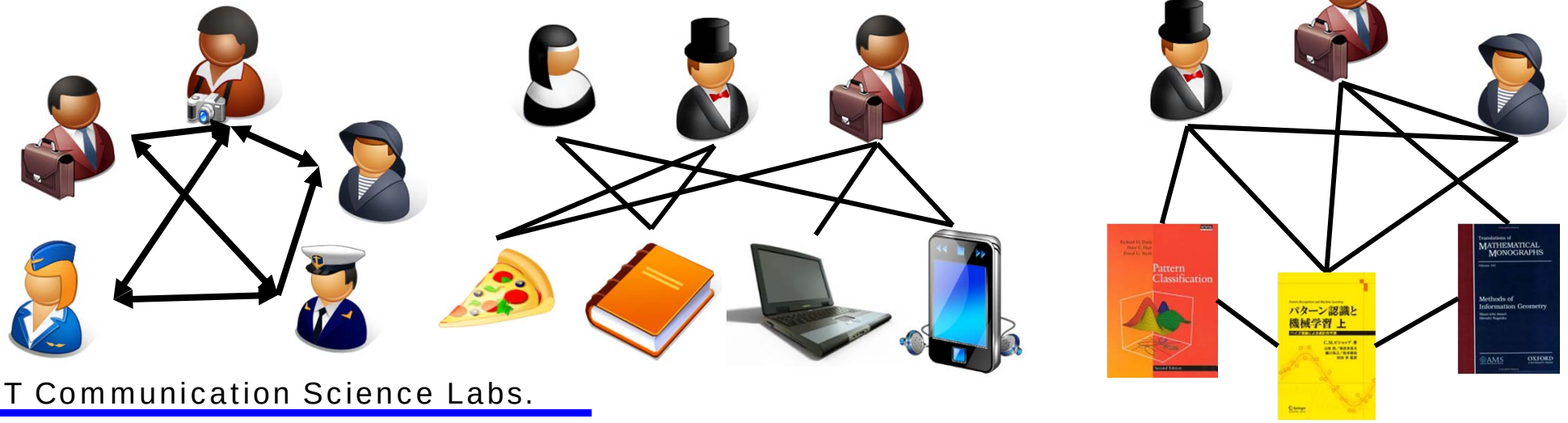
非定常な時系列関係データの解析に関する研究

NTT コミュニケーション科学基礎研究所
石黒　勝彦

“関係データ”

オブジェクト間の関係の有無・あるいは強弱などをまとめたデータを関係データ(relational data)と呼びます

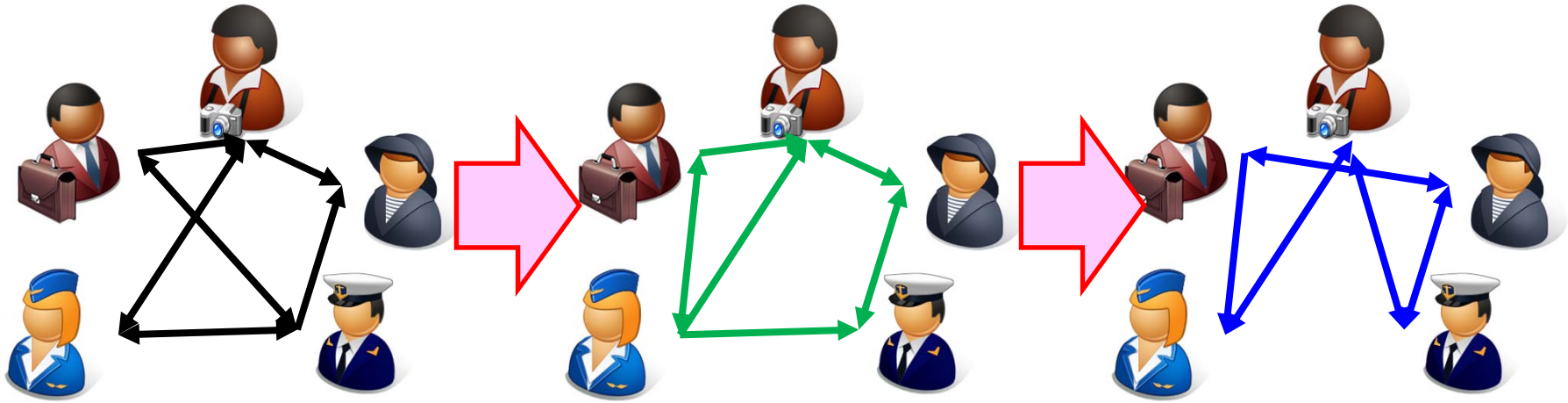
- SNSにおける友達コミュニティ抽出
- オンラインショッピング履歴に基づくレコメンド
- 論文と著者の組み合わせによる研究トピック解析



“時系列”関係データ

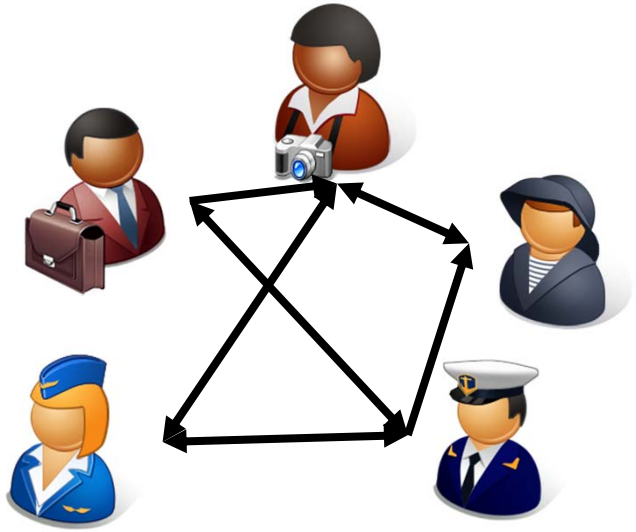
実世界にある“関係”の中には、時系列にしたがって変化するものがたくさんあります

- SNS上の友達関係は簡単に変化します
- CMが当たると突然商品が売れ出します
- 研究プロジェクトや異動で共著者は変わります



関係データは行列に見えます

例: SNSのメンバー間の友達リンク

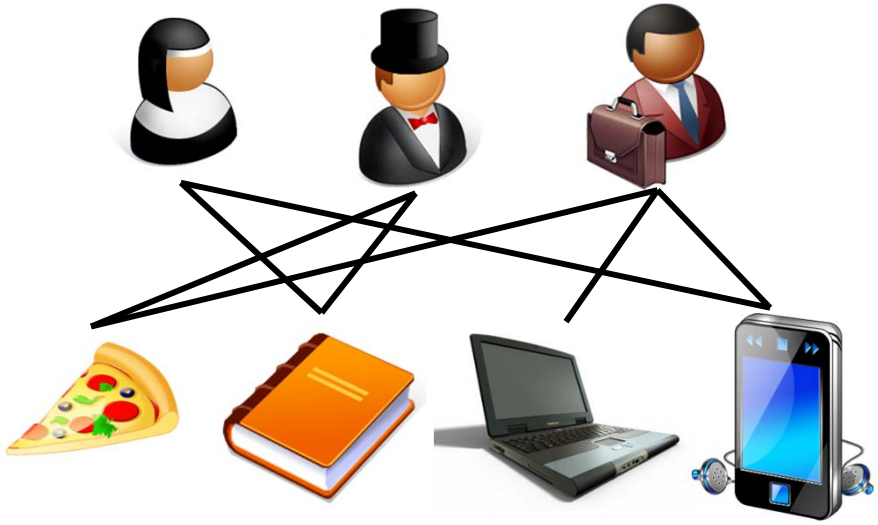


フォローしたメンバー i

0	1	0	1	0
0	0	1	0	1
0	1	0	0	0
1	0	1	0	1
0	1	0	1	0

フォローされたメンバー j

例: オンラインショップの購買履歴

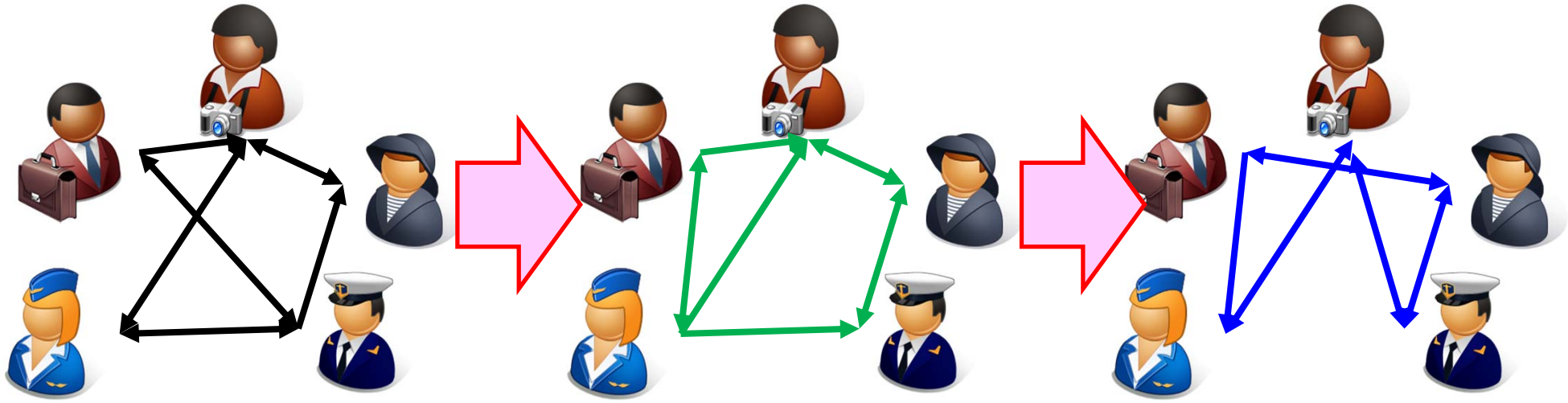


チェックした顧客 i

0	1	0	1
1	1	0	0
1	0	1	1

チェックされたアイテム j

時系列関係データは
それの時系列なので

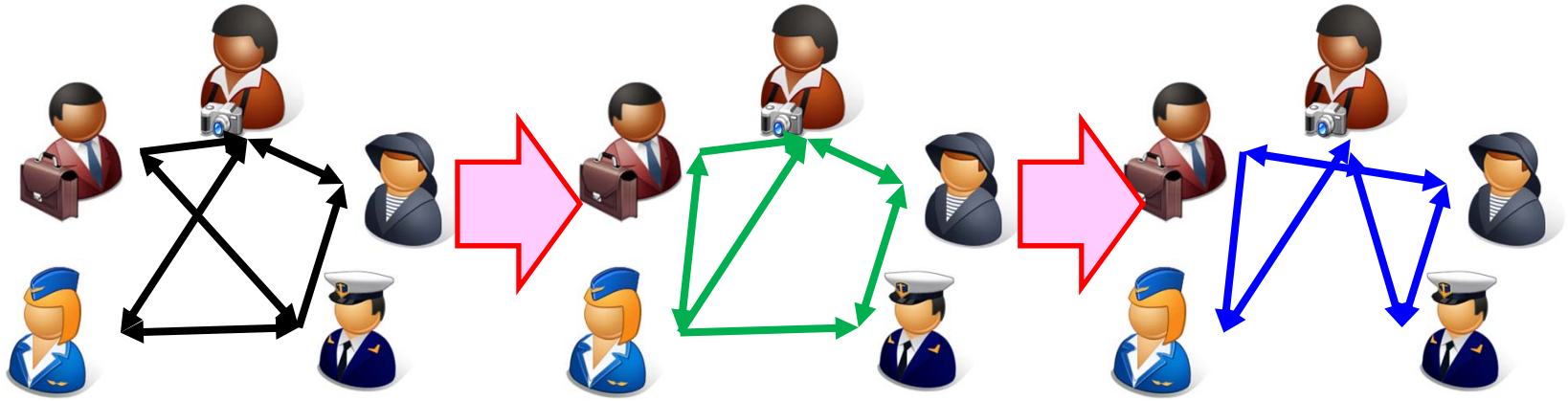


0	1	0	1	0
0	0	1	0	1
0	1	0	0	0
1	0	1	0	1
0	1	0	1	0

0	1	0	0	1
0	0	1	0	0
0	1	0	1	0
0	0	1	0	0
1	1	0	1	0

0	0	0	0	1
0	0	0	1	1
1	0	0	1	0
0	1	1	0	0
0	1	0	0	0

時系列関係データは必ずテンソルです



時刻という特殊な軸をもった
テンソルデータ
→ それに応じた“モデル”が必要

					0 0 0 0 1
				0	1 0 0 1 1
0	1	0	1	0	0 0
0	0	1	0	1	0 0
0	1	0	0	0	0 0
1	0	1	0	1	0
0	1	0	1	0	

フォローしたメンバー i

フォローされたメンバー j

時刻 t

本発表の構成

- 時間に従って変化する関係データの解析手法について

キーワード:
時系列関係データ,ノンパラメトリックベイズ,
HDP-HMM, IRM, クラスタリング

- 関係データの解析
  - 大きな2つの分類
  - IRMによるクラスタリング
- 時系列関係データの解析
  - 実データに見る非定常・非連続性
  - dIRMによる時間発展クラスタリング
- 時系列関係データ解析の“次の”課題

関係データを解析して テクニカルにどう利用したいか？

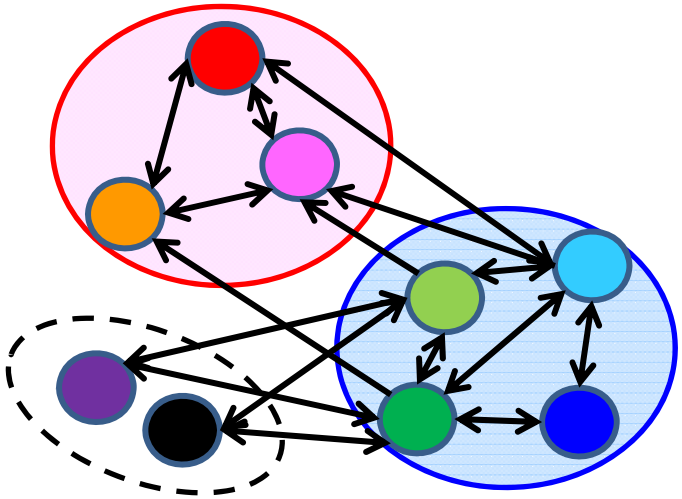
NTT

- 与えられたデータの性質を調べる　こっち！！
  - 内部にどんなクラスタ・コミュニティがあるか
  - クラスタやコミュニティ同士の関係はどうなっているか
  - 生成モデルが多い

- 与えられたデータから学習して他に活かす
  - 欠損データの補完
  - 未来における関係の予測
  - テンソル分解（≒低ランク近似）や識別モデルが多い

（静的な）関係データで オブジェクトのクラスタリング

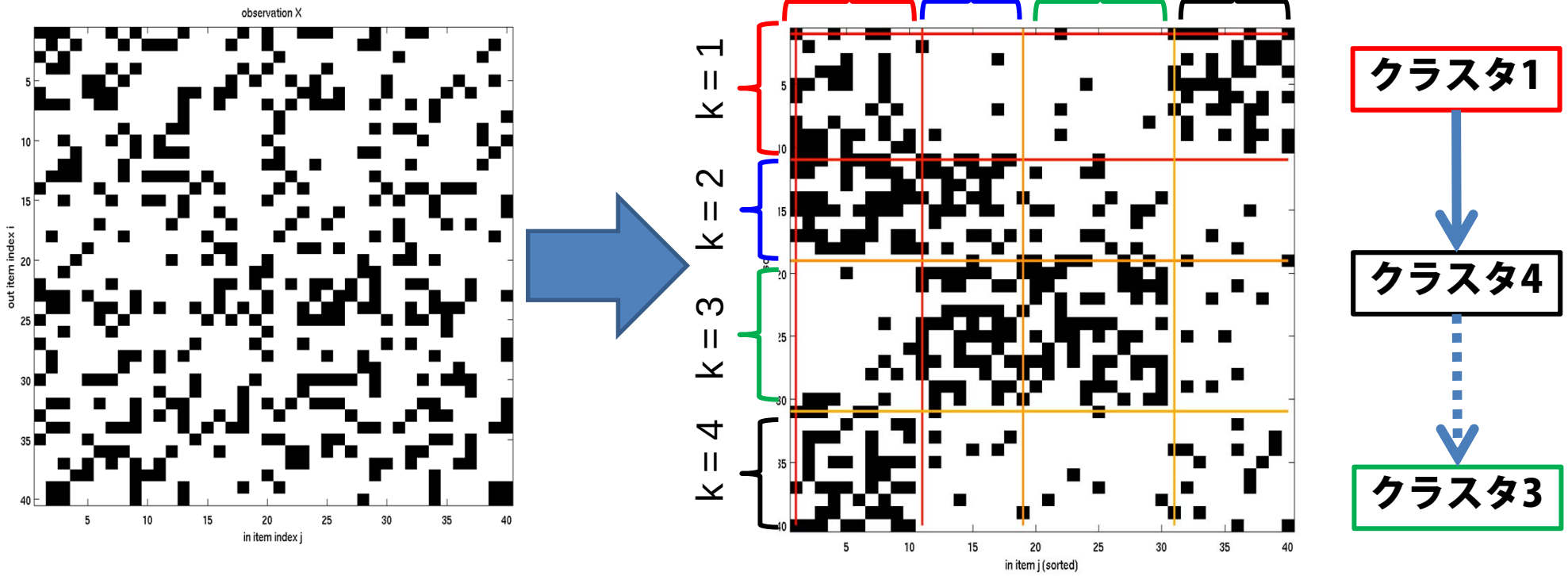
- 似たようなリンクを持つオブジェクトをまとめて、オブジェクトのクラスタ（コミュニティ）を抽出します
- 人間関係のグループ（派閥）発見や、遺伝子の機能的分類に応用できます



Infinite Relational Model (IRM)

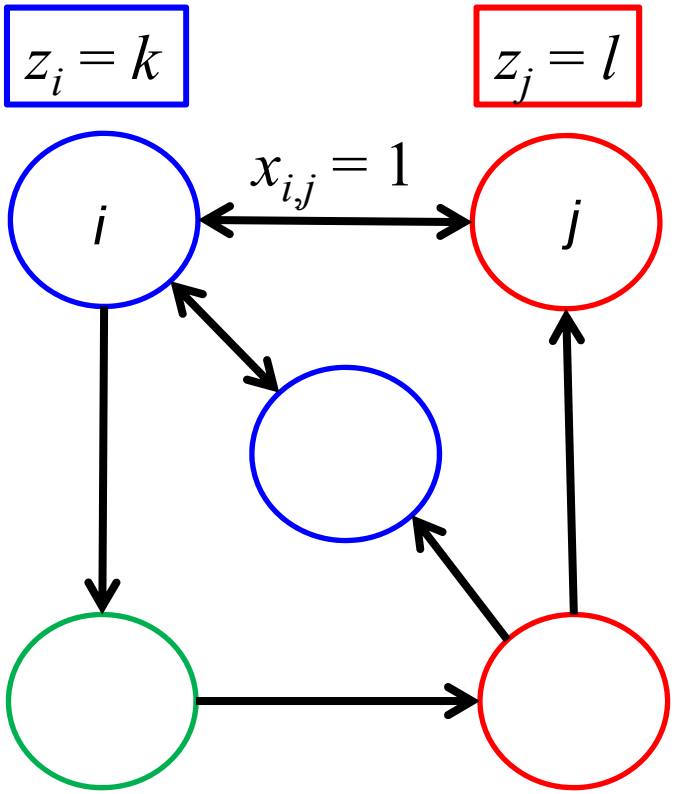
(Kemp et al., 2006)

同じような関係を持つオブジェクトを
適切な数のクラスタに集約する手法です。
テンソル“分割”手法といえるでしょう



IRMの問題定義

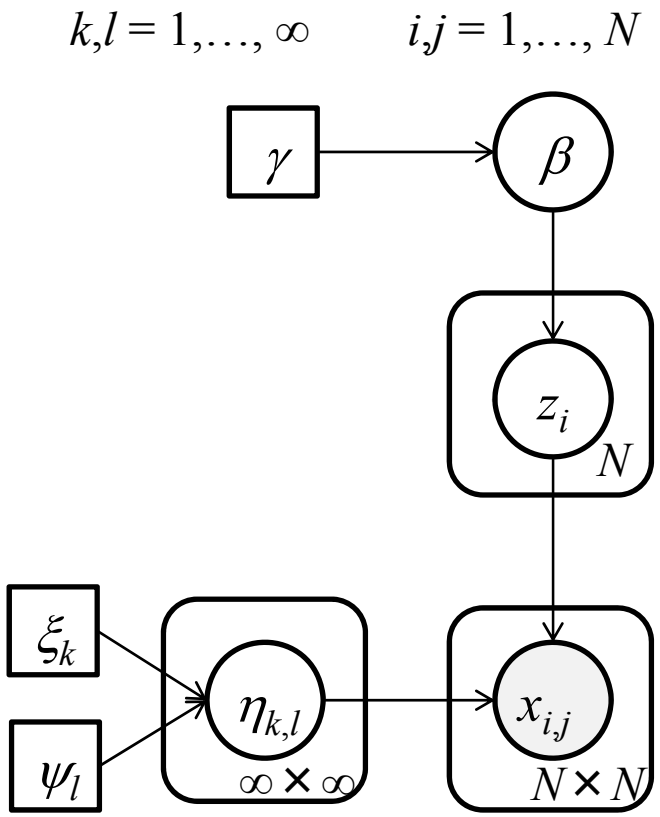
観測された関係データXを用いて、
クラスタリングの隠れ変数Zとパラメータを決めます



$$x_{i,j} = \{0, 1\} \sim \mathrm{p}(\eta_{zi,\, zj})$$

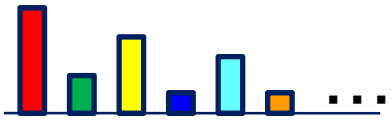
$$\{\eta_{k,l}\} =$$

0.2	0.7	0.1	0.3
⋮	…		



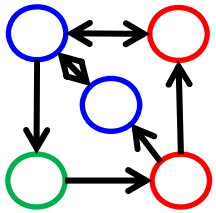
無限次元のクラスタ混合割合ベクトル

$\beta \sim \mathrm{Stick}(\gamma)$



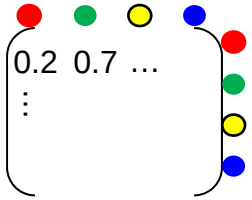
オブジェクトのクラスタリング隠れ変数

$z_i \sim \mathrm{Multinomial}(\beta)$



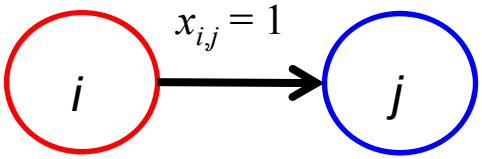
クラスタ間の接続確率

$\eta_{k,l} \sim \mathrm{Beta}(\xi_k, \psi_l)$



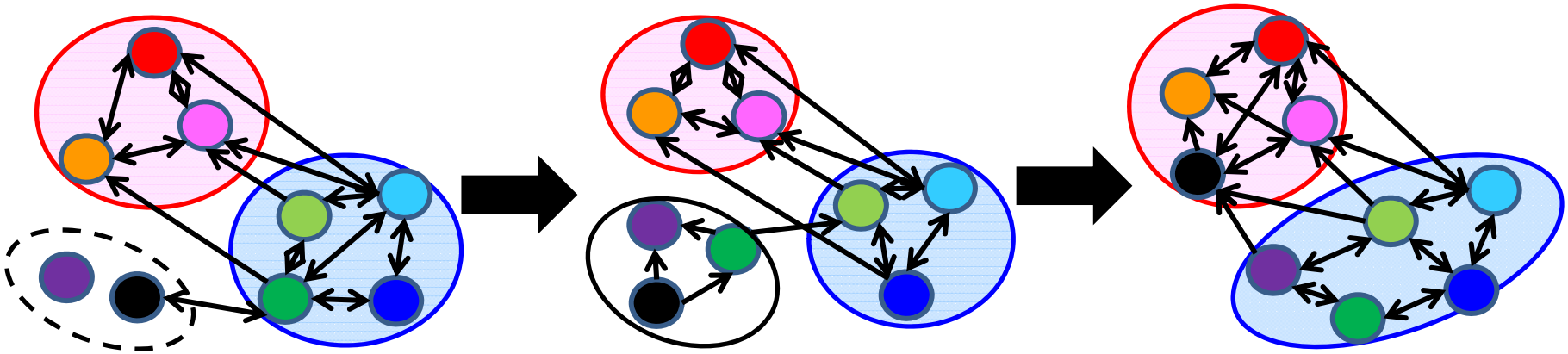
観測された関係データ

$x_{i,j} \sim \mathrm{Bernoulli}(\eta_{zi,\,zj})$



続いて時系列関係データでの オブジェクトの クラスタリングを考えます

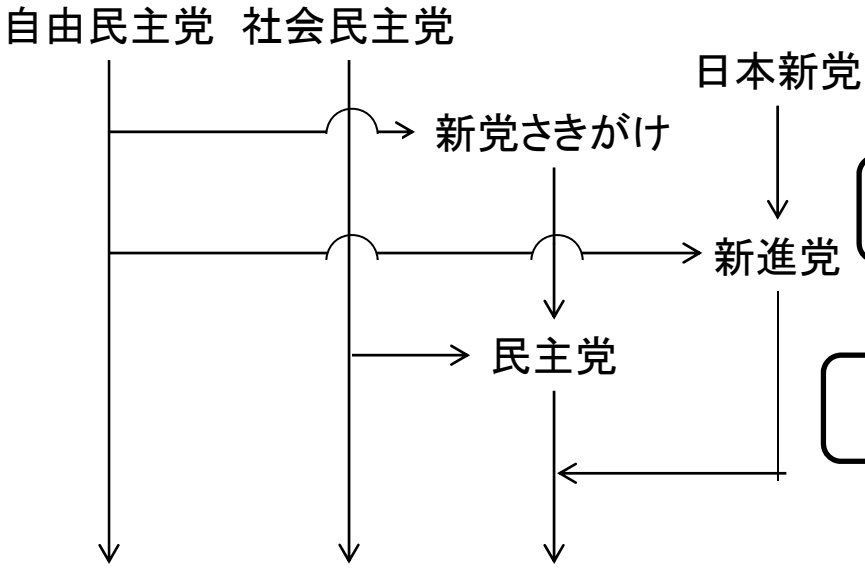
- 時間方向でクラスタを同定しながら、各クラスタの隆盛や生成・消滅を解析できます



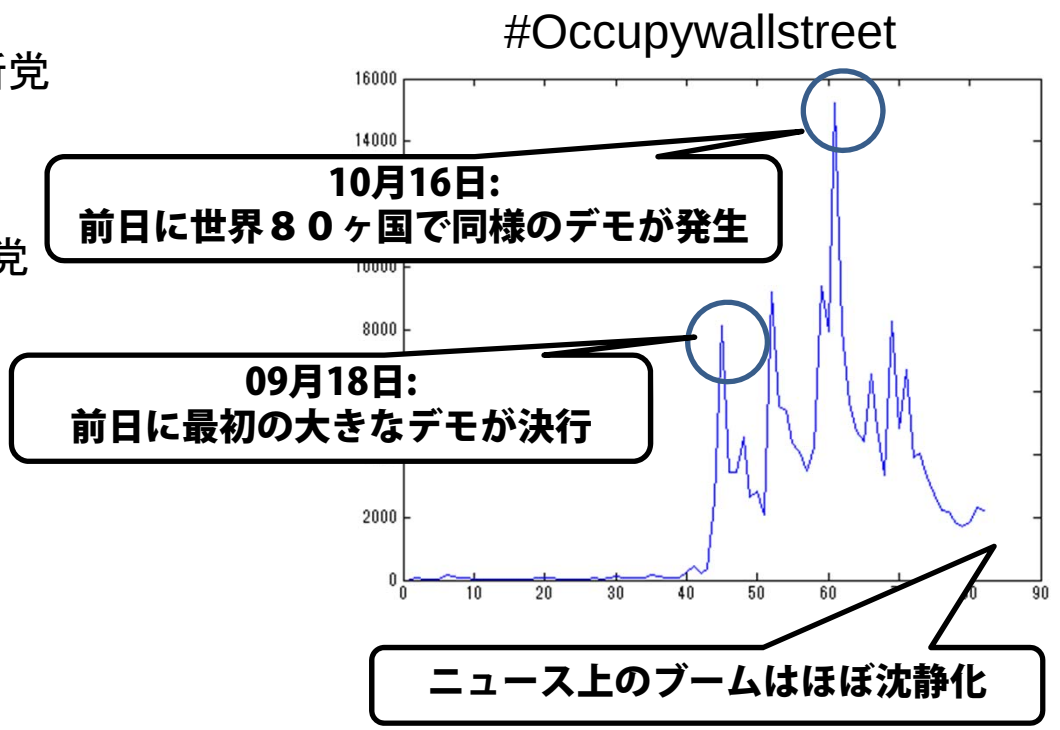
現実の時系列関係データを見ると？

連続時間の類似性を持ちながら、
非連続、非定常な時間変化を見せます

人のコミュニティは組織の再編成で
予兆なく分裂・統合を起こします

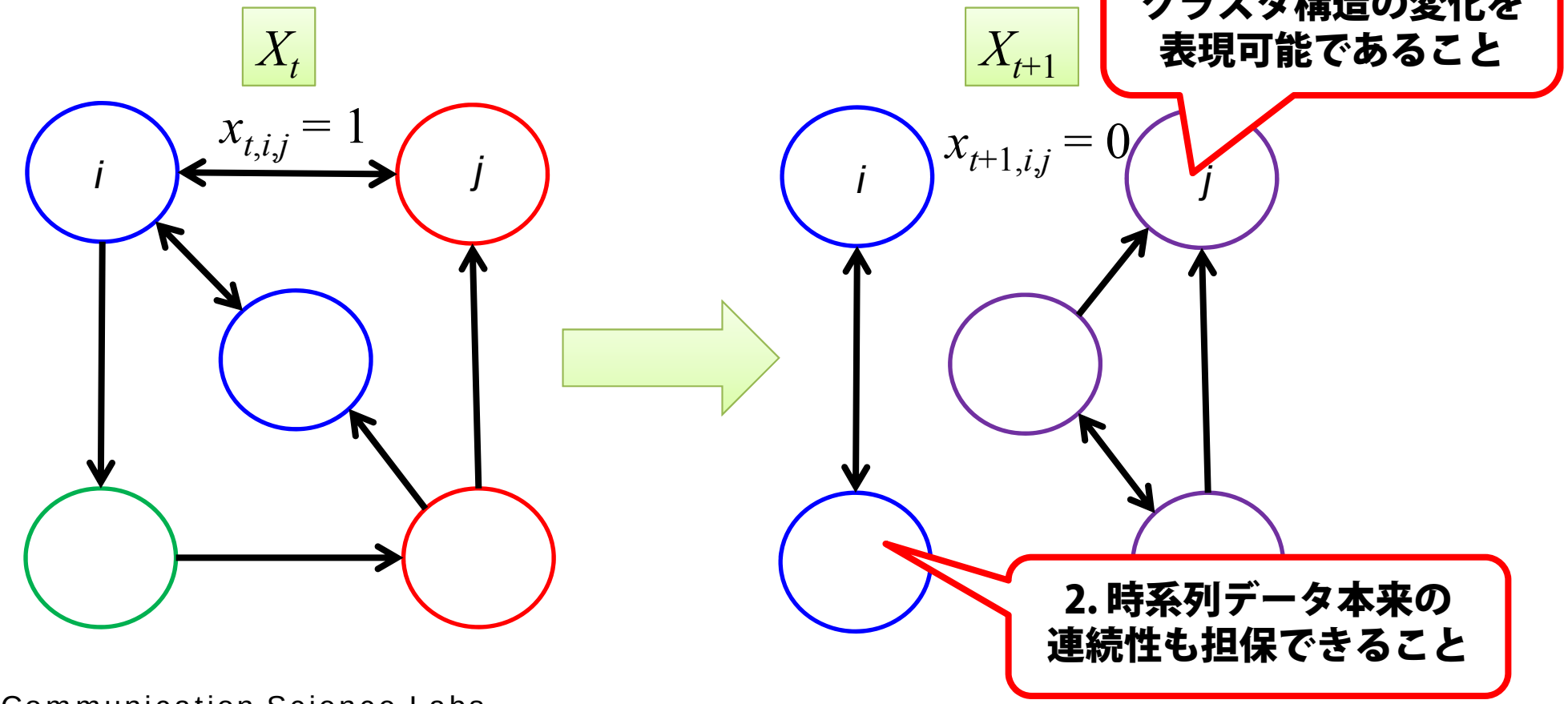


ウェブページやツイートの“ブーム”
は短命なクラスタを作ります

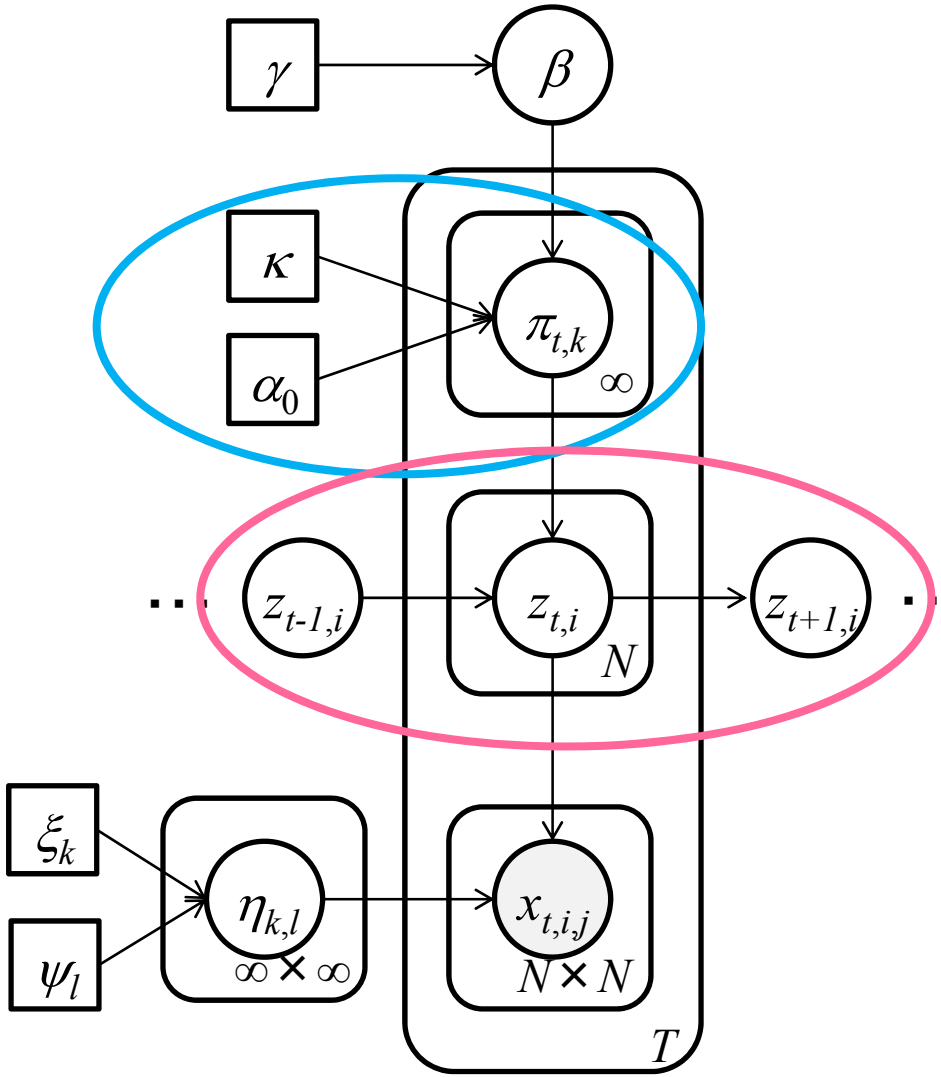


dynamic IRM (Ishiguro et al., 2010)

非連続かつ非定常な変化も含む時系列関係データのための IRM拡張モデルを考案しました

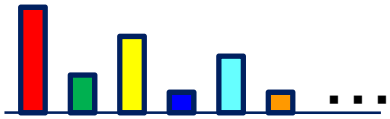


$t = 1,\ldots, T$　　$k,l = 1,\ldots, \infty$　　$i,j = 1,\ldots, N$



時系列全体での無限個のクラスタ割合

$\beta \sim \mathrm{Stick}(\gamma)$



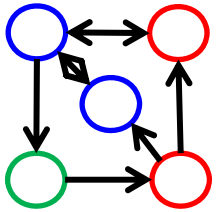
時刻依存のクラスタ間HMM遷移確率

$\pi_{t,k} \sim \mathrm{DP}(\alpha_0+\kappa,\ (\alpha_0\beta+\kappa\delta_k) / (\alpha_0+\kappa))$

$\pi_{t,}$ =

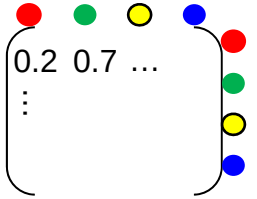
オブジェクトのクラスタリング隠れ変数

$z_{t,i} \sim \mathrm{Multinomial}(\pi_t, z_{t-1,i})$



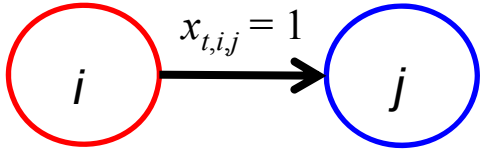
クラスタ間の接続確率

$\eta_{k,l} \sim \mathrm{Beta}(\xi_k, \psi_l)$



観測された関係データ

$x_{t,i,j} \sim \mathrm{Bernoulli}(\eta_{zt,i,\ zt,j})$

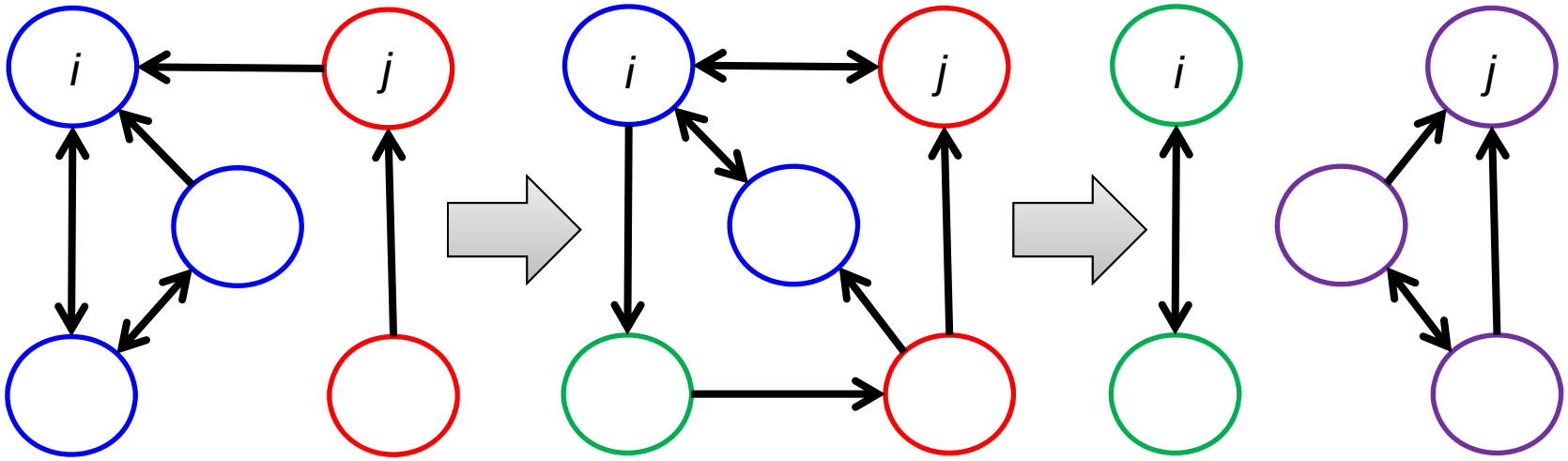


時間発展のモデル化(1/2)

ノードの所属クラスタはHMMで離散的に変化します

$$z_{t,i} \sim \text{Multinomial}(\pi_t, z_{t-1,i})$$

$$\pi_{t,\bullet} = \begin{pmatrix} 0.6 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} \qquad \pi_{t,\bullet} = \begin{pmatrix} 0.1 \\ 0.3 \\ 0.1 \\ 0.5 \end{pmatrix}$$

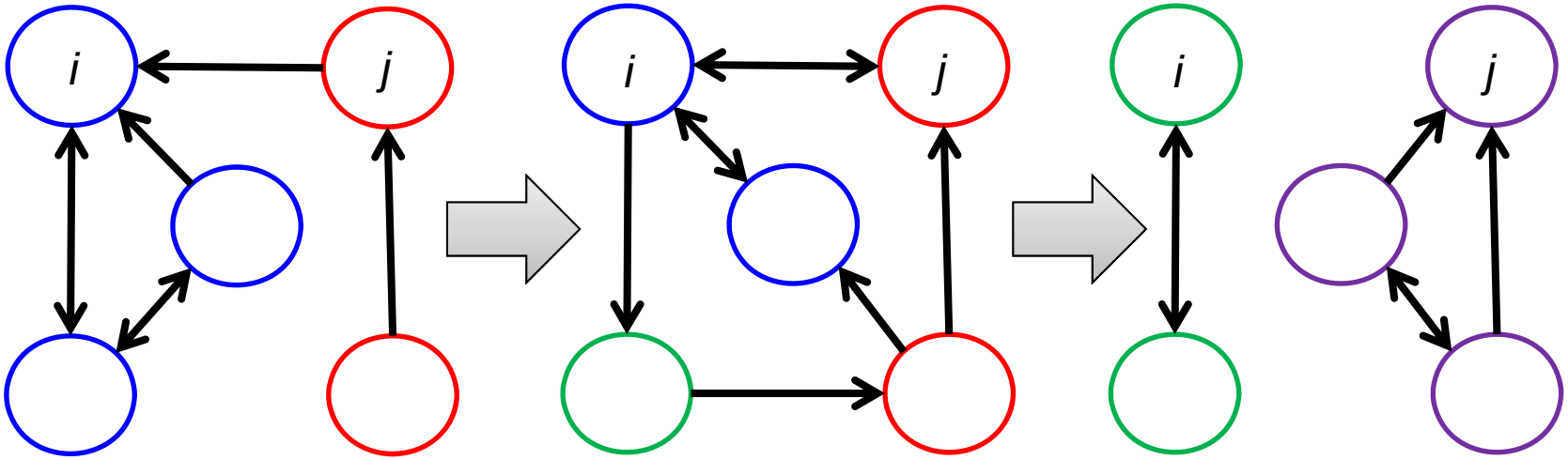


時間発展のモデル化(2/2)

非定常なクラスタの生成・消滅・結合・分裂などを表現するため、クラスタ遷移確率は時間ごとに異なるものとします

$\pi_{t,\bullet} = \begin{pmatrix} 0.6 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix}$ $\pi_{t+1,\bullet} = \begin{pmatrix} 0.1 \\ 0.09 \\ 0.8 \\ 0.01 \end{pmatrix}$

$\pi_{t,\bullet} = \begin{pmatrix} 0.05 \\ 0.35 \\ 0.1 \\ 0.5 \end{pmatrix}$ $\pi_{t+1,\bullet} = \begin{pmatrix} 0.01 \\ 0.3 \\ 0.28 \\ 0.41 \end{pmatrix}$



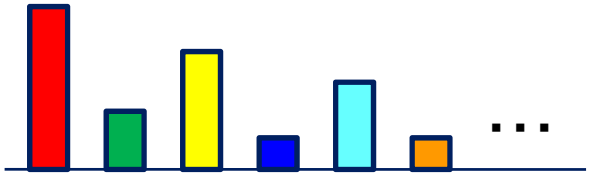
sticky prior + HDP-HMM

(cf. Fox et al., 2008)

時系列データ本来の時間連続性と非定常な遷移を
バランスできるノンパラメトリックベイズモデルを使います

時系列全体でのクラスタ混合割合（無限次元）

$$\beta \sim \mathrm{Stick}(\gamma)$$

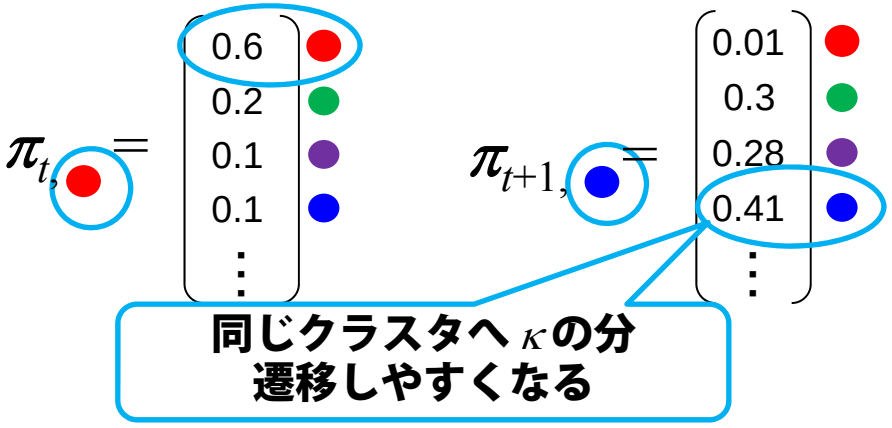


クラスタ間遷移確率はβを元にして、
時刻ごと、クラスタごとにサンプリングします

※DP(Dirichlet Process)≒無限次元のディリクレ分布

$$\pi_{t,k} \sim \mathrm{DP}(\alpha_0+\kappa, (\alpha_0\beta+\kappa\delta_k) / (\alpha_0+\kappa))$$

δ_k: k番目だけ"1"のベクトル

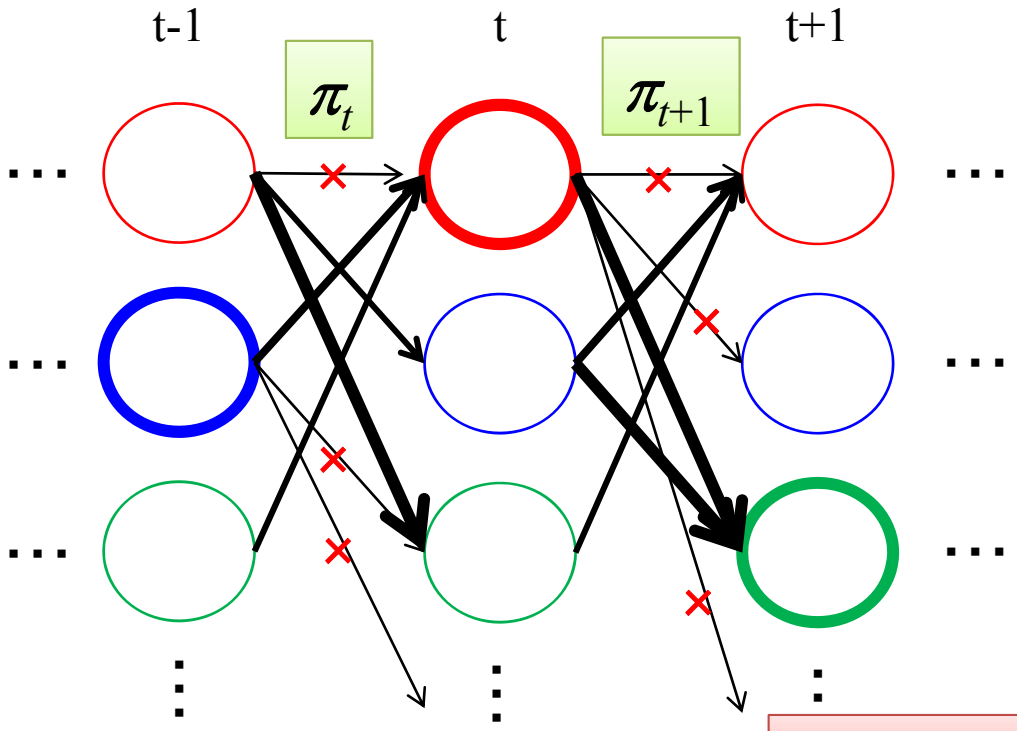


推論方法

- Stick-breaking processの有限近似解
  - ☺ 計算が早い、分かり易い計算式
  - ☹ 近似解、導出が難しい
- Chinese Restaurant Franchise (CRF) + Gibbs
  - ☺ 導出・実装が簡単、大域解（漸近的に）
  - ☹ 収束がかなり遅い
- Slice sampler　今回はこれ！！
  - ☺ 収束はそこそこ、大域解（漸近的に）
  - ☹ 導出が難しい

Slice samplerによる系列サンプリング(van Gael et al., 2008)

確率的にクラスタ数を有限個に打ち切ることで
forward-backwardによる効率的な系列サンプリングができます



$K = K^* < \infty$

auxiliary variableで打ち切ります

$u_{t,i} \sim \mathrm{Uniform}(\pi_{t,z_{t-1},i,z_{t,i}})$

$u_{t,i} < \pi_{t,k,l}$ → possible path

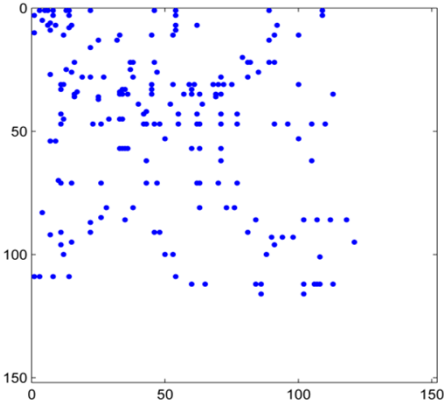
$u_{t,i} >= \pi_{t,k,l}$ → forbidden path

可能なパスが有限個になるので
forward-backwardで
効率的に系列をサンプリングできます

実験データ: Enronの社内 E-mail network (Klimat&Yang, 2004)

- 2001年(破たんの年)の毎月のE-mailのやり取りを記録

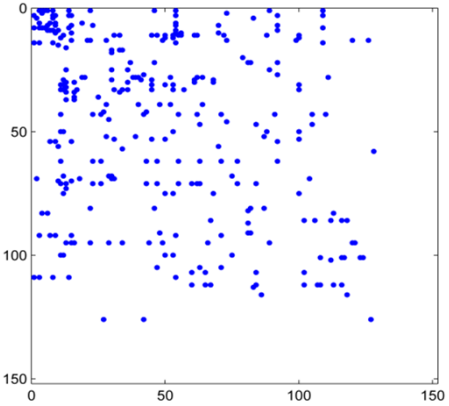
2001/01



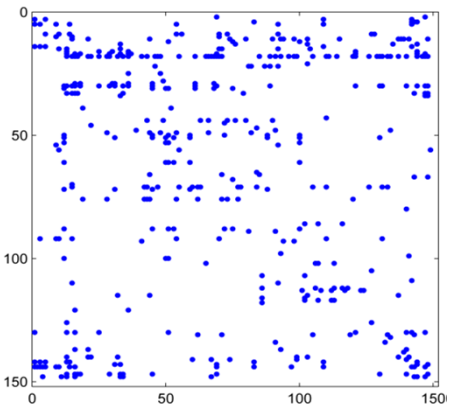
送信した社員
(オブジェクト *i*)

受信した社員
(オブジェクト *j*)

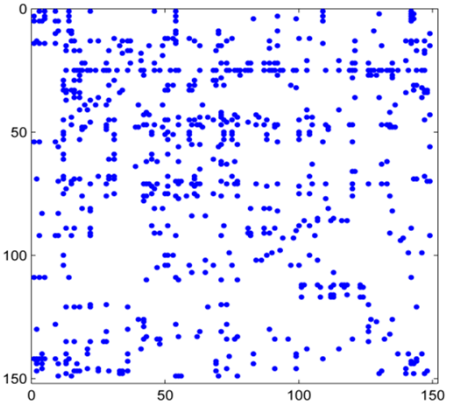
2001/04



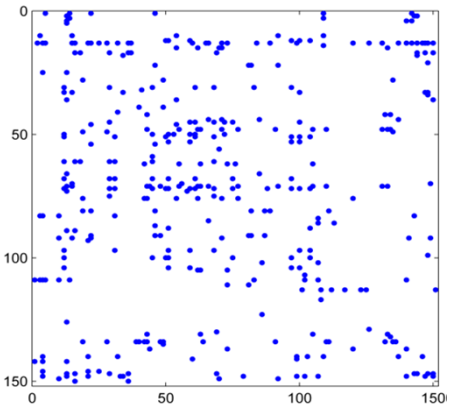
2001/08



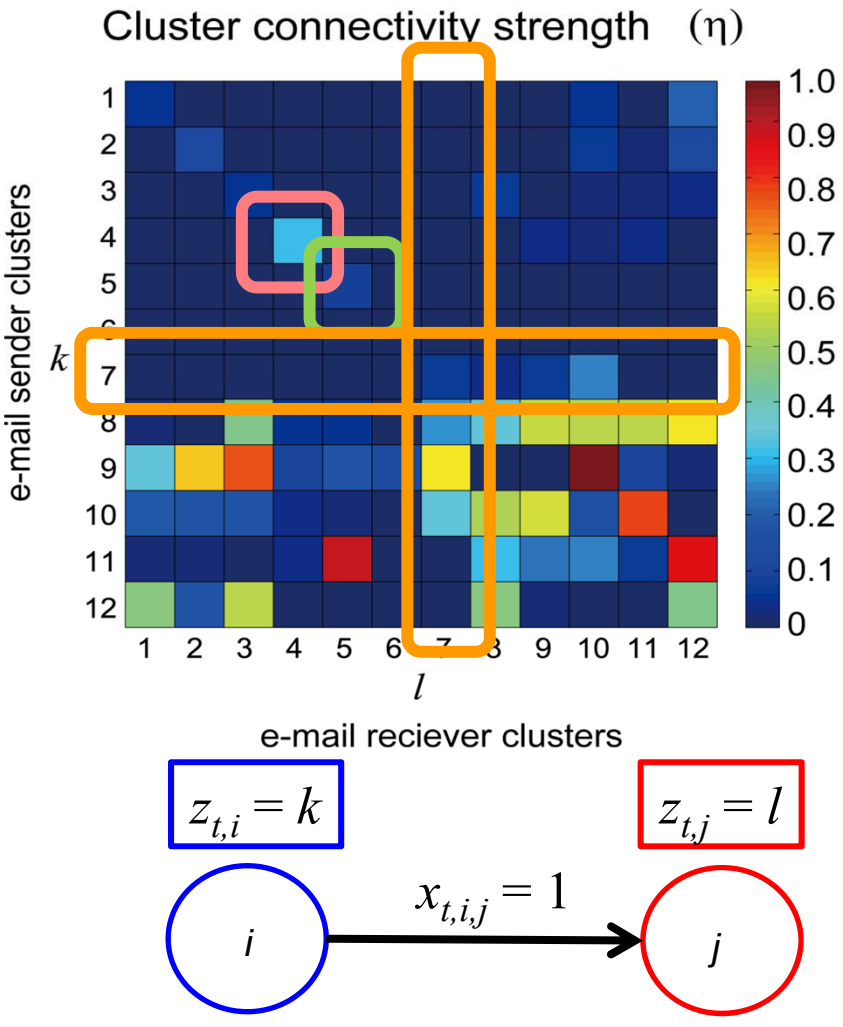
2001/10

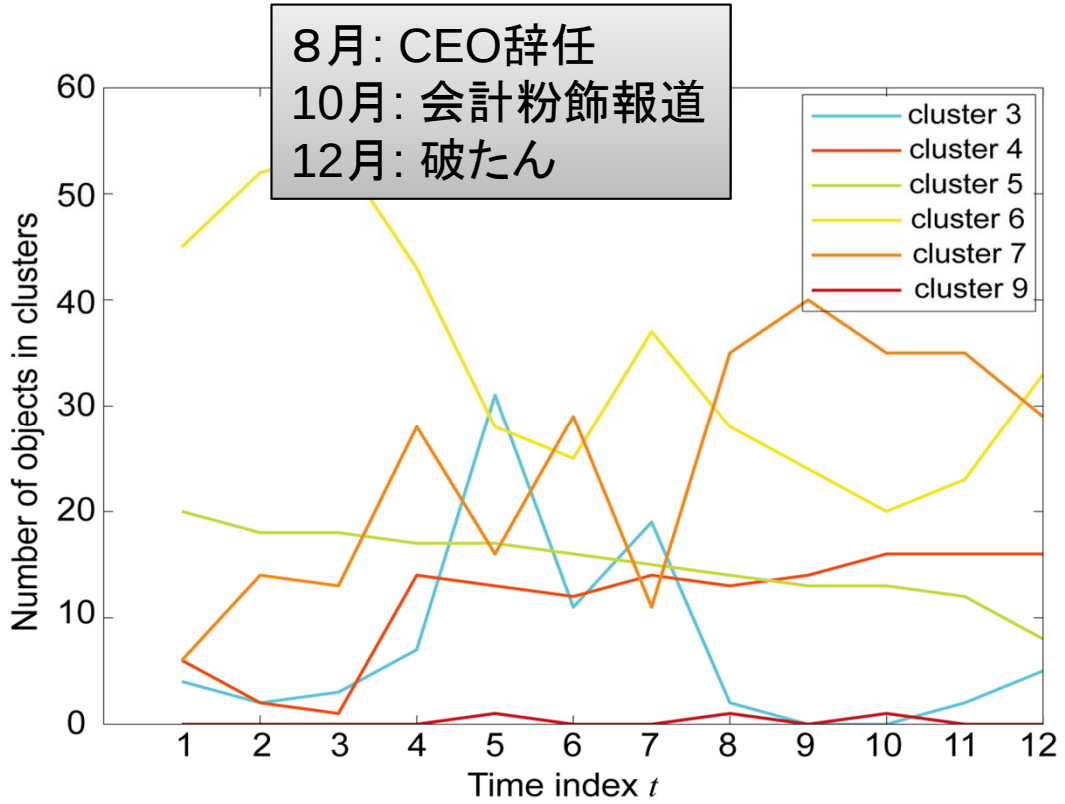


2001/12



実験結果(1/2)



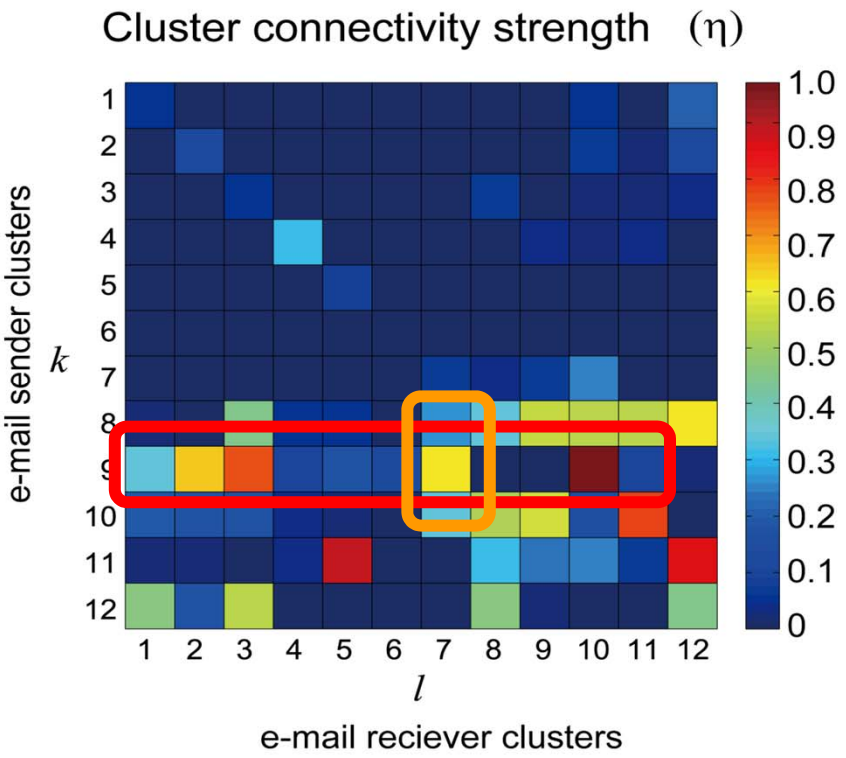


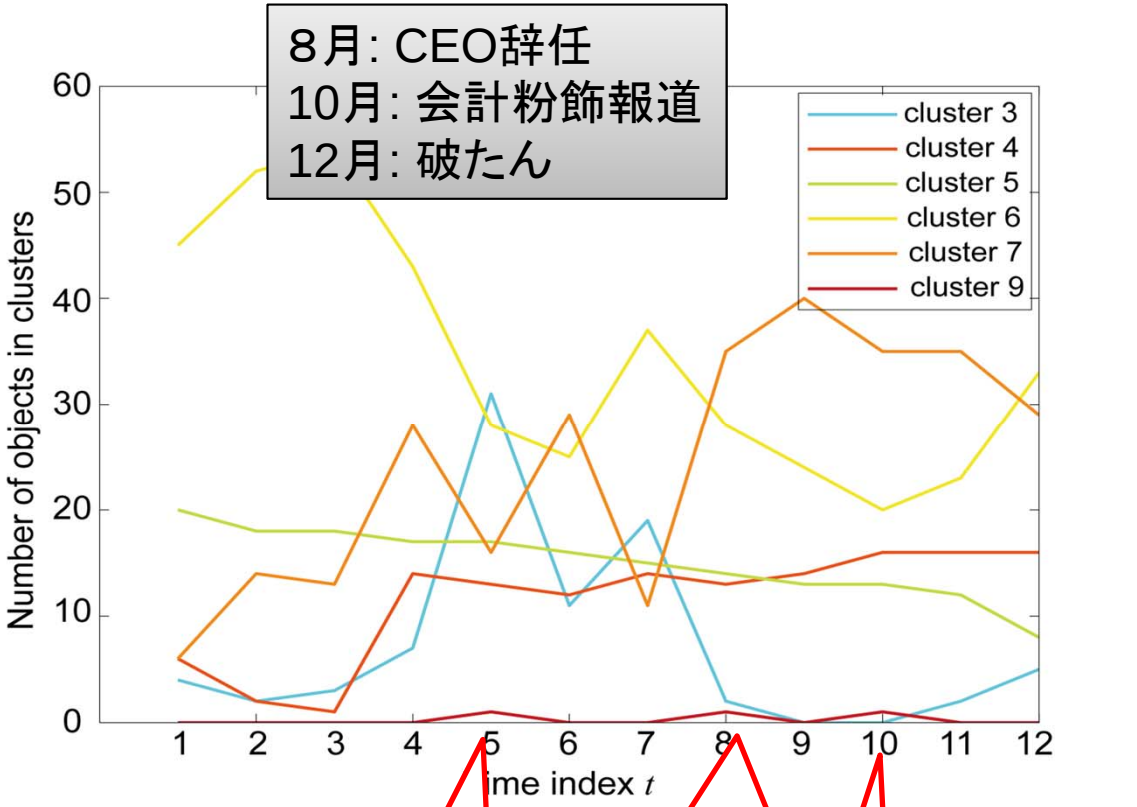
クラスタ４：規制ビジネスコミュニティ

クラスタ５：財務・金融コミュニティ

クラスタ７：管理職およびその関係者

実験結果(2/2)





クラスタ７：管理職およびその関係者

クラスタ９：管理職クラスタに多くメールを送った「時の人」

the CEO of Enron America

the founder, the chairman

the COO

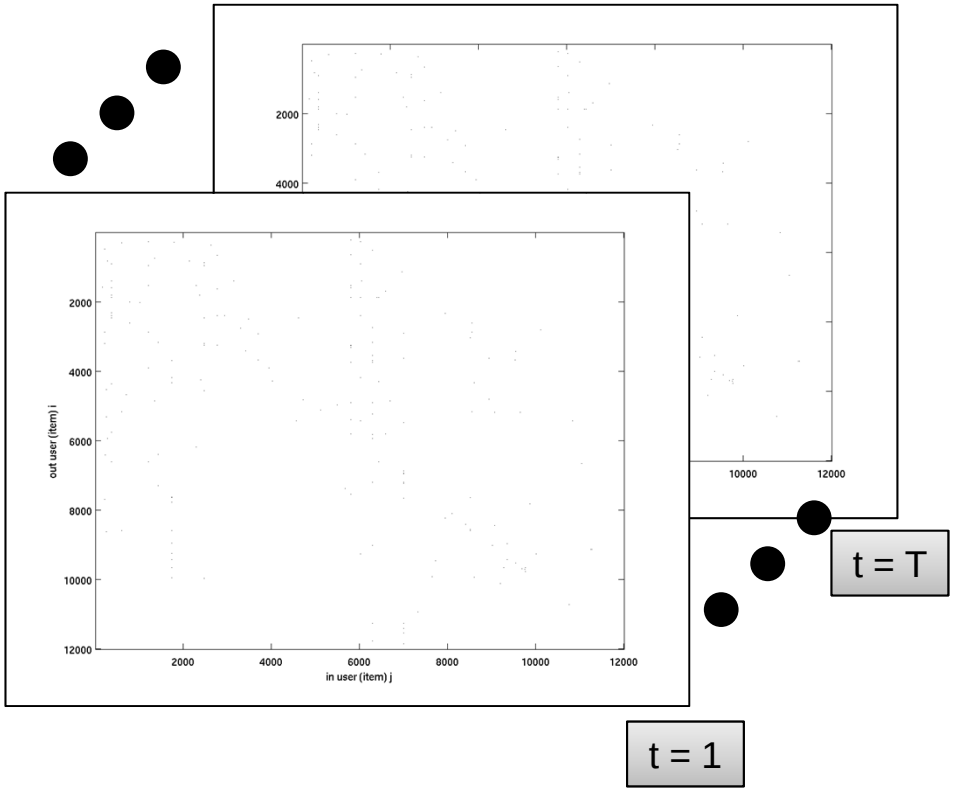
まだまだ問題は残っています・・・

特にweb関係のデータはスパースです。
全部のデータをクラスタリングする必要ってあるのでしょうか？

問題1：ほとんどの部分は“空”
→☹ クラスタリングしても特に嬉しいことがない

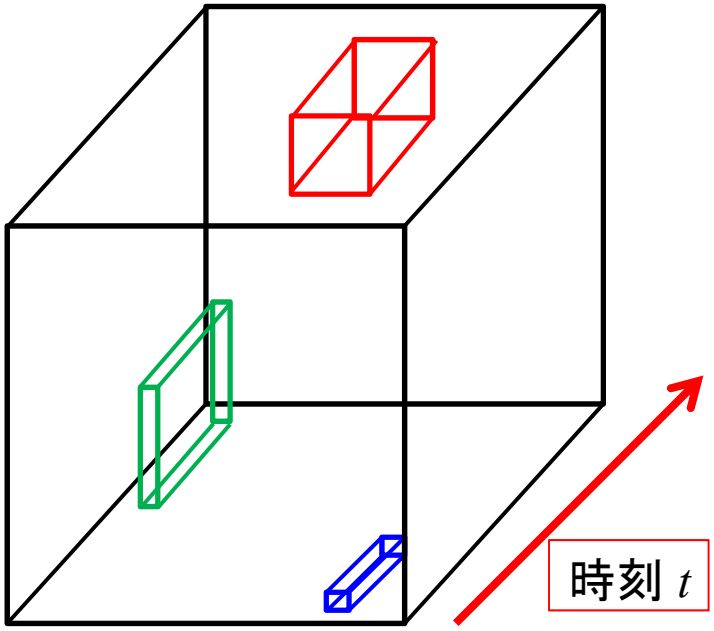
問題2：どのオブジェクトもほとんどリンクを
持ってないという点では同じ
→ ☹ 解釈不可能なクラスタばかり

動画投稿サイトのユーザ間お気に入り関係



時系列関係データの
サブセットクラスタリング

結局、大部分のデータエントリーには興味がなく、
“キモ”の部分の要約情報だけあればいい
→ 全データでなく、重要な部分だけのクラスタリング



ポイント1： 時刻によってクラスタリングすべき
オブジェクト集合が変化する

ポイント2：“ローカルな重要部分”のクラスタ構造の
パラメータもおそらく異なる

ポイント3：全体を通して情報をあまり落とさずに
クラスタリングの整合性をとりたい

ポスターで第一歩を発表します：まずは行列レベルです

特徴（観測量）の取捨選択を行いながらクラスタリングを行う
サブセットクラスタリング法の時系列・関係データ拡張

(cf. Hoff, 2005; Guan et al., 2011)

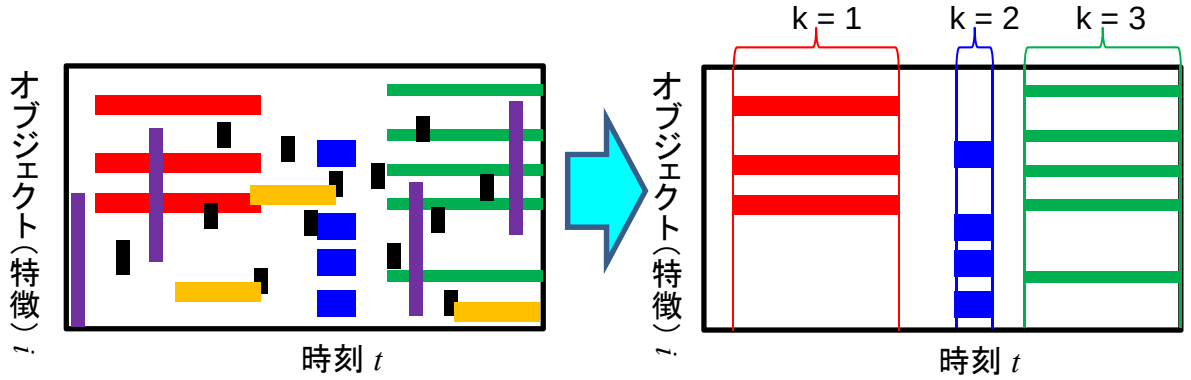
D-144

動的サブセットクラスタリング

石黒　勝彦、上田　修功、澤田　宏 (NTT)

目的: 時系列データから、**時間特異的に重要な部分だけを**
クラスタリングしたい

例えば:
ある期間で特徴的な購買パターン（ブーム）を抽出
大量のセンサーデータから、役立つ部分だけをまとめて発見



まとめ

- 非定常・非線形・非連続な時間変化を見せる関係データのため、ノンパラメトリックベイズに基づく生成モデルをご紹介しました

- 時系列関係データのfuture workの一つとして、サブセットクラスタリングという問題を提示させていただきました

- テンソル分解手法の応用可能性、あるいはサブセットクラスタリング問題について、ご教示・議論をいただければ幸いです

参考文献

- Kemp et al., “Learning Systems of Concepts with an Infinite Relational Model”, AAAI, 2006.
- Ishiguro et al., “Dynamic Infinite Relational Model for Time-varying Relational Data Analysis”, NIPS, 2010.
- Fox et al., “An HDP-HMM for Systems with State Persistence ”, ICML, 2008.
- van Gael et al., “Beam Sampling for the Infinite Hidden Markov Model”, ICML, 2008.

- B. Klimat and Y. Yang. “The enron corpus: A new dataset for email classification research”, ECML, 2004.

- Hoff, “Subset Clustering of Binary Sequences, with an Application to Genomic Abnormality Data”, Biometrics, 2005.
- Guan et al., “A Unified Probabilistic Model for Global and Local Unsupervised Feature Selection”, ICML, 2011.

ご清聴ありがとうございました

- 非定常な時系列関係データの解析に関する研究
- 石黒　勝彦
(NTT コミュニケーション科学基礎研究所)
- E-mail: ishiguro.katsuhiko@lab.ntt.co.jp
- Twitter: @k_ishiguro