Information-theoretic Learning and Data Mining

Kenji Yamanishi
The University of Tokyo, JAPAN

Oct. 18th, 2009

Contents

1. Information-theoretic Learning and Novelty Detection

Data Mining Issues

Mining from large amount of dynamic and heterogeneous data is a critical issue

Data

WEB

logs

cellar

RFID

media

sensor

Dynamic and Non-stationary Data

Distributed and Heterogeneous Data

Data MIning

＝Machine Learnig + Pre/Post Processing

Patten Discovery

Future Prediction

Anomaly Detection

Knowledge Discovery

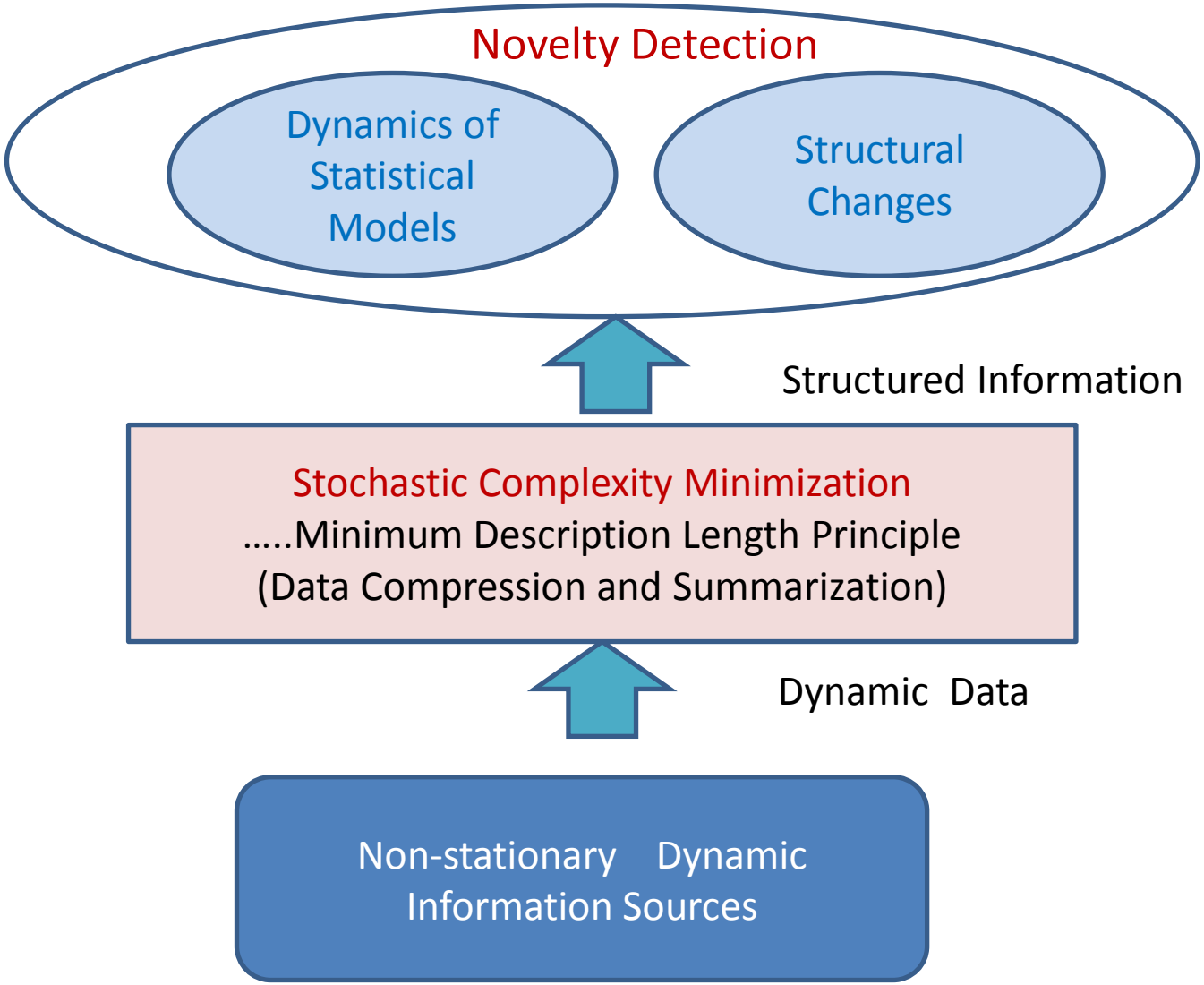
Marketing, CRM, SCM , ERP, Community Mgmnt.

Risk Managment

Fraud Detection, Failure Detection, Compliance Mgmnt, Security

Information-theoretic Learning and Data Mining

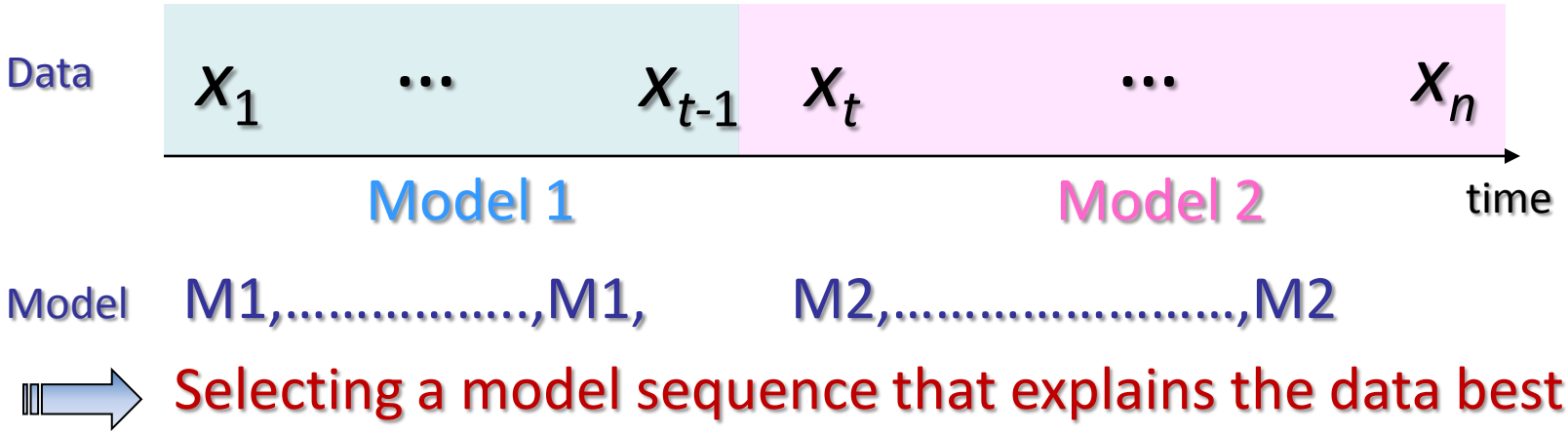
Structural changes from dynamic data are discovered using stochastic complexity minimization



2. Theory of Dynamic Model Selection

Concept of Dynamic Model Selection

Issue : How do you track changes of statistical models *under non-stationary environments*?



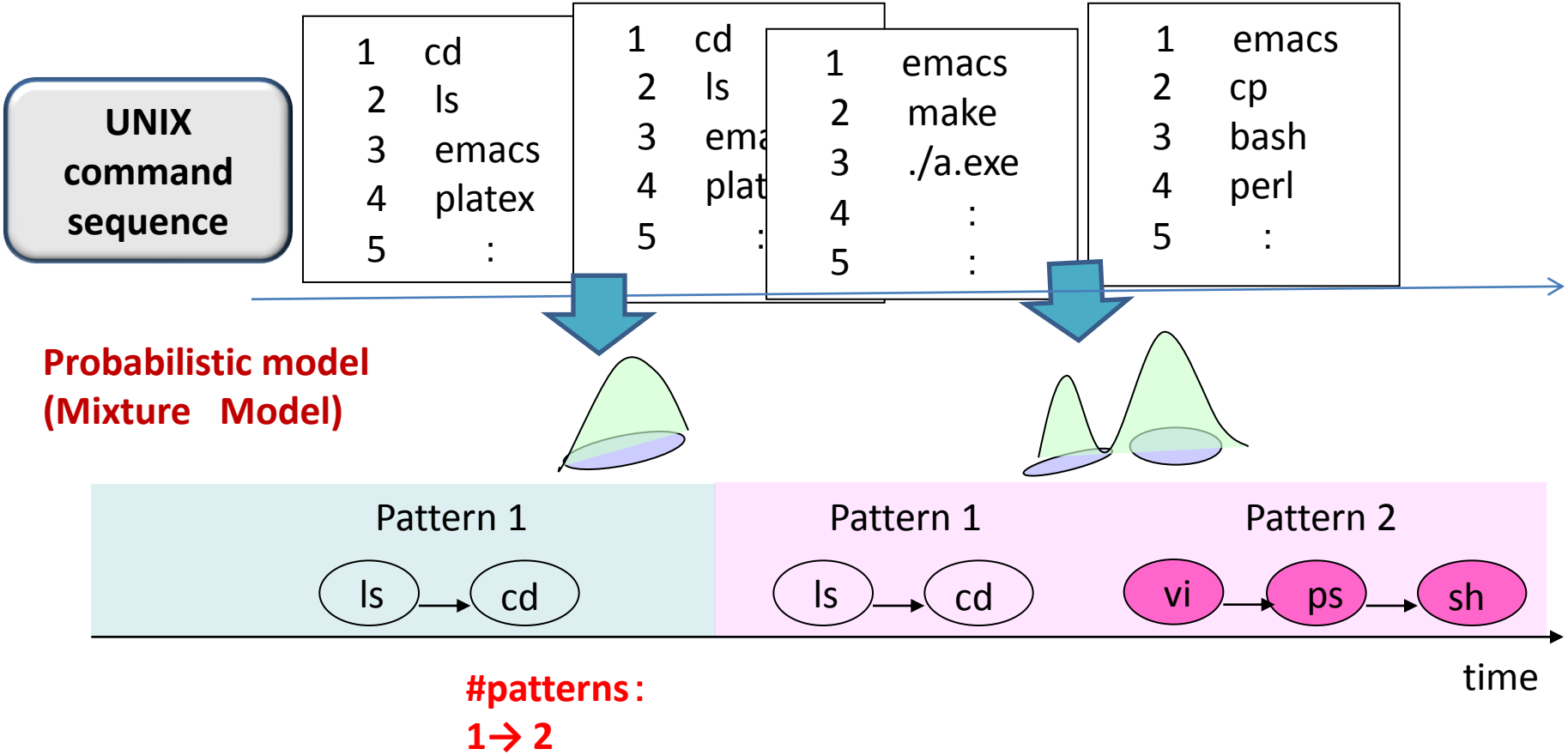
Selecting a model sequence that explains the data best

Methodologies

- Discounting Learning
- Windowing
- Dynamic Model Selection(DMS)

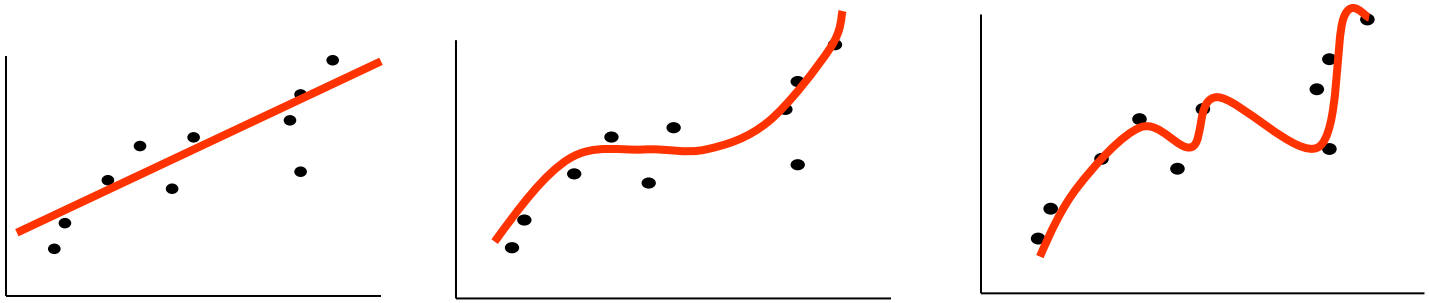
Ex.1: UNIX Command Pattern Analysis

Tracking changes of the mixture size leads to detection of a new pattern



Ex.2 Curve Fitting

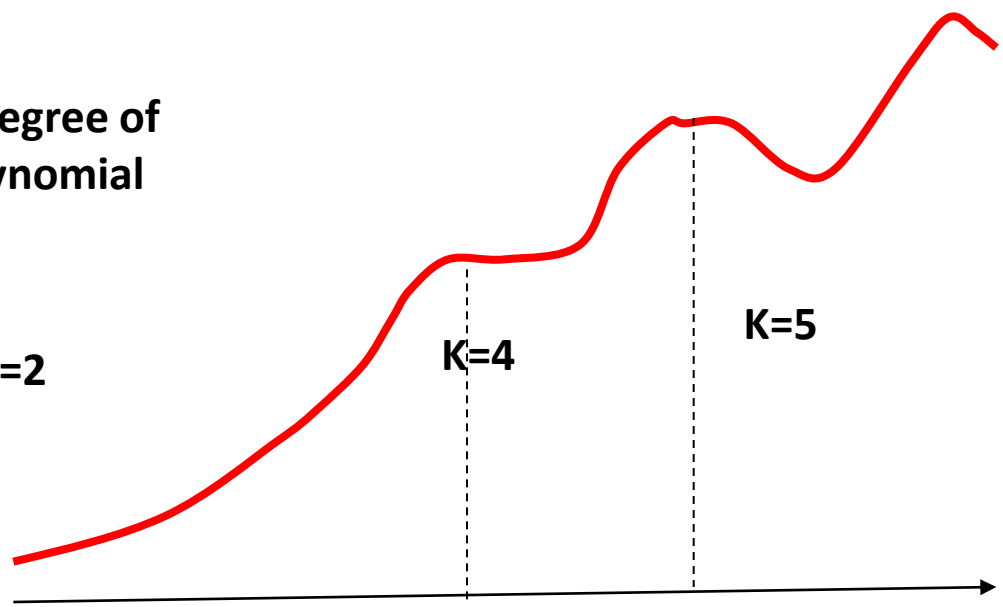
Static Model Selection



$\theta_0 + \theta_1 x$ $\qquad$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ $\qquad$ $\theta_0 + \theta_1 x + \cdots + \theta_d x^d$

Dynamic Model Selection

K: degree of polynomial



Degree of curve changes over time

Tracking curve degrees which will change over time

Related Work

- **Static model selection**

  AIC [Akaike 1973], BIC [Schwarz 1978],
  MDL [Rissanen 1978], MML [Wallace & Freeman 1978]

  – Under stationarity assumption

- **Tracking the best expert** [Herbster & Warmuth 1998]

  Derandomization [Vovk 1997][Cesanianchi Lugosi 2006]

  – Best model (="Best expert") may change over time

  – Cannot track the sequence of best models itself

- MDL-Based Dynamic Model Selection under non-stationarity assumption [Yamanishi and Maruyama 2007]

- **Switching** [Ervan, Grunwald, Rooij 2008]

  – Convergence as with MDL and convergence rate as with AIC

- **Changing dependency detection** [Fearnhead& Liu 2007]

MDL-based Information Criterion for DMS

Evaluating goodness of model sequence from the

Minimum Description Length (MDL) Principle [Rissanen 1978]

$$x^n = x_1 \cdots x_n$$ **Data Sequence**

$$k^n = k_1 \cdots k_n$$ **Model Sequence**

$$\ell_A(x^n : k^n)$$ **Code-length required by *A* to encode *x^n* and *k^n* under prefix condition**

$$\sum_{x^n, k^n} 2^{-\ell(x^n, k^n)} \leq 1$$ **(Kraft's Inequality)**

Minimize w.r.t. k^n for given x^n

- **Batch DMS**
- **Sequential DMS**

Predictive Distribution

Dynamics of model changes is described using predictive dist.

Class of probability models

$$\mathcal{P}_k = \{P(x^n|\theta, k) : \ \theta \in \Theta_k\} \quad (n = 1, 2, \cdots)$$

$$\dim \Theta_1 < \cdots < \dim \Theta_k < \dim \Theta_{k+1} < \cdots$$

k: **model**

Predictive distribution of x_t given x^{t-1}:

$$P(x_t|x^{t-1} : k_t) = P(x_t|\hat{\theta}_{t-1} : k_t) \qquad (\text{plug} - \text{in dist.})$$

$\hat{\theta}_{t-1}$: m.l.e. of θ from $x^{t-1} = x_1, \ldots, x_{t-1}$.

$$P(x_t|x^{t-1} : k_t) = \int P(x_t|\theta) P(\theta|x^{t-1} : k_t) d\theta \quad (\text{Bayes predictive dist.})$$

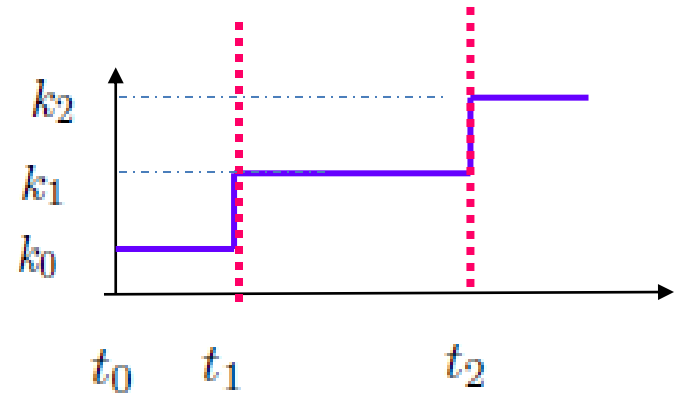
$P(\theta|x^{t-1})$: posterior prob. density of θ from x^{t-1}.

$$P(x_t|x^{t-1} : k_t) = \frac{P(x_t \cdot x^{t-1}|\hat{\theta}(x_t \cdot x^{t-1}) : k_t)}{\sum_x P(x \cdot x^{t-1}|\hat{\theta}(x \cdot x^{t-1}) : k_t)} \quad (\text{SNML dist.})$$

SNML: Sequentially Normalized Maximum Likelihood

Switching Distribution

- $x^n = x_1 \cdots x_n$: data sequence
- m: ♯ of change points
- t_i $(i = 0, \ldots, m)$: the ith change point
  $(t_0 = 1,\ t_{m+1} = n + 1)$
- k_i $(i = 0, \ldots, m)$: the ith model value
- $\mathrm{s} = (t_0, k_0)(t_1, k_1), \ldots, (t_m, k_m)$ model sequence



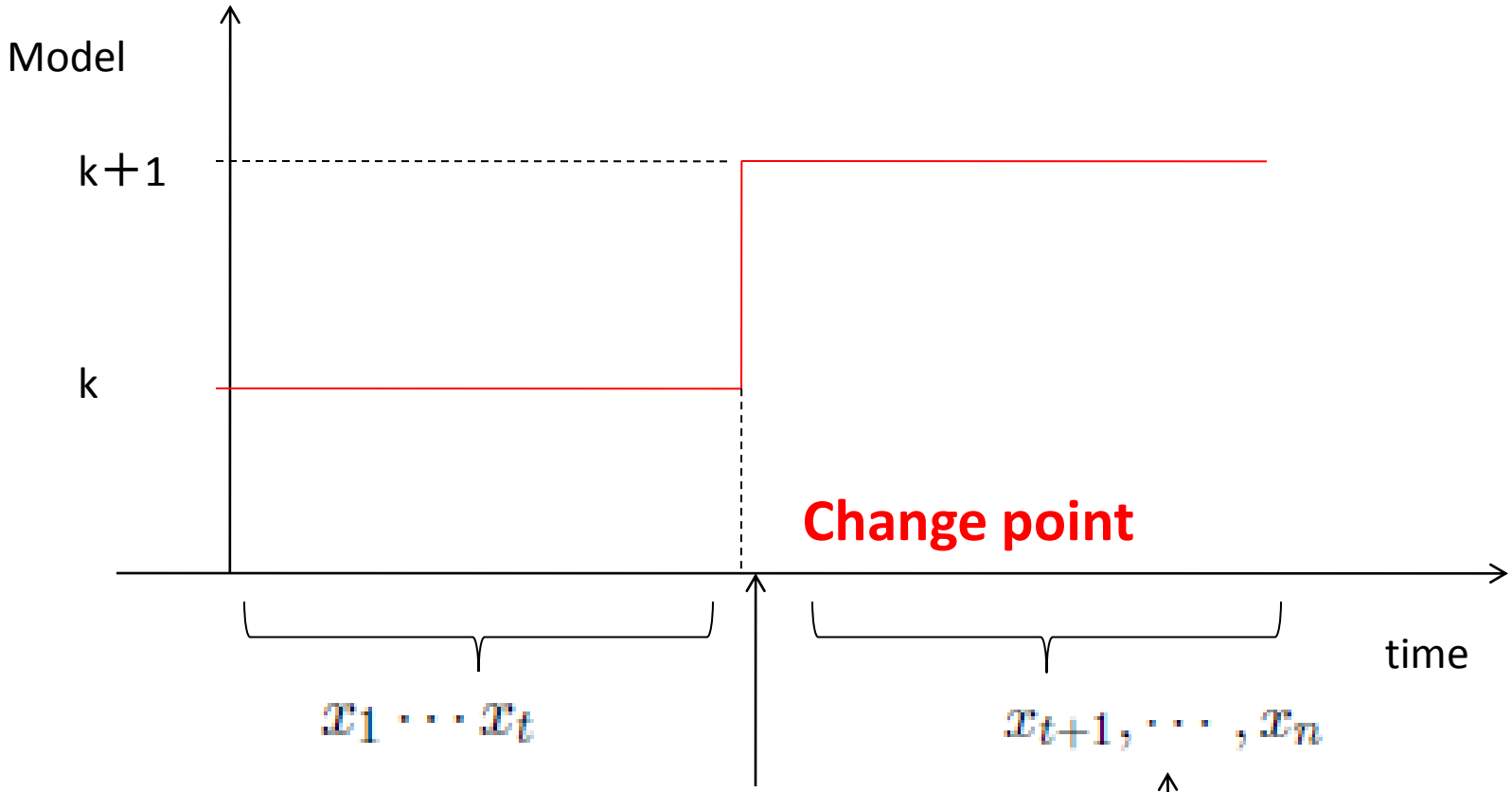
$m = 2$

Switching distribution associated with s:

$$P_{SW}(x_t | x^{t-1} : \mathrm{s}) = \begin{cases} P(x_t | x^{t-1} : k_0) & t_0 \le t \le t_1 - 1 \\ P(x_t | x^{t-1} : k_1) & t_1 \le t \le t_2 - 1 \\ P(x_t | x^{t-1} : k_2) & t_2 \le t \le t_3 - 1 \\ \cdots & \cdots \end{cases}$$

$$P_{SW}(x^n | \mathrm{s}) = \prod_{t=1}^{n} P_{SW}(x_t | x^{t-1} : \mathrm{s})$$

Description of Model Transition



$$-\log p(k+1|k)$$

Code length for model transition

$$\sum_{i=1}^{t} -\log p(x_i|x^{i-1} : k_i)$$

Predictive code length

$$\sum_{i=t+1}^{n} -\log p_{k+1}(x_i|x^{i-1} : k_i)$$

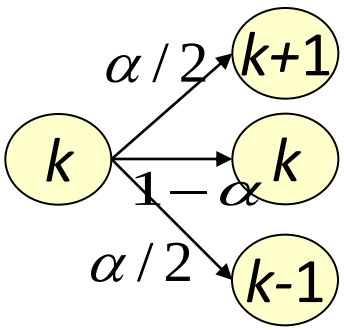
Predictive code length

DMS(Dynamic Model Selction) Criterion

Assumption: model probabilistically transits

Model transition probability $P_t(k_t = k \mid k^{t-1}, \alpha)$, α : parameter

$$P(k_t \mid k_{t-1} : \alpha) = \begin{cases} 1-\alpha, & \text{if } k_t = k_{t-1} \text{ and } k_{t-1} \neq 1, K, \\ 1-\dfrac{\alpha}{2}, & \text{if } k_t = k_{t-1} \text{ and } k_{t-1} = 1, K, \\ \dfrac{\alpha}{2}, & \text{if } k_t = k_{t-1} \pm 1 \end{cases}$$



DMS（Dynamic Model Selection) Criterion：

$$\ell(x^n : k^n) = \underbrace{\sum_{t=1}^{n} -\log P(x_t \mid x^{t-1} : k_t)}_{\text{Predictive code-length for data}} + \underbrace{\sum_{t=1}^{n} -\log P_t(k_t \mid k^{t-1}, \hat{\alpha})}_{\text{Predictive code- length for model seq.}}$$

Input: $x^n = x_1, \ldots x_n$ $\hat{\alpha}$: estimate of α

Output: $k^n = k_1, \ldots k_n$ minimizing $\ell(x^n : k^n)$

Coding Bound for Batch DMS for Switching Dist.

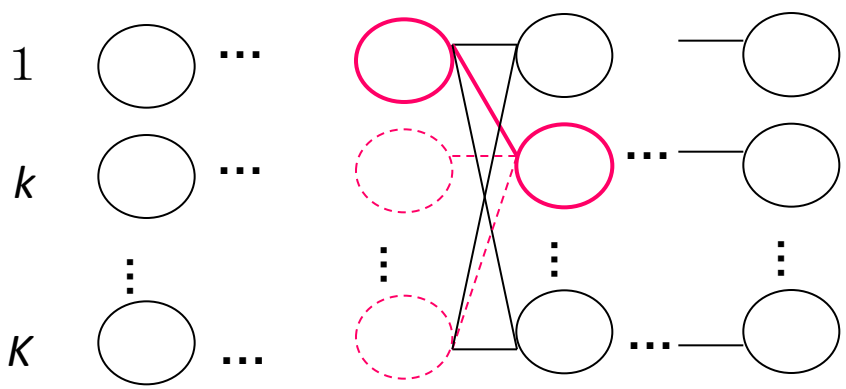
Theorem 1: [Yamanishi and Maruyama IEEE Trans IT 07] *There exists a batch DMS algorithm (DMS1) for switching distributions that runs in time* O(Kn^2) *and the total code-length is upper bounded by*

$$\min_{m} \min_{(t_0,k_0),\cdots,(t_m,k_m)} \left\{ \underbrace{\sum_{j=0}^{m} \sum_{t=t_j}^{t_{j+1}-1} -\log P(x_t|x^{t-1}:k_i)}_{\text{Data complexity for switching}} + \underbrace{nH\left(\frac{m}{n}\right) + \frac{1}{2}\log n + m + o(\log n)}_{\text{Model seq. complexity}} \right\}$$

POINT 1: **Optimal path search based on *Dynamic Programming***

POINT 2: ***Krichevsky and Trofimov Estimation* of model transition probabilities**

$$\hat{\alpha}_t = \frac{n_{k,t} + \frac{1}{2}}{t}$$



At each time, at each state select an optimal path from path sets selected at the latest time

Sequential DMS Problem

[Yamanishi and Sakurai WITMSE 2010]

$$\mathcal{S}_n = \{\mathbf{s} = (t_0, k_0), \ldots, (t_m, k_m) : t_0 = 1 < \cdots < t_m < n, \; k_m \in \{1, \ldots, K\}, \; m \in \mathbf{N}\}$$

model sequence set

$$\mathcal{S}_{(a,b)} = \{\mathbf{s} \in \mathcal{S}_n : \; a \le t_i \le b \; (i = 1, \cdots, m)\}$$

model sequence set restricted to (a,b)

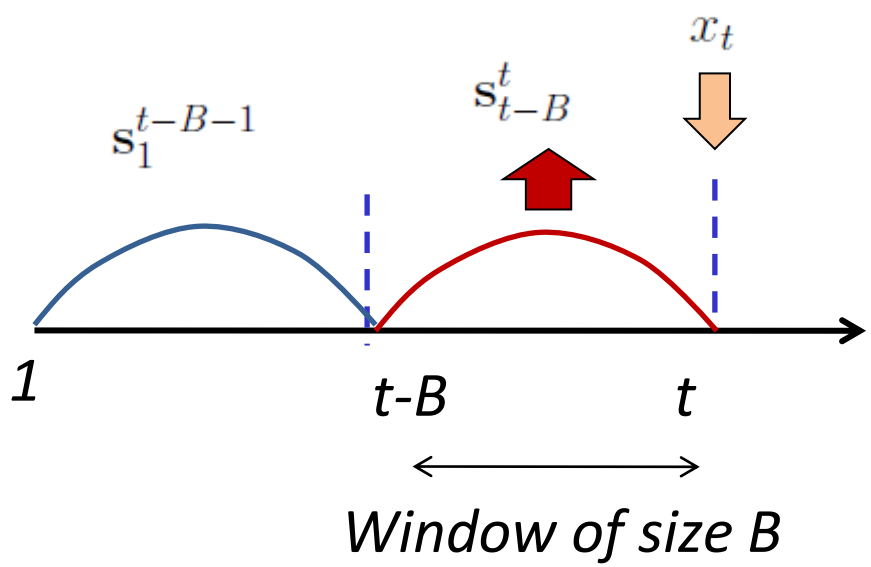
$B > 0$: window size

At each time t,

Input: x_t

Given: $\mathbf{s}_1^{t-B-1} \in \mathcal{S}_{(1,t-B-1)}$,

Output: $\mathbf{s}_{t-B}^{t} \in \mathcal{S}_{(t-B,t)}$



so that the DMS criterion is minimum:

$$-\log P_{SW}(x^t | \mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t}) + \ell(\mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t}).$$

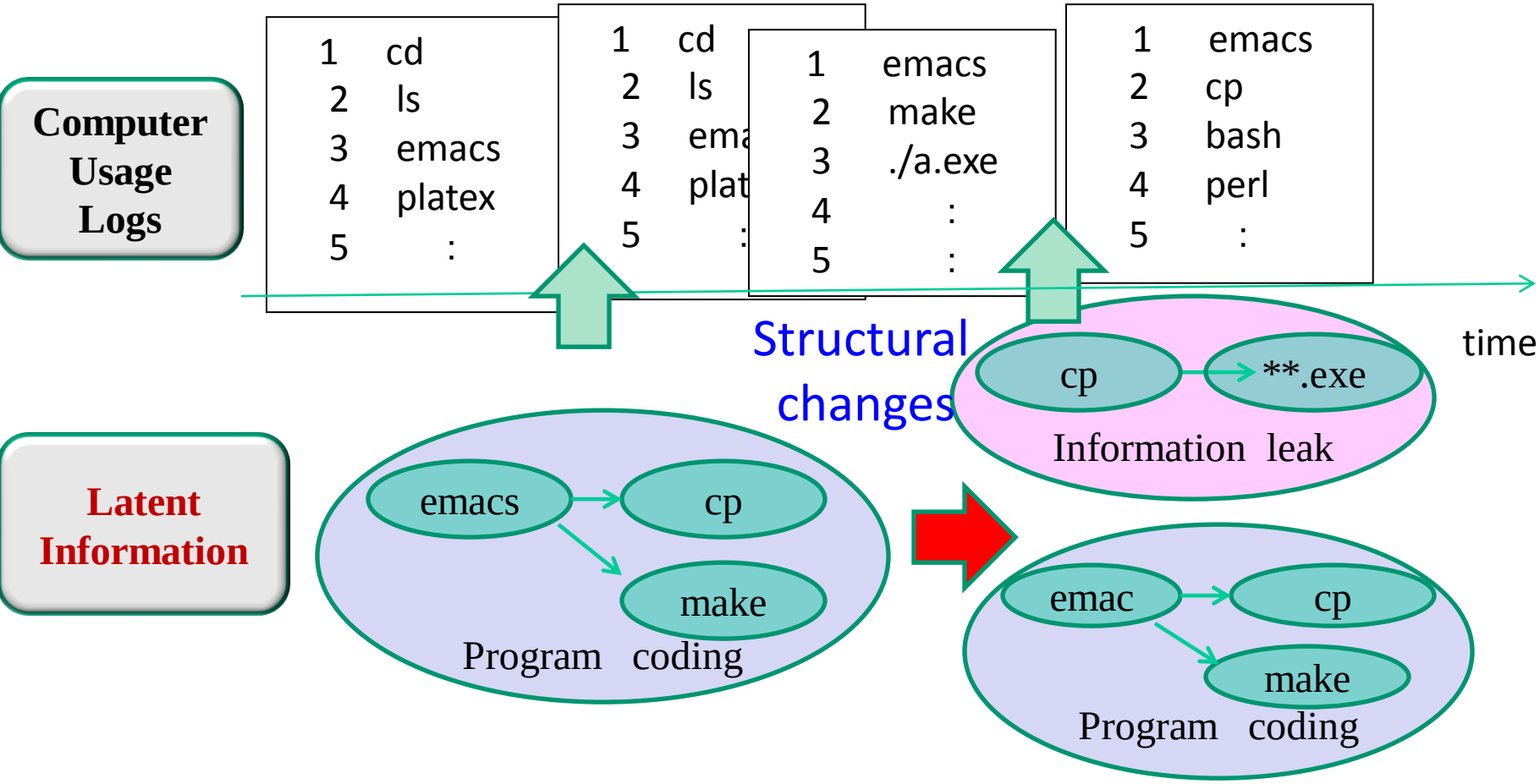
Data code-length for switching dist.

Predictive code-length of model seq.

3. Applications of DMS to Data Mining

3.1. Masquerade Detection

Tracking the emergence of a new pattern leads to the discovery of a masquerader's behavior

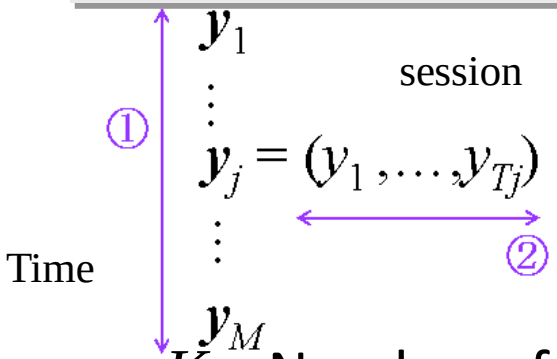


Model

[Maruyama & Yamansihi ITW04]

$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M$: command patterns are modeled using **HMM Mixture**

Wish to detect masquerade by tracking changes in mixture size



command session $\boldsymbol{y}_j$ = (ps, cd, ls, cp, …)

$= (y_1, \ldots y_{Tj})$ length T_j

$$P(\boldsymbol{y}_j \mid \theta) = \sum_{k=1}^{K} \pi_k P_k(\boldsymbol{y}_j \mid \theta_k)$$

K: Number of different behavior patterns (Mixture size)

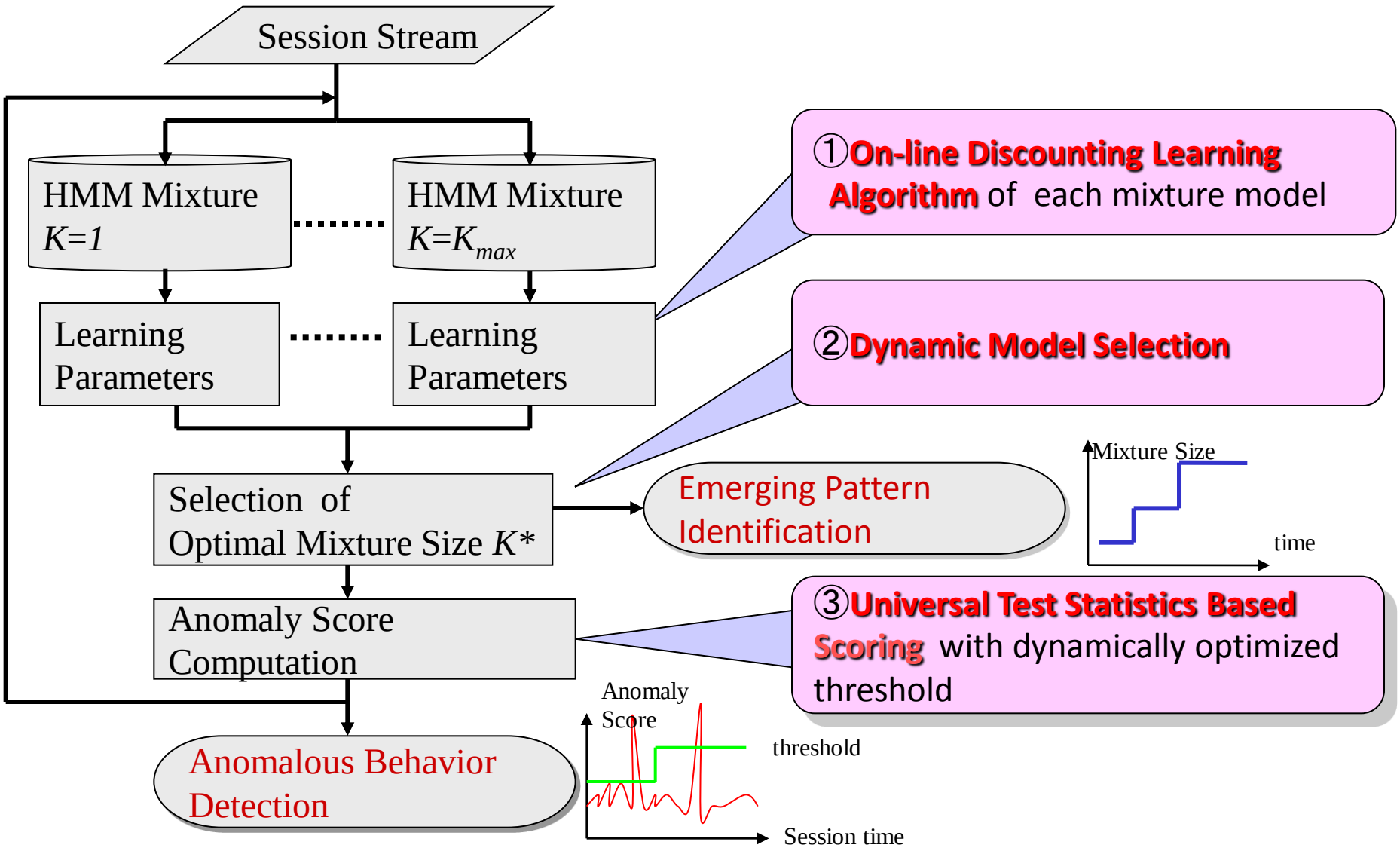
Each behavior pattern is represented by 1-st order HMM

$$P_k(\boldsymbol{y}_j \mid \theta_k) = \sum_{(x_1, \ldots, x_{T_j})} \gamma_k(x_1) \prod_{t=1}^{T_j - 1} a_k(x_{t+1} \mid x_t) \prod_{t=1}^{T_j} b_k(y_t \mid x_t)$$

$(x_1, \ldots, x_{T_j})$: hidden states⇒Markov

γ_k : initial prob. a_k : transition prob. b_k : event prob.

Flow of Behavior Change Detection

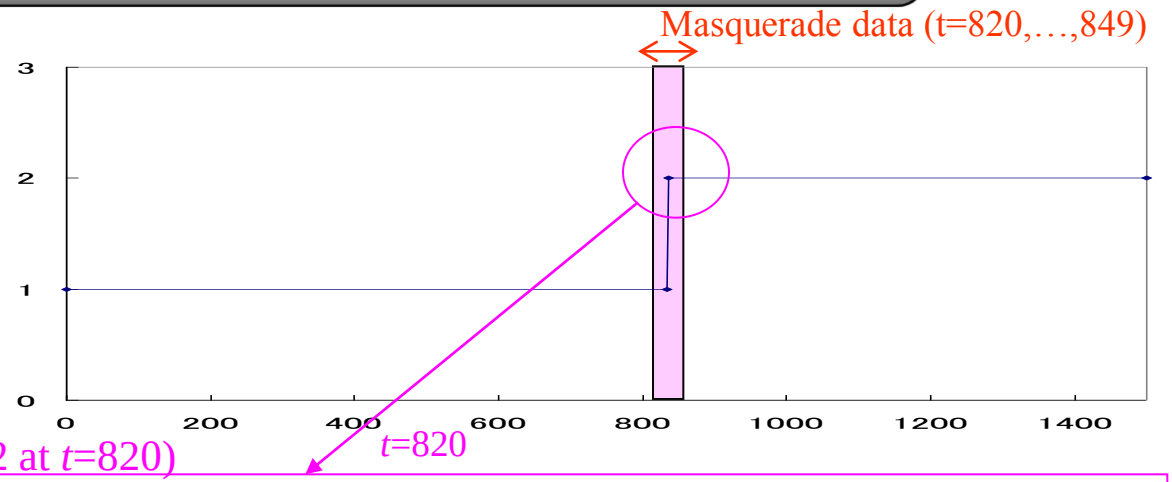


Masquerade Pattern Detection

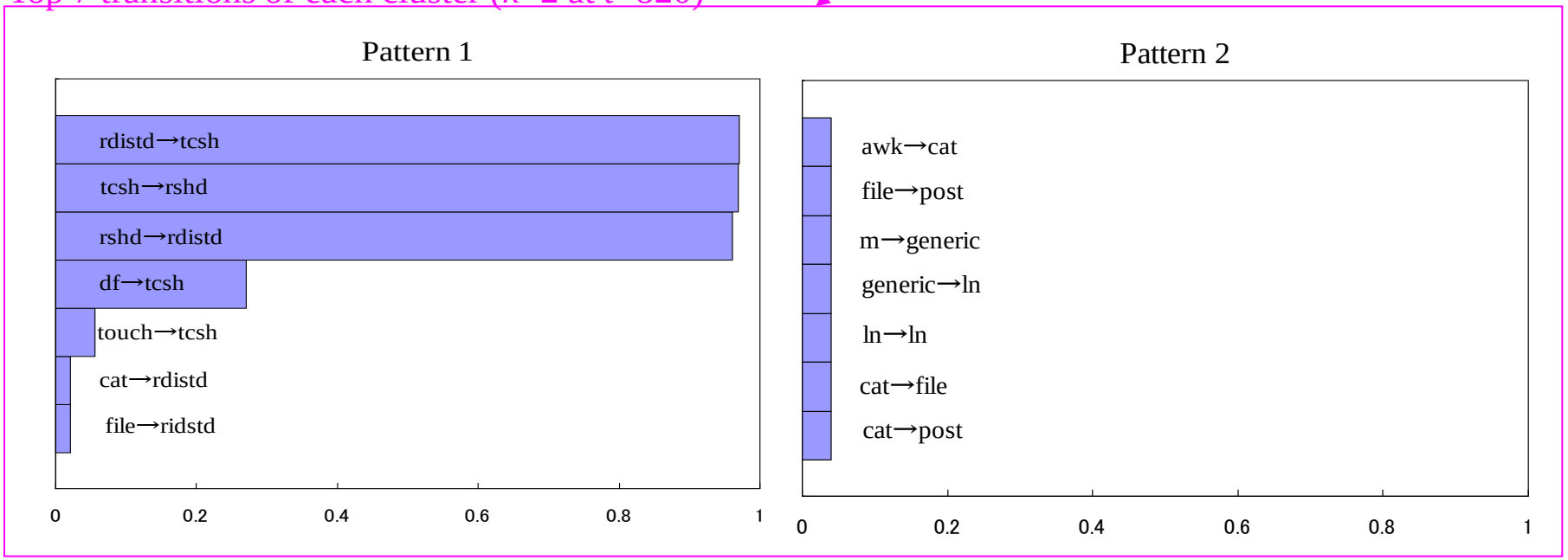
Masquerade's pattern is largely deviated from the normal user's pattern (mainly operating "remote shells").

Ex. User30

T-DMS1 detects the masquerader's command pattern at t=820



Top 7 transitions of each cluster (*k*=2 at *t*=820)



3.2 Failure Detection

[Yamansihi & Maruama KDD2005]

- Syslog….A sequence of events collected using the BSD syslog protocol
- Includes information crucial to network failure

$y_1, \ldots, y_M$: syslog patterns are also modeled using **HMM Mixture**

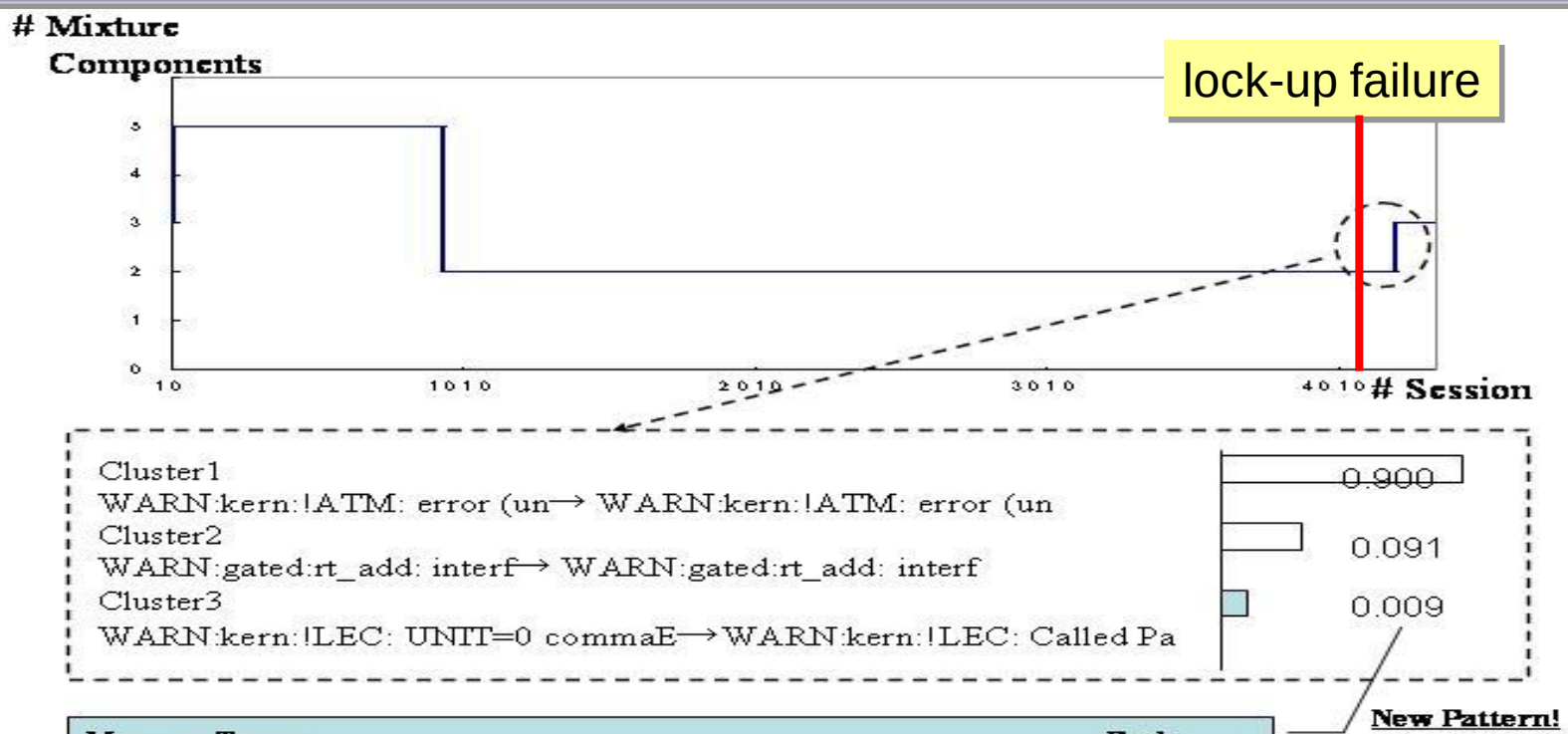
Wish to identify emerging failure patterns by tracking changes in mixture size.

ID	Time stamp	Event Severity	Att1	Att2	Message
##	Nov 13 00:06:23:	ERR	bridge:	!brdgursrv:	queue is full. discarding a message.
##	Nov 13 10:15:00:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/1 Lock-Up!!
##	Nov 13 10:15:10:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/2 Lock-Up!!
##	Nov 13 10:15:20:	WARN:	INTR:	ether2atm:	Ethernet Slot 2L/3 Lock-Up!!

Emerging Pattern Identification

- **Goal:** Track a new behavior pattern caused by a network failure

Identify a new pattern by detecting the change point of mixture size



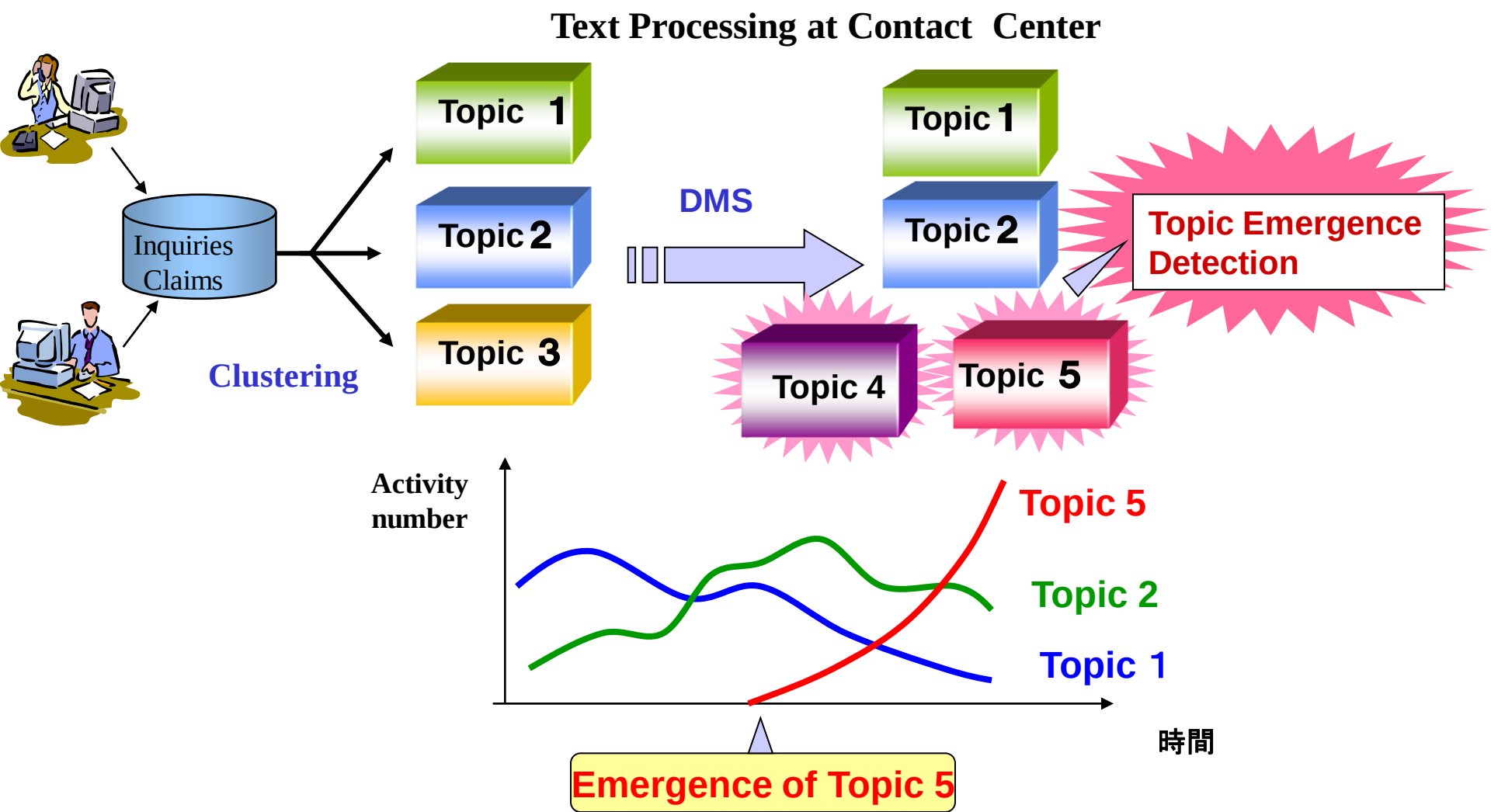
Message Trans.	Prob.
WARN:kern:!LEC: UNIT=0 comma→WARN:kern:!LEC: Called Pa	0.954
ERR :bridge:!brdgursrv: q→ ERR :bridge:!brdgursrv: q	0.751
ERR :gated:krt_ifread: io→ ERR :bridge:!brdgursrv: q	0.734
WARN:kern:!LEC: Multicast→ WARN:kern:!LEC: Control D	0.691

The component with smallest occurrence probability corresponds to a new pattern

3-3. Topic Dynamics Detection

[Morinaga and Yamanishi KDD2004]

Tracking changes of topic organization



Dynamic Topic Modeling with Gaussian Mixture

Document vector: tf or tf-idf

$$x = (tf(w_1), \cdots, tf(w_d)) \qquad \text{or} \qquad x = (tf \cdot idf(w_1), \cdots, tf \cdot idf(w_d))$$

Gaussian mixture topic model: Component = Topic

$$p(x \mid \theta : K) = \sum_{i=1}^{K} \pi_i \phi(x \mid \mu_i, \Sigma_i)$$

$$= \sum_{i=1}^{K} \pi_i \times \frac{1}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)\Sigma_i^{-1}(x - \mu_i) \right)$$

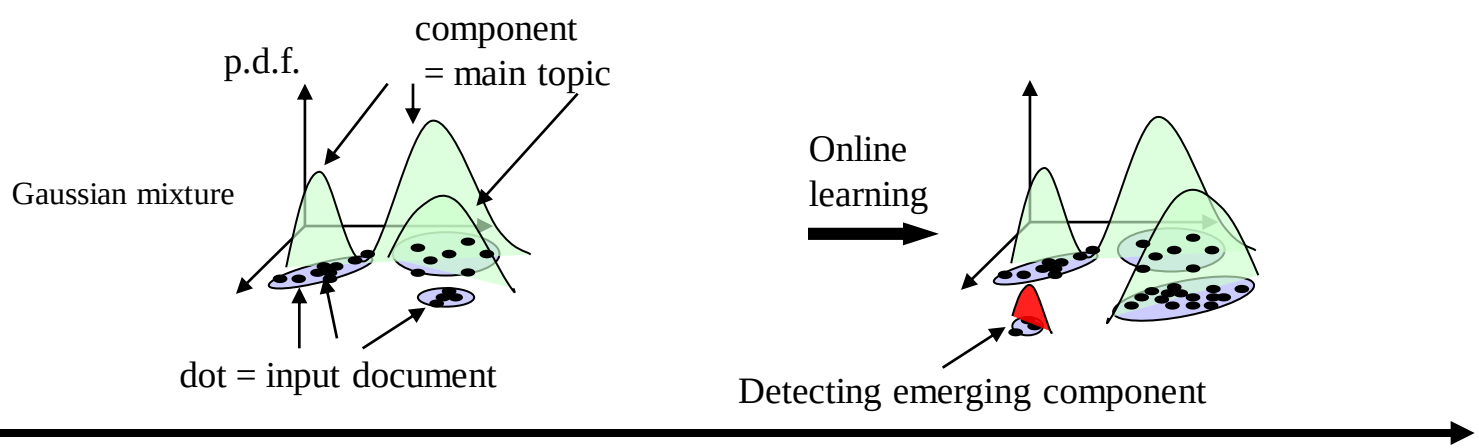
where $\pi_i > 0 \ \ (i = 1, \cdots, K)$ and $\sum_{i=1}^{K} \pi_i = 1$.

Parameters:

K : number of components / topics

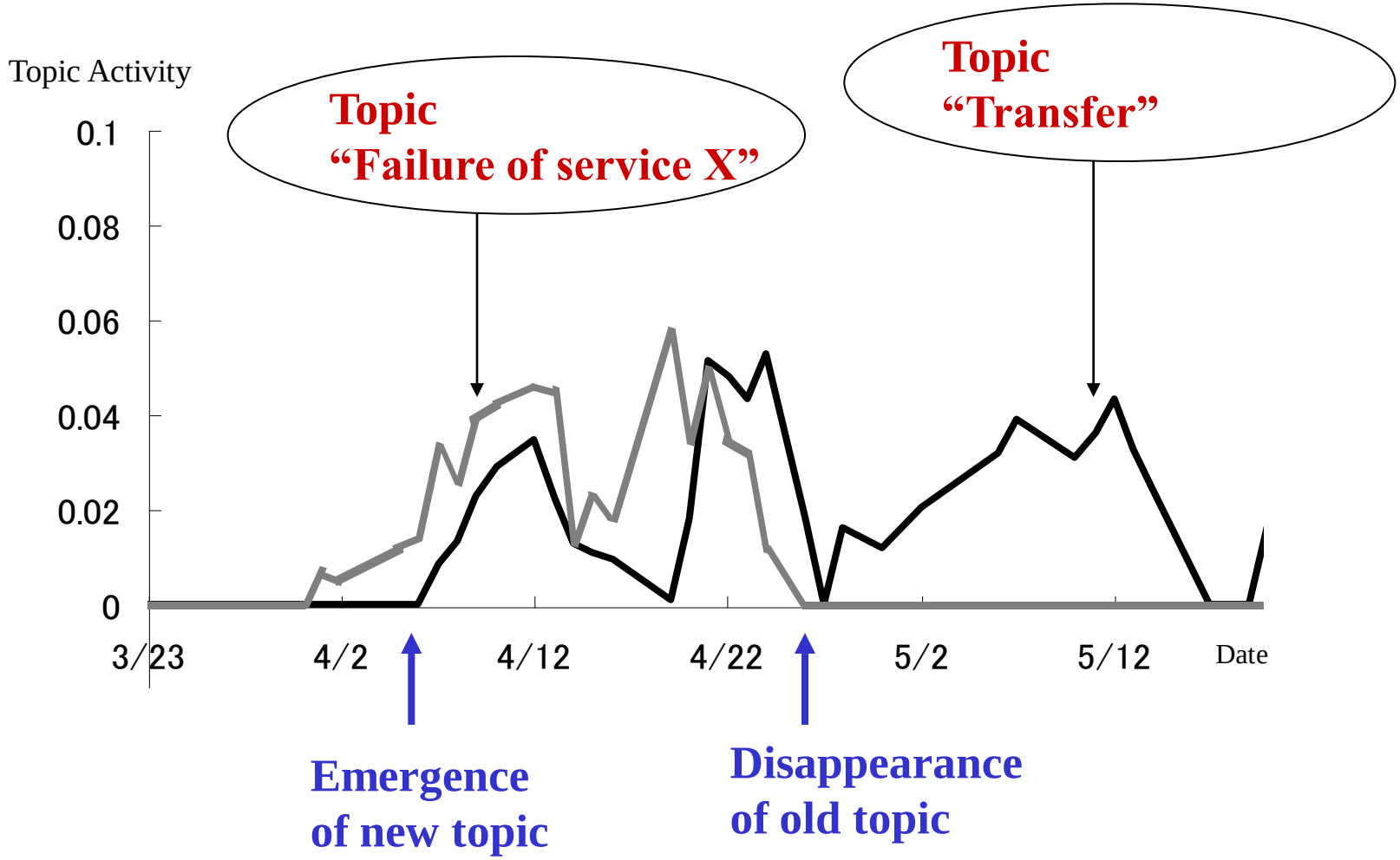
π_i : weight of the i - th component / how likely the i - th topic apears

μ_i, Σ_i : how each component / topic is distributed



Topic Dynamics Detection at Contact Center

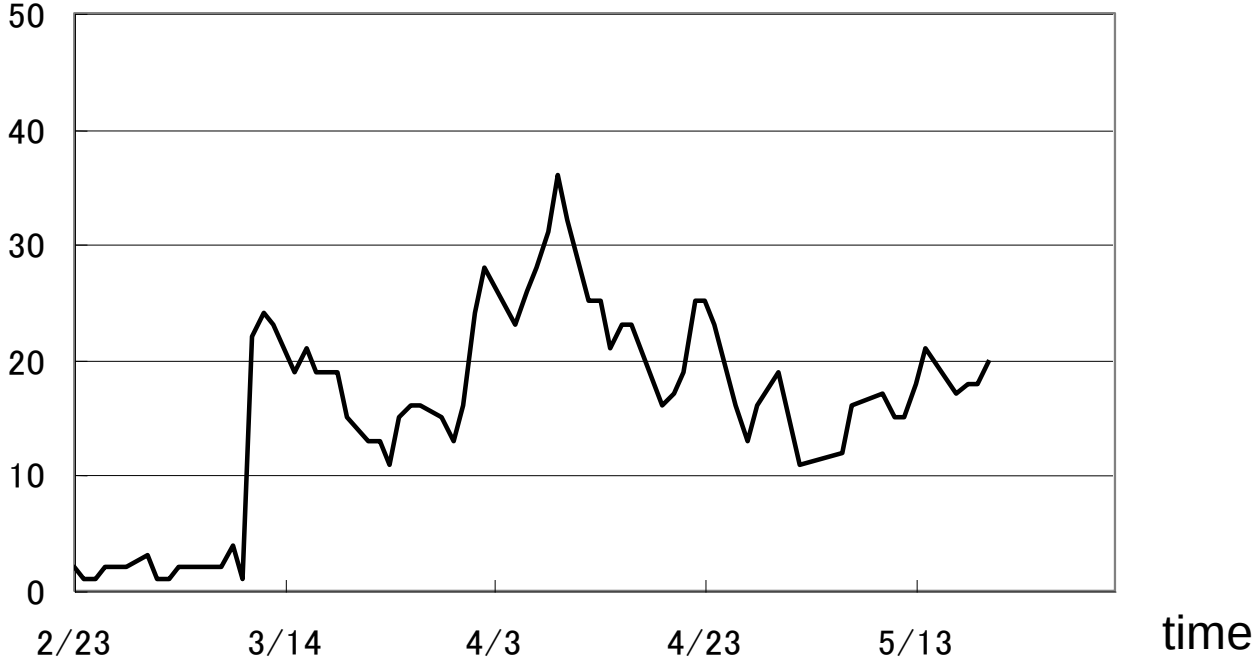
Topic structure changes in contact center text stream



Changes of Topic Number

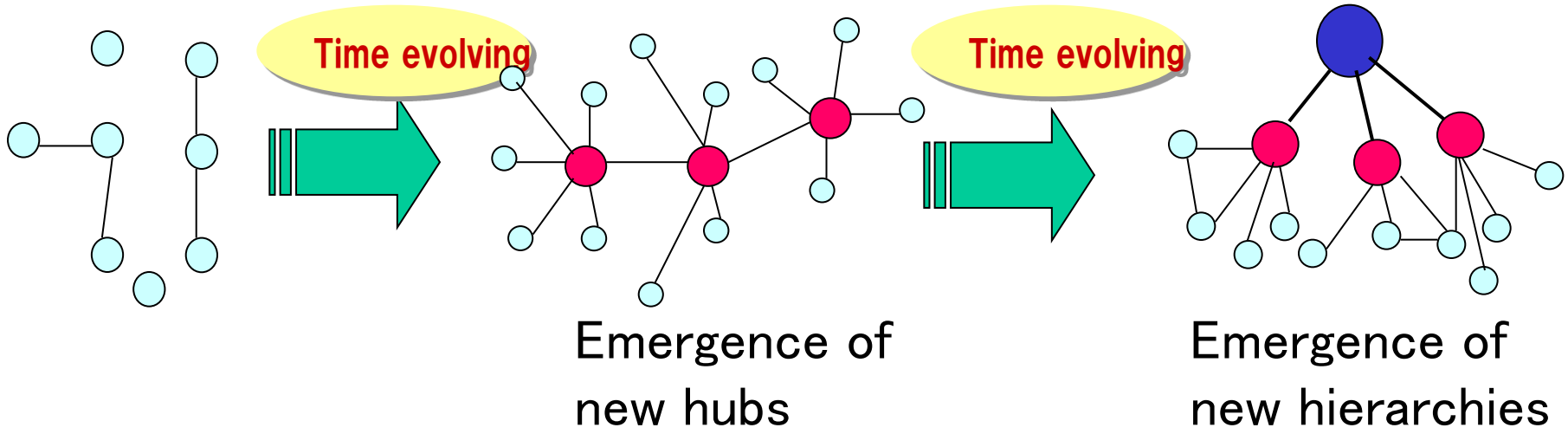
- Total number of inquiries at a contact center is 1202

Changes of number of main topics over time



3-4. Network Structure Mining

Tracking network hierarchy changes leads to new hub detection



ex. SNS networks

Detection of influencers, criminal groups, new communities

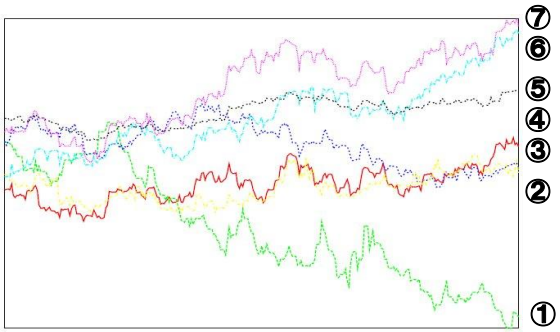
Other examples:

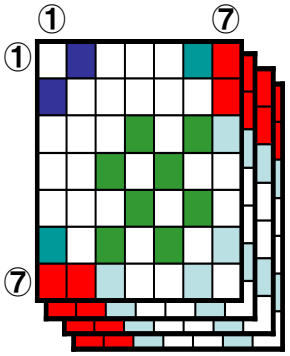
Physical networks, coauthor network, copurchase network

Graph Structure Change Detection

[Hirose, Yamanishi, Nakata, Fujimaki KDD2009]

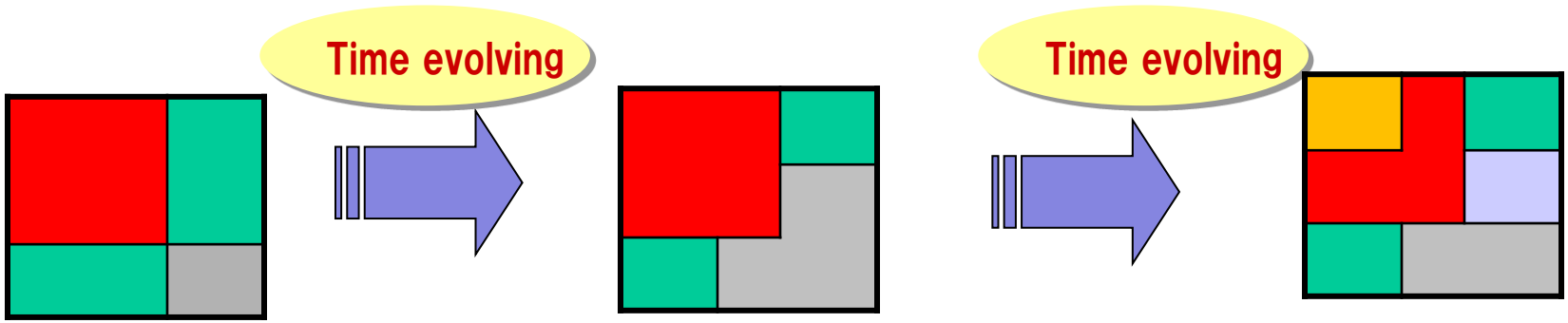
Tracking a new graph cluster leads to discovery of a new community





Time series of correlation matrix

…



Change of partitioning structure

Emergence of a new cluster

4. Summary

- Novelty detection from dynamic data is a challenging issue in data mining.
- Dynamic model selection based on MDL principle is a novel method to detect structural changes in a data stream
- Applications cover masquerade detection, failure detection, topic emergence detection, network structure mining and are applicable to a wide range of areas.

References

1) K.Yamanishi, J.Takeuchi, G.Williamas, and P.Milne: On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery Journal, pp:275-300, May 2004, Volume 8, Issue 3.

2) J.Takeuchi and K.Yamanishi: A Unifying Framework for Detecting Outliers and Change-points from Time Series. IEEE Transactions on Knowledge and Data Engineering, 18:44, pp: 482-492, 2006.

3)S.Morinaga and K.Yamanishi: Tracking Dynamics of Topic Trends Using a Finite Mixture Model. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004), ACM Press, 2004.

4) K.Yamanishi and Y.Maruyama: Dynamic Syslog Mining for Network Failure Monitoring. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2005), pp: 499-508, ACM Press, 2005.

5) K.Yamanishi and Y.Maruyama: Dynamic Model Selection with Its Applications to Novelty Detection. IEEE Transactions on Information Theory, pp : 2180-2189, VOL 53, NO 6, June, 2007.

6) S.Hirose, K.Yamanishi, T.Nakata, R.Fujimaki: Network Anomaly Detection based on Eigen Equation Compression. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2009), 2009.

7) K.Yamanishi and E.Sakurai: Extensions and Probabilistic Analysis of Dynamic Model Selection.2010 Workshop on Information Theoretic Methods for Science and Engineering(WITMSE 2010), Tampere, Finland, August 2010.