Trails and Networks: Higher-order networks, Trail Clustering

Mihovil Bartulovic

mbartulovic@cmu.edu

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

What are trails? (1)

- Graph theory: A trail in a walk with no repeated edge. The length of a trail is constrained by the number of edges.
- Trail is a path of an ego through time and space
  - people, ideas, diseases etc.
- It is a time-ordered sequence, i.e., a sequence of observations taken at different times.

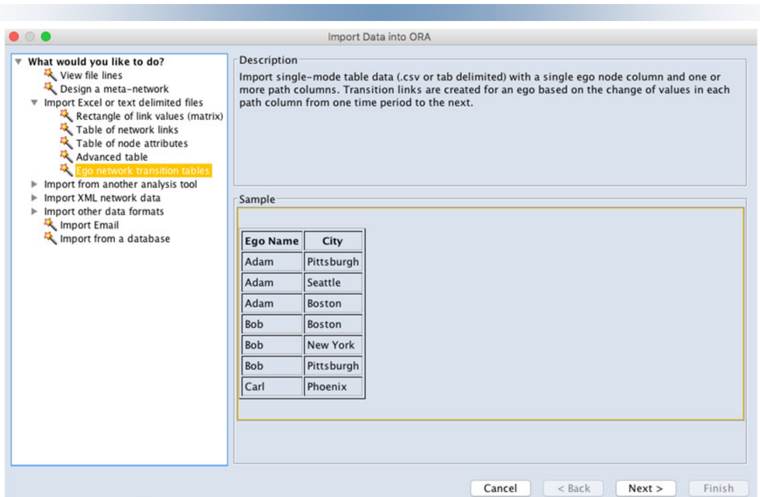
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

What are trails? (2)

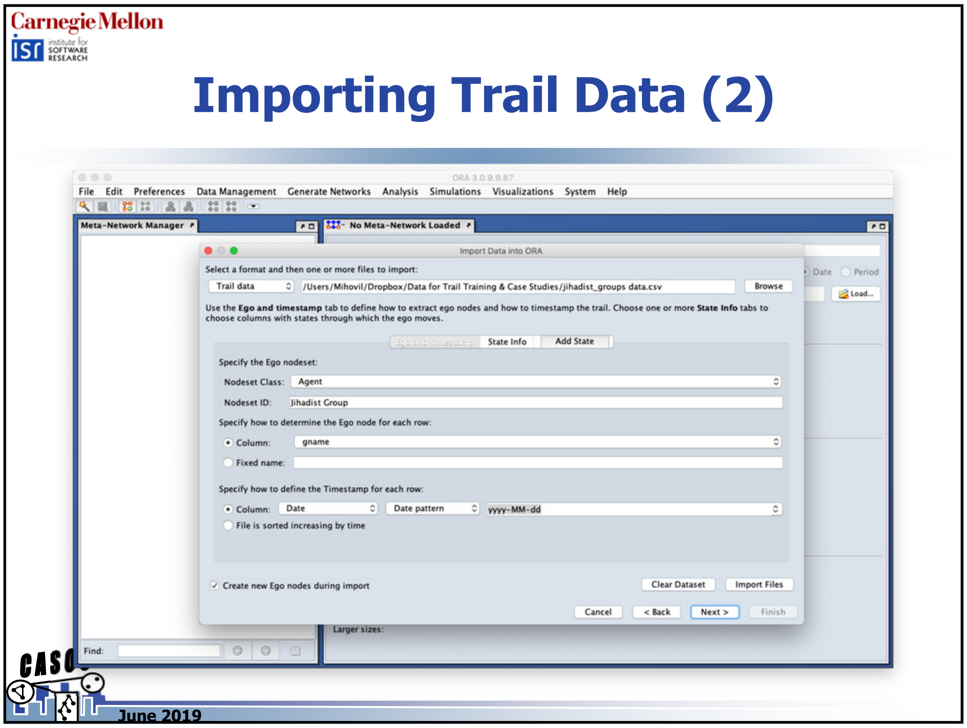
- Question 1: How can networks be generated from trail data?
- Question 2: Can we always use classic network metrics on networks created from trails?

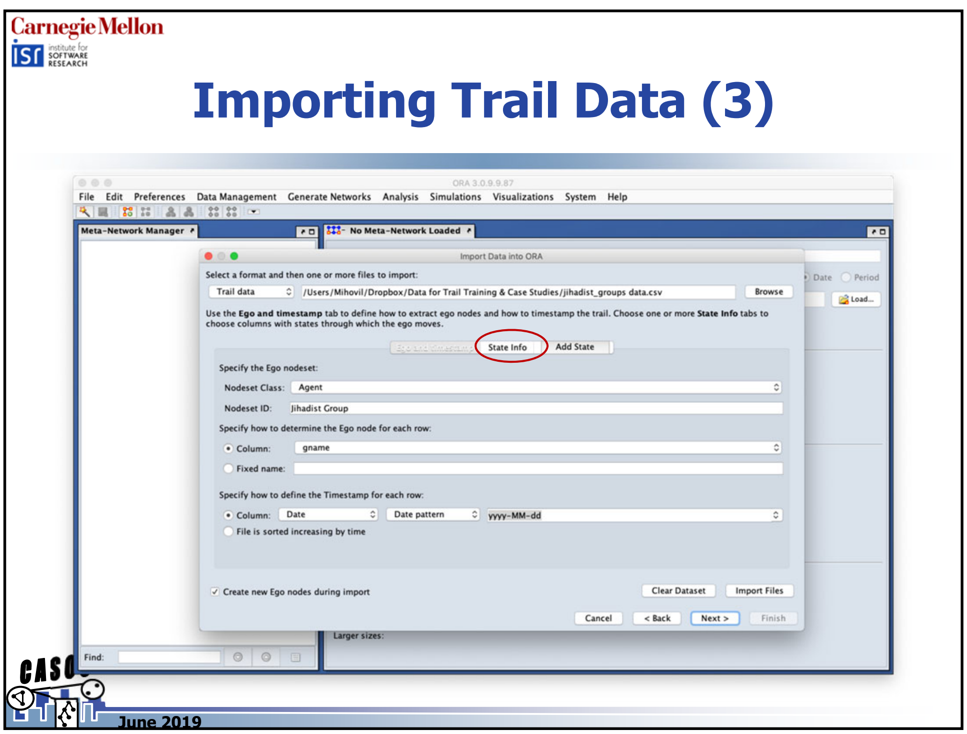
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Importing Trail Data (1)



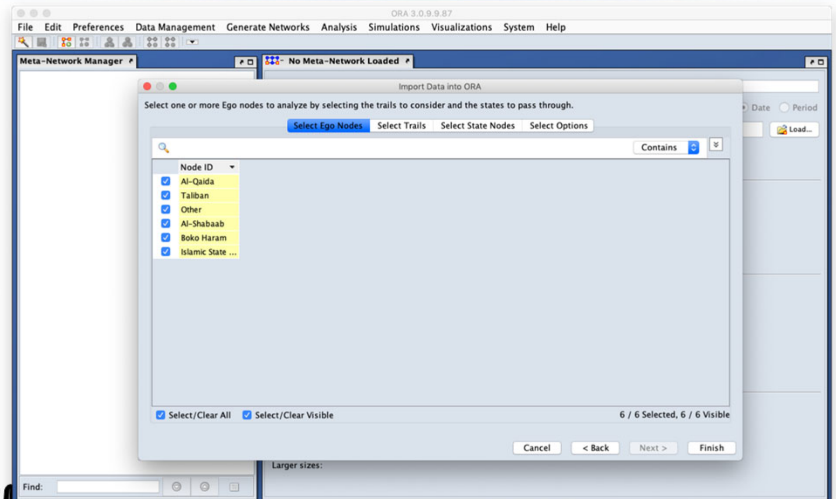
June 2019


Carnegie Mellon
isr institute for SOFTWARE RESEARCH
Importing Trail Data (2)
ORA 3.0.9.9.87
File Edit Preferences Data Management Generate Networks Analysis Simulations Visualizations System Help
Meta-Network Manager
No Meta-Network Loaded
Import Data into ORA
Select a format and then one or more files to import:
Trail data
/Users/Mihovil/Dropbox/Data for Trail Training & Case Studies/jihadist_groups data.csv
Browse
Use the Ego and timestamp tab to define how to extract ego nodes and how to timestamp the trail. Choose one or more State Info tabs to choose columns with states through which the ego moves.
State Info
Add State
Specify the Ego nodeset:
Nodeset Class: Agent
Nodeset ID: Jihadist Group
Specify how to determine the Ego node for each row:
Column: gname
Fixed name:
Specify how to define the Timestamp for each row:
Column: Date
Date pattern
yyyy-MM-dd
File is sorted increasing by time
Create new Ego nodes during import
Clear Dataset
Import Files
Cancel
< Back
Next >
Finish
Date
Period
Load...
Larger sizes:
Find:
June 2019


Carnegie Mellon
isr institute for SOFTWARE RESEARCH
Importing Trail Data (3)
ORA 3.0.9.9.87
File Edit Preferences Data Management Generate Networks Analysis Simulations Visualizations System Help
Meta-Network Manager
No Meta-Network Loaded
Import Data into ORA
Select a format and then one or more files to import:
Trail data
/Users/Mihovil/Dropbox/Data for Trail Training & Case Studies/jihadist_groups data.csv
Browse
Use the Ego and timestamp tab to define how to extract ego nodes and how to timestamp the trail. Choose one or more State Info tabs to choose columns with states through which the ego moves.
State Info
Add State
Specify the Ego nodeset:
Nodeset Class: Agent
Nodeset ID: Jihadist Group
Specify how to determine the Ego node for each row:
Column: gname
Fixed name:
Specify how to define the Timestamp for each row:
Column: Date
Date pattern
yyyy-MM-dd
File is sorted increasing by time
Create new Ego nodes during import
Clear Dataset
Import Files
Cancel
< Back
Next >
Finish
Date
Period
Load...
Larger sizes:
Find:
June 2019

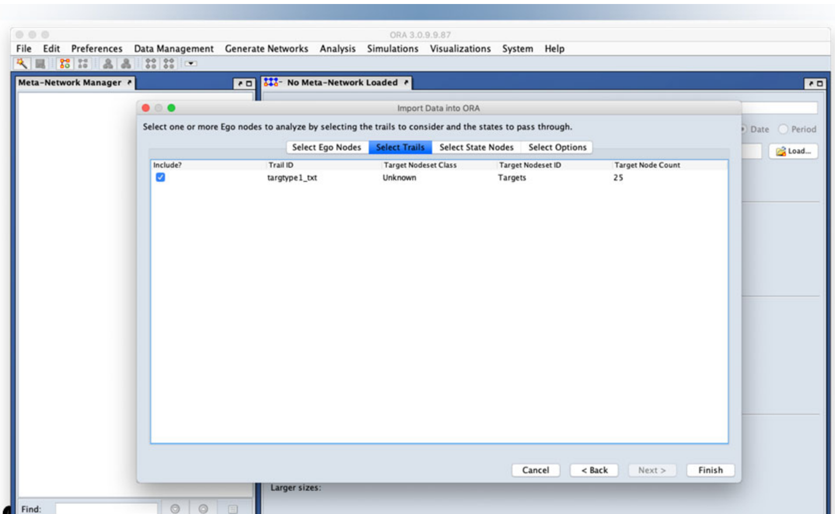
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

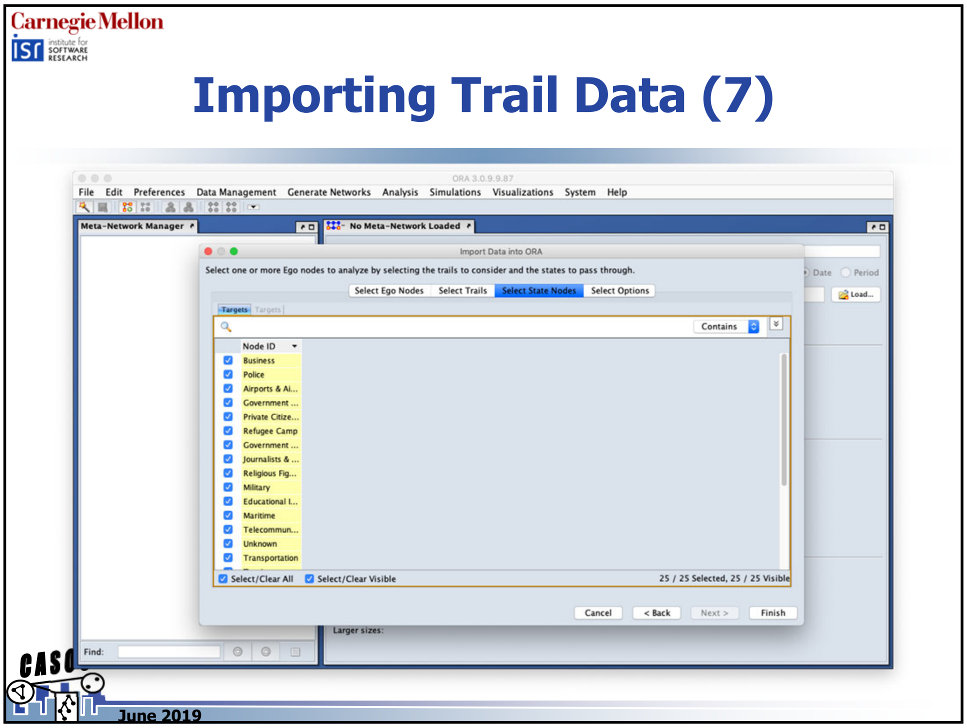
Importing Trail Data (5)

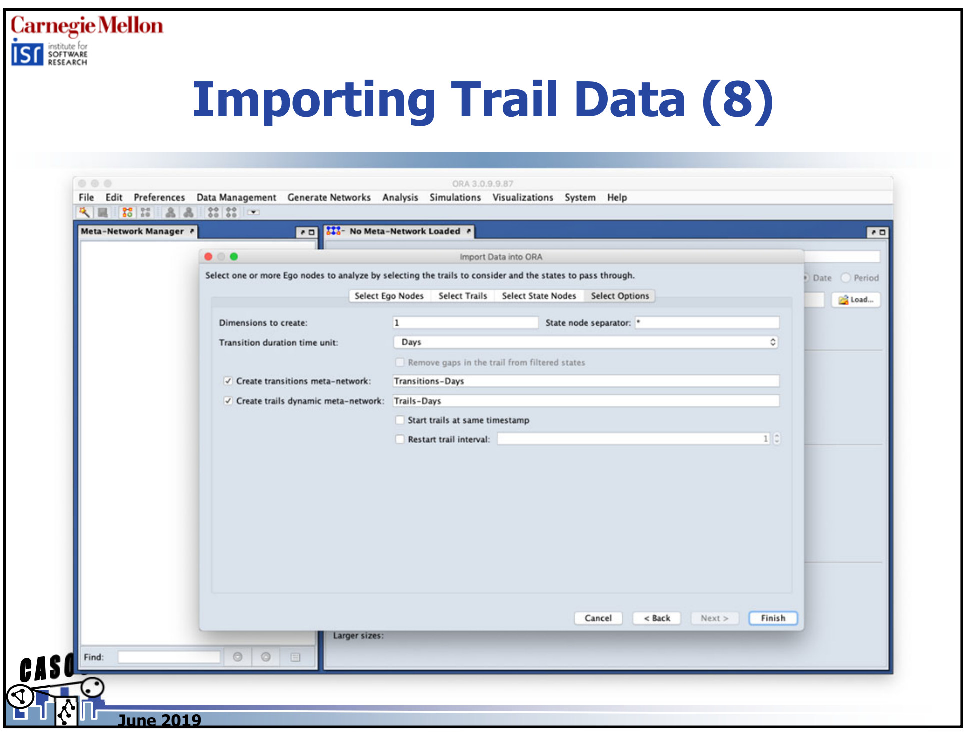


Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Importing Trail Data (6)




Carnegie Mellon
isr institute for SOFTWARE RESEARCH
Importing Trail Data (7)
File Edit Preferences Data Management Generate Networks Analysis Simulations Visualizations System Help
Meta-Network Manager
No Meta-Network Loaded
Import Data into ORA
Select one or more Ego nodes to analyze by selecting the trails to consider and the states to pass through.
Select Ego Nodes Select Trails Select State Nodes Select Options
Contains
Node ID
Business
Police
Airports & Ai...
Government ...
Private Citize...
Refugee Camp
Government ...
Journalists & ...
Religious Fig...
Military
Educational I...
Maritime
Telecommun...
Unknown
Transportation
Select/Clear All Select/Clear Visible
25 / 25 Selected, 25 / 25 Visible
Cancel < Back Next > Finish
Date Period
Load...
Larger sizes:
Find:
June 2019


Carnegie Mellon
isr institute for SOFTWARE RESEARCH
Importing Trail Data (8)
File Edit Preferences Data Management Generate Networks Analysis Simulations Visualizations System Help
Meta-Network Manager
No Meta-Network Loaded
Import Data into ORA
Select one or more Ego nodes to analyze by selecting the trails to consider and the states to pass through.
Select Ego Nodes Select Trails Select State Nodes Select Options
Dimensions to create: 1
State node separator: *
Transition duration time unit: Days
Remove gaps in the trail from filtered states
Create transitions meta-network: Transitions-Days
Create trails dynamic meta-network: Trails-Days
Start trails at same timestamp
Restart trail interval: 1
Cancel < Back Next > Finish
Date Period
Load...
Larger sizes:
Find:
June 2019

Carnegie Mellon

Importing Trail Data (9)

- Data is imported both as a sequence of "per time slice" networks and aggregated transitional networks (number of transitions ego has between two nodes)
  - "Per time slice" networks → Looms
  - Aggregated transitional networks → Markov Chains

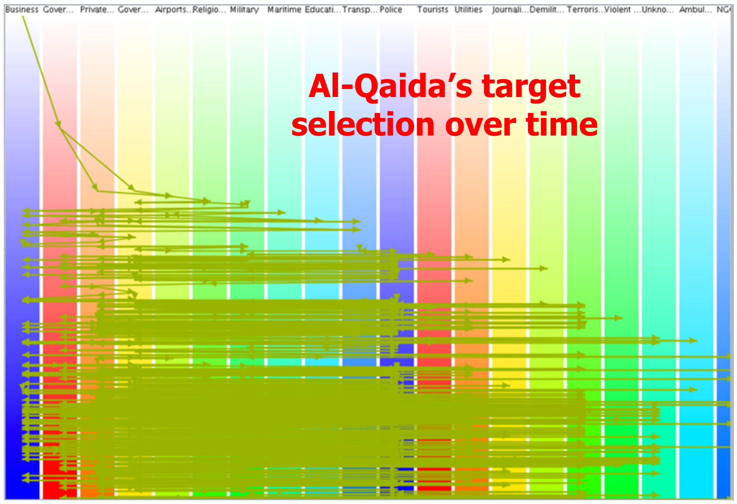
Carnegie Mellon

Looms (1)

- Visualization depends on what we wish to observe
- Good indicator of timeline
- Sometimes cluttered

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Looms (2)



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Networks From Trails (1)

- Question 1: How can networks be generated from trail data?
  - Markov Chains - network of transitional probabilities (or cumulative weights) among nodes i.e. each node represents a location or an individual

Networks From Trails (2)

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	F1	F2	F3	F2	F1	F2
Trail 2	F2	F3	F4	F2	F1	F1
Trail 3	F2	F3	F1	F1	F2	F3

	F1	F2	F3	F4
F1	2	3	0	0
F2	2	0	4	0
F3	1	1	0	1
F4	0	1	0	0

Traffic flow network

$$P(F_i \to F_j) = \frac{N(F_i \to F_j)}{\sum_j N(F_i \to F_j)}$$

	F1	F2	F3	F4
F1	0.4	0.6	0	0
F2	0.33	0	0.67	0
F3	0.33	0.33	0	0.33
F4	0	1	0	0

Markov transition network

From Trails to Transitional Networks

- Observe ego's transitions from one state to another
- Aggregate the observed transitions
- Create probabilities from the aggregated values

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Why do we care about high dimensional networks?

- Both sequential and “memory” property of the data has to be accounted for
  - network-analytic methods make the fundamental assumption that paths are transitive, i.e. the existence of paths from a to b and from b to c implies a transitive path from a via b to c.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

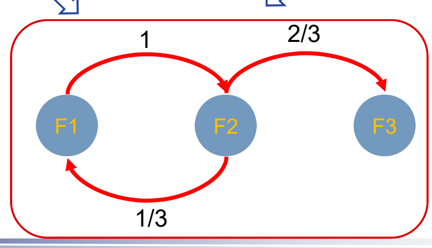
Example 1 – Function Calling

Function Caller	Function Called
F2	F3
F2	F1
F2	F3
F1	F2
F1	F2

Time

Function Caller	Function Called
F1	F2
F2	F1
F1	F2
F2	F3
F2	F3

Time



We lost the temporal component!

Why do we care about high dimensional networks?

- Agent's paths and previous actions matter
  - First-order network is built by taking the number of transitions between pairs of nodes as edge weights (or scaled to transitional probabilities)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Why do we care about high dimensional trails?

- Agent's paths and previous actions matter
  - First-order network is built by taking the number of trails between pairs of nodes as edge weights (or scaled to transitional probabilities) → PROBLEM!!
    - Same nodes could be used by different entities coming from different nodes following their own path
  - Solution → splitting the "crossroad" nodes
    - We care about where ego comes from
    - More accurate simulation of movement patterns observed in the original data

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Example 2 - Jihadist Groups (1)

Group Name	Target
ISIL	Business
Al-Qaida	Police
ISIL	Military
Al-Qaida	Military
Al-Qaida	Government (General)
ISIL	NGO
…	…

Time

CASOS
June 2019

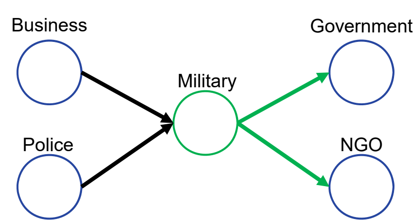
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Example 2 - Jihadist Groups (2)

Group Name	Target
ISIL	Business
Al-Qaida	Police
ISIL	Military
Al-Qaida	Military
Al-Qaida	Government
ISIL	NGO
…	…

Time

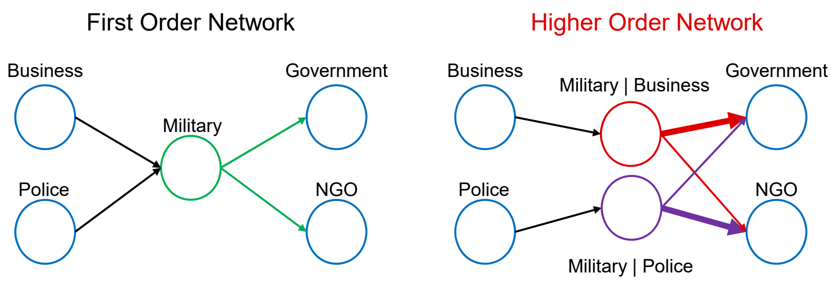
First Order Network



CASOS
June 2019

Carnegie Mellon

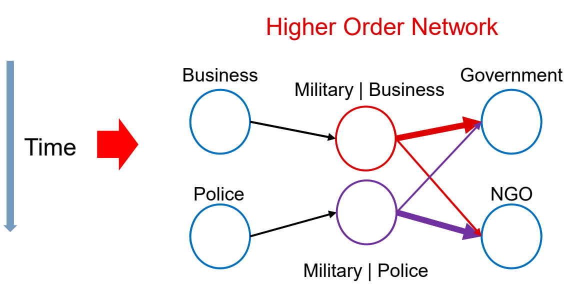
Example 2 - Jihadist Groups (3)



Carnegie Mellon

Example 2 - Jihadist Groups (4)

Group Name	Target
ISIL	Business
Al-Qaida	Police
ISIL	Military
Al-Qaida	Military
Al-Qaida	Government
ISIL	NGO
...	...



More informative and
better representation of the data!

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Higher Order Networks (1)

- Rethinking the building blocks of a network:
  - Instead of using a node to represent a single entity, we break down the node into different higher order nodes that carry different dependency relationships (each node can now represent a series of entities)
  - Military | Business and Military | Police → the edges can now involve multiple different targets as entities and carry different weights → second-order dependencies.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Higher Order Networks (2)

- Out-edges are in the form of $i|h \rightarrow k$ instead of $i \rightarrow k$, transitional probability from node $i|h$ to node j is

$$P\left(X_{t+1} = j | X_t = (i|h)\right) = \frac{N(i|h \rightarrow j)}{\sum_k N(i|h \rightarrow k)}$$

- Movement depends on the current node and on one or more other entities in the new network representation

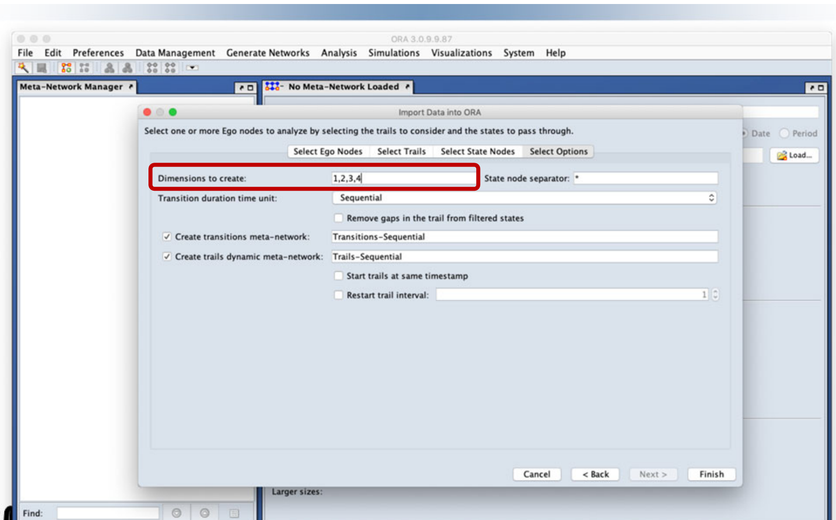
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Higher Order Networks (3)

- This new representation is consistent with conventional networks and compatible with existing network analysis methods
  - We need to be careful when using the network metrics and have full graph of how network is created and what edges represent!
- PROBLEM – How to determine optimal order of the Higher Order Network?
  - Statistical analysis, Maximum likelihood, …

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Importing High-Dimensional Trails



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Trail Clustering (1)

- Data from domains such as protein sequences retail transactions intrusion detection and web logs have an inherent sequential nature
- Clustering of such data sets is useful for various purposes
  - For example clustering of sequences from commercial data sets may help marketer identify different customer groups based upon their purchasing patterns

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Trail Clustering (2)

- Let us have a dataset of n trails to be clustered
- Let us have a set $C = \{c_1, c_2, \ldots, c_k\}$ of k corpora with $|\mathrm{c_j}| = \mathrm{N_j}$ trails within each corpora
- A trail will be denoted by i $(i = 1, \ldots, n)$. Each trail is characterized by a sequence of states x_i from a finite set $\boldsymbol{X}$.
- Let $x = (x_1, \ldots, x_n)$ denote a sample of size n. Let x_{it} denote the state of the trail i at position t.
- We assume discrete time from 0 to T_i $(t = 0,1, \ldots, T_i)$.
- Thus, the vector $\boldsymbol{x}_i$ denotes the consecutive states x_{it}, with $t = 0, \ldots, T_i$. The sequence $\mathbf{x_i} = (\mathrm{x_{i0}}, \mathrm{x_{i1}}, \ldots, \mathrm{x_{iT_i-1}}, \mathrm{x_{iT_i}})$ can be extremely difficult to characterize and describe, due to its varying dimension $(T_i + 1)$.

Carnegie Mellon

Trail Clustering (3)

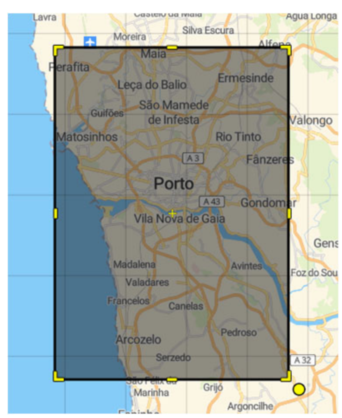
$$\begin{array}{ll} \underset{c_j \in \mathcal{C}}{\arg\min} & \mathcal{D}(c_j, \mathbf{x_i}) \\ \text{subject to} & \mathcal{C} = \{c_1, c_2, \ldots, c_k\}, \\ & x_{i,T_i - t} \in \mathcal{X}, t \in \langle 0, T_i \rangle \end{array}$$

- $D(\cdot,\cdot)$ cost function taking form of inverse similarity coefficient or distance metric

Carnegie Mellon

Trail Clustering Example (1)

- Taxi trip location data from Porto, Portugal
- (Latitude, Longitude) pairs over time per taxi trip



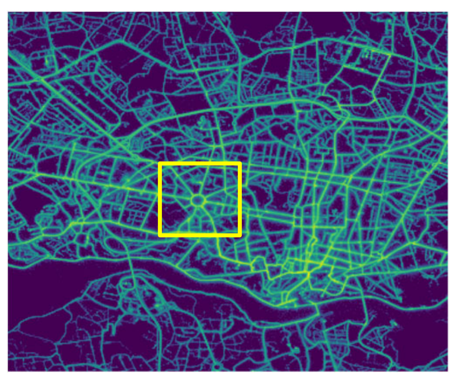
Carnegie Mellon

Trail Clustering Example (2)

Carnegie Mellon

Trail Clustering Example (2)

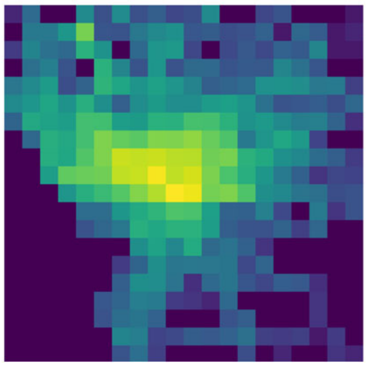
- Data split:
  - Corpus 1: Taxi trips that use roundabout (6565 trails)
  - Corpus 2: Everything else (3435 trails)



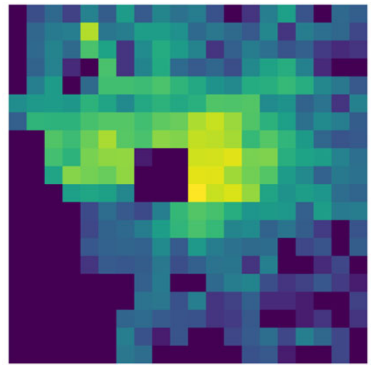
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Trail Clustering Example (1)

With roundabout corpus



No roundabout corpus



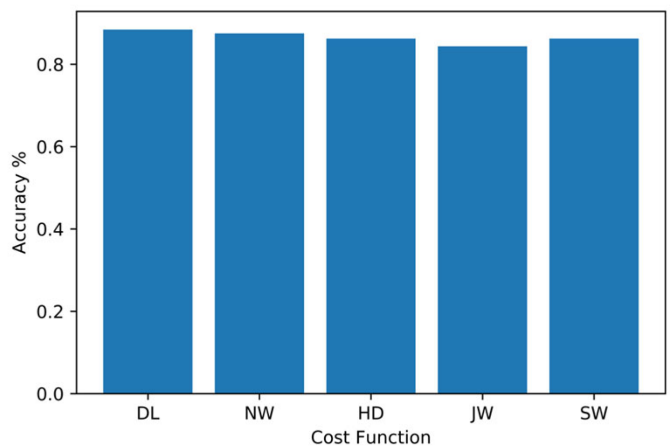
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Trail Clustering Example (1)

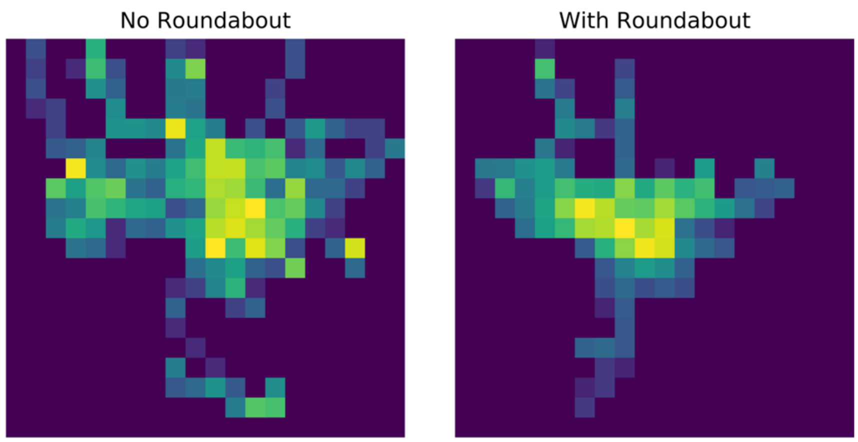
- Cost functions:
  - Damerau–Levenshtein distance (DL)
  - Hamming Distance (HD)
  - Jaro–Winkler distance (JW)
  - Needleman–Wunsch algorithm (NW)
  - Smith–Waterman algorithm (SW)

- Test data:
  - 320 trails to cluster

Trail Clustering

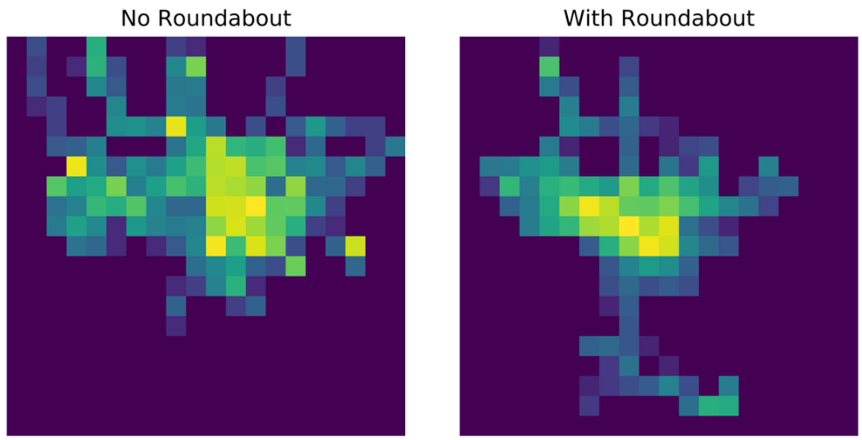


DL heatmap



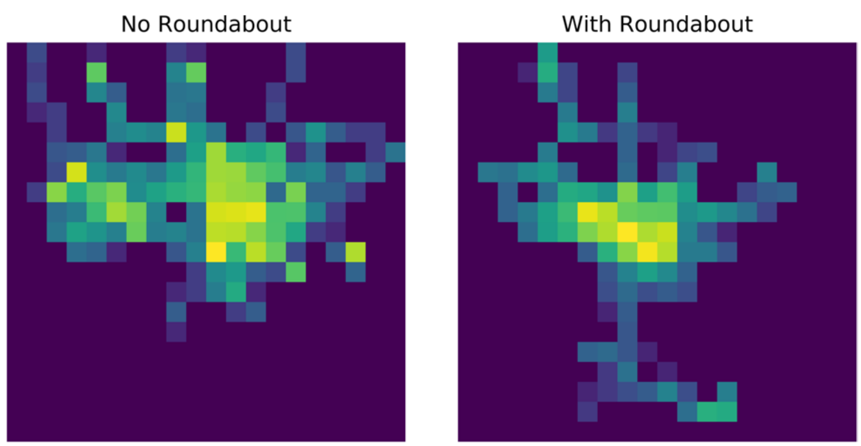
Carnegie Mellon

HD heatmap



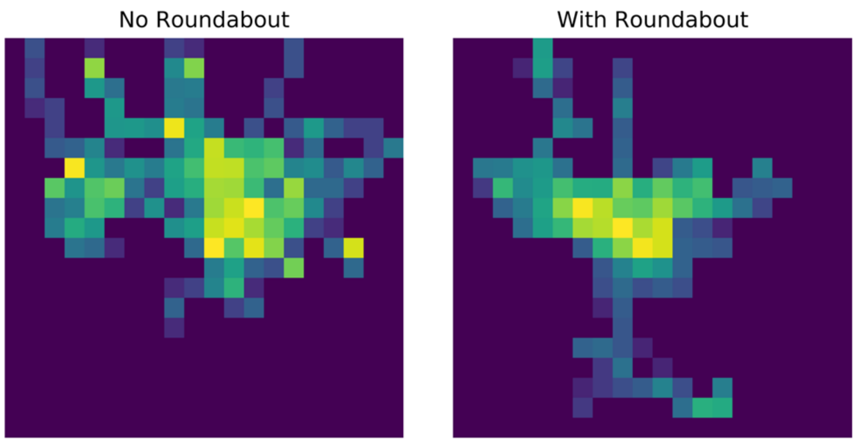
Carnegie Mellon

JW heatmap



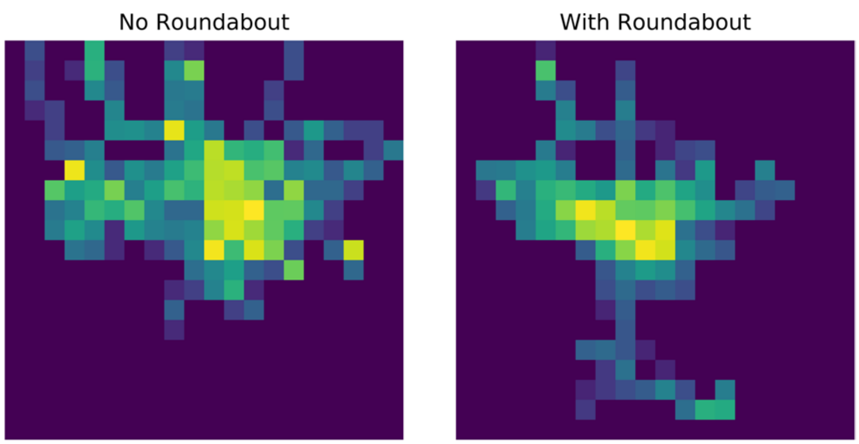
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

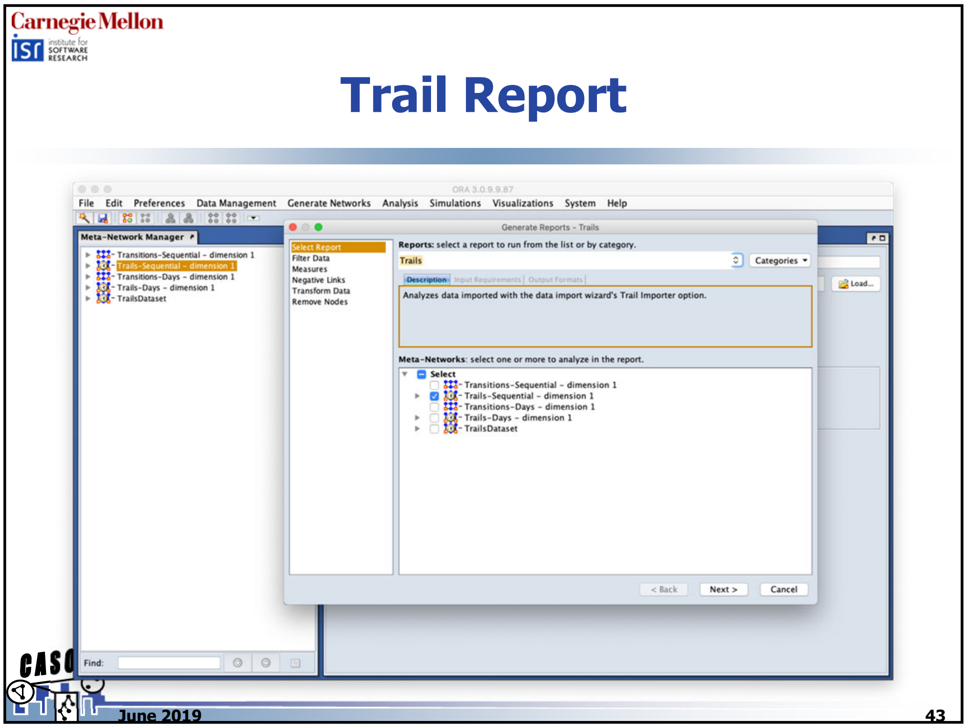
NW heatmap



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

SW heatmap




Carnegie Mellon
isr institute for SOFTWARE RESEARCH
Trail Report
File Edit Preferences Data Management Generate Networks Analysis Simulations Visualizations System Help
Meta-Network Manager
Transitions-Sequential – dimension 1
Trails-Sequential – dimension 1
Transitions-Days – dimension 1
Trails-Days – dimension 1
TrailsDataset
Generate Reports - Trails
Select Report
Filter Data
Measures
Negative Links
Transform Data
Remove Nodes
Reports: select a report to run from the list or by category.
Trails
Categories
Description
Analyzes data imported with the data import wizard's Trail Importer option.
Meta-Networks: select one or more to analyze in the report.
Select
Transitions-Sequential – dimension 1
Trails-Sequential – dimension 1
Transitions-Days – dimension 1
Trails-Days – dimension 1
TrailsDataset
< Back
Next >
Cancel
Find:
Load...
June 2019
43