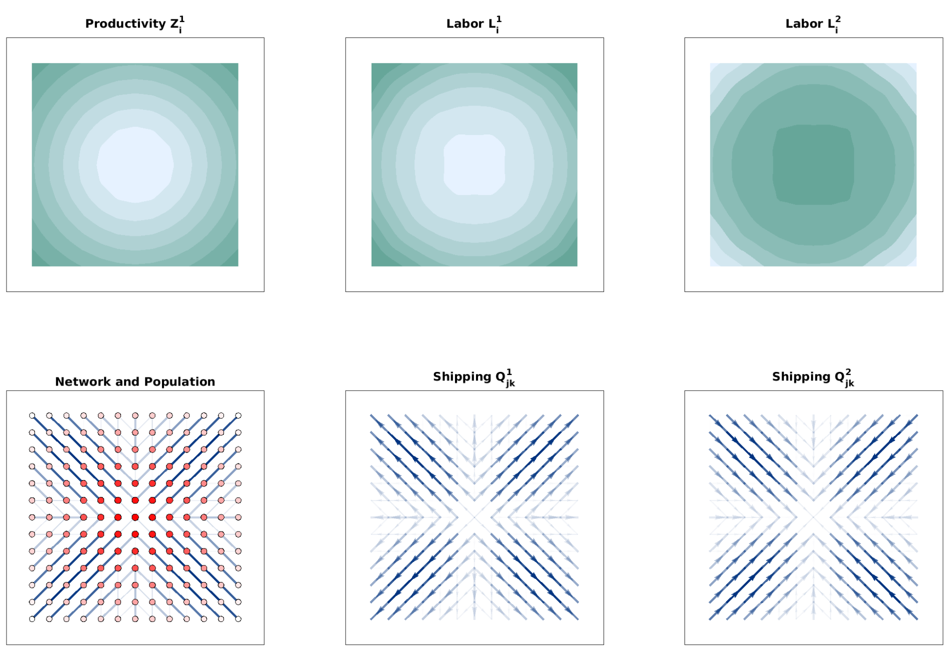
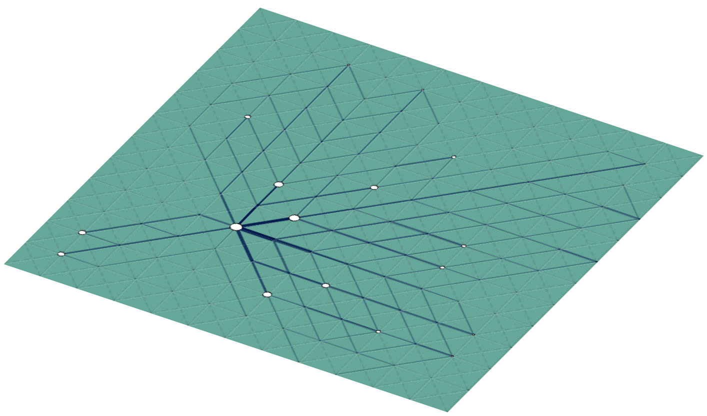
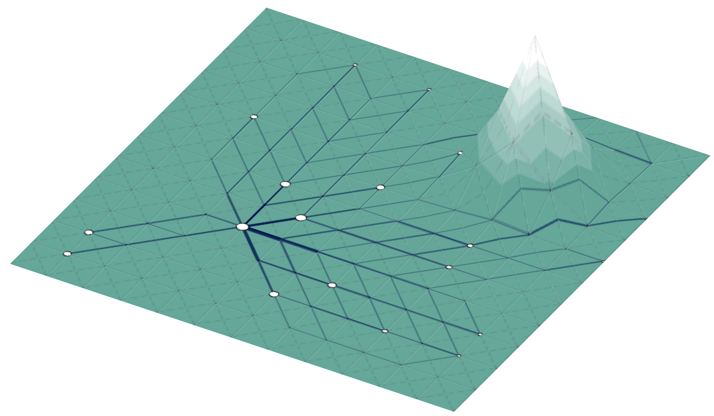


Figure A.11: The Optimal Transport Network under Alternative Building Costs

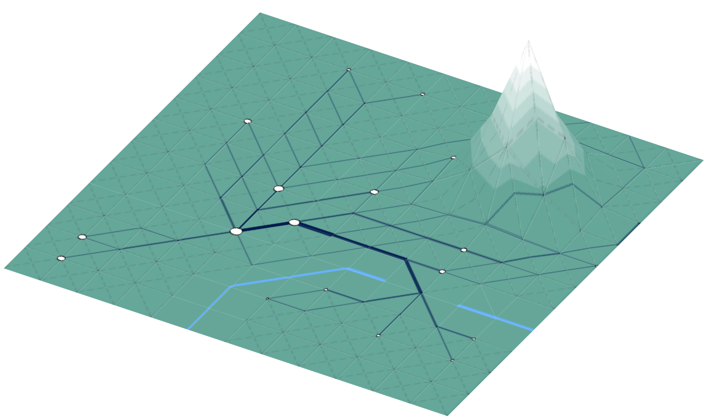
(a) Baseline Geography



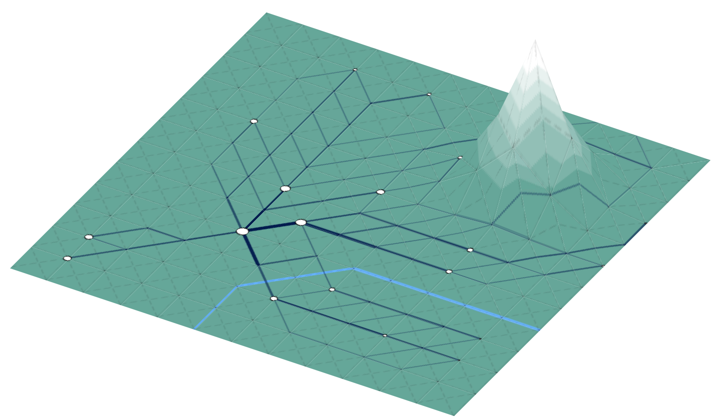
(b) Adding a (Gaussian) Mountain



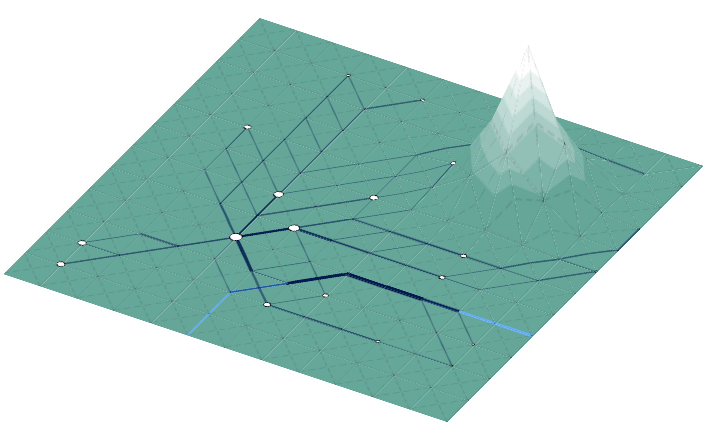
(c) Adding a River and an Exogenous Bridge



(d) Adding Endogenous Bridges



(e) Adding Water Transport



(f) Non-Convex Case ($\gamma = 2$; $\beta = 1$) with Annealing

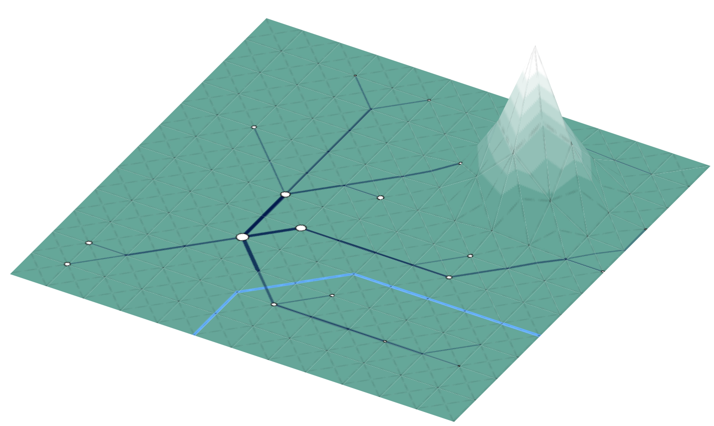
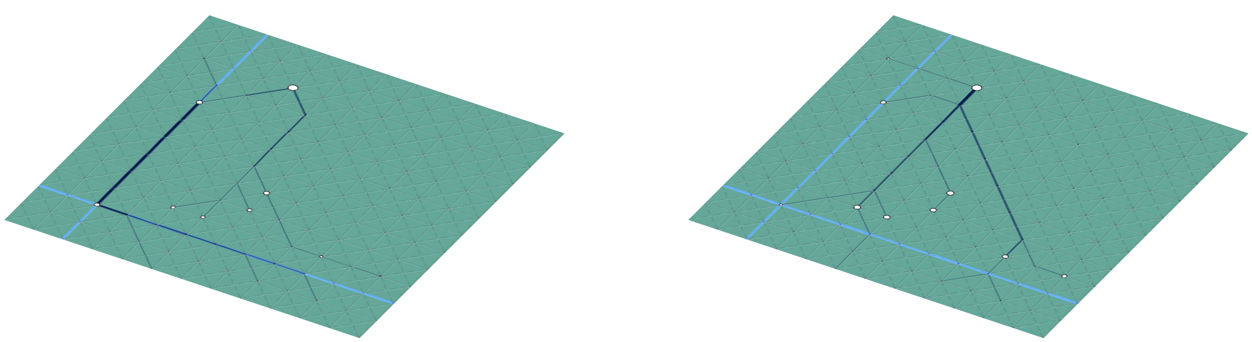


Figure A.12: Arrival of a New Transport Technology and Network Reoptimization

(a) Initial Geography Dependence on Water Transport

(b) Allowing for Cheap Land Transport



B Proofs

Proposition 1. *(Convexity of the Planner's Problem) Given the network investments $\{I_{jk}\}$, the optimal-transport and neoclassical allocation problems are convex if $Q\tau_{jk}(Q, I_{jk})$ is convex in $Q \in \mathbb{R}_+$ for all j and $k \in \mathcal{N}(j)$. The full planner's problem from Definition* (1) *(resp. Definition* (2)*) is a convex (resp. quasiconvex) optimization problem if $Q\tau_{jk}(Q, I)$ is convex in $Q \in \mathbb{R}_+$ and $I \in \mathbb{R}_+$ for all j and $k \in \mathcal{N}(j)$. When the transport technology is given by* (5) *and* (6)*, this property is ensured by $\beta \geq \gamma$.*

Proof. Consider the planner's problem from Definition 1. We can write the problem as

$$\max_{\left\{C_j, \left\{C_j^n, \left\{Q_{jk}^n, I_{jk}\right\}_{k \in \mathcal{N}(j)}\right\}\right\}_{\forall j}} f_0 = \sum_j \omega_j L_j U\left(\frac{C_j}{L_j}, \frac{H_j}{L_j}\right)$$

subject to: (i) availability of traded commodities,

$$f_1 = C_j - C_j\left(C_j^1, .., C_j^N\right) \leqslant 0 \textit{ for all } j;$$

(ii) the balanced-flows constraint,

$$f_2 \equiv C_j^n + \sum_{k \in N(j)} Q_{jk}^n\left[1 + \tau_{jk}\left(Q_{jk}^n, I_{jk}\right)\right] - z_j^n F\left(L_j^n, V_j^{1n}, .., V_j^{Kn}\right) - \sum_{i \in N(j)} Q_{ij}^n \leq 0 \text{ for all } j,\, n;$$

(iii) the network-building constraint,

$$\sum_j \sum_{k \in \mathcal{N}(j)} I_{jk} \leq K;$$

and conditions (iv)-(v) in the text. Since constraints (iii)-(v) are linear, we just need f_0 to be concave and f_1 and f_2 to be convex. Since $\frac{\partial U}{\partial c} < 0$, f_0 is concave. $C_j\left(\left\{C_j^n\right\}\right)$ is concave, it is concave, hence f_1 is convex. If $Q\tau_{jk}(Q, I)$ is convex in then f_2 is the sum of linear and convex functions, hence it is convex.

Consider now the planner's problem with labor mobility from Definition 2. Because U is homothetic we can express it as $U = G(U_0(c, h))$, where G is an increasing continuous function and U_0 is homogeneous of degree 1. Therefore, imposing the change of variables $U_j = L_j G^{-1}(u)$, the planner's problem can be restated as

$$\max_{C_j, \left\{C_j^n, L_j^n, \bar{V}_j^n, \left\{Q_{jk}^n, \kappa_{jk}^n\right\}_{k \in \mathcal{N}(j)}\right\}, U_j, L_j} \min_j \left\{\frac{U_j}{L_j}\right\}$$

subject to the convex restrictions (i)-(v) above as well as

$$U_j \leq U_0(C_j, H_j) \textit{ for all } j;$$

The objective function is quasiconcave because $\frac{U_j}{L_j}$ is quasiconcave and the minimum of quasiconcave functions is quasiconcave. In addition, all the restrictions are convex. Arrow and Enthoven (1961) then implies that the Karush-Kuhn-Tucker conditions are sufficient if the gradient of the objective function is different from zero at the candidate for an optimum, and here the gradient never vanishes.

Finally, assuming (5) and (6), we have

$$Q_{jk}^n \tau\left(Q_{jk}^n, G(I_{jk})\right) = \frac{\delta_{jk}^\gamma}{I_{jk}^\gamma}\left(Q_{jk}^n\right)^{\beta+1},$$

which is convex if $\beta \geqslant \gamma$.

□

Proposition 2. *When the transport technology is given by* (5) *and* (6) *, $\gamma > \beta$ and there is a single commodity, the optimal transport network is a disjoint union of trees.*

Proof. TBC.

$\square$

Definition 3. The decentralized equilibrium without labor mobility consists of quantities $c_j, h_j, C_j, C_j^n, L_j^n, \bar{V}_j^n, \{Q_{jk}^n\}_{k\in\mathcal{N}(j)}$, goods prices $\{p_j^n\}_n, p_j^C, p_j^H$ in each location j and factor prices $w_j, \{r_j^m\}_m$ in each location j such that:

(i)(a) consumers optimize:

$$\{c_j, h_j\} = \arg\max_{c_j^0, h_j^0} U\left(c_j^0, h_j^0\right)$$
$$p_j^C c_j^0 + p_j^H h_j^0 = e_j \equiv w_j + t_j,$$

where e_j are expenditures per worker in j and where p_j^C is the price index associated with $C_j\left(c_j^1, .., c_j^N\right)$ at prices $\{p_j^n\}_n$ and t_j is a transfer per worker located in j. The set of transfers satisfy

$$\sum_j t_j L_j = \Pi$$

where Π adds up the aggregate returns to the portfolio of fixed factors and the government tax revenue,

$$\Pi = \sum_j p_j^H H_j + \sum_j \sum_m r_j^m V_j^m + \sum_j \sum_{k\in\mathcal{N}(j)} \sum_n t_{jk}^n p_k^n Q_{jk}^n;$$

(i)(b) firms optimize:

$$L_j^n, \bar{V}_j^n = \arg\max_{L_j^{n0}, \bar{V}_j^{n0}} p_j^n F_j^n\left(L_j^{n0}, \bar{V}_j^{n0}\right) - w_j L_j^{n0} - \sum_m r_j^m V_j^{mn0};$$

(i)(c) transport companies optimize

$$p_k^n\left(1 - t_{jk}^n\right) \le p_j^n\left(1 + \tau_{jk}^n\right), \ = \text{if } Q_{jk}^n > 0;$$

(i)(d) producers of final commodities optimize:

$$\{C_j^n\} = \arg\max_{C_j^{n0}} C_j\left(\{C_j^{n0}\}\right) - \sum_j p_j^n C_j^{n0};$$

as well as the market-clearing and non-negativity constraints (i), (ii), (iv), and (v) from Definition 1.

If, in addition, labor is mobile, then the decentralized equilibrium also consists of utility u and employment $\{L_j\}$ such that

$$L_j u = U_j\left(c_j, h_j\right)$$

whenever $L_j > 0$, and condition (vii) from Definition 2 holds. If labor is mobile, we further impose equal ownership across workers regardless of location: $b_j = \frac{1}{L}$.

Proposition 3. *(First and Second Welfare Theorems) If the sales tax on shipments of product n from j to k is $1 - t_{jk}^n = \frac{1+\tau_{jk}^n}{1+\left(\varepsilon_{Q,jk}^n+1\right)\tau_{jk}^n}$ then: (i) if labor is immobile, the competitive allocation coincides with the planner's problem under specific planner's weights ω_j. Conversely, the planner's allocation can be implemented by a market allocation with specific transfers b_j; and (ii) if labor is mobile, the competitive allocation coincides with the planner's problem. In either case, the price of good n in location j, p_j^n, equals the multiplier on the balanced-flows constraint in the planner's allocation, P_j^n.*

Proof. Under the tax scheme in the proposition, condition (i)(c) from the market allocation is equivalent to the first-order condition (7) from the planner's problem. Without labor mobility, the rest of the allocation corresponds to a standard neoclassical economy with convex technologies and preferences where the welfare theorems hold. Specifically, the first-order conditions from the firm optimization problem in the market allocation (i)(b) coincide with the planner's problem letting w_j and r_j^m be the multipliers of the planner's constraints (iv). Because C_j is homogeneous of degree 1, the first-order conditions over C_j^n in the market allocation is $p_j^C \frac{\partial C_j}{\partial C_j^n} = p_j^n$. Letting P_j^n be the multiplier of the balanced-flows constraint (ii) in the planner's allocation, this condition coincides with the planner's first-order conditions if $P_j^n = p_j^n$. Since the market clearing constraints are the same in the market's and the planner's allocation, the planner's allocation coincides with the market if the planner's weights are such that the planner's FOC for C_j coincide with the market. This is the case if the weight ω_j from the planner's problem equals the inverse of the multiplier on the budget constraint from the consumer's optimization problem (i)(a) in the market allocation. To find that weight, using that U we can write $U = G(U_0(c,h))$, where U_0 is homogeneous of degree 1. Then, the planner's allocation coincide with the market's under weights

$$\omega_j = \frac{e_j}{G'(U_0(c_j,h_j))\,U_0(c_j,h_j)},$$

where e_j is the expenditure per worker and c_j, h_j are the consumption per worker of the traded and non-traded good in the market allocation. If U is homogeneous of degree one, then $\omega_j = P_j^U$, where P_j^U is the price index associated with $U(c_j,h_j)$ at the market equilibrium prices p_j^C, p_j^H. In the opposite direction, given arbitrary weights ω_j, the market allocation implements the planner's under the transfers $t_j = P_j^C c_j + P_j^H h_j - W_j$ constructed using the quantities $\{c_j, h_j\}$ from the planner's allocation and the multipliers $\{P_j^C, P_j^H\}$ and W_j corresponding to the constraints (i) and (iv) of the planner's problem, respectively.

For the case with labor mobility, note that adding up the planner's first-order conditions on c_j and h_j gives

$$\begin{aligned} P_j^H h_j + P_j^C c_j &= \left(\frac{\tilde{\omega}_j}{L_j}\right)\left[U_C(c_j,h_j)\,c_j + U_H(c_j,h_j)\,h_j\right] \\ &= W_j - W^L \end{aligned}$$

where $\tilde{\omega}_j$ is the multiplier of the labor-mobility constraint (vi) of the planner's problem, W_j is the multiplier of the local labor-market clearing constraint (iv) in the planner's problem, and W^L is the multiplier of the aggregate labor-market clearing constraint (vii) in the planner's problem. The second equality follows from the first-order condition with respect to L_j in the planner's problem. Then, the market allocation and the planner's solution coincide if, in the market allocation, $e_j = w_j + Constant$, implying $t_j = \frac{\Pi}{L}$, as we have assumed.

$\square$