Personalization

Want to scale? Think P2P

Anne-Marie Kermarrec

LinkedIn
VK
skype
LiveJournal
foursquare
g+
vimeo
YouTube
Insta
facebook
g+
twitter
myBlogger
flickr
Bēh
instagram
Pinterest
tumblr.
yelp
Like

A cry for personalization

Why is personalization so difficult?

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

Scalable personalization infrastructures

KNN computation over large data

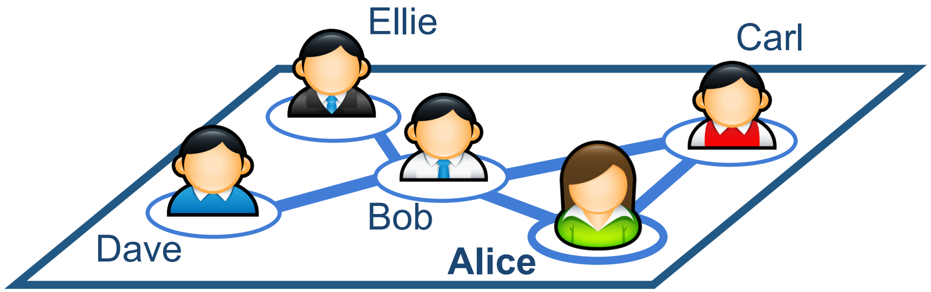
Basic building block for many applications

- Similarity search
- Machine learning
- Data mining
- Image processing
- **Collaborative filtering**

KNN-based user-centric collaborative filtering

Provide each user with her k closest neighbors

(Users owns a profile, the system has its favorite similarity metric)



Use this topology for

- personalized notifications
- recommendation

Dealing with truly big data

Want to scale? Think P2P

Do not look exhaustively

The key to scalability in KNN graph construction

Consider a partial set of candidates

Sampling-based approach

P2P KNN graph construction

Which nodes are close? → Similarity metric

How to discover them? → Sampling

Which nodes are close?

Model

U(sers) × *I*(tems) (items)

Profile(u) = vector of liked/shared/viewed items

Cosine similarity metric

$$Similarity\,(n,p) = \frac{n.p}{\|n\|\,\|p\|}$$

Jaccard metric

$$Jaccard(n,p) \;= \left|\frac{n \cap p}{n \cup p}\right|$$

Minimal information: **no tag, no user's input, generic**

How to discover them: Gossip-based computing

Each node maintains a set of neighbors (c entries)

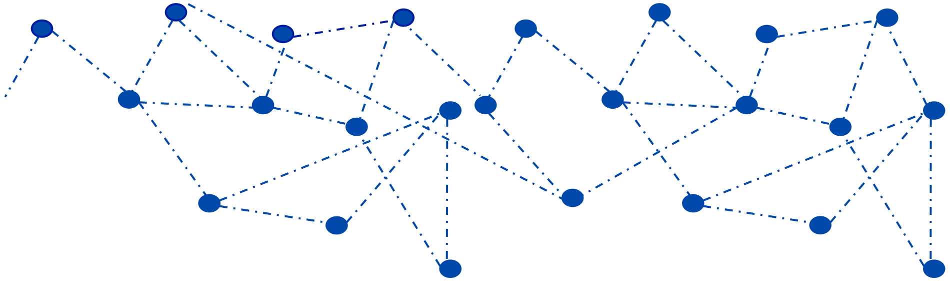
Peer exchange

Shuffle

Result → random graph

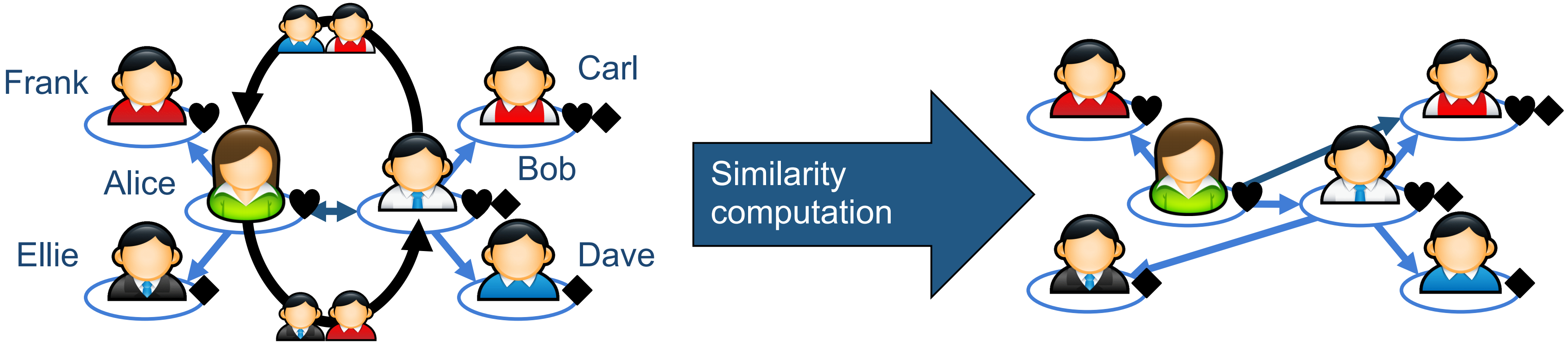
Highly resilient against churn, partition

Small diameter



[JGKVV, ACM TOCS 2007]

KNN construction

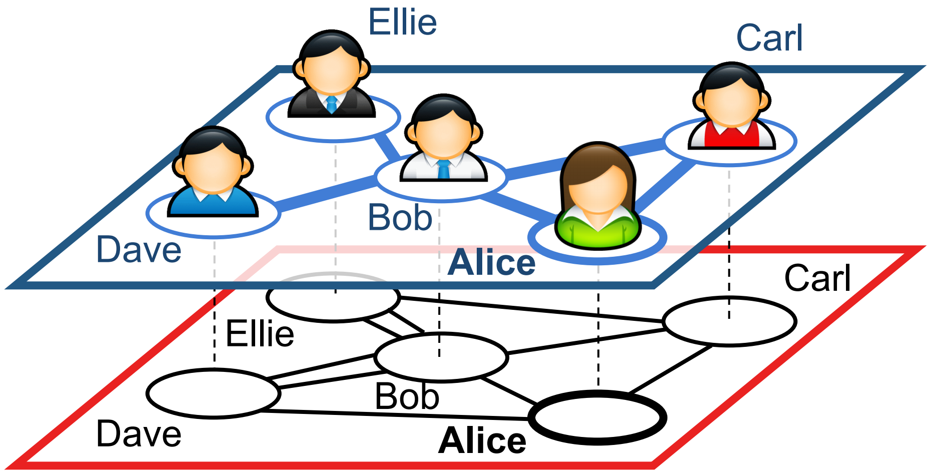


Similarity computation

❶ exchange of neighbors lists

❷ neighborhood optimization

Decentralized KNN selection



clustering layer
gossip-based
topology clustering

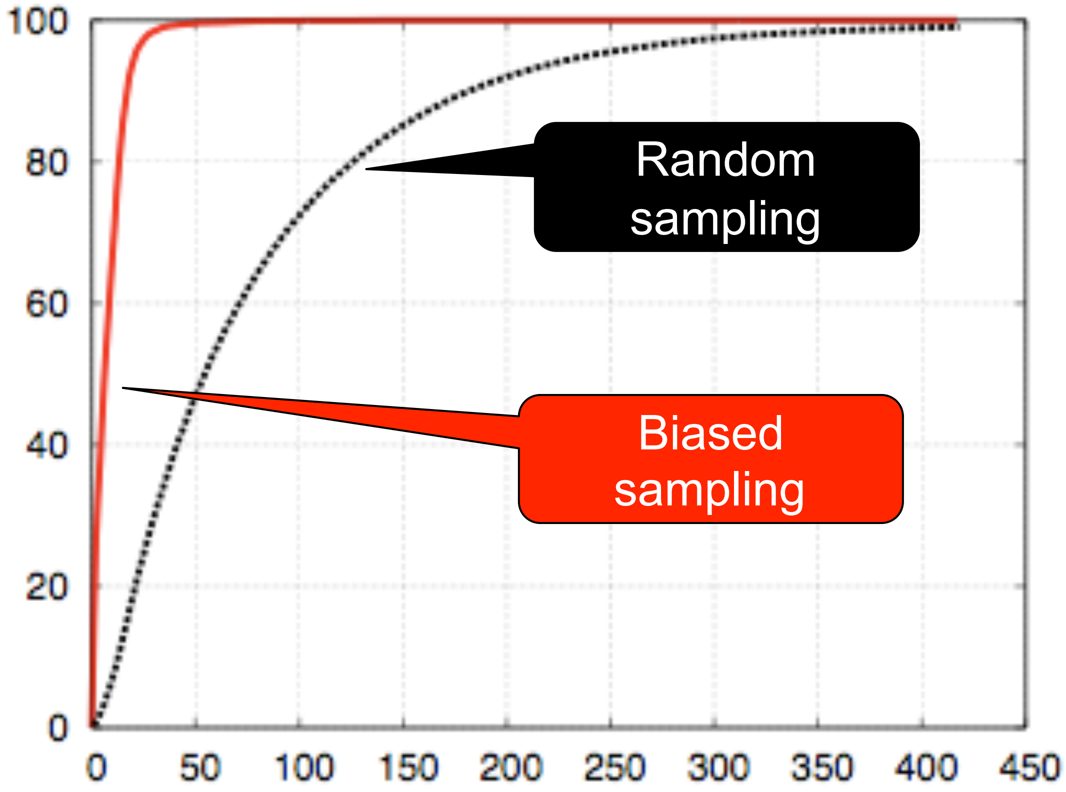
RPS layer providing
random sampling

— *Random link* ▬ *Interest-based link*

[FGKL Middleware 2010]

Convergence

c current neighbors versus the c closest



Random sampling

Biased sampling

Cycles

Applications

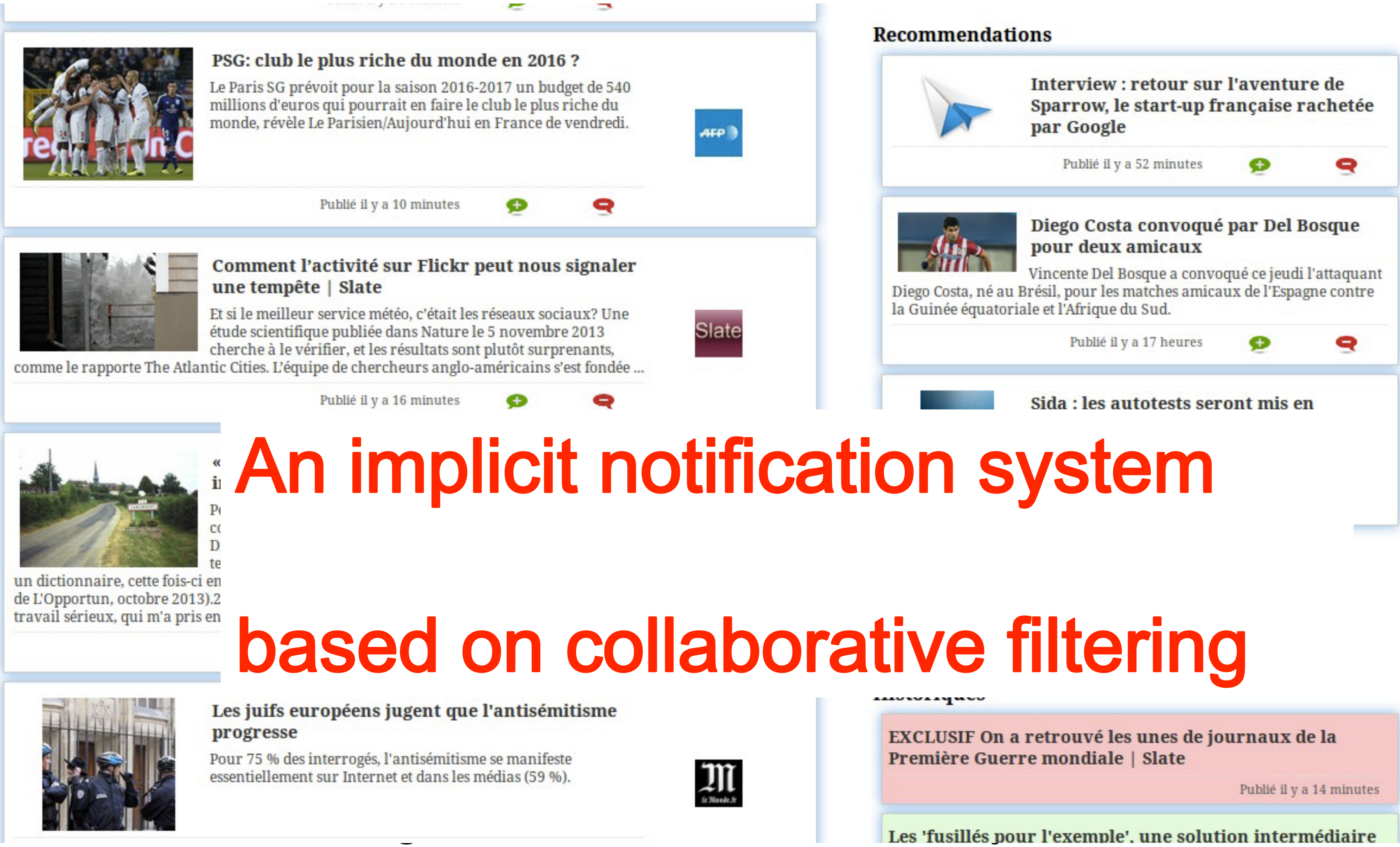
- **Decentralized news recommendation** [BFGJK, IPDPS 2013]
- Top-K [BGKL, ACM TODS 2011] [BGK, ACM TOIT 2014]
- Geo recommendation [BKKT, ICDCS 2012]

DECENTRALIZED NEWS RECOMMENDER

Notification is taking over

An implicit notification system

based on collaborative filtering



PSG: club le plus riche du monde en 2016 ?

Le Paris SG prévoit pour la saison 2016-2017 un budget de 540 millions d'euros qui pourrait en faire le club le plus riche du monde, révèle Le Parisien/Aujourd'hui en France de vendredi.

Publié il y a 10 minutes

Comment l'activité sur Flickr peut nous signaler une tempête | Slate

Et si le meilleur service météo, c'était les réseaux sociaux? Une étude scientifique publiée dans Nature le 5 novembre 2013 cherche à le vérifier, et les résultats sont plutôt surprenants, comme le rapporte The Atlantic Cities. L'équipe de chercheurs anglo-américains s'est fondée ...

Publié il y a 16 minutes

« i

Pe co D te un dictionnaire, cette fois-ci en de L'Opportun, octobre 2013).2 travail sérieux, qui m'a pris en

Les juifs européens jugent que l'antisémitisme progresse

Pour 75 % des interrogés, l'antisémitisme se manifeste essentiellement sur Internet et dans les médias (59 %).

Recommendations

Interview : retour sur l'aventure de Sparrow, le start-up française rachetée par Google

Publié il y a 52 minutes

Diego Costa convoqué par Del Bosque pour deux amicaux

Vincente Del Bosque a convoqué ce jeudi l'attaquant Diego Costa, né au Brésil, pour les matches amicaux de l'Espagne contre la Guinée équatoriale et l'Afrique du Sud.

Publié il y a 17 heures

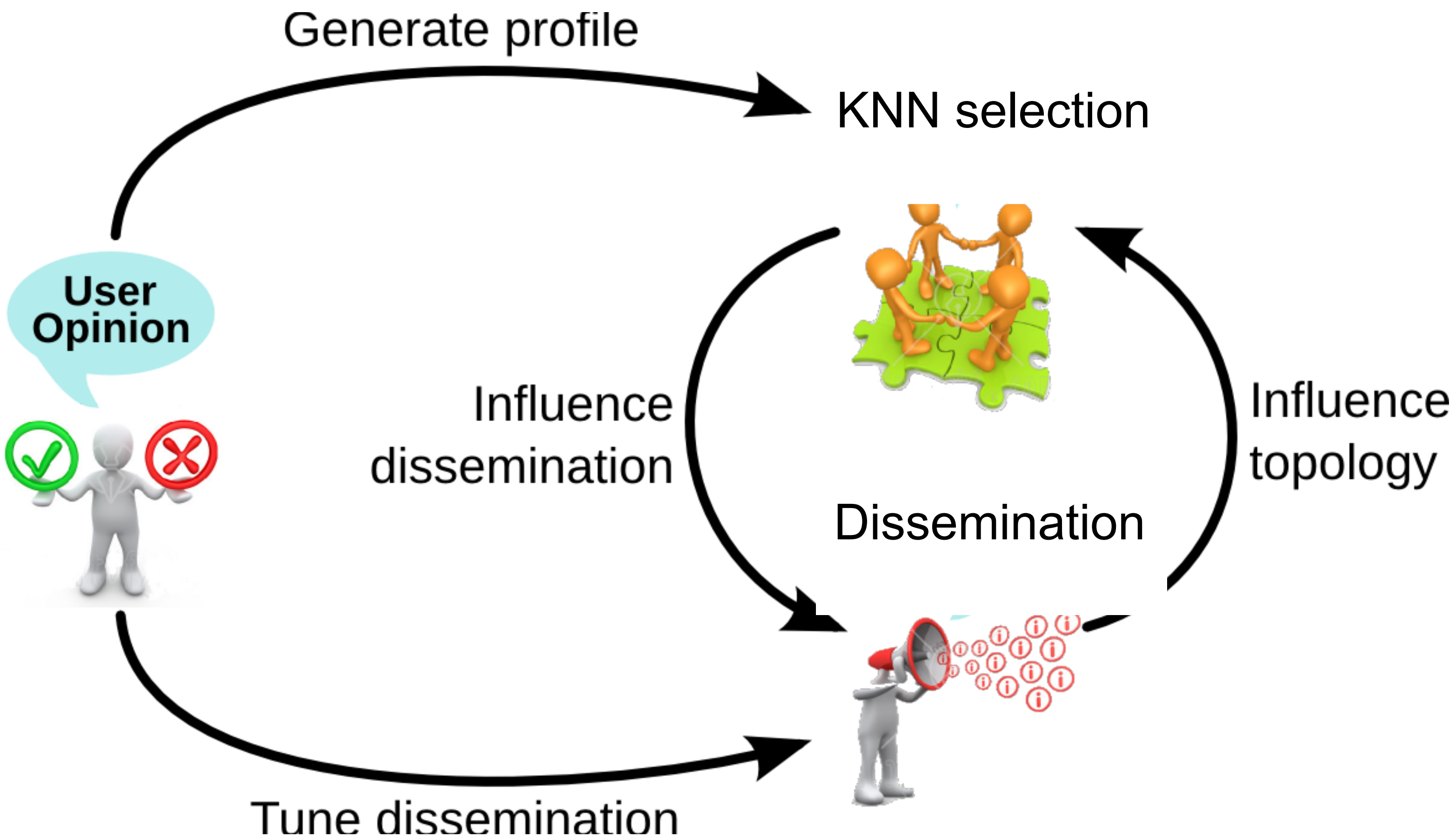
Sida : les autotests seront mis en

EXCLUSIF On a retrouvé les unes de journaux de la Première Guerre mondiale | Slate

Publié il y a 14 minutes

Les 'fusillés pour l'exemple', une solution intermédiaire

WhatsUp in a nutshell



Generate profile

KNN selection

User Opinion

Influence dissemination

Influence topology

Dissemination

Tune dissemination

Dissemination: orientation and amplification

Orientation: **to whom**?

Exploit:
Forward
To friends

Explore:
Forward to
random
users

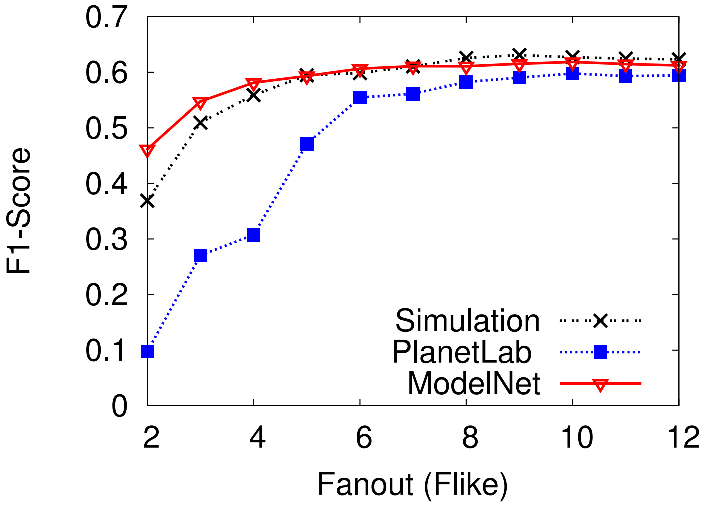
Amplification: **to how many**?

Increase
Fanout
(Log(n))

Decrease
Fanout
(1)

WhatsUp in action on the survey (480 users)

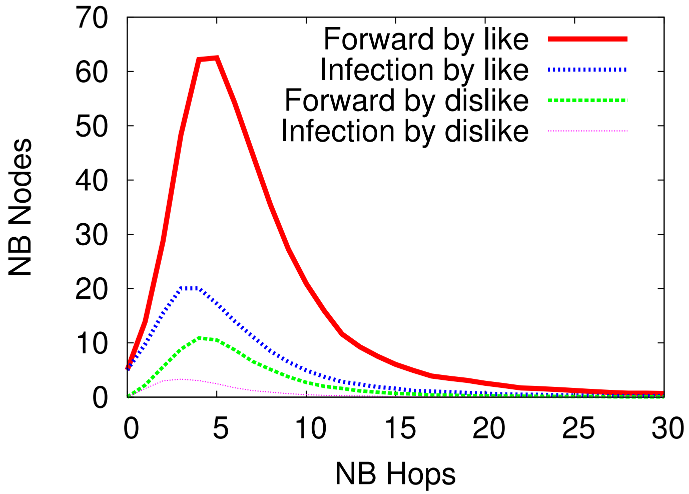
	Precision	Recall	F1-Score	Messages
Gossip (f=4)	0.34	0.99	0.51	2.3 M
Cosine-CF	0.50	0.65	0.57	5,9k
Whatsup (f=10)	**0.471**	**0.83**	**0.60**	**2,4k**



Orientation (survey)

News items received through a dislike forward

Number of dislikes	0	1	2	3	4
Fraction of liked news	54%	31%	10%	3%	2%



WhatsUp versus Pub/Sub

Approach	Precision	Recall	F1-Score
Pub/Sub	0.40	1.0	0.58
WhatsUp	0.47	0.83	0.60

WhatsUp versus cascading

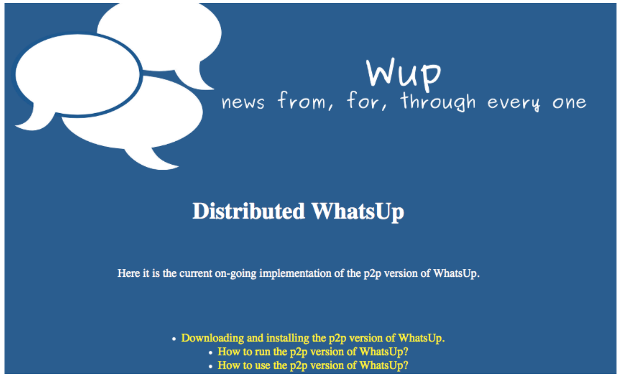
Approach	Precision	Recall	F1-Score
Cascading	0.57	0.09	0.16
WhatsUp	0.56	0.57	0.57

Take away message

Personalization is needed

Decentralization is healthy

Gossip-based computing is one (the) way to go



Wup

news from, for, through every one

Distributed WhatsUp

Here it is the current on-going implementation of the p2p version of WhatsUp.

- Downloading and installing the p2p version of WhatsUp.
- How to run the p2p version of WhatsUp?
- How to use the p2p version of WhatsUp?

STRONGER GARANTEES

Privacy matters

- Obfuscation
- Anonymous routing
- Threshold protocol
- Differentially private protocol
- Landmark-based protocol

http://131.254.213.98:8080/wup/

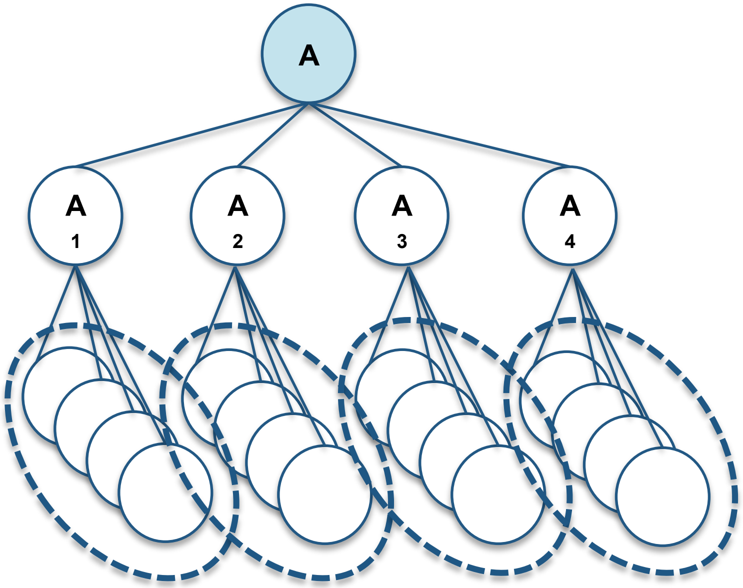
Operational prototype

Tested on 500 users @ TrentoRise last year

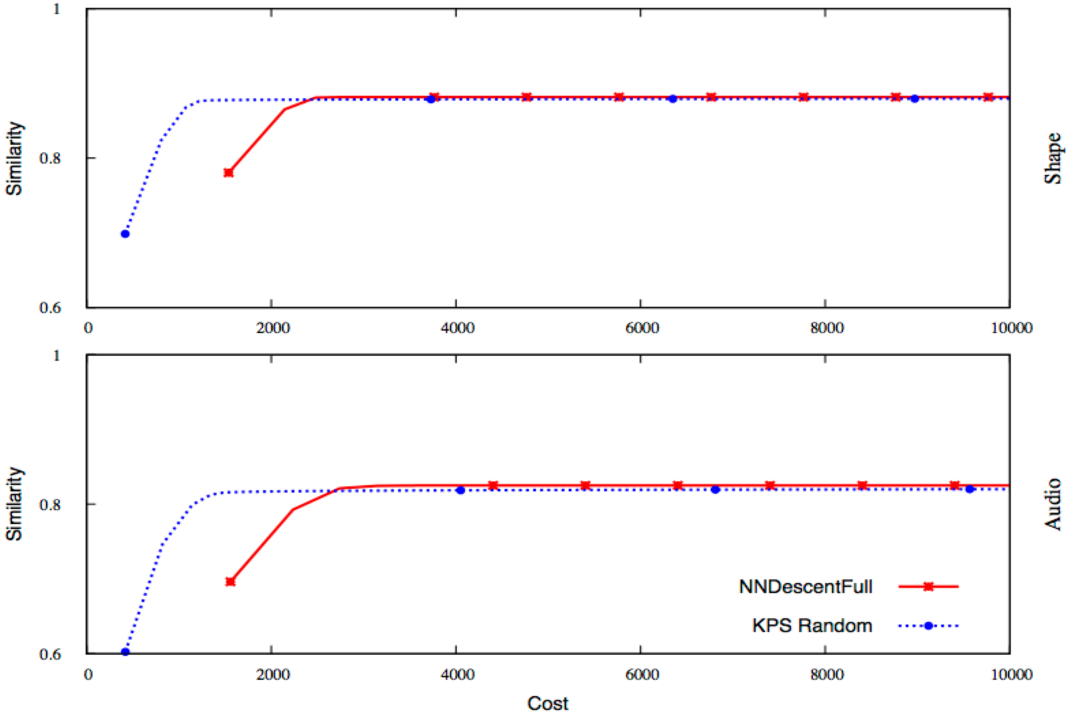
TRY IT ☺

For those who are afraid of P2P

Turns out to be an effective centralized algorithm too.



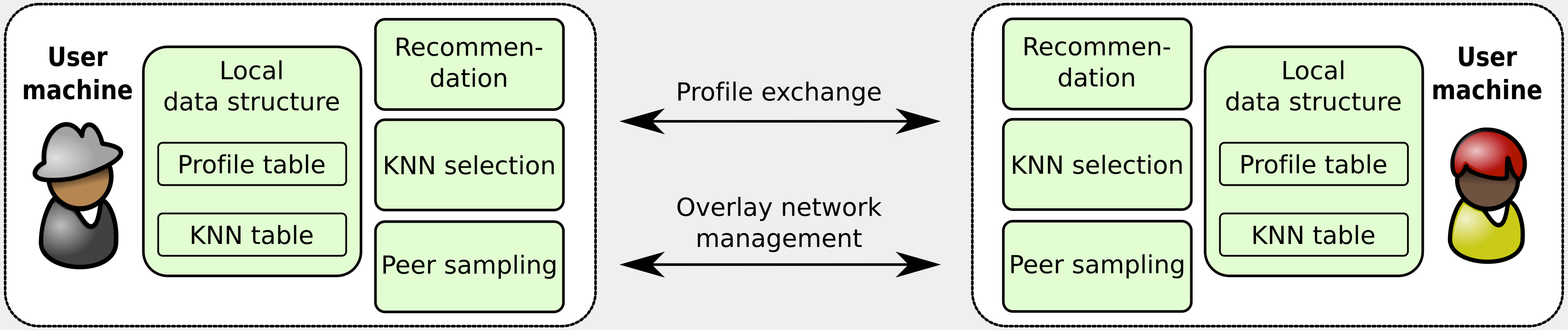
Candidate set: neighbors of neighbors
+ Random candidates for dynamics



Comparison with [Dong&al, 2012]

Hybrid recommendation engine

Decentralized approach



User machine

Local data structure

Profile table

KNN table

Recommendation

KNN selection

Peer sampling

Profile exchange

Overlay network management

Recommendation

KNN selection

Peer sampling

Local data structure

Profile table

KNN table

User machine

Data structures

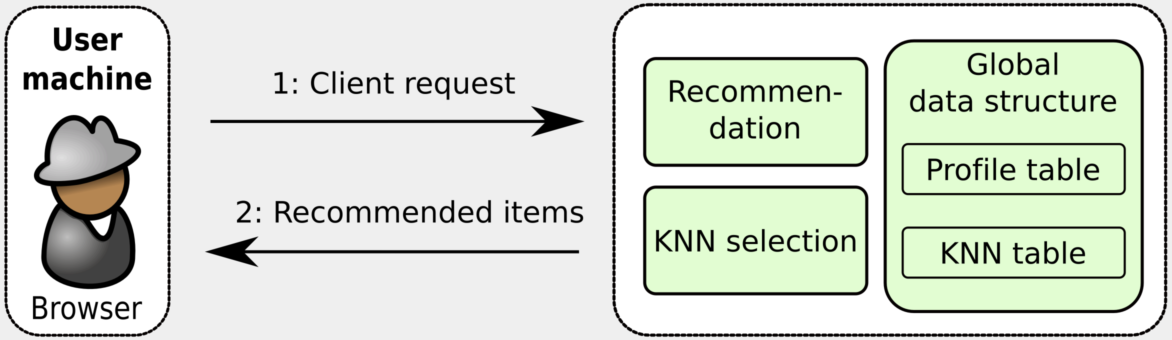
Profile table

uid	P(uid) = {list of iid}

KNN table

uid	Knn(uid) ={list of uid}

Centralized approach



User machine

Browser

1: Client request

2: Recommended items

Recommendation

KNN selection

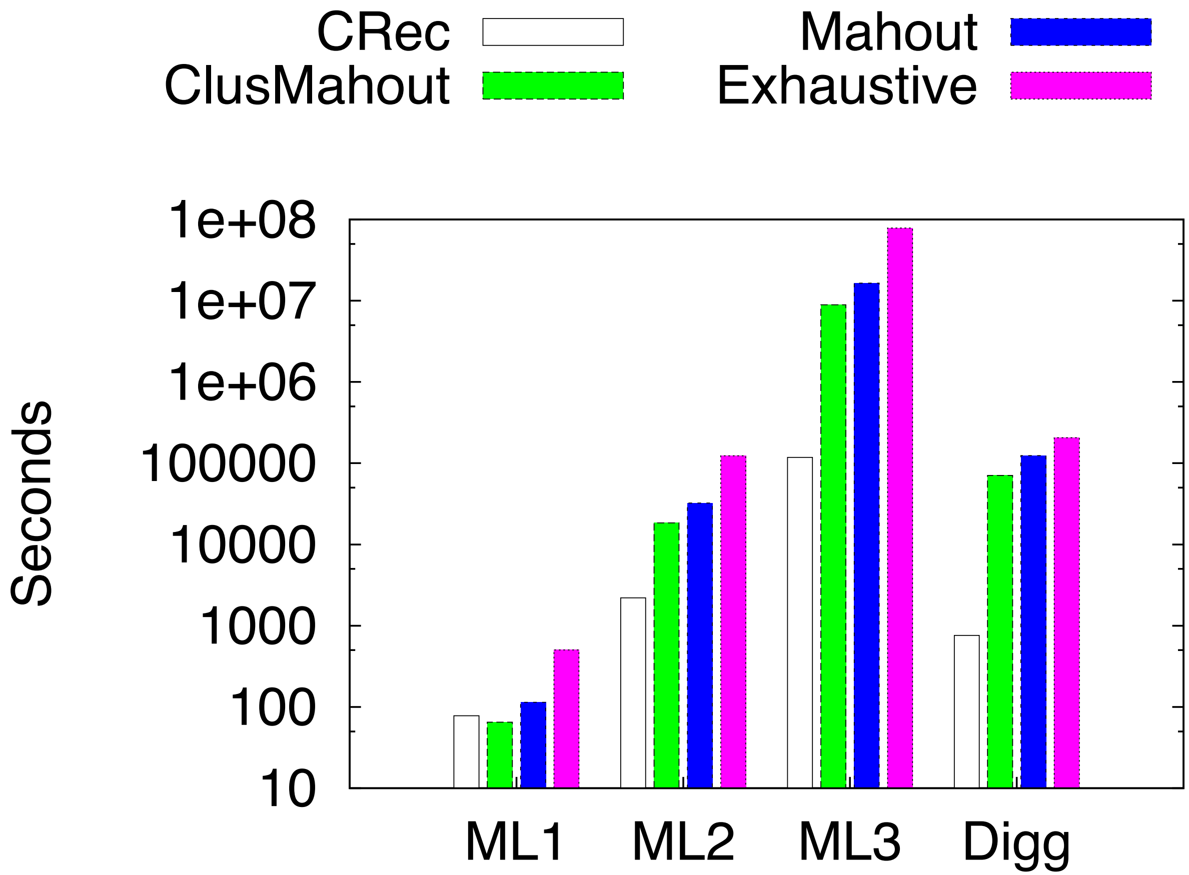
Global data structure

Profile table

KNN table

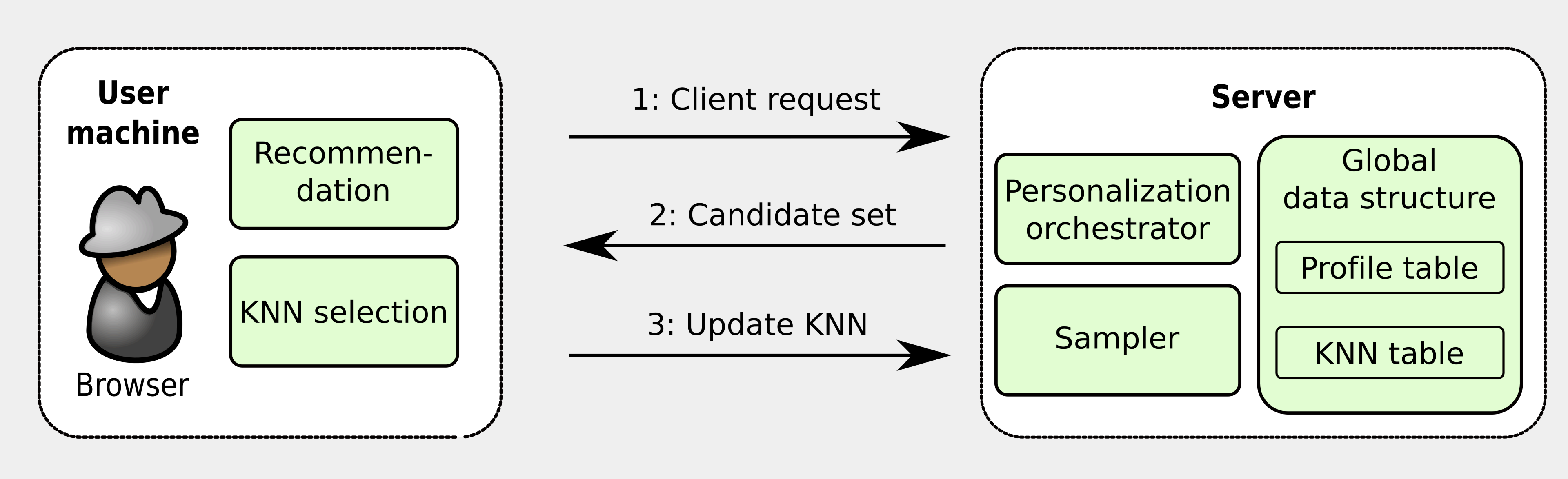
Cost

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



Grid 5000 implementation

HyRec: Taking the best of both worlds



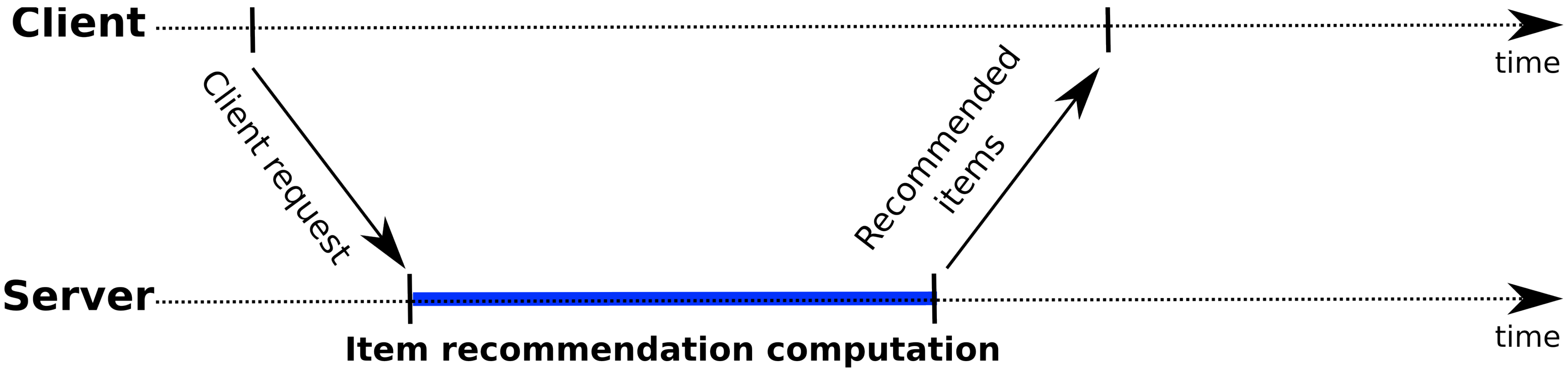
Online KNN selection

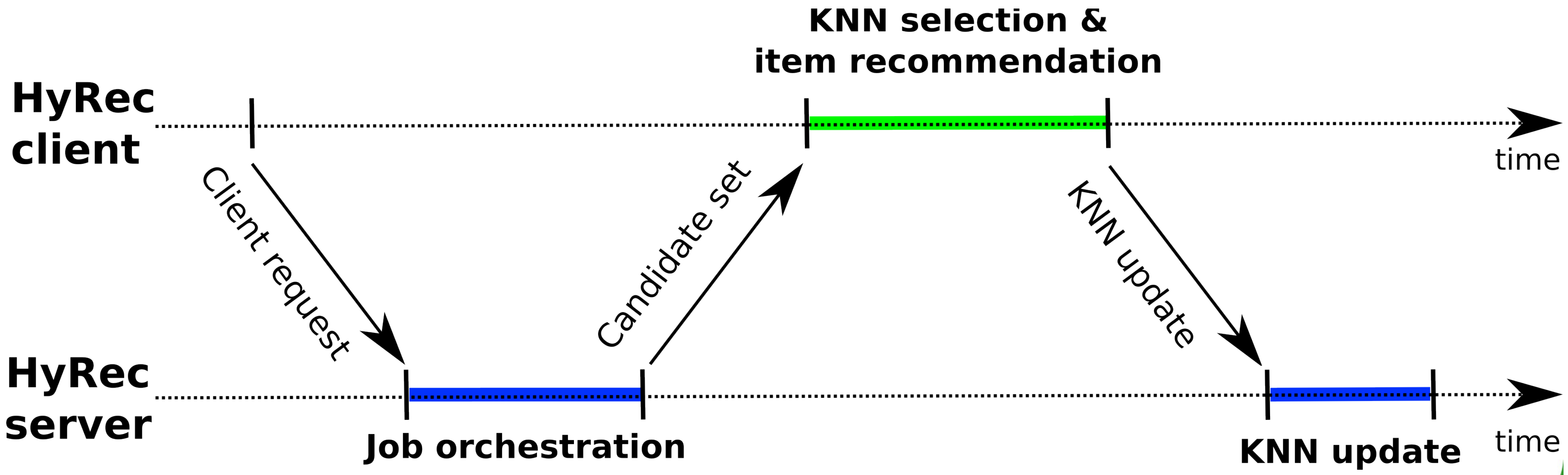
Candidate set (k): k^2 users for quick convergence, k random (biased) for dynamics

No data stored at the client

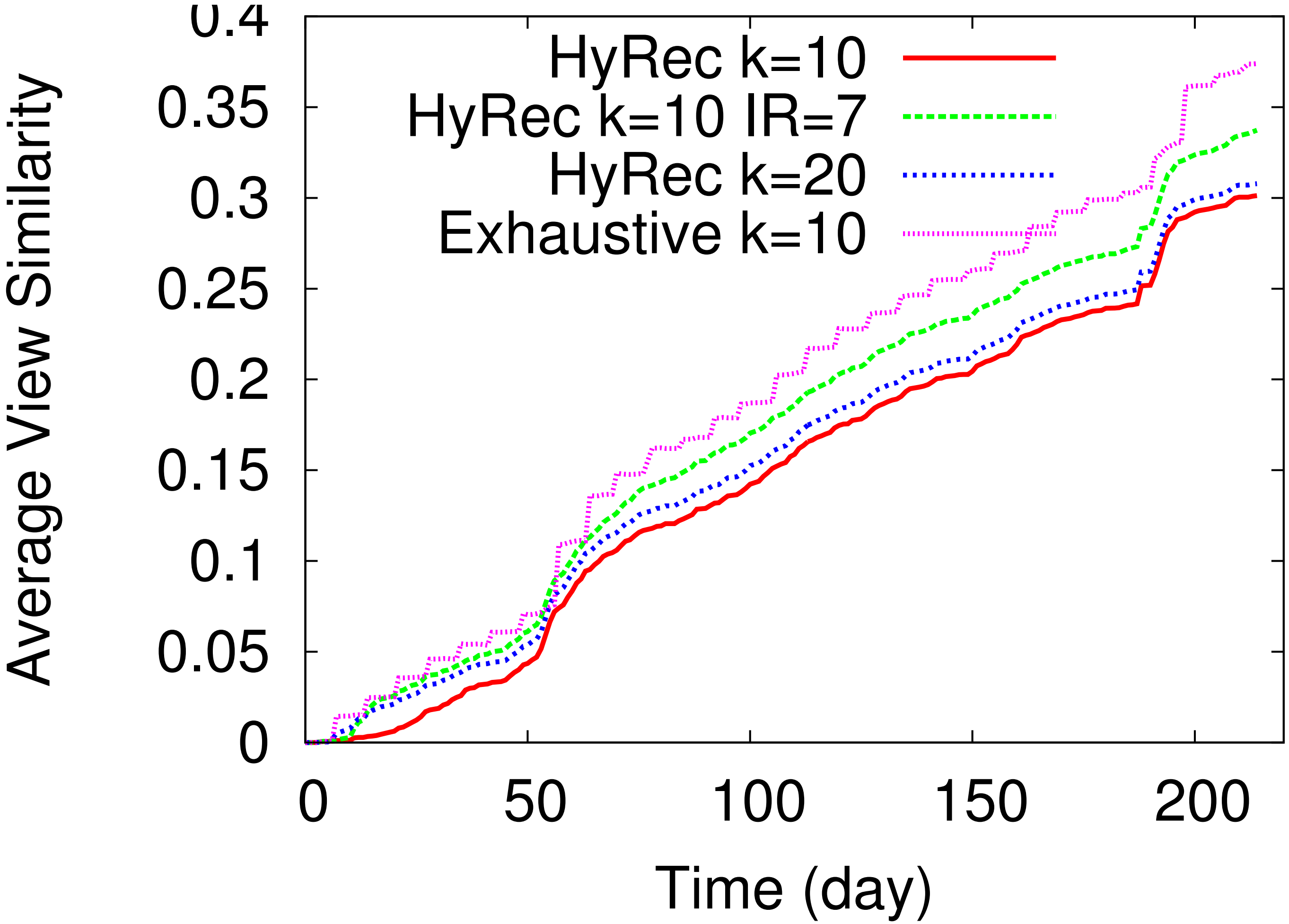
Recommendation: R most popular items

HyRec client: Javascript (widget) running in the browser

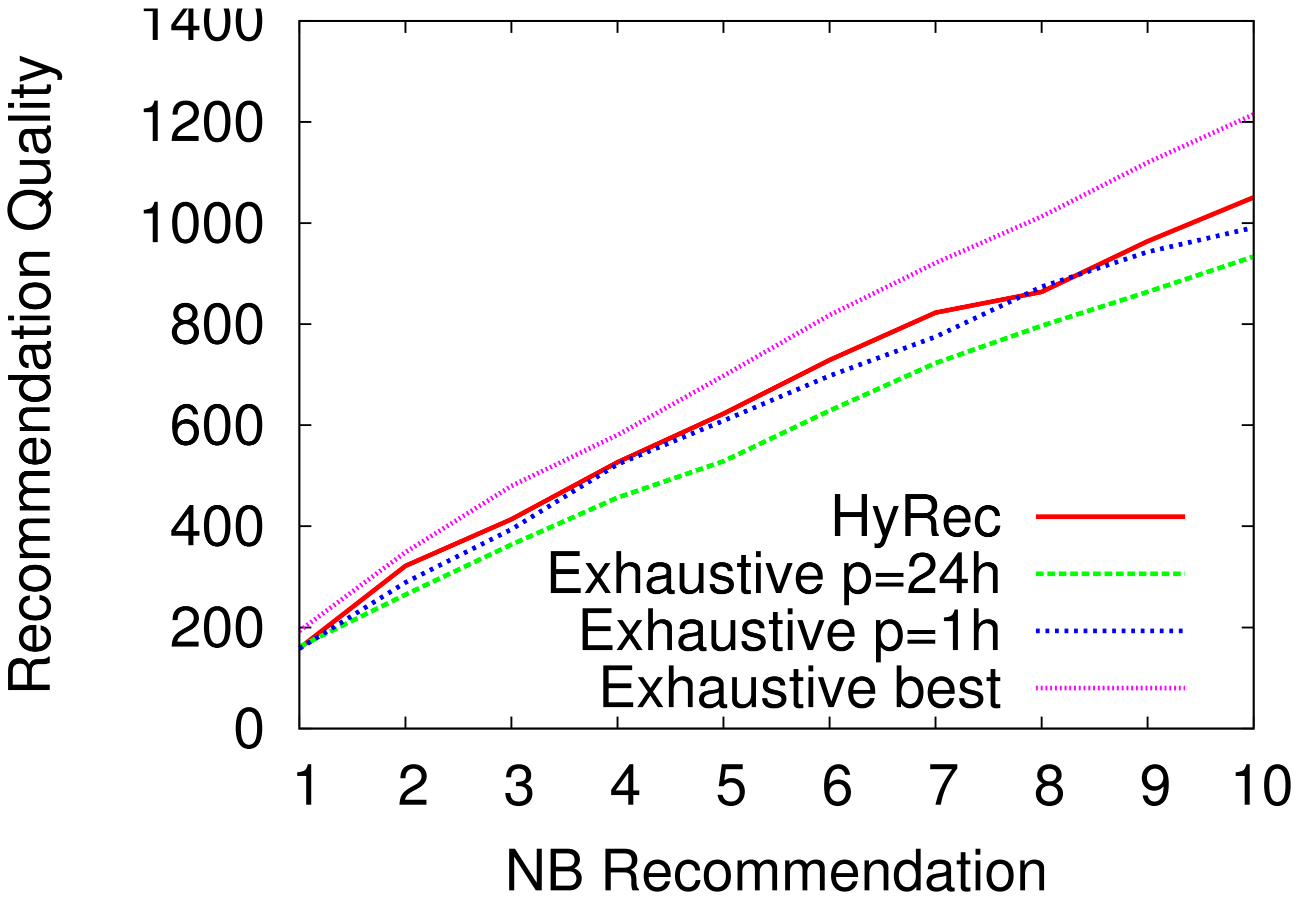

Client
time
Client request
Recommended items
Server
time
Item recommendation computation


KNN selection & item recommendation
HyRec client
time
Client request
Candidate set
KNN update
HyRec server
time
Job orchestration
KNN update

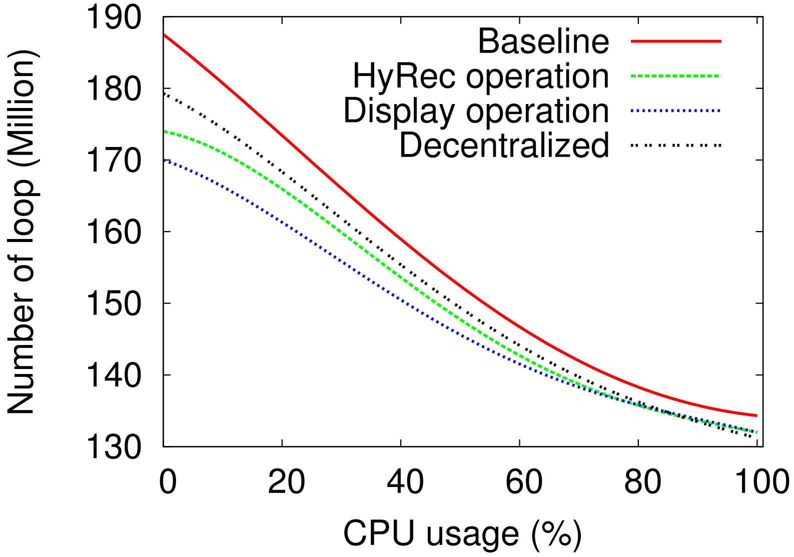
View similarity (MovieLens)



Recommendation quality

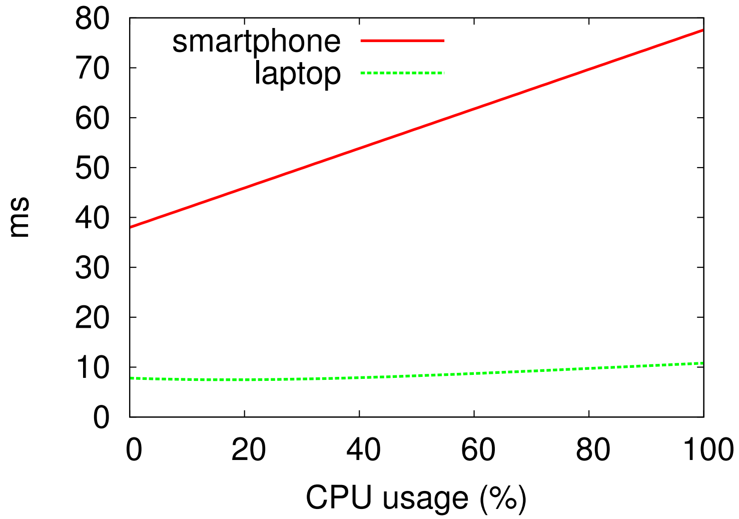


HyRec versus the client load



Impact of HyRec

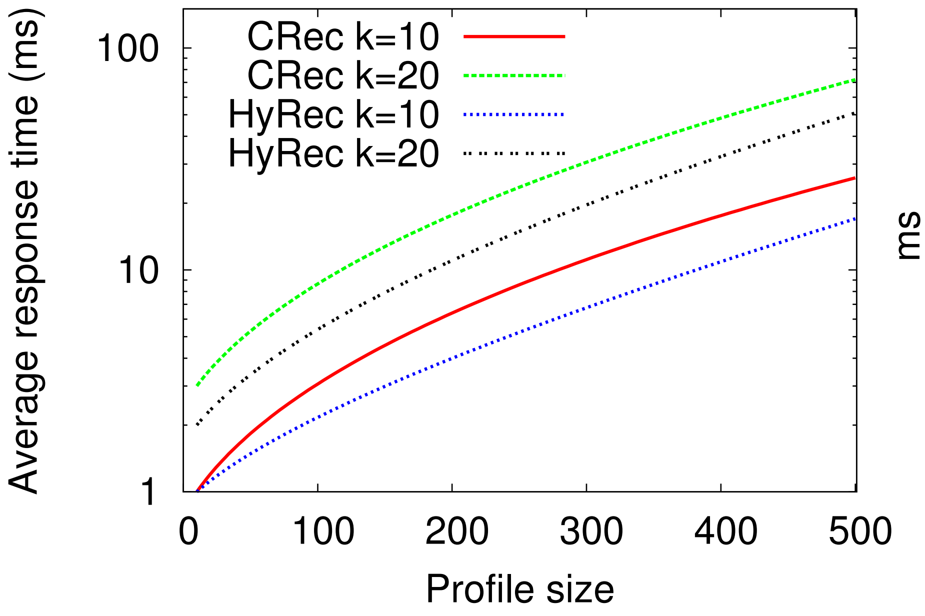
Negligible disruption of HyRec



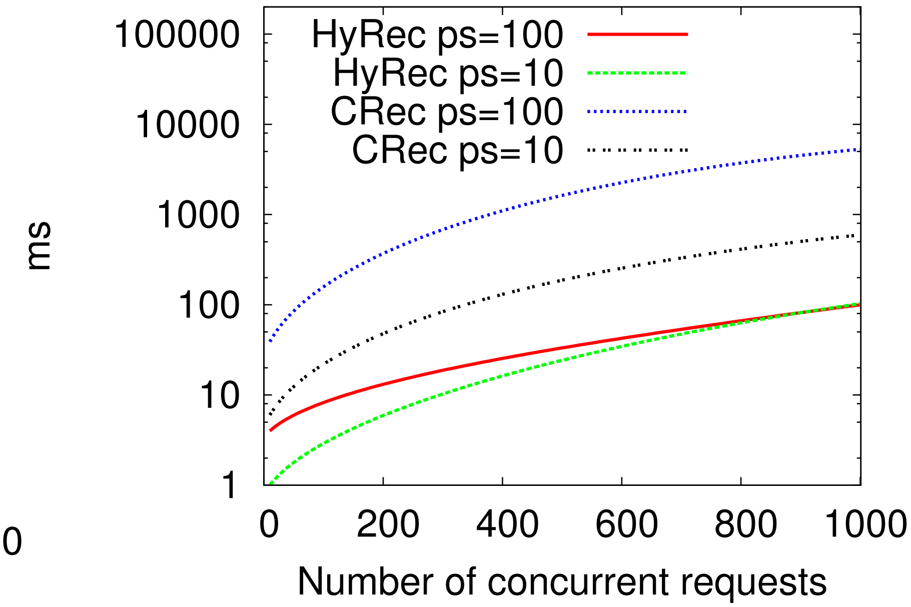
Impact of the client load

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

HyRec versus a centralized recommender



Impact of the profile size



Impact of the request stress

Take away message

P2P design is crucial

Leveraging clients machine has a significant impact on scalability

Enable any content provider to implement personalization

To take away

Personalization is crucial

P2P in a design mindset

Thank you