Guided by Your Stars: Two-Mode Segmentation Using Large-Scale Online Product Rating Networks

Today's Talk

- Motivation
- Proposed Model
- Simulation Study
- Application: Amazon Online Review
- Conclusions

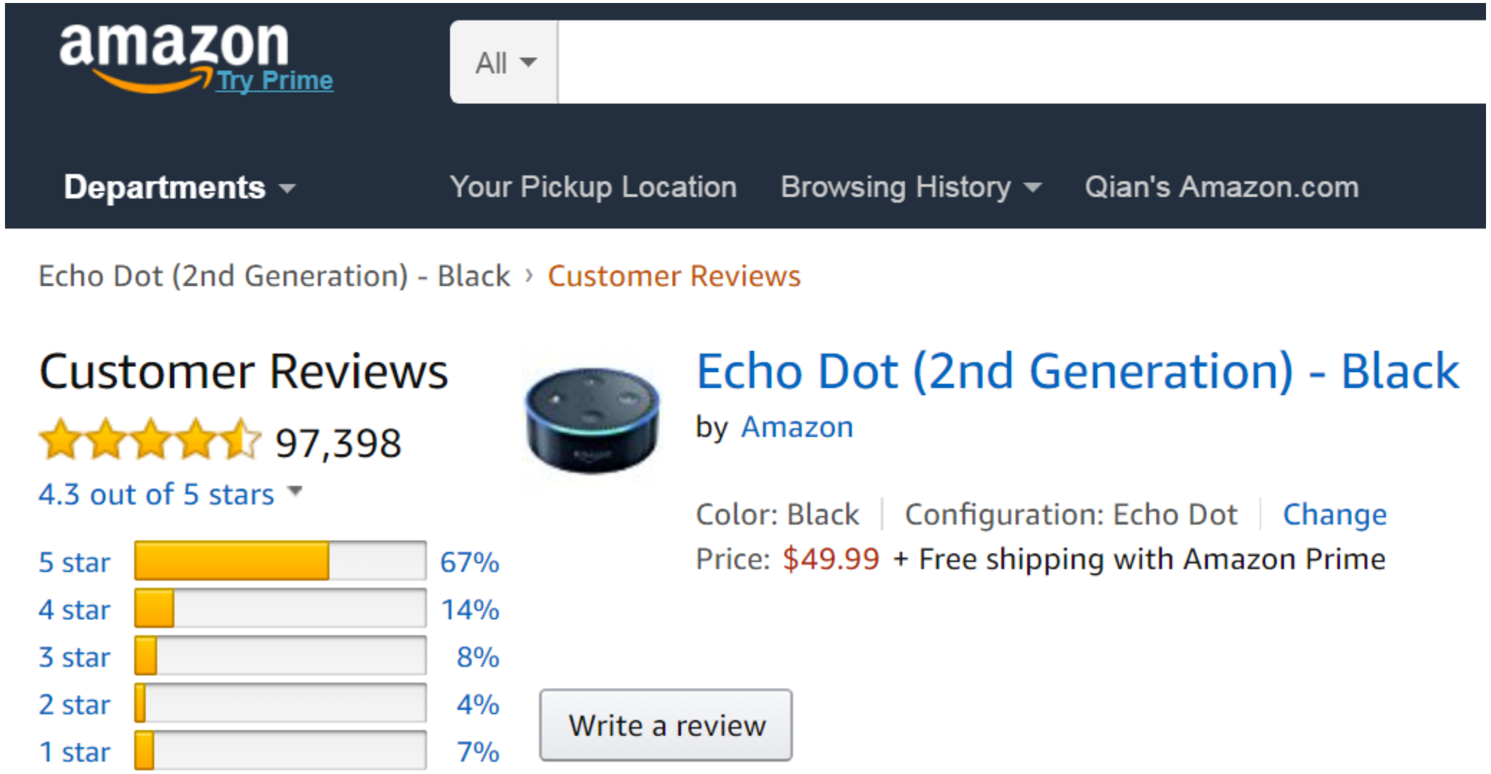
Two-Mode Segmentation is Useful for Ecommerce

Ecommerce

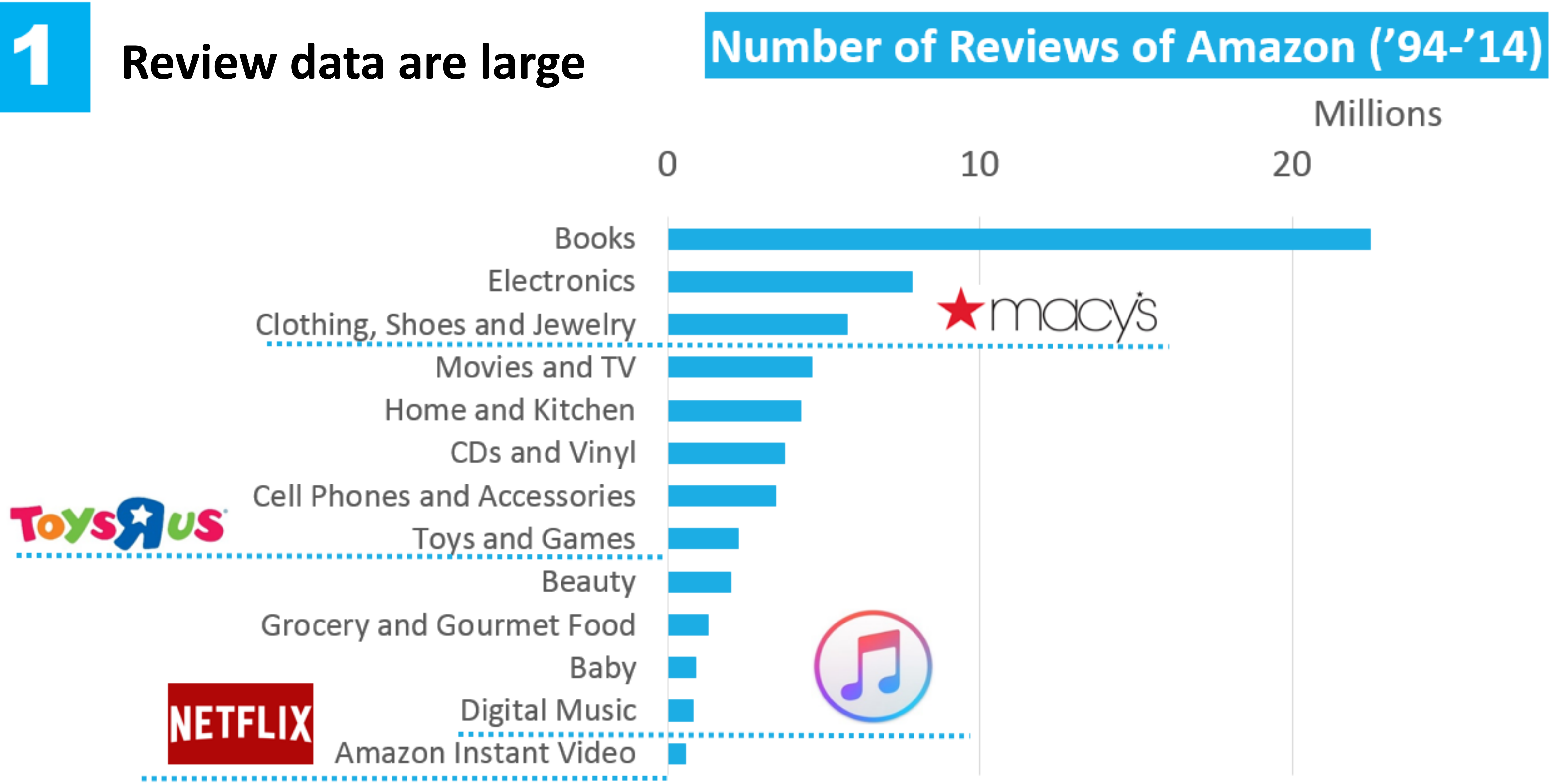
Two-Mode Segmentation Classifies Users and Products

Online Product Reviews are Emerging Data Sources

- **Free**
- **Unsolicited**
- **User-Generated**
- **Actual purchasing behavior and experience**

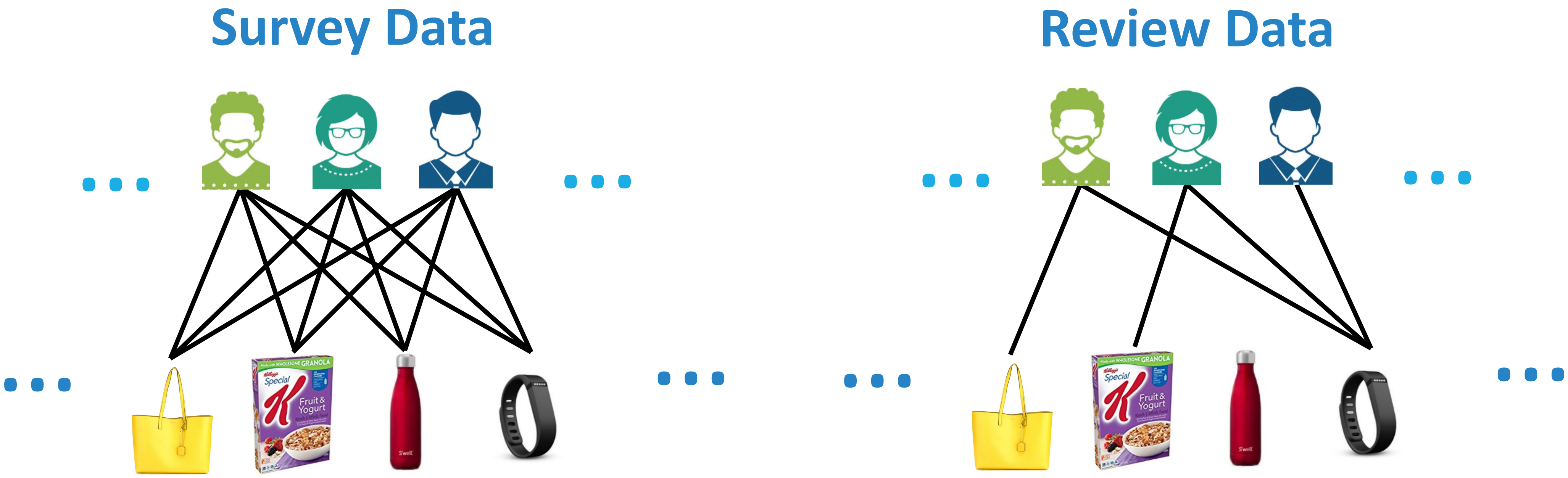


1 Review data are large



Massive Review Data: Modeling and Computational Challenges

2 **Review data are sparse**



Massive Review Data: Modeling and Computational Challenges

3 **Heterogeneous users, some review data are fake**

Source: https://www.cbsnews.com/news/amazon-sues-over-1000-people-for-writing-and-selling-fake-reviews/

Massive Review Data: Modeling and Computational Challenges

3 **Heterogeneous users, some review data are fake**

Source: https://www.nbclosangeles.com/news/local/Fake-Reviews-on-Yelp-Facebook-Google-447796103.html

Massive Review Data: Modeling and Computational Challenges

1 **Large size**

2 **Sparse nature**

3 **User Heterogeneity/ Fake Product Ratings**

4 **Ordinal Review Data**

Comparison with Existing Models

	Large Dataset	Sparse Dataset	User Heteroge-neity	Ordinal Rating
Two-mode Clustering Approaches			?	✓

- **Most focus on continuous, binary and count responses**
  Vichi 2001; DeSarbo et. al 2004; Rocci and Vichi 2008; Govaert and Nadif 2003 and 2010; Pledger and Arnold 2014
- **One recent work focuses on ordinal responses**
  Mathechou et. al, 2016, *psychometrika*

Survey Data

Comparison with Existing Models

		Large Dataset	Sparse Dataset	User Heteroge-neity	Ordinal Rating
Two-mode Clustering Approaches				?	✓
Network Approaches	**Projection-Based**	✓	✓		
	Modularity-Based	✓	✓		
	MDL-Based SBMs	✓	✓		

- Projection-Based Methods: e.g., Newman 2001; Barabasi et al. 2002; Zhou et al. 2007
- Modularity-based Methods: e.g., Girvan and Newman 2004; Newman 2004
- MDL-Based Stochastic Block Models (SBM): e.g., Peixoto 2013; Rosvall and Bergstrom 2008

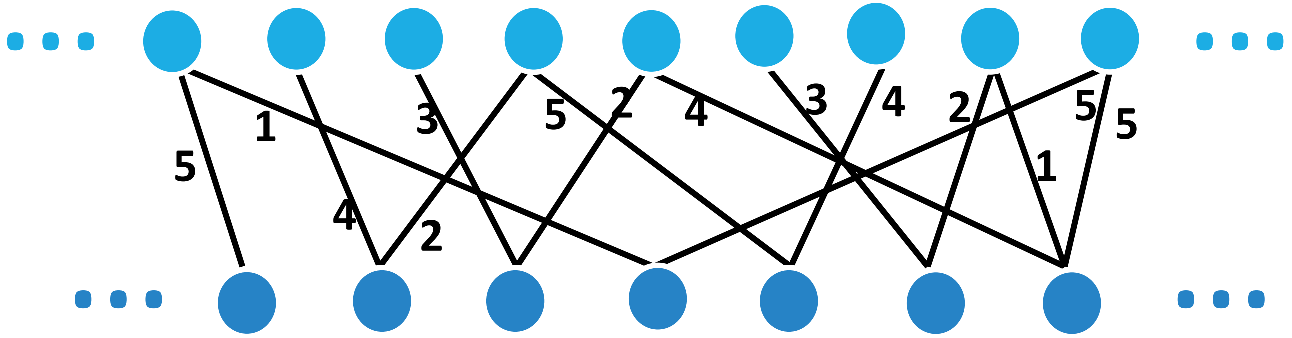
Comparison with Existing Models

		Large Dataset	Sparse Dataset	User Heteroge-neity	Ordinal Rating
Two-mode Clustering Approaches					✓
Network Approaches	**Projection-Based**	✓	✓		
	Modularity-Based	✓	✓		
	MDL-Based SBMs	✓	✓		
	Proposed Model	✓	✓	✓	✓

The Proposed Integrated Two-Mode Network-Based Methodology

Notations

User Nodes: $i \epsilon \{1, \dots, N\}$



Product Nodes: $j \epsilon \{1, \dots, M\}$

Adjacency Matrix: Review or Not

$$\boldsymbol{E} = (E_{ij})_{N \times M} \qquad E_{ij} \in \{0,1\}$$

Ordinal Weight Matrix: Ratings

$$\mathbf{Y} = (Y_{ij})_{N \times M} \qquad Y_{ij} \in \{1, \dots, 5\}$$

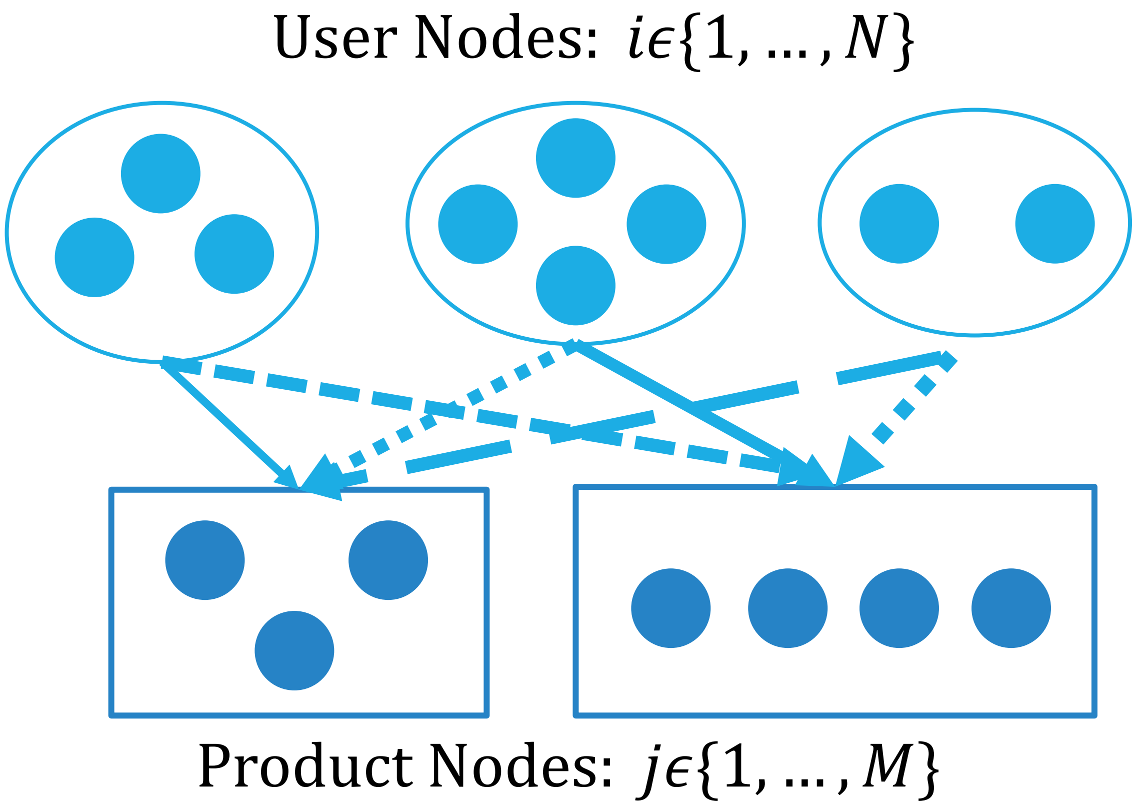
(Wasserman and Faust 1994; Hanneman and Riddle 2005; Grewal et al. 2006; Ransbotham et al. 2012)

The Proposed Integrated Two-Mode Network-Based Methodology

Notations

Membership of User: $Z_i^u \in \{1, \dots, K\}$

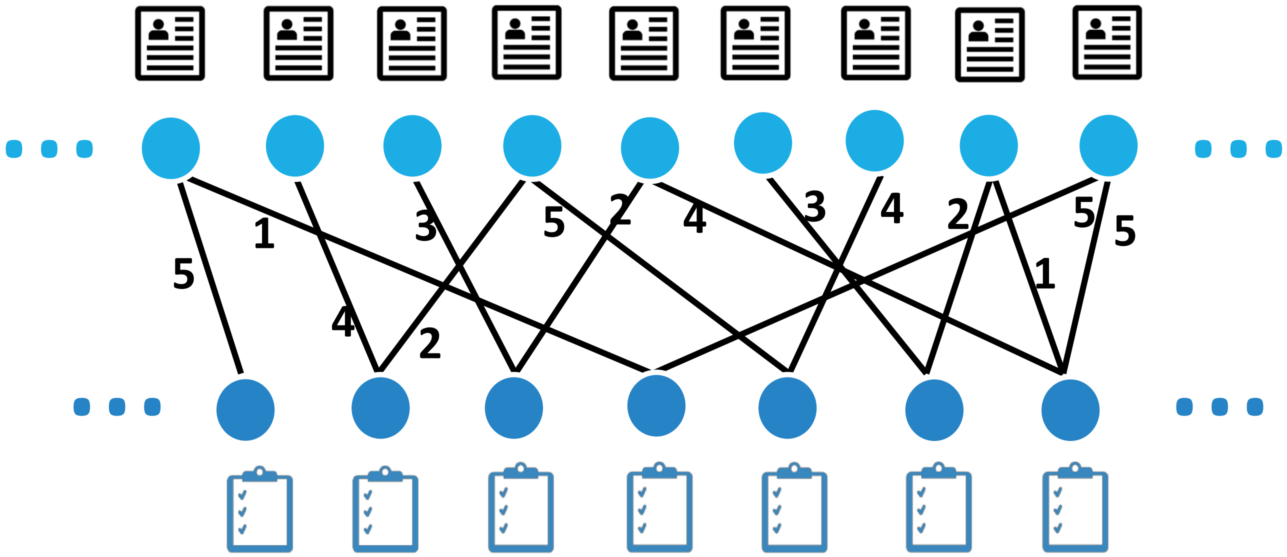
Membership of Product: $Z_j^p \in \{1, \dots, L\}$



The Proposed Integrated Two-Mode Network-Based Methodology

Notations

User Nodes: $i \epsilon \{1, \ldots, N\}$



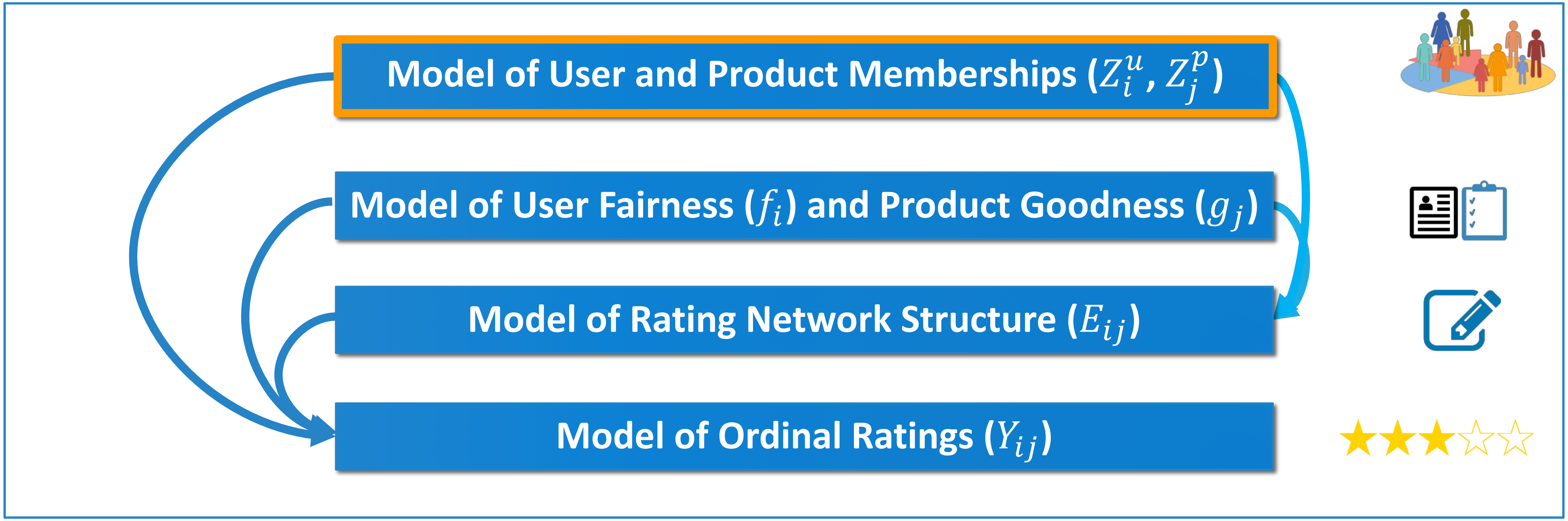
Product Nodes: $j \epsilon \{1, \ldots, M\}$

User Fairness: $f_i \in [0{,}100\%]$

Product Goodness: $g_j \in [0{,}5]$

(Kumar et al. 2016; Rayana and Akoglu 2015; Wang et al. 2011; Kleinberg 1998)

The Proposed Integrated Two-Mode Network-Based Methodology



The Proposed Integrated Two-Mode Network Based Methodology

- ✓ **User Heterogeneity**
- ✓ **Sparse nature**
- ✓ **Large size**
- ✓ **Ordinal Responses**

Model of User and Product Memberships (Z_i^u, Z_j^p)

Model of User Fairness (f_i) and Product Goodness (g_j)

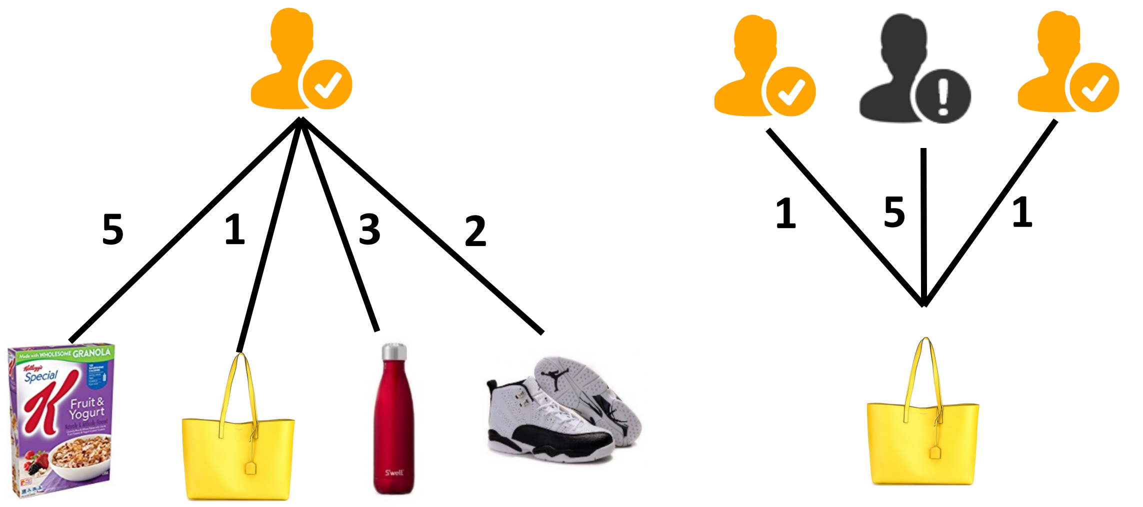
Model of Rating Network Structure (E_{ij})

Model of Ordinal Ratings (Y_{ij})

★★★☆☆

Model of User Fairness and Product Goodness

f_i: **How fair user i is in assigning ratings** 100% 100% 20% 100%

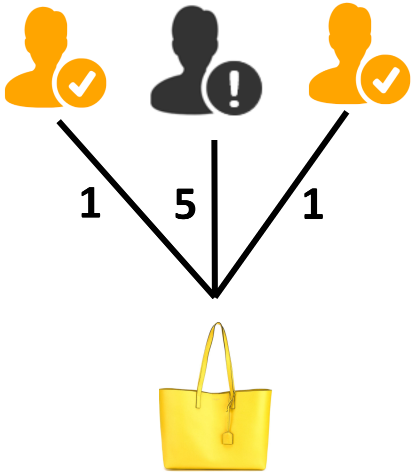


g_j: **The true quality of product j** 5 1 3 2 1

(Kumar et al. 2016; Akoglu et al. 2013; Lim et al., 2010; Wang et al. 2011; Kleinburg 1998)

Model of User Fairness and Product Goodness

$f_i^{(t)}$: **100% 20% 100%**



$g_j^{(t+1)}$: **1**

Goodness

$$g_j^{(t+1)} = \frac{1}{\sum_i E_{ij}} \sum_{i:\, E_{ij}=1} Y_{ij} \times f_i^{(t)}$$

$$g_j \in [0, 5]$$

t: the number of iterations

(Kumar et al. 2016; Akoglu et al. 2013; Lim et al., 2010; Wang et al. 2011; Kleinburg 1998)

Model of User Fairness and Product Goodness

$f_i^{(t+1)}$: **100%**

$g_j^{(t+1)}$: **5** **1** **3** **2**

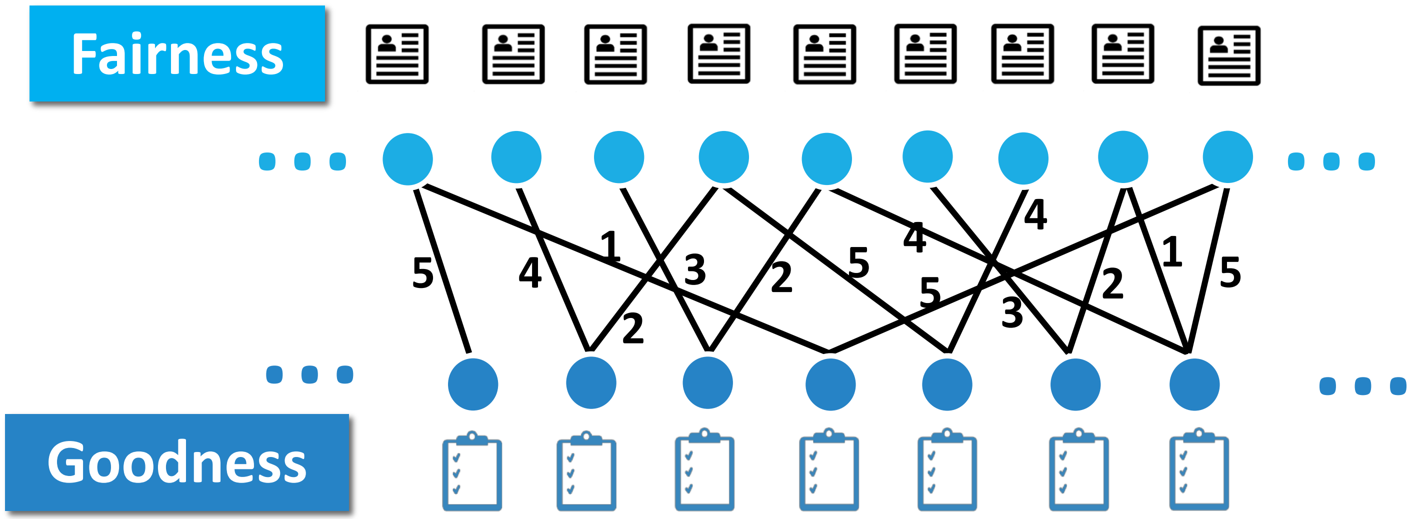
Fairness

$$f_i^{(t+1)} = 1 - \frac{1}{\sum_j E_{ij}} \sum_{j:\, E_{ij}=1} \frac{1}{5} \left| Y_{ij} - g_j^{(t+1)} \right|$$

$$f_i \in [0,100\%]$$

(Kumar et al. 2016; Akoglu et al. 2013; Lim et al., 2010; Wang et al. 2011; Kleinburg 1998)

Model of User Fairness and Product Goodness



$f_i^{(t+1)}$

$g_j^{(t+1)}$

Model of User and Product Memberships

User Membership

Membership of user node i, $i \epsilon \{1, \ldots, N\}$, follows a multinomial distribution:

$$Z_i^u \sim_{iid} \text{Multinomial}(1; \pi_1, \ldots, \pi_K)$$

- K denotes the total number of segments for user nodes
- $Z_i^u \epsilon \{1, \ldots, K\}$, $\boldsymbol{Z^u} = (Z_i^u)$
- $\pi_1, \ldots, \pi_K$ denote user's membership probabilities

Model of User and Product Memberships

Product Membership

Membership of product node j, j$\epsilon\{1, \ldots, M\}$, follows a multinomial distribution:

$$Z_j^p \sim_{iid} \text{Multinomial}(1; \varphi_1, \ldots, \varphi_L)$$

- L denotes the total number of segments for product nodes
- $Z_j^p \epsilon\{1, \ldots, L\}$, $\boldsymbol{Z^p} = (Z_j^p)$
- $\varphi_1, \ldots, \varphi_L$ denote product's membership probabilities

Model of Review Behavior

Whether or not user i reviews product j follows a Bernoulli distribution:

$$E_{ij}|Z_i^u = k, Z_j^p = l \sim_{ind} \text{Bernoulli}\left(\mathrm{P}_{ij}(\boldsymbol{\theta}) = \mathrm{P}_{ij}\left(\theta_{kl}^0, \boldsymbol{\theta}_k^u, \boldsymbol{\theta}_l^p\right)\right)$$

Intercept

Segment-level User Effects

Segment-level Product Effects

$\boldsymbol{X}_i^u$: **User attributes**

- User fairness
- Total # of reviews given by user i
- Helpfulness of user i

$\boldsymbol{X}_j^p$: **Product attributes**

- Product goodness
- total # of reviews received by product j, etc.

$$\mathrm{P}_{ij}\left(\theta_{kl}^0, \boldsymbol{\theta}_k^u, \boldsymbol{\theta}_l^p\right) = \frac{e^{\theta_{kl}^0 + \left(\boldsymbol{X}_i^u\right)' \boldsymbol{\theta}_k^u + \left(\boldsymbol{X}_j^p\right)' \boldsymbol{\theta}_l^p}}{1 + e^{\theta_{kl}^0 + \left(\boldsymbol{X}_i^u\right)' \boldsymbol{\theta}_k^u + \left(\boldsymbol{X}_j^p\right)' \boldsymbol{\theta}_l^p}}$$

Model of Ordinal Review Ratings ★★★☆☆

We model review rating using an ordistic model:

Ordistic Weight Parameters

$$Y_{ij} = y_{ij} | Z_i^u = k, Z_j^p = l, E_{ij}{=}1 \sim_{iid} F_{kl}\left(y_{ij}; \boldsymbol{\mu}, \boldsymbol{w}\right)$$

$$F_{kl}\left(y_{ij}; \boldsymbol{\mu}, \boldsymbol{w}\right) = \frac{\exp(\mu_{kl,y} w_{kl} - \frac{1}{2}\mu_{kl,y}^2)}{\sum_{r=1}^{R} \exp(\mu_{kl,r} w_{kl} - \frac{1}{2}\mu_{kl,r}^2)}$$

$$w_{kl} = \beta_{kl}^0 + (\boldsymbol{X}_i^u)' \boldsymbol{\beta}_k^u + \left(\boldsymbol{X}_j^p\right)' \boldsymbol{\beta}_l^p$$

Ordistic Intercepts

Segment-level User Effects

Segment-level Product Effects

(Rennie and Srebro 2005; Chen et al 2017)

Model Estimation

Likelihood Function

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{E},\boldsymbol{Y}) = \prod_{i}\prod_{j}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\{\pi_k\varphi_l\left[\mathrm{P}_{ij}\left(\theta_{kl}^{0},\boldsymbol{\theta}_{k}^{u},\boldsymbol{\theta}_{l}^{p}\right)F_{kl}\left(y_{ij};\boldsymbol{\mu},\boldsymbol{w}\right)\right]^{E_{ij}}\left[1-\mathrm{P}_{ij}\left(\theta_{kl}^{0},\boldsymbol{\theta}_{k}^{u},\boldsymbol{\theta}_{l}^{p}\right)\right]^{1-E_{ij}}\right\}$$

Maximum Likelihood Estimation (MLE)

$$\max_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{E},\boldsymbol{Y}) = \max_{(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{\varphi})}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{w},\boldsymbol{\pi},\boldsymbol{\varphi}\,|\boldsymbol{E},\boldsymbol{Y})$$

Model Estimation

EM Algorithm?

Let $\boldsymbol{Z} = (\boldsymbol{Z^u}, \boldsymbol{Z^p}) = \left((Z_1^u, \dots, Z_N^u), (Z_1^p, \dots, Z_M^p)\right)$ denote the latent memberships, then the complete-data likelihood can be derived:

$$\mathcal{L}_c(\boldsymbol{\Theta}|\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{Y}) = \prod_i \prod_j \left\{ \pi_{Z_i^u} \varphi_{Z_j^p} \left[\mathrm{P}_{ij}\left(\theta^0_{Z_i^u Z_j^p}, \boldsymbol{\theta}^u_{Z_i^u}, \boldsymbol{\theta}^p_{Z_j^p}\right) F_{Z_i^u Z_j^p}\left(y_{ij}; \boldsymbol{\mu}, \boldsymbol{w}\right)\right]^{E_{ij}} \left[1 - \mathrm{P}_{ij}\left(\theta^0_{Z_i^u Z_j^p}, \boldsymbol{\theta}^u_{Z_i^u}, \boldsymbol{\theta}^p_{Z_j^p}\right)\right]^{1-E_{ij}} \right\}$$

Model Estimation

EM Algorithm?

Iterate between E-step and M-step until convergence:

- **Expectation Step (E-step)**

Compute the expected value of the complete log likelihood function

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}\right) = E_{\boldsymbol{Z}|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}}\left(\log\mathcal{L}_c(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y})\right),$$

which is a lower bound on $\log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{E},\boldsymbol{Y}\right)$

- **Maximization Step (M-step)**

Maximize the lower bound to obtain $\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}\right)$

Model Estimation

EM Algorithm?

However, in E-step, the lower bound is intratable:

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}\right) = E_{\boldsymbol{Z}|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}}\left(\log\mathcal{L}_c(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y})\right)$$

Maximum Likelihood Estimation (EM Algorithm) is Impossible:
A critical computing challenge for complex and large-scale models!

Model Estimation

Solution: Variational EM

- **Variational E-step**

  1) Derive a tractable evidence lower bound (ELBO) on $\log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)} \middle| \boldsymbol{E}, \boldsymbol{Y}\right)$ through

     ✓ Jensen's inequality

     ✓ A variational family of distributions $q(\boldsymbol{Z}; \boldsymbol{u}, \boldsymbol{v})$ (e.g., the mean-field family)

     to approximate $p\left(\boldsymbol{Z} \middle| \boldsymbol{E}, \boldsymbol{Y}; \boldsymbol{\Theta}^{(t)}\right)$

  2) Maximize the ELBO w.r.t. variational parameters

$$(\boldsymbol{u}, \boldsymbol{v})^{(t+1)} = \underset{(\boldsymbol{u}, \boldsymbol{v})}{\operatorname{argmax}} \operatorname{ELBO}\left(\boldsymbol{\Theta}^{(t)}, \boldsymbol{u}, \boldsymbol{v}\right)$$

- **M-step**

  Maximize the ELBO updated from E-step to obtain $\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \operatorname{ELBO}\left(\boldsymbol{\Theta}, \boldsymbol{u}^{(t+1)}, \boldsymbol{v}^{(t+1)}\right)$

Model Estimation

More Details about Variational EM: Specifying the form of $q(\boldsymbol{Z}; \boldsymbol{u}, \boldsymbol{v})$

- Use a mean-field family which is fully factorized by multinomial distributions
- Variational parameters $\boldsymbol{u}_i = (u_{i,1}, \dots, u_{i,K})$ and $\boldsymbol{v}_j = (v_{j,1}, \dots, v_{j,L})$

$$q(\boldsymbol{Z}; \boldsymbol{u}, \boldsymbol{v}) = q(\boldsymbol{Z}^{\boldsymbol{u}}; \boldsymbol{u})q(\boldsymbol{Z}^{\boldsymbol{p}}; \boldsymbol{v}) = \prod_{i=1}^{N} P(Z_i^u; \boldsymbol{u}_i) \prod_{j=1}^{M} P\left(Z_j^p;\ \boldsymbol{v}_j\right)$$

$$Z_i^u \sim_{iid} \mathrm{Multinomial}(1; \boldsymbol{u}_i)$$

$$Z_j^p \sim_{iid} \mathrm{Multinomial}(1; \boldsymbol{v}_j)$$

Model Estimation

More Details about Variational EM: Jensen's inequality and ELBO

$$\log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{E},\boldsymbol{Y}\right) = \log E_q\left(\frac{\mathcal{L}_c\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y}\right)}{q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})}\right) \geq \boxed{E_q\left(\log\mathcal{L}_c\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y}\right)\right) - E_q\left(\log q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})\right)}$$

$$= \mathrm{ELBO}\left(\boldsymbol{\Theta}^{(t)},\boldsymbol{u},\boldsymbol{v}\right)$$

$$\mathrm{ELBO}\left(\boldsymbol{\Theta}^{(t)},\boldsymbol{u},\boldsymbol{v}\right) = \log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{E},\boldsymbol{Y}\right) - \mathrm{KL}\left(q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})||p\left(\boldsymbol{Z}\middle|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}\right)\right)$$

$$\mathrm{KL}\left(q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})||p\left(\boldsymbol{Z}\middle|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}\right)\right) \equiv \int q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})\frac{q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})}{p\left(\boldsymbol{Z}\middle|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}\right)}d\boldsymbol{Z}$$

Model Estimation

More Details about Variational EM: Jensen's inequality and ELBO

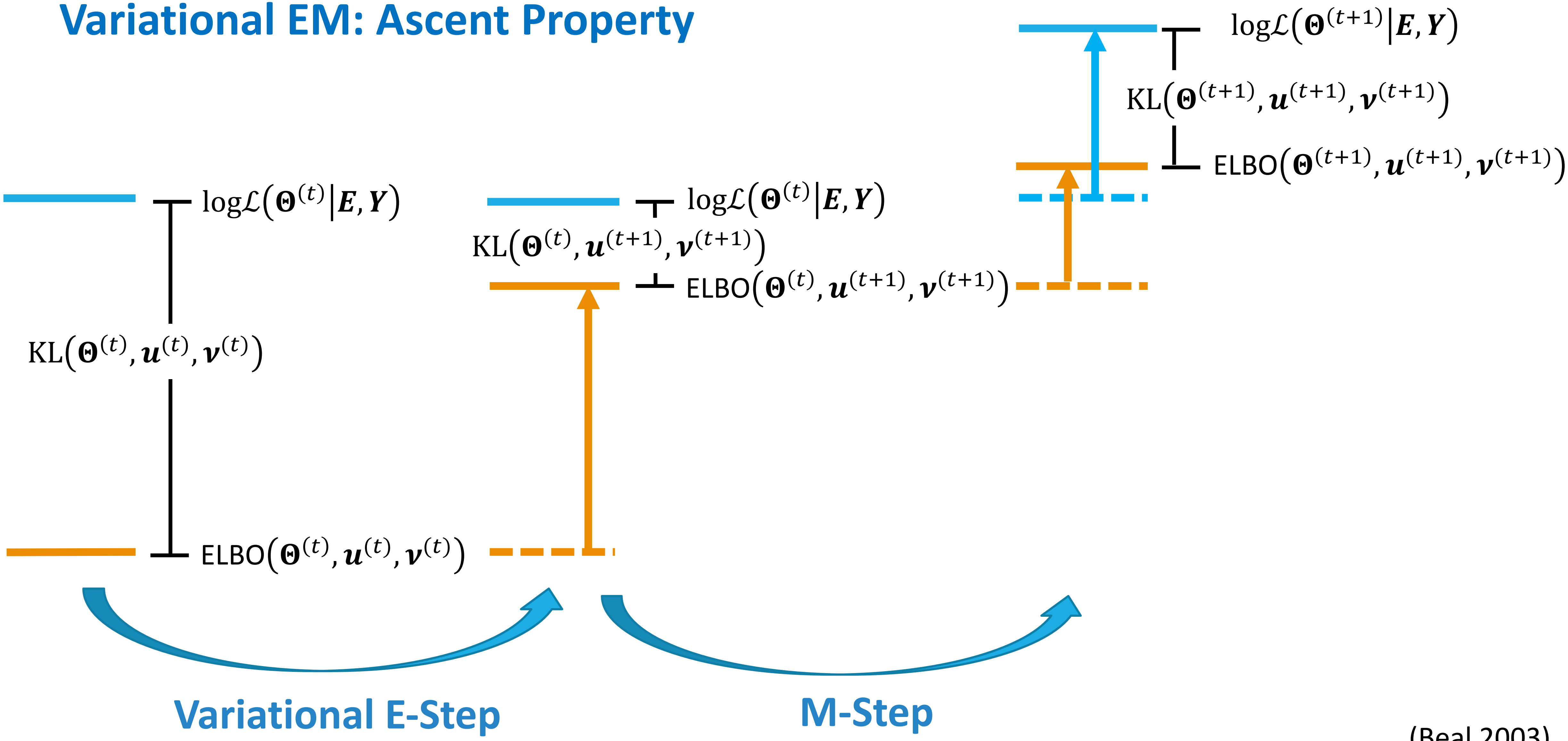
$$\log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{E},\boldsymbol{Y}\right) = \log E_q\left(\frac{\mathcal{L}_c\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y}\right)}{q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})}\right) \geq \boxed{E_q\left(\log\mathcal{L}_c\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{Z},\boldsymbol{E},\boldsymbol{Y}\right)\right) - E_q\left(\log q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})\right)}$$

$$= \mathrm{ELBO}\left(\boldsymbol{\Theta}^{(t)},\boldsymbol{u},\boldsymbol{v}\right)$$

$$\mathrm{ELBO}\left(\boldsymbol{\Theta}^{(t)},\boldsymbol{u},\boldsymbol{v}\right) = \log\mathcal{L}\left(\boldsymbol{\Theta}^{(t)}\middle|\boldsymbol{E},\boldsymbol{Y}\right) - \mathrm{KL}\left(q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})||p\left(\boldsymbol{Z}\middle|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}\right)\right)$$

$$\max_{(\boldsymbol{u},\boldsymbol{v})}\mathrm{ELBO}\left(\boldsymbol{\Theta}^{(t)},\boldsymbol{u},\boldsymbol{v}\right) \quad \Longleftrightarrow \quad \min_{(\boldsymbol{u},\boldsymbol{v})}\mathrm{KL}\left(q(\boldsymbol{Z};\boldsymbol{u},\boldsymbol{v})||p\left(\boldsymbol{Z}\middle|\boldsymbol{E},\boldsymbol{Y};\boldsymbol{\Theta}^{(t)}\right)\right)$$

Variational EM: Ascent Property



(Beal 2003)

Model Estimation

Further Speed Boosting: Stochastic Optimization with Mini-Batches

- It is computationally wasteful when a single update of parameters $\boldsymbol{\Theta}$ require an update of all individual-level latent parameters $(\boldsymbol{u}, \boldsymbol{v})$

- Solution:
  - Use a randomly selected subset – a mini-batch – for each iteration
  - **Variational E-step**: update variational parameters $(\boldsymbol{u}, \boldsymbol{v})$ only for the individual within the mini-batch
  - **M-step**: update $\boldsymbol{\Theta}$ using only updated variational parameters within the mini-batch

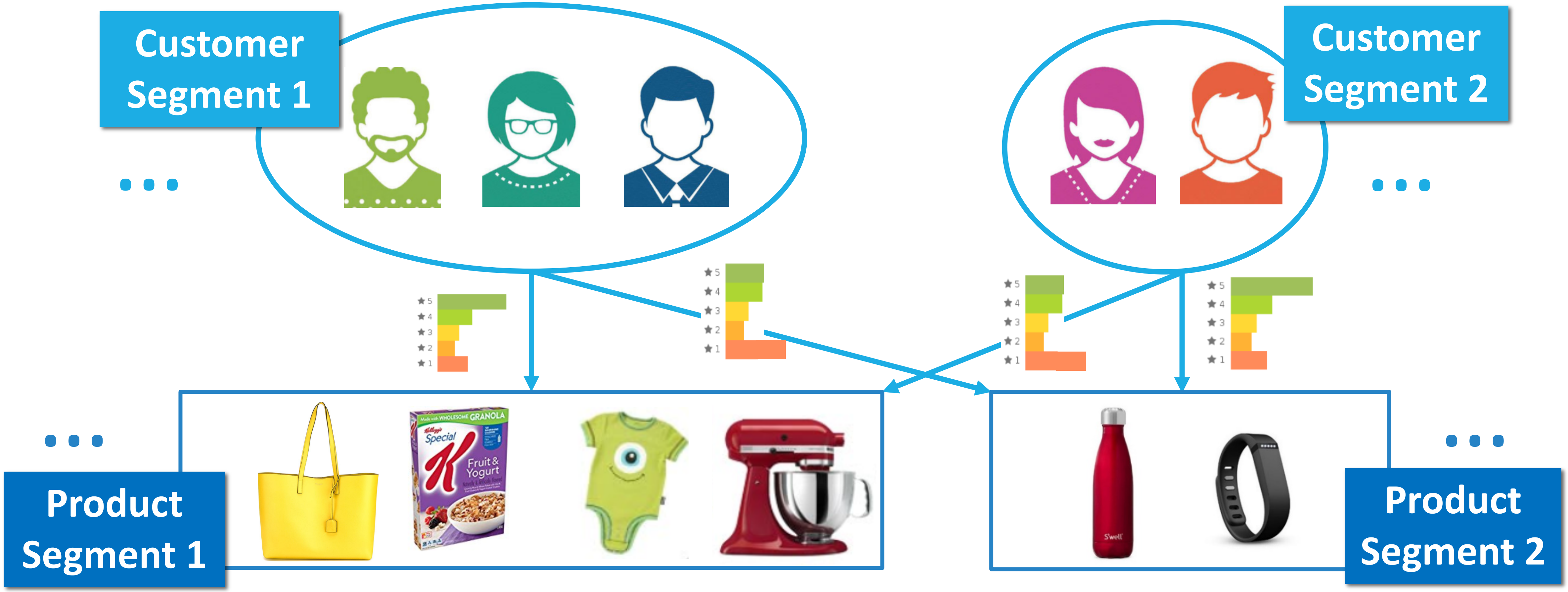
Model Estimation

Segmentation $\widehat{Z_i^u}$ and $\widehat{Z_j^p}$

Model Selection Criteria

- To determine the optimal number of segments (K, L)
- Integrated Classification Likelihood (ICL) criterion (Biernacki et al., 2000)

Two-Mode Segmentation Classifies Users and Products



Simulation

We plan to show our model outperforms the benchmarks in terms of:

- ❑ Review data recovery
- ❑ Segment recovery
- ❑ Parameter recovery
- ❑ Computational efficiency

Application: Amazon Review Data

Dataset: Amazon 5-core Review Data

Category: Clothing, Shoes and Jewelry - Fashion

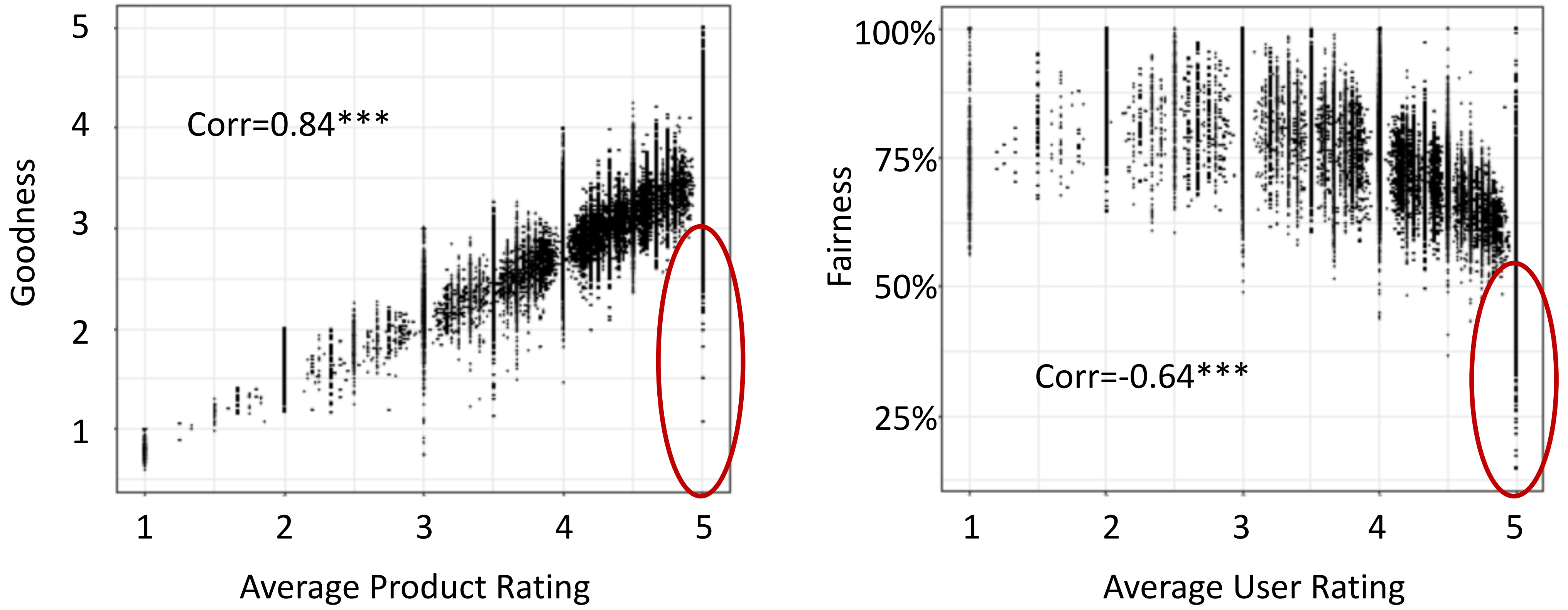
Time: Year 2014

Size:

- 19500 products
- 27500 users
- 101000 reviews

(McAuley et al. 2015; He and McAuley 2016)

Computed Fairness and Goodness



***p-value<0.01

Conclusions

- **Provide an efficient segmentation methodology for Ecommerce companies**
  - Incorporate machine-learned user fairness and product goodness
  - Design a stochastic variational EM algorithm for computational efficiency

- **Extend latent-class based segmentation methods to network modeling**
- **Advance the applications of network analysis in marketing science**

Thank You!
Q&A