Resolving Author Name Homonymy to Improve Resolution of Structures in Co-author Networks

Theresa Velden, Asif-ul Haque, Carl Lagoze
Cornell University

JCDL 2011

name homonymy := same name for
different individuals

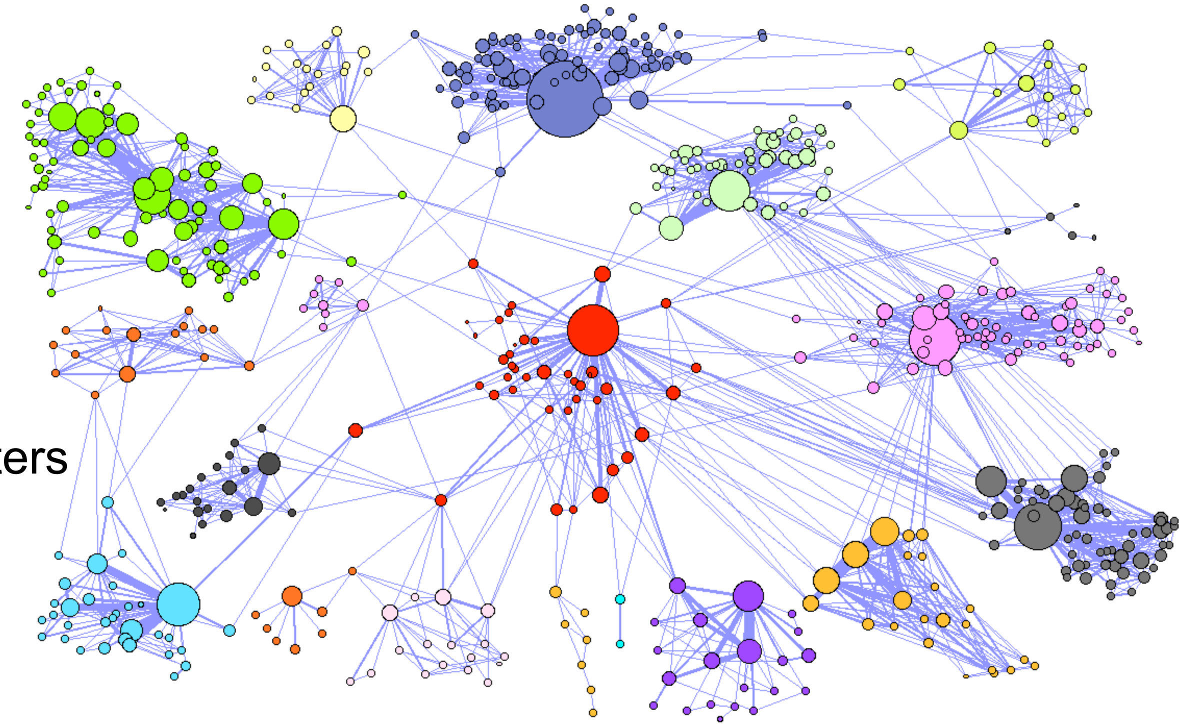
e.g.: J.H. Kim, or M. Smith

Outline

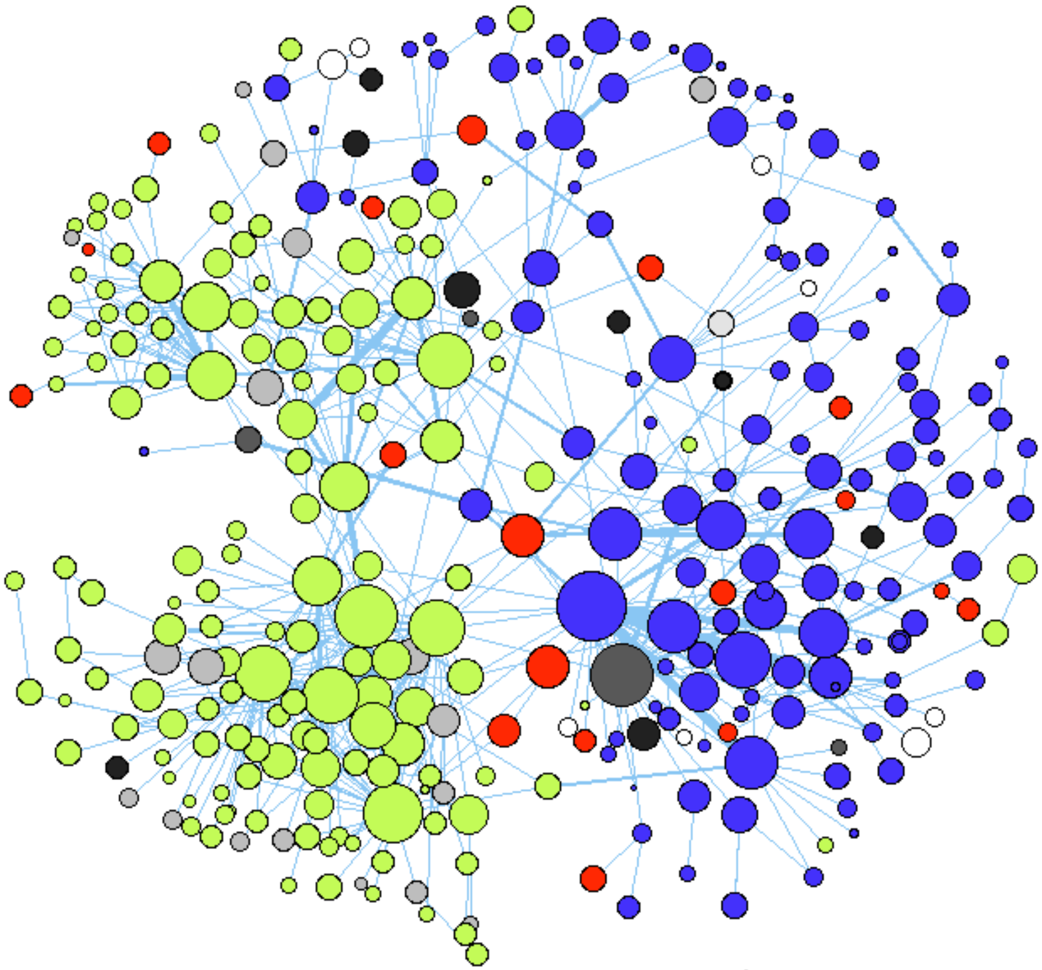
- Motivation
- Approach
  - Disambiguation Algorithm
  - Case study data set
  - Mesoscopic network structure & ground truth sample
- Results
- Conclusions

Motivation

- Increasing interest in structural analysis of co-author networks to study patterns and temporal dynamics of scientific collaboration
- **Meso-scopic analysis:** Clustering exposes modular substructure of co-author networks
- Our work: compare between scientific fields:
  - internal structure of co-author clusters
  - collaboration patterns between co-author clusters



Motivation



Global group collaboration network of a research specialty

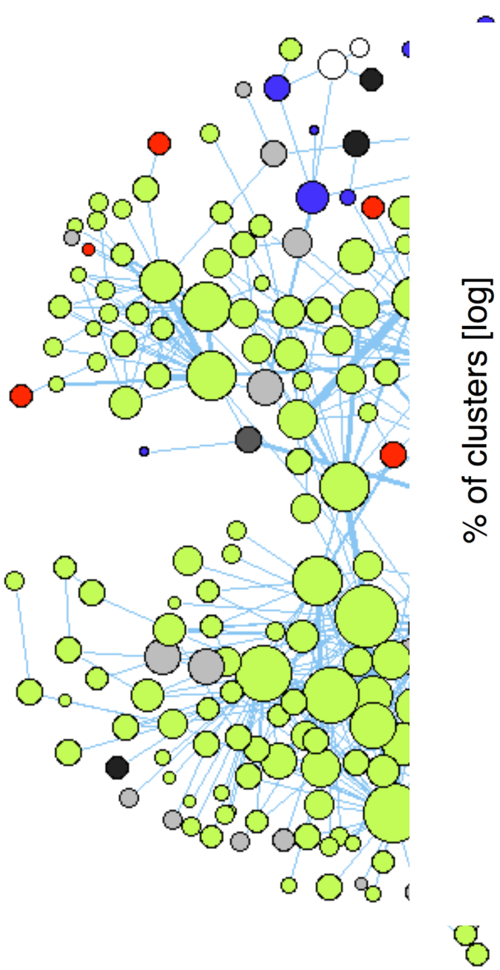
- Asian
- European
- North American

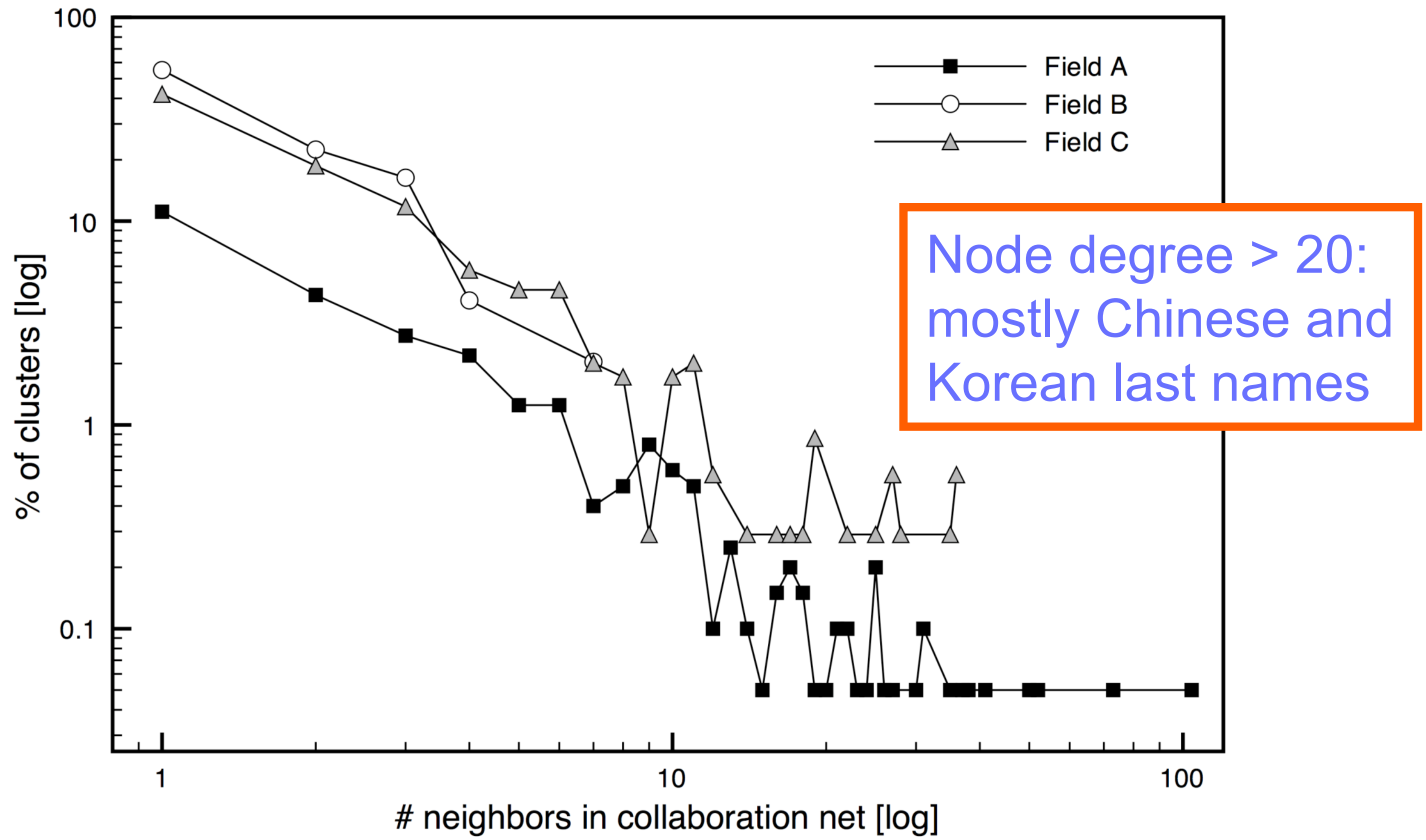
WoS ISI data: 1987-2008
authors identified by initials and last name

→ coauthor network with about 18,000 authors

Velden, Haque, Lagoze, Scientometrics 85(1), 2010

Motivation





Velden, Haque, Lagoze, Scientometrics 85(1), 2010

Motivation

- Conclusion: suspect relevant network distortion by name homonymy

- Goal of this study:
  - assess network distortion introduced by name homonymy
  - develop and evaluate a simple disambiguation algorithm that
    - uses minimal features (wide applicability)
    - scales for use on large data sets

Approach: algorithm

- Data features used:
  - **co-author names** by itself very effective: I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K. Sung, and J.-H. Lee. Information Processing and Management, 45:84−97, 2009; also: H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. In JCDL 2004
  - **self-citation;** high precision reported: D. M. McRae-Spencer and N. R. Shadbolt. In JCDL, 2006

→ for each author name grow connected components of authoring instances (publications) using co-author overlap ≥ 1 and self-citation as merge criteria

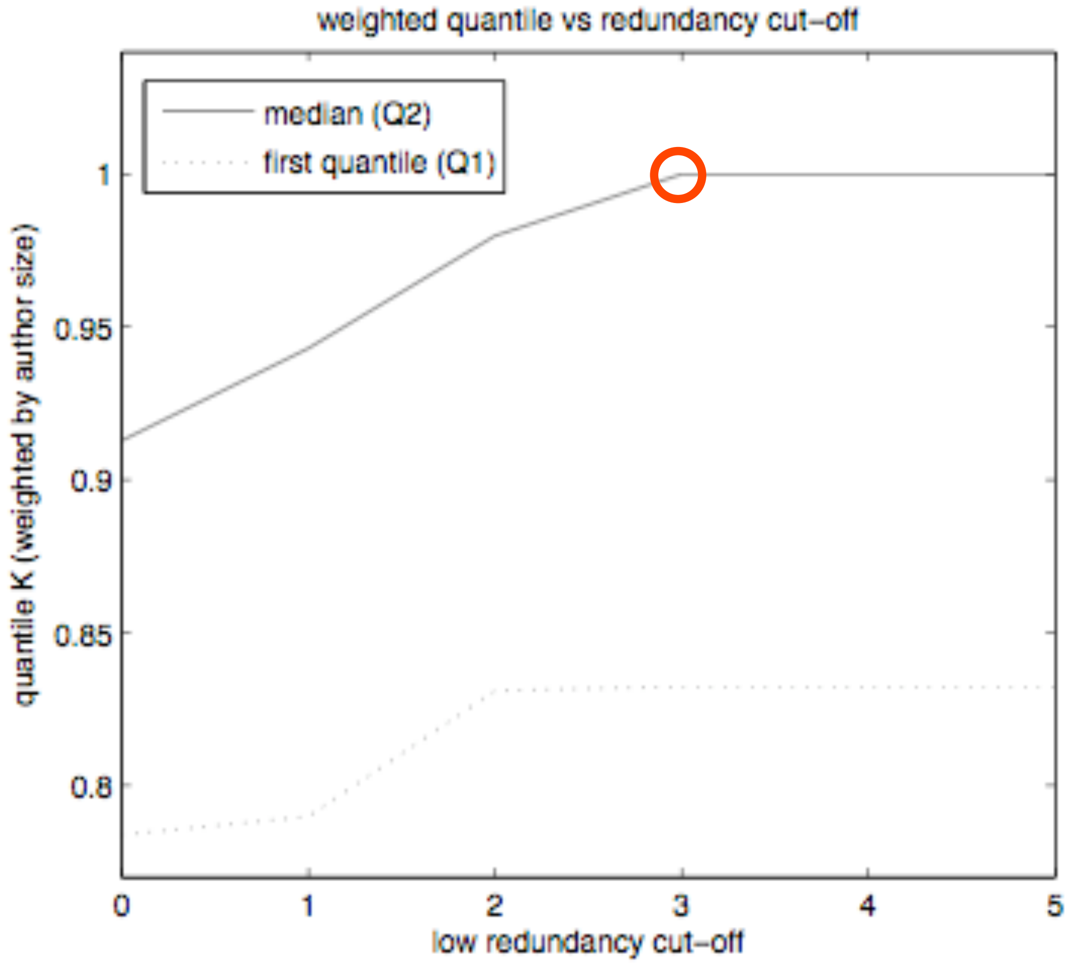
Approach: algorithm

- However, beneficial to entirely exclude less common last names from disambiguation attempt…

- **Cut-off parameter based on commonality** (ambiguity) of coauthor name:
  - ‘raw name redundancy’ r_n: counting occurrence of unique initials for each last name
  - derived from data set
  - same name commonality metric as Bhattacharya and Getoor, ACM Trans. Knowl.Discov. Data, 1, March 2007

Approach: cut-off parameter

Semi-supervised:
cut-off parameter
for name redundancy
empirically determined
from training data



Approach: K-metric

Ferreira, A. Veloso, M. Goncalves, and A. Laender. JCDL, 2010

N: nodes in article graph
i: empirical clustering (algorithm)
j: theoretical clustering (groundtruth)

Average clustering purity:

$$\mathbf{ACP} = \frac{1}{N} \sum_{i=1}^{e} \sum_{j=1}^{t} \frac{n_{ij}^2}{n_i}$$

Average author purity (fragmentation):

$$\mathbf{AAP} = \frac{1}{N} \sum_{j=1}^{t} \sum_{i=1}^{e} \frac{n_{ij}^2}{n_j}$$

$$\mathbf{K} = \sqrt{\mathbf{ACP} \times \mathbf{AAP}}$$

→ use K weighted by # of publications

Approach: case study data set

- From a comparative study of collaboration patterns in research specialties in chemistry
- Web of Science (Thomson Reuters) data
- Time range: 1987-2008, 22 years
- 29,905 publications
- Co-author network (undisambiguated): 18,419 nodes
- Giant component size: 93.7%
- Co-authors per paper: mean 3.8, median 3 (max 34)

Approach: case study data set

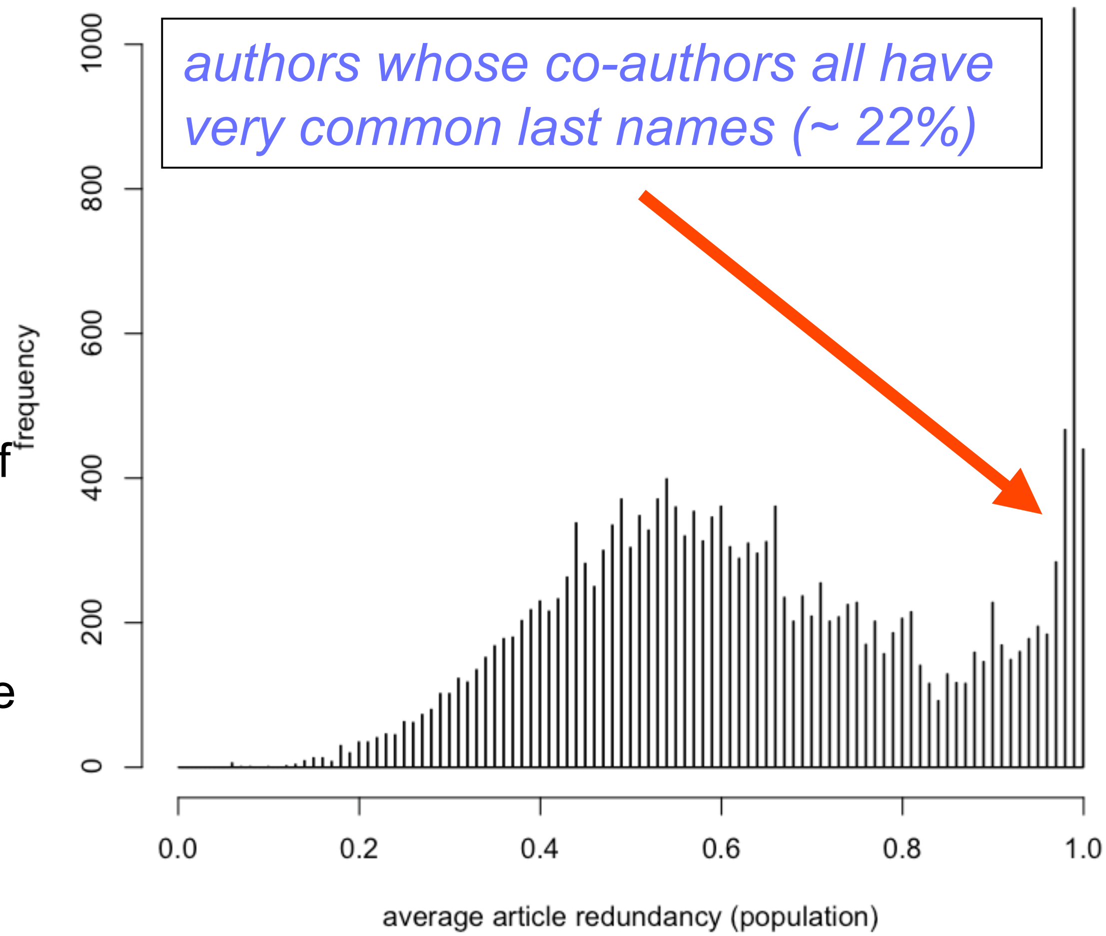
name redundancy s_n
of a last name L:

$$s_n(L) = \mathbf{Pr}[X \leq r_n(L)]$$

with $r_n(L)$: raw name redundancy

article redundancy := product of name redundancies of all co-authors

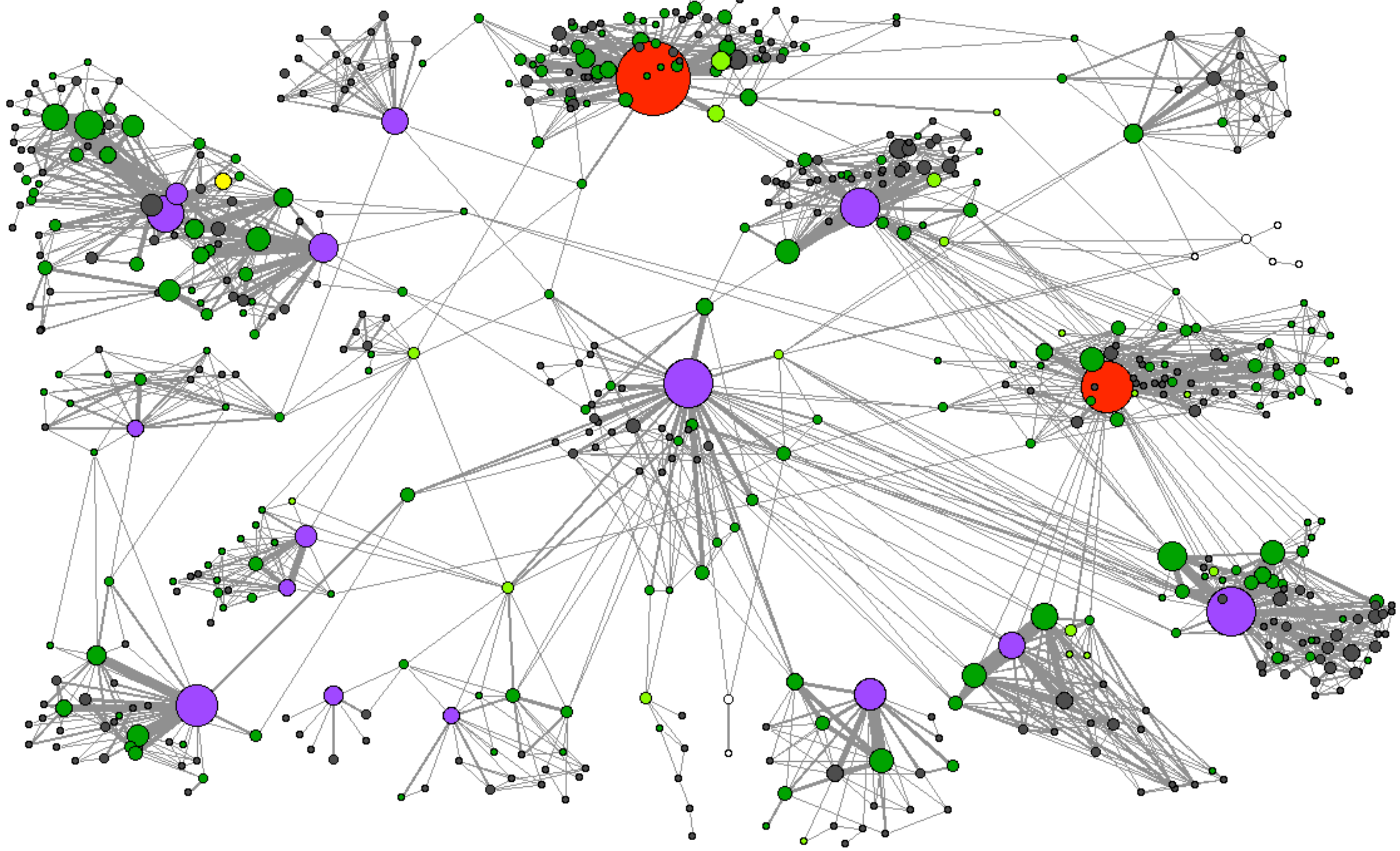
average article redundancy: the average of article redundancies for an (undisambiguated) author



Approach: mesoscopic network structure

- classification of nodes by cluster-internal and cluster-external links

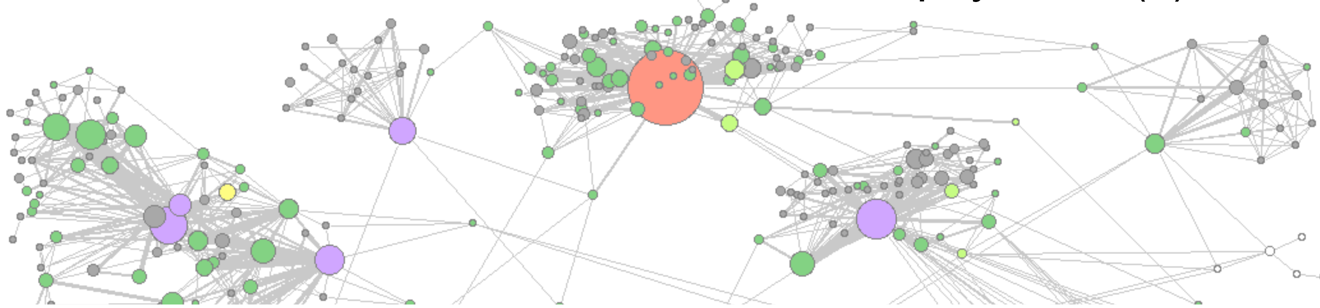
Guimera, M. Sales-Pardo, and L. Amaral. Nature physics, 3(1):63–69, 2007



Approach: mesoscopic network structure

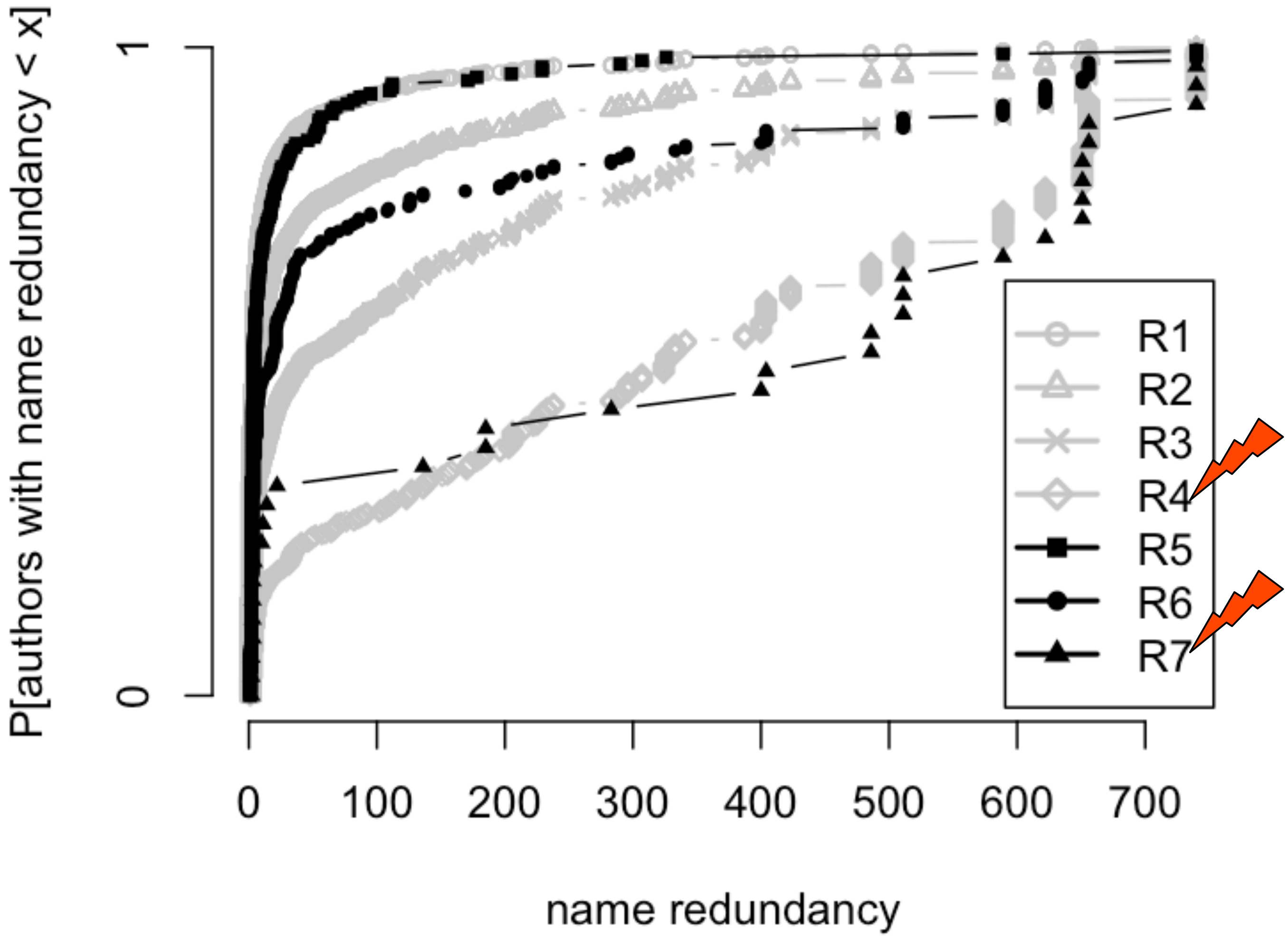
- classification of nodes by cluster-internal and cluster-external links

Guimera, M. Sales-Pardo, and L. Amaral. Nature physics, 3(1):63–69, 2007



	Node Role	Characterization	Proportion in Population
Non Hubs	R1	'ultra-peripheral nodes'	30.3%
	R2	'peripheral nodes'	48.4%
	R3	'connector nodes'	14.8%
	R4	'satellite connector nodes'	3.6%
Hubs	R5	'provincial hubs'	1.1%
	R6	'connector hubs'	1.5%
	R7	'global hubs'	0.2%

Approach: node role specific distortion



Heavy tail of high redundancy names

Approach: ground truth sample

- Statistical representative ground truth sample stratified by node role

	Node Role	Number in Population	Number in Ground-truth Sample	Proportion of Population Sampled
Non Hubs	R1	5167	102	2.0%
	R2	8245	102	1.2%
	R3	2527	102	4.0%
	R4	611	89	14.6%
Hubs	R5	195	72	36.9%
	R6	257	77	30.0%
	R7	34	28	82.4%

Sample size to allow determination of error with 10% accuracy (95% confidence interval); training data set: sampled an additional 33% for each stratum

Results: network distortion

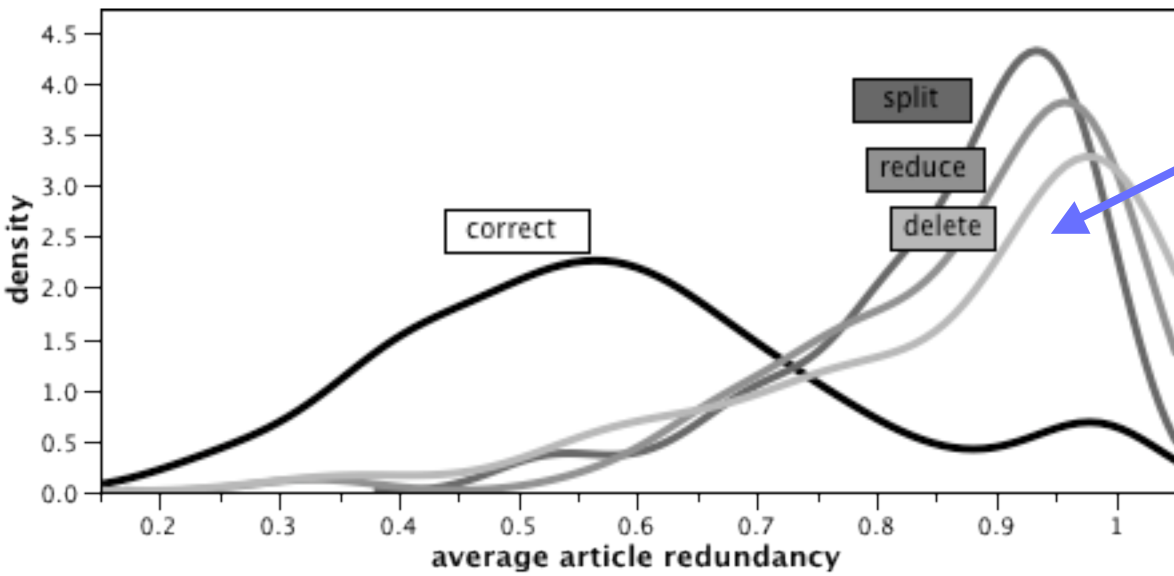
Error for ground truth sample of authors

	R1 [%]	R2 [%]	R3 [%]	R4 [%]	R5 [%]	R6 [%]	R7 [%]
correct	98.0	80.4	51.5	22.5	88.9	72.7	32.1
reduce	0	7.8	11.9	16.9	6.9	10.4	28.6
split	1.0	3.9	10.9	11.2	4.2	13.0	17.9
delete	1.0	7.8	25.7	49.4	0	3.9	21.4

Results: network distortion

Error for ground truth sample of authors

	R1 [%]	R2 [%]	R3 [%]	R4 [%]	R5 [%]	R6 [%]	R7 [%]
correct	98.0	80.4	51.5	22.5	88.9	72.7	32.1
reduce	0	7.8	11.9	16.9	6.9	10.4	28.6
split	1.0	3.9	10.9	11.2	4.2	13.0	17.9
delete	1.0	7.8	25.7	49.4	0	3.9	21.4



author teams with exclusively very common last names

Results: algorithm performance

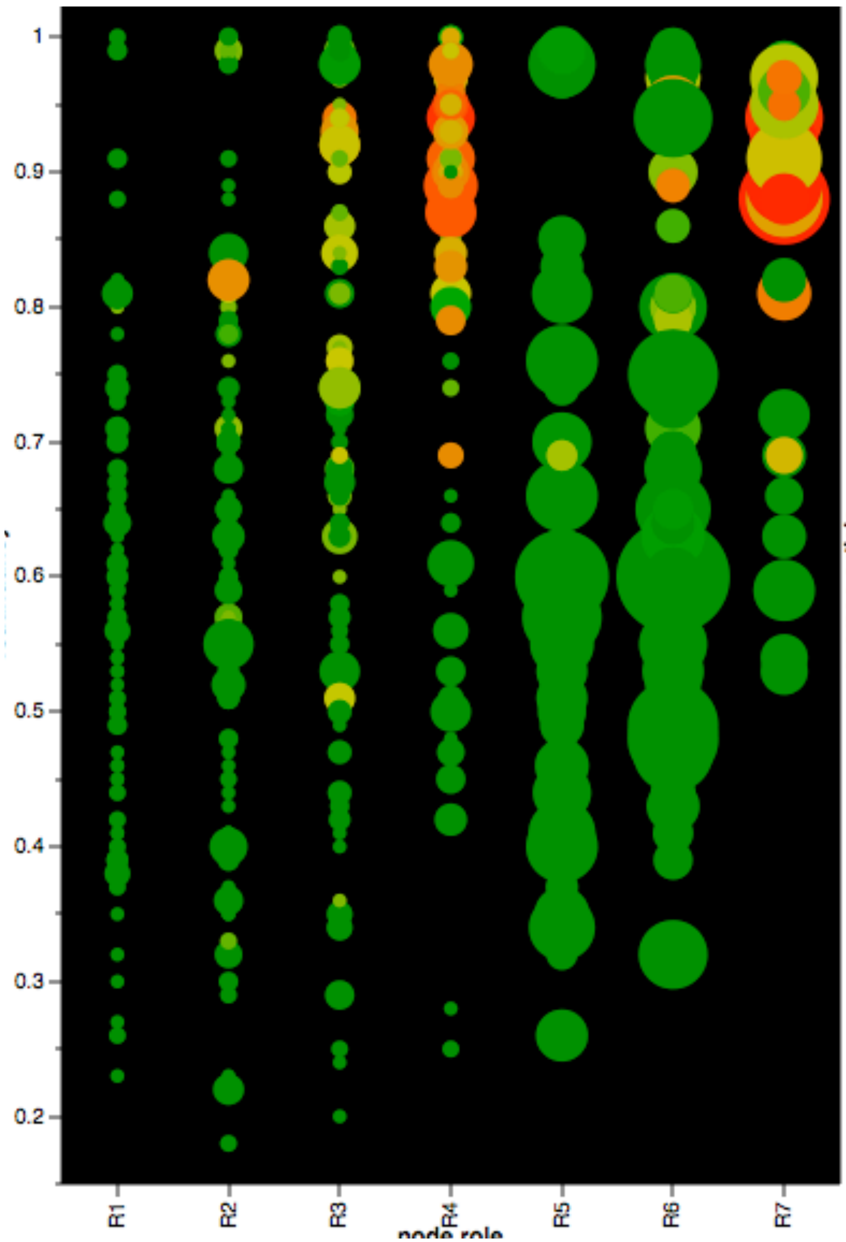
weighted k (571 authors in groundtruth)

	median		25%	
	nondis	dis	nondis	dis
R1	1.00	1.00	1.00	1.00
R2	1.00	1.00	1.00	1.00
R3	0.85	1.00	0.65	0.89
R4	0.50	1.00	0.40	0.89
R5	1.00	1.00	1.00	1.00
R6	1.00	1.00	1.00	0.98
R7	0.54	0.93	0.28	0.89

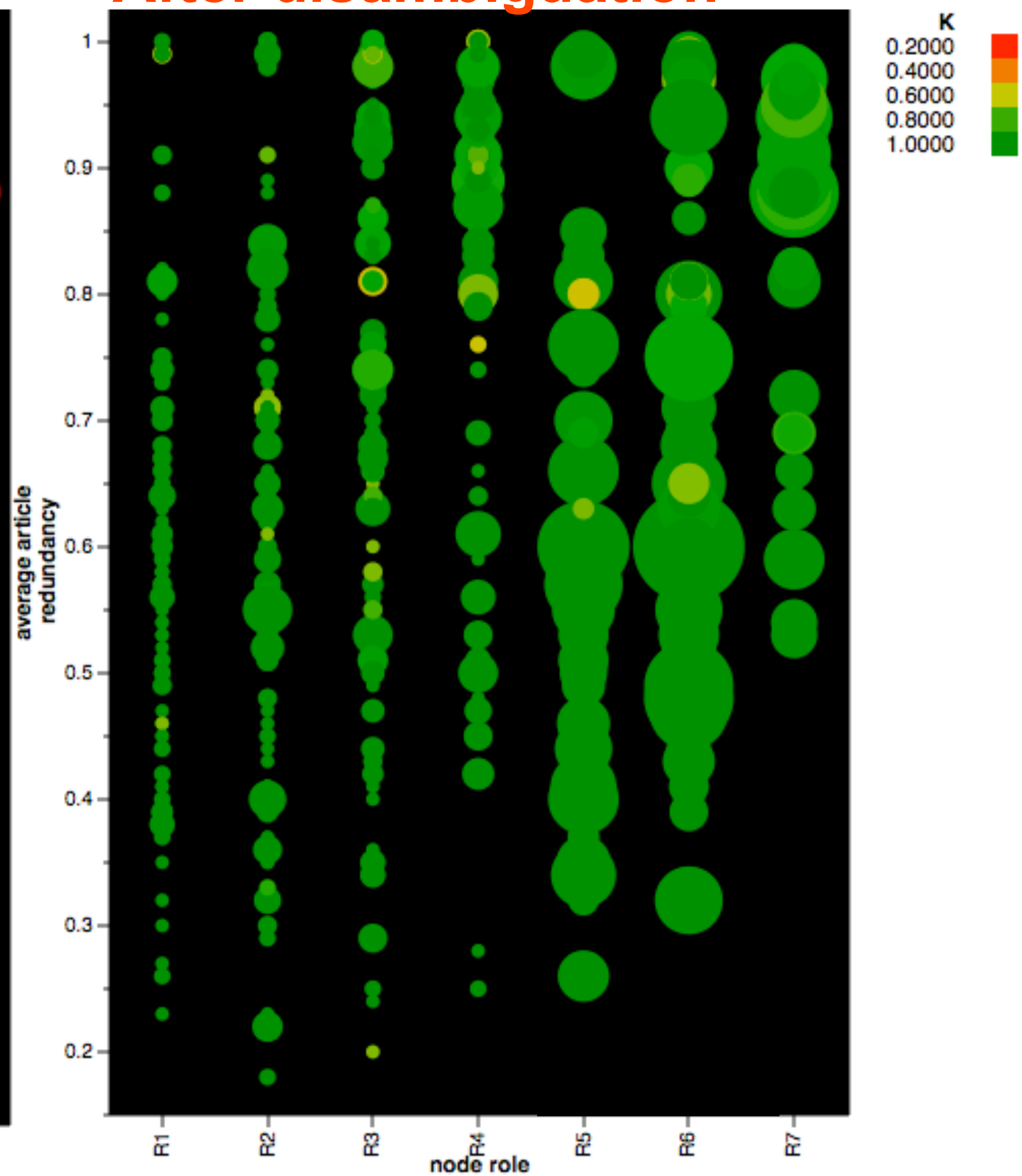
Remaining error:
oversplitting (15.9%), over-merging (2.6%), oversplitting & overmerging (4.6%)

Results: algorithm performance

Before

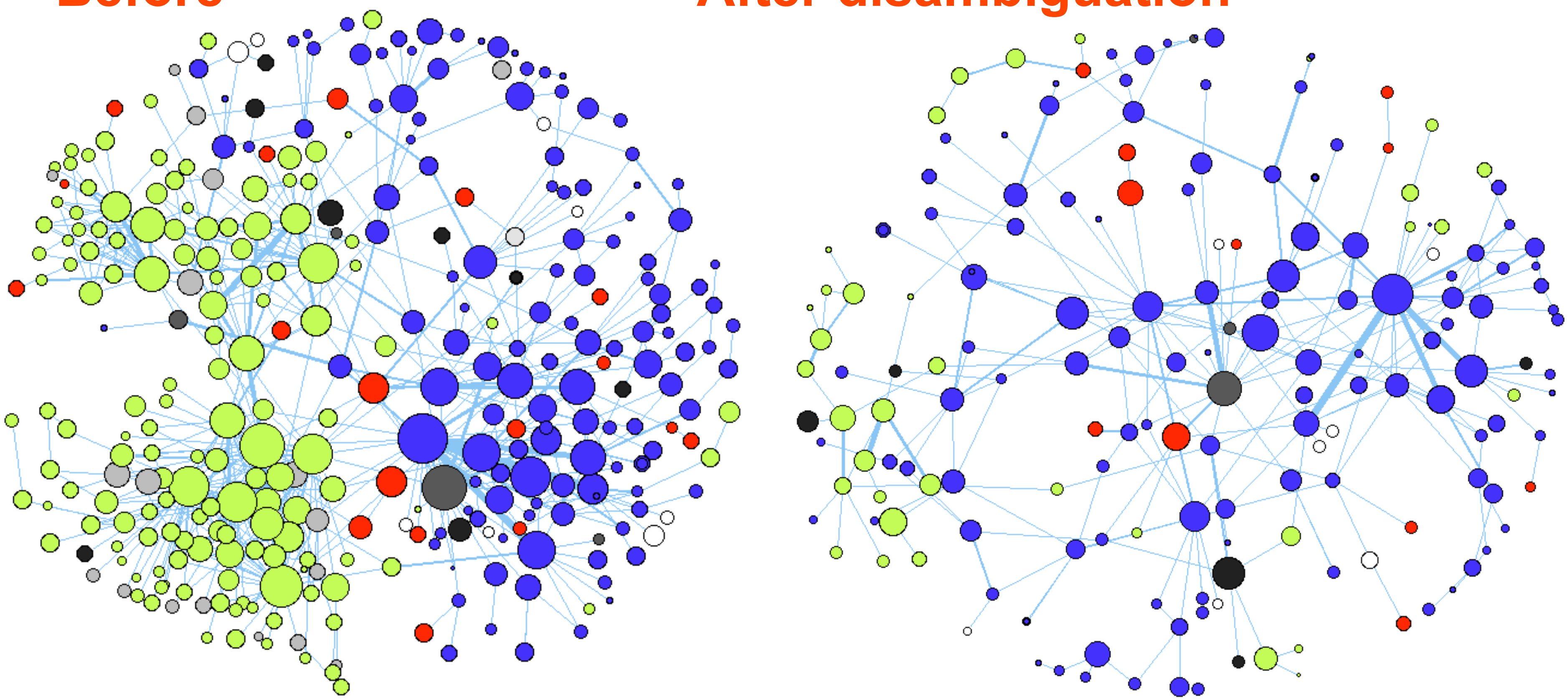


After disambiguation



Results: Collaboration Network

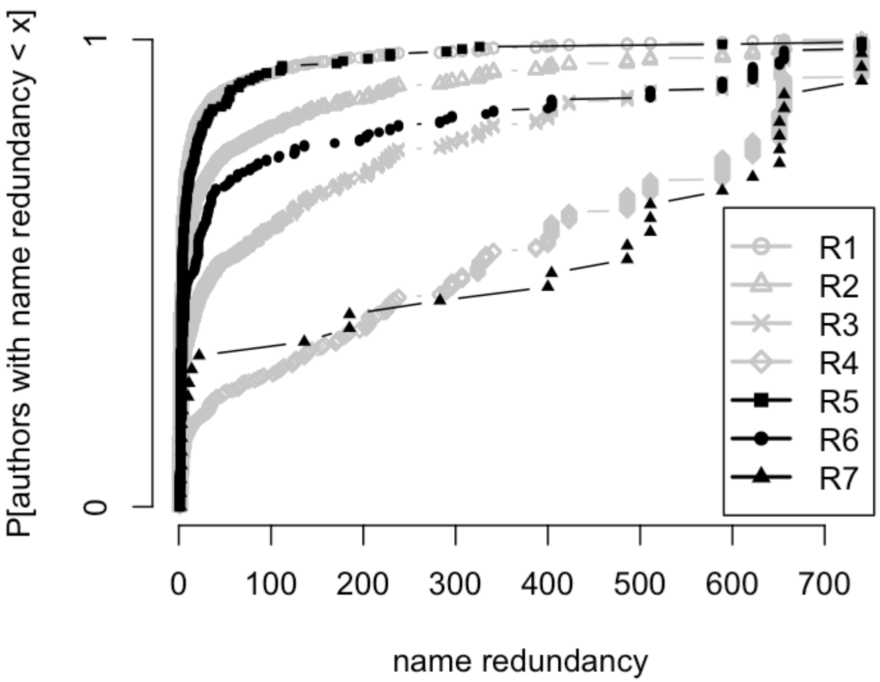
Before **After disambiguation**



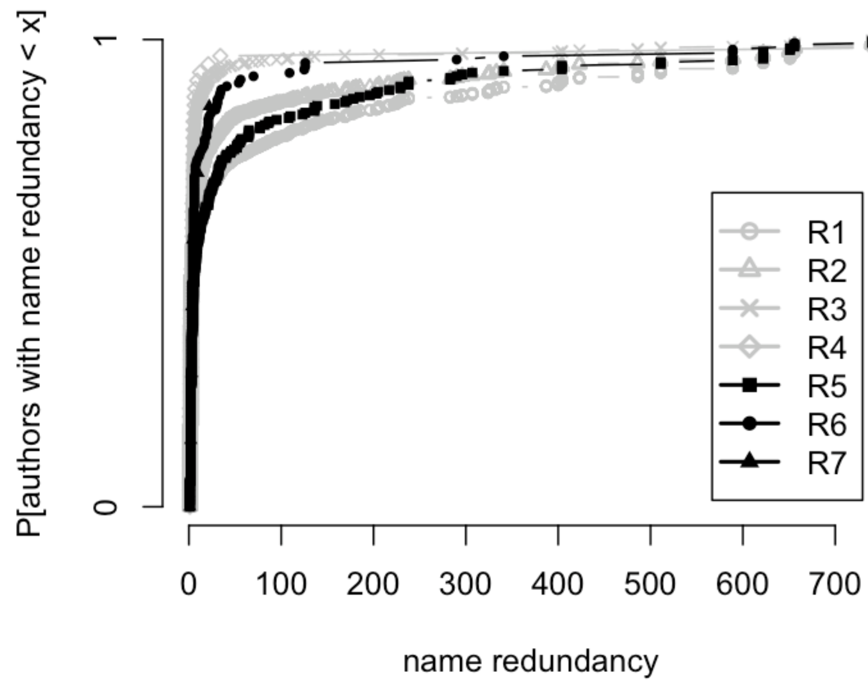
proportion of Asian affiliated author clusters: reduced from 43% to 19%
average node degree decrease from 3.9 to 2.8

Results: assessing distortion without groundtruth

Before



After disambiguation



Conclusions

- Homonymy introduces significant network distortion, especially for cluster interconnectivity
- Algorithm effectively reduces error using co-author names, selfcitations, name commonality
- Advantages of algorithm: scalability, broad applicability
- New approach to assessing distortion without (expensive) ground truth: differences between node role classes w.r.t. distribution of the commonality of last names

Thank you!

ground truth data online: http://arxiv.org/abs/1106.2473

contact: tav6@cornell.edu