NTT

トピックモデルの応用：関係データ、ネットワークデータ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- 文書や著者の間に「関係」のネットワーク（グラフ）が想定されるデータセットが対象です

- お互いの関係をどのようにモデルに取り入れるかがポイントです

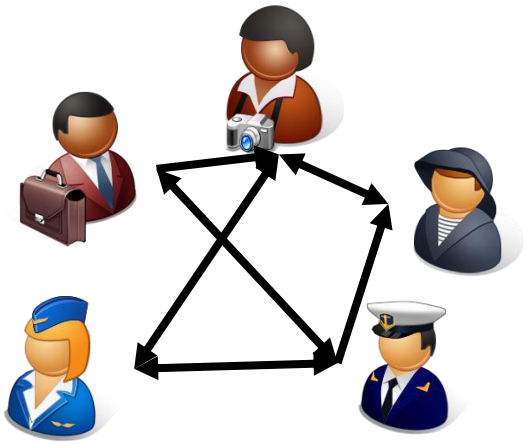
関係データ

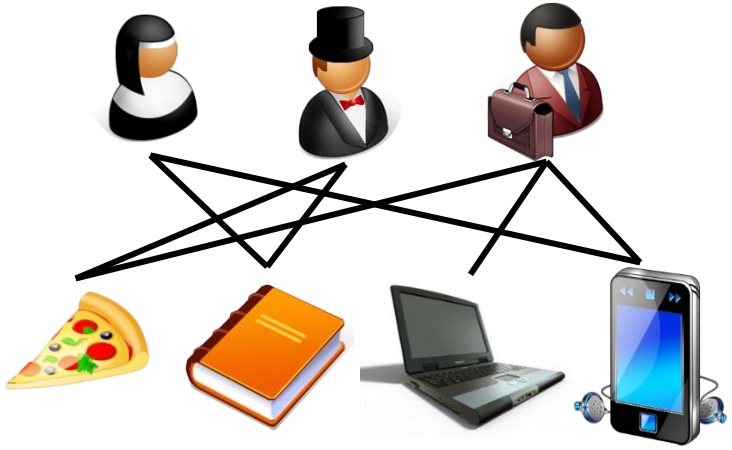
- 複数のオブジェクト（ノード）の間にリンク(エッジ）があってつながっているデータです
- 数学的には、いわゆる「グラフ」です

$$G = (V, E)$$

V(vertex): オブジェクト、ノード

E(edge): リンク、エッジ

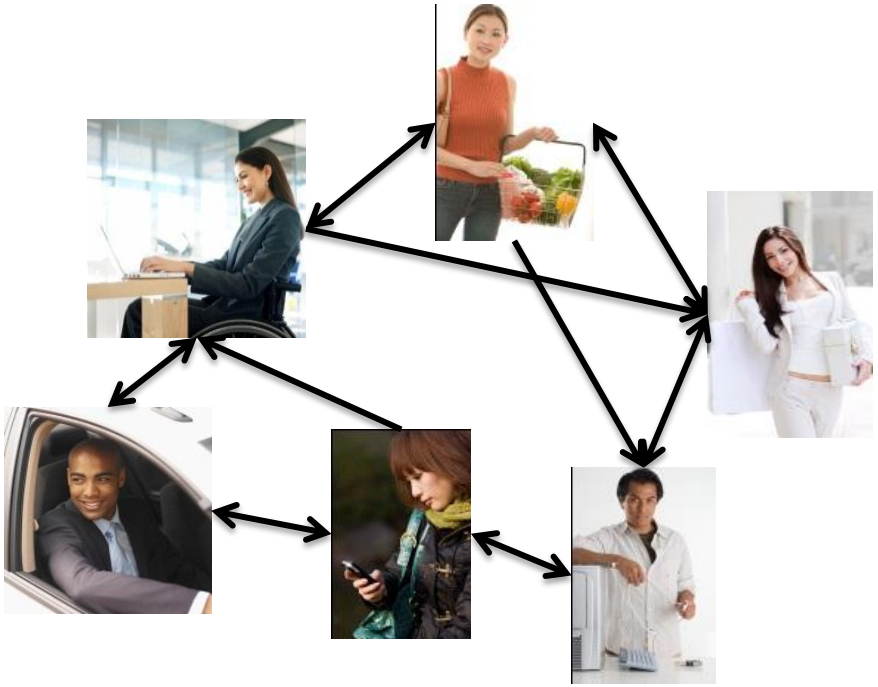




どんな関係データがありますか？

- ソーシャルネットワークサービス(SNS)上の友達関係、フォロー関係

$$G = (V, E) = (\text{ユーザ}, \text{フォロー})$$



SNS内のコミュニティ発見

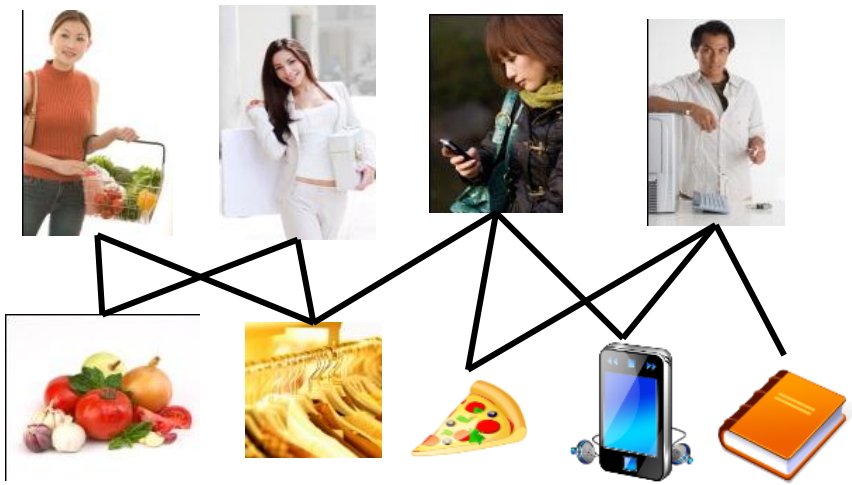
影響力の大きなユーザの発見

口コミ情報の伝搬範囲の最大化

どんな関係データがありますか？

- ネットショッピングなどの購買データ

$$G = (V, E) = (\text{ユーザ} \times \text{商品}, \text{販売実績})$$



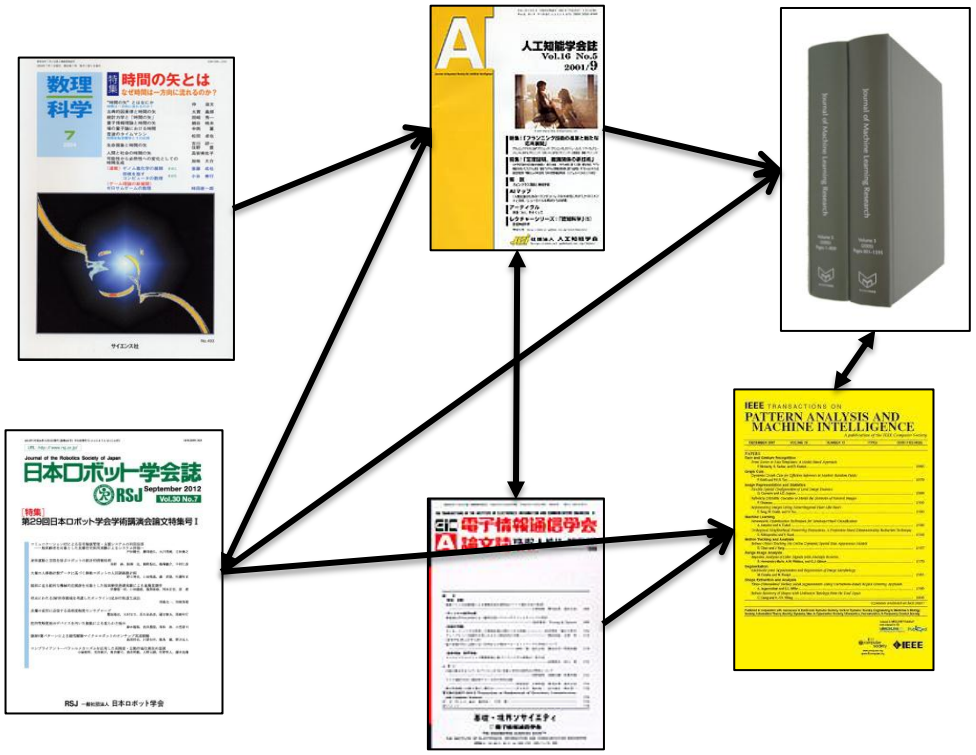
（販売実績に基づく）顧客のセグメント解析

商品のレコメンデーション（協調フィルタリング）

どんな関係データがありますか？

- 特許・技術論文の引用関係

$$G = (V, E) = (\text{特許・論文}, \text{引用・参照})$$



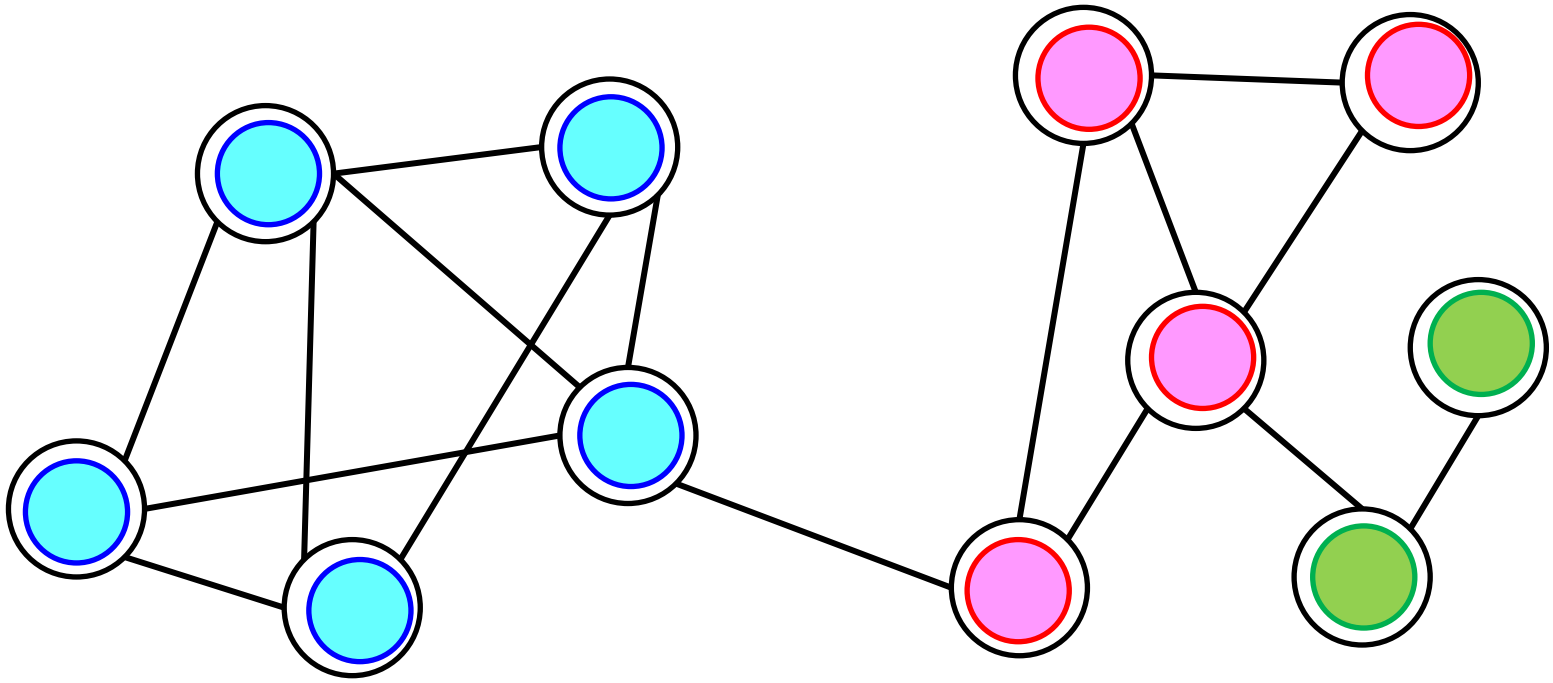
関連特許の検索

関連論文のおすすめ

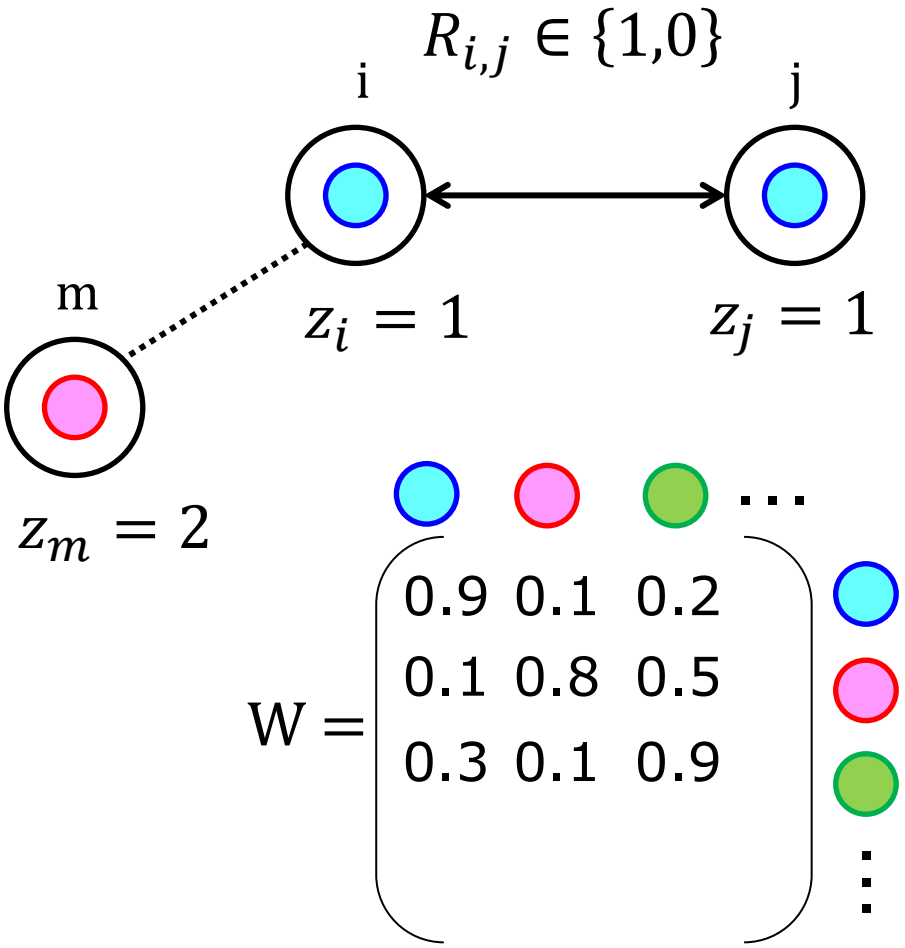
論文・特許の影響力可視化

関係データモデリング手法の例：無限関係モデル(IRM) [Kemp, 2006]

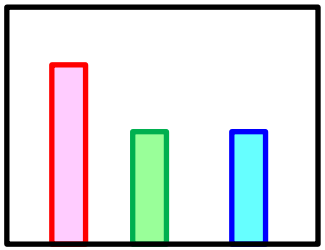
- シンプルで有効性の高い関係データモデル
- グラフのリンク構造から、オブジェクトをクラスタリング（カテゴライズ）してくれます



IRMの生成モデル



$\boldsymbol{\alpha} \sim \mathrm{Stick}(\gamma)$



$z_i = k \sim \mathrm{Mult}(\boldsymbol{\alpha})$

$w_{k,l} \sim \mathrm{Beta}(a, b)$

$R_{i,j} \sim \mathrm{Bernoulli}\left(w_{z_i,z_j}\right)$

→ 1　　⋯→ 0

IRMの問題点：
グラフの構造だけしか使わない

- 各オブジェクトは様々な情報・特徴をもっているはず → 使わない手はない

- ユーザの性別・年齢・プロフィール文
- 商品の値段・成分・キャッチコピー
- 論文（特許）の内容・請求項・キーワード

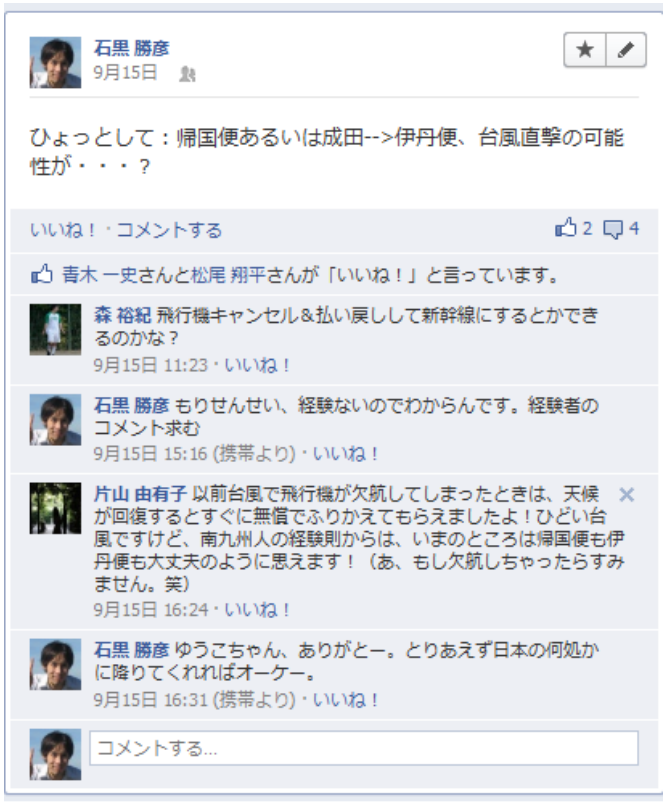
トピックモデル！！

Relational Topic Models
[Chang and Blei, 2009]

Chang and Blei,
"Relational Topic Models for Document Networks",
in Proc. AISTATS, 2009.

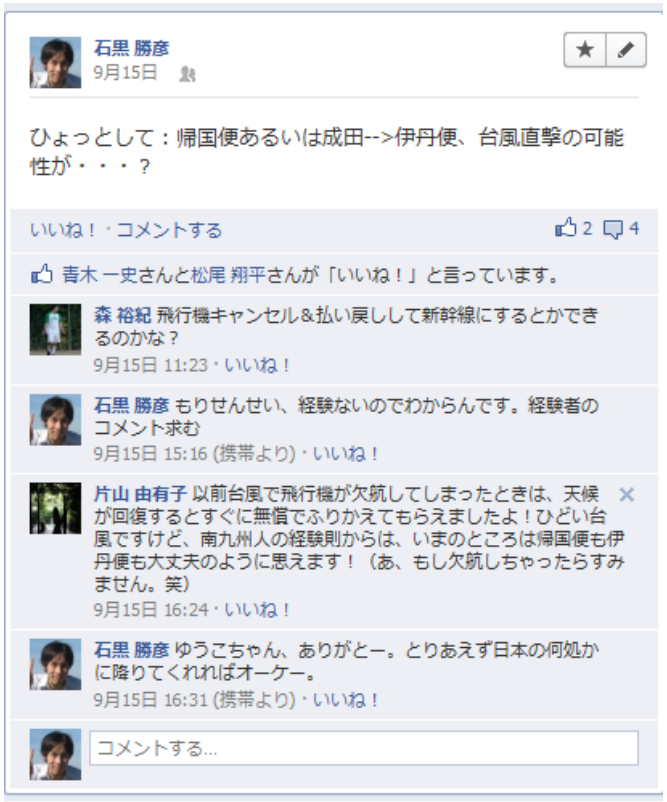
文書をリンクする情報は世の中沢山あります

- SNSでの返信、ブログの引用、特許の関連文献、論文のreference, …



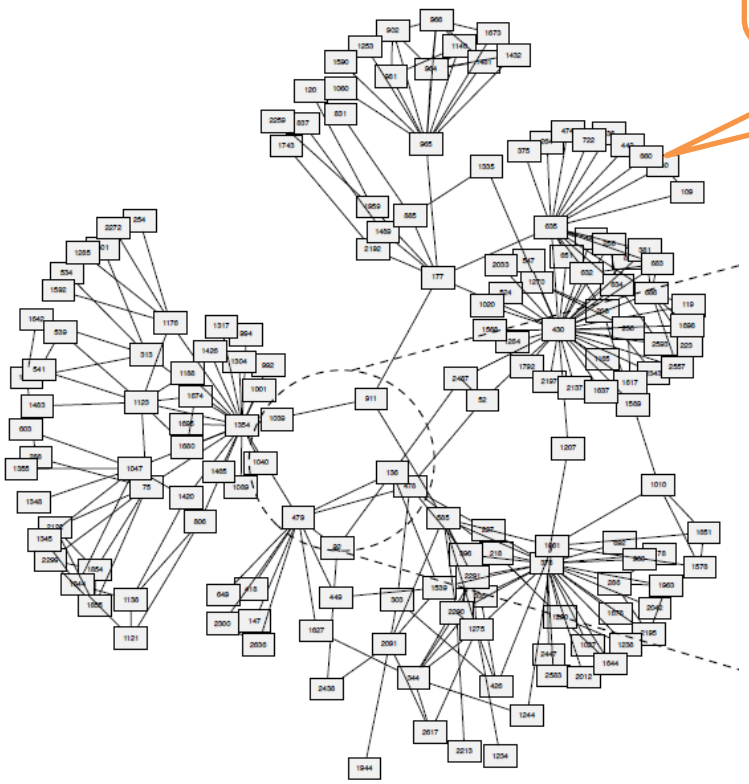
モデル化したくなります

- 関連する論文や、リツイートしたくなるようなつぶやきを自動的に発見できます



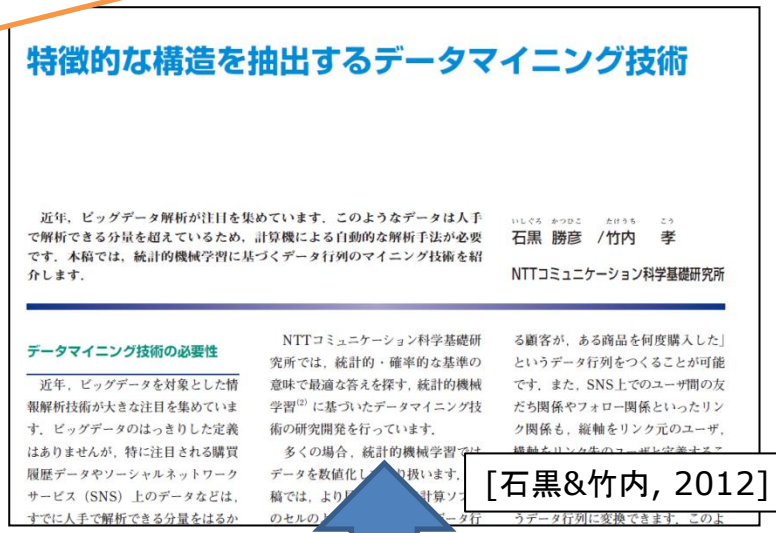
典型的なデータ構造のイメージ

論文引用ネットワーク



[Chang and Blei, 2009]

リンク＝引用した・された
オブジェクト＝文書（論文）：BoW表現



特徴的な構造を抽出するデータマイニング技術

近年，ビッグデータ解析が注目を集めています．このようなデータは人手で解析できる分量を超えているため，計算機による自動的な解析手法が必要です．本稿では，統計的機械学習に基づくデータ行列のマイニング技術を紹介します．

石黒 勝彦 /竹内 孝

NTTコミュニケーション科学基礎研究所

データマイニング技術の必要性

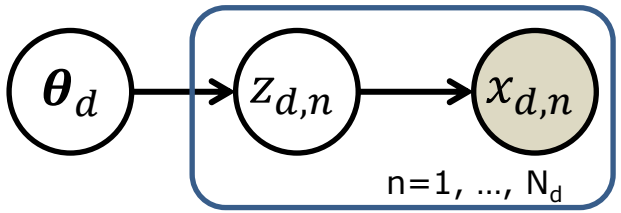
近年，ビッグデータを対象とした情報解析技術が大きな注目を集めています．ビッグデータのはっきりした定義はありませんが，特に注目される購買履歴データやソーシャルネットワークサービス（SNS）上のデータなどは，すでに人手で解析できる分量をはるか

NTTコミュニケーション科学基礎研究所では，統計的・確率的な基準の意味で最適な答えを探す，統計的機械学習[2]に基づいたデータマイニング技術の研究開発を行っています．

多くの場合，統計的機械学習では

る顧客が，ある商品を何度購入した」というデータ行列をつくることが可能です．また，SNS上でのユーザ間の友だち関係やフォロー関係といったリンク関係も，縦軸をリンク元のユーザ，

[石黒&竹内, 2012]



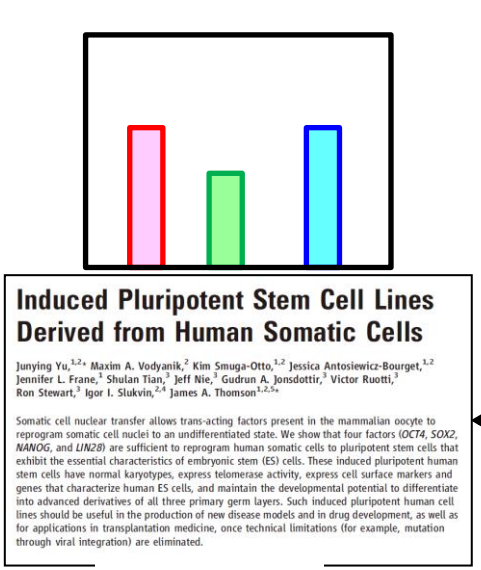
提案法: Relational Topic Model (RTM)

- 「リンク」を活かしたトピックモデル
  - 文書の中身だけでなく、文書間のリンクの生成過程も同時に確率モデル化
  - 具体的には論文や特許データを想定

- 文書のリンク推定: 論文の内容(BoW)から、関連がある論文を発見
- 文書のトピック推定: 特許の引用情報から、自分の特許とのバッティング度合を推定

手法のアイデア

- 内容（トピック）が似ている→引用（リンク）が発生する
- 文書のもつトピック分布の類似度に応じて、文書の間のリンク発生確率が変わる

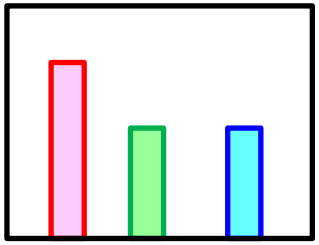


Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

Junying Yu,[1,2,*] Maxim A. Vodyanik,[2] Kim Smuga-Otto,[1,2] Jessica Antosiewicz-Bourget,[1,2] Jennifer L. Frane,[1] Shulan Tian,[3] Jeff Nie,[3] Gudrun A. Jonsdottir,[3] Victor Ruotti,[3] Ron Stewart,[3] Igor I. Slukvin,[2,3] James A. Thomson[1,2,3,*]

Somatic cell nuclear transfer allows trans-acting factors present in the mammalian oocyte to reprogram somatic cell nuclei to an undifferentiated state. We show that four factors (*OCT4, SOX2, NANOG,* and *LIN28*) are sufficient to reprogram human somatic cells to pluripotent stem cells that exhibit the essential characteristics of embryonic stem (ES) cells. These induced pluripotent human stem cells have normal karyotypes, express telomerase activity, express cell surface markers and genes that characterize human ES cells, and maintain the developmental potential to differentiate into advanced derivatives of all three primary germ layers. Such induced pluripotent human cell lines should be useful in the production of new disease models and in drug development, as well as for applications in transplantation medicine, once technical limitations (for example, mutation through viral integration) are eliminated.

[Yu, 2007]



Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors

Kazutoshi Takahashi[1] and Shinya Yamanaka[1,2,*]

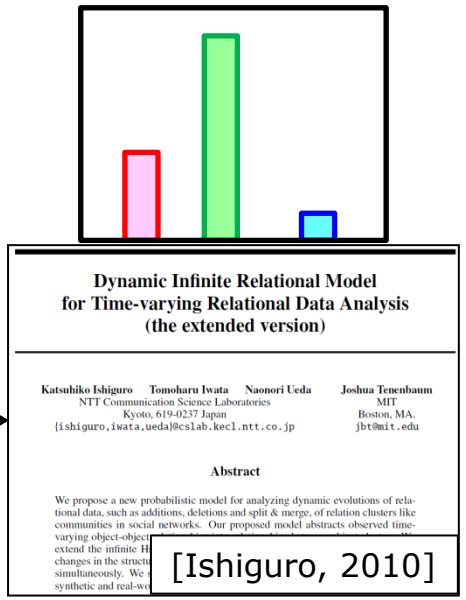
[1] Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
[2] CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan
*Contact: yamanaka@frontier.kyoto-u.ac.jp
DOI 10.1016/j.cell.2006.07.024

SUMMARY

Differentiated cells can be reprogrammed to an embryonic-like state by transfer of nuclear contents into oocytes or by fusion with embryonic stem (ES) cells. Little is known about factors

or by fusion with ES cells (Cowan et al., 2005; Tada et al., 2001), indicating that unfertilized eggs and ES cells contain factors that can confer totipotency or pluripotency to somatic cells. We hypothesized that the factors that play important roles in the maintenance of ES cell identity also play pivotal roles in the induction of pluripotency in

[Takahashi & Yamanaka, 2006]



Dynamic Infinite Relational Model for Time-varying Relational Data Analysis (the extended version)

Katsuhiko Ishiguro Tomoharu Iwata Naonori Ueda
NTT Communication Science Laboratories
Kyoto, 619-0237 Japan
{ishiguro,iwata,ueda}@cslab.kecl.ntt.co.jp

Joshua Tenenbaum
MIT
Boston, MA.
jbt@mit.edu

Abstract

We propose a new probabilistic model for analyzing dynamic evolutions of relational data, such as additions, deletions and split & merge, of relation clusters like communities in social networks. Our proposed model abstracts observed time-varying object-object

extend the infinite H

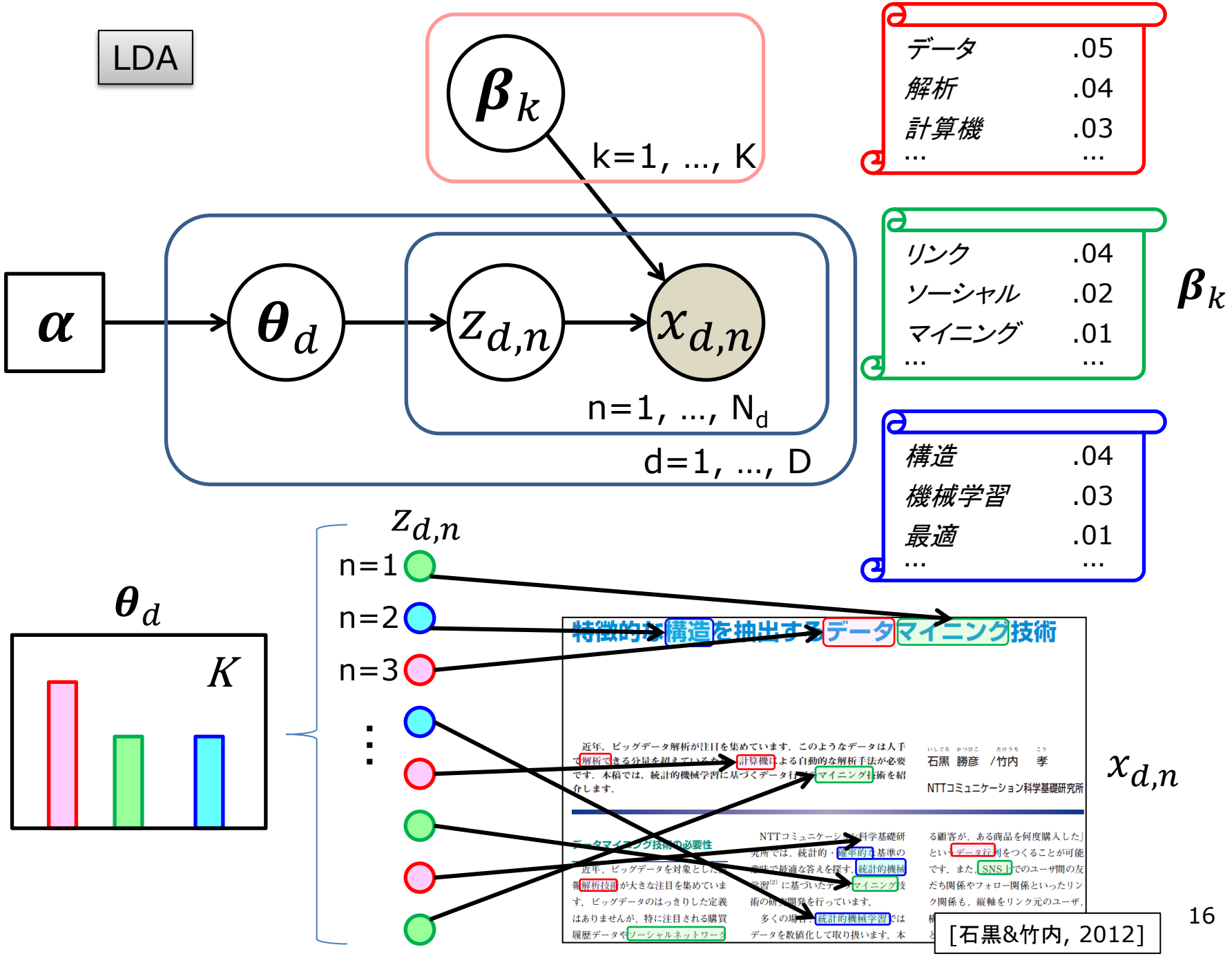
changes in the structu

simultaneously. We

synthetic and real-wo

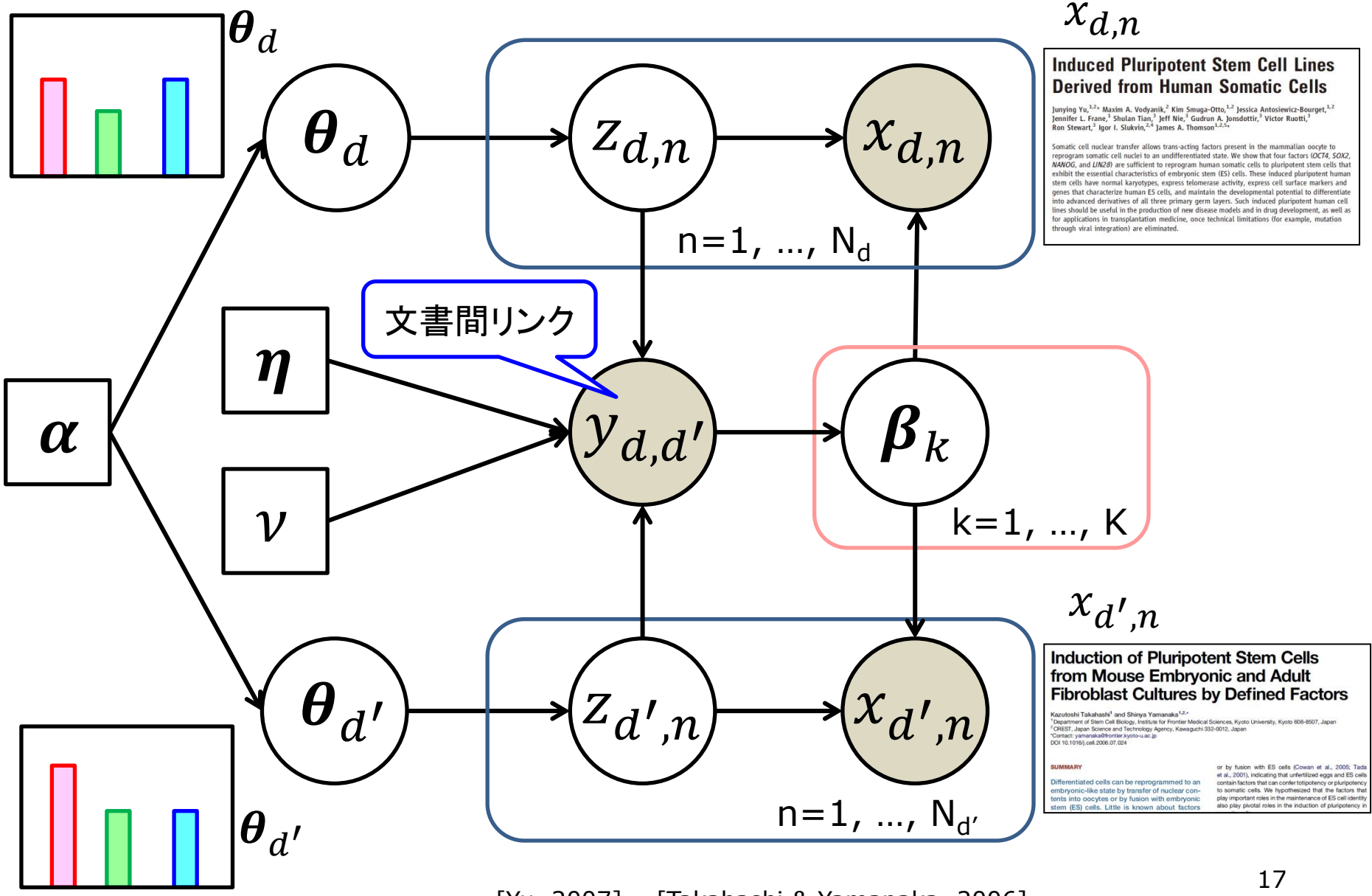
[Ishiguro, 2010]

LDA



[石黒&竹内, 2012]

Relational Topic Model（d, d'に関するプレートは省略）



$x_{d,n}$

Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

Junying Yu,[1,2,*] Maxim A. Vodyanik,[2] Kim Smuga-Otto,[1,2] Jessica Antosiewicz-Bourget,[1,2] Jennifer L. Frane,[1] Shulan Tian,[3] Jeff Nie,[3] Gudrun A. Jonsdottir,[3] Victor Ruotti,[3] Ron Stewart,[3] Igor I. Slukvin,[2,4] James A. Thomson[1,2,5,*]

Somatic cell nuclear transfer allows trans-acting factors present in the mammalian oocyte to reprogram somatic cell nuclei to an undifferentiated state. We show that four factors (*OCT4*, *SOX2*, *NANOG*, and *LIN28*) are sufficient to reprogram human somatic cells to pluripotent stem cells that exhibit the essential characteristics of embryonic stem (ES) cells. These induced pluripotent human stem cells have normal karyotypes, express telomerase activity, express cell surface markers and genes that characterize human ES cells, and maintain the developmental potential to differentiate into advanced derivatives of all three primary germ layers. Such induced pluripotent human cell lines should be useful in the production of new disease models and in drug development, as well as for applications in transplantation medicine, once technical limitations (for example, mutation through viral integration) are eliminated.

$x_{d',n}$

Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors

Kazutoshi Takahashi[1] and Shinya Yamanaka[1,2,*]
[1]Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
[2]CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan
*Contact: yamanaka@frontier.kyoto-u.ac.jp
DOI 10.1016/j.cell.2006.07.024

SUMMARY

Differentiated cells can be reprogrammed to an embryonic-like state by transfer of nuclear contents into oocytes or by fusion with embryonic stem (ES) cells. Little is known about factors

or by fusion with ES cells (Cowan et al., 2005; Tada et al., 2001), indicating that unfertilized eggs and ES cells contain factors that can confer totipotency or pluripotency to somatic cells. We hypothesized that the factors that play important roles in the maintenance of ES cell identity also play pivotal roles in the induction of pluripotency in

[Yu, 2007]　[Takahashi & Yamanaka, 2006]

生成モデル

for 文書d = 1, 2, …, D

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment $z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$

word observation $x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

for 文書ペア d = 1, 2, …, D, d′ = 1, 2, …, D

doc-doc link observation

$$y_{d,d'}|\boldsymbol{z}_d,\boldsymbol{z}_{d'},\boldsymbol{\eta},\nu\sim\mathrm{Bernoulli}\left(\psi\left(y_{d,d'}|\boldsymbol{z}_d,\boldsymbol{z}_{d'},\boldsymbol{\eta},\nu\right)\right)$$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

文書-文書リンクの接続確率

- 各文書のトピックヒストグラム（の平均）を使う
  → 内容の要約情報を計算

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d}\mathbf{z}_{d,n}$$

$z_{d,n}$をK次元ベクトルとして見ています

シグモイドモデル $\psi\left(y_{d,d'}|\mathbf{z}_d,\mathbf{z}_{d'},\boldsymbol{\eta},\nu\right)=\sigma(\boldsymbol{\eta}^T(\bar{\mathbf{z}}_d\circ\bar{\mathbf{z}}_{d'})+\nu)$

指数モデル $\psi\left(y_{d,d'}|\mathbf{z}_d,\mathbf{z}_{d'},\boldsymbol{\eta},\nu\right)=\exp(\boldsymbol{\eta}^T(\bar{\mathbf{z}}_d\circ\bar{\mathbf{z}}_{d'})+\nu)$

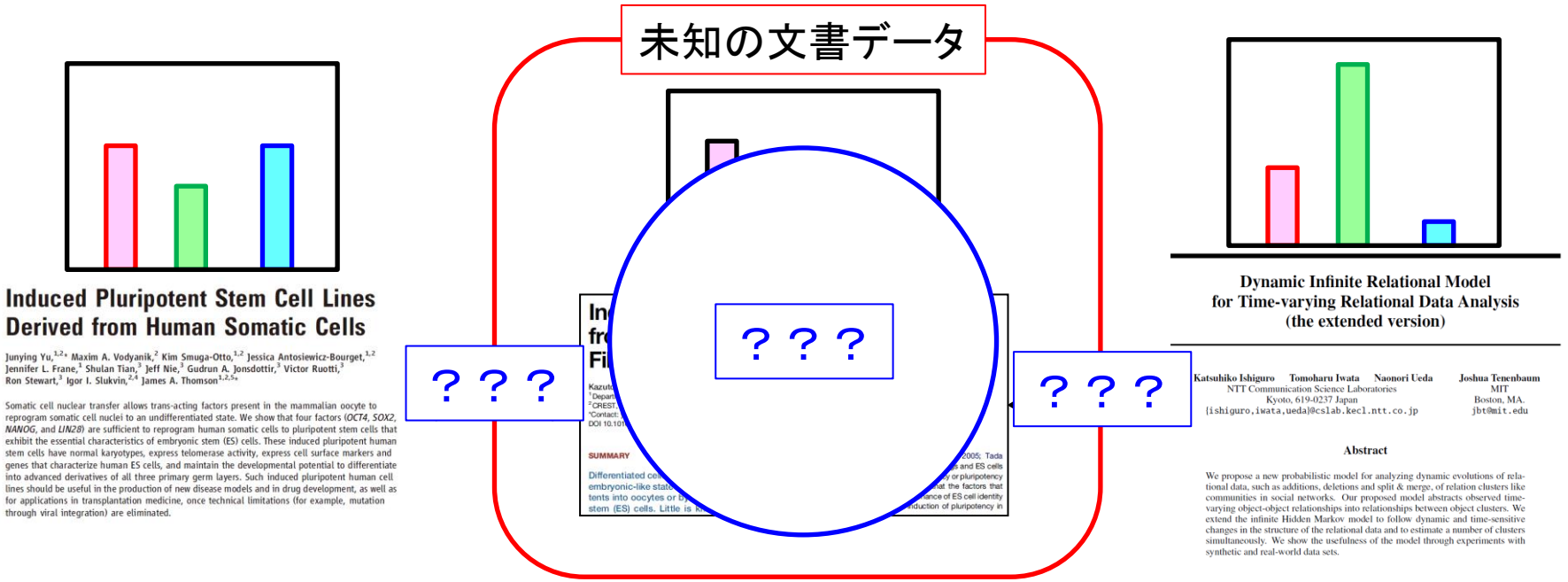
隠れ変数、パラメータの推定

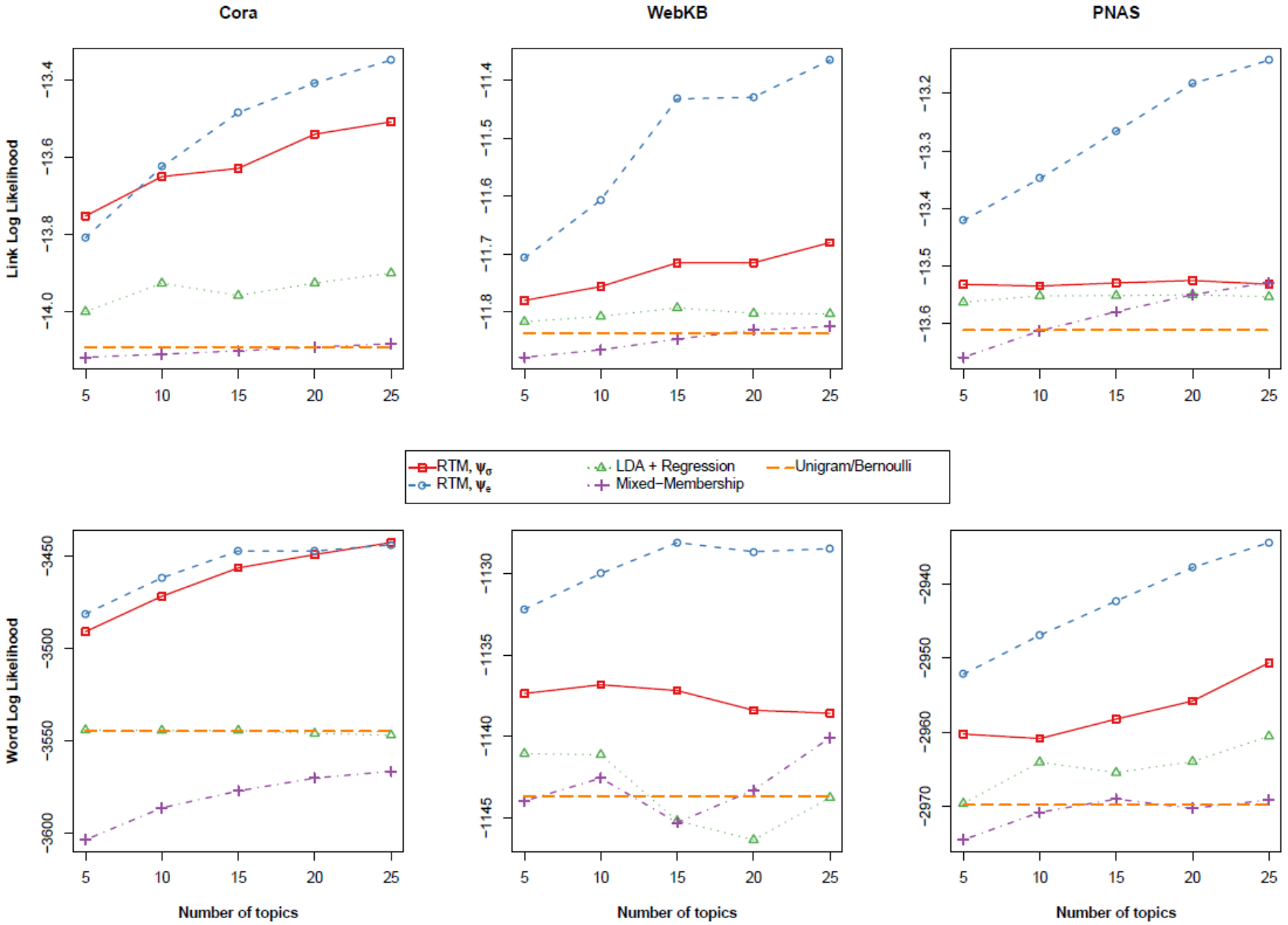
- 論文中では変分ベイズ(VB)による解法が導出されています

- 詳細はひとまず割愛します・・・

予測

- 学習が完了した提案モデルは、2種類の予測タスクに利用できます
  - リンク予測タスク
  - 内容（トピック）予測タスク





[Chang and Blei, 2009]

赤、青：提案法（詳細が少し違う）緑：トピックモデル→リンク予測
紫：文書情報を無視　オレンジ：文書情報と関係情報を別々にモデル化

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo Rates of convergence of the Hastings and Metropolis algorithms **Possible biases induced by MCMC convergence diagnostics** Bounding convergence time of the Gibbs sampler in Bayesian image restoration Self regenerative Markov chain Monte Carlo Auxiliary variable methods for Markov chain Monte Carlo with applications **Rate of Convergence of the Gibbs Sampler by Gaussian Approximation** Diagnosing convergence of Markov chain Monte Carlo algorithms	**RTM (ψ_e)**
Exact Bound for the Convergence of Metropolis Chains Self regenerative Markov chain Monte Carlo **Minorization conditions and convergence rates for Markov chain Monte Carlo** Gibbs-markov models Auxiliary variable methods for Markov chain Monte Carlo with applications Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models Mediating instrumental variables A qualitative framework for probabilistic inference Adaptation for Self Regenerative MCMC	**LDA + Regression**

[Chang and Blei, 2009]

Relational Topic Model: まとめ

- 文書と文書の間にリンクがあるデータセットのモデル化
- 文書のトピックが似ているとリンクが張られやすくなるようにモデルを立てている
- リンク予測や内容予測、お勧め論文など

他の関係データトピックモデル

- Liu et al., “Topic-link LDA: Joint models of topic and author community”, in Proc. ICML, 2009.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Kemp, 2006] Kemp et al., “Learning Systems of Concepts with an Infinite Relational Model”, in Proc. AAAI, 2006.
- [Chang and Blei, 2009] Chang and Blei, “Relational Topic Models for Document Networks”, in Proc. AISTATS, 2009.
- [Takahashi & Yamanaka, 2006] Takahashi and Yamanaka, “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors”, Cell, Vol. 126, pp. 663-676, 2006.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [Ishiguro, 2010] Ishiguro et al, “Dynamic Infinite Relational Model for Time-varying Relational Data Analysis”, in Proc. NIPS, 2010.

引用及び参考文献

- [Yu, 2007] Yu et al., “Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells”, Science, Vol. 318, pp. 1917-1920, 2007.