<Your Name>


# Case Study: Trails

Binxuan Huang
binxuanh@cs.cmu.edu

CASOS
Societal Computing, School of Computer Science, Carnegie Mellon University





- Introduction of trail, Markov transition network
- Case study
- Current research progress
- Hands on




<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Introduction

- Trail: a path of an object through time and space

| Time | 4 pm@Apr. 1 | 3 pm@Apr. 2 | 9 am@Apr. 3 | 1 pm@Apr. 3 | 2 pm@Apr. 4 | 4 pm@Apr. 5 |
|---|---|---|---|---|---|---|
| Trail 1 | L1 | L2 | L3 | L2 | L1 | L2 |
| Trail 2 | L2 | L3 | L4 | L2 | L1 | L1 |
| Trail 3 | L2 | L3 | L1 | L1 | L2 | L3 |

| | L1 | L2 | L3 | L4 |
|---|---|---|---|---|
| L1 | 2 | 3 | 0 | 0 |
| L2 | 2 | 0 | 4 | 0 |
| L3 | 1 | 1 | 0 | 1 |
| L4 | 0 | 1 | 0 | 0 |

Traffic flow network

$$P(L_i \rightarrow L_j) = \frac{N(L_i \rightarrow L_j)}{\sum_j N(L_i \rightarrow L_j)}$$

| | L1 | L2 | L3 | L4 |
|---|---|---|---|---|
| L1 | 0.4 | 0.6 | 0 | 0 |
| L2 | 0.33 | 0 | 0.67 | 0 |
| L3 | 0.33 | 0.33 | 0 | 0.33 |
| L4 | 0 | 1 | 0 | 0 |

Markov transition network

CASOS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Introduction

- Adding BEGIN/END locations to capture more information

| Time | 4 pm@Apr. 1 | 4 pm@Apr. 1 | 3 pm@Apr. 2 | 9 am@Apr. 3 | 1 pm@Apr. 3 | 2 pm@Apr. 4 | 4 pm@Apr. 5 | 4 pm@Apr. 5 |
|---|---|---|---|---|---|---|---|---|
| Trail 1 | BEGIN | L1 | L2 | L3 | L2 | L1 | L2 | END |
| Trail 2 | BEGIN | L2 | L3 | L4 | L2 | L1 | L1 | END |
| Trail 3 | BEGIN | L2 | L3 | L1 | L1 | L2 | L3 | END |

Traffic flow network

| | B | L1 | L2 | L3 | L4 | E |
|---|---|---|---|---|---|---|
| B | 0 | 1 | 2 | 0 | 0 | 0 |
| L1 | 0 | 2 | 3 | 0 | 0 | 1 |
| L2 | 0 | 2 | 0 | 4 | 0 | 1 |
| L3 | 0 | 1 | 1 | 0 | 1 | 1 |
| L4 | 0 | 0 | 1 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 |

Markov transition network

| | B | L1 | L2 | L3 | L4 | E |
|---|---|---|---|---|---|---|
| B | 0 | 0.33 | 0.67 | 0 | 0 | 0 |
| L1 | 0 | 0.4 | 0.6 | 0 | 0 | 1 |
| L2 | 0 | 0.29 | 0 | 0.57 | 0 | 0.14 |
| L3 | 0 | 0.25 | 0.25 | 0 | 0.25 | 0.25 |
| L4 | 0 | 0 | 1 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 |

CASOS



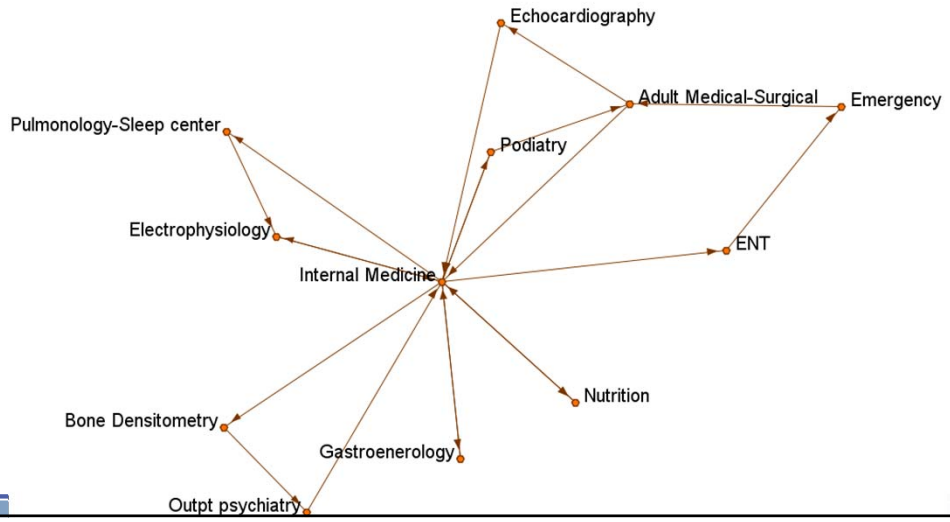
CASOS

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Case study

- Case study: one patient health trail
-



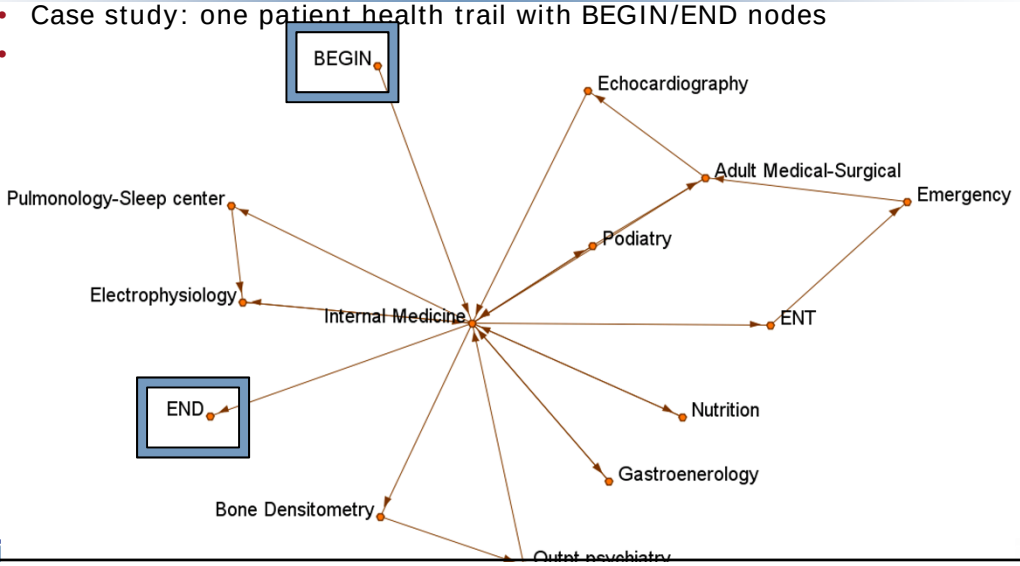

CASOS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Case study

- Case study: one patient health trail with BEGIN/END nodes
-




CASOS



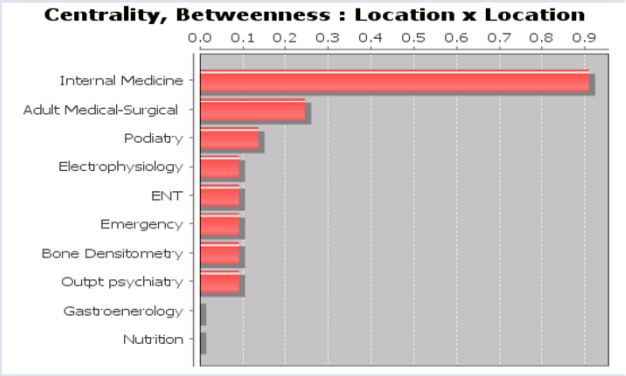
CASOS

<Your Name>

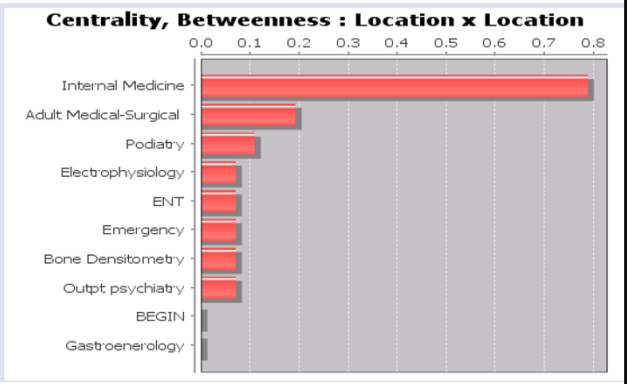
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Case study

- Compare between one patient health trail with/without BEGIN/END
- Without BEGIN/END




With BEGIN/END




CASOS

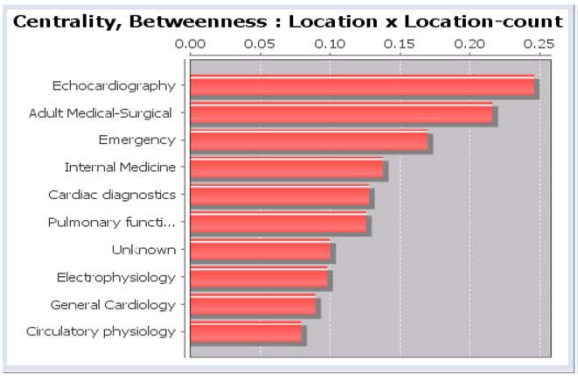


Carnegie Mellon
isr institute for SOFTWARE RESEARCH

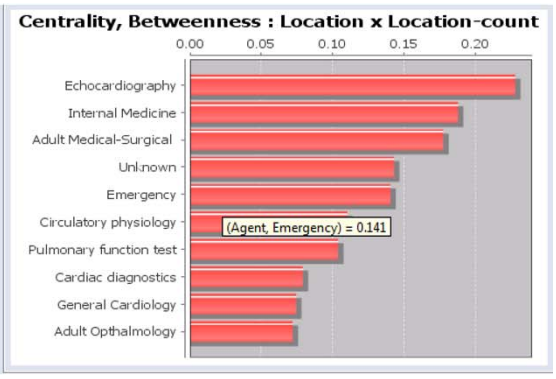
# Case study

- Compare between group of health trails with/without BEGIN/END

Without BEGIN/END




With BEGIN/END




CASOS



CASOS

<Your Name>

Carnegie Mellon isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

| Time | 4 pm@Apr. 1 | 3 pm@Apr. 2 | 9 am@Apr. 3 | 1 pm@Apr. 3 | 2 pm@Apr. 4 | 4 pm@Apr. 5 |
|---|---|---|---|---|---|---|
| Trail 1 | L1 | L2 | L3 | L2 | L1 | L2 |
| Trail 2 | L2 | L3 | L4 | L2 | L1 | L1 |
| Trail 3 | L2 | L3 | L1 | L1 | L2 | L3 |

Lower time resolution ↓ Broken point

| Time | Apr. 1 | Apr. 2 | Apr. 3 | Apr. 3 | Apr. 4 | Apr. 5 |
|---|---|---|---|---|---|---|
| Trail 1 | L1 | L2 | L3 | L2 | L1 | L2 |
| Trail 2 | L2 | L3 | L4 | L2 | L1 | L1 |
| Trail 3 | L2 | L3 | L1 | L1 | L2 | L3 |

CASOS

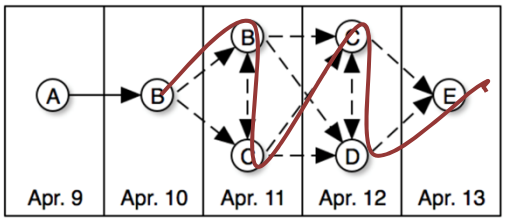


Carnegie Mellon isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- For a location sequence in a broken point: Find a location sequence with maximum transition probability product
  - $p(l_1 \to l_2 \to \cdots \to l_n) = p(l_2|l_1)p(l_3|l_2) \ldots p(l_n|l_{n-1})$
- Relation with Traveling Salesman Problem: visit each location exactly once and find the minimum travelling cost(NP-hard)
  - $-logp(l_1 \to l_2 \to \cdots \to l_n) = \sum_{i=2}^{n} -logp(l_n|l_{n-1})$

| Time | 2016.4.9 | 2016.4.10 | 2016.4.11 | 2016.4.11 | 2016.4.12 | 2016.4.12 | 2016.4.13 |
|---|---|---|---|---|---|---|---|
| Location | A | B | B | C | C | D | E |




CASOS



CASOS

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- Four algorithms
  - Random: Randomly pick the location order
  - Greedy: At each step, select next location with maximum transition probability.
  - Global greedy: First find a location pair with highest transition probability, then traverse to the source and target locations.
  - Exact algorithm: enumerate all the possible routes.

- Partition trails when time intervals are larger than a threshold.

| Time | Apr. 1 | Apr. 1 | Apr. 2 | Apr. 3 | June 3 | June 4 | June 5 | June 5 |
|---|---|---|---|---|---|---|---|---|
| Trail 1 | BEGIN | L1 | L2 | L3 | L2 | L1 | L2 | END |

| Time | Apr. 1 | Apr. 1 | Apr. 2 | Apr. 3 | Apr.3 | June 3 | June 3 | June 4 | June 5 | June 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| Trail 1 | BEGIN | L1 | L2 | L3 | END | BEGIN | L2 | L1 | L2 | END |




Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- Datasets:
  - Health record data
    - Location: health service
    - Agent: patient
    - Record: (patient, health service, date)

| | # of Records | # of Agents | # of Locations |
|---|---|---|---|
| Health data | 94885 | 5055 | 115 |

Use 814 unbroken trails as testing data

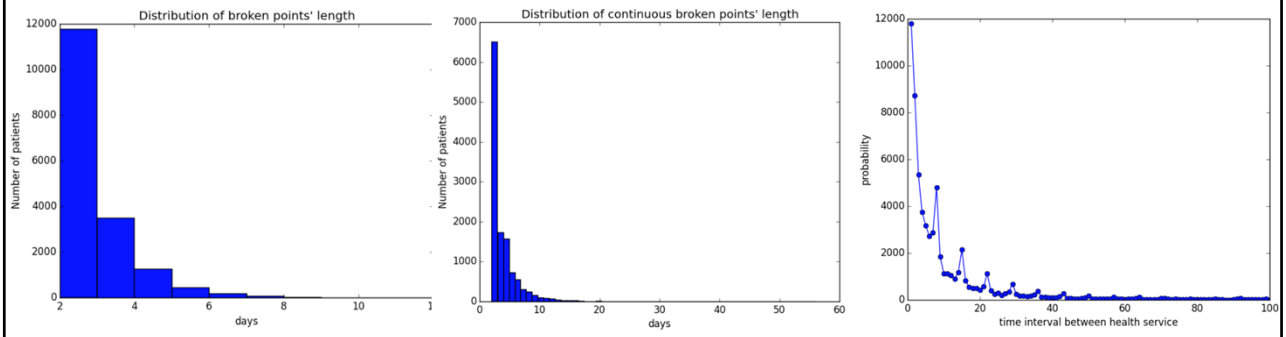



<Your Name>

Carnegie Mellon isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- Statistics of health dataset






Carnegie Mellon isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- Experiments setup
  - Artificially change the time resolution of 814 unbroken health trails
  - $timestamp' = \lfloor \frac{timestamp}{\text{resolution}} \rfloor * \text{resolution}.$

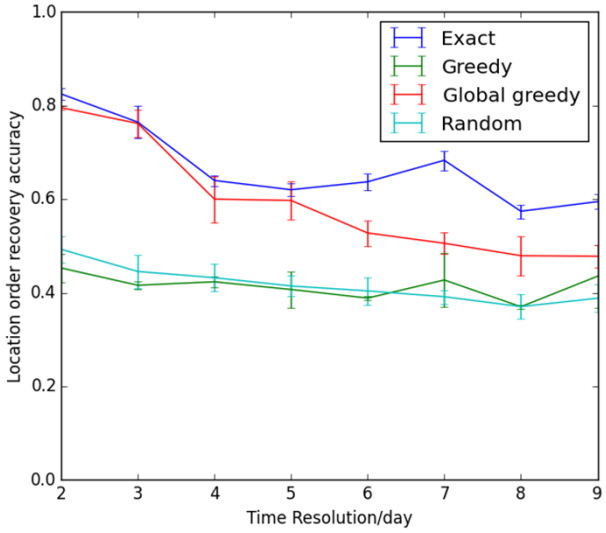
| Time resolution(day) | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| # of broken trails | 164 | 213 | 250 | 263 | 268 | 278 | 294 | 297 |
| # of broken points | 178 | 251 | 296 | 319 | 340 | 353 | 374 | 389 |
| Avg. length of broken points | 2.073 | 2.223 | 2.284 | 2.426 | 2.482 | 2.586 | 2.591 | 2.627 |
| Avg. length of cont. broken points | 2.121 | 2.364 | 2.449 | 2.650 | 2.749 | 2.916 | 2.909 | 3.041 |
| Max. length of broken points | 4 | 5 | 6 | 5 | 6 | 6 | 6 | 7 |

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

Comparison between four algorithms



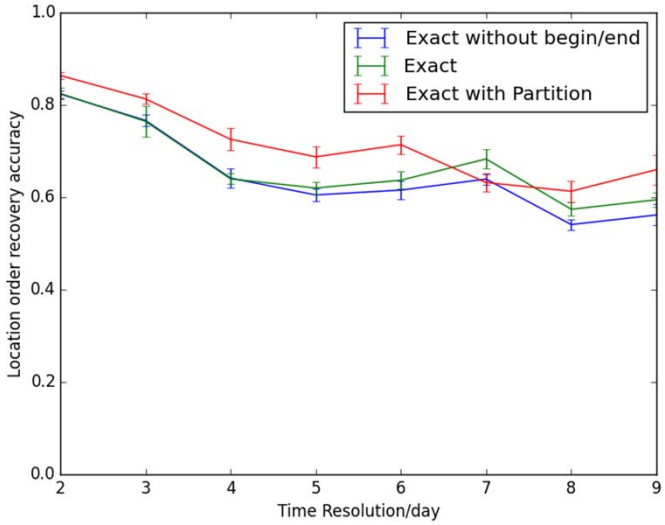

CASOS

15

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Current progress: Overcome the Time Resolution Issue in Trails

- The effects of BEGIN/END nodes and partitions




CASOS

16

CASOS

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Hands on

- Data preparation: program function call data

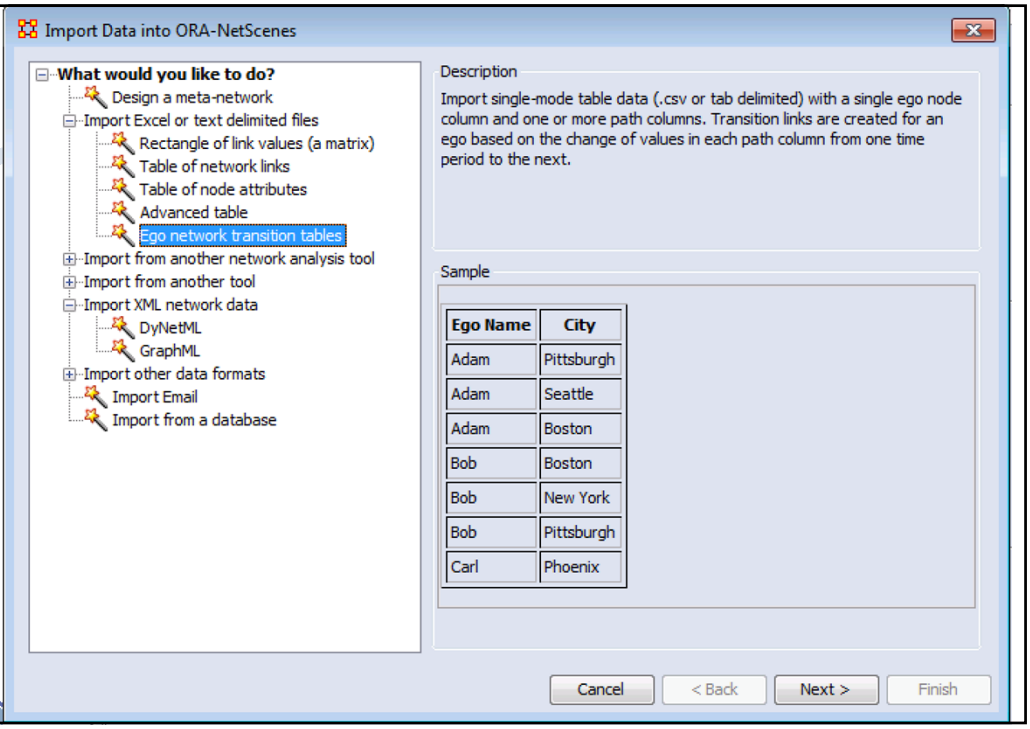
| id | timestamp | location |
|---|---|---|
| 0 | 553987065457672.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987065574331.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987065768508.00 | class java.lang.Class |
| 0 | 553987065819048.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987100679470.00 | class java.net.URL |
| 0 | 553987102005655.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987102202112.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987102260252.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987102331691.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987102384890.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987102425550.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987476519118.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987476619437.00 | class java.net.URL |
| 0 | 553987476664276.00 | class edu.umd.cs.findbugs.PluginLoader |
| 0 | 553987476741035.00 | class java.lang.String |
| 0 | 553987476797655.00 | class edu.umd.cs.findbugs.PluginLoader |

CASOS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

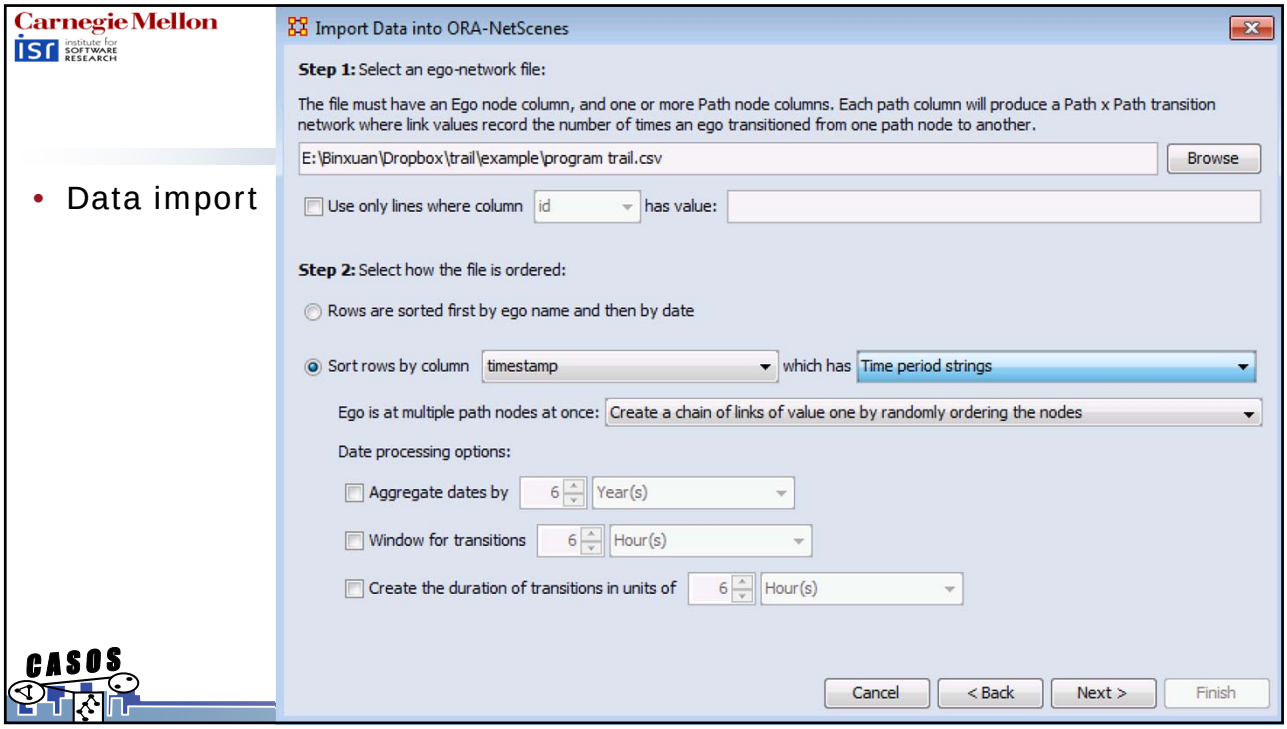
- Data import



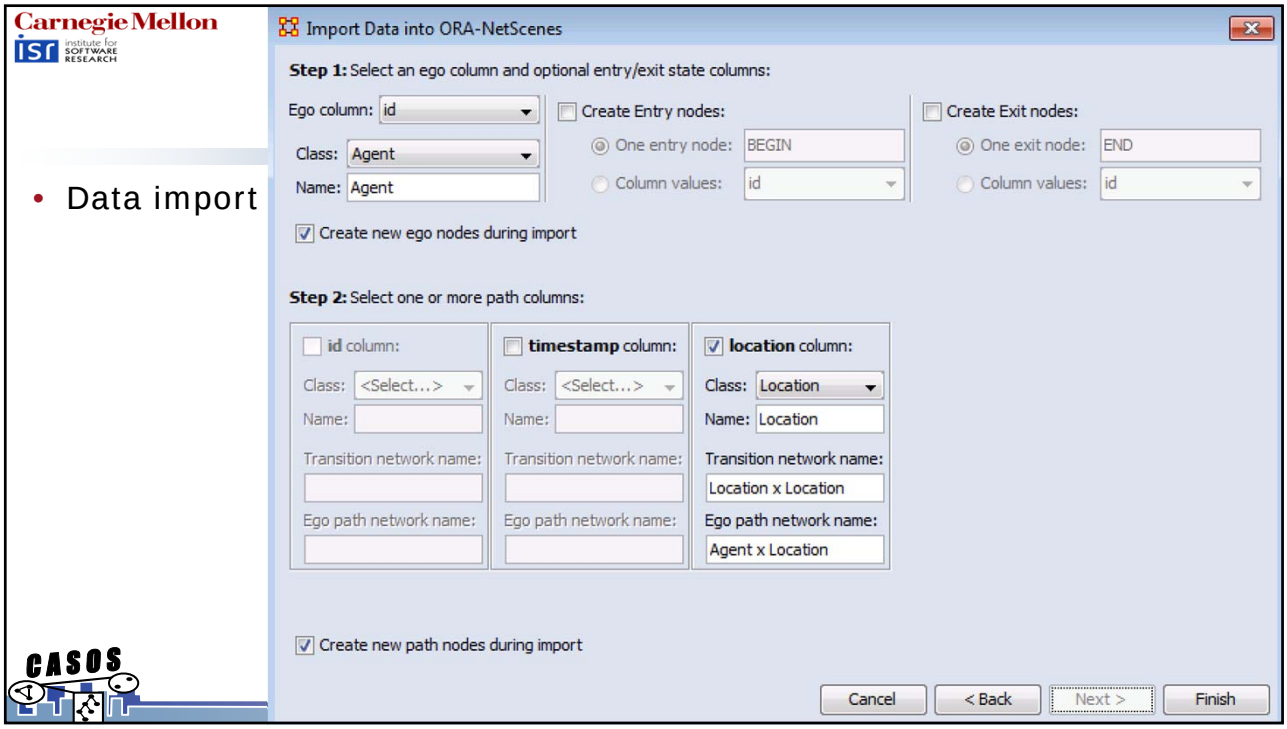

CASOS

<Your Name>

- Data import




- Data import

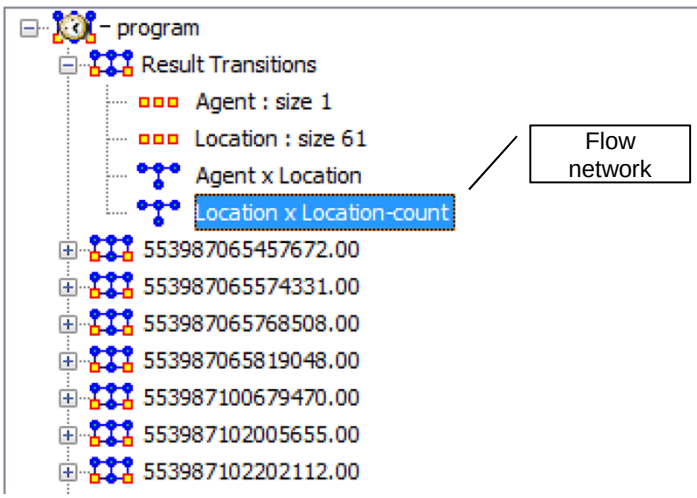


<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Hands on

- Data transform



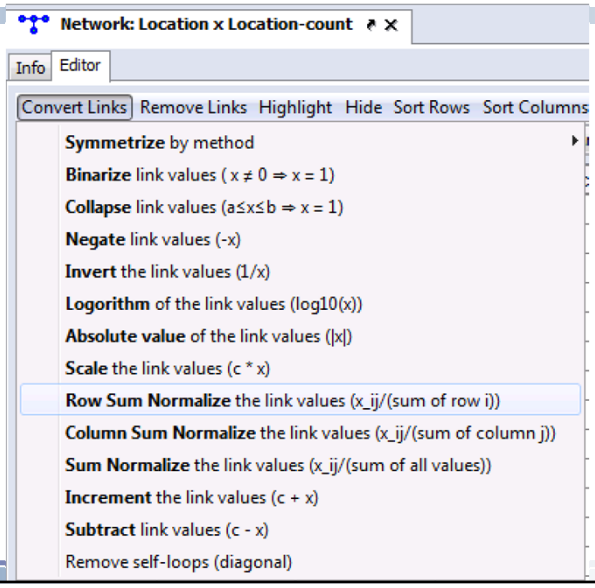

CASOS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Hands on

- Data transform




CASOS



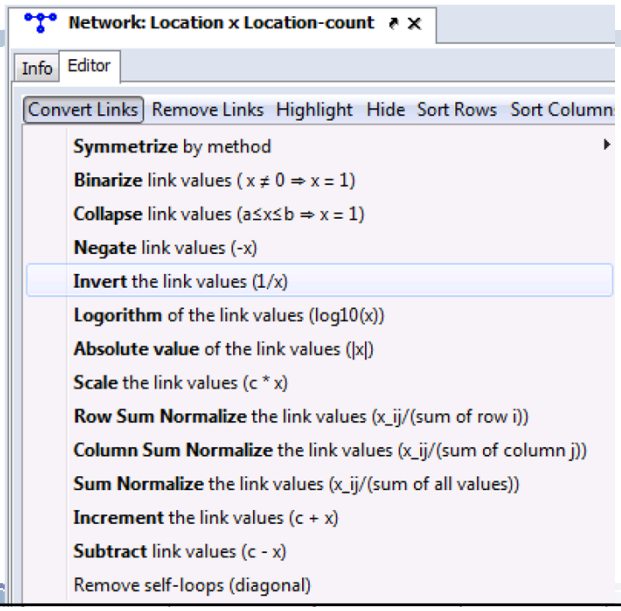
CASOS

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Hands on

- Data transform



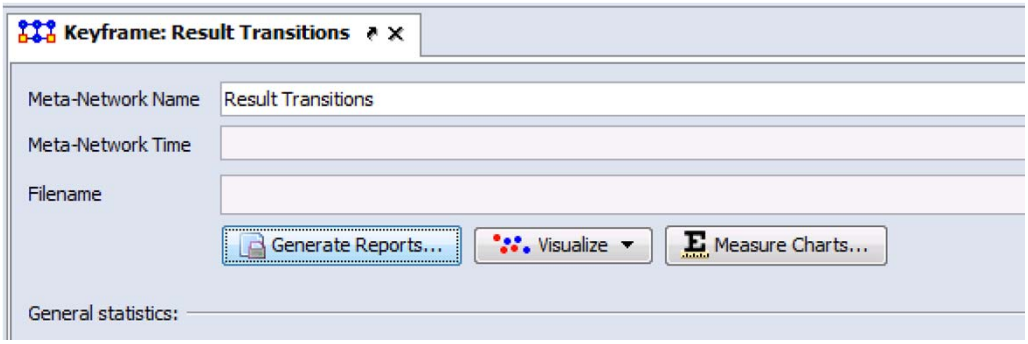

CASOS

23

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# Hands on

- Analysis




CASOS



CASOS