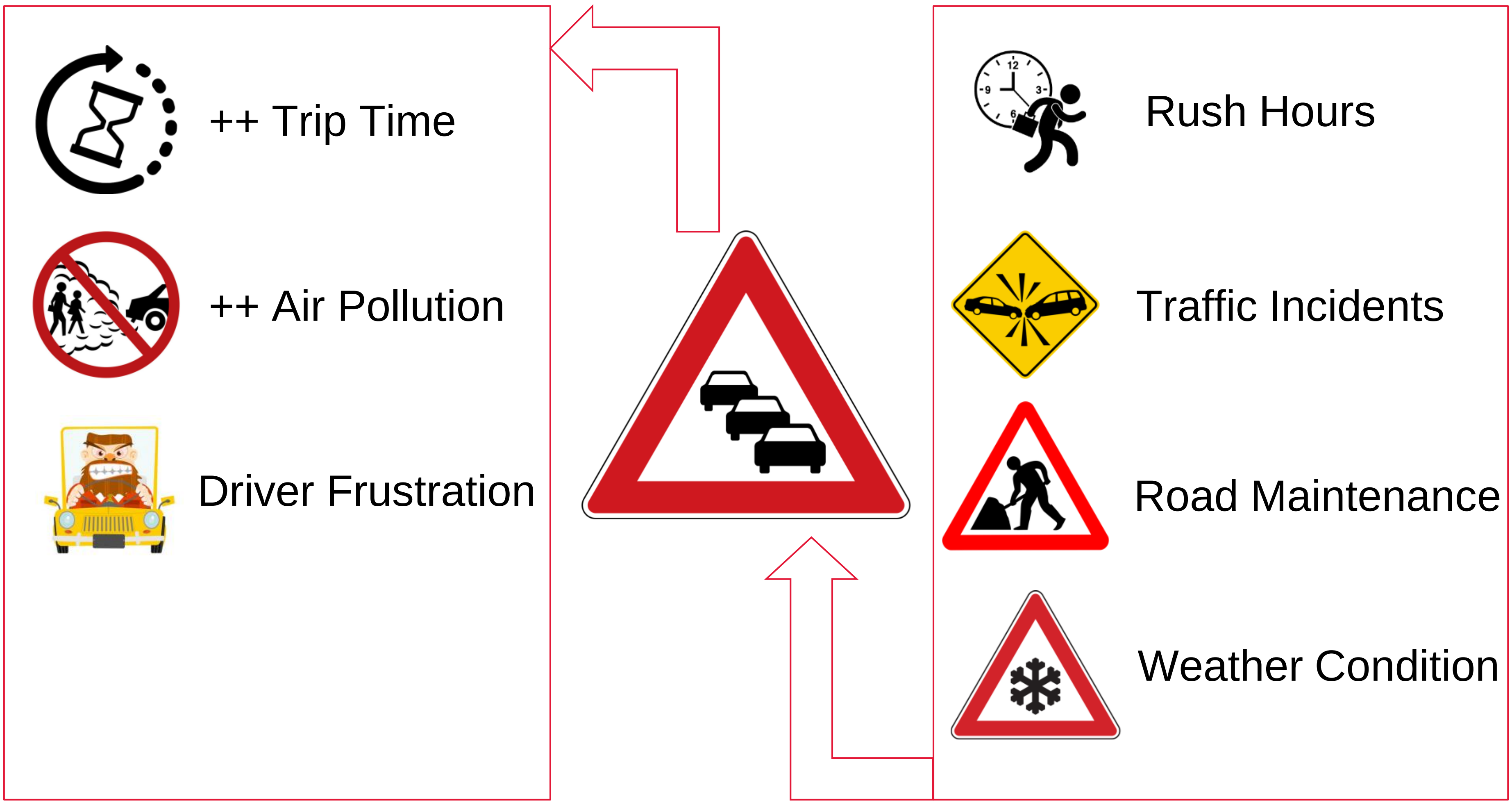
Network-aware Multi-agent Reinforcement Learning for Adaptive Navigation of Vehicles in a Dynamic Road Network

M.A.Sc. Thesis of Fazel Arasteh
York University, Toronto, Canada

Motivation

Traffic Congestion

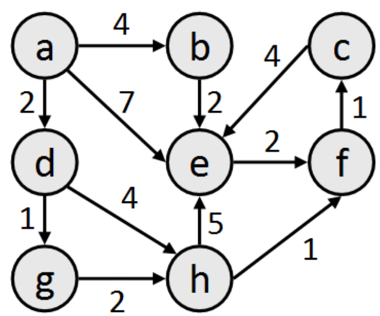


Expensive Solution: Construct More Roads

Economical Solution: Algorithmic Solution

Base Method: Shortest Path First Algorithm (SPF)

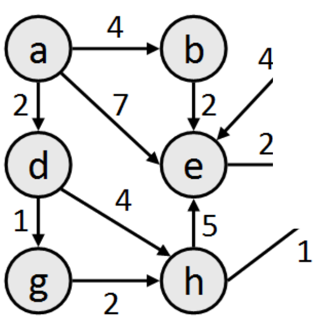
SPF



Static Network: Constant Travel Times

No Traffic Jam

SPF

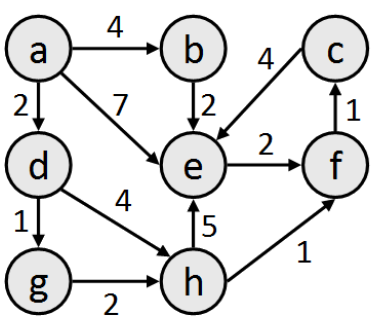


Optimal

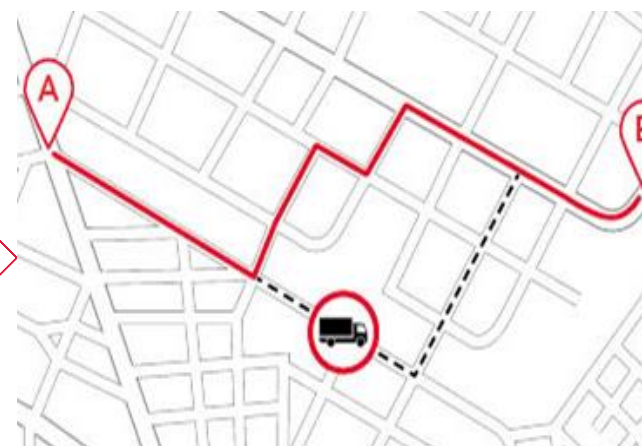
Dynamic Network: Changing Travel Times

Traffic Jam

SPF

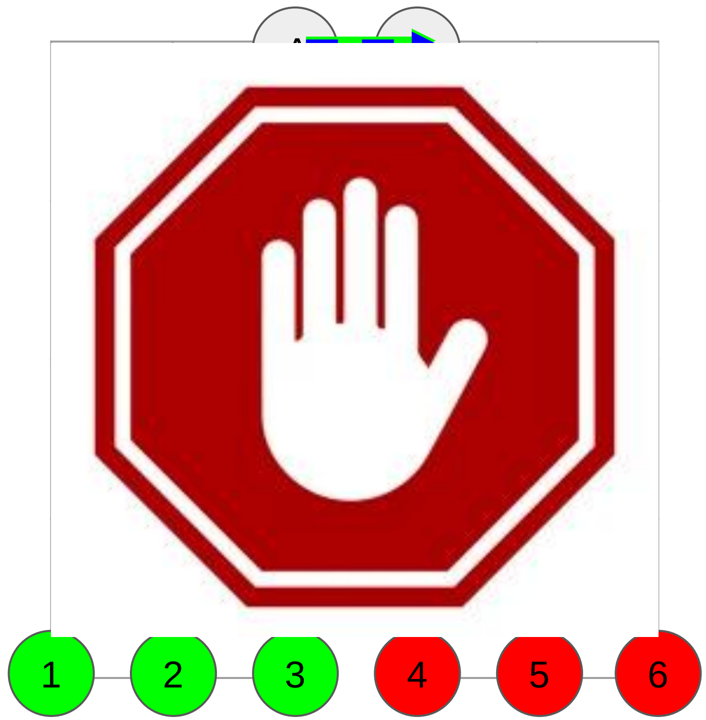


Travel Time Prediction



Limitation 1: Inaccurate Long-term Travel Time Predictions

Limitation 2: Greedy SPF! No Collaboration!



Vehicle Navigation Problem

Traffic Jam

Fleet of Vehicles

Improve Travel Time

Collaboration

Problem Definition

Vehicle Navigation Problem

Given:

Road network of the controlled area:

$$W=\{R,I\}$$

State of road network *W* at time *t* which is the *expected travel time* in every road at time *t:*

$$S_W^t$$

$$E(\text{Travel-Time}(r)) \mid r \in R$$

A set of *origin-destination trips* with *time-label* τ indicating when the trip starts:

$$Trips=\{(o, d, \tau) \mid o \in R,\ d \in I\}$$

Vehicle Navigation Problem

Task:

Generate a path for each trip:

$$Path(trip) \mid trip \in Trips$$

Objective:

Minimizing the average travel time for all the trips:

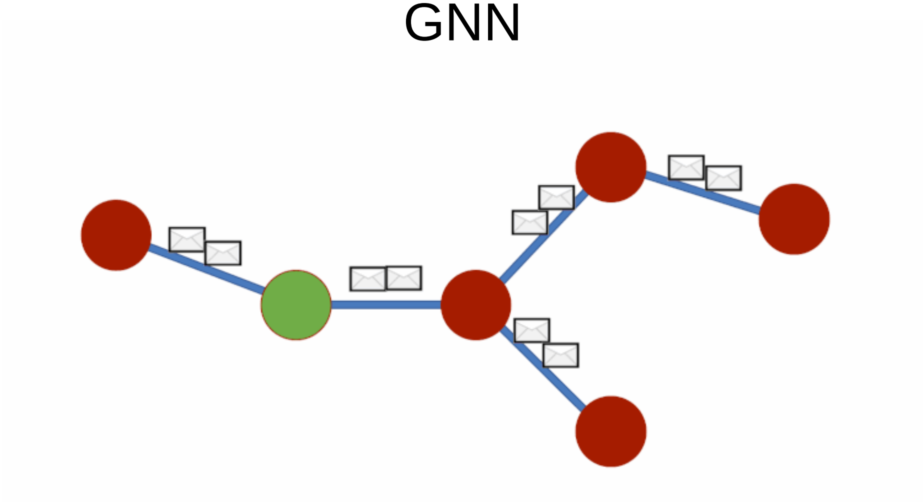
$$\tau' = \textit{finish time of trip } (o, d, \tau)$$
$$\textit{Travel-Time}(trip) = \tau' - \tau$$

$$\boldsymbol{AVTT = \sum_{trip \in Trips} \textit{Travel-Time}(trip) / |Trips|}$$

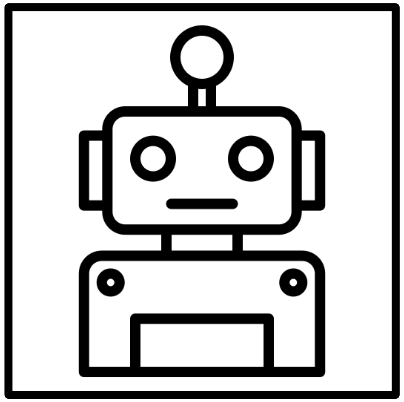
Minimize AVTT

Method: GNN + RL

GNN

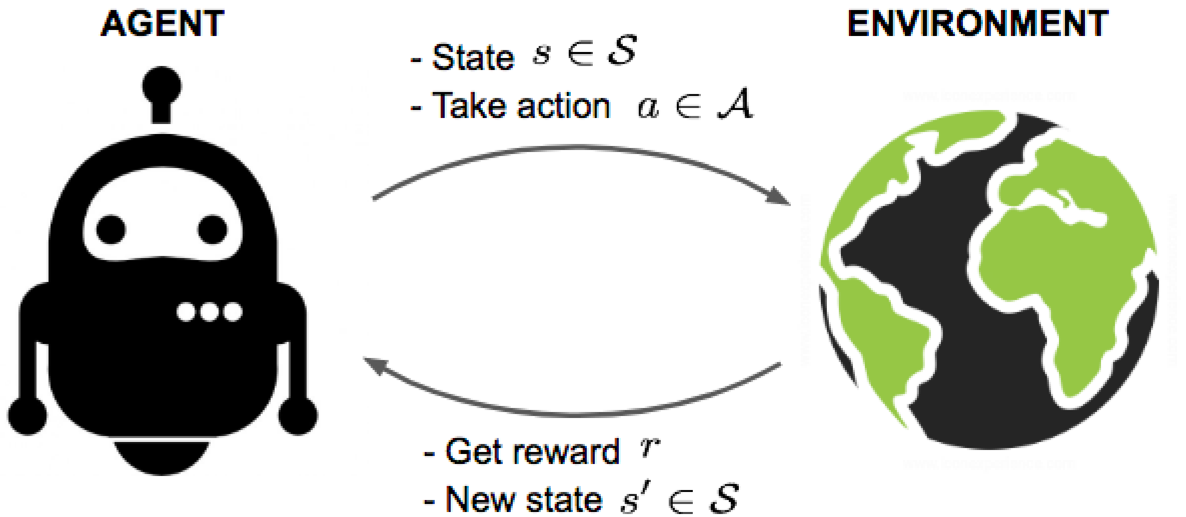


RL

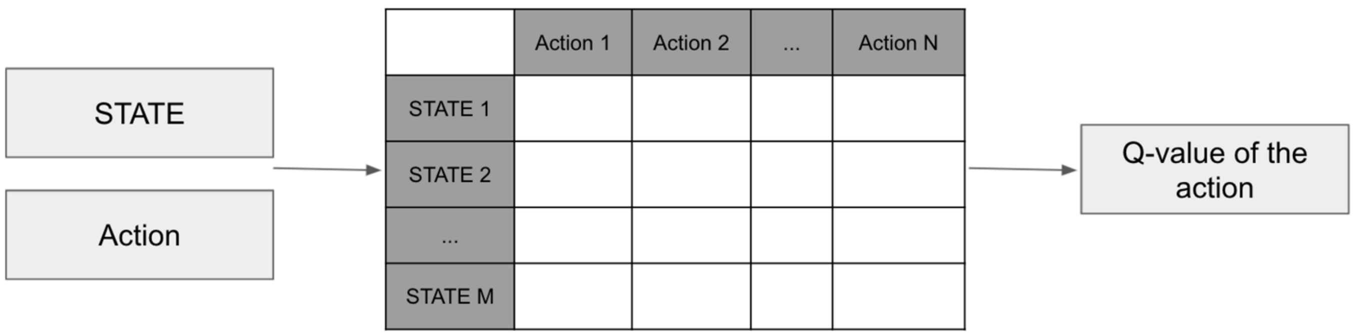


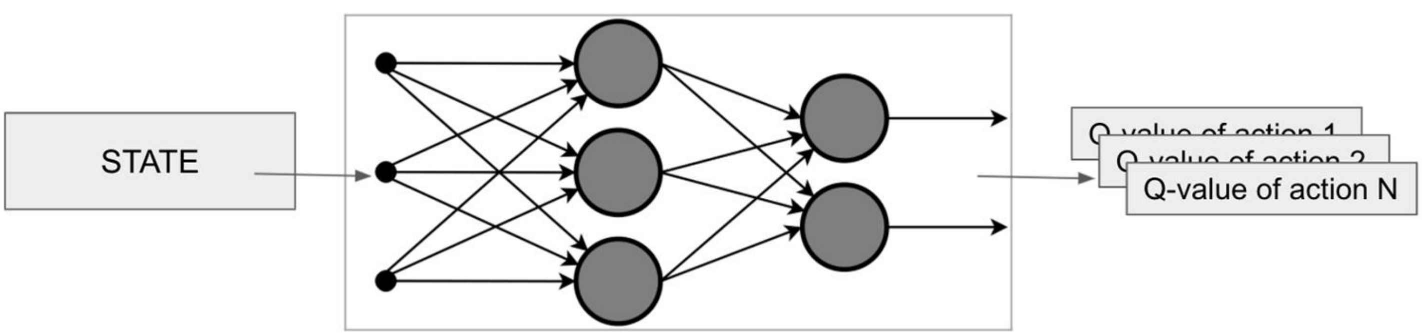
Background

Reinforcement Learning



Q-learning, DQN

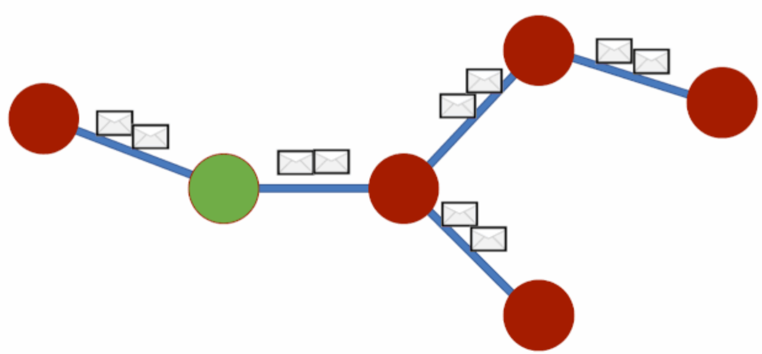




Graph Neural Networks

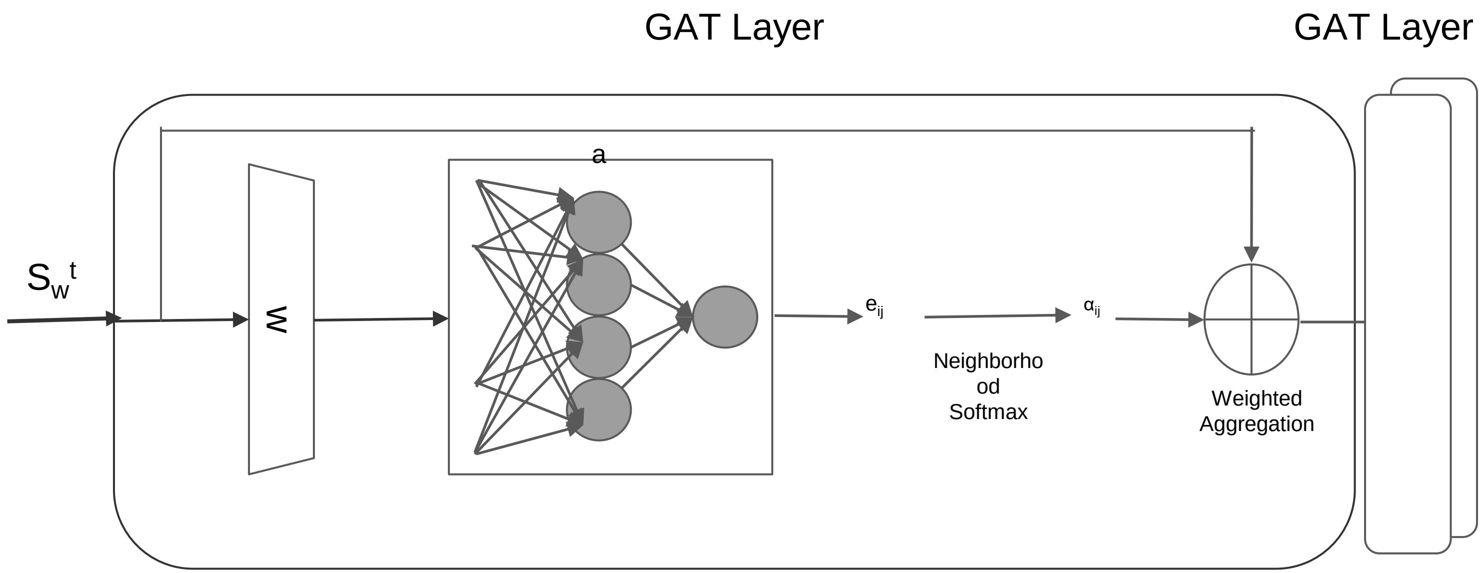
0	1	0	0
0	0	0	0
0	1	0	1
0	0	0	0
0	0	1	0
1	1	0	0

1	1	0	1	0	0
1	1	1	0	1	0
0	1	1	0	0	1
1	0	0	1	1	0
0	1	0	1	1	1
0	0	0	0	1	1



0.1	10	0	0.5	2	14	0.9	0
5	2	8	4	3	1	0	0
0	1	0	0.4	0	1	0	0
0	10	0	0	0	1	0	0
0	1	9	0	0	1	0	0
0	1	0	5.5	0	1	3.5	0

Graph Attention Networks



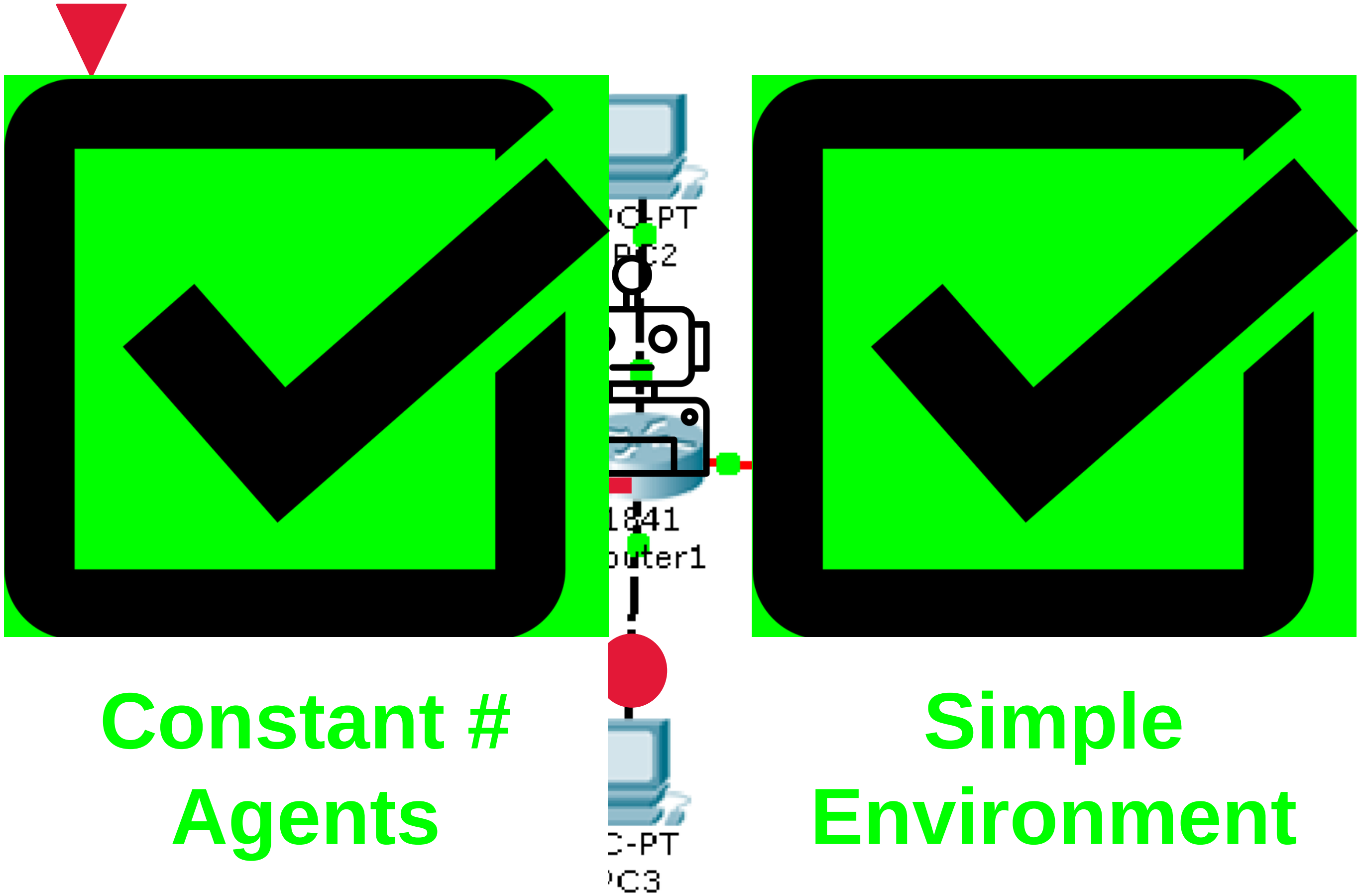
Methodology

Vehicles as Agents

Too Many Agents!

Complex Environment!

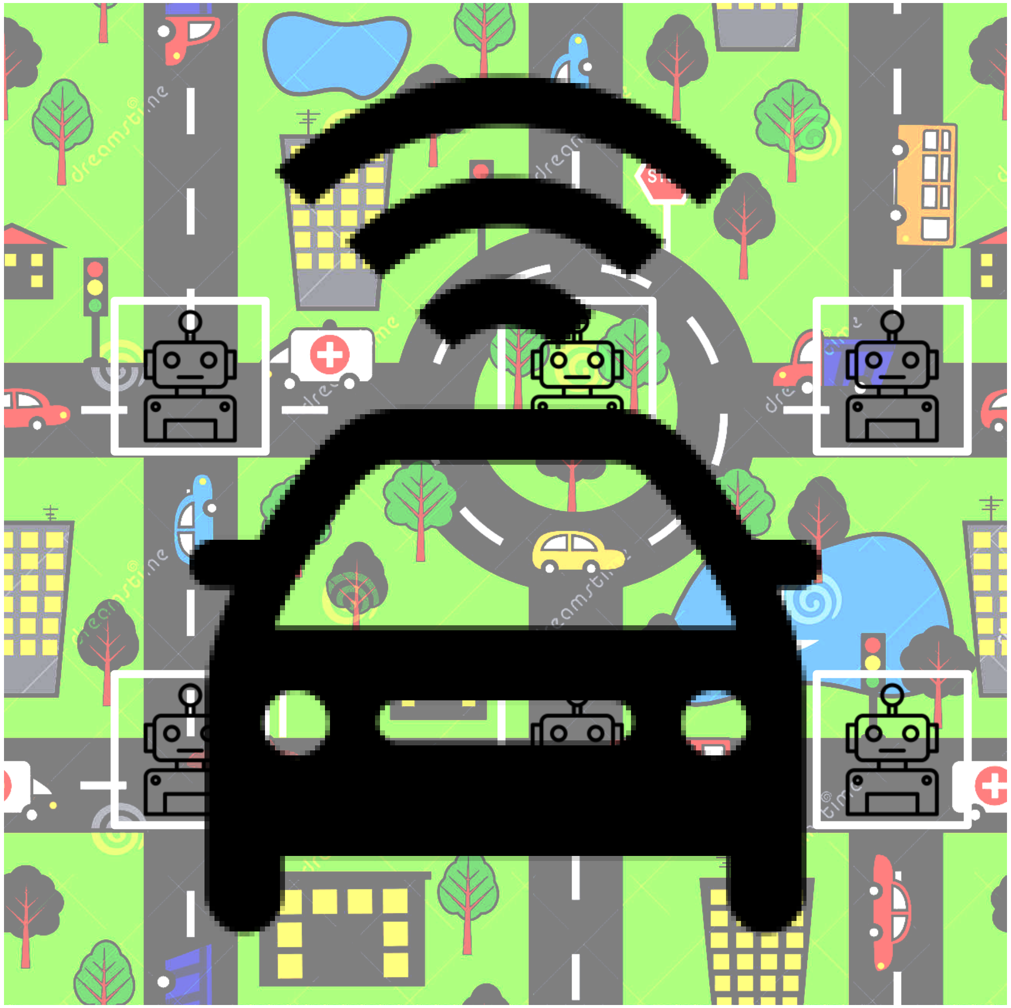
Packet Routing Problem



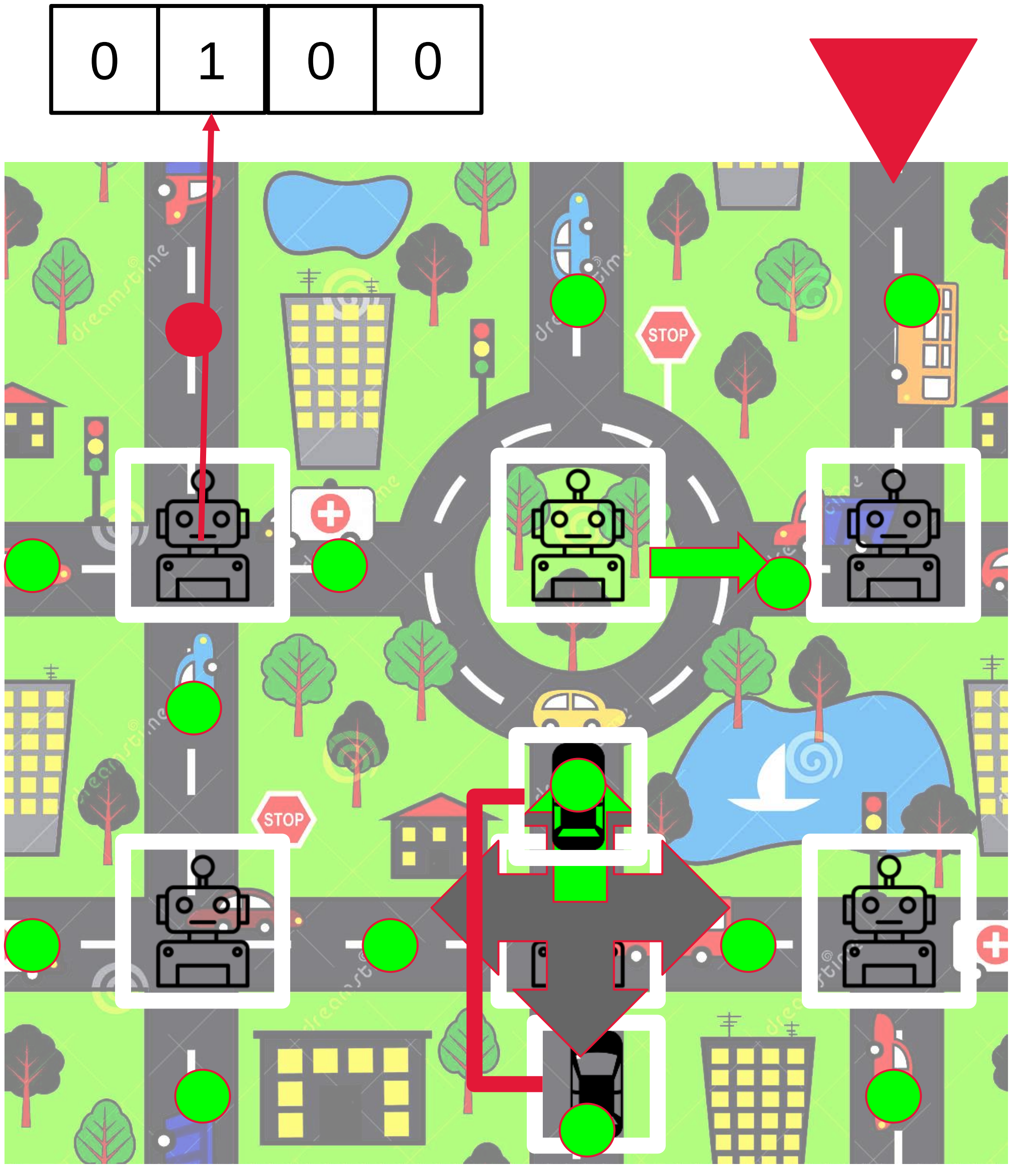
Constant # Agents

Simple Environment

Road Network - IP Network Analogy

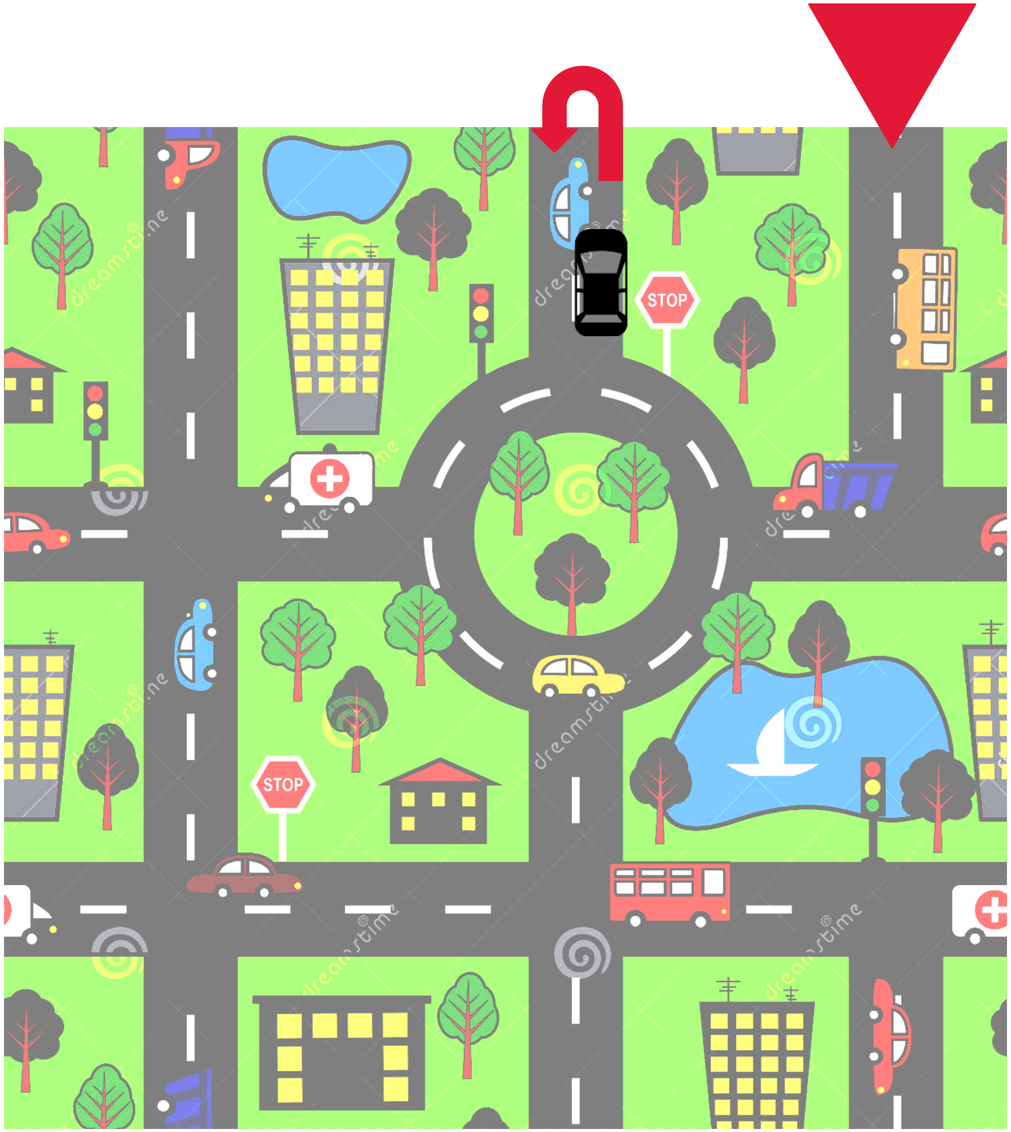


MARL Formulation

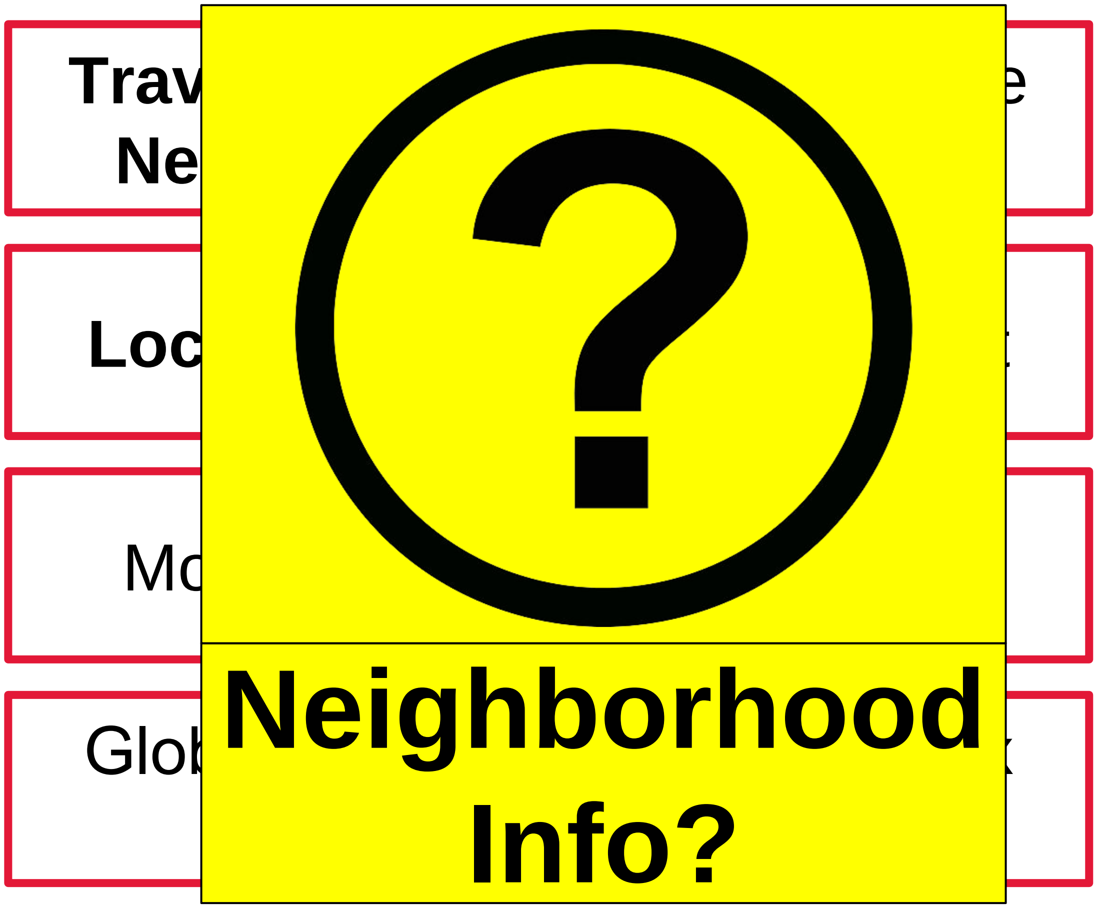


Agents	Intersection
State	Agent ID+ S_W + D
Action	Next Road
Next State	Next Agent ID+ S_W + D
Reward	-△T

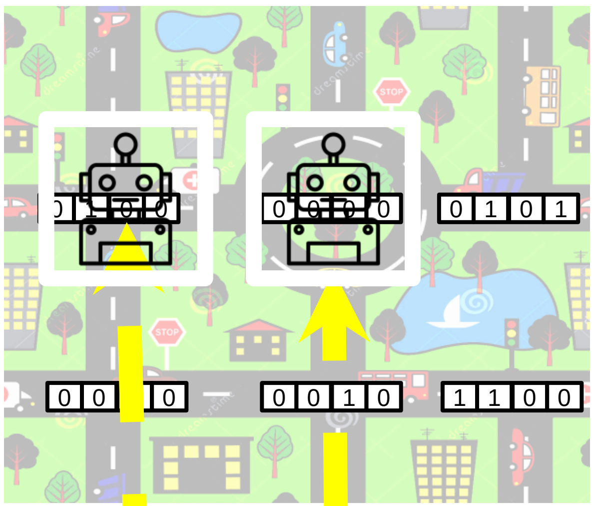
Closed Controlled Area



Challenge 1: Huge Irrelevant Network State



Solution: Graph Attention Networks



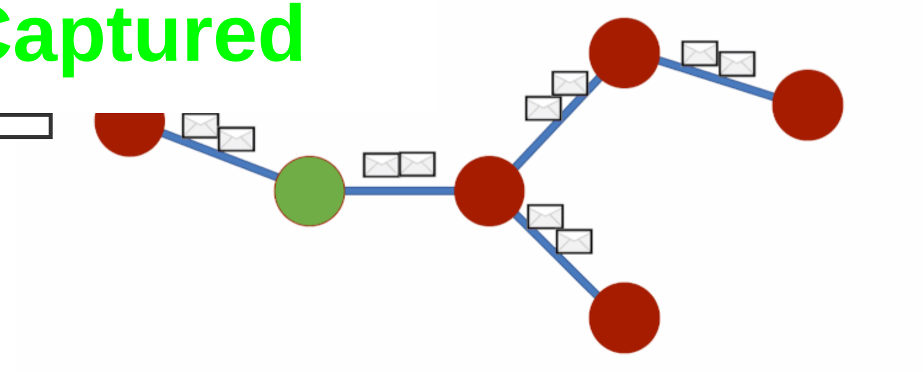
S_W

A

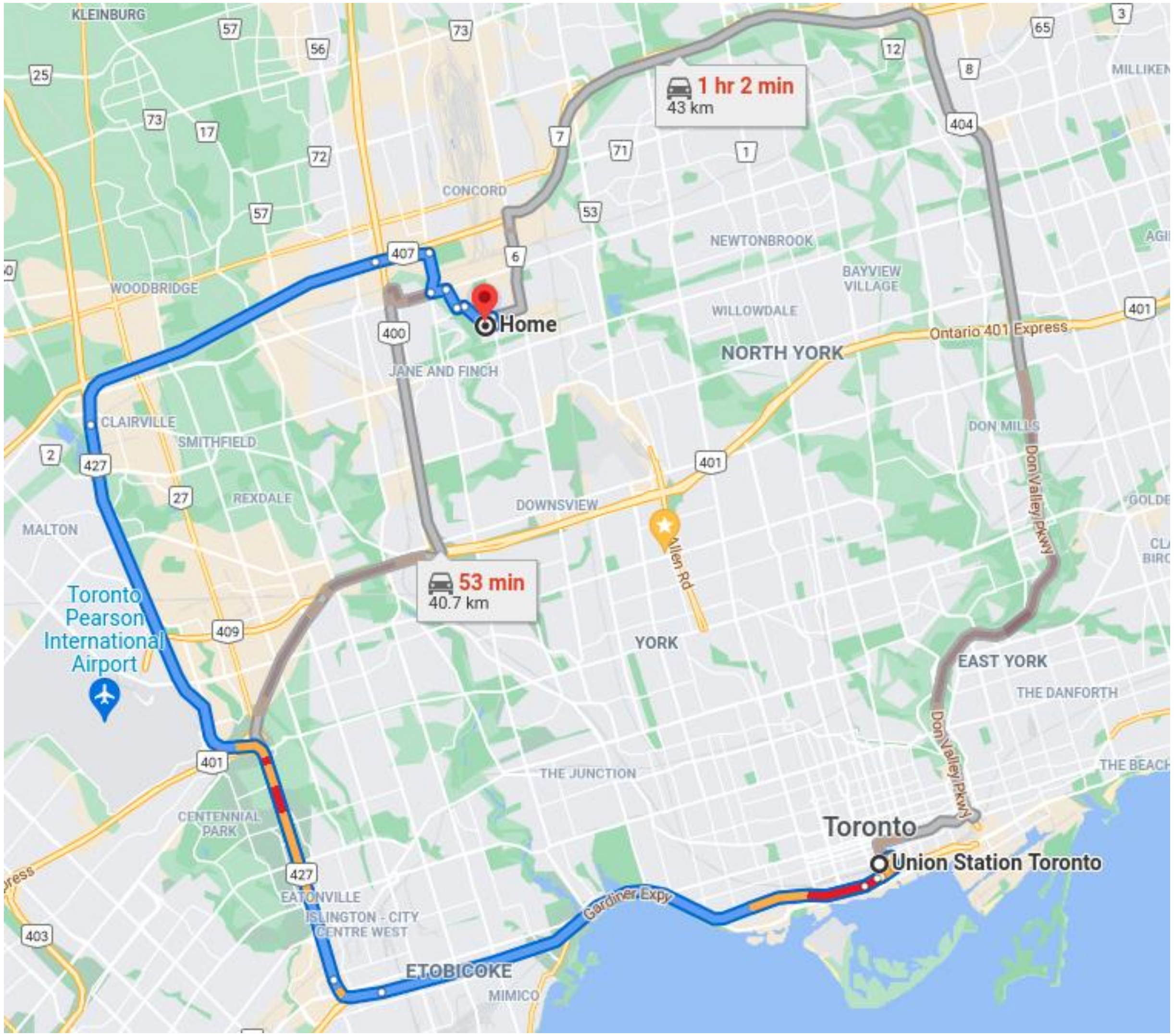
1	0	1	0	0
1	1	0	1	0
1	1	0	0	1
0	0	1	1	0
1	0	1	1	1
0	0	0	1	1

0.1	[illegible]	0	0.5	2	14	0.9	0
5	2	8	4	[illegible]	1	0	0
0	1	0	0.4	0	1	0	0
0	10	0	0	0	1	0	0
0	1	9	0	0	1	0	0
0	1	0	5.5	0	1	3.5	0

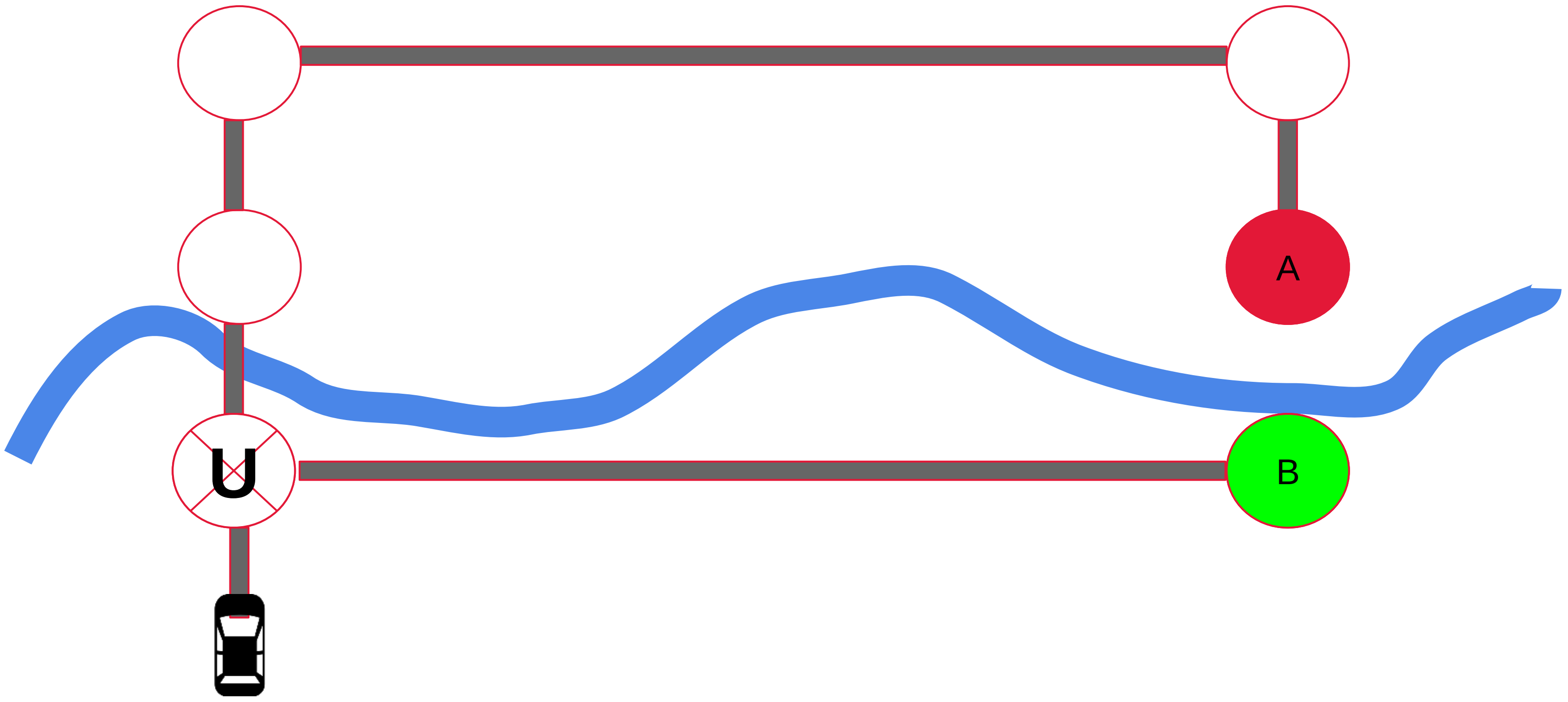
Neighborhood Captured



Locality of Access: Intuition for Routing



Exception: Disconnected Near Intersections!



Challenge 2: Intersection IDs

Normalized Co One-hot

x=-0.12		0	0	0	0	0	0	0	0

Preserves Loc

the Locality of Access

2 dimens

dimensional data

Hard for NN to separate→
No Convergence

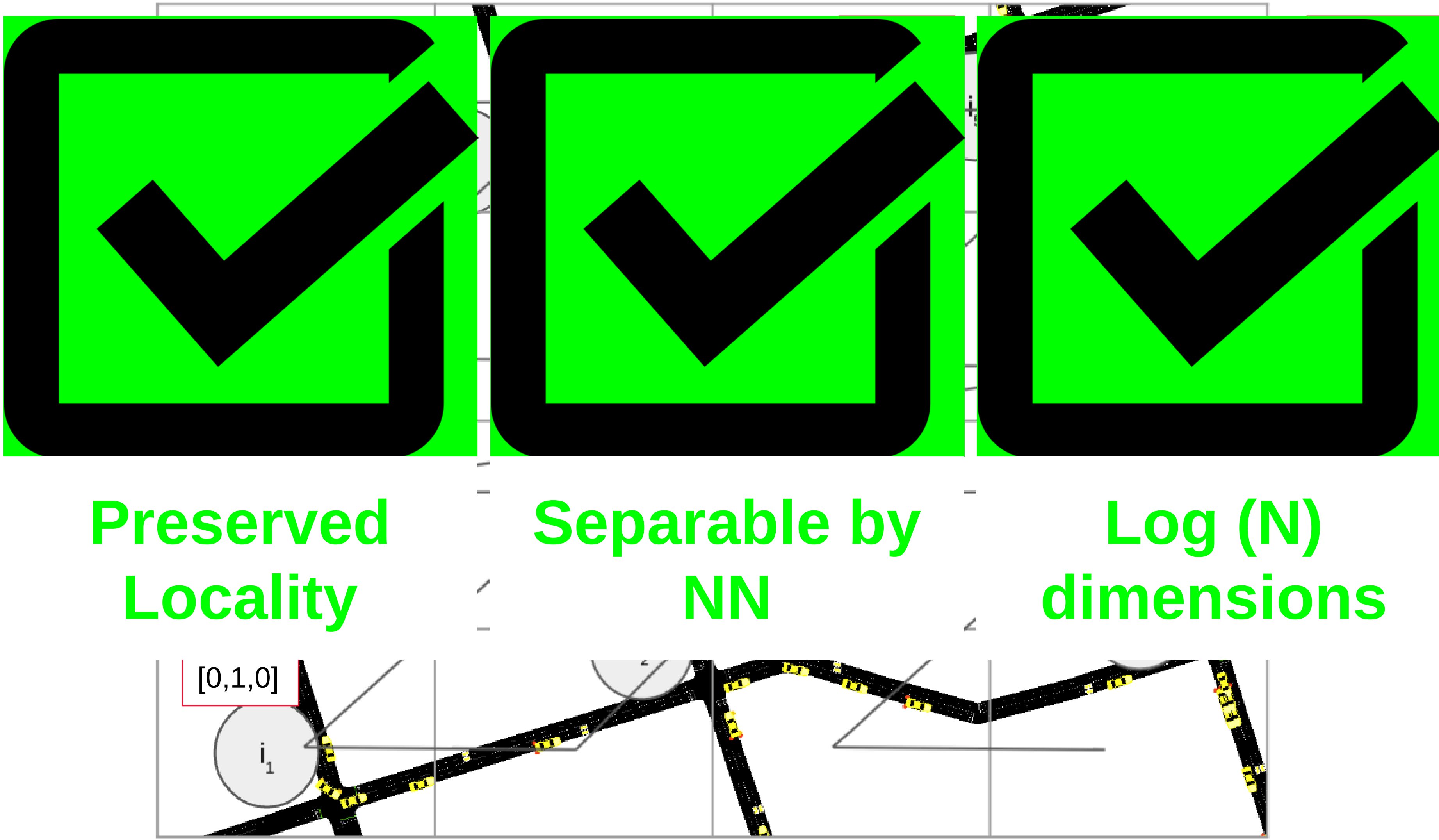
Easy for NN to separate

How to have both?

Solution: Space Filtering (e.g. Z-Order)

	x: 0 000	1 001	2 010	3 011	4 100	5 101	6 110	7 111
y: 0 000	000000	000001	000100	000101	010000	010001	010100	010101
1 001	000010	000011	000110	000111	010010	010011	010110	010111
2 010	001000	001001	001100	001101	011000	011001	011100	011101
3 011	001010	001011	001110	001111	011010	011011	011110	011111
4 100	100000	100001	100100	100101	110000	110001	110100	110101
5 101	100010	100011	100110	100111	110010	110011	110110	110111
6 110	101000	101001	101100	101101	111000	111001	111100	111101
7 111	101010	101011	101110	101111	111010	111011	111110	111111

Solution: Space Filtering (e.g. Z-Order)

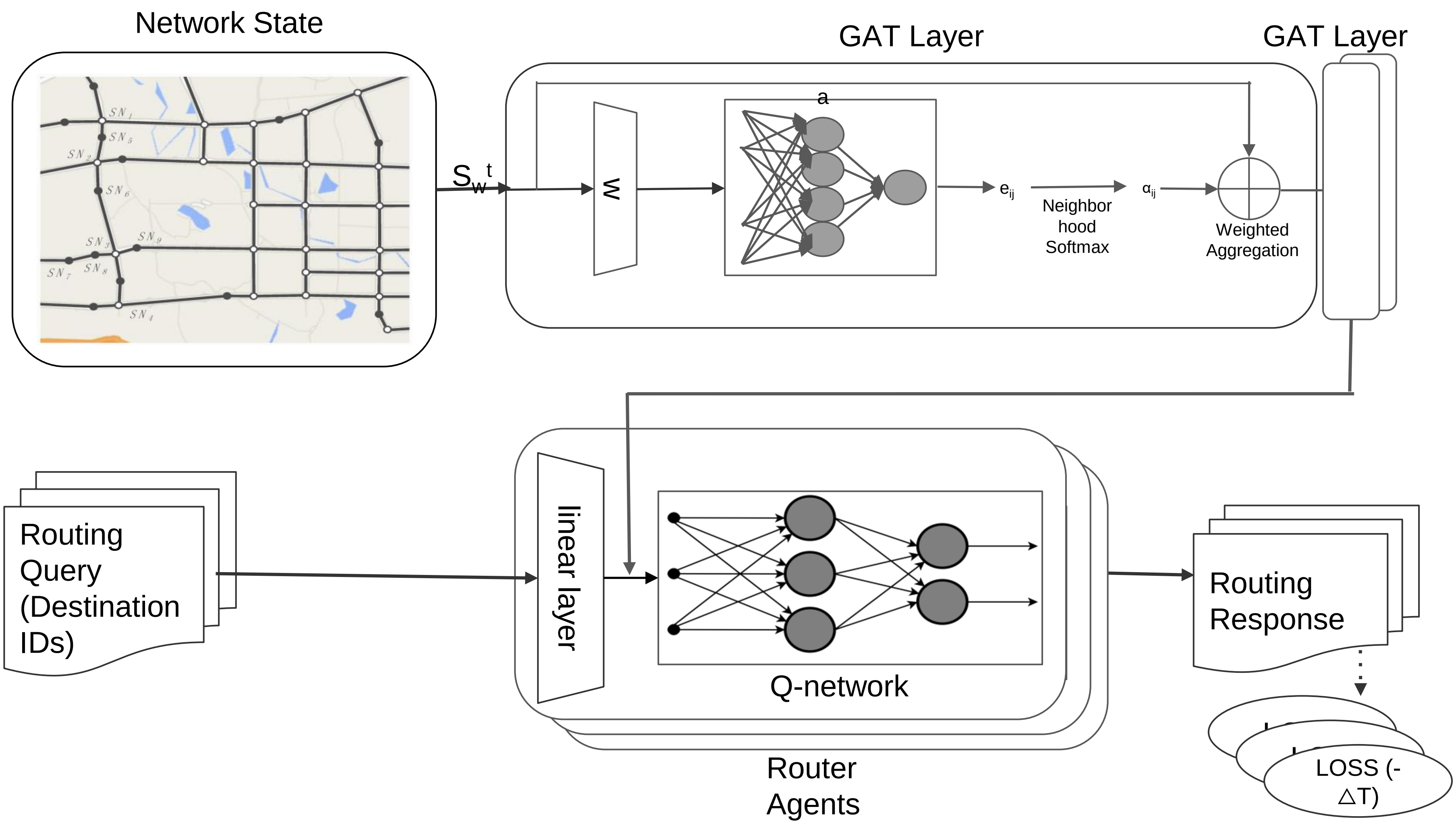


Preserved Locality

Separable by NN

Log (N) dimensions

Model Architecture



Reward Function Justification: End2End travel time prediction

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\,(R_{t+1}+\gamma \max_{a\in A_{t+1}} Q(S_{t+1},a)-Q(S_t,A_t))$$

$\gamma=1,\ \alpha=1$

$$Q(S_t,A_t) \leftarrow R_{t+1}+ \max_{a\in A_{t+1}} Q(S_{t+1},a)$$

S_{t+N} terminal

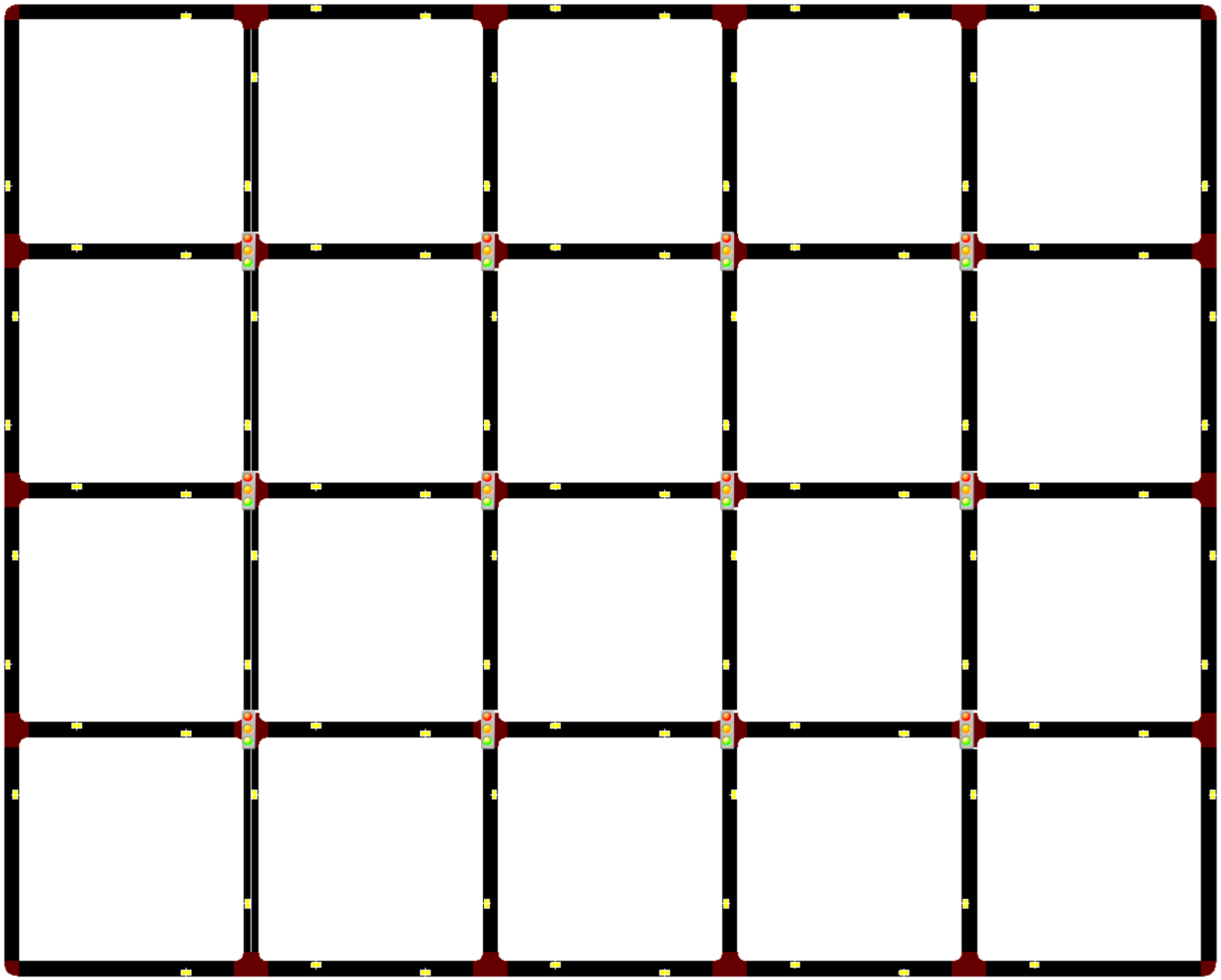
$$Q(S_t,A_t) \leftarrow R_{t+1}+R_{t+2}+\ldots+R_{t+N}$$

$r= -\Delta T$

$$Q(S_t,A_t) \leftarrow -\Delta T_1- \Delta T_2 - \ldots - \Delta T_N = -\text{ Travel Time}$$

Experimental Evaluation

Datasets: Grid Network



Datasets: Downtown Toronto

Datasets: Traffic Demand

Uniform Demand

Biased Demand

Dynamics Simulation

Traffic Jam

traffic-state-change-period

congestion-epsilon

congestion-speed-factor

traffic-state-change-period

Base Lines, Algorithm Versions

Travel Time Shortest Path First (SPF)

Travel Time Shortest Path First with Rerouting (SPFWR)

Q-routing

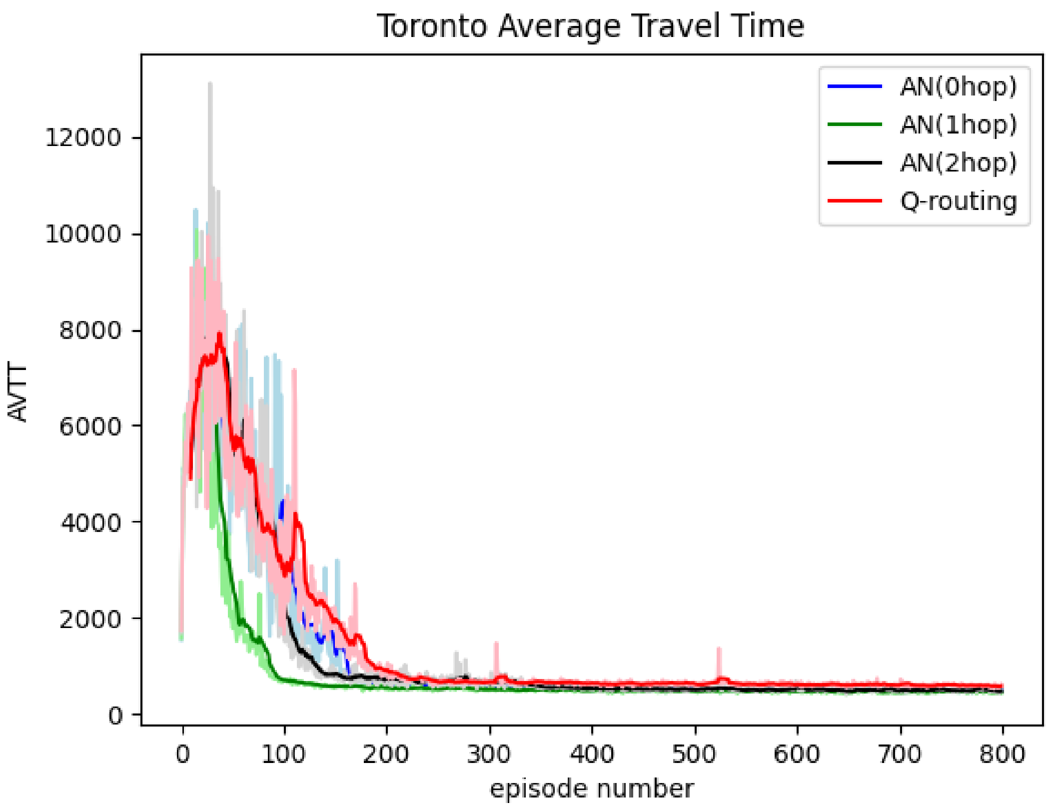
AN(0hop), AN(1hop), AN(2hop)

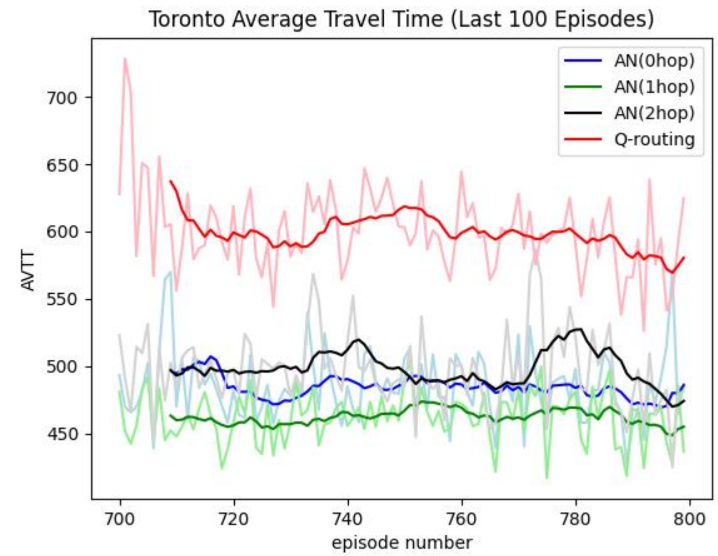
Evaluation Metric

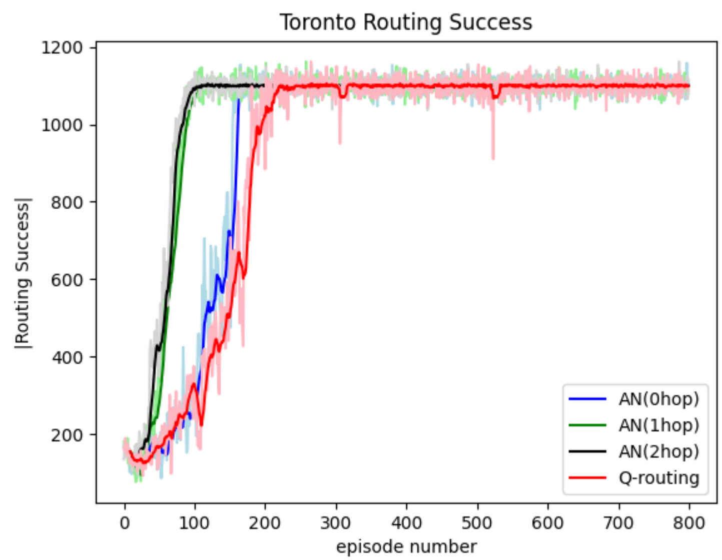
Average Travel Time (AVTT)

Routing Success (RS)

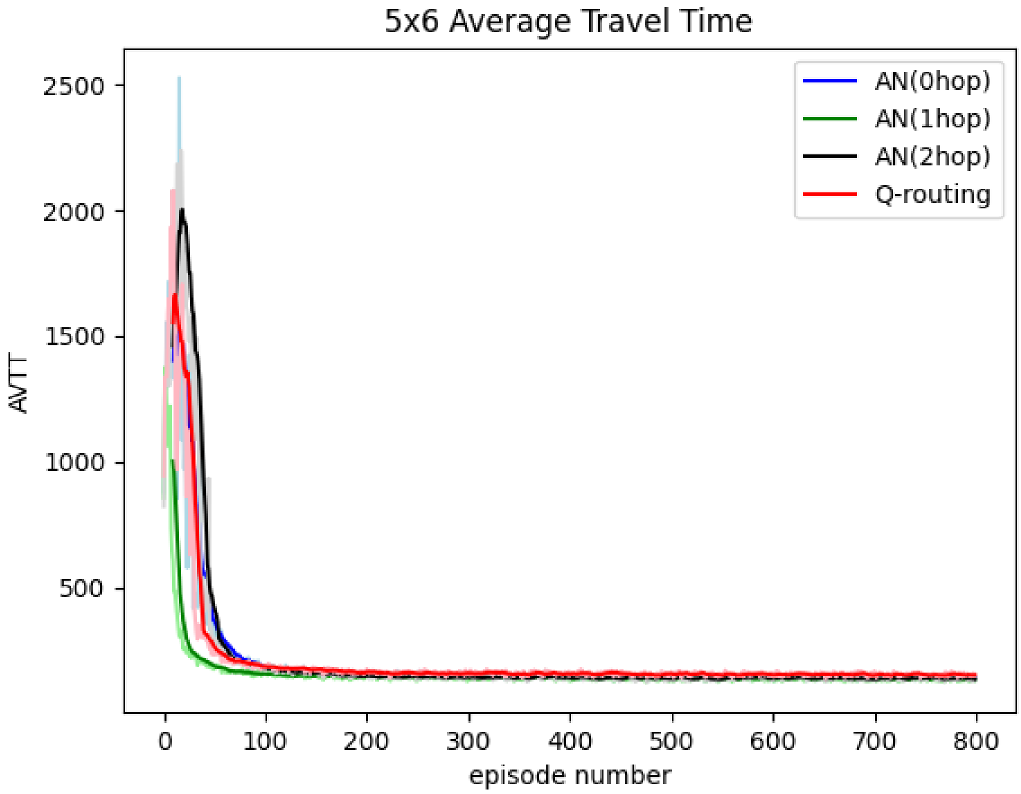
Performance Evaluation (Online Training)

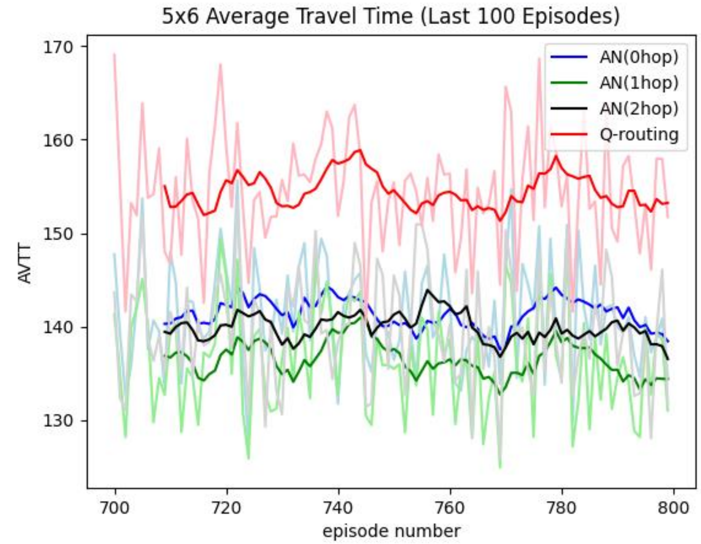


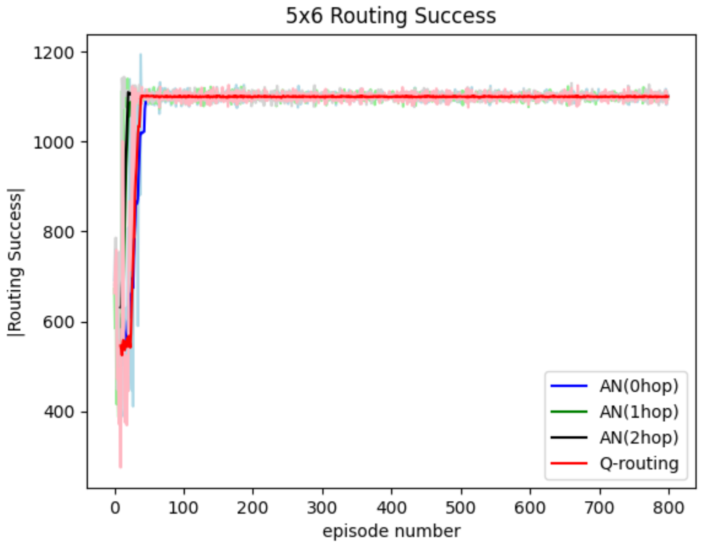




Performance Evaluation (Online Training)







Performance Evaluation (Offline Testing Settings)

2000 Uniform Trips

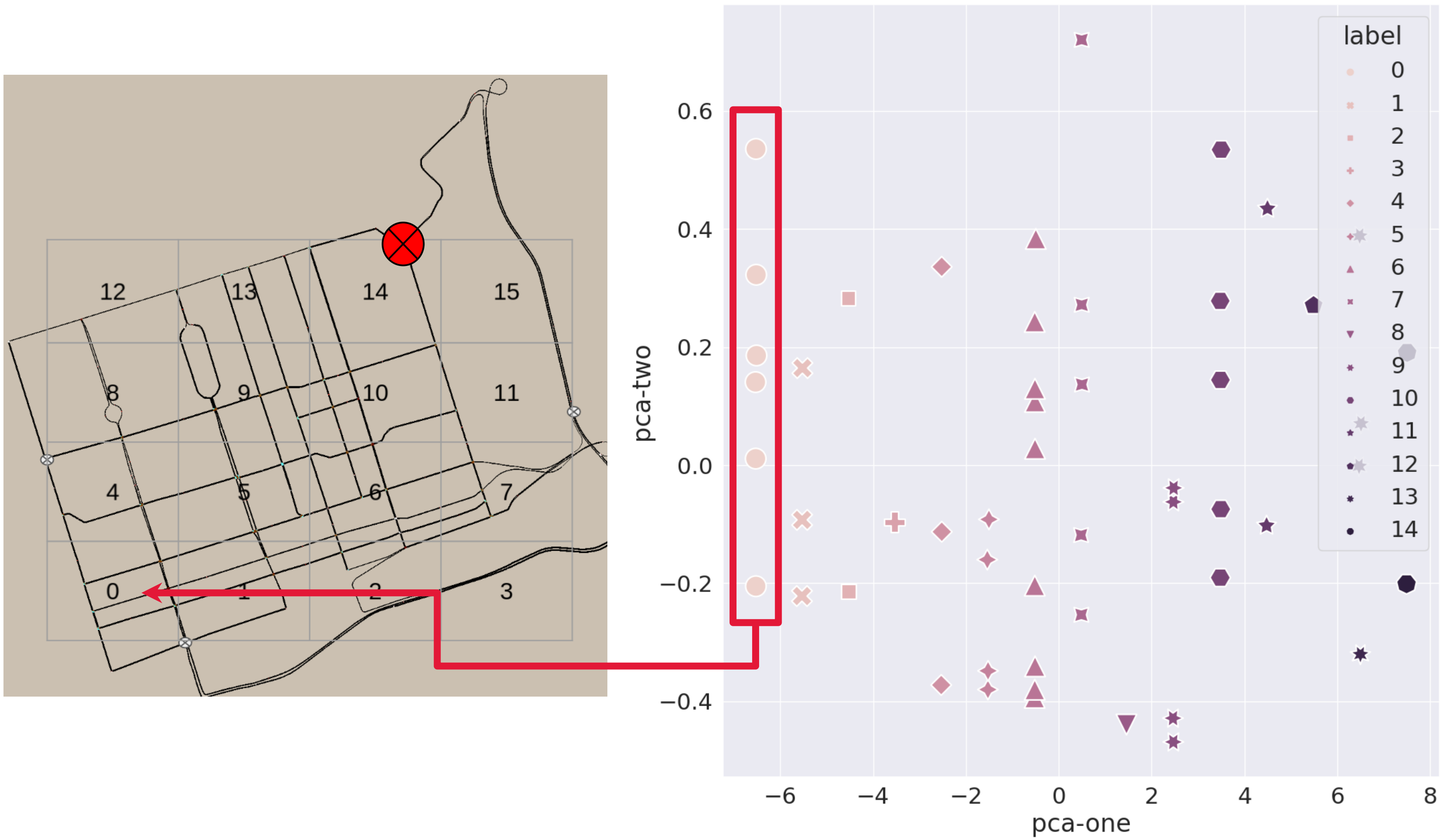
200 Biased Trips

Routing Success = 2200

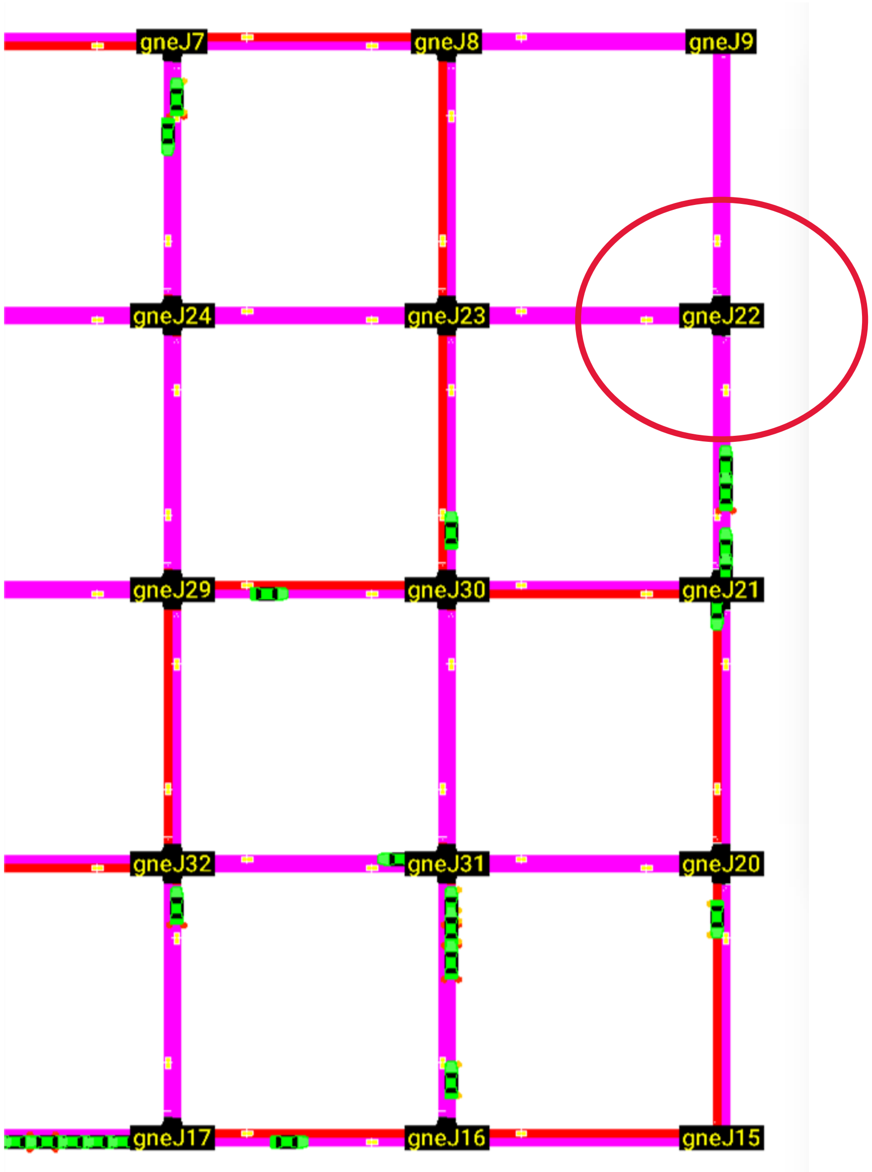
Performance Evaluation (Offline Testing Results)

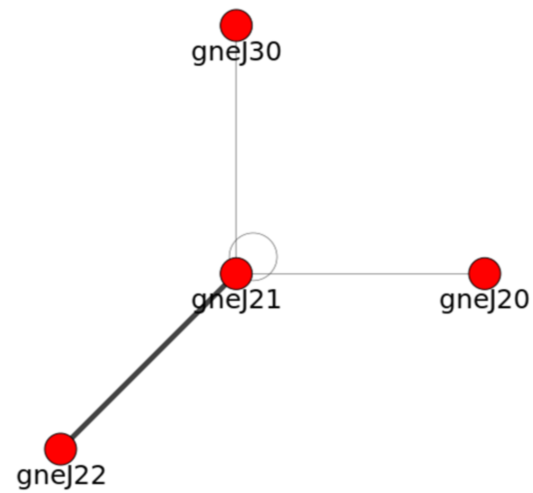
	5x6 Network	D.T Toronto
SPF	173.4	551.7
SPFWR	205.1	**475.6**
QR	159.6	∞
AN(0hop)	143.7	477.6
AN(1hop)	**138.4**	476.4
AN(2hop)	145.4	479.3

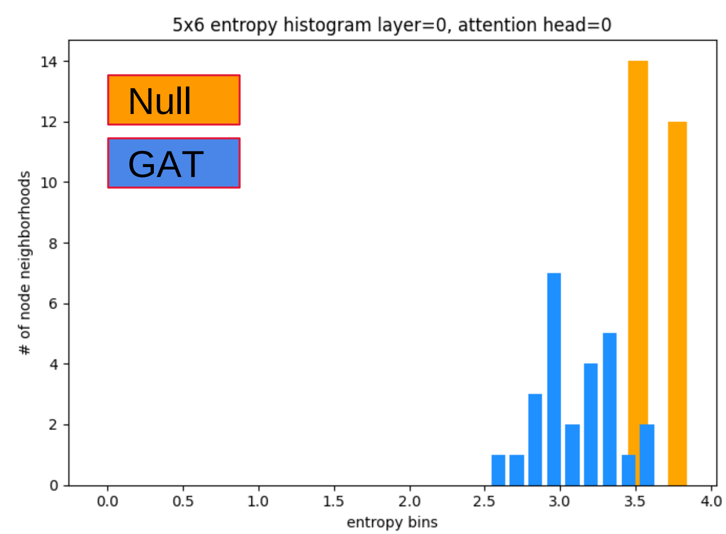
Locality of Access Evaluation



Attention Evaluation

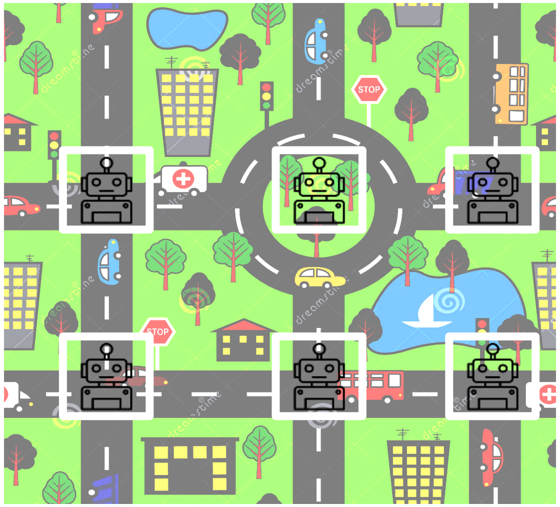


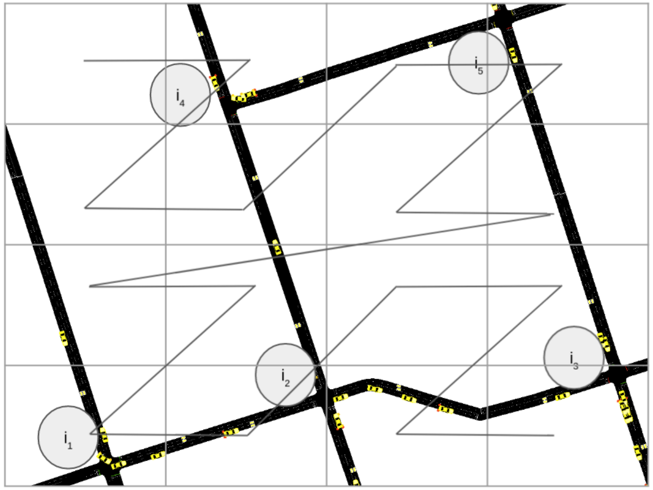




Conclusion, Limitations & Future Work

Conclusion





Limitations & Future Work

Limitations:

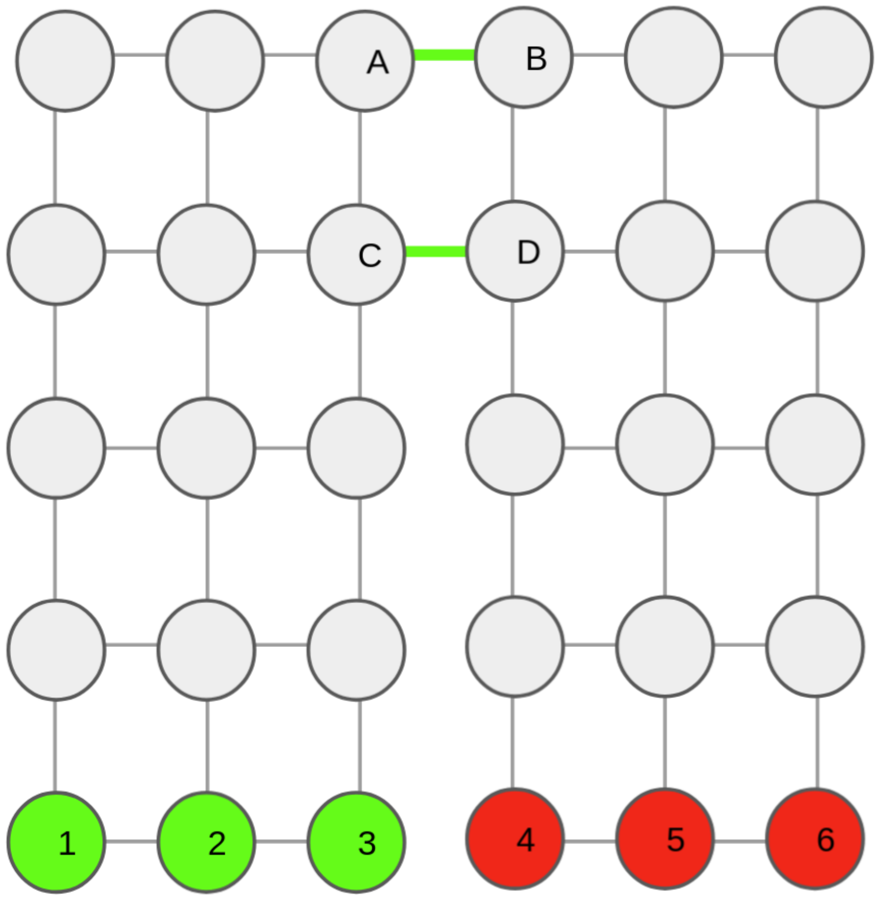
1. Reliability
2. Scalability
3. Network State Capturing

Future Work:

1. Shared Policies
2. Hierarchical Routing
3. Traffic Signal Control

Thank You

Collaborative Policies



Reward Function Justification: End2End travel time prediction

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\,(R_{t+1}+\gamma\,\max_{a\in A_{t+1}} Q(S_{t+1},a)-Q(S_t,A_t))$$

$\gamma = 1,\ \alpha = 1$

$$Q(S_t,A_t) \leftarrow R_{t+1} + \max_{a\in A_{t+1}} Q(S_{t+1},a)$$

S_{t+N} terminal

$$Q(S_t,A_t) \leftarrow R_{t+1}+R_{t+2}+\ldots+R_{t+N}$$

$r = -\Delta T$

$$Q(S_t,A_t) \leftarrow -\Delta T_1 - \Delta T_2 - \ldots - \Delta T_N = -\text{ Travel Time}$$