ネットワークとテキストデータに対するトピックモデリング

五十嵐　未来

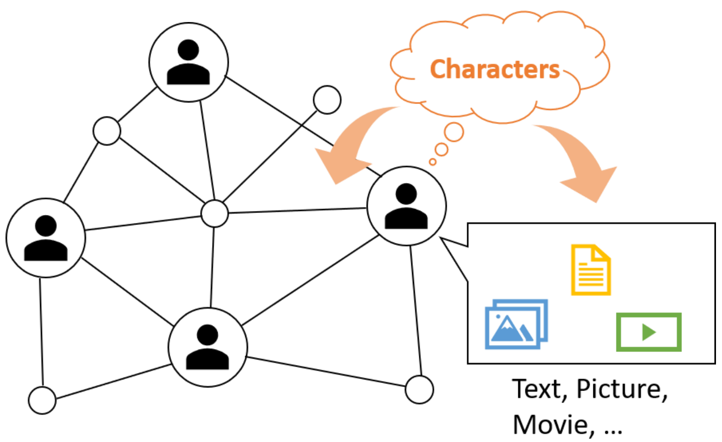
Introduction (self)

- 五十嵐未来（25歳・独身・仙台出身）
- 趣味は音楽、ヨガを始めようとしている

- 経済学研究科D2・データ科学国際共同大学院
- DC1取ったところが人生のピーク
- ジャーナルreject＆コンペ落選＆研究進まない
- 8月からMarylandへ
- Doctor良いことないっすね

Introduction (research)

- マーケティングの目的：「personalityを知ること」
  - ➢ デモグラフィック・アンケート・行動データが用いられる
  - ➢ ソーシャルメディア上における社会ネットワークの形成とコンテンツの生成
  - ➢ 非構造かつ大規模・情報量豊富なデータ
- 本研究の目的：
  社会ネットワークとUGCを考慮した
  “特性”を推定する統計モデルの提案



Model specification

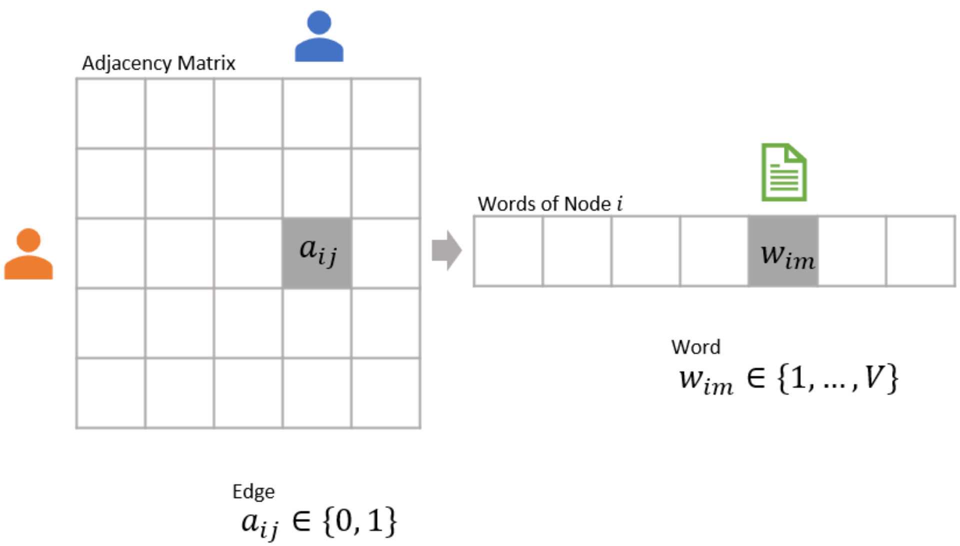
Data

- 隣接行列 A (0: not connected, 1: connected)

$$a_{ij} \in \{0, 1\}, \qquad i, j = 1, \ldots, D$$

- Bag of words W (1: baseball, 2: book, ⋯, V: iPhone)

$$w_{im} \in \{1, \ldots, V\}, \qquad m = 1, \ldots, M_i$$



Model specification

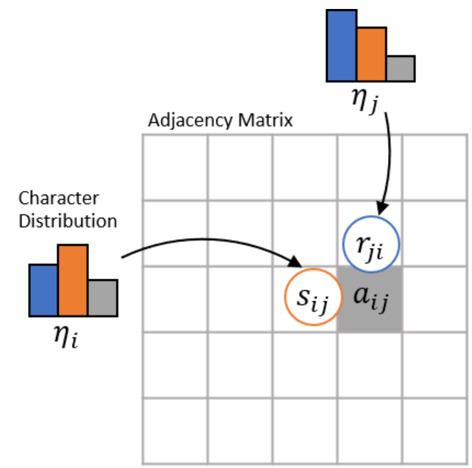
Network

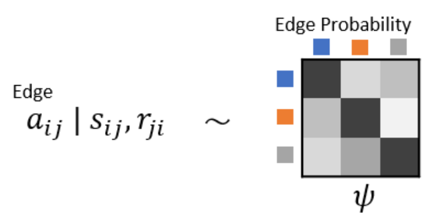
- エッジ $i \to j$について、送り手 i と 受け手 j は**特性分布** (η) に従う潜在特性 (s_{ij}, r_{ji}) を持つ

$$s_{ij} \sim Categorical(\eta_i), \qquad r_{ji} \sim Categorical(\eta_j)$$
$$s_{ij}, r_{ji} \in \{1, \ldots, K\}$$
$$\sum_{k=1}^{K} \eta_{ik} = 1,\ \forall i, \qquad \eta_{ik} \geq 0,\ \forall k$$

- s_{ij} と r_{ji} が与えられれば、エッジ a_{ij} は **エッジ確率** (ψ) に従って生成される

$$a_{ij} | s_{ij}, r_{ji} \sim Bernoulli(\psi_{s_{ij}, r_{ji}}),$$
$$0 \leq \psi_{kk'} \leq 1,\ \forall k, k'$$





Model specification

Text

- ノード i の m 番目の単語は潜在特性 (x_{im}) と潜在トピック (z_{im}) を持ち、特性分布 (η) と**トピック分布** (θ) に従う

$$x_{im} \sim Categorical(\eta_i), \qquad x_{im} \in \{1, \ldots, K\}$$

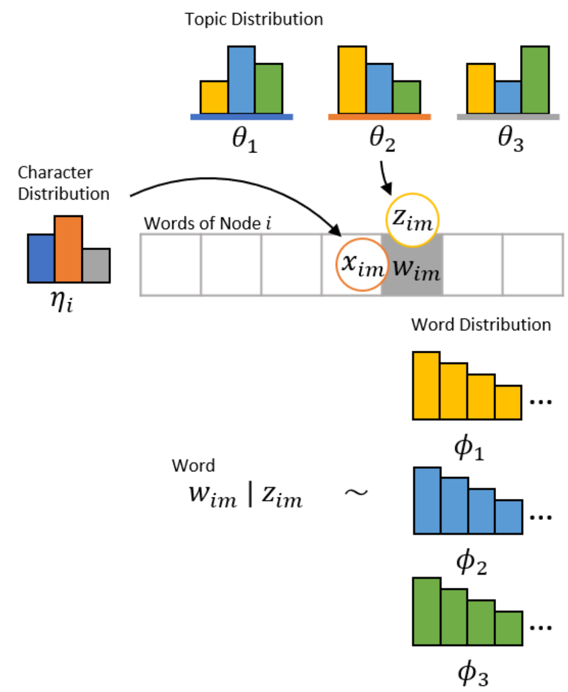
$$z_{im}|x_{im} \sim Categorical(\theta_{x_{im}}), \qquad z_{im} \in \{1, \ldots, L\}$$

$$\sum_{l=1}^{L} \theta_{kl}, \ \forall k, \qquad \theta_{kl} \geq 0, \ \forall l$$

- z_{im}が与えられれば、単語 w_{im} は**単語分布** (ϕ) に従って生成される

$$w_{im}|z_{im} \sim Categorical(\phi_{z_{im}})$$

$$\sum_{v=1}^{V} \phi_{lv}, \ \forall l, \qquad \phi_{lv} \geq 0, \ \forall v$$



Estimation

共役性に従って事前分布を設定

尤度	事前分布	（完全条件付き） 事後分布
$P(s_{ij}\|\eta_i) = Categorical(\eta_i)$ $P(r_{ij}\|\eta_i) = Categorical(\eta_i)$ $P(x_{im}\|\eta_i) = Categorical(\eta_i)$	$P(\eta_i\|\gamma) = Dirichlet(\gamma)$	$P(\eta_i\|s_i, r_i, x_i, \gamma) =$ $Dirichlet(N_i + M_i + \gamma_k)$
$P(a_{ij}\|s_{ij}, r_{ji}, \psi) =$ $Bernoulli(\psi_{s_{ij}, r_{ji}})$	$P(\psi_{kk'}\|\delta, \epsilon) = Beta(\delta, \epsilon)$	$P(\psi_{kk'}\|A, S, R, \delta, \epsilon) =$ $Beta(n_{kk'}^{(p)} + \delta, n_{kk'}^{(m)} + \epsilon)$
$P(z_{im}\|x_{im}, \theta) =$ $Categorical(\theta_{x_{im}})$	$P(\theta_k\|\alpha) = Dirichlet(\alpha)$	$P(\theta_k\|X, Z, \alpha) =$ $Dirichlet(M_k + \alpha)$
$P(w_{im}\|z_{im}, \phi) =$ $Categorical(\phi_{z_{im}})$	$P(\phi_l\|\beta) = Dirichlet(\beta)$	$P(\phi_l\|W, Z, \beta) =$ $Dirichlet(M_l + \beta)$

Estimation

崩壊型ギブスサンプリングを使ってパラメータを推定

パラメータを積分消去し、潜在変数の条件付き事後分布は以下のように導出される (Igarashi & Terui 2019)：

$$P(s_{ij}=k, r_{ji}=k' | a_{ij}, A_{\setminus ij}, S_{\setminus ij}, R_{\setminus ji}, X, \gamma, \delta, \epsilon)$$
$$= \frac{N_{ik\setminus ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\setminus ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\setminus ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\setminus ji} + M_{jt} + \gamma_t\right)} \times \frac{\left(n^{(p)}_{kk'\setminus ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(m)}_{kk'\setminus ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(p)}_{kk'\setminus ij} + n^{(m)}_{kk'\setminus ij} + \delta_{kk'} + \epsilon_{kk'}}$$

$$P(x_{im}=k, z_{im}=l | w_{im}=v, W_{\setminus im}, S, R, X_{\setminus im}, Z_{\setminus im}, \alpha, \beta, \gamma)$$
$$= \frac{M_{lv\setminus im} + \beta_v}{\sum_u \left(M_{lu\setminus im} + \beta_u\right)} \times \frac{M_{kl\setminus im} + \alpha_l}{\sum_q \left(M_{kq\setminus im} + \alpha_q\right)} \times \frac{N_{ik} + M_{ik\setminus im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\setminus im} + \gamma_t\right)}$$

Estimation

上記のサンプリング式から得られるサンプルを用いてパラメータは点推定される:

$$\widehat{\eta}_{ik} = \frac{1}{G-b} \sum_{g=b+1}^{G} \frac{N_{ik}^{(g)} + M_{ik}^{(g)} + \gamma_k}{\sum_t \left(N_{it}^{(g)} + M_{it}^{(g)} + \gamma_t \right)}$$

$$\widehat{\psi}_{kk'} = \frac{1}{G-b} \sum_{g=b+1}^{G} \frac{n_{kk'}^{(p,g)} + \delta_{kk'}}{n_{kk'}^{(p,g)} + n_{kk'}^{(m,g)} + \delta_{kk'} + \epsilon_{kk'}}$$

$$\widehat{\theta}_{kl} = \frac{1}{G-b} \sum_{g=b+1}^{G} \frac{M_{kl}^{(g)} + \alpha_l}{\sum_q \left(M_{kl}^{(g)} + \alpha_q \right)}$$

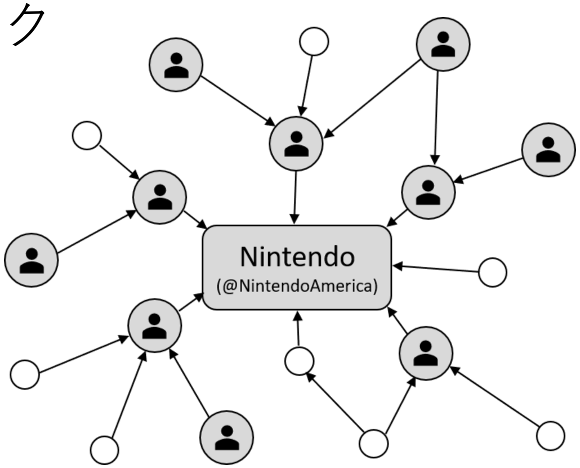
$$\widehat{\phi}_{lv} = \frac{1}{G-b} \sum_{g=b+1}^{G} \frac{M_{lv}^{(g)} + \beta_v}{\sum_u \left(M_{lu}^{(g)} + \beta_u \right)}.$$

特性の数 (K) とトピックの数 (L) は WAIC (Watanabe 2010) によって探索する

Dataset

実証分析で用いるTwitterデータは以下で構成される：

- 任天堂アカウントを中心とするネットワーク
（2018年5月1日におけるフォロー関係）
- タイムライン上に投稿されたTweets
（2017年9月1日から2018年2月28日）



サンプリングと前処理を施した後の要約統計量：

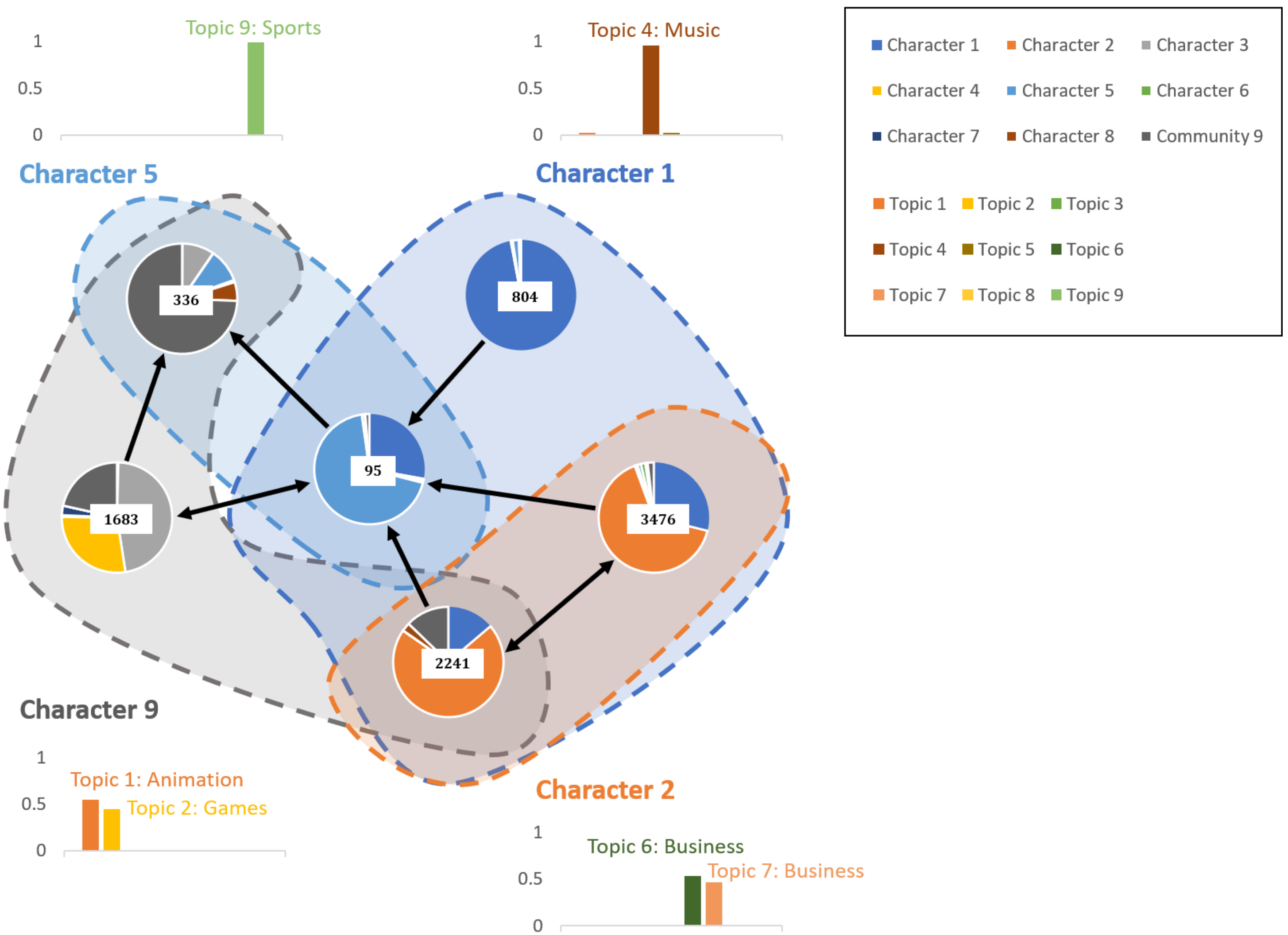
D (nodes)	V (words)	Ave. links (sparsity)	Ave. words (sparsity)
3,500	9,001	19.7 links (0.56%)	59.3 words (1.69%)

Empirical results

各トピックで頻出する上位10個の単語

Topic 1 (Animation)	Topic 2 (Game)	Topic 3 (E-sports)	Topic 4 (Music)	Topic 5 (Every life)	Topic 6 (Business)	Topic 7 (Business)	Topic 8 (Streaming Broadcasting)	Topic 9 (Sports)
podernfamili	vgc	hori	vevo	leed	trapadr	growthhack	nonfollow	zeldathon
gamedesign	savvi	mkleosaga	spinrilla	cto	digitalmarket	gdpr	teamemmmmsi	dokkan
criticalrol	gamedesign	wnf	lube	momlif	ddrive	socialmediamarket	twitchkitten	htgawm
blackclov	steinsgat	mdva	suav	dogsoftwitt	contentmarket	iartg	roku	orton
hunterxhunter	nyxl	hyrulesaga	drippin	beck	smm	smm	wizebot	oiler
jojosbizarreadventur	xenovers	cfl	ahscult	austria	amread	gainwithpyewaw	ryzen	sdlive
fursuitfriday	acnl	nood	wshh	hemp	bigdata	asmsg	airdrop	horford
tfc	artstat	qanba	ouija	tock	gdpr	ifb	dg	herewego
amiga	firer	zeku	foodporn	crowdfir	gainwithxtiandela	digitalmarket	freebiefriday	rozier
sml	tamagotchi	junedecemb	sizzl	monaco	fiverr	css	streamersconnect	earnhistori

Empirical results



Prediction

- 設定

  Network (19.7 edges/node) $\xrightarrow{10\%}$ テストデータ

  Network (19.7 edges/node) $\xrightarrow{90\%}$ 訓練データ

  Text (59.3 words/node) $\xrightarrow{100\%}$ 訓練データ

- 予測確率

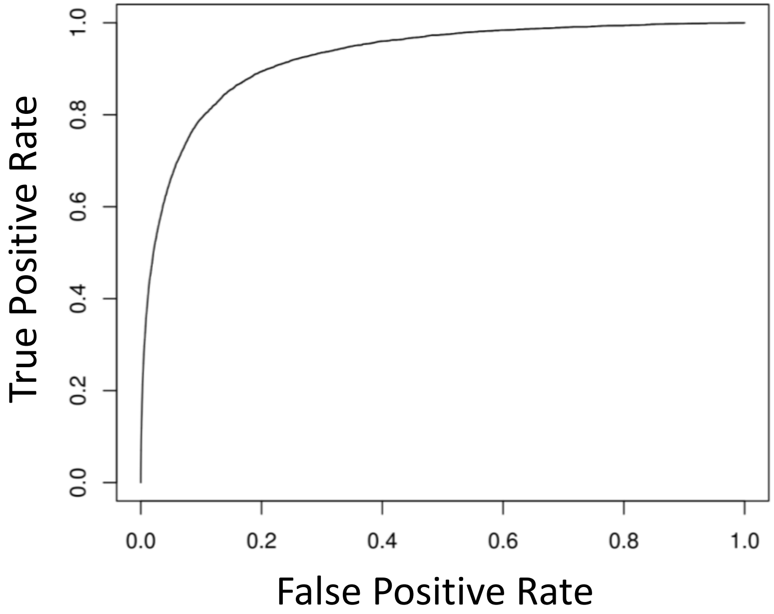
$$P(a_{ij} = 1) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \widehat{\eta}_{ik} \cdot \widehat{\eta}_{jk'} \cdot \widehat{\psi}_{kk'}$$

- 評価方法
  - Area Under the Curve (AUC)
  - Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(FN + TN)(TN + FP)(FP + TP)}}$$

Prediction

- ROC curve & AUC



- AUC: 0.93 (Perfect -> 1.00, At Random -> 0.5)
- 予測性能の良さを示す

Prediction

- 混同行列

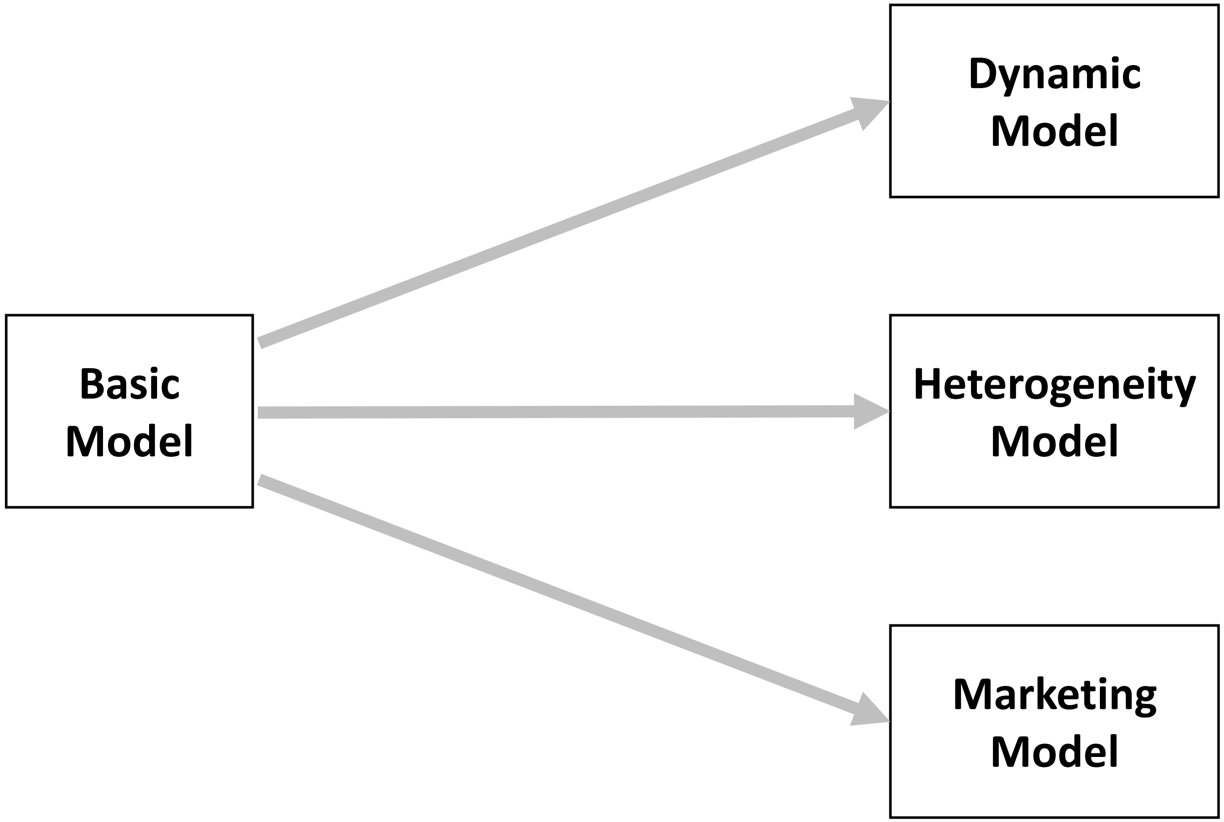
		Prediction	
		Link	Non-link
Data	Link	2,041	4,786
	Non-link	7,079	1,211,094

  - Cutoff: 0.08
  - MCC: 0.254
  - True Positive Rate: $\frac{2{,}041}{2{,}041+4{,}786} \approx$ 29.9%
- モデルの予測性能
  - リンクの予測に対して十分な精度
  - 改善の余地あり
(ex. modeling sparsity, Airoldi et al 2008; Latouche et al 2011)

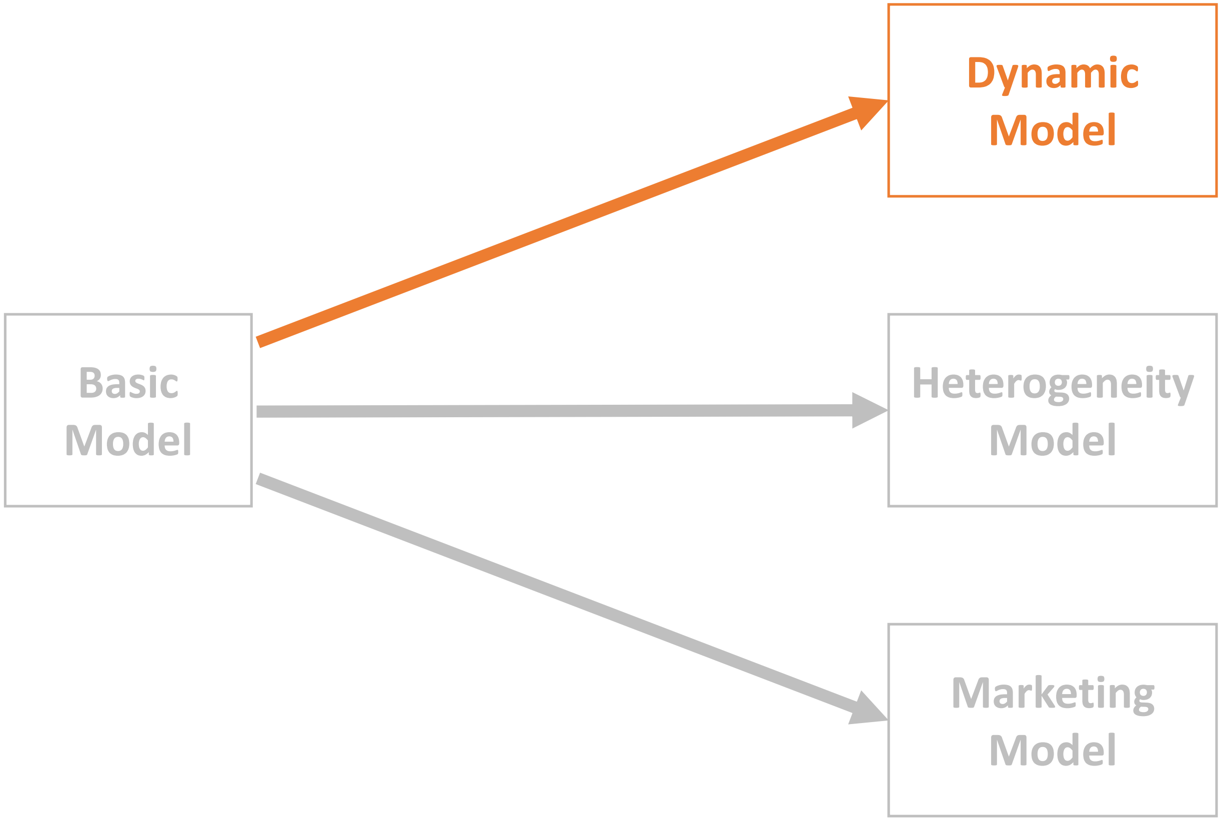
Conclusion (basic)

- ソーシャルメディア上の**社会ネットワークとテキスト**データから**人々の特性**を推定する
  - トピックモデルを提案
- 人々は自分の特性に応じてネットワークの形成とテキストコンテンツの生成を行っている
  - 特性分布 (η_i) が反映している
- 実証分析
  - ネットワークとテキストから意味のある特性を抽出
  - ネットワーク情報と単語トピックにより特性に意味を付与
  - テストデータ予測において十分な精度を示す

Where is next ?



Where is next ?



Introduction (dynamics)

- ネットワークとテキストは時間で変化している

  - “特性” と “トピック” が発展する

- 本研究の目的：
  時間発展するネットワークとテキストの**動的変化**を捉える統計モデルを提案する

Model specification

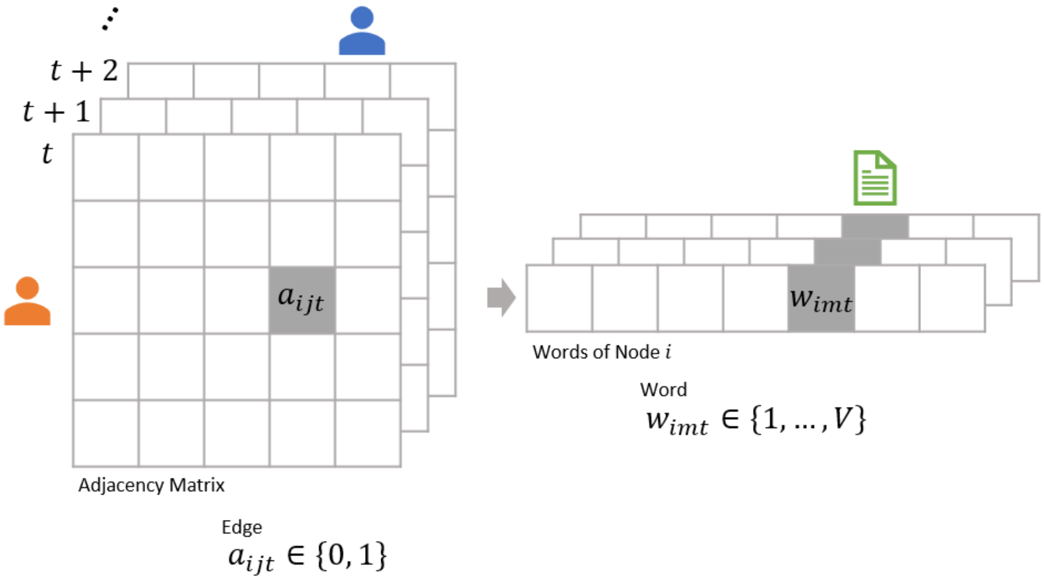
Data

- 隣接行列 A (0: not connected, 1: connected)

$$a_{ijt} \in \{0,1\}, \qquad i,j = 1,\ldots,D, \quad t = 1,\ldots,T$$

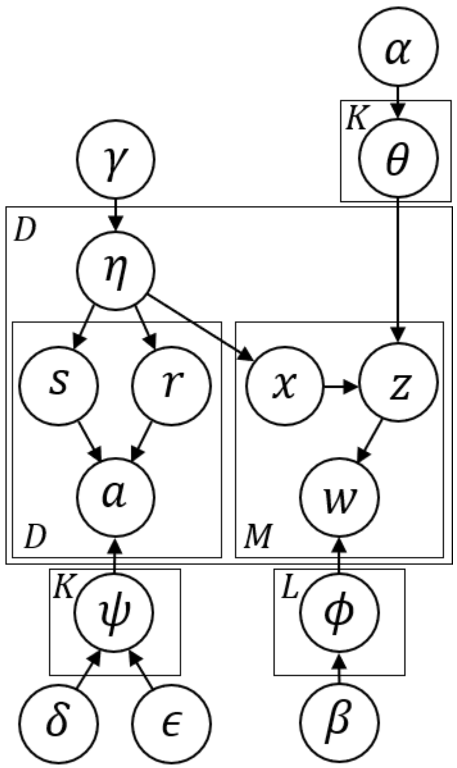
- Bag of words W (1: baseball, 2: book, ⋯, V: iPhone)

$$w_{imt} \in \{1,\ldots,V\}, \qquad m = 1,\ldots,M_{it}, \quad t = 1,\ldots,T$$

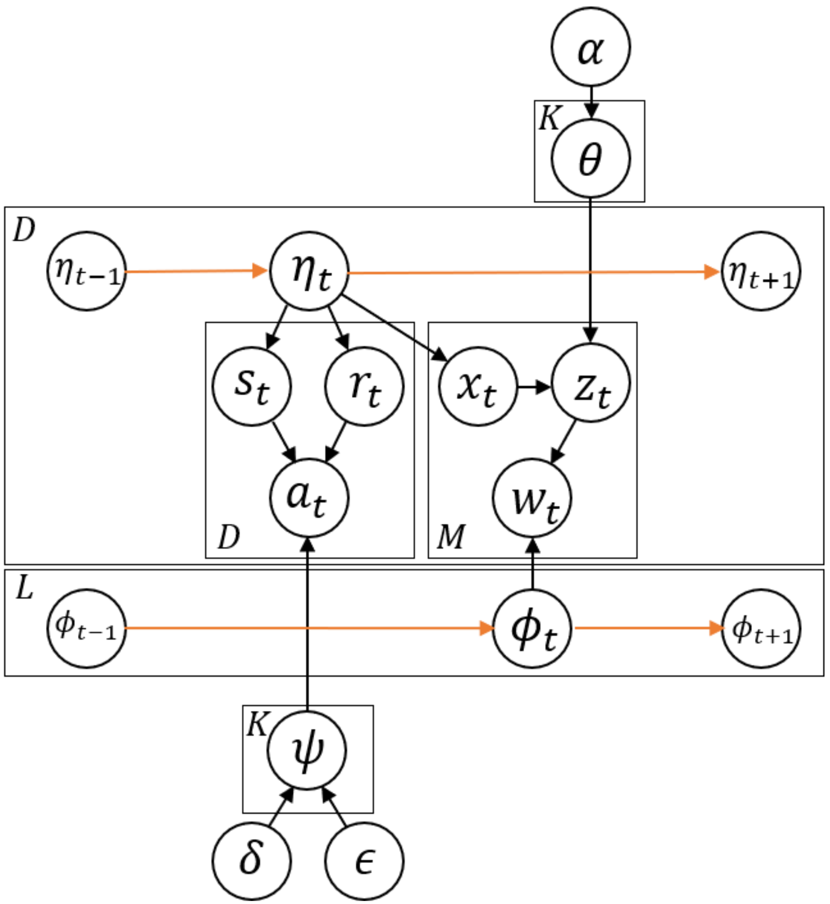


Model specification

提案モデル (Static)



提案モデル (Dynamic)



Model specification

Network

- **特性分布** (η_t) がガウスノイズと共に発展していく (random walk)

$$\eta_{it}|\eta_{it-1} \sim N_K\left(\eta_{it-1}, \sigma_\eta^2 I\right)$$

- 時点 t におけるエッジ $i \rightarrow j$について、送り手 i と受け手 j は正規化特性分布に従う潜在特性 (s_{ijt}, r_{jit}) を持つ

$$s_{ijt} \sim Categorical\left(\pi\left(\eta_{it}\right)\right), \qquad r_{jit} \sim Categorical\left(\pi\left(\eta_{jt}\right)\right)$$

$$\pi(x) = \frac{exp(x_k)}{\sum_{k'} exp(x_{k'})}$$

- s_{ijt} と r_{jit} が与えられれば、エッジ a_{ijt} は**エッジ確率** (ψ) に従って生成される

$$a_{ijt}|s_{ijt}, r_{jit} \sim Bernoulli\left(\psi_{s_{ijt}, r_{jit}}\right)$$

Model specification

Text

- 時点 t におけるノード i の m 番目の単語は潜在特性 (x_{imt}) と潜在トピック (z_{imt}) を持ち、特性分布 (η_t) と**トピック分布** (θ)に従って生成される

$$x_{imt} \sim Categorical(\pi\left(\eta_{it}\right)), \qquad z_{imt}|x_{imt} \sim Categorical(\theta_{x_{imt}})$$

- **単語分布** (ϕ_t) がガウスノイズと共に発展していく

$$\phi_{lt}|\phi_{lt-1} \sim N_V\left(\phi_{lt-1}, \sigma_\phi^2 I\right)$$

- z_{imt} が与えられれば、単語 w_{imt}は正規化単語分布に従って生成される

$$w_{imt}|z_{imt} \sim Categorical\left(\pi\left(\phi_{z_{imt}t}\right)\right)$$

Estimation

- 静的モデルでは、共役事前分布を設定し、（崩壊型）ギブスサンプラーを導出することが出来た

尤度： $s_{ij}|\eta_i \sim Categorical\left(\eta_i\right)$　事前分布： $\eta_i \sim Dirichlet\left(\gamma\right)$

事後分布： $\eta_i|\cdot \sim Dirichlet\left(\cdot\right)$

- 動的モデルでは事前分布が共役でない

尤度： $s_{ijt}|\eta_{it} \sim Categorical\left(\pi\left(\eta_{it}\right)\right)$　事前分布： $\eta_{it} \sim N\left(\eta_{it-1}, \sigma_\eta^2 I\right)$

事後分布：事前分布と同じ形で事後分布を導出できない

→ **変分ベイズ**

- 変分ベイズは、KL情報量の意味で真の事後分布と最も近い変分事後分布を探索する

$$\beta = \{\eta_{1:T}, \phi_{1:T}, \psi, \theta, s_{1:T}, r_{1:T}, x_{1:T}, z_{1:T}\}$$
$$q(\beta|data) = \arg\min_q \; KL[q(\beta)||p(\beta|data)]$$
$$\text{s.t. } q(\beta) \text{ is factorizable}$$

Estimation

平均場族の仮定

$$q(\beta) = \prod_{i=1}^{D} \{q(\eta_{i1}, \dots, \eta_{iT})\} \times \prod_{l=1}^{L} \{q(\phi_{l1}, \dots, \phi_{lT})\} \times \prod_{k=1}^{K} \left\{ q(\theta_k) \prod_{k'=1}^{K} q(\psi_{kk'}) \right\}$$
$$\times \prod_{t=1}^{T} \left\{ \prod_{i=1}^{D} \left[\prod_{j=1}^{D} q(s_{ijt}) q(r_{jit}) \right] \left[\prod_{m=1}^{M_{it}} q(x_{imt}) q(z_{imt}) \right] \right\}$$

- $q(\eta_{i1}, \dots, \eta_{iT})$ と $q(\phi_{l1}, \dots, \phi_{lT})$ はこれ以上分解すべきでない
→ 結合分布に時間依存性を仮定しているから

変分ベイズ + **カルマンフィルタ** (Blei & Lafferty 2006).

モデル

$$\begin{cases} \eta_{it} | \eta_{it-1} \sim N(\eta_{it-1}, \sigma_\eta^2 I) \\ s_{ijt} | \eta_{it} \sim Categorical\,(\pi(\eta_{it})) \end{cases}$$

→

近似

$$\begin{cases} \eta_{it} | \eta_{it-1} \sim N(\eta_{it-1}, \sigma_\eta^2 I) \\ \hat{\eta}_{it} | \eta_{it} \sim N(\eta_{it}, \rho_\eta^2 I) \end{cases}$$

変分観測量

- カルマンフィルタは線形ガウスフィルタリングに対するベイズ最適解を閉じた形で与える
- $\eta_{1:T}$ の推定をカルマンフィルタで、$\hat{\eta}_{1:T}$ の推定を変分ベイズで行う

Estimation

フィルタ分布

$$q(\eta_{it}|\hat{\eta}_{i1:t}) = N\left(\mu_{it}, \lambda_{it}^2 I\right)$$

$$\mu_{it} = \left(\frac{\rho_\eta^2}{\lambda_{it}^2 + \sigma_\eta^2 + \rho_\eta^2}\right)\mu_{it-1} + \left(1 - \frac{\rho_\eta^2}{\lambda_{it}^2 + \sigma_\eta^2 + \rho_\eta^2}\right)\hat{\eta}_{it}, \quad \lambda_{it}^2 = \left(\frac{\rho_\eta^2}{\lambda_{it}^2 + \sigma_\eta^2 + \rho_\eta^2}\right)(\lambda_{it-1}^2 + \sigma_\eta^2)$$

$$q(\phi_{lt}|\hat{\phi}_{l1:t}) = N\left(\pi_{lt}, \omega_{lt}^2 I\right)$$

$$\pi_{lt} = \left(\frac{\rho_\phi^2}{\omega_{lt}^2 + \sigma_\phi^2 + \rho_\phi^2}\right)\pi_{lt-1} + \left(1 - \frac{\rho_\phi^2}{\omega_{lt}^2 + \sigma_\phi^2 + \rho_\phi^2}\right)\hat{\phi}_{lt}, \quad \omega_{lt}^2 = \left(\frac{\rho_\phi^2}{\omega_{lt}^2 + \sigma_\phi^2 + \rho_\phi^2}\right)(\omega_{lt-1}^2 + \sigma_\phi^2)$$

平滑化分布

$$q(\eta_{it}|\hat{\eta}_{i1:T}) = N\left(\tilde{\mu}_{it}, \tilde{\lambda}_{it} I\right)$$

$$\tilde{\mu}_{it} = \left(1 - \frac{\lambda_{it}^2}{\lambda_{it}^2 + \sigma_\eta^2}\right)\mu_{it} + \left(\frac{\lambda_{it}^2}{\lambda_{it}^2 + \sigma_\eta^2}\right)\tilde{\mu}_{it+1}, \quad \tilde{\lambda}_{it}^2 = \lambda_{it}^2 + \left(\frac{\lambda_{it}^2}{\lambda_{it}^2 + \sigma_\eta^2}\right)^2 \left(\tilde{\lambda}_{it+1}^2 - (\lambda_{it}^2 + \sigma_\eta^2)\right)$$

$$q(\phi_{lt}|\hat{\phi}_{l1:T}) = N\left(\tilde{\pi}_{lt}, \tilde{\omega}_{lt} I\right)$$

$$\tilde{\pi}_{lt} = \left(1 - \frac{\omega_{lt}^2}{\omega_{lt}^2 + \sigma_\phi^2}\right)\pi_{lt} + \left(\frac{\omega_{lt}^2}{\omega_{lt}^2 + \sigma_\phi^2}\right)\tilde{\pi}_{lt+1}, \quad \tilde{\omega}_{lt}^2 = \omega_{lt}^2 + \left(\frac{\omega_{lt}^2}{\omega_{lt}^2 + \sigma_\phi^2}\right)^2 (\tilde{\omega}_{lt+1}^2 - (\omega_{lt}^2 + \sigma_\phi^2))$$

Estimation

Evidence Lower Bound (ELBO)

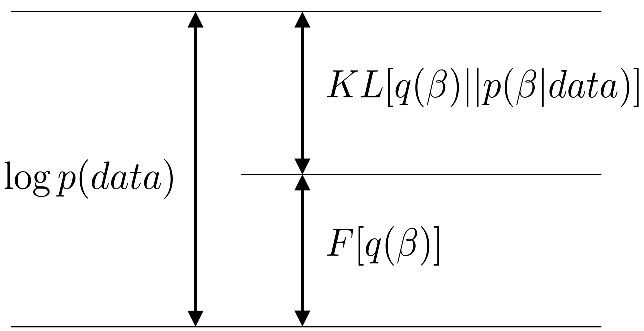
- KL情報量の最小化とELBOの最大化が数学的に等価である

$$\arg\min_{q} \; KL[q(\beta)||p(\beta|data)]$$

$$log(data) = F[q(\beta)] + KL[q(\beta)||p(\beta|data)]$$

$$F[q(\beta)] = \int q(\beta) \log \frac{p(data, \beta)}{q(\beta)} d\beta$$

$$\arg\max_{q} \; F[q(\beta)]$$



Estimation

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log(a_{1:T}, w_{1:T}) \geq \int \sum_{s,r,x,z} q(\beta_{1:T}) \log \frac{p(a_{1:T}, w_{1:T}, \eta_{1:T}, \phi_{1:T}, \psi, \theta)}{q(\eta_{1:T}, \phi_{1:T}, \psi, \theta)} d\eta d\phi d\psi d\theta \\
= \mathbb{E}_{q(s,r,\psi)} \left[\log p(a_{1:T}|s_{1:T}, r_{1:T}, \psi)\right] + \mathbb{E}_{q(\psi)} \left[\log \frac{p(\psi)}{q(\psi)}\right] \\
+ \mathbb{E}_{q(s,r,x,\eta)} \left[\log \frac{p(s_{1:T}|\eta_{1:T})p(r_{1:T}|\eta_{1:T})p(x_{1:T}|\eta_{1:T})}{q(s_{1:T})q(r_{1:T})q(x_{1:T})}\right] + \mathbb{E}_{q(\eta)} \left[\log \frac{p(\eta_{1:T})}{q(\eta_{1:T})}\right] \\
+ \mathbb{E}_{q(z,\phi)} \left[\log p(w_{1:T}|z_{1:T}, \phi_{1:T})\right] + \mathbb{E}_{q(\phi)} \left[\log \frac{p(\phi_{1:T})}{q(\phi_{1:T})}\right] \\
+ \mathbb{E}_{q(x,z,\theta)} \left[\log \frac{p(z_{1:T}|x_{1:T}, \theta)}{q(z_{1:T})}\right] + \mathbb{E}_{q(\theta)} \left[\log \frac{p(\theta)}{q(\theta)}\right]
\end{aligned}
$$

ELBOを各パラメータに対して変分することで停留点を見つける

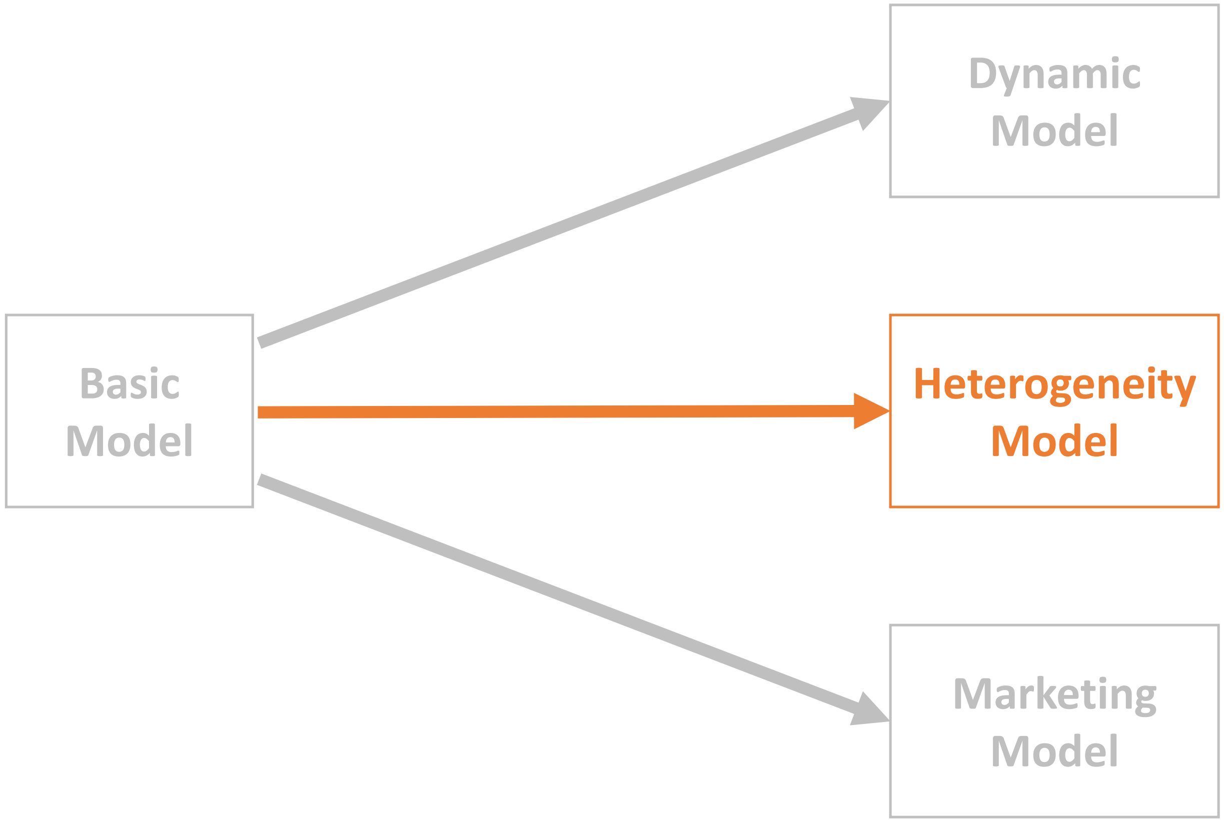
→ ELBOが収束するまで**変分パラメータを更新する**

変分観測量 $(\hat{\eta}_t, \hat{\phi}_t)$ を使って時間発展パラメータ (η_t, ϕ_t) に対する
フィルタリングと平滑化を行う

Conclusion (dynamic)

- 時間発展するネットワークとテキストを性質を捉える動的トピックモデルを提案
- **特性分布** (η_t) と**単語分布** (ϕ_t) がガウスノイズと共に発展していくことを仮定
- **変分ベイズ**と**カルマンフィルタ**の組み合わせによる推定
  - 線形ガウス状態空間モデルを構築するために変分観測量を導入し、カルマンフィルタ・平滑化によって最適解を計算
  - ELBOの変分によって求めた停留点で変分パラメータを更新

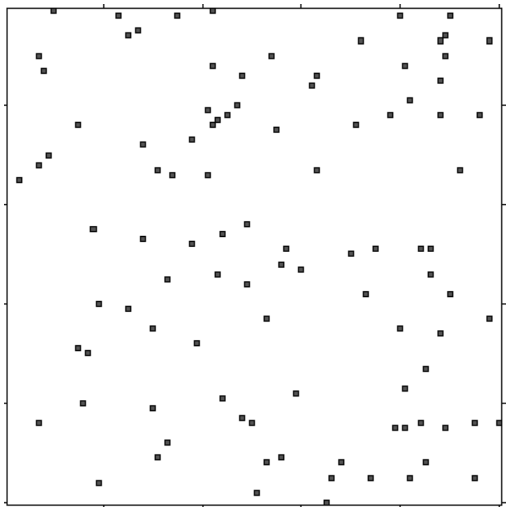
Where is next ?



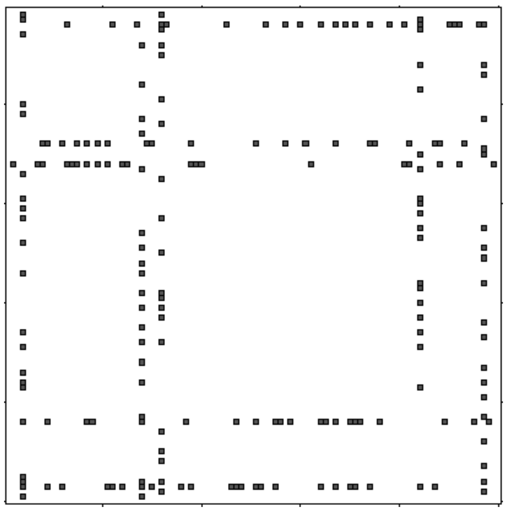
Introduction (heterogeneity)

- 社会ネットワークにはノードに**異質的構造**がある
  - 少数のノードが多くのエッジを集める（ハブノード）
  - 他大多数のノードは少数の関係性しか持たない

ランダムスパース



規則的なスパース



→エッジ確率を異質にする (ψ -> $\psi^{(i,j)}$)

Model specification

- Basicモデル

$$a_{ij}|s_{ij}=k, r_{ji}=k' \sim Bernoulli(\psi_{kk'})$$

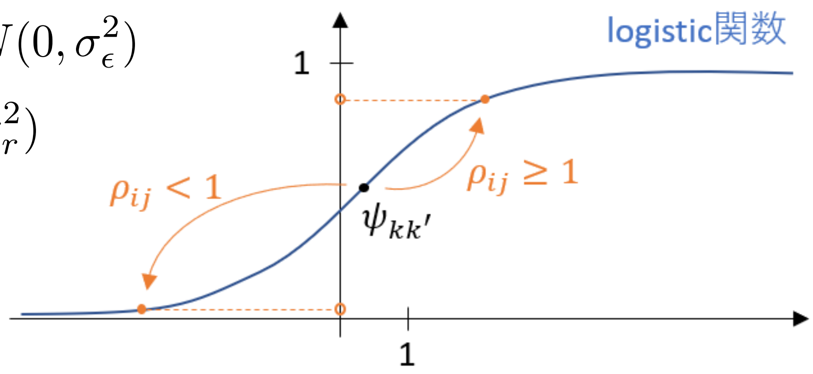
- Heteroモデル
  - ネットワーク生成過程

$$a_{ij}|s_{ij}=k, r_{ji}=k' \sim Bernoulli\left(logistic(\rho_{ij}\cdot\psi_{kk'})\right)$$

  - 分解モデル

$$\rho_{ij}=\lambda_i^{(s)}+\lambda_j^{(r)}+\epsilon_{ij}, \quad \epsilon_{ij}\sim N(0,\sigma_\epsilon^2)$$

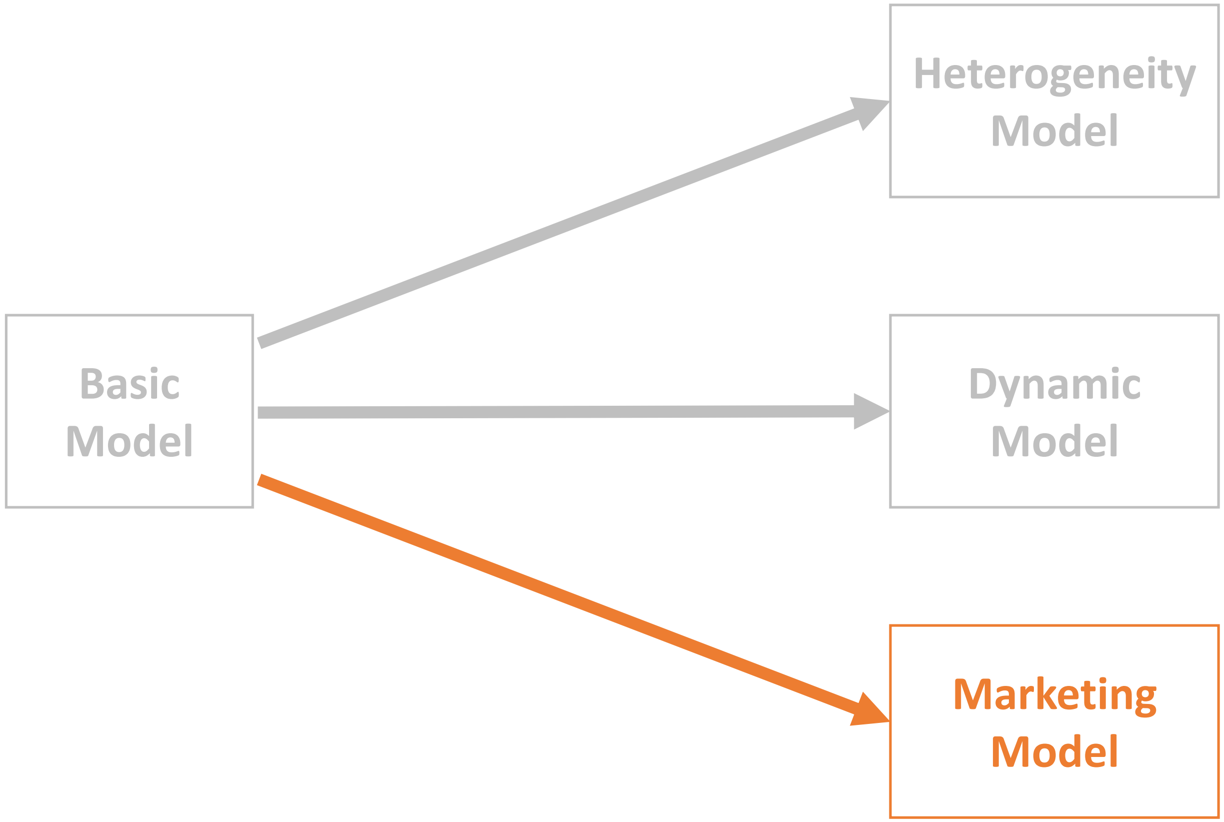
$$\lambda_i^{(s)}\sim N(0,\sigma_s^2), \quad \lambda_j^{(r)}\sim N(0,\sigma_r^2)$$



Estimation

- MCMC？
  - $\psi_{kk'}, \rho_{ij}$ について共役でない
  - その他のパラメータについては共役
  - MH-Gibbsサンプリング
- 変分ベイズ？

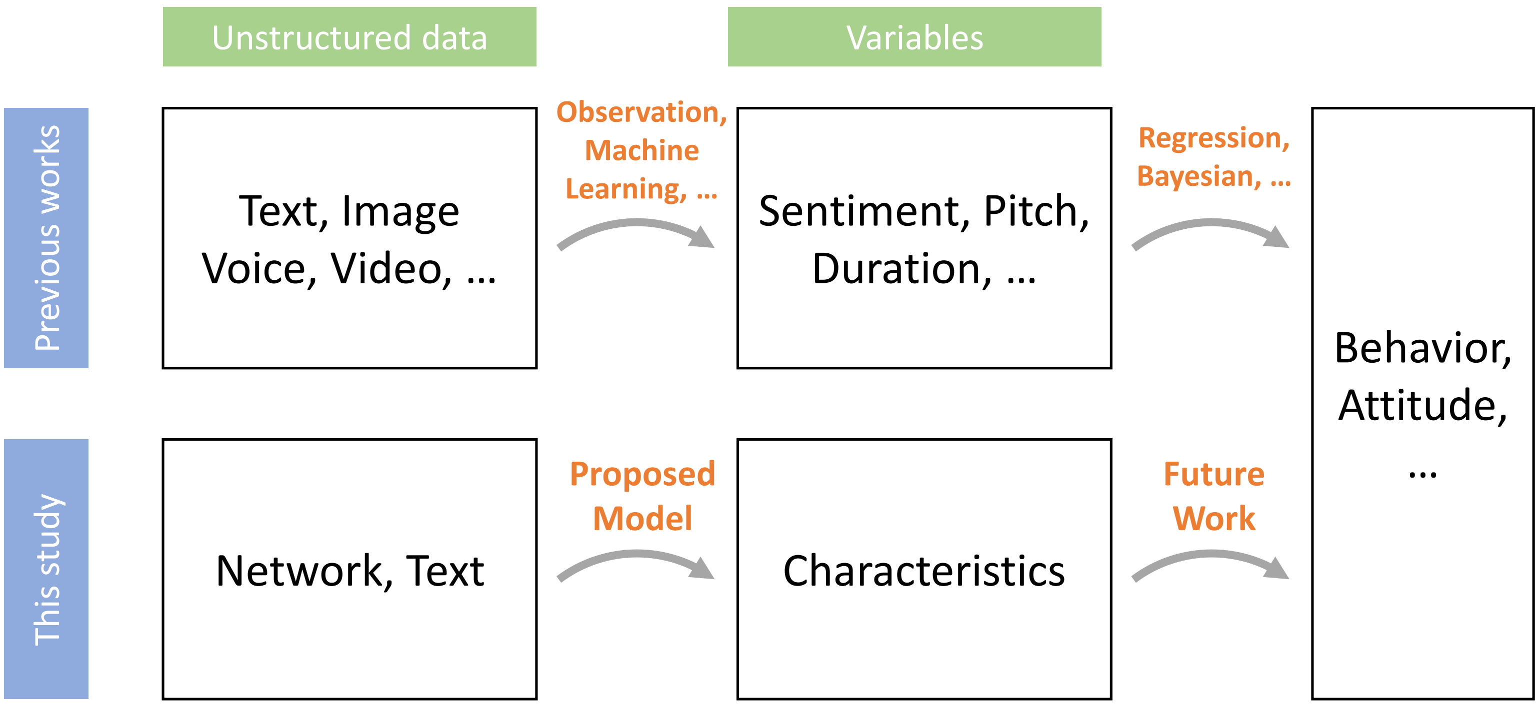
Where is next ?



Introduction (marketing)

マーケティングの目的：“Personalityを知ること”

→ 購買行動への影響



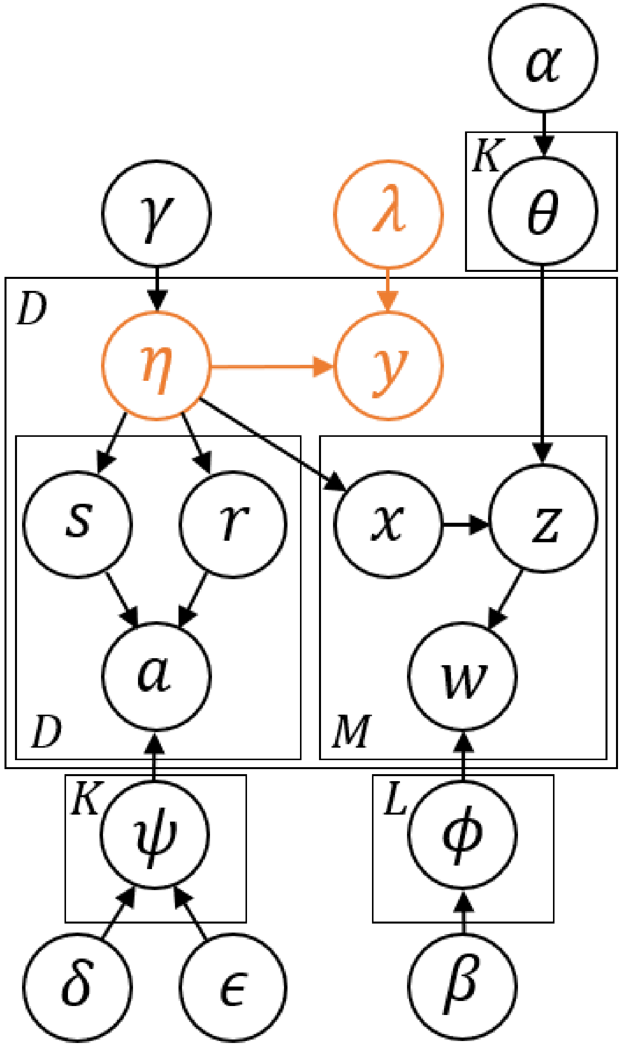
Model specification

教師ありトピックモデル

$$y_i = \boldsymbol{\lambda}^T \boldsymbol{\eta}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

y_i ：購買行動変数

→ ネットワーク・テキスト・購買行動を考慮した特性



Model specification

説明変数の候補

$$y_i = \boldsymbol{\lambda}^T \boldsymbol{\eta}_i + ?? + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

- マーケティング変数（**価格**、**特別陳列**など）

  → 企業が管理可能な変数、購買への影響を知りたい主たる変数

- ネットワーク変数（Hetero modelの$\boldsymbol{\lambda_i^{(s)}}$, $\boldsymbol{\lambda_i^{(r)}}$など）

  → エッジを多く結ぶ・結ばれる人が購買にどう影響するか

- テキスト変数（商品に関する**口コミ**など）

  → 自分自身の周りで投稿された口コミに影響を受けて購買行動が変化

Model specification

階層モデリング

$$y_i = \boldsymbol{\lambda}_i^T \boldsymbol{\eta}_i + ?? + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$
$$\lambda_i = \bar{\lambda} \boldsymbol{z}_i + e_i, \quad e_i \sim N(0, \sigma_e^2 I)$$

z_iとして以下のような変数が候補となる

- 消費者属性（性別、年代、職業、年収など）
- ブランドロイヤルティ
- 購買傾向（を示す情報）

→ λ_iの原因となるような変数、事後的に解釈がつく変数

Conclusion (whole)

- Basic model
  - ➢ネットワークとテキストの生成過程をモデリング
  - ➢Social Media上の人々の特性を推定
- Dynamic model
  - ➢時間で発展するネットワークとテキストをモデリング
- Heterogeneity model
  - ➢社会ネットワークに存在する異質性をモデリング
  - ➢エッジ確率に階層構造を仮定し、異質な「つながりやすさ」を表現
- Marketing model
  - ➢購買行動も考慮しながら消費者の特性を推測するモデル

ベイズモデリングを使えば様々な変数・モデル構造を仮定できる