

時系列関係データにおける 非定常な潜在構造の推定

NTT コミュニケーション科学基礎研究所

石黒 勝彦

ishiguro.katsuhiko@lab.ntt.co.jp

何の数字でしょうか？

約1500万人



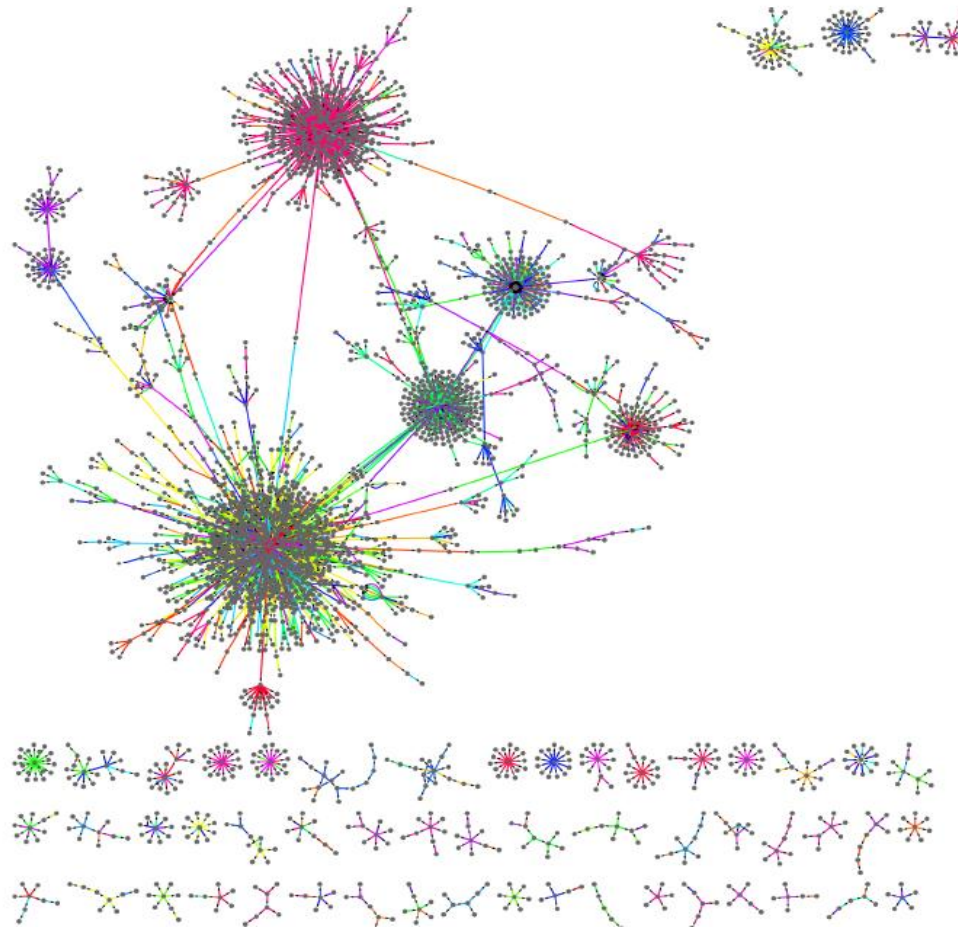
約700万人



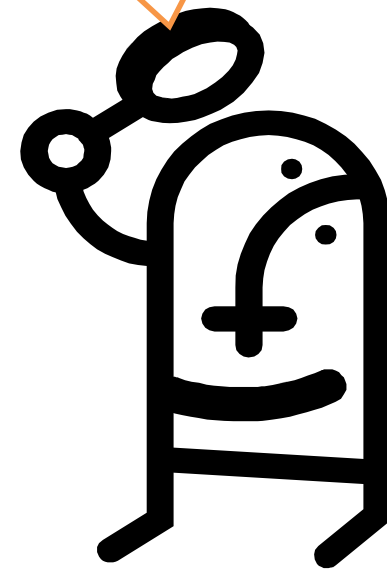
日本国内のSNS利用者数（2011年4月）
（ネットレイティングス社調べ）



SNSは巨大な “関係データ”です

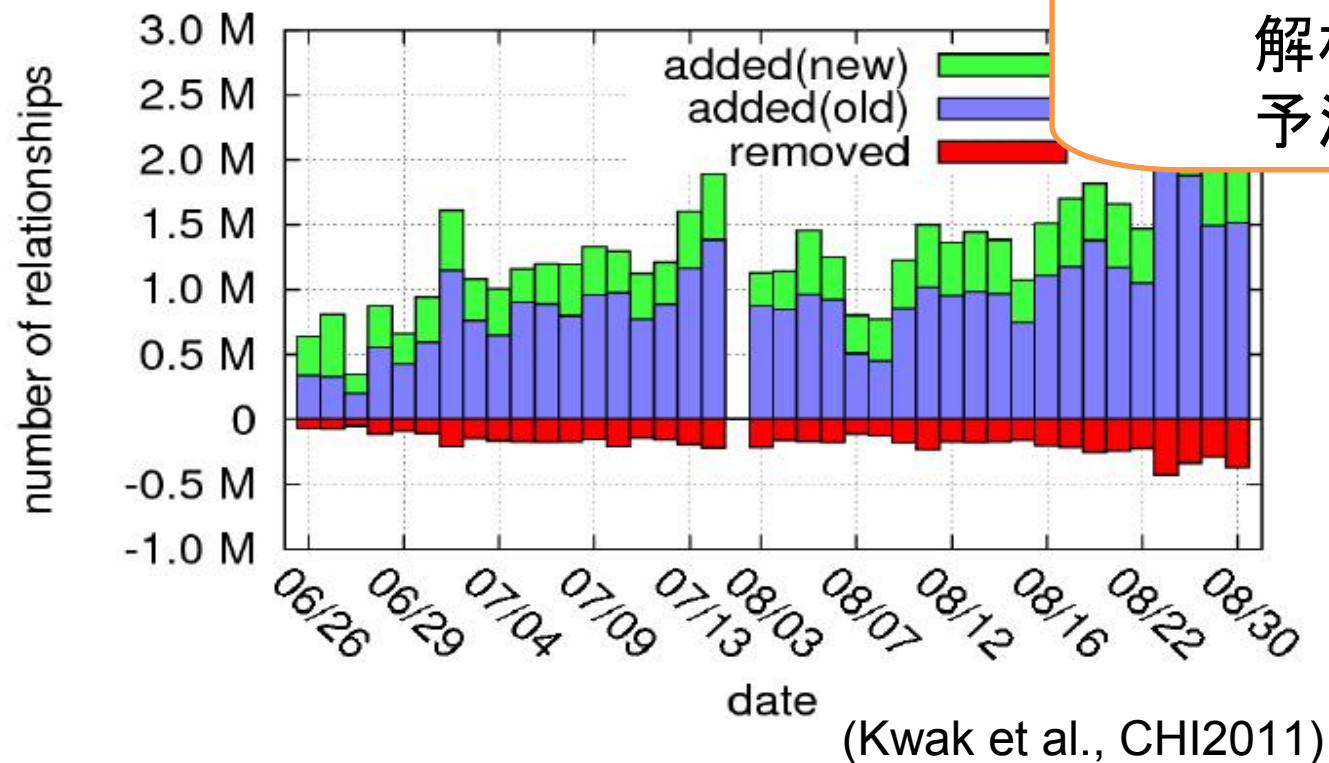


沢山の関係を
解析したい！
予測したい！

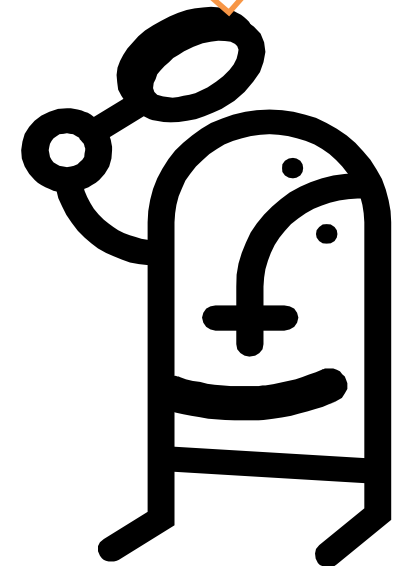


“Retweet Tree” (Kwak et al., WWW2010)

本発表の興味： “関係”は時間で変わります



関係の時間変化こそ
解析したい！
予測したい！

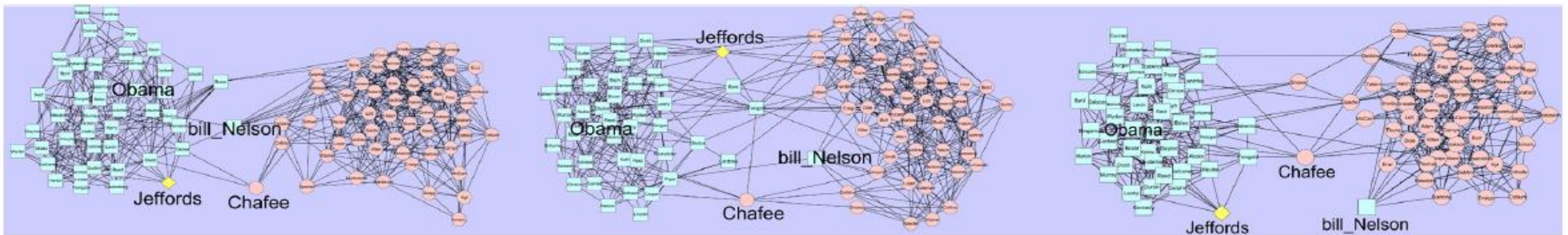


概要

- 時系列関係データ
 - データの定義、解析手法のマップ
- 時系列関係データの潜在構造推定：研究例1
 - 複雑な潜在構造、連続な時間発展
- 時系列関係データの潜在構造推定：研究例2
 - シンプルな潜在構造、非定常な時間発展

(時系列)関係データ

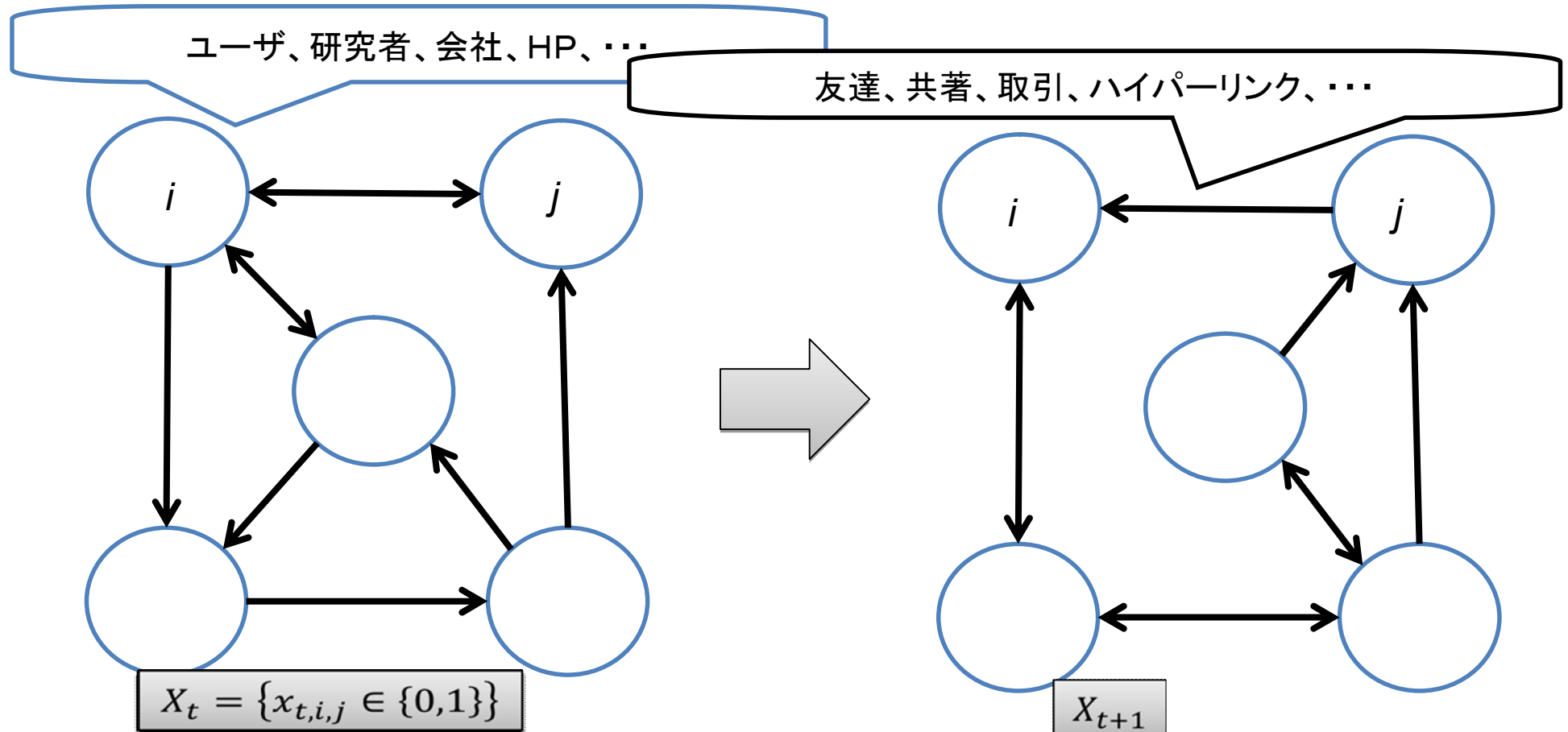
- オブジェクト間の関係の有無を表現したデータです
 - SNS上の友人関係、論文の引用関係、HP間のハイパーリンクなど
- “時系列”関係データは、静的な潜在構造の推定とともに、構造の時間変化という面白い問題も提供してくれます



“dynamic senator network” (Ahmed & Xing, 2009.)

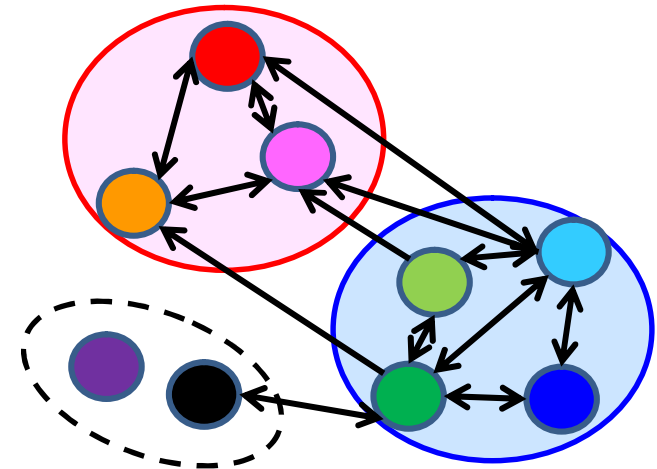
最もシンプルな形態の 時系列関係データ

オブジェクトをノード、関係をエッジとする
時間発展ネットワーク(グラフ)データとして解析可能です



本発表のスコープ： ノードのクラスタリング

- ノードのもつリンク関係をもとにして、時系列関係データ上から類似ノードを抽出します
 - ユーザコミュニティ抽出
 - 遺伝子の機能的分類
- ☹ 巨大データへのスケーラビリティは目をつぶります



時系列関係データの クラスタリング手法マップ

講演者の主観による

離散
(非定常)

時間
発展
モデル

連続
(なめらか)

Ishiguro et al., NIPS2010
重複なし無限クラスタリング
離散発展モデル(HMM)

Yang et al., SDM2009.
重複なしクラスタリング
離散発展モデル(HMM)

Kim&Han, VLDB2009.
時間・空間の局所近傍クリーク数え上げ
時刻間の相互情報量最大化

Ahmed&Xing, PNAS2009.
ノード間コネクション強度の推定
L1 sparsityによるクラスタ発見
連続発展モデル(差分最小平滑化)



Fu et al., ICML2009.
重複ありクラスタリング
連続発展モデル(State-space)

Tang et al., SIGKDD2008.
multi-modeネットワーク
mode内で重複なしクラスタリング
連続発展モデル(差分最小平滑化)

simple

潜在クラスタ構造モデルの複雑さ

complicated

dMMSB法

(dynamic Mixed Membership Stochastic Block model)

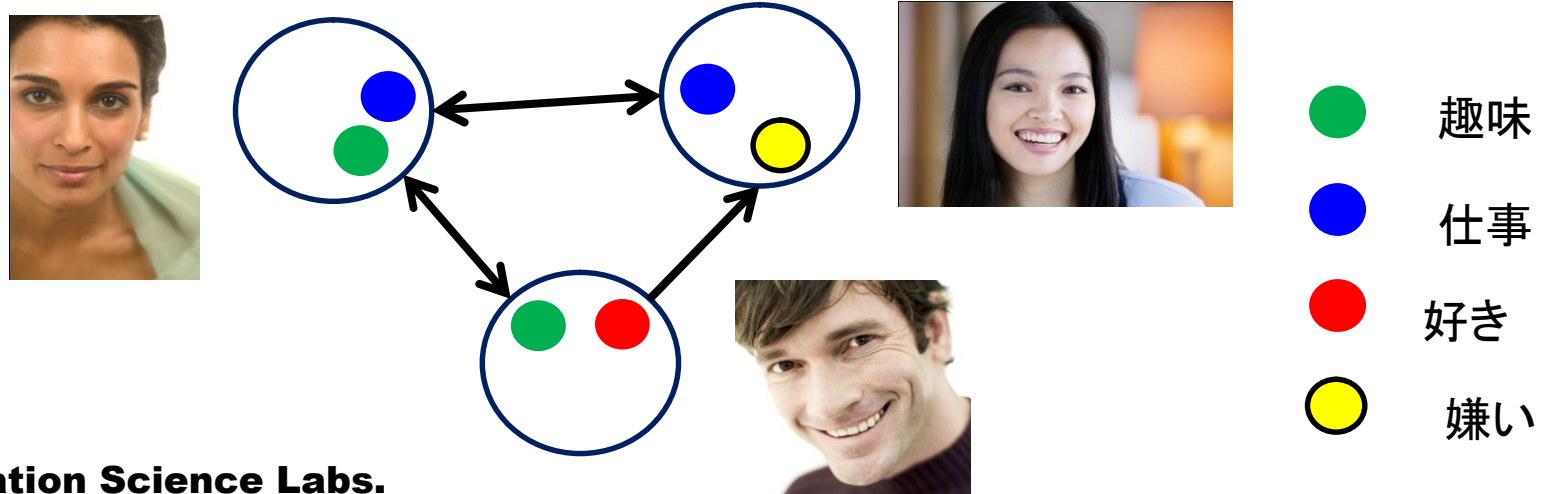
【潜在構造：複雑】
【時間発展：連続】

Fu et al.,
“Dynamic Mixed Membership Blockmodel for Evolving Networks”,
Proc. ICML, 2009.

モデルの着眼点： “関係”は複数の要因から (Airoldi et al., 2008)

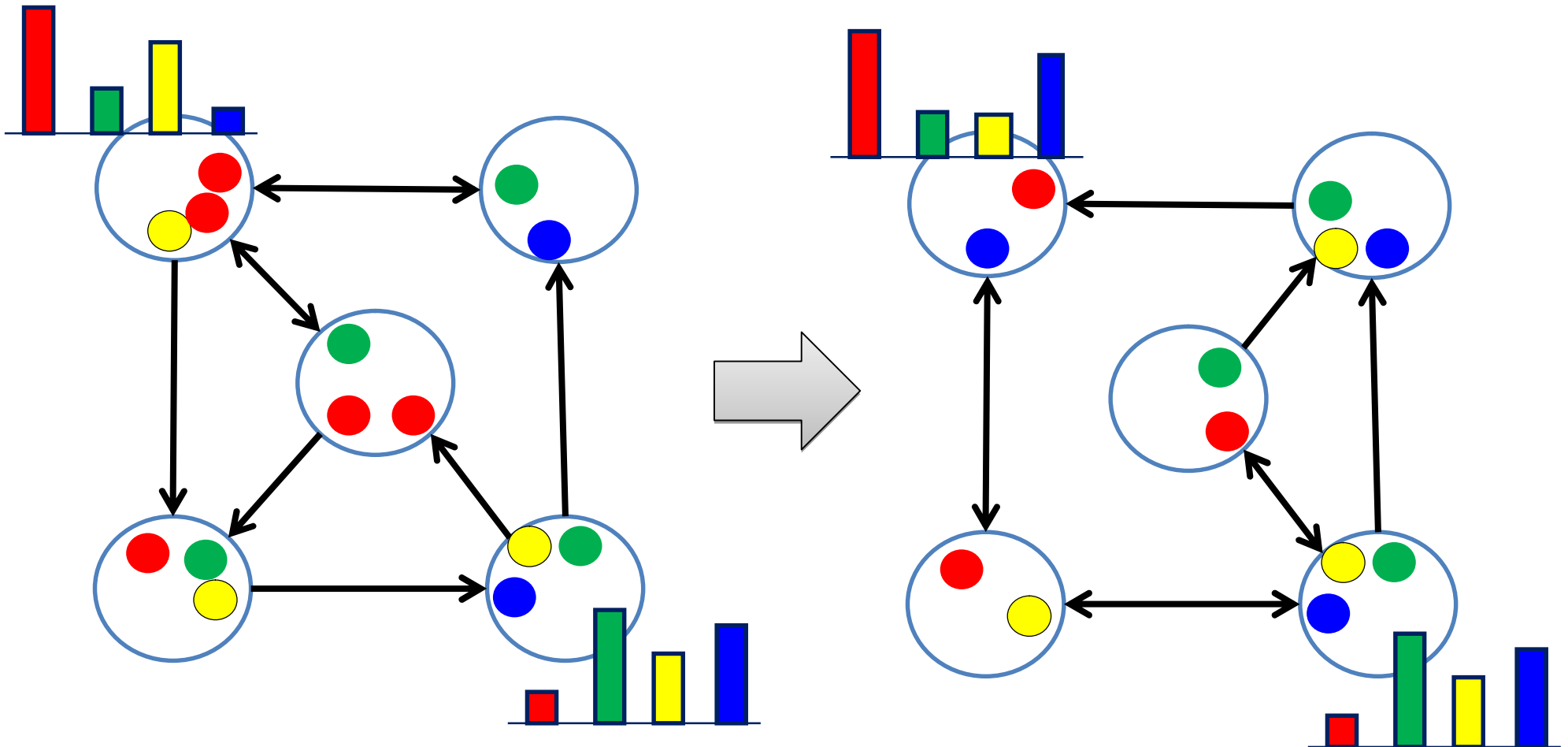
- 観測された関係データの背後に、複数の種類のroll(役回り)が潜在している
- 各ノードは、相手に応じてrollを使い分ける

例：人間関係において、人は複数の顔(側面)を切り替えています



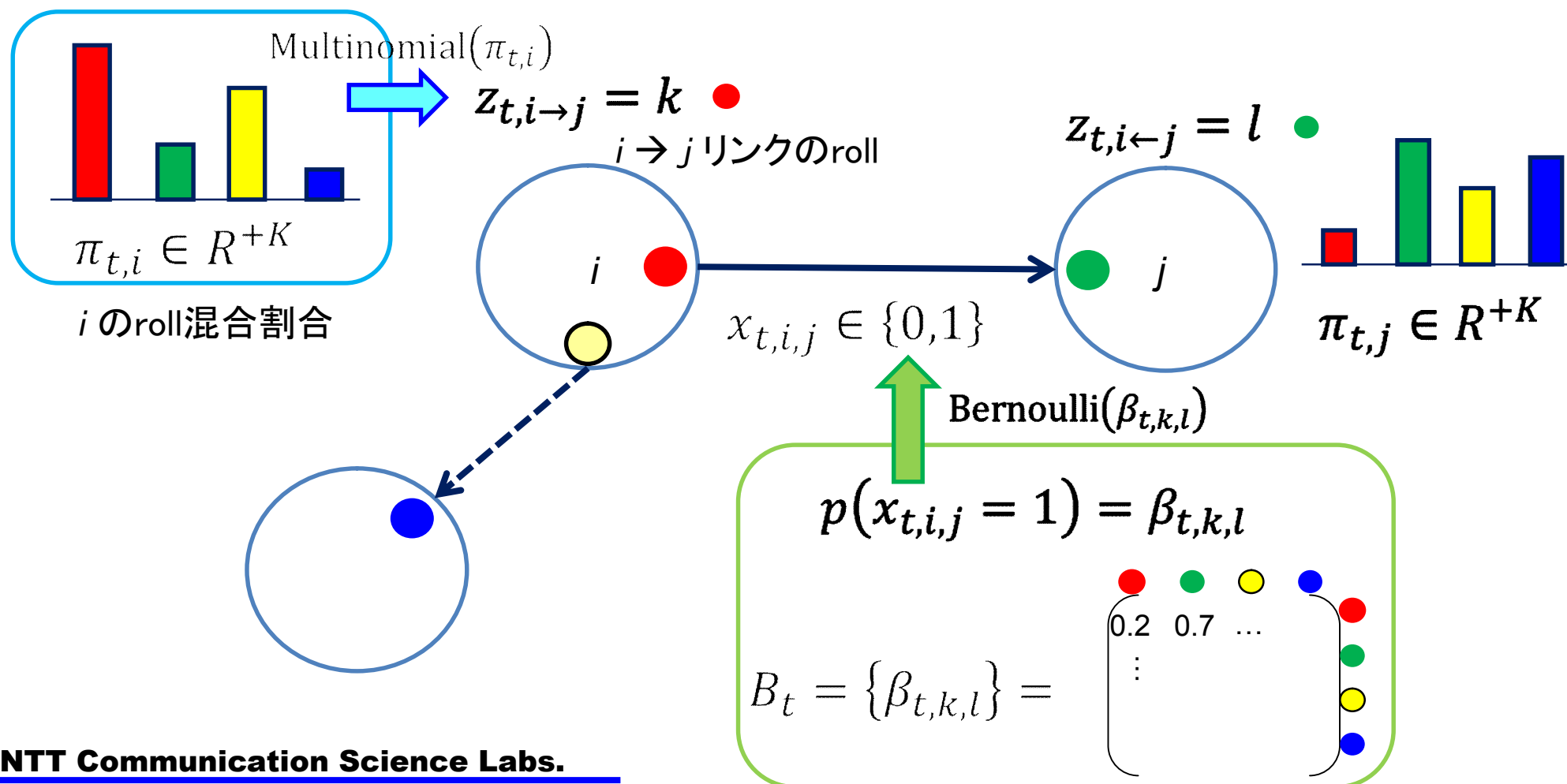
dMMSBの概要

1. 複数のrollの混合確率によるノードのソフトクラスタリングを行います
2. 時間発展は全体で連続的(なめらか)な発展を仮定します



潜在構造のモデル化

すべてのノード間リンクに、役回り(roll)による接続確率を仮定します

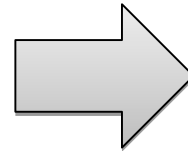


時間発展のモデル化

rollの混合確率や接続確率を
状態空間モデルでなめらかに時間発展させます

$$\eta_t \sim N(b\eta_{t-1}, \psi)$$

$$\beta_{t,k,l} \sim \text{LogisticN}(\eta_t, S_t)$$

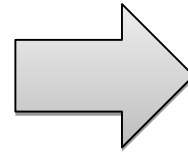


$$\eta_{t+1} \sim N(b\eta_t, \psi)$$

$$\beta_{t+1,k,l} \sim \text{LogisticN}(\eta_{t+1}, S_{t+1})$$

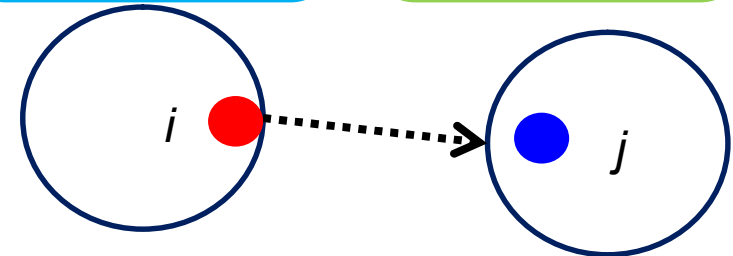
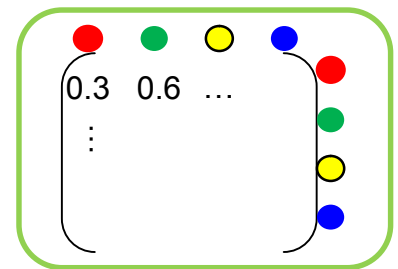
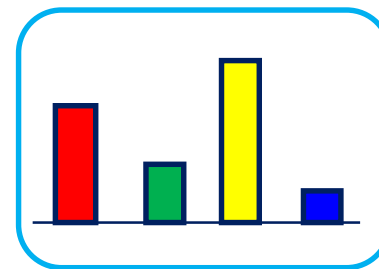
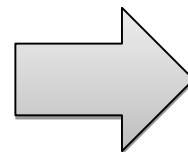
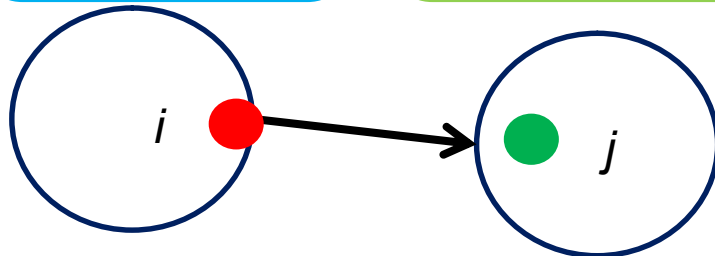
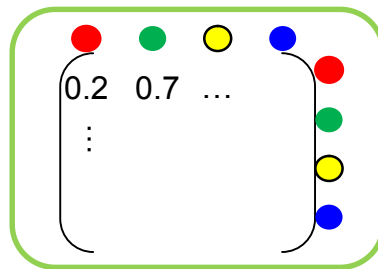
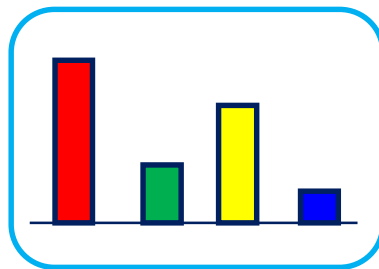
$$\mu_t \sim N(A\mu_{t-1}, \Phi)$$

$$\pi_{t,i} \sim \text{LogisticN}(\mu_t, \Sigma_t)$$



$$\mu_{t+1} \sim N(A\mu_t, \Phi)$$

$$\pi_{t+1,i} \sim \text{LogisticN}(\mu_{t+1}, \Sigma_{t+1})$$



生成モデル

解法: 変分ベイズ

パラメータの初期値

$$\eta_1 \sim N(\iota, \psi)$$

$$\mu_1 \sim N(\nu, \Phi)$$

roleごとの接続確率

$$\eta_t \sim N(b\eta_{t-1}, \psi)$$

$$\beta_{t,k,l} \sim \text{LogisticN}(\eta_t, S_t)$$

⊙ 実際の実験、推論では時刻不変に近似しています

ノードのrole混合確率

$$\mu_t \sim N(A\mu_{t-1}, \Phi)$$

$$\pi_{t,i} \sim \text{LogisticN}(\mu_t, \Sigma_t)$$

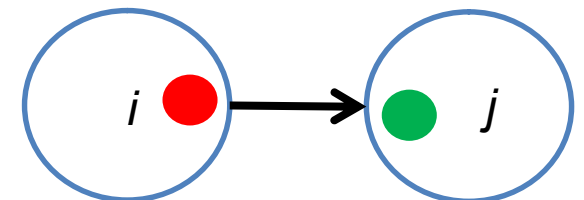
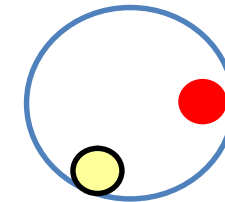
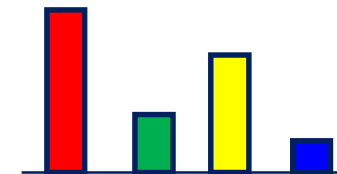
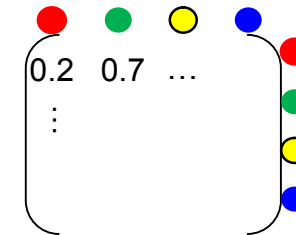
関係ごとのroleアサイン

$$z_{t,i \rightarrow j} \sim \text{Multinomial}(\pi_{t,i})$$

$$z_{t,i \leftarrow j} \sim \text{Multinomial}(\pi_{t,j})$$

観測された関係データ

$$x_{t,i,j} \sim \text{Bernoulli}(\beta_{t,z_{t,i \rightarrow j}, z_{t,i \leftarrow j}})$$

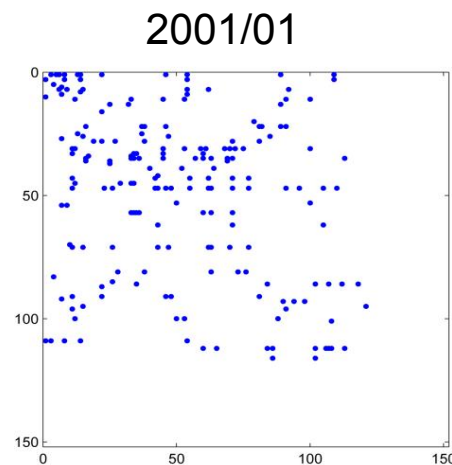


実験データ: Enronの社内 E-mail network (Klimat&Yang, 2004)



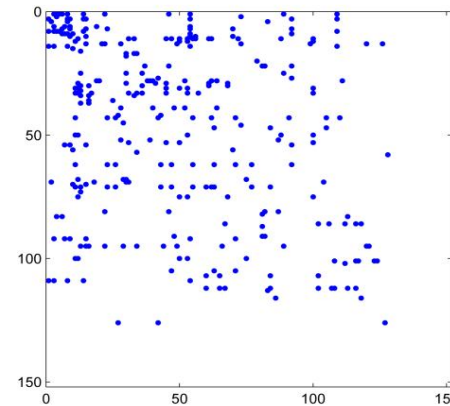
- 2001年(破たんの年)の毎月のE-mailのやり取りを記録

送信した社員
(ノード i)

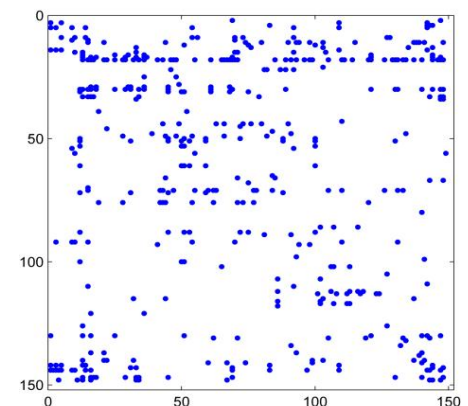


受信した社員
(ノード j)

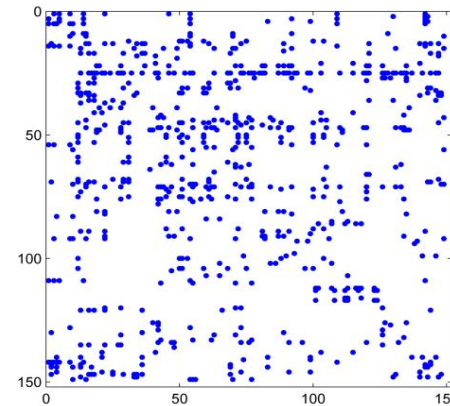
2001/04



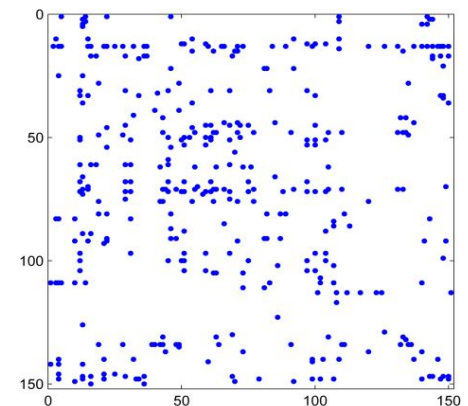
2001/08



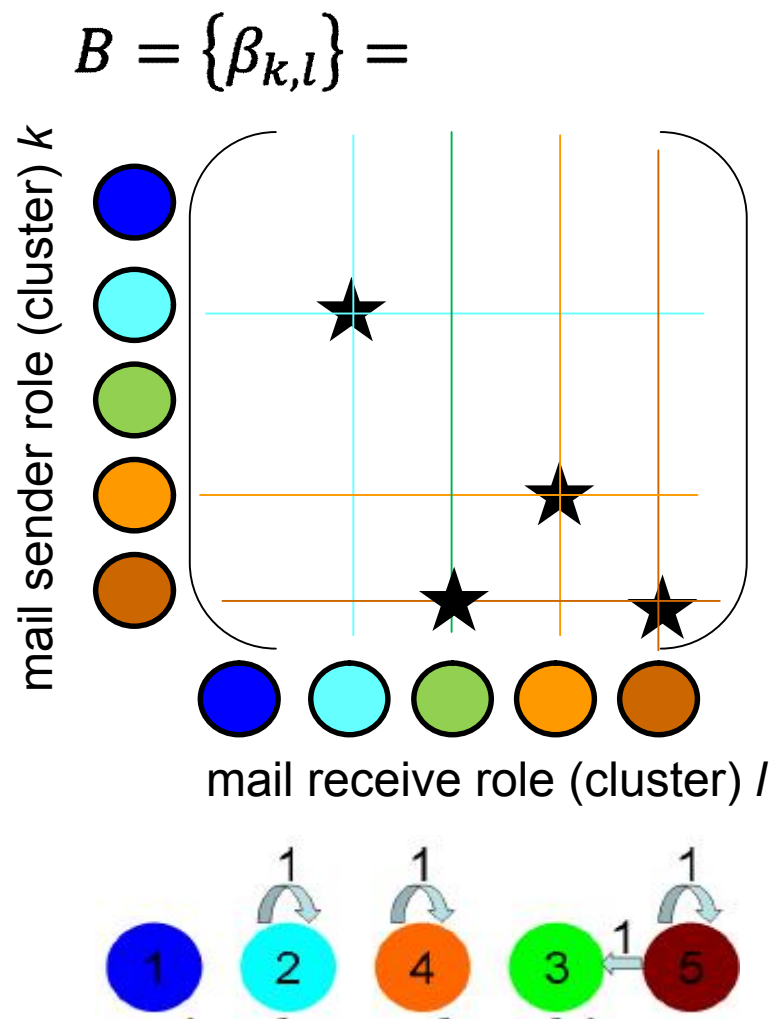
2001/10



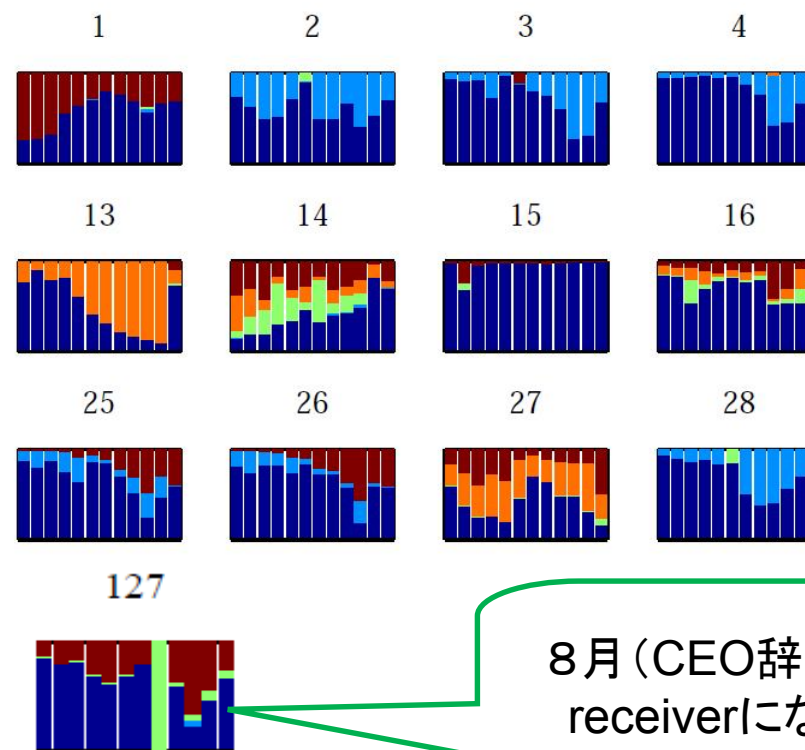
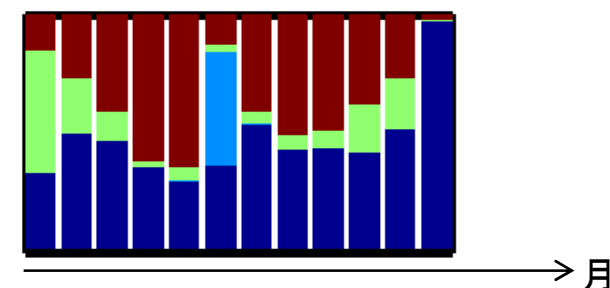
2001/12



実験結果



月ごとの
roll混合比率



8月 (CEO辞任の月) に
receiverになっている

the founder, chairman

dIRM法 **(dynamic Infinite Relational Model)**

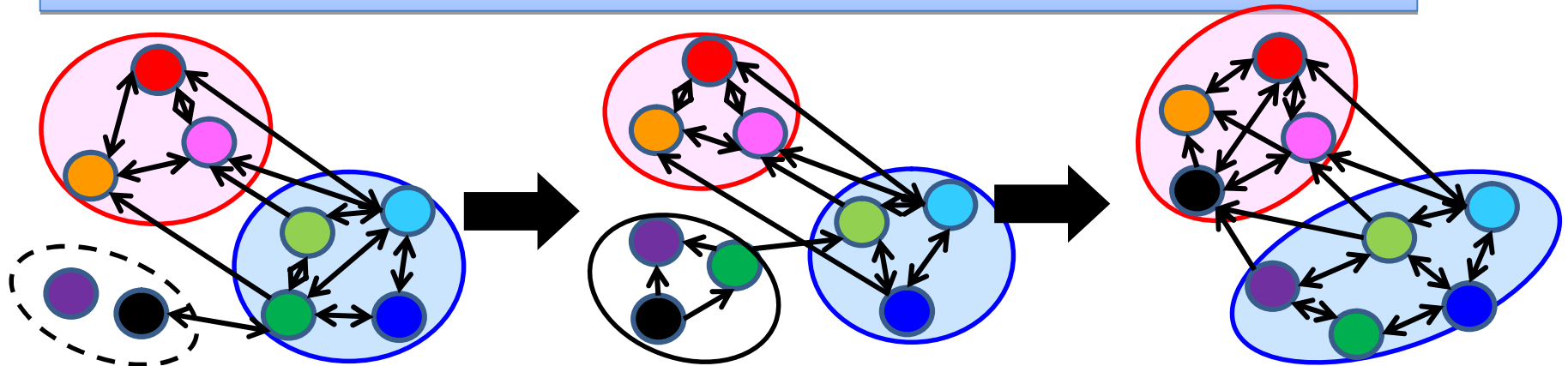
【潜在構造：シンプル】
【時間発展：離散】

Ishiguro et al.,
“Dynamic Infinite Relational Model for Time-varying Relational Data Analysis”,
Proc. NIPS, 2010.

モデルの着眼点： “関係の時間変化”は非定常

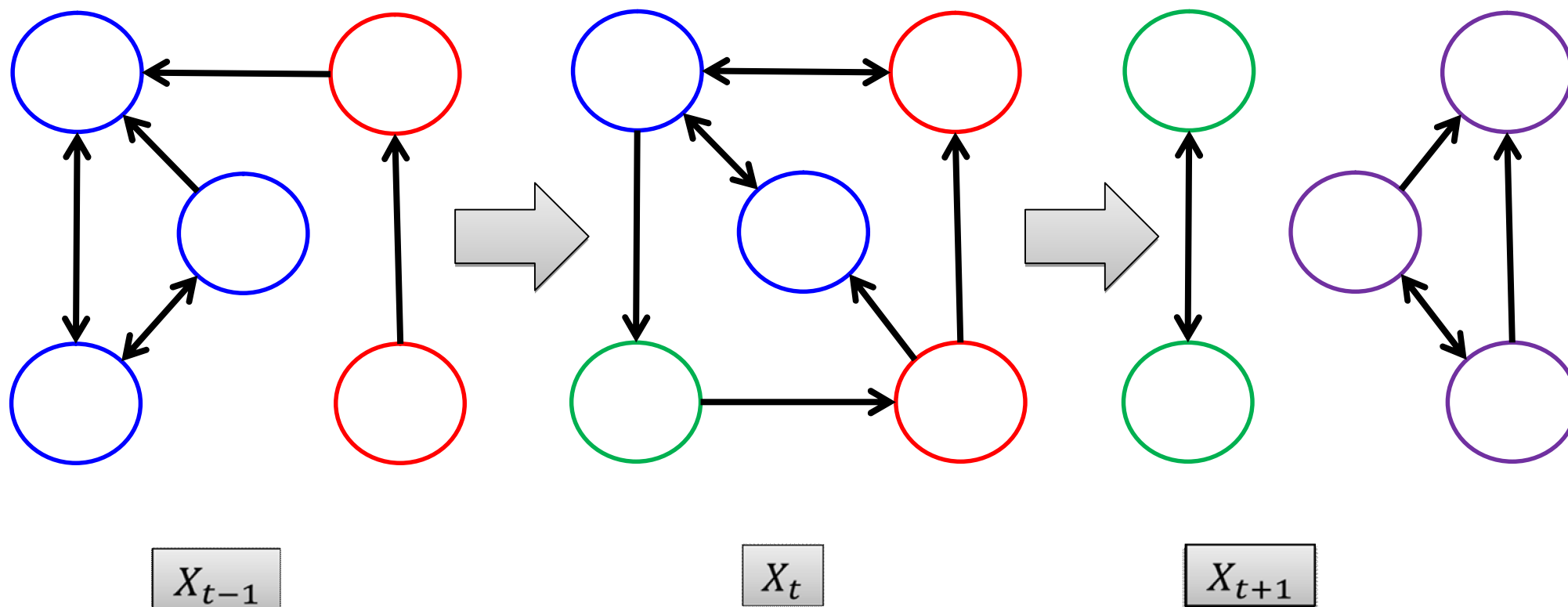
- 関係データは、非連続・瞬間的な構造変化を見せることがある
- 関係を大きく変えるような突然の変化を前提としたモデルも必要？

例： 組織内のコミュニティは部門の再編成で突然分裂・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



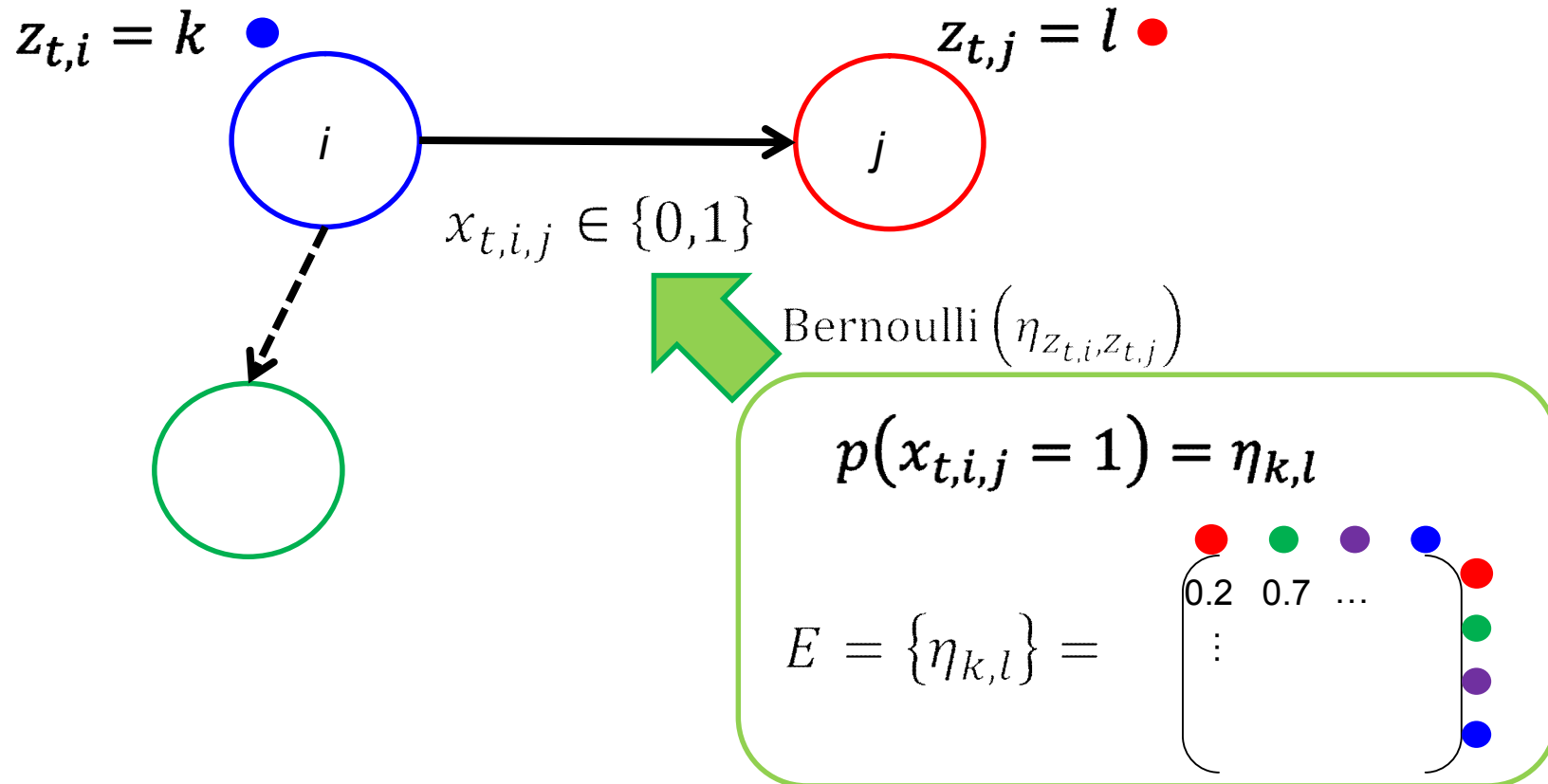
dIRM法の概要

1. 未知数のクラスタへのハードクラスタリングを行います
2. 時間発展は非定常的で急激な変化を許容します



潜在構造のモデル化

dMMSBと違い、各ノードごとに所属クラス(role)は一つだけです



時間発展のモデル化(1/2)

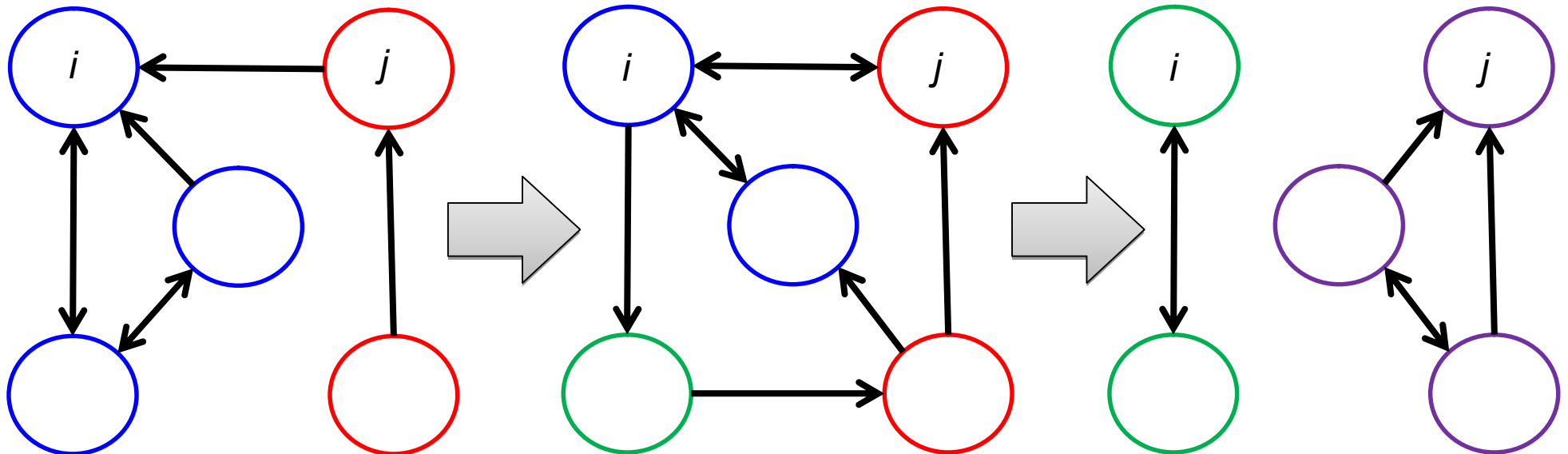
ノードの所属クラスはHMMで離散的に変化します

$$z_{t,i} | z_{t-1,i} = k \sim \text{Multinomial}(\pi_{t,k})$$

$\pi_t, = \bullet$

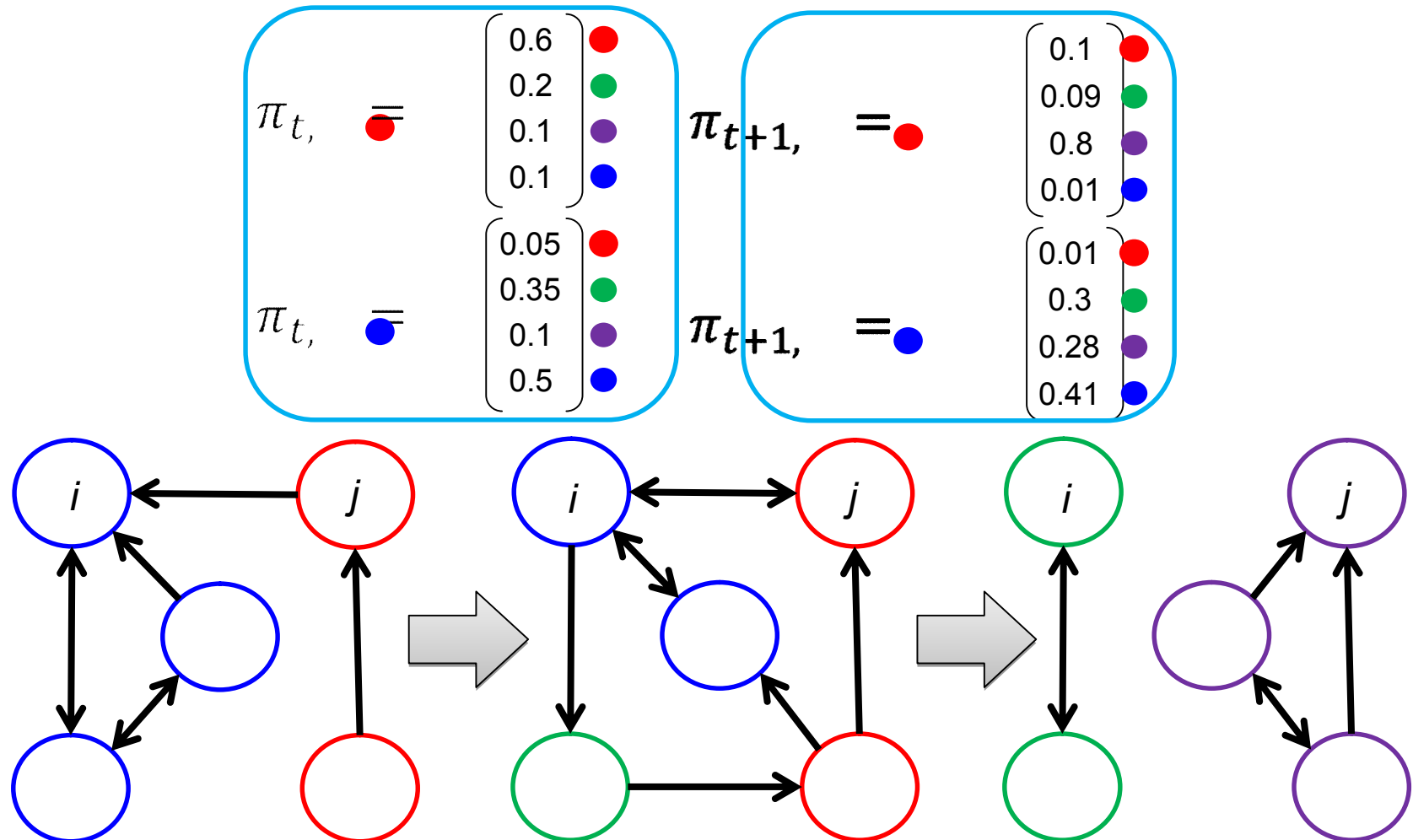
$$\begin{bmatrix} 0.6 \\ 0.2 \\ 0.1 \\ 0.1 \end{bmatrix} \begin{matrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{matrix} = \pi_t$$

$$\begin{bmatrix} 0.1 \\ 0.3 \\ 0.1 \\ 0.5 \end{bmatrix} \begin{matrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{matrix}$$



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に、
クラスタ遷移確率は時間ごとに異なるものと仮定します



非定常と連続性をバランス: sticky infinite HMM (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確率とを
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

$$\beta \sim \text{Stick}(\gamma) \quad \beta_k = v_k \prod_{l=1}^{k-1} (1 - v_l) \\ v_k \sim \text{Beta}(1, \gamma)$$

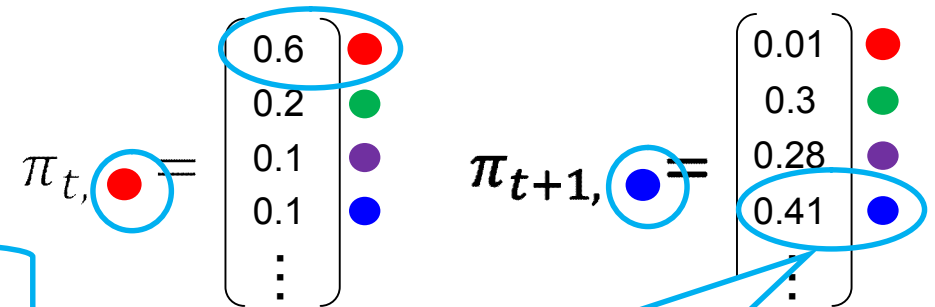
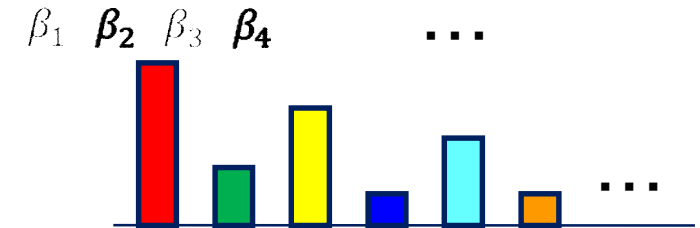
クラスタ間遷移確率

(オリジナルのsticky iHMMを時刻依存に拡張)

※DP(Dirichlet Process)≡無限次元のディリクレ分布

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

δ_k : k番目だけ"1"のベクトル



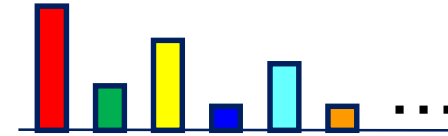
同じクラスタへ遷移しやすくなる

生成モデル

解法: Gibbsサンプリング

時系列全体でのクラスタ割合

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



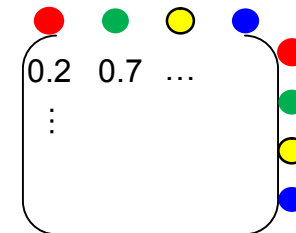
時刻依存の
クラスタ間遷移確率

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

$$\pi_{t,\bullet} = \begin{pmatrix} 0.6 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} \begin{matrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{matrix}$$

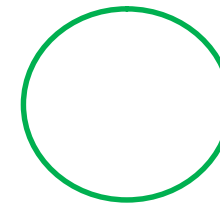
クラスタ間の接続確率

$$\eta_{k,l} \sim \text{Beta}(\xi_k, \eta_l)$$



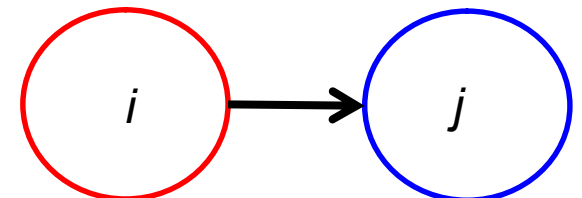
ノードのクラスタアサイン

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

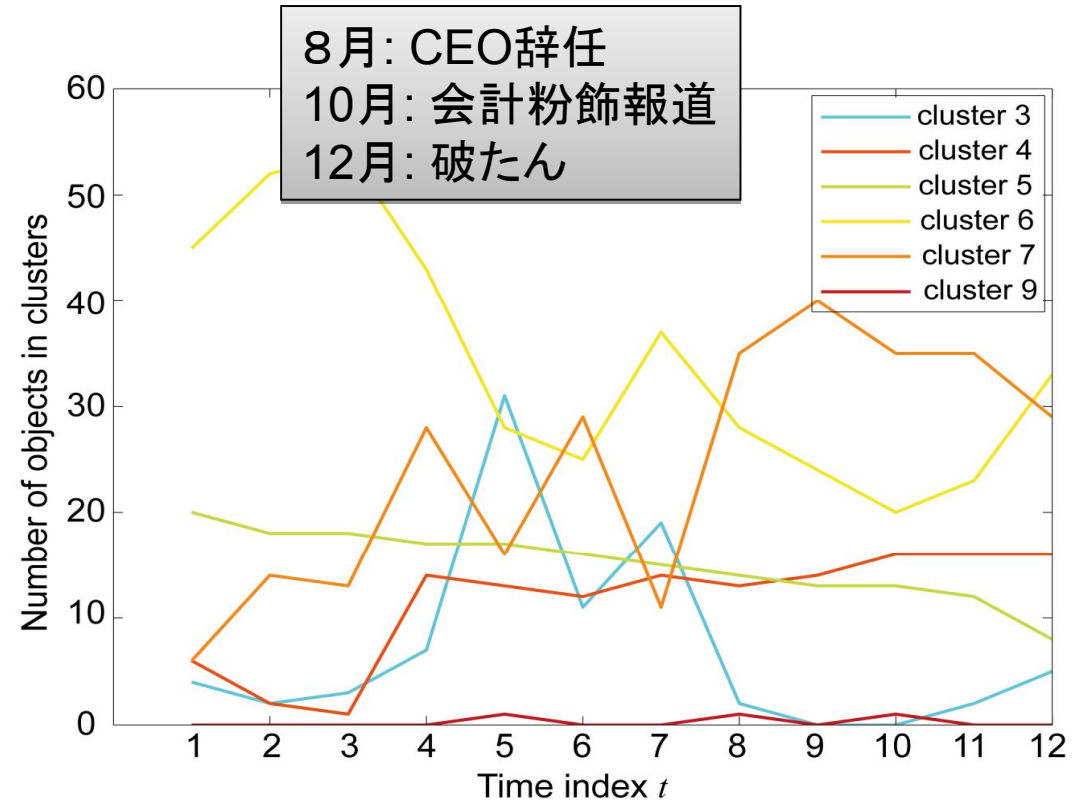
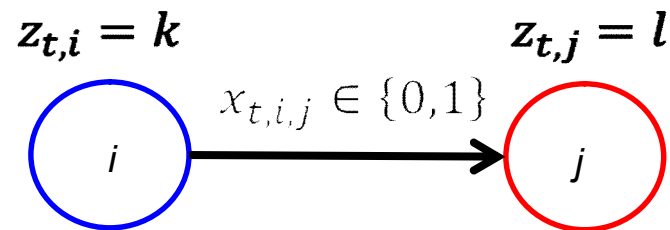
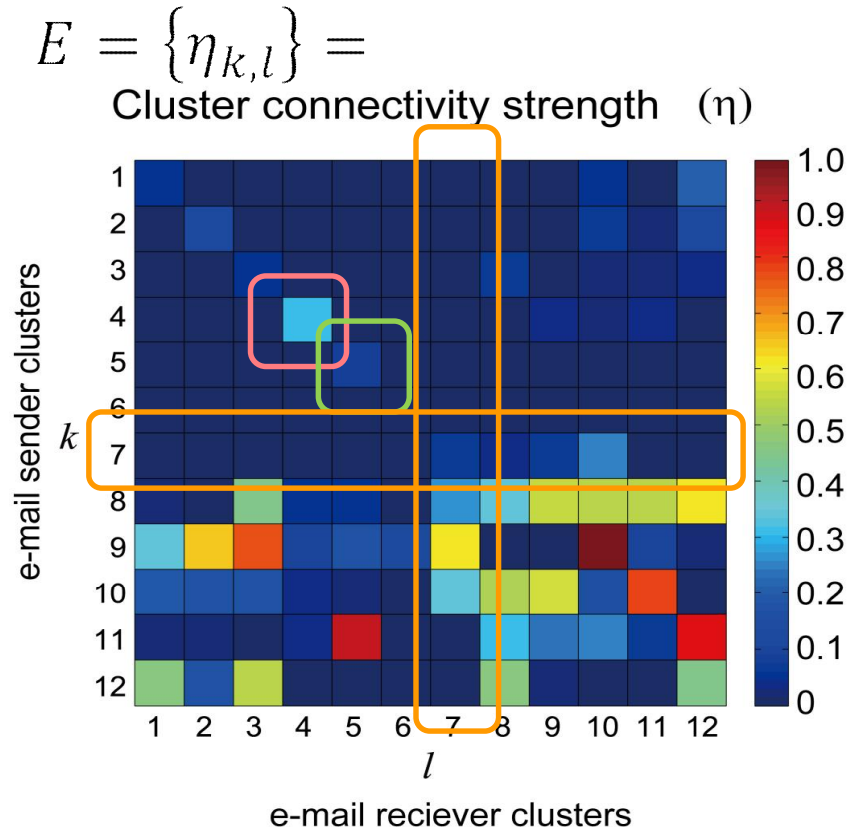


観測された関係データ

$$x_{t,i,j} \sim \text{Bernoulli}(\eta_{z_{t,i}, z_{t,j}})$$



実験結果(1/2)

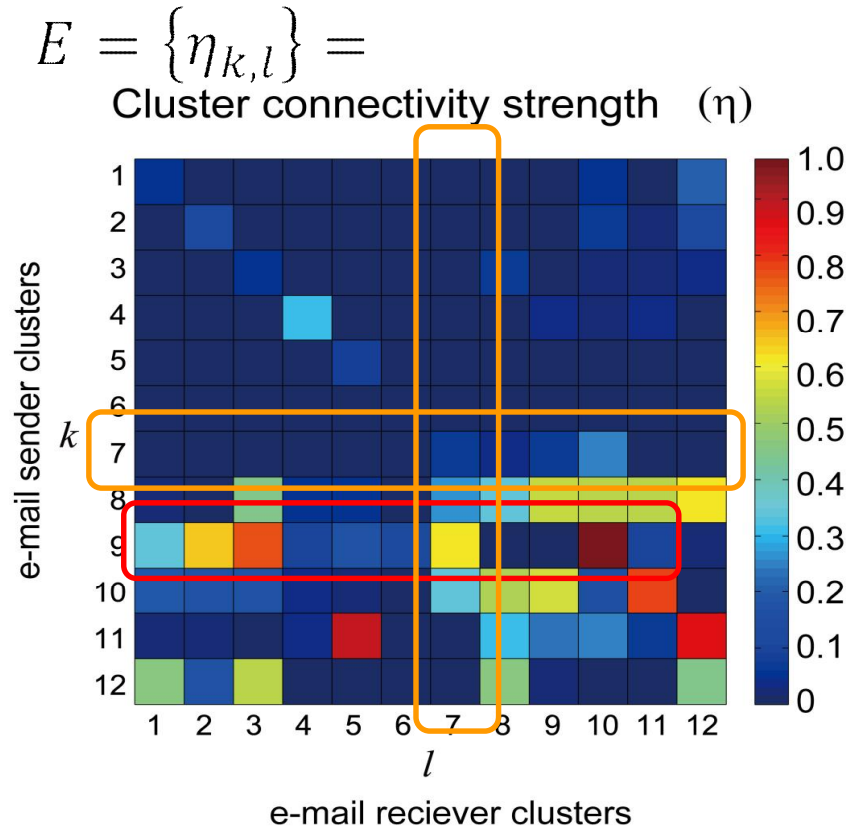


クラスタ4: 規制ビジネスコミュニティ

クラスタ5: 財務・金融コミュニティ

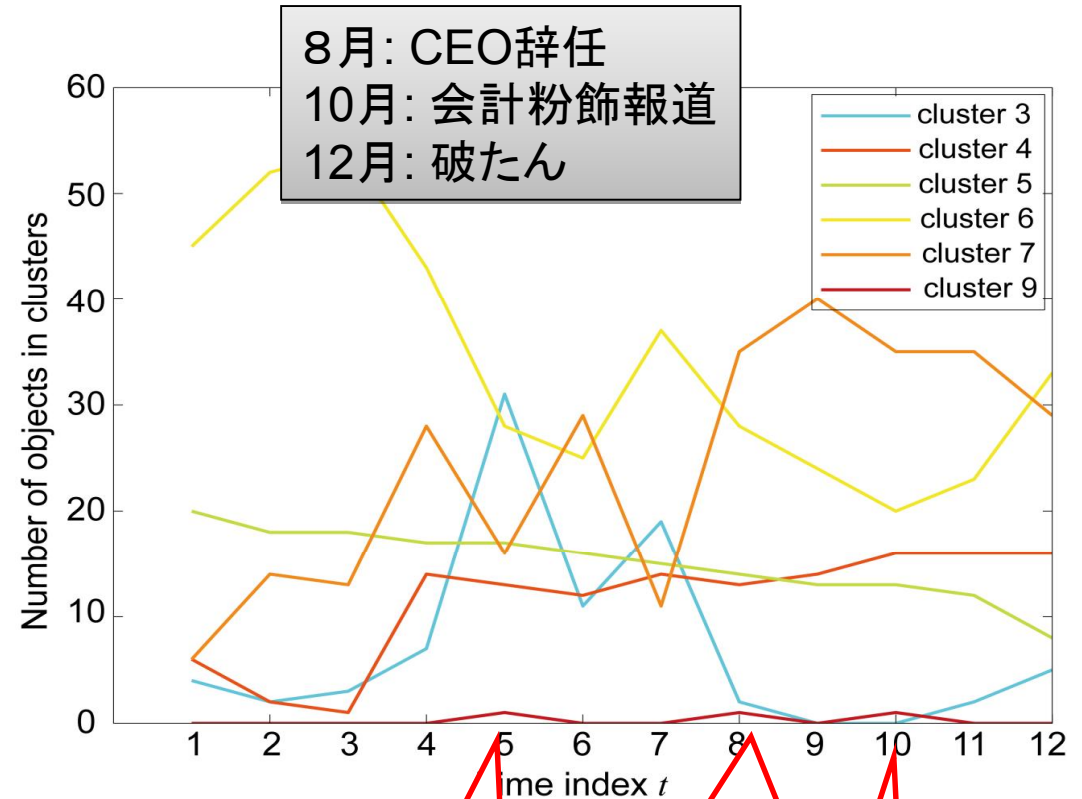
クラスタ7: 管理職およびその関係者

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

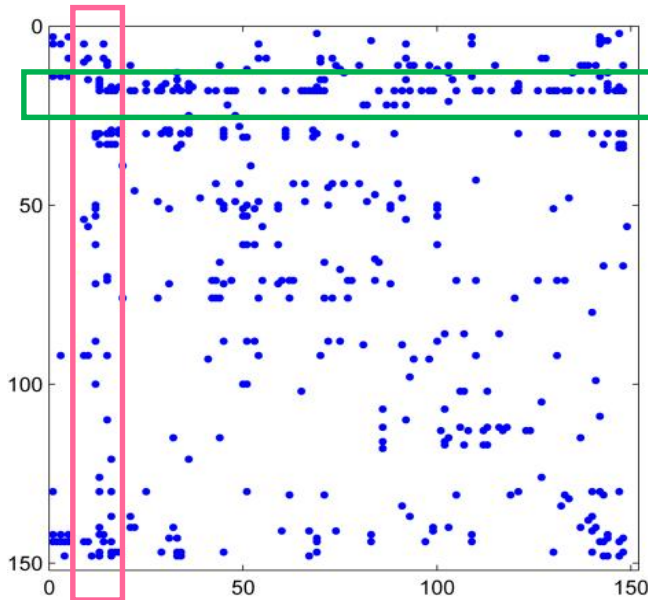
the founder, the
chairman

the COO

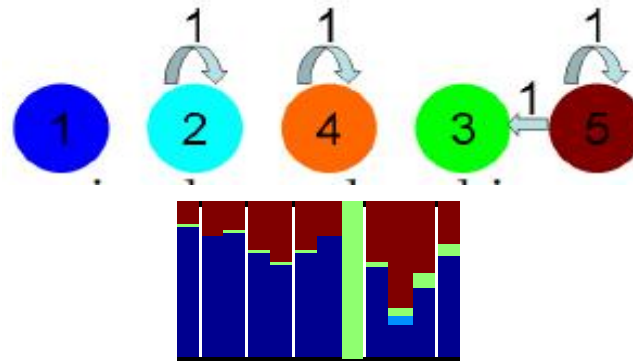
dMMSBとdIRMの違い

8月の
Enron founder (chairman)

2001/08



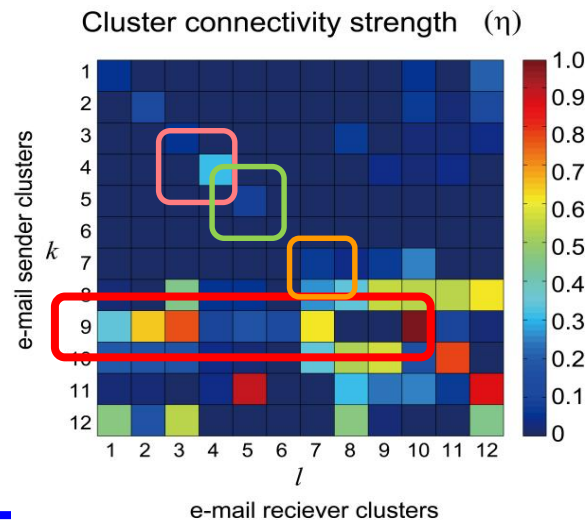
dMMSB



・茶色クラスタ(roll)からの受信に着目

・クラスタ(roll)と実世界の関係については特に言及されず

dIRM



・管理職クラスタなどへの送信に注目

・実際のコミュニティに対応するような社員クラスタが得られた

おわりに

- 時系列関係データのクラスタリングについてお話しさせていただきました
- 2つの両極端な研究例を詳しくご紹介しました。得られる結果も異なってくることを説明しました
- 色々な手法がありますが、目的に応じて使い分けていただくのが良いと考えます

参考文献

- Kwak et al., “What is Twitter, a Social Network or a News Media?”, Proc. WWW, 2010.
- Kwak et al., “Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter”, Proc. CHI, 2011.
- Ahmed ana Xing, “Recovering time-varying networks of dependencies in social and biological studies”, PNAS, 2009.
- Fu et al., “Dynamic Mixed Membership Blockmodel for Evolving Networks”, Proc. ICML, 2009.
- Tang et al., “Community Evolution in Dynamic Multi-mode Networks”, Proc. SIGKDD, 2008.
- Airoldi et al., “Mixed-membership stochastic blockmodels”, JMLR, 2008.

参考文献

- Ishiguro et al., “Dynamic Infinite Relational Model for Time-varying Relational Data Analysis”, Proc. NIPS, 2010.
- Yang et al., “A Bayesian Approach toward Finding Communities and their Evolutions in Dynamic Social Networks”, Proc. SDM, 2009.
- Kim and Han, “A Particle-and-density Based Evolutionary Clustering Method for Dynamic Networks”, Proc. VLDB, 2009.
- B. Klimat and Y. Yang. “The enron corpus: A new dataset for email classification research”, Proc ECML, 2004.
- Fox et al., “An HDP-HMM for Systems with State Persistence ”, Proc. ICML, 2008.

参考文献

- Myers and Leskovec, “On the Convexity of Latent Social Network Inference”, Proc. NIPS, 2010.
- Guo et al., “Recovering Temporally Rewiring Networks: a Model-based Approach”, Proc. ICML, 2007.
- Tang et al., “Small-world behavior in time-varying graphs”, Phy. Rev. E, 2010.