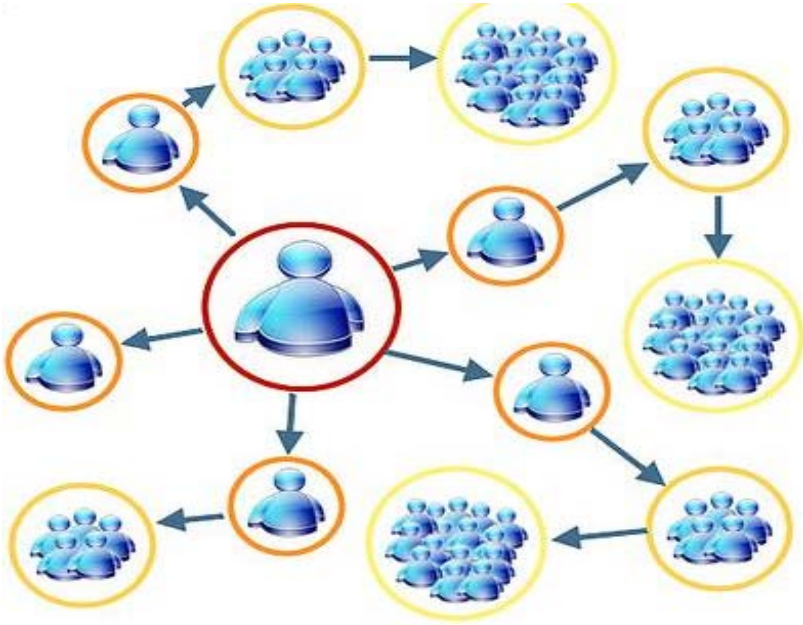
Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies

Iraklis Varlamis
Department of Informatics and Telematics,
Harokopio University of Athens, Greece
varlamis@hua.gr

George Tsatsaronis
Department of Computer and Information Science,
Norwegian University of Science &Technology, Norway
gbt@idi.ntnu.no

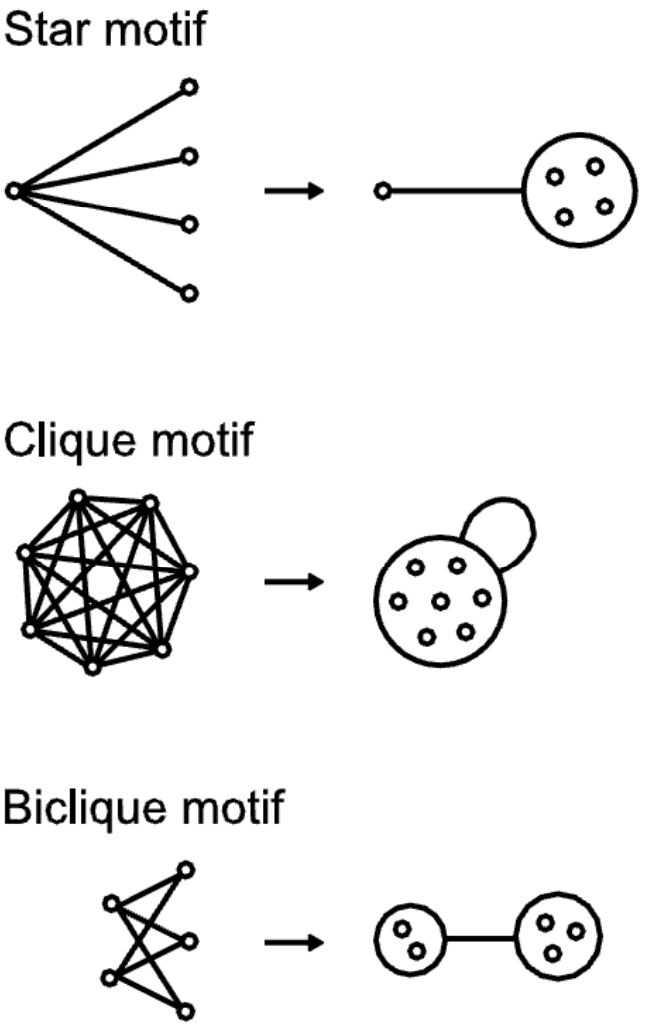


Contents

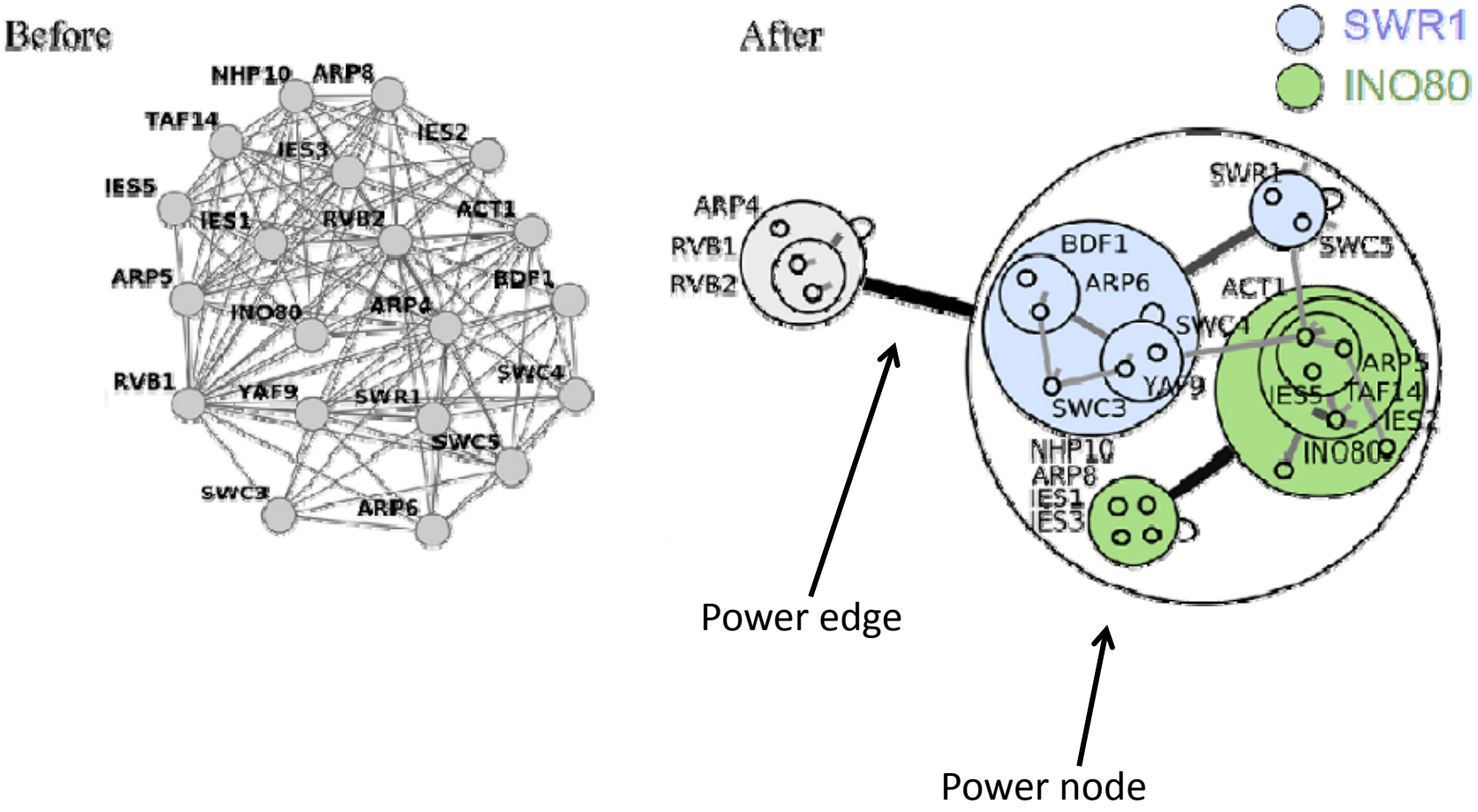
- Preliminaries
  - Power Graphs
  - Mapping Co-authorship graphs to Power Graphs
- Co-authorship patterns
- Experiments
- Conclusions

Power Graph Analysis

- A technique transferred from the biomedical domain
- Offers lossless representation of the original graph
- Offers graph size reduction especially for dense graphs
- Supported motifs: star, clique, bi-clique



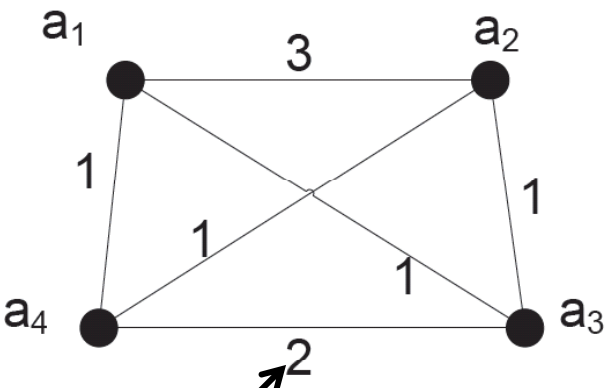
From Graphs to Power Graphs



Co-authorship graphs

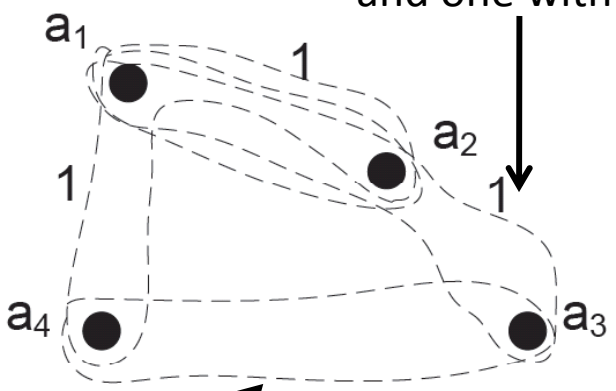
- We can either define hyper-edges or convert them to sets of simple edges

$paper-id$	$Authors$
p_1	a_1,a_2,a_3
p_2	a_1,a_2,a_4
p_3	a_1,a_2
p_4	a_3,a_4
p_5	a_3,a_4



Authors a_3 & a_4 have written two papers together

and one with a third co-author



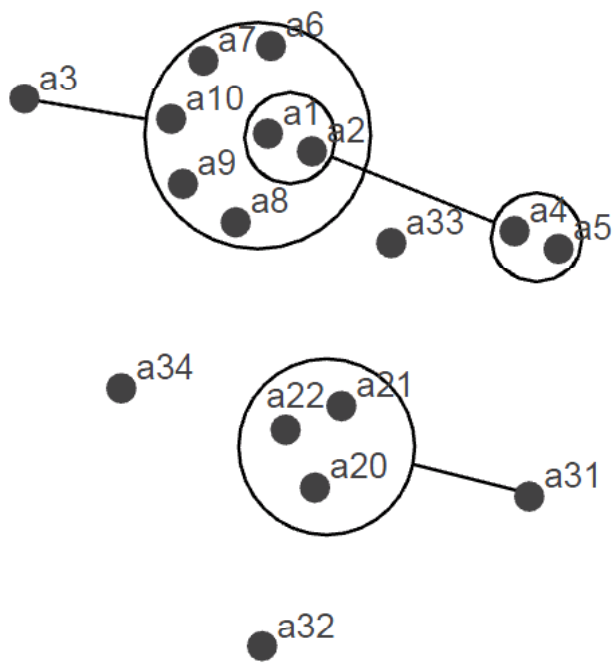
One as a pair

Co-authorship Power Graphs

- In analogy to protein networks, which contain proteins as nodes and edges that represent their interactions, the co-authorship graphs contain information about authors and their co-operations
- Finding reoccurring structural motifs in co-authorship graphs is a step towards discovering author communities
  - A power node contains authors that have published together
  - A star links a single author with all her co-authors
  - Bi-cliques link group of authors that have published together
  - Power node inclusion (nested power nodes) denotes subgroups of authors within a power node that collaborate more frequently

Example

Edges	
a1	a3
a1	a4
a1	a5
a2	a3
a2	a4
a2	a5
a3	a6
a3	a7
a3	a8
a3	a9
a3	a10
a20	a31
a20	a32
a21	a31
a21	a33
a22	a31
a22	a34



Biclique

- Authors a1 and a2 have exactly the same co-authors (a3, a4, a5) but have never cooperated
- The same holds for authors a3, a4 and a5 who have never worked together

Star

- all the co-authors of a31 form a star. Authors a20 to a22 are of potential interest for future cooperation

Inclusion

- author a3 has collaborated with a1, a2 and a6 to a10. Authors of the two nested power nodes should be checked for collaboration

Experimental Setup

- We employed the DBLP Computer Science Bibliography
- We focus on select subsets of publications (on specific domain areas) and

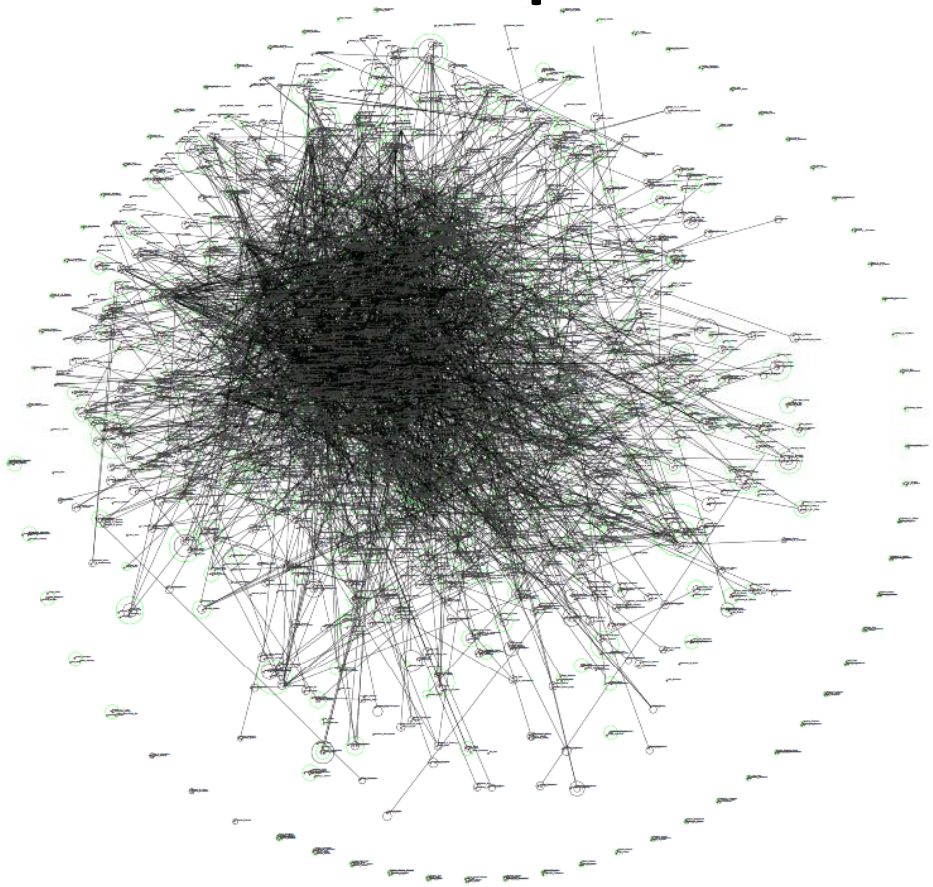
  (a) create the initial co-authorship graph from the selected subset of publications,

  (b) generate the Power Graph from the initial graph,

  (c) prune the weakest components of the Power Graph in order to improve the readability of the result,

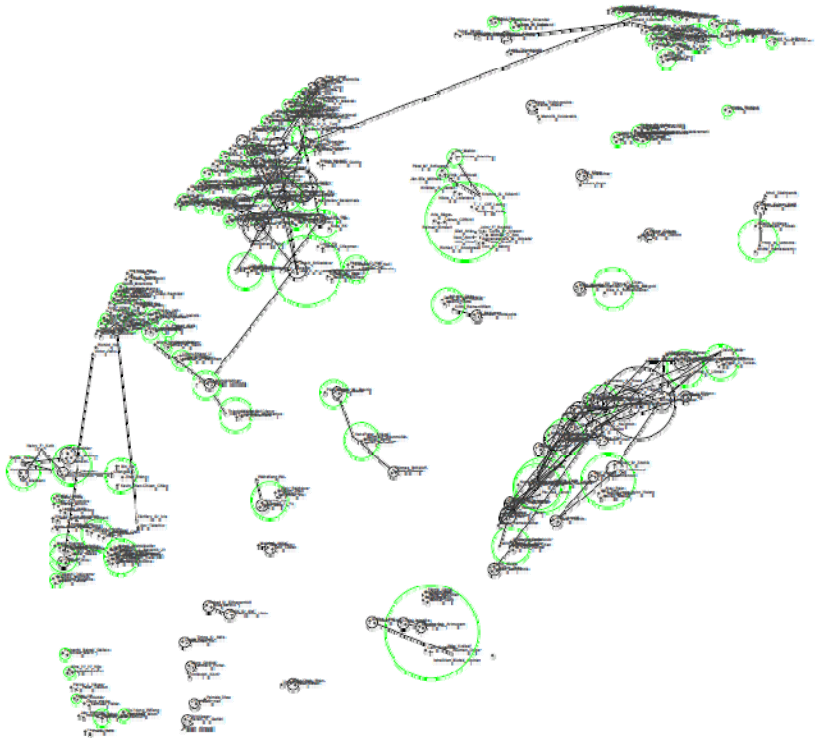
  (d) present in details the most interesting structures in each power graph.

Top-5 DB conferences



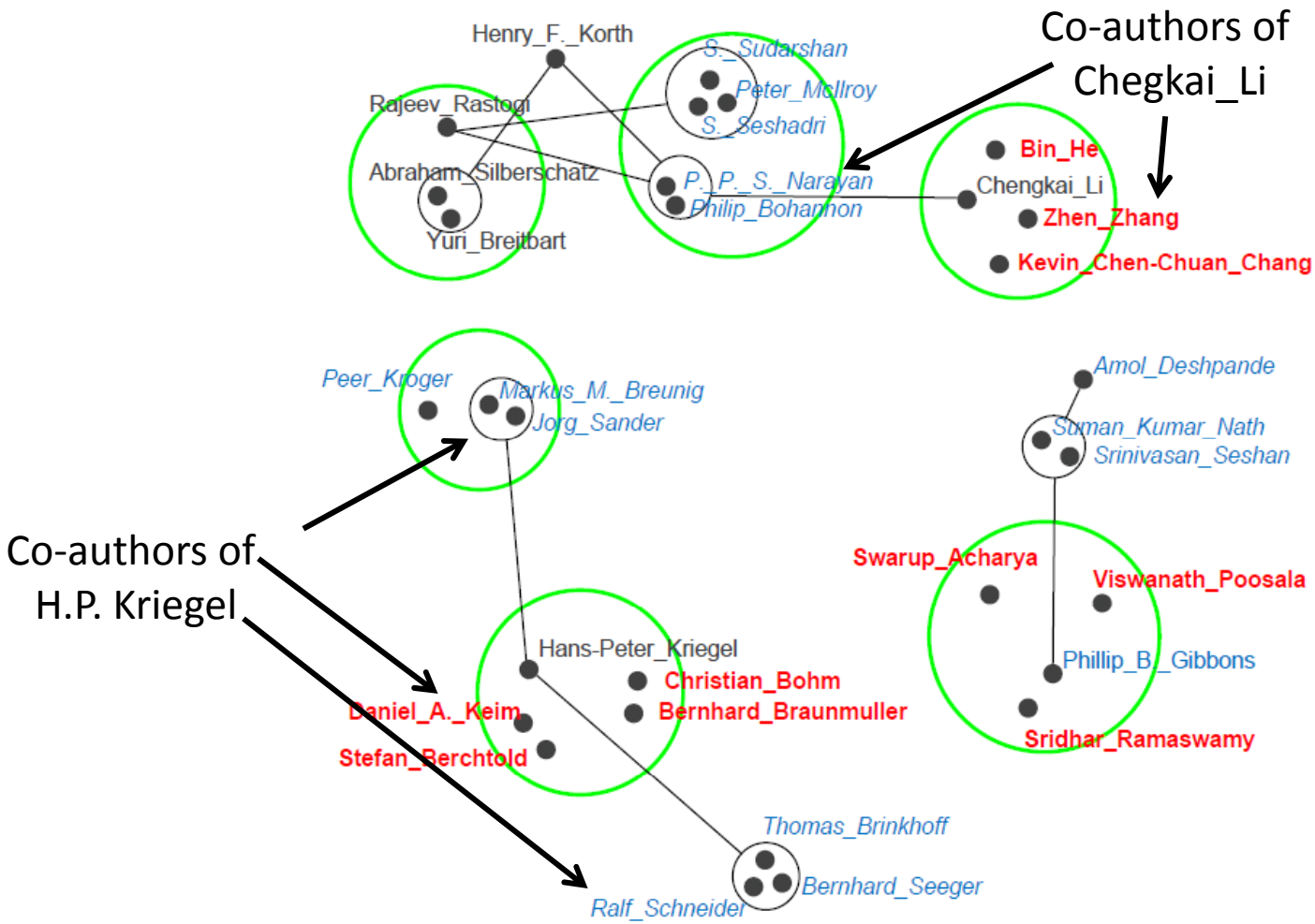
The complete Power Graph for the author of the top-5 Database conferences (SIGMOD, VLDB, PODS, ICDE, ICDT)

11,369 papers written by 10,524 authors (since 1969)



Pruned Power Graph
(weak Power Edges have been removed)

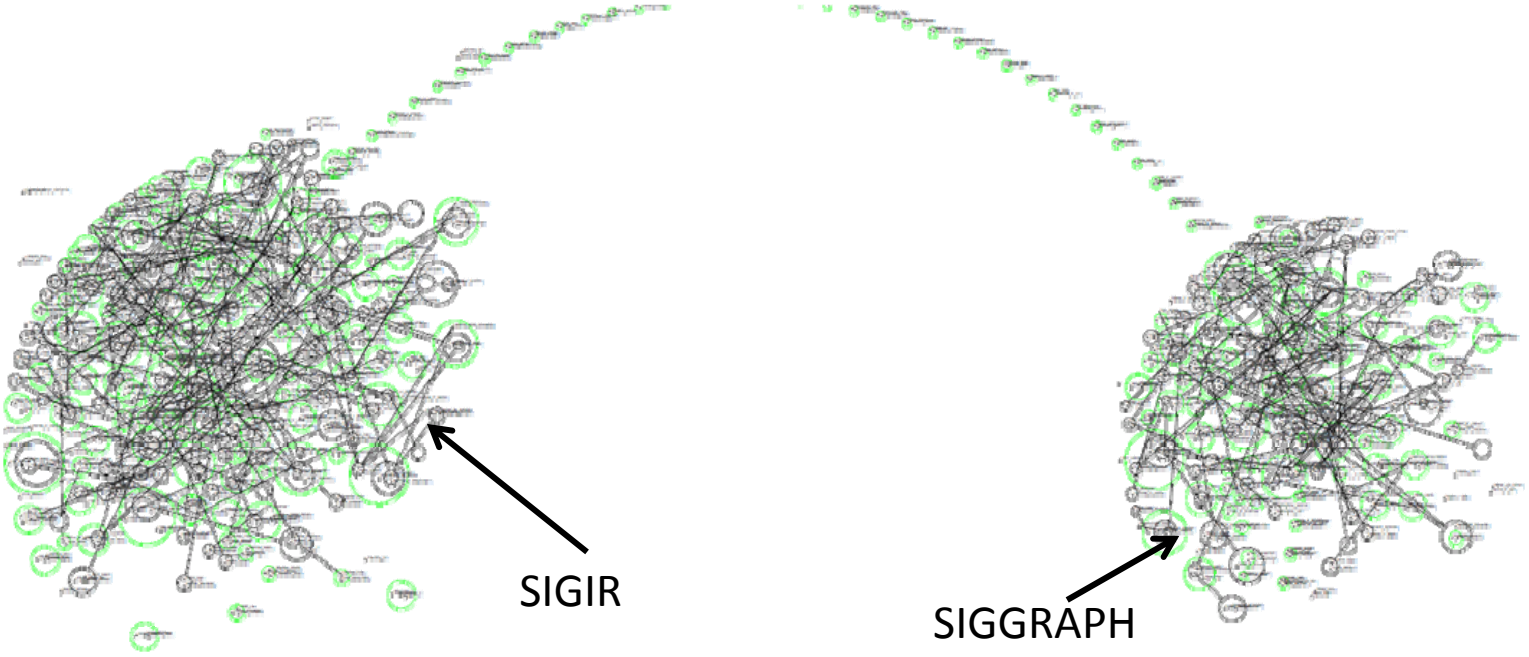
Top-5 DB conferences: zoom in



Potential synergies must be examined between authors in red and authors in blue in each sub-graph

Multidisciplinary graph

- Choose papers that have been published in SIGGRAPH (computer graphics) and SIGIR (information retrieval)
- 3,568 papers written by 5,386 authors



Pruned Power Graph

Measuring content similarity

- Potential research synergies can be ranked according to the content similarity of the respective publication records for each author

TOP CANDIDATE PAIRS, RANKED BY SIMILARITY

Using only top-5 DB conferences

Author A	Author B	similarity
Daniel_Deutch	Anat_Eyal	0.222
Kristen_LeFevre	Alexandre_V._Evfimievski	0.179
Ming-Chuan_Wu	Steve_Herbert	0.156
Babu_Krishnaswamy	Aleksandras_Surna	0.154
Alon_Y._Halevy	Chen_Li	0.152
Ming-Chuan_Wu	Aleksandras_Surna	0.148
Jorg_Sander	Daniel_A._Keim	0.139
Conor_Cunningham	Steve_Herbert	0.138
Sandeepan_Banerjee	Anand_Manikutty	0.136
Yanif_Ahmad	Magdalena_Balazinska	0.135

Wrote a paper on CIDR 2005

Conclusions

- a novel approach for the organization and the efficient presentation of bibliographic database contents
- use of a graph reduction method that facilitates
  - the efficient visualization of the dense co-authorship graph,
  - The identification of potential research synergies based on the analysis of the Power Graph, and
  - the ranking of potential coauthor pairs by similarity of interests
- It is on our next plans to apply the same approach to more bibliographic networks as well as to other social networks, and to study the evolution of the graphs over time based on the comparison of different graph snapshots taken in different years
- G. Tsatsaronis, I. Varlamis, S. Torge, M. Reimann, K. Nørvåg, M. Schroeder and M. Zschunke, "***How to Become a Group Leader? or Modeling Author Types based on Graph Mining***", International Conference on Theory and Practice of Digital Libraries **(TPDL 2011)**, September 2011, Berlin, Germany.