Inferring Traffic Cascading Patterns

Yuxuan Liang, Zhongyuan Jiang, Yu Zheng

v-yuxuli@microsoft.com

Microsoft Research
微软亚洲研究院

XIDIAN UNIVERSITY 1931

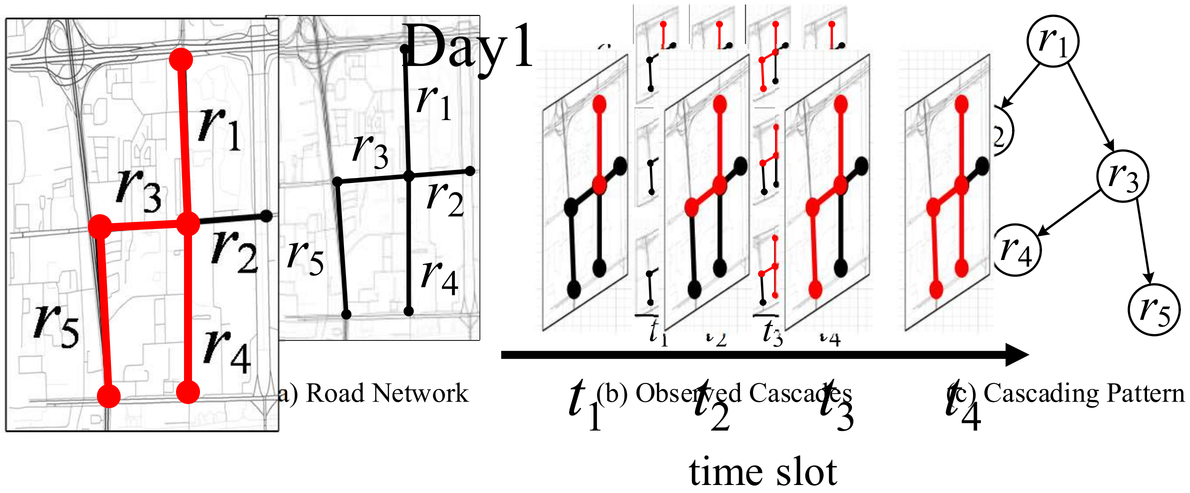
SIGSPATIAL 2017

Traffic Cascading Patterns

Traffic congestion in a city's road network usually spreads or relieves through the traffic cascading patterns.

Here we give it a **definition**

- A cascading pattern G is considered as a time-evolving underlying network over which traffic congestion spreads in a given time span
- We assume the pattern in a specific time span (i.e., morning rush hours) is uniform



What we do - Inferring Traffic Cascading Patterns

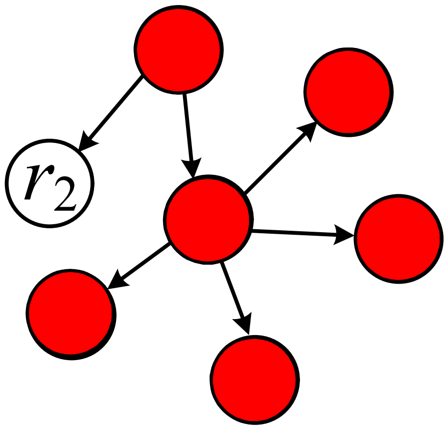
We propose a data-driven generative probabilistic approach to infer the cascading patterns.

Motivation

- Predict future traffic conditions
- Identify bottlenecks of road networks

Assumption

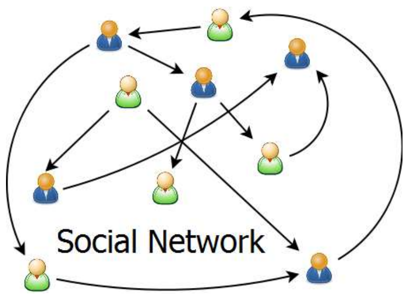
- We assume the traffic congestion spreads in the format of directed tree(no-loops, no many-to-one spreading)

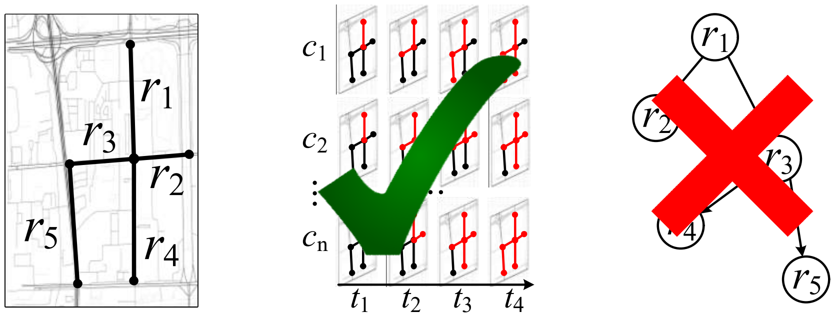


Obviously, r_3 is the bottleneck of the road network

Challenges

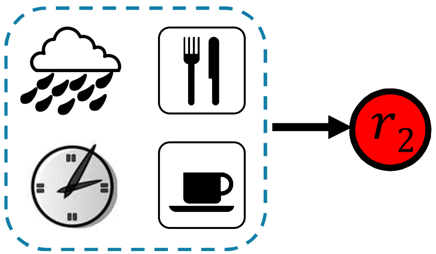
- Implicit interaction





(a) Road Network (b) Observed Cascades (c) Cascading Pattern

- Multiple sources
  - Neighboring traffic
  - Surrounding environment



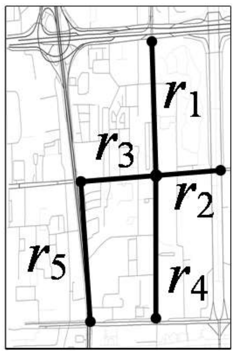
- Geospatial correlations
  - Waiting time (temporal distance)
  - Spatial distance

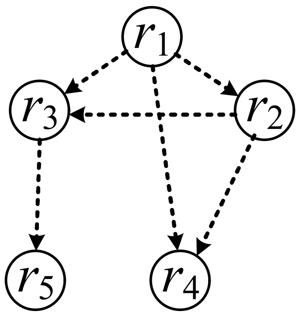
Preliminary

Definition 1 (*Road Network*)

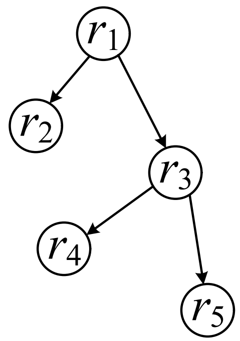
Definition 2 (*Traffic Condition*): binary status, congested or smooth at a fixed time slot

Definition 3 (Cascading Patterns): A cascading pattern G is considered as an time-evolving underlying *network* over which traffic congestion spreads in a given time span. The directed edges connecting two different road segments as *casual links*.





Road Connectivity Graph



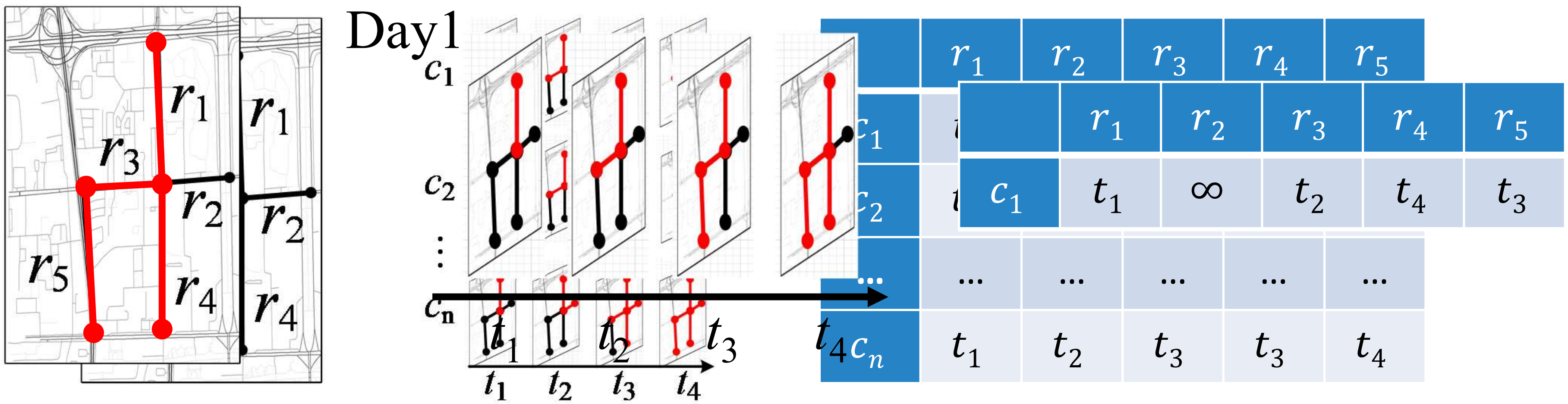
Cascading Pattern

Preliminary - Cascade Extraction

Definition 4 (***Cascade***): A cascade is a n-D vector extracted from one day's observed traffic conditions.

$c = \{t_1, t_2, \dots, t_n\}$,where t_i records when r_i gets congested and $t_i \in R$.

In fact, most cascades hit not all road segments, so we set $t_i = \infty$ when r_i is not hit by a cascade.

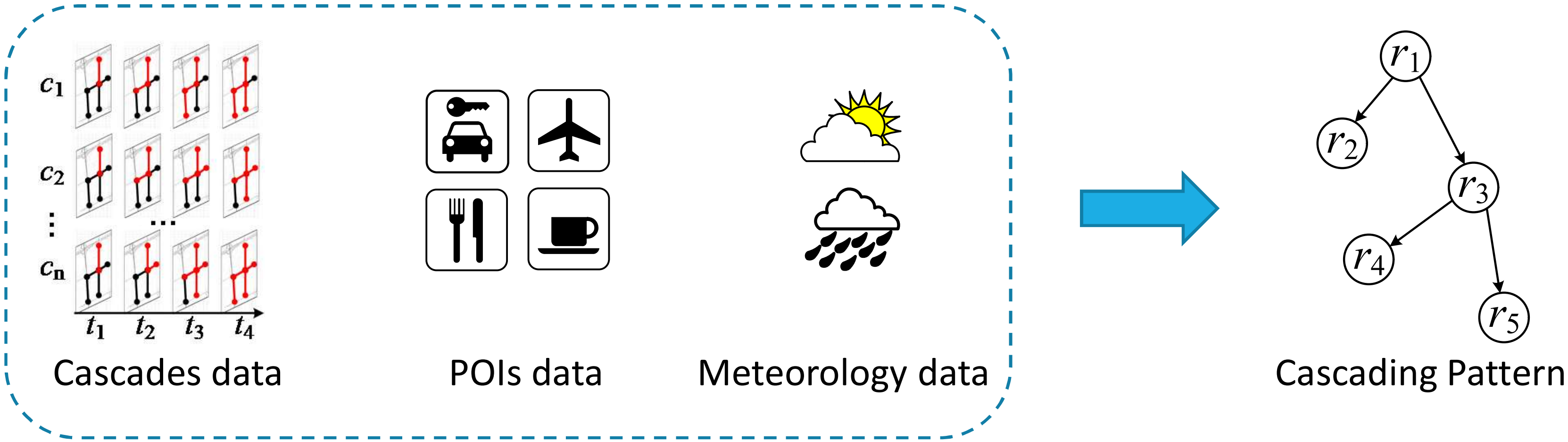


Problem Statement

Given

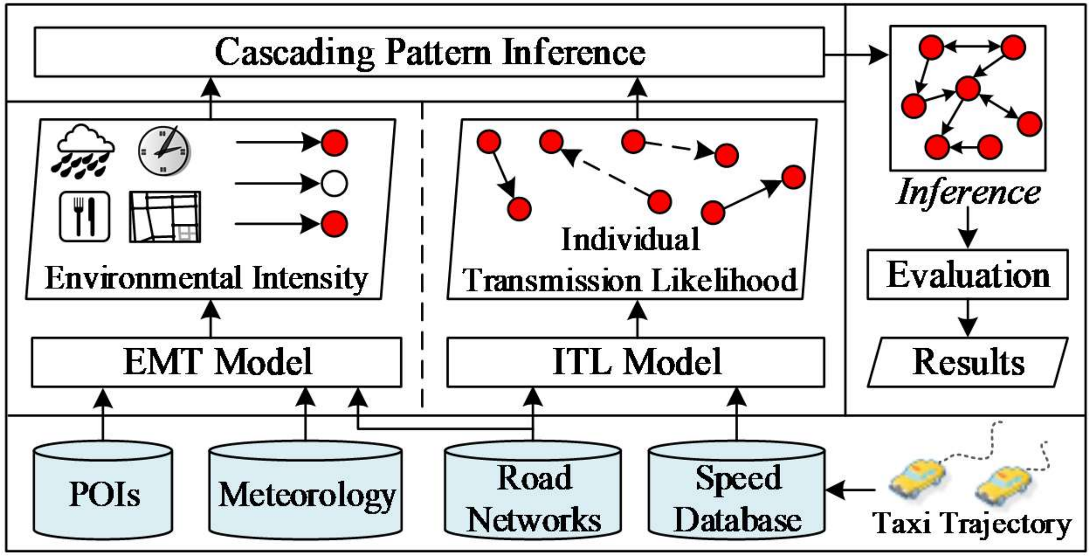
- a road network
- a time span including several time intervals
- Discrete traffic conditions on road segments at each time interval during a period of m days

We extract a set of cascades $C = \{c_1, c_2, \dots, cm\}$



Framework

- Data acquisition
- Multiple sources modeling
- Cascading pattern construction
- Evaluation



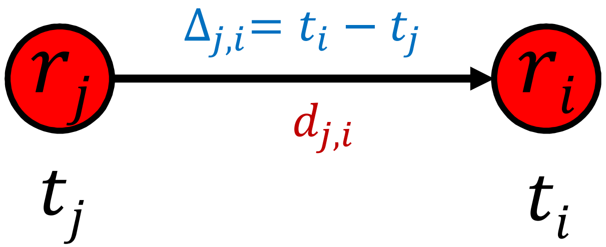
Multiple Source Modeling

Neighboring Traffic

Individual transmission likelihood model based on

- Waiting time (temporal distance)
- Road distance (spatial distance)

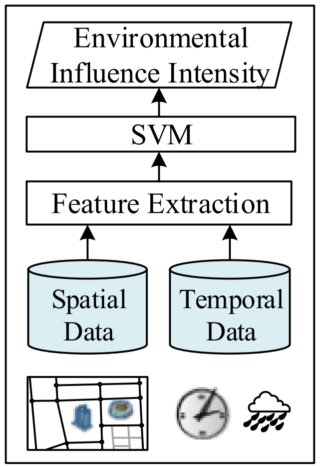
$$f(t_i | t_j; a_{j,i}, \lambda) \propto e^{-a_{j,i}(\Delta_{j,i} + \lambda * d_{i,j})}$$



Environmental Factors

Supervised learning approach

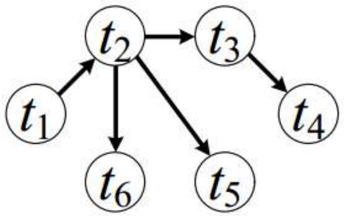
- Feature extraction
  - Temporal features: time of day, meteorology
  - Spatial features: POIs, road network
- SVM-based method



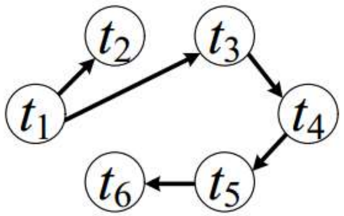
Cascading Pattern Construction

Model formulation

- Suppose we have a cascade $c = \{t_1, t_2, \dots, t_6\}$, where $t_i > t_{i-1}$. There are some possible propagation trees and 2 simple examples as follows.

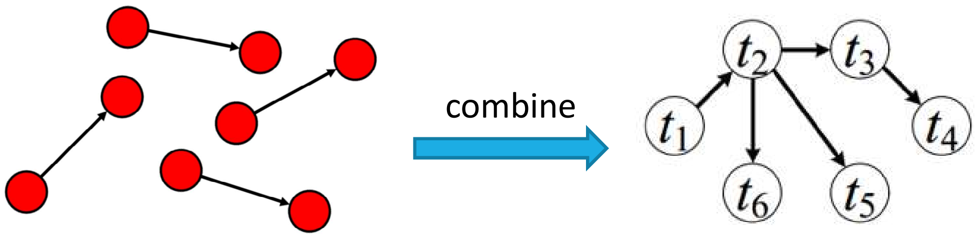


Propagation Tree 1



Propagation Tree 2

- Given a propagation tree T, the likelihood of a cascade c:



Individual transmission likelihood

Given tree structure

$$f(c|T) = \prod_{(j,i)\in E_T} f\left(t_i | t_j; a, \lambda\right)$$

Cascading Pattern Construction

Model formulation

- Given a propagation tree T, the likelihood of a cascade c:
- Given a network G considering all possible trees T, the likelihood of a cascade c:

$$f(c|T) \longrightarrow f(c|G) = \sum_{T \in \mathrm{T}_c(G)} f(c|T)\underline{f(T|G)}$$

For a specific cascade, this part is equal.

simplify

$$f(c|G) \propto \sum_{T \in \mathrm{T}_c(G)} f(c|T)$$

Conditional independence

Joint likelihood of C

$$f(C|G) = \prod_{c \in C} f(c|G)$$

Optimization target: $\boxed{\hat{G}} = \arg\max_{|G| \leq \boxed{k}} f(C|G)$ difficult to optimize

$\hat{G}$ best explains the observed cascades,
and the maximization is over all possible graphs G at most k edges

Alternative Optimization (CasInf)

$$\hat{G} = \underset{|G| \le k}{\arg\max} f(C|G)$$

1. matrix tree theorem

2. Combine environmental intensity

3. Log-likelihood

$$\hat{G} = \underset{|G| \le k}{\arg\max} F(C|G)$$

- Alternative target
- It satisfies submodularity, a natural diminishing returns property.

Optimization: greedy algorithm

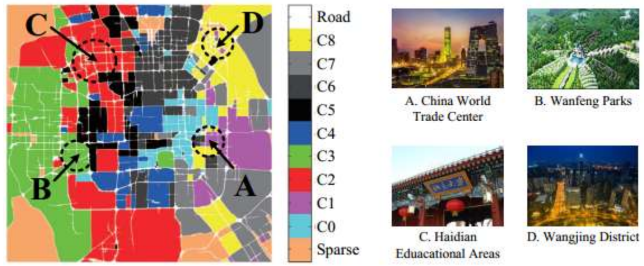
Algorithm 1 Approximate algorithm for *CasInf*

Input:
k : the number of edges in cascading patterns we infer.
C : the set of cascades obtained in a time span.
$\mathcal{D}$: a constant denoting the spatial constraint.
Output: G: the inferred cascading pattern.

```
G ← K̄;
P ← all pairs (j, i): ∃c ∈ C with t_j < t_i and d_{j,i} < D
while |G| ≤ k do
  for all (j, i) ∈ P\G do
    δ_{j,i} = 0;
    for all c : t_j < t_i do
      ε_g ← environmental intensity inference
      w_c(m, n) ← weight of (m, n) in G ∪ (j, i);
      for all m : t_m < t_i and m ≠ j do
        δ_{c,j,i} = δ_{c,j,i} + w_c(m, i);
      end for
      δ_{j,i} = δ_{j,i} + log((δ_{c,j,i} + w_c(j,i)) / δ_{c,j,i});
    end for
  end for
  (j*, i*) ← argmax_{(j,i)∉G} δ_{j,i};
  G ← G ∪ (j*, i*);
end while
```

In each iteration, we select

$$e_i = \underset{e \in G_i \backslash G_{i-1}}{\arg\max} \; F(C|G_{i-1} \cup e) - F(C|G_{i-1}).$$



Settings

Datasets

- Taxi Trajectories data
- Road networks
- Meteorological data
- POIs

Datasets		Values
Taxi Trajectories	Number of Taxies	32,670
	Time Spans	03/01/2015-06/30/2015 11/01/2015-03/31/2016
Road Networks	Number of Segments	249,080
	Number of Nodes	186,266
	Total Length	25,638 km
POIs	Number of POIs	651,016
	Number of Categories	20
Meteorological Data	Number of Stations	16
	Time Spans	01/01/2015-12/31/2016

Ground Truth

- Occurring probability of each casual links
- Score of a cascading pattern

$$prob_{j,i} = \frac{\#hit_{j,i}(\gamma)}{m_j}$$

$$score@k(\hat{G}) = \frac{\sum_{(j,i)\in E_{\hat{G}}} prob_{j,i}}{k}$$

Baselines & Results

Diffusion network inference

- $NetInf$
- $stNetInf$
- $MultiTree$

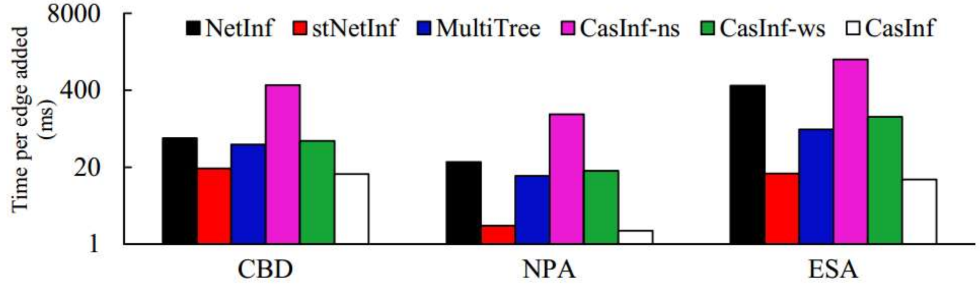
Frequent subgraph mining

- Frequency-Based Method (*FBM*)
- *STC-DBN*: the state-of-the-art approach

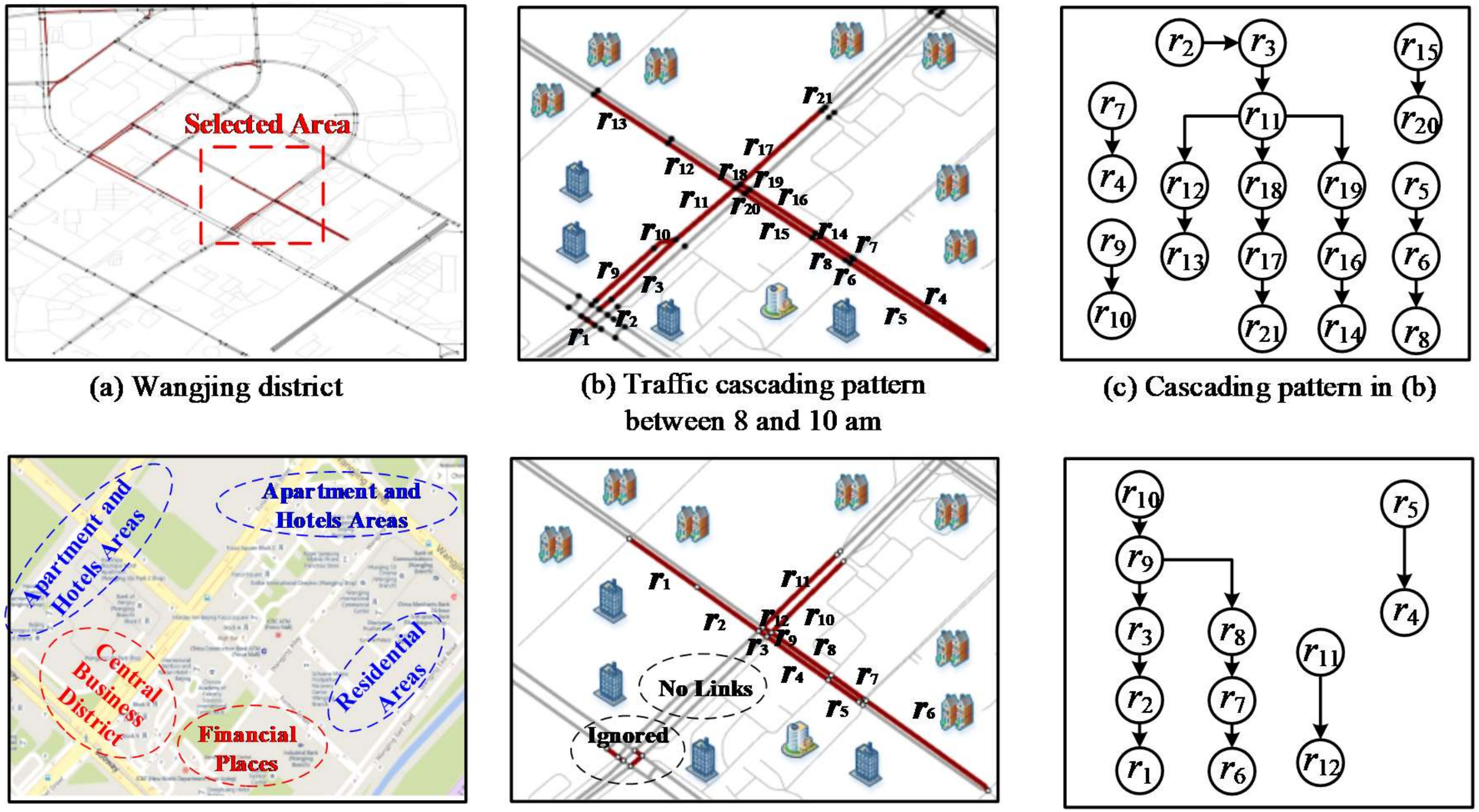
Methods	CBD	NPA	ESA	Overall
NetInf	0.270	0.119	0.116	0.168
stNetInf	0.308	0.201	0.394	0.301
MultiTree	0.311	0.140	0.141	0.197
FBM	0.287	0.193	0.171	0.217
STC-DBN	0.307	0.198	0.225	0.243
CasInf-gd	0.359	0.258	0.488	0.368
CasInf-td	0.336	0.199	0.203	0.246
CasInf-ne	0.363	0.298	0.515	0.392
CasInf-ni	0.197	0.129	0.215	0.180
CasInf	**0.384**	**0.317**	**0.545**	**0.415**

Variant of $CasInf$

- $CasInf$-*gd*
- $CasInf$-*td*
- $CasInf$-*ne*
- $CasInf$-*ni*



Case Study



(a) Wangjing district

(b) Traffic cascading pattern between 8 and 10 am

(c) Cascading pattern in (b)

(d) Selected area

(e) Traffic cascading pattern between 7 and 9 pm

(f) Cascading pattern in (e)

Conclusion

- We propose a data-driven approach to infer the traffic cascading patterns from multiple spatio-temporal datasets
- More precisely, we propose *ITL* model, *EMT* model and cascading pattern construction model for the cascading pattern inference

Thanks

Yuxuan Liang
v-yuxuli@microsoft.com

Released Codes & Papers