Online Appendix - Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System

Benjamin Faber*

February 21, 2014

Abstract

This appendix proceeds in five sections. Appendix 1 presents a simple multi-region core-periphery model based on Helpman and Krugman (1985). Appendix 2 presents estimation results concerning the proportion and characterization of complier counties that drive the local average connection effects estimated in the paper. Appendix 3 presents additional estimation and robustness results for both average NTHS connection effects and their heterogeneity with respect to pre-existing county characteristics. Appendix 4 describes the datasets and construction of variables. Appendix 5 describes the construction of the least cost path and Euclidean spanning tree networks.

*Department of Economics, University of California Berkeley; Email: benfaber@econ.berkeley.edu.

Appendix 1: Core-Periphery Model

The model is based on Helpman and Krugman (1985) and adapted to a setting with multiple and *ex ante* asymmetric regions. In addition, I introduce capital as an input to industrial production and allow this factor to be mobile across regions as in Martin and Rogers (1995). This serves to adapt the original cross-country model without factor mobility to a within country setting with partial factor mobility without altering the original set of microeconomic forces at play. The exposition closely follows the footlose capital model in Baldwin *et al.* (2003).

The economy is populated by a continuum of agents who are distributed over R regions. There are two sectors of production, labeled agriculture (A) and industry (M), and two factors of production labeled labor (L) and capital (K). The former is assumed to be immobile across regions, while the latter is mobile. Mobile stocks of capital are owned by workers, and returns to capital are repatriated across regions.

Preferences

The representative consumer in each region has two-tier preferences, where the upper tier is a Cobb-Douglas nest of consumption of agriculture (which will be the numeraire good) and a composite of industrial varieties. Industrial goods enter as a constant elasticity of substitution (CES) sub-utility function defined over a continuum of industrial varieties i(i=1,2,...N). Consumer utility in region j(j=1,2,...R) is given by:

$$U_j = C_{Mj}^{\mu} C_{Aj}^{1-\mu} \qquad C_{Mj} = \left(\int_{i=0}^{N} c_{ij}^{1-1/\sigma} \right)^{\frac{1}{1-1/\sigma}} \qquad 0 < \mu_{hk} < 1 < \sigma$$

C_{Mj} and C_{Aj} are consumption of industry and agriculture in region j respectively, c_{ij} is consumption of manufacturing variety i in region j, μ is the expenditure share on industry, and σ is the elasticity of substitution between varieties. Standard utility maximization yields a constant division of expenditure between sectors and CES demand for an industrial variety i in region j:

$$c_{ij} = \frac{p_{ij}^{-\sigma}}{\int_{i=0}^{N} p_{ij}^{1-\sigma} di} \mu Y_j$$

Y_j is total regional factor income of labor (L_j) and capital (K_j), with wage rate w_j and capital return π_j:

$$Y_j = w_j L_j + \pi_j K_j$$

Technology

The numeraire agricultural sector requires a_A units of labor to make one unit of A. It is subject to perfect competition, constant returns to scale and faces no trade costs. Marginal cost pricing implies that $p_{Aj} = a_A w_j$ and costless trade equalizes prices and wages across regions so that

$p_{Aj} = p_A$ and $w_j = w$ as long as some positive fraction of A is produced in every region.[1] The industrial sector M is subject to increasing returns, Dixit-Stiglitz monopolistic competition and iceberg trade costs. Each firm of a continuum of industrial producers requires one fixed cost unit of capital K, and a_M units of L to produce a unit of M. This implies a cost function $\pi + wa_M x$, where x is firm level output. It is costless to ship industrial goods within a region, but $\tau_{jk} - 1$ units of the good are used up in transportation between two regions j and k. It is assumed that τ_{jk}=τ_{kj}. It proves convenient to define $\phi_{jk} = \tau_{jk}^{1-\sigma}$ as the "freeness" of trade ranging from 0 (prohibitive costs) to 1 (costless trade). Dixit-Stiglitz monopolistic competition and the above demand imply that mill pricing is optimal for industrial firms, so that the price ratio of a variety in an export region k over its local market price in j is τ_{jk}. For a variety i produced in region j but also sold in another region k this is:

$$p_{ij} = \frac{wa_M}{1 - 1/\sigma}, \;\; p_{ik} = \tau_{ik} \frac{wa_M}{1 - 1/\sigma}$$

Equilibrium

Because the marginal cost of industrial firms depends on the immobile factor whose price is pinned down by costless trade in the numeraire sector, industrial f.o.b. prices are equalized across regions and consumer prices differ only by transport costs. As capital enters as fixed cost component in industrial production, this also implies that capital returns are equal to the operating profit of a typical variety. Under Dixit-Stiglitz competition, this is equal to the value of sales divided by σ: $\pi = px/\sigma$. Normalizing the price of agriculture to be the numeraire and choosing units of A such that $p_A = a_A = w = 1$, we can use demand and mill pricing to solve for the equilibrium returns to the mobile factor:[2]

$$\pi_j = \left(\sum_k \frac{\phi_{jk} S_{Yk}}{\sum_m \phi_{mk} S_{Nm}} \right) \frac{\mu}{\sigma} \frac{Y}{K}$$

S_Y represents regional shares of total expenditure, and S_N are regional shares of the mass of total industrial varieties. Y and K stand for total expenditure and the total capital endowment across all regions respectively. Given repatriation of capital returns to immobile owners, regional expenditure shares are a deterministic function over regional shares of capital owners and labor endowments, S_K and S_L respectively:

$$S_{Yj} = \left(1 - \frac{\mu}{\sigma}\right) S_{Lj} + \frac{\mu}{\sigma} S_{Kj}$$

Because capital is freely mobile across regions, there are two possible types of equilibria: core-periphery outcomes where S_N can be 0 or 1, and interior location equilibria. Given all regions

[1] Factor price equalization in the Helpman and Krugman (1985) structure implies a focus on firm relocation as the adjustment channel to equalize profits across regions. An alternative adjustment channel arises in the absence of factor price equalization through wage adjustments across regions (captured by so called wage equations in this literature). See Chapter 12 in Combes *et al.* (2008) for a discussion and formalization of these alternative adjustment channels.

[2] Industrial sales in market j become $\sum_k p_{jk} x = \left(\sum_k \frac{\phi_{jk} S_{Yk}}{\sum_m \phi_{mk} S_{Nm}} \right) \mu \frac{Y}{K}$. Capital returns in market j are thus $\pi_j = \frac{1}{\sigma} \sum_k p_{jk} x$.

maintain some positive fraction of industrial activity, capital returns are equalized so that the long run equilibrium location condition is given by $\pi_j = \pi$ for $0 < S_{Nj} < 1$. The profit equation coupled with inter-regional profit equalization yield a system of R equations that can be solved for an R∗1 vector of regional industrial production shares as a function of an R∗R bilateral trade cost matrix and an R∗1 vector of regional expenditure shares that are in turn determined by regional endowments.[3]

Predictions

The empirical estimations of the paper are based on the comparison of changes of economic outcomes among peripheral county regions that were connected to new NTHS routes relative to non-connected peripheral counties. Given that in general equilibrium it would be a strong assumption that non-connected regions are not affected at all by the network, the most basic policy scenario thus requires at least three regions. Consider two initially identical peripheral regions and one larger metropolitan core region, denoted by superscripts P1, P2 and C respectively, that are identical in terms of tastes, technology, and initial bilateral trade costs. Geometrically, one can think of this scenario as three regions located on the endpoints of an equilateral triangle. The profit equation in the first peripheral region becomes:

$$\pi^{P1} = \left(\frac{S_Y^{P1}}{S_N^{P1} + \phi\left(1 - S_N^{P1}\right)} + \phi \frac{S_Y^{C}}{S_N^{C} + \phi\left(1 - S_N^{C}\right)} + \phi \frac{S_Y^{P2}}{S_N^{P2} + \phi\left(1 - S_N^{P2}\right)} \right) \frac{\mu}{\sigma} \frac{Y}{K}$$

Profits in the core region are isomorphic, and profits in the second peripheral region are given by $\pi^{P2} = 1 - \pi^{P1} - \pi^{C}$. Initial peripheral symmetry implies that $S_Y^{P1} = S_Y^{P2}$, and ϕ is the identical bilateral trade freeness between all three regions at an initial period. We now introduce asymmetric trade integration in the most convenient way. Let $\alpha\phi$ denote the bilateral trade freeness between peripheral region 1 and the core region after a negative bilateral trade cost shock, while ϕ is the unchanged initial trade freeness between all regions. Initially, α=1, while after the trade cost shock takes effect, α is in the range $1 < \alpha < (1/\phi)$. Using peripheral symmetry $S_Y^{P1} = S_Y^{P2}$, introducing the asymmetric trade cost shock $\left(\phi^{P1} = \alpha\phi\right)$, and solving for the equilibrium difference of industrial activity between connected and non-connected peripheral regions subject to profit equalization $\pi^{P1} = \pi^{P2} = \pi^{C}$, we get:

$$S_N^{P1} - S_N^{P2} = \left(\left(\frac{1}{2} - \frac{3}{2} S_Y^{C} \right) \frac{(\alpha - 1)\,\phi}{1 - \alpha\phi} + 1 \right) \frac{1 + \alpha\phi - 2\phi^2}{(1 - \phi)\,(1 + \alpha\phi - 2\phi)} - \frac{3\phi}{1 + \alpha\phi - 2\phi} - 1$$

This provides a closed form solution for peripheral differences in industrial activity as a function of relative market sizes, initial levels of trade costs, and the degree of asymmetric trade integration. At the initial α=1 position, perfect symmetry between peripheral regions leads to $S_N^{P1} - S_N^{P2} = 0$. The question is what happens to industrial production in the connected peripheral county relative to the non-connected one after the trade cost shock materializes. The derivative of interest is $\frac{\partial\left(S_N^{P1} - S_N^{P2}\right)}{\partial\alpha}$. The sign of this derivative in principle depends on the extent of the pre-existing core-

[3]Notice that total profits must be equal to total payments to capital. Also, $\frac{\mu}{\sigma}Y$ must equal to profits. This leaves us with R-1 independent equations. Using those and the fact that the sum of the shares S_N must sum to one, we can solve for S_N in all markets as a function of exogenous trad costs and exogenous expenditure shares.

periphery gradient summarized in S_Y^C, the level of pre-existing trade integration ϕ, as well as the extent of asymmetric trade integration captured by α. It is clear from the expression above that for any given scenario of core-periphery integration $1 < \alpha < (1/\phi)$ and initial trade costs ϕ, the difference in industrial production shares between the integrating and the non-integrating periphery becomes more negative as the core-periphery size asymmetry (summarized by S_Y^C) increases. Using this insight, one can solve for the necessary degree of the core-periphery gradient at which $\frac{\partial\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha} < 0$ holds for any combination of initial trade costs and trade cost shock asymmetry. From the expression above, this is the case as long as the metropolitan region is at least twice the size of an individual peripheral region.

Prediction 1: *Falling trade costs between a sufficiently uneven core-periphery pair of regions lead to a reduction of industrial production in the integrating periphery relative to a non-integrating peripheral control region.*

Descriptive statistics in Table 1 of the paper indicate that the model's size asymmetry threshold is clearly exceeded when comparing non-targeted peripheral counties to the targeted metropolitan city regions. The intuition behind this results is as follows. Equilibrium profits are a positive function of access to consumer expenditure, and decreasing in access to competing industrial producers. The former enters as agglomeration force and the latter as a dispersion force. On one hand, lower trade costs decrease the relative disadvantage of higher product market competition in the larger market because the relative increase in competition is stronger for the smaller region. On the other hand, lower trade costs also decrease the market access advantage of the larger region because the relative increase in market access is stronger for the smaller region. The microfoundation of the home market channel is that falling trade costs attenuate the dispersion force at a faster rate than the agglomeration force.

The model makes additional predictions about county level changes to overall GDP and agricultural output. Aggregate GDP moves in parallel to industrial output, but less than proportional because labor formerly used in industry remains productive in the region. Prediction 1 thus holds for aggregate production, but we expect a lower point estimate on the elasticity of peripheral GDP to trade cost reductions compared to industrial activity. Conversely, the reallocation of labor to the agricultural numeraire sector implies that falling trade costs have the opposite effect on agricultural output growth.[4]

Prediction 2: *The negative effect of integration holds, but to a lesser extent, for total regional production, and is reversed in sign for agricultural production.*

In addition to the predictions on the average effects of integration among peripheral counties, the richness of the empirical setting also allows to test how the home market channel should affect peripheral counties differently. The first cross-derivative prediction is that the home market effect should be more pronounced among peripheral counties whose initial level of trade costs *vis-à-vis* the core region is lower: $\frac{\partial^2\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha\partial\phi} < 0$. For a given trade cost reduction, the marginal effect on industrial and aggregate production should be more negative at higher initial levels of ϕ.

[4]To see this more clearly, we can write differences in total production and differences in agricultural output as a function of differences in industrial production as follows: $GDP^{P1} - GDP^{P2} = \left(S_N^{P1} - S_N^{P2}\right)\frac{\mu}{\sigma}Y + \left(S_L^{P1} - S_L^{P2}\right)L$ and $GDP_{Ag}^{P1} - GDP_{Ag}^{P2} = \left(S_L^{P1} - S_L^{P2}\right)L - \left(S_N^{P1} - S_N^{P2}\right)\mu\frac{\sigma-1}{\sigma}Y$.

Prediction 3: *The negative effects of integration on industrial and total production are more pronounced among peripheral regions with initially lower trade costs to the larger core region.*

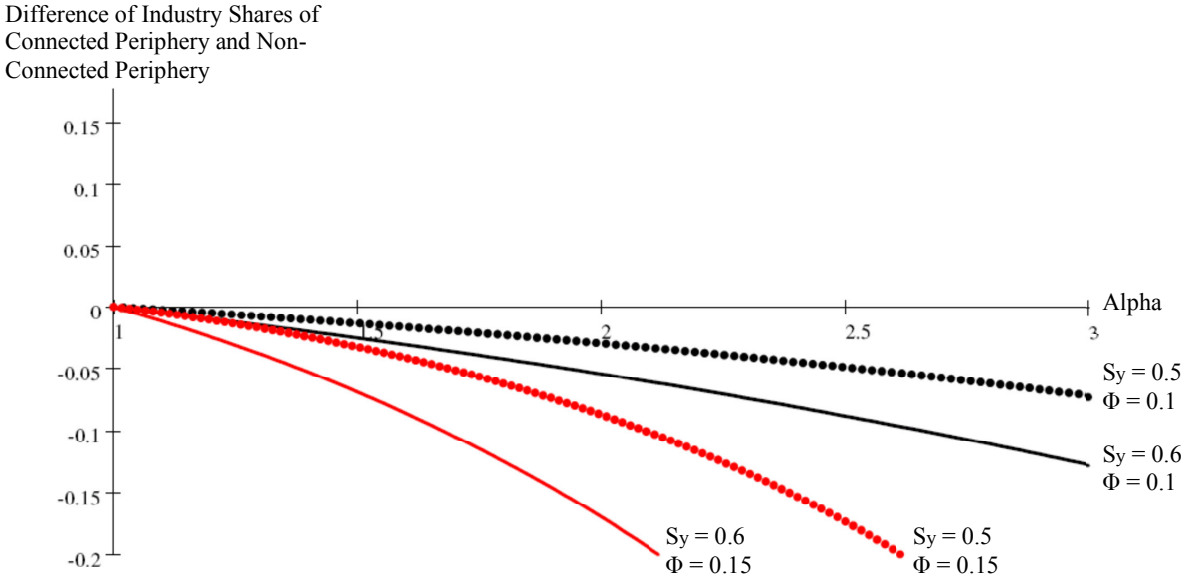
This interaction effect is related to what the trade literature has referred to as home market magnification (Baldwin *et al.*, 2003). The intuition is that falling trade costs attenuate the peripheral location advantage of less market crowding at a faster rate than the metropolitan market access advantage, so that at lower initial trade costs between core and periphery a given trade cost reduction will require a larger relocation of industrial production to equalize the rate of capital return. The second cross-derivative prediction is that, holding initial trade freeness constant, the home market effect is stronger among peripheral counties whose size differential *vis-à-vis* the core is more pronounced: $\frac{\partial^2\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha\partial S_Y^C} < 0$.

Prediction 4: *The negative effects of integration on industrial and total production are more pronounced among peripheral regions with an initially stronger market size differential to the core region.*

The prediction that the home market channel should operate more strongly among smaller peripheral regions is also intuitive. Falling trade costs weaken the dispersion force at a faster rate than the agglomeration force, so that for a larger core-periphery size gradient, and thus higher initial levels of agglomeration and dispersion forces, a given trade cost reduction requires more industrial concentration in the core to equalize profits.

Figure 1.1 provides a graphical illustration of Predictions1-4.

Figure 1.1: Plotting Predictions 1-4



The x-axis displays the degree to which the policy treatment lowers the trade cost of the connected peripheral region to the core region relative to the non-connected peripheral region. The axis starts at the initially identical trade freeness *vis-à-vis* the metropolitan core (α=1). The y-axis displays the difference of industrial production shares between the connected and the non-connected peripheral regions. Sy is the share of total expenditure located in the metropolitan region, and Φ is the initial trade freeness parameter between all regions.

Appendix 2: Local Average Connection Effects

The instrumental variable estimates presented in the paper represent the local average treatment effect (LATE) of network connections among peripheral counties whose treatment status is affected by location along the all-China least cost spanning tree network. The evaluation literature refers to the latter category as "compliers", as opposed to "always taker" counties that were connected despite their location away from the spanning tree paths.[5]

Descriptive statistics and the pattern of coefficient estimates discussed in the paper suggest that planners targeted economically prosperous counties on the way between targeted city regions. In this empirical context, the concern addressed in this additional set of estimations is that least cost spanning tree location might have affected actual highway placements only for a subset of remote and economically stagnant counties on the way between targeted nodes, so that the estimated local average NTHS connection effects might systematically differ from population average effects.

While it is not possible to identify the complier status of individual counties in the county sample, it is possible to estimate the proportion of compliers among the treated counties as well as their observable characteristics (Abadie, 2003; Angrist and Pischke, 2008). The proportion of compliers among all actually treated NTHS counties is given by:

$$P\left(C_{1i} > C_{0i} | C_i = 1\right) = \frac{P\left(C_i = 1 | C_i > C_{0i}\right) P\left(C_{1i} > C_{0i}\right)}{P\left(C_i = 1\right)}$$

$$= \frac{P\left(z_i = 1\right)\left(E\left(C_i | z_i = 1\right) - E\left(C_i | z_i = 1\right)\right)}{P\left(C_i = 1\right)}$$

where C_i is actual NTHS connection status of county i, C_{1i} and C_{0i} are the connection status in cases where the instrument predicts treatment or not, P and E are probability and expectation operators, and z_i is the treatment value of county i as predicted by the instrument. The second equality makes use of the two facts that the total size of the complier group is given by the Wald first stage, and that by independence $P\left(z_i = 1 | C_{1i} > C_{01}\right) = P\left(z_i = 1\right)$. The proportion of compliers among treated counties can then be expressed as the product of the first stage estimate and the proportion of predicted treatments, divided by the proportion of actually treated counties. As presented in the first Column of the Table 2.1, this proportion is estimated to be 22% for both least cost path as well as the Euclidean spanning tree instruments.

The critical question is to what extent these complying counties could be systematically different from the rest of the treated counties. In the case of binary treatments and binary dependent characteristics we know that the relative likelihood of compliers falling into the binary observable category is given by the ratio of the first stage Wald estimated for a particular subgroup over the full sample first stage estimate. To this end, Columns 2-6 of Table 2.1 report first stage point estimates in stated order for counties with above mean 1997 levels of population, urban population, the share of urban population, GDP, and GDP per capita.

If the concern was true that the estimated local average treatment effects are unrepresentative of the population average effects, then we would expect the first stage predictive power of the in-

[5] An implicit assumption is that there are no "defiers" in this context, as location along advantageous construction routes does not cause counties not to be on the network.

struments to differ significantly across observable pre-existing county characteristics. As discussed above, this would constitute evidence of significantly different likelihoods of observables among always takers as opposed to complying counties. In particular, in the present setting one would be concerned that the instrumental variable connection predictors would have a lower estimated effect on actual NTHS route placements among the large, urbanized, and rich county groups represented in Columns 2-6.

The reported results provide evidence against this concern. In particular, the first stage point estimates do not significantly differ from the full sample first stage estimate when estimated for different subsamples of counties as indicated across the columns. Figure 2.1 then takes a closer cartographic inspection of actual as opposed to predicted route placements to offer two plausible explanations for the absence of clear observable differences between compliers and always takers.

The figure depicts two snapshots at the county level of the PR China in which counties are color coded according to their nominal levels of GDP in 1997. Both cases compare actual NTHS route placements to predictions from the least cost path spanning tree instrument. Case A illustrates the first point. It is evident that the least cost path algorithm is subject to prediction errors for both bigger urban and smaller rural counties, even in cases where no obvious incentive for deviations from the least cost path is evident. This has to do with the fact that entire bilateral route segments might have been differently picked by the instrument as opposed to NTHS routes, and most importantly, because planners built many more bilateral routes than the minimum number of edges that the spanning tree algorithm picks.

Case B illustrates the second point. When planners do seem to have deviated from the least cost path for an obvious reason (e.g. connection of a prefecture level capital city on the way as indicated in the figure), then this deviation for one "important" county leads to prediction errors for several "unimportant" counties on the way to this target. Both of these features that are evident from the GIS snapshots tend to work against any systematic correlation between the predictive power of the instrument in the first stage and the potential heterogeneity of the highway effect.

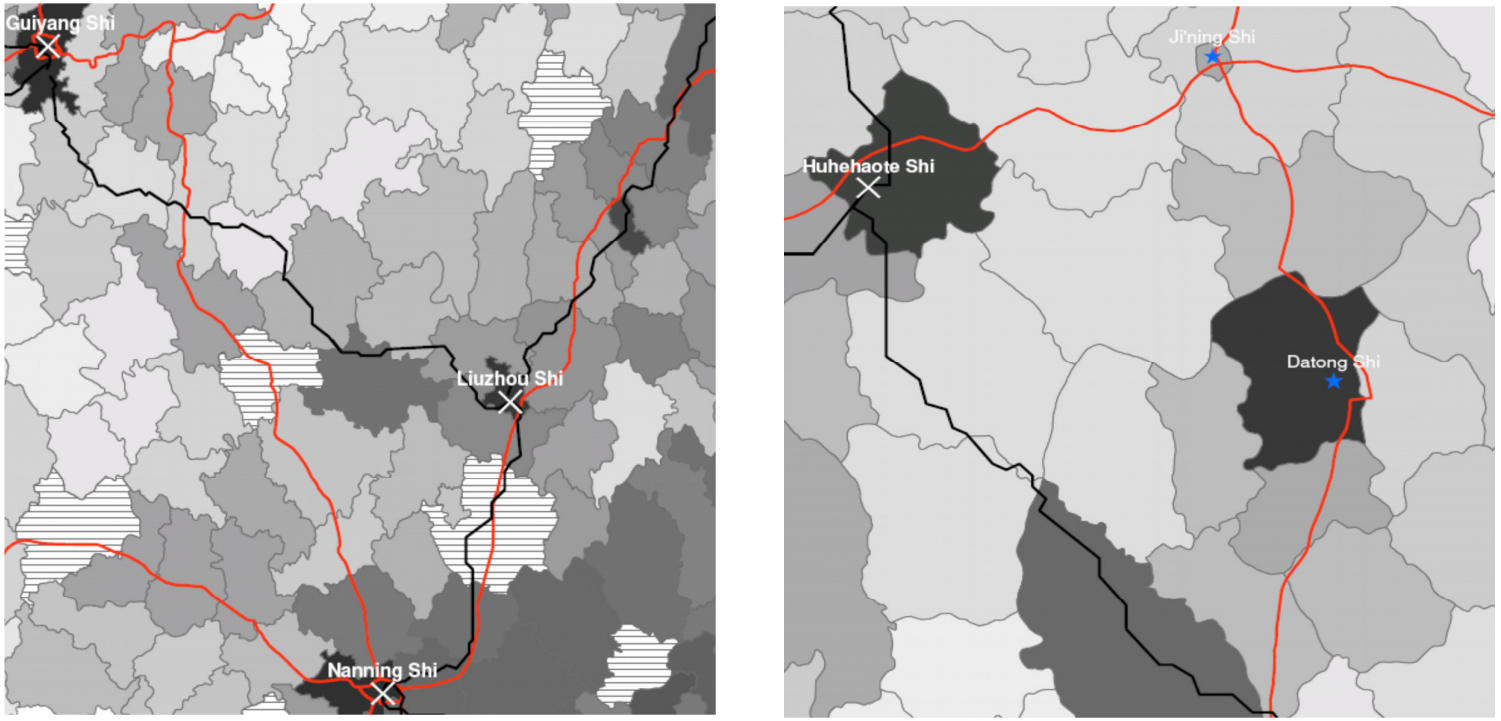
Table 2.1: Estimated Proportion of Compliers and Relative Likelihoods of Observable Characteristics

	(1) Full Sample	(2) Pop 97	(3) Urban Pop 97	(4) %Urban Pop 97	(5) GDP 97	(6) GDP Cap 97
Panel A: LCP IV						
Connect 1st Stage Point Estimate	0.418***	0.383***	0.432***	0.494***	0.399***	0.433***
	(0.0601)	(0.0821)	(0.0704)	(0.0599)	(0.0873)	(0.0869)
F-Statistic p-value [Coef=0.418]		0.677	0.839	0.214	0.832	0.864
Obs	1367	650	662	633	673	664
Estimated Proportion of Compliers Among Treated Counties	0.222					
Panel B: Euclid IV						
Connect 1st Stage Point Estimate	0.314***	0.354***	0.375***	0.328***	0.365***	0.337***
	(0.0492)	(0.0690)	(0.0822)	(0.0776)	(0.0784)	(0.0712)
F-Statistic p-value [Coef=0.314]		0.567	0.462	0.860	0.521	0.750
Obs	1367	650	662	633	673	664
Estimated Proportion of Compliers Among Treated Counties	0.221					

Each point estimate stems from a separate regression. The table presents first stage point estimates for regressions of binary NTHS connections on spanning tree connections and province fixed effects across different county samples. All regressions include province fixed effects. LCP IV stands for the least cost path spanning tree instrument. Euclid IV stands for the straight line spanning tree instrument. The first column presents the full sample first stage estimate. The following columns (in stated order) present this estimate for counties with above median 1997 levels of population, urban population, shares of urban population, GDP, and GDP per capita. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Figure 2.1: Cartographic Inspection of the Instrument

Case A **Case B**



The network in red color depicts actual NTHS expressway routes. The network in black color depicts the least cost path spanning tree network. Crosses indicate targeted metropolitan nodes. Counties are color coded according to their nominal levels of GDP in 1997, where darker colors represent higher values. Striped areas indicate missing 1997 GDP data.

Appendix 3: Additional Estimation and Robustness Results

This section presents additional estimation and robustness results. Table 3.1 presents estimation results when replacing the binary NTHS connection identifier with log distance to the nearest NTHS segment among non-targeted peripheral counties. The presented results confirm the main findings reported in Section 4 of the paper with point estimates of the opposite sign as expected.

Table 3.2 presents estimation results for a series of additional robustness specifications concerning the average NTHS connection effects on county growth discussed in the paper. The first row of results reproduces the baseline estimates of the NTHS connection effect for industrial output growth, non-agricultural output growth, GDP growth, and local government revenue growth for the preferred specification with both instruments and the full set of pre-existing county controls.

The second row of results addresses the concern that the geographical characteristics used in the construction of the least cost path instrument could directly affect county growth and thereby lead to a violation of the exclusion restriction. To address this concern, the table reports results after including the average terrain slope gradient, the percentage of water coverage, the percentage of wetlands coverage, and the percentage of developed land coverage as additional county controls. The NTHS connection effect estimates are unaffected by the inclusion of these additional controls, and show a very slight increase.

The third row of results addresses the concern that location along least costly paths might be subject to stronger endogeneity concerns in mountainous provinces where valleys provide natural advantages for settlements and economic development. The presented results are estimated after excluding the mountain provinces of Gansu, Qinghai, Sichuan, Tibet, and Xinjiang. The exclusion of these regions also address the concern that due to the mountainous terrain a new long distance railway route was built following closely the route of the NTHS between Golmud and Lasa over the same period. The NTHS connection effects are confirmed in sign and statistical significance for all dependent variables when estimated on the restricted county sample.

The fourth row of results addresses the concern that least costly route locations between the major city regions of China are likely to be correlated with historical trade routes. To this end, I obtained geo-referenced routes for the Northern and the Southern routes of the trans-Asian Silk road from the Old World Trade Routes (OWTRAD) Project.[6] The Southern routes of the Silk Road are sometimes referred to as the Tea Horse Road instead. Figure 3.1 provides an illustration of the NTHS network and the Silk Road routes. Reported estimation results include the log distance to the nearest Silk Road segment as an additional county control. The baseline NTHS coefficients are hardly affected by the inclusion of this additional control, indicating that the baseline controls for pre-existing political and economic characteristics have effectively captured county proximity to historical trade routes.

The fifth row addresses the concern that the spanning tree instruments might be picking up county locations with preferential market access positions in the preceding period. These locations could be especially well suited for the process of urbanization and decentralization that Baum Snow *et al.* (2012) have found to be the case for prefecture level central city districts in China during the 1990s. To this end, I compute each county's log market potential in 1997 following Harris (1954)

[6] See www.ciolek.com/owtrad.html.

as the distance weighted sum of all other county seat populations in China where the weights are equal to inverse distances. The fact that the baseline point estimates are virtually unaffected by the inclusion of this additional control suggests that omitted differences in pre-existing market access are not confounding the IV estimates.

The sixth row of results addresses the concern that the initial period output levels among NTHS connected peripheral counties might have been inflated by construction activity already underway in 1997. The concern is that the significant negative effects are driven in part by this inflation of the initial levels of economic activity. To address this concern, I include a dummy indicator for road construction underway in 1997 that I collect from the 1998 Atlas source described in the Data Appendix below. The inclusion of this additional county control hardly affects the baseline point estimates of the NTHS connection effects, indicating that road construction activity underway in 1997 did not lead to a spurious negative growth effect among NTHS connected counties.

The final two rows of Table 3.2 address the concern that the included county controls and province fixed effects do not sufficiently address the potential concern that county location on a spanning tree instrument is correlated with proximity to the coastline. To test for this concern, I include log distance to the nearest coastline, or alternatively a dummy for county location within 50km of the nearest coastline as additional control variables. The fact that the baseline point estimates are virtually unaffected by the inclusion of these additional variables provides evidence against this concern.

Figure 3.2, Table 3.3, and Table 3.4 present additional results about sample selection concerns in the light of administrative boundary changes. Boundary changes over the estimation period 1997-2006 are likely to be correlated with local economic changes over the period. Because estimation results are based on historically consistent administrative units, this gives rise to two econometric concerns. First, if boundary changes mainly occur close to growing cities, then estimation results could be based on an unrepresentative subsample of counties. A second concern is that the NTHS network might itself have affected the propensity of boundary changes among connected peripheral counties relative to the control group. In an extreme scenario we could imagine that the network had large positive effects on a significant fraction of peripheral counties, but those effects drop out of the estimation sample due to the induced boundary changes of increased local growth.

To investigate the first concern, Figure 3.2 and Table 3.3 report to what extent boundary changes are concentrated in particular parts of China and in proximity to large or small city centers. The map suggests that boundary changes are not concentrated along the the Eastern coastline and appear to occur in all parts of China, including the West and the South. In turn, Table 3.3 reports that boundary changes do not appear to have a higher propensity of occurrence in proximity to large or small cities. To address this question more systematically, the final column present estimation results that suggest that counties that reported boundary changes over the period 1997-2006 did not systematically differ in terms of distance weighted access to urban populations across Chinese counties in 1997.[7]

Table 3.4 addresses the second concern that NTHS connections might themselves have affected the propensity for boundary changes. The results suggest that this does not appear to have been

[7]Following Harris (1954), I compute distance weighted access to urban populations as the weighted sum of all other counties' urban populations in 1997 with weights equal to inverse distances to the origin county.

the case. In particular, while NTHS placements appear to be positively correlated to boundary changes in the unconditional OLS regression, this association becomes a statistical zero after controlling for county characteristics, and it turns negative and statistically insignificant in the IV estimations with or without controls. In summary, these additional results provide reassurance against the concern that sample selection might be driving the empirical findings.

Table 3.5 reports additional results of the estimations on interaction effects comparing two stage least squares (2SLS) estimates with limited information maximum likelihood (LIML) estimates when using both spanning tree networks to instrument for NTHS connections and its interaction terms with respect to pre-existing county characteristics. The concern addressed in these estimations is that the drop in the first stage F-statistics among specifications with interaction effects could lead to weak instrument bias. Panel A of the table reports 2SLS results, and Panel B reports LIML results for identical specifications. The table indicates that LIML coefficient estimates on the main NTHS effect and its interaction terms are slightly higher across both dependent variables (industrial output growth and GDP growth). Given that the LIML estimator has been shown to be less affected by weak instrument bias, this finding provides some reassurance against this concern.[8]

Finally, the system of government revenue collection underwent significant reforms in 1994 under the so called tax sharing system (Wong, 2000, Qiao *et al.*, 2008; Lin, 2009). According to these accounts, the main change to the government revenue system was the introduction of the so called tax sharing system (TSS) (fenshuizhi) in 1994, which among Chinese bureaucrats had the objective slogan of "raising the two ratios": the revenue-to-GDP ratio, and the ratio of central government's share to total revenue (Wong, 2000).[9]The question is how these reforms could affect the validity of the placebo estimations in Table 4 of the paper. For this to be a concern, it would have to be the case that the reforms affect changes in local government revenue growth among counties on the spanning tree instruments more negatively relative to counties not located on the spanning tree instruments, conditional on province fixed effects, and controlling for differences in 1990 urban populations, industrialization, skilled labor force shares, city status and prefecture level capital status. The three above accounts in the literature do not raise such a concern.[10]

[8]See for example Angrist and Pischke (2008, Section 4.6) for a discussion 2SLS and LIML estimates in the context of weak instrument concerns.

[9]Further on this from Qiao *et al.* (2008, pp. 114): "*The key measures in the TSS included the introduction of a value added tax as the major revenue source and the setting up of uniform tax-sharing rates for major taxes including VAT, which replaced the previous fixed amount remittance scheme in the FRS. An important measure of the TSS was to split up the old tax service in two and set up a national tax services (NTSs) in all provinces to collect central taxes and shared taxes, and a separate local tax services (LTSs) for the collection of the taxes assigned to local governments.*"

[10]In fact, all three accounts report suggestive evidence that the recentralization of the revenue system in 1994 led to higher inequality in fiscal resources across Chinese regions because the TSS abolished some of the redistributive provisions of the previous system. For example, from Qiao *et al.* (2008, pp. 115): "*In summary, the review of the basic data shows that equity in the geographical distribution of fiscal resources, in the aggregate, has become worse.*" In this paper's empirical context –where the key concern is that route placements pick up richer and more prosperous locations– one would expect this feature (to the extent not already captured by the controls), to work in the opposite direction of the reported findings in Table 4 of the paper.

Table 3.1: Effects of Log Distance to NTHS among Peripheral Counties

Dependent Variables		(1) OLS No Controls	(2) OLS With Controls	(3) LCP IV No Controls	(4) LCP IV With Controls	(5) Euclid IV No Controls	(6) Euclid IV With Controls	(7) Both IVs No Controls	(8) Both IVs With Controls
	lnDistHwy	0.000119	-0.00872	0.0987*	0.0954	0.130	0.135	0.112**	0.113*
Change ln(IndGVA)		(0.0157)	(0.0188)	(0.0545)	(0.0674)	(0.0817)	(0.0906)	(0.0533)	(0.0615)
1997-2006	Obs	1302	1280	1302	1280	1302	1280	1302	1280
	R^2	0.241	0.255						
	lnDistHwy	0.000770	-0.00750	0.0929*	0.0939	0.133*	0.145*	0.109**	0.115**
Change ln(NonAgGVA)		(0.0128)	(0.0146)	(0.0487)	(0.0588)	(0.0692)	(0.0781)	(0.0495)	(0.0566)
1997-2006	Obs	1285	1262	1285	1262	1285	1262	1285	1262
	R^2	0.268	0.284						
	lnDistHwy	0.00184	0.0214	0.0670	0.138*	0.149*	0.233***	0.0994*	0.177***
Change ln(GovRevenue)		(0.0156)	(0.0169)	(0.0588)	(0.0724)	(0.0789)	(0.0853)	(0.0559)	(0.0667)
1997-2006	Obs	1290	1285	1290	1285	1290	1285	1290	1285
	R^2	0.274	0.332						
	lnDistHwy	-0.0156	-0.0111	0.0439	0.0639	0.0805	0.113	0.0589	0.0845*
Change ln(GDP)		(0.0103)	(0.0116)	(0.0385)	(0.0434)	(0.0624)	(0.0685)	(0.0433)	(0.0480)
1997-2006	Obs	1297	1272	1297	1272	1297	1272	1297	1272
	R^2	0.230	0.265						
	lnDistHwy	-0.00769	-0.00589	-0.00535	0.00231	0.00556	0.00979	-0.00108	0.00542
Change ln(AgGVA)		(0.00982)	(0.0103)	(0.0322)	(0.0402)	(0.0328)	(0.0381)	(0.0284)	(0.0340)
1997-2006	Obs	1335	1313	1335	1313	1335	1313	1335	1313
	R^2	0.203	0.208						
	lnDistHwy	-0.00320	0.000279	-0.0208***	-0.0172*	-0.0138	-0.0129	-0.0181**	-0.0154
Change ln(Population)		(0.00219)	(0.00245)	(0.00799)	(0.0103)	(0.0112)	(0.0127)	(0.00796)	(0.00941)
1997-2006	Obs	1337	1314	1337	1314	1337	1314	1337	1314
	R^2	0.235	0.271						

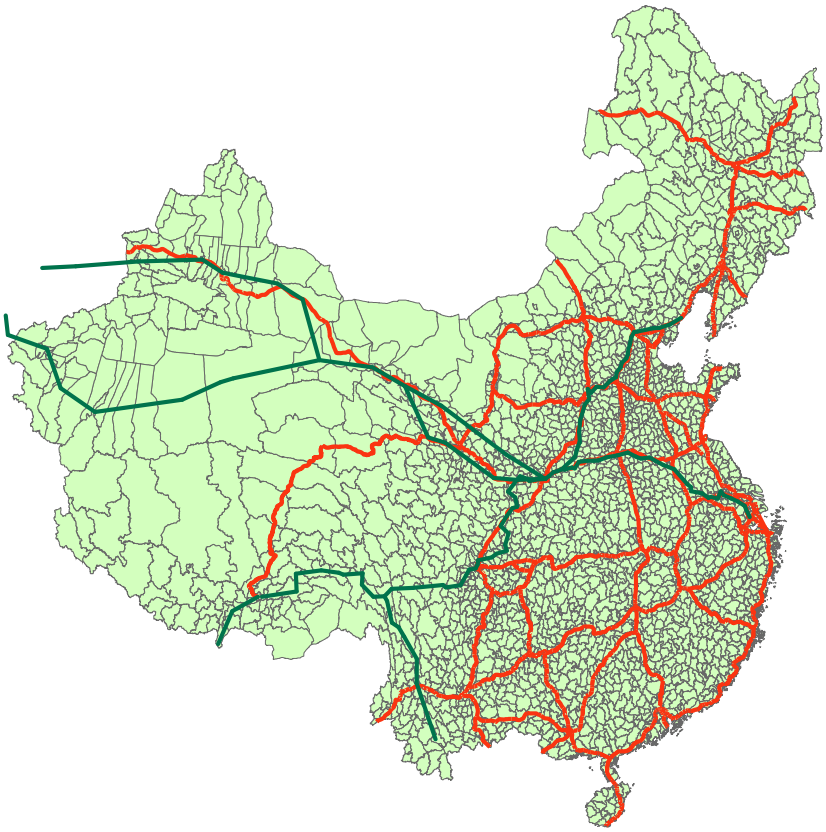
Each point estimate stems from a separate regression. All regressions include province fixed effects. LCP IV stands for the least cost path spanning tree instrument. Euclid IV stands for the straight line spanning tree instrument. No controls columns refer to regressions on NTHS treatment and log county distance to the nearest targeted city node. With Controls indicates a full set of 1990 county controls (city status and prefecture capital dummies, log urban population, share of agricultural employment, and share of above compulsory school attainment in 20+ population). The dependent variables in order as listed are county level industry gross value added, manufacturing plus services gross value added, local government revenue, total GDP, agricultural gross value added, and population. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 3.2: Additional Robustness Specifications

Robustness specifications		(1) Change ln(IndGVA) 1997-06	(2) Change ln(NonAgGVA) 1997-06	(3) Change ln(GDP) 1997-06	(4) Change ln(GovRevenue) 1997-06
Baseline estimates	Connect	-0.297***	-0.268***	-0.203**	-0.257***
		(0.108)	(0.0969)	(0.0886)	(0.0996)
	Obs	1280	1262	1272	1285
Control for direct effects of geographical variables used in least cost path construction	Connect	-0.308***	-0.277***	-0.209**	-0.236**
		(0.110)	(0.0988)	(0.0908)	(0.100)
	Obs	1280	1262	1272	1285
Exclude mountain provinces and Golmud-Lasa Railway	Connect	-0.363***	-0.369***	-0.297***	-0.216*
		(0.122)	(0.100)	(0.0940)	(0.123)
	Obs	1043	1032	1040	1039
Control for log distance to historical trade routes	Connect	-0.300***	-0.272***	-0.206**	-0.251**
		(0.106)	(0.0951)	(0.0874)	(0.102)
	Obs	1280	1262	1272	1285
Control for market access in 1997	Connect	-0.289**	-0.255***	-0.208**	-0.262***
		(0.114)	(0.0961)	(0.0879)	(0.0921)
	Obs	1280	1262	1272	1285
Control for construction underway in 1997	Connect	-0.296***	-0.266***	-0.201**	-0.256**
		(0.111)	(0.0969)	(0.0897)	(0.0996)
	Obs	1280	1262	1272	1285
Control for log distance to coast	Connect	-0.295***	-0.267***	-0.203**	-0.260***
		(0.107)	(0.0959)	(0.0881)	(0.0992)
	Obs	1280	1262	1272	1285
Control for location within 50km of coast	Connect	-0.299***	-0.269***	-0.205**	-0.256**
		(0.109)	(0.0978)	(0.0900)	(0.100)
	Obs	1280	1262	1272	1285

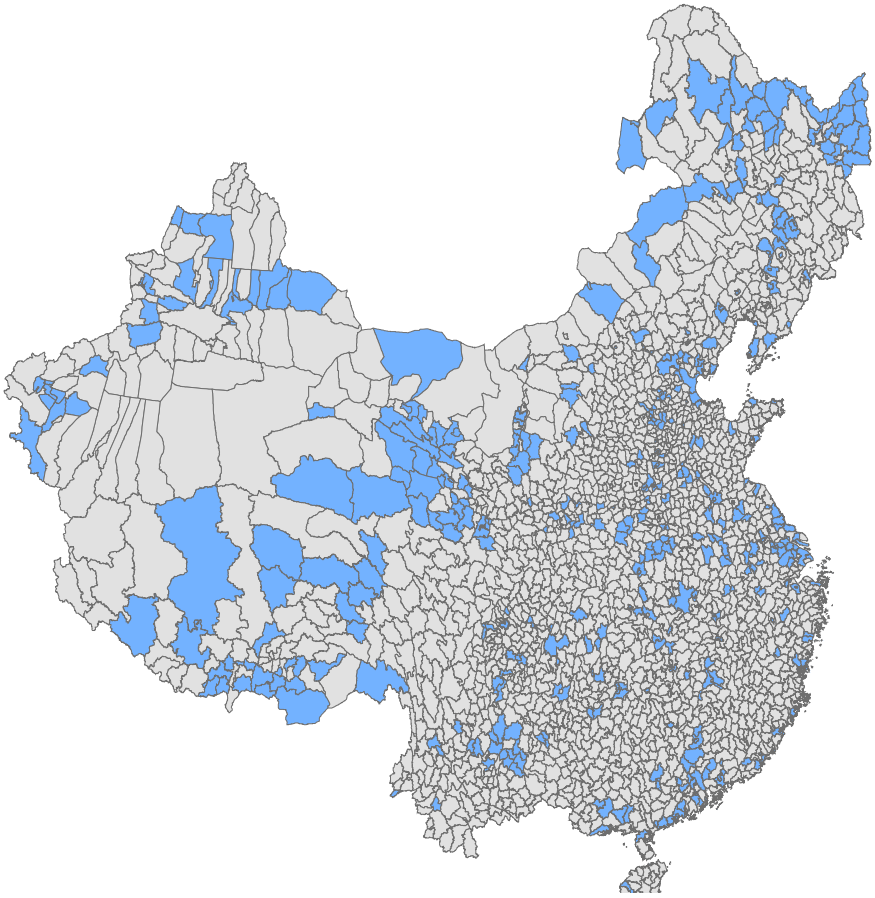
Each point estimate stems from a separate regression. All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage IV estimates using the least cost path and the Euclidean spanning tree networks as instruments for NTHS connections. The dependent variables in order of the columns as listed are log changes of county level industrial gross value added, non-agricultural gross value added, total GDP, and local government revenue. Controls for geographical characteristics used in the construction of the least cost path spanning tree instrument are average county slope, and county percentage of wetland water, or developed coverage. Mountainous provinces refer to Gansu, Qinghai, Sichuan, Tibet, and Xinjiang. Historical trade routes are the Northern and Southern routes of the Silk Road (see map presented below). The control for market access in 1997 is the log of a county's market potential according to Harris (1954), i.e. it is the weighted sum of all other county seat populations in China, where the weight is equal to the inverse of bilateral distances. Control for construction in 1997 refers to a dummy variable indicating all counties with reported “under construction” expressway routes in 1997. The final two rows add controls for log distance to the nearest coastline and county location within 50km of the coastline respectively. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Figure 3.1: The Northern and Southern Routes of the Silk Road



The network in red color depicts the completed NTHS network in 2007. The green routes represent the Northern and Southern Routes of the Silk Road.

Figure 3.2: Boundary Changes during Estimation Period



Boundary changes over the estimation period 1997-2006 are colored in blue. A boundary change is defined as a change in area under administration that exceeds 5% of the initial area in 1997.

Table 3.3: Are Boundary Changes More Likely Near Cities?

Dependent variables:	(1) Distance (km) to Nearest Targeted Centers	(2) Distance (km) to Nearest Prefecture Capitals	(3) Distance (km) to Nearest City	(4) Log Distance Weighted Access to Urban Population in 1997
Boundary Change	27.73	1.572	0.405	-0.110
	(17.58)	(6.711)	(6.660)	(0.0663)
Obs	1,677	1,677	1,677	1,677
R^2	0.005	0.000	0.000	0.011

The table reports mean differences between peripheral counties that do or do not reported boundary changes during the estimation period 1997-2006. Standard errors are clustered at the provice level. ***1%, **5%, and *10% significance levels.

Table 3.4: Do NTHS Connections Lead to Boundary Changes?

Dependent variable: Boundary Change Indicator	(1) OLS No controls	(2) OLS With Controls	(3) IV No Controls	(4) IV With Controls
Connect	0.0601**	0.0297	0.0219	-0.0531
	(0.0239)	(0.0248)	(0.0545)	(0.0662)
Obs	1,677	1,644	1,677	1,644

All regressions include province fixed. Columns 2 and 4 include the full set of county controls. The dependent variable is a dummy that takes the value of 1 if a peripheral county reported a boundary change over the estimation period 1997-2006. IV stands for the least cost path spanning tree instrument. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 3.5: Comparing 2SLS and LIML Estimates of Interaction Effects

Dependent variable:	Change ln(IndGVA) 1997-2006		Change ln(GDP) 1997-2006	
	(1)	(2)	(1)	(2)
Panel A: 2SLS				
Connect	-0.297***	-3.876***	-0.203**	-3.496***
	(0.108)	(1.333)	(0.0886)	(0.948)
Connect*ln(DistNode)		0.680***		0.623***
		(0.232)		(0.161)
Connect*Emp90Dum		0.400		0.396**
		(0.247)		(0.196)
Obs	1280	1280	1272	1272
First stage F-Stat	18.886	2.047	17.425	2.147
Panel B: LIML				
Connect	-0.297***	-3.914***	-0.205**	-3.540***
	(0.108)	(1.353)	(0.0893)	(0.970)
Connect*ln(DistNode)		0.686***		0.631***
		(0.235)		(0.165)
Connect*Emp90Dum		0.406		0.404**
		(0.250)		(0.200)
Obs	1280	1280	1272	1272
First stage F-Stat	18.886	2.047	17.425	2.147

All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage estimates using the least cost path and the Euclidean spanning tree networks to instrument for NTHS connections as well as their reported interaction terms. lnDistNode is log county distance to the nearest targeted city node. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Appendix 4: Data Appendix

GIS Data

Geo-referenced administrative boundary data for the year 1999 were obtained from the ACASIAN Data Center at Griffith University in Brisbane, Australia. These data provide a county-level geographical information system (GIS) dividing the surface of mainland China into 2341 county level administrative units, 349 prefectures, and 33 provinces. Chinese administrative units at the county level are subdivided into county level cities (shi), counties (xian), and urban wards of prefecture level cities (shixiaqu).

Administrative units in China are identified by a system of guo biao codes that allows the matching of records across the GIS and socioeconomic datasets. In addition to guo biao codes, the combination of prefecture and county names were used to double check the consistent matching of administrative units across the datasets. I use reported data on the county area under administration in km^2 from the Provincial Statistical Yearbook series to identify significant boundary changes over time. The historically consistent county sample for estimations on changes 1997-2006 are defined as counties without administrative area changes in excess of 5%. For the placebo falsification test that is estimated on the identical county sample for both the pre- and post-NTHS periods, 1990-1997 and 1997-2007, the same threshold is applied to changes for both periods.

Geo-referenced NTHS highway routes as well as Chinese transport network data were obtained from the ACASIAN Data Center. NTHS highway routes were digitized on the basis of a collection of high resolution road atlas sources published between 1998 and 2007 that is listed below.

(1) China Newest Public Road Atlas (1998), Ha Na Bin Map Publishing Company
(2) China Road Atlas (2002), Shadong Map Publishing Company
(3) China Public Road Atlas (2002), Shandong Map Publishing Company
(4) China Expressway Atlas (2003), People's Transport Press
(5) China Transportation Network Atlas (2003), Guangdong Map Publishing Company
(6) China Road Atlas (2003), Xue Yuan Map Publishing Company
(7) China Automobile Map (2003), China World Map Publishing Company
(8) Chinese People's Road Atlas (2005), Globe Publishing Company
(9) China Road Atlas (2007), Shandong Map Publishing Company

These atlas sources made it possible to classify NTHS segments into three categories: opened to traffic before mid-1997 (10% of NTHS), opened to traffic between mid-1997 and end of 2003 (81% of NTHS), and opened to traffic after the end of 2003 (9% of NTHS). In particular, Source (1) was used to digitize a baseline layer of NTHS routes that were in place by mid year in 1997, and Source (8) was used to digitize a baseline layer of NTHS routes that were in place by the end of 2003. These baseline route maps were then cross-referenced with route information provided in the remaining listed atlas sources. In cases where the remaining atlas sources were at odds with the information of the baseline maps (i.e. routes present in 1997 but not in 2000 or thereafter, or routes present in 2003 but not thereafter), a decision was taken on the basis of the majority of sources (for the 1997 layer), or after tracking down highway openings through press releases on highway opening ceremonies for a small number of cases where Sources (8) and (9) were at odds.

Finally, land cover and elevation data that are used in the construction of least cost path highway routes were obtained from the US Geological Survey Digital Chart of the World (DCW) project, and complemented by higher resolution Chinese hydrology data from the ACASIAN data center. The higher resolution hydrology data from ACASIAN was used to assure that rivers were not interrupted by grid cells coded as mostly covered by land in the lower resolution raster data on land cover obtained from the DCW.

Socio-Economic Data

The Provincial Statistical Yearbook series report production approach county GDP broken up into primary, secondary, and tertiary gross value added. Value added is reported as gross output value less intermediate inputs and value added tax. Traditionally, construction is included together with manufacturing under the secondary industrial sector. The county level data are collected from local establishments under the supervision of the provincial governments, and the Provincial Statistical Yearbooks constitute a separate process of data collection from the national Statistical Yearbook series that is undertaken by the National Bureau of Statistics.[11]

The reported production output data collected by local governments in principle cover the entirety of producing establishments located in the area of the county authority. This is in contrast to central government production statistics that are based on a cut-off of 5 million Yuan annual revenues for so called directly reporting industrial enterprises.[12] The data are collected by teams of local bureaucrats in the form of surveys that are filled out by the establishments located in the jurisdiction of the county.

The Provincial Statistical Yearbooks also provide local government revenues that are reported from the revenue accounts of local authorities. Government revenues mainly consist of industrial and commercial taxes (including value added tax) as well as corporate income taxes (Lin, 2009). The population records contained in the yearbook series refer to locally registered populations under the household registration system.

The CITAS data from the 1990 Population Census provide county level data on population broken up by urban and non-urban, education, and employment shares at the county level. The 1990 Census was the fourth census conducted by the National Bureau of Statistics, and the information therein was recorded on the basis of household questionnaires that were collected locally. Population figures refer to registered county level populations, and agricultural employment shares, as well as above compulsory schooling shares of the population are computed using the county totals and subtotals reported in the CITAS data. The control variable for urban population in 1990 is registered residents in urban wards taken from the CITAS population records.[13]

[11] In the case of multiple central city wards (shixiaqu) of a prefecture level city, these are treated as one county level administrative unit.

[12] This difference is sometimes cited as one of the reasons for discrepancies between the national and the sum of province level economic accounts.

[13] This definition of urban population used in the control variable does not directly correspond with the official Chinese administrative definition. Traditionally two characteristics are used to classify urban residents for Chinese official use. The first is that at least one person of the household holds a "non-agricultural occupation" (industry or services). The second is that the sub-county level administrative unit ("zhen"=ward) is classified urban as opposed to rural.

Appendix 5: Construction of Spanning Trees

This section describes the construction of the least cost path and Euclidean spanning tree networks depicted in Figures 2 and 3 in the paper. The stated objectives of the NTHS in 1992 were to connect all provincial capitals and cities with an urban registered population above 500,000 and connect targeted nodes to border segments as part of the Asian Highway Network in border provinces. In the 1990 Chinese Population Census, 54 cities correspond to these criteria.[14]

The following computation steps have been executed in ESRI's ArcGIS software. To construct the least cost path spanning tree network depicted in Figure 2 of the paper, I adapt a simple construction cost function from the transport engineering literature (Jha *et al.*, 2001; Jong and Schonfeld, 2003).[15]

$$c_i = 1 + slope_i + 25 * Developed_i + 25 * Water_i + 25 * Wetland_i$$

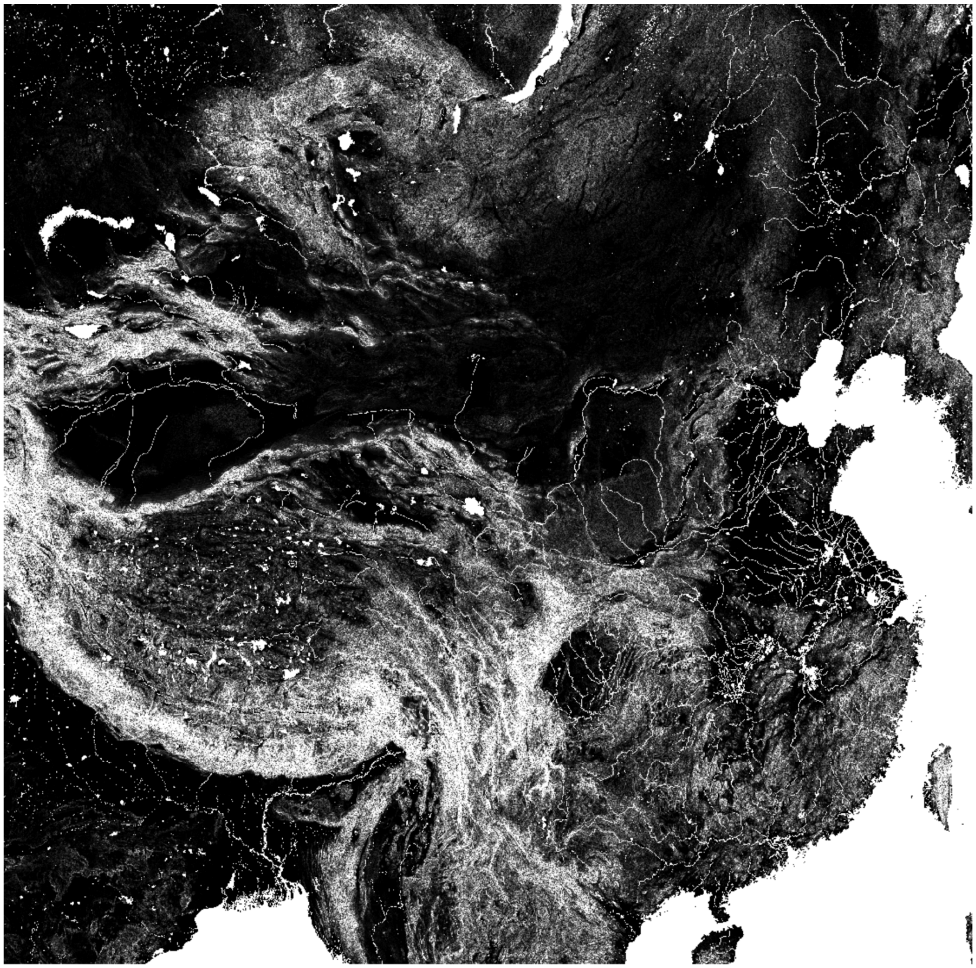
c_i is the cost of crossing a pixel of land i, $Developed_i$ indicates whether the pixel is covered by built structures, $Slope_i$ is i's average slope gradient, $Water_i$ and $Wetland_i$ are dummies indicating whether i is covered by water or wetland. This simple specification of the construction cost function implies that shorter and flatter routes will be preferred, while high costs are assigned to crossing water bodies, wetlands, or built structures. I use the remote sensing data on land cover and elevation described in the previous section to compute c_i for a continuous grid of land parcels covering the PR China. For computational feasibility, I reclassify the original resolutions of the elevation and land cover grids from 30 arc seconds (approximately 0.82x0.82 km^2) to 2x2 km^2 grid cells.[16] This yields an isotropic cost surface grid covering the PR China in a rectangle of approximately 4.7 million 2x2 km grid cells. Figure 5.1 provides a graphical illustration of this construction cost surface.

[14]The records of the 1990 Population Census became available for administrative use in 1991, and constituted the highest quality and most recent information about population registries at the ward level ("zhen") across China at the time of decision making for the NTHS in 1992. According to the Chinese administrative definition, the urban registered population of a central city is the sum of households with urban occupations in the wards ("zhen") of the central city county level units (shixiaqu) of the municipality. 1990 Census population and occupation data at the sub-county ward level was provided by the ACASIAN Data Center from archival records held at Griffith University library.

[15]The choice of the cost factor to be 25 is informed by empirical estimates of the per lane mile construction cost of highways relative to bridges. See for example WSDT (2002).

[16]To assure the continuation of rivers, the reclassified grid cells were classified as covered by water if any of the area was classified as water in the higher resolution grid.

Figure 5.1: Construction Cost Raster



The figure depicts the construction cost raster used as input into the least cost path algorithm. The color scale ranges from white (very high cost of crossing a parcel of land) to black (very low cost of crossing a square km parcel of land). The cost assignment is based on land gradient (slope) as well as land cover (water, wetlands, and developed land), and described in more detail in the text.

I then proceed to construct least cost highway paths between all 1431 $\left(\frac{54*53}{2}\right)$ possible bilateral pairs of targeted city nodes. To achieve this, I follow the accumulative cost minimization procedure pioneered by Douglas (1994). The first step is to compute Dijkstra's (1959) optimal route algorithm to identify the least costly path between each one of the 54 nodes and every cell center of the grid covering the PR China's surface. To calculate the cost of moving from the center of an origin cell to the center of one of the eight directly adjacent cells, there are two types of cost functions subject to which Dijkstra's algorithm is computed:

$$c_{od1} = \frac{c_o + c_{d1}}{2}\pi, \text{ and}$$

$$c_{od2} = \sqrt{2}\frac{c_o + c_{d2}}{2}\pi$$

where c_{od1} is the cost of moving from the origin cell to one of four horizontally or vertically adjacent cells, c_{od2} is the cost of moving to one of four diagonally adjacent cells, c_o, c_{d1} and c_{d2} are the assigned construction costs of the respective cells, and π is the km cell resolution. These

computations result in 54 separate cumulative cost rasters, each containing about 4.7 million 2x2 km pixels covering the PR China. Each cell is assigned the accumulative cost associated with Dijkstra's optimal route solution between any origin cell on the raster to one of the 54 nodal destinations. In addition, the computations yield 54 separate directional backlink raster files, assigning a code between 1-8 to each cell on the grid that indicates the moving direction from any cell to one of its eight adjacent cells on the identified least cost path to the particular targeted node of the grid.

The accumulative cost rasters and directional backlink rasters for each of the 54 nodes ultimately enable me to construct 1431 hypothetical least cost highway construction paths between all possible nodal connections. In the second stage, I then extract the aggregate construction cost of each possible bilateral connection in order to compute Kruskal's minimum spanning tree algorithm. This algorithm identifies 53 least cost connections that connect each of the 54 targeted cities on a single network. This yields an all-China spanning tree network.

The final step to constructing the network depicted in Figure A.3 in the paper is to apply the least cost path algorithm to find least costly connections between capitals of border provinces and segments of China's border. Least costly paths to any segment of the border within the same compass quadrant (NE, SE, SW, NW) as NTHS routes were constructed without imposing *ex ante* restrictions on the end points located on the border.

To construct the straight line spanning tree network depicted in Figure 3 of the paper, the first step is to compute great circle distances between all possible 1431 bilateral connections of the network, which is done by applying the Haversine formula to bilateral coordinate pairs. I then compute Kruskal's algorithm to identify the minimum number of edges that connect all targeted cities subject to the minimum aggregate distance impedance on the network. To account for the fact that Chinese planners construct many more than the minimum spanning tree connections, I re-run Kruskal's algorithm after dividing China into North-Center-South, as well subject to East-Center-West geographical areas.[17] These two additional estimations add 9 bilateral routes in addition to the 53 connections that resulted from the all-China estimation. The final step is to include minimum great circle distance connections from provincial capitals of border provinces to the nearest border segment within the same compass quadrants as NTHS routes.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, *113*(2), 231–263.

Angrist, J., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton Univ Pr.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G., & Robert-Nicoud, F. (2003). *Economic geography and public policy.* Princeton University Press.

Baum-Snow, N., Brandt, L., Henderson, J., Turner, M., & Zhang, Q. (2012). Roads, railroads and decentralization of Chinese cities. *mimeo, Brown University*.

Combes, P., Mayer, T., & Thisse, J. (2008). *Economic geography: The integration of regions and nations.* Princeton University Press.

[17] I define these geographical areas on the basis of six geographic regions with official administrative recognition in China: East, North, North-East, North-West, South-Central, and South-West.

Douglas, D. (1994). Least-cost path in gis using an accumulated cost surface and slopelines. *Cartographica The International Journal for Geographic Information and Geovisualization*, *31*(3), 37–51.

Harris, C. D. (1954). The, market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers*, *44*(4), 315–348.

Jha, M., McCall, C., & Schonfeld, P. (2001). Using GIS, genetic algorithms, and visualization in highway development. *Computer-Aided Civil and Infrastructure Engineering*, *16*(6), 399–414.

Jong, J., & Schonfeld, P. (2003). An evolutionary model for simultaneously optimizing three-dimensional highway alignments. *Transportation Research Part B: Methodological*, *37*(2), 107–128.

Lin, S. (2009). The rise and fall of China's government revenue.

Martin, P., & Rogers, C. A. (1995). Industrial location and public infrastructure. *Journal of International Economics*, *39*(3-4), 335-351.

Qiao, M.-V. J., B., & Xu, Y. (2008). The tradeoff between growth and equity in decentralization policy: China's experience. *Journal of Development Economics*, *86*(1), 112–128.

Wong, C. (2000). Central-local relations revisited the 1994 tax-sharing reform and public expenditure management in China. *China Perspectives*, 52–63.

WSDT. (2002). *Highway construction cost comparison survey final report.* Washington State Department of Transport.