

Thematically Focused Search in Web 2.0 Folksonomies

Sergej Sizov, Olaf Görlitz

University of Koblenz-Landau

Germany

Vannevar Bush's Memex (1945) & Mass Intelligence (2008)



can this huge mass of users give us new insights that would not be possible by considering individual contributions ?!

The mass makes the difference?



Explore / Tags / **www** / clusters

[web](#), [internet](#), [website](#), [design](#), [site](#), [blog](#),
[webdesign](#), [google](#), [flickr](#), [screenshot](#)

➔ [See more in this cluster...](#)

[canada](#), [banff](#), [alberta](#), [rockies](#), [mountains](#),
[hotel](#), [mountain](#), [fairmont](#), [bowriver](#), [lakelouise](#)

➔ [See more in this cluster...](#)

BibSonomy :: tag ::

A blue social bookmark and publication sharing system.

- related tags

+ web + article + 2007 + conference + search + wwwbook + internet + 2006 + semanticWeb + research + workshop + searchengine + history + web2.0 + studienarbeit + semanticweb + society + wwwkap1 + tagging + w3c + wwwkap17 + social + xml + semantic + blog + rezeption + semwebss06 + scale-free-networks + informationRetrieval + google + pkipper + webcrawler + book + folksonomy + navigation + kommunikation + clustering + browser + swss06-01 + ajax + library + statistics + graph + tv + mediennutzung + selektion + collaboration + bookmark + background + metadata



del.icio.us / search

Search results for **www**

related tags: [web](#) [webdesign](#) [css](#) [reference](#) [design](#) [internet](#) [software](#) [programming](#) [html](#) [tools](#)



WIKIPEDIA
Die freie Enzyklopädie

YouTube
Broadcast Yourself™

[Neues Konto](#) | [Konto](#) | [Protokoll](#) | [Hilfe](#) | [Anmelden](#) | [Site:](#)

Startseite

Videos

Kanäle

Community

Videos ▾

Suchen

[Einstellungen](#)

Video hochladen

YouTube will be undergoing scheduled maintenance, starting around 7:00 pm PDT.

“www” Videoergebnisse 1 - 20 von etwa 80.100.000

Anzeige:

Suchen nach: **Videos** | [Kanäle](#)

Sortieren nach:
Relevanz | [Datum](#)

Hochgeladen: [Heute](#) | [Diese Woche](#) | **Insgesamt**

Flickr example – what you get ..



Sie sind nicht angemeldet [Anmelden](#) [Hilfe](#)

[Startseite](#) [Die Tour](#) [Registrieren](#) [Entdecken](#) ▾

[Suchen](#) ▾

Suchen

Fotos

[Gruppen](#)

[Personen](#)

SUCHEN

[Erweiterte Suche](#)

[Nach Kamera suchen](#)

Volltext Nur Tags

✔ Wir haben 5 Ergebnisse für Fotos für **athos** und **fire** gefunden.

[Als Diashow anzeigen](#) (⌂)

Anzeigen: [Relevanteste](#) • [Neueste](#) • [Interessanteste](#) Zeige: [Details](#) • [Thumbnails](#)



Mt. Athos Fire Bread

Hochgeladen: 3. Oktober 2007



Von **Let them eat cake.**

[Mehr Fotos](#) oder ihr [Profil](#) ansehen.

[bread](#), [germanvillage](#), [nikonf3](#), [columbusoh](#)

...



Monastery Grigoriou

Hochgeladen: 1. Februar 2008



Von **Makednos**

[Mehr Fotos](#) oder sein [Profil](#) ansehen.

[door](#), [wood](#), [blue](#), [trees](#) ...

Sponsoren-Links

[Mount Athos Wein](#)

Den griechischen Rotwein-Klassiker günstig bestellen bei Hawesko.
www.hawesko.de

[Athos günstig kaufen](#)

Riesenauswahl zu Niedrigpreisen. eBay Käuferschutz bis € 200,00.
www.ebay.de

[Fire](#)

Lassen Sie sich das nicht entgehen. Schnäppchenpreise gibt's bei uns.
www2.superangebote24.de

[Alodisa - Ihr Spezialist für Räucherwerk](#)

Wie bieten Ihnen feinstes Räucherwerk, Räucherkohle und Ikonen und weitere christliche und...

Query: "Athos fire"

У Светој српској царској лаври Хиландару на Светој Гори Атонској, у ноћи између 3. и 4. марта 2004. године, избио је пожар великих размера.



Flickr: queries with low recall



Sie sind nicht angemeldet [Anmelden](#) [Hilfe](#)

[Startseite](#) [Die Tour](#) [Registrieren](#) [Entdecken](#) ▾

[Suchen](#) ▾

Suchen

Fotos

[Gruppen](#)

[Personen](#)

bamberg aula

SUCHEN

[Erweiterte Suche](#)
[Nach Kamera suchen](#)

Volltext Nur Tags

! Wir konnten keine Ergebnisse zu für **bamberg** und **aula** finden.

Möchten Sie stattdessen nach [germany](#), [deutschland](#), [bavaria](#), [bayern](#) or [franken](#) suchen?

Sie [Anmelden](#) | [Kostenlosen Account einrichten](#)

Entdecken [Places](#) | [Letzte 7 Tage](#) | [Dieser Monat](#) | [Beliebte Tags](#) | [Creative Commons](#) | [Suchen](#)

Hilfe [Community-Richtlinien](#) | [Hilfeforum](#) | [FAQ](#) | [Sitemap](#) | [E-Mail-Hilfe](#)

[Flickr Blog](#) | [Über Flickr](#) | [Nutzungsbedingungen](#) | [Datenschutz](#) | [Copyright-Richtlinien](#) | [Missbrauch melden](#)

a **YAHOO!**company

[繁體中文](#) | [Deutsch](#) | [English](#) | [Español](#) | [Français](#) | [한글](#) | [Italiano](#) | [Português](#)

Copyright © 2008 Yahoo! Alle Rechte vorbehalten.

Flickr: queries with low recall (2)



Sie sind nicht angemeldet [Anmelden](#)

[Startseite](#) [Die Tour](#) [Registrieren](#) [Entdecken](#) ▾

[Such](#)

Suchen

Fotos [Gruppen](#) [Personen](#)

SUCHEN

[Erweiterte Suche](#)
[Nach Kamera suchen](#)

Volltext Nur Tags

✓ Wir haben 2 **Ergebnisse** zu für **bamberg** und **dominikanerbau** gefunden.

[Diashow](#)

Anzeigen: [Relevanteste](#) • [Neueste](#) • [Interessanteste](#) Zeige: [Details](#) • [Thumbnails](#)



P1410173

Hochgeladen: 26. November 2007



Von **dmonniaux**

[dmonniaux](#) – [Fotostream](#) oder [Profil](#).

[bamberg, dominikanerbau](#)

ADVERTISEMENT

[\[Zum Anzeigen klicken\]](#)

Outline

Problem formalization

Thematically focused search and ranking

Distributed setting – pro & contra

Evaluation

Formalizing the problem

Collaborative content sharing framework:

users $u \in U$ tags $t \in T$ resources $r \in R$

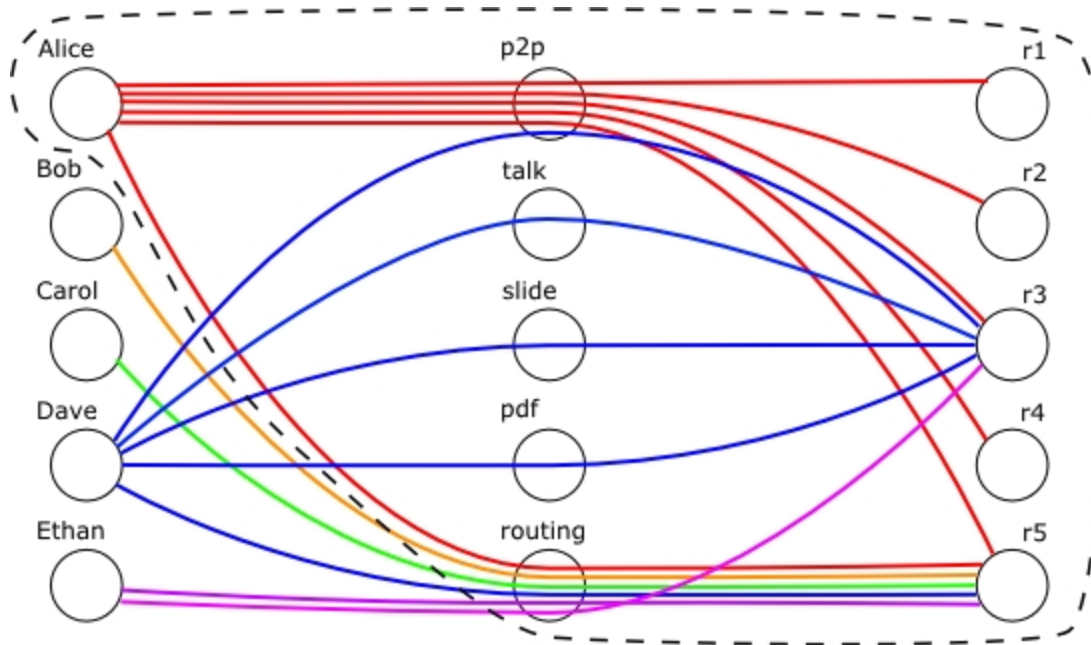
Information cloud: $T := (Y^*, f), f(t) : Y^* \subseteq Y, Y^* \rightarrow [0..1]$

- ◆ user-centric: $T_u := (Y_u, f), Y_u \subseteq u \times T \times R,$
- ◆ resource-centric: $T_r := (Y_r, f), Y_u \subseteq U \times T \times r,$
- ◆ community-specific: $T_{U^*} := (Y_{U^*}, f), Y_{U^*} \subseteq U^* \times T \times R$
- ◆ collection-specific $T_{R^*} := (Y_{R^*}, f), Y_{R^*} \subseteq U \times T \times R^*$
- ◆ arbitrary $T_{U^*R^*} := (Y_{U^*R^*}, f), Y_{U^*R^*} \subseteq U^* \times T \times R^*$
.. e.g. obtained by traversing the hypergraph up to certain depth

Common recommender scenarios:

- ◆ Given a user, recommend photos which may be of interest.
- ◆ Given a user, recommend users they may like to contact.
- ◆ Given a user, recommend groups they may like to join.

The IR background – constructing feature vectors



$$if(i) = (a_i, b_i)$$

$$iif(i) = \left(\log \frac{|J|}{|J^*|}, \log \frac{|K|}{|K^*|} \right),$$

$j \in J^*/k \in K^* :$
 $(i, j, k) \in Y^*$

$$weight(i) = iif(i) \cdot if(i)^T$$

.. defined analogously to tf-idf

tag features:

$$iif \cdot if(p2p) = \left(\log \frac{5}{2}, \log \frac{5}{5} \right) \cdot (6, 6)^T$$

$$iif \cdot if(talk) = \left(\log \frac{5}{1}, \log \frac{5}{1} \right) \cdot (1, 1)^T$$

$$iif \cdot if(slide) = \left(\log \frac{5}{1}, \log \frac{5}{1} \right) \cdot (1, 1)^T$$

$$iif \cdot if(pdf) = \left(\log \frac{5}{1}, \log \frac{5}{1} \right) \cdot (1, 1)^T$$

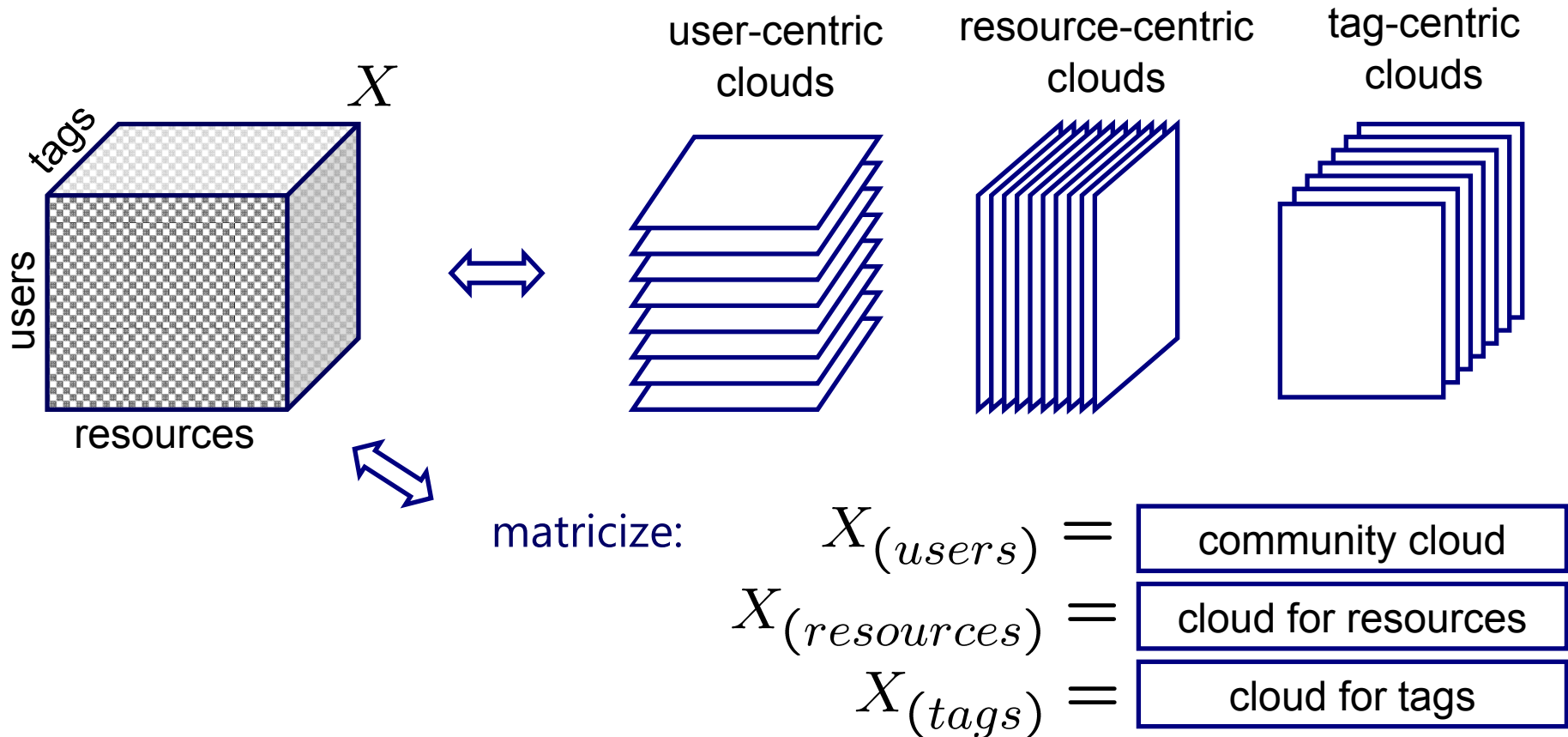
$$iif \cdot if(routing) = \left(\log \frac{5}{5}, \log \frac{5}{1} \right) \cdot (6, 6)^T$$

Further dimensions of interest:

- favorites
- group membership
- contact lists
- comments on other's resources

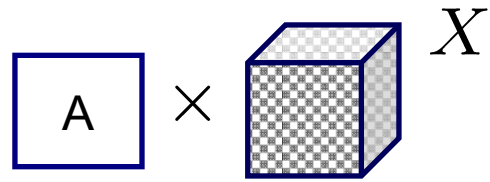
Generalization: the tensor model

idea: using multi-dimensional arrays for representing relationships



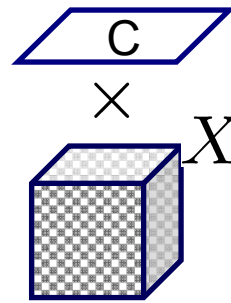
slide shows 3rd order tensor for common Web 2.0 dimensions, but can (and should) be extended by other relationships (favorites, comments, groups..)

Tensors: mode-n matrix multiplication, Tucker decomposition,..



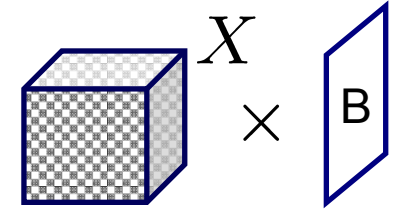
$$Y = X \times_1 A$$

$$Y_{(:,k)} = X_{(:,k)} \times A^T$$



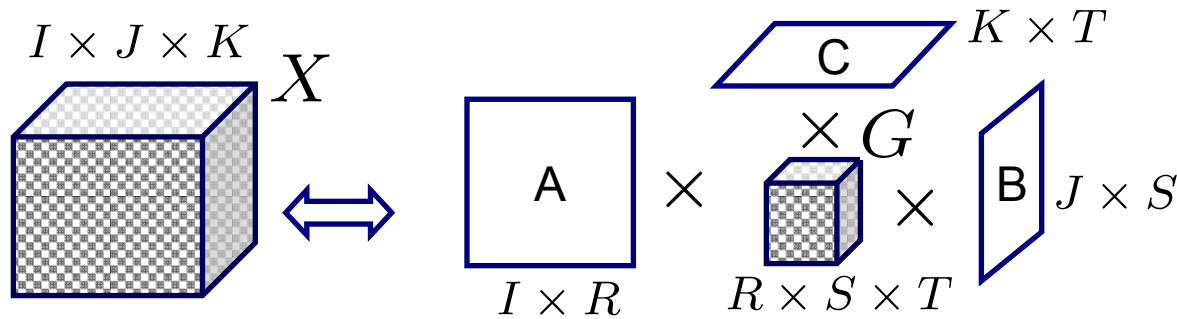
$$Y = X \times_3 B$$

$$Y_{(i,:)} = X_{(i,:)} \times C^T$$



$$Y = X \times_2 B$$

$$Y_{(:,j)} = X_{(:,j)} \times B^T$$



general idea: decompose tensor in order to identify significant "factors" along each dimension (multi-dimensional methods analogously to LSI, PCA)

our current approach: input-tensor R is decomposed into $R=UxDxV$, V contains the orthogonal mapping of R into space of real tensors with target dimension (i.e. V is also a matrix). Restrict the resulting feature vectors to 5-10 most significant dimensions, analogously to LSI.

User-centered focusing

1. Compute characteristic feature vectors for resources, tags, contacts, or favorites of the given user
2. Construct appropriate decision model (centroid, naive bayes, SVM, etc.)
3. Explore the tagging cloud around the user, order matches according wrt estimated utility function (cosine similarity, classification confidence, etc.)
4. Return the top-k result set (e.g. top-10, top-20) to the user

Datasets

- ◆ Flickr dataset (2004-2005)
 - ◆ 319,686 users,
 - ◆ 1,607,879 tags,
 - ◆ 28,153,045 resources,
 - ◆ 112,900,000 tag assignm.
- ◆ Del.icio.us dataset (2003-2006)
 - ◆ 532,924 users,
 - ◆ 2,481,698 tags,
 - ◆ 17,262,480 resources,
 - ◆ 140,126,586 tag assignm.

Evaluation: apriori method

- remove a certain fraction of relationships (e.g. group participation, comments, ..) from the test cloud
- test the ability of the recommender to reconstruct missing relationships (i.e. to place them within top-k of the result set)

Results: user-focused recommendations

recommending favorites

User representation	Training:10 prec@10	Training:10 prec@20
<i>Random</i>	0.167	0.167
<i>User items</i>	0.259	0.268
<i>Commented items</i>	0.236	0.221
<i>Favorites</i>	0.872	0.727
<i>Combined</i>	0.854	0.713
	Training:20 prec@10	Training:20 prec@20
<i>Random</i>	0.167	0.167
<i>Commented items</i>	0.255	0.248
<i>Favorites</i>	0.918	0.851
<i>Combined</i>	0.899	0.828
	Training:40 prec@10	Training:40 prec@20
<i>Random</i>	0.167	0.167
<i>Commented items</i>	0.265	0.266
<i>Favorites</i>	0.933	0.903
<i>Combined</i>	0.914	0.876

recommending contacts

User representation	prec@5	prec@10
<i>Random</i>	0.167	0.167
<i>User Items</i>	0.574	0.472
<i>Commented Items</i>	0.576	0.473
<i>Favorites</i>	0.535	0.455
<i>Contacts (training 10)</i>	0.604	0.497
<i>Contacts (training 20)</i>	0.611	0.498

tensor based recommendation: consistently better accuracy
in preliminary experiments, now under evaluation

Decentralized setting: pro & contra argumentation



- ☹ multiple accounts for different resource types
- ☹ space limitations (e.g. max 200 photos in Flickr)
- ☹ censorship, rank manipulations
- ☹ single point of failure

Distributed Tagging System ?

- ◆ tag any kind of personal data on the Desktop
- ◆ share and browse tagged data in a P2P network

Our implementation: Tagster

open source, available at <http://isweb.uni-koblenz.de>

Meta Methods: General Model

given: set of methods $V = \{v_1, \dots, v_k\}$, confidence grades $res(v_i, d)$ for document d

Meta result (restrictivity by thresholds t_1 and t_2 , tuning by weights $w(v_i)$):

$$Meta(d) = \begin{cases} +1 & \text{if } \sum_i res_i(d) \cdot w(v_i) > t_1 \\ -1 & \text{if } \sum_i res_i(d) \cdot w(v_i) < t_2 \\ 0 & \text{otherwise} \end{cases}$$

Special cases:

- “Unanimous Decision”
- “Voting”
- “Weighted Average” (e.g., weighted by some quality estimator)

Collaborative organization of document collections

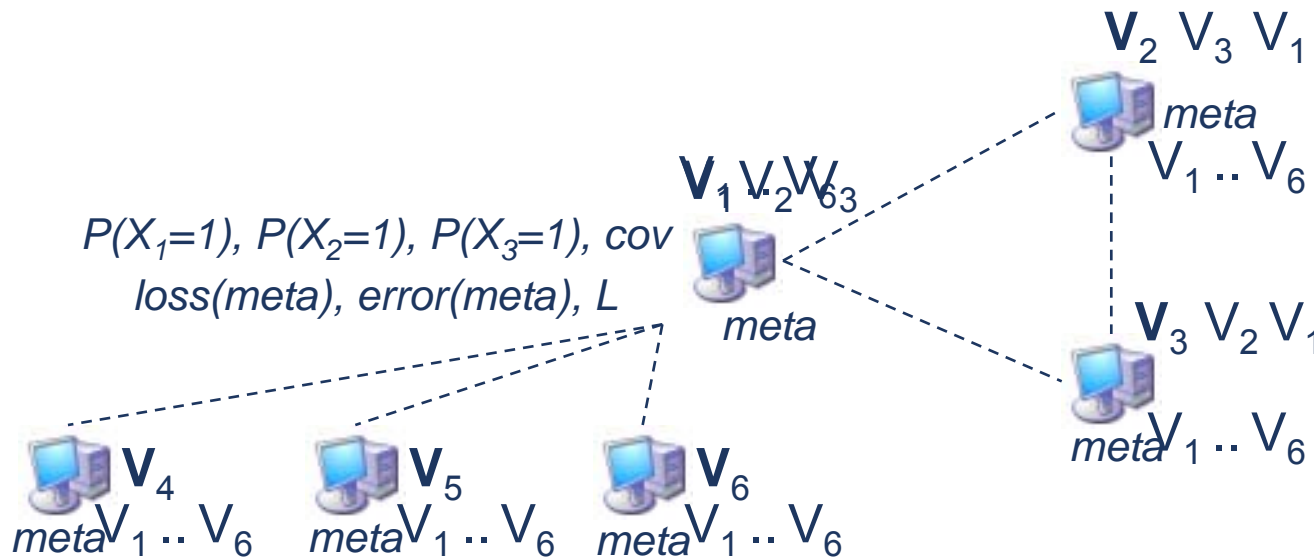
Given: set of methods $V = \{v_1, \dots, v_L\}$, „unanimous decision“

$$X_i = \begin{cases} 1 & \text{if } v_i \text{ assigns document correctly} \\ 0 & \text{otherwise} \end{cases}$$

$$P(X_1 = 1, \dots, X_L = 1) = P(X_1 = 1) \cdot \prod_{i=1}^{L-1} \frac{P(X_i = 1)P(X_{i+1} = 1) + \text{cov}(X_i, X_{i+1})}{P(X_i = 1)}$$

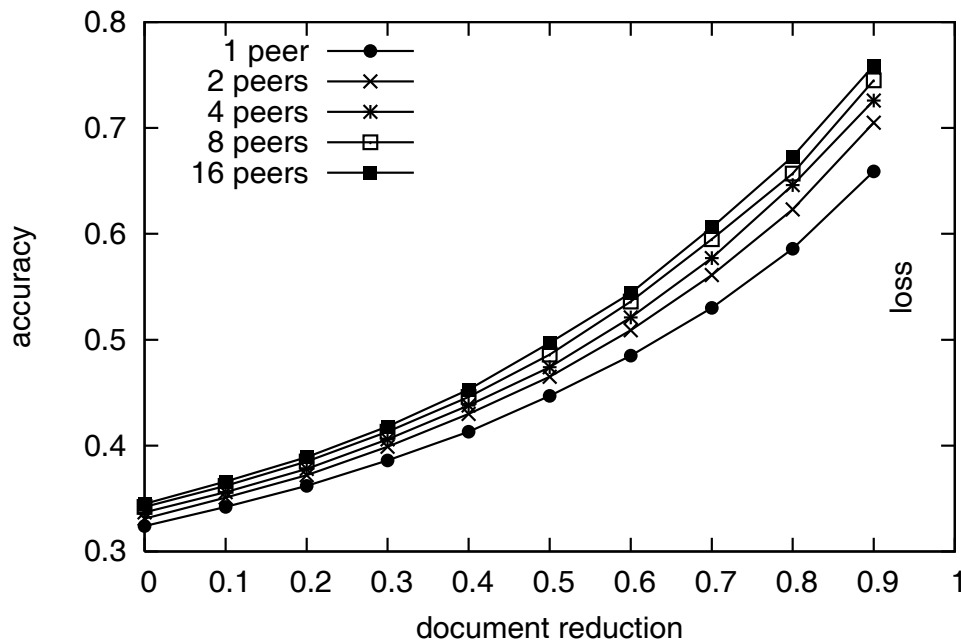
$$\text{error}(\text{meta}) = P(X_1 = 0, \dots, X_L = 0 | X_1 = \dots = X_L)$$

$$\text{loss}(\text{meta}) = 1 - P(X_1 = \dots = X_L)$$

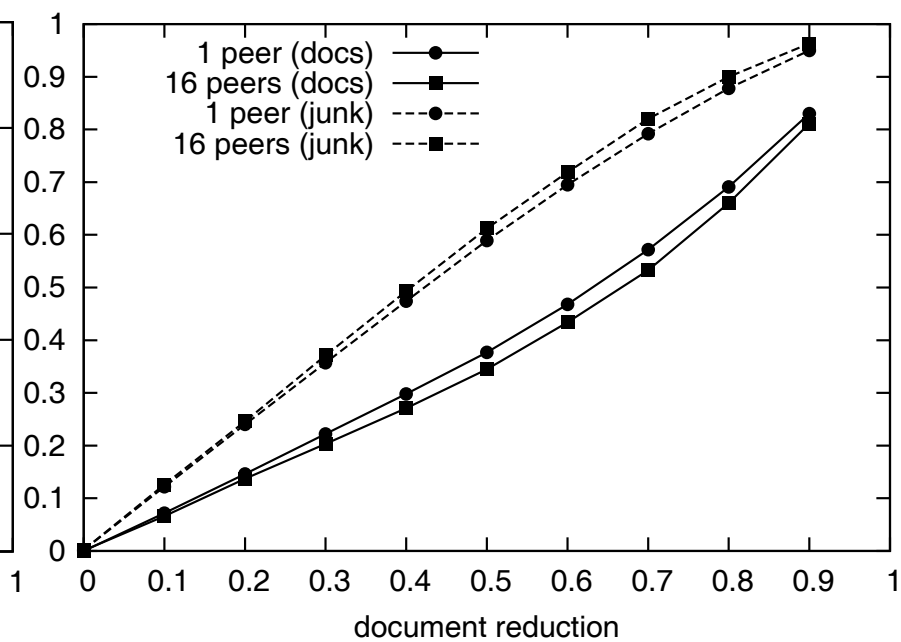


Decentralized Collaboration: Results

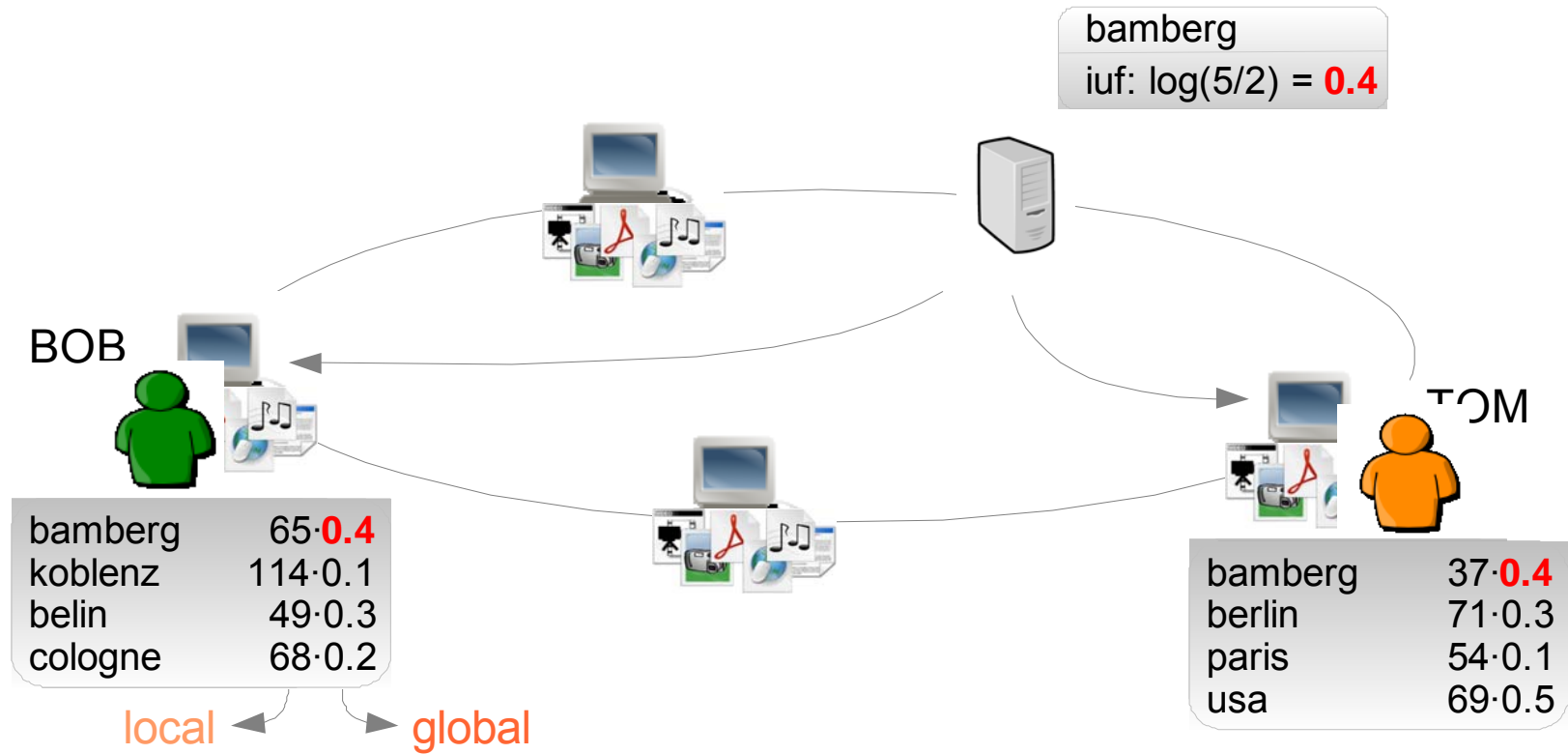
Accuracy: del.icio.us, Junk=1/2



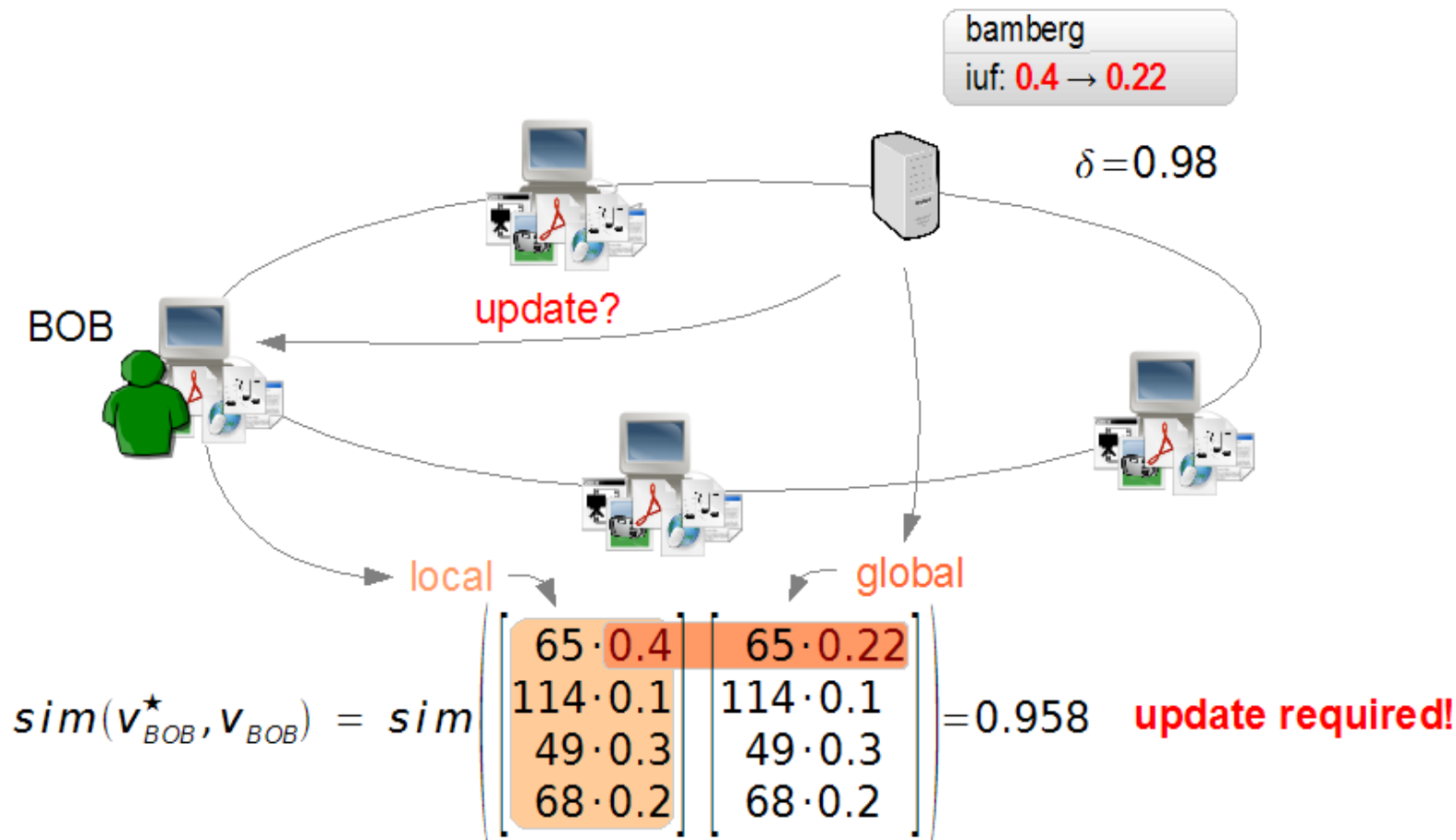
Junk Reduction and Document Loss: del.icio.us, Junk=1/2



Distributed scenario: an example



Distributed scenario: example (2)



The PINTS approach

- ◆ Index peers monitor feature vector accuracy for their tags
- ◆ Compare feature approximation with the tag's true *iuf* value



BOB's approximation:

$$\mathbf{v}_{BOB}^*(\theta) = \begin{pmatrix} tf(t_1) \cdot (a_{t_1} \cdot \theta + b_{t_1}) \\ \vdots \\ tf(t_m) \cdot (a_{t_m} \cdot \theta + b_{t_m}) \\ \vdots \\ tf(t_N) \cdot (a_{t_N} \cdot \theta + b_{t_N}) \end{pmatrix}$$



index peer's view:

$$\mathbf{v}_{BOB, t_m}^\circ(\theta) = \begin{pmatrix} tf(t_1) \cdot (a_{t_1} \cdot \theta + b_{t_1}) \\ \vdots \\ tf(t_m) \cdot iuf_{t_m}^{true} \\ \vdots \\ tf(t_N) \cdot (a_{t_N} \cdot \theta + b_{t_N}) \end{pmatrix}$$

- ◆ Index peer needs to know the other approximations
- ◆ Vector similarity must be above threshold δ

$$sim(\mathbf{v}^*, \mathbf{v}_{t_m}^\circ) > \delta$$

$$sim(\mathbf{v}^*, \mathbf{v}_{t_m}^\circ) = \frac{\mathbf{v}^* \cdot \mathbf{v}_{t_m}^\circ}{\|\mathbf{v}^*\| \|\mathbf{v}_{t_m}^\circ\|}$$

The PINTS approach (2)

$$\text{sim}(v^*, v_{t_m}^\circ) = \frac{v^* \cdot v_{t_m}^\circ}{\|v^*\| \|v_{t_m}^\circ\|}$$

$$v^* \cdot v_{t_m}^\circ = \sum_{t_i \neq t_m} (tf(t_i)^2 \cdot (a_{t_i} \cdot \theta + b_{t_i})^2) + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m}) \cdot iuf_{t_m}^{true}$$

$$\|v^*\| = \sqrt{\sum_{t_i \neq t_m} (tf(t_i)^2 \cdot (a_{t_i} \cdot \theta + b_{t_i})^2) + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m})^2}$$

$$\|v_{t_m}^\circ\| = \sqrt{\sum_{t_i \neq t_m} (tf(t_i)^2 \cdot (a_{t_i} \cdot \theta + b_{t_i})^2) + tf(t_m)^2 \cdot (iuf_{t_m}^{true})^2}$$

$$A_{t_m} = \sum_{t_i \neq t_m} (tf(t_i)^2 \cdot a_{t_i}^2) \quad B_{t_m} = \sum_{t_i \neq t_m} (tf(t_i)^2 \cdot a_{t_i} \cdot b_{t_i}) \quad C_{t_m} = \sum_{t_i \neq t_m} (tf(t_i)^2 \cdot b_{t_i}^2)$$

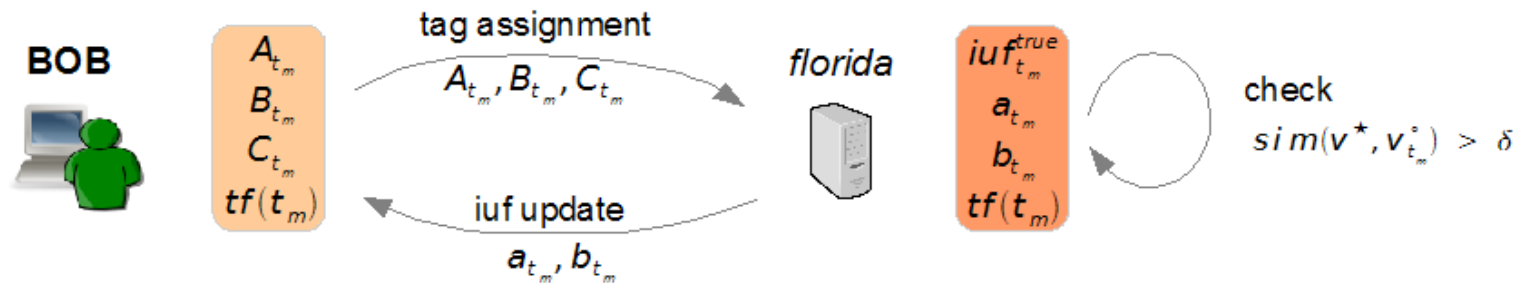
$$v^* \cdot v_{t_m}^\circ = A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m}) \cdot iuf_{t_m}^{true}$$

$$\|v^*\| = \sqrt{A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m})^2}$$

$$\|v_{t_m}^\circ\| = \sqrt{A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (iuf_{t_m}^{true})^2}$$

PINTS: update strategy

$$\text{sim}(v^*, v_{t_m}^\circ) = \frac{v^* \cdot v_{t_m}^\circ}{\|v^*\| \|v_{t_m}^\circ\|}$$

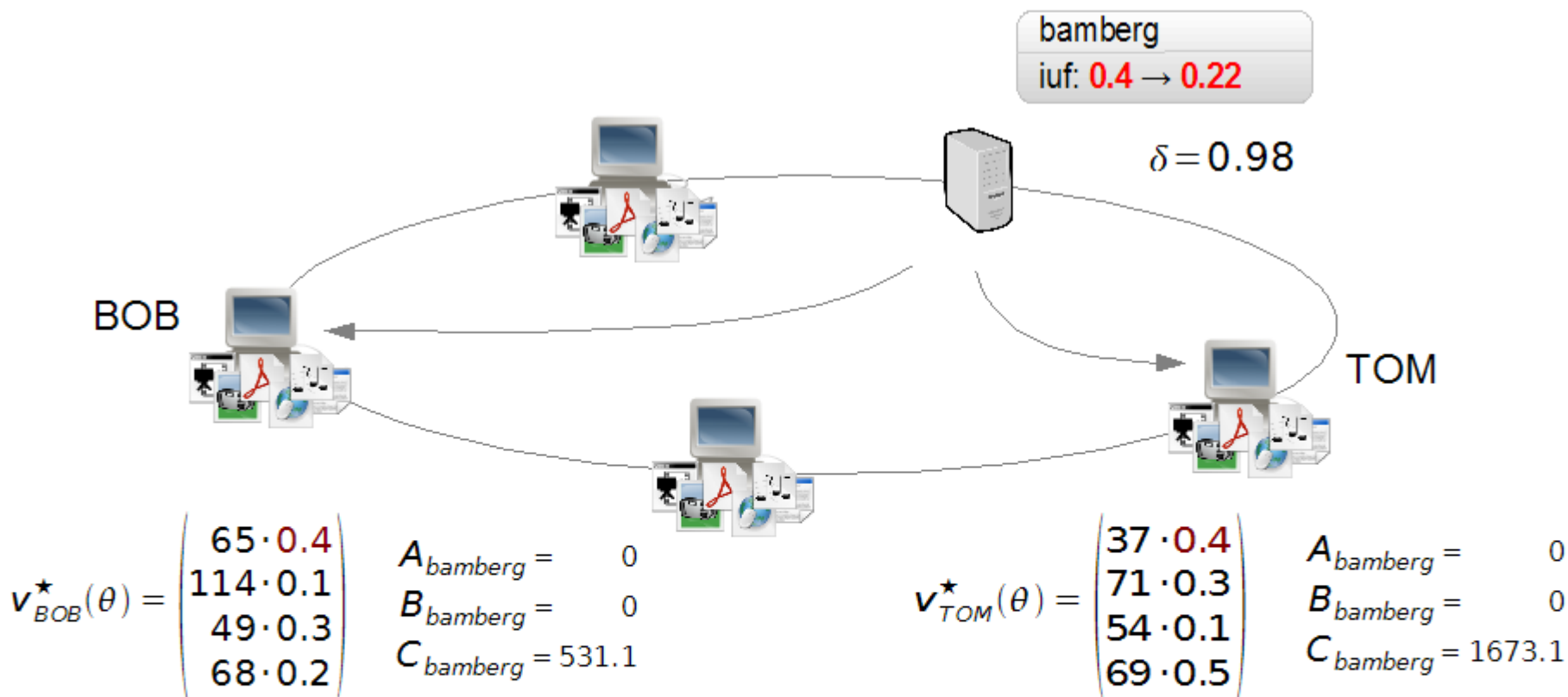


$$v^* \cdot v_{t_m}^\circ = A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m}) \cdot iuf_{t_m}^{true}$$

$$\|v^*\| = \sqrt{A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (a_{t_m} \cdot \theta + b_{t_m})^2}$$

$$\|v_{t_m}^\circ\| = \sqrt{A_{t_m} \theta^2 + 2 B_{t_m} \theta + C_{t_m} + tf(t_m)^2 \cdot (iuf_{t_m}^{true})^2}$$

PINTS updates: Beispiel



$$sim(v_{BOB}^*, v_{BOB, bamberg}^\circ) = 0.958$$

update required

$$sim(v_{TOM}^*, v_{TOM, bamberg}^\circ) = 0.989$$

no update required

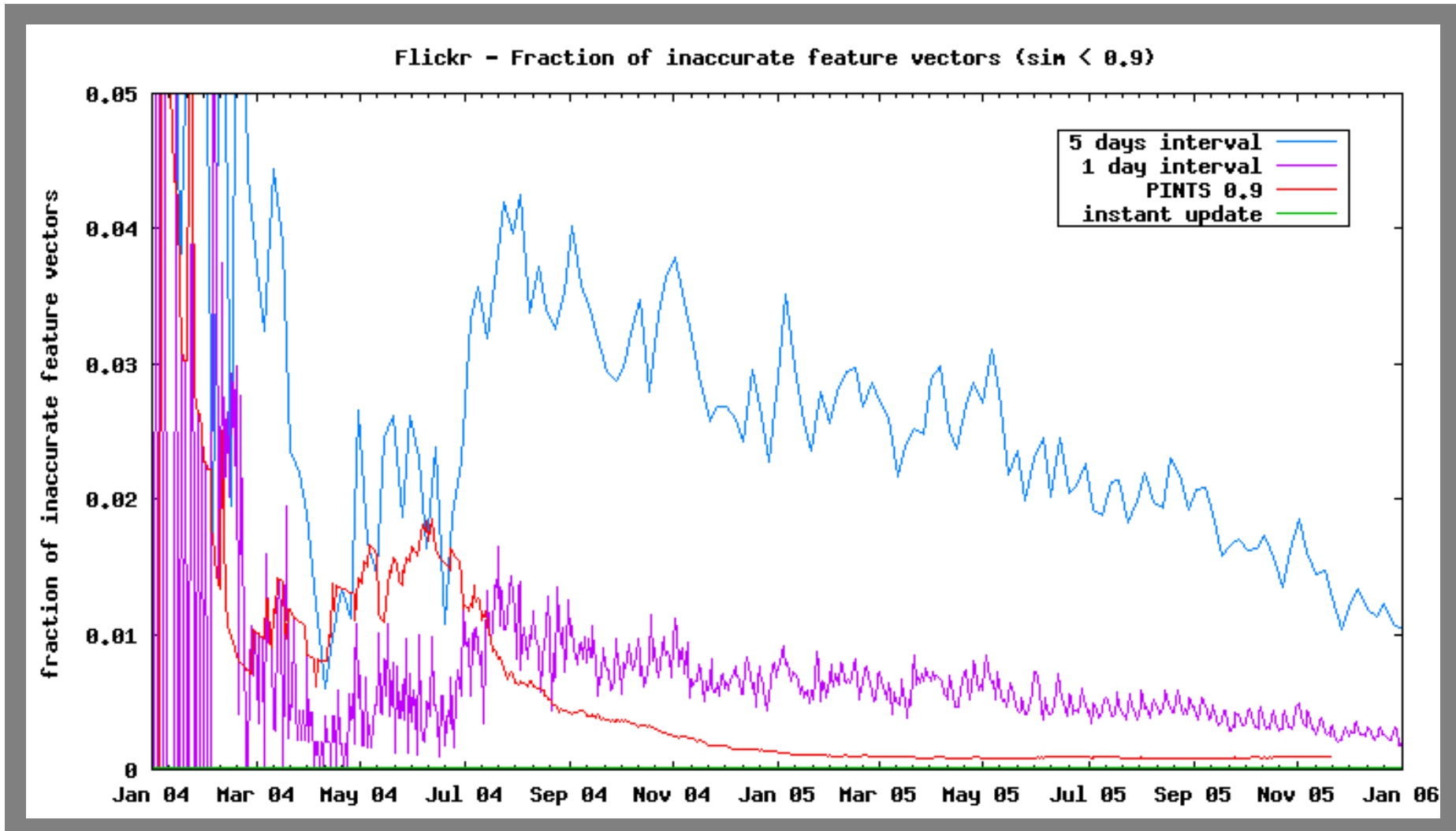
Objectives

- ◆ check if (and how frequent) specified thresholds violated
- ◆ compare message complexity for various methods

Methodology

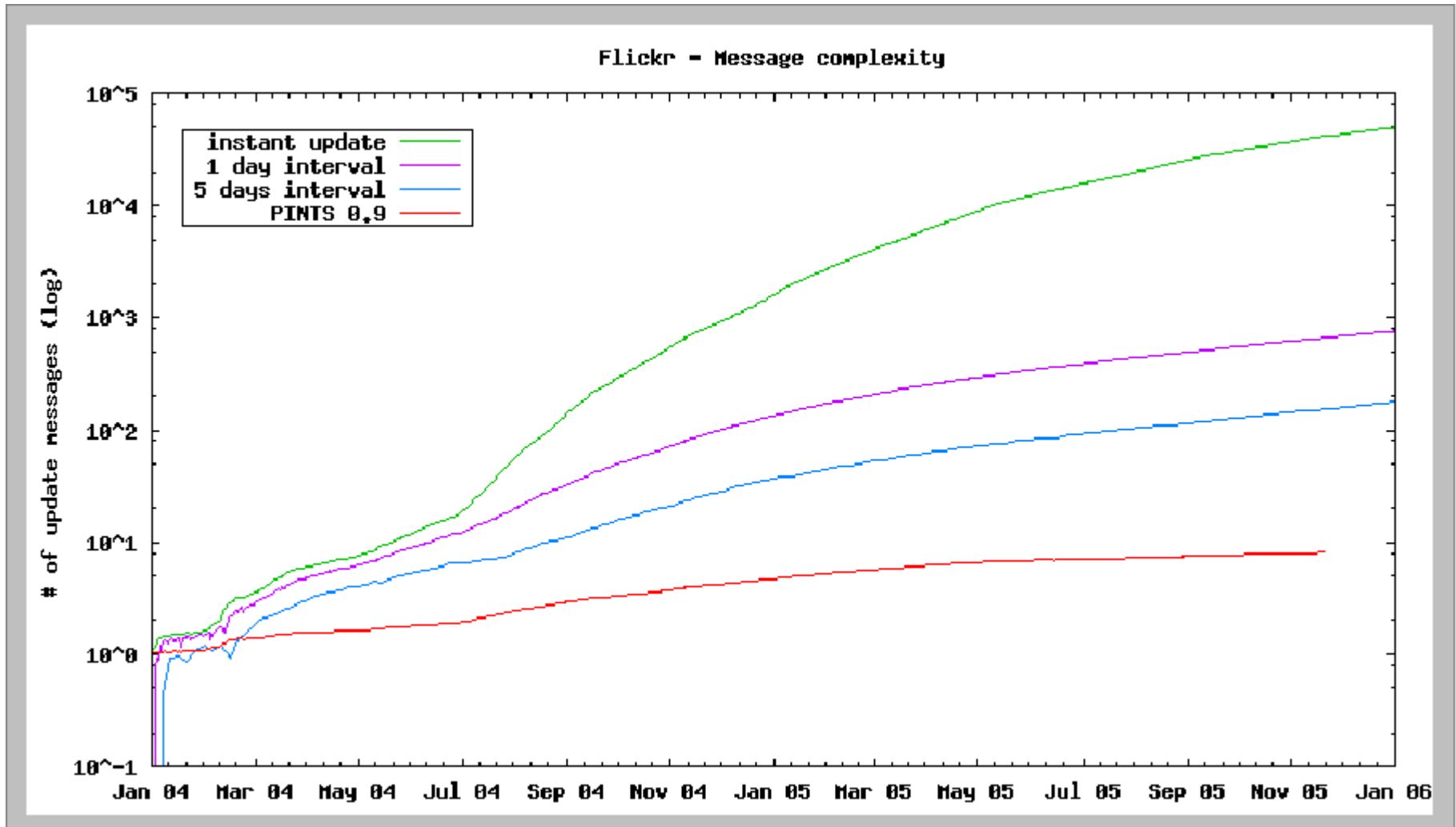
- ◆ use real world tagging traces (time-ordered tag assignm.)
 - flickr.com (~320k users, ~1.6m tags, ~28.2m resources)
 - del.icio.us (~533k users, ~2.3m tags, ~17.3m resources)
- ◆ replay tagging traces in P2P simulation
- ◆ measure inaccurate feature vectors, message complexity
- ◆ evaluate against interval-based update

PINTS Evaluation (2)



PINTS: higher accuracy than interval updates

PINTS Evaluation (2)



PINTS: high accuracy at low message complexity

Conclusions

Conclusions

Focused search & recommendation in Web 2.0 folksonomies:

- ◆ IR-like problem formalization
- ◆ Personal and social aspects/dimensions are important
- ◆ Multi-dimensional setting helps to improve accuracy
- ◆ Can be realized for centralized and decentralized architectures

Future work

- ◆ bridging the semantic gap between low-level and high-level features
- ◆ decentralized computations on large sparse matrices (e.g P2P based PageRank or HITS estimation)
- ◆ evaluation methodology for Web 2.0 applications
- ◆ better understanding of Web 2.0 evolution patterns

thank you