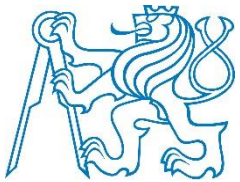
Recommender Systems

Ing. Tomáš Řehořek

tomas.rehorek@fit.cvut.cz

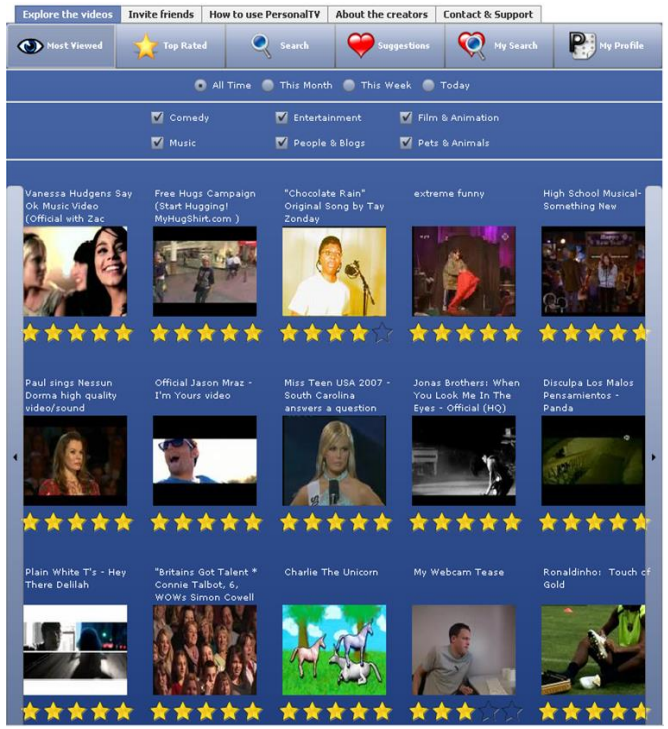


Computational Intelligence Group (CIG)
Department of Theoretical Computer Science
Faculty of Information Technology (FIT)
Czech Technical University (CTU) in Prague

Modgen, s.r.o.

Recommender Systems: Quick Intro

- Software systems providing **suggestions** for **items** to be of use to a **user**
  - Based on her past behavior (item purchases, ratings with stars, views…)
  - Based on behavior of other users of the system



Content-based Recommendation

- Uses meta-data about users/items in the system
- Attempts to recommend items similar to those liked by the user in the past

Collaborative Filtering

- Most prominent family of algorithms in Recommender Systems
- Analyses **purchase/rating history** in the whole userbase, recommends items liked by similar users

Collaborative Filtering

	Movie 1	Movie 2	Movie 3	…	Movie *n*
User 1	2	?	?		?
User 2	?	5	4		?
User 3	4	?	1		3
⋮					
User *m*			2		

INPUT

- Set of ***users*** $U = \{u_1, \dots, u_m\}$,
- Set of ***items*** $I = \{i_1, \dots, i_n\}$,
- Set of ***ratings*** $R = \{r_1, \dots, r_k\}$
  - $r_i \in U \times I \times \mathbb{R}$ (explicit ratings) or
  - $r_i \in U \times I \times \{0,1\}$ (implicit ratings)
- ***Target user*** $u_x \in U$

OUTPUT

- ***Top-N recommendations***
  - Set $\{i_1^{rec}, \dots, i_N^{rec}\} \subset I$ of items that the user will most likely to appreciate
- ***Rating predictions***
  - Predict ratings of unknown couples from $U \times I$

Our Research

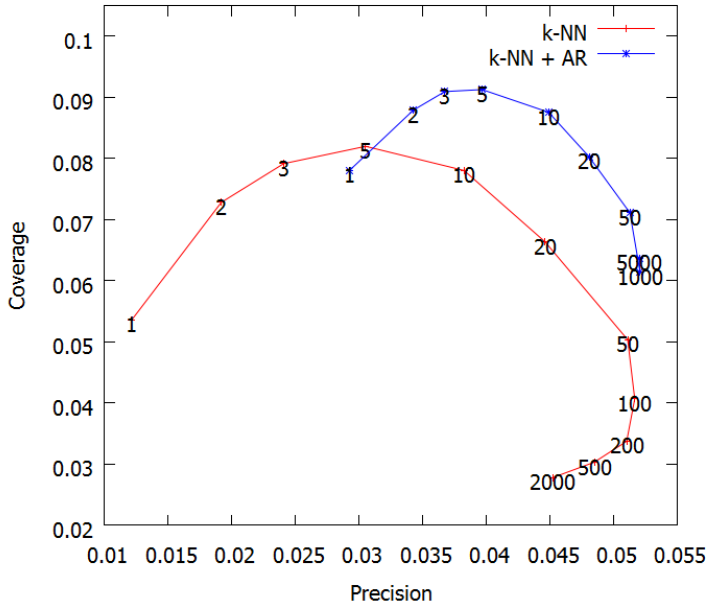
II. Matrix-Based Algorithms

- User-based *k*-Nearest Neighbors
- Item-based *k*-Nearest Neighbors
- Matrix Factorization

UserID	Item1	Item2	Item3	Item4	Item5	Item6	Item7	···
1	?	3	?	?	4	1	3	···
2	3	?	4	?	5	?	1	···
3	?	2	5	5	3	1	?	···
4	2	?	?	?	2	?	5	···
5	3	?	?	5	4	?	2	···
6	1	4	?	4	?	3	?	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

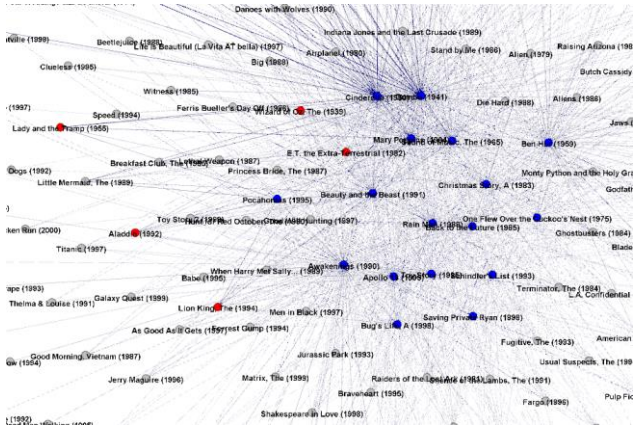
I. Multi-Objective Optimization

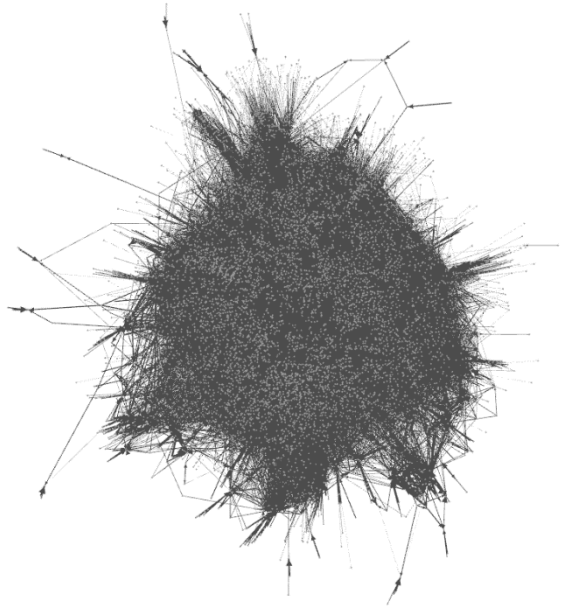
- Trade-offs between the accuracy and the diversity of items being recommender



III. Rule-Based and Graph-based Algorithms

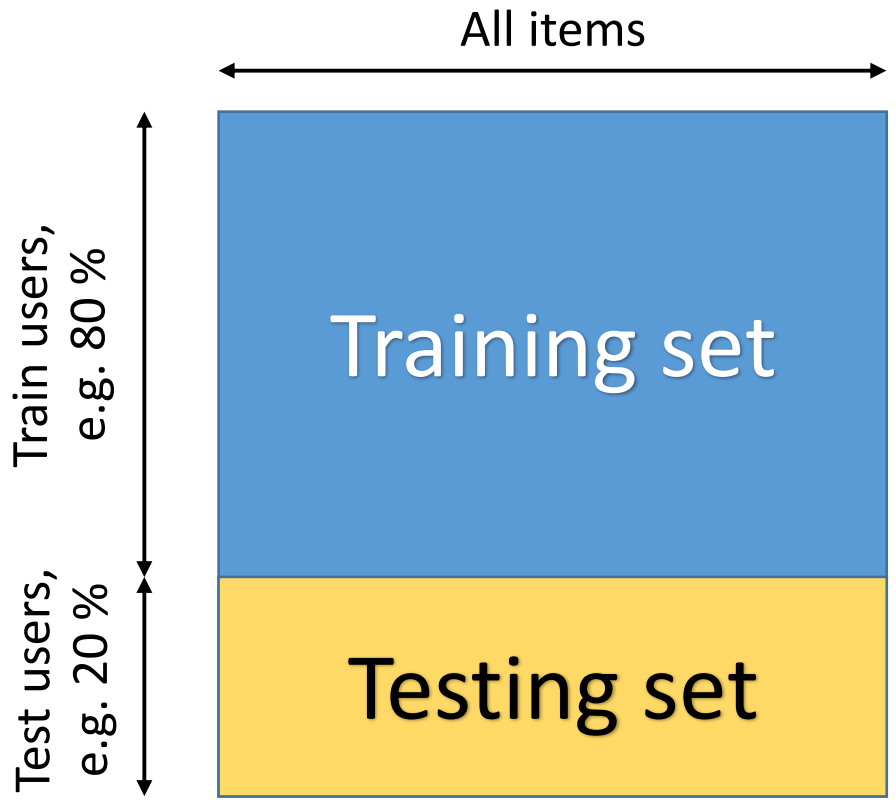
- Association Rules
- Sequential Patterns





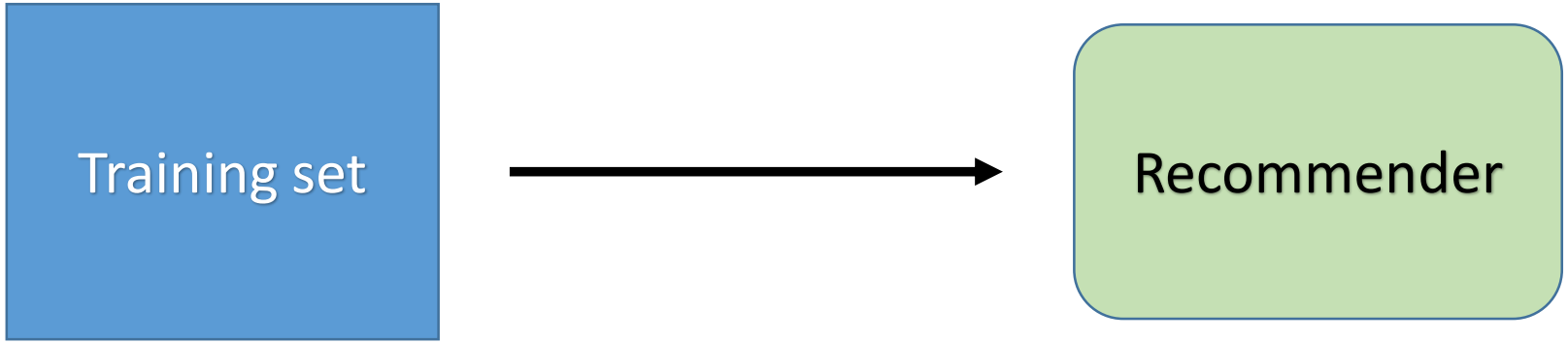
Split Validation of a Recommender System

- To validate the Recommender, one may divide the **users** into:
  - **Training set**, which is fully submitted to the Recommender
  - **Testing set**, submitted only partially and used to evaluate the Recommender

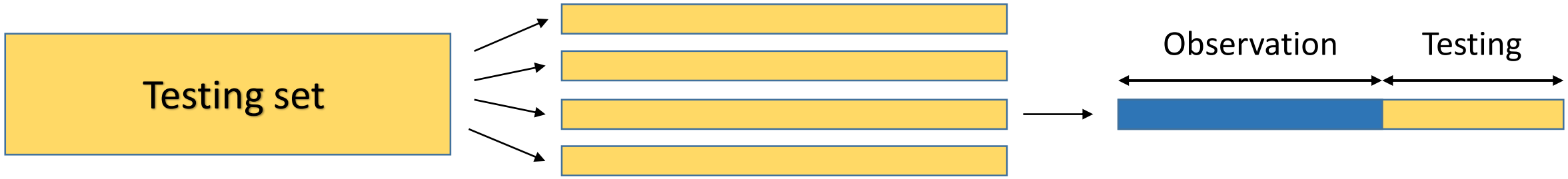


Split Validation of a Recommender System

- From the **training set**, all the items are submitted to the system

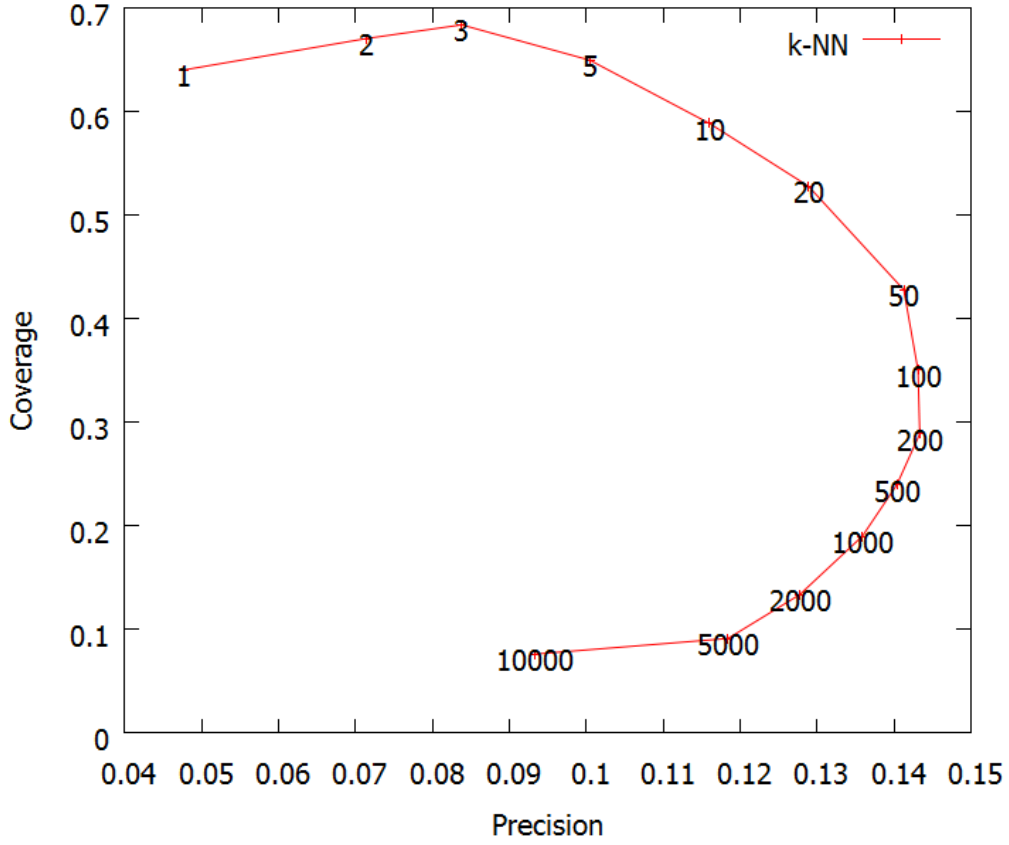


- From the **testing set**, for **each user**, the set of ratings in divided into:
  - **Observation subset** – the ratings/purchases **submitted to the system**
  - **Testing subset** – the ratings/purchases used to **evaluate the system**

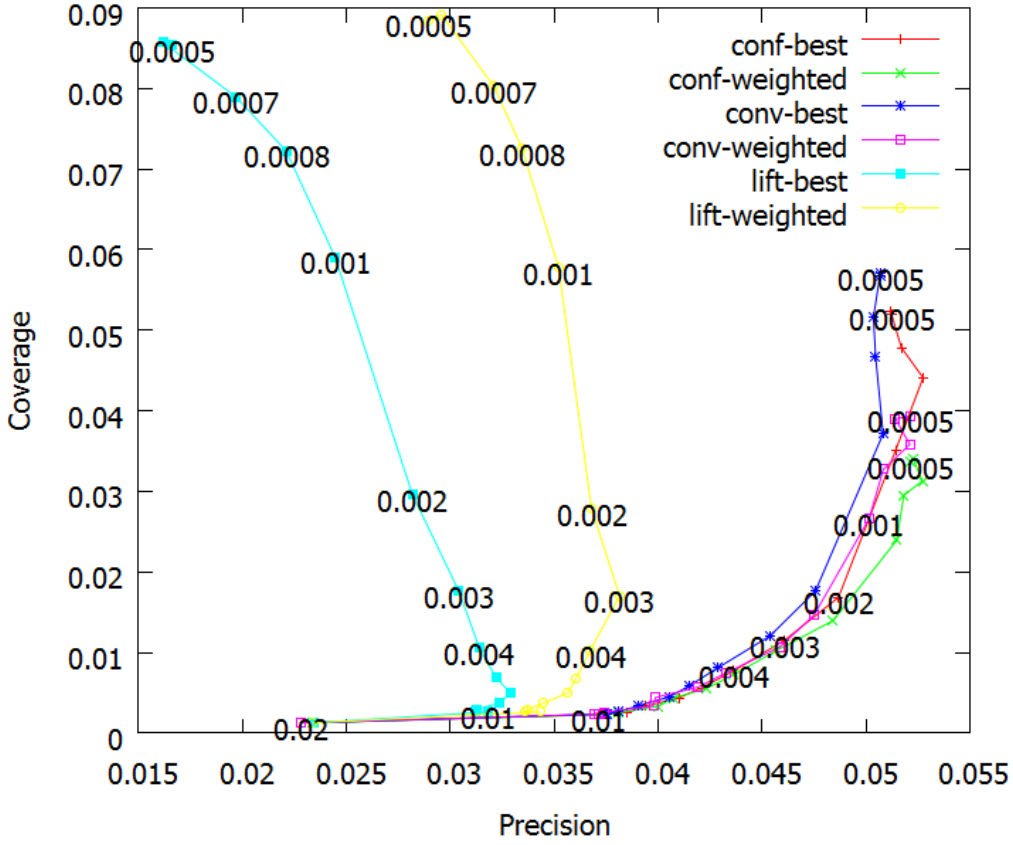


Our Research: Multi-Objective Optimization

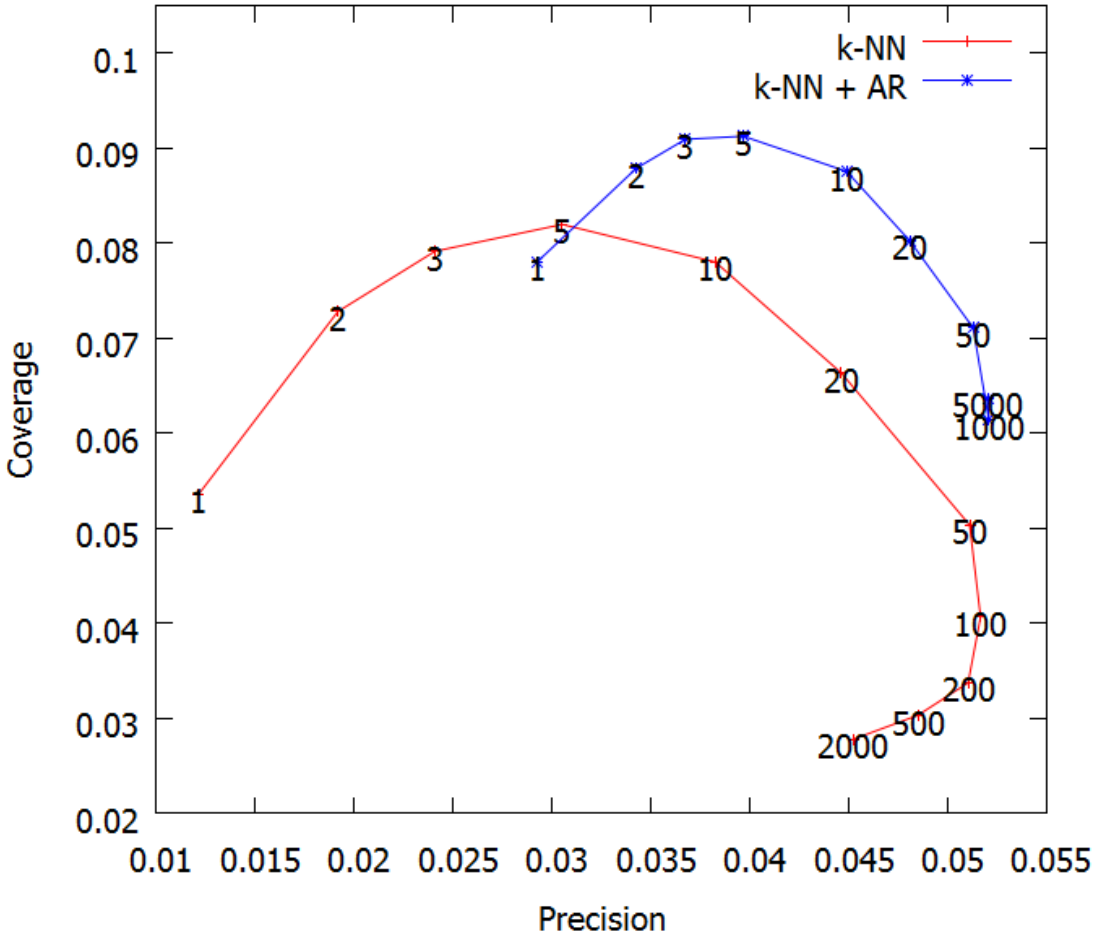
k-Nearest Neighbors

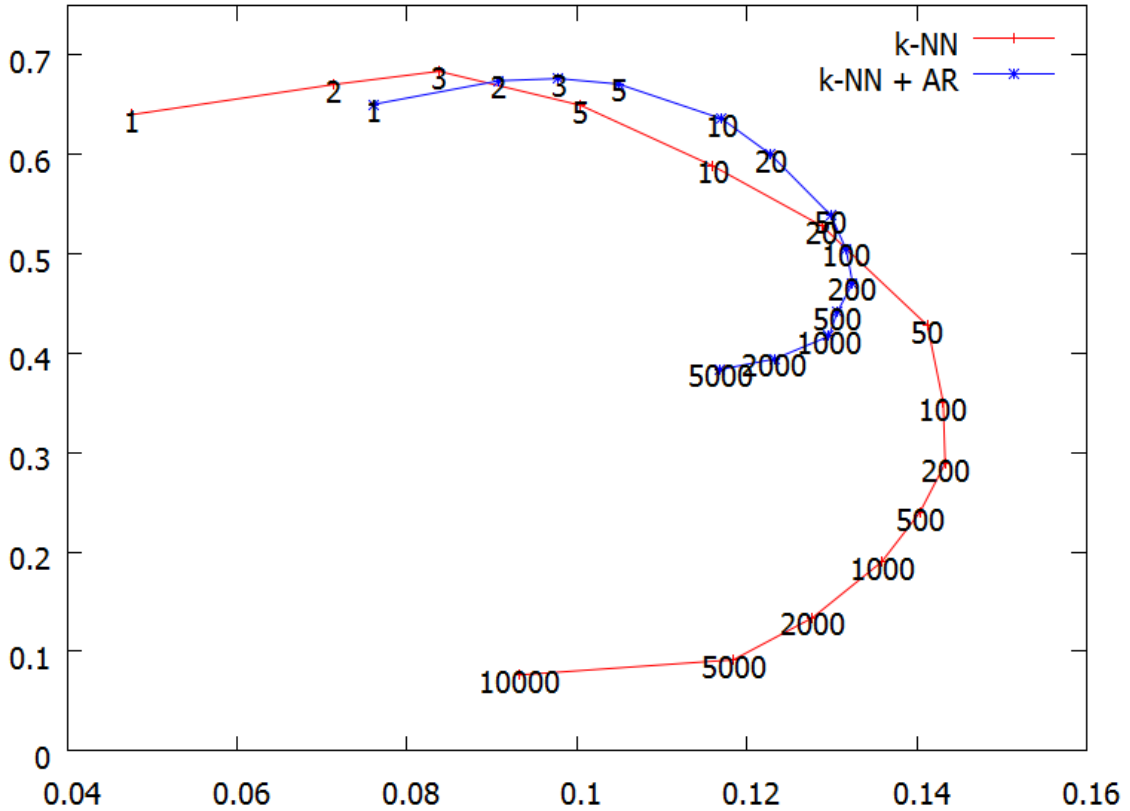


Association Rules

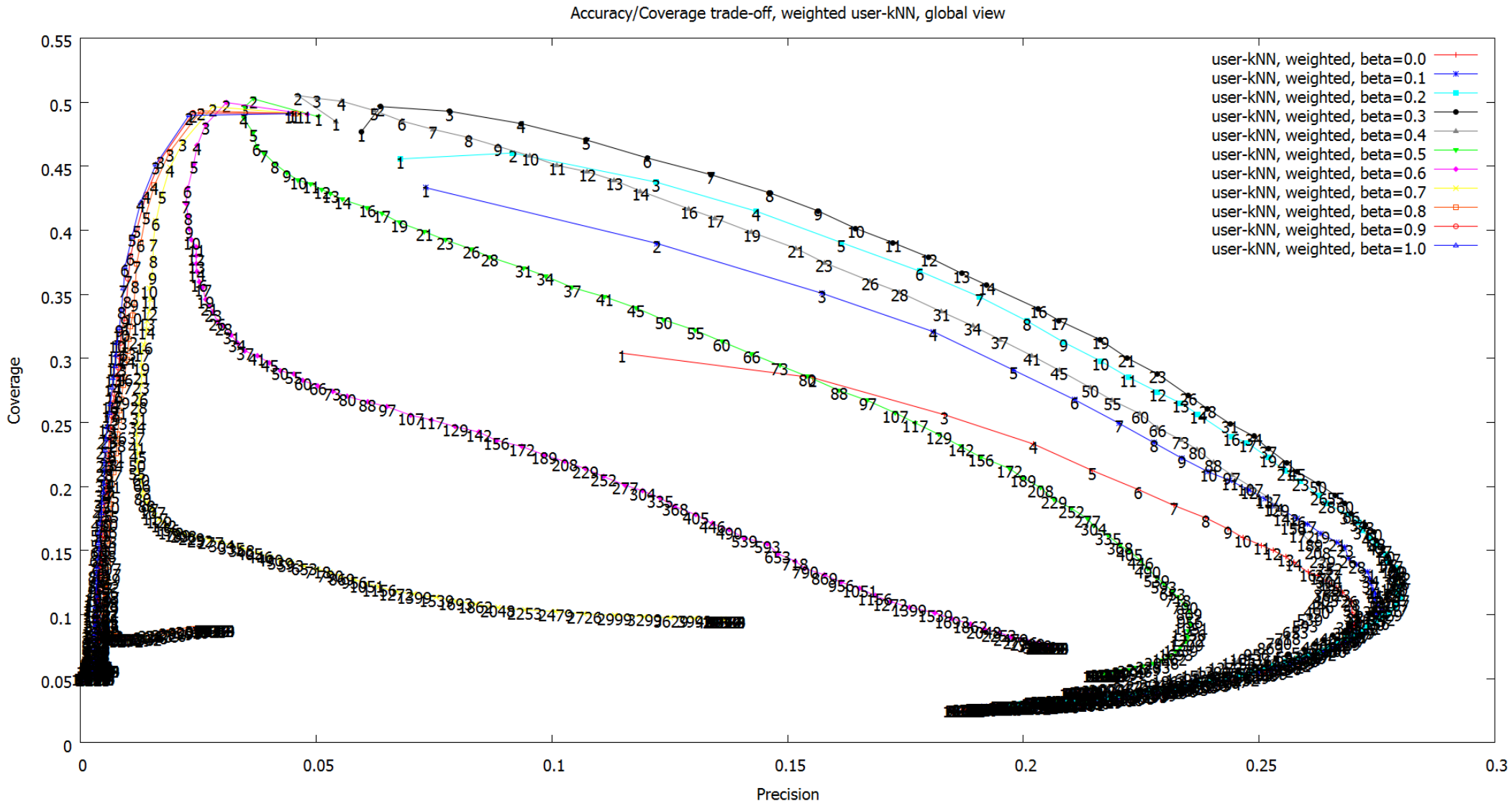


Our Research: Model Ensembles

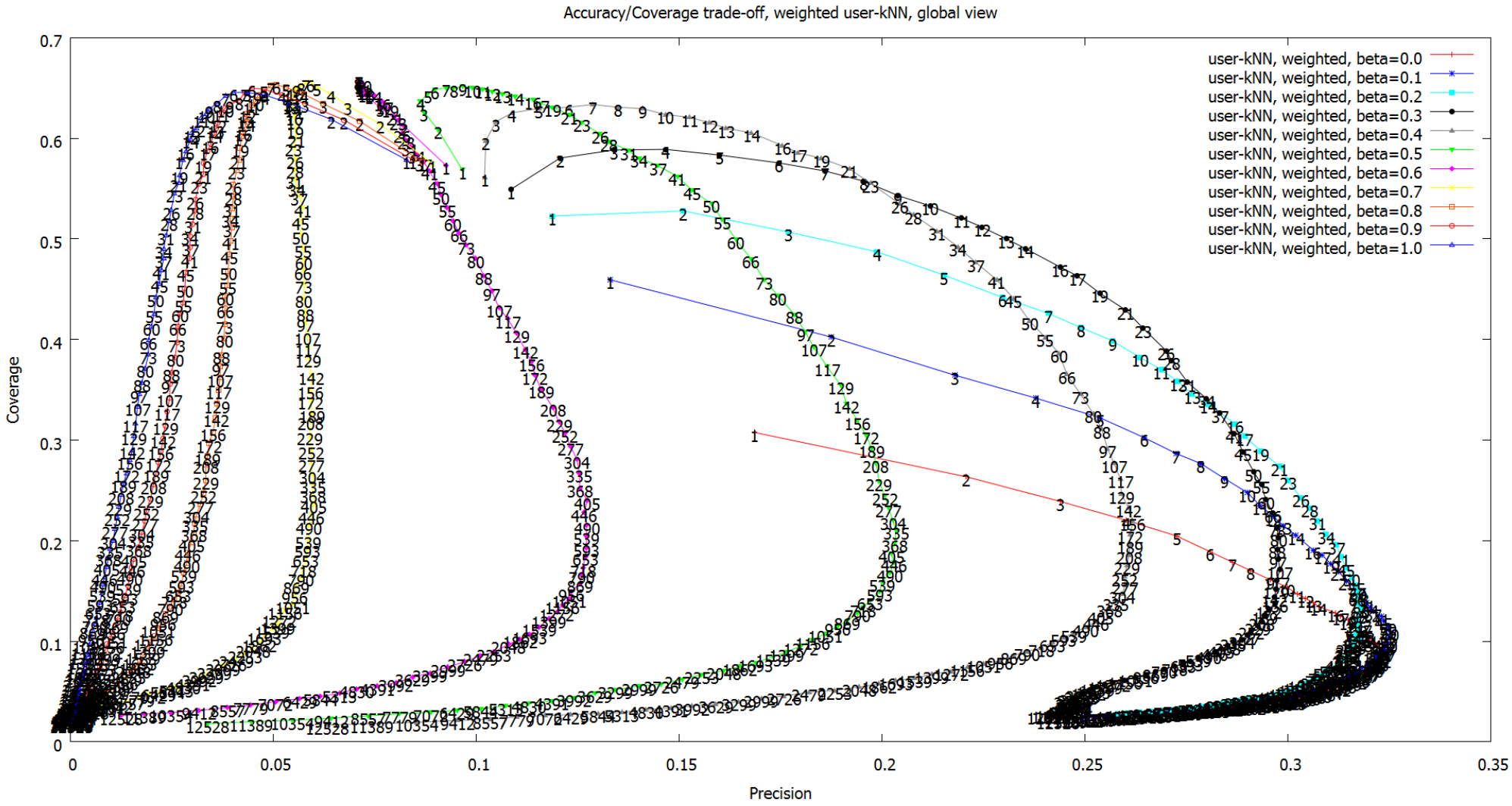




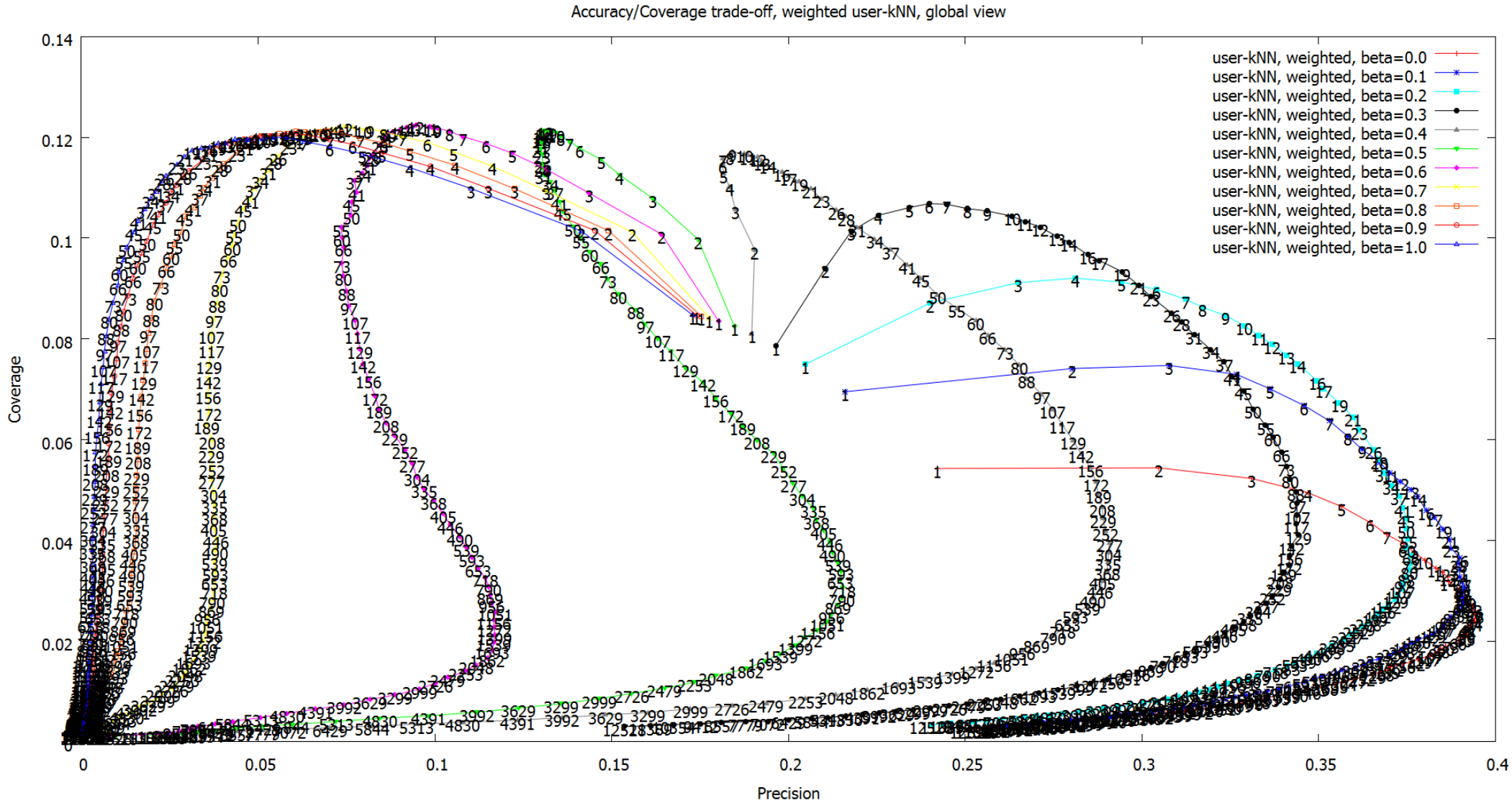
Our Research: Bestseller Penalization (MovieLens1M)



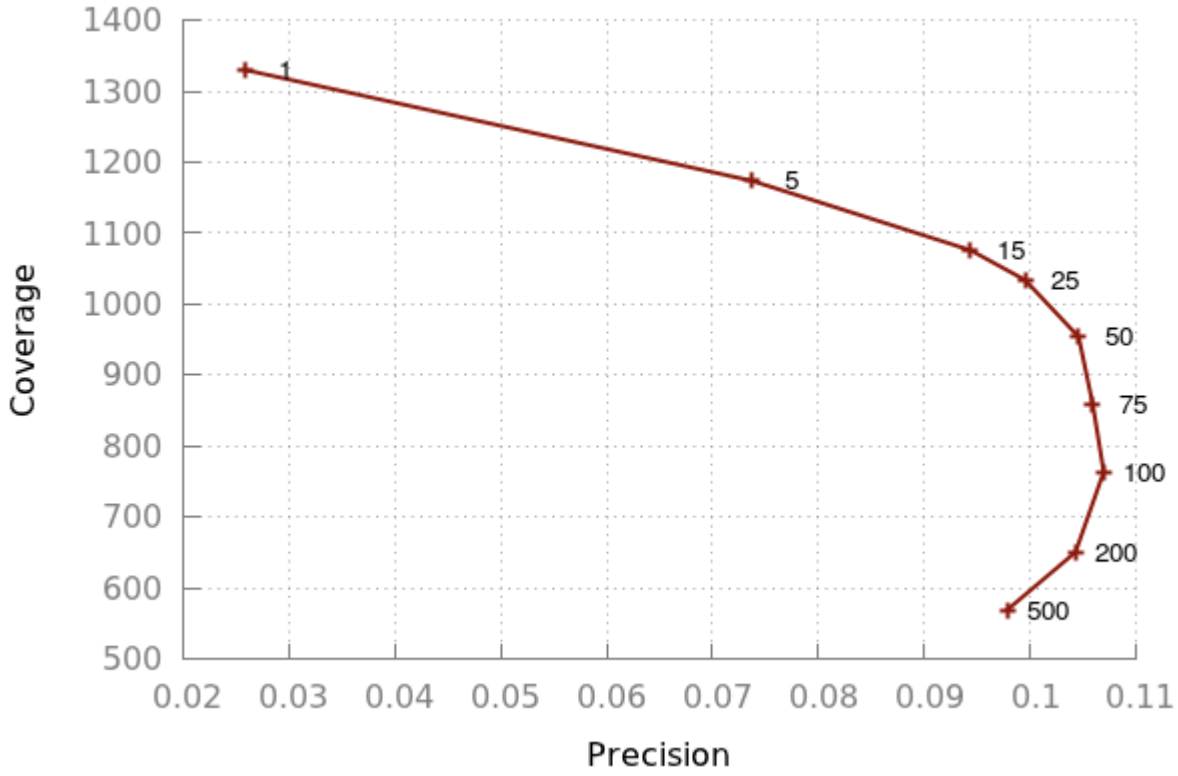
Our Research: Bestseller Penalization (MovieLens10M)



Our Research: Bestseller Penalization (Libimseti.cz)



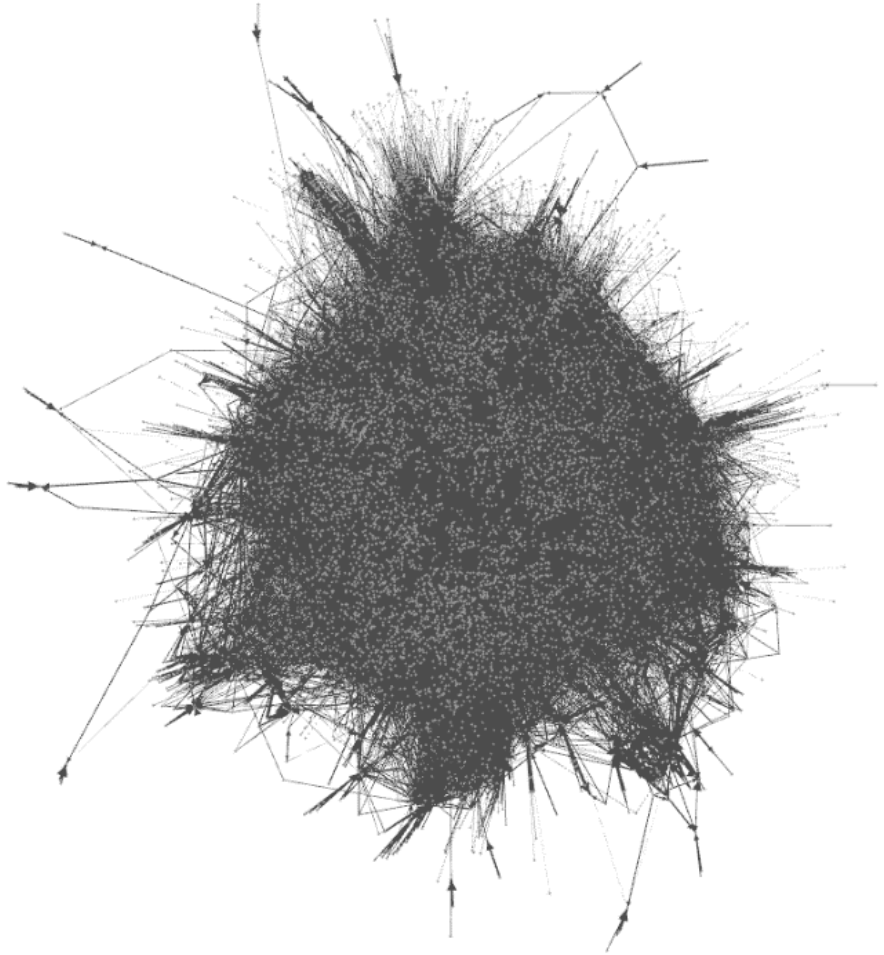
Business Impact?

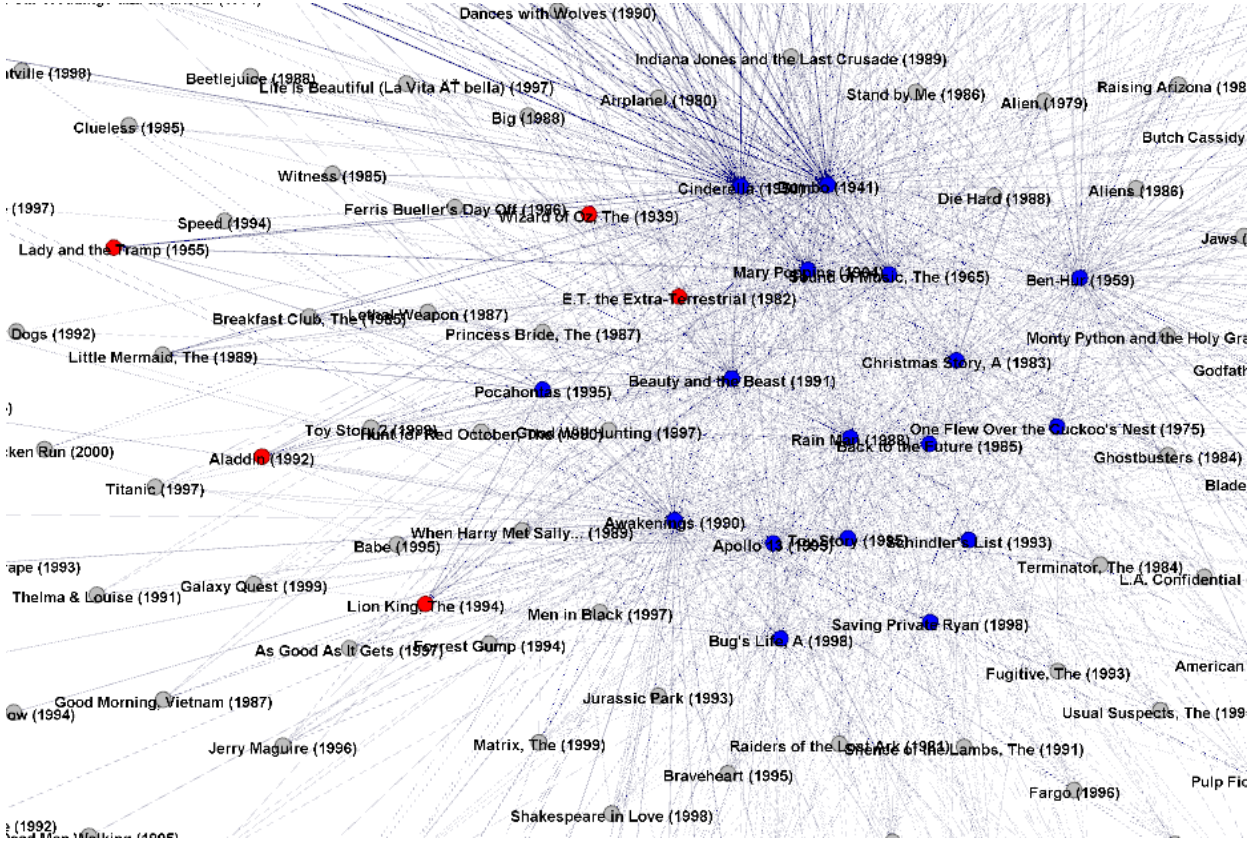


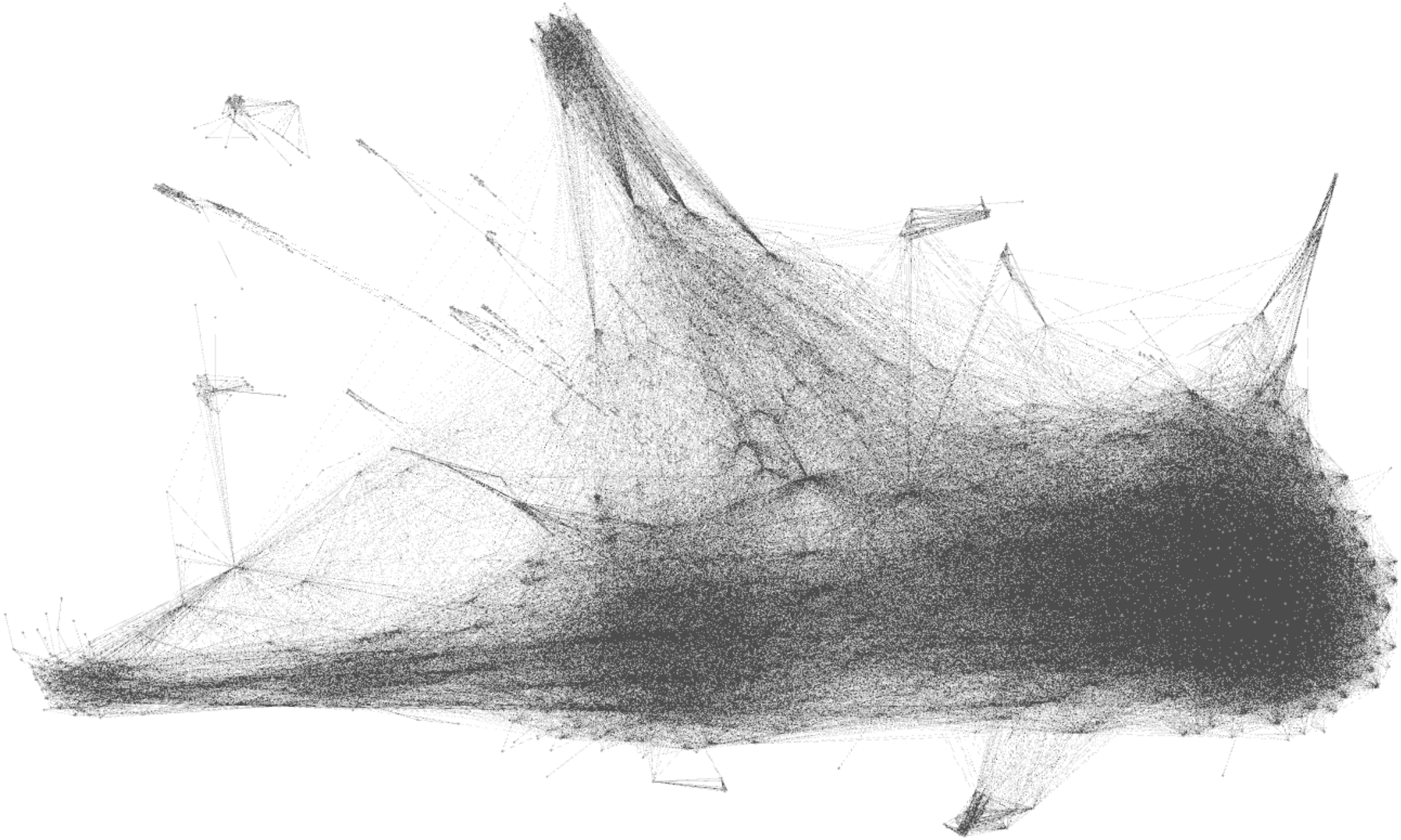
2.	n25	339	(8.79%)
3.	n100	209	(5.42%)
4.	n500	207	(5.37%)
5.	n5	120	(3.11%)
6.	old	99	(2.57%)

isthereanydeal.com

Live demo: Visualizing Association Rules







Recommendation API

- Recommendation as a cloud service
- Domain-independent
  - IPTV, VoD, Eshops, Cultural events, Sport facilities…
- Fast and scalable implementation of CF algorithms
  - Parallel Association Rules
  - Parallel k-NN
  - Parallel Matrix Factorization
- Real-time model updates
- Adjusting recommendation to fit business needs
  - Filtering, Boosting

Live demo: Boosting and Filtering

https://modgen.net:8766/goout/

User 86118115275

Items purchased

Count: **19**

!itemId	!Purchase time	!Total purchases	name	schedule	score	tag	type
event-163790	**2013-11-21 17:47:29**	**690**	Lucie, větší než malé množství lásky	1395615540.0	None	{'muzikal', 'divadlo', 'show'}	event
event-165498	**2013-11-21 17:47:53**	**42**	Strakonický dudák	1400194740.0	None	{'drama', 'divadlo'}	event
event-17737	**2013-11-21 17:48:49**	**323**	Vražda v salónním coupé	1386889140.0	None	{'top', 'komedie', 'divadlo'}	event
event-17749	**2013-11-21 17:49:03**	**308**	Drobečky z perníku	1392677940.0	None	{'divadlo', 'tragikomedie'}	event
event-17737	**2013-11-21 17:50:47**	**323**	Vražda v salónním coupé	1386889140.0	None	{'top', 'komedie', 'divadlo'}	event
event-5810	**2013-11-21 17:51:03**	**131**	39 stupňů	1386889140.0	None	{'drama', 'divadlo'}	event
event-212343	**2013-11-21 17:51:31**	**8**	Zahradní slavnost	1386889140.0	None	{'drama', 'divadlo'}	event
event-184929	**2013-11-21 17:52:10**	**225**	Jedenácté přikázání	1403737140.0	None	{'komedie', 'divadlo'}	event
event-184929	**2013-11-21 17:53:47**	**225**	Jedenácté přikázání	1403737140.0	None	{'komedie', 'divadlo'}	event
event-5810	**2013-11-21 17:53:50**	**131**	39 stupňů	1386889140.0	None	{'drama', 'divadlo'}	event
event-100659	**2013-11-21 17:53:59**	**560**	Kapka medu pro Verunku	1401663540.0	None	{'pro-deti', 'muzikal', 'pohadka', 'divadlo'}	event
event-166156	**2013-11-21 17:54:05**	**232**	Sejdeme se Na Cibulce	1387321140.0	None	{'debata', 'divadlo', 'show'}	event
event-137165	**2013-11-21 17:54:31**	**185**	Růžové brýle	1391900340.0	None	{'drama', 'divadlo'}	event
event-134719	**2013-11-21 17:54:47**	**21**	Vlny	1390258740.0	None	{'drama', 'divadlo', 'show'}	event
event-8363	**2013-11-21 17:54:57**	**31**	Příběhy ze Starého zákona II	1386889140.0	None	{'drama', 'divadlo', 'literatura'}	event
event-111031	**2013-11-21 17:55:08**	**256**	Aspects of Alice	1388444340.0	None	{'cerne', 'divadlo'}	event
event-140055	**2013-11-21 17:55:35**	**155**	Faust – Muž mezi Bohem a Ďáblem	1393109940.0	None	{'cerne', 'divadlo'}	event
event-209912	**2013-11-21 17:56:41**	**68**	Strauss Gala	1386889140.0	None	{'klasicka', 'koncerty'}	event
performer-15694	**2013-11-21 17:56:41**	**174**	Symfonický orchestr Českého rozhlasu	None	0	{'klasicka', 'kapely'}	performer

Items recommended

Filter

```
'schedule' > 1386279637 and "komedie" not in 'tag'
```

Booster

```
if "drama" in 'tag' then 2 else 1
```

Submit

Result

Count: 10
Time: 155.82 ms

!itemId	!Total purchases	name	schedule	score	tag	type
event-176612	**105**	Dva	1390863540.0	None	{'drama', 'divadlo'}	event
event-5147	**47**	César a Drana	1387234740.0	None	{'drama', 'divadlo'}	event
event-5368	**30**	Šoa	1390517940.0	None	{'drama', 'divadlo'}	event
event-11271	**30**	Poledí úděl	1390863540.0	None	{'drama', 'divadlo'}	event
event-198430	**81**	Vánoční koncert Strauss gala	1386889140.0	None	{'klasicka', 'koncerty'}	event
event-108968	**40**	Amerikana III	1400799540.0	None	{'balet', 'tanec', 'divadlo'}	event
event-142810	**113**	Babička	1390690740.0	None	{'drama', 'divadlo'}	event
event-202294	**29**	Ateliér Open Space 1	1387493940.0	None	{'workshop', 'divadlo'}	event
event-8444	**66**	Carmen	1399330740.0	None	{'opera', 'divadlo'}	event
event-30166	**42**	Česká mše vánoční – Dětská opera Praha	1387666740.0	None	{'klasicka', 'divadlo', 'koncerty'}	event

Parallel Computing Challanges in Recommender Systems

- Online Recommender Systems must satisfy many tough requirements:
  - Handling click-streams of **millions of users** in **real-time**
  - 100 ms to build recommendations based on database of 1.000.000 users?
  - Datasets for online portals are as large as 250 GB of compressed rating data!

Sample problems

Find Nearest Neighbors

Given vector $u \in \mathbb{R}^n$ and set $U = \{u'_1, \dots, u'_m\}$, find subset of k vectors from U which are most similar to u

Find Frequent Itemsets

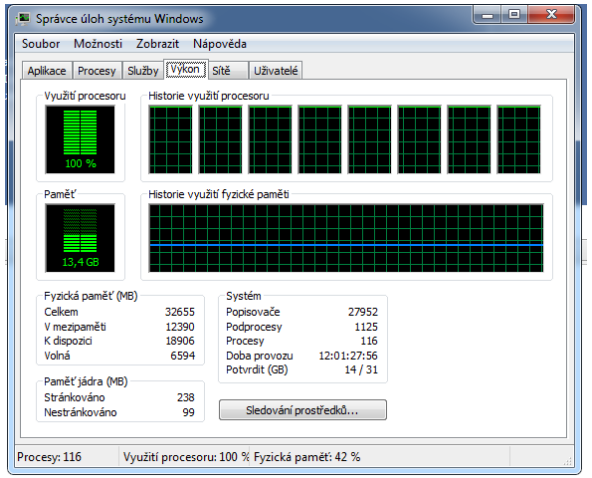
Given sets $I = \{i_1, \dots, i_n\}$ and $U = \{u_1, \dots, u_m\}$, $u_i \subseteq I$, find set $X \subseteq 2^I$ such that $x \in X$ iff

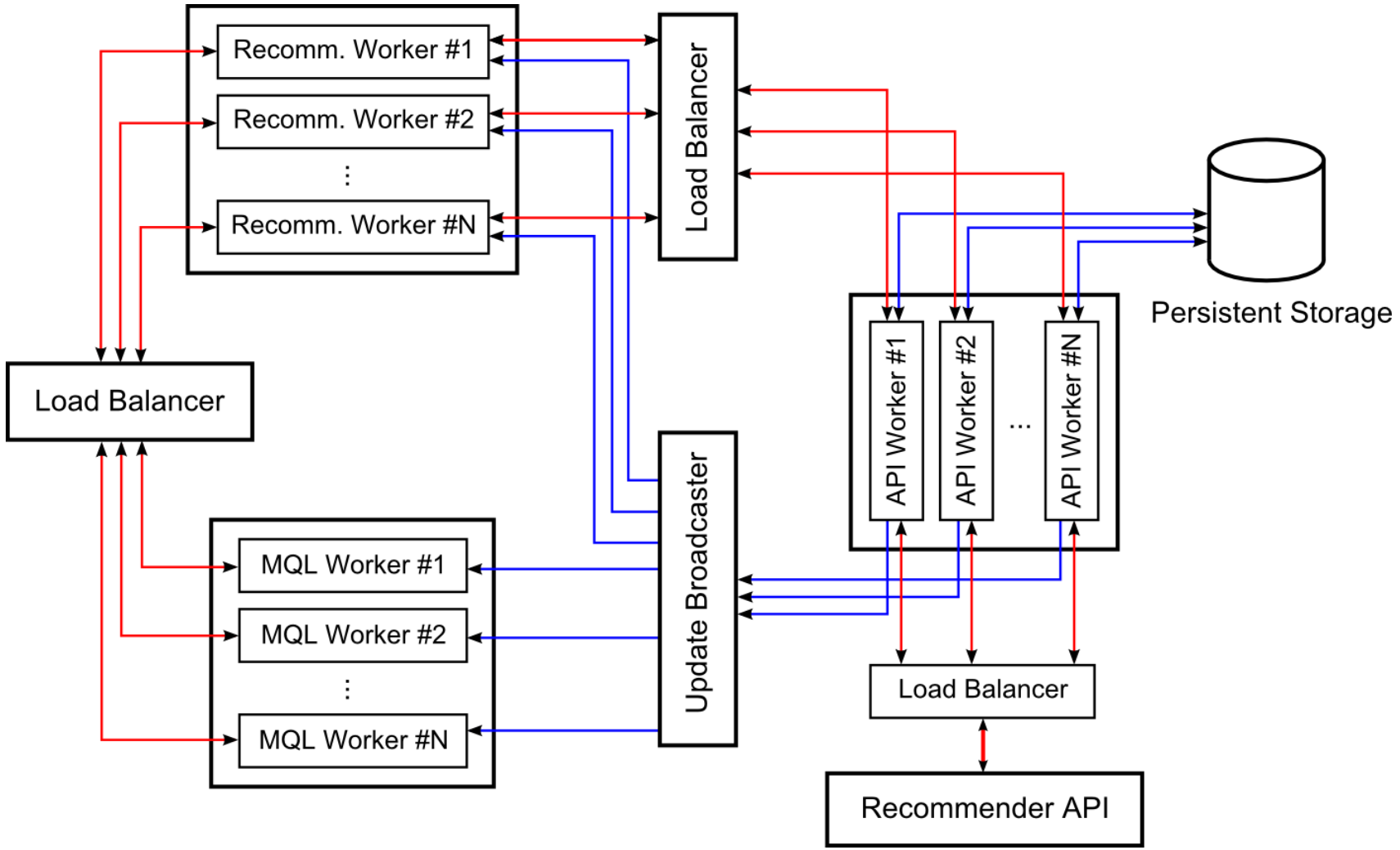
$$\frac{|\{u \in U | x \subseteq u\}|}{|I|} \geq s_{min}$$

for some $s_{min} \in [0,1]$

- for $n, m = 10^6$ or even more (Youtube, last.fm)
- within small, fixed amount of time (several 10s of milliseconds?)

Parallel Computing: Modgen Architecture





Thank you for your attention!

Ing. Tomáš Řehořek
tomas.rehorek@modgen.net