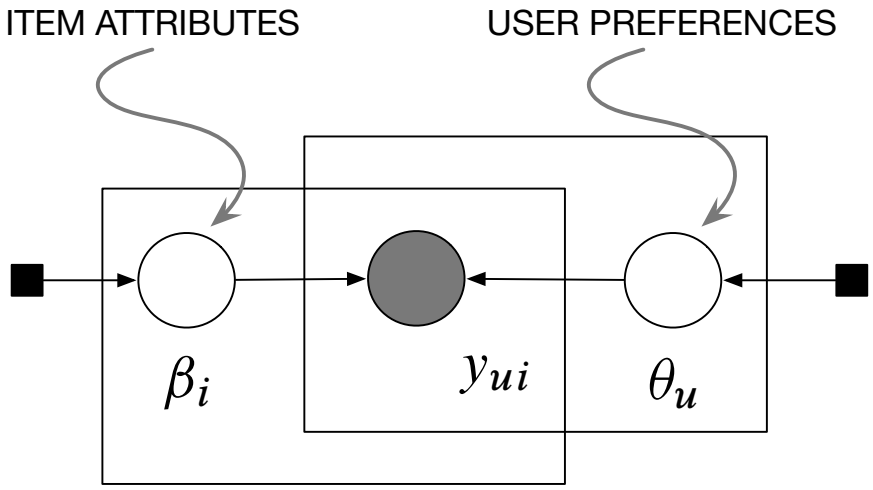
Causal Inference and Recommendation Systems

David M. Blei
Departments of Computer Science and Statistics
Data Science Institute
Columbia University

- In: Ratings or click data
- Out: A system that provides recommendations

Classical solution: Matrix factorization



$$\theta_u \sim f(\cdot)$$
$$\beta_i \sim g(\cdot)$$
$$y_{ui} \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- Users have *latent preferences* θ_u.
  Items have *latent attributes* β_i.
- A users rating comes from an exponential family,
  e.g., Poisson, Bernoulli, Gaussian.

Koren et al., *Computer* 2009

Classical solution: Matrix factorization

"All Things Airplane"

Flying Solo
Crew-Only 787 Flight Is Approved By FAA
All Aboard Rescued After Plane Skids Into Water at Bali Airport
Investigators Begin to Test Other Parts On the 787
American and US Airways May Announce a Merger This Week

"Personal Finance"

In Hard Economy for All Ages Older Isn't Better It's Brutal
Younger Generations Lag Parents in Wealth-Building
Fast-Growing Brokerage Firm Often Tangles With Regulators
The Five Stages of Retirement Planning Angst
Signs That It's Time for a New Broker

Example components from *New York Times* click data

- Condition on click data $\mathbf{y}$ to estimate the posterior $p(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{y})$.
  - estimates of user preferences and item attributes
- Form predictions with the posterior predictive distribution.
- This is the backbone of many recommender methods.

Salakhutdinov and Mnih, *International Conference on Machine Learning* 2008
Gopalan et al., *Uncertainty in Artificial Intelligence* 2015

Today

- Two related ideas that build on matrix factorization.
  - **Causal inference for recommendation** (new; feedback wanted!)
  - **Modeling user exposure in recommendation**
- Both ideas use the theory around causal inference.
  - Key idea: The **exposure model** (aka the assignment mechanism)

Liang et al., ACM Conference on Recommendation Systems, 2016

Causal Inference for Recommendation

(with Dawen Liang and Laurent Charlin)

Causal inference and recommendation

- **Causal inference**
  - Expose a unit to a treatment
  - What would happen if a patient received a treatment?
  - Biased data from observational studies
- **Recommendation**
  - Expose a user to an item
  - What would happen if a user was recommended an item?
  - Biased data from logged user behavior

Causal recommendation

- Treat the inference of **user preferences** as a causal inference
- We want preferences θ_i to answer (from attributes β_j):

  What if user i saw item j? How would user i like it?

- This leads to a different approach from classical matrix factorization.

A potential outcomes perspective

- Consider a single user i and fix the attributes of all items β_j.
- Assume a potential outcomes model

$$
\begin{aligned}
a_{ij} &\sim p_{ij} \\
y_{ij}(0) &\sim \delta_0(\cdot) \\
y_{ij}(1) &\sim \text{exp-fam}(\theta_i^\top \beta_j)
\end{aligned}
$$

- Usual inference about θ_i is causal if we have **ignorability**,

$$
a_{ij} \perp\!\!\!\perp (y_{ij}(0), y_{ij}(1))
$$

(This is typically the main issue behind observational data.)

Rubin, 1974, 1975
Imbens and Rubin, 2015

A potential outcomes perspective

- Clearly, $a_{ij} \perp\!\!\!\perp y_{ij}(0)$, because $y_{ij}(0)$ is known.
- But is $a_{ij} \perp\!\!\!\perp y_{ij}(1)$?
  - Probably not. Users seek out items that they will like.
- This biases our estimates of the users' preferences.

Marlin and Zemel, 2009
Schnabel et al., 2016

Causal recommendation

- In ratings data we observe
  - which items each user saw a_{ij} (binary).
  - how they rated those items, $y_{ij}(1)$ when $a_{ij} = 1$
  - and $y(0)$, trivially
- Fit an **exposure model**, how users find items.
  (This is also called the *assignment mechanism*.)
- Fit preferences by using the exposure model to correct for self-selection

Causal recommendation

- Fit the exposure model using Poisson factorization,

$$a_{ij} \sim \text{Poisson}(\pi_i^\top \lambda_j)$$

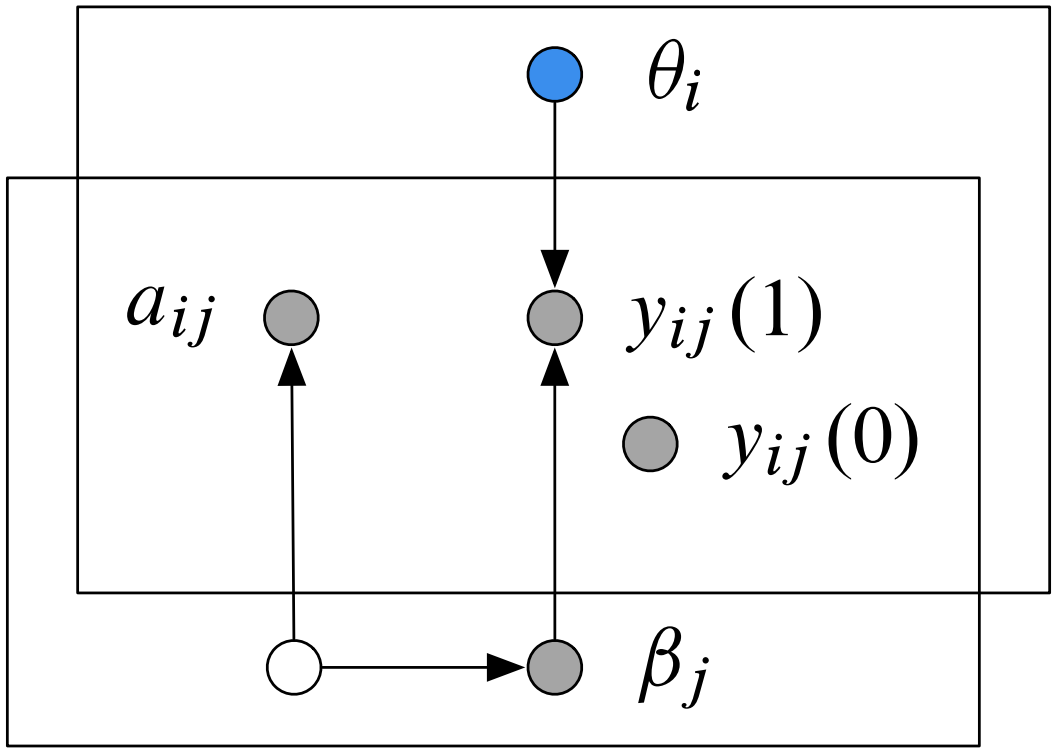
- Use **inverse propensity weighting** to estimate user preferences θ_i,

$$\hat{\theta} = \text{MAP}(\{a_{ij}, y_{ij}(a_{ij}), 1/p(a_{ij} = 1)\})$$

- Intuition: if a user sees and likes a difficult-to-find movie then we upweight its influence on her preferences.

Why propensity weighting works

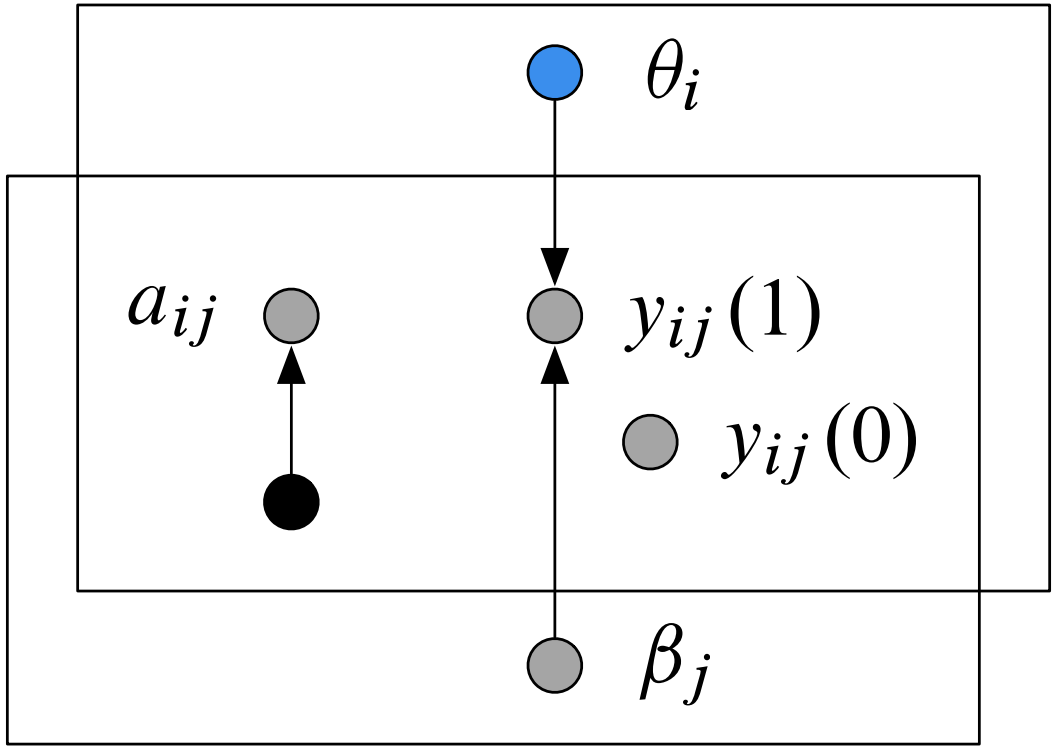
Suppose our data come from this model:



Then inference about preferences θ_i are not causal inferences.

Why propensity weighting works

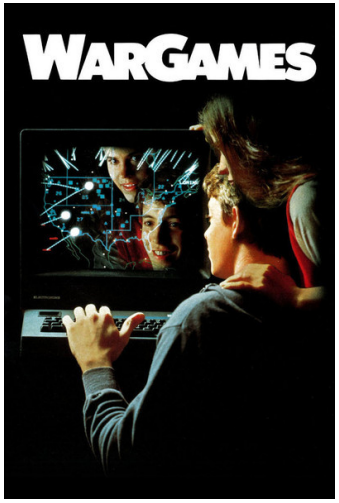
We want data to come from this model:

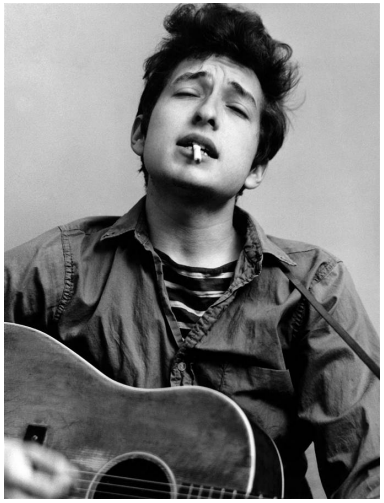


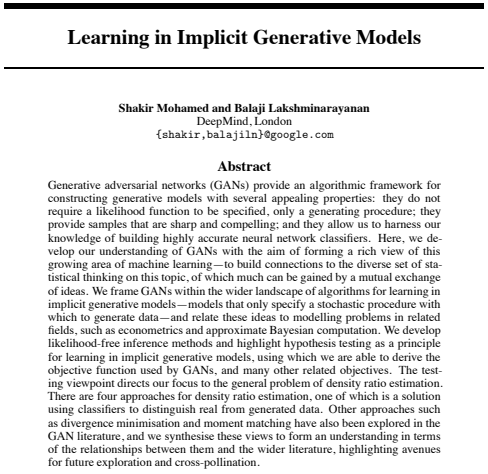
The exposure probabilities are known and independent of β_i.

Solution: Importance sampling, divide each sample by $1/p(a_{ij})$.

Data sets







Learning in Implicit Generative Models

Shakir Mohamed and Balaji Lakshminarayanan
DeepMind, London
{shakir,balajiln}@google.com

Abstract

Generative adversarial networks (GANs) provide an algorithmic framework for constructing generative models with several appealing properties: they do not require a likelihood function to be specified, only a generating procedure; they provide samples that are sharp and compelling; and they allow us to harness our knowledge of building highly accurate neural network classifiers. Here, we develop our understanding of GANs with the aim of forming a rich view of this growing area of machine learning—to build connections to the diverse set of statistical thinking on this topic, of which much can be gained by a mutual exchange of ideas. We frame GANs within the wider landscape of algorithms for learning in implicit generative models—models that only specify a stochastic procedure with which to generate data—and relate these ideas to modelling problems in related fields, such as econometrics and approximate Bayesian computation. We develop likelihood-free inference methods and highlight hypothesis testing as a principle for learning in implicit generative models, using which we are able to derive the objective function used by GANs, and many other related objectives. The testing viewpoint directs our focus to the general problem of density ratio estimation. There are four approaches for density ratio estimation, one of which is a solution using classifiers to distinguish real from generated data. Other approaches such as divergence minimisation and moment matching have also been explored in the GAN literature, and we synthesise these views to form an understanding in terms of the relationships between them and the wider literature, highlighting avenues for future exploration and cross-pollination.

- **MovieLens 1M, 10M**: Users rating movies
- **Yahoo R3**: Users listening to music
- **Arxiv**: Users downloading PDFs (exposure = seeing the abstract)

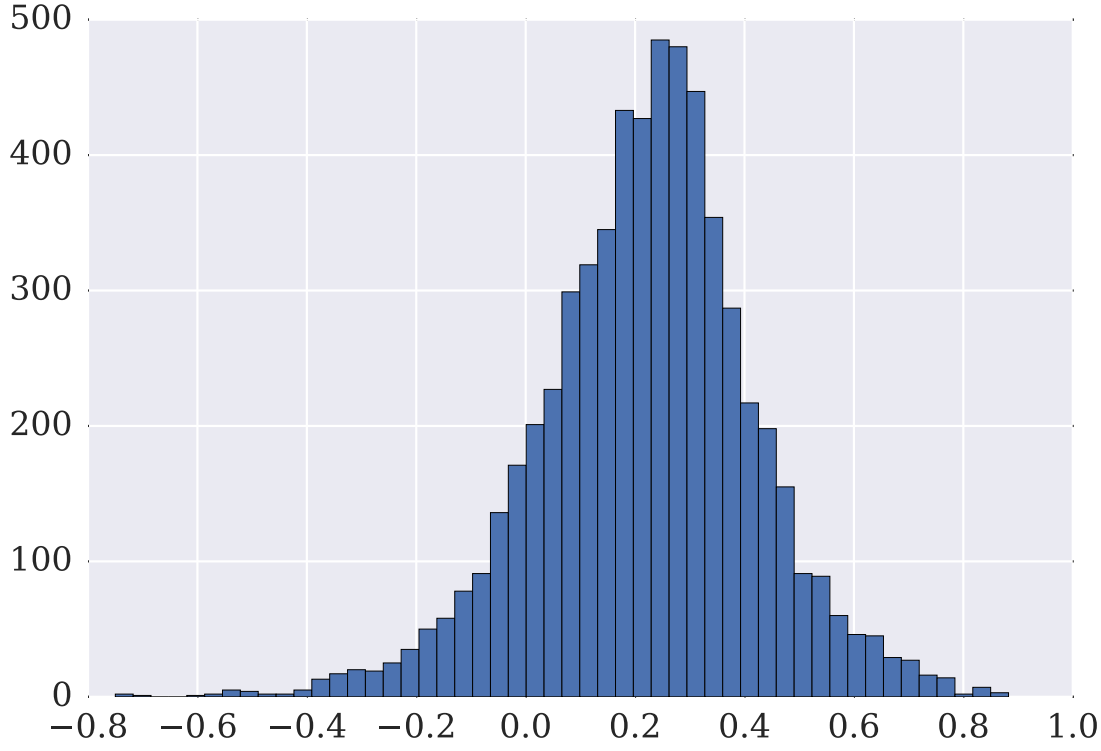
How good is the exposure model?

Model	ML-1M	ML-10M	Yahoo-R3	ArXiv
Popularity	-1.39	-1.64	-1.81	-3.83
Poisson factorization	**-0.97**	**-1.08**	**-1.58**	**-2.71**

Tran et al., Arxiv, 2016

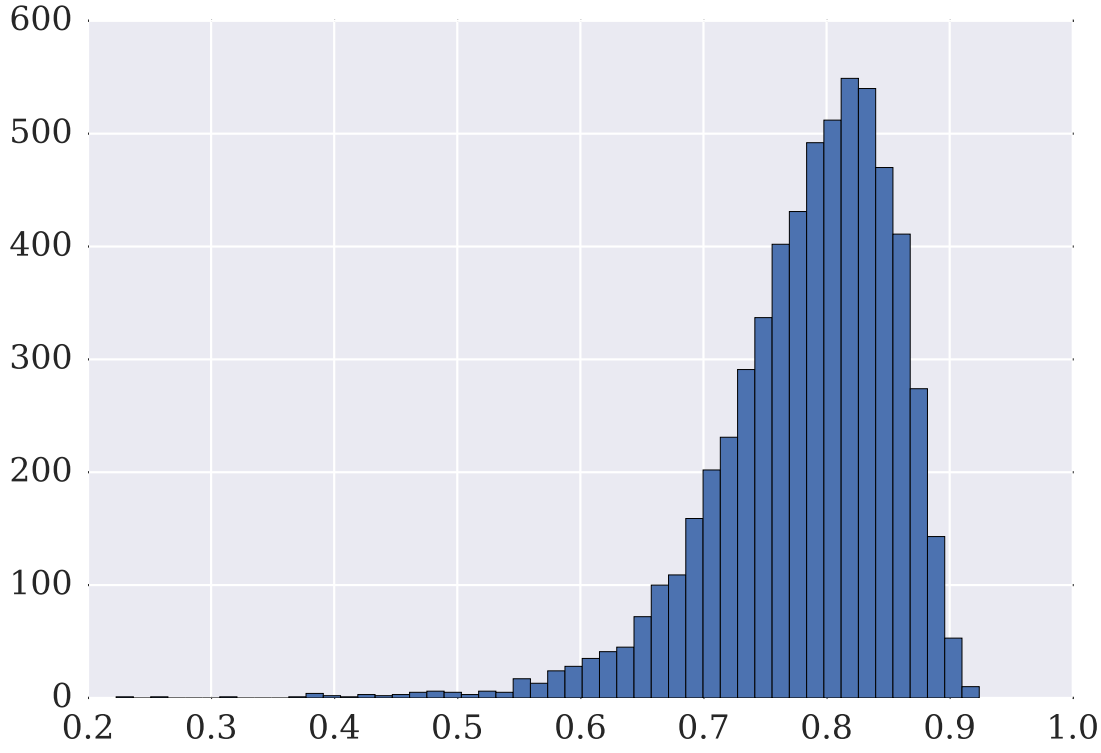
Golapan et al., *Uncertainty in Artificial Intelligence*, 2015

Self-selection on ML 1M



Correlation of the log odds of assignment to the observed rating

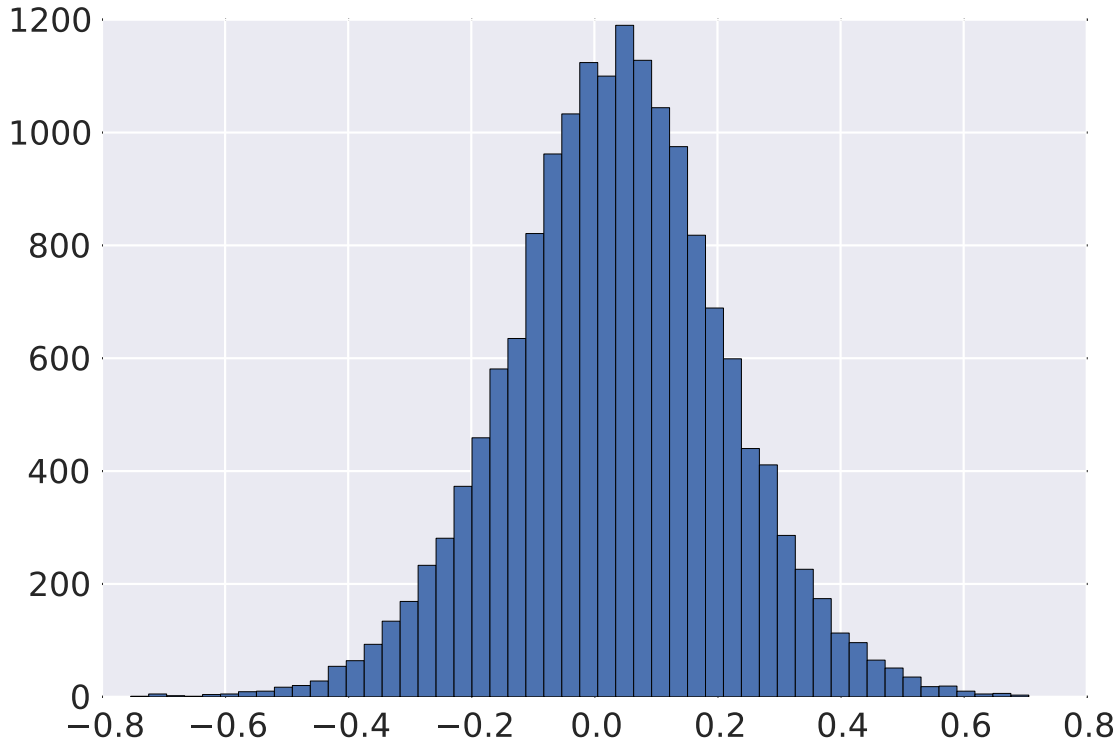
Self-selection on ML 1M



Correlation of the log odds of assignment to the predicted rating

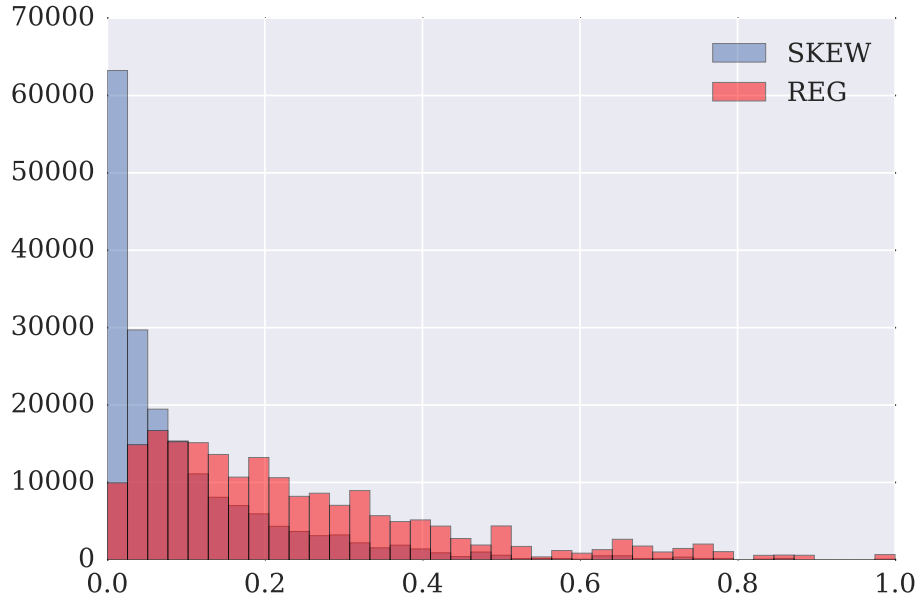
(Workshop confession: What does this mean?)

Different confounding on the Arxiv



Correlation of the log odds of assignment to the predicted rating

Aside: Two test sets



Histogram of popularities of the standard and skewed test set

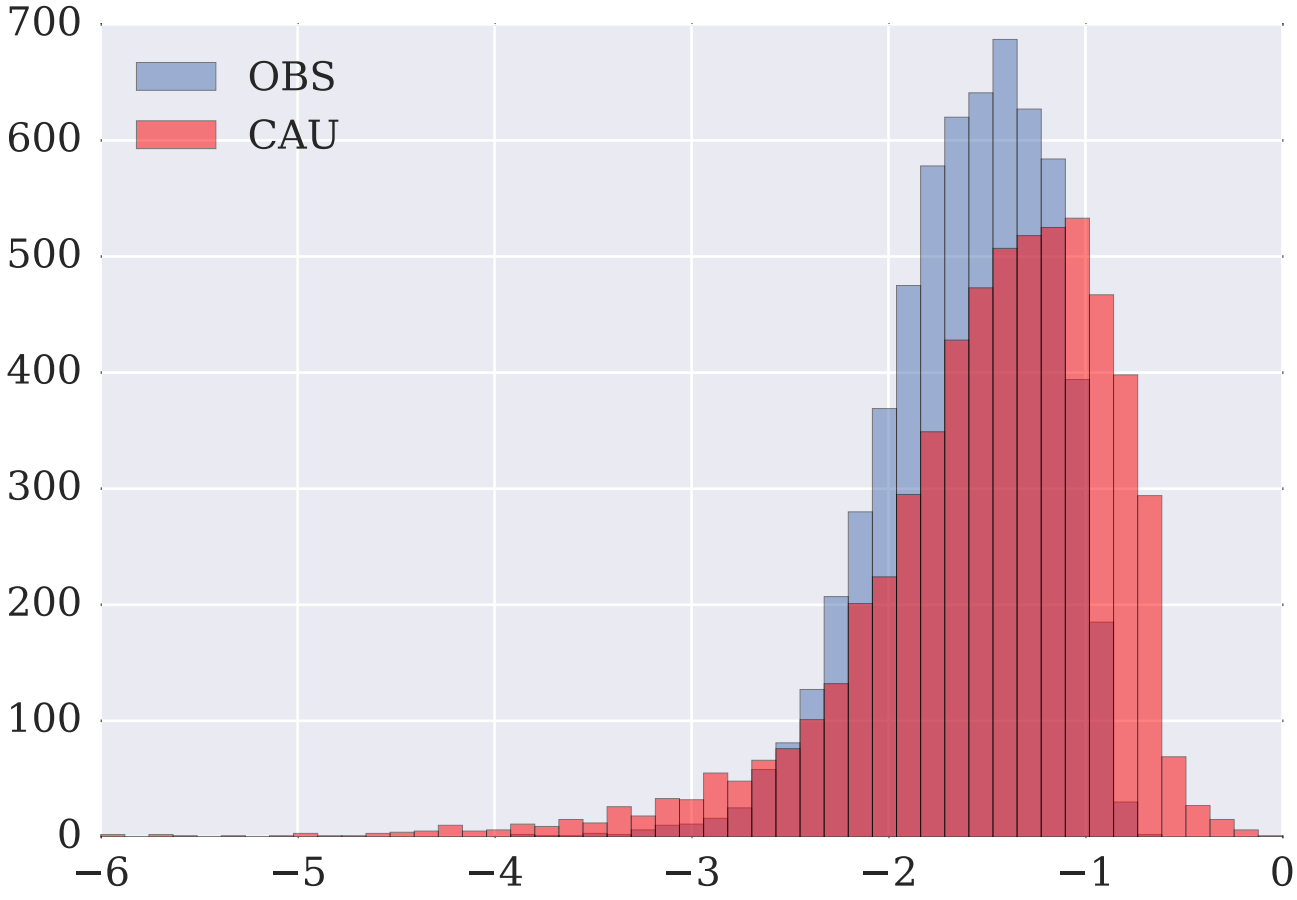
- A test set is biased by the same process that created the training set.
- We created “skewed” test sets, where any item has equal probability.
- Main idea: Test set distribution is different from the training distribution.

Causal inference improves predictions

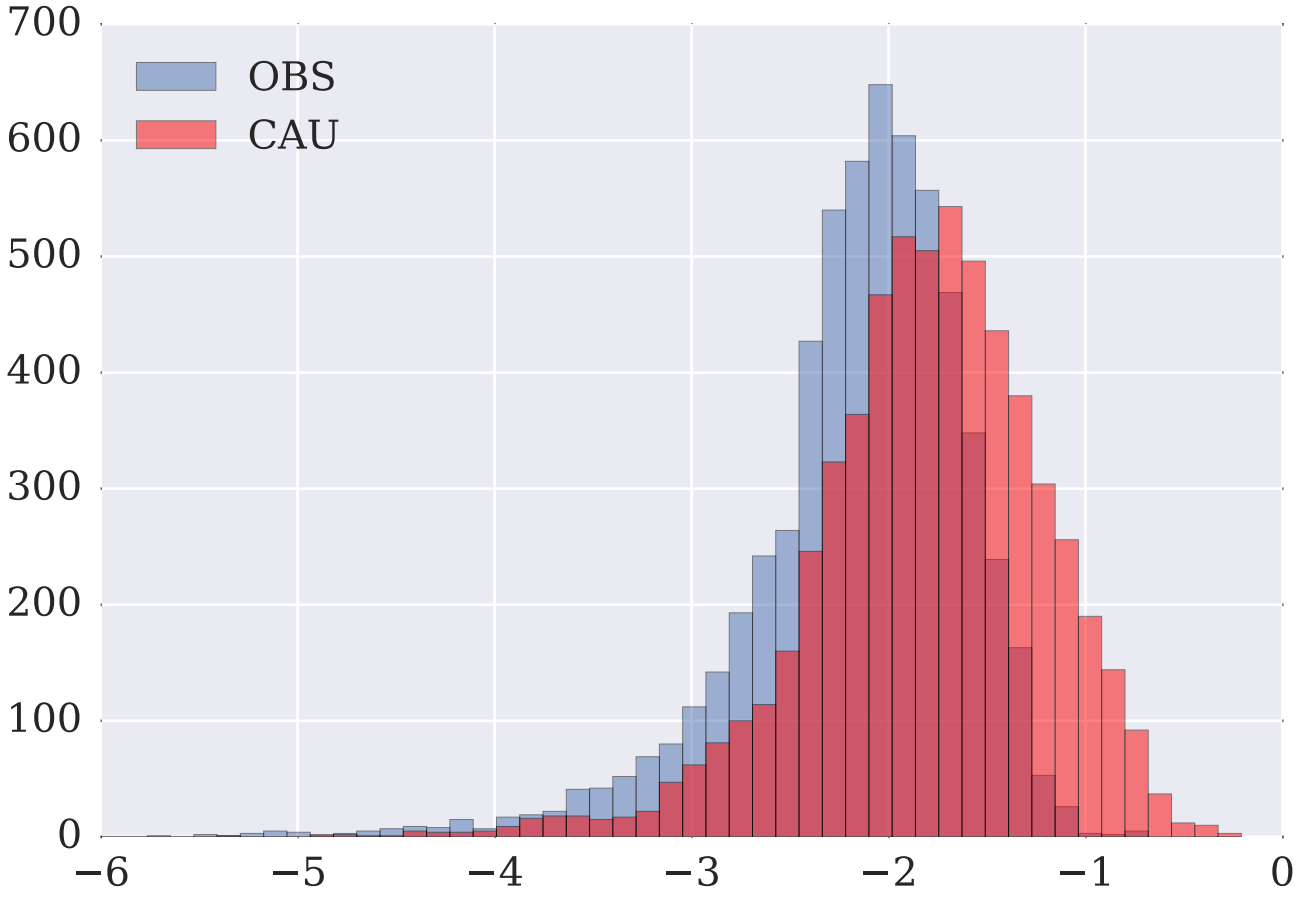
		ML-1M		ML-10M		Yahoo-R3		ArXiv	
		REG	SKEW	REG	SKEW	REG	SKEW	REG	SKEW
Pop	OBS	-1.50	-2.07	-1.62	-2.59	-1.58	-1.75	-1.61	-1.65
	CAU	-1.61	-1.95	-1.67	**-1.89**	-1.51	-1.56	-1.74	-1.76
PF	OBS	-1.50	-2.07	-1.62	-2.59	-1.58	-1.75	-1.61	-1.65
	CAU	**-1.48**	**-1.84**	**-1.51**	-1.96	**-1.49**	**-1.55**	**-1.60**	**-1.62**

Predictive log tail probability (bigger is better)

ML-1M (regular test set)

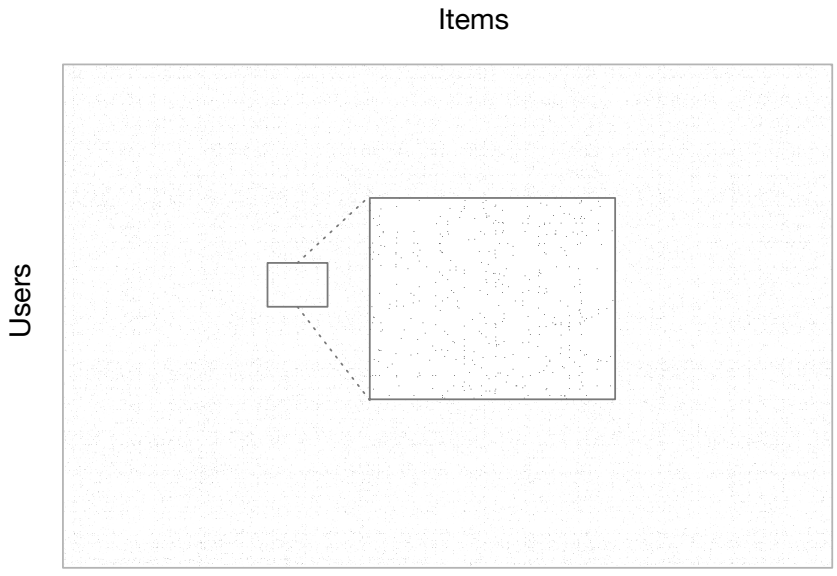


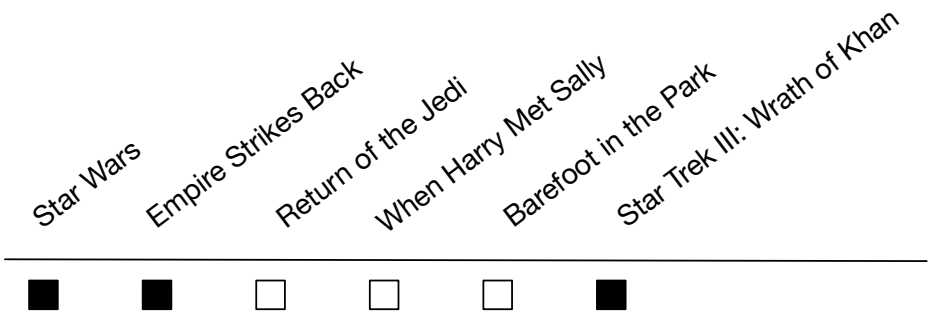
ML-1M (skewed test set)



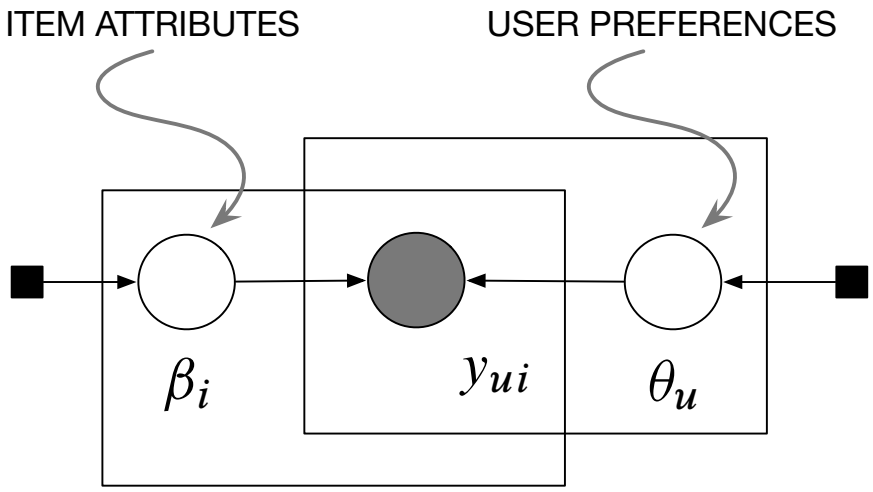
Modeling User Exposure in Recommendation

(with Dawen Liang, Laurent Charlin, and James McInerney)





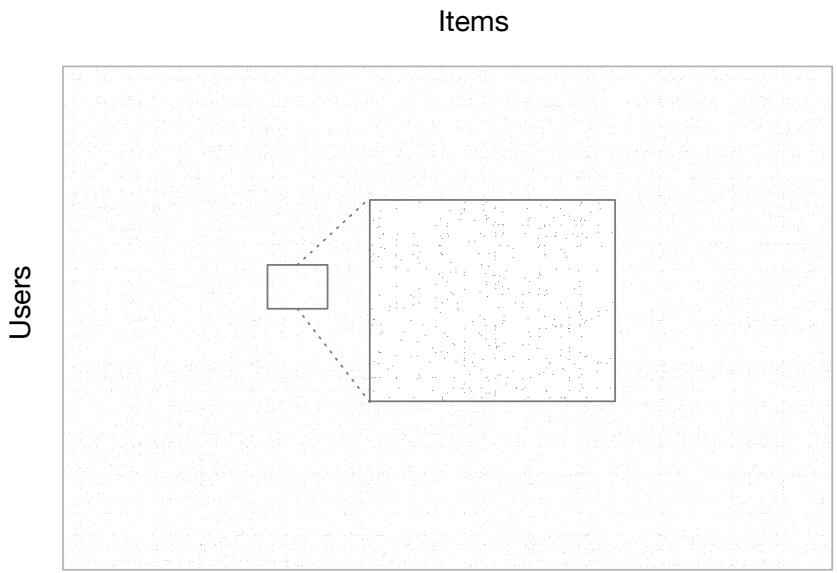
- **Implicit data** is about users interacting with items
  - clicks, likes, purchases, ...
- Less information than explicit data (e.g. ratings), but more prevalent
- Challenge: We only observe positive signal.

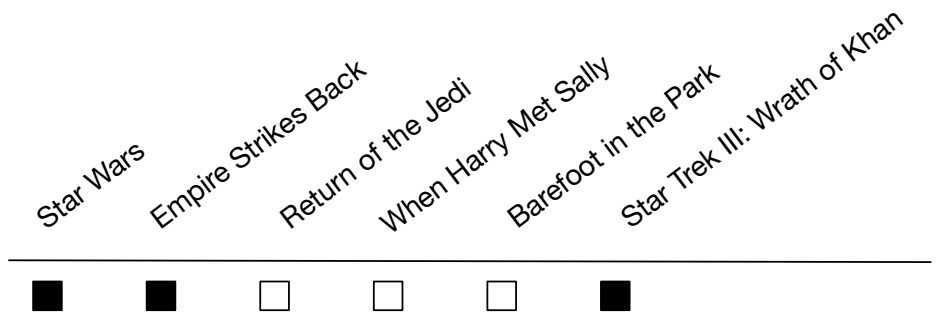


$$\theta_u \sim f(\cdot)$$
$$\beta_i \sim g(\cdot)$$
$$y_{ui} \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- Classical solution: **matrix factorization**
- Users are associated with *latent preferences* θ_u.
  Items are associated with *latent attributes* β_i.
- Whether a user clicked on an item comes from an exponential family, e.g., Poisson, Bernoulli, Gaussian.

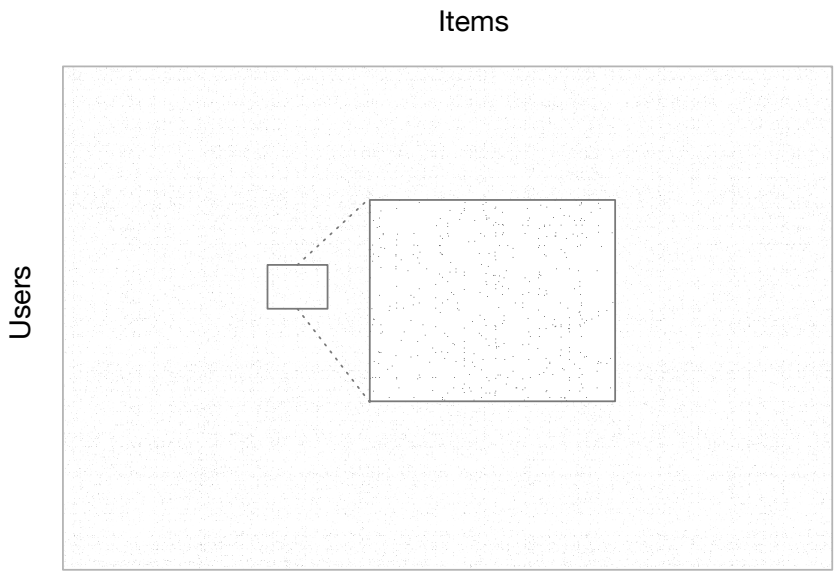
Koren et al., *Computer* 2009

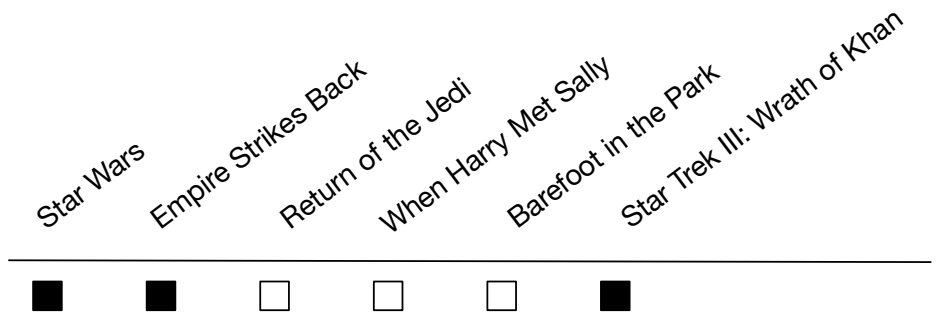




- Tacit assumption—*each user considered clicking on each item.*
- I.e., every zero indicates that a user decided not to click on an item.
- This is false! Users are only **exposed** to a subset of items.

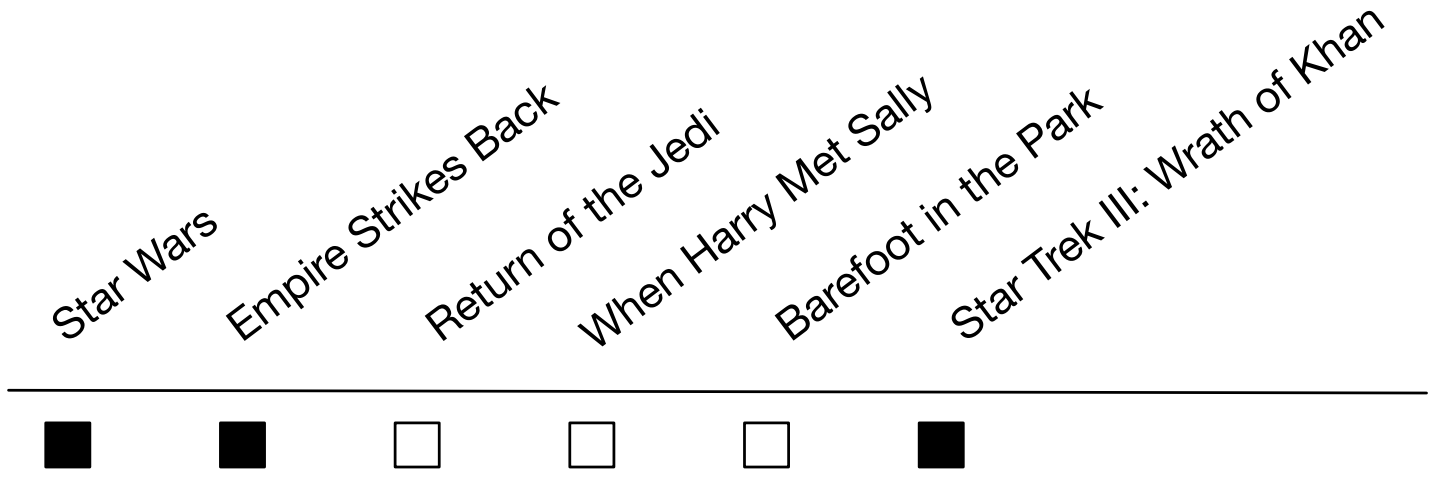
Hu et al., *International Conference on Data Mining* 2008
Rendle et al., *Uncertainty in Artificial Intelligence* 2009



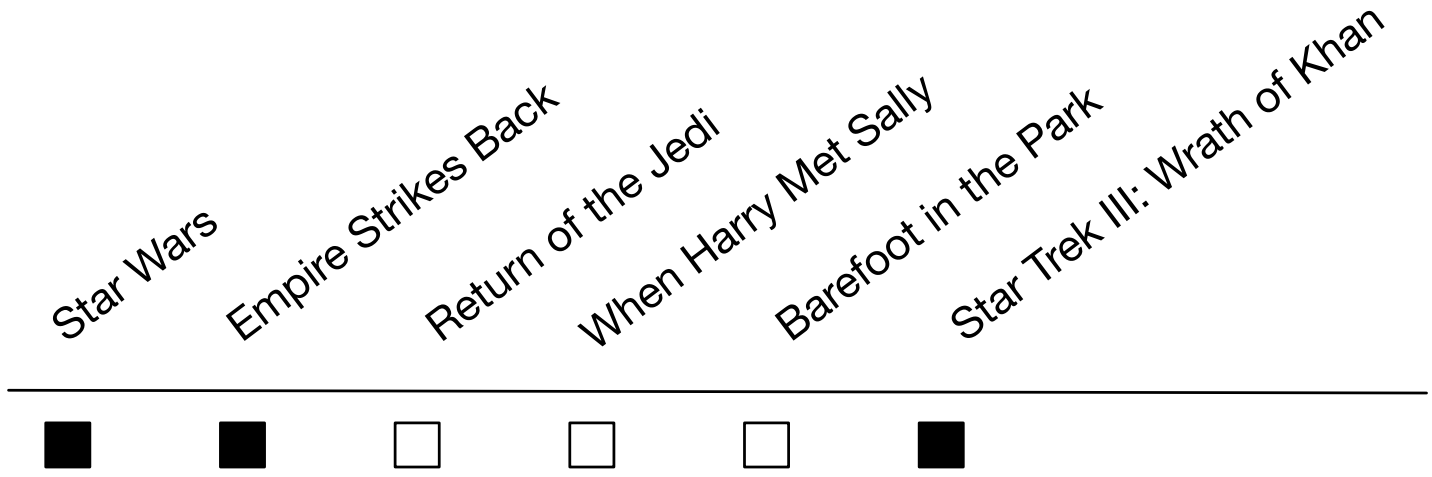


- This issue biases estimates of preferences and attributes.
- In practice, researchers correct for this bias by downweighting the 0's.
- This gives state-of-the-art results, but it is ad-hoc.

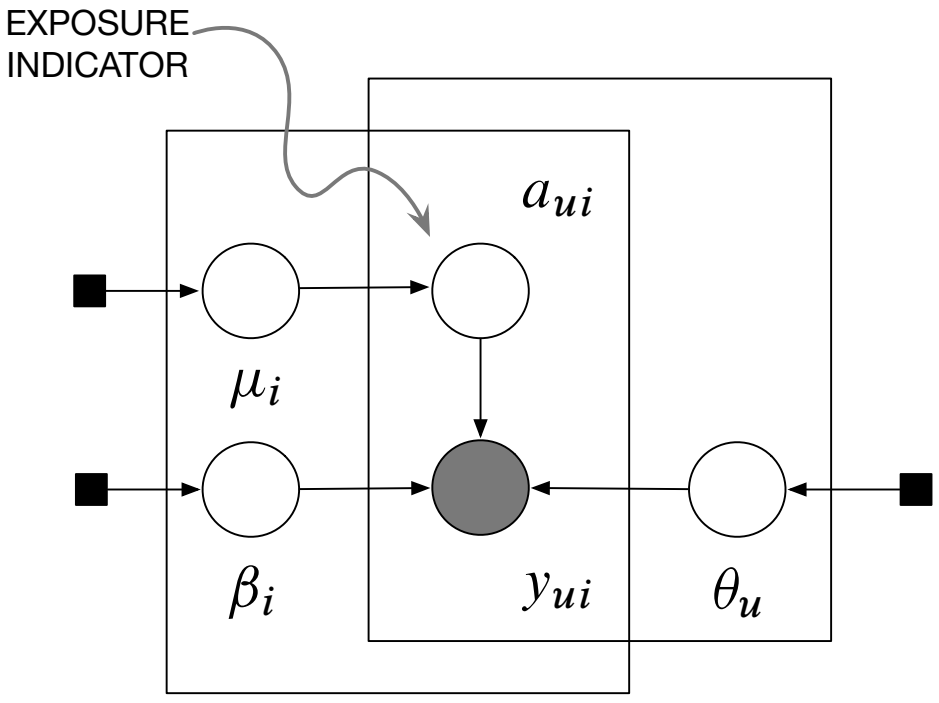
Hu et al., *International Conference on Data Mining* 2008
Rendle et al., *Uncertainty in Artificial Intelligence* 2009



- Idea: Model clicks from a two-stage process.
- First, a user is *exposed* to an item.
- Then, she decides whether to click on it.
- (Inspired by, but different from, potential outcomes.)

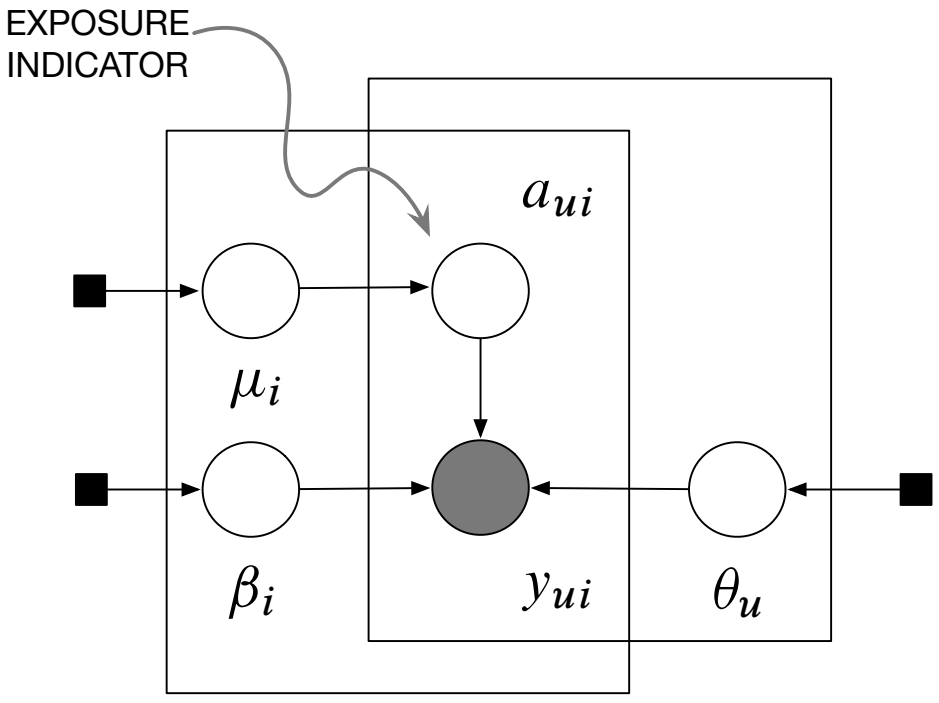


- Challenge: Exposure is (partly) latent.
- Clicks mean that the user was exposed.
- But non-clicks can arise in two ways:
  - The user didn't see the item.
  - The user chose not to click on it.



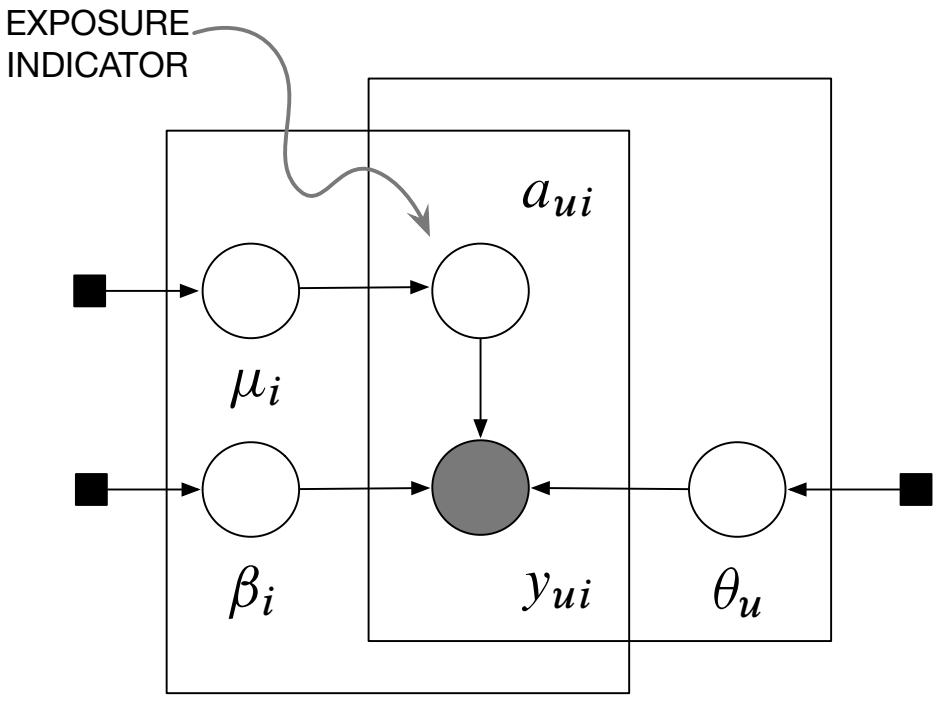
$$a_{ui} \sim \text{Bern}(\mu_i)$$
$$y_{ui} \,|\, a_{ui} = 0 \sim \delta_0$$
$$y_{ui} \,|\, a_{ui} = 1 \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- **Exposure MF**: an exposure model and a click model
- First generate whether the user was exposed to the item.
- Conditional on the exposure, generate whether the user clicked.



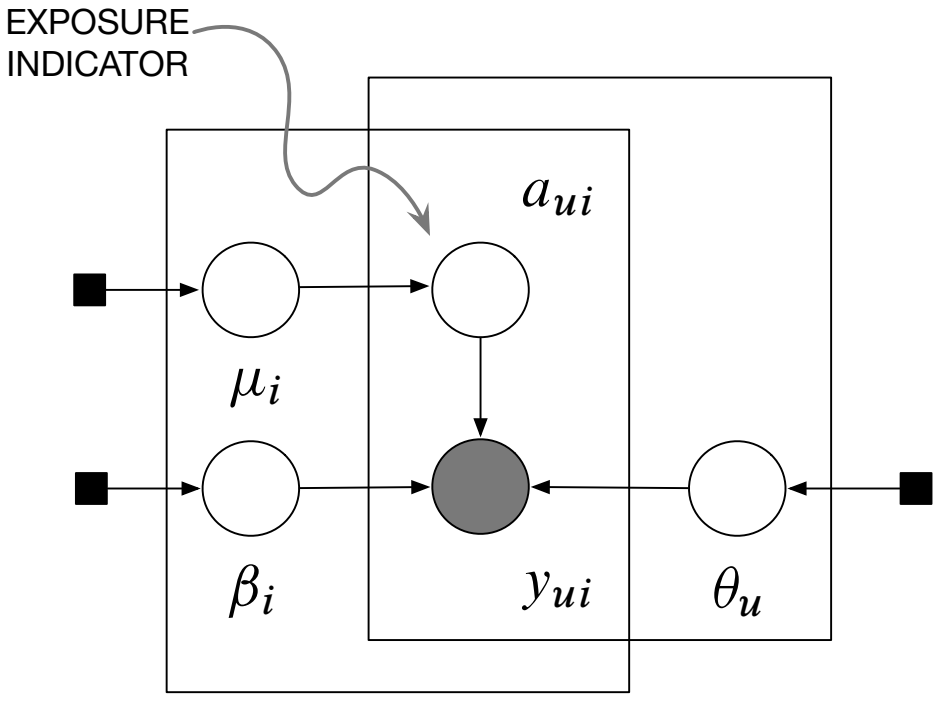
$$a_{ui} \sim \text{Bern}(\mu_i)$$
$$y_{ui} \,|\, a_{ui} = 0 \sim \delta_0$$
$$y_{ui} \,|\, a_{ui} = 1 \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- Popularity-based exposure: $a_{ui} \sim \text{Bernoulli}(\mu_i)$
- Location-based exposure: $a_{ui} \sim \text{Bernoulli}(\sigma(x_u^\top \ell_i))$
- Others: social-network exposure, author-preference exposure, ...



$$a_{ui} \sim \text{Bern}(\mu_i)$$
$$y_{ui} \,|\, a_{ui} = 0 \sim \delta_0$$
$$y_{ui} \,|\, a_{ui} = 1 \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- Given data $\mathbf{y}$ we find MAP estimates of the parameters $\mathcal{M} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\mu}\}$
- Recall that the exposure indicator is *sometimes* latent.
- We use expectation maximization.

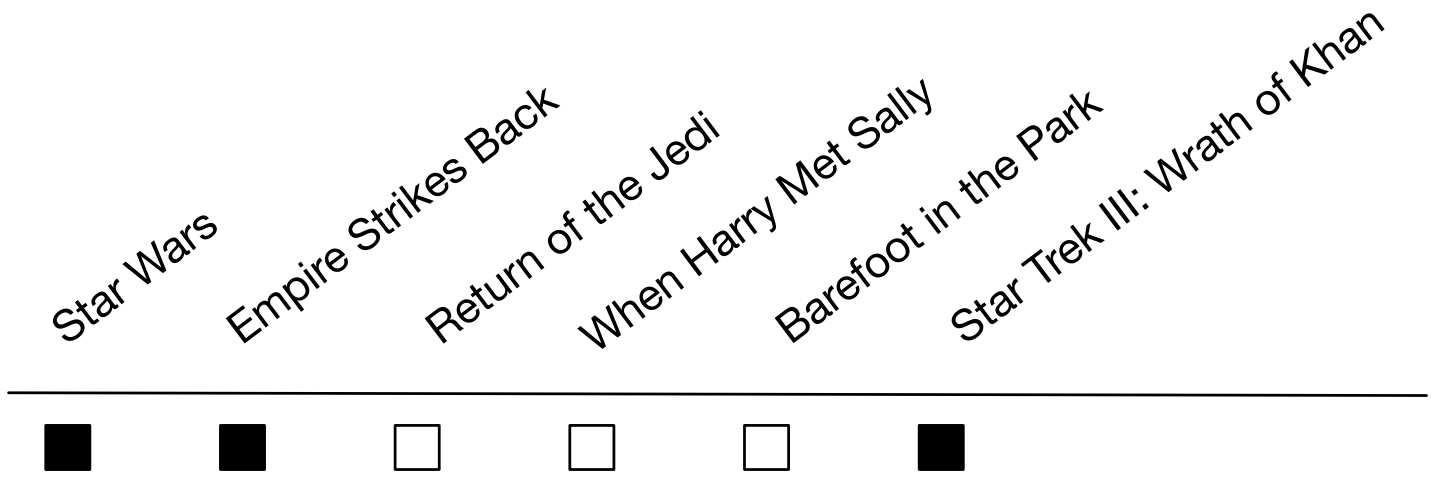


$$a_{ui} \sim \text{Bern}(\mu_i)$$
$$y_{ui} \,|\, a_{ui} = 0 \sim \delta_0$$
$$y_{ui} \,|\, a_{ui} = 1 \sim \text{exp-fam}\left(\theta_u^\top \beta_i\right)$$

- *E-step*: Estimate the posterior of each unclicked exposure,

$$\pi_{ui} = p(a_{ui} = 1 \,|\, y_{ui} = 0, \mathcal{M}).$$

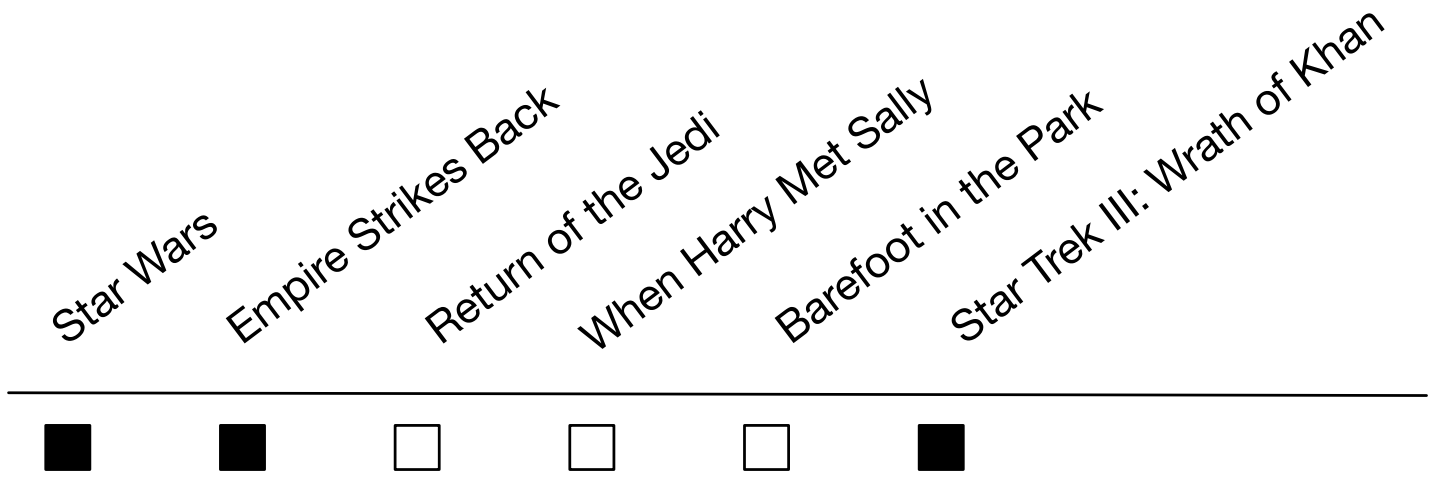
- *M-step*: Estimate $\mathcal{M}$ from the expected complete log likelihood of $\mathbf{y}$.



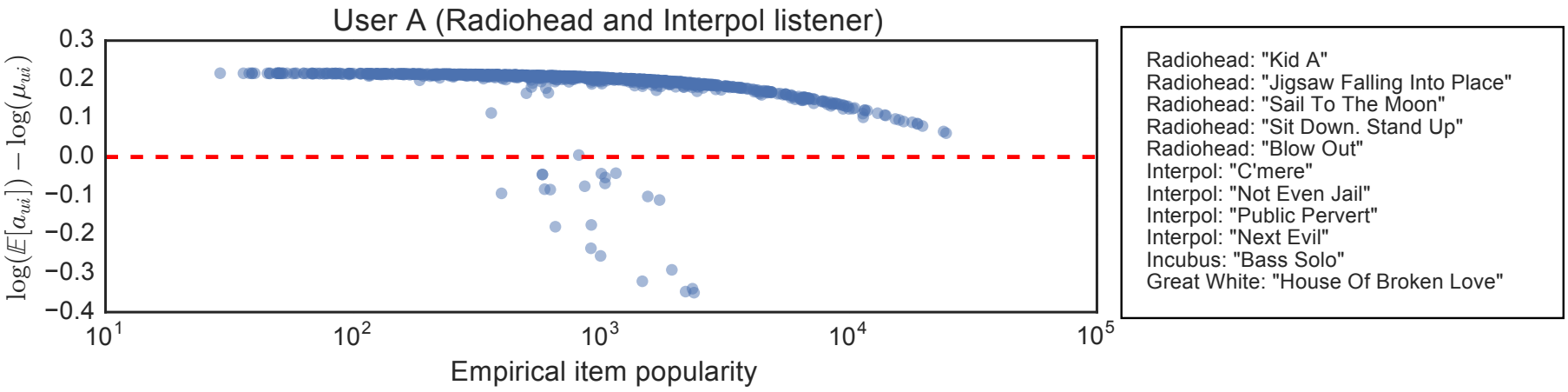
- E-step: Exposure prior interacts with probability of *not clicking*,

$$p(a_{ui} = 1 \mid y_{ui} = 0, \mathcal{M}) \propto p(a_{ui} = 1 \mid \mathcal{M})\, p(y_{ui} = 0 \mid a_{ui} = 1, \mathcal{M})$$

- In this example:
  - Posterior exposure to *Return of the Jedi* is smaller than the prior.
  - Posterior exposure to *When Harry Met Sally* is larger.



- M-step: Weighted factorization
  - Zeros weighted by the posterior probability of exposure.
  - Related to WMF, but emerges from a probabilistic perspective
  - And, not all 0s are reweighted by the same amount
- E.g., the zero for *Return of the Jedi* does not pull the user's preferences.
- (Exposure M-step: MLE with expected sufficient statistics.)



- This user likes Interpol and Radiohead.
- Songs that she might like are down-weighted more than usual. This distinguishes the algorithm from WMF.
- But, it's a modeling choice. (The alternative could also be plausible.)

	TPS	**Mendeley**	**Gowalla**	**ArXiv**
Type	Music	Papers	Venues	Clicks
# of users	220K	45K	58K	38K
# of items	23K	76K	47K	45K
# interactions	14.0M	2.4M	2.3M	2.5M
$\%$ interactions	0.29%	0.07%	0.09%	0.15%

We studied several large data sets of implicit data.

	TPS		**Mendeley**		**ArXiv**	
	WMF	ExpoMF	WMF	ExpoMF	WMF	ExpoMF
Recall@20	0.195	**0.201**	0.128	**0.139**	0.143	**0.147**
NDCG@100	0.255	**0.263**	0.149	**0.159**	0.154	**0.157**
MAP@100	0.092	**0.109**	0.048	**0.055**	0.051	**0.054**

Weighted MF vs. popularity-based exposure

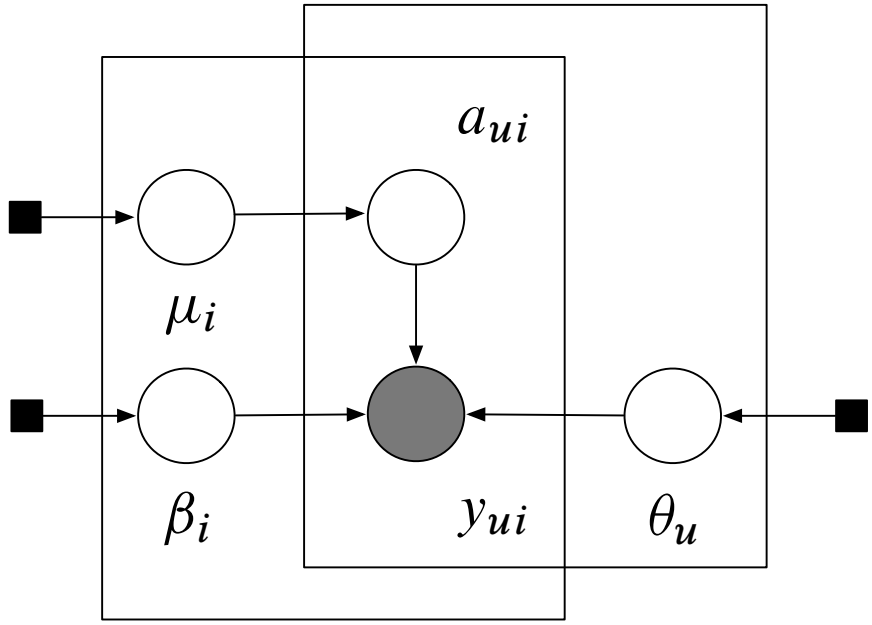
	WMF	Popularity Exposure	Location Exposure
Recall@20	0.122	0.118	**0.129**
NDCG@100	0.118	0.116	**0.125**
MAP@100	0.044	0.043	**0.048**

WMF vs. popularity-based vs. location-based exposure (Gowalla)

Related work:

- Causal inference and potential outcomes
- Weighted matrix factorization
- Missing data and recommender systems
- Spike and slab models

Pearl, *Causality*
Imbens and Rubin, *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*
Gelman et al., *Bayesian Data Analysis*
Hu et al., *International Conference on Data Mining* 2008
Marlin et al., *Uncertainty in Artificial Intelligence* 2007



Summary:

- We developed a latent exposure process for modeling implicit data.
- Enables us to consider two models—one of clicks and one of exposure.
- Outperforms weighted matrix factorization; opens the door to new methods
- See our paper at WWW 2016.

Causal Inference and Recommendation Systems

- Learning the assignment mechanism is useful for causal inference
  - It helps us better learn user preferences
  - It helps us better model implicit data
- Other work on recommender systems from our group
  - Hierarchical Poisson factorization
  - Social networks and recommender systems
  - Dynamic recommender systems
  - Combining content and clicks
  - Embeddings for recommendation

Gopalan et al., *Uncertainty in Artificial Intelligence* 2015
Chaney et al., *ACM Conference on Recommendation Systems* 2015
Charlin et al., *ACM Conference on Recommendation Systems* 2015
Gopalan et al., *Neural Information Processing Systems* 2014
Gopalan et al., *Artificial Intelligence and Statistics* 2014
Wang and Blei, *Knowledge Discovery and Data Mining* 2013
Liang et al., *ACM Conference on Recommendation Systems* 2016

Why it's not crazy to model integer data with a Normal distribution.

- OK, it is a little.
- Let $\mu = \theta_u^\top \beta_i$.
  Consider the Poisson with parameter $\exp\{\mu\}$.

$$p(y) \propto \exp\{y\mu - \exp\{\mu\} - \log y!\}$$

  Now consider a (unit-variance) Gaussian

$$p(y) \propto \exp\{y\mu - \mu^2/2 - y^2/2\}$$

- These are different distributions. But both contain
  - $y\mu$
  - A "penalty" for large parameters
  - A "penalty" for large variables