The Web Alter-Ego project

Rachid Guerraoui (EPFL) & Anne-Marie Kermarrec (Inria)

Google Focused Award

Personalization is now ubiquitous

Why is personalization challenging?

- **Huge volume** of data: small portion of interest
- Dynamic and diverse interests
- Interesting stuff does not come always from friends
- Classical notification systems do not filter enough or too much

KNN-based collaborative filtering

The Web-Alter ego project

Extracting like-minded Internet users should be a basic Web service

Goals of Web Alter-Ego : cross-apps KNN-based collaborative filtering

1. Provides an efficient scalable infrastructure
2. Provides privacy guarantees

TEAM
Nitin Chiluka (postdoc Inria)
Nupur Mittal (PhD student Inria)
Rhicheek Patra (PhD student EPFL)
Antoine Rault (PhD student Inria)
Masha Taziki (PhD student EPFL)
Jingjing Wang(PhD student EPFL)

Main results so far

[1] A. Boutet, D. Frey, R. Guerraoui, A.-M. Kermarrec, and R. Patra. *Hyrec: Leveraging browsers for scalable recommenders.* In **ACM/IFIP/USENIX MIDDLEWARE 2014**.

[2] R. Guerraoui, A.-M. Kermarrec, R. Patra, and M. Taziki. *D2P: Distance-Based Differential Privacy in Recommenders.* In Volume 8 Issue 8, **PVLDB**, 2015

[3] D. Frey, R. Guerraoui, A.-M. Kermarrec, A. Rault (INRIA) F. Taïani, J. Wang. *Hide & Share: Landmark-based Similarity for Private KNN Computation*. **IEEE/IFIP DSN 2015**

HyRec: Leveraging Browsers for Scalable Recommenders

Antoine Boutet, Davide Frey, Rachid Guerraoui, Anne-Marie Kermarrec, Rhicheek Patra

Middleware 2014

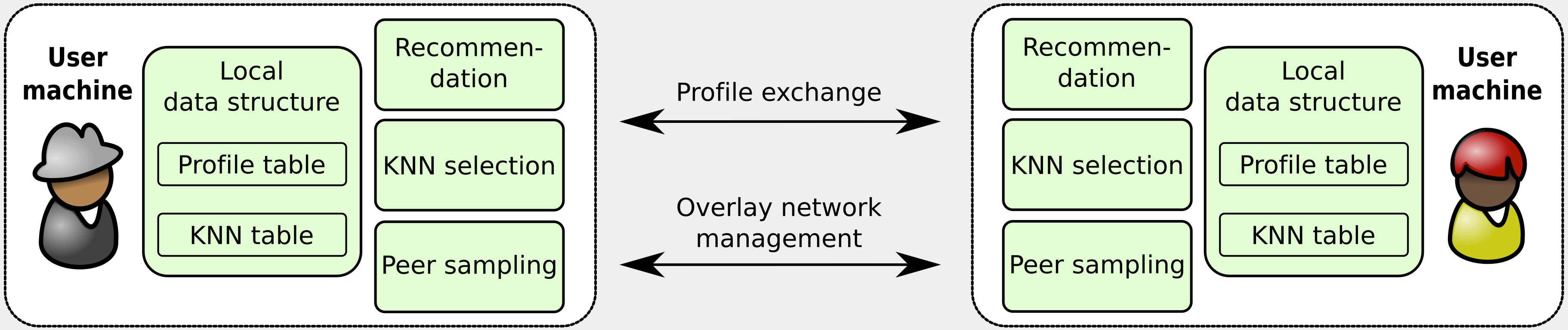
Personalization

Personalization schemes are resource greedy

- Fully decentralized systems, scalable but difficult to manage
- Centralized systems need huge computational power

Democratizing personalization is also crucial for small web content providers

Decentralized approach

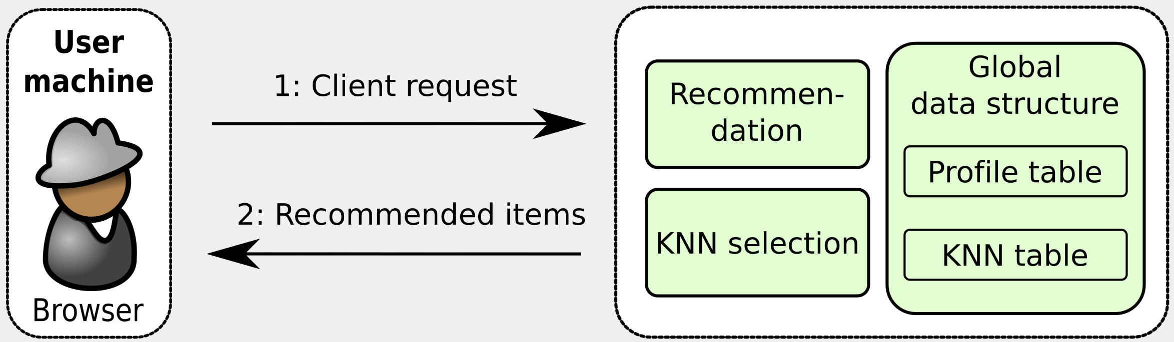


Data structures

Profile table	
uid	P(uid) = {list of iid}

KNN table	
uid	Knn(uid) ={list of uid}

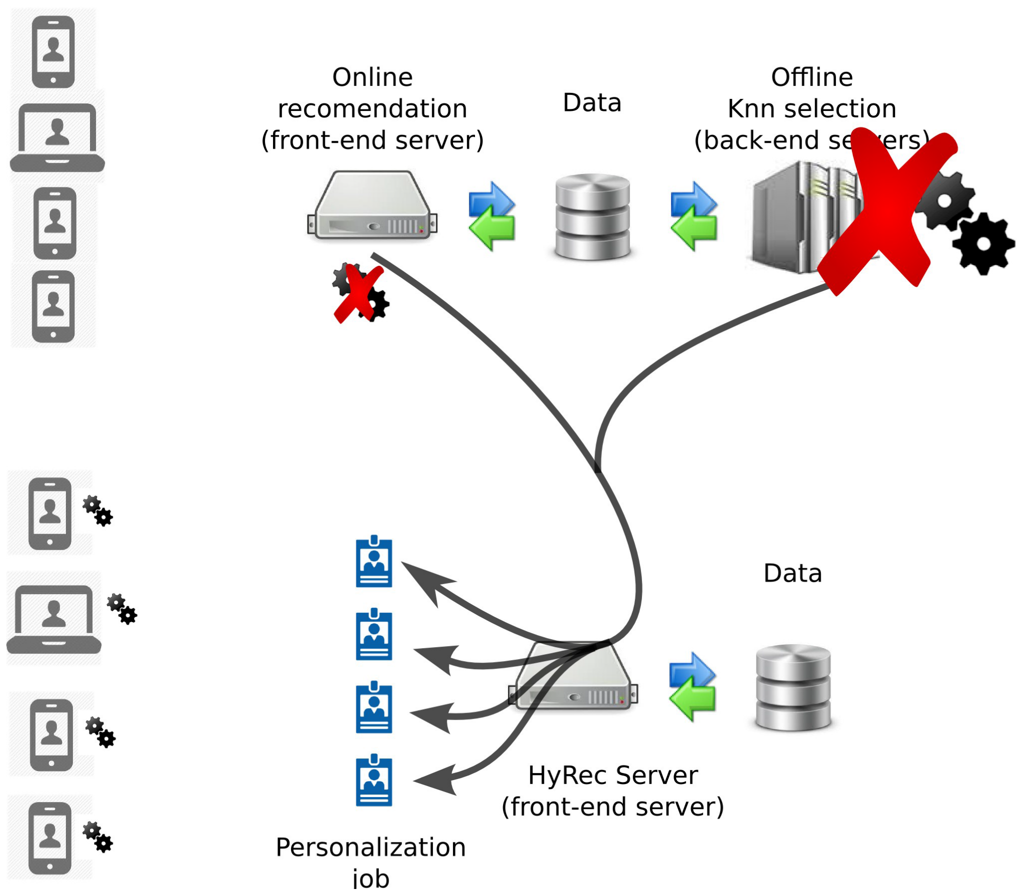
Centralized approach



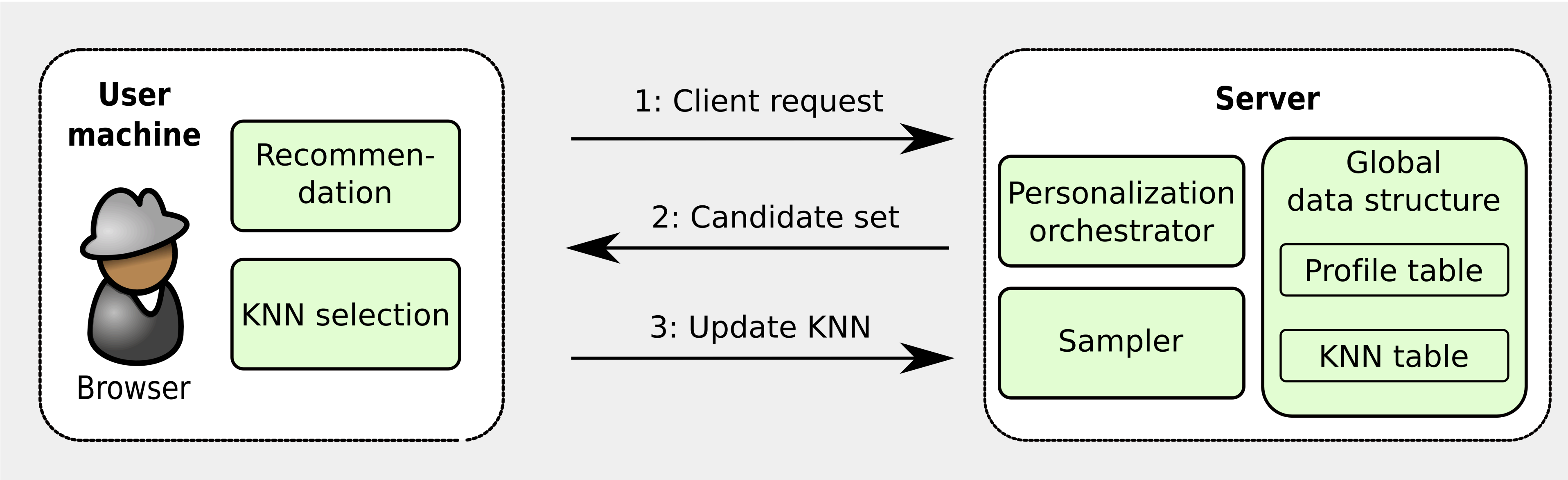
HyRec’s challenge

Traditional centralized architecture

HyRec architecture



HyRec: tasks to offload



No data stored at the client

Javascript (Interaction with the server's api)

- KNN computation
- Compute recommendations

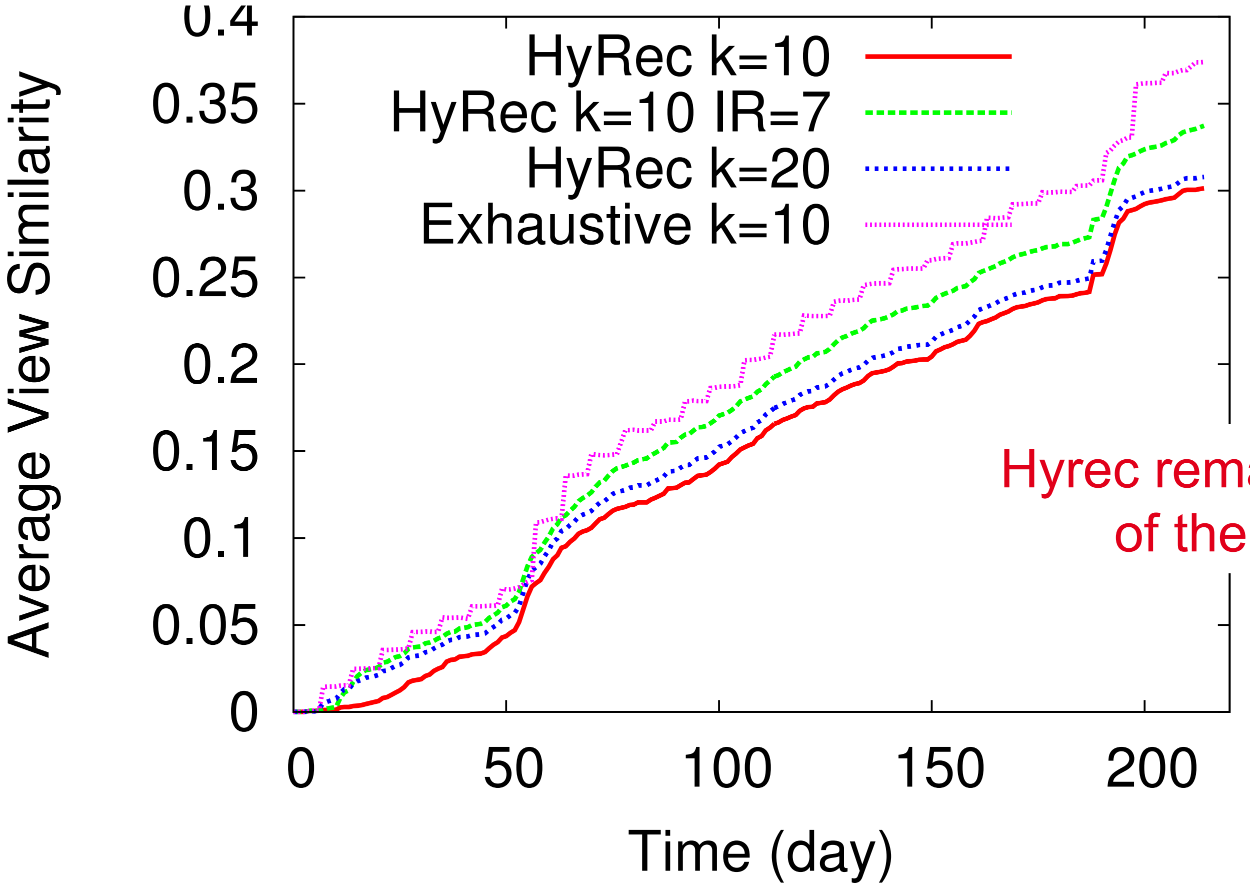
Sample: Identify the candidate set (Two-hop neighborhood + k random)

Orchestrator :

- Personalization job (json) containing profile + profiles of users in the CS
- Update the knn table

View similarity

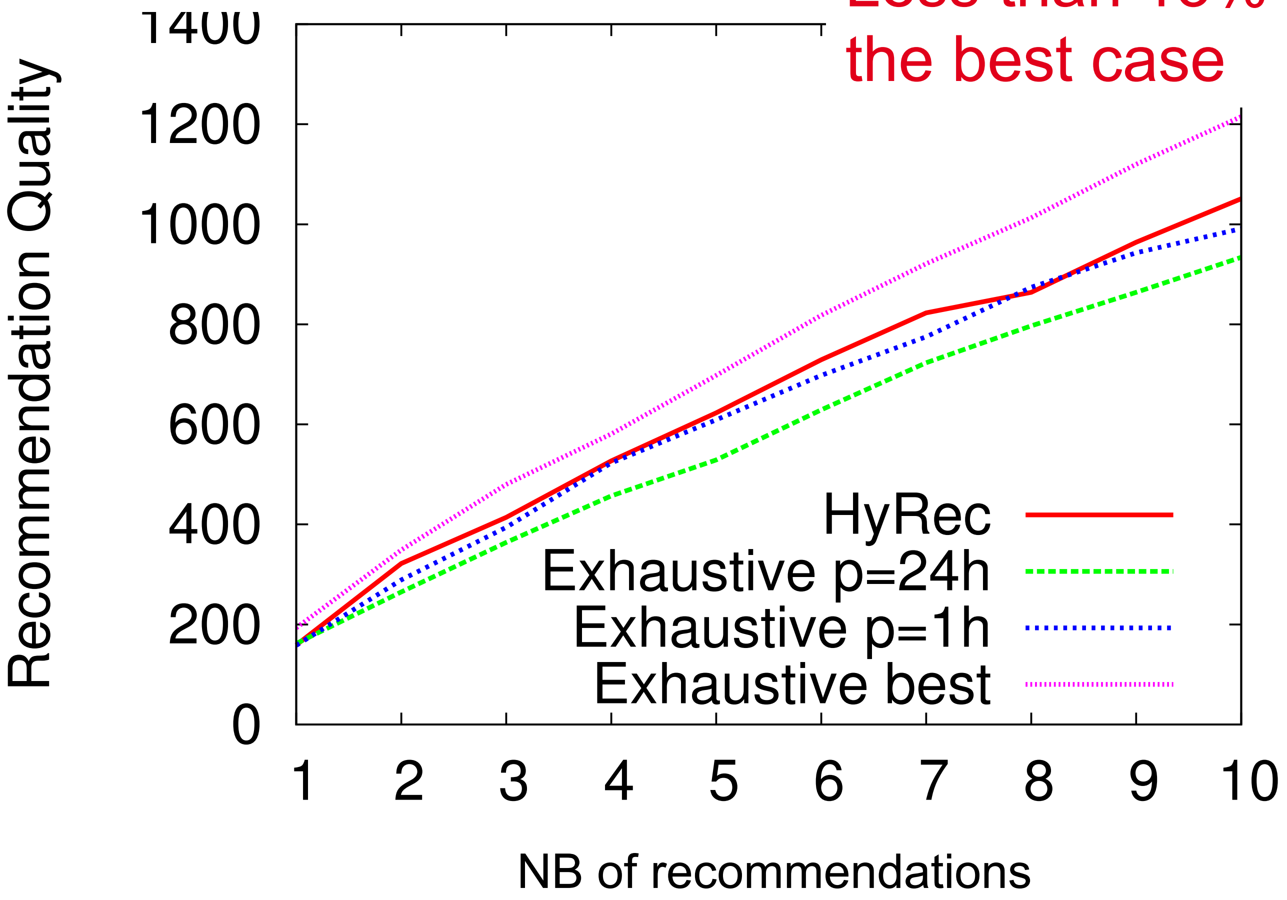
Dataset	Users	Items	Ratings
MovieLens1	943	1700	100,000
MovieLens2	6,040	4000	1,000,000
MovieLens3	69,878	10,000	10,000,000
Digg	59,167	7724	782,807



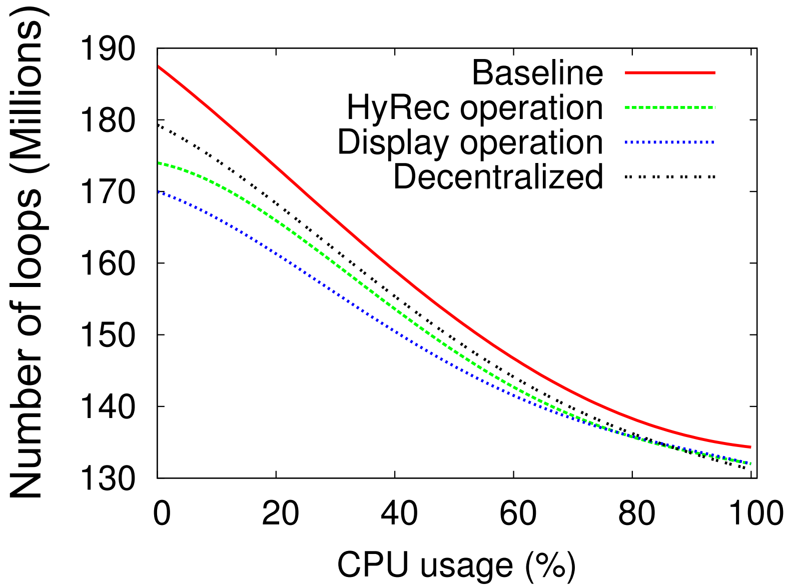
Hyrec remains within 20% of the ideal KNN

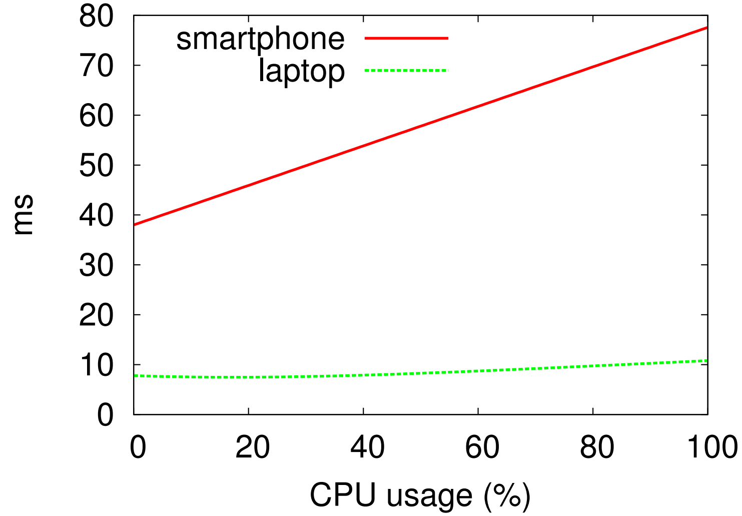
Recommendation quality

Less than 13% below the best case



HyRec versus the client load





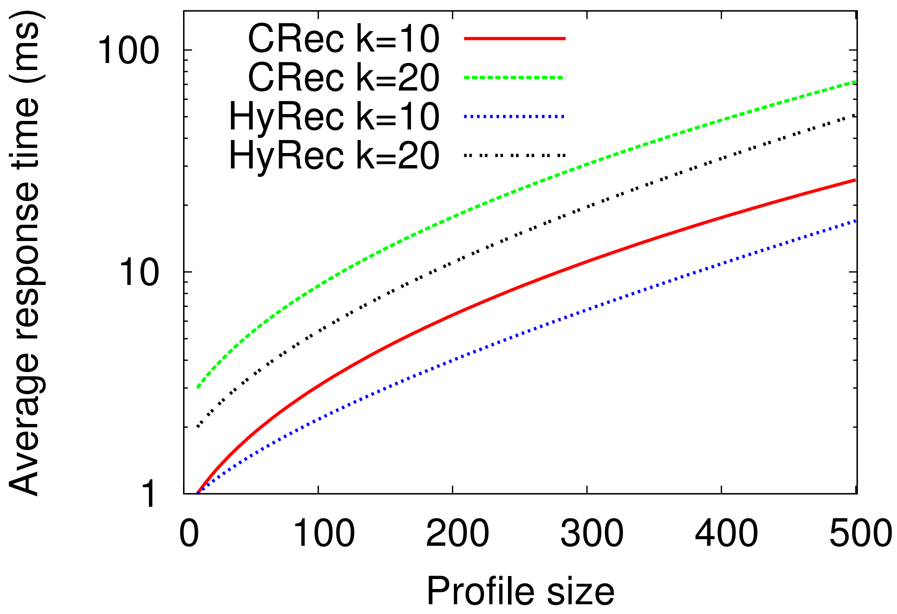
Impact of HyRec

Negligible disruption of HyRec

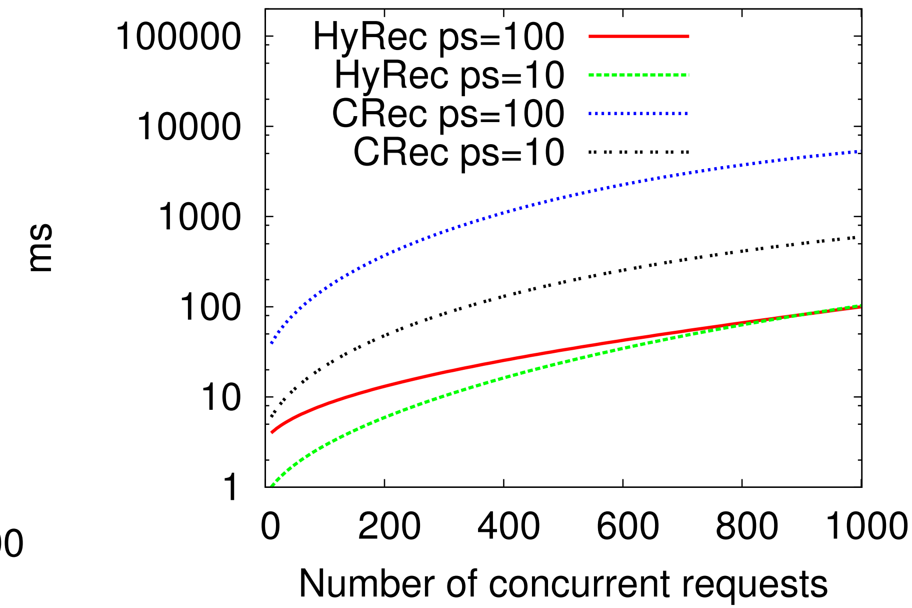
Impact of the client load

50% load
<60ms on smartphone
<10ms on laptop

HyRec versus a centralized recommender



Impact of the profile size



Impact of the number of requests

Take away message

Scalable recommendation engines

Decentralized algorithms design

Hybrid infrastructures

D2P: Distance-Based Differential Privacy in Recommenders.

R. Guerraoui, A.-M. Kermarrec, R. Patra, and M. Taziki.

VLDB 2015

About privacy

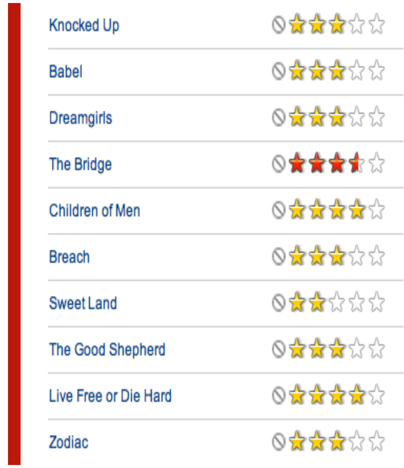
Ex: Netflix challenge 2 and IMDB (Internet Movie Database)

« privacy expert Larry Ponemon says that Netflix could have likely avoided the matter altogether by using a technique called "data masking" that would have randomized its data set while still keeping the data relevant to developers »

Problem statement

1) Collaborative filtering relies on users profiles

2) Privacy guarantees needed



D2P: Distance-based Differential Privacy protocol: probabilistic substitution techniques to create the Alter-ego profile

Differential Privacy [Dwork 2006]

$Prob(Q(D))/Prob(Q(D+/-1)) \leq e^{\varepsilon}$

$Prob(R|true\ world = D)/Prob(R|true\ world = D+-1) \leq e^{\varepsilon}$

The released result R gives minimal evidence about whether or not any given individual contributed to the data set.

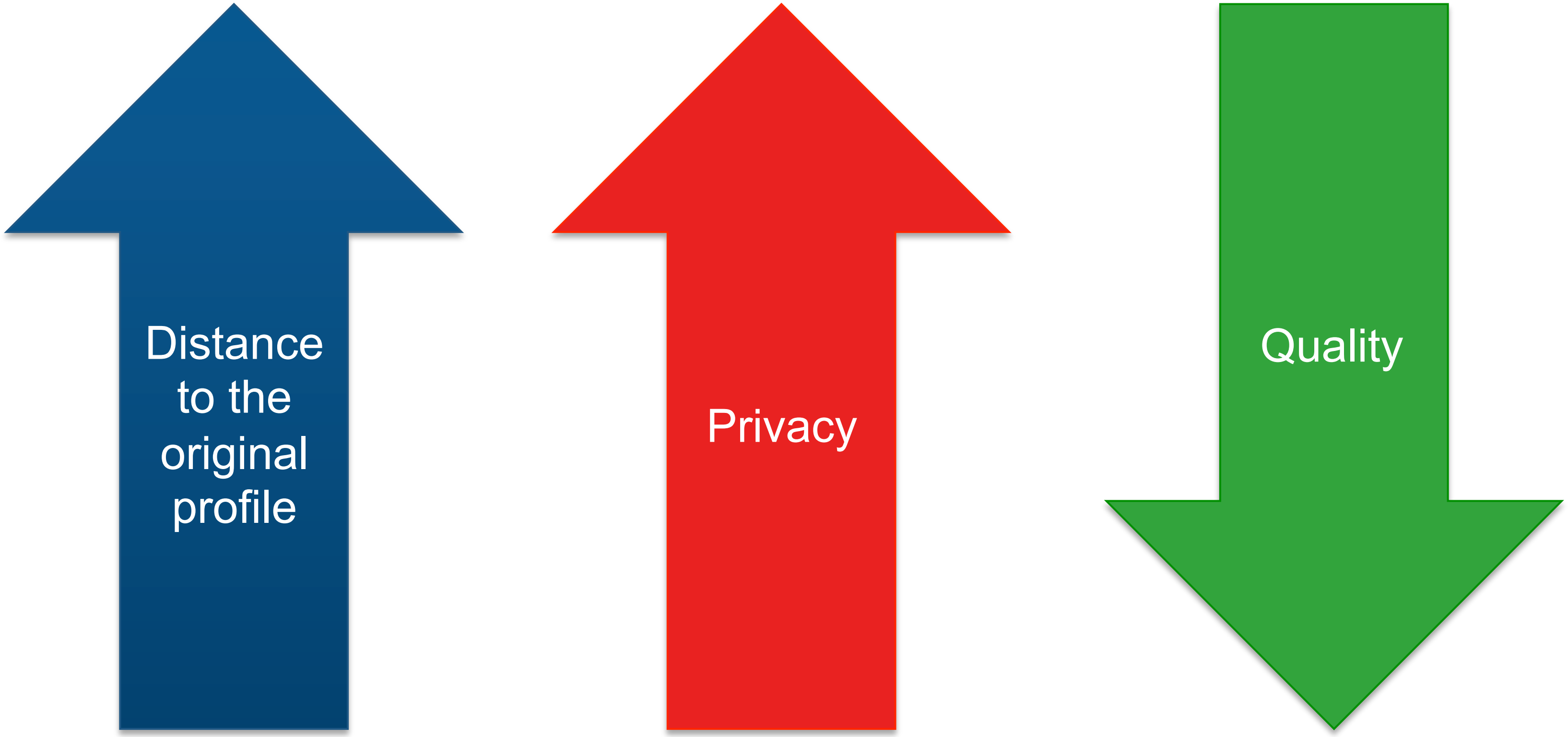
Adding (Laplacian) noise

DP2: DP applied to recommenders

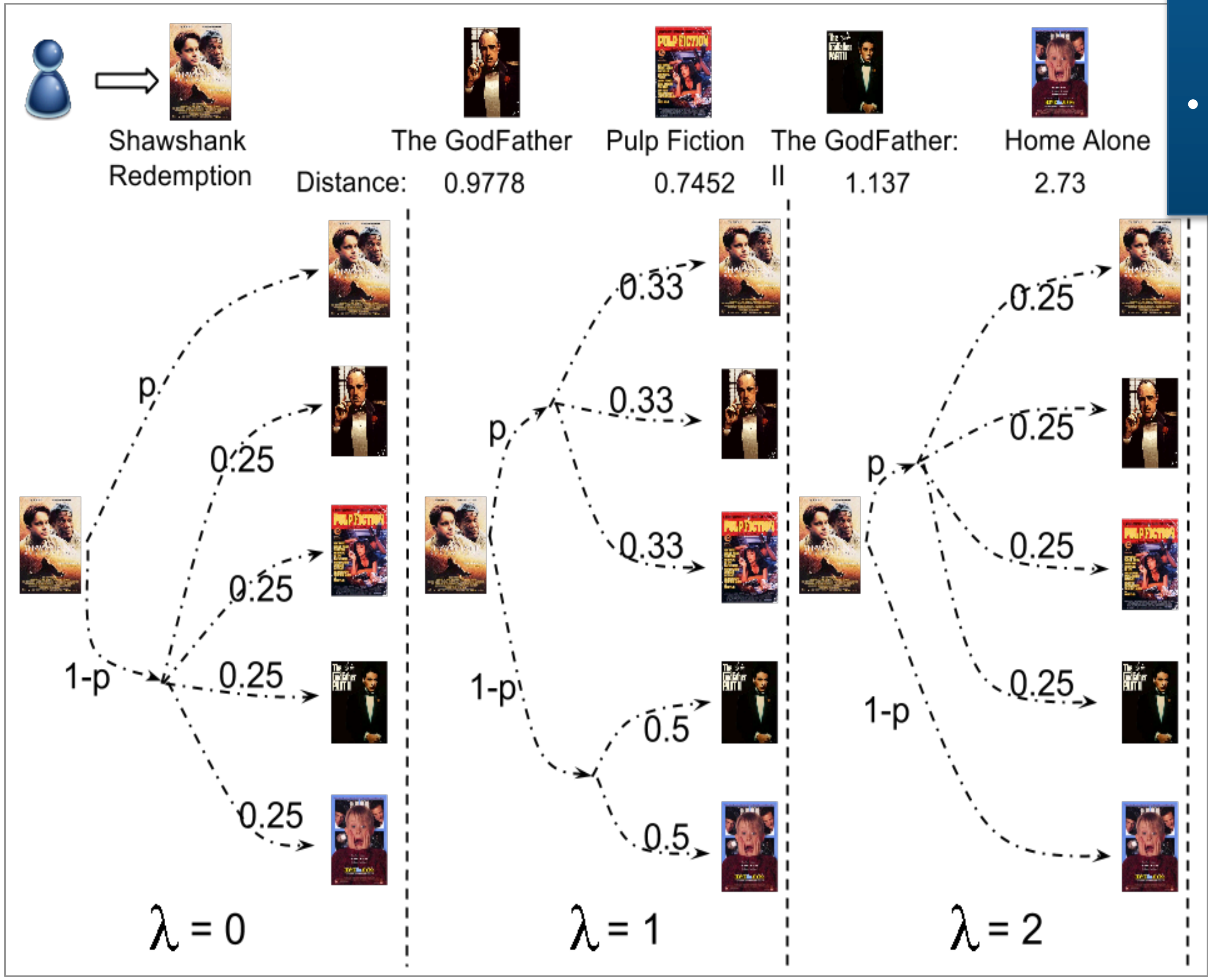
- **DP**: Avoid any user to guess, based on her recommandations whether some other users has one item *I* in her profile
- **D2P**: And any item within some distance λ from *I*

D2P builds an alter-ego profile where some items are probabilistically replaced

Technical challenge: trade-off



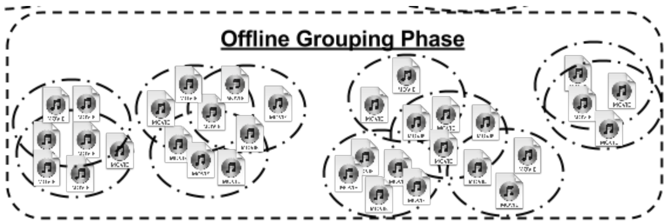
Example



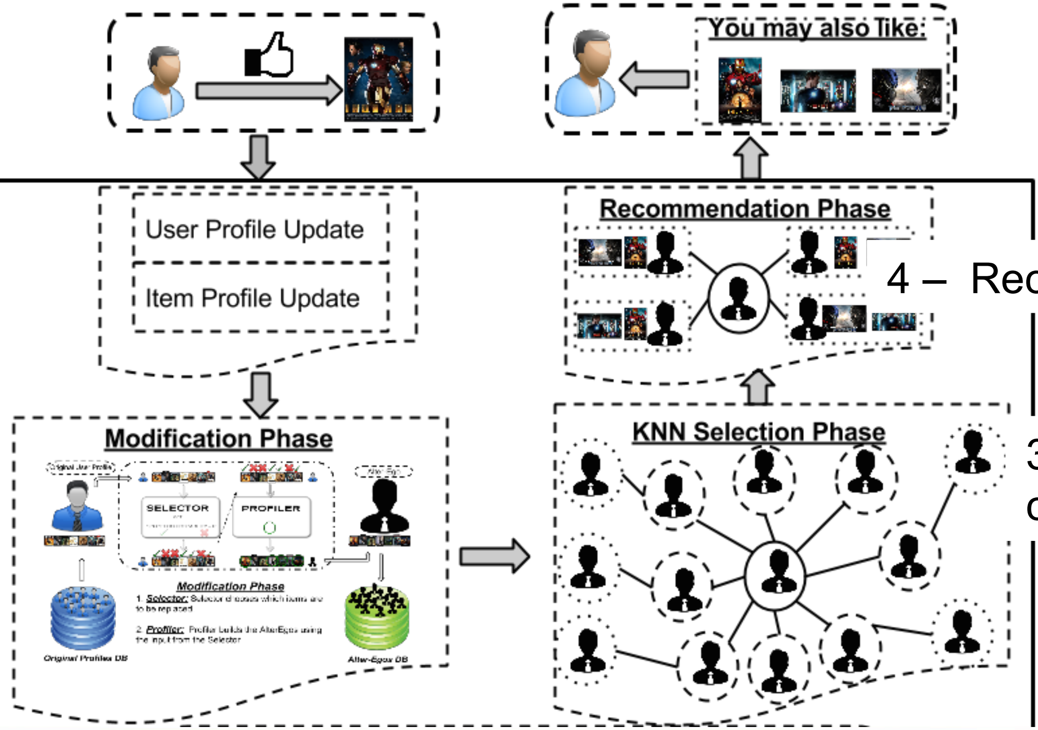
D2P selects
- movies with distance less than an upper bound with prob. p,
- random movies with prob. 1-p

D2P Recommender

1- A group G_i, contains all items with distance less than λ from i



Distance between items (i and j) = (1/cos_sim(i,j)) - 1



4 – Recommendations

3 – KNN computation

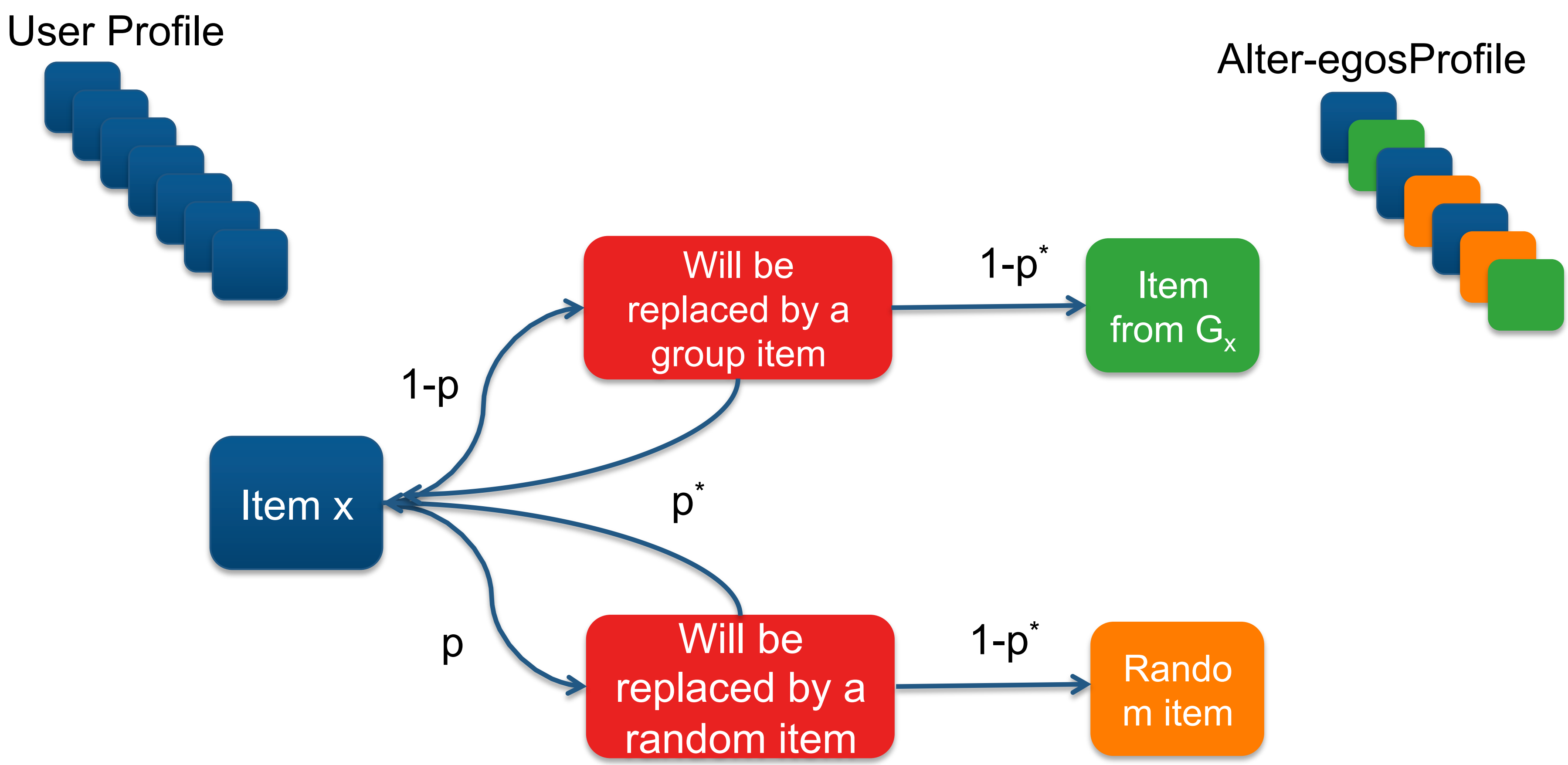
2 - Create Alter-egos profile for each user (item substitution)

D2P Components

- *Selector:* This component *decides* whether to replace an item with a *close* item or *any* item.

- *Profiler:* This component builds the *Alter-Ego* profiles by *replacing* the items based on Selector's decision.

Construction of the alter-ego profile



Distance-based Differential Privacy

For any two adjacent profile sets D_1 and D_2, where U denotes any arbitrary user, S denotes any possible subset of elements and GRP(S) denotes union of element-wise groups of items in subset S, then any mechanism R is private if the following inequality holds:

$$\frac{Pr[\mathcal{R}(\mathcal{D}_1,\mathcal{U}) \in \mathcal{GRP}_\lambda(\mathcal{S})]}{Pr[\mathcal{R}(\mathcal{D}_2,\mathcal{U}) \in \mathcal{GRP}_\lambda(\mathcal{S})]} \leq e^\epsilon$$

We show (Theorem 1) that a mechanim M relying on Alter-egos profile is an (ε, λ) mechanism

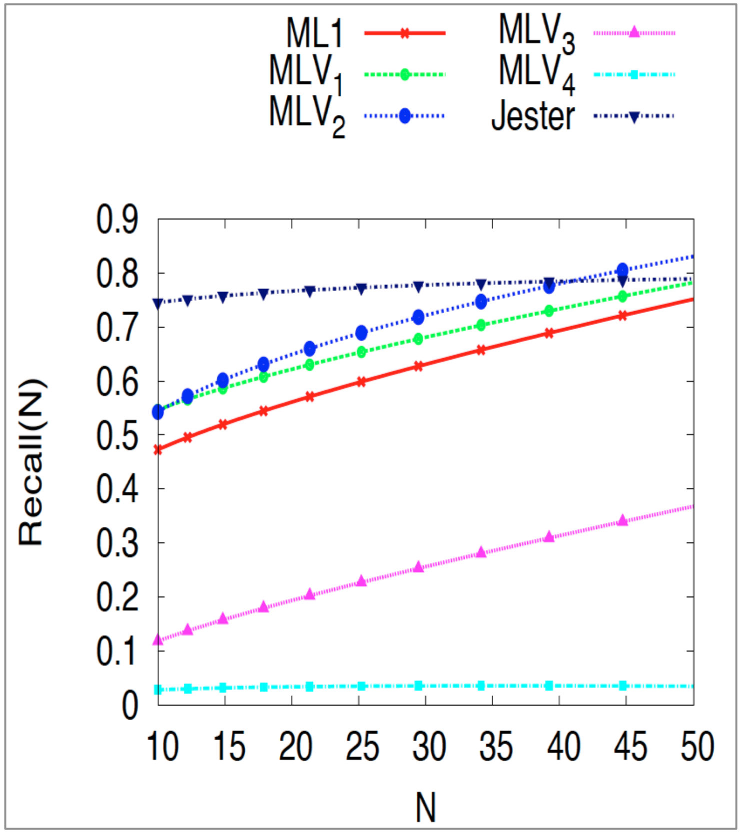
Experimental evaluation

Experimental setup

- Training set (80%) – Test set (20%)
- **Metrics**
  - Precision = $T_p/(T_p+F_p)$
  - Recall = $T_p/(T_p+F_p)$
- **Datasets**
  - MovieLens (100k ratings, 943 users, 1602 movies)
  - Jester (4.1M ratings, 73 421 users, 100 jokes) – 500 users

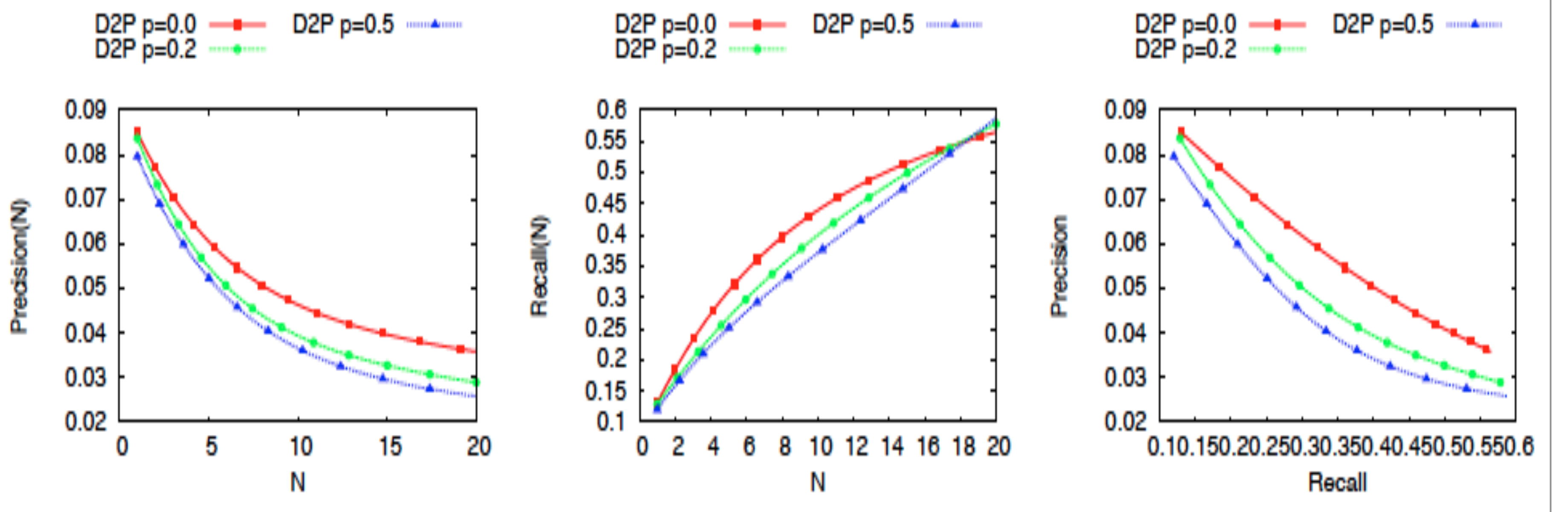
Impact of Rating Density

Dataset	#Users	#Items	Ratings	RD(%)
Jester	500	100	36000	71.01
ML1	940	1680	99647	6.31
MLV_1	470	840	76196	19.3
MLV_2	470	840	16187	4.1
MLV_3	470	840	6317	1.6
MLV_4	470	840	750	0.19



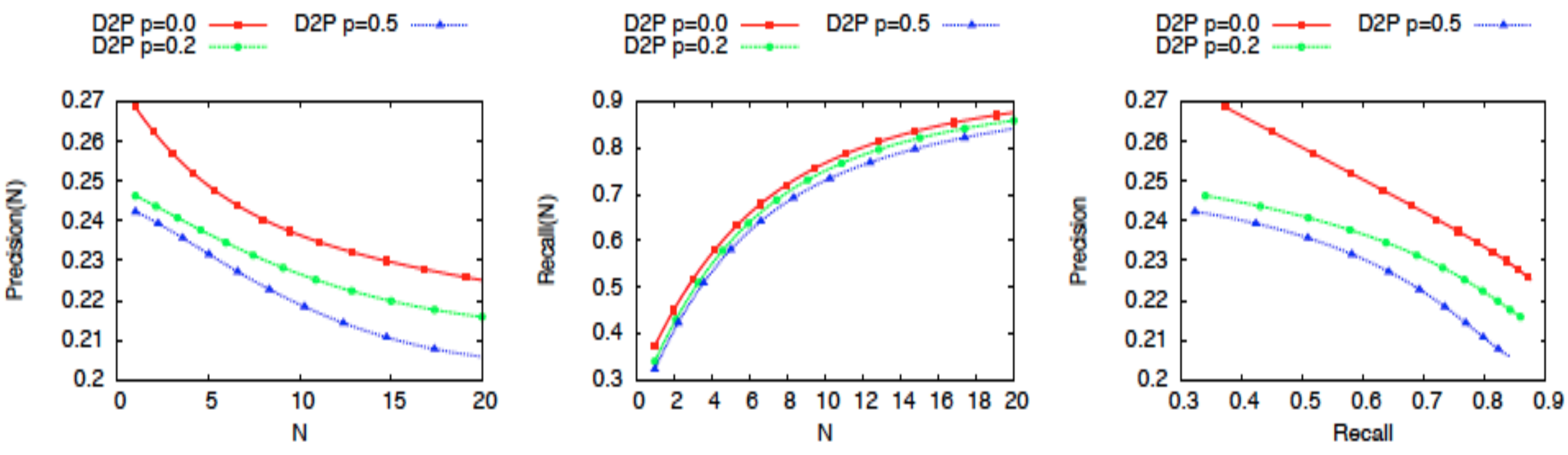
Effect of Selector probability p (MovieLens)

The lower p (fewer random substitutions) the better the recommendation quality



(a) *Precision@N Comparison.* (b) *Recall@N Comparison.* (c) *Precision-Recall Comparison.*

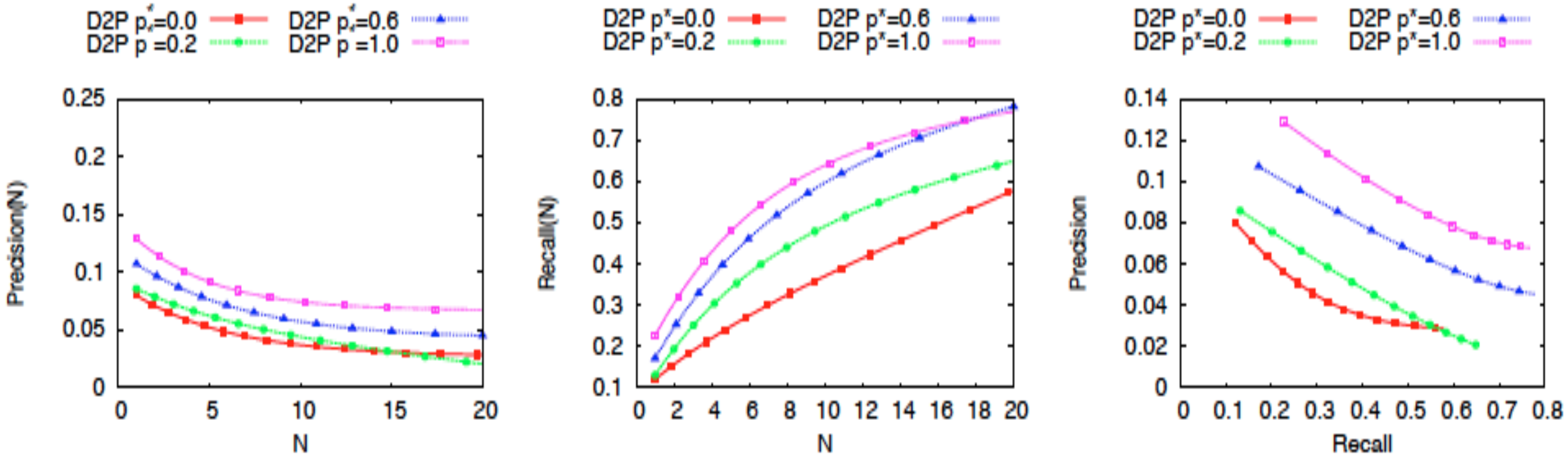
Effect of Selector Probability p (jester)



(a) *Precision@N Comparison.* (b) *Recall@N Comparison.* (c) *Precision-Recall Comparison.*

Effect of Profiler Probability (p*) (MovieLens)

The higher p* (the closer to the true profile) the better the recommendation quality

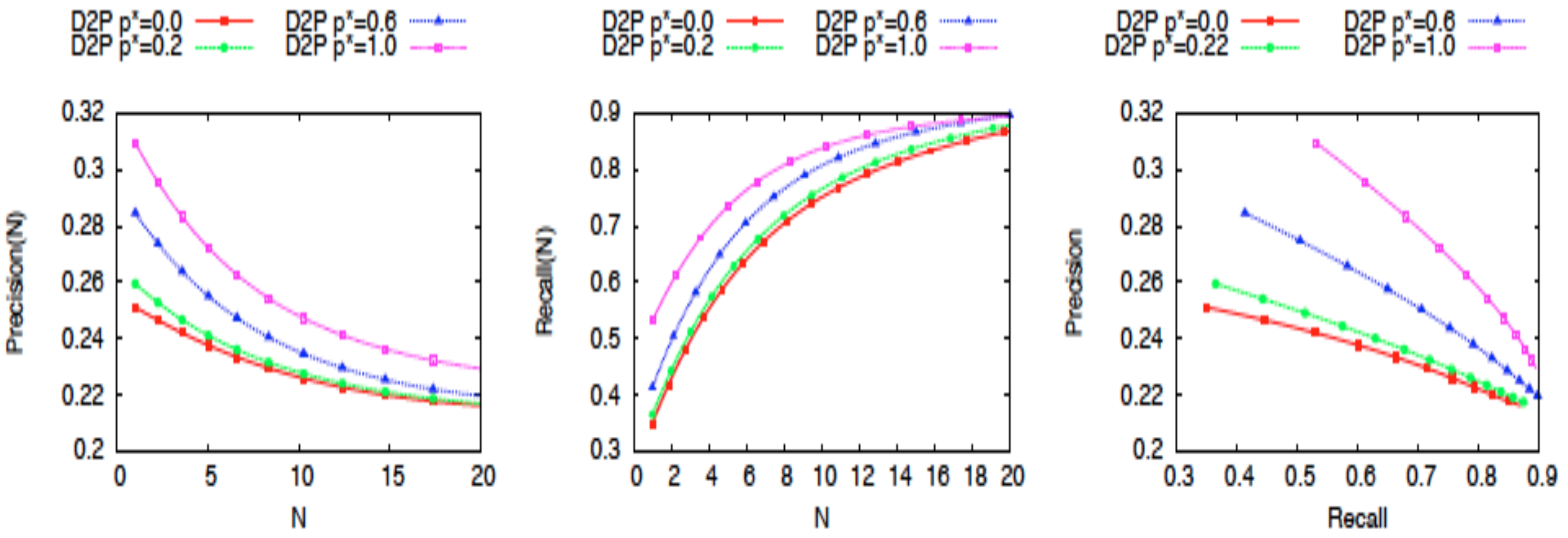


(a) *Precision@N Comparison.*

(b) *Recall@N Comparison.*

(c) *Precision-Recall Comparison.*

Effect of Profiler Probability p* (jester)



(a) *Precision@N Comparison.* (b) *Recall@N Comparison.* (c) *Precision-Recall Comparison.*

Overhead

- We compare the overhead of our system with the overhead in [1]

DP2 improves wrt efficiency

Datasets	$\mathcal{D}2\mathcal{P}$ Overhead			DP_δ Overhead
	RL	Online	Offline	Offline
$ML1$	196ms	32ms	4.54s	120s
$Jester$	24ms	12ms	162ms	740ms

[1]. McSherry, Frank, and Ilya Mironov. "Differentially private recommender systems: building privacy into the net." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

To take away

Low-overhead solution

Extension of differential privacy to recommenders

Future plans in Web Alter-Ego

- Anonymous recommenders
- Quantifying the privacy impact of a click
- Impact of cross-applications

THANK YOU