UNIVERSITY OF ALBERTA

# Building, Exploring and Querying Social Networks

Denilson Barbosa

denilson@cs.ualberta.ca

CWRCshop April 2010

# Goals

- Social network analysis
- **Ego-centric**: individuals and how they stand in the network
- **Socio-centric**: the network as a whole
- **Dynamic** : observe the network over time, and fit models to describe the changes

©2005 COVERLY 4-12 SPEEDBUMP.com
...AND THAT HOUND WENT TO OBEDIENCE SCHOOL WITH A SPANIEL WHO STAYED AT A KENNEL WITH A COLLIE WHO LIVES NEXT DOOR TO KEVIN BACON'S MUTT!...
SIX PEDIGREES OF SEPARATION

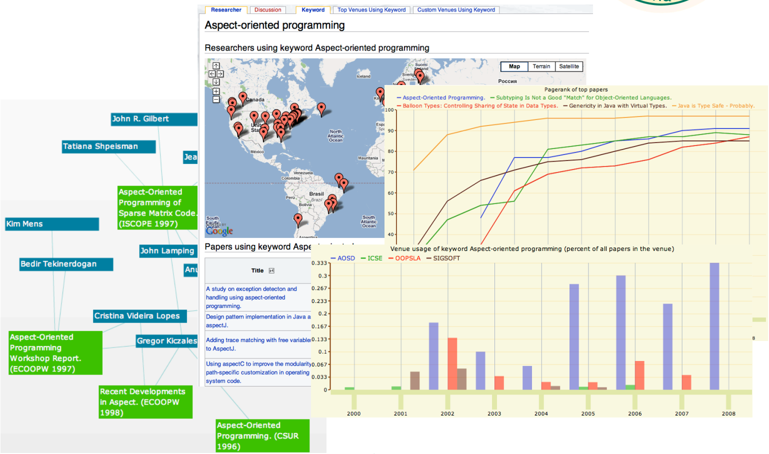
Copyright ©2005 Creators Syndicate, Inc.

# What CS brings to to social network analysis

- **Systematic collection, extraction, integration and analysis** of networks (nodes/edges)

- **Visualization, querying** and **exploration** of social networks

- **Scalability**
  - Many success stories for **static network analysis** (e.g., PageRank)



Researcher | Discussion | Keyword | Top Venues Using Keyword | Custom Venues Using Keyword

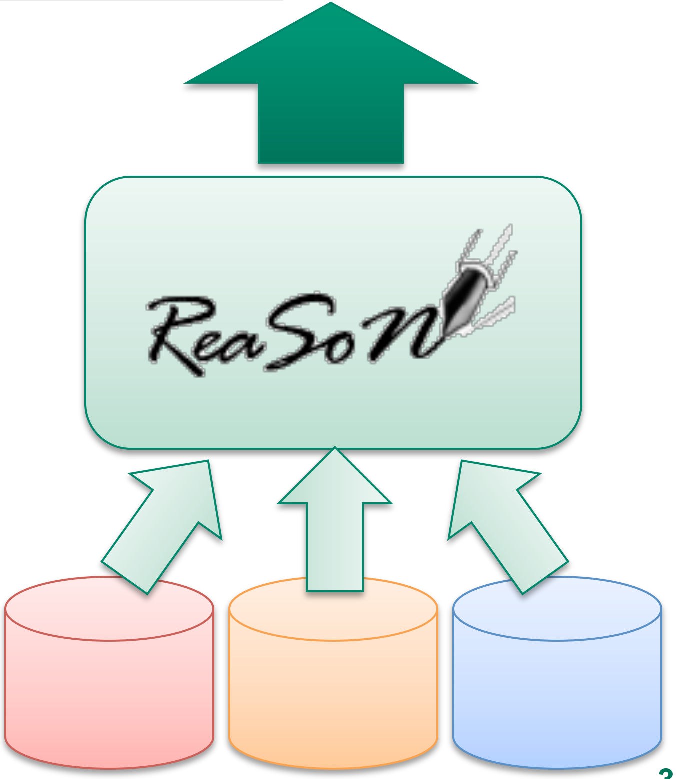
Aspect-oriented programming

Researchers using keyword Aspect-oriented programming

Papers using keyword Aspe

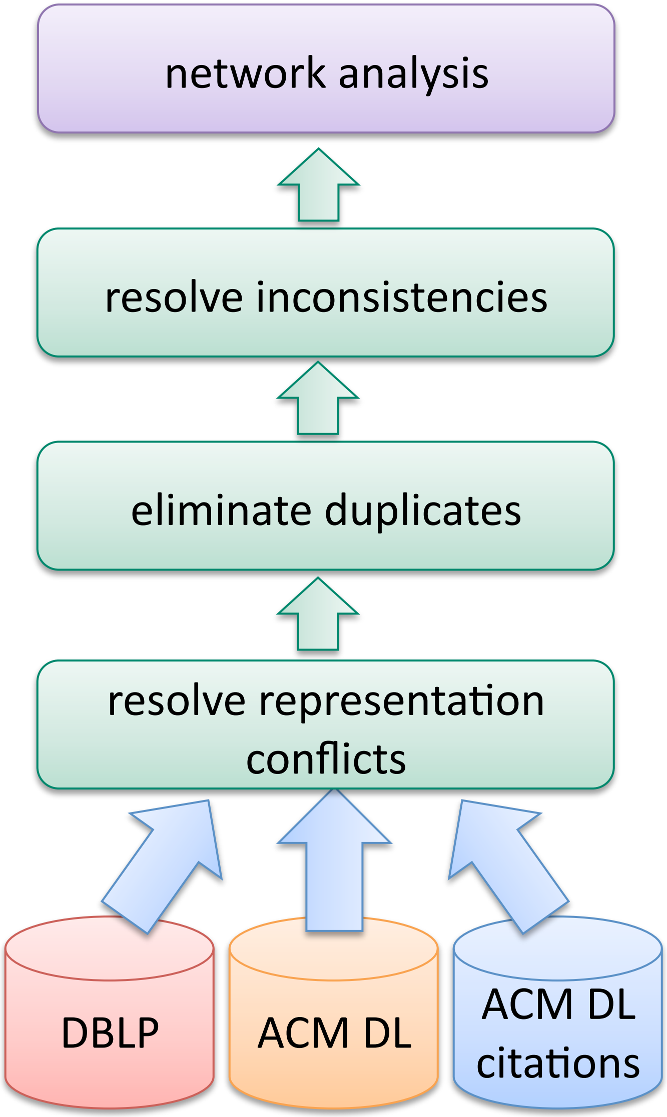
| Title |
|---|
| A study on exception detecton and handling using aspect-oriented programming. |
| Design pattern implementation in Java a aspectJ. |
| Adding trace matching with free variable to AspectJ. |
| Using aspectC to improve the modularity path-specific customization in operating system code. |

John R. Gilbert
Tatiana Shpeisman
Jea
Aspect-Oriented Programming of Sparse Matrix Code. (ISCOPE 1997)
Kim Mens
John Lamping
Bedir Tekinerdogan
Anu
Cristina Videira Lopes
Gregor Kiczales
Aspect-Oriented Programming Workshop Report. (ECOOPW 1997)
Recent Developments in Aspect. (ECOOPW 1998)
Aspect-Oriented Programming. (CSUR 1996)
Map
Terrain
Satellite
Canada
Iceland
Россия
North Atlantic Ocean
South Atlantic Ocean
South Pacific Ocean
México
Venezuela
Colombia
Perú
Brasil
Bolivia
Mauritania
Google
Pagerank of top papers
Aspect-Oriented Programming.
Subtyping Is Not a Good "Match" for Object-Oriented Languages.
Balloon Types: Controlling Sharing of State in Data Types.
Genericity in Java with Virtual Types.
Java is Type Safe - Probably.
100
90
80
70
60
50
40
Venue usage of keyword Aspect-oriented programming (percent of all papers in the venue)
AOSD
ICSE
OOPSLA
SIGSOFT
0.333
0.3
0.267
0.233
0.2
0.167
0.133
0.1
0.067
0.033
0
2000
2001
2002
2003
2004
2005
2006
2007
2008



# ReaSoN workflow

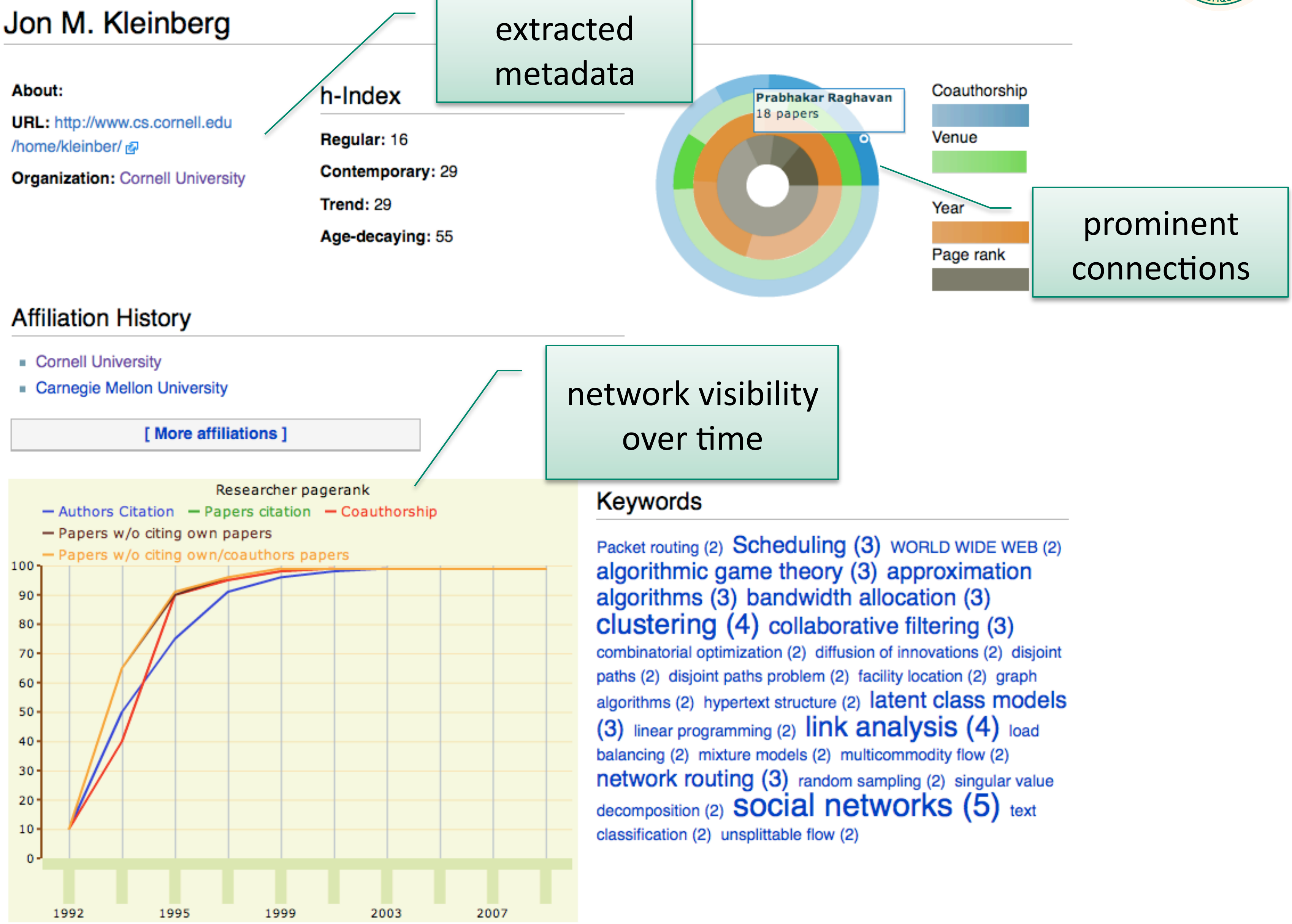
- ReaSoN networks:
  - Citations among publications (368K publications, 1.3M citations)
  - Citations among researchers (248K researchers, 8M citations)
  - Citations among venues (2,895 venues, 135K citations)
  - Joint co-authorship of publications (379K researchers, 2.2M citations)



# Built-in pages in ReaSoN

- **"Ego-centric"** visualization of actors
- ReaSoN visualizes several networks at once
  - Actors (e.g., researchers) participate in many networks (e.g., co-authorship, citation, etc.)
  - Time-versioned networks (e.g., citations by year)
- Every node in any social network has a default page
  - Shows the most prominent information about the node and links to the most prominent connections in the various networks
  - Builds on MediaWiki/Annoki extensions (wikimap, wiEGO, discussion page, versions, etc.)

# Researcher page



## Jon M. Kleinberg

**About:**

**URL:** http://www.cs.cornell.edu/home/kleinber/

**Organization:** Cornell University

### h-Index

**Regular:** 16

**Contemporary:** 29

**Trend:** 29

**Age-decaying:** 55

### Affiliation History

- Cornell University
- Carnegie Mellon University

[ More affiliations ]

### Keywords

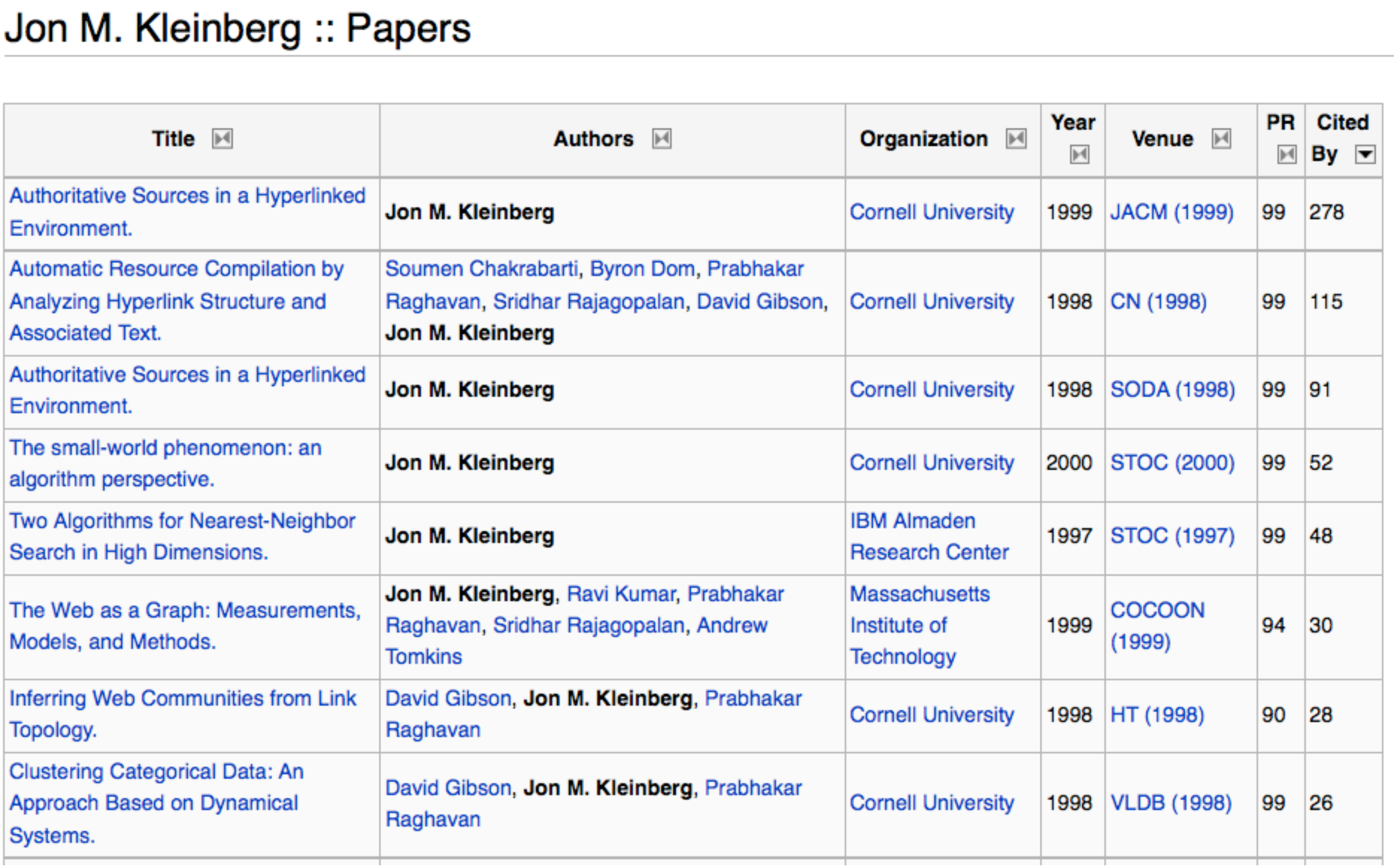
Packet routing (2) Scheduling (3) WORLD WIDE WEB (2) algorithmic game theory (3) approximation algorithms (3) bandwidth allocation (3) clustering (4) collaborative filtering (3) combinatorial optimization (2) diffusion of innovations (2) disjoint paths (2) disjoint paths problem (2) facility location (2) graph algorithms (2) hypertext structure (2) latent class models (3) linear programming (2) link analysis (4) load balancing (2) mixture models (2) multicommodity flow (2) network routing (3) random sampling (2) singular value decomposition (2) social networks (5) text classification (2) unsplittable flow (2)

extracted metadata

prominent connections

network visibility over time

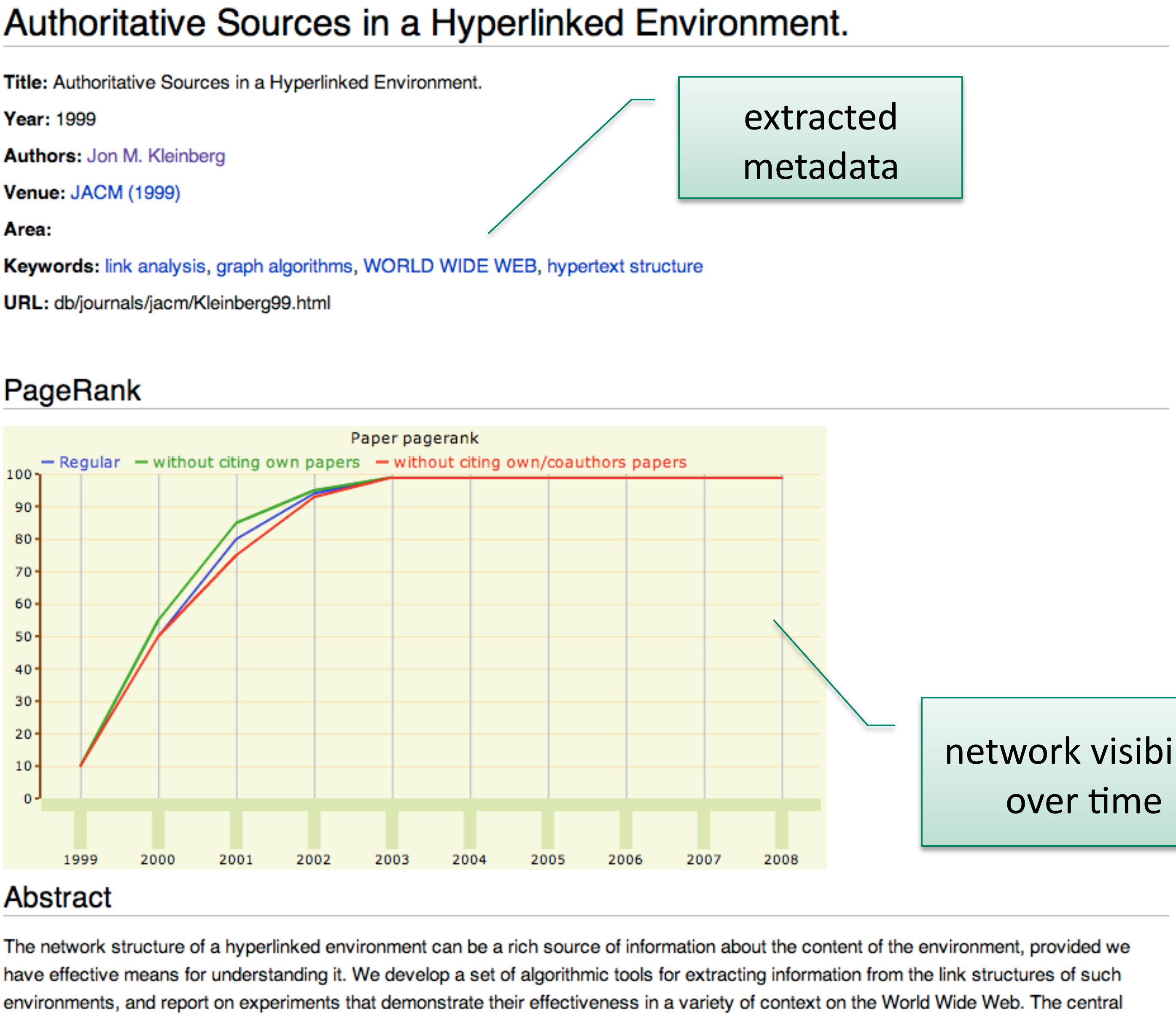
# Publications of a researcher



## Jon M. Kleinberg :: Papers

| Title ⧓ | Authors ⧓ | Organization ⧓ | Year ⧓ | Venue ⧓ | PR ⧓ | Cited By ▾ |
|---|---|---|---|---|---|---|
| Authoritative Sources in a Hyperlinked Environment. | **Jon M. Kleinberg** | Cornell University | 1999 | JACM (1999) | 99 | 278 |
| Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. | Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, **Jon M. Kleinberg** | Cornell University | 1998 | CN (1998) | 99 | 115 |
| Authoritative Sources in a Hyperlinked Environment. | **Jon M. Kleinberg** | Cornell University | 1998 | SODA (1998) | 99 | 91 |
| The small-world phenomenon: an algorithm perspective. | **Jon M. Kleinberg** | Cornell University | 2000 | STOC (2000) | 99 | 52 |
| Two Algorithms for Nearest-Neighbor Search in High Dimensions. | **Jon M. Kleinberg** | IBM Almaden Research Center | 1997 | STOC (1997) | 99 | 48 |
| The Web as a Graph: Measurements, Models, and Methods. | **Jon M. Kleinberg**, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins | Massachusetts Institute of Technology | 1999 | COCOON (1999) | 94 | 30 |
| Inferring Web Communities from Link Topology. | David Gibson, **Jon M. Kleinberg**, Prabhakar Raghavan | Cornell University | 1998 | HT (1998) | 90 | 28 |
| Clustering Categorical Data: An Approach Based on Dynamical Systems. | David Gibson, **Jon M. Kleinberg**, Prabhakar Raghavan | Cornell University | 1998 | VLDB (1998) | 99 | 26 |

# Publication page



## Authoritative Sources in a Hyperlinked Environment.

**Title:** Authoritative Sources in a Hyperlinked Environment.

**Year:** 1999

**Authors:** Jon M. Kleinberg

**Venue:** JACM (1999)

**Area:**

**Keywords:** link analysis, graph algorithms, WORLD WIDE WEB, hypertext structure

**URL:** db/journals/jacm/Kleinberg99.html

## PageRank

## Abstract

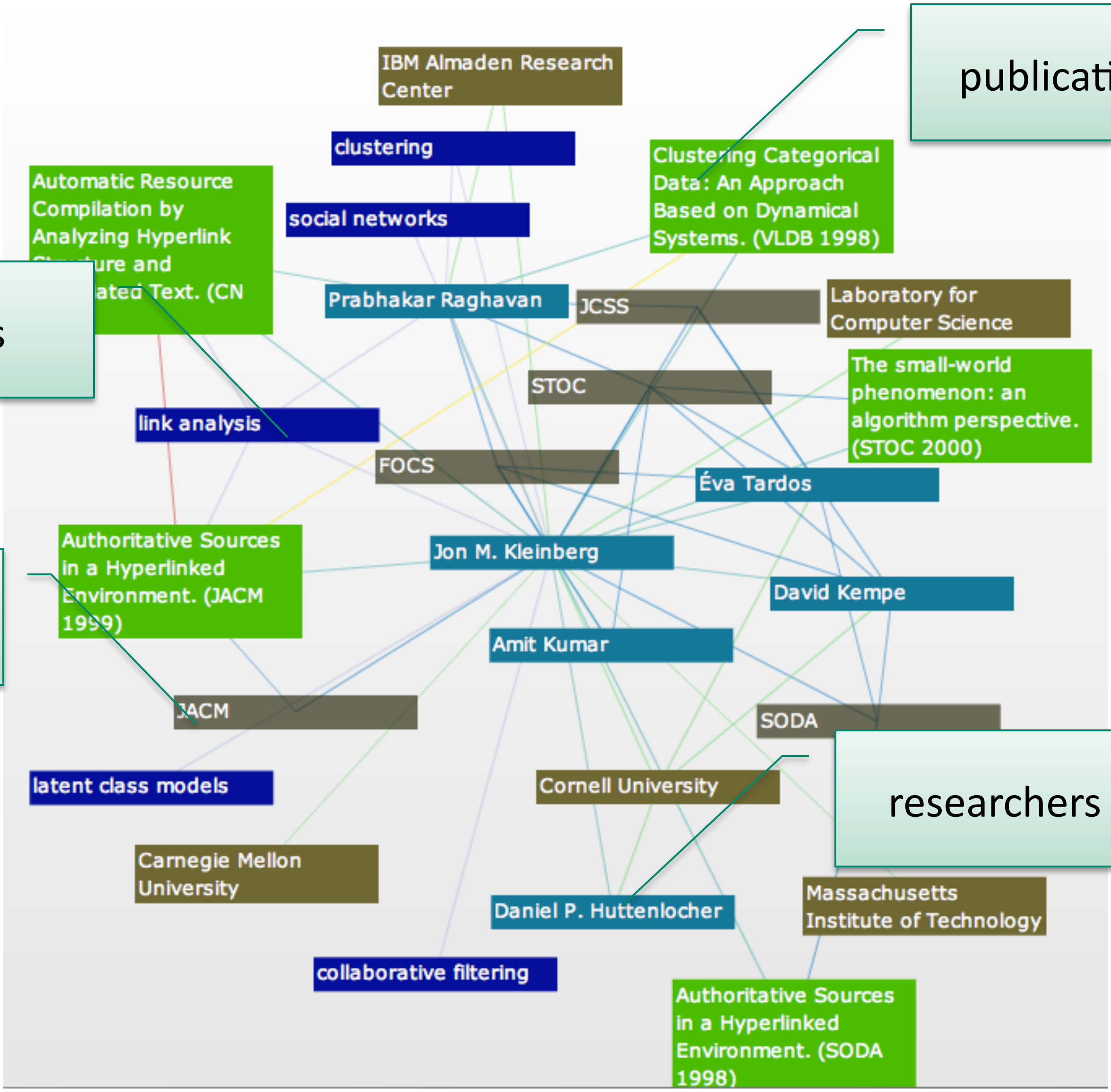
The network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. We develop a set of algorithmic tools for extracting information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of context on the World Wide Web. The central

extracted metadata

network visibility over time

# Interactive wikiMap visualization

- Available from most actor pages


IBM Almaden Research Center
clustering
social networks
Automatic Resource Compilation by Analyzing Hyperlink Structure and
ated Text. (CN
Clustering Categorical Data: An Approach Based on Dynamical Systems. (VLDB 1998)
Prabhakar Raghavan
JCSS
Laboratory for Computer Science
STOC
The small-world phenomenon: an algorithm perspective. (STOC 2000)
link analysis
FOCS
Éva Tardos
Authoritative Sources in a Hyperlinked Environment. (JACM 1999)
Jon M. Kleinberg
David Kempe
Amit Kumar
JACM
SODA
latent class models
Cornell University
Carnegie Mellon University
Daniel P. Huttenlocher
Massachusetts Institute of Technology
collaborative filtering
Authoritative Sources in a Hyperlinked Environment. (SODA 1998)

publications

topics

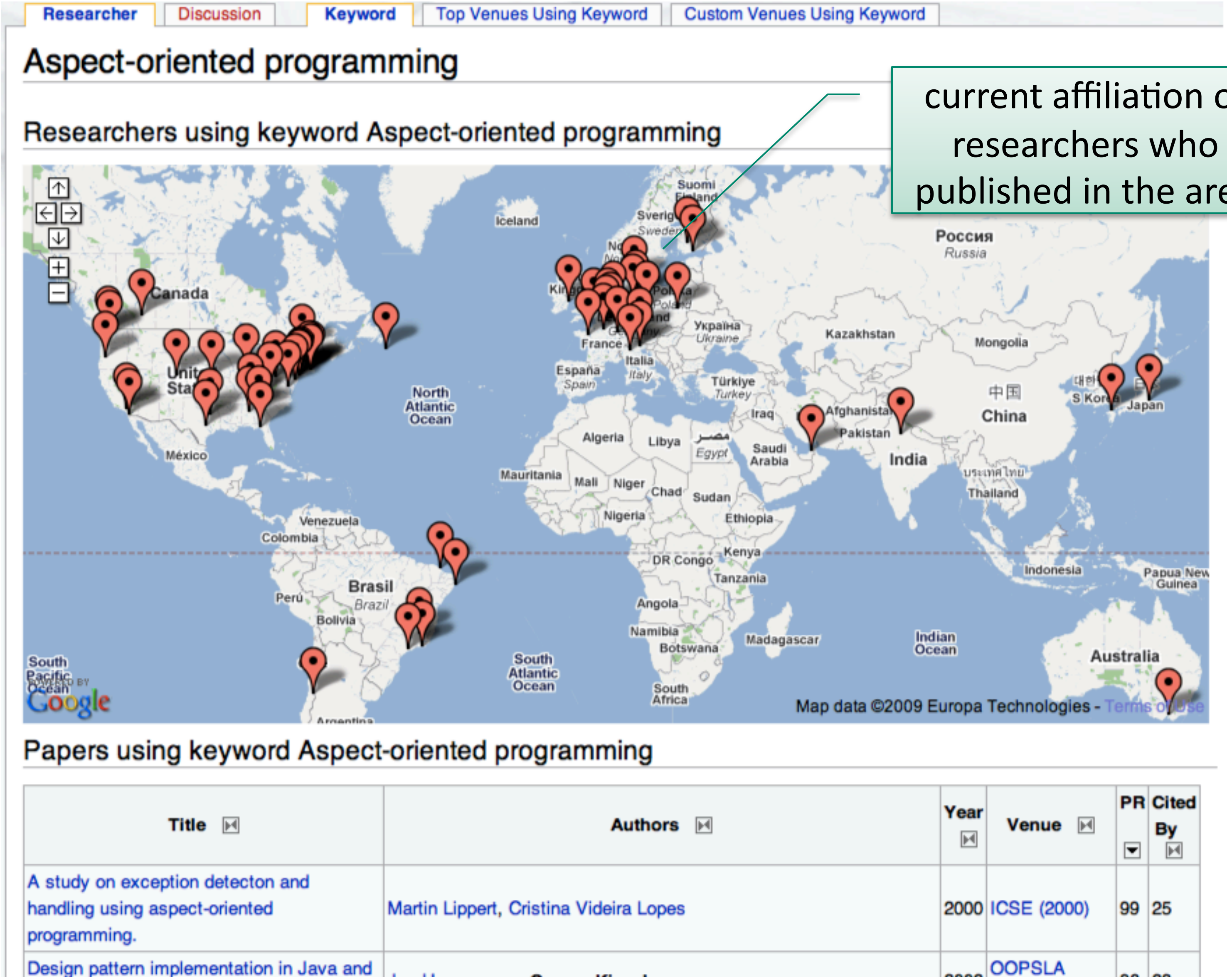
venues

researchers

# Dynamic pages in ReaSoN

- Pre-defined **exploration interfaces**
  - Geo-referenced visualization of paper by topic (keywords)
  - Visibility analysis over time for publications, venues, and organizations (author affiliations)
- Results of **user-specified queries…**

# Exploring topics/keywords



## Aspect-oriented programming

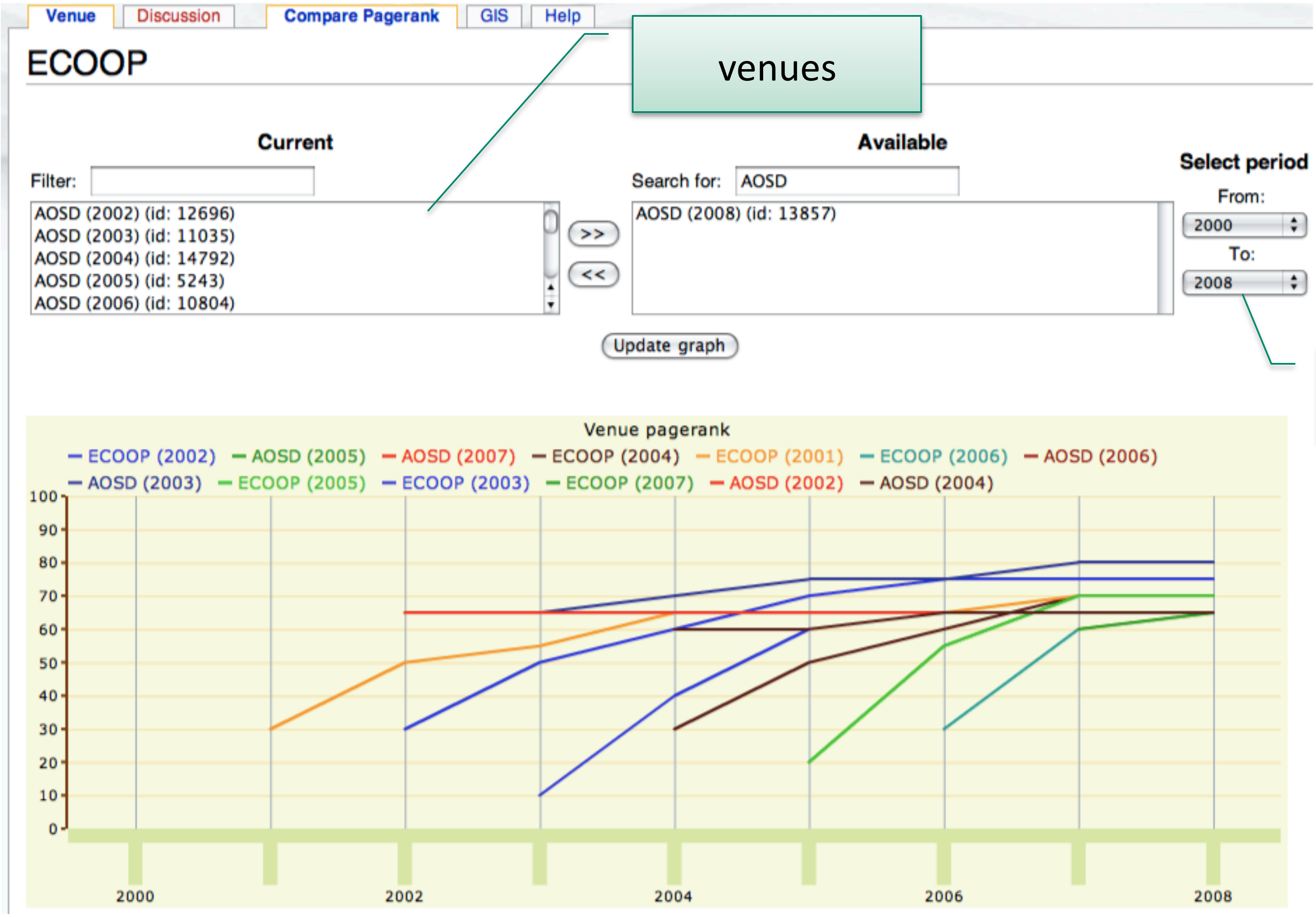
### Researchers using keyword Aspect-oriented programming

### Papers using keyword Aspect-oriented programming

| Title | Authors | Year | Venue | PR | Cited By |
|---|---|---|---|---|---|
| A study on exception detecton and handling using aspect-oriented programming. | Martin Lippert, Cristina Videira Lopes | 2000 | ICSE (2000) | 99 | 25 |
| Design pattern implementation in Java and | | | OOPSLA | | |

current affiliation of researchers who published in the area

# Visibility over time of multiple venues



ECOOP

Current

Filter:

AOSD (2002) (id: 12696)
AOSD (2003) (id: 11035)
AOSD (2004) (id: 14792)
AOSD (2005) (id: 5243)
AOSD (2006) (id: 10804)

Available

Search for: AOSD

AOSD (2008) (id: 13857)

Select period

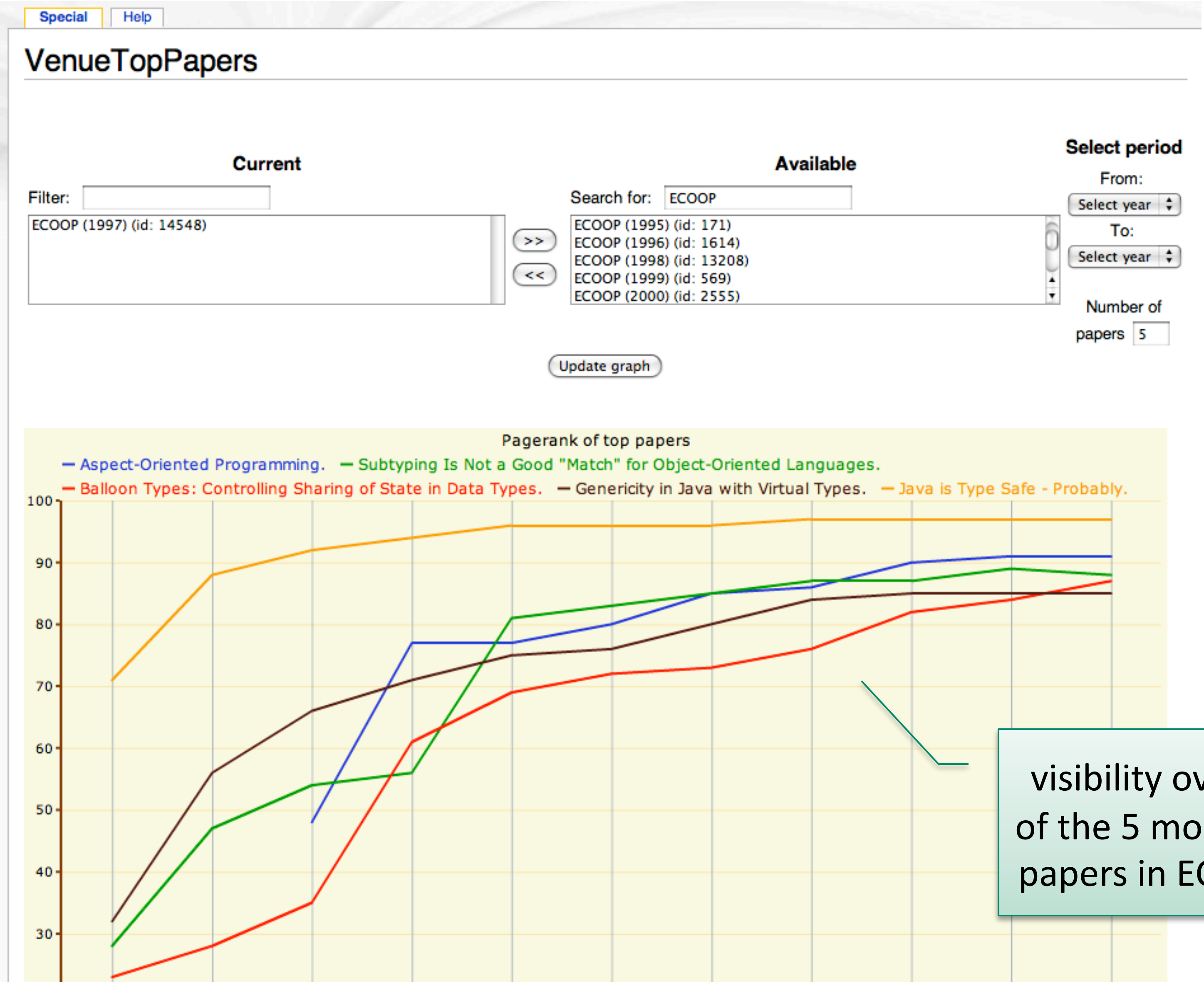
From: 2000

To: 2008

venues

range

# Most visible publications in venue



Special | Help

VenueTopPapers

**Current**

Filter:

ECOOP (1997) (id: 14548)

>> <<

**Available**

Search for: ECOOP

ECOOP (1995) (id: 171)
ECOOP (1996) (id: 1614)
ECOOP (1998) (id: 13208)
ECOOP (1999) (id: 569)
ECOOP (2000) (id: 2555)

**Select period**

From: Select year

To: Select year

Number of papers: 5

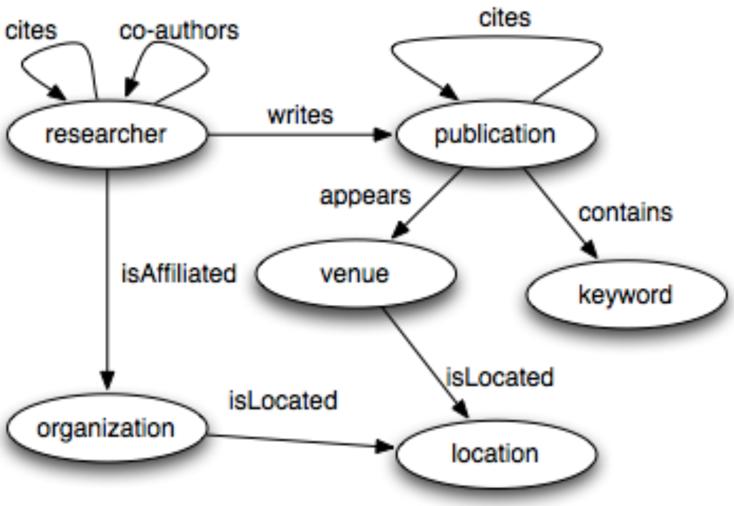
Update graph

Pagerank of top papers
Aspect-Oriented Programming.
Subtyping Is Not a Good "Match" for Object-Oriented Languages.
Balloon Types: Controlling Sharing of State in Data Types.
Genericity in Java with Virtual Types.
Java is Type Safe - Probably.
100
90
80
70
60
50
40
30

visibility over time of the 5 most visible papers in ECOOP 97

# Dynamic pages in ReaSoN

- Pre-defined **exploration interfaces…**
- Results of **user-specified queries**
  - Simple SQL-like query language
  - Data model:
    - Actors (relations)
    - Properties (attributes)
    - Connections (binary relations)
  - Selection criteria over properties
  - Subset of FO queries (CQs with inequalities)
  - Ranking based on visibility



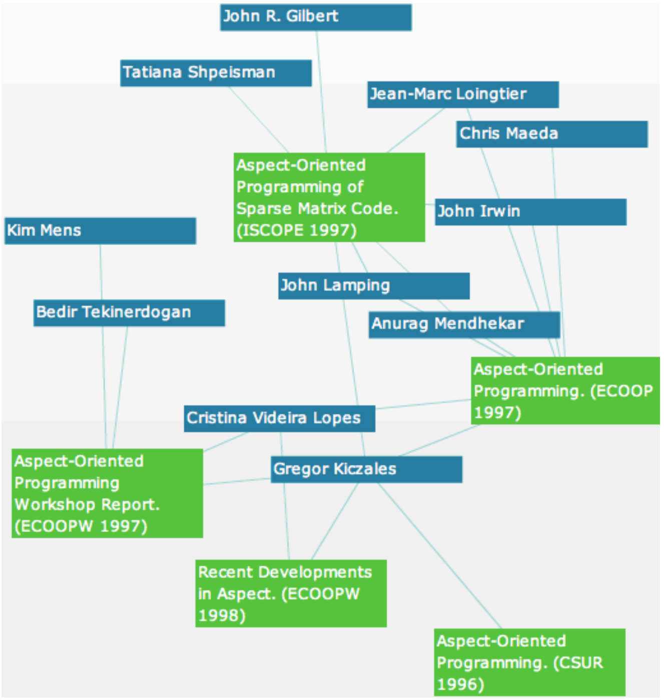
ReaSoN data model

# Queries in ReaSoN

$$\begin{array}{l} [\texttt{SELECT}|\texttt{MAP}|\texttt{EXPLORE}] \quad a_i.p_1, \ldots, a_j.p_n \\ \texttt{FROM} \quad \langle\text{Actor}\rangle\ a_1, \ldots \langle\text{Actor}\rangle\ a_k \\ \texttt{WHERE} \quad \langle\text{Predicate}\rangle\ [\texttt{AND}\ \langle\text{Predicate}\rangle]* \end{array}$$

- **SELECT** queries return tables
- **MAP** queries return Google maps with links to authors
- **EXPLORE** queries return wikiMaps

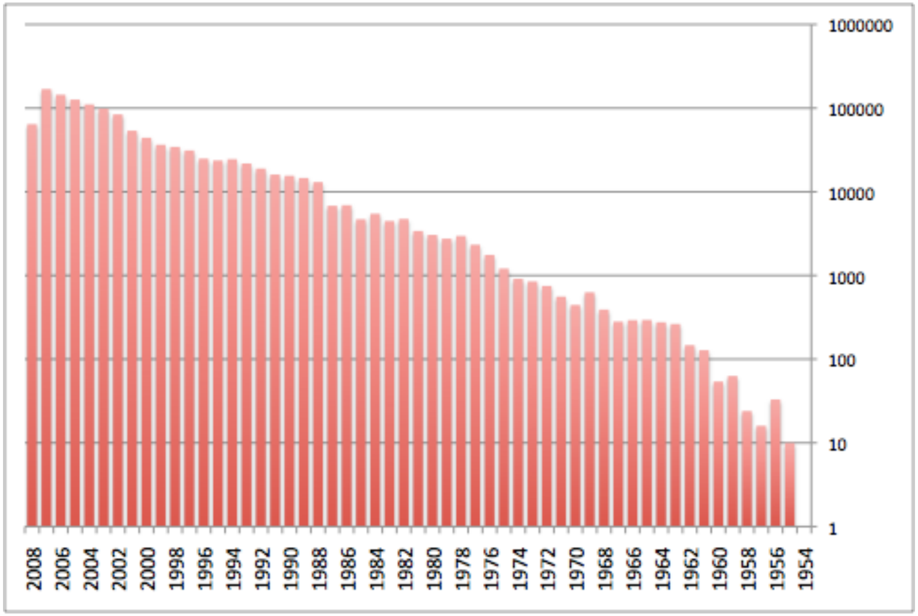
```
EXPLORE r1.name, p1.title, p1.venue
FROM publication p1, researcher r1,
     researcher r2
WHERE writes(r1,p1) AND
      writes(r2,p1) AND
      r2.name=’Gregor Kiczales’ AND
      p1.title><’Aspect’ AND
      p1.year<2000
```

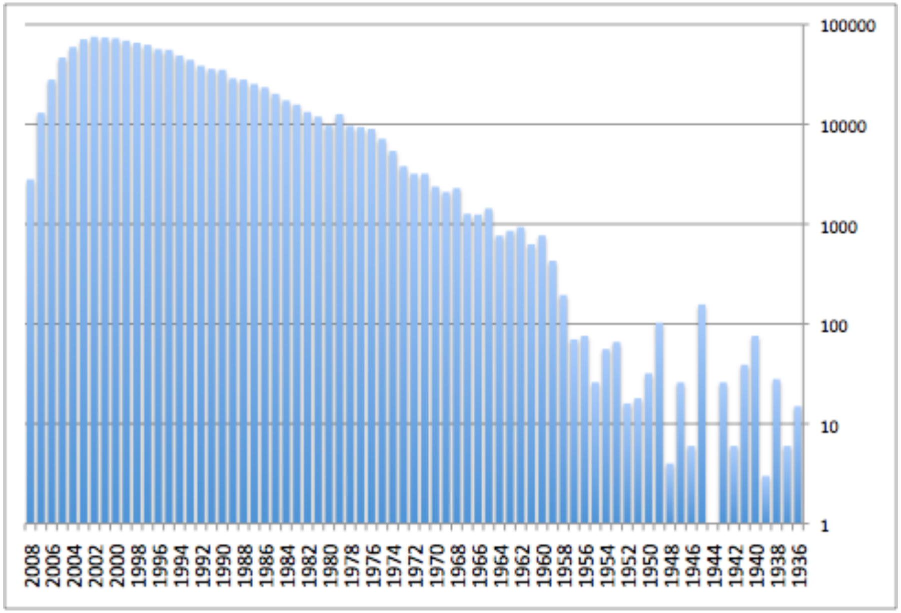

John R. Gilbert
Tatiana Shpeisman
Jean-Marc Loingtier
Chris Maeda
Aspect-Oriented Programming of Sparse Matrix Code. (ISCOPE 1997)
John Irwin
Kim Mens
John Lamping
Bedir Tekinerdogan
Anurag Mendhekar
Aspect-Oriented Programming. (ECOOP 1997)
Cristina Videira Lopes
Aspect-Oriented Programming Workshop Report. (ECOOPW 1997)
Gregor Kiczales
Recent Developments in Aspect. (ECOOPW 1998)
Aspect-Oriented Programming. (CSUR 1996)

# Preliminary socio-centric results

- Our first networks
  - Citations among publications, researchers, across venues
    - With and without nepotistic citations
  - Joint co-authorship of publications
- Some stats
  - 485,267 publications
  - 379,188 researchers
  - 1,301,365 citations
  - 3,793 venues (2,355 conferences and 623 journals)
  - 1,865 organizations (1,153 containing “Univ” in their name)

# CS people forget easily



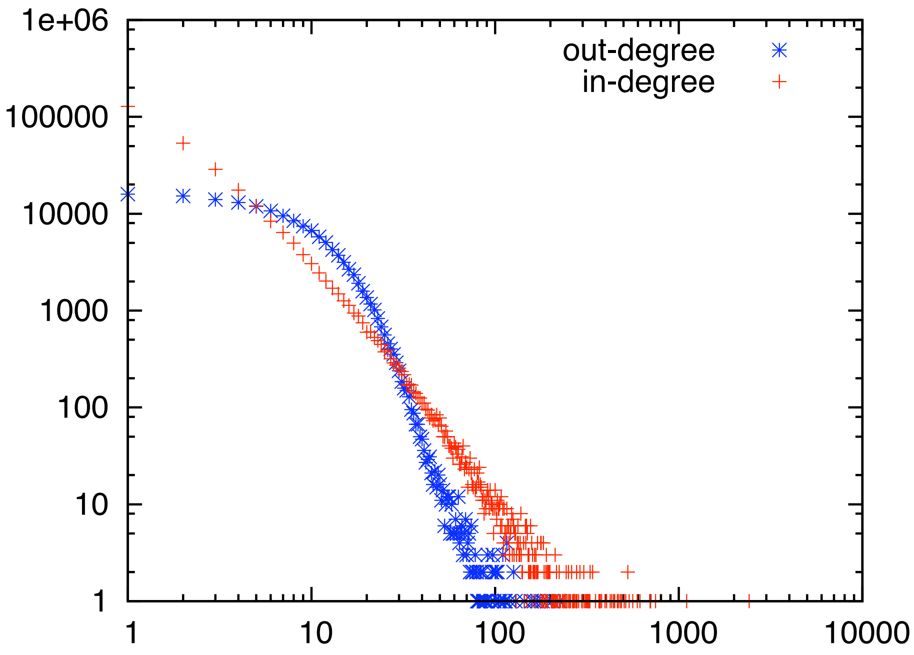


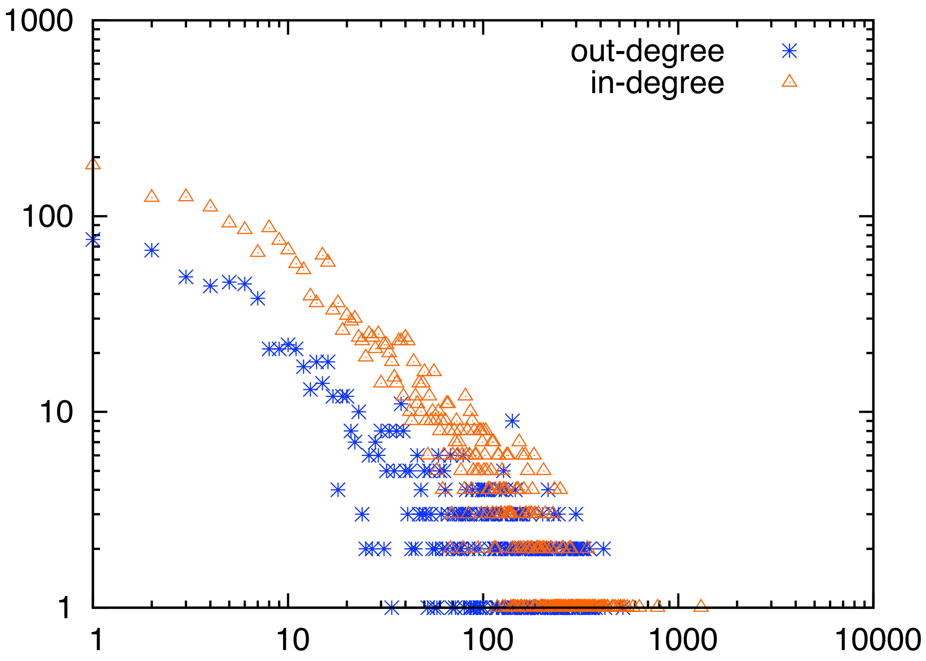
outgoing citations

incoming citations

- 3/4 of the citations refer to papers within 5 of the paper making the citation

# We live in a power-law world

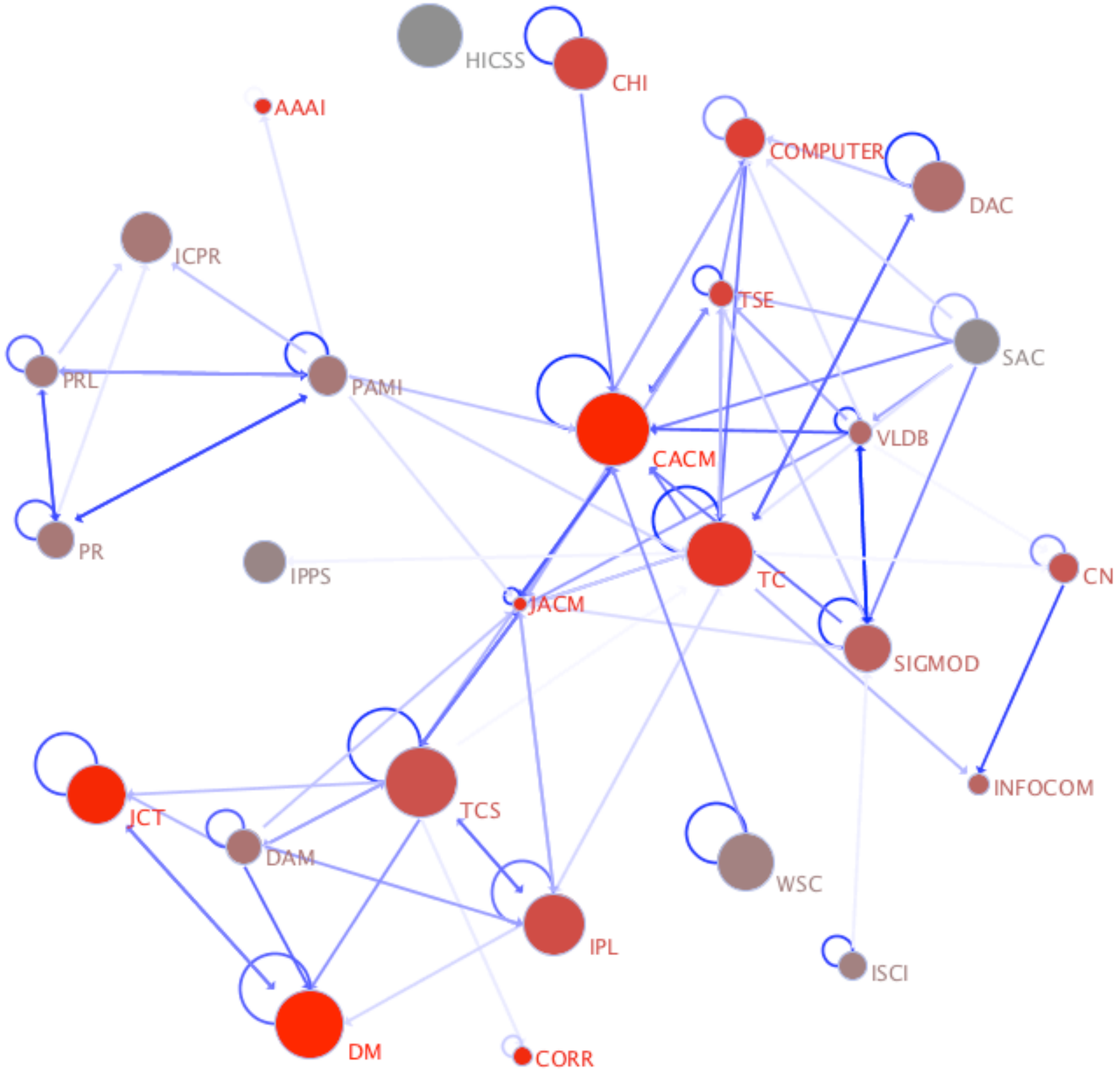




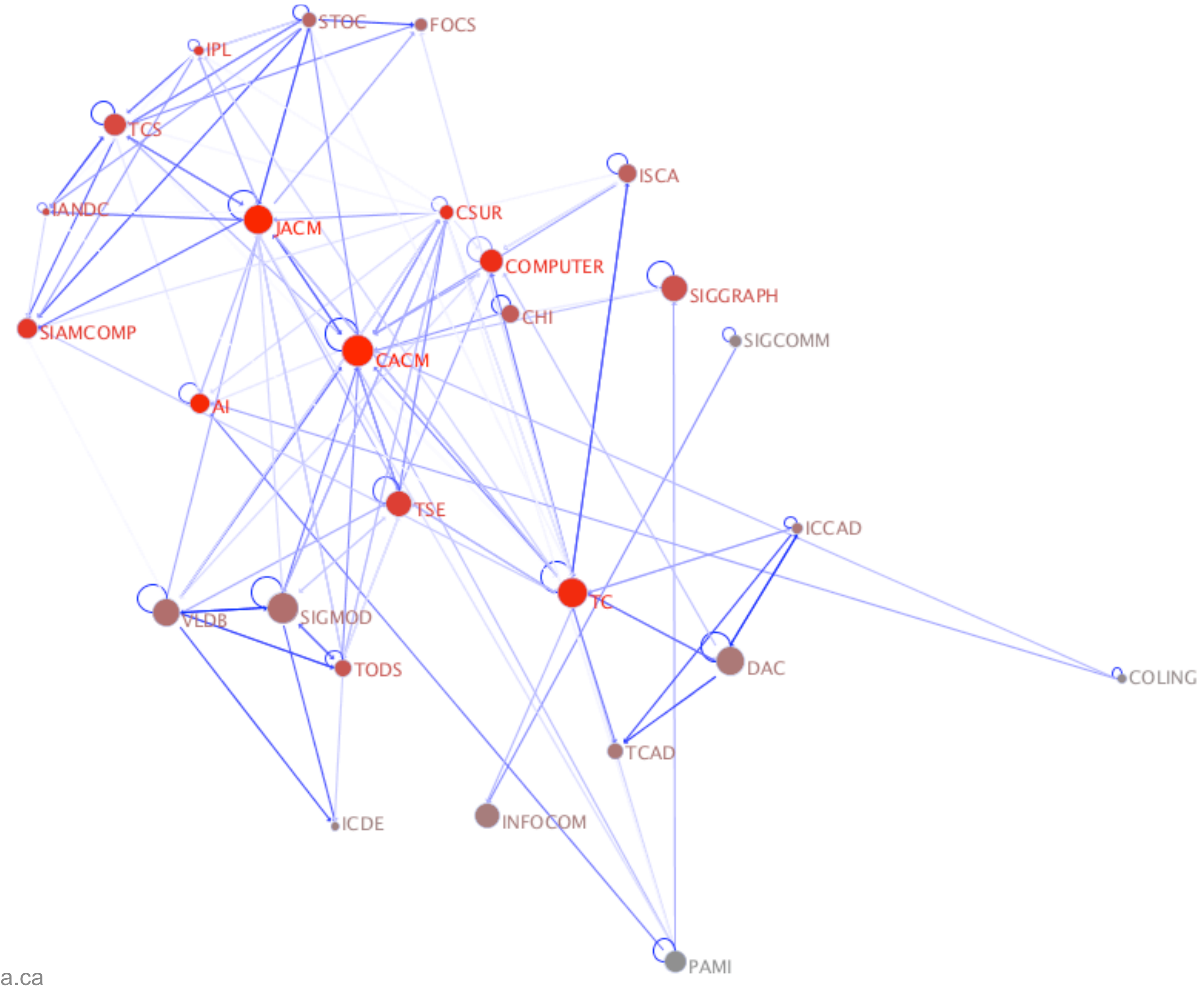
**Paper citation network**:
most paper cite < 20 other papers and are never cited

**Venue citation network**:
there are a few highly visible venues (e.g., top conferences) and many obscure one

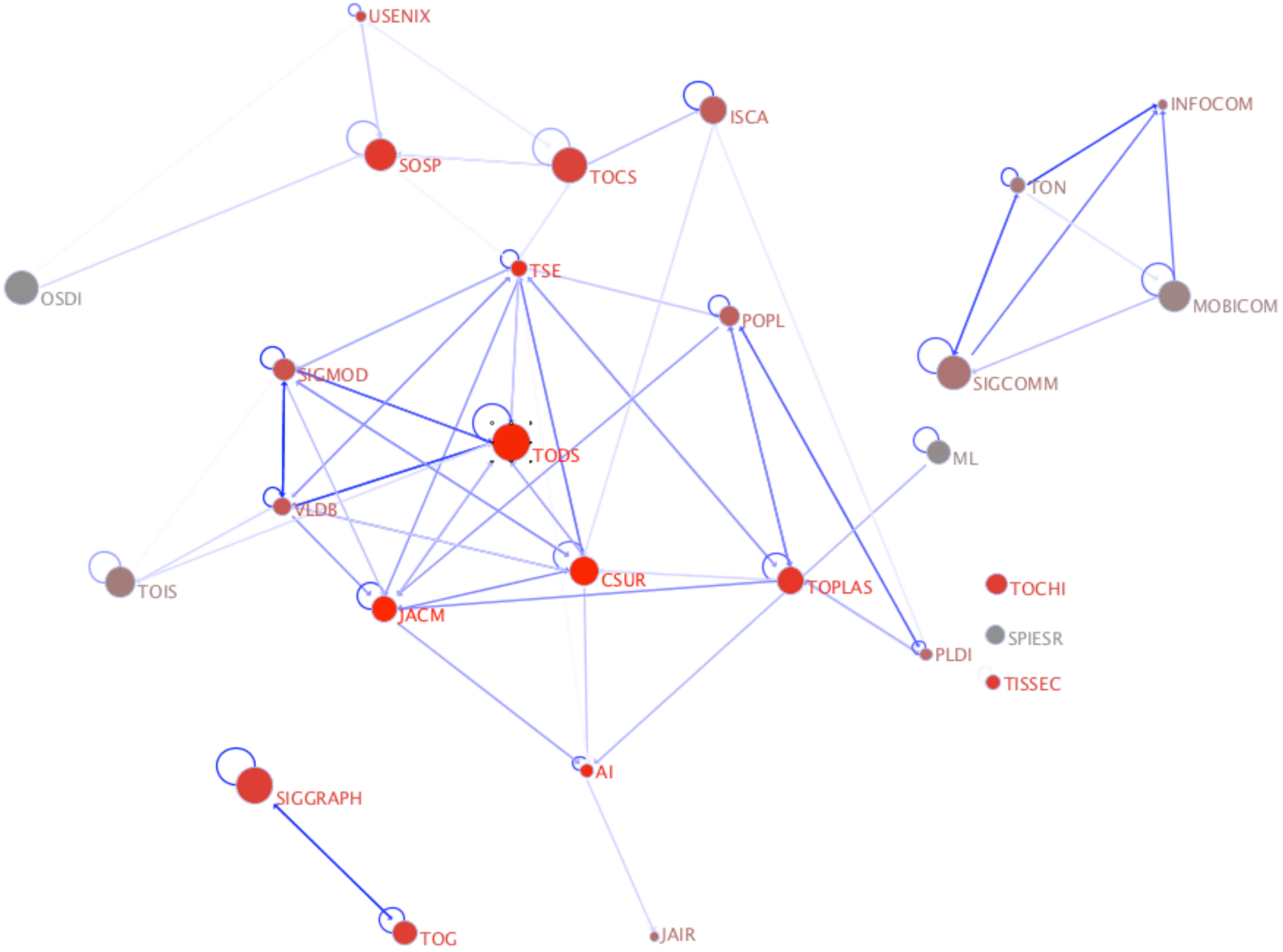
# The top 1% CS venues in SIZE



# The top 1% CS venues in CITATIONS



# The top 1% CS venues in IMPACT FACTOR


USENIX
ISCA
INFOCOM
SOSP
TOCS
TON
OSDI
TSE
MOBICOM
POPL
SIGMOD
SIGCOMM
TODS
ML
VLDB
TOIS
CSUR
TOPLAS
TOCHI
JACM
SPIESR
PLDI
TISSEC
AI
SIGGRAPH
TOG
JAIR