Training and Testing of Recommender Systems on Data Missing Not at Random

Harald Steck

Bell Labs, Murray Hill

at KDD, July 2010

Overview

Real-World Problem:

Make personalized recommendations to users that they find "relevant":

1. from <u>all</u> items (in store)
2. pick a few for each user
3. with the goal: each user finds recommended items "relevant".

eg "relevant" = 5-star rating in Netflix data

approx.

Define Data Mining Goal (how to test):

- off-line test with historical rating data
- high accuracy
  - RMSE on <u>observed</u> ratings (popular)
  - nDCG on <u>observed</u> ratings [Weimer et al. '08]

approx.

Find (approximate) solution to Goal defined above:

- choose model(s)
- appropriate training-objective function
- efficient optimization method

Overview

Real-World Problem:

Make personalized recommendations to users that they find "relevant":

1. from <u>all</u> items (in store)
2. pick a few for each user
3. with the goal: each user finds recommended items "relevant".

eg "relevant" = 5-star rating in Netflix data

this talk

approx.

Define Data Mining Goal (how to test):

- off-line test with historical rating data
- high accuracy
    - RMSE on <u>observed</u> ratings (popular)
    - nDCG on <u>observed</u> ratings [Weimer et al. '08]

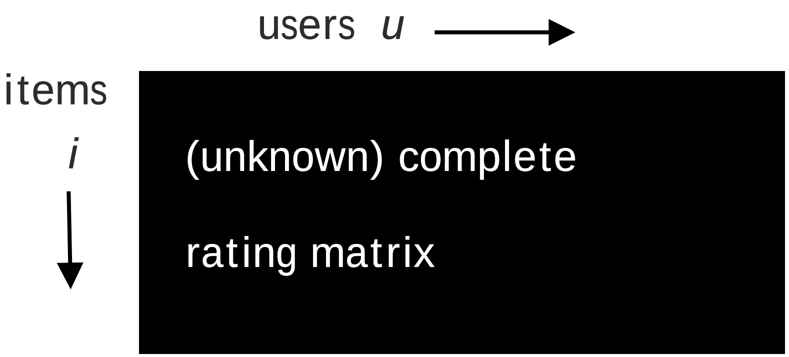
approx.

Find (approximate) solution to Goal defined above:

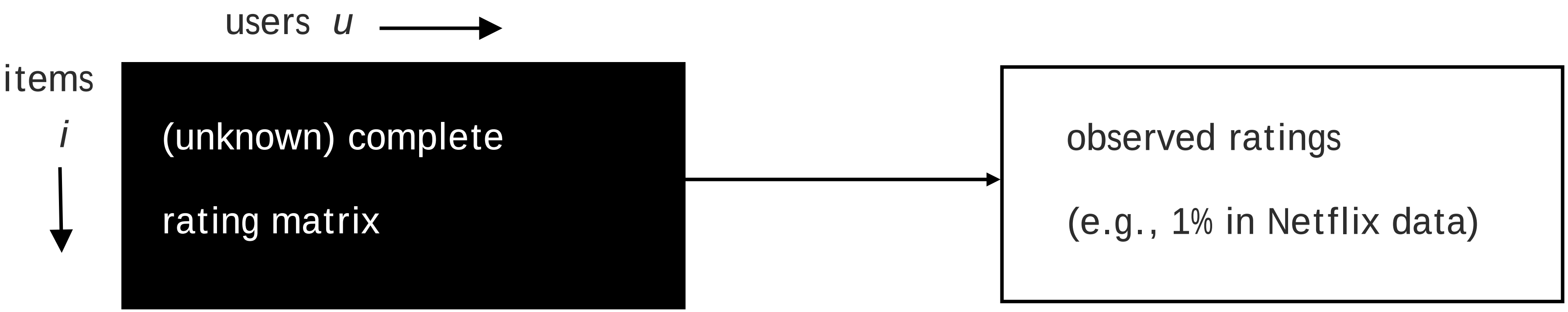
- choose model(s)
- appropriate training-objective function
- efficient optimization method

Alcatel·Lucent

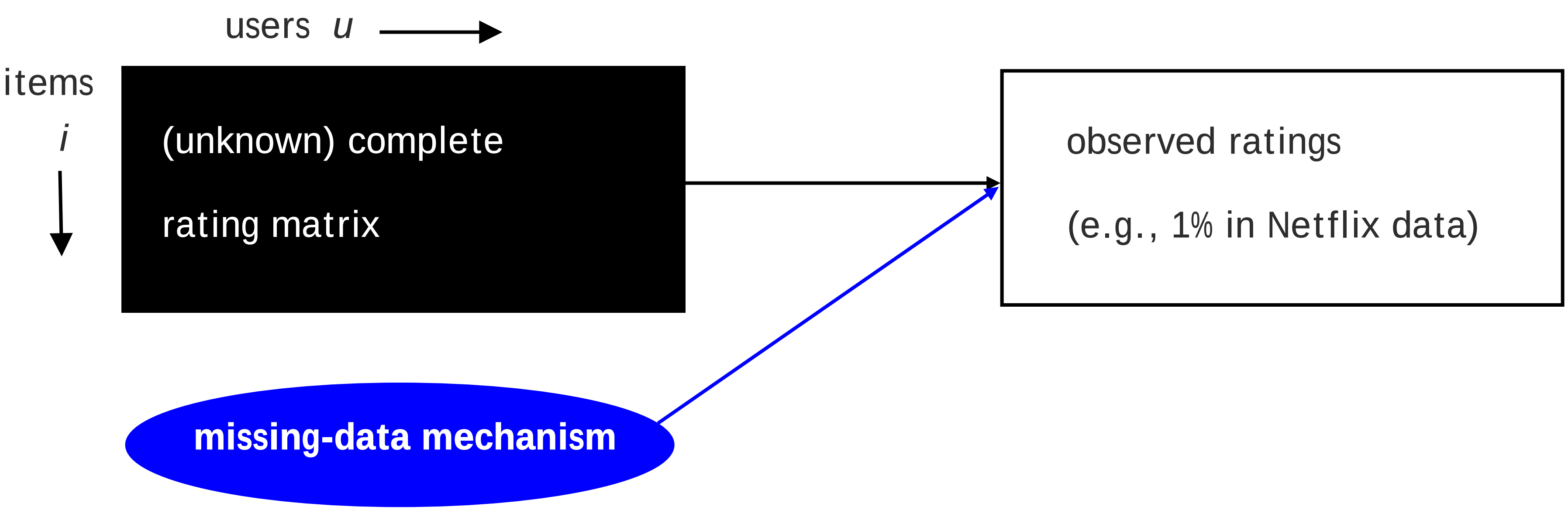
Data



Data


users u
items
i
(unknown) complete
rating matrix
observed ratings
(e.g., 1% in Netflix data)

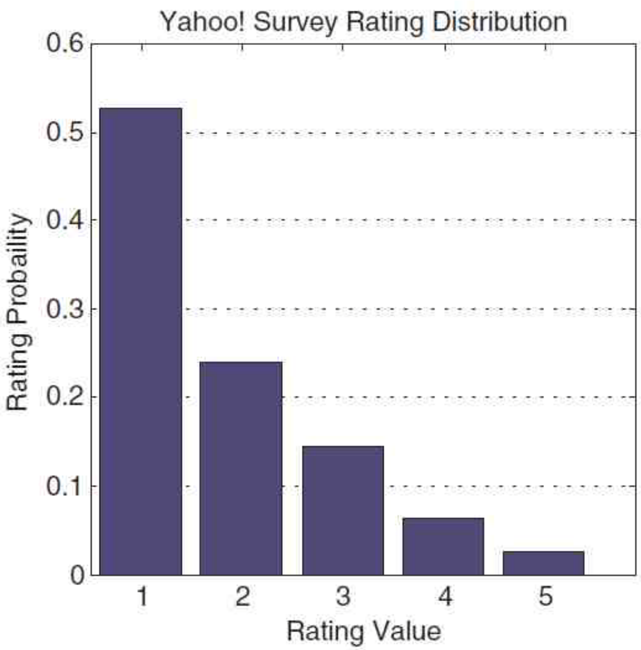
Data

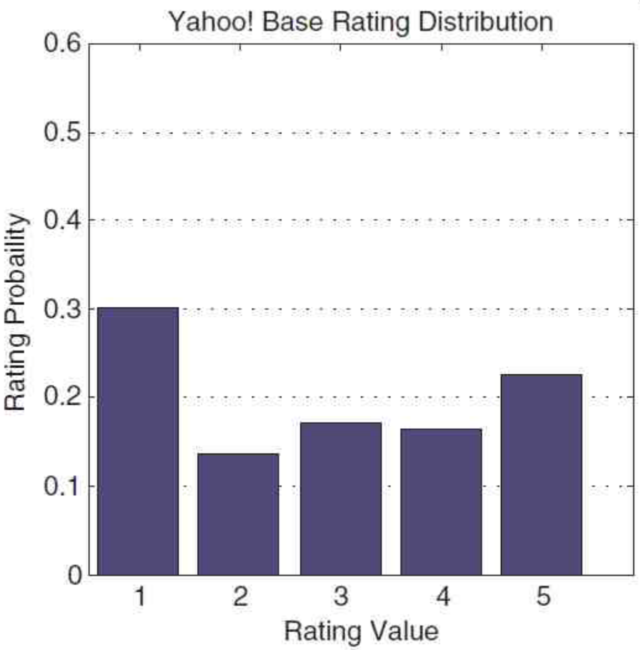


- (General) missing-data mechanism cannot be ignored [Rubin ’76; Marlin et al. ’09,’08,’07].
- Missing at random [Rubin ’76; Marlin et al. ’09,’08,’07]:
  - Rating value has **<u>no</u>** effect on probability that it is missing
  - Correct results obtained by ignoring missing ratings.

Ratings are missing not at random (MNAR): Empirical Evidence

Graphs from [Marlin & Zemel '09]:





Survey: ask users to rate a <u>random</u> list of items: approximates **complete** data

Typical Data: users are <u>free to choose</u> which items to rate -> available data are **MNAR** :

instead of giving low ratings, users tend to not give a rating at all.

Overview

Real-World Problem:

Make personalized recommendations to users that they find “relevant”:

1. from all items (in store)
2. pick a few for each user
3. with the goal: each user finds recommended items “relevant”.

this talk

approx.

Define Data Mining Goal (how to test):

- off-line test with historical rating data
- high accuracy
  - RMSE, nDCG,… on observed ratings
  - Top-k Hit-Rate,… on all items

approx.

Find (approximate) solution to Goal defined above:

- choose model(s)
- appropriate training-objective function
- efficient optimization method

Alcatel·Lucent

Test Performance Measures on MNAR Data

- many popular performance measures cannot readily deal with missing ratings
- only a few from among all items can be recommended
- **Top-*k* Hit Rate w.r.t. all items:**
  - $$\frac{\#\ \text{relevant items in top } k}{\#\ \text{relevant items}} = \text{recall}$$
  - $$\frac{\#\ \text{relevant items in top } k}{k} = \text{precision}$$

Alcatel·Lucent

Test Performance Measures on MNAR Data

- most popular performance measures cannot readily deal with missing ratings
- only a few from among all items can be recommended
- **Top-*k* Hit Rate w.r.t. all items:**

  - $$\mathrm{TOPK}_u(k) = \frac{\#\ \text{relevant items in top } k}{\#\ \text{relevant items}} = \text{recall}$$

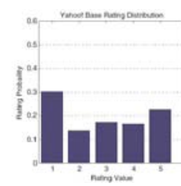
  - $$\frac{\#\ \text{relevant items in top } k}{k} = \text{precision}$$

- when comparing different rec. sys. on fixed data and fixed *k*: recall $\propto$ precision

- under mild assumption:

  recall on MNAR data = unbiased estimate of recall on (unknown) complete data

Assumption: The **relevant** ratings are missing at random.



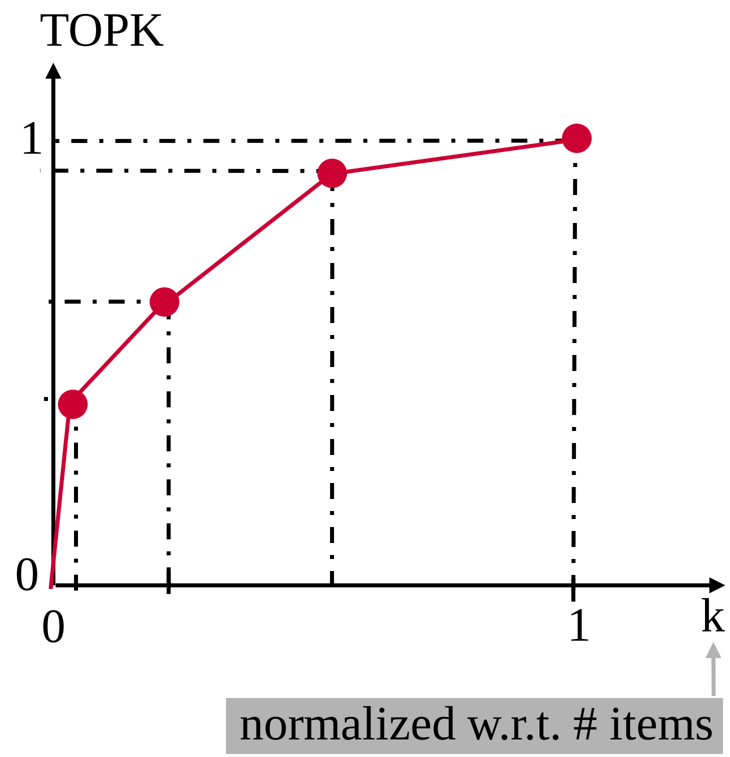
Alcatel·Lucent

Test Performance Measures on MNAR Data

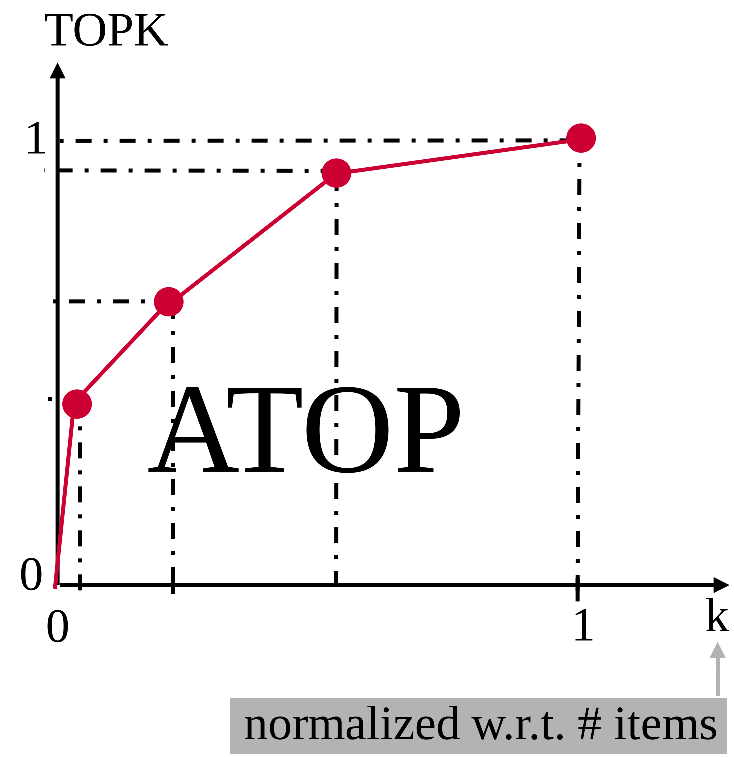
<u>Top-*k* Hit-Rate:</u>

- depends on *k*

- ignores ranking



Test Performance Measures on MNAR Data



Top-*k* Hit-Rate:

- depends on *k*
- ignores ranking

Area under TOPK curve (**ATOP**):

- independent of *k*
- in [0,1], larger is better
- captures ranking of all items
- agrees with area under ROC curve in leading order if # relevant items << # items
- unbiased estimate from MNAR data for unknown complete data under above assumption

Overview

Real-World Problem:

Make personalized recommendations to users that they find "relevant":

1. from all items (in store)
2. pick a few for each user
3. with the goal: each user finds recommended items "relevant".

this talk

approx.

Define Data Mining Goal (how to test):

- off-line test with historical rating data
- high accuracy
    - TOPK, ATOP,... on all items

approx.

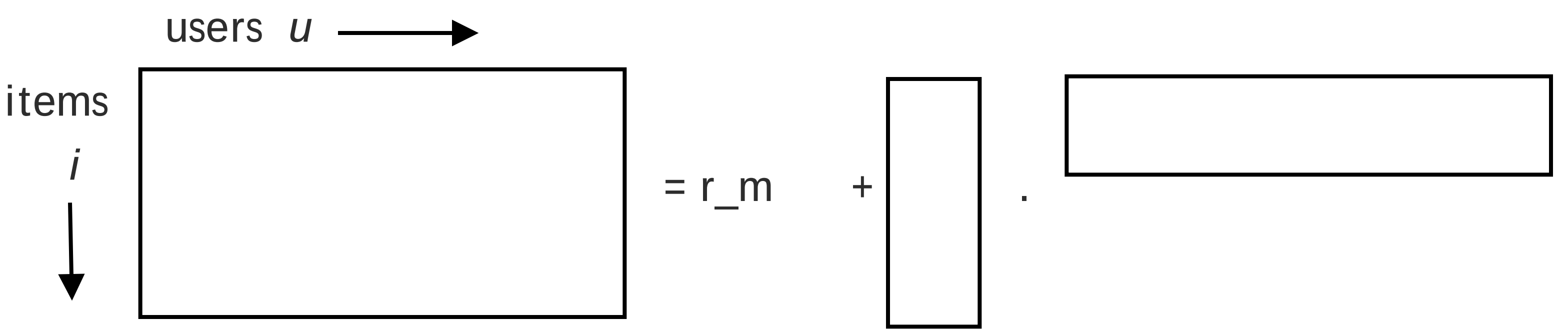
Find (approximate) solution to Goal defined above:

- choose model(s)
- appropriate training objective function
- efficient optimization

Low-rank Matrix Factorization Model

Matrix of predicted ratings:

$$\hat{R} = r_m + PQ^\top$$



- rating offset: r_m
- rank of matrices P,Q: dimension of low-dimensional latent space, eg d_0 = 50

Training Objective Function: AllRank

minimal modification of usual least squares problem:

- account for **all** items per user: observed and missing ratings $R^{\mathrm{o\&i}}_{i,u}$

- imputed value for missing ratings: r_m

- create balanced training set: weights (1 if observed, w_m if missing)

- (usual) regularization of matrix elements: lambda

$$\sum_{\text{all } u} \sum_{\text{all } i} W_{i,u} \cdot \left\{ \left(R^{\mathrm{o\&i}}_{i,u} - (r_m + PQ^\top)_{i,u} \right)^2 + \lambda \left(\sum_{d=1}^{d_0} P^2_{i,d} + Q^2_{u,d} \right) \right\}$$

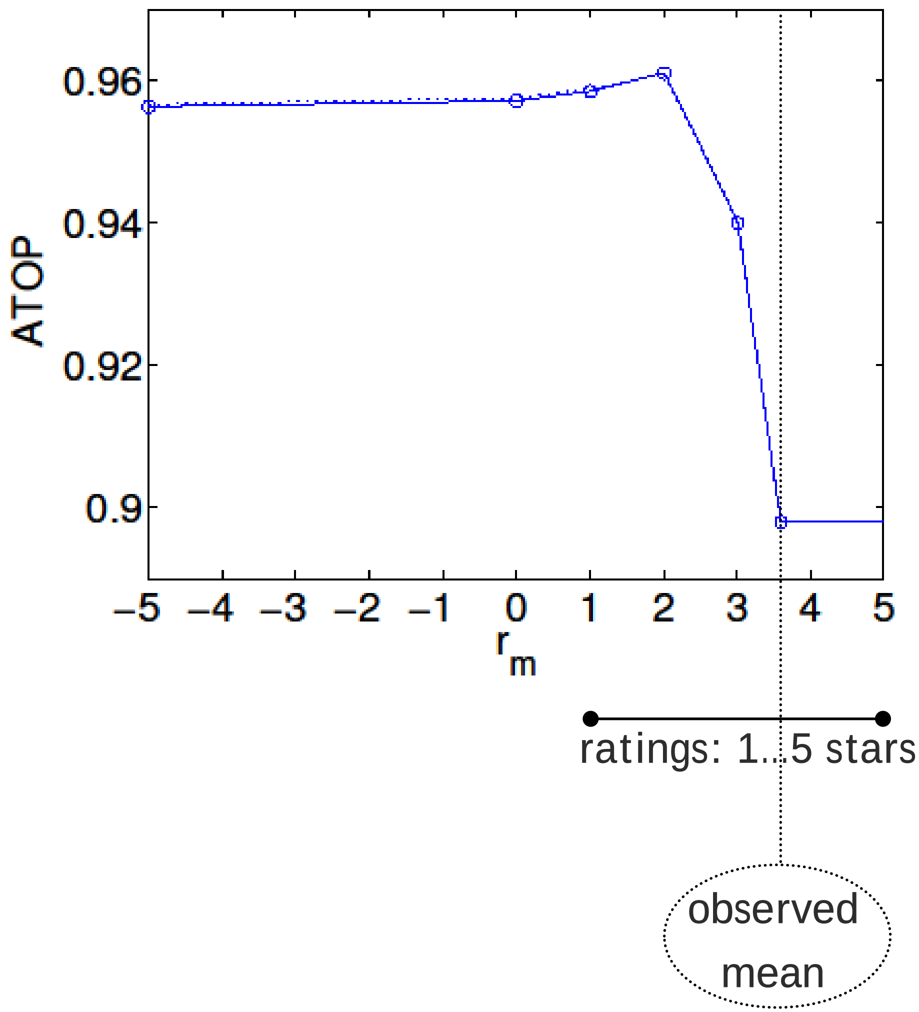
Efficient Optimization:

- gradient descent by alternating least squares

- tuning parameters r_m, w_m, lambda have to be optimized as well (eg w.r.t. ATOP)

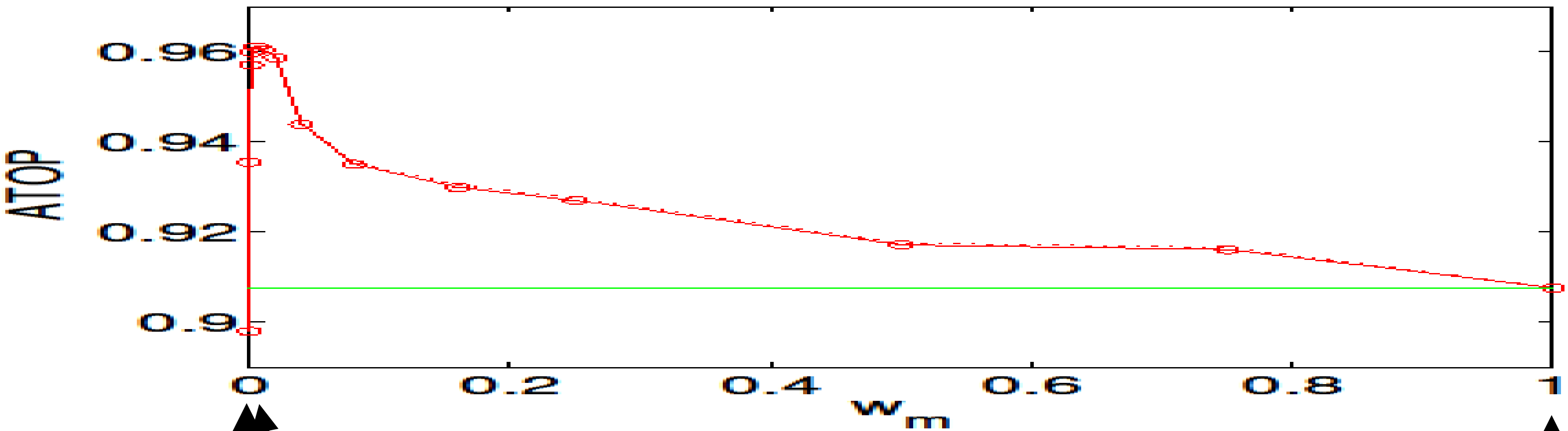
Experimental Results on Netflix Data: Imputed Rating Value r_m

- optimum for imputed value exists
- optimal r_m ≈2
- optimal r_m may be interpreted as mean of missing ratings
- exact imputation value < 2 is not critical
- imputed value < observed mean



Alcatel·Lucent

Experimental Results on Netflix Data: Weight of Missing Ratings w_m



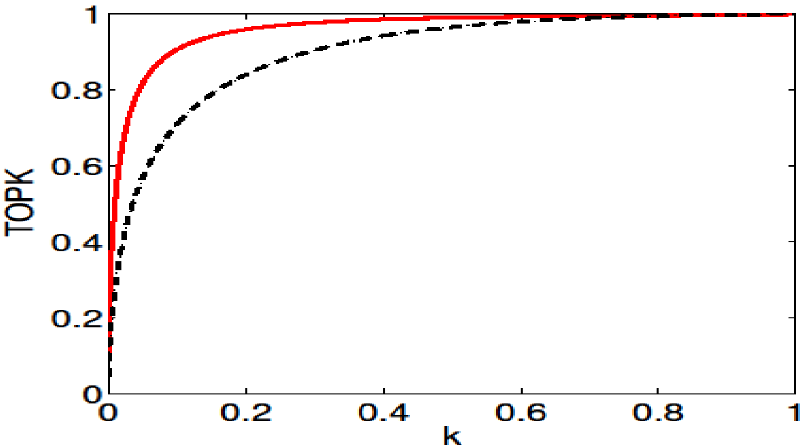
w_m=1: standard SVD (plus penalty term), like in Latent Semantic Analysis

w_m ≈ 0.005 is **optimal**; compare to fraction of observed ratings = 0.01

w_m=0: ignores missing ratings, and is **worst** w.r.t. ATOP

Alcatel·Lucent

Experimental Results on Netflix Data: Top-k Hit-Rate

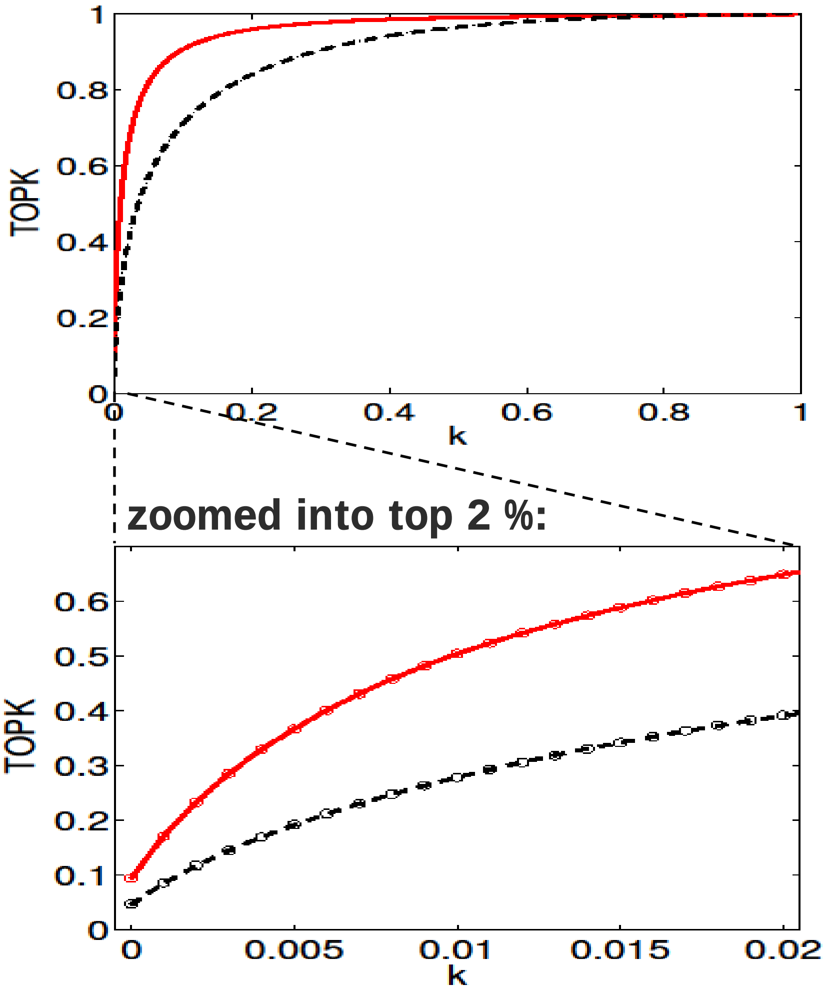


Comparison of Approaches:

— AllRank (RMSE = 1.106)

– – ignore missing ratings (RMSE = 0.921)

Experimental Results on Netflix Data: Top-k Hit-Rate

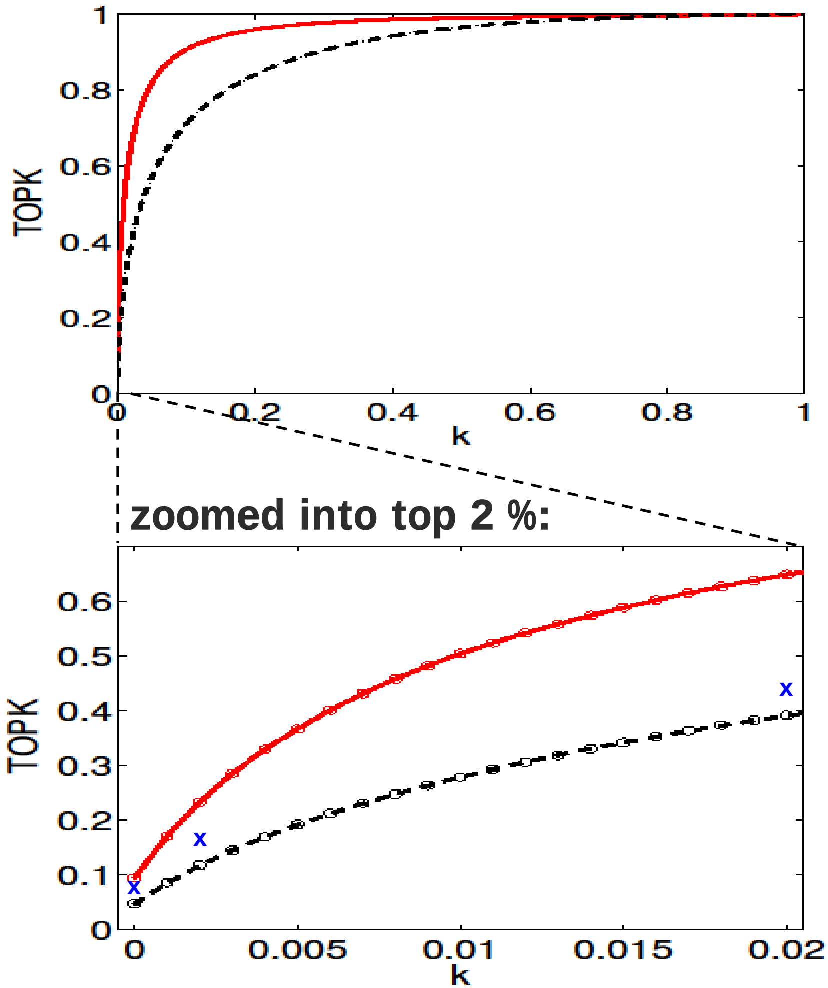


Comparison of Approaches:

— AllRank (RMSE = 1.106)

-- ignore missing ratings (RMSE = 0.921)

Experimental Results on Netflix Data: Top-k Hit-Rate



Comparison of Approaches:

— AllRank (RMSE = 1.106)

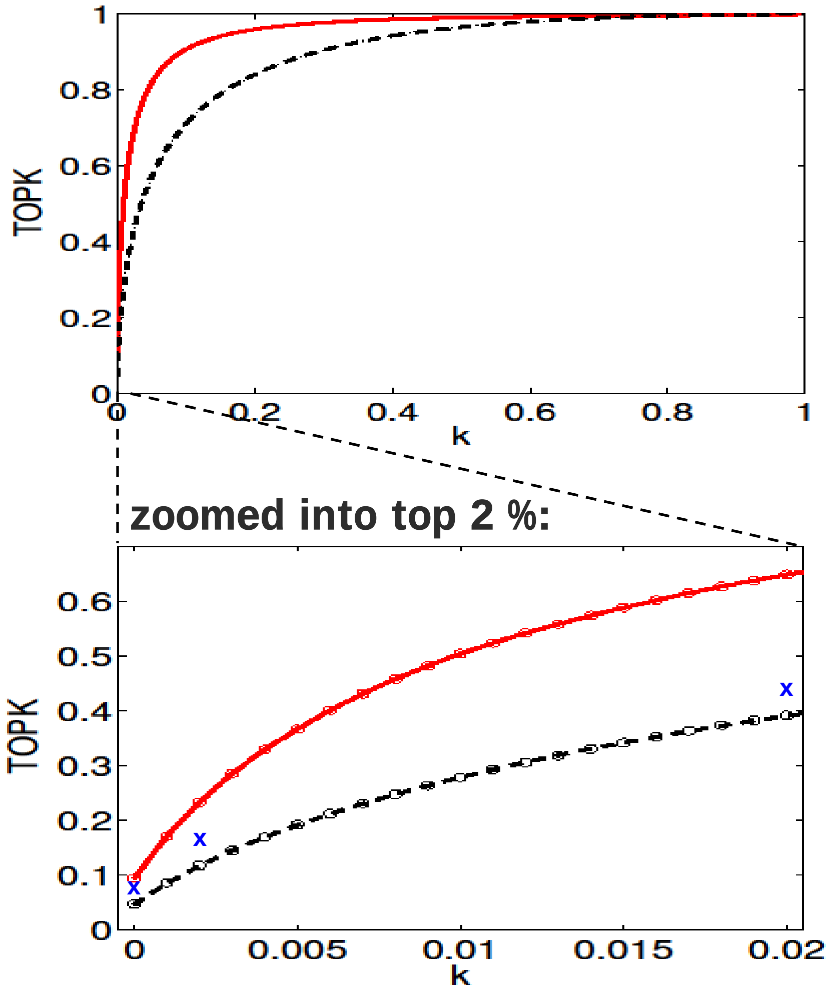
- - ignore missing ratings (RMSE = 0.921)

x integrated model [Koren '08] (RMSE = 0.887)
(trained to minimize RMSE)

39 % .... 50 % larger Top-k Hit-Rate: AllRank vs. integrated model

Alcatel·Lucent

Experimental Results on Netflix Data: Top-k Hit-Rate



Comparison of Approaches:

— AllRank (RMSE = 1.106)

- - ignore missing ratings (RMSE = 0.921)

x integrated model [Koren '08] (RMSE = 0.887)
(trained to minimize RMSE)

Large increase in Top-k Hit-Rate when accounting also for missing ratings when training on MNAR data.

39 % .... 50 % larger Top-k Hit-Rate: AllRank vs. integrated model

Alcatel·Lucent

Related Work

explicit feedback data (ratings):

- improved RMSE on observed data also increases Top-k Hit-Rate on all items [Koren ’08]
- ratings are missing not at random:
  - improved models: conditional RBM, NSVD1/2, SVD++ [Salakhutdinov ’07; Paterek ’07; Koren ‘08]
  - test on “complete” data, train multinomial mixture model on MNAR data [Marlin et al. ’07,’09]

implicit feedback data (clickstream data, TV consumption, tags, bookmarks, purchases, …):

- [Hu et al. ’07; Pan et al. ’07]:
  - binary data, **only positives are observed** -> missing ones assumed negatives
  - trained matrix-factorization model with weighted least-squares objective function
  - claimed difference to **explicit feedback data: latter provides positive and negative** observations

Conclusions and Future Work

- considered **explicit** feedback data missing not at random (**MNAR**)
- test performance measures:
  - **close to real-world problem**
  - **unbiased** on MNAR data (under mild assumption)
  - (Area under) Top-k Hit Rate, ...
- efficient surrogate objective function for training:
  - **AllRank**: **accounting for missing ratings** leads to **large improvements in Top-k Hit-Rate**

Future Work:

- better test performance measures, training objective functions and models
- results obtained w.r.t. RMSE need not hold w.r.t. Top-k Hit-Rate on MNAR data, eg collaborative filtering vs content based methods

www.alcatel-lucent.com