Toward Personalized Query Expansion

Marin Bertier Rachid Guerraoui Anne-Marie Kermarrec
Vincent Leroy

March 27, 2009

Outline

Motivation

Gossple Overview

Personalized Network

TagMap & TagRank

Evaluation

Conclusion

Motivation

- "web 2.0" websites, users generate content
- social websites (facebook . . .)
- collaborative filtering (last.fm, amazon . . .)
- P2P networks

$\Rightarrow$ users want to **share** . . .

Motivation

. . . but they want **anonymity**

- Facebook policy
- concerns about Google storing personal data
- Big Brother syndrome

Gossple

- decentralized system
  - no one owns all the information
  - scales (computation, storage, reactivity)
- users share information
  - personalized network
  - better than explicit social network
- preserves anonymity
  - no association between data and user

System model

Folksonomy:

- user annotated content
- $IS = \{(u, i, t)\}, u \in U(users), i \in I(items), t \in T(tags)$
- Delicious, Youtube, Flickr, CiteULike . . . are folksonomies

User Vincent:

Radiohead Rock, Music

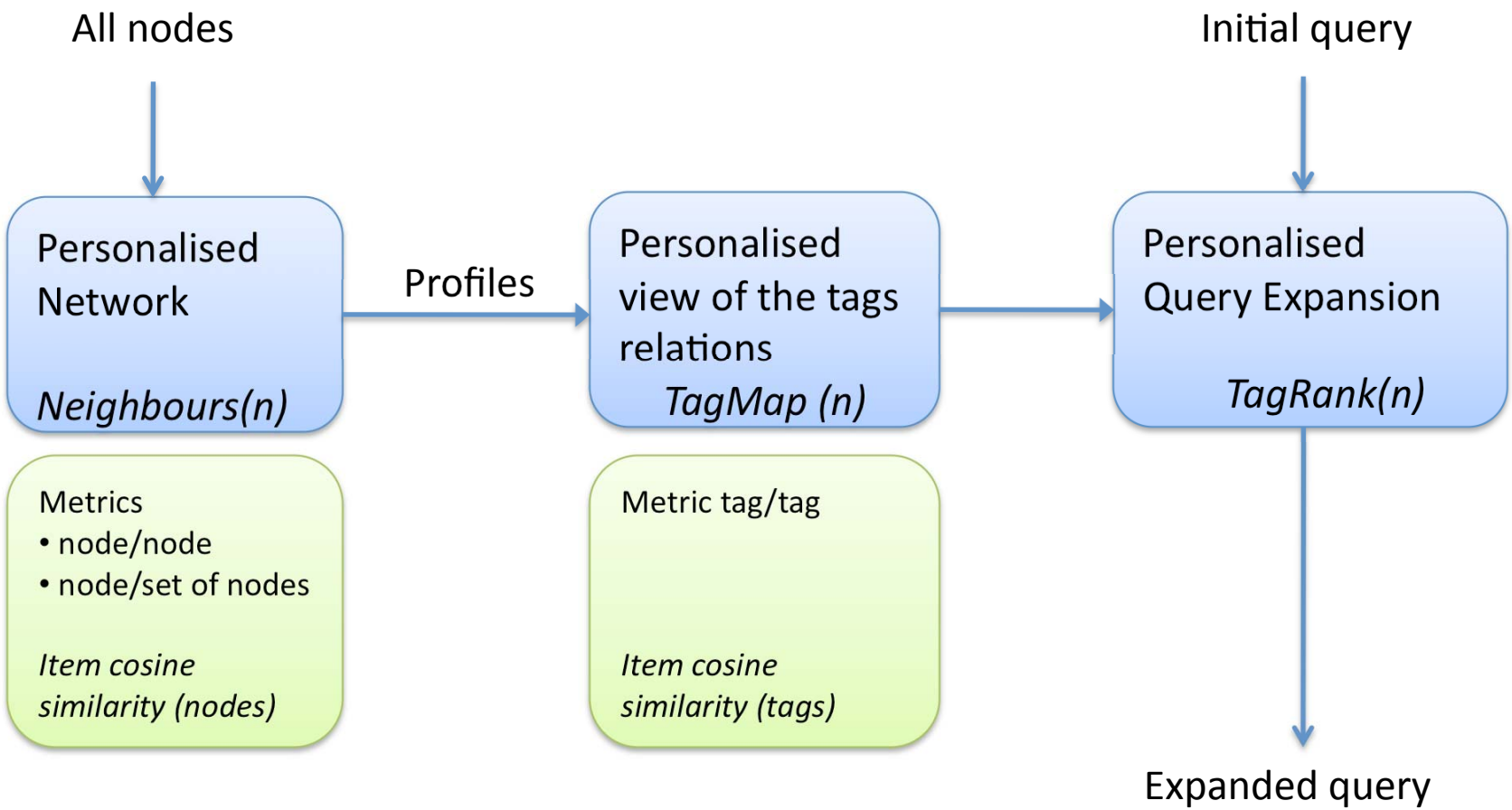
Gentoo Linux, Informatics

Dell Informatics, Computer

Query expansion

- transform user query
  - add new terms
  - weight terms
- personalized system (apple, jaguar, latex. . .)
- find new relevant items (recall)
- remain precise (precision)

Gossple Overview



Personalized Network

- view of the users that are the “closest”
- taken into account for query expansion
- preserve anonymity

Cosine Similarity

- Score between 2 users based on tagging behavior
- item cosine similarity:
  $ItemCos(u_1, u_2) = \frac{|Item(\{u_1\}) \bigcap Item(\{u_2\})|}{\sqrt{|Item(\{u_1\})| \times |Item(\{u_2\})|}}$
- normalized overlap
- shared items increase score
- non shared items decrease score

$\Rightarrow$ favor precise interests

Multi-Interest Clustering

- neighbors only reflect dominant interests
- need to rate a full set of users
- trade off between amount of shared items and distribution

$$SetItemVect(set) = \sum_{p \in set} \frac{(ItemVect(p) \otimes ItemVect(n))}{\|ItemVect(p)\|}$$

$$SetScore(n, set) = SetItemVector(set) \cdot ItemVect(n) \times \cos(SetItemVector(set), ItemVect(n))^b$$

Gossip Protocols

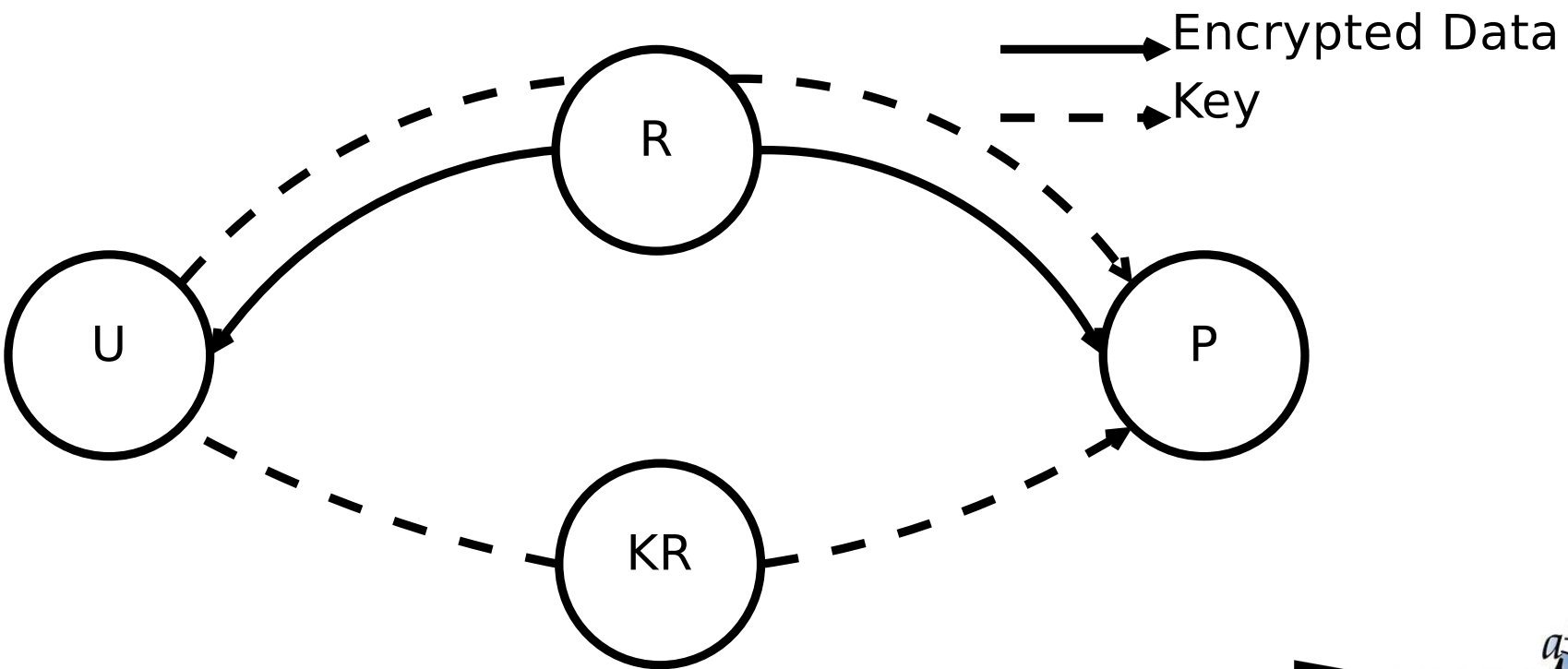
- RPS provides dynamic random view
- exchange information with peers in view
- fast clustering convergence
- churn resilience

Anonymity

Gossip on behalf

- user u represented by proxy p
- u and p communicate through r with encryption
- p builds u's personalized network

$\Rightarrow$ no association between u and his profile



TagMap

- represent tags proximity
- personalized for each user
- $IS_{u_k} = \{(u, i, t)\}$ with $u \in PersNet(u_k) \bigcup \{u_k\}$
- from IS_{u_k}, item cosine similarity between tags

TagMap

	Music	BritPop	Classic	Bach	Oasis
Music	1	0.9	0.3	0.3	0
BritPop		1	0	0	0.41
Classic			1	1	0
Bach				1	0
Oasis					1

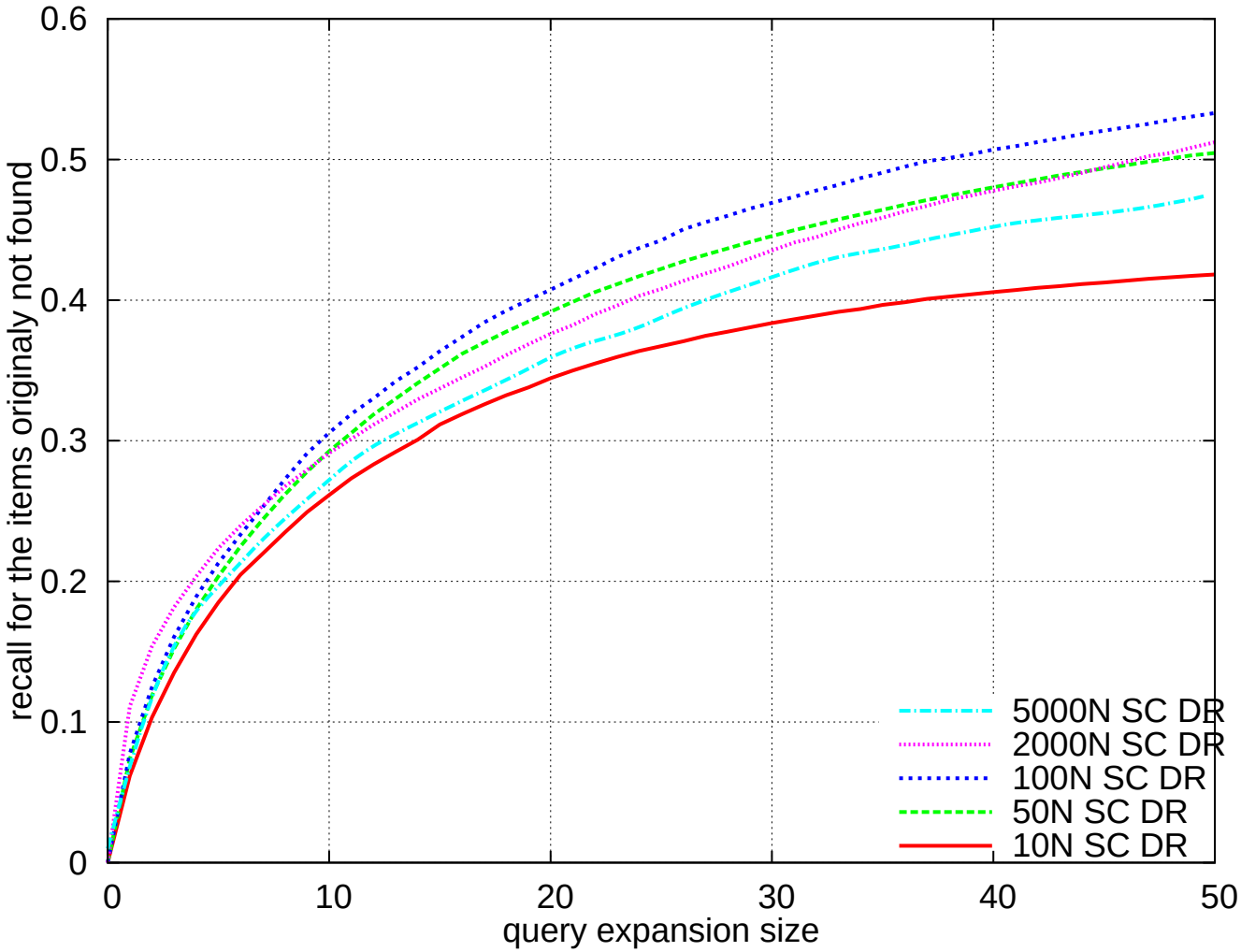
TagRank

- folksonomies suffer from sparsity
- TagMap only provides "one-hop" distance between tags
- query expansion:
  - represent TagMap as a graph
  - do random walks from query terms (PageRank random surfer model)
  - $\Rightarrow$ tag centrality w.r. to a query

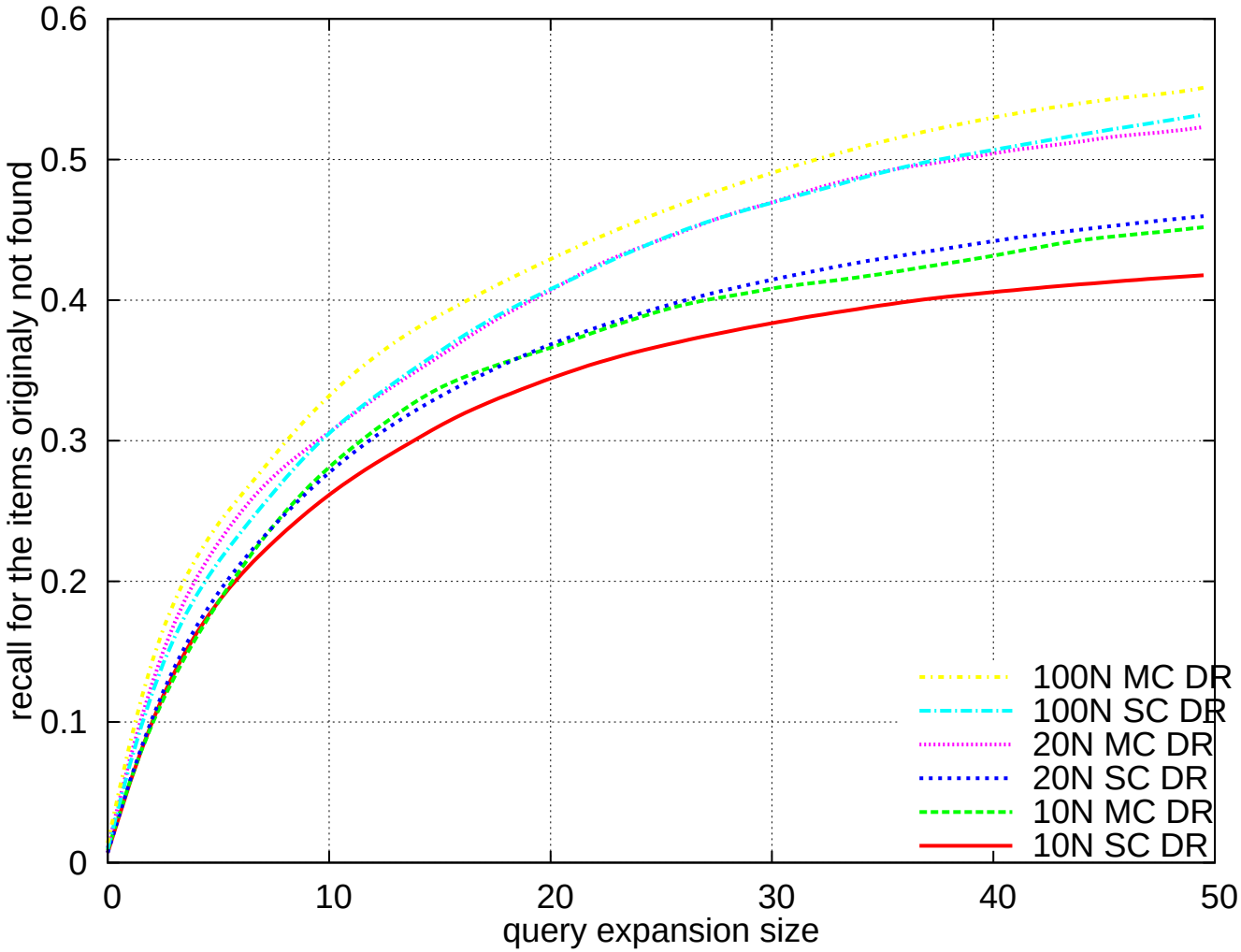
Evaluation protocol

- traces from real websites
  - Delicious (20000 users)
  - CiteULike (30000 users)
- remove 1 item from profile
- generate personalized network & TagMap
- query=tags that were associated with the item
- measure recall (item$\in$ resultset) & precision (relative to item rank)

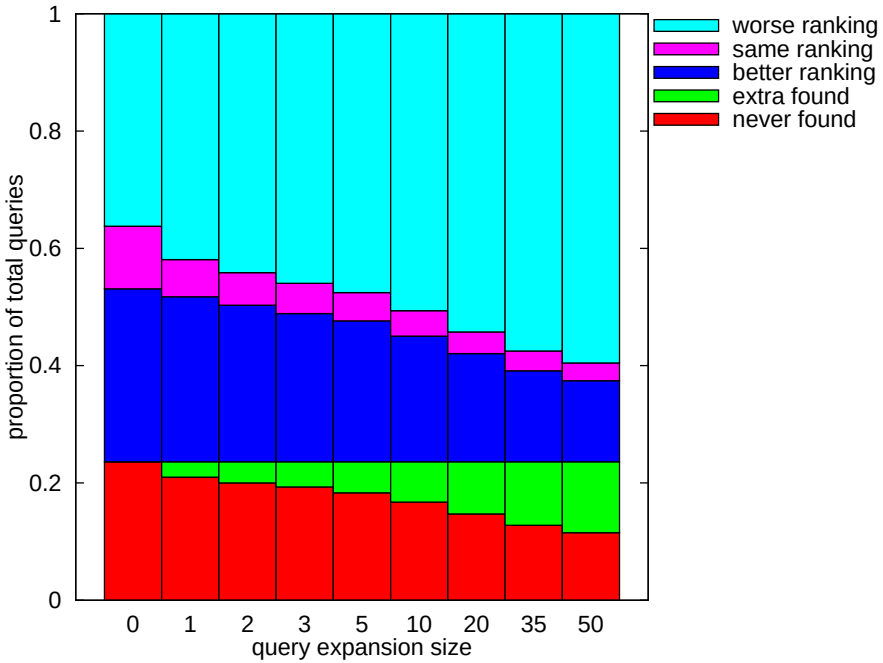
Personalized Query Expansion

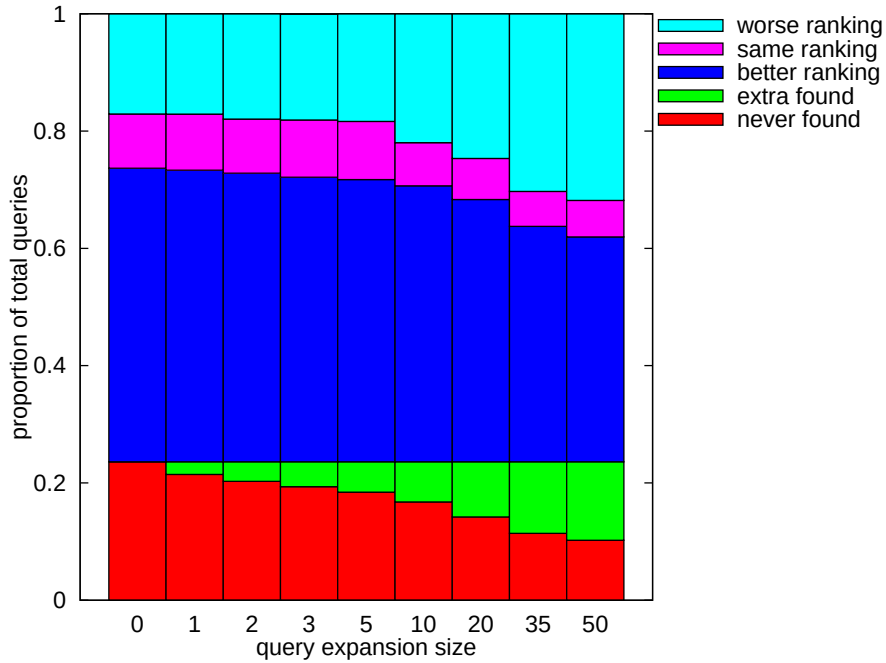


Multi-Interest Clustering



TagRank


worse ranking
same ranking
better ranking
extra found
never found
proportion of total queries
query expansion size
0
0.2
0.4
0.6
0.8
1
0 1 2 3 5 10 20 35 50


worse ranking
same ranking
better ranking
extra found
never found
proportion of total queries
query expansion size
0
0.2
0.4
0.6
0.8
1
0 1 2 3 5 10 20 35 50

Conclusion

- efficient personalized query expansion
- anonymity
- explore similarity metrics
- decentralized search
- recommendations
- ...

Acknowledgement

We warmly thank Vivien Quéma for his help at early stages of this work.