

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

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

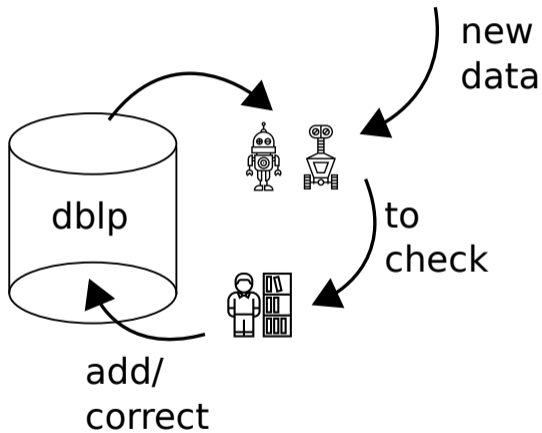
TPDL September 2018, Porto




What is dblp?

- ▶ Computer science + ϵ
- ▶ 4 million records
- ▶ 2 million author profiles

- ▶ Computer science + ϵ
- ▶ 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
```



[+] Wei Wang   

> Home > Persons

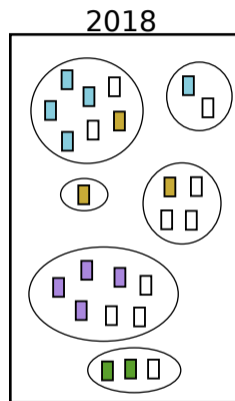
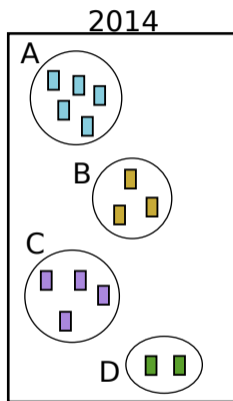
 

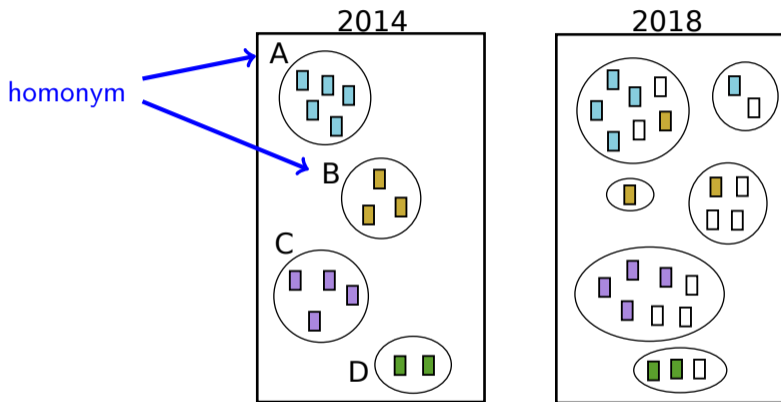
This is just a *disambiguation page*, and is not intended to be the bibliography of an actual person. The links to all actual bibliographies of persons of the same or a similar name can be found below. Any publication listed on this page has not been assigned to an actual author yet. If you know the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

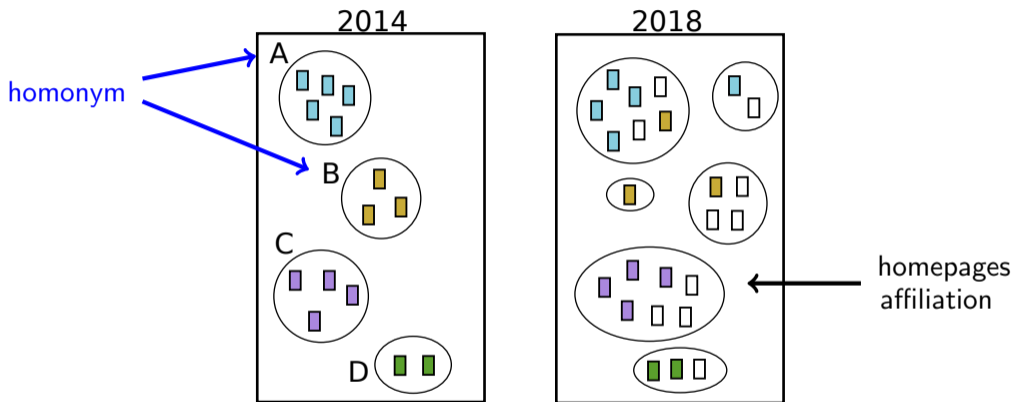
[\[-\] Other persons with the same name](#)

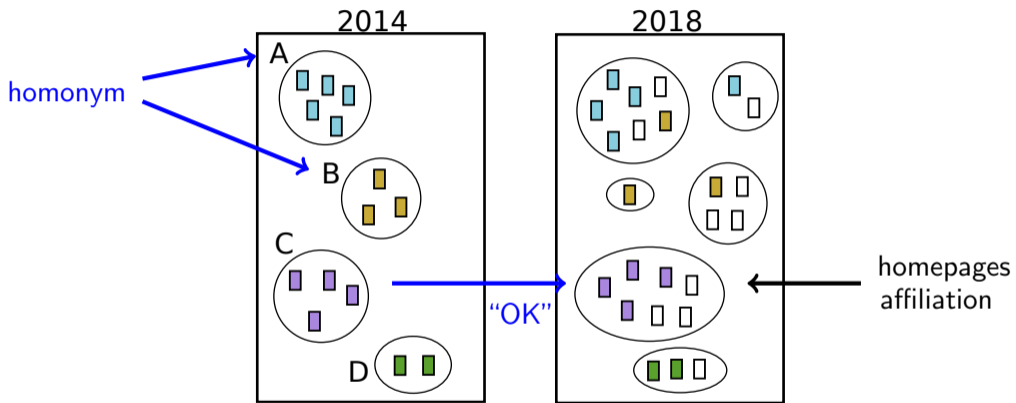
- Wei Wang ⁰⁰⁰¹ — University of Waterloo, David R. Cheriton School of Computer Science, ON, Canada
- Wei Wang ⁰⁰⁰²  — Nanjing University, State Key Laboratory for Novel Software Technology, Nanjing, China; National University of Singapore, ECE Department, Singapore
- Wei Wang ⁰⁰⁰³ — State University 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.

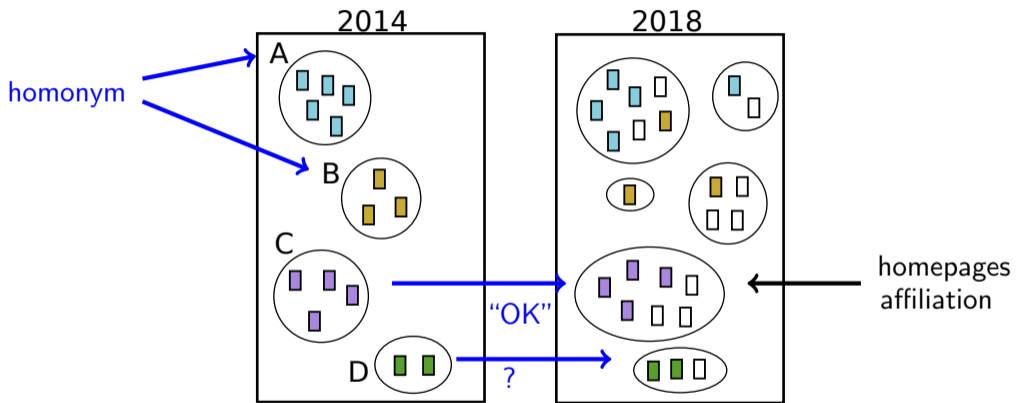
Task: find profiles that are most likely
homonyms.

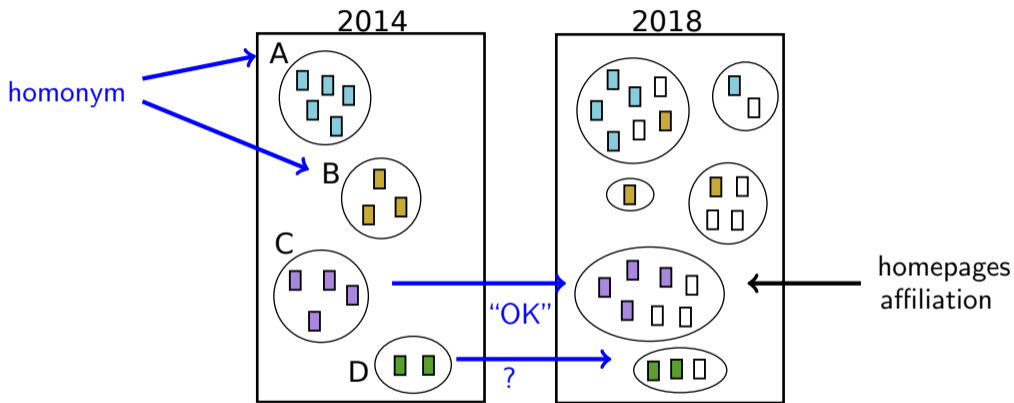






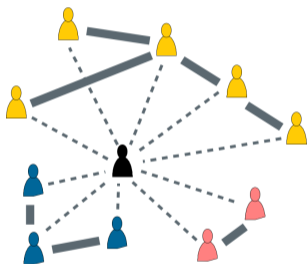




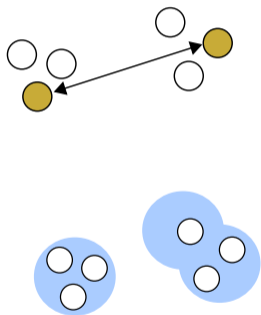


homonyms: 2802
 no homonyms: 21576

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	

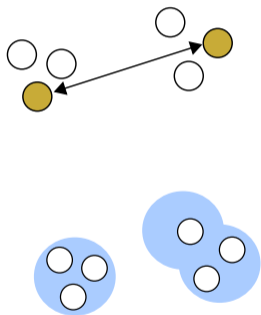


- ▶ #Coauthor groups
- ▶ Size of 5 largest cluster
- ▶ Unevenness of size distribution



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

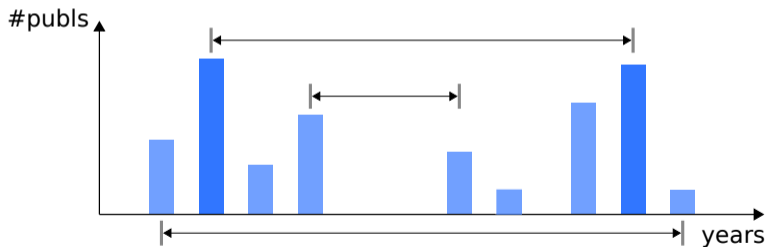
- ▶ Diameter
- ▶ Greedy cover number
- ▶ 5/25/50/75/95 percentile
 - ▶ Pairwise distance
 - ▶ Distance to centroid



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

- ▶ Diameter
- ▶ Greedy cover number
- ▶ 5/25/50/75/95 percentile
 - ▶ Pairwise distance
 - ▶ Distance to centroid
- ▶ #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	0.130 ± 0.06	0.799 ± 0.01
BC	0.818 ± 0.17	0.057 ± 0.02	0.106 ± 0.03	0.197 ± 0.05	0.842 ± 0.01
BT	0.542 ± 0.18	0.051 ± 0.03	0.092 ± 0.05	0.138 ± 0.06	0.786 ± 0.01
BV	0.745 ± 0.04	0.232 ± 0.05	0.350 ± 0.07	0.372 ± 0.06	0.815 ± 0.01
BY	0.781 ± 0.02	0.153 ± 0.01	0.256 ± 0.02	0.314 ± 0.02	0.820 ± 0.00
BTV	0.709 ± 0.01	0.268 ± 0.01	0.389 ± 0.02	0.393 ± 0.01	0.832 ± 0.00
All	0.793 ± 0.01	0.424 ± 0.01	0.552 ± 0.01	0.541 ± 0.01	0.890 ± 0.00

Person	Score	#Pe	#CaA	#CoR	#Vim	#SCC	H	diam	cover	d(x,y)	d(x,c)	v(diam,ycover)	v(d(x,y))	v(d(x,c))	years	ys	yag	ym	
Xia Zhou (276)	0.9949	210	655	7316	142	27	0.31	1.57	171	0.99	0.88	1.20	69	0.93	0.67	30	21	10	8
Xiao Jiang (276)	0.9855	225	384	956	155	24	0.44	1.55	179	0.99	0.88	1.22	63	0.92	0.66	16	17	1	2
A.A. Dogov (2)	0.9815	4	5	2	2	0.81	0.90	2	0.90	0.29	1.10	2	1.10	0.33	40	2	40	46	
M. E. Mitchell (2)	0.9811	2	5	6	2	2	0.72	0.90	2	0.96	0.30	1.07	2	1.07	0.32	47	2	47	47
Perthard David (2)	0.9810	2	3	1	2	2	0.92	0.93	2	0.93	0.27	1.03	2	1.03	0.33	52	2	52	52
J. Underwood (2)	0.9797	2	6	10	2	2	0.65	0.98	2	0.98	0.28	1.11	2	1.11	0.26	20	2	20	25
Jing Zhang (273)	0.9796	213	440	1670	130	23	0.37	1.53	154	0.99	0.85	1.15	59	0.92	0.62	24	21	4	7
Yang Zhang (273)	0.9796	321	495	1682	147	24	0.30	1.53	231	0.98	0.82	1.19	66	0.93	0.66	18	19	1	5
E. Baigun (2)	0.9794	5	4	2	2	0.97	1.04	2	1.04	0.36	1.09	2	1.09	0.33	35	2	35	35	
M. Diao (2)	0.9789	2	7	9	2	2	0.99	0.89	2	0.89	0.28	1.09	2	1.09	0.34	31	2	31	31
Y. Sakamoto (2)	0.9783	2	15	55	2	2	0.92	0.77	2	0.77	0.22	1.03	2	1.03	0.32	32	2	32	32
E. Wilson (2)	0.9778	2	7	9	2	2	0.99	0.87	2	0.87	0.28	1.03	2	1.03	0.28	43	2	43	43
E. Hogg (2)	0.9762	2	26	105	2	2	0.96	0.72	2	0.72	0.21	1.01	2	1.01	0.30	26	2	26	26
Karoline Wood (2)	0.9762	2	3	1	2	2	0.92	0.91	2	0.91	0.26	1.06	2	1.06	0.39	21	2	21	31
E. Lund (2)	0.9760	2	18	100	2	2	0.65	1.04	2	1.04	0.31	0.98	2	0.98	0.30	45	2	45	45
Huang Di Sun (2)	0.9758	2	11	25	2	2	0.99	0.92	2	0.92	0.27	1.22	2	1.22	0.38	12	2	12	12
J. Bartlett (2)	0.9756	2	19	24	2	2	0.88	1.12	2	1.12	0.35	1.09	2	1.09	0.33	34	2	34	34
Y. Chen (2)	0.9754	42	243	1438	37	35	0.89	1.00	0.98	0.85	1.13	39	0.94	0.68	26	19	6	3	
Robert J. Martin (2)	0.9752	2	2	1	2	1	0.80	1.02	2	1.02	0.31	1.03	2	1.03	0.36	51	2	51	51
Mohamed Elshorbagy (2)	0.9752	2	3	1	2	2	0.92	0.96	2	0.96	0.30	1.04	2	1.04	0.31	37	2	37	37
Natalia Dine (2)	0.9747	2	15	49	2	2	1.00	0.65	2	0.65	0.18	1.16	2	1.16	0.33	16	2	16	16
C. Wood (2)	0.9747	2	6	6	2	2	1.00	1.09	2	1.09	0.33	1.06	2	1.06	0.31	33	2	33	33
R. P. Wood (2)	0.9743	2	3	1	2	2	0.92	0.78	2	0.78	0.22	1.03	2	1.03	0.31	36	2	36	36
R. Wagner (2)	0.9743	2	3	1	2	2	0.92	0.89	2	0.89	0.28	1.03	2	1.03	0.37	32	2	32	32
Jeffrey Moore (2)	0.9742	2	6	7	2	2	0.92	0.87	2	0.87	0.25	1.00	2	1.00	0.30	44	2	44	44
Kotenko Nikolayev (2)	0.9739	2	19	81	2	2	1.00	0.83	2	0.83	0.24	0.88	2	0.88	0.26	54	2	54	54
Charles R. Parker (2)	0.9738	2	9	16	2	2	0.89	0.98	2	0.98	0.31	1.09	2	1.09	0.30	37	2	37	37

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