MARKET FOR CONGESTION: VCG TRAFFIC TOLLS IN BEIJING

RICHARD XU, BONNIE LIU

Contents

1. Introduction

With increasing urbanization and car ownership ensuing economic prosperity, traffic congestion is becoming a pervasive problem worldwide with dire economic, societal, and environmental consequences. Global congestion costs are estimated at about 1 trillion USD. Congestion precipitates wasted time, fuel, and carbon dioxide emissions. The Beijing government has attempted myriad solutions, including a license plate lottery (which caused long queues due to supply shortages) and rotational driving permission (cars with license plates ending in certain specified digits are prohibited from entering the roads on a specified day of the week). New York's taxi medallion

Date: November 5, 2020.

cap and trade system has similarly caused soaring prices and distributive inequality. Inadequacy of non-market solutions demands a reconsideration of implementing traffic tolls to counter the market distortion.

In this paper, we seek to model a market for congestion and propose a VCG-resemblant mechanism to optimally price highway traffic. Currently, Beijing government levies tolls on ring roads that is either a flat rate or proportional to the distance traveled. However, this pricing mechanism fails to dynamically account for real-time congestion as it is time-invariant. Alternatively, the government should consider levying tolls based on travel time instead of distance, which could further deter traffic during peak congestion times. The mechanism is premised upon the rationale that drivers should be charged based on the slow-down they cause others. In essence, the VCG mechanism requires drivers to internalize the negative externality it imposes on other drivers, and embodies a Pigouvian philosophy that deters socially harmful behavior by imposing a penalty that is the pecuniary equivalent of the inconvenience inflicted upon other market participants. In particular, marginal cost pricing, which charges each network user for the congestion effects caused by its presence, eliminates inefficiencies of selfish routing. The VCG nature predicts that it would charge higher prices during rush hours / peak traffic times, and that it would charge industrial trucks more than smaller vehicles, as trucks takes up more space and slows down traffic more.

In particular, Beijing's dilemma distinguishes itself from that of other cities in its unique urban design and roads infrastructure, i.e. its set of concentric ring roads around the city center connected by radial roads. By construction, the ring roads allow drivers to circumvent business districts and neighborhoods. Hence, within practical routing constraints, the reward/punishment must be commensurate to sufficiently motivate socially optimal behavior and align independent optimization with collective social optimization.

2. Current Research

In a traffic network, the latency of a path is the time required to travel from one end to the other. Given graph $G = (V, E)$, with $n = |V|$, the latency of edge $e \in E$ is a function $l_e : R^+ \cup \{0\} \to R^+ \cup \{0\}$. In the congestion model, latency $l_e(f_e)$ is a nonnegative and nondecreasing function of the flow f_e on the edge. In quotidian terms, the more traffic on the road, the longer time it takes for each driver to reach their destination. A traffic pattern that minimizes average latency is a system optimal flow. On the collective level, a system optimal flow f minimizes $\sum_{e \in E} l_e(f_e) f_e$, which is the sum of the product of the latency times flow for all edges on the network. On the individual driver level, a self-oriented driver traveling from s to t chooses a path P that minimizes the latency experienced: given that all other network traffic is fixed as f, the traveler minimizes $\sum_{e \in P} l_e(f_e)$. Each user engaging in such optimization results in the Nash flow, which is a Nash equilibrium for the routing game (where action space is set of all $s - t$ paths), as each player individually minimizes latency given decisions of other users, as reflected in the flow.

Traffic tolls have been traditionally established in the literature as an effective method to motivate homogeneous users to choose paths for system optimal flow while selfishly minimizing individual latency and toll. Under the marginal cost framework, the optimal toll is equivalent to the marginal increase in aggregate latency caused by the increased flow from this additional driver, yielding $\tau_e = l'_e(f_e) f_e$. In the more complicated and realistic scenario of heterogeneous population where each driver has an idiosyncratic preference and valuation of time, scholars have considered a multiplier $\alpha(a)$ that represents agent a's valuation of time. In this framework, user a seeks a path P that minimizes the "cost function," $\sum_{e \in P} \alpha(a) l_e(f_e) + \tau_e$. This cost function aggregates additively the time cost (first term) and the monetary cost (second term); thus, $\alpha(a)$ can be construed of as the individual weight of time in relation to money. This multiplier is likely to be higher for higher income populations (e.g. white collar workers) than lower income populations, as the former experiences higher opportunity cost of time.

Unsurprisingly, the Nash flow is drastically different from a system optimal flow, which is the precisely cause of the congestion externality. Therefore, much of the literature explores how to induce selfish users of the network to follow a traffic pattern that minimizes the average latency experienced by the users. Fleischer et. al (2004) proved the existence of tolls to induce multicommodity, heterogeneous network users that independently choose routes minimizing their own linear function of tolls versus latency to produce minimum average latency. Fleischer (2005) further showed that linear tolls are necessary and sufficient for such purposes: the optimal toll on each edge needs to be no more than the latency of the maximum latency path in the minimum average latency flow times the maximum valuation of time on that edge. Fleischer presents that the total toll paid by a user (across all edges in a path) is bounded by the same quantity, which is always less than $\alpha_{max} l_{max} n$. The authors show that in series parallel graphs, the set of paths traveled in Nash flow with optimal tolls is independent of the valuations of time of the users, and is determined by $\tilde{f}$ only.

Fleischer models the set of users as the set [0, 1] ordered by their increasing valuation of time. Given congestion-aversion function $\alpha : [0, 1] \to R^+$ and toll vector $\tau : E \to \cup\{0\}$, denote the Nash flow by f_α^τ. The authors prove that linear tolls bounded by $1 + \alpha(1)L$ suffice to induce a minimum latency flow as a Nash flow, where l_{max} is the maximum latency of an edge in $\tilde{f}$, and $L = \max_{p \in P} \sum_{e \in P} l_e(1)$. Furthermore, this linear bound also holds in the multiple source, single sink setting, and to induce any given acyclic flow, in addition to the system optimal flow. Furthermore, this theorem can be strengthened by bounding tolls separately on each edge: replacing T with $T_e := \min\{\alpha(a) | e \in \gamma(a)$ in Nash flow of optimal flows$\}$ in map γ for edge e. This suggests that the toll for agent a is bounded by the latency of maximum latency path times their own valuation of time. This theorem additionally proves the existence of tolls to induce any pre-specified acyclic flow. Fleischer (2005) additionally bounds total toll in the following theorem: Let τ be a set of tolls that induce the system optimal flow, then there exists tolls τ' that induce the system optimal flow such that user a pays at most $\alpha(a)L$.

Meanwhile, Cole et. al (2006) explore limitations and potential downsides of marginal cost pricing. The principle of marginal cost pricing assumes that taxes cause no disutility to network users, which is warranted only when collected taxes are feasibly returned, directly or indirectly, to the users. When the assumption fails to hold, while marginal cost pricing can minimize latency and increase cost, it also causes disutility. The authors reach the following conclusions: A) In every network with linear latency functions, marginal cost taxes do not improve the cost of a flow at Nash equilibrium. B) The maximum-possible benefit of taxes is no more than that of edge removals, since levying a sufficiently large tax on an edge effectively removes it from the network C) For every network with linear latency functions—not merely worst-case examples—taxes cannot decrease the cost of a flow at Nash equilibrium beyond what can be achieved by removing edges. Cole et. al (2003) show that edges of a single-commodity network can be priced so that optimal routing is Nash equilibrium, even for heterogeneity, and offer a mathematical condition for the function $\alpha 4$ such that disutility from the tax is small.

Cramton et. al (2019) describe a market for road use based on efficient scheduling, routing, and pricing. Under the design, road use is priced dynamically by marginal demand during constrained times and locations. In unconstrained times and locations, a nominal fee is paid for road use to recover costs, as in other utilities. Transport is scheduled based on forward prices and then routed in real time based on real-time road-use prices.

3. Practical Analysis

3.1. **Obtaining the Data.** We obtain traffic data using the Baidu Maps API. This is the only source of Beijing map data we were able to find. The API can be accessed using HTTP request, as explained in the following sequence diagram.

Beijing is structured as multiple concentric "rings" connected by major avenues in different spatial directions. We omdel the Beijing as concentric squares and find the name of each of the major highways. The names are stored in the file `road.txt`.

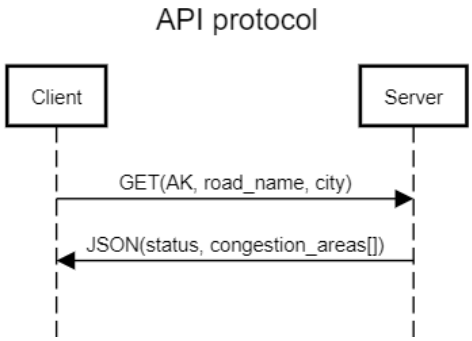


FIGURE 1. Sequence diagram for API request

After obtaining an API key from Baidu, the program `get_traffic_data.py` queries the traffic data. The script is run every 30 minutes using an AWS EC2 instance using the command

```
nohup python get_traffic_data.py >> log.out
```

After that is cleaning the data. The we receive the data in a dictionary, and sometimes there is no traffic data. That can be one of two cases:

- There is indeed no congestion,
- The service gathering traffic data is down.

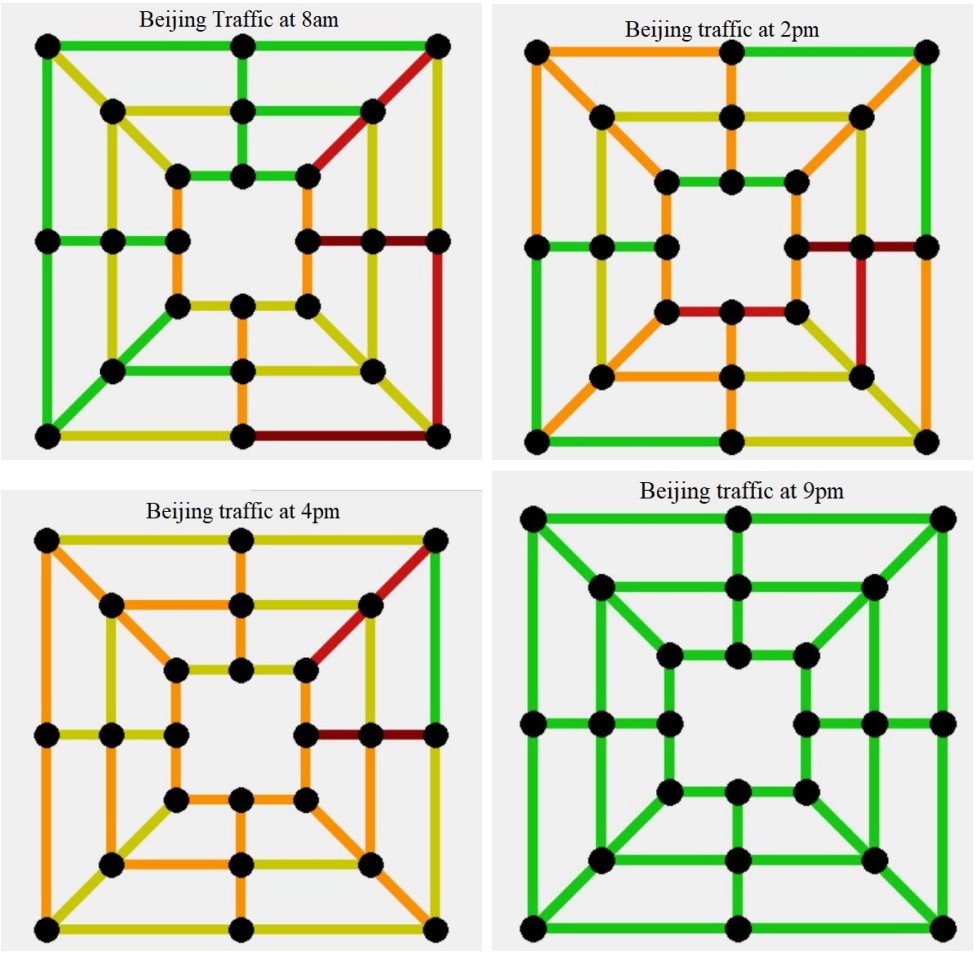
It is fairly easy to notice which is which, since in the second case almost all the roads are labeled as no data. We proceed to clean the data.

3.2. **Visualizing the Data.** We visualize Bejiing as three concentric squares, representing the 2nd, 3rd and 4th ring of the city where most of the traffic is. We then use the `graphics` package to first plot the vertices and edges representing where the roads and the points where the radial avenues reach the rings. For example, the road "Cai Hu Ying South", romanized, is the road connecting the south-west vertices together. From there, we parse the traffic data, found in the parser in file `draw.py`, and draw out the lines.

Due to time constraint, we captured all the pictures and self-labeled their time. The colors represent traffic speed in the following order:

- Green: No congestion
- Yellow: 20-30 km/h
- Orange: 10-20 km/h
- Red: 5-10 km/h
- Dark Red: 0-5 km/h

All the code for drawing these diagrams are stored in `draw.py`. Below are the visualized graphs.



We observe that

(1) Beijing traffic is congested during all business hours, in addition to the peak hours.

(2) Congestion is indeed a major issue in Beijing. This is a typical day, and for multiple roads at multiple times, the traffic is slower than 5 km/h.

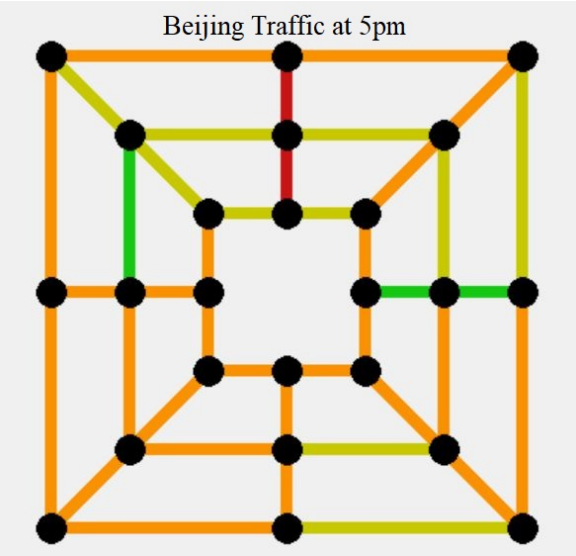


FIGURE 2. Beijing Congestion Level at Different Times of Day

(3) Surprisingly, the moment when traffic is most congested is not at 5pm. This is likely because the traffic is more managed then. This is corroborated by how the traffic is more uniform at 5pm while at 2pm there are roads with significant congestion and roads nearby that are green.

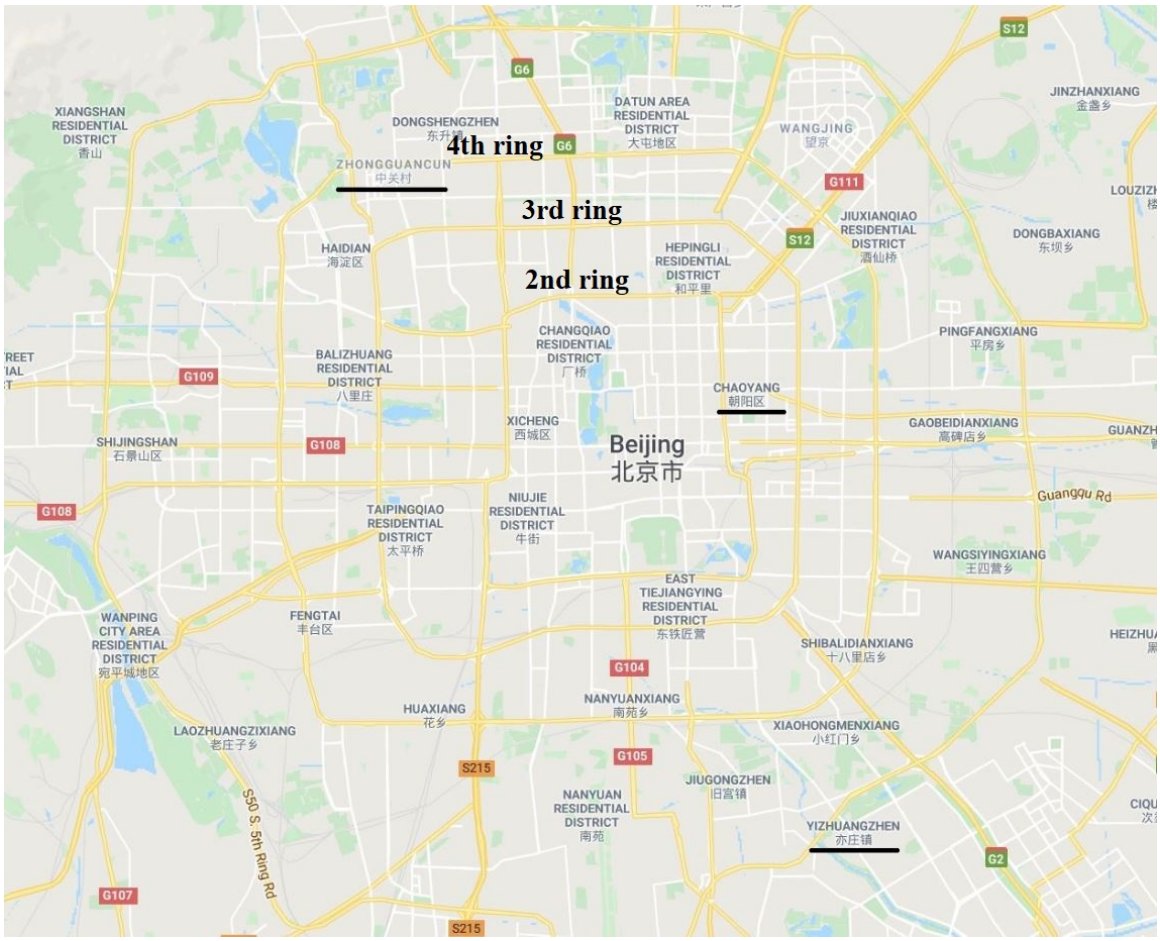
3.3. **Implementing the Toll.** Because of time constraint we did not have time to fully implement an automated version of the traffic, but we were able to calculate some of the tolls by hand. Here is the procedure we used.

- Given the traffic speed, we used the graphs presented in Hall '96 to estimate the traffic density of the roads. In their paper about price of anarchy on networks, Zhang et. al outlined a way to obtain demand of each road and flow capacity using the traffic speed, but we did not have time to implement the algorithm.
- Next, we used the slope of the graphs in Hall '96 to estimate the additional slowdown caused by a car on the road and price accordingly.

Surprisingly, the VCG-inspired toll on marginal slowdown was easier to calculate than the linear toll proposed by Fleischer et. al. The VCG-inspired toll is local, while

to find the optimal linear toll, we need to look at the whole network effect. Despite this difference, both run in polynomial time in the size of the graph, and the graph is fairly small. Therefore, speed should no longer be an issue once we automate the process. Some of the toll results can be found in the subsequent section.

3.4. **Case studies of Different Trips.** To better explain the effect of the toll on the population, we perform case studies on some model trips across the city. These trips are motivated by people Bonnie knew from the city. One challenge comes from estimating a person's willingness to pay for the highways, since it is unclear what type of alternatives they have. We use the subway as our baseline alternative. The subway costs 5 RMB (equivalent to 0.7 USD) per ride and is quite reliable in Beijing, so it is a reasonable alternative.



- Alice is a researcher at the Chinese Academy of Sciences who lives in the Chaoyang district. She is traveling from her house to the university. Using the subway takes 55 minutes. Driving takes 29 minutes without traffic but 40 minutes during peak traffic, and the proposed tolls were 0 RMB and 18.9 RMB, respectively. Being a typical graduate student, Alice's willingness to pay for the drive is likely less than 80 RMB/h, and would take the subway during peak traffic.
- Bella is a programmer working in Zhong Guan Cun who lives across the city in the East Tiejiangying residential district. Using the subway takes 96 minutes. Driving takes 34 minutes without traffic and 65 minutes during peak traffic, and the tolls were 0 RMB and 28.3 RMB. The toll is relatively low because the 3rd ring is not particularly crowded at peak hour (mostly yellow lines). Bella will likely pay to drive, because her willingness to pay for her time may be more than 54.8 RMB/h.

We see that the toll is able to deter some individuals from driving and opting to take the subway. One major drawback this policy is that it can increase inequality, since poor people will have a higher value for money and lower willingness to pay for the roads. We are continuing to think about how to remedy this issue.

4. Policy Discussions

The congestion toll proposed in this paper interacts with broader policy questions. Alternatives include government subsidies for public transportation; however, given that Beijing currently expansive and subsidized subway and bus systems, effects of additional subsidies may not be salient enough to motivate the switch from private vehicles to public transportation. Congestion due to high urban density may also be alleviated by the government's plan to expand the city outward to Xiong'An New Economic Zone. Mandating public and private institutions like government bureaus, universities, and industries to relocate their campus to the outskirts of the city, and

providing housing subsidies in suburban areas are Beijing's bandaid solutions to its exacerbating urbanization.

Policy-makers must additionally consider potential regressive distributional effects. Since the toll is indiscriminately imposed at uniform rates regardless of wealth, the lower/medium income populations may experience disproportionate adverse effects compared to the affluent. With data from the 2010 Beijing Household Travel Survey conducted by the Beijing Transportation Research Center (BTRC), an agency of the Beijing municipal government, Linn, Wang, and Xie (2016) consider the congestion pricing scheme currently under consideration in Beijing, which charges highway traffic within third ring road 8 RMB per entrance, and which exempts buses and provides a 90 percent discount for residents within the third ring. They find that a very small proportion of motorized trips would be subject to the full congestion charge. The directly affected individuals tend to have higher household incomes and are wealthier than individuals not directly affected, reflecting the fact that individuals who drive to work in Beijing are relatively wealthy. The Suits index for the congestion charge is 0.027, indicating that the congestion charge is slightly progressive.

5. Conclusion and Future Directions

We found that the toll can be an effective way to reduce congestion in the city, and reviewed previous literature about the efficacy of different types of tolls on the overall flow of a network. Then, we dived deep in to the toll systems and found different toll implementations to solve the issue. Next, we put theory into practice by constructing a data pipeline that gathers traffic information, visualizing those information and explaining how the toll system would be calculated when implemented in real life. Finally, we examine the distributional and related effects of this policy.

For future directions, we aim to continue the code and create more detailed figures for the city. Unlike major cities in the US, the literature surrounding Beijing is relatively limited and we struggled to find relevant plots about, for example income distribution and traffic situation, in the literature. Having a framework to gather

data automatically can be extended to create these plots, which can be useful for other research. Next, we hope to finish writing the automated toll system started in this paper. The system shows promise but the implementation of linear tolls as described in the papers has been a challenge.

Overall, it has been an enriching project to do that led us to multiple interesting places, and I believe this can be a fascinating area for further research.

All code used for this project can be found in the accompanied .zip file.

6. References

F. L. Hall. Traffic stream characteristics. In N.H. Gartner, C.J. Messer, and A.K. Rathi, editors, Traffic Flow Theory. US Federal Highway Administration, 1996.

R. Cole, Y. Dodis, and T. Roughgarden. How much can taxes help selfish routing? Journal of Computer and System Sciences, 72(3):444–467, 2006.

L. Fleischer, K. Jain, and M. Mahdian. Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion

R. Cole, Y. Dodis, and T. Roughgarden. Pricing network edges for heterogeneous selfish users. In Proc. 35th Symp. on Theory of Computing, pages 521–530, 2003.

T. Harks, G. Schafer, and M. Sieg. Computing flow-inducing network tolls. Technical Report 36-2008, Institut fur Mathematik, Technische Universit at Berlin, Germany, 2008.

L. Fleischer. Linear tolls suffice: New bounds and algorithms for tolls in single source networks. Theoretical Computer Science, 348(2-3):217–225, 2005.