zalando

EVALUATING RECOMMENDER SYSTEMS

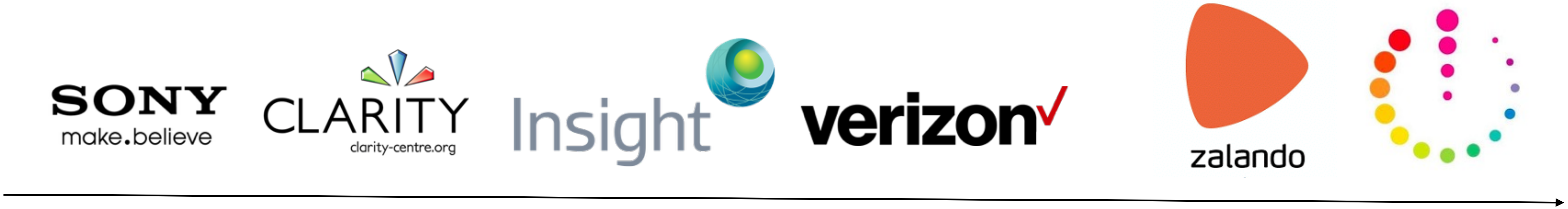
ACCURACY AND BEYOND

GITHUB.COM/HCORONA/AICS-2016

HUMBERTO CORONA
@TOTOPAMPIN

24-10-2016

ABOUT ME


SONY
make.believe
CLARITY
clarity-centre.org
Insight
verizon
zalando

REFERENCES

[1] Humberto Jesús Corona Pampín, Houssem Jerbi, and Michael P. O'Mahony. "Evaluating the Relative Performance of Neighbourhood-Based Recommender Systems." Spanish Conference of Information Retrieval, 2014

[2] Humberto Jesús Corona Pampín, Houssem Jerbi, and Michael P. O'Mahony. "Evaluating the Relative Performance of Collaborative Filtering Recommender Systems." Journal of Universal Computer Science 21.13 (2015): 1849-1868.

ZALANDO

https://www.zalando.co.uk/women-street-style/
https://www.zalando.co.uk/men-street-style/

RECOMMENDER SYSTEMS

Enable **content discovery**
by learning the user preferences and
exploiting the wisdom of the crowd.

EVALUATION

EVALUATION METRICS

RMSE

PRECISION

RECALL

F-1

DIVERSITY

POPULARITY

CATALOG COVERAGE

PER USER ITEM COVERAGE

UNIQUENESS

EVALUATION METRICS, ACCURACY

RMSE

PRECISION

RECALL

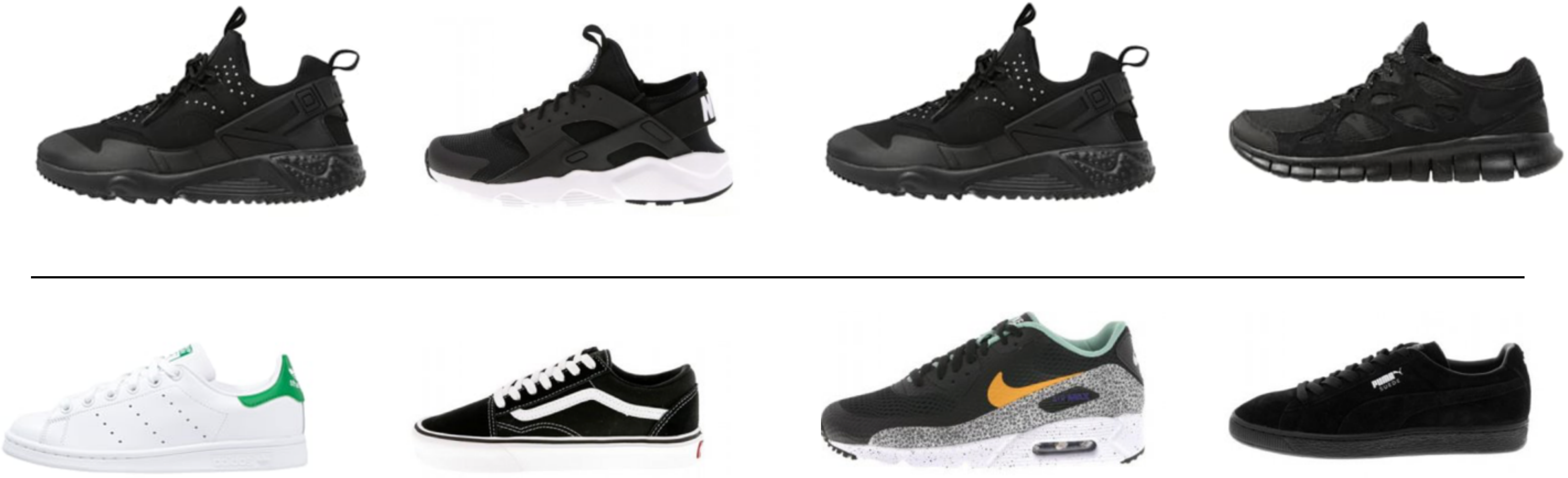
F-1

EVALUATION METRICS, BEYOND ACCURACY

- DIVERSITY
- POPULARITY
- CATALOG COVERAGE
- PER USER ITEM COVERAGE
- UNIQUENESS

EVALUATION METRICS

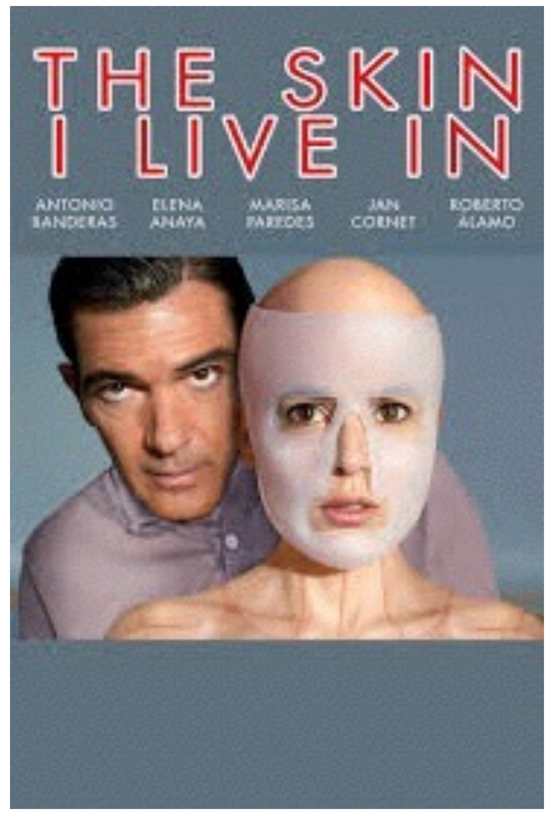
DIVERSITY

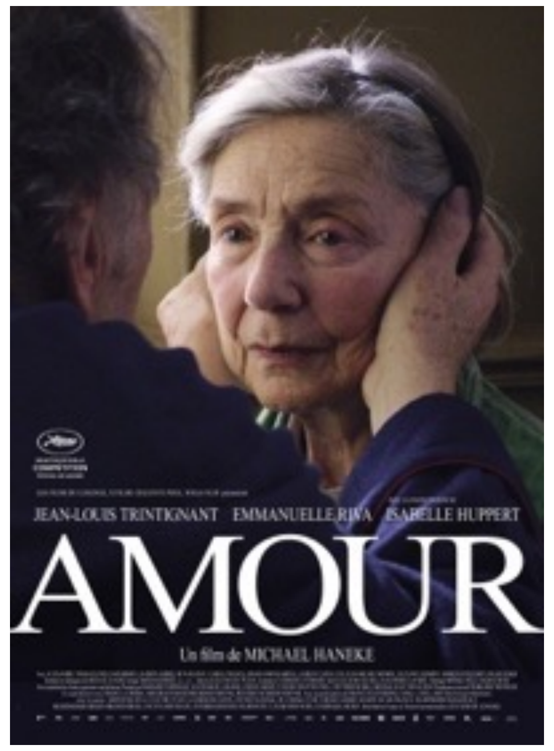


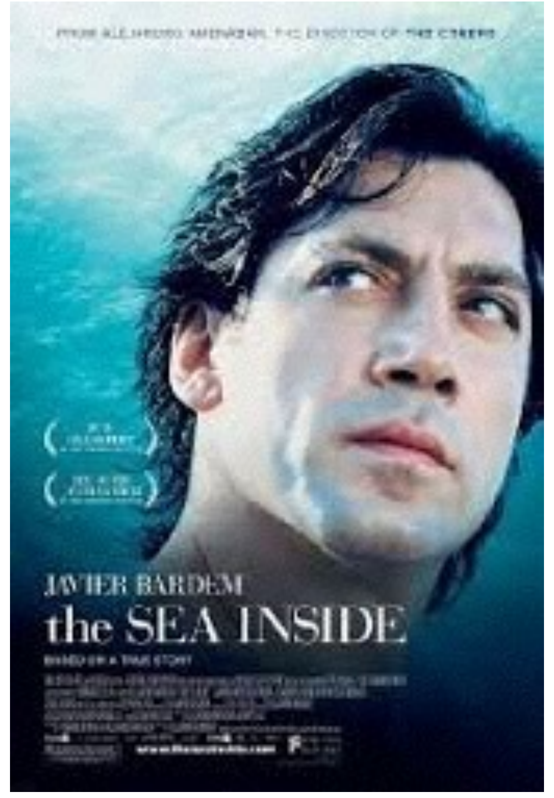
EVALUATION METRICS

POPULARITY

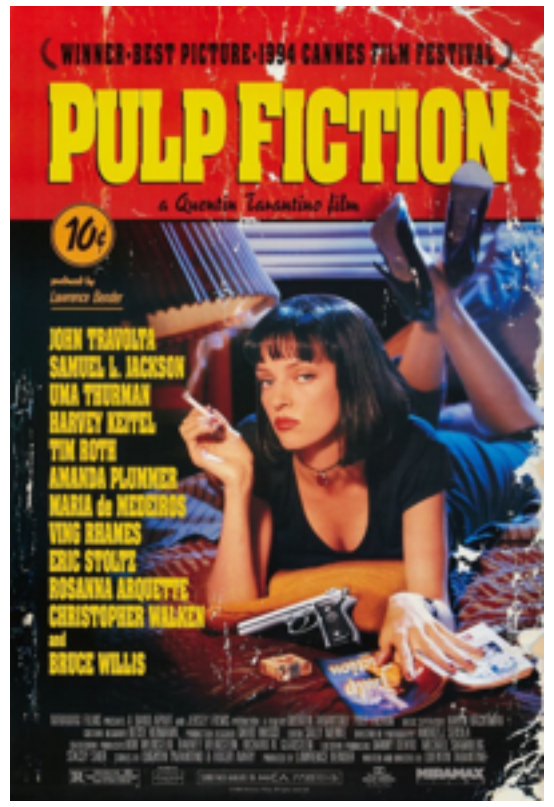
WINNER BEST ACTOR COLIN FIRTH VENICE FILM FESTIVAL
COLIN FIRTH JULIANNE MOORE
A SINGLE MAN
A FILM BY TOM FORD




JEAN-LOUIS TRINTIGNANT EMMANUELLE RIVA ISABELLE HUPPERT
AMOUR
Un film de MICHAEL HANEKE



"A POWERFUL FILM ABOUT THE FEROCITY OF A MOTHER'S LOVE"
WINNER
"A HEART-SWELLING, BREATHTAKING PIECE OF CINEMA"
Mommy



WELCOME TO A WORLD WITHOUT RULES
BALE CAINE LEDGER OLDMAN ECKHART GYLLENHAAL FREEMAN
THE DARK KNIGHT
JULY 18

EVALUATION METRICS

PER USER ITEM COVERAGE

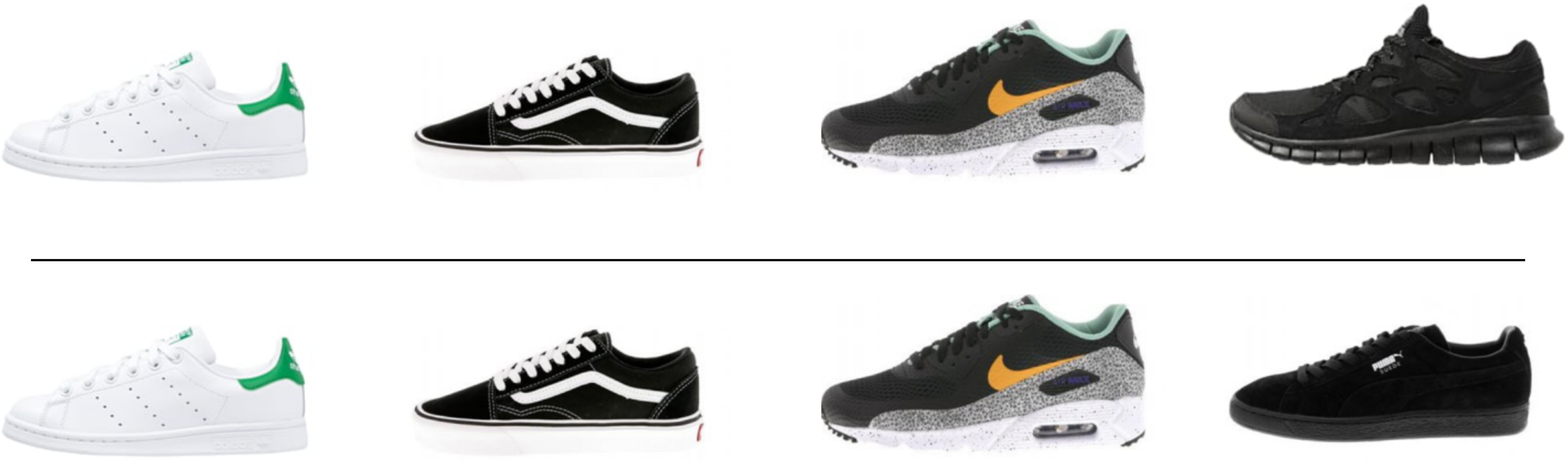
The proportion of items, across the catalog, which are candidates for recommendations.

CATALOG COVERAGE

Proportion of items which ever get recommended.

EVALUATION METRICS

UNIQUENESS



EVALUATION METRICS

RMSE

PRECISION

RECALL

F-1

DIVERSITY

POPULARITY

CATALOG COVERAGE

PER USER ITEM COVERAGE

UNIQUENESS

EVALUATION METRICS

PRECISION

RECALL

F-1

DIVERSITY

POPULARITY

CATALOG COVERAGE

PER USER ITEM COVERAGE

UNIQUENESS

ARE UKNN AND IKNN REALLY THAT DIFFERENT?

A COMPARATIVE ANALYSIS

EXPERIMENT DESIGN

THE DATA

MOVIELENS - 100K

MOVIELENS - 1M

TRAINING DATA

TESTING DATA

10 ITEMS TEST SET

THE MODELS

UKNN

IKNN

UKNN [20, 200]

IKNN FIXED

EVALUATION

ACCURACY

BEYOND ACCURACY

THE ALGORITHMS

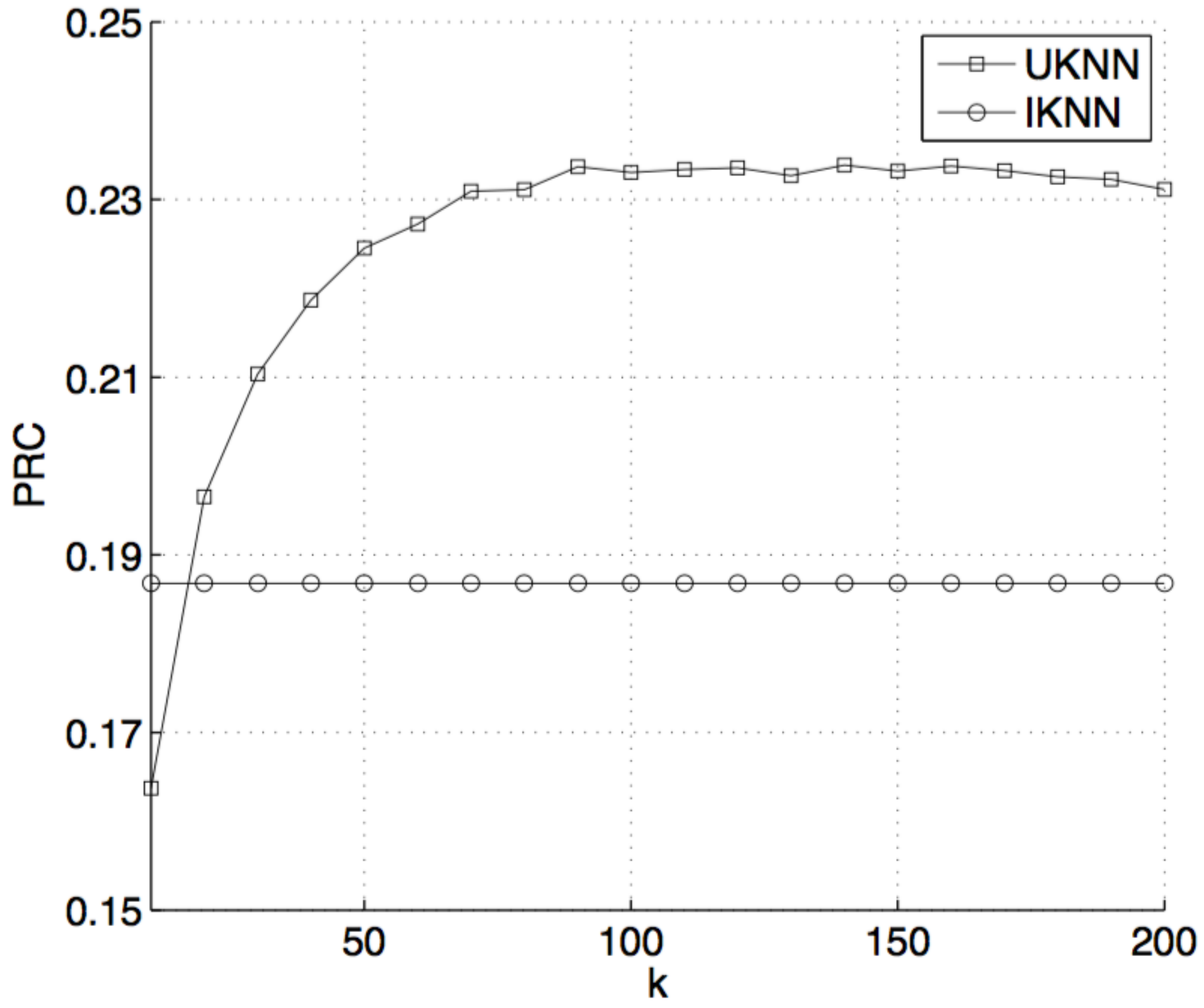
USER BASED COLLABORATIVE FILTERING (UKNN)

- Find similar users
- word of mouth
- The neighbours paradigm
- Scales with number of users

ITEM-BASED COLLABORATIVE FILTERING (IKNN)

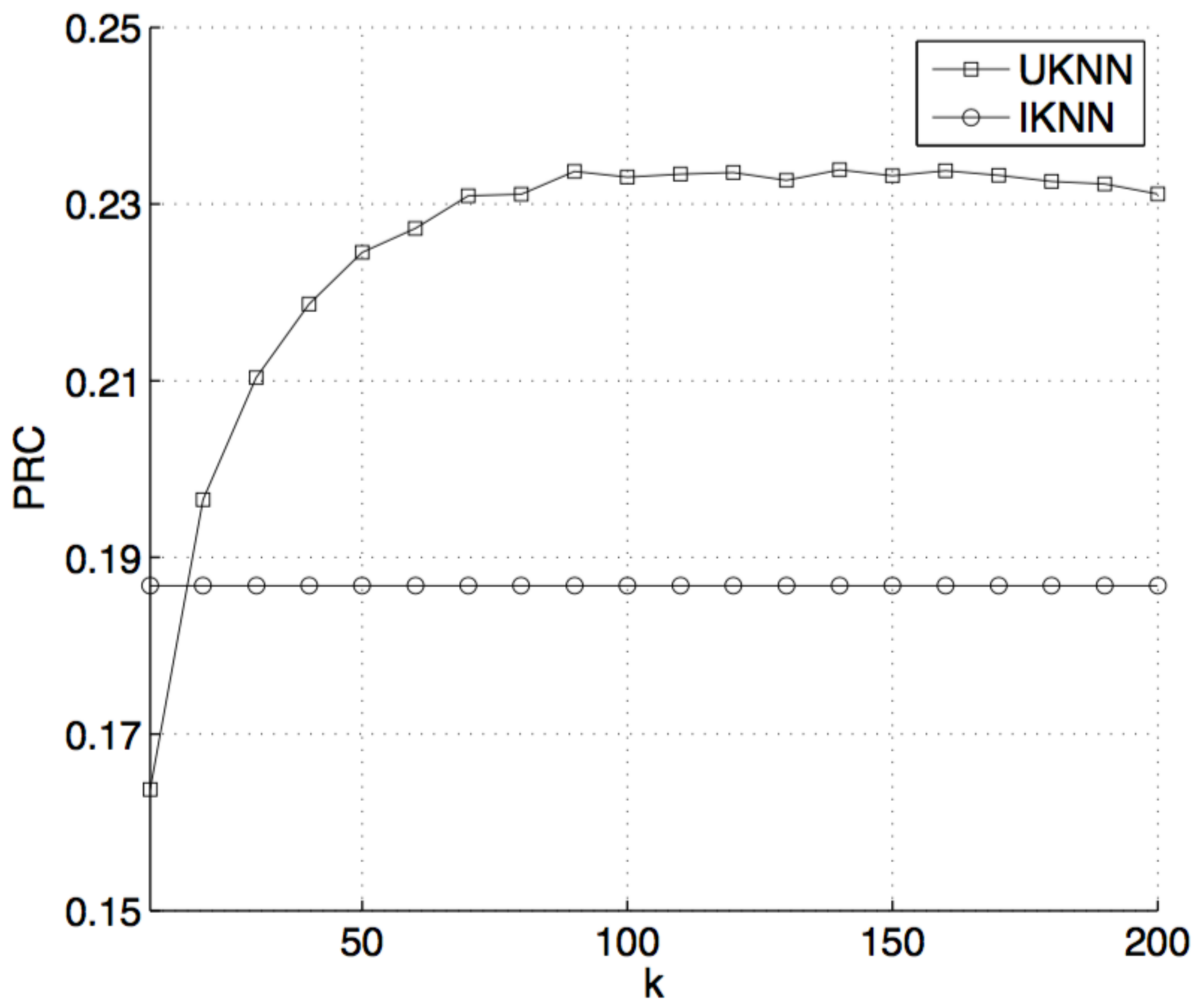
- Find similar items
- Scalable
- Widely used

RESULTS

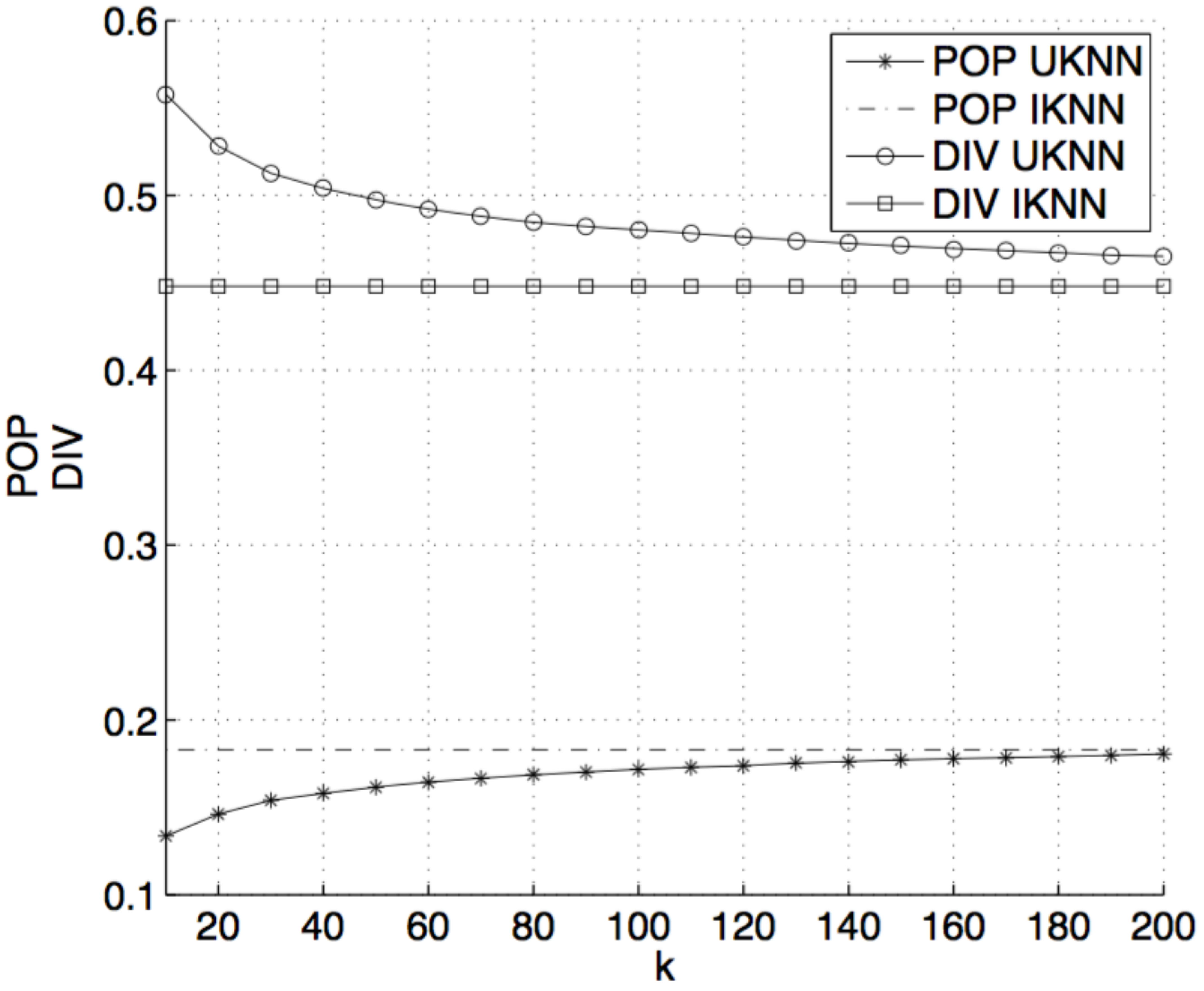


(a) Precision vs. neighbourhood size.

RESULTS



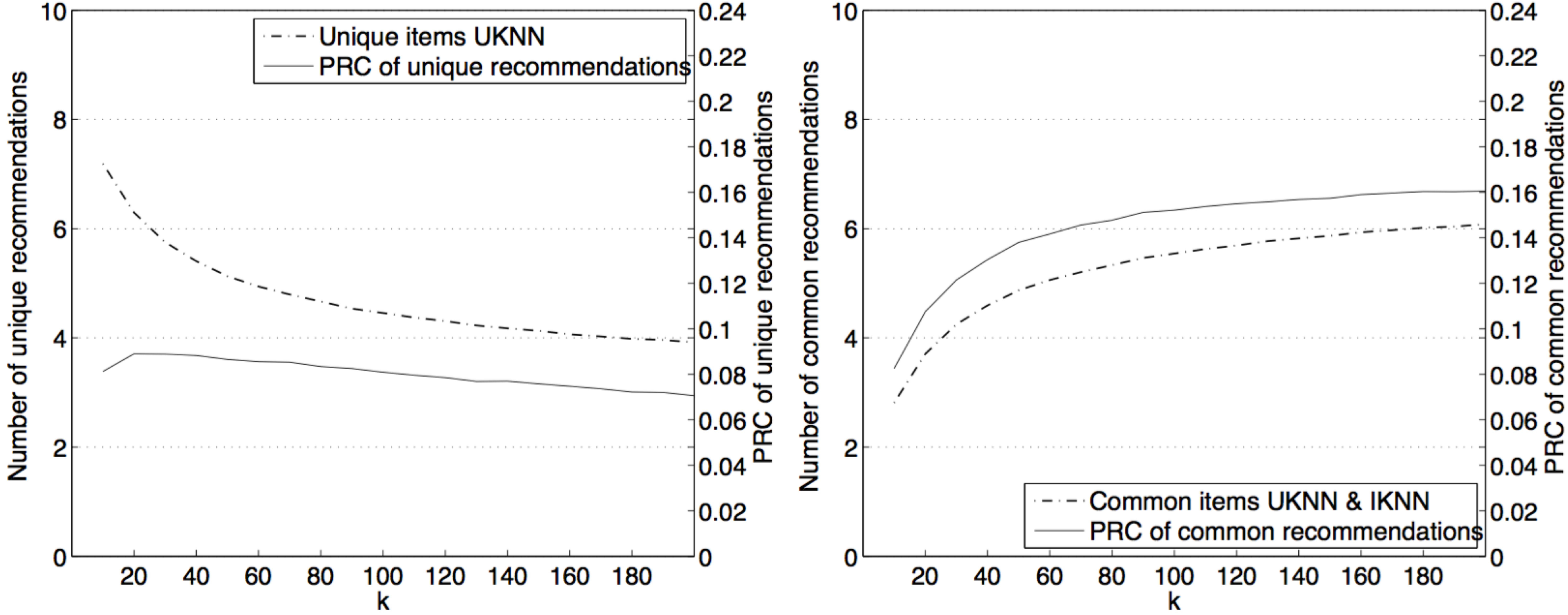
(a) Precision vs. neighbourhood size.



(b) POP and DIV vs. neighbourhood size.

Insert footnote

RESULTS



(c) Number and precision of unique *UKNN* recommendations vs. neighbourhood size.

(d) Number and precision of common recommended items vs. neighbourhood size.

SUMMARY

LESSONS LEARNED

- One size fits all is not true, **never, ever!**
- Use many metrics, even if you don’t optimise for them
  - **They help understanding what is the model doing**
- Use various datasets (if you want to publish a paper) - **Do results generalise?**
- Understand what is the best proxy or dataset for your evaluation goal.

CONCLUSIONS

- User-based (UKNN) and item-based (UKNN) collaborative filtering algorithms have a high inverse correlation between popularity and diversity.
- Smaller neighbourhood sizes (for UKNN) lead to more unique, less popular, and more diverse recommendations.
- Recommend a common set of items at large neighbourhood sizes.

- Matrix factorisation approach (WMF) leads to more accurate and diverse recommendations, while being less biased toward popularity.
- item-based collaborative filtering (IKNN) has significantly better catalog coverage.

zalando

EVALUATING RECOMMENDER SYSTEMS

ACCURACY AND BEYOND

GITHUB.COM/HCORONA/AICS-2016

HUMBERTO CORONA
@TOTOPAMPIN

24-10-2016

EXPERIMENT II

A BIAS ANALYSIS

EXPERIMENT DESIGN

THE DATA

- FACEBOOK DATASET
- MOVIELENS - HETREC
- LASTFM - HETREC

- TRAINING DATA
- TESTING DATA
- 10 FOLD CROSSVALIDATION

THE MODELS

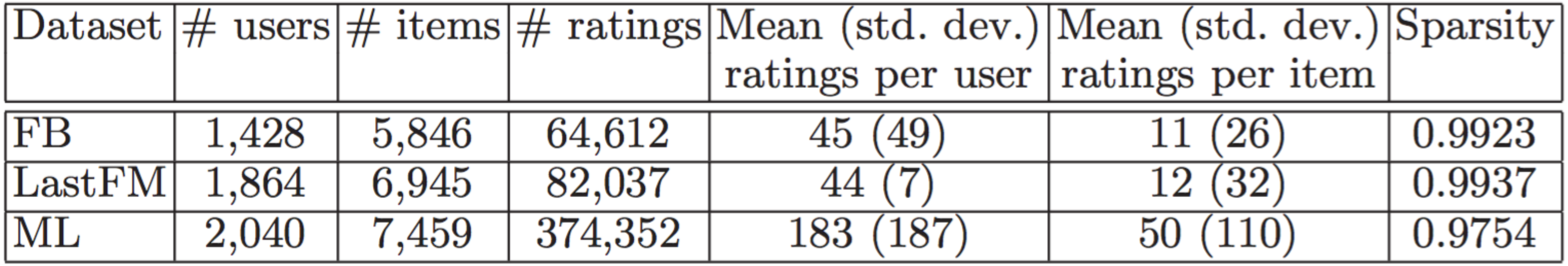
- UKNN
- IKNN
- WMF

- ACCURACY OPTIMISATION

EVALUATION

- ACCURACY
- BEYOND ACCURACY
- SIGNIFICANCE

THE DATASETS



Dataset	# users	# items	# ratings	Mean (std. dev.) ratings per user	Mean (std. dev.) ratings per item	Sparsity
FB	1,428	5,846	64,612	45 (49)	11 (26)	0.9923
LastFM	1,864	6,945	82,037	44 (7)	12 (32)	0.9937
ML	2,040	7,459	374,352	183 (187)	50 (110)	0.9754

Table 1: Summary statistics for the datasets after pre-processing.

FACEBOOK DATASET	MUSIC / BANDS
LASTFM - HETREC	MUSIC / BANDS
MOVIELENS - HETREC	MOVIES

THE ALGORITHMS

USER BASED COLLABORATIVE FILTERING (UKNN)

- Find similar users
- word of mouth
- The neighbours paradigm
- Scales with number of users

ITEM-BASED COLLABORATIVE FILTERING (IKNN)

- Find similar items
- Scalable
- Widely used

MATRIX FACTORISATION (WEIGHTED)

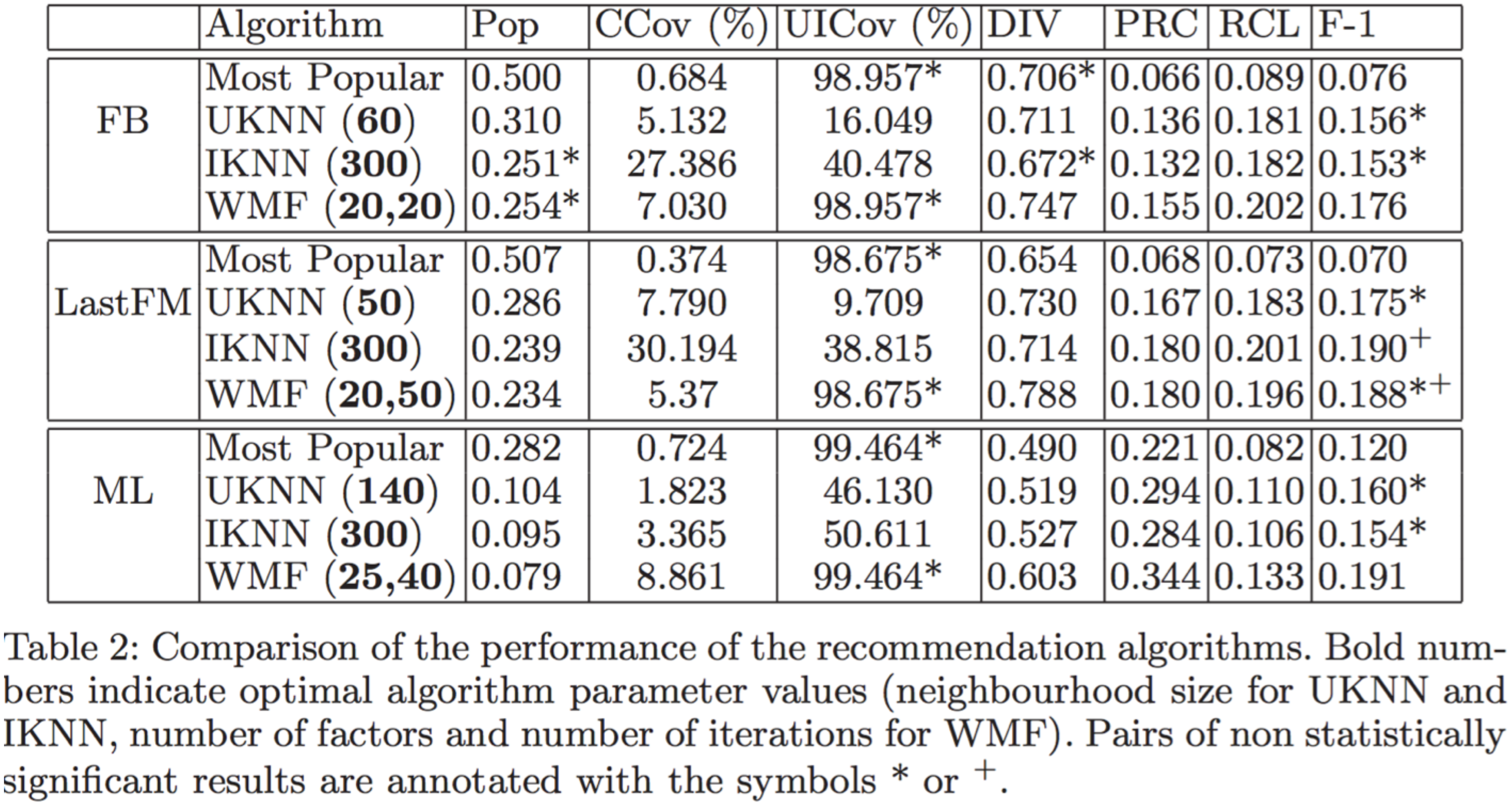
- Latent Factors
- Really good accuracy
- Scalable
- Parallel computing
- Very accurate

EVALUATION METRICS

- **PRECISION**: Out of the items recommended, how many are good recommendations?
- **RECALL**: How many of the items the user likes are being recommended?
- **F-1**: Mixes the properties of Precision and Recall into a single metric

- **DIVERSITY**: How different are the items in the list of the recommendations?
- **POPULARITY:** How popular are the items recommended
- **(PER USER) ITEM COVERAGE:** Proportion of items that are *candidates* for recommendations
- **CATALOG COVERAGE:** The proportion of items of the catalog that ever get recommended
- **UNIQUENESS:** How many items in two recommendation lists are different from each other?

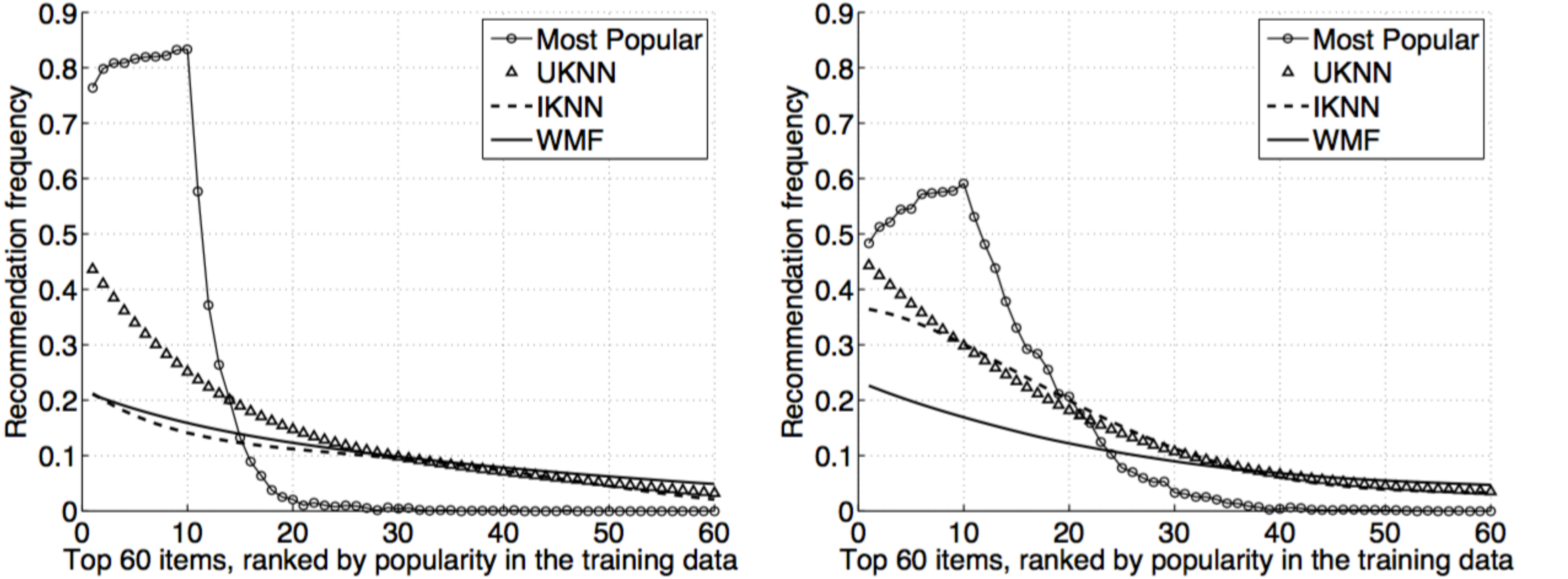
RESULTS



	Algorithm	Pop	CCov (%)	UICov (%)	DIV	PRC	RCL	F-1
FB	Most Popular	0.500	0.684	98.957*	0.706*	0.066	0.089	0.076
	UKNN (**60**)	0.310	5.132	16.049	0.711	0.136	0.181	0.156*
	IKNN (**300**)	0.251*	27.386	40.478	0.672*	0.132	0.182	0.153*
	WMF (**20,20**)	0.254*	7.030	98.957*	0.747	0.155	0.202	0.176
LastFM	Most Popular	0.507	0.374	98.675*	0.654	0.068	0.073	0.070
	UKNN (**50**)	0.286	7.790	9.709	0.730	0.167	0.183	0.175*
	IKNN (**300**)	0.239	30.194	38.815	0.714	0.180	0.201	0.190^{+}
	WMF (**20,50**)	0.234	5.37	98.675*	0.788	0.180	0.196	0.188*$^{+}$
ML	Most Popular	0.282	0.724	99.464*	0.490	0.221	0.082	0.120
	UKNN (**140**)	0.104	1.823	46.130	0.519	0.294	0.110	0.160*
	IKNN (**300**)	0.095	3.365	50.611	0.527	0.284	0.106	0.154*
	WMF (**25,40**)	0.079	8.861	99.464*	0.603	0.344	0.133	0.191

Table 2: Comparison of the performance of the recommendation algorithms. Bold numbers indicate optimal algorithm parameter values (neighbourhood size for UKNN and IKNN, number of factors and number of iterations for WMF). Pairs of non statistically significant results are annotated with the symbols * or $^{+}$.

RESULTS - POPULARITY BIAS



(a) Facebook dataset

(b) MovieLens dataset

Figure 1: Recommendation frequency of the 60 most popular items. For clarity, *UKNN*, *IKNN* and *WMF* are approximated by a 5-degree polynomial function.

RESULTS - OTHER PROPERTIES

- **Accuracy**: WMF performs best in terms of F-1 for the Facebook and MovieLens datasets, while the accuracy of the UKNN and IKNN algorithms are similar.

- **Per-user item coverage**
  - WMF algorithm considers almost every item as a candidate (UICov > 98%).
  - The UKNN algorithm (by definition) only items which are in the user’s neighbourhood can be considered as recommendation candidates. IKNN was seen to outperform UKNN in all datasets in terms of

- **Coverage**: the IKNN algorithm, performs significantly better than the other algorithms, covering up to 30% of the item catalog - Up to 6 times more items than the UKNN and WMF algorithms.

- **Diversity:** the WMF algorithm performs better, with a performance around 9% higher on average than the best neighbourhood-based approach

RESULTS - CONSISTENCY

- Important to evaluate in different datasets.
- MovieLens dataset, (3 times more dense than the Facebook and LastFM datasets), the catalog coverage of the IKNN algorithm is ~ 10 times smaller than for the LastFM and Facebook datasets.