

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

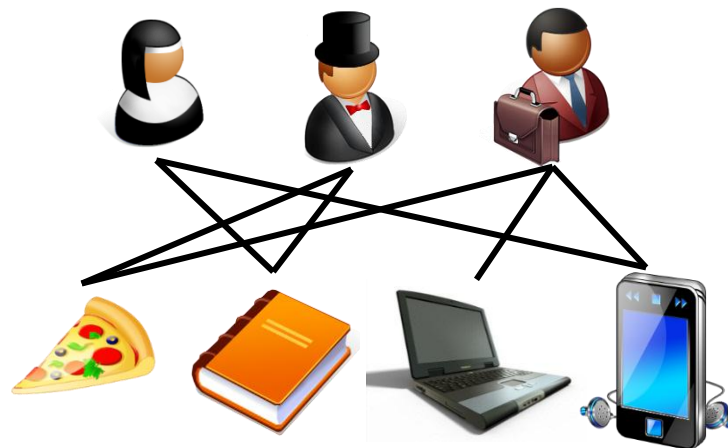
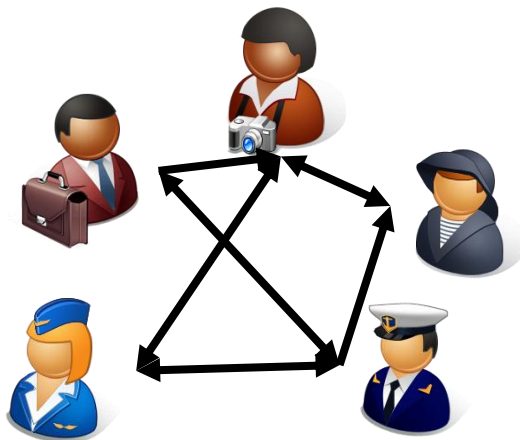
関係データ

- 複数のオブジェクト(ノード)の間にリンク(エッジ)があつ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

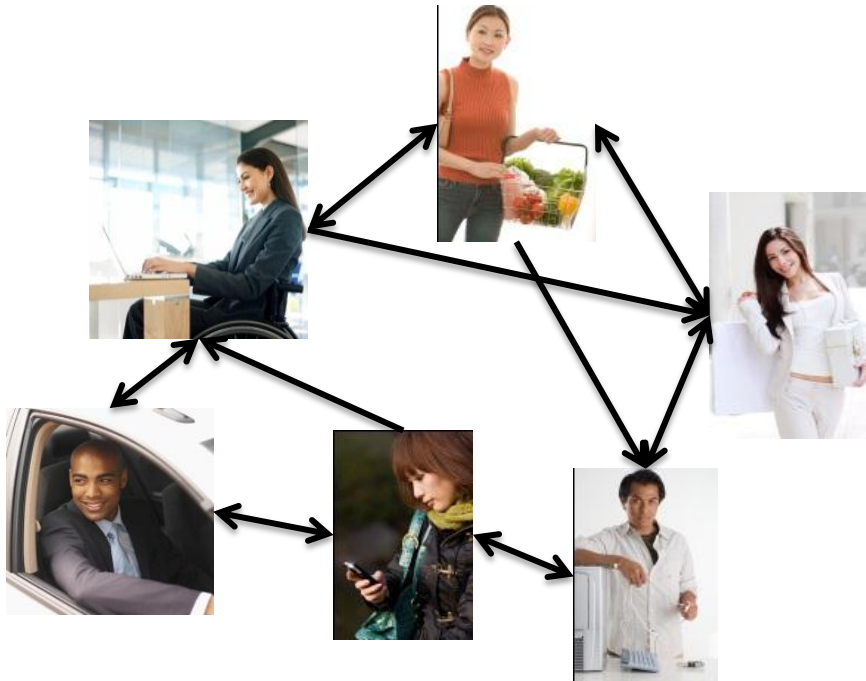
E(edge): リンク、エッジ



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

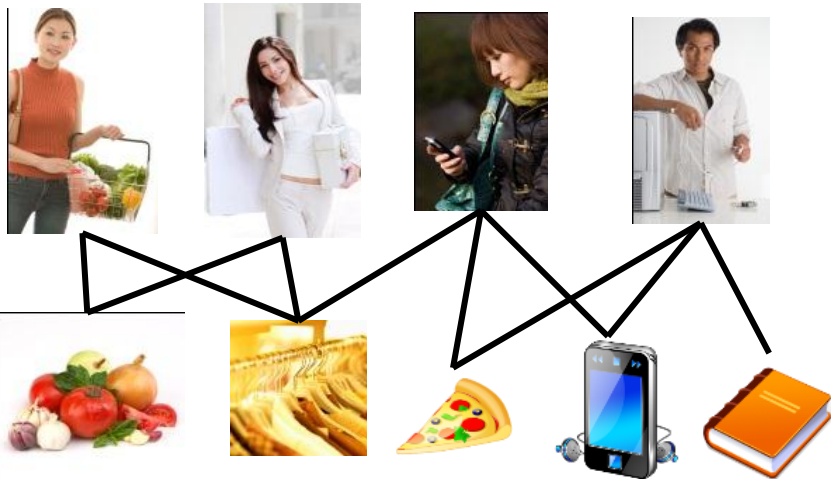
影響力の大きなユーザの発見

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

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



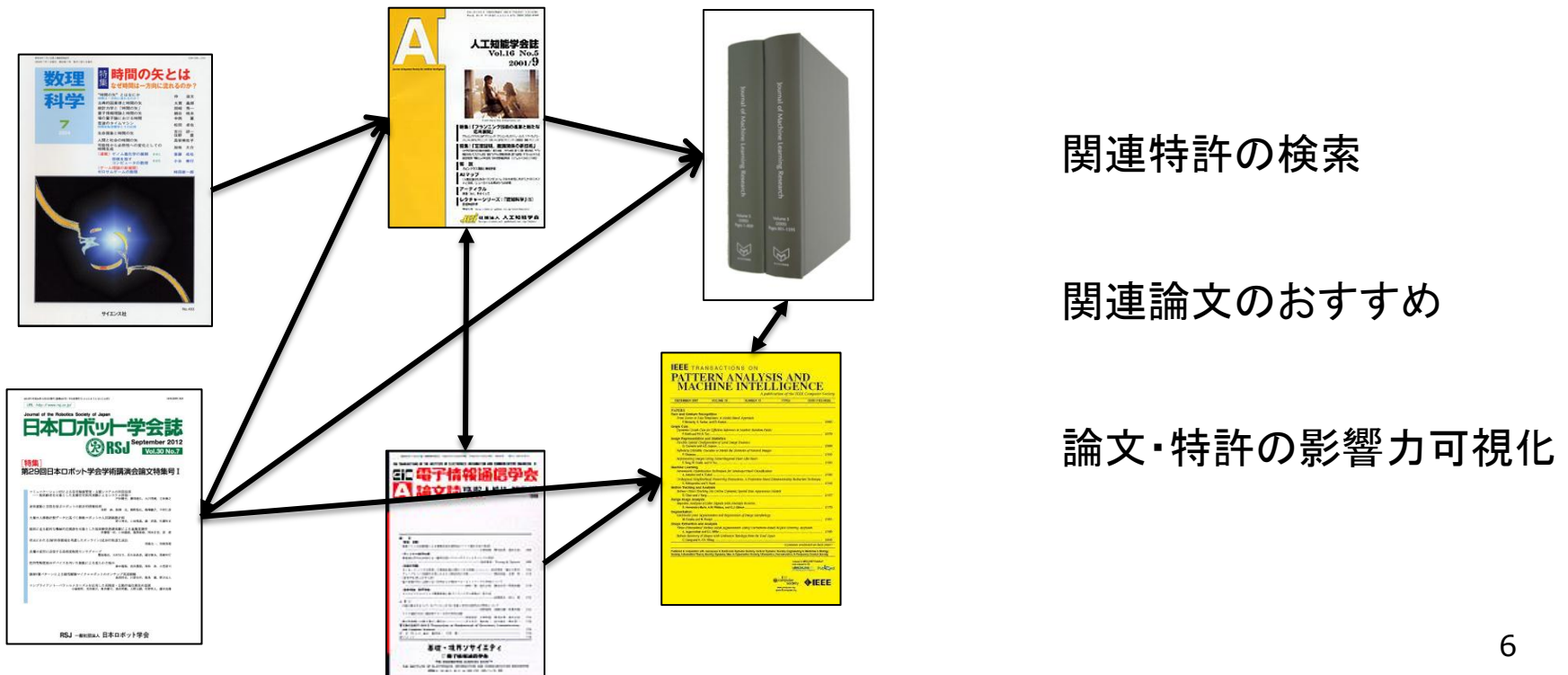
(販売実績に基づく)顧客のセグメント解析

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？

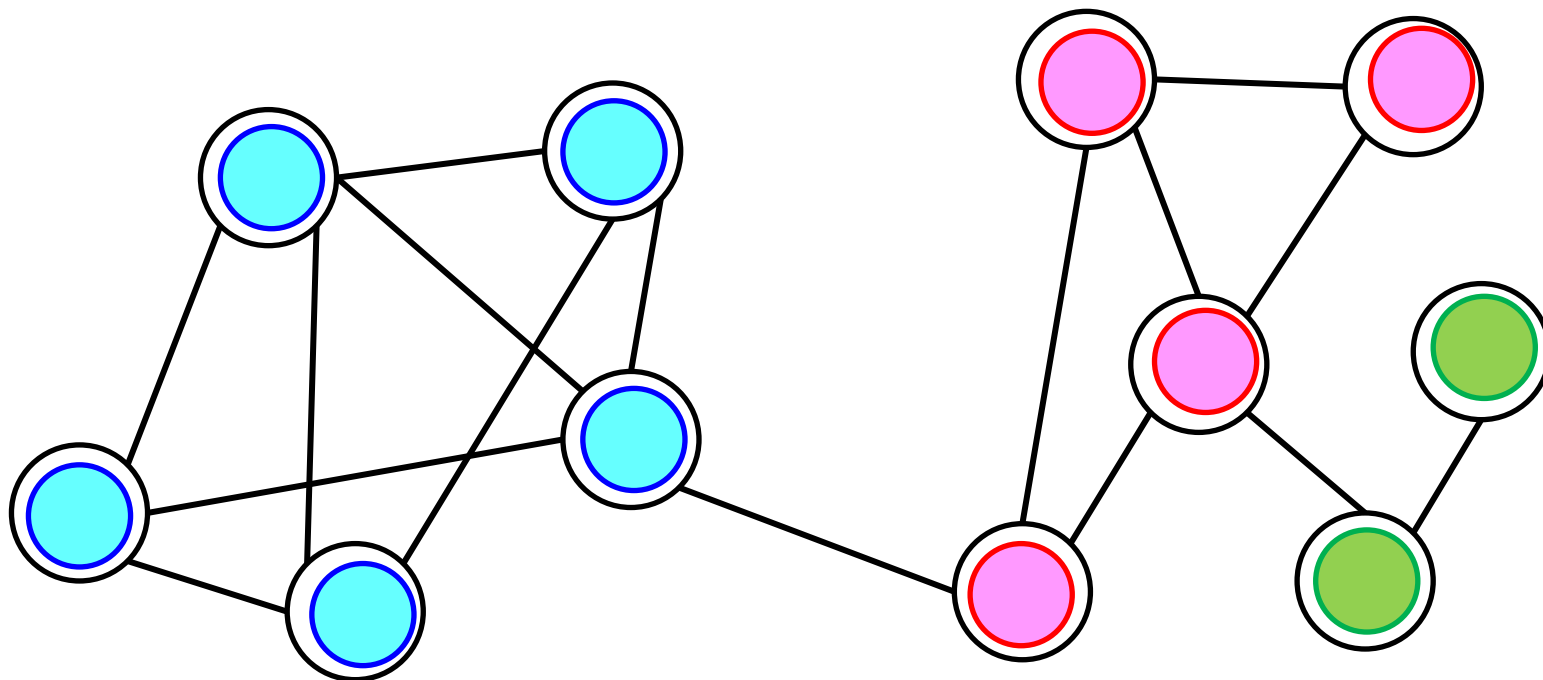
- 特許・技術論文の引用関係

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

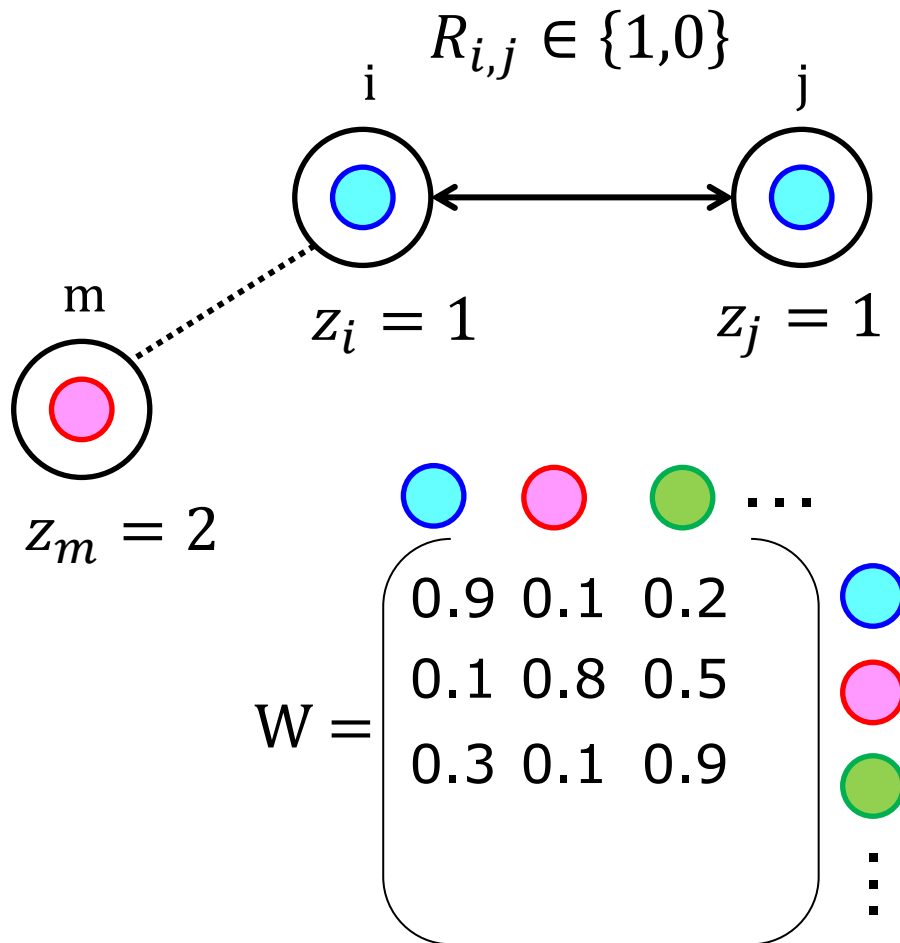


関係データモデリング手法の例： 無限関係モデル(IRM) [Kemp, 2006]

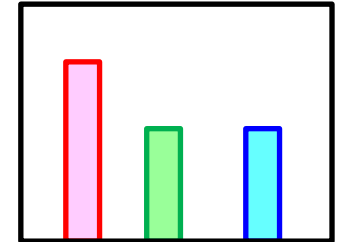
- シンプルで有効性の高い関係データモデル
- グラフのリンク構造から、オブジェクトをクラスタリング(カテゴライズ)してくれます



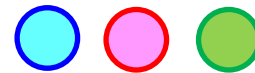
IRMの生成モデル



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



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



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

$$R_{i,j} \sim \text{Bernoulli}(w_{z_i, z_j})$$

$\longrightarrow 1$ $\cdots \cdots \cdots \longrightarrow 0$

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, ...

石黒 勝彦
9月15日

ひょっとして：帰国便あるいは成田-->伊丹便、台風直撃の可能性が・・・？

いいね！・コメントする 2 4

👍 青木 一史さんと松尾 翔平さんが「いいね！」とっています。

森 裕紀 飛行機キャンセル&払い戻して新幹線にするとかできるのかな？
9月15日 11:23 · いいね！

石黒 勝彦 もりせんせい、経験ないのでわかりません。経験者のコメント求む
9月15日 15:16 (携帯より) · いいね！

片山 由有子 以前台風で飛行機が欠航してしまったときは、天候が回復するとすぐに無償でふりかえてもらえましたよ！ひどい台風ですけど、南九州人の経験則からは、いまのところは帰国便も伊丹便も大丈夫のように思えます！（あ、もし欠航しちゃったらすみません。笑）
9月15日 16:24 · いいね！

石黒 勝彦 ゆうごちゃん、ありがとー。とりあえず日本の何処かに降りてくれればオーケー。
9月15日 16:31 (携帯より) · いいね！

コメントする...

REFERENCES

- Adhikary, S., and Eilers, M. (2005). Transcriptional regulation and transformation by Myc proteins. *Nat. Rev. Mol. Cell Biol.* 6, 635–645.
- Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* 17, 126–140.
- Baudino, T.A., McKay, C., Pendeville-Samain, H., Nilsson, J.A., Maclean, K.H., White, E.L., Davis, A.C., Ihle, J.N., and Cleveland, J.L. (2002). c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes Dev.* 16, 2530–2543.
- Boyer, L.A., Lee, T.J., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956.
- Bromberg, J.F., Wrzeszczynska, M.H., Devgan, G., Zhao, Y., Pestell, R.G., Albanese, C., and Darnell, J.E., Jr. (1999). Stat3 as an oncogene. *Cell* 98, 295–303.
- Burdon, T., Stracey, C., Chambers, I., Nichols, J., and Smith, A. (1999). Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. *Dev. Biol.* 210, 30–43.

- van Leeuwen, S., Taketo, M.M., Roberts, (2002). Apc modulates embryonic stem-cell self-renewal by regulating the dosage of beta-catenin signaling. *Development* 129, 1077–1084.
- Li, Y., McClintick, J., Zhong, L., Edenberg, R.J. (2005). Murine embryonic stem cell self-renewal is promoted by SOCS-3 and inhibited by the zinc finger protein Klf4. *Blood* 105, 635–637.
- Lin, T., Chao, C., Saito, S., Mazur, S.J., Miao, L., and Xu, Y. (2004). p53 induces differentiation of embryonic stem cells by suppressing Nanog expression. *Published online December 26, 2004.* 10.1002/stem.1000
- Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Tropea, G., George, J., Leong, B., Liu, J., Lau, J., Ng, K., et al. (2007). Oct4 and Nanog transcription network regulates pluripotency in human embryonic stem cells. *Nat. Genet.* 38, 431–440.
- Martin, G.R. (1981). Isolation of a pluripotent cell line from human embryos cultured in medium conditioned by teratocarcinoma. *Proc. Natl. Acad. Sci. USA* 78, 763–767.
- Maruyama, M., Ichisaka, T., Nakagawa, M. (2007). Efficient generation of human induced pluripotent stem cells by direct transcription of pluripotency-associated genes in somatic cells. *Nat. Methods* 4, 473–480.
- Matsuda, T., Nakamura, T., Nakao, K., and Yokota, T. (1999). STAT3 activation

[Takahashi & Yamanaka, 2006]

Cell 126, 663–676, August 25, 2006

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

The screenshot shows a Twitter thread. At the top, a user named 石黒 勝彦 (Ishikawa Katsuhiko) posted on 9月15日: "ひょっとして：帰国便あるいは成田-->伊丹便、台風直撃の可能性が・・・？". Below the tweet, there are interaction buttons for "いいね！" (2 likes) and "コメントする" (4 replies). A reply from 青木 一史 and 松尾 翔平 says "いいね！" and "と言っています". Another reply from 森 裕紀 asks "飛行機キャンセル&払い戻して新幹線にするとかできるのかな？" (9月15日 11:23). A reply from 石黒 勝彦 says "もりせんせい、経験ないのでわかりません。経験者のコメント求む" (9月15日 15:16). A reply from 片山 由有子 says "以前台風で飛行機が欠航してしまったときは、天候が回復するとすぐに無償でふりかえてもらえましたよ！ひどい台風ですけど、南九州人の経験則からは、いまのところは帰国便も伊丹便も大丈夫のように思えます！（あ、もし欠航しちゃったらすみません。笑）" (9月15日 16:24). A reply from 石黒 勝彦 says "ゆうごちゃん、ありがとー。とりあえず日本の何処かに降りてくれればオーケー。" (9月15日 16:31). At the bottom, there is a "コメントする..." input field.

REFERENCES

- Adhikary, S., and Eilers, M. (2005). Transcriptional regulation and transformation by Myc proteins. *Nat. Rev. Mol. Cell Biol.* 6, 635–645.
- Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* 17, 126–140.
- Baudino, T.A., McKay, C., Pendergill-Samain, H., Nilsson, J.A., Maclean, K.H., White, E.L., Davis, A.C., Ihle, J.N., and Cleveland, J.L. (2002). c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes Dev.* 16, 2530–2543.
- Boyer, L.A., Lee, T.J., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956.
- Bromberg, J.F., Wrzeszczynska, M.H., Devgan, G., Zhao, Y., Pestell, R.G., Albanese, C., and Darnell, J.E., Jr. (1999). Stat3 as an oncogene. *Cell* 98, 295–303.
- Burdon, T., Stracey, C., Chambers, I., Nichols, J., and Smith, A. (1999). Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. *Dev. Biol.* 210, 30–43.

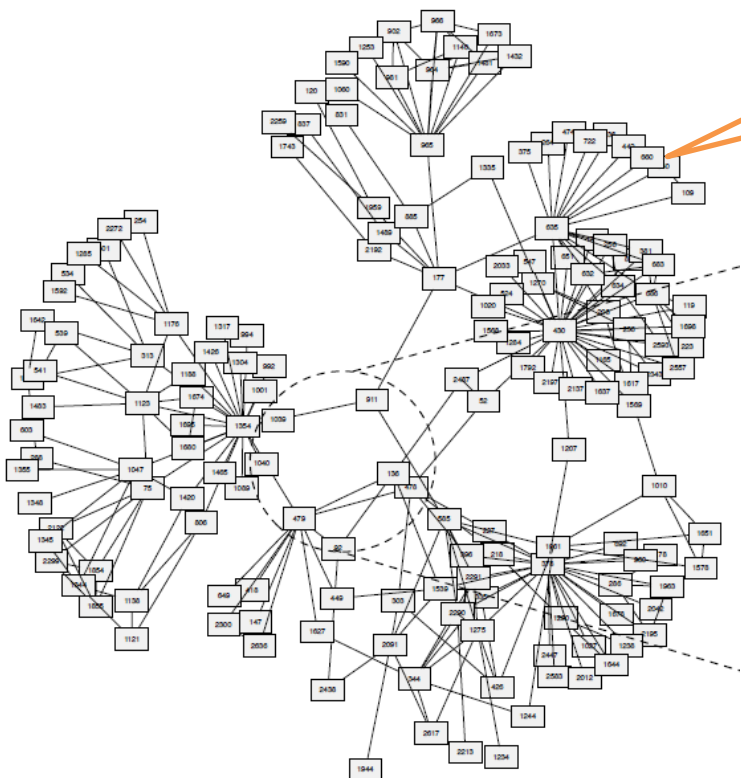
- van Leeuwen, S., Taketo, M.M., Roberts, (2002). Apc modulates embryonic stem-cell lineage by regulating the dosage of beta-catenin signaling.
- Li, Y., McClintick, J., Zhong, L., Edenberg, R.J. (2005). Murine embryonic stem cells are promoted by SOCS-3 and inhibited by the zinc finger protein Klf4. *Blood* 105, 635–637.
- Lin, T., Chao, C., Saito, S., Mazur, S.J., and Xu, Y. (2004). p53 induces differentiation of embryonic stem cells by suppressing Nanog expression. Published online December 26, 2004. *10*
- Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Treutlein, G., George, J., Leong, B., Liu, J., Zhou, J., Zhang, W., et al. (2006). Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* 38, 431–440.
- Martin, G.R. (1981). Isolation of a pluripotent mouse embryo cultured in medium conditioned by embryonic stem cells. *Proc. Natl. Acad. Sci. USA* 78, 763–767.
- Maruyama, M., Ichisaka, T., Nakagawa, M. (2007). Different roles for sox15 and sox2 in transcriptional regulation of embryonic stem cells. *J. Biol. Chem.* 282, 280–284.
- Matsuda, T., Nakamura, T., Nakao, K., and Yokota, T. (1999). STAT3 activation

[Takahashi & Yamanaka, 2006]

Cell 126, 663–676, August 25, 2006

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

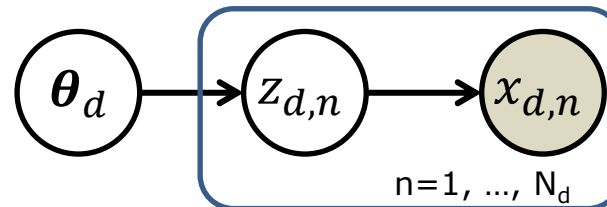
近年、ビッグデータを対象とした情報解析技術が大きな注目を集めています。ビッグデータのはっきりした定義はありませんが、特に注目される購買履歴データやソーシャルネットワークサービス (SNS) 上のデータなどは、すでに人手で解析できる分量をはるかに

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る、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データを数値化した上で扱います。本稿では、より人間に近い感覚でデータのセルの

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のが可能です。また、SNS上でのユーザー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ー、横軸をリンク先のユーザーと定義する

[石黒&竹内, 2012]

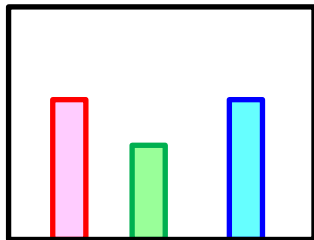


提案法: Relational Topic Model (RTM)

- 「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

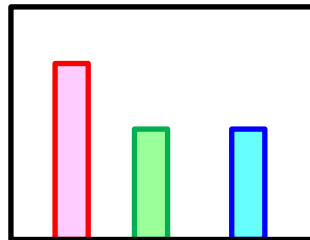


Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

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,² Igor I. Slukvin,^{2,4} James A. Thomson^{1,2,5,6}

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 *LIN28B*) 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

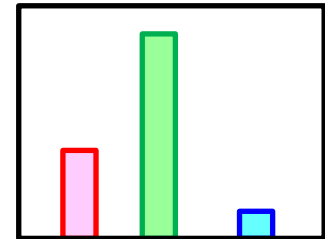
Kazutoshi Takahashi¹ and Shinya Yamanaka^{1,2,*}

¹Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
²CREST, 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., 2007), 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)

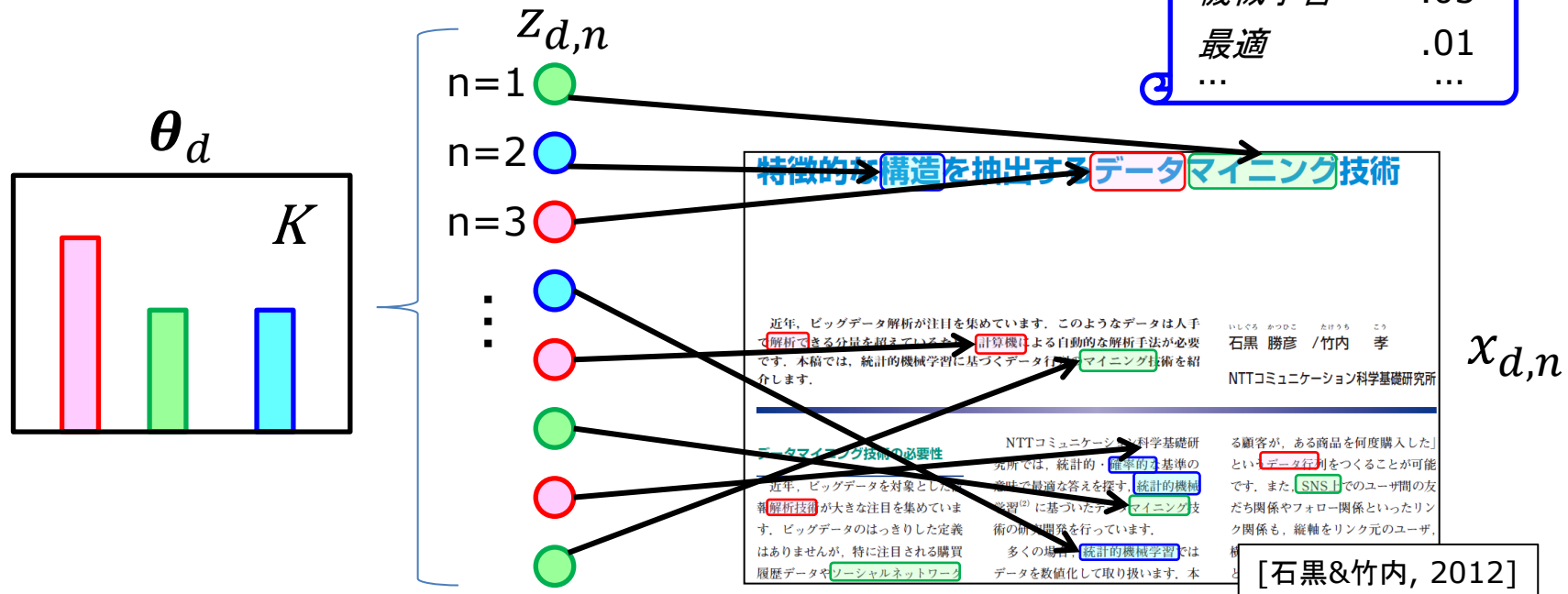
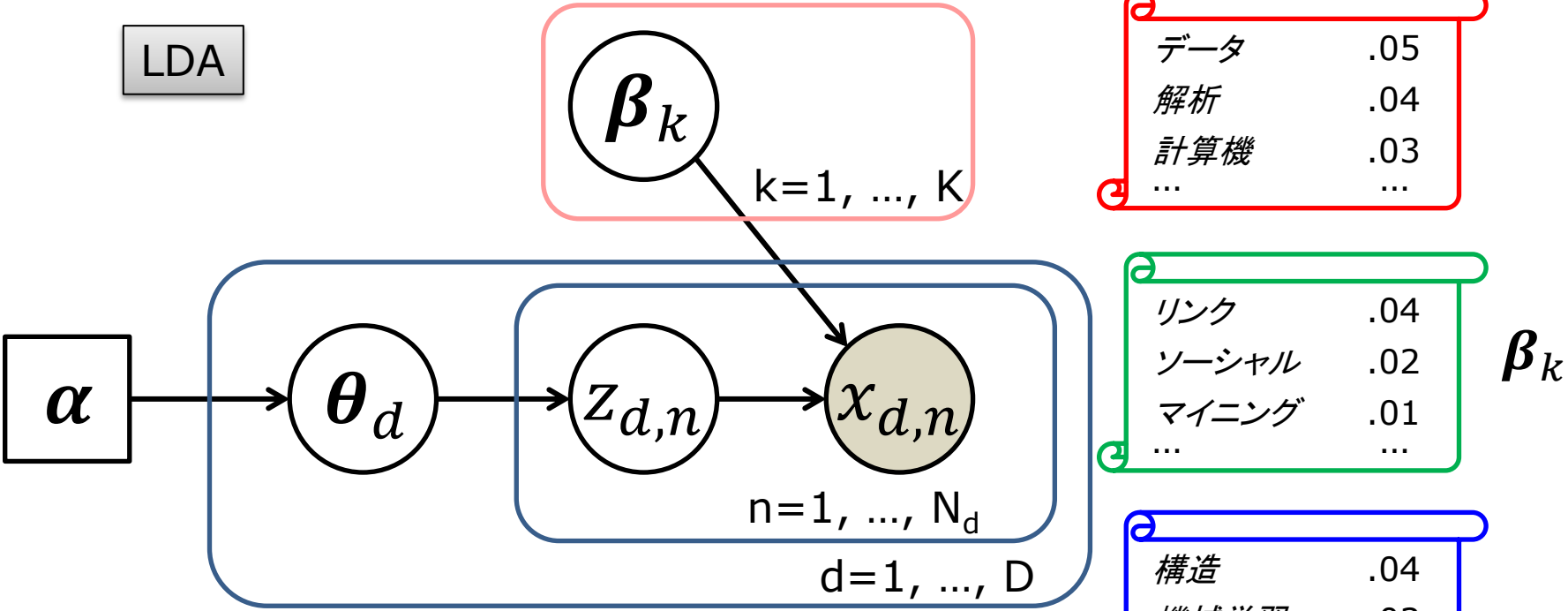
Katsuhiko Ishiguro Tomoharu Iwata Naomori Ueda Joshua Tenenbaum
 NTT Communication Science Laboratories MIT
 Kyoto, 619-0237 Japan Boston, MA, USA
 {ishiguro,iwata,ueda}@cs.lab.kecl.ntt.co.jp jbt@mit.edu

Abstract

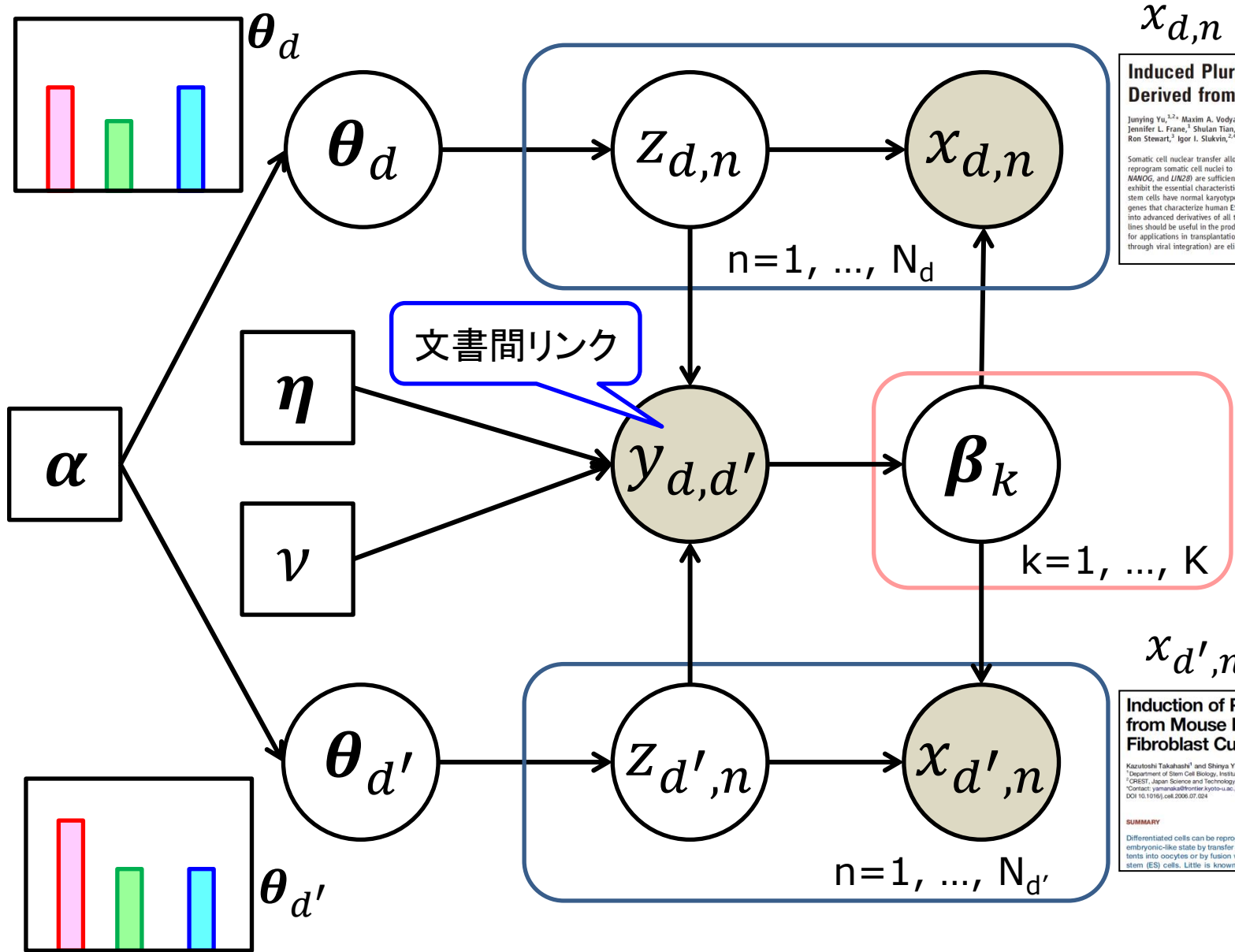
We propose a new probabilistic model for analyzing dynamic evolutions of relational data, such as additions, deletions and split & merge, of relation clusters like communities in social networks. Our proposed model abstracts observed time-varying object-object relations into a dynamic infinite HMM. We extend the infinite HMM to handle changes in the structure simultaneously. We show that our model can capture synthetic and real-world

[Ishiguro, 2010]

LDA



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



Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

Junyong 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,² Igor I. Slukvin,^{1,4} 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.

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

Kazutoshi Takahashi¹ and Shinya Yamanaka^{1,2,*}
¹Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
²CREST, 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., 2007), indicating that fertilized 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

生成モデル

for 文書 $d = 1, 2, \dots, D$

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

for 単語 $n = 1, 2, \dots, N_d$

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

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

for 文書ペア $d = 1, 2, \dots, D, d' = 1, 2, \dots, D$

doc-doc link observation

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

for トピック $k = 1, 2, \dots, K$

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

$$\bar{\mathbf{z}}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} \mathbf{z}_{d,n} \quad \mathbf{z}_{d,n} \text{を} K \text{次元ベクトルとして見えています}$$

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

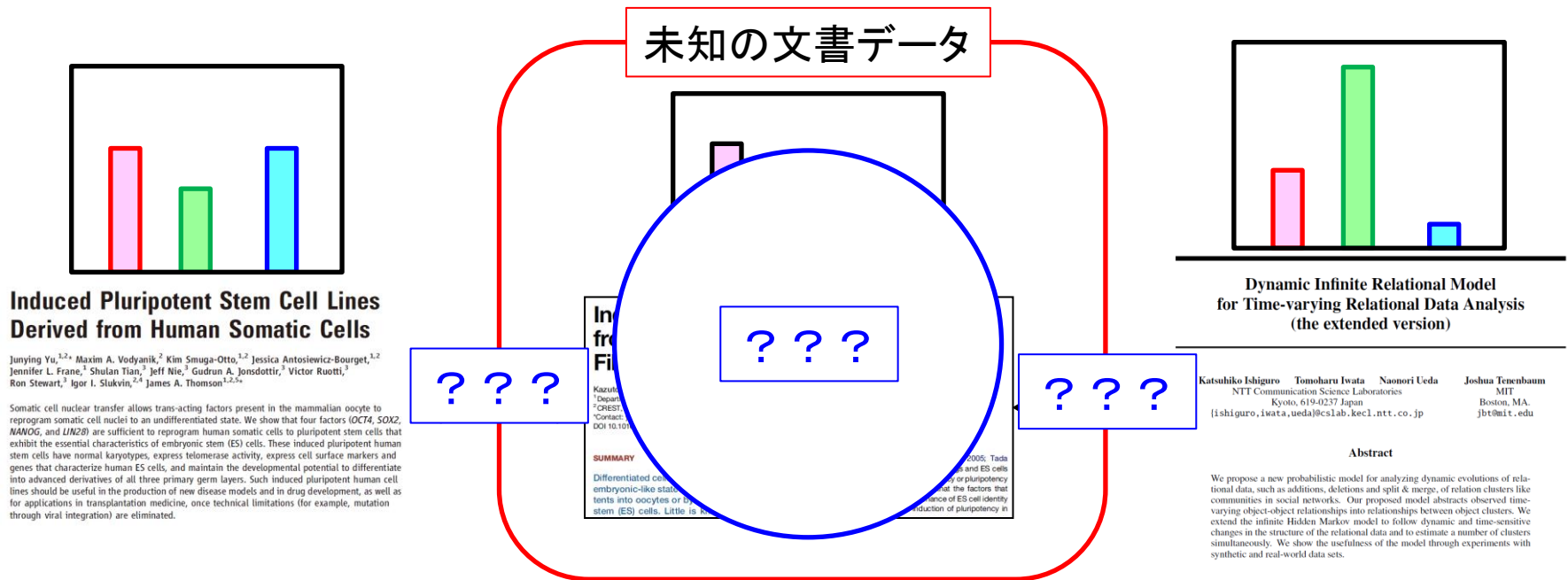
指数モデル $\psi(y_{d,d'} | \mathbf{z}_d, \mathbf{z}_{d'}, \boldsymbol{\eta}, \nu) = \exp(\boldsymbol{\eta}^T (\bar{\mathbf{z}}_d \circ \bar{\mathbf{z}}_{d'}) + \nu)$

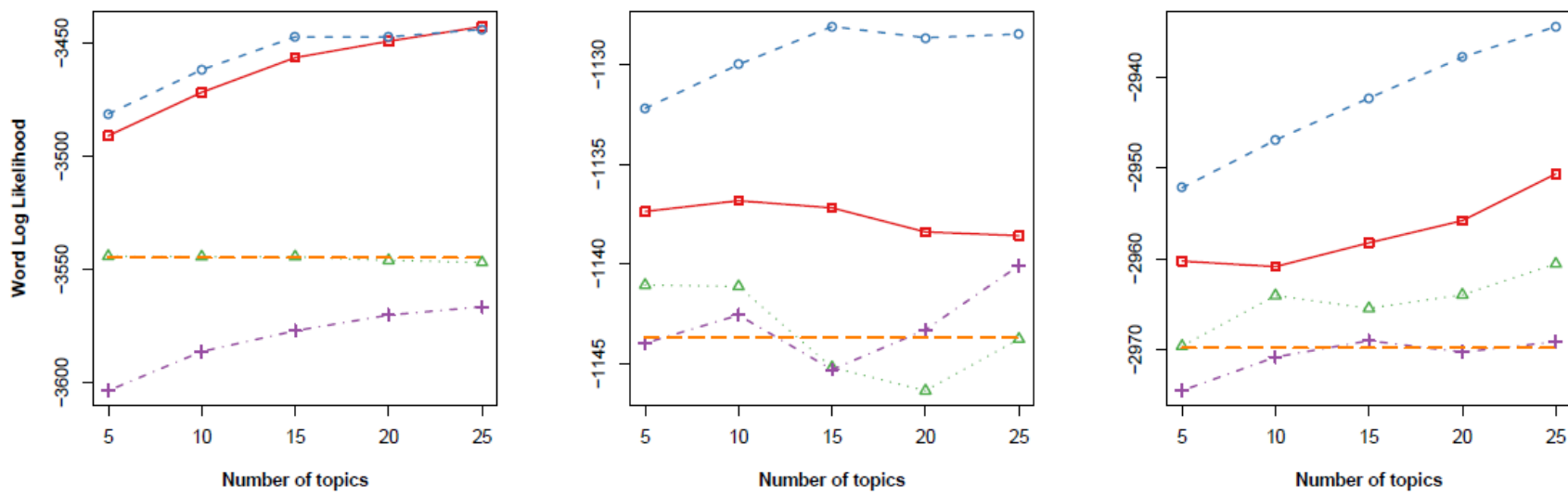
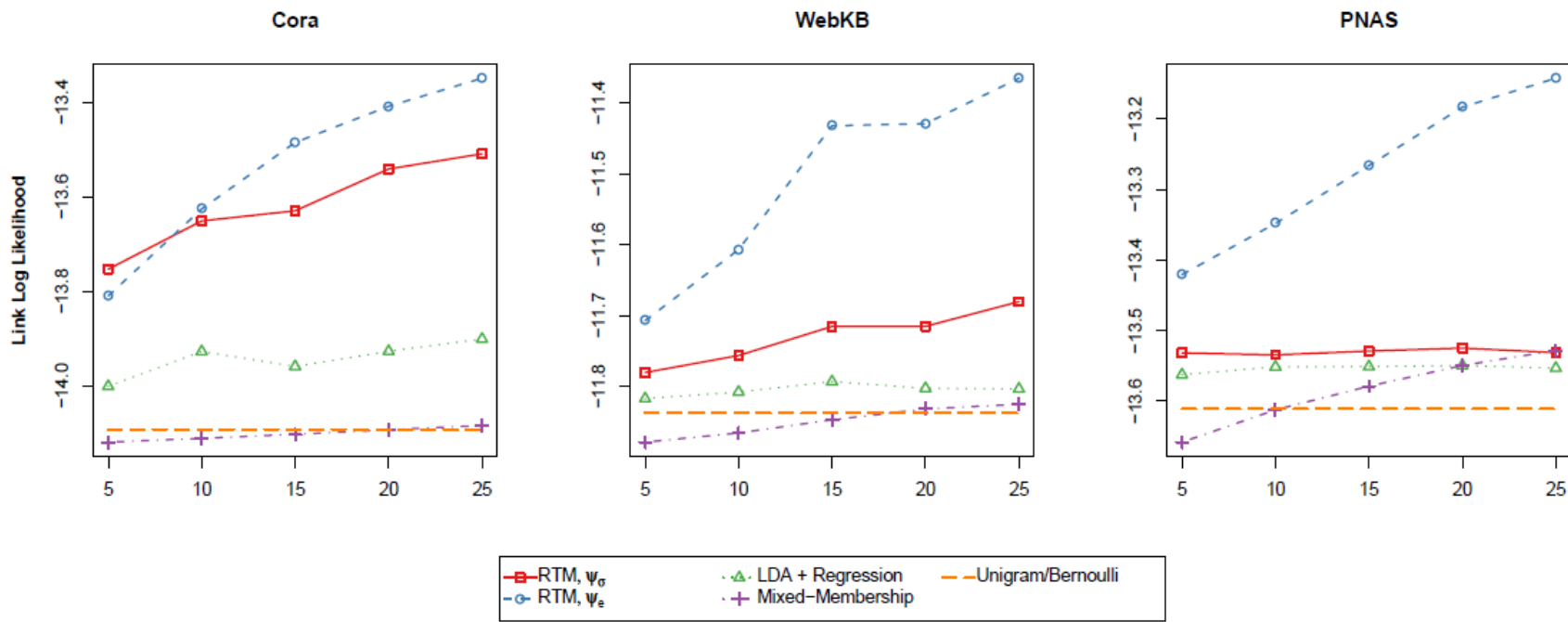
隠れ変数、パラメータの推定

- 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

<p><i>Markov chain Monte Carlo convergence diagnostics: A comparative review</i></p>	
<p>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</p>	RTM (ψ_e)
<p>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</p>	LDA + Regression

[Chang and Blei, 2009]

Relational Topic Model: まとめ

- 文書と文書の間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.