Analysis and Modeling of New York City Taxis

Dr. Chenguang Zhu
February 9, 2016
Overview

• Data of all trips made by yellow taxis in 2013
  • Vehicle and driver information
  • Pickup and drop-off time / location
  • Fare, tip, surcharge
  • Distance

• Public dataset released by New York City Taxi & Limousine Commission
Big Data System

• Cloud-based system for transportation data

• Input data:
  • GPS
  • App
  • Smart transit card

• Functions:
  • Visualization
  • Analysis
  • Query
  • Emulation

• Demo
Data statistics

• 173,185,091 trips (~474K trips per day)

• 43,191 drivers, 14,144 vehicles

• Total fare: $2,138,595,152

• Total distance: 500,558,422 miles (2.6 round trips from Earth to Sun)
Destination of rides

Weekday mornings (6:30AM ~ 9:30AM)
Destination of rides

Weekday evenings (5:30PM ~ 8:30PM)
Destination of rides

Weekends (9:00AM ~ 6:00PM)
Periodicity

• Regularity in taxi system
  • Weekly
  • Similarity among weekdays and weekends

• Observed at different times in different locations

• Reflected in various facets (number of trips, distance, etc.)
Pickups in Midtown Manhattan

Number of pickups

Hour

Memorial Day

Memorial Weekend

2013-05-06 to 2013-05-31

2013-06-01 to 2013-06-30
Distance of trips in Midtown Manhattan

![Graph showing the distance of trips in Midtown Manhattan for various dates between May 5, 2013, and June 30, 2013.](chart)
Tip ratio in Midtown Manhattan
Pickups in Financial District

![Graph showing number of pickups over time for various dates in 2013.]
Pickups in Brooklyn

[Graph showing the number of pickups over time for each day from September 2nd to October 27th, 2013.]
Periodicity

- Periodicity is inherent in taxi systems
- Appears in different locations, times, and aspects

Implications
- Recognize general patterns from a typical week
- Predict demand with higher accuracy
- Quickly spot irregularity → weather conditions, special events
Impact of severe weather

• New York City is in the northeast of US

• Exposed to severe natural phenomena: blizzards, storms, ...

• Transportation system impacted by harsh weather

• Quantify the impact
2013 February Nor’easter

• Feb. 8 ~ 9, 2013

• Heavy snowfall and hurricane-force winds

• 11.4 inches (29cm) of snow at Central Park
2013 February Nor’easter

Normal Friday nights in Jan.-Mar. 2013

Night on Feb. 8, 2013
Number of taxis

- Nor'easter on Feb. 8 and 9
- Precipitation during Nor'easter

Nor'easter on Feb. 8 and 9:
- 7 hours
- 15 hours
- 12 hours
Impact of Socio-cultural Factors: New York Knicks

- Famous basketball team in NBA
- Home stadium located at Madison Square Garden
- Game results and tipping behaviors
<table>
<thead>
<tr>
<th>Date &amp; Start Time</th>
<th>Opponent</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/03/2013 7:30PM</td>
<td>San Antonio Spurs</td>
<td>Knicks won</td>
</tr>
<tr>
<td>01/07/2013 7:30PM</td>
<td>Boston Celtics</td>
<td>Knicks lost</td>
</tr>
<tr>
<td>01/11/2013 8:00PM</td>
<td>Chicago Bulls</td>
<td>Knicks lost</td>
</tr>
<tr>
<td>01/30/2013 7:30PM</td>
<td>Orlando Magic</td>
<td>Knicks won</td>
</tr>
</tbody>
</table>
Post-game trips

• Compare with trips on other weekdays in January
  • Started from 10:30PM to midnight
  • Originated from within 2 blocks of Madison Square Garden

Tip ratio = \( \frac{\text{Tips}}{\text{Total payment} - \text{tips}} \)

<table>
<thead>
<tr>
<th>Game result</th>
<th>Average tip ratio per trip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knicks won</td>
<td>19.0%</td>
</tr>
<tr>
<td>No game</td>
<td>18.3%</td>
</tr>
<tr>
<td>Knicks lost</td>
<td>18.1%</td>
</tr>
</tbody>
</table>

Unpaired t-test between win and lose: \( p\text{-value}=0.01224 \)
Summary

• Analysis on New York City taxi data

• Periodicity

• Weather impact

• Relationship between socio-cultural factors and taxi trips

• Next: model taxi trips to enhance systematic efficiency
Modeling Taxi Trip Assignment
Demand-supply system
Goal

• Assume start and end location/time of each trip are fixed

• Reassign *all* trips to possibly *fewer* taxis

• Reduce idle time
Our model

• Based on network flow

• Spatial-temporal network

• Vehicles $\rightarrow$ flow    Trips $\rightarrow$ edges with capacity limit
Discretization

- **Time discretization**
  - 3:00PM
  - 3:01PM
  - 3:02PM
  - 3:03PM

- **Location discretization**
  - 36 regions
  - District boundary, main roads
Spatial-temporal representation

• Construct a node for each discretized time & location

• Node $p_{t,l}$: the taxi is in region $l$ at time $t$

<table>
<thead>
<tr>
<th>Time</th>
<th>3:00PM</th>
<th>3:01PM</th>
<th>3:02PM</th>
<th>3:03PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region A</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Region B</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Region C</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
1. An idle taxi can stay in the same region for the next minute
   - Construct edge $p_{t,l} \to p_{t+1,l}$
2. An idle taxi can move to another region $l'$
   • Get the duration $t_{l \rightarrow l'}$ from data
   • Construct edge $p_{t,l} \rightarrow p_{t+t_{l \rightarrow l'}, l'}$
3. A taxi takes a trip
   • Start: region $l$ at $t$; end: region $l'$ at $t'$
   • Construct edge $p_{t,l} \rightarrow p_{t',l'}$
Capacity limit

- Consider each taxi as a flowing particle
- Upper-limit and lower-limit of capacity for each edge
Capacity limit

• Edges for idle taxis have lower-limit 0 and upper-limit $\infty$

![Diagram showing capacity limits forRegion A, Region B, and Region C at 3:00PM, 3:01PM, 3:02PM, and 3:03PM.](image-url)
Capacity limit

- Each trip is assigned to exactly one taxi
- Trip edge has both upper-limit and lower-limit the same as the number of associated trips

```
Region A
```

```
Region B
```

```
Region C
```

```
3:00PM 3:01PM 3:02PM 3:03PM
```

```
0,∞ 0,∞ 0,∞ 0,∞
```

```
0,∞ 0,∞ 0,∞ 0,∞
```

```
1,1 0,∞ 0,∞ 0,∞
```

```
0,∞ 0,∞ 0,∞ 0,∞
```
Source and sink

- Construct source $S$ and sink $T$
- Connect $S$ to trip-starting nodes and trip-ending nodes to $T$
- Each taxi hypothetically starts from $S$ and ends at $T$
Connection to network flow

One working taxi $\Leftrightarrow$ One unit flow from $S$ to $T$

Feasible flow $\Leftrightarrow$ All trip requests are satisfied

Minimize the number of required taxis $\Leftrightarrow$ Minimize feasible flow

Push-relabel Algorithm $O(|V|^2 |E|)$
Idle time as cost

- For all idle edges, associate the length of idle time as cost
- All other edges have cost 0
- Minimum cost minimum feasible flow

Orlin et. al. $O(|E| \log |V|(|E| + |V| \log |V|))$
Assign trips

- After flow values are determined, employ breadth-first-search (BFS) on edges with flow from $S$ to $T$

- Each path determines the actions of a new taxi
Experiment

• We evaluated our model on taxi trips in New York City on May 15, 2013 (Wed.)

• 12-hour shift from 4AM to 4PM

• 214,805 trips by 12,366 taxis
Implementation

• We built the flow network: 51,842 nodes, 624,067 edges

• Ran on a single machine with 2.3 GHz Intel Core i7 and a memory of 16GB 1600 MHz DDR3

• Minimum feasible flow algorithm (FF): 28.334s

• Minimum cost minimum feasible flow algorithm (MCFF): 259.753s
## Result

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Real data</th>
<th>FF</th>
<th>MCFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of taxis</td>
<td>12,366</td>
<td>8,883</td>
<td>8,981</td>
</tr>
<tr>
<td>Number of trips per taxi</td>
<td>17.4</td>
<td>24.2</td>
<td>23.9</td>
</tr>
<tr>
<td>Idle time per taxi (hours)</td>
<td>4.1</td>
<td>3.5</td>
<td>2.8</td>
</tr>
<tr>
<td>Earnings per taxi ($)</td>
<td>252.6</td>
<td>351.6</td>
<td>347.8</td>
</tr>
</tbody>
</table>
Working time-line
Learn from the models

• Explore patterns of assignment given by models

• Important for:
  • **Taxi operators**: spot inefficiencies in current taxi system
  • **Individual driver**: arrange trips in a smarter way
Airport trips

- Airports are far from residential / business district
  - JFK is ~18 miles from Manhattan
Airport trips

- Airports are far from residential / business district
  - JFK is ~18 miles from Manhattan

- Airport trips are popular
  - 2.2% of all observed trips start from or end at JFK

- After dropping off a passenger at JFK, should the driver wait at the airport, or leave the airport without passengers?
Airport trips

• In real data, **49.2%** of taxis sending passengers to JFK come back with new passengers from the airport

• **67.0%** for FF, **64.0%** for MCFF

• Takeaway message: wait for your next passenger at JFK
Conclusion

Big data for taxi systems

Analysis

• Periodicity
• Impact of weather conditions
• Impact of socio-cultural factors

Modeling

• Trip reassignment
• Network flow model
• Scalable
• Enhance systematic efficiency
Thank you!