



Case Study: Trails

Binxuan Huang
binxuanh@cs.cmu.edu

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



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



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



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

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

	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

Traffic flow network
Markov transition network

CASOS Summer Institute 2016 3

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

	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

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

	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

Traffic flow network
Markov transition network

CASOS Summer Institute 2016 4



Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Case study

- Case study: one patient health trail
-

Diagram illustrating a patient health trail centered on Internal Medicine, connected to various other medical specialties and services:

- Echocardiography
- Adult Medical-Surgical
- Emergency
- ENT
- Nutrition
- Gastroenerology
- Outpt psychiatry
- Bone Densitometry
- Electrophysiology
- Pulmonology-Sleep center

CASOS

5

Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Case study

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

Diagram illustrating a patient health trail centered on Internal Medicine, connected to various other medical specialties and services, with specific nodes highlighted as BEGIN and END:

- Echocardiography
- Adult Medical-Surgical
- Emergency
- ENT
- Nutrition
- Gastroenerology
- Outpt psychiatry
- Bone Densitometry
- Electrophysiology
- Pulmonology-Sleep center

CASOS

6



Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

Case study

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

Centrality, Betweenness : Location x Location

Location	Value
Internal Medicine	0.85
Adult Medical-Surgical	0.25
Podiatry	0.15
Electrophysiology	0.10
ENT	0.08
Emergency	0.05
Bone Densitometry	0.05
Output psychiatry	0.05
Gastroenterology	0.02
Nutrition	0.01

Centrality, Betweenness : Location x Location

Location	Value
Internal Medicine	0.75
Adult Medical-Surgical	0.20
Podiatry	0.12
Electrophysiology	0.08
ENT	0.06
Emergency	0.04
Bone Densitometry	0.04
Output psychiatry	0.04
Gastroenterology	0.02
BEGIN	0.01

CASOS CASOS Summer Institute 2016 7

Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

Case study

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

Centrality, Betweenness : Location x Location-count

Location	Value
Echocardiography	0.22
Adult Medical-Surgical	0.20
Emergency	0.17
Internal Medicine	0.15
Cardiac diagnostics	0.14
Pulmonary functi...	0.13
Unl-known	0.10
Electrophysiology	0.09
General Cardiology	0.08
Circulatory physiology	0.07

Centrality, Betweenness : Location x Location-count

Location	Value
Echocardiography	0.20
Internal Medicine	0.18
Adult Medical-Surgical	0.17
Unl-known	0.15
Emergency	0.14
Circulatory physiology	0.13
Pulmonary function test	0.11
Cardiac diagnostics	0.10
General Cardiology	0.09
Adult Ophthalmology	0.08

CASOS CASOS Summer Institute 2016 8



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 CASOS Summer Institute 2016 9

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
 - $p(l_1 \rightarrow l_2 \rightarrow \dots \rightarrow l_n) = p(l_2|l_1)p(l_3|l_2) \dots p(l_n|l_{n-1})$
- Relation with Traveling Salesman Problem: visit each location exactly once and find the minimum travelling cost(NP-hard)
 - $-\log p(l_1 \rightarrow l_2 \rightarrow \dots \rightarrow l_n) = \sum_{i=2}^n -\log p(l_i|l_{i-1})$

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 10



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

CASOS

CASOS Summer Institute 2016 11

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

CASOS

CASOS Summer Institute 2016 12



Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Current progress: Overcome the Time Resolution Issue in Trails

- Statistics of health dataset

The first chart, 'Distribution of broken points' length', is a histogram with 'Number of patients' on the y-axis (0 to 12000) and 'days' on the x-axis (2 to 11). The distribution is highly skewed towards lower values. The second chart, 'Distribution of continuous broken points' length', is a histogram with 'Number of patients' on the y-axis (0 to 7000) and 'days' on the x-axis (0 to 60). It also shows a strong right skew. The third chart is a line plot with 'probability' on the y-axis (0 to 12000) and 'time interval between health service' on the x-axis (0 to 100). It shows a sharp peak at day 0 and a long tail extending to 100 days.

CASOS

CASOS Summer Institute 2016 13

Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Current progress: Overcome the Time Resolution Issue in Trails

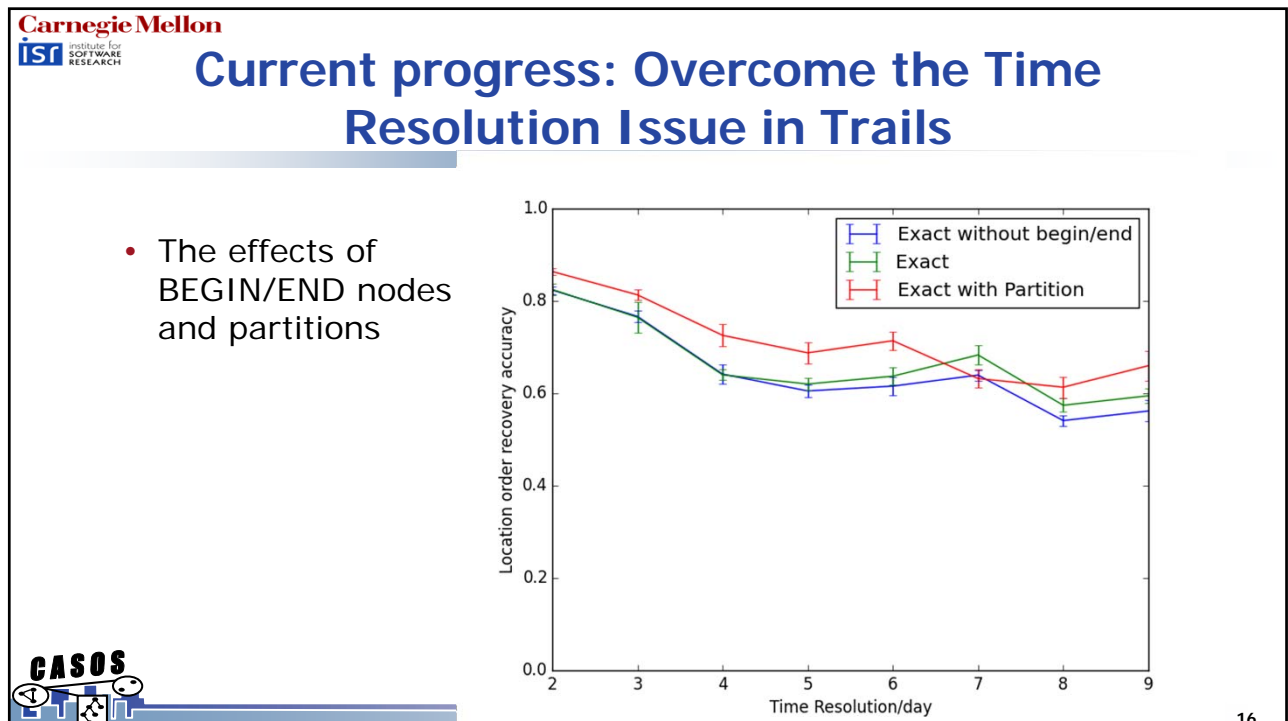
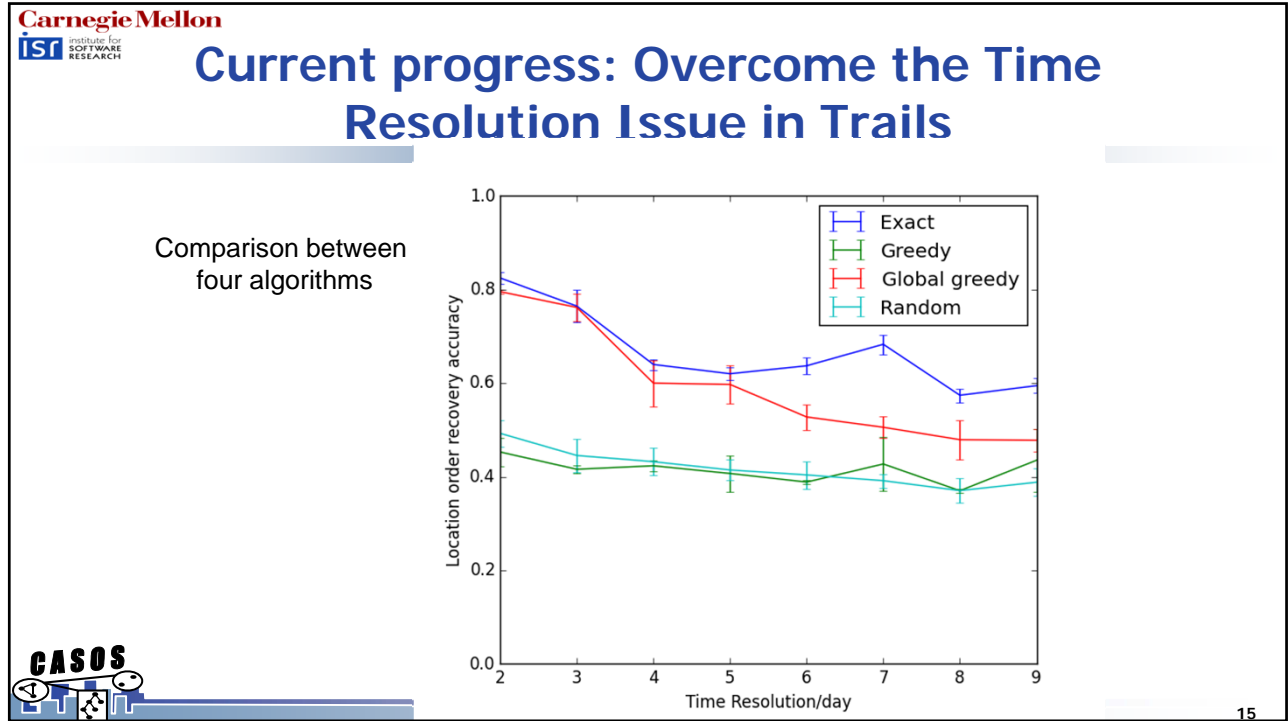
- Experiments setup
 - Artificially change the time resolution of 814 unbroken health trails
 - $timestamp' = \lfloor \frac{timestamp}{resolution} \rfloor * resolution.$

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

CASOS

CASOS Summer Institute 2016 14





Carnegie Mellon
ISRI 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 Summer Institute 2016 17

Carnegie Mellon
ISRI Institute for Software Research

- Data import

Import Data into ORA-NetScenes

What would you like to do?

- Design a meta-network
- Import Excel or text delimited files
 - Rectangle of link values (a matrix)
 - Table of network links
 - Table of node attributes
 - Advanced table
 - Ego network transition tables
- Import from another network analysis tool
- Import from another tool
- Import XML network data
 - DyNetML
 - GraphML
- Import other data formats
 - Import Email
 - Import from a database

Description

Import single-mode table data (.csv or tab delimited) with a single ego node column and one or more path columns. Transition links are created for an ego based on the change of values in each path column from one time period to the next.

Sample


Ego Name	City
Adam	Pittsburgh
Adam	Seattle
Adam	Boston
Bob	Boston
Bob	New York
Bob	Pittsburgh
Carl	Phoenix

Cancel < Back Next > Finish



Carnegie Mellon ISRI Institute for Software Research

- Data import



Import Data into ORA-NetScenes

Step 1: Select an ego-network file:
The file must have an Ego node column, and one or more Path node columns. Each path column will produce a Path x Path transition network where link values record the number of times an ego transitioned from one path node to another.
E:\Binxuan\Dropbox\trail\example\program trail.csv Browse

Use only lines where column **id** has value:

Step 2: Select how the file is ordered:

Rows are sorted first by ego name and then by date

Sort rows by column **timestamp** which has **Time period strings**

Ego is at multiple path nodes at once: **Create a chain of links of value one by randomly ordering the nodes**

Date processing options:

Aggregate dates by **6** Year(s)


Window for transitions **6** Hour(s)

Create the duration of transitions in units of **6** Hour(s)

Cancel < Back Next > Finish

Carnegie Mellon ISRI Institute for Software Research

- Data import



Import Data into ORA-NetScenes

Step 1: Select an ego column and optional entry/exit state columns:

Ego column: **id** Create Entry nodes: One entry node: **BEGIN** Create Exit nodes: One exit node: **END**

Class: **Agent** Column values: **id** Column values: **id**

Name: **Agent** Create new ego nodes during import

Step 2: Select one or more path columns:

<input type="checkbox"/> id column:	<input type="checkbox"/> timestamp column:	<input checked="" type="checkbox"/> location column:
Class: <Select...>	Class: <Select...>	Class: Location
Name: <input type="text"/>	Name: <input type="text"/>	Name: Location
Transition network name: <input type="text"/>	Transition network name: <input type="text"/>	Transition network name: Location x Location
Ego path network name: <input type="text"/>	Ego path network name: <input type="text"/>	Ego path network name: Agent x Location

Create new path nodes during import

Cancel < Back Next > Finish



Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Hands on

- Data transform

program

- Result Transitions
 - Agent : size 1
 - Location : size 61
 - Agent x Location
 - Location x Location-count
- 553987065457672.00
- 553987065574331.00
- 553987065768508.00
- 553987065819048.00
- 553987100679470.00
- 553987102005655.00
- 553987102202112.00

Flow network

CASOS

CASOS Summer Institute 2016 21

Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Hands on

- Data transform

Network: Location x Location-count

Info Editor

Convert Links Remove Links Highlight Hide Sort Rows Sort Columns

Symmetrize by method

- Binarize link values ($x \neq 0 \Rightarrow x = 1$)
- Collapse link values ($a \leq x \leq b \Rightarrow x = 1$)
- Negate link values ($-x$)
- Invert the link values ($1/x$)
- Logarithm of the link values ($\log_{10}(x)$)
- Absolute value of the link values ($|x|$)
- Scale the link values ($c * x$)
- Row Sum Normalize the link values ($x_{ij}/(\text{sum of row } i)$)**
- Column Sum Normalize the link values ($x_{ij}/(\text{sum of column } j)$)
- Sum Normalize the link values ($x_{ij}/(\text{sum of all values})$)
- Increment the link values ($c + x$)
- Subtract link values ($c - x$)
- Remove self-loops (diagonal)

CASOS

22



Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Hands on

- Data transform

Network: Location x Location-count

Info Editor

Convert Links Remove Links Highlight Hide Sort Rows Sort Column

- Symmetrize by method
- Binarize link values ($x \neq 0 \Rightarrow x = 1$)
- Collapse link values ($a \leq x \leq b \Rightarrow x = 1$)
- Negate link values ($-x$)
- Invert the link values ($1/x$)**
- Logarithm of the link values ($\log_{10}(x)$)
- Absolute value of the link values ($|x|$)
- Scale the link values ($c * x$)
- Row Sum Normalize the link values ($x_{ij}/(\text{sum of row } i)$)
- Column Sum Normalize the link values ($x_{ij}/(\text{sum of column } j)$)
- Sum Normalize the link values ($x_{ij}/(\text{sum of all values})$)
- Increment the link values ($c + x$)
- Subtract link values ($c - x$)
- Remove self-loops (diagonal)

CASOS

23

Carnegie Mellon
ISR Institute for SOFTWARE RESEARCH

Hands on

- Analysis

Keyframe: Result Transitions

Meta-Network Name: Result Transitions

Meta-Network Time:

Filename:

Generate Reports... Visualize Measure Charts...

General statistics:

CASOS

CASOS Summer Institute 2016

24

