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

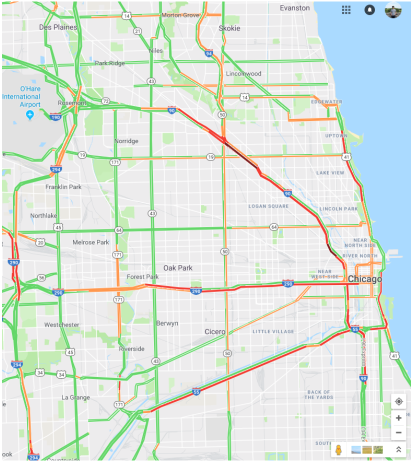
1. Massive data: opportunities and challenges

2. Statistical Origin-Destination Demand Estimation

3. Mobility Data Analytics Center (big MAC)

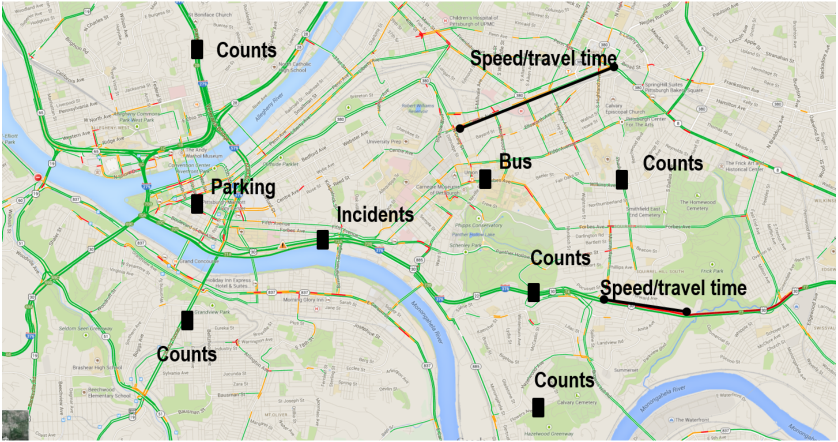
Smart decision making?

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

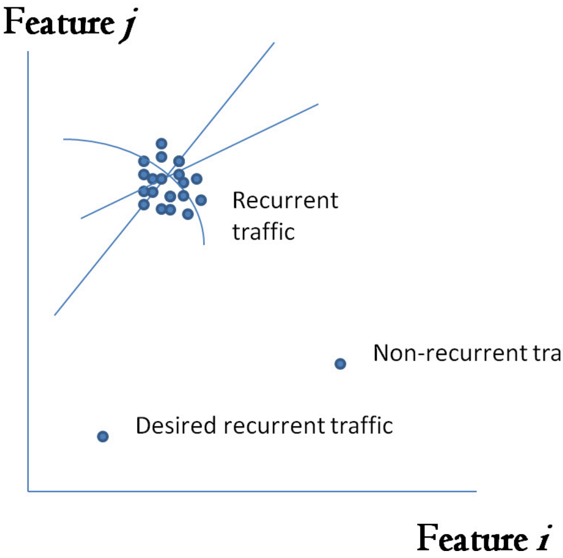


Massive data: useful but challenging

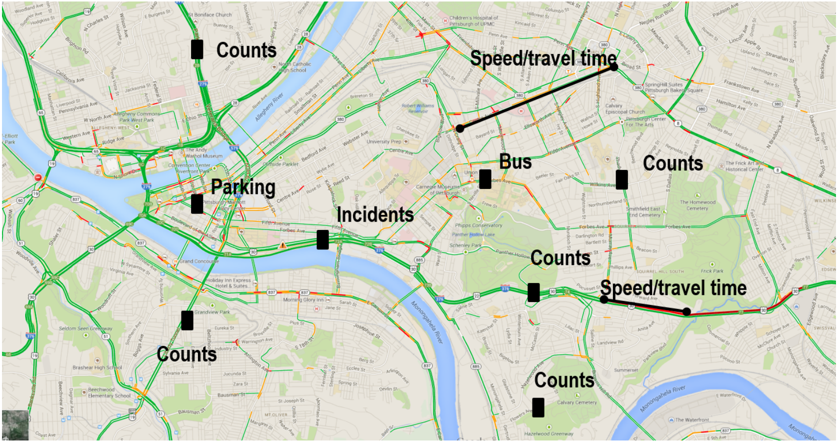
Fusion. Bias. Sparsity. Computation. Unexplored space.



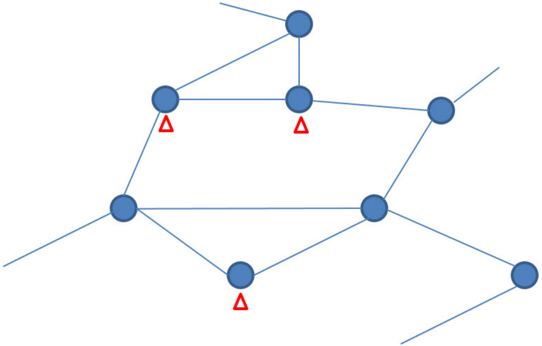
Unexplored space



A possible solution: data + physics



A generic infrastructure network



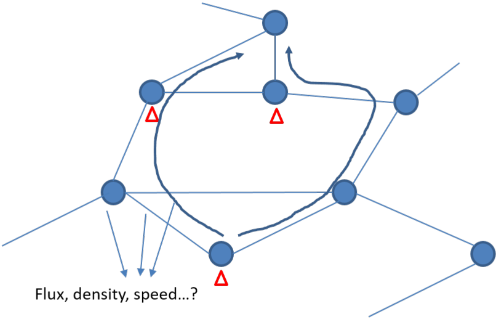
An infrastructure node

Users: human beings

Goods: water, energy, vehicle...

A sensor

A generic infrastructure network

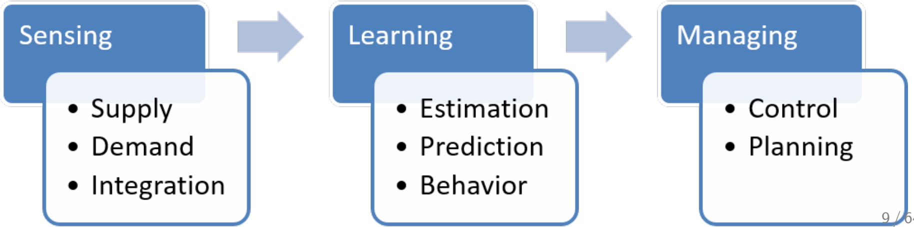


Final goals: evaluation and intervention

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

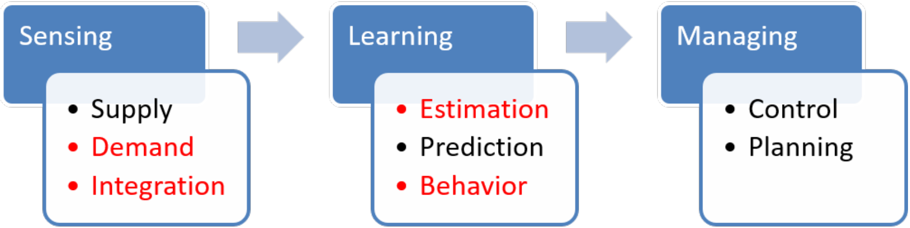
Sensing-Learning-Managing

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

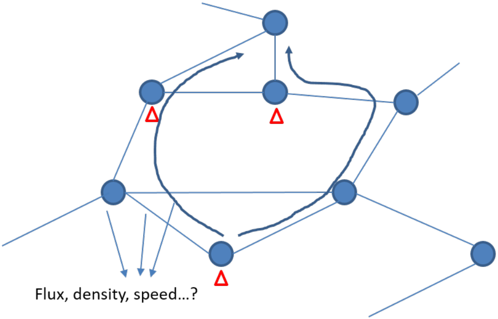


Sustainable mobility

- 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 $(x, f, c) = G(\text{supply}, \text{demand})$

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)

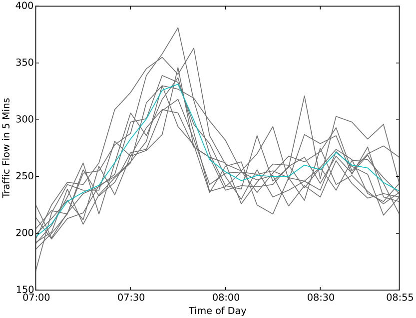
Traffic Assignment

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)

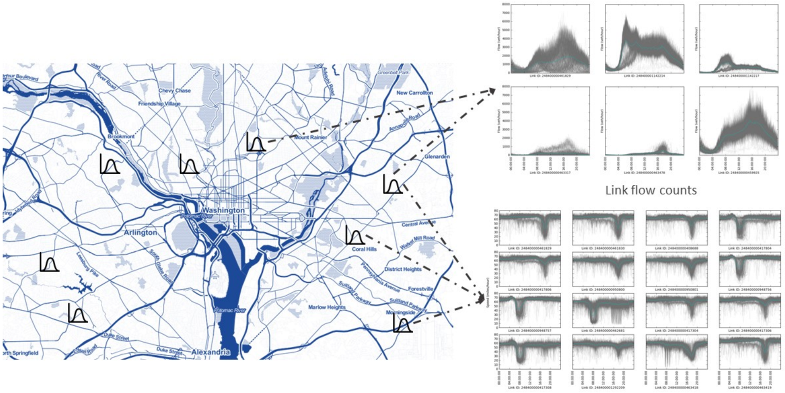
Probabilistic traffic demand: $Q \sim \mathcal{N}(q, \Sigma_q)$

GESTA - cont.

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

GESTA - cont.

Sensing: $X_m = X + \epsilon_e$

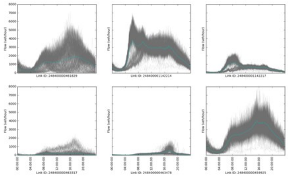


Overview of DC network

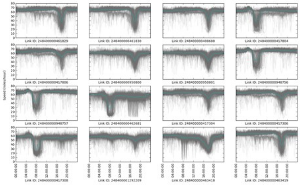
Link flow speeds

GESTA - cont.

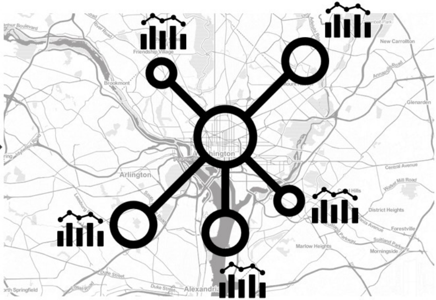
System states: $C = C(X, F)$



Link flow counts

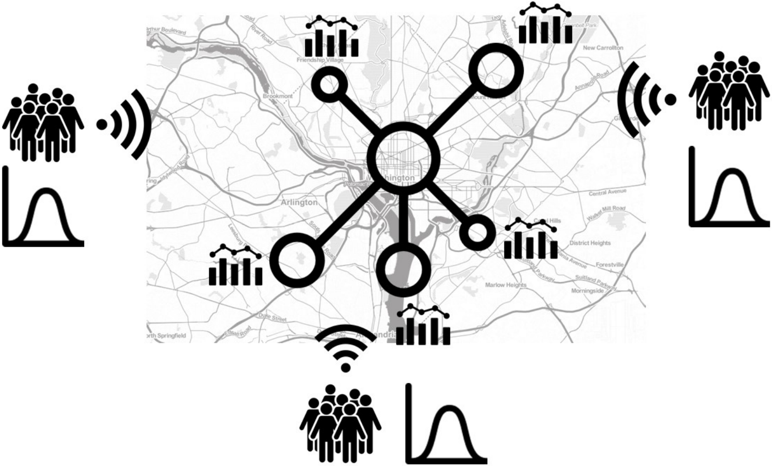


Link flow speeds

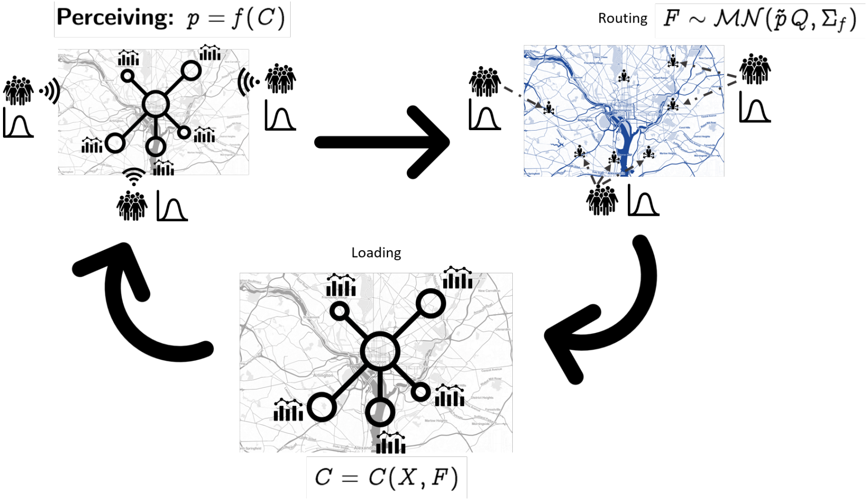


GESTA - cont.

Perception: $p = f(C)$



GESTA - cont.


Perceiving: $p = f(C)$
Routing $F \sim \mathcal{MN}(\tilde{p}Q, \Sigma_f)$
Loading
$C = C(X, F)$

GESTA - cont.

$$
\begin{aligned}
\textit{Level } 1: \quad & X_m = X + \epsilon_e && \text{(Unknown Error)} \\
& \epsilon_e \sim \mathcal{N}(\mathbf{0}, \Sigma_e) \\
\textit{Level } 2: \quad & X = \Delta F \\
& F \sim \mathcal{MN}(\tilde{p}Q, \Sigma_f) && \text{(Route choice variation)} \\
\textit{Level } 3: \quad & Q \sim \mathcal{N}(q, \Sigma_q) && \text{(Demend variation)}
\end{aligned}
$$

GESTA - cont.

GESTA features:

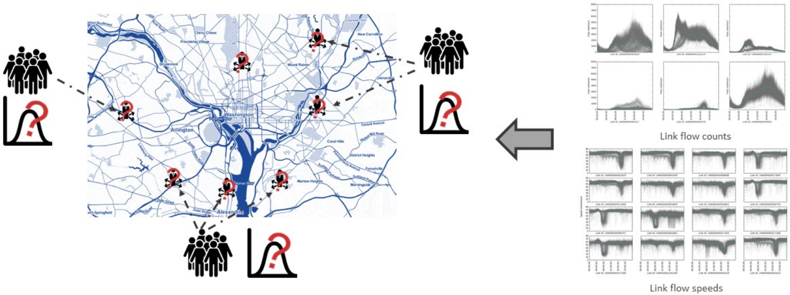
- Daily traffic condition is not in equilibrium
- Statistical equilibrium is built in a probability space
- Link/path flow variance = demand variance + choice variance

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

$$\min_q L(x^o, Aq) \tag{1}$$

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

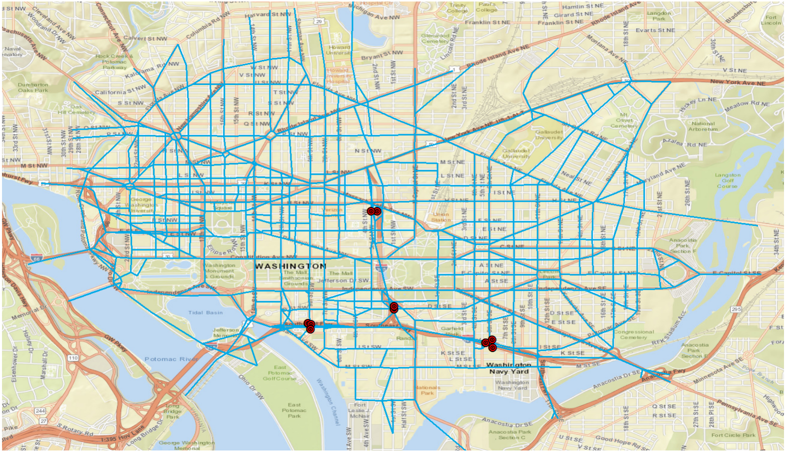
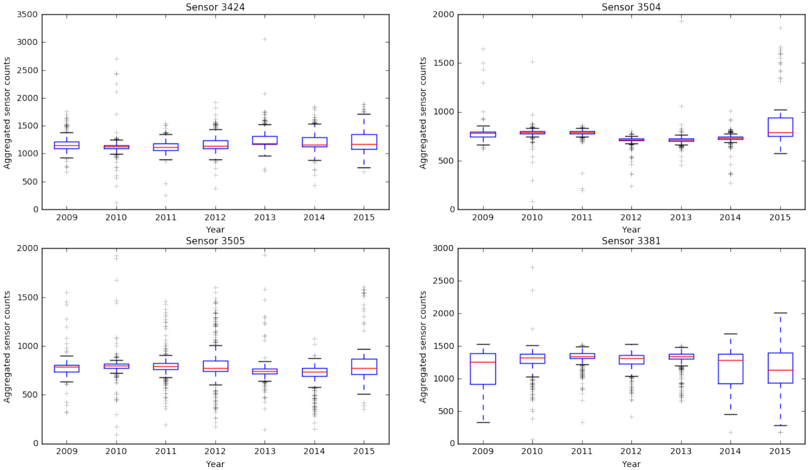


Figure: The Washington D.C. Downtown network

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

$$\min_{q,\Sigma_q} R(X^o, AQ) \tag{2}$$

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, Σ_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

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

Where Φ^+ 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:

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

Methods:

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

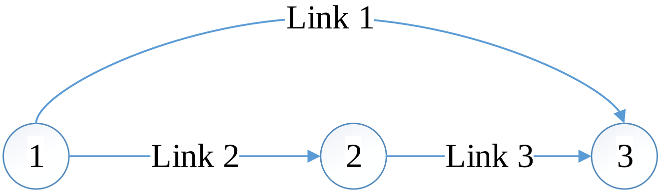
Observerbility

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

A small example - cont.

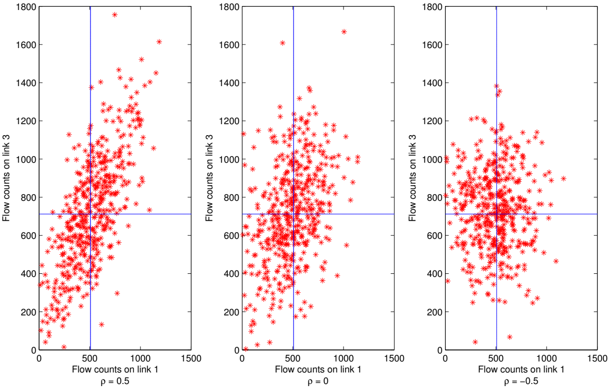


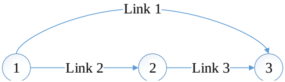
Figure: Synthesized "true" link flow data for different correlation ρ

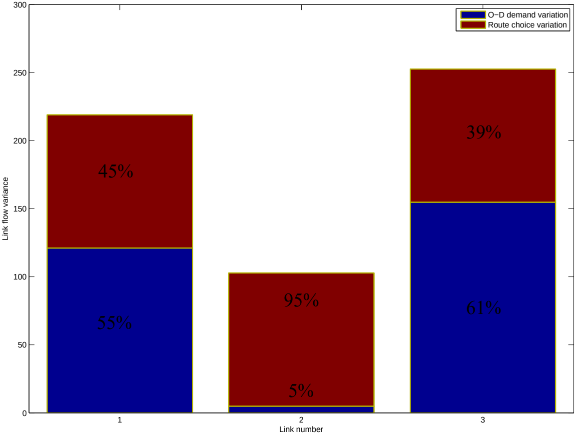
A small example - cont.

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

True ρ	Settings	$\hat{q}_{1\to3}$	$\hat{q}_{2\to3}$	$\hat{\sigma}^2_{1\to3}$	$\hat{\sigma}^2_{2\to3}$	$\hat{\rho}$	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

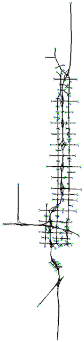
A small example - cont.



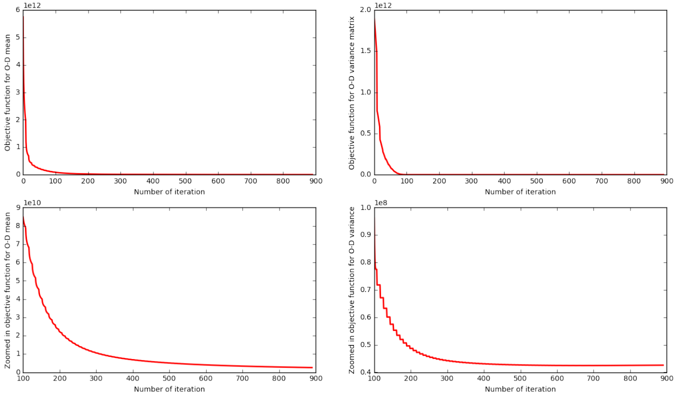


SR-41 Corridor

- 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.

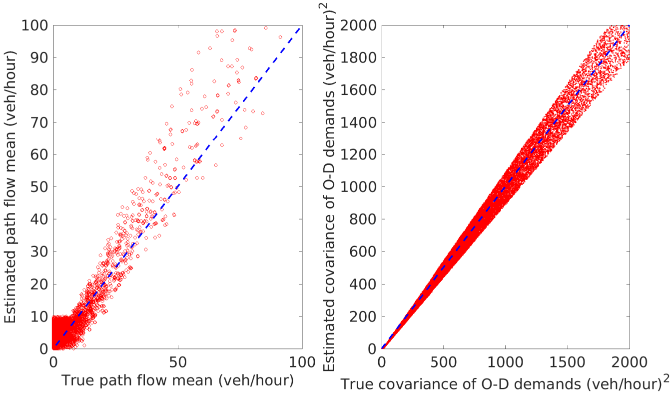


Figure: Estimated and "true" OD demand (Left: mean; Right: covariance)

SR-41 - cont.

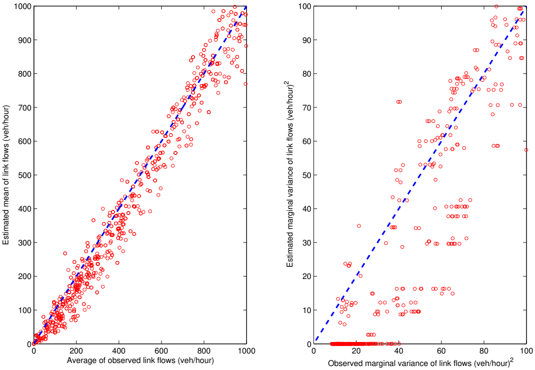
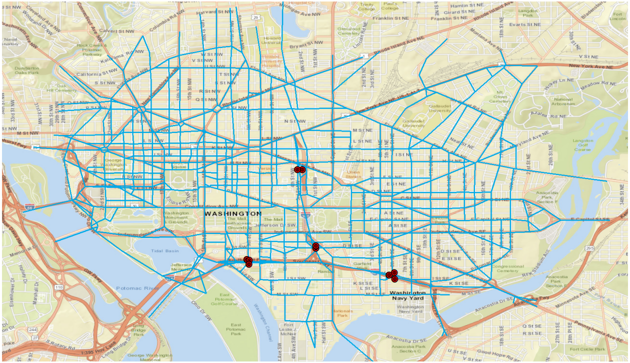


Figure: Estimated and “observed” link flow (Left: mean; Right: variance of the marginal distributions)

Washington D.C. Downtown network



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

Washington D.C. Downtown network - cont.

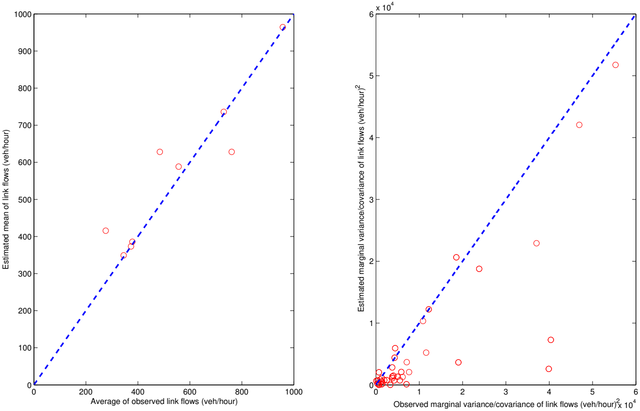


Figure: Estimated and observed link flow during the morning peak (Left: mean; Right: variance/covariance)

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

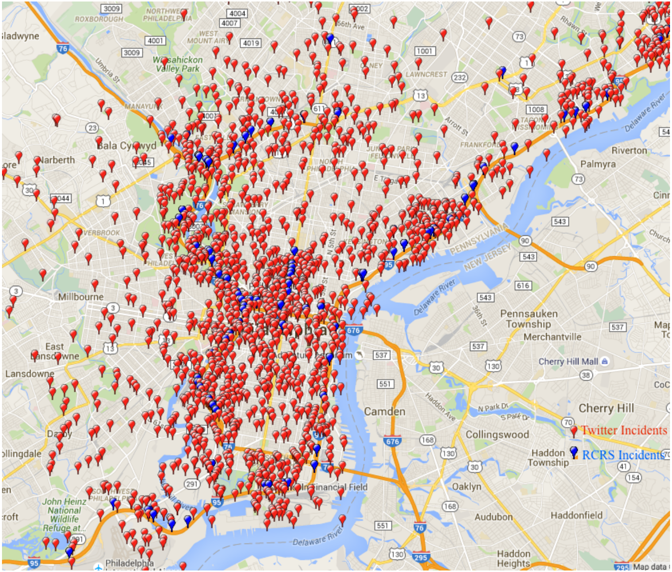
Social media (e.g., Twitter)

Mobility Data Analytics Center (big MAC)

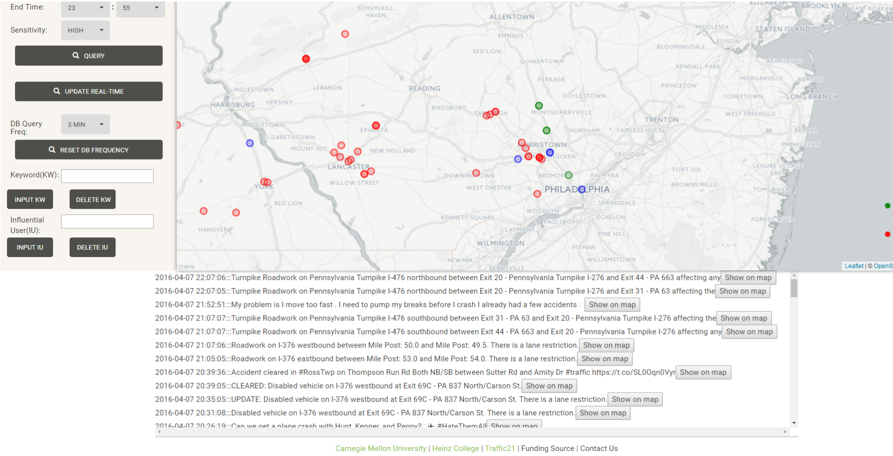
Using data analytics, quantitative techniques, and domain knowledge to address real-world problems

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

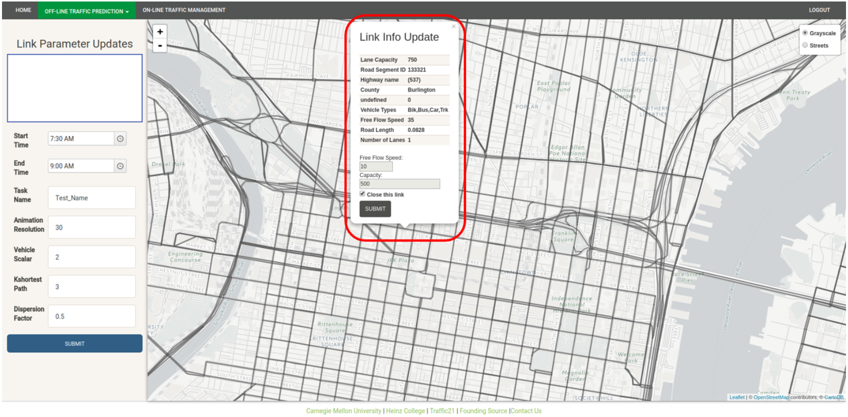
Twitter-based incident detection



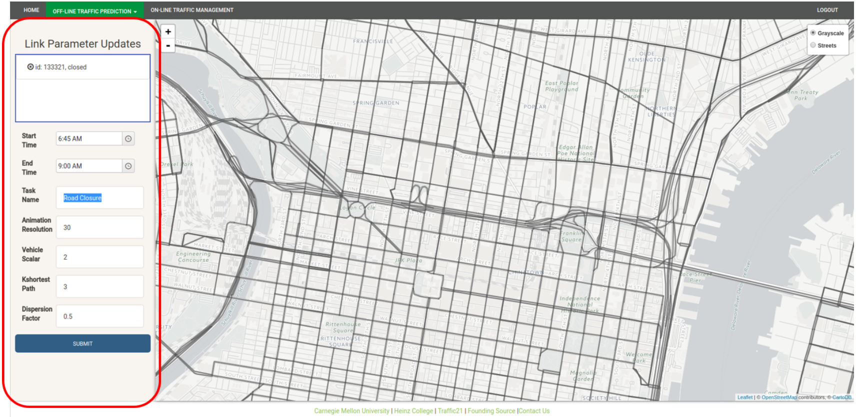
Twitter-based incident detection



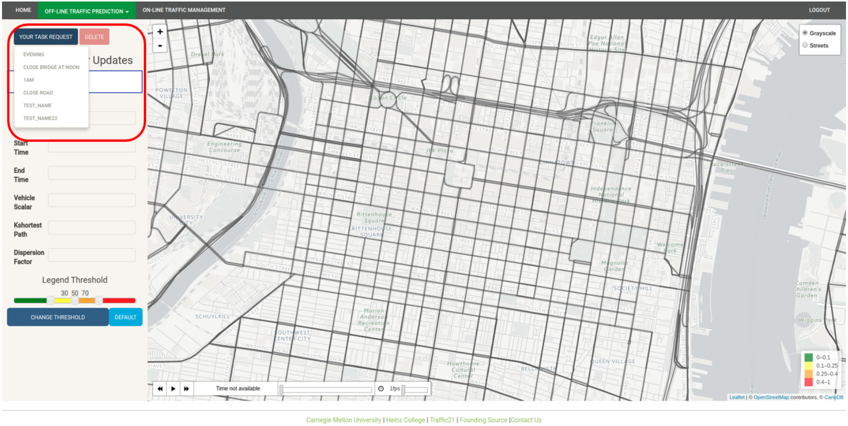
Off-line dynamic network analysis



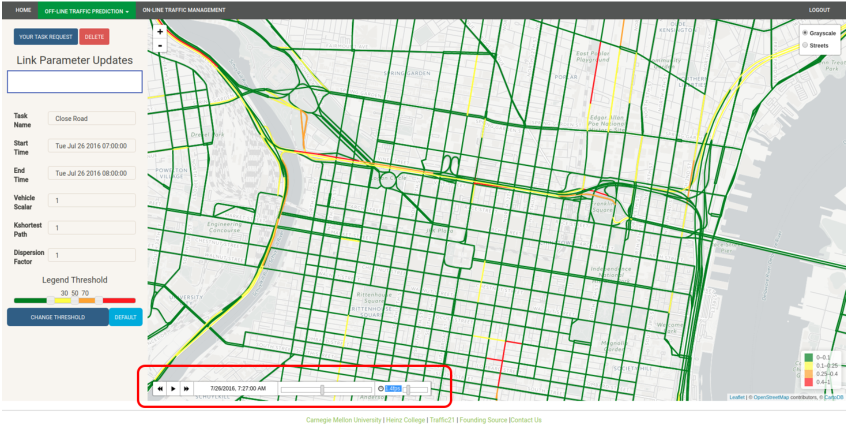
Off-line dynamic network analysis



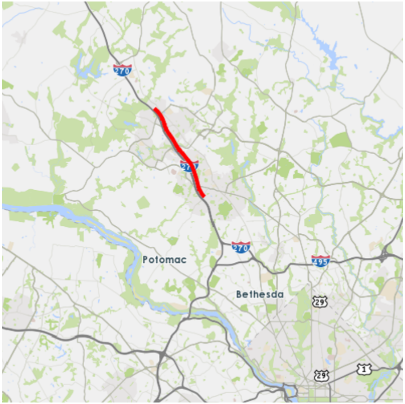
Off-line dynamic network analysis



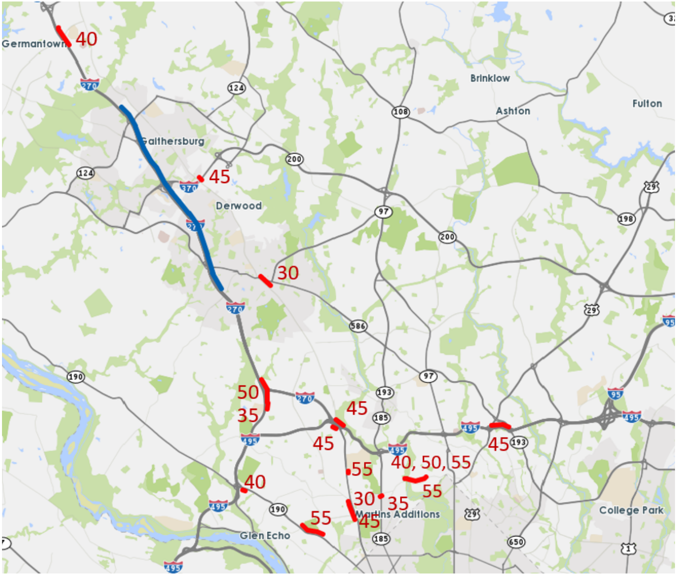
Off-line dynamic network analysis



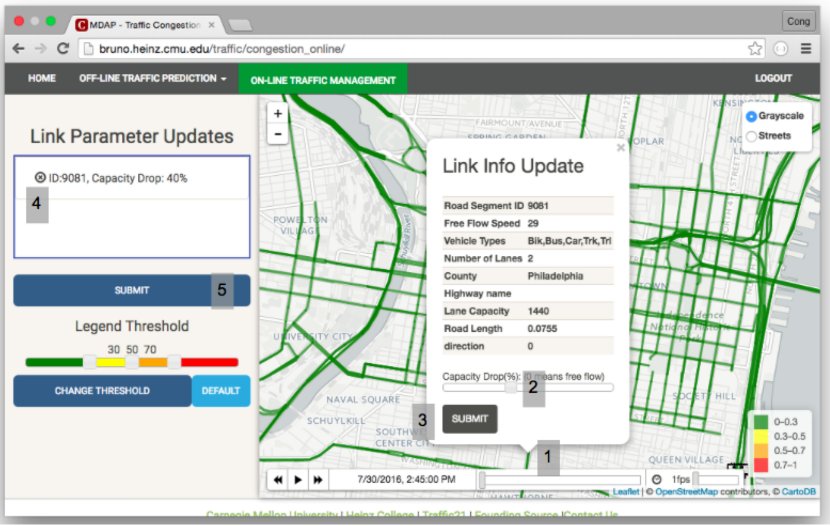
Real-time traffic operation: traffic prediction



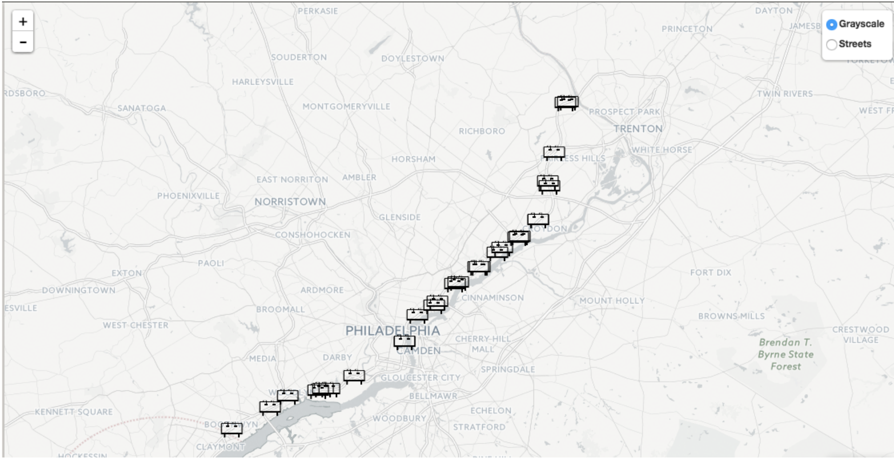
Real-time traffic operation: traffic prediction



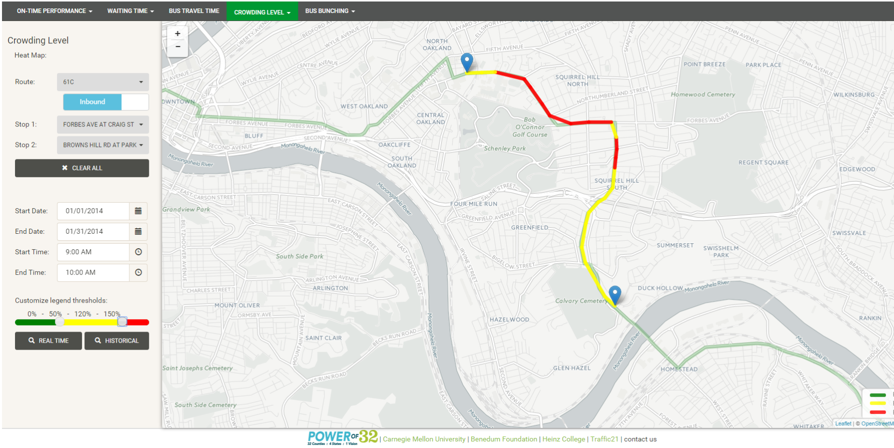
Real-time traffic operation: demand management



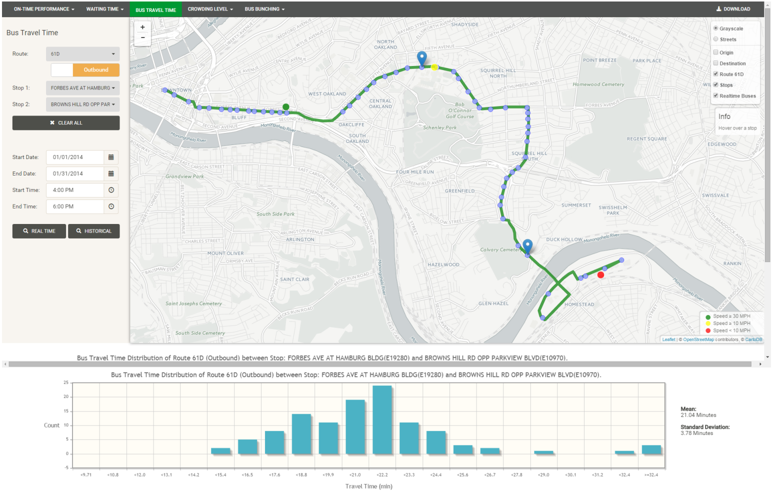
Real-time traffic operation: demand management



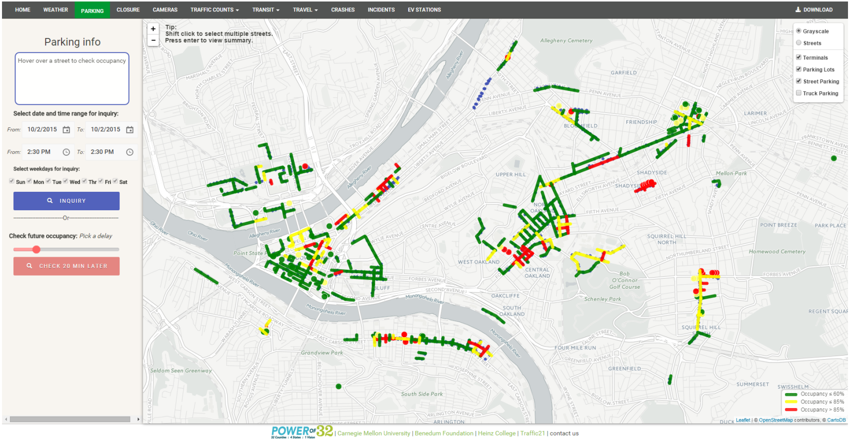
Public transport



Public transport



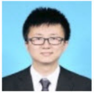
Parking

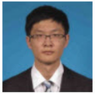

HOME
WEATHER
PARKING
CLOSURE
CAMERAS
TRAFFIC COUNTS
TRANSIT
TRAVEL
CRASHES
INCIDENTS
EV STATIONS
DOWNLOAD
Parking info
Hover over a street to check occupancy
Select date and time range for inquiry:
From: 10/2/2015 To: 10/2/2015
From: 2:30 PM To: 2:30 PM
Select weekdays for inquiry:
Sun Mon Tue Wed Thr Fri Sat
INQUIRY
Or
Check future occupancy: Pick a delay
CHECK 20 MIN LATER
Tip:
Shift click to select multiple streets.
Press enter to view summary.
Grayscale
Streets
Terminals
Parking Lots
Street Parking
Truck Parking
Occupancy ≤ 60%
Occupancy ≤ 85%
Occupancy > 85%
Leaflet | © OpenStreetMap contributors, © CartoDB
POWER OF 32 | Carnegie Mellon University | Benedum Foundation | Heinz College | Traffic21 | contact us

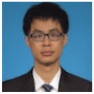
Team

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

PENNDOT
PITA
CLAUDE WORTHINGTON BENEDUM FOUNDATION
U.S. Department of Transportation Federal Highway Administration
A Division of Traffic 21 Reseaerch Institute traffic21
Metro21
Technologies for Safe and Efficient Transportation
A U.S. DOT UNIVERSITY TRANSPORTATION CENTER
National Science Foundation WHERE DISCOVERIES BEGIN