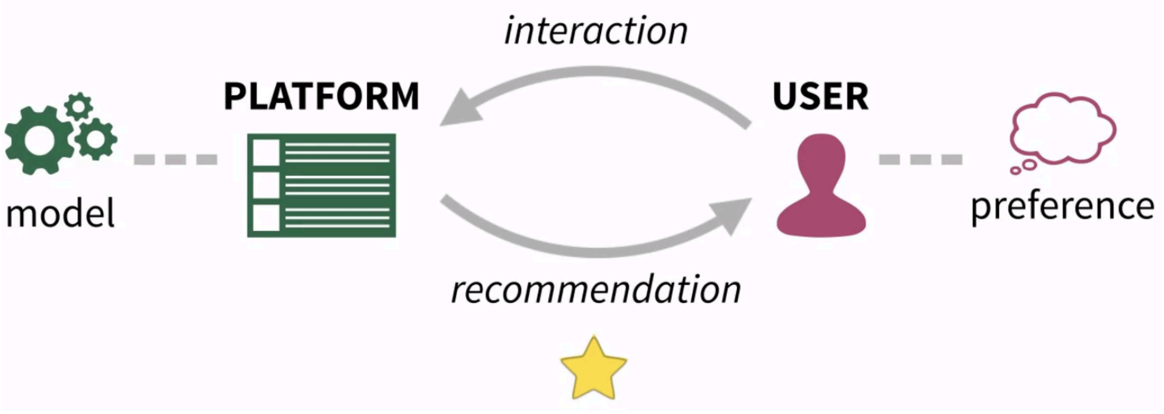
How Recommendation System Hurt Users with Minority Preferences

1/27/21 12:28 PM

Speaker: Allison Chaney (Assistant Professor, Duke University, Fuqua School of Business)
Date: 1/27/2021

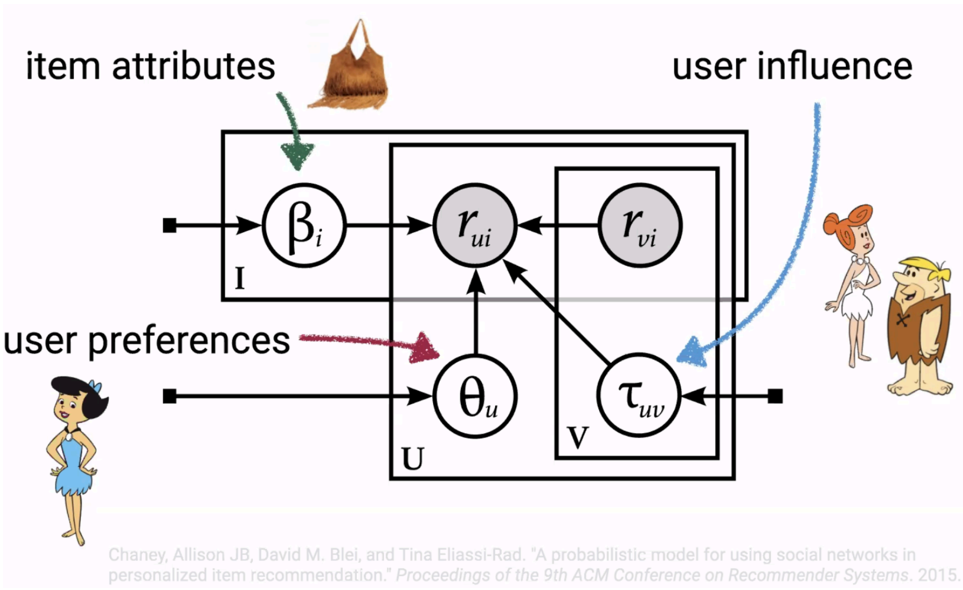
Business Benefits
- Customer retention (less likely to switch to competitor)
- Increasing average order value (more likely to add extra items to cart)
- Improved engagement



Algorithmic Confounding
- The feedback loop always recommend similar products for previous interaction (not necessary true preference)

Why do we need to understand algorithmic confounding?
- Customer perspective:
  - Always get same thing, not what I want…
- Company perspective:
  - Most recommendation algorithms are essentially segmenting the users of a platform
  - Poor segmentation
- Technical perspective:
  - Using confounded data lead to bad training and evaluation protocols
  - Model assumption not equal to true data generating

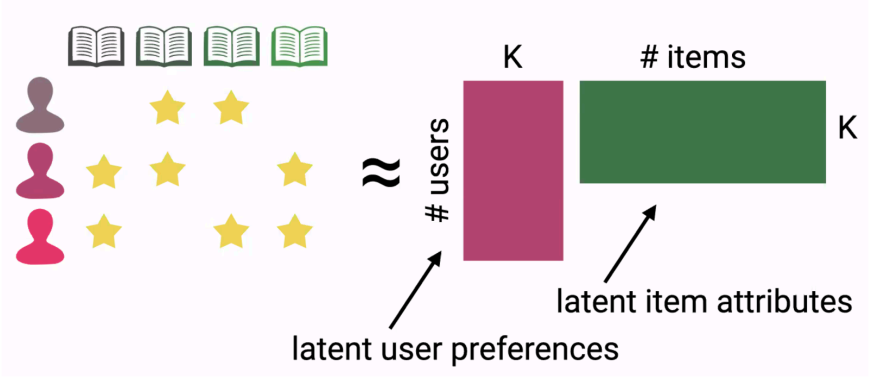
Motivation: From Social Poisson Factorization



- Etsy: simple model did better than the full version model

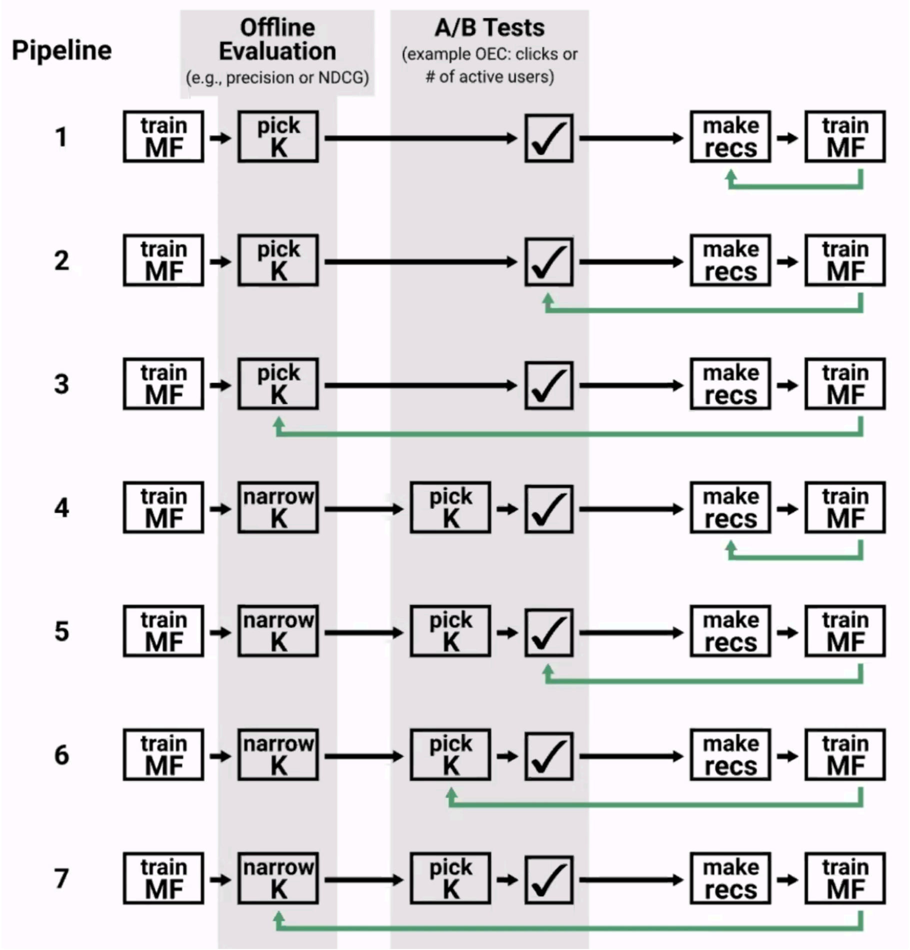
Findings
- Cause homogenization of user behavior
- Users experience losses in utility, losses are distributed unequally
- Feedback loop amplifies the impact of recommendation systems on the distribution of item consumption

Research
Matrix Factorization



Evaluating Recommendation Systems
1. Offline evaluation
   Metrics: nDCG, precision
2. A/B tests
   Metrics: # active users, CTR

Experiment set-up:

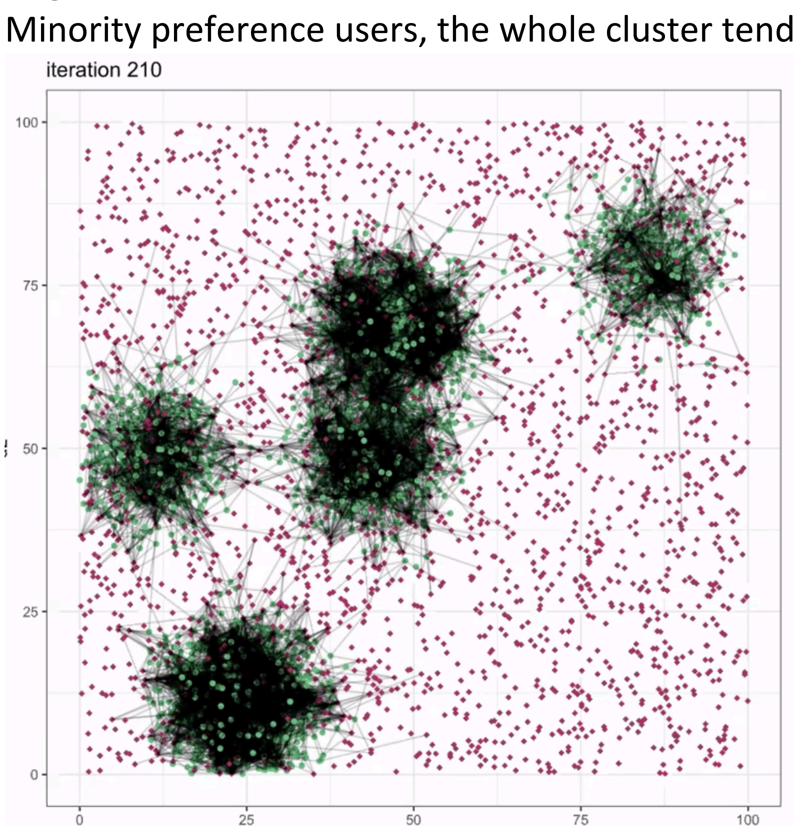


Method: Simulations
- Preference space K: 2 or 3 dimensions
- Items: 10K initially, +3K per iteration
- Users: 5000 initially, +5 per iteration
- Segments: none, 5, 10
- 10 random seeds per setting
- Train first at iteration 100, then consider updating every 50 iterations
- New items recommended 10% of the time

Evaluation Criteria
- Offline: nDCG
- A/B test criterion: user activity (have they consumed anything in the last 20 time periods?)

Choice Model
- Consideration set size of 5
- When full, two options:
  - Consume the highest utility item
  - Reject the lowest utility item
- Probabilistic based on highest utility item

Utility is calculated by the distance metrics
Simulation Visualization
- Users in smaller clusters (in terms of number of users) are tended to be grouped into larger clusters
- Minority preference users, the whole cluster tend to be pushed away from its center



Why are some pipelines better?
- Pipeline 3 and 7 are significantly better
- Other pipelines after 1st and 2nd segments, are hurting customer activity
- Distance for preference and recommendation are pushed away except for P3 and P7

Conclusion
1. Poorly tuned models disproportionately hurt users with minority preferences by recommending them content further from their preferences
2. Repeated A/B tests should be used with care: they can weaken overall performance (and contribute to the above effects), but can improve quality with poorly tuned models