A long-exposure photograph of a highway at night, showing vibrant light trails from cars in various colors (blue, white, red, yellow, green) against a dark background. The trails are blurred and radiate from the top center towards the bottom corners.

# Statistical inference of probabilistic O-D demand using day-to-day traffic data

Sean Qian  
Assistant Professor, CEE & Heinz, CMU  
Northwestern, May. 31, 2018

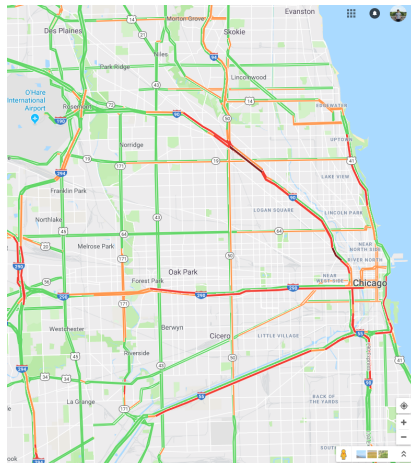
# Outline

- 1 Massive data: opportunities and challenges
- 2 Statistical Origin-Destination Demand Estimation
- 3 Mobility Data Analytics Center (big MAC)



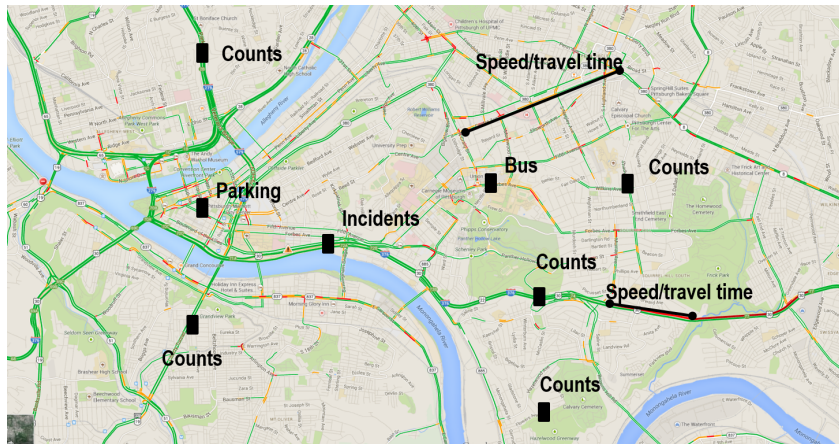
# Smart decision making?

- Incident management
- Infrastructure retrofit
- Ride-sourcing impact/regulation
- Parking pricing
- ...



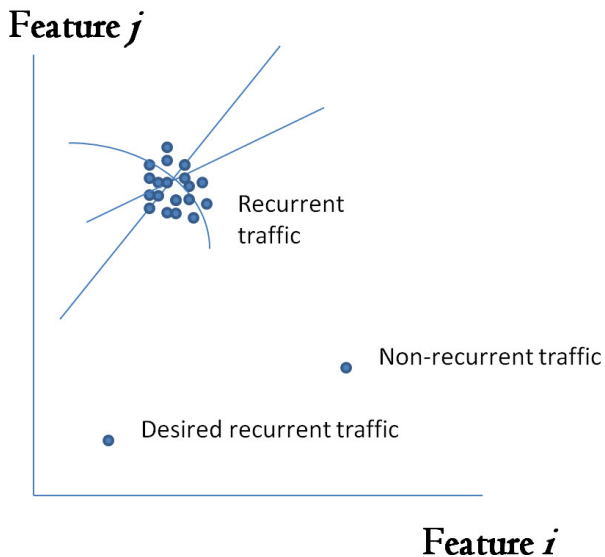
# Massive data: useful but challenging

Fusion. Bias. Sparsity. Computation. Unexplored space.

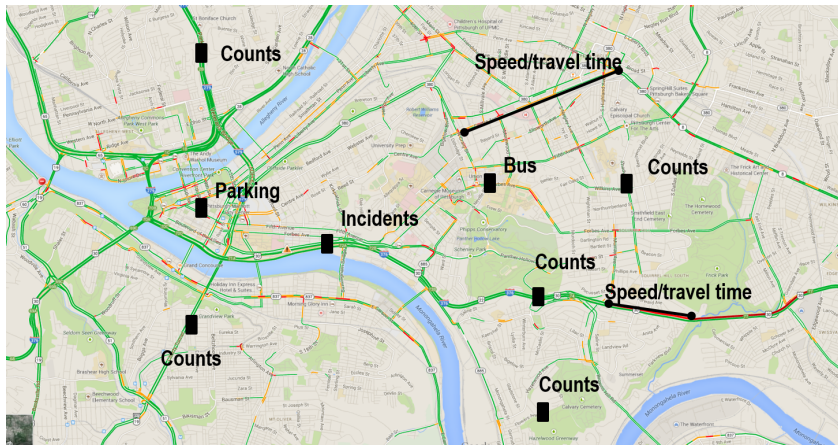




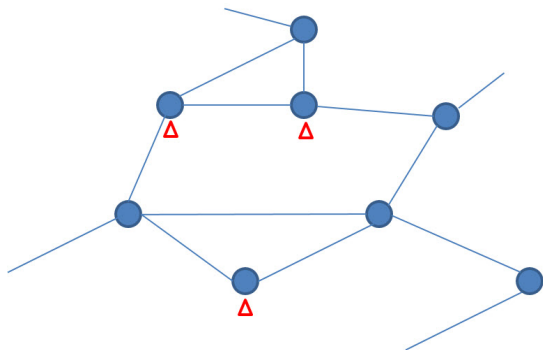
# Unexplored space



# A possible solution: data + physics



# A generic infrastructure network



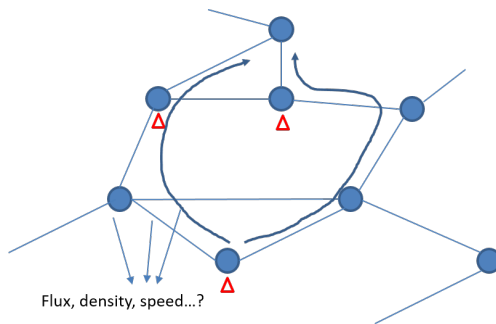
● An infrastructure node

Users: human beings

Goods: water, energy, vehicle...

△ A sensor

# A generic infrastructure network



Final goals: evaluation and intervention

- 1 Sensing** in sampled locations/time
- 2 Infer** features of users, goods and infrastructure
- 3 Predict** spatio-temporal distributions and system performance
- 4 Make decisions:** manage supply and demand

# Sensing-Learning-Managing

## 1 Sensing

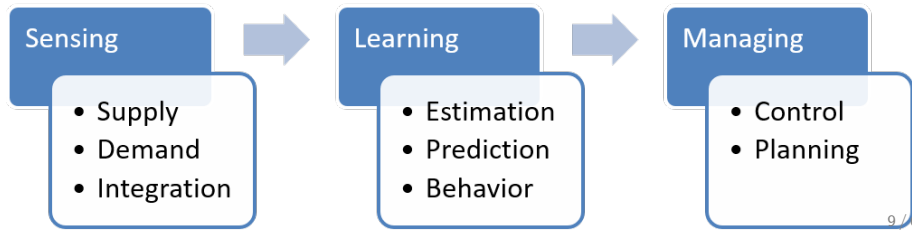
- Supply: network features, planned and unplanned incidents, weather, etc.
- Passengers and vehicles: roadway, parking, transit, bikes, pedestrian, etc.

## 2 Learning

- Behavior: choices of time, routes, modes and parking
- Data mining: best estimation and prediction

## 3 Decision making

- Short-term control
- long-term planning

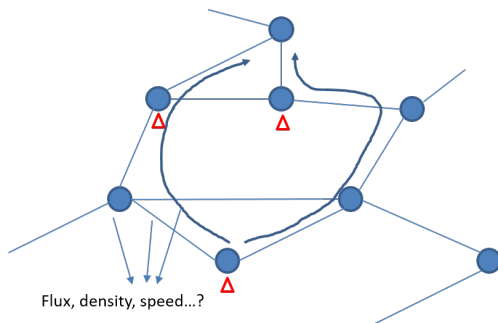


# Sustainable mobility

- Minimal congestion
- Resilient
- Safe
- Environmentally friendly



# Concepts...



Final goals: evaluation and intervention

- $x$ : link flow (flux, density, speed...)
- $f$ : path flow (flux, density, speed...)
- $c$ : system states (cost, time, emissions...)

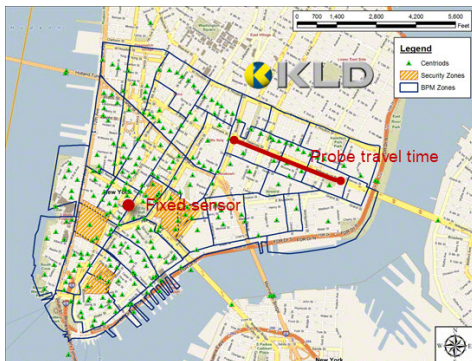
Given  $x^o, f^o, c^o$  and supply, learn  $(x, f, c) = G(\text{supply, demand})$

# ODE: Behavioral model $G$

- Use OD demand  $q$  to approximate demand
- Define user behavior  $G$

$$G : (\text{supply}; q) \mapsto (x, f, c)$$

- Given  $x^o, f^o, c^o$  and supply, estimate  $q$
- Calibrate  $G$ , estimate/predict  $(x, f, c)$





# Basic Notations

## Supply:

- Transportation network  $N$
- $A$  links, finite flow capacity  $C_a$  of link  $a$
- $K$  routes, a route  $k$  contains different set of links

## Demand:

- O-D: origin destination demand  $q_{rs}$ , indicating the number of travelers from  $r$  to  $s$ .
- Associate an OD with multiple routes, flow rate  $f_{rs}^k$
- Behavior: route choice

## Observation:

- Link flow counts ( $x^o$ )
- Link travel time ( $c^o$ )

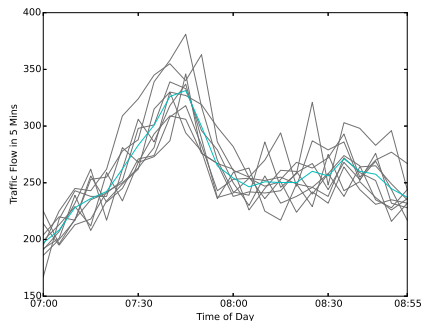
# Traffic Assignment

## Traditional Model:

- $TA : (N; q) \mapsto (x, f, c)$

## Challenge:

- Data Variation
  - Variance-covariance of observed data
  - Variance-covariance of  $(x, f, c)$



# Statistical Traffic Assignment

- Make the best use of data: mean and variance
- $(x, f, q) \rightarrow (X, F, Q)$
- Statistical equilibrium: a new behavioral model

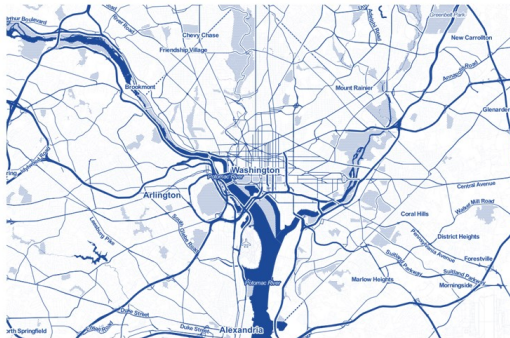
# Generalized Statistical Traffic Assignment (GESTA)

First, we work on  $G$

$$G : (N; Q) \mapsto (X, F, C)$$

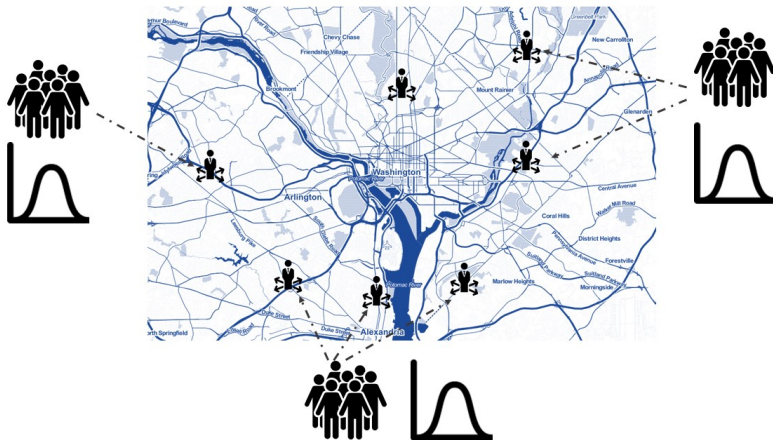
# Generalized Statistical Traffic Assignment (GESTA)

Probabilistic traffic demand:  $Q \sim \mathcal{N}(q, \Sigma_q)$



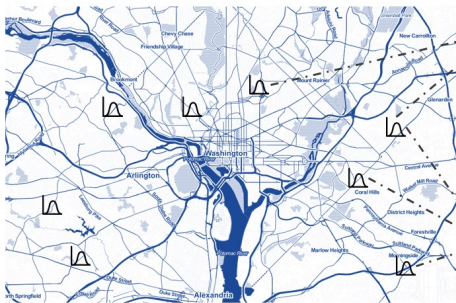
## GESTA - cont.

Stochastic routing:  $F \sim MN(\tilde{p}Q, \Sigma_f)$

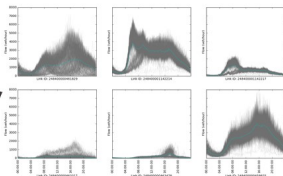


## GESTA - cont.

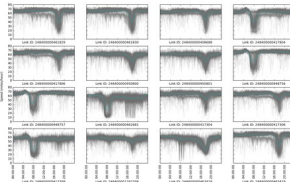
Sensing:  $X_m = X + \epsilon_e$



Overview of DC network



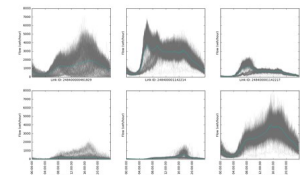
Link flow counts



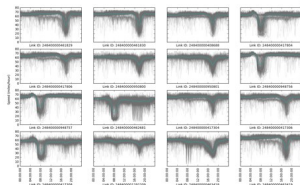
Link flow speeds

## GESTA - cont.

System states:  $C = C(X, F)$



Link flow counts



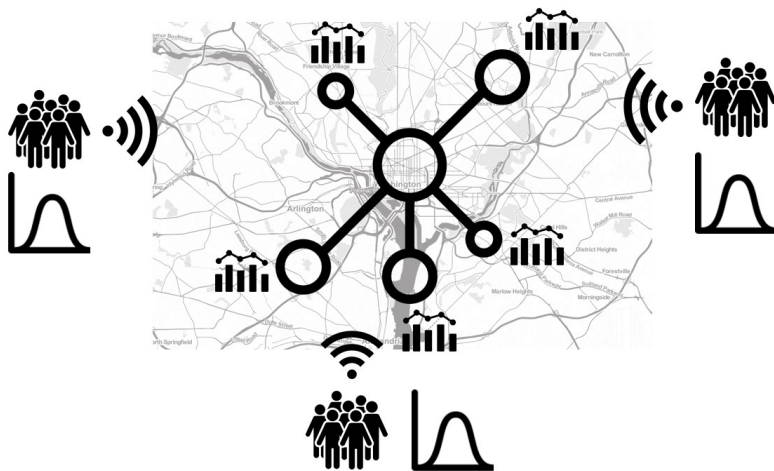
Link flow speeds



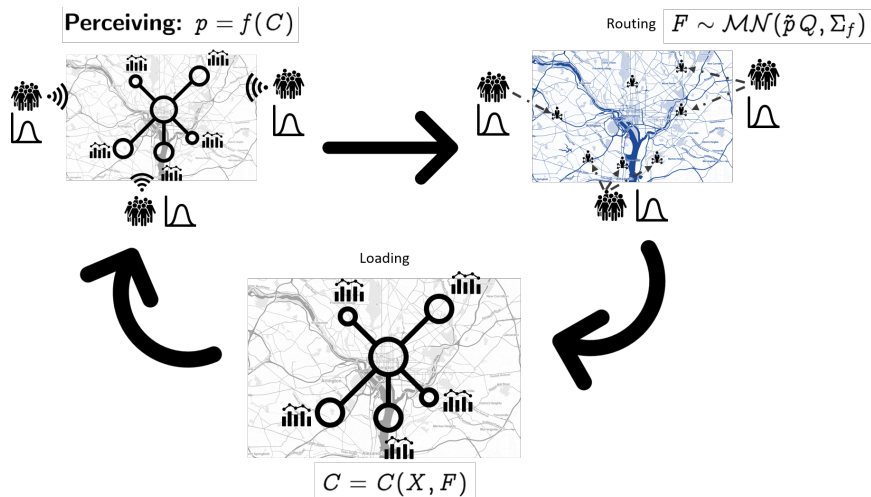


## GESTA - cont.

Perception:  $p = f(C)$



## GESTA - cont.



## GESTA - cont.

*Level 1 :*       $X_m = X + \epsilon_e$       (Unknown Error)

$$\epsilon_e \sim \mathcal{N}(\mathbf{0}, \Sigma_e)$$

*Level 2 :*       $X = \Delta F$

$F \sim \mathcal{MN}(\tilde{p} Q, \Sigma_f)$       (Route choice variation)

*Level 3 :*       $Q \sim \mathcal{N}(q, \Sigma_q)$       (Demend variation)

## GESTA - cont.

### GESTA features:

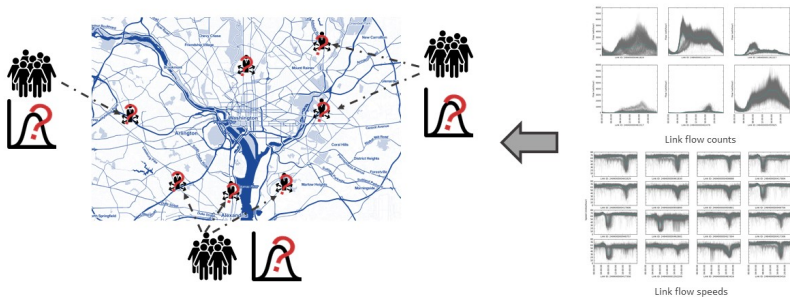
- Daily traffic condition is not in equilibrium
- Statistical equilibrium is built in a probability space
- Link/path flow variance = demand variance + choice variance

Wei Ma, Sean Qian. (2017) "On the Variance of Recurrent Traffic Flow for Statistical Traffic Assignment", Transportation Research Part C, Vol.81, pp.57-82.

# ODE: Learn GESTA

Now we know  $G : (N; Q) \mapsto (X, F, C)$

How can we learn  $X, F, C, Q$  from  $X^o, F^o, C^o$



# Review

## Deterministic O-D estimation problem

$$\min_q L(x^o, Aq) \quad (1)$$

where  $A$  is the assignment matrix,  $q$  is O-D demand, and  $x^o$  is the observed link flows.

### Estimation Methods:

- Entropy maximizing models
- Generalized least square
- Maximum likelihood estimator

# New challenges

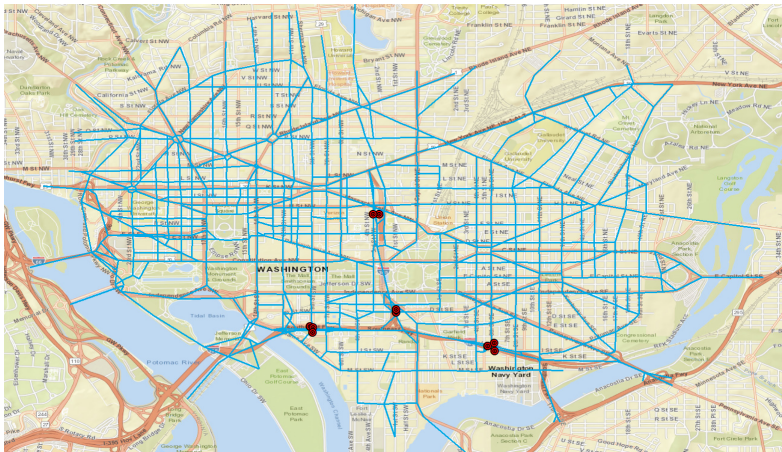
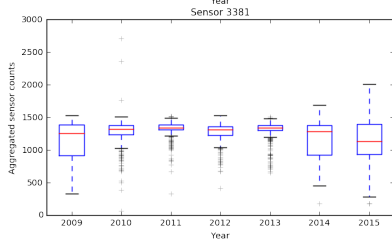
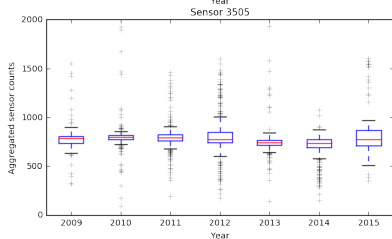
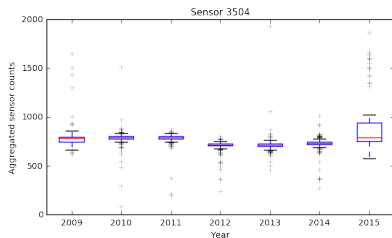
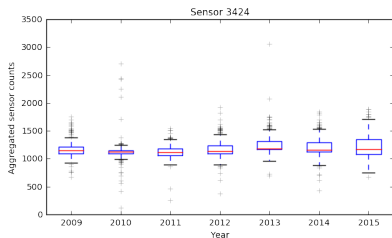


Figure: The Washington D.C. Downtown network

# New challenges - cont'd





# Challenges

- Data Variation
  - Multi-day data
  - Variance-covariance of  $Q, X, F, C$
- Scalability
- Observability

# Data Variation

## Idea:

- Estimate the probabilistic O-D demand

### Probabilistic O-D estimation problem

$$\min_{q, \Sigma_q} R(X^o, A Q) \quad (2)$$

where  $R$  is the risk function,  $A$  is the assignment matrix,  $Q$  is the random vector of O-D demand, and  $X^o$  is the random vector of observed link flows.

# Scalability and Observability

*Since the model gets more complicated:*

## **Scalability:**

- Is it still possible to scale to large networks?

# Scalability and Observability

*Since the model gets more complicated:*

## **Scalability:**

- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.

# Scalability and Observability

*Since the model gets more complicated:*

## **Scalability:**

- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.

## **Observability:**

- Can we still estimate OD using a small fraction of observations?

# Scalability and Observability

*Since the model gets more complicated:*

## **Scalability:**

- Is it still possible to scale to large networks?
- Solution: approximate high-dimensional probability distribution using data, instead of Bayesian inference.

## **Observability:**

- Can we still estimate OD using a small fraction of observations?
- Solution: sparsity analysis when highly underestimated

# Objective

## How to estimate:

- OD mean and cov:  $q, \Sigma_q$
- flow mean and cov:  $c, \Sigma_c, x, \Sigma_x, f, \Sigma_f$
- Route choice probability  $p$

## Such that GESTA:

- Best fits data collected over many years
- Scales easily
- Has fairly good observability

# IGLS Framework

- Iterative Generalized Least Square: EM like algorithm
- Separates the probabilistic OD estimation problem into two sub-problems:
  - Estimate OD mean vector
  - Estimate OD variance/covariance matrix
- Newton-Raphson step



# Estimate OD mean

## Traditional? with new statistical insights

$$\begin{aligned} \min_f \quad & n (\Delta^o f - \hat{x}^o)^T \left( \hat{\Sigma}_x^o \right)^{-1} (\Delta^o f - \hat{x}^o) \\ \text{s.t.} \quad & f \in \Phi^+ \end{aligned} \tag{3}$$

Where  $\Phi^+$  is the feasible set of  $f$ , such as Probit-based GESTA.

### Methods:

- Heuristic method, Yang (1995)
- Single level method, Shen & Wynter (2012)

# Estimate OD variance/covariance matrix

## Sparse penalization:

$$\begin{aligned}
 \min_{\Sigma_q} \quad & \|S_x^o - \Sigma_x^o\|_F^2 + \lambda \|\Sigma_q\|_1 \\
 \text{s.t.} \quad & \Sigma_x^o = \Delta^o \Sigma_{f|q} (\Delta^o)^T + \Delta^o \tilde{p} \Sigma_q \tilde{p}^T (\Delta^o)^T \\
 & \Sigma_q \succeq 0
 \end{aligned} \tag{4}$$

## Methods:

- Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Nesterov 2005)

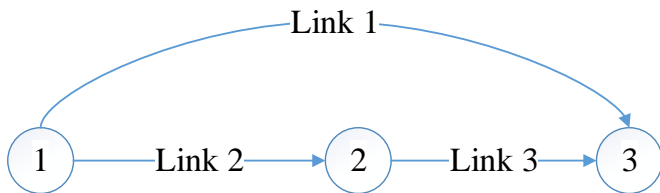
# Observerbility

**The statistical risk of the OD mean estimator under IGLS -  
the statistical risk of the deterministic OD**

**is bounded, and declines w.r.t. sample size**

Wei Ma, Sean Qian. (2018) "Statistical inference of probabilistic origin-destination demand using day-to-day traffic data", Transportation Research Part C, Vol.88, pp.227-256.

## A small example



- OD:  $1 \rightarrow 3$ ,  $2 \rightarrow 3$
- Observation: link 1 and link 3
- 500 days

# A small example - cont.

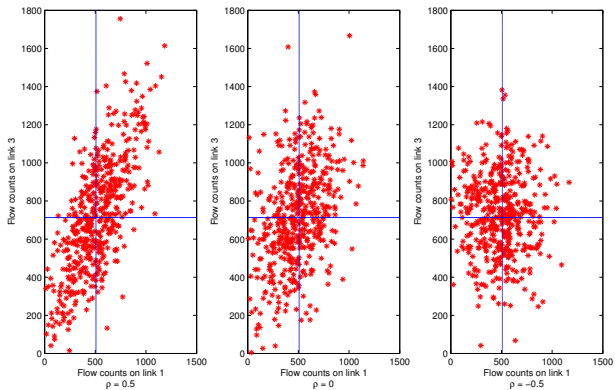


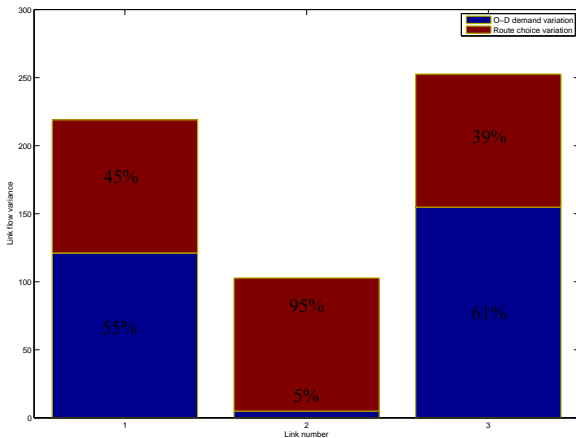
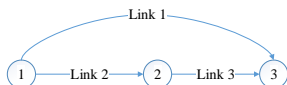
Figure: Synthesized “true” link flow data for different correlation  $\rho$

# A small example - cont.

**Table:** Results of probabilistic ODE on the three-link toy network (no historic O-D demand information is used)

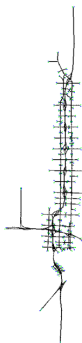
True $\rho$	Settings	$\hat{q}_{1 \rightarrow 3}$	$\hat{q}_{2 \rightarrow 3}$	$\hat{\sigma}_{1 \rightarrow 3}^2$	$\hat{\sigma}_{2 \rightarrow 3}^2$	$\hat{\rho}$	RMPSE	KL-distance
	True value	700	500	175	125	NA	NA	NA
0.5	w/o EC - w/o Lasso	722.17	500.41	186.69	134.21	0.56	3.62%	3.64
	Logit - w/o Lasso	682.36	499.63	207.94	134.21	0.50	2.08%	1.17
	Probit - w/o Lasso	699.50	499.63	200.94	134.21	0.52	0.07%	0.01
0	w/o EC - w/o Lasso	715.91	500.46	143.05	138.74	0.03	1.87%	0.74
	Logit - w/o Lasso	681.28	500.46	162.49	138.75	0.02	2.21%	1.01
	Probit - w/o Lasso	700.30	500.46	152.15	138.75	0.03	0.06%	0.01
	Logit - w/ Lasso	681.28	500.46	144.52	128.75	0.00	2.21%	1.01
	Probit - w/ Lasso	700.02	500.46	132.27	128.75	0.00	0.05%	0.004
-0.5	w/o EC - w/o Lasso	703.41	499.06	173.34	132.60	-0.41	0.43%	0.04
	Logit - w/o Lasso	681.05	499.06	184.13	132.60	-0.39	2.23%	1.47
	Probit - w/o Lasso	701.71	499.06	174.19	132.60	-0.41	0.23%	0.02

# A small example - cont.



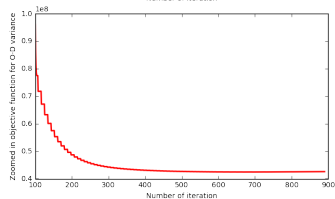
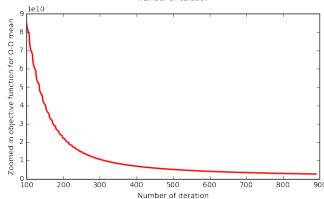
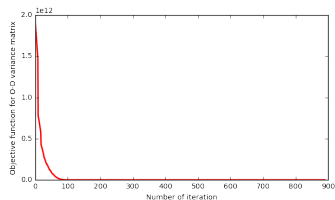
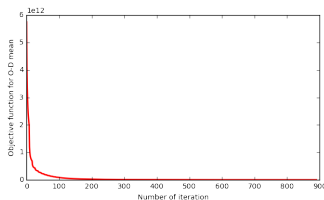
## SR-41 Corridor

- 2,413 links and 7,110 O-D pairs
- 10% of O-D pairs (randomly chosen) are mutually correlated with a correlation randomly drawn from 0 to 0.5
- Randomly choose 50% of the links on the network to be observed for 1,000 days





## SR-41 - cont.



The entire process of 900 iterations takes 486 minutes, but the estimate is reasonably good within approximately 300 minutes.

## SR-41 - cont.

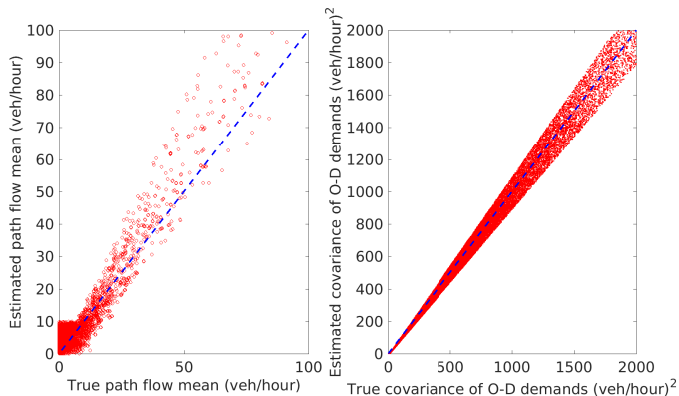


Figure: Estimated and “true” OD demand (Left: mean; Right: covariance)

## SR-41 - cont.

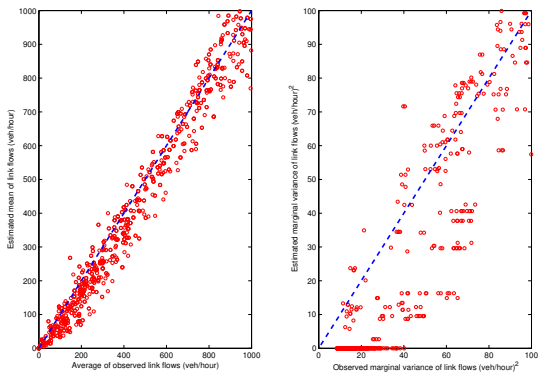
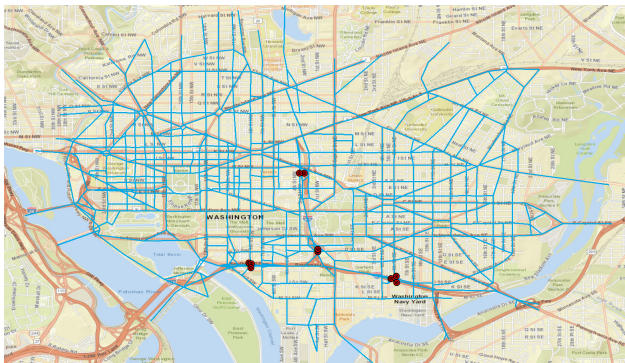


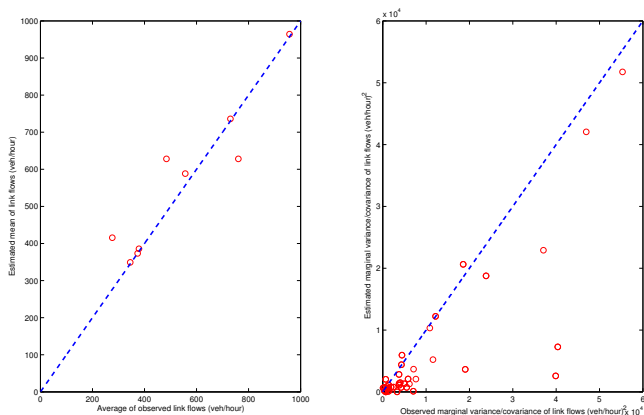
Figure: Estimated and “observed” link flow (Left: mean; Right: variance of the marginal distributions)

# Washington D.C. Downtown network



- 984 road junctions
- 2,585 road segments
- 4,900 O-D pairs

# Washington D.C. Downtown network - cont.



**Figure:** Estimated and observed link flow during the morning peak (Left: mean; Right: variance/covariance)

# What's next

## Unsupervised learning:

- Weekdays/weekends
- Seasonal behavior

## Hypothesis test and variance analysis

- Recurrent-nonrecurrent pattern detection
- Real-time subgraph anomaly detection

## Extensions:

- Prior for variance/covariance matrix
- Other data sets, e.g., speeds
- Dynamic OD demand
- Multi-modal

# MAC data sets in Pittsburgh

## **GIS, demographics, economics, weather**

### **Traffic counts**

- Highways, major arterials

### **Traffic time/speed**

- INRIX, HERE, Uber Movement, AVI, BT

### **Transit**

- APC-AVL, Park-n-ride, incidents

### **Parking**

- Transactions of on-street meters and occupancy of garage

### **Incidents**

- RCRS/PD/911/311/PTC/PennDOT Crash/Road closures/Events

### **Social media (e.g., Twitter)**

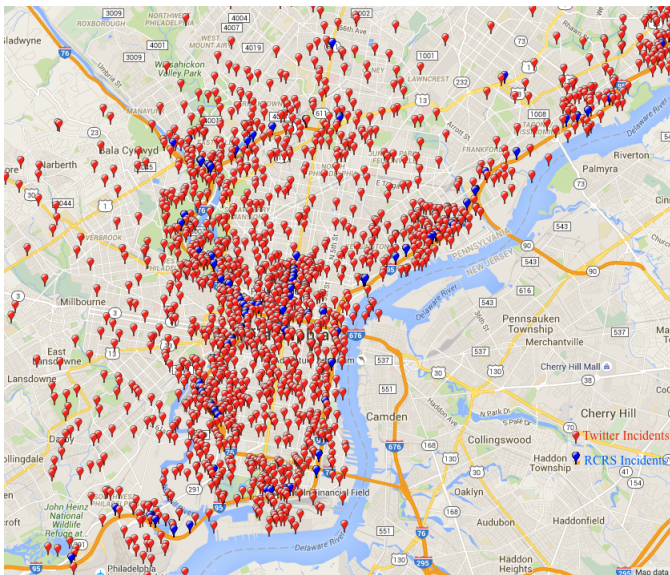
# Mobility Data Analytics Center (big MAC)

Using data analytics, quantitative techniques, and domain knowledge to address real-world problems

- Twitter-based incident detection
- Off-line dynamic network analysis
- Real-time traffic operation
- Parking
- Public transit



# Twitter-based incident detection



# Twitter-based incident detection

End Time: 23 : 55

Sensitivity: HIGH

QUERY

UPDATE REAL-TIME

DB Query Freq: 3 MIN

RESET DB FREQUENCY

Keyword(KW):

INPUT KW DELETE KW

Influential User(IU):

INPUT IU DELETE IU

2016-04-07 22:07:06::Turnpike Roadwork on Pennsylvania Turnpike I-476 northbound between Exit 20 - Pennsylvania Turnpike I-276 and Exit 44 - PA 663 affecting any [Show on map](#)

2016-04-07 22:07:05::Turnpike Roadwork on Pennsylvania Turnpike I-476 northbound between Exit 20 - Pennsylvania Turnpike I-276 and Exit 31 - PA 63 affecting the [Show on map](#)

2016-04-07 21:52:51::My problem is I move too fast. I need to pump my breaks before I crash I already had a few accidents [Show on map](#)

2016-04-07 21:07:07::Turnpike Roadwork on Pennsylvania Turnpike I-476 southbound between Exit 31 - PA 63 and Exit 20 - Pennsylvania Turnpike I-276 affecting the [Show on map](#)

2016-04-07 21:07:07::Turnpike Roadwork on Pennsylvania Turnpike I-476 southbound between Exit 44 - PA 663 and Exit 20 - Pennsylvania Turnpike I-276 affecting any [Show on map](#)

2016-04-07 21:07:06::Roadwork on I-376 westbound between Mile Post: 50.0 and Mile Post: 49.5. There is a lane restriction. [Show on map](#)

2016-04-07 21:05:05::Roadwork on I-376 eastbound between Mile Post: 53.0 and Mile Post: 54.0. There is a lane restriction. [Show on map](#)

2016-04-07 20:39:36::Accident cleared in #RossTwip on Thompson Run Rd Both NB/SB between Sutter Rd and Amity Dr #traffic https://t.co/SL00qn0Vyr [Show on map](#)

2016-04-07 20:39:05::CLEARED: Disabled vehicle on I-376 westbound at Exit 69C - PA 837 North/Carson St. [Show on map](#)

2016-04-07 20:35:05::UPDATE: Disabled vehicle on I-376 westbound at Exit 69C - PA 837 North/Carson St. There is a lane restriction. [Show on map](#)

2016-04-07 20:31:08::Disabled vehicle on I-376 westbound at Exit 69C - PA 837 North/Carson St. There is a lane restriction. [Show on map](#)

2016-04-07 20:26:19::Can we get a plane crash with Clint Kapper and Penny? #HateThemAll [Show on map](#)

# Off-line dynamic network analysis

Link Parameter Updates

Start Time: 7:30 AM

End Time: 9:00 AM

Task Name: Test\_Name

Animation Resolution: 30

Vehicle Spiker: 2

Kahortest Path: 3

Dispersion Factor: 0.5

SUBMIT

Link Info Update

Lane Capacity: 750

Road Segment ID: 133321

Highway name: 937

County: Burlington

undefined: 0

Vehicle Types: Sk,Bus,Ca,Tr

Free Flow Speed: 35

Road Length: 0.0628

Number of Lanes: 1

Free Flow Speed: 30

Capacity: 500

Close this link

SUBMIT

Legend: Grayscale, Streets

Carnegie Mellon University | Heinz College | Traffic21 | Founding Source | Contact Us

Legend | © OpenStreetMap contributors, © OpenStreetMap contributors

# Off-line dynamic network analysis

HOME OFF-LINE TRAFFIC PREDICTION ON-LINE TRAFFIC MANAGEMENT LOGOUT

### Link Parameter Updates

id: 133321, closed

Start Time: 6:45 AM

End Time: 9:00 AM

Task Name: Road Closure

Animation Resolution: 30

Vehicle Scaler: 2

Kahortest Path: 3

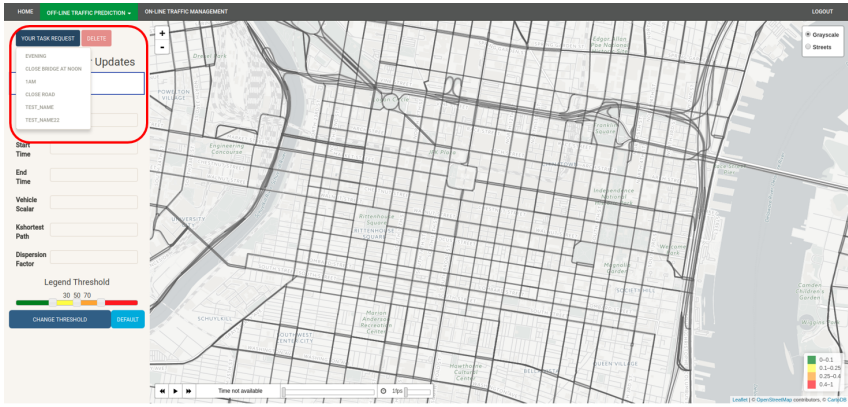
Dispersion Factor: 0.5

SUBMIT

Legend: Originals Streets

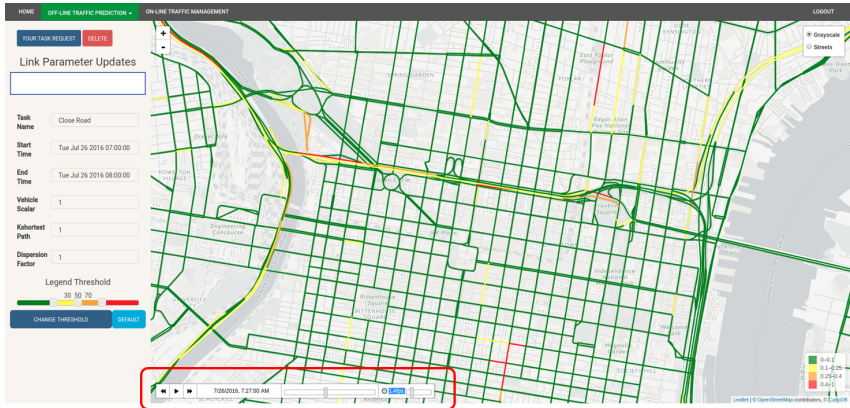
Carnegie Mellon University | Heinz College | Traffic21 | Founding Source | Contact Us

# Off-line dynamic network analysis



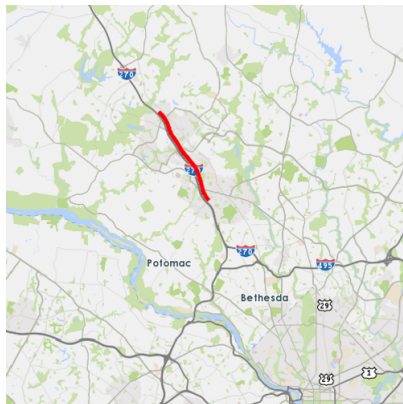
Carnegie Mellon University | Heinz College | Traffic21 | Founding Source | Contact Us

# Off-line dynamic network analysis

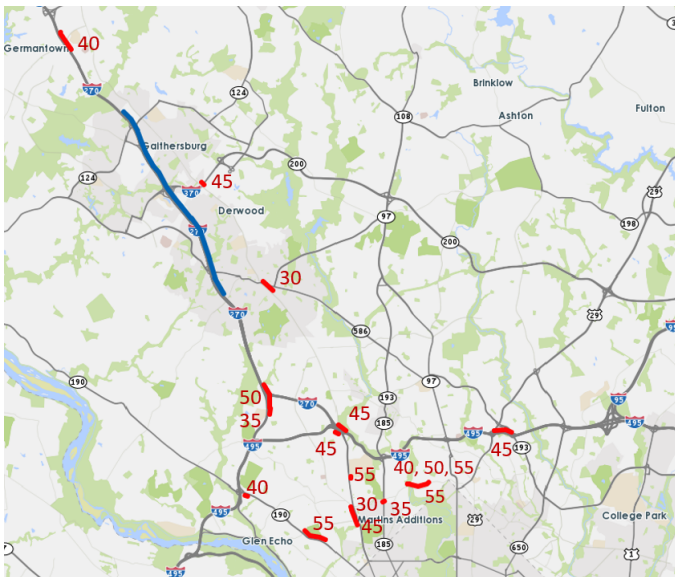


Carnegie Mellon University | Heinz College | Traffic21 | Founding Source | Contact Us

# Real-time traffic operation: traffic prediction



# Real-time traffic operation: traffic prediction





# Real-time traffic operation: demand management

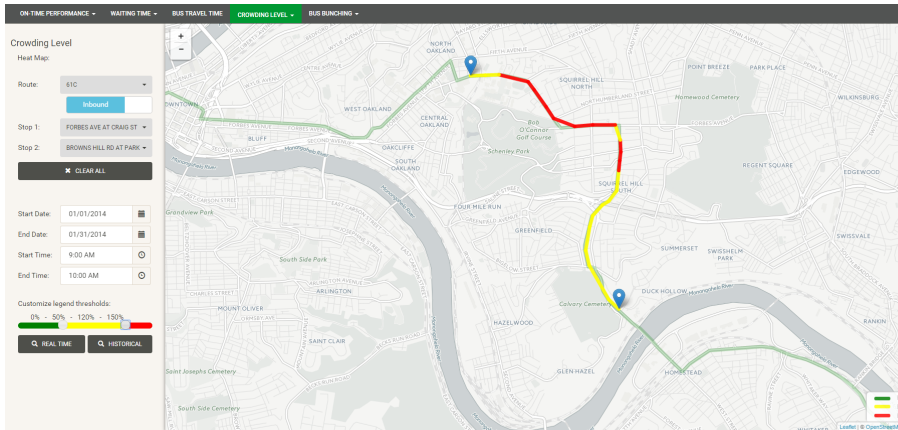
The screenshot displays the MDAP - Traffic Congestion online interface. The browser address bar shows the URL: `bruno.heinz.cmu.edu/traffic/congestion_online/`. The interface is divided into several sections:

- Navigation:** HOME, OFF-LINE TRAFFIC PREDICTION, ON-LINE TRAFFIC MANAGEMENT (highlighted), and LOGOUT.
- Link Parameter Updates:** A form for updating link parameters. It shows "ID:9081, Capacity Drop: 40%" and a value of "4" in a text input field. A "SUBMIT" button is visible with a "5" next to it.
- Legend Threshold:** A color scale legend ranging from 30 to 70. The scale is currently set to 50. A "CHANGE THRESHOLD" button and a "DEFAULT" button are present.
- Link Info Update:** A popup window showing details for Road Segment ID 9081:
  - Road Segment ID: 9081
  - Free Flow Speed: 29
  - Vehicle Types: Bik,Bus,Car,Trk,Tri
  - Number of Lanes: 2
  - County: Philadelphia
  - Highway name:
  - Lane Capacity: 1440
  - Road Length: 0.0755
  - direction: 0
  - Capacity Drop(%): 2 (with "0 means free flow")
 A "SUBMIT" button is located at the bottom of this popup.
- Map:** A map of Philadelphia showing traffic congestion levels. The map is overlaid with green lines representing road segments. A legend on the right indicates congestion levels: 0-0.3 (green), 0.3-0.5 (yellow), 0.5-0.7 (orange), and 0.7-1.0 (red). The map also shows a "Grayscale" and "Streets" toggle.
- Map Controls:** Navigation arrows, a timestamp "7/30/2016, 2:45:00 PM", a scale of "1fps", and map data sources: "Leaflet | © OpenStreetMap contributors, © CartoDB".

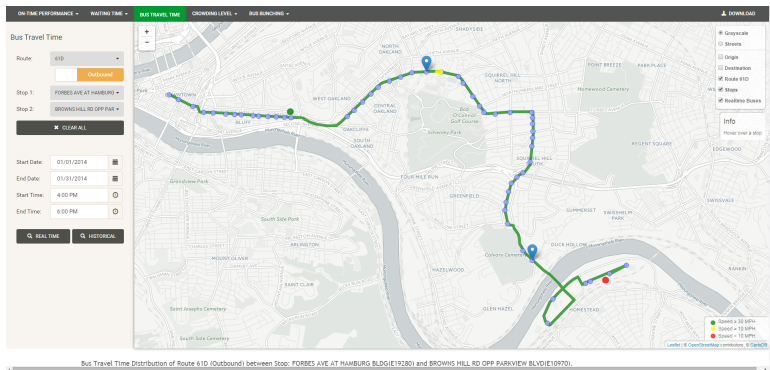
# Real-time traffic operation: demand management



# Public transport



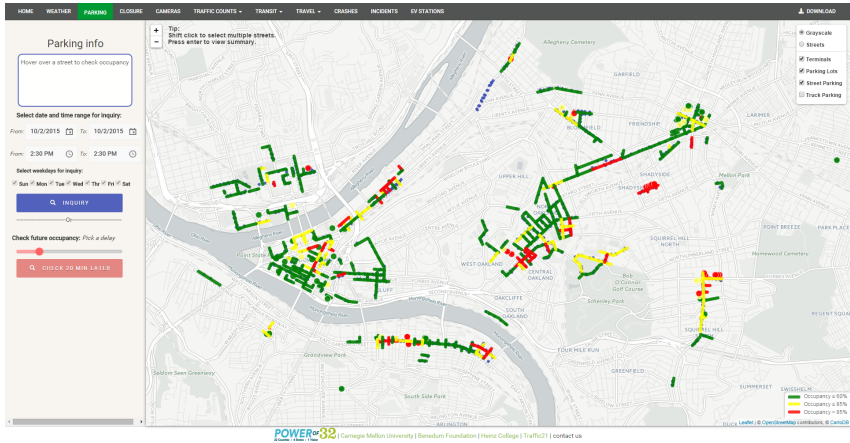
# Public transport



Bus Travel Time Distribution of Route 61D (Outbound) between Stop: FORBES AVE AT HAMBURG BLDG(E19280) and BROWNS HILL RD OPP PARKVIEW BLVD(E10970).



# Parking



# Team

## CURRENT MEMBERS



**Prof. Sean Qian**  
Director



**Xidong Pi**  
PhD Candidate



**Wei Ma**  
PhD Candidate



**Pinchao Zhang**  
PhD Candidate



**Shuguan Yang**  
PhD Candidate



**Matthew Battifarano**  
PhD Student



**Weiran Yao**  
PhD Student



**Rick Grahn**  
PhD Student

## ALUMNI



**Dr. Yiming Gu**  
Senior Scientist  
United Technologies  
Research Center



**Zhangning Hu**  
Software Engineer  
Google



**Zach Sussman**  
Software Engineer  
Apple



**Juncheng Zhan**  
Undergraduate  
SCS CMU

*Thanks! Questions and comments?*

*Sean Qian, seanqian@cmu.edu*

MAC: mac.heinz.cmu.edu



**CLAUDE  
WORTHINGTON  
BENEDUM  
FOUNDATION**



A Division of Traffic 21 Research Institute  
**traffic21**  
TRANSPORTATION RESEARCH CENTER OF CHANGING MOBILITIES

