Part 3

Extension of time-series: tensor analysis

Yasushi Sakurai (Kumamoto University)
Yasuko Matsubara (Kumamoto University)
Christos Faloutsos (Carnegie Mellon University)

Kumamoto U CMU CS

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

  Multi-Aspect Non-linear Time-series

  M A N T

Kumamoto U CMU CS

Outline

- Tensor decomposition
  - Motivation
  - Basic approaches
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

  Multi-Aspect Non-linear Time-series

  M A N T

Kumamoto U CMU CS

Motivation 1: Why "matrix"?

- Why matrices are important?

Kumamoto U CMU CS

Examples of Matrices: Graph - social network

	John	Peter	Mary	Nick	...
John	0	11	22	55	...
Peter	5	0	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Examples of Matrices: cloud of n-d points

	chol#	blood#	age	..	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Examples of Matrices: Market basket

- **market basket** as in Association Rules

	milk	bread	choc.	wine	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Examples of Matrices: Documents and terms

	data	mining	classif.	tree	...
Paper#1	13	11	22	55	...
Paper#2	5	4	6	7	...
Paper#3	...	...	...	...	...
Paper#4	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Examples of Matrices: Authors and terms

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Examples of Matrices: sensor-ids and time-ticks

	temp1	temp2	humid.	pressure	...
t=1	13	11	22	55	...
t=2	5	4	6	7	...
t=3	...	...	...	...	...
t=4	...	...	...	...	...
...	...	...	...	...	...

Kumamoto U CMU CS

Motivation 2: Why tensors?

- Q: what is a tensor?

Kumamoto U CMU CS

Motivation 2: Why tensors?

- A: N-D generalization of matrix:

sigmod' 07	data	mining	query	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation 2: Why tensors?

- A: N-D generalization of matrix:

sigmod' 05
sigmod' 06
sigmod' 07

	data	mining	query	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Tensors are useful for 3 or more modes

Terminology: 'mode' (or 'aspect'):

Mode#3

data	mining	query	tree	...
13	11	22	55	...
5	4	6	7	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Mode#2

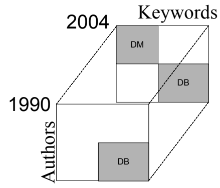
Mode (== aspect) #1

Motivating Applications

- Why matrices are important?
- Why tensors are useful?
  - P1: social networks
  - P2: web mining

P1: Social network analysis

- Traditionally, people focus on static networks and find community structures
- We plan to monitor the change of the community structure over time



P2: Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)
    - context-sensitive hypergraph analysis

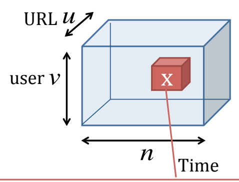
Tensors for time-series analysis

- Time-stamped events
  - *e.g., web clicks*

Time	URL	User
08-01-12:00	CNN.com	Smith
08-02-15:00	YouTube.com	Brown
08-02-19:00	CNET.com	Smith
08-03-11:00	CNN.com	Johnson
...	...	...

Represent as
M^th order tensor (M=3)

$\mathcal{X} \in \mathbb{N}^{u \times v \times n}$

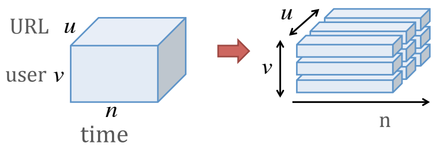


Element x: # of events

e.g., 'Smith', 'CNN.com', 'Aug 1, 10pm'; 21 times

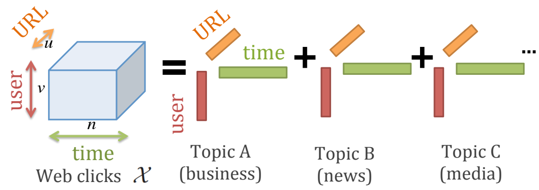
Tensors for time-series analysis

- Individual-sequence mining
- Create a set of (u * v) sequences of length (n)
- Apply the mining algorithm for each sequence



Tensors for time-series analysis

- Multi-aspect time-series analysis



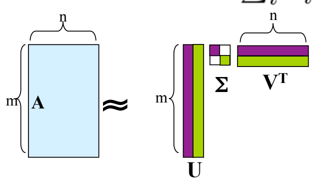
Outline

- Tensor decomposition
  - Motivation
  - Basic approaches
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

Multi-Aspect Non-linear Time-series

Reminder: SVD

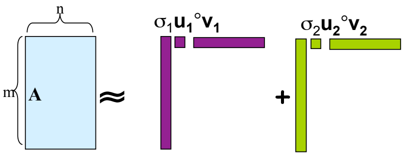
$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \textstyle\sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

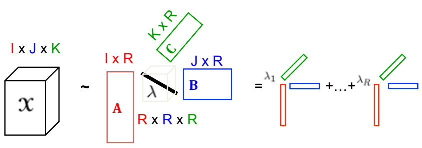
Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \textstyle\sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

Goal: extension to >=3 modes



$$\mathcal{X} \approx [\![\lambda\,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r\, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Kumamoto U — CMU CS

Main points:

- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)
- Details follow

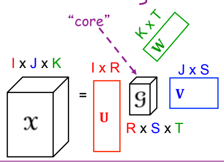
Kumamoto U — CMU CS

Specially Structured Tensors

- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathrm{U} \times_2 \mathrm{V} \times_3 \mathrm{W} = \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \equiv [\![\mathcal{G} \,; \mathrm{U}, \mathrm{V}, \mathrm{W}]\!]$$

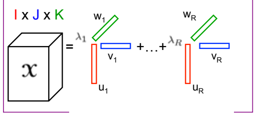
Our Notation



- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \equiv [\![\lambda \,; \mathrm{U}, \mathrm{V}, \mathrm{W}]\!]$$

Our Notation



Details — CMU CS

Specially Structured Tensors

- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathrm{U} \times_2 \mathrm{V} \times_3 \mathrm{W} = \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \equiv [\![\mathcal{G} \,; \mathrm{U}, \mathrm{V}, \mathrm{W}]\!]$$

In matrix form:

$$\mathrm{X}_{(1)} = \mathrm{U}\mathrm{G}_{(1)}(\mathrm{W} \otimes \mathrm{V})^{\mathsf{T}}$$
$$\mathrm{X}_{(2)} = \mathrm{V}\mathrm{G}_{(2)}(\mathrm{W} \otimes \mathrm{U})^{\mathsf{T}}$$
$$\mathrm{X}_{(3)} = \mathrm{W}\mathrm{G}_{(3)}(\mathrm{V} \otimes \mathrm{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathrm{W} \otimes \mathrm{V} \otimes \mathrm{U})\mathrm{vec}(\mathcal{G})$$

- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \equiv [\![\lambda \,; \mathrm{U}, \mathrm{V}, \mathrm{W}]\!]$$

In matrix form:

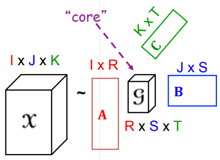
Let $\Lambda = \mathrm{diag}(\lambda)$

$$\mathrm{X}_{(1)} = \mathrm{U}\,\Lambda\,(\mathrm{W} \odot \mathrm{V})^{\mathsf{T}}$$
$$\mathrm{X}_{(2)} = \mathrm{V}\,\Lambda\,(\mathrm{W} \odot \mathrm{U})^{\mathsf{T}}$$
$$\mathrm{X}_{(3)} = \mathrm{W}\,\Lambda\,(\mathrm{V} \odot \mathrm{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathrm{W} \odot \mathrm{V} \odot \mathrm{U})\,\lambda$$

Kumamoto U — CMU CS

Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- $\mathcal{G}$: how groups relate to each other

Kumamoto U — CMU CS

Intuition behind core tensor

- 2-d case: co-clustering
- [Dhillon et al. Information-Theoretic Co-clustering, KDD'03]

Kumamoto U — CMU CS

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \end{array}$$

eg, terms x documents

$$\begin{array}{cc} & k \\ m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \end{array} \begin{array}{cc} & l \\ k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \end{array} \begin{array}{cc} & n \\ l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} \end{array} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

Kumamoto U CMU CS

med. doc cs doc

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

med. terms

cs terms

common terms

term group x doc. group

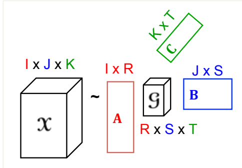
$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

doc x doc group

term x term-group

Kumamoto U CMU CS

Tucker Decomposition



$$\mathcal{X} \approx [\![\mathcal{G} \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Given A, B, C, the optimal core is:

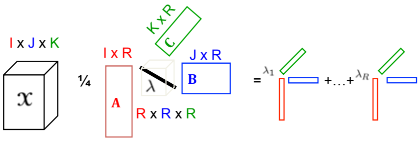
$$\mathcal{G} = [\![\mathcal{X} \,; \mathbf{A}^\dagger, \mathbf{B}^\dagger, \mathbf{C}^\dagger]\!]$$

- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal array decomposition
- **A**, **B**, and **C** generally assumed to be orthonormal (generally assume they have full column rank)
- $\mathcal{G}$ is not diagonal
- Not unique

Recall the equations for converting a tensor to a matrix

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathcal{G})$$

CANDECOMP/PARAFAC Decomposition



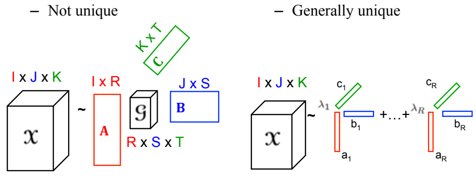
$$\mathcal{X} \approx [\![\lambda \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

- CANDECOMP = Canonical Decomposition (Carroll & Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- Core is diagonal (specified by the vector λ)
- Columns of **A**, **B**, and **C** are not orthonormal
- If R is *minimal*, then R is called the **rank** of the tensor (Kruskal 1977)
- Can have rank($\mathcal{X}$) > min{I,J,K}

Kumamoto U CMU CS

Tucker vs. PARAFAC Decompositions

- Tucker
  - Variable transformation in each mode
  - Core G may be dense
  - A, B, C generally orthonormal
  - Not unique
- PARAFAC
  - Sum of rank-1 components
  - No core, i.e., superdiagonal core
  - A, B, C may have linearly dependent columns
  - Generally unique



Kumamoto U CMU CS

Tensor tools - summary

- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- To solve: Alternating Least Squares

- Toolbox: from Tamara Kolda:

http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/

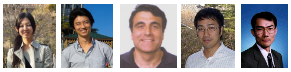
Kumamoto U CMU CS

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

Multi-Aspect Non-linear Time-series

M A N T



[Matsubara+ KDD'12]

Fast Mining and Forecasting of Complex Time-Stamped Events

Yasuko Matsubara (Kyoto University)
Yasushi Sakurai (NTT)
Christos Faloutsos (CMU)
Tomoharu Iwata (NTT)
Masatoshi Yoshikawa (Kyoto University)

Motivation

- Complex time-stamped events
  {timestamp + multiple attributes}

e.g., web click events:
{timestamp, URL, user ID, access devices, http referrer,…}

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	CNET.com	Smith	mac
2012-08-03-11:00	CNN.com	Johnson	ipad
…	…	…	…

Motivation

Q1. Are there any topics ?
- news, tech, media, sports, etc…

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	CNET.com	Smith	mac
2012-08-03-11:00	CNN.com	Johnson	ipad
…	…	…	…

e.g.,
CNN.com, CNET.com -> news topic
YouTube.com -> media topic

Motivation

Q2. Can we group URLs/users accordingly?

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	CNET.com	Smith	mac
2012-08-03-11:00	CNN.com	Johnson	ipad
…	…	…	…

e.g.,
CNN.com & CNET.com (related to news topic)
Smith & Johnson (related to news topic)

Motivation

Q3. Can we forecast future events?
- How many clicks from 'Smith' tomorrow?
- How many clicks to 'CNN.com' over next 7 days?

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	CNET.com	Smith	mac
2012-08-03-11:00	CNN.com	Johnson	ipad
2012-08-05-12:00	CNN.com	Smith	iphone
2012-08-05-19:00	CNET.com	Smith	iphone

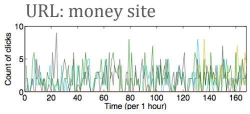
future clicks?

Motivation

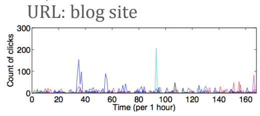
Web click events – can we see any trends?

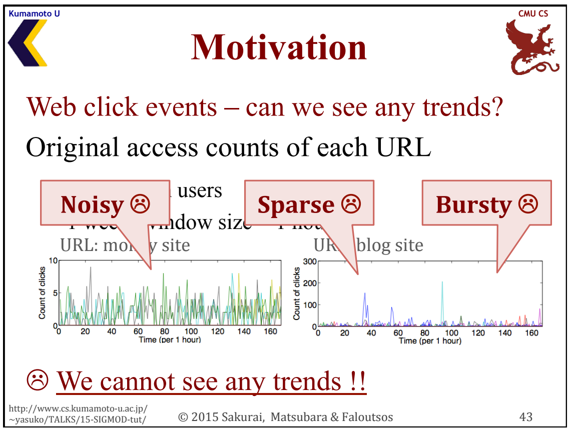
Original access counts of each URL
- 100 random users
- 1 week (window size = 1 hour)

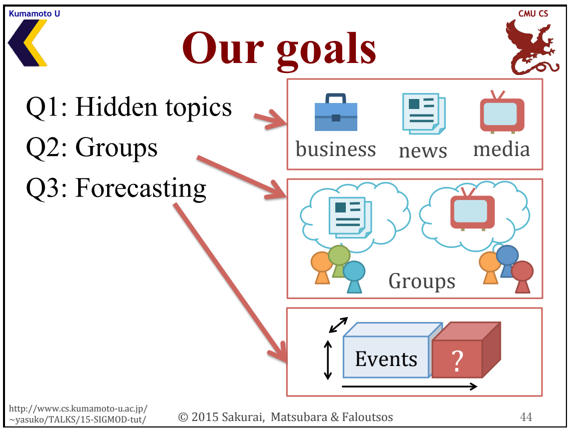
URL: money site

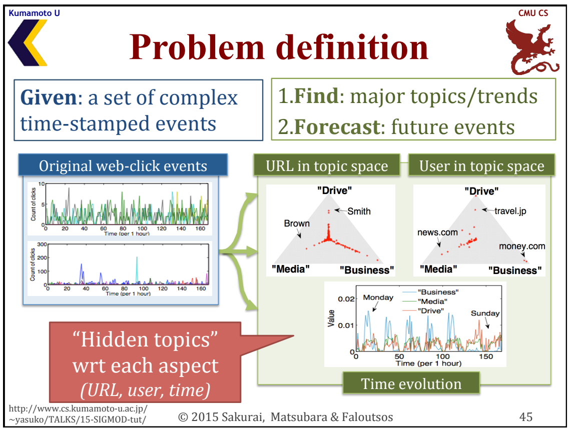


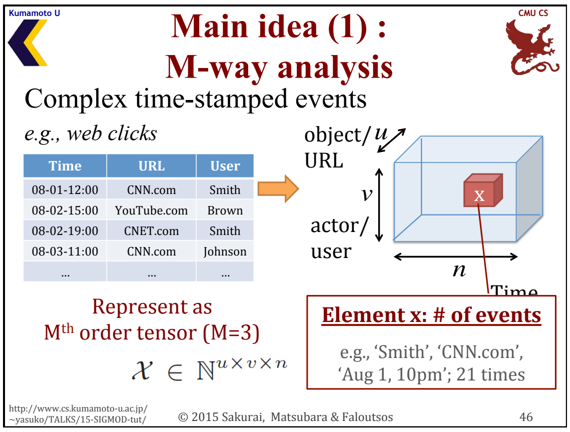
URL: blog site

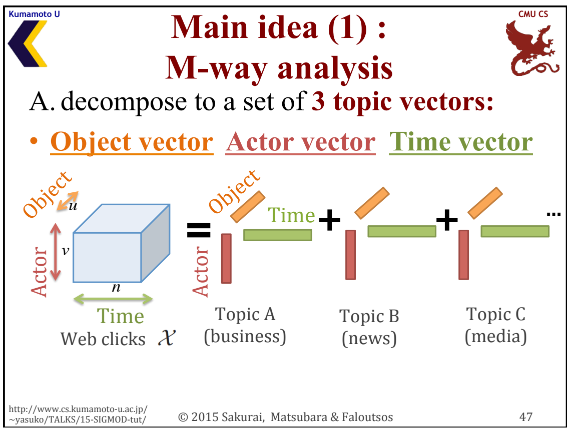


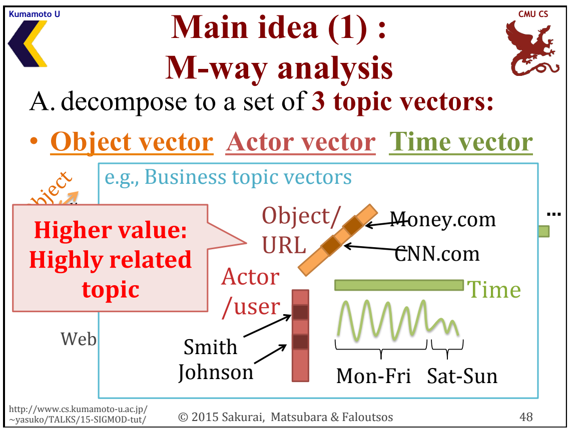

Kumamoto U
CMU CS
Motivation
Web click events – can we see any trends?
Original access counts of each URL
Noisy ☹
Sparse ☹
Bursty ☹
users
URL: money site
URL: blog site
Count of clicks
Time (per 1 hour)
☹ We cannot see any trends !!
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
43

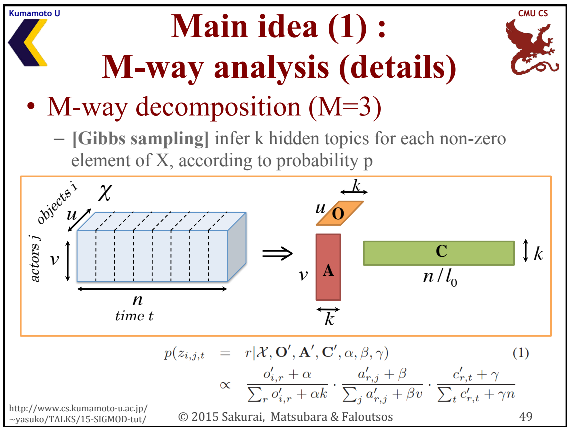

Kumamoto U
CMU CS
Our goals
Q1: Hidden topics
Q2: Groups
Q3: Forecasting
business
news
media
Groups
Events
?
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
44

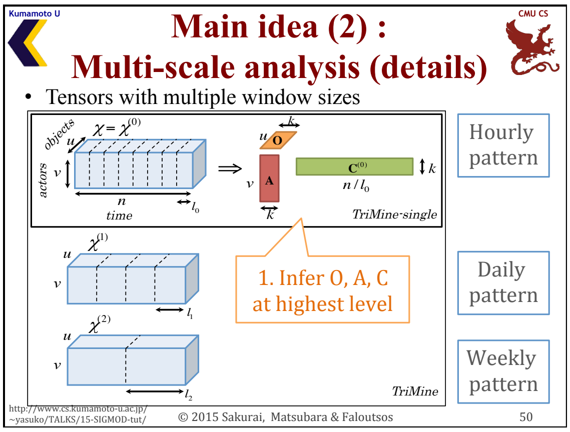

Kumamoto U
CMU CS
Problem definition
Given: a set of complex time-stamped events
1.Find: major topics/trends
2.Forecast: future events
Original web-click events
URL in topic space
User in topic space
"Drive"
Smith
Brown
"Media"
"Business"
travel.jp
news.com
money.com
Monday
Sunday
Time evolution
"Hidden topics" wrt each aspect (URL, user, time)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
45

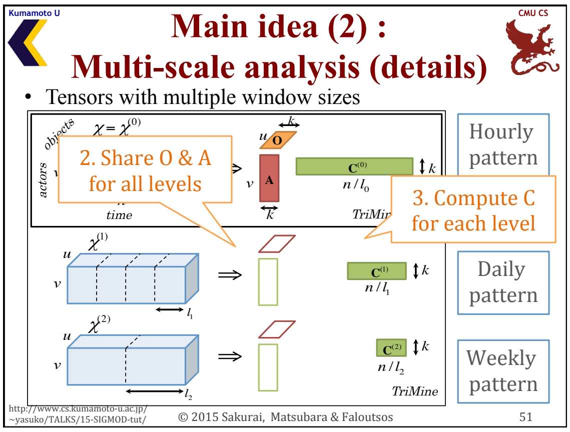

Kumamoto U
CMU CS
Main idea (1) : M-way analysis
Complex time-stamped events
e.g., web clicks
Time | URL | User
08-01-12:00 | CNN.com | Smith
08-02-15:00 | YouTube.com | Brown
08-02-19:00 | CNET.com | Smith
08-03-11:00 | CNN.com | Johnson
... | ... | ...
object/URL u
actor/user v
n
Time
x
Represent as M^{th} order tensor (M=3)
$\mathcal{X} \in \mathbb{N}^{u \times v \times n}$
Element x: # of events
e.g., 'Smith', 'CNN.com', 'Aug 1, 10pm'; 21 times
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
46

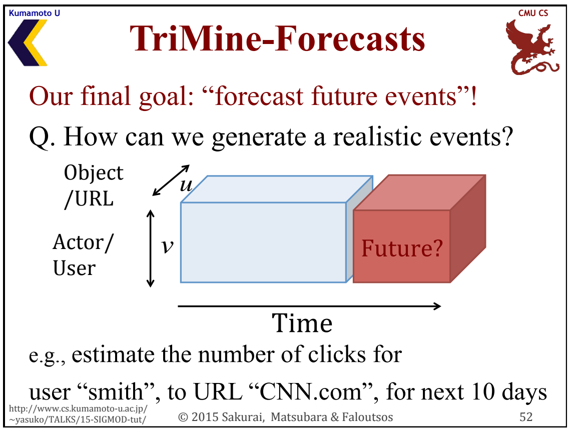

Kumamoto U
CMU CS
Main idea (1) : M-way analysis
A. decompose to a set of 3 topic vectors:
• Object vector Actor vector Time vector
Object u
Actor v
n
Time
Web clicks $\mathcal{X}$
=
Object
Actor
Time
+
+
...
Topic A (business)
Topic B (news)
Topic C (media)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
47

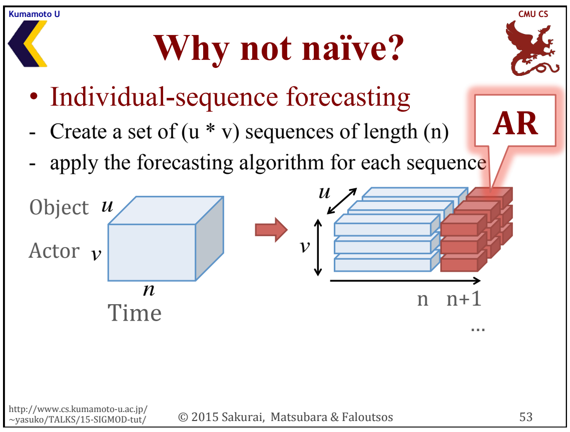

Kumamoto U
CMU CS
Main idea (1) : M-way analysis
A. decompose to a set of 3 topic vectors:
• Object vector Actor vector Time vector
e.g., Business topic vectors
Higher value: Highly related topic
Object/URL
Money.com
CNN.com
Actor /user
Time
Web
Smith
Johnson
Mon-Fri
Sat-Sun
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
48

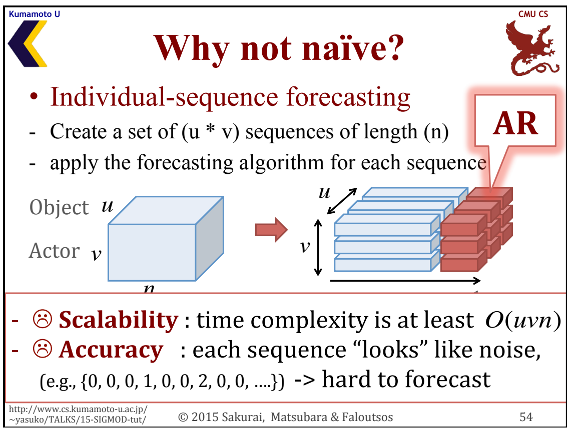

Kumamoto U
CMU CS
Main idea (1) :
M-way analysis (details)
• M-way decomposition (M=3)
– [Gibbs sampling] infer k hidden topics for each non-zero element of X, according to probability p
objects i
actors j
time t
n
u
v
k
O
A
C
n / l0
(1)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
49

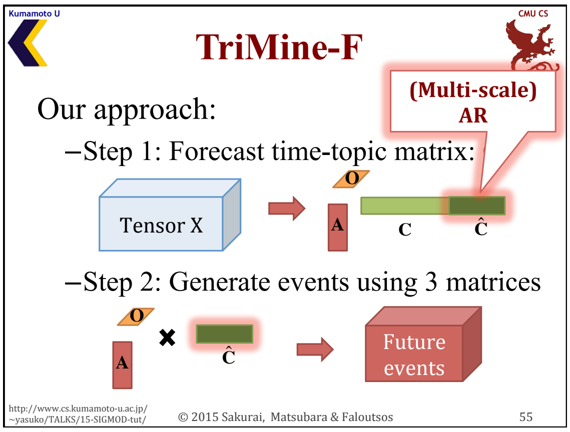

Kumamoto U
CMU CS
Main idea (2) :
Multi-scale analysis (details)
• Tensors with multiple window sizes
objects
actors
time
TriMine-single
Hourly pattern
Daily pattern
Weekly pattern
1. Infer O, A, C at highest level
TriMine
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
50

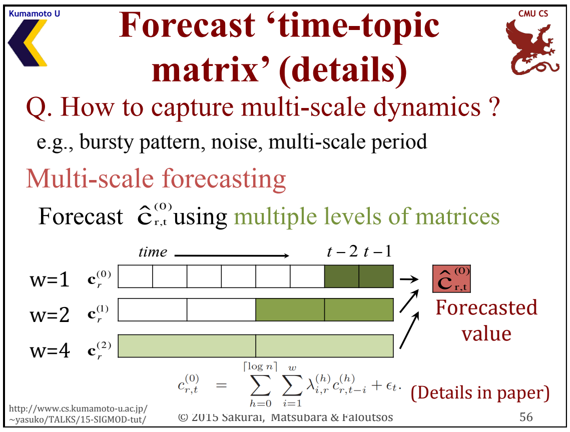

Kumamoto U
CMU CS
Main idea (2) :
Multi-scale analysis (details)
• Tensors with multiple window sizes
objects
actors
time
2. Share O & A for all levels
3. Compute C for each level
Hourly pattern
Daily pattern
Weekly pattern
TriMine
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
51

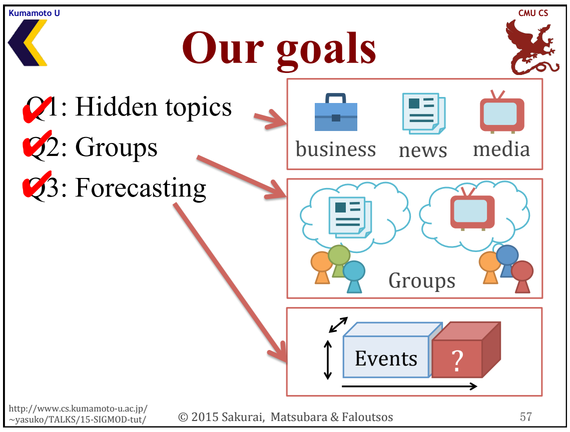

Kumamoto U
CMU CS
TriMine-Forecasts
Our final goal: "forecast future events"!
Q. How can we generate a realistic events?
Object /URL
Actor/ User
u
v
Future?
Time
e.g., estimate the number of clicks for
user "smith", to URL "CNN.com", for next 10 days
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
52

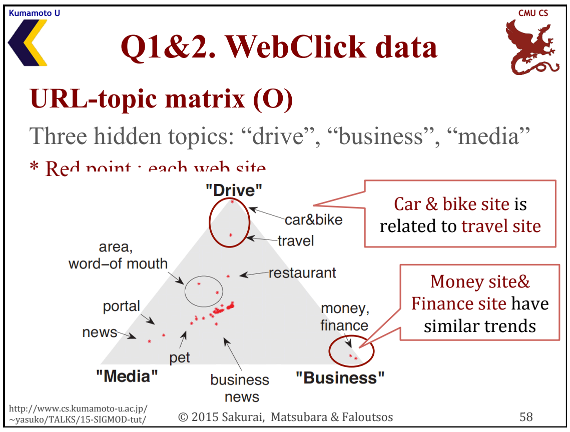

Kumamoto U
CMU CS
Why not naïve?
• Individual-sequence forecasting
- Create a set of (u * v) sequences of length (n)
- apply the forecasting algorithm for each sequence
AR
Object u
Actor v
n
Time
n n+1
...
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
53

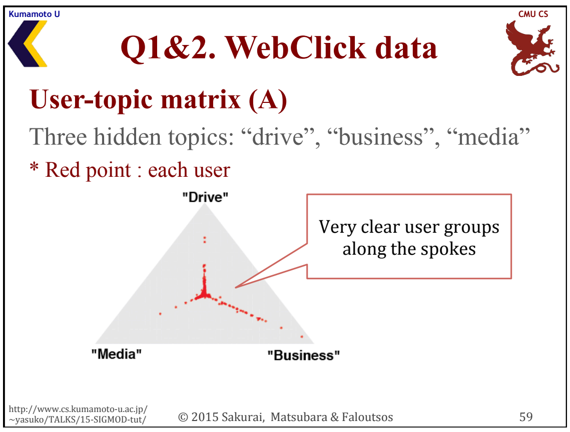

Kumamoto U
CMU CS
Why not naïve?
• Individual-sequence forecasting
- Create a set of (u * v) sequences of length (n)
- apply the forecasting algorithm for each sequence
AR
Object u
Actor v
- ☹ Scalability : time complexity is at least O(uvn)
- ☹ Accuracy : each sequence "looks" like noise,
(e.g., {0, 0, 0, 1, 0, 0, 2, 0, 0, ...}) -> hard to forecast
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
54

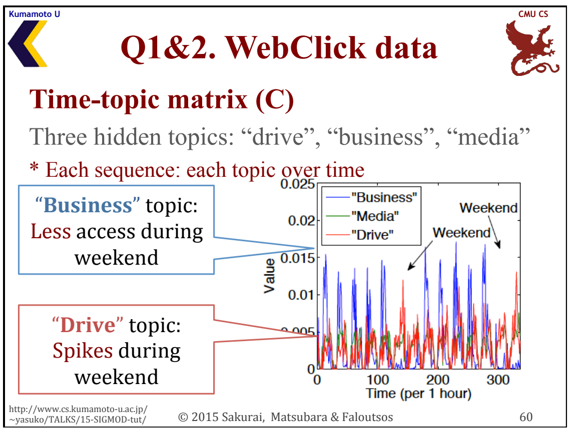

Kumamoto U
CMU CS
TriMine-F
Our approach:
–Step 1: Forecast time-topic matrix:
(Multi-scale) AR
Tensor X
O
A
C
Ĉ
–Step 2: Generate events using 3 matrices
O
A
Ĉ
Future events
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
55


Kumamoto U
CMU CS
Forecast 'time-topic matrix' (details)
Q. How to capture multi-scale dynamics ?
e.g., bursty pattern, noise, multi-scale period
Multi-scale forecasting
Forecast $\hat{c}^{(0)}_{r,t}$ using multiple levels of matrices
time
$t-2$ $t-1$
w=1 $c^{(0)}_r$
w=2 $c^{(1)}_r$
w=4 $c^{(2)}_r$
$\hat{c}^{(0)}_{r,t}$
Forecasted value
$c^{(0)}_{r,t} = \sum_{h=0}^{\lceil \log n \rceil} \sum_{i=1}^{w} \lambda^{(h)}_{i,r} c^{(h)}_{r,t-i} + \epsilon_t.$
(Details in paper)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
56


Kumamoto U
CMU CS
Our goals
Q1: Hidden topics
Q2: Groups
Q3: Forecasting
business
news
media
Groups
Events
?
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
57


Kumamoto U
CMU CS
Q1&2. WebClick data
URL-topic matrix (O)
Three hidden topics: "drive", "business", "media"
* Red point : each web site
"Drive"
car&bike
travel
area, word–of mouth
restaurant
portal
news
pet
money, finance
business news
"Media"
"Business"
Car & bike site is related to travel site
Money site& Finance site have similar trends
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
58


Kumamoto U
CMU CS
Q1&2. WebClick data
User-topic matrix (A)
Three hidden topics: "drive", "business", "media"
* Red point : each user
"Drive"
"Media"
"Business"
Very clear user groups along the spokes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
59

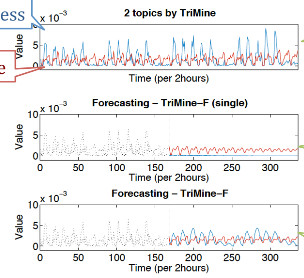

Kumamoto U
CMU CS
Q1&2. WebClick data
Time-topic matrix (C)
Three hidden topics: "drive", "business", "media"
* Each sequence: each topic over time
"Business" topic: Less access during weekend
"Drive" topic: Spikes during weekend
"Business"
"Media"
"Drive"
Weekend
Weekend
Value
0.025
0.02
0.015
0.01
0.005
0
0
100
200
300
Time (per 1 hour)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
60

Q3. Forecasting accuracy

- Benefit of multiple time-scale forecasting

business

drive



Original sequence of matrix (C)

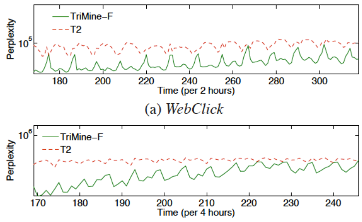
Forecast C' using single level -> failed

Multi-scale forecast -> captured cyclic patterns

Q3. Forecasting accuracy

Temporal perplexity (entropy for each time-tick)

Lower perplexity: higher predictive accuracy



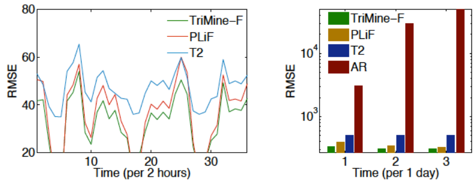
(a) *WebClick*

(b) *Ondemand TV* T2: [Hong et al. KDD'11]

Q3. Forecasting accuracy

Accuracy of event forecasting

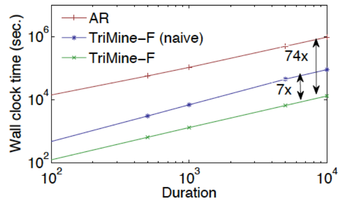
RMSE between original and forecasted events (lower is better)



PLiF [Li et al.VLDB'10] , T2: [Hong et al.KDD'11]

Q3. Scalability

- Computation cost (vs. AR)

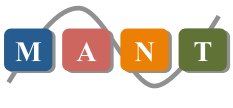


- TriMine provides a reduction in computation time (up to 74x)

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

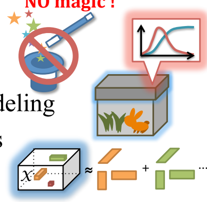
Multi-Aspect Non-linear Time-series



Non-linear tensor analysis

New research directions

1. Automatic mining
2. Non-linear (gray-box) modeling
3. Large-scale tensor analysis

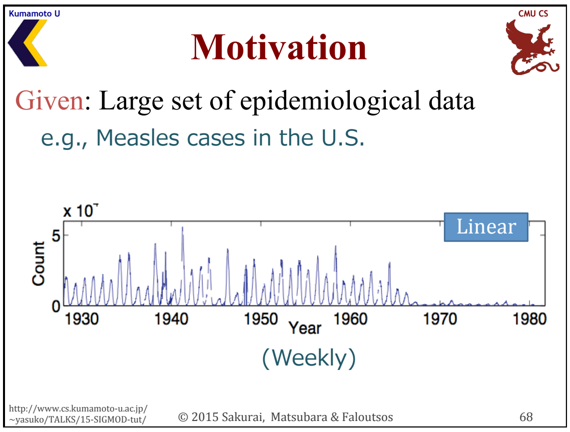


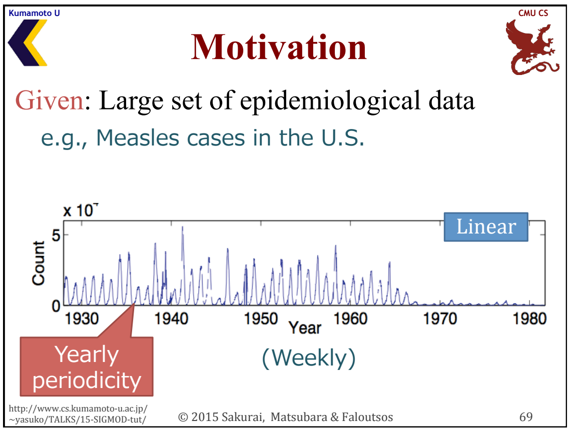
Put all together

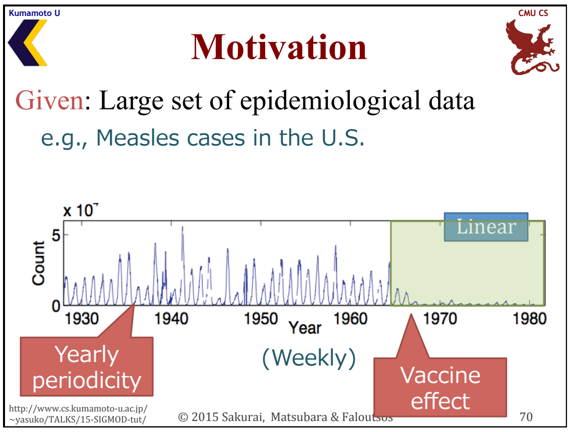
New challenge: MANT analysis

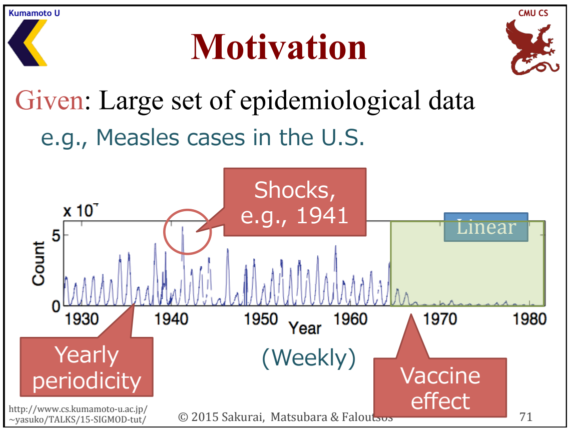
Multi-Aspect Non-linear Time-series

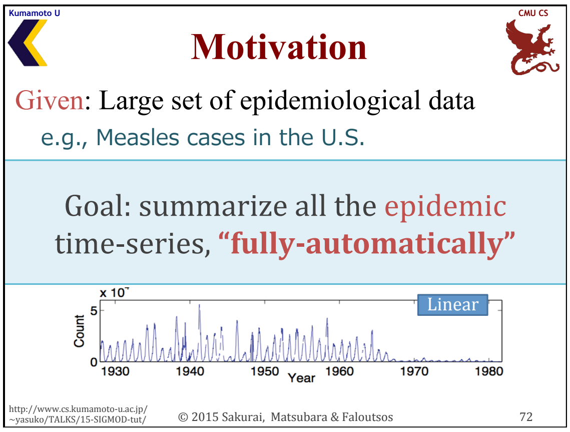

[Matsubara+ KDD'14]
FUNNEL: Automatic Mining of Spatially Coevolving Epidemics
Yasuko Matsubara, Yasushi Sakurai (Kumamoto University)
Willem G. van Panhuis (University of Pittsburgh)
Christos Faloutsos (CMU)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
67

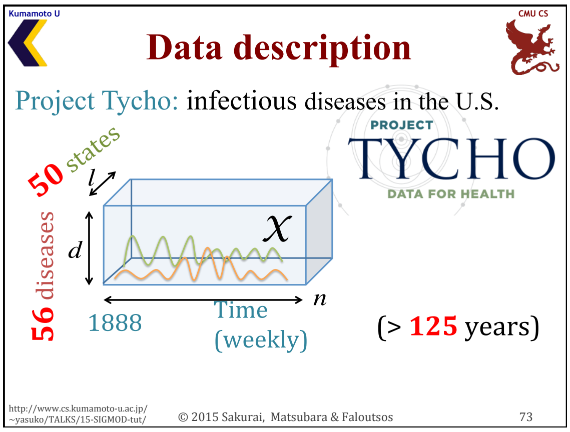

Kumamoto U
CMU CS
Motivation
Given: Large set of epidemiological data
e.g., Measles cases in the U.S.
Linear
Count
Year
1930
1940
1950
1960
1970
1980
(Weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
68

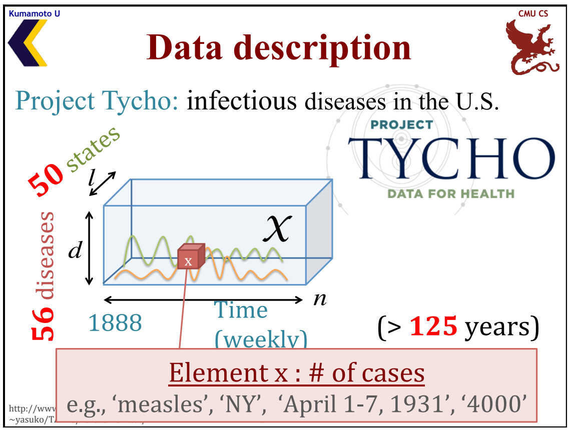

Kumamoto U
CMU CS
Motivation
Given: Large set of epidemiological data
e.g., Measles cases in the U.S.
Linear
Count
Year
(Weekly)
Yearly periodicity
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
69

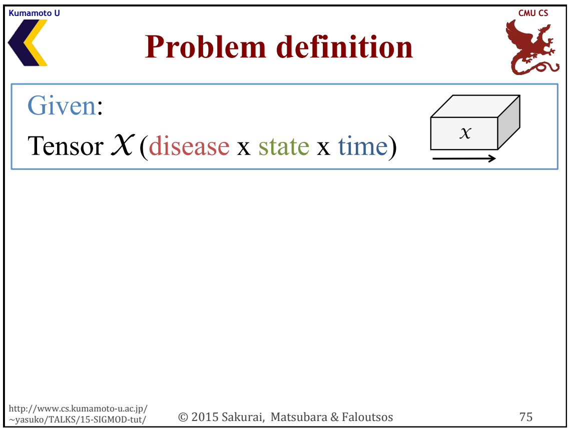

Kumamoto U
CMU CS
Motivation
Given: Large set of epidemiological data
e.g., Measles cases in the U.S.
Linear
Count
Year
(Weekly)
Yearly periodicity
Vaccine effect
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
70

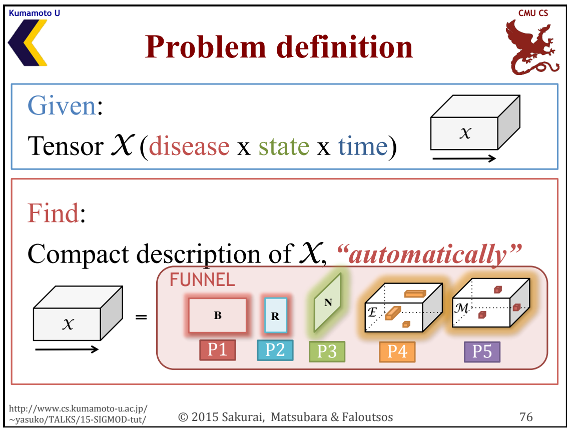

Kumamoto U
CMU CS
Motivation
Given: Large set of epidemiological data
e.g., Measles cases in the U.S.
Shocks, e.g., 1941
Linear
Count
Year
(Weekly)
Yearly periodicity
Vaccine effect
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
71

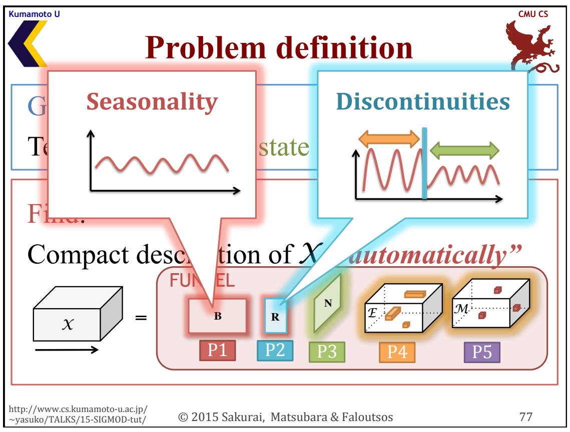

Kumamoto U
CMU CS
Motivation
Given: Large set of epidemiological data
e.g., Measles cases in the U.S.
Goal: summarize all the epidemic time-series, "fully-automatically"
Linear
Count
Year
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
72

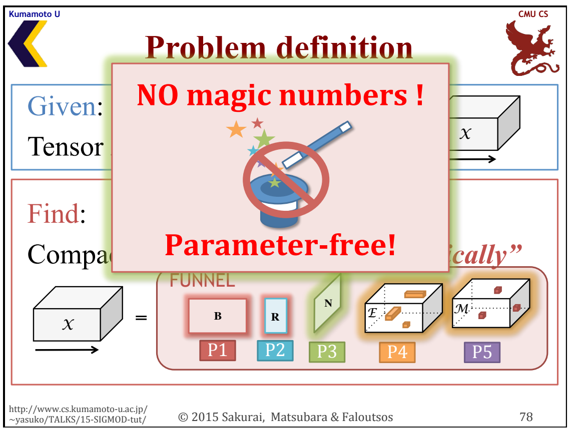

Kumamoto U
CMU CS
Data description
Project Tycho: infectious diseases in the U.S.
PROJECT
TYCHO
DATA FOR HEALTH
50 states
l
56 diseases
d
X
n
1888
Time
(weekly)
(> 125 years)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
73

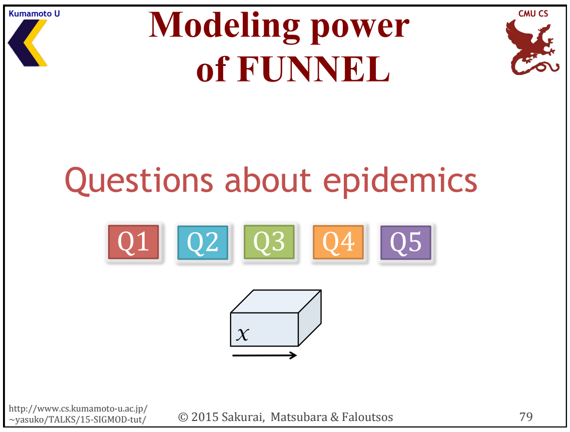

Kumamoto U
CMU CS
Data description
Project Tycho: infectious diseases in the U.S.
PROJECT
TYCHO
DATA FOR HEALTH
50 states
l
56 diseases
d
X
x
n
1888
Time
(weekly)
(> 125 years)
Element x : # of cases
e.g., 'measles', 'NY', 'April 1-7, 1931', '4000'

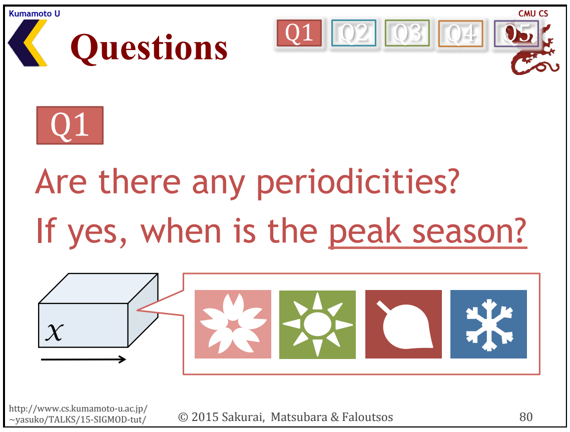

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
X
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
75

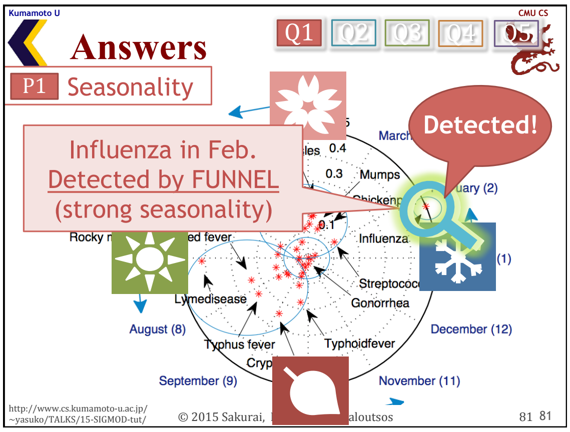

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
Find:
Compact description of X, "automatically"
X
=
FUNNEL
B
R
N
P1
P2
P3
P4
P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
76

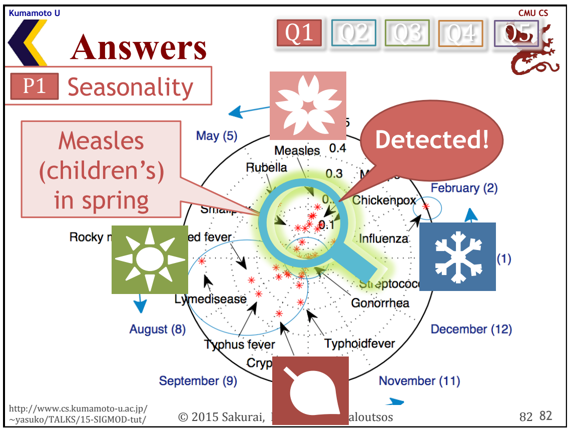

Kumamoto U
CMU CS
Problem definition
Seasonality
Discontinuities
state
Compact desc tion of X automatically"
FUN EL
B
R
N
P1
P2
P3
P4
P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
77

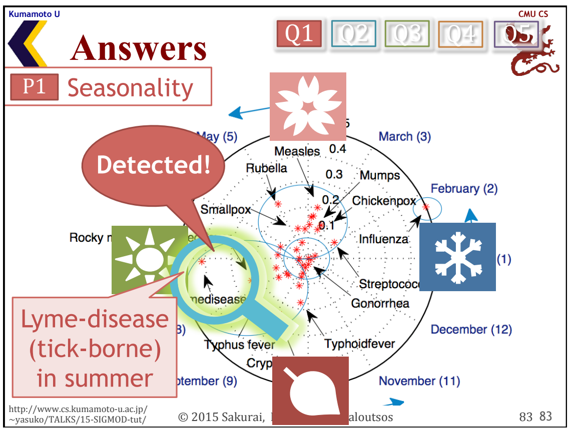

Kumamoto U
CMU CS
Problem definition
Given:
Tensor
NO magic numbers !
Parameter-free!
Find:
Compa
cally"
FUNNEL
B
R
N
P1
P2
P3
P4
P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
78

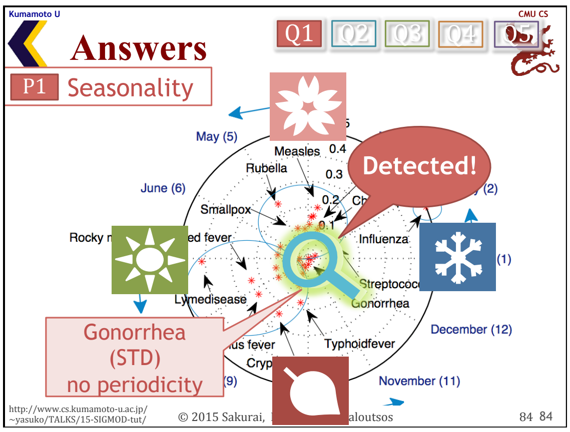

Kumamoto U
CMU CS
Modeling power
of FUNNEL
Questions about epidemics
Q1 Q2 Q3 Q4 Q5
x
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
79

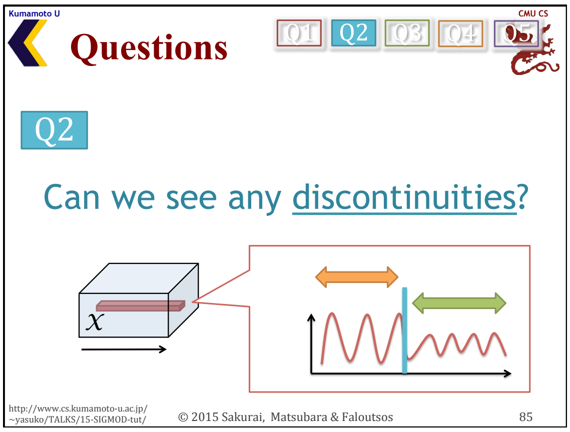

Kumamoto U
CMU CS
Questions
Q1 Q2 Q3 Q4 Q5
Q1
Are there any periodicities?
If yes, when is the peak season?
x
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
80

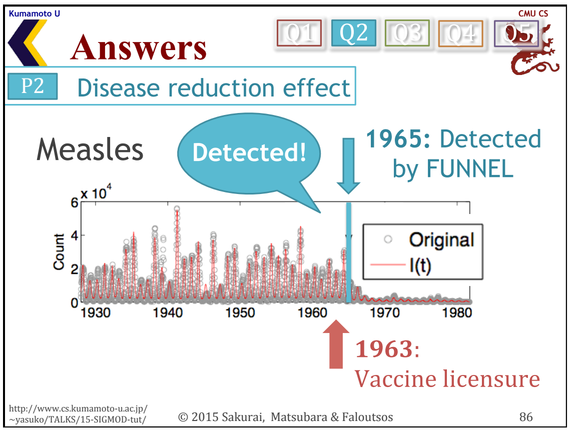

Kumamoto U
CMU CS
Answers
Q1 Q2 Q3 Q4 Q5
P1 Seasonality
Influenza in Feb.
Detected by FUNNEL
(strong seasonality)
Detected!
Mumps
Influenza
Streptococ
Gonorrhea
Typhoidfever
Typhus fever
Lymedisease
August (8)
September (9)
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai,
81 81

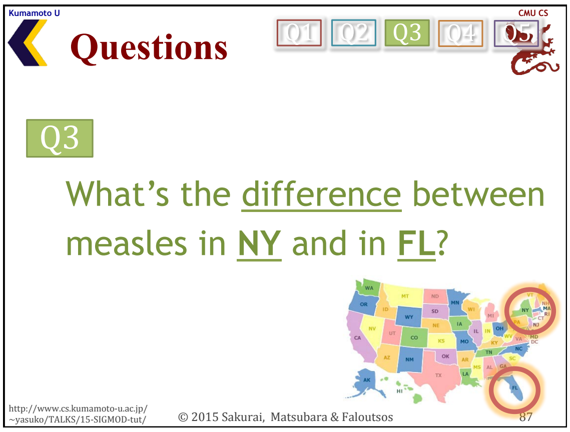

Kumamoto U
CMU CS
Answers
Q1 Q2 Q3 Q4 Q5
P1 Seasonality
Measles
(children's)
in spring
Detected!
May (5)
Measles
Rubella
February (2)
Chickenpox
Influenza
Gonorrhea
Typhoidfever
Typhus fever
Lymedisease
August (8)
September (9)
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai,
82 82

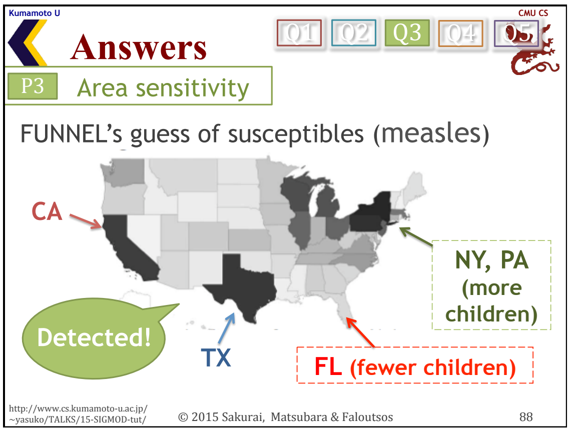

Kumamoto U
CMU CS
Answers
Q1 Q2 Q3 Q4 Q5
P1 Seasonality
Detected!
Lyme-disease
(tick-borne)
in summer
March (3)
February (2)
Measles
Rubella
Mumps
Chickenpox
Smallpox
Influenza
Streptococ
Gonorrhea
Typhoidfever
Typhus fever
December (12)
November (11)
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai,
83 83

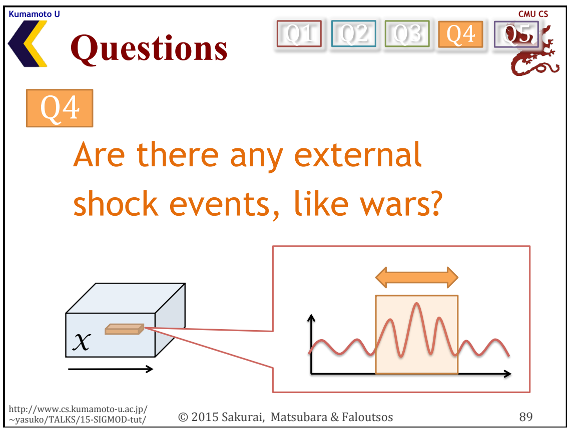

Kumamoto U
CMU CS
Answers
Q1 Q2 Q3 Q4 Q5
P1 Seasonality
Detected!
Gonorrhea
(STD)
no periodicity
May (5)
June (6)
Measles
Rubella
Smallpox
Influenza
Streptococ
Gonorrhea
Typhoidfever
Lymedisease
December (12)
November (11)
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai,
84 84

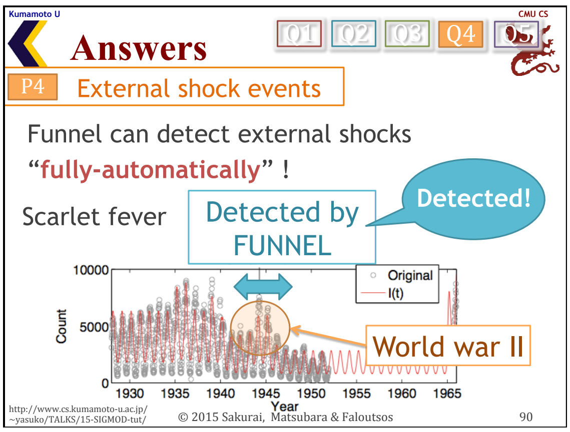

Kumamoto U
Questions
Q1 Q2 Q3 Q4 Q5
CMU CS
Q2
Can we see any discontinuities?
x
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
85

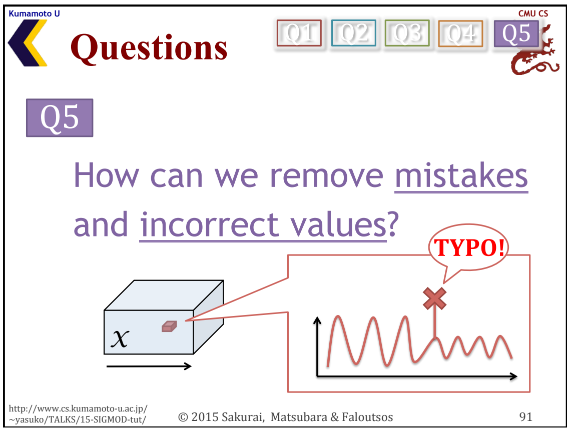

Kumamoto U
Answers
Q1 Q2 Q3 Q4 Q5
CMU CS
P2 Disease reduction effect
Measles
Detected!
1965: Detected by FUNNEL
x 10^4
Count
Original
I(t)
1930 1940 1950 1960 1970 1980
1963: Vaccine licensure
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
86

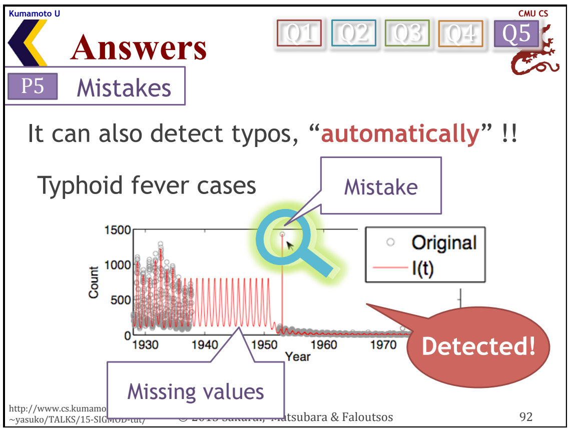

Kumamoto U
Questions
Q1 Q2 Q3 Q4 Q5
CMU CS
Q3
What's the difference between measles in NY and in FL?
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
87

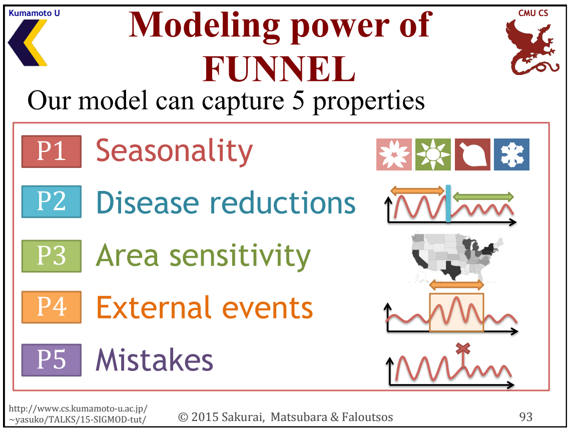

Kumamoto U
Answers
Q1 Q2 Q3 Q4 Q5
CMU CS
P3 Area sensitivity
FUNNEL's guess of susceptibles (measles)
CA
NY, PA (more children)
Detected!
TX
FL (fewer children)
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
88

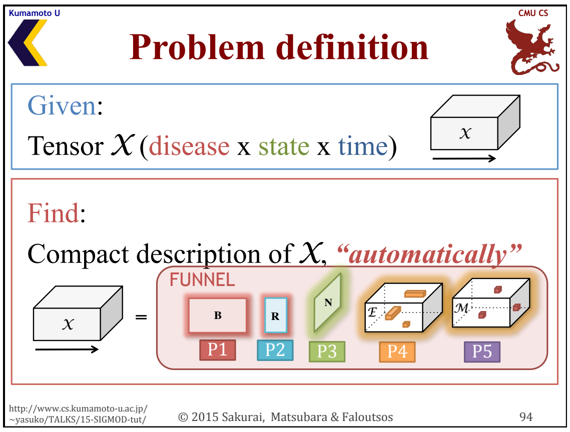

Kumamoto U
Questions
Q1 Q2 Q3 Q4 Q5
CMU CS
Q4
Are there any external shock events, like wars?
x
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
89

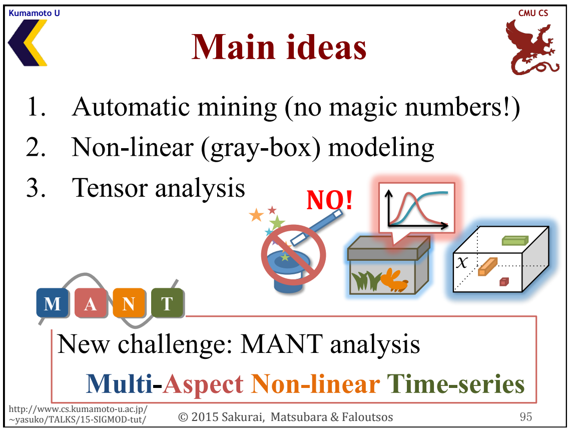

Kumamoto U
Answers
Q1 Q2 Q3 Q4 Q5
CMU CS
P4 External shock events
Funnel can detect external shocks "fully-automatically" !
Scarlet fever
Detected by FUNNEL
Detected!
10000
5000
0
Count
Original
I(t)
World war II
1930 1935 1940 1945 1950 1955 1960 1965
Year
http://www.cs.kumamoto-u.ac.jp/
~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
90

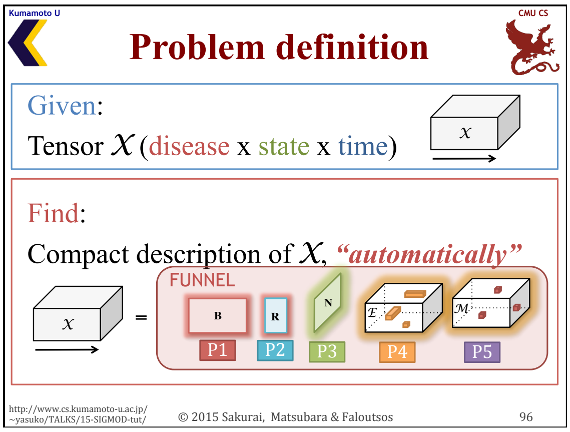

Kumamoto U
CMU CS
Questions
Q1 Q2 Q3 Q4 Q5
Q5
How can we remove mistakes and incorrect values?
TYPO!
x
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
91

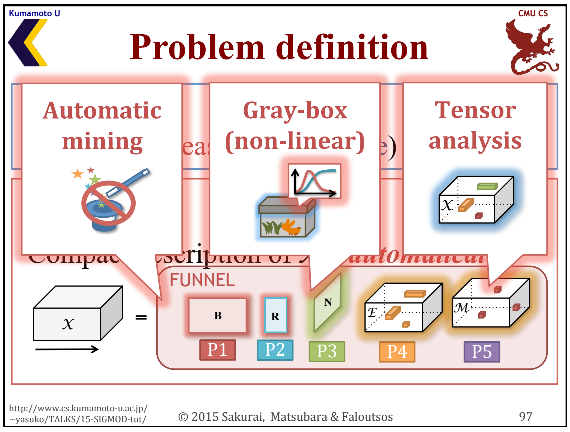

Kumamoto U
CMU CS
Answers
Q1 Q2 Q3 Q4 Q5
P5 Mistakes
It can also detect typos, "automatically" !!
Typhoid fever cases
Mistake
Original
I(t)
Count
Year
Detected!
Missing values
92

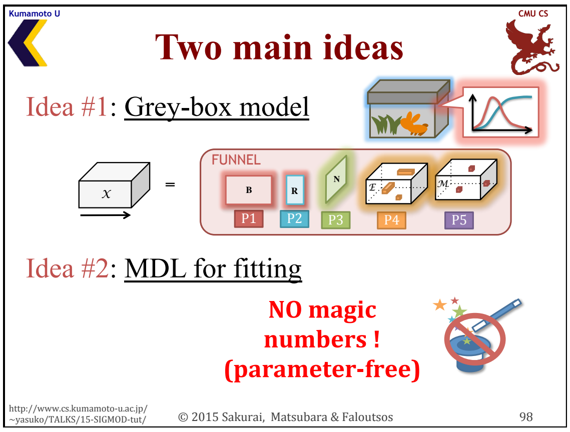

Kumamoto U
CMU CS
Modeling power of FUNNEL
Our model can capture 5 properties
P1 Seasonality
P2 Disease reductions
P3 Area sensitivity
P4 External events
P5 Mistakes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
93

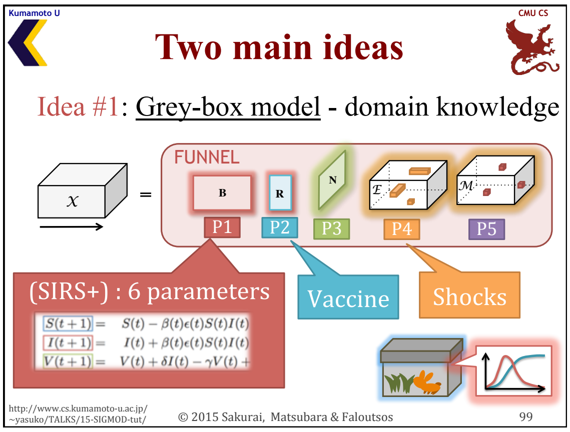

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
Find:
Compact description of X, "automatically"
FUNNEL
B R N
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
94

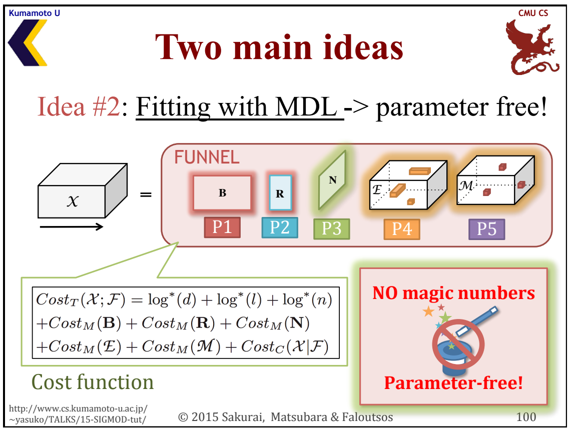

Kumamoto U
CMU CS
Main ideas
1. Automatic mining (no magic numbers!)
2. Non-linear (gray-box) modeling
3. Tensor analysis
NO!
M A N T
New challenge: MANT analysis
Multi-Aspect Non-linear Time-series
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
95

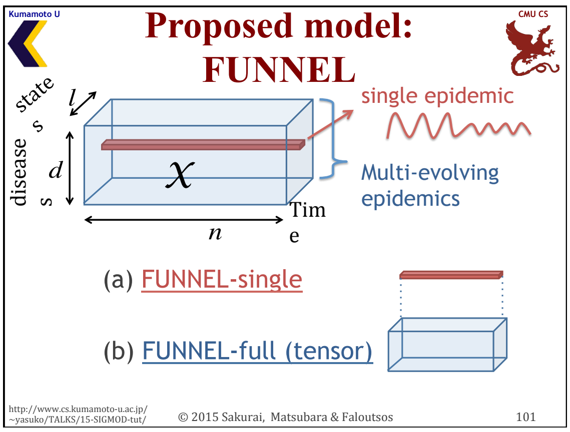

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
Find:
Compact description of X, "automatically"
FUNNEL
B R N
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
96

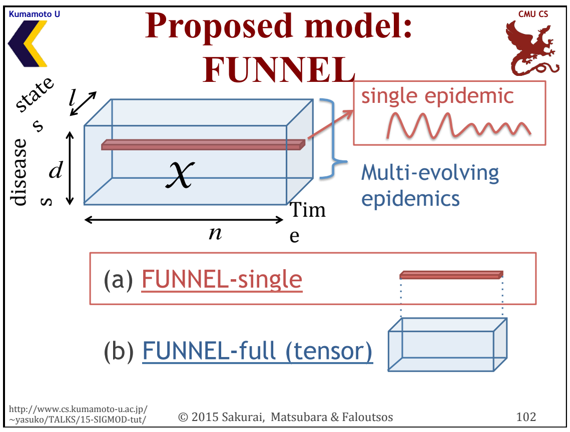

Kumamoto U
CMU CS
Problem definition
Automatic mining
Gray-box (non-linear)
Tensor analysis
FUNNEL
B
R
N
P1
P2
P3
P4
P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
97

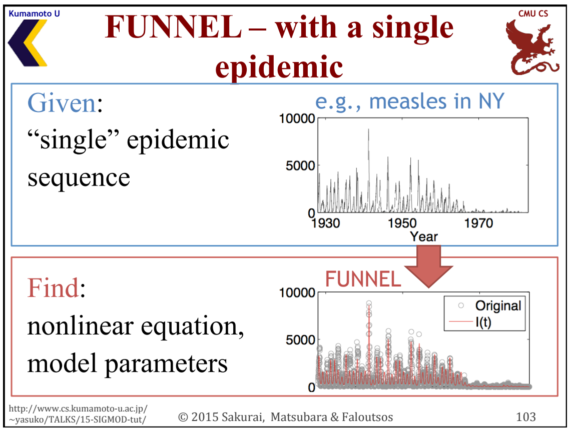

Kumamoto U
CMU CS
Two main ideas
Idea #1: Grey-box model
FUNNEL
B
R
N
P1
P2
P3
P4
P5
Idea #2: MDL for fitting
NO magic numbers !
(parameter-free)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
98


Kumamoto U
CMU CS
Two main ideas
Idea #1: Grey-box model - domain knowledge
FUNNEL
B
R
N
P1
P2
P3
P4
P5
(SIRS+) : 6 parameters
Vaccine
Shocks
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
99

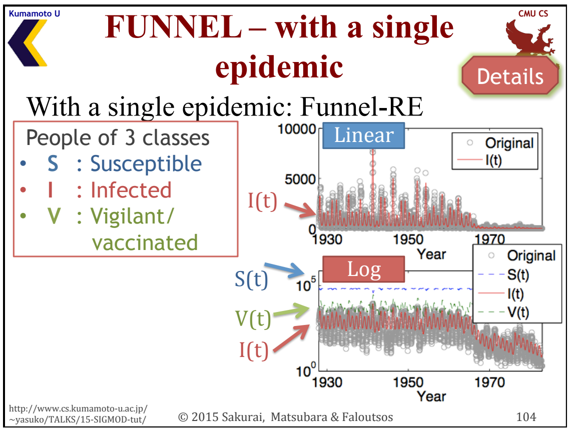

Kumamoto U
CMU CS
Two main ideas
Idea #2: Fitting with MDL -> parameter free!
FUNNEL
B
R
N
P1
P2
P3
P4
P5
$Cost_T(\mathcal{X};\mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n)$
$+Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N})$
$+Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$
Cost function
NO magic numbers
Parameter-free!
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
100


Kumamoto U
CMU CS
Proposed model: FUNNEL
states
diseases
Time
single epidemic
Multi-evolving epidemics
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
101

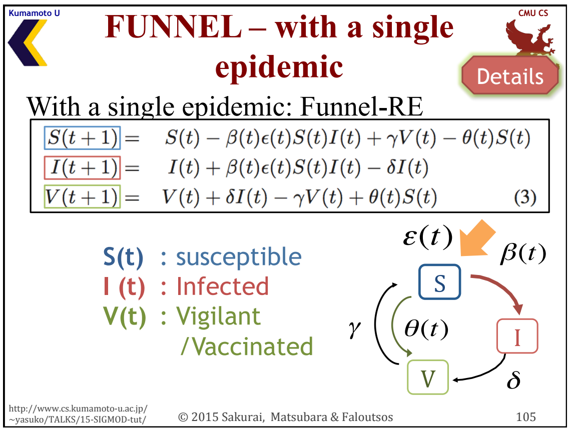

Kumamoto U
CMU CS
Proposed model: FUNNEL
states
diseases
Time
single epidemic
Multi-evolving epidemics
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
102



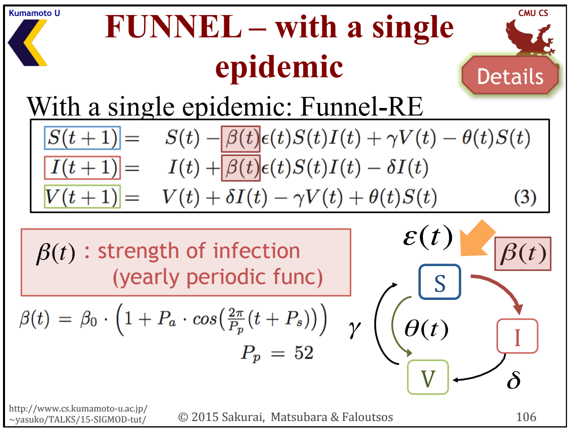
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Given:
"single" epidemic sequence
e.g., measles in NY
10000
5000
0
1930 1950 1970
Year
Find:
nonlinear equation, model parameters
FUNNEL
Original
I(t)
10000
5000
0
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
103



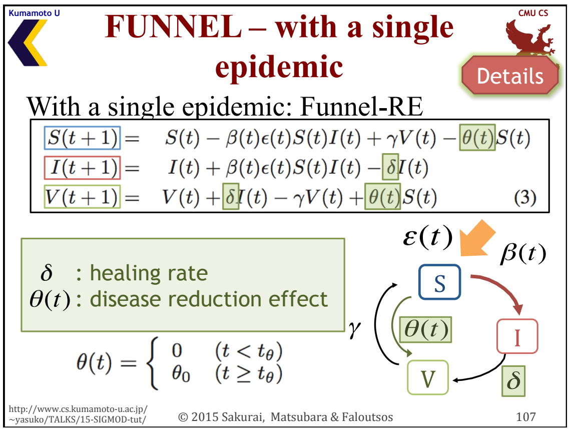
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
People of 3 classes
• S : Susceptible
• I : Infected
• V : Vigilant/ vaccinated
Linear
Original
I(t)
10000
5000
0
I(t)
1930 1950 1970
Year
Log
S(t)
V(t)
I(t)
Original
S(t)
I(t)
V(t)
10^5
10^0
1930 1950 1970
Year
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
104



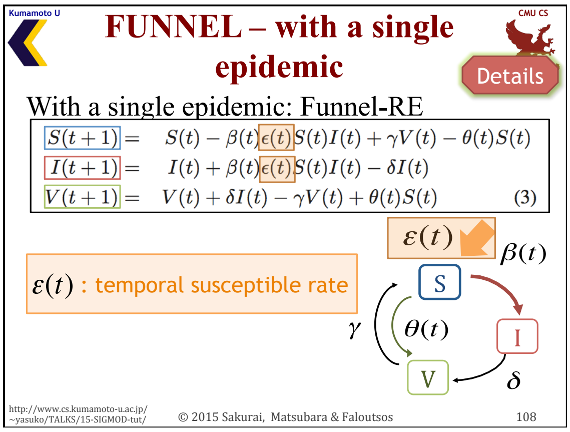
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
S(t) : susceptible
I (t) : Infected
V(t) : Vigilant /Vaccinated
$\varepsilon(t)$ $\beta(t)$ S I V γ $\theta(t)$ δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
105



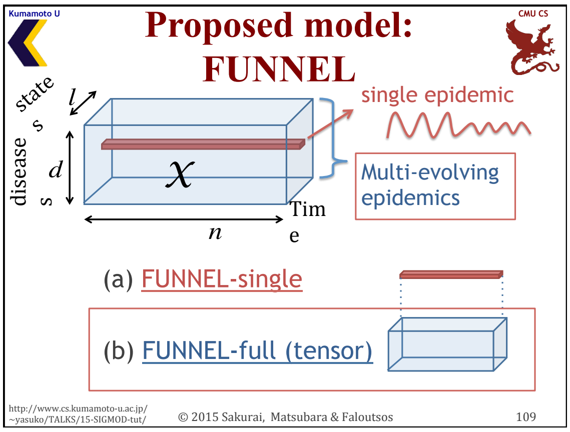
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
$\beta(t)$: strength of infection (yearly periodic func)
$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)$
$P_p = 52$
$\varepsilon(t)$ $\beta(t)$ S I V γ $\theta(t)$ δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
106



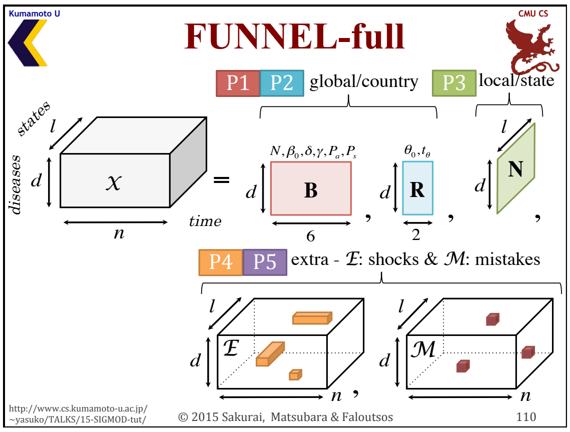
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
δ : healing rate
$\theta(t)$: disease reduction effect
$\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$
$\varepsilon(t)$ $\beta(t)$ S I V γ $\theta(t)$ δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
107



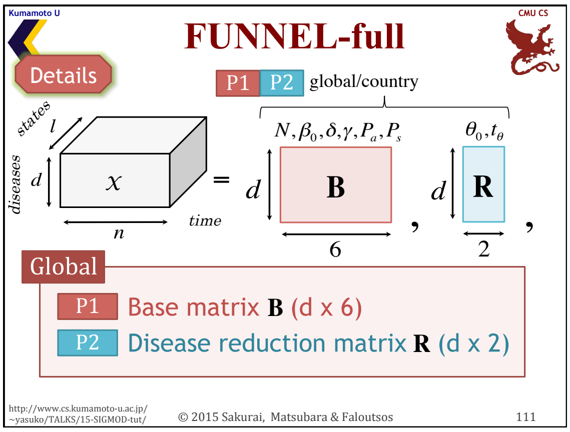
Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
$\varepsilon(t)$: temporal susceptible rate
$\varepsilon(t)$ $\beta(t)$ S I V γ $\theta(t)$ δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
108



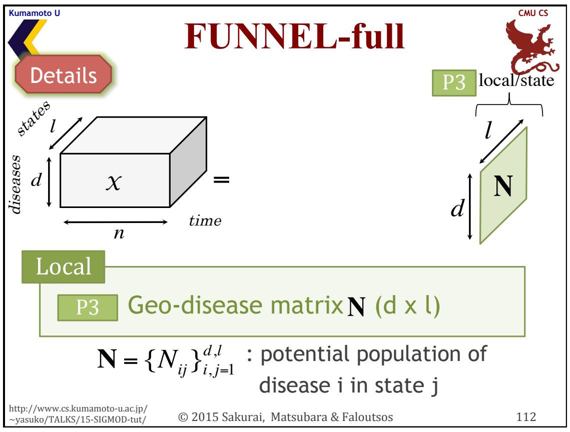
Kumamoto U
CMU CS
Proposed model: FUNNEL
single epidemic
Multi-evolving epidemics
states
diseases
Time
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
109



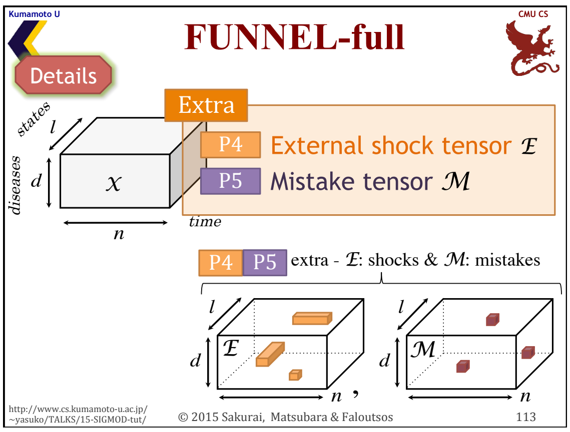
Kumamoto U
CMU CS
FUNNEL-full
P1 P2 global/country
P3 local/state
states
diseases
time
$N, \beta_0, \delta, \gamma, P_a, P_s$
θ_0, t_θ
B
R
N
P4 P5 extra - $\mathcal{E}$: shocks & $\mathcal{M}$: mistakes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
110



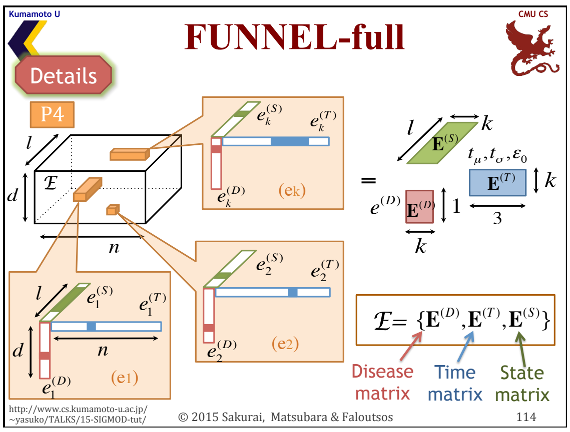
Kumamoto U
CMU CS
FUNNEL-full
Details
P1 P2 global/country
states
diseases
time
$N, \beta_0, \delta, \gamma, P_a, P_s$
θ_0, t_θ
B
R
Global
P1 Base matrix B (d x 6)
P2 Disease reduction matrix R (d x 2)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
111



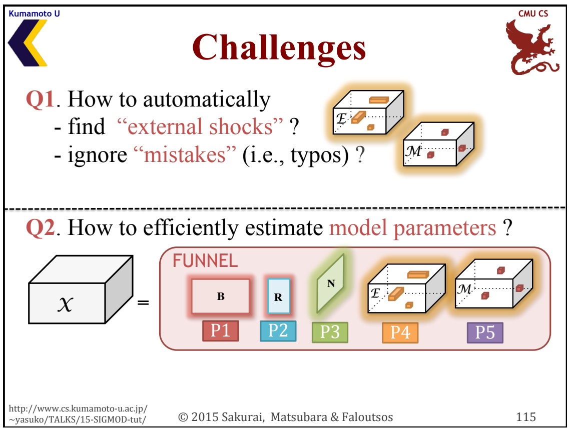
Kumamoto U
CMU CS
FUNNEL-full
Details
P3 local/state
states
diseases
time
N
Local
P3 Geo-disease matrix N (d x l)
$\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$: potential population of disease i in state j
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
112

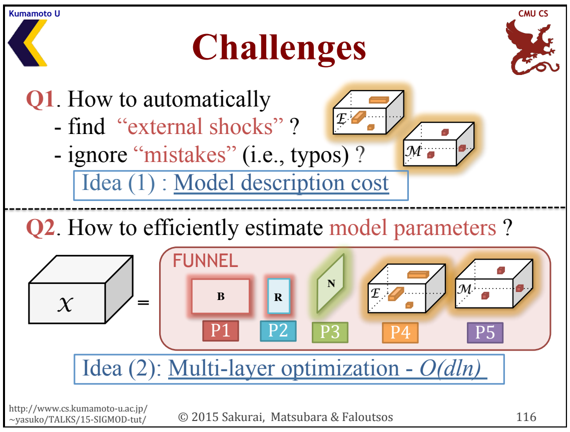


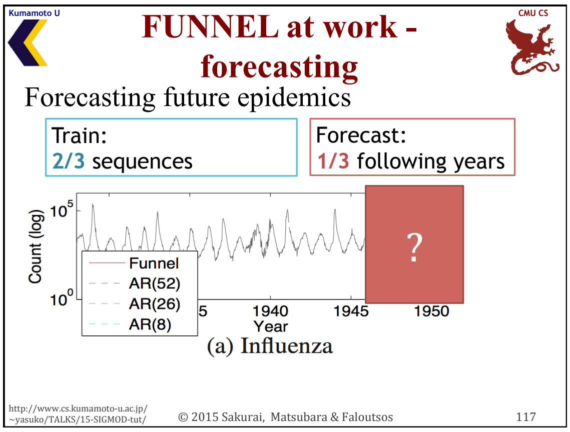
Kumamoto U
CMU CS
FUNNEL-full
Details
Extra
P4 External shock tensor $\mathcal{E}$
P5 Mistake tensor $\mathcal{M}$
states
diseases
time
P4 P5 extra - $\mathcal{E}$: shocks & $\mathcal{M}$: mistakes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
113

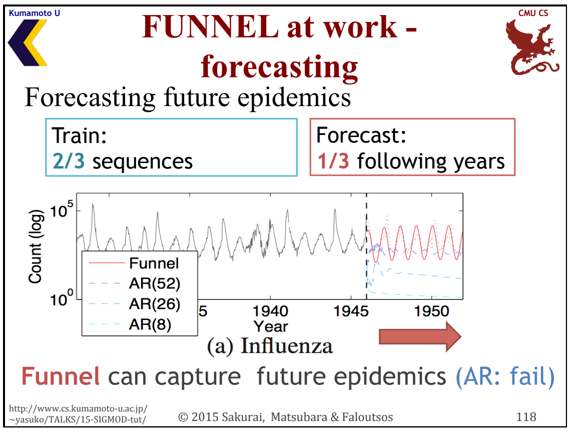


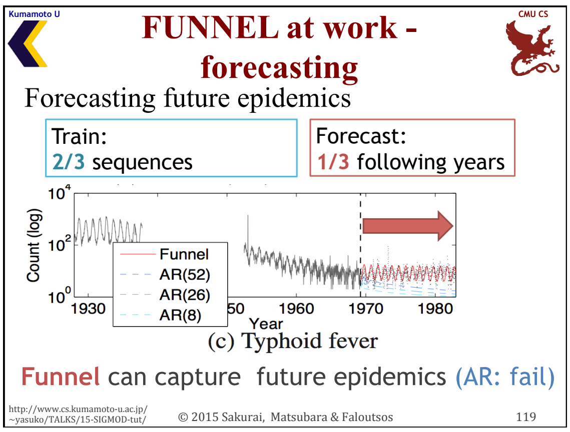
Kumamoto U
CMU CS
FUNNEL-full
Details
P4
(ek)
(e1)
(e2)
$t_\mu, t_\sigma, \varepsilon_0$
$\mathcal{E} = \{\mathbf{E}^{(D)}, \mathbf{E}^{(T)}, \mathbf{E}^{(S)}\}$
Disease matrix
Time matrix
State matrix
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
114

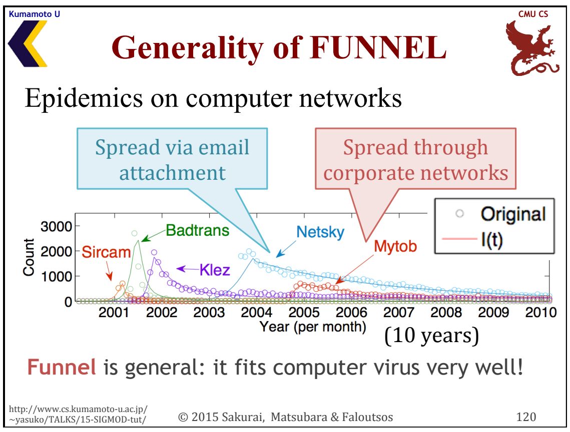

Kumamoto U
CMU CS
Challenges
Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?
Q2. How to efficiently estimate model parameters ?
FUNNEL
P1
P2
P3
P4
P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
115


Kumamoto U
CMU CS
Challenges
Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?
Idea (1) : Model description cost
Q2. How to efficiently estimate model parameters ?
FUNNEL
P1
P2
P3
P4
P5
Idea (2): Multi-layer optimization - O(dln)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
116


Kumamoto U
CMU CS
FUNNEL at work - forecasting
Forecasting future epidemics
Train: 2/3 sequences
Forecast: 1/3 following years
Count (log)
Funnel
AR(52)
AR(26)
AR(8)
1940
1945
1950
Year
?
(a) Influenza
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
117


Kumamoto U
CMU CS
FUNNEL at work - forecasting
Forecasting future epidemics
Train: 2/3 sequences
Forecast: 1/3 following years
Count (log)
Funnel
AR(52)
AR(26)
AR(8)
1940
1945
1950
Year
(a) Influenza
Funnel can capture future epidemics (AR: fail)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
118


Kumamoto U
CMU CS
FUNNEL at work - forecasting
Forecasting future epidemics
Train: 2/3 sequences
Forecast: 1/3 following years
Count (log)
Funnel
AR(52)
AR(26)
AR(8)
1930
1960
1970
1980
Year
(c) Typhoid fever
Funnel can capture future epidemics (AR: fail)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
119


Kumamoto U
CMU CS
Generality of FUNNEL
Epidemics on computer networks
Spread via email attachment
Spread through corporate networks
Original
I(t)
Count
Sircam
Badtrans
Klez
Netsky
Mytob
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year (per month)
(10 years)
Funnel is general: it fits computer virus very well!
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/
© 2015 Sakurai, Matsubara & Faloutsos
120

Kumamoto U | CMU CS

Part 3 Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Matrices and tensors provide elegant theory and algorithms
- MANT analysis

Multi-Aspect Non-linear Time-series

http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/ © 2015 Sakurai, Matsubara & Faloutsos 121

Kumamoto U | CMU CS

References

- Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha: Information-theoretic co-clustering. KDD 2003: 89-98
- T. G. Kolda, B. W. Bader and J. P. Kenny. *Higher-Order Web Link Analysis Using Multilinear Algebra*. In: ICDM 2005, Pages 242-249, November 2005.
- Jimeng Sun, Spiros Papadimitriou, Philip Yu. *Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams*, Proc. of the Int. Conf. on Data Mining (ICDM), Hong Kong, China, Dec 2006

http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/ © 2015 Sakurai, Matsubara & Faloutsos 122

Part 3

Extension of time-series: tensor analysis

Yasushi Sakurai (Kumamoto University)
Yasuko Matsubara (Kumamoto University)
Christos Faloutsos (Carnegie Mellon University)

http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/ © 2015 Sakurai, Matsubara & Faloutsos 123