Homonym Detection in Curated Bibliographies: Learning from dblp's Experience

Marcel R. Ackermann, Florian Reitz
LZI Schloss Dagstuhl (dblp), Germany

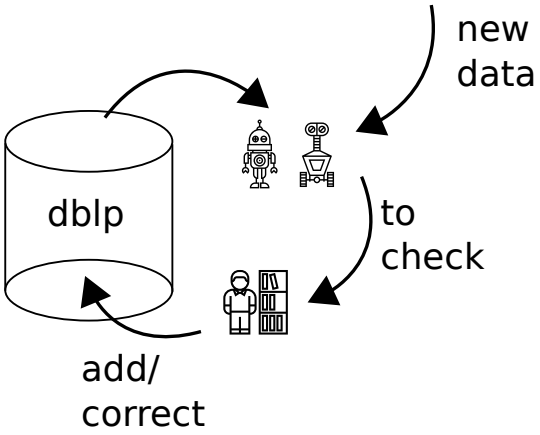
TPDL September 2018, Porto

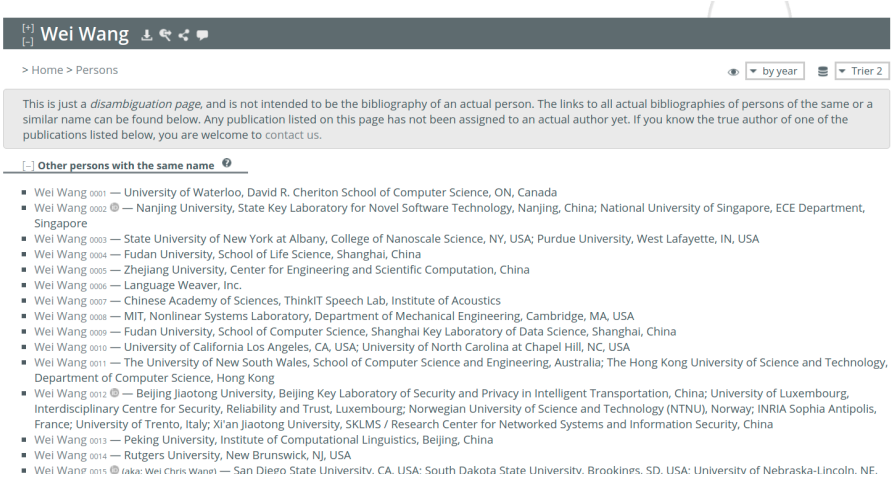
What is dblp?

- Computer science $+ \epsilon$
- 4 million records
- 2 million author profiles

- Computer science $+ \epsilon$
- 4 million records
- 2 million author profiles

```
author: Marcel R. Ackermann
author: Florian Reitz
title: Homonym Detection in Bibliographies
year: 2018
journal: CoRR
ee: http://arxiv.org/abs/1806.06017
```


new
data
dblp
to
check
add/
correct



[+] [–] Wei Wang

> Home > Persons

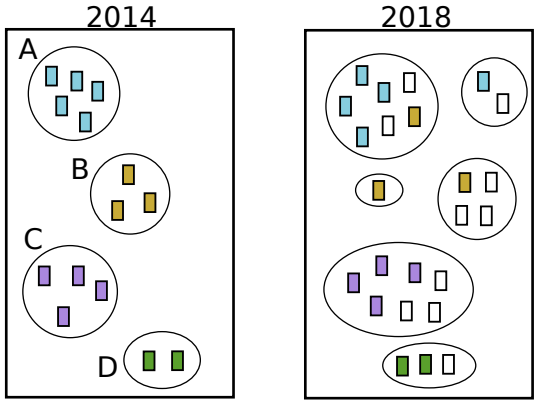
by year | Trier 2

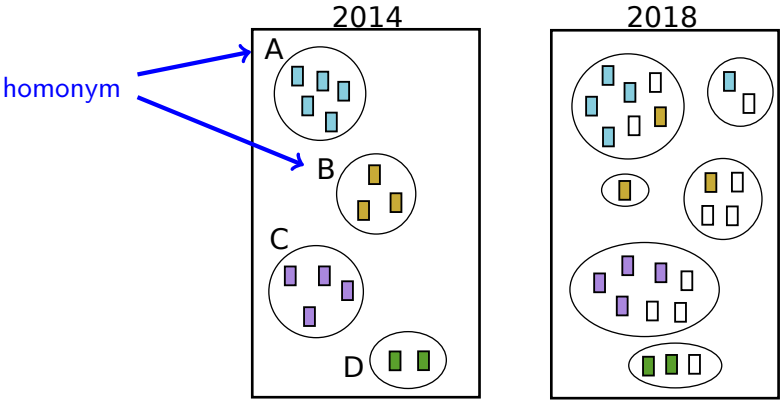
This is just a *disambiguation page*, and is not intended to be the bibliography of an actual person. The links to all actual bibliographies of persons of the same or a similar name can be found below. Any publication listed on this page has not been assigned to an actual author yet. If you know the true author of one of the publications listed below, you are welcome to contact us.

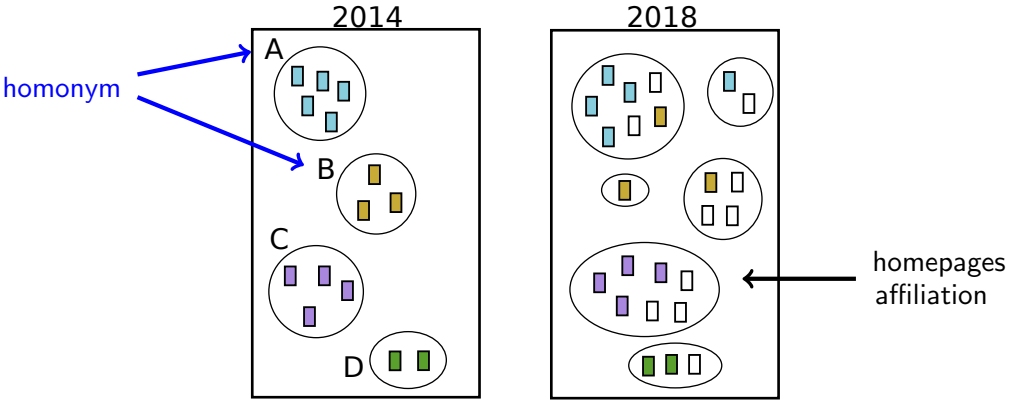
[–] **Other persons with the same name**

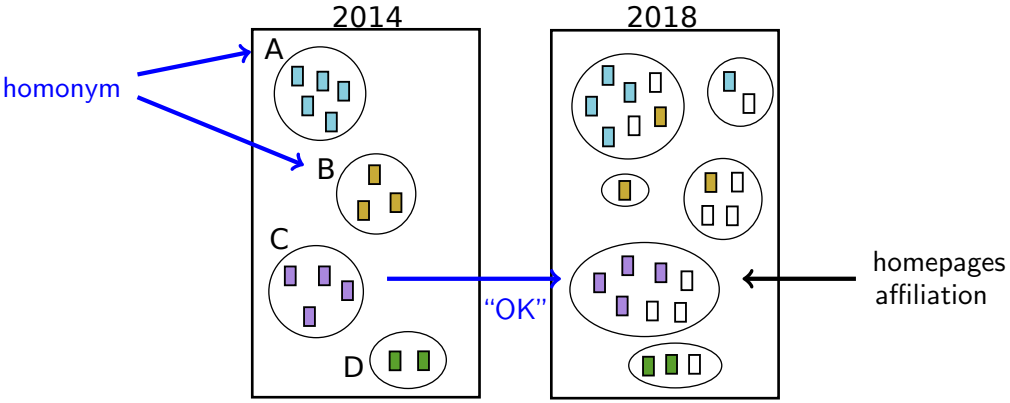
- Wei Wang 0001 — University of Waterloo, David R. Cheriton School of Computer Science, ON, Canada
- Wei Wang 0002 — Nanjing University, State Key Laboratory for Novel Software Technology, Nanjing, China; National University of Singapore, ECE Department, Singapore
- Wei Wang 0003 — State University of New York at Albany, College of Nanoscale Science, NY, USA; Purdue University, West Lafayette, IN, USA
- Wei Wang 0004 — Fudan University, School of Life Science, Shanghai, China
- Wei Wang 0005 — Zhejiang University, Center for Engineering and Scientific Computation, China
- Wei Wang 0006 — Language Weaver, Inc.
- Wei Wang 0007 — Chinese Academy of Sciences, ThinkIT Speech Lab, Institute of Acoustics
- Wei Wang 0008 — MIT, Nonlinear Systems Laboratory, Department of Mechanical Engineering, Cambridge, MA, USA
- Wei Wang 0009 — Fudan University, School of Computer Science, Shanghai Key Laboratory of Data Science, Shanghai, China
- Wei Wang 0010 — University of California Los Angeles, CA, USA; University of North Carolina at Chapel Hill, NC, USA
- Wei Wang 0011 — The University of New South Wales, School of Computer Science and Engineering, Australia; The Hong Kong University of Science and Technology, Department of Computer Science, Hong Kong
- Wei Wang 0012 — Beijing Jiaotong University, Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, China; University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust, Luxembourg; Norwegian University of Science and Technology (NTNU), Norway; INRIA Sophia Antipolis, France; University of Trento, Italy; Xi'an Jiaotong University, SKLMS / Research Center for Networked Systems and Information Security, China
- Wei Wang 0013 — Peking University, Institute of Computational Linguistics, Beijing, China
- Wei Wang 0014 — Rutgers University, New Brunswick, NJ, USA
- Wei Wang 0015 (aka: Wei Chris Wang) — San Diego State University, CA, USA; South Dakota State University, Brookings, SD, USA; University of Nebraska-Lincoln, NE,

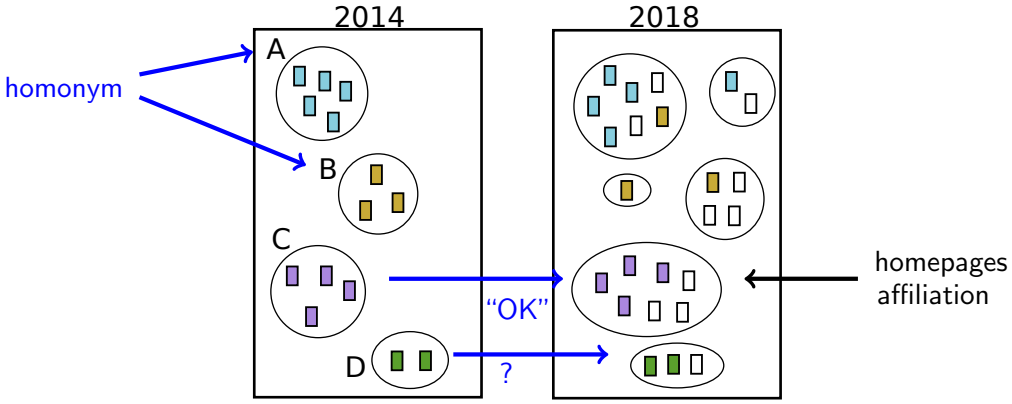
Task: find profiles that are most likely homonyms.

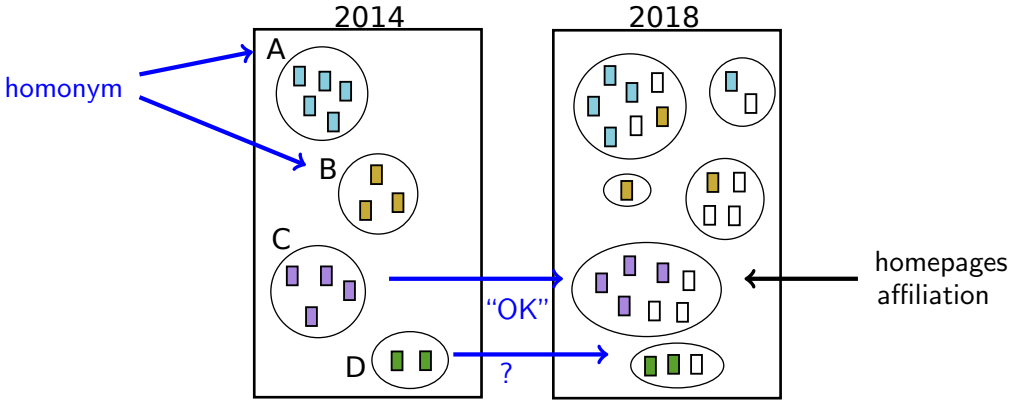

2014
A
B
C
D
2018


homonym
2014
A
B
C
D
2018


2014
2018
homonym
A
B
C
D
homepages
affiliation

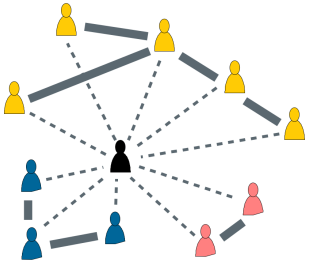

2014
2018
homonym
A
B
C
D
"OK"
homepages
affiliation


2014
2018
homonym
A
B
C
D
"OK"
?
homepages
affiliation

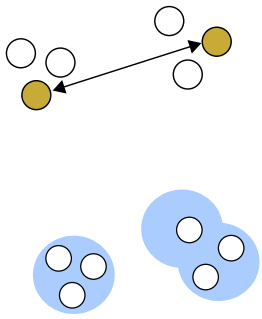


homonyms: 2802
no homonyms: 21576

group	dims	features
B	3	basics: #publications, #coauthors, #co-relations
C	7	local coauthor community coherence
T	12	semantic similarity of publication titles
V	13	semantic similarity of publication venues
Y	4	years of activity
total	39	

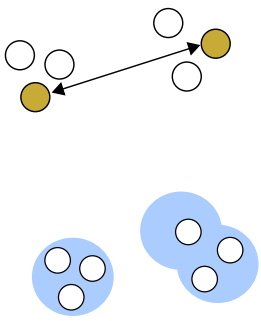


- #Coauthor groups
- Size of 5 largest cluster
- Unevenness of size distribution



Titles: doc2vec in $\mathbb{R}^{150}$ (cos distance)

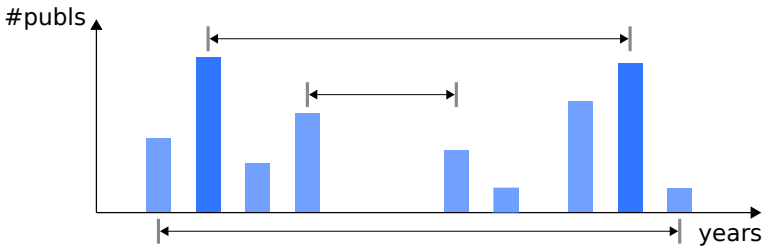
- Diameter
- Greedy cover number
- 5/25/50/75/95 percentile
  - Pairwise distance
  - Distance to centroid



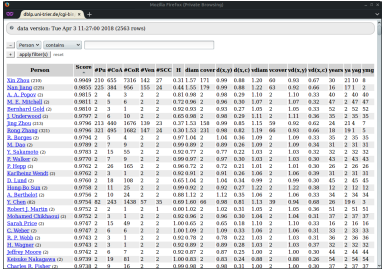
Venues: doc2vec in $\mathbb{R}^{150}$ (cos distance)

- Diameter
- Greedy cover number
- 5/25/50/75/95 percentile
  - Pairwise distance
  - Distance to centoid
- #Different venues

- #Active years
- Span: earlies to most recent
- Largest gap
- Distance between largest modes



feat.	precision	recall	F1-score	**MCC**	AUROC
B	0.823 ± 0.31	0.024 ± 0.01	0.047 ± 0.02	$\mathbf{0.130} \pm 0.06$	0.799 ± 0.01
BC	0.818 ± 0.17	0.057 ± 0.02	0.106 ± 0.03	$\mathbf{0.197} \pm 0.05$	0.842 ± 0.01
BT	0.542 ± 0.18	0.051 ± 0.03	0.092 ± 0.05	$\mathbf{0.138} \pm 0.06$	0.786 ± 0.01
BV	0.745 ± 0.04	0.232 ± 0.05	0.350 ± 0.07	$\mathbf{0.372} \pm 0.06$	0.815 ± 0.01
BY	0.781 ± 0.02	0.153 ± 0.01	0.256 ± 0.02	$\mathbf{0.314} \pm 0.02$	0.820 ± 0.00
BTV	0.709 ± 0.01	0.268 ± 0.01	0.389 ± 0.02	$\mathbf{0.393} \pm 0.01$	0.832 ± 0.00
All	$\mathbf{0.793} \pm 0.01$	$\mathbf{0.424} \pm 0.01$	$\mathbf{0.552} \pm 0.01$	$\mathbf{0.541} \pm 0.01$	$\mathbf{0.890} \pm 0.00$



top 100

- 74 true positive
- 12 false positive
- 14 ???

- Top-k precision acceptable
- Title: semantic similarity not helpful
- Venue: semantic similarity **most** helpful
- Improve vectorization.

- Top-k precision acceptable
- Title: semantic similarity not helpful
- Venue: semantic similarity **most** helpful
- Improve vectorization.

Thank you!

@dblp_org