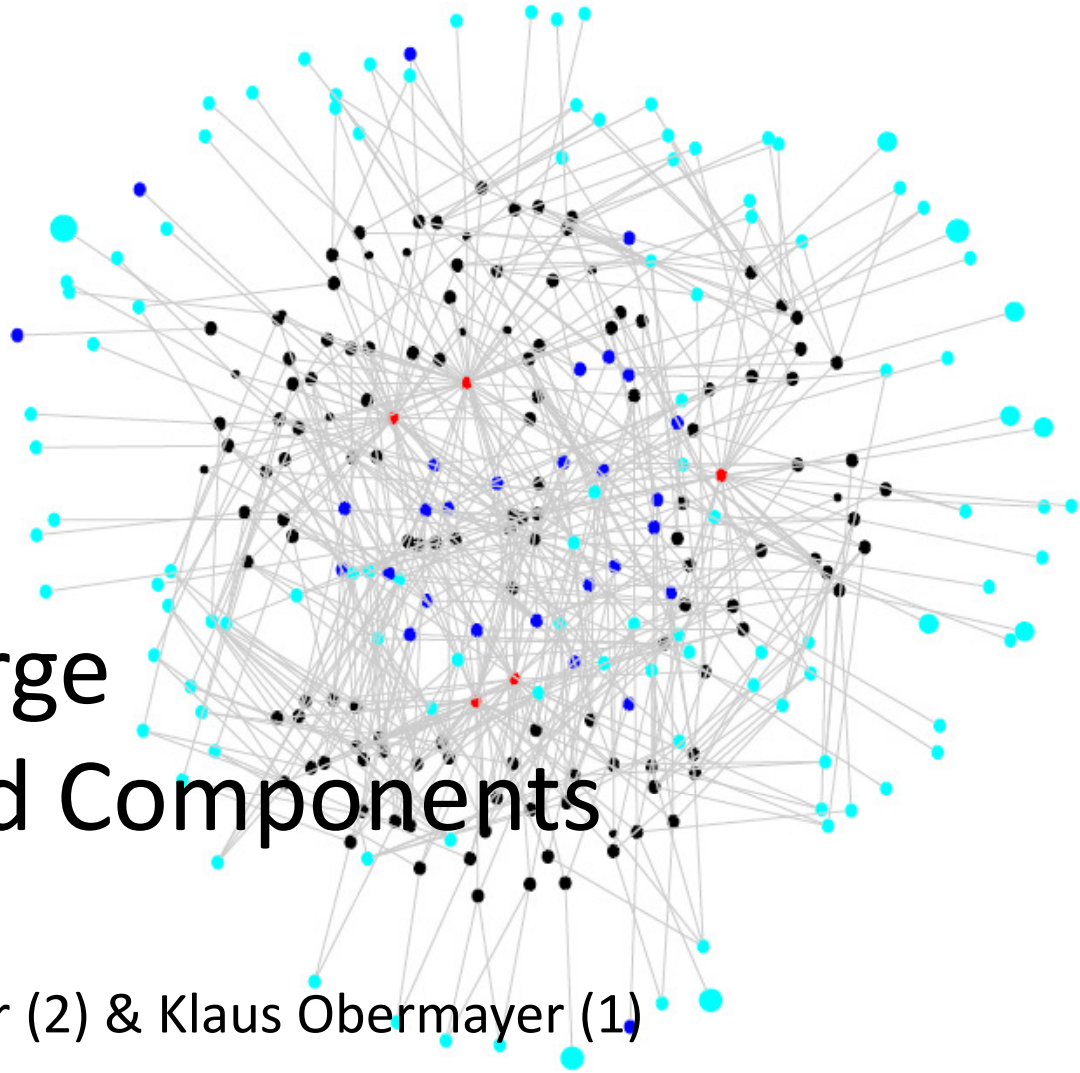


Tag Spam Creates Large Non-Giant Connected Components

Nicolas Neubauer (1), Robert Wetzker (2) & Klaus Obermayer (1)

Neural Information Processing Group (1), DAI Lab (2)

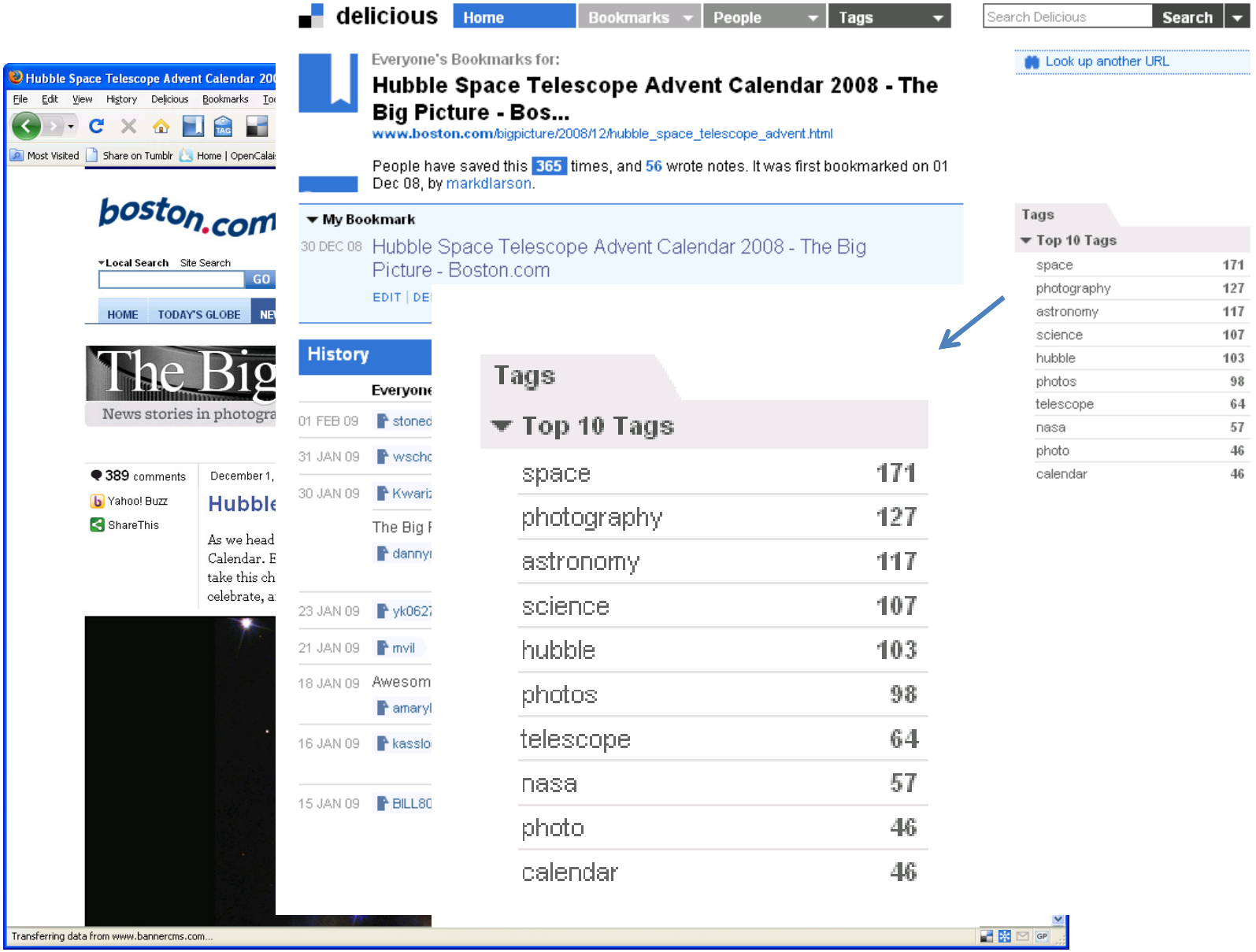
Technische Universität Berlin

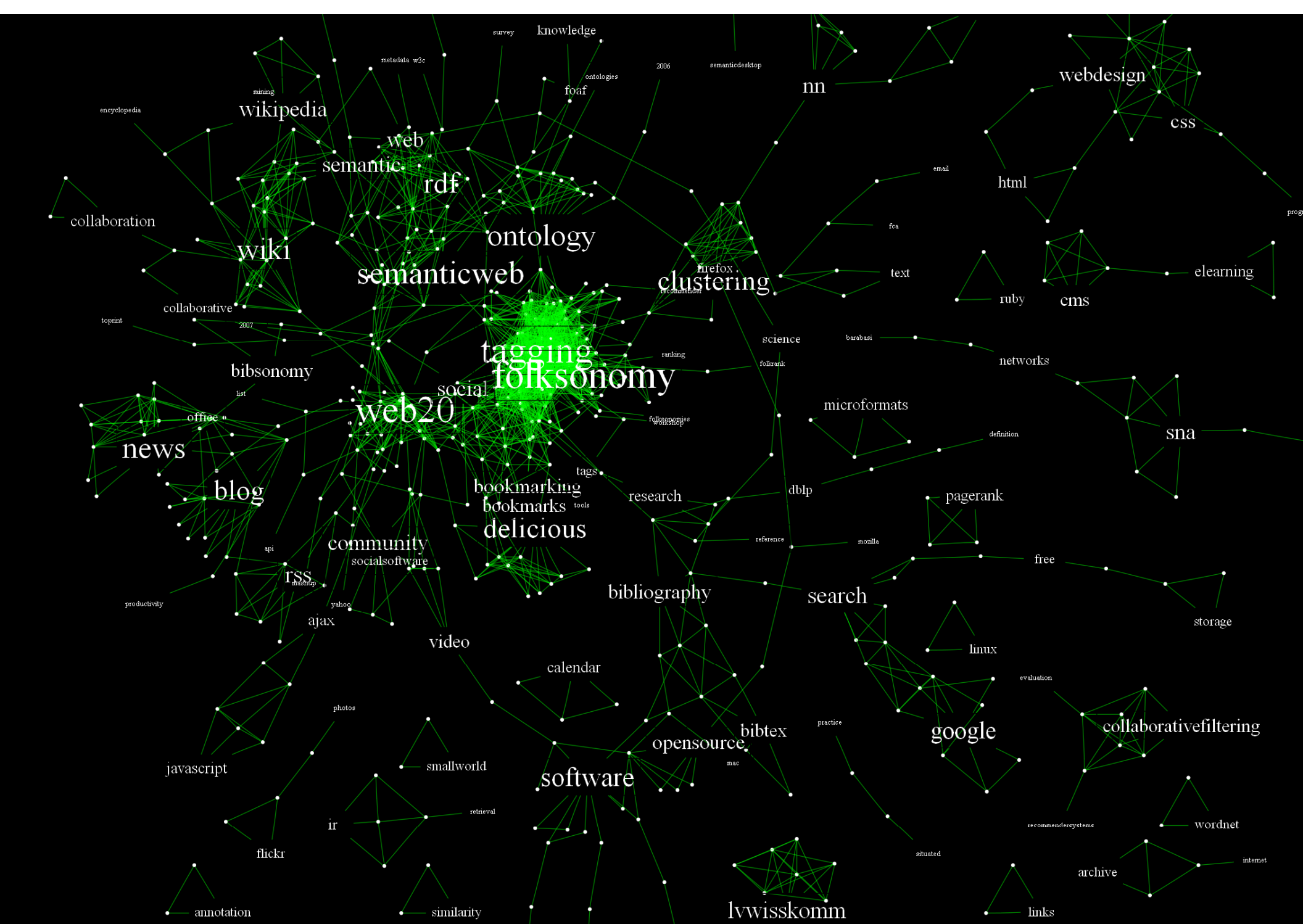
AIRWeb@WWW'09, 21.4.2009

Overview

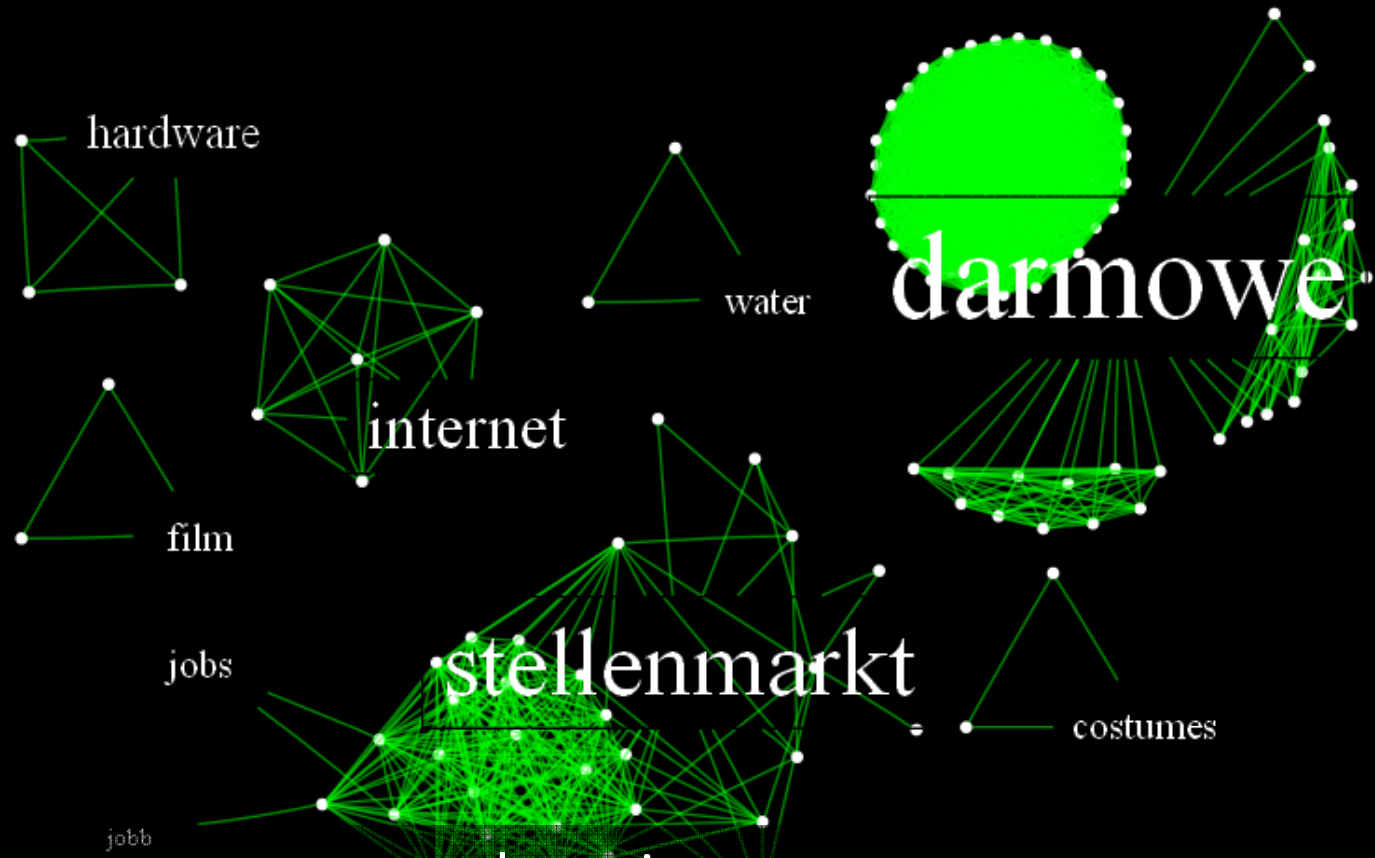
1. Spam in Social Bookmarking Systems
2. Hyperincident Connected Components
3. Document/User and Tag/User Graphs
4. Conclusions

Social Tagging





Edges: Top 2000 similarities between top 800 documents (no spam) - Bibsonomy



- Some tag spam targets search engines
  - Top entry for a given tag might indicate relevance
- Other tag spam targets users
  - Sites with certain tags might lure users into visiting them
- Spammers behave so radically different it shows in the resulting network structures

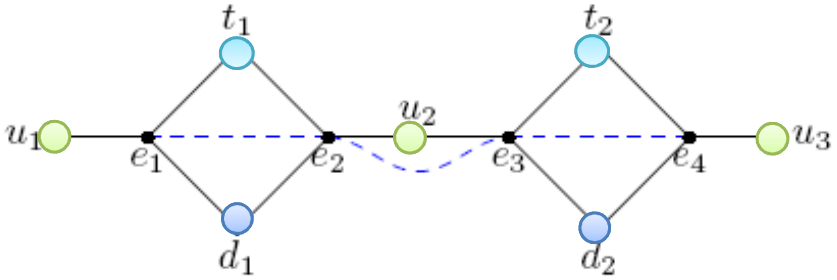
Edges: Top 2000 similarities between top 800 documents (spam) - Bibsonomy

Overview

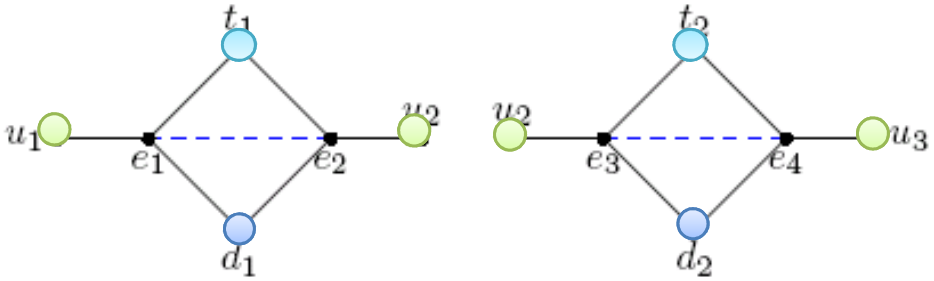
1. Spam in Social Bookmarking Systems
2. Hyperincident Connected Components
3. Document/User and Tag/User Graphs
4. Conclusions

Hyperincident Connectivity

- Tagging data can be interpreted as a hypergraph, defined by hyperedges (d,u,t) for a document d being tagged with tag t by a user u
- Two edges are incident if they share a node (i.e., d, u, or t)
  - In all examined datasets, everything was basically connected to everything
- Definition: Two edges are 2-hyperincident if they share at least two nodes
- 2-hyperincident connected components:
  Components of edges between paths of 2-hyperincident edges exist

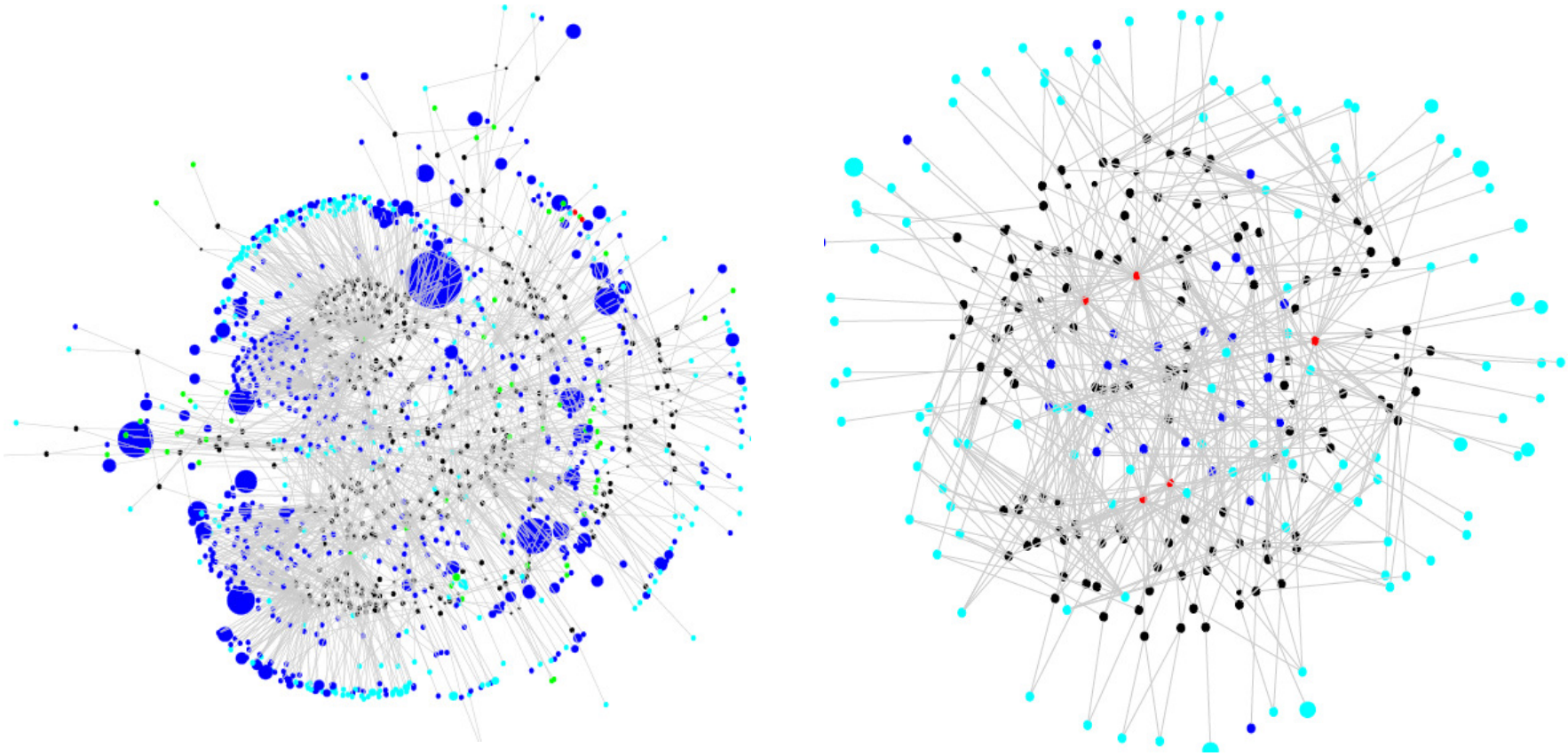


Blue, dotted lines indicate incident edges



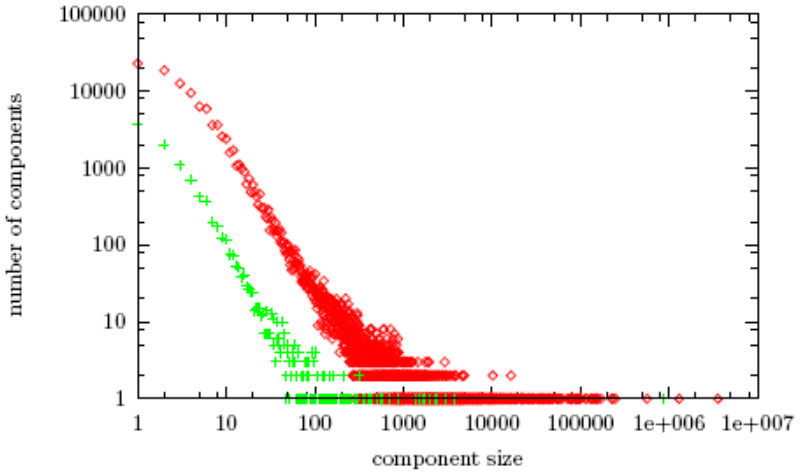
Blue, dotted lines indicate 2-hyperincident edges

Expanding 2-hyperincident edges around a user

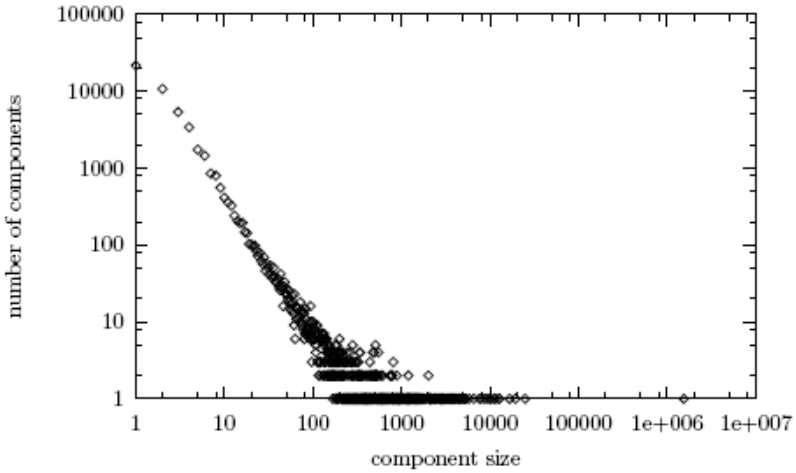


- Starting from a legitimate user, we had to stop at a limit of discovered nodes (here: 2000)
- Starting from spam users, we often found closed sets of connected nodes
- We did not find such components for legitimate users

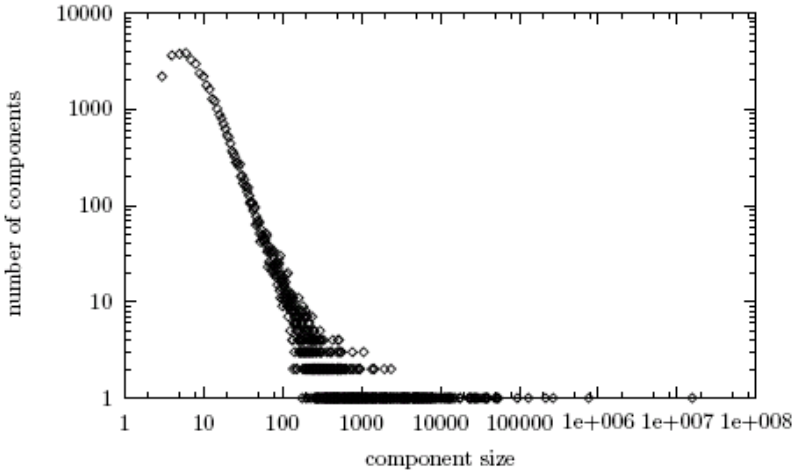
Distribution of Component Sizes



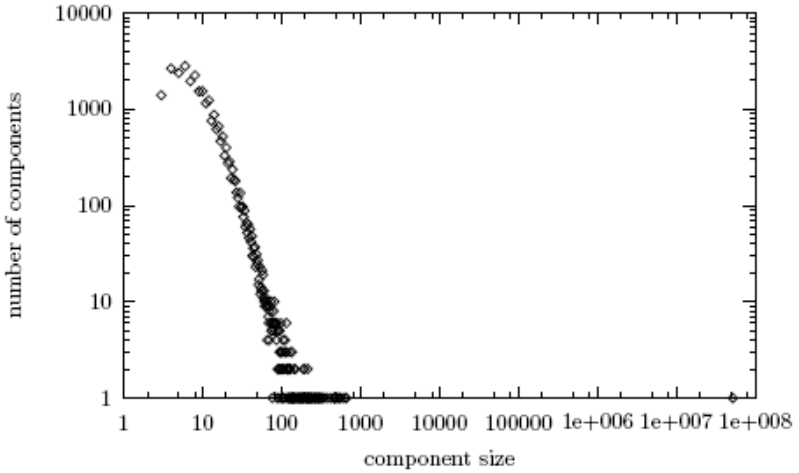
(a) Bibsonomy (green: without spam)



(b) CiteULike



(c) Delicious

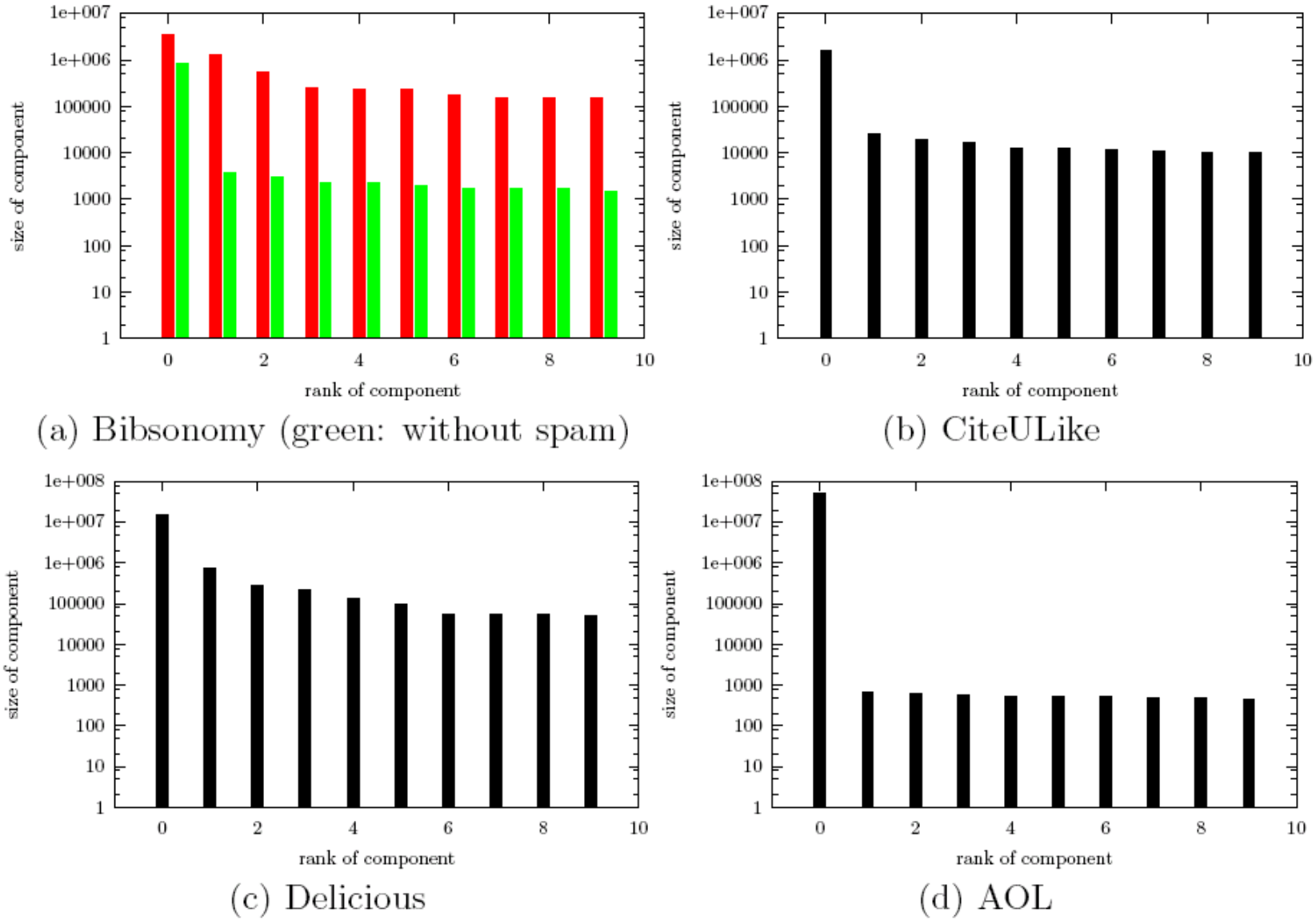


(d) AOL

x=number of components of size y (log/log)

Neubauer&Obermayer: Hyperincident Connected Components of Tagging Networks, HyperText 2009, in press

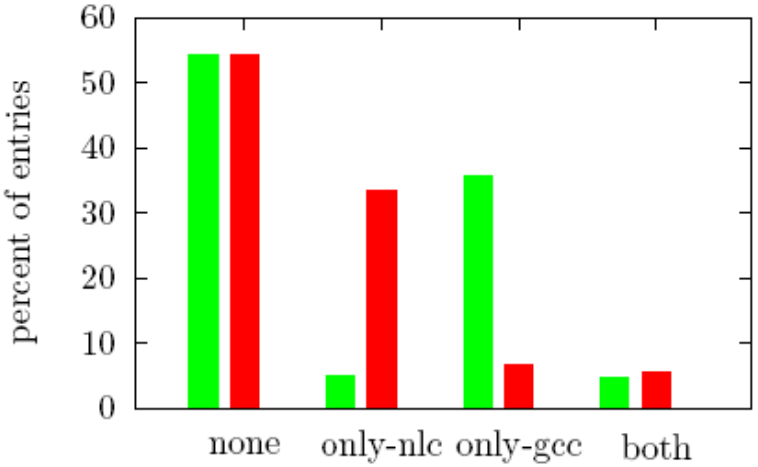
Distribution of Large Components‘ Sizes



x = rank of component, y = number of edges in component

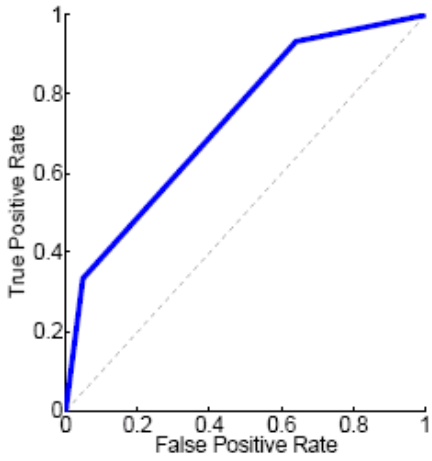
Spam Detection

- Users in nlc/gcc are likely to be spammers/non-spammers
- Are spammers/non-spammers also likely to live in nlc/gcc?
- Yes
  - although many users from both classes do neither.



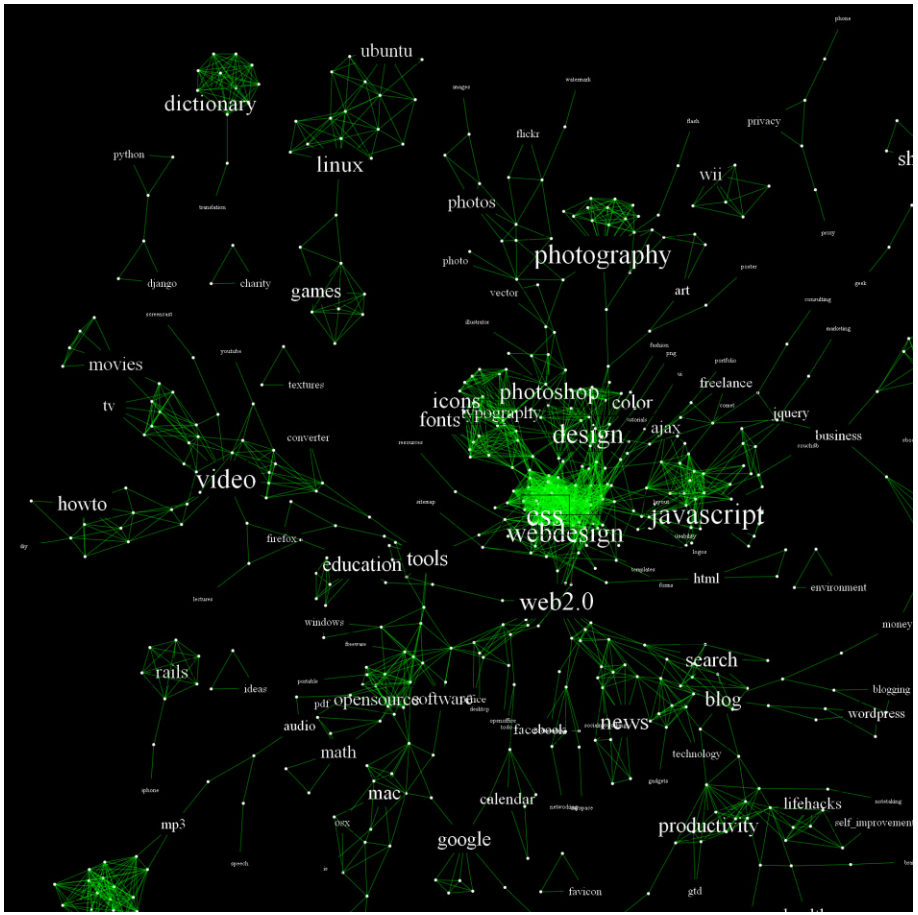
Distribution of users over components

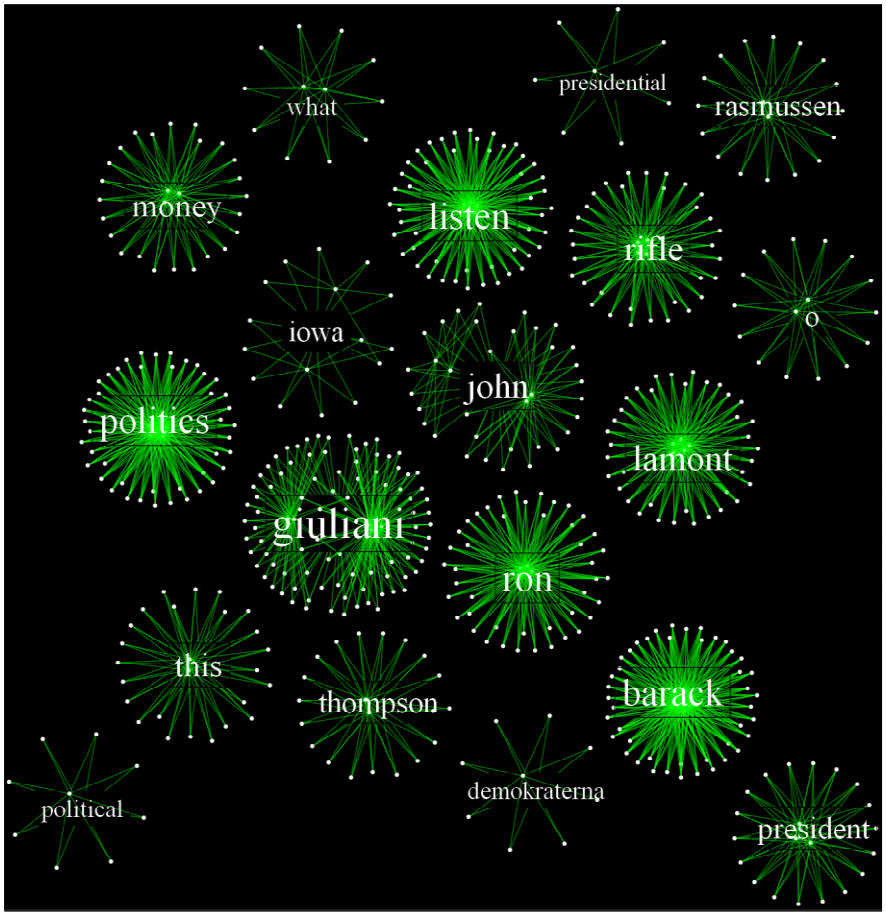
- Simple classification heuristic:
  - If user is only in nlc-> spam = 1
  - If user is only in gcc -> spam = 0
  - otherwise-> spam = 0.5
  - Note that users can be in more than one component
- Area under ROC curve (AUC - balanced accuracy) of .73



ROC curve of simple classifier

Largest and Next-largest 2-HCC for one Month of Delicious Tags





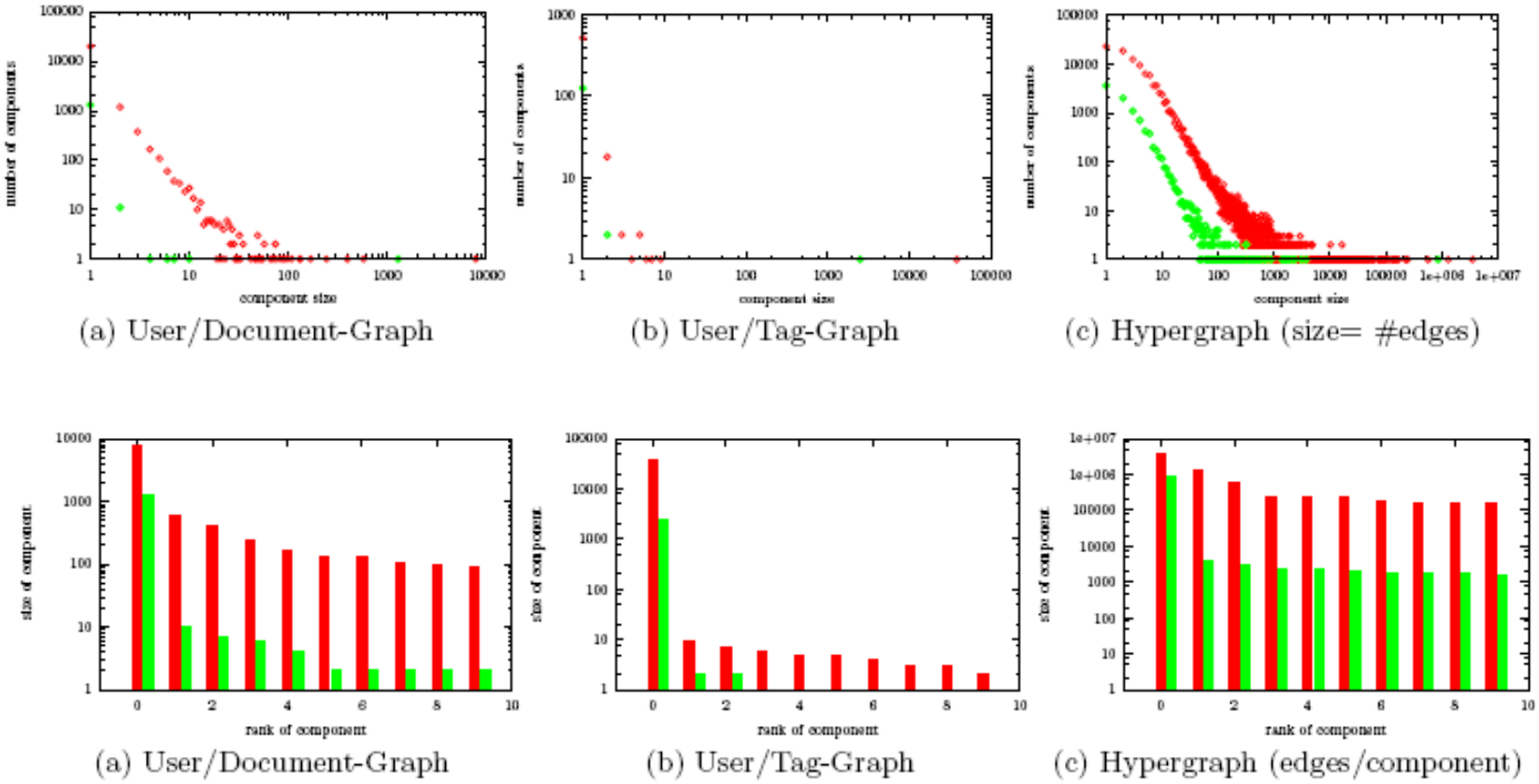
Overview

1. Spam in Social Bookmarking Systems
2. Hyperincident Connected Components
3. Document/User and Tag/User Graphs
4. Conclusions

Doubting Hyper-Incident Connectivity

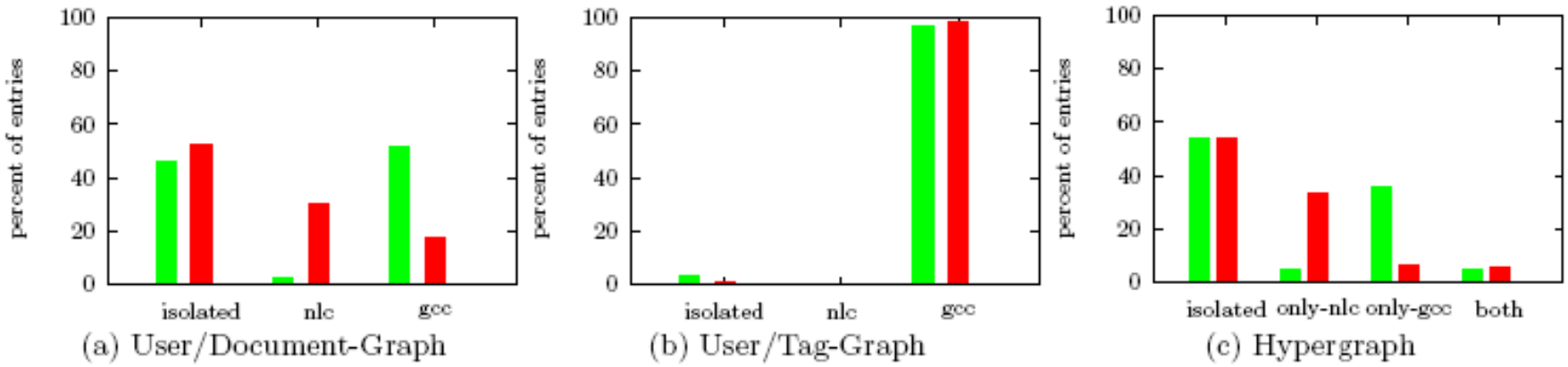
- “Nice result, but probably mostly based on documents”
- Short story: Right.
  - Long story: Tags do have a bit of influence here.
- Question: What happens if we examine connectivity on the document/user graph, ie edges=(d,u) for (d,u,t) in hypergraphs?
  - And what happens if we do the same for the tag/user graph?

Connectivity Structure (Bibsonomy)



- We see a the distribution of component sizes in the user/document graph closely resembles the one found in the entire hypergraph
- The tag/document graph is mostly connected

User Distribution

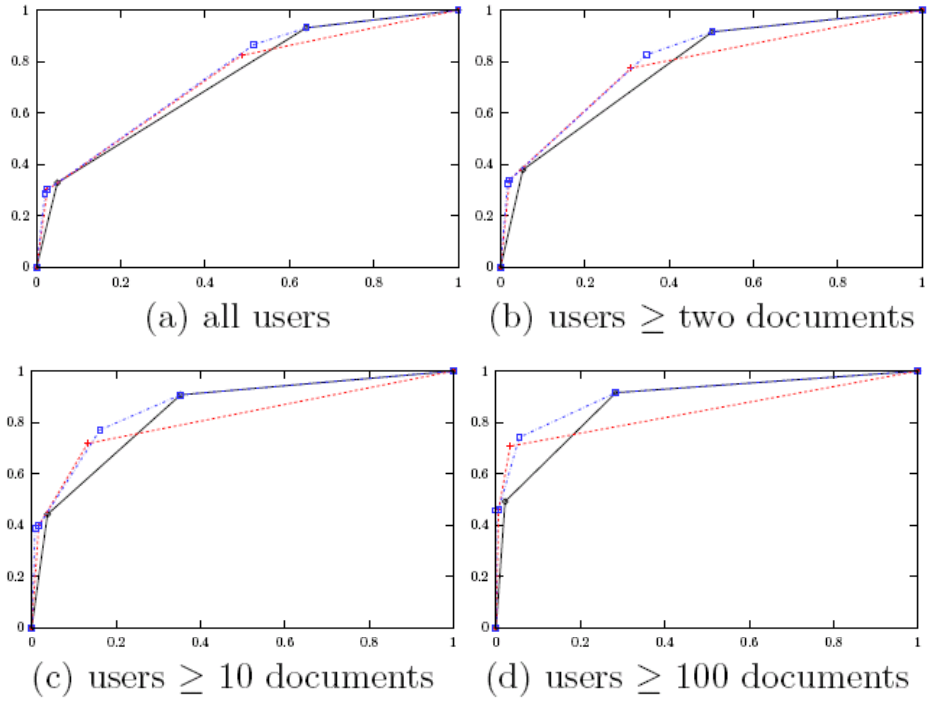


- Accordingly, membership information on the user/document graph is comparably informative, while the tag/document graph is useless

Spam Detection

New spam detection experiments:

- applied above heuristic on document/user graph (red)
- compared to original approach (black)
- new heuristic (blue):
  new maximum spam score for users being in nlc in both graphs
- also examined effect of #documents/user



ROC curves for all three heuristics

Results:

- Hypergraph and document/user graph connectivity provide similar, but sometimes complementary information
- Entire approach works better when users have more documents

	min # docs/user			
	0	1	10	100
User/Document	0.73	0.78	0.81	0.84
User/Tag	0.49	0.49	0.50	0.50
Hypergraph	0.73	0.78	0.84	0.88
Combined	0.76	0.81	0.87	0.91

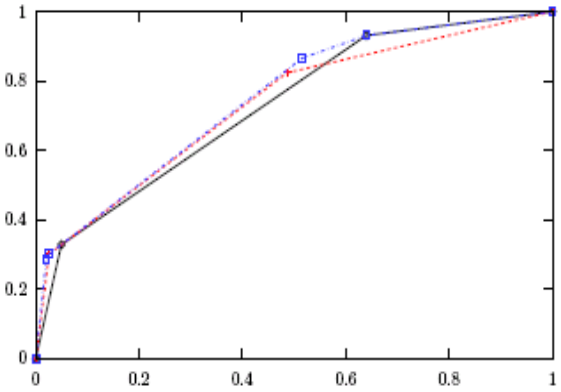
AUC values

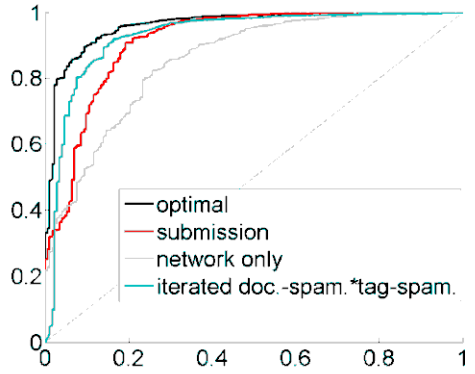
Overview

1. Spam in Social Bookmarking Systems
2. Hyperincident Connected Components
3. Document/User and Tag/User Graphs
4. Conclusions

Final Results & Discussion

	Requirements	
	Feature extraction on resources or references	Previous Labels
Content analysis	X	X
Reference analysis	X	X
User Similarity		X
Structural Analysis		





- Accuracy decreases, but so do domain dependence and requirements on available information
- Addition to other, more specialized approaches
- Stand-alone baseline when more specialized approaches are not available
- Although a large part of connectivity is related to documents, tags do play a subtle role.
- Next : Exploring temporal evolution & even stricter notions of connectivity