Linking Records in a Dynamic World

Pei Li

University of Milan – Bicocca

Joint work w. Xin Luna Dong, Andrea Maurino, Divesh Srivastava

Some Statistics from DBLP*

- Top 10 authors with most number of papers
  - Wei Wang (476 papers)
- Top 5 authors with most number of co-authors
  - Wei Wang (656 co-authors)
- Top 10 authors with most number of conference papers within the same year
  - Wei Wang (75 conf. papers in 2006)

*http://www2.research.att.com/~marioh/dblp.html

(last updated on March 13th 2009)

Some Statistics from DBLP

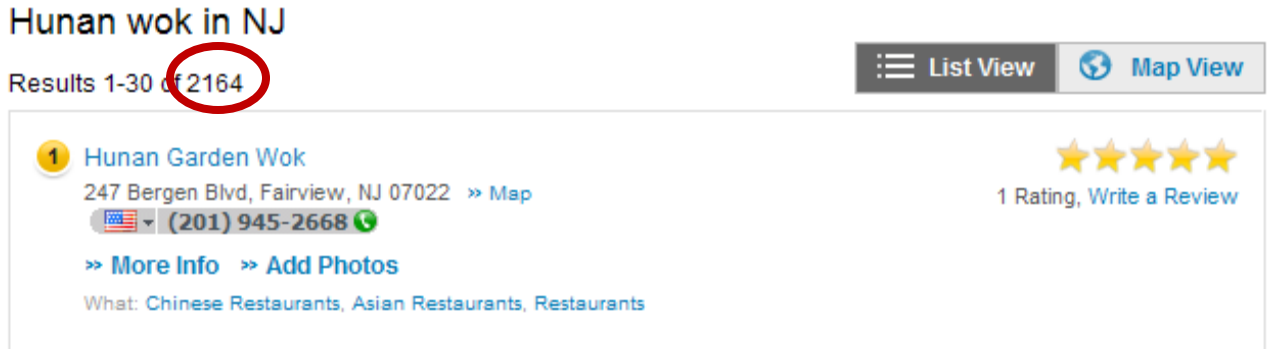
- Wei Wang - Language Weaver, Inc.
- Wei Wang - The Chinese University of Hong Kong, Mechanical and Automation Engineering
- Wei Wang - IBM China Research Laboratory, Haidian District, Beijing
- Wei Wang - Center for Engineering and Scientific Computation, Zhejiang University, China
- Wei Wang - Fudan University, Shanghai, China
- Wei Wang - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
- Wei Wang - University of North Carolina at Chapel Hill

-How many Wei Wang's are there?
-What are their authoring histories?

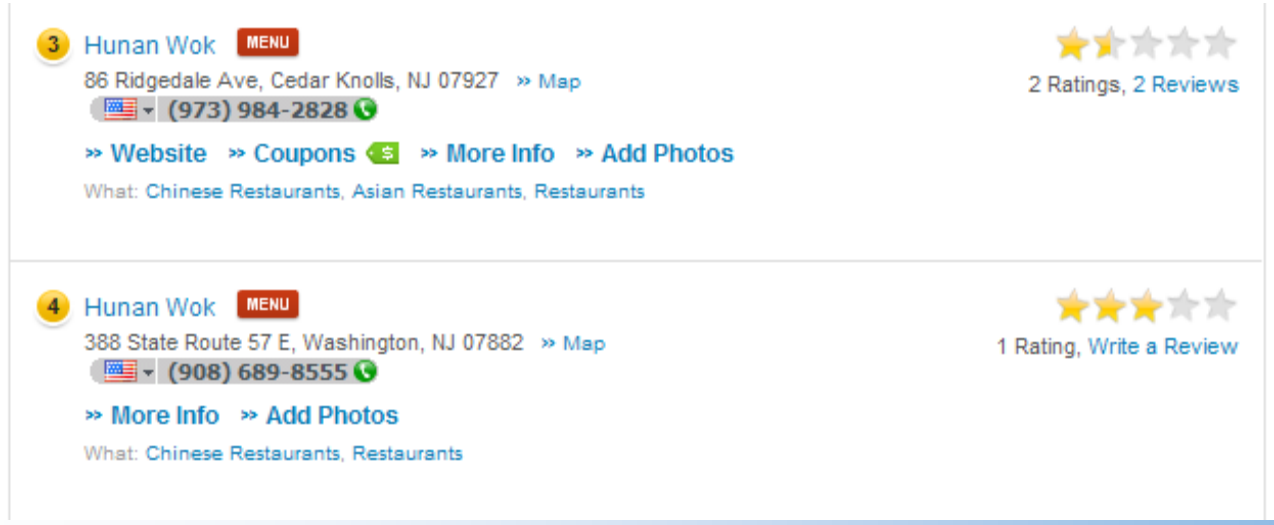
Ask

550	Wei Wang, Tun Lu, Tiejiang Liu, Qi Wang, Ning Gu: Service deployment optimization oriented collaborative dynamic double buffer pool in clouds. CSCWD 2011: 408-415
549	Changqing Liu, Yingguang Li, Wei Wang, Weiming Shen: A feature-based NC machining time forecasting model. CSCWD 2011: 593-598
548	Yi Chen, Wei Wang, Ziyang Liu: Searching, Analyzing and Exploring Databases. DASFAA (2) 2011: 467-469
547	Qianlu Lin, Chuan Xiao, Muhammad Aamir Cheema, Wei Wang: Finding the Sites with Best Accessibilities to Amenities. DASFAA (2) 2011: 58-72
546	H. Howie Huang, Nan Zhang, Wei Wang, Gautam Das, Alexander S. Szalay: Just-in-Time Analytics on Large File Systems. FAST 2011: 217-230
545	Yi Chen, Wei Wang, Ziyang Liu: Keyword-based search and exploration on databases. ICDE 2011: 1380-1383

Some Statistics from YellowPages



-Are there any business chains?
-If yes, which businesses are their members?



Record Linkage

- **Record linkage** takes a set of records as input and discovers which records refer to the same real-world entity.

- Existing record-linkage techniques (surveyed in [Elmagarmid, 07], [Koudas, 06])
  - Focus on different representations of the same value
  - E.g., *IBM* vs. *International Business Machines*

Diversity in a Dynamic World

- In reality, we observe value diversity of entities
  - Values can evolve over **time**
    - Catholic Healthcare (1986 - 2012) → Dignity Health (2012 -)
  - **Different members** of the same group can have diversity

ID	Name	Address	Phone	URL
001	F.B. Insurance	Vernon 76384 TX	877 635-4684	txfb-ins.com
002	F.B. Insurance #1	Lufkin 75901 TX	936 634-7285	txfb.org
003	F.B. Insurance #5	Cibolo 78108 TX	877 635-4684	

  - Some sources may provide **erroneous data**

ID	Name	URL	Source
001	Meekhof Tire Sales & Service Inc	www.meekhoftire.com	Src. 1
002	Meekhof Tire Sales & Service Inc	www.napaautocare.com	Src. 2

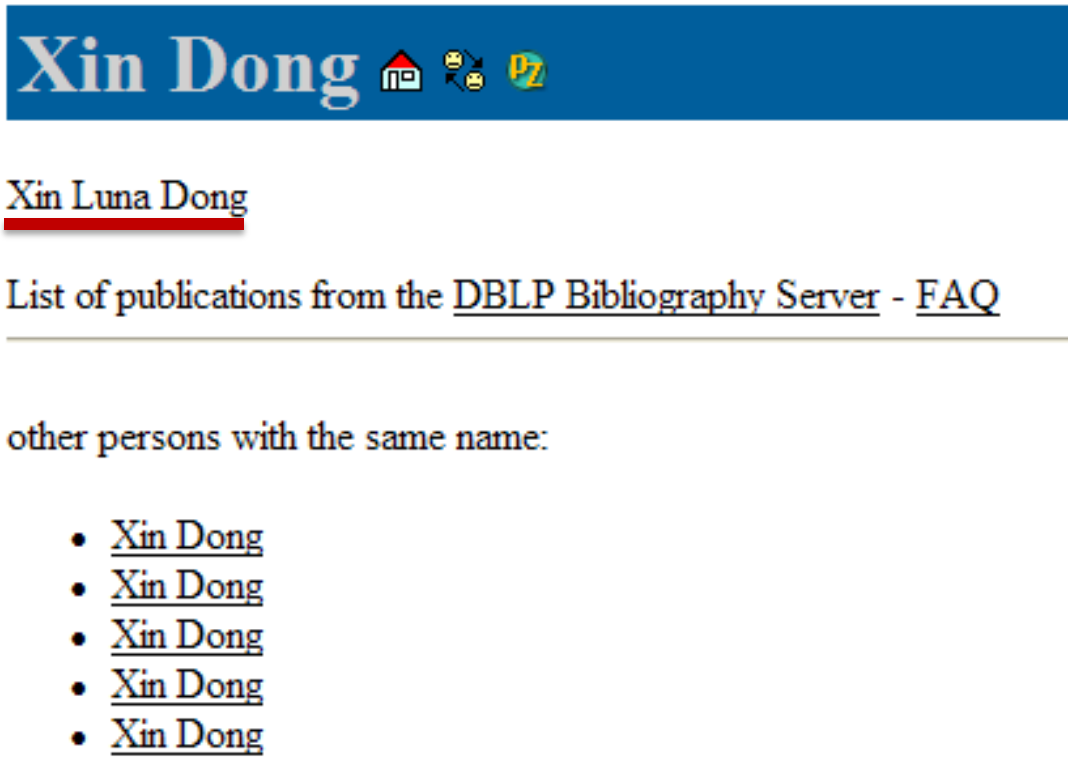
Diversity in a Dynamic World

- Record linkage in a dynamic world
  - Tolerance to high diversity of values
    - over time **- linking temporal records**
    - among different members of the same group **- linking group members**

Linking Temporal Records

Real-life Stories from Luna (I)

- Luna's DBLP entry



Xin Dong

Xin Luna Dong

List of publications from the DBLP Bibliography Server - FAQ

other persons with the same name:

- Xin Dong
- Xin Dong
- Xin Dong
- Xin Dong
- Xin Dong

Real-life Stories from Luna (II)

	2011
36	Anish Das Sarma, Xin Luna Dong, Alon Y. Halevy: Data integration with dependent sources. EDBT 2011: 401-412
35	Xin Luna Dong, Divesh Srivastava: Large-scale copy detection. SIGMOD Conference 2011: 1205-1208
34	Su Chen, Xin Luna Dong, Laks V. S. Lakshmanan, Divesh Srivastava: We challenge you to certify your updates. SIGMOD Conference 2011: 481-492
33	Xiaoping Sun, Xin Dong: Special Issue: Semantics, Knowledge and Grids. Concurrency and Computation: Practice and Experience 23(9): 863-865 (2011)
	2010
32	Xin Luna Dong, Felix Naumann: Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA, June 6, 2010 WebDB 2010
31	Alvaro Pinto, Zhe Zhang, Xin Dong, Senem Velipasalar, Mehmet Can Vuran, Mustafa Cenk Gursoy: Energy Consumption and Latency Analysis for Wireless Multimedia Sensor Networks. GLOBECOM 2010: 1-5
30	Xin Dong, Mehmet Can Vuran: Vision Graph Construction in Wireless Multimedia Sensor Networks. GLOBECOM 2010: 1-5
29	Xin Dong, Laure Berti-Equille, Yifan Hu, Divesh Srivastava: Global Detection of Complex Copying Relationships Between Sources. PVLDB 3(1): 1358-1369 (2010)
28	Songtao Guo, Xin Dong, Divesh Srivastava, Remi Zajac: Record Linkage with Uniqueness Constraints and Erroneous Values. PVLDB 3(1): 417-428 (2010)
27	Xin Dong, Laure Berti-Equille, Yifan Hu, Divesh Srivastava: SOLOMON: Seeking the Truth Via Copying Detection. PVLDB 3(2): 1617-1620 (2010)
26	Xin Luna Dong, Felix Naumann: 13th international workshop on the web and databases: WebDB 2010. SIGMOD Record 39(3): 37-39 (2010)

Real-life Stories from Luna (III)

- Lab visiting

Dong Xin

List of publications from the DBLP Bibliography Server - FAQ

Ask others: ACM DL/Guide - CSB - MetaPress - Google - Bing - Yahoo

		2011
40	Bib TeX ML	Peixiang Zhao, Xiaolei Li, Dong Xin, Jiawei Han: Graph cube: on warehousing and OLAP multidimensional networks. SIGMOD Conference 2011: 853-864
39	Bib TeX ML	Bahman Bahmani, Kaushik Chakrabarti, Dong Xin: Fast personalized PageRank on MapReduce. SIGMOD Conference 2011: 973-984
38	Bib TeX ML	Yeye He, Dong Xin: SEISA: set expansion by iterative similarity aggregation. WWW 2011: 427-436
37	Bib TeX ML	Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng, Dong Xin: EntityTagger: automatically tagging entities with descriptive phrases. WWW (Companion Volume) 2011: 19-20
		2010
36	Bib TeX ML	Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti, Arnd Christian König, Dong Xin: Query portals: dynamically generating portals for entity-oriented web queries. SIGMOD Conference 2010: 1171-1174
35	Bib TeX ML	Dong Xin, Yeye He, Venkatesh Ganti: Keyword++: A Framework to Improve Keyword Search Over Entity Databases. PVLDB 3 (1): 711-722 (2010)

r1: Xin Dong
R. Polytechnic Institute

r2: Xin Dong
University of Washington

r3: Xin Dong
University of Washington

r4: Xin Luna Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r6: Xin Luna Dong
AT&T Labs-Research

-How many authors?

-What are their authoring histories?

1991 2004 2005 2006 2007 2008 2009 2010 2011

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r1: Xin Dong
R. Polytechnic Institute

r4: Xin Luna Dong
University of Washington

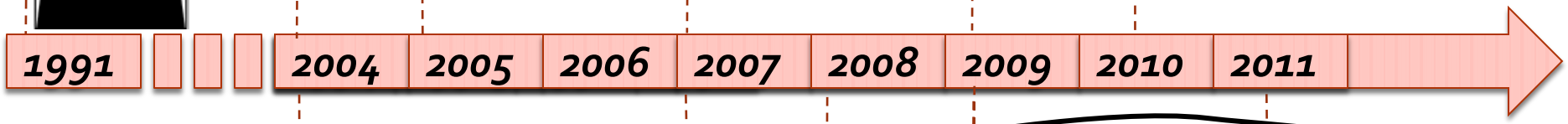
r2: Xin Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r3: Xin Dong
University of Washington

r6: Xin Luna Dong
AT&T Labs-Research

-Ground Truth



3 authors

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r1: Xin Dong
R. Polytechnic Institute

r4: Xin Luna Dong
University of Washington

r2: Xin Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r3: Xin Dong
University of Washington

r6: Xin Luna Dong
AT&T Labs-Research

-Solution 1:

-requiring high value consistency

1991 2004 2005 2006 2007 2008 2009 2010 2011

5 authors
false negative

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

r1: Xin Dong
R. Polytechnic Institute

r4: Xin Luna Dong
University of Washington

r2: Xin Dong
University of Washington

r5: Xin Luna Dong
AT&T Labs-Research

r3: Xin Dong
University of Washington

r6: Xin Luna Dong
AT&T Labs-Research

-Solution 2:
-matching records w. similar names

1991 2004 2005 2006 2007 2008 2009 2010 2011

2 authors
false positive

r11: Dong Xin
Microsoft Research

r8:Dong Xin
University of Illinois

r12: Dong Xin
Microsoft Research

r9: Dong Xin
Microsoft Research

r10: Dong Xin
University of Illinois

r7: Dong Xin
University of Illinois

Opportunities

Continuity of history

Smooth transition

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Seldom erratic changes

Intuitions

ID	Name	Affiliation	Co-authors	Year
r1	Xin Dong	R. Polytechnic Institute	Wozny	1991
r2	Xin Dong	University of Washington	Halevy, Tatarinov	2004
r7	Dong Xin	University of Illinois	Han, Wah	2004
r3	Xin Dong	University of Washington	Halevy	2005
r4	Xin Luna Dong	University of Washington	Halevy, Yu	2007
r8	Dong Xin	University of Illinois	Wah	2007
r9	Dong Xin	Microsoft Research	Wu, Han	2008
r10	Dong Xin	University of Illinois	Ling, He	2009
r11	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
r5	Xin Luna Dong	AT&T Labs-Research	Das Sarma, Halevy	2009
r6	Xin Luna Dong	AT&T Labs-Research	Naumann	2010
r12	Dong Xin	Microsoft Research	He	2011

Less reward on the same value over time

Less penalty on different values over time

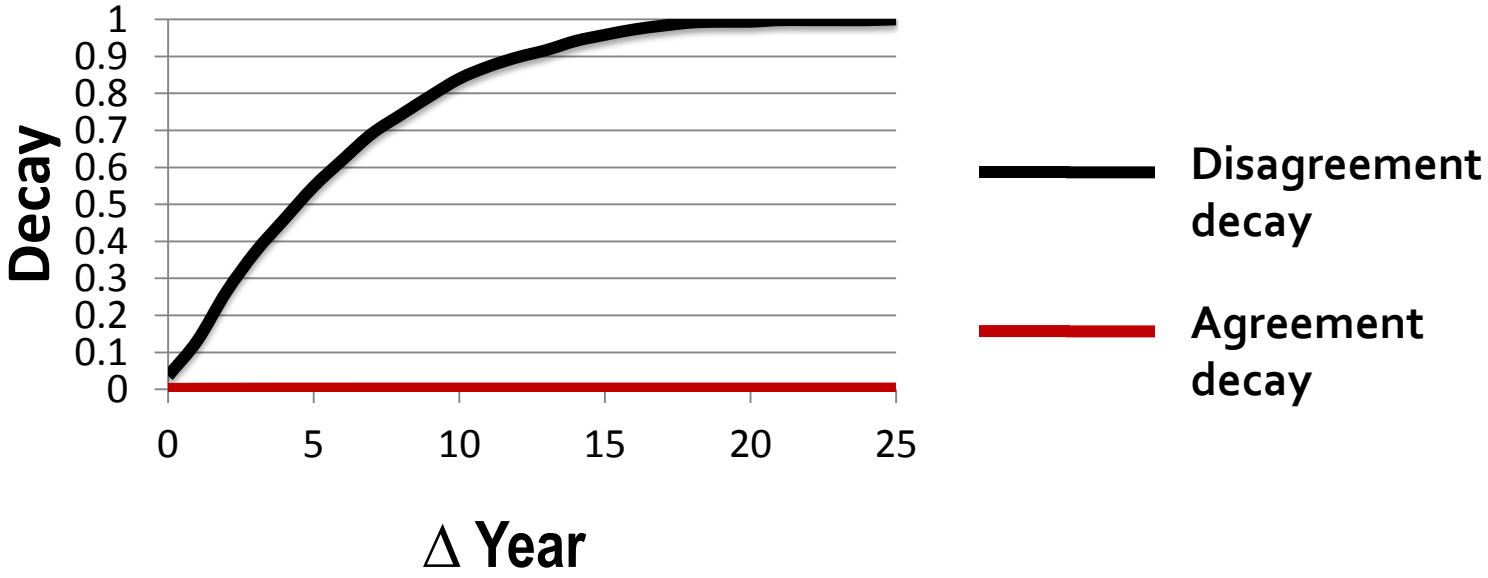
Consider records in time order for clustering

Problem Statement

- Input: a set of records R, in the form of (x_1, …, x_n, t)
  - t: time stamp
  - x_i: value of attribute A_i at time t
- Output: clustering of R such that
  - records in the same cluster refer to the same entity
  - records in different clusters refer to different entities

Overview of Our Solution

- Apply **time decay** in record similarity
  - Decay allows tolerance on value evolution
  - E.g. Decay of address learnt from European Patent data



- Consider **time order** of records in clustering
  - Accumulate evidence over time and make global decisions

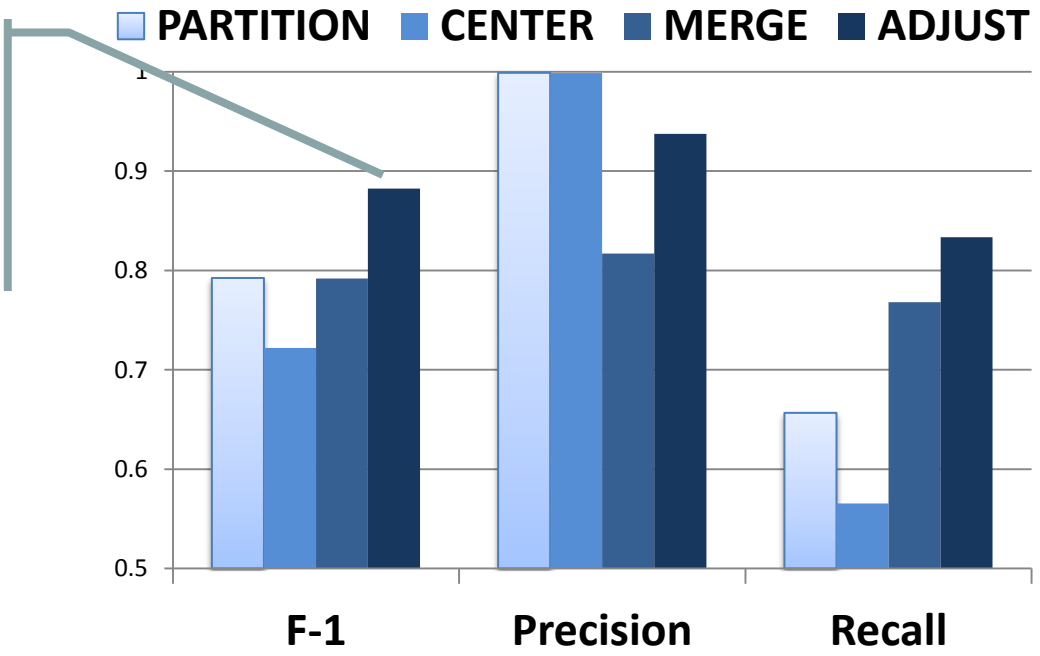
Experiment Setting

- Implementation
  - Baseline: PARTITION, CENTER, MERGE
  - Our approaches: EARLY, LATE, ADJUST
- Comparison: Precision/Recall/F-measure
  - Precision = |TP|/(|TP|+|FP|)
  - Recall =|TP|/(|TP|+|FN|)
  - F-measure = 2PR/(P+R)

Accuracy on Patent Data

- Data set: a benchmark of European patent data set
  - 1871 records, 359 entities, in 1978-2003
  - Compare name & affiliation
- Golden standard: http://www.esf-ape-inv.eu/

Adjust improves over baseline by 11-22%

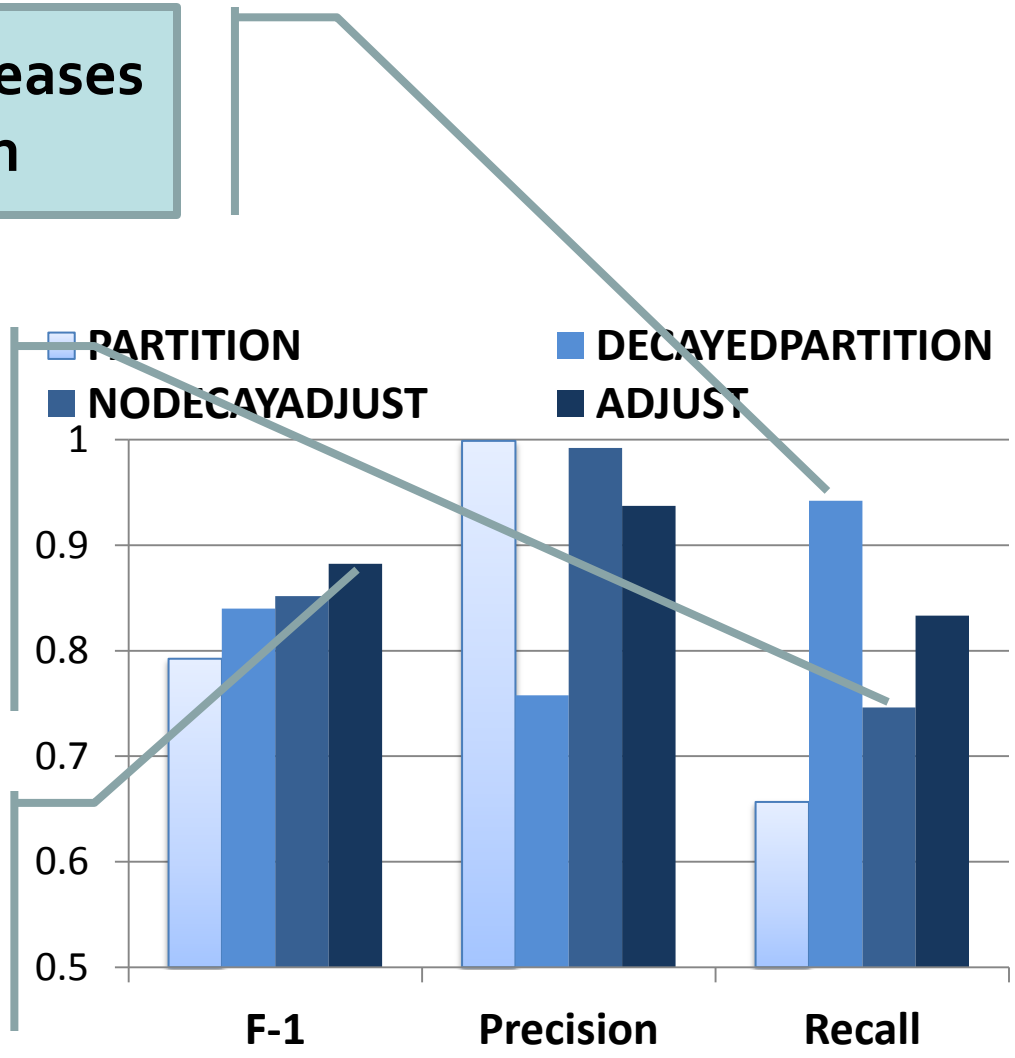


Contribution of Decay and Temporal Clustering

Applying decay in itself increases recall by sacrificing precision

Temporal clustering increases recall moderately without reducing precision much

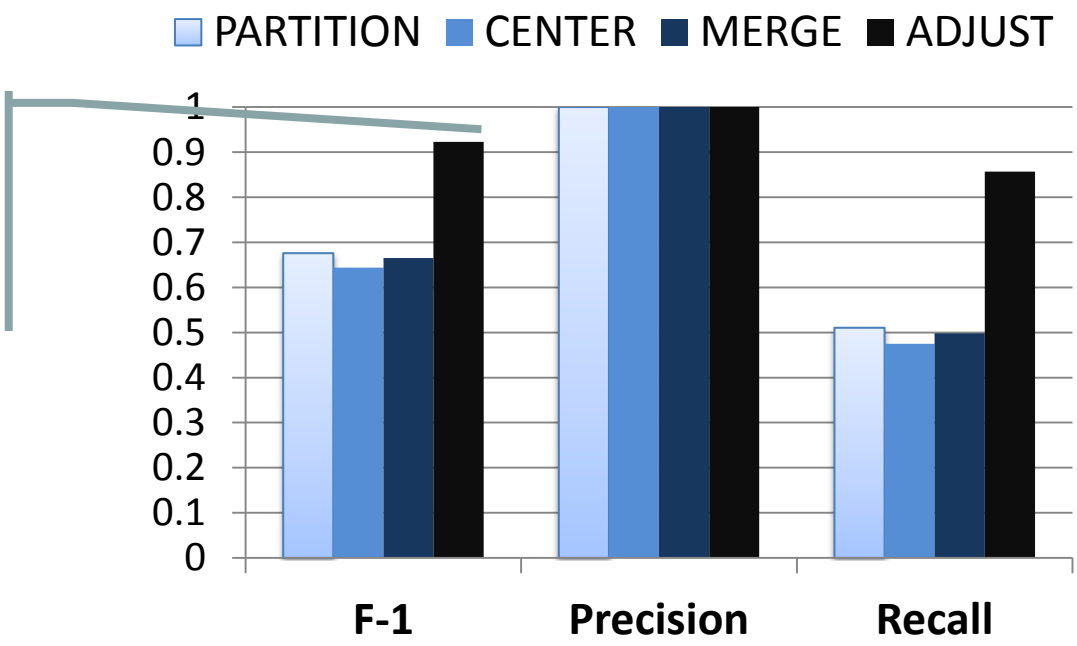
Combining both obtains the best results



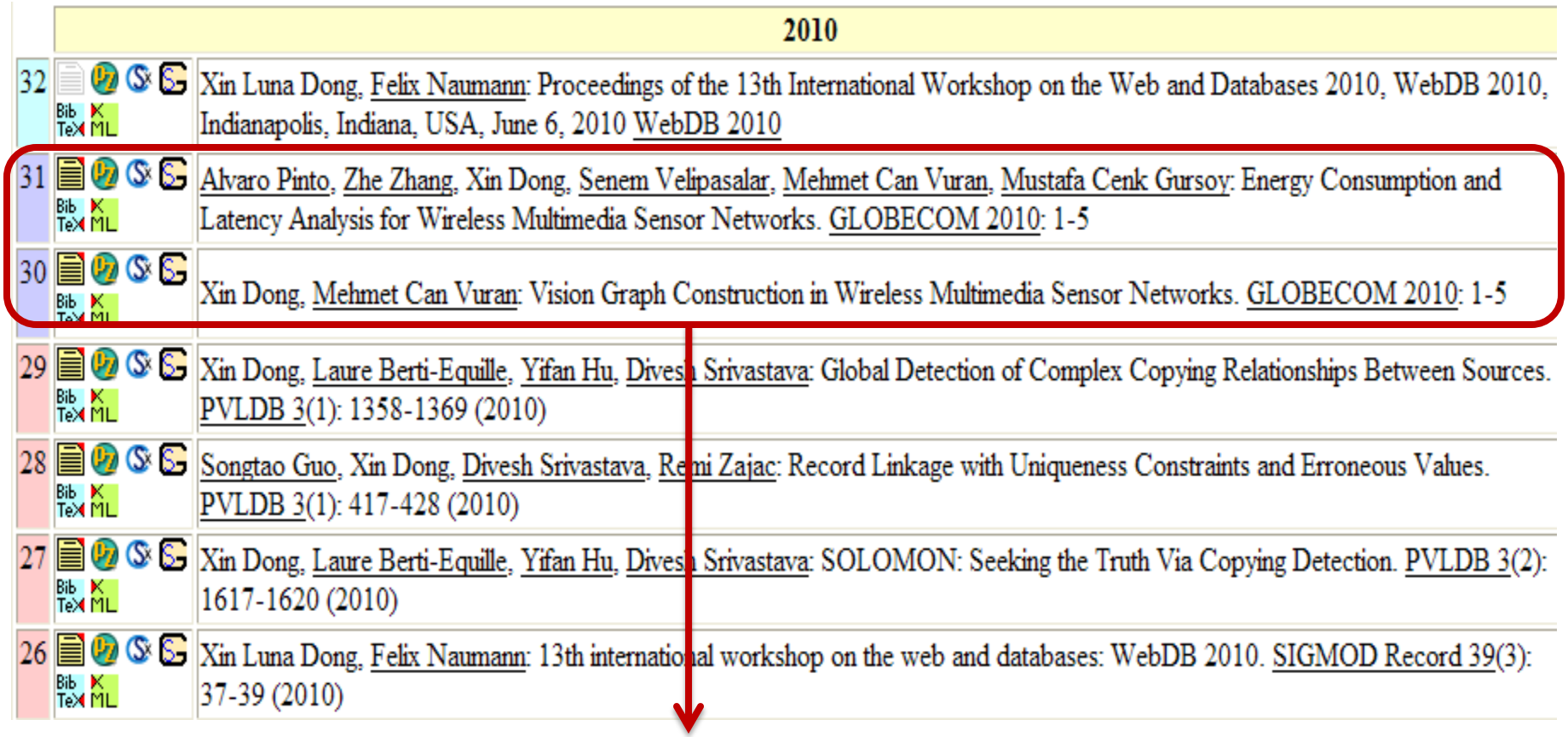
Accuracy on DBLP Data – Xin Dong

- Data set: Xin Dong data set from DBLP
  - 72 records, 8 entities, in 1991-2010
  - Compare name, affiliation, title & co-authors
- Golden standard: by manually checking

Adjust improves over baseline by 37-43%



Error We Fixed



2010

32 Xin Luna Dong, Felix Naumann: Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA, June 6, 2010 WebDB 2010

31 Alvaro Pinto, Zhe Zhang, Xin Dong, Senem Velipasalar, Mehmet Can Vuran, Mustafa Cenk Gursoy: Energy Consumption and Latency Analysis for Wireless Multimedia Sensor Networks. GLOBECOM 2010: 1-5

30 Xin Dong, Mehmet Can Vuran: Vision Graph Construction in Wireless Multimedia Sensor Networks. GLOBECOM 2010: 1-5

29 Xin Dong, Laure Berti-Equille, Yifan Hu, Divesh Srivastava: Global Detection of Complex Copying Relationships Between Sources. PVLDB 3(1): 1358-1369 (2010)

28 Songtao Guo, Xin Dong, Divesh Srivastava, Remi Zajac: Record Linkage with Uniqueness Constraints and Erroneous Values. PVLDB 3(1): 417-428 (2010)

27 Xin Dong, Laure Berti-Equille, Yifan Hu, Divesh Srivastava: SOLOMON: Seeking the Truth Via Copying Detection. PVLDB 3(2): 1617-1620 (2010)

26 Xin Luna Dong, Felix Naumann: 13th international workshop on the web and databases: WebDB 2010. SIGMOD Record 39(3): 37-39 (2010)

Records with affiliation University of Nebraska–Lincoln

We Only Made One Mistake

2010

36 Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti, Arnd Christian König, Dong Xin: Query portals: dynamically generating portals for entity-oriented web queries. SIGMOD Conference 2010: 1171-1174

35 Dong Xin, Yeye He, Venkatesh Ganti: Keyword++: A Framework to Improve Keyword Search Over Entity Databases. PVLDB 3 (1): 711-722 (2010)

2009

34 Surajit Chaudhuri, Venkatesh Ganti, Dong Xin: Exploiting web search to generate synonyms for entities. WWW 2009: 151-160

33 Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti, Arnd Christian König, Dong Xin: Exploiting web search engines to search structured databases. WWW 2009: 501-510

32 Xu Ling, Xin He, Dong Xin: Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics 25 (5): 571-577 (2009)

31 Hongyan Liu, Xiaoyu Wang, Jun He, Jiawei Han, Dong Xin, Zheng Shao: Top-down mining of frequent closed patterns from very high dimensional data. Inf. Sci. 179(7): 899-924 (2009)

30 Tianyi Wu, Dong Xin, Qiaozhu Mei, Jiawei Han: Promotion Analysis in Multi-Dimensional Space. PVLDB 2(1): 109-120 (2009)

29 Surajit Chaudhuri, Venkatesh Ganti, Dong Xin: Mining Document Collections to Facilitate Accurate Approximate Entity Matching. PVLDB 2(1): 395-406 (2009)

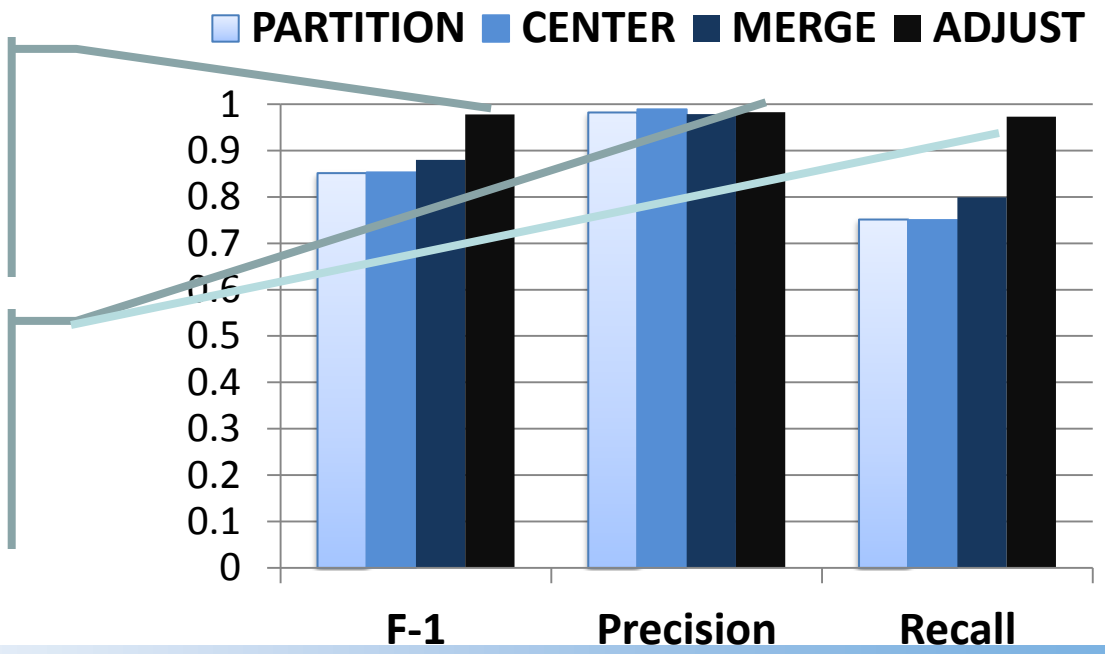
Author's affiliation on Journal papers are out of date

Accuracy on DBLP Data (Wei Wang)

- Data set: Wei Wang data set from DBLP
  - 738 records, 18 entities + potpourri, in 1992-2011
  - Compare name, affiliation & co-authors
- Golden standard: from DBLP + manually checking

Adjust improves over baseline by 11-15%

High precision (.98) and high recall (.97)



Mistakes We Made

Wei Wang

List of publications from the DBLP Bibliography Server - FAQ

other persons with the same name:

- Wei Wang - School of Life Science, Fudan University, China
- Wei Wang - Nonlinear Systems Laboratory, Department of Mechanical Engineering, MIT
- Wei Wang - University of Maryland Baltimore County
- Wei Wang - University of Naval Engineering
- Wei Wang - ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences
- Wei Wang - Rutgers University, New Brunswick, NJ, USA
- Wei Wang - Purdue University Indianapolis
- Wei Wang - INRIA Sophia Antipolis, Sophia Antipolis, France
- Wei Wang - Institute of Computational Linguistics, Peking University
- Wei Wang - National University of Singapore
- Wei Wang - Nanyang Technological University, Singapore
- Wei Wang - Computer and Electronics Engineering, University of Nebraska Lincoln, NE, USA
- Wei Wang - The University of New South Wales, Australia
- Wei Wang - Language Weaver, Inc.
- Wei Wang - The Chinese University of Hong Kong, Mechanical and Automation Engineering
- Wei Wang - Center for Engineering and Scientific Computation, Zhejiang University, China
- Wei Wang - Fudan University, Shanghai, China
- Wei Wang - University of North Carolina at Chapel Hill

1 record @ 2006

72 records @ 2000-2011

Mistakes We Made

2006

10 Sining Liu, Brian King, Wei Wang: A CRT-RSA Algorithm Secure against Hardware Fault Attacks. DASC 2006: 51-60

9 Sining Liu, F. Bowen, Brian King, Wei Wang: Elliptic curves cryptosystem implementation based on a look-up table sharing scheme. ISCAS 2006

Purdue University

2005

8 Shaoqiang Bi, Warren J. Gross, Wei Wang, Asim J. Al-Khalili, M. N. S. Swamy: An Area-Reduced Scheme for Modulo 2n-1 Addition/Subtraction. IWSOC 2005: 396-399

7 Wei Wang, Xiaolin Zhang, Chenyang Yang, M. N. S. Swamy, M. Omair Ahmad: RRNS Quasi-Chaotic Coding and Its FPGA Implementation. SNPD 2005: 274-280

Univ. of Western Ontario

2004

6 Shaoqiang Bi, Wei Wang, Asim J. Al-Khalili: Modulo deflation in $(2^n+1, 2^n, 2^n-1)$ converters. ISCAS (2) 2004: 429-432

5 Wei Wang, M. N. S. Swamy, M. Omair Ahmad: RNS Application for Digital Image Processing. IWSOC 2004: 77-80

4 Wei Wang, M. N. S. Swamy, M. Omair Ahmad: Novel Design and Fpga Implementation of Da-rns Fir Filters. Journal of Circuits, Systems, and Computers 13(6): 1233-1250 (2004)

2003

3 Wei Wang, M. N. S. Swamy, M. Omair Ahmad: Moduli selection in RNS for efficient VLSI implementation. ISCAS (4) 2003: 512-515

2002

2 Wei Wang, M. N. S. Swamy, M. Omair Ahmad: A new architecture of RRNS error-correcting QC encoder/decoder and its FPGA implementation. ISCAS (5) 2002: 813-816

Concordia University

Errors We Fixed … despite some mistakes

- 546 records in potpourri
  - Correctly merged 63 records to existing Wei Wang entries
  - Wrongly merged 61 records
    - 26 records: due to missing department information
    - 35 records: due to high similarity of affiliation
      - E.g., Northwest University of Science & Technology
        Northeast University of Science & Technology
  - Precision and recall of .94 w. consideration of these records

Linking Group Members

1 **Taco Casa** MENU
619 McFarland Blvd, Northport, AL 35476 » Map
(205) 339-5977
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants

2 Taco Casa MENU
603 15th St E, Tuscaloosa, AL 35401 » Map
(205) 345-0751
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants

4 Taco Casa MENU
4629 Highway 280, Birmingham, AL 35242 » Map
(205) 980-9699
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants

5 Taco Casa MENU
1060 Fairfax Park, Tuscaloosa, AL 35406 » Map
(205) 752-0990
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants

6 Elva's Taco Casa MENU
807 Royalty Ave, Odessa, TX 79761 » Map
(432) 333-2831
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants

7 Taco Casa MENU
701 W California St, Gainesville, TX 76240 » Map
(940) 665-7477
» Website » More Info » Add Photos
What: Mexican Restaurants, Fast Food Restaurants, Restaurants

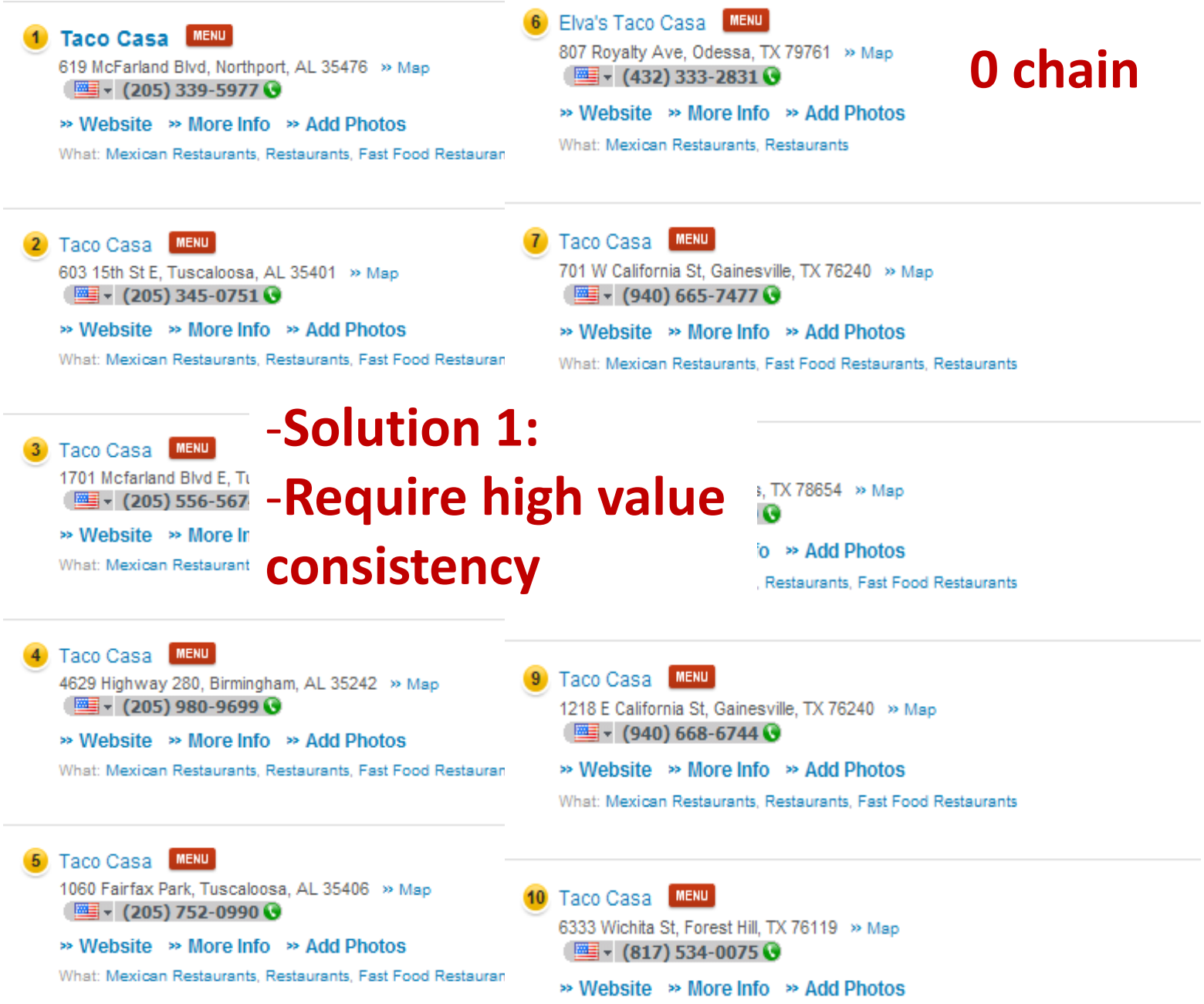
9 Taco Casa MENU
1218 E California St, Gainesville, TX 76240 » Map
(940) 668-6744
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants

10 Taco Casa MENU
6333 Wichita St, Forest Hill, TX 76119 » Map
(817) 534-0075
» Website » More Info » Add Photos

-Are there any business chains?

-If yes, which businesses are their members?


1 Taco Casa MENU
619 McFarland Blvd, Northport, AL 35476 » Map
(205) 339-5977
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
2 Taco Casa MENU
603 15th St E, Tuscaloosa, AL 35401 » Map
(205) 345-0751
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
3 Taco Casa MENU
1701 Mcfarland Blvd E, Tuscaloosa, AL 35404 » Map
(205) 556-5674
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
4 Taco Casa MENU
4629 Highway 280, Birmingham, AL 35242 » Map
(205) 980-9699
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
5 Taco Casa MENU
1060 Fairfax Park, Tuscaloosa, AL 35406 » Map
(205) 752-0990
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
6 Elva's Taco Casa MENU
807 Royalty Ave, Odessa, TX 79761 »
(432) 333-2831
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants
2 chains
7 Taco Casa MENU
701 W California St, Gainesville, TX 76240 » Map
(940) 665-7477
» Website » More Info » Add Photos
What: Mexican Restaurants, Fast Food Restaurants, Restaurants
8 Taco Casa MENU
1400 Fm 1431, Marble Falls, TX 78654 » Map
(830) 693-7789
» Website » More Info » Add Photos
Restaurants, Restaurants, Fast Food Restaurants
-Ground Truth
9 Taco Casa MENU
1218 E California St, Gainesville, TX 76240 » Map
(940) 668-6744
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
10 Taco Casa MENU
6333 Wichita St, Forest Hill, TX 76119 » Map
(817) 534-0075
» Website » More Info » Add Photos


1 Taco Casa MENU
619 McFarland Blvd, Northport, AL 35476 » Map
(205) 339-5977
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
6 Elva's Taco Casa MENU
807 Royalty Ave, Odessa, TX 79761 » Map
(432) 333-2831
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants
0 chain
2 Taco Casa MENU
603 15th St E, Tuscaloosa, AL 35401 » Map
(205) 345-0751
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
7 Taco Casa MENU
701 W California St, Gainesville, TX 76240 » Map
(940) 665-7477
» Website » More Info » Add Photos
What: Mexican Restaurants, Fast Food Restaurants, Restaurants
-Solution 1:
-Require high value consistency
3 Taco Casa MENU
1701 Mcfarland Blvd E, Tu
(205) 556-567
» Website » More In
What: Mexican Restaurant
, TX 78654 » Map
o » Add Photos
, Restaurants, Fast Food Restaurants
4 Taco Casa MENU
4629 Highway 280, Birmingham, AL 35242 » Map
(205) 980-9699
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
9 Taco Casa MENU
1218 E California St, Gainesville, TX 76240 » Map
(940) 668-6744
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
5 Taco Casa MENU
1060 Fairfax Park, Tuscaloosa, AL 35406 » Map
(205) 752-0990
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
10 Taco Casa MENU
6333 Wichita St, Forest Hill, TX 76119 » Map
(817) 534-0075
» Website » More Info » Add Photos


1 Taco Casa MENU
619 McFarland Blvd, Northport, AL 35476 » Map
(205) 339-5977
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
6 Elva's Taco Casa MENU
807 Royalty Ave, Odessa, TX 79761 » Map
(432) 333-2831
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants
1 chain
2 Taco Casa MENU
603 15th St E, Tuscaloosa, AL 35401 » Map
(205) 345-0751
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
7 Taco Casa MENU
701 W California St, Gainesville, TX 76240 » Map
(940) 665-7477
» Website » More Info » Add Photos
What: Mexican Restaurants, Fast Food Restaurants, Restaurants
3 Taco Ca
1701 Mc
» Webs
What: Mexican Restaurants, Restaurants, Fast Food Restauran
-Solution 2:
-Match records w. same name
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
4 Taco Casa MENU
4629 Highway 280, Birmingham, AL 35242 » Map
(205) 980-9699
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
9 Taco Casa MENU
1218 E California St, Gainesville, TX 76240 » Map
(940) 668-6744
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restaurants
5 Taco Casa MENU
1060 Fairfax Park, Tuscaloosa, AL 35406 » Map
(205) 752-0990
» Website » More Info » Add Photos
What: Mexican Restaurants, Restaurants, Fast Food Restauran
10 Taco Casa MENU
6333 Wichita St, Forest Hill, TX 76119 » Map
(817) 534-0075
» Website » More Info » Add Photos

Challenges

Erroneous values

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Ca	00	AL	tacocasa.com
r3				tacocasa.com, tacocasatexas.com
		00	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

Scalability
6.8M Records

Different local values

Two-Stage Linkage I

- **Stage I: Identify cores** containing listings very likely to belong to the same chain
  - Require strong robustness in presence of possibly erroneous values → Graph theory
  - High Scalability

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Casa	900	AL	tacocasa.com
r3	Taco Casa	900	AL	tacocasa.com, tacocasatexas.com
r4	Taco Casa	900	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

Two-Stage Linkage II

- **Stage II: Cluster** cores and remaining records into chains.
  - Collect strong evidence from cores and leverage in clustering
  - No penalty on local values

Reward strong evidence

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Casa	900	AL	tacocasa.com
r3	Taco Casa	900	AL	tacocasa.com, tacocasatexas.com
r4	Taco Casa	900	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

Two-Stage Linkage II

- **Stage II: Cluster** cores and remaining records into chains.
  - Collect strong evidence from cores and leverage in clustering
  - No penalty on local values

Reward strong evidence

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Casa	900	AL	tacocasa.com
r3	Taco Casa	900	AL	tacocasa.com, tacocasatexas.com
r4	Taco Casa	900	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

Two-Stage Linkage II

- **Stage II: Cluster** cores and remaining records into chains.
  - Collect strong evidence from cores and leverage in clustering
  - No penalty on local values

Apply weak evidence

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Casa	900	AL	tacocasa.com
r3	Taco Casa	900	AL	tacocasa.com, tacocasatexas.com
r4	Taco Casa	900	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

Two-Stage Linkage II

- **Stage II: Cluster** cores and remaining records into chains.
  - Collect strong evidence from cores and leverage in clustering
  - No penalty on local values

No penalty on local values

ID	name	phone	state	URL domain
r1	Taco Casa		AL	tacocasa.com
r2	Taco Casa	900	AL	tacocasa.com
r3	Taco Casa	900	AL	tacocasa.com, tacocasatexas.com
r4	Taco Casa	900	AL	
r5	Taco Casa	900	AL	
r6	Taco Casa	701	TX	tacocasatexas.com
r7	Taco Casa	702	TX	tacocasatexas.com
r8	Taco Casa	703	TX	tacocasatexas.com
r9	Taco Casa	704	TX	
r10	Elva's Taco Casa		TX	tacodemar.com

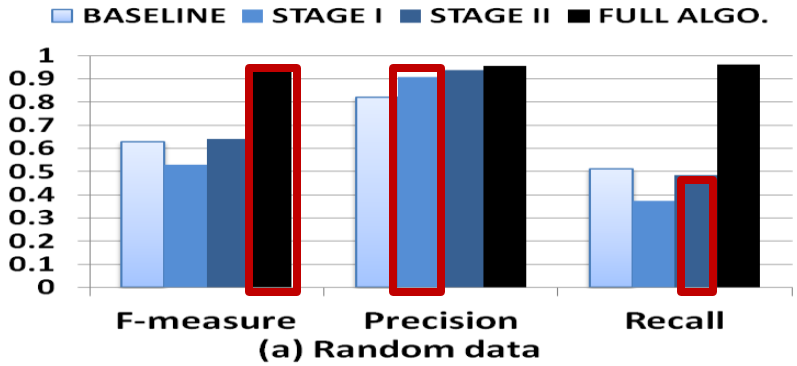
Experimental Evaluation I

- **Data set**
  6.8M records from *YellowPages.com*
- 6.9 hours
  - 2.2 hrs for Stage I (core generation)
  - 4.7 hrs for Stage II (clustering)
- 80K chains and 1M records in chains

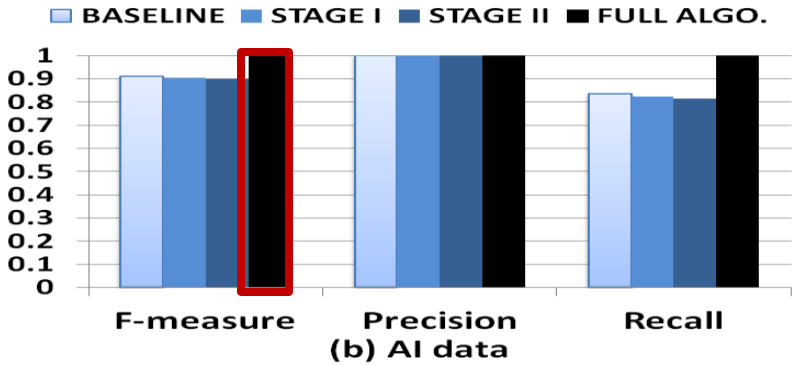
Chain name	# Stores
USPS - United States Post Office	12,776
SUBWAY	11,278
State Farm Insurance	8,711
McDonald's	7,450
U-Haul Neighborhood Dealer	7,105
Edward Jones	6,781

Experimental Evaluation II

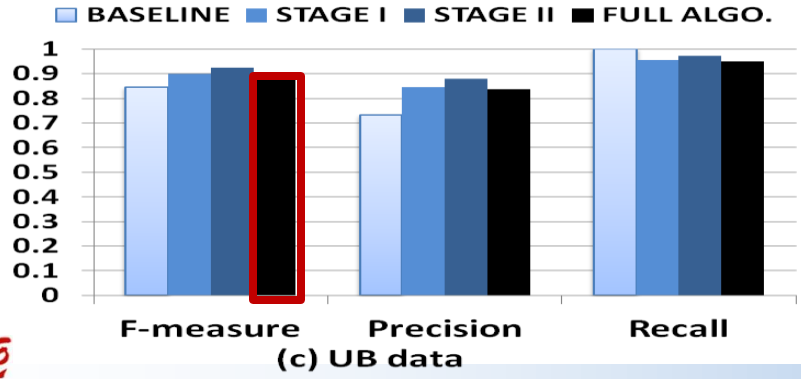
Sample	#Records	#Chains	Chain size	#Single-biz records
Random	2062	30	[2, 308]	503
AI	2446	1	2446	0
UB	322	7	[2, 275]	5
FBIns	1149	14	[33, 269]	0



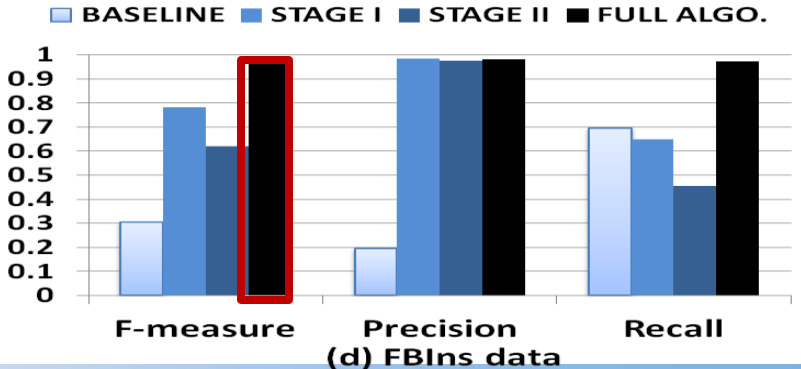
(a) Random data



(b) AI data



(c) UB data



(d) FBIns data

Related Work

- Traditional record linkage techniques
  - Record similarity computation
    - Classification [Fellegi,69], Distance [Dey,08], Rule [Hernandez,98]
  - Record clustering
    - Transitive rule [Hernandez,98], Optimization [Wijaya,09]
- Temporal linkage
  - Periodical behavior patterns [Yakout,10]
  - Rule-based linkage [Burdick, 11]
- Two-stage clustering
  - K-means based clustering [Larsen, 99]
  - Probabilistic model [Liu, 02]
  - Bootstrapping [Yoshida, 10]

Conclusions

- In some applications record linkage needs to be tolerant with **value diversity**
- When linking **temporal records**, **time decay** allows tolerance on evolving values
- When linking **group members**, **two-stage linkage** allows leveraging strong evidence and allows tolerance on different local values

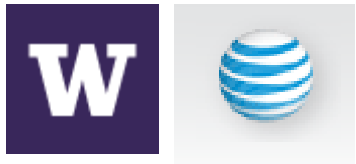
Thanks!

Contact: pei.li@disco.unimib.it

Disagreement Decay

- Intuition: different values over a long time is not a strong indicator of referring to different entities.
  - University of Washington (01-07)

  AT&T Labs-Research (07-date)



- Definition (***Disagreement decay***)
  - Disagreement decay of attribute A over time Δt is the probability that an entity changes its A-value within time Δt.

Agreement Decay

- Intuition: the same value over a long time is not a strong indicator of referring to the same entities.
  - Adam Smith: (1723-1790)

    Adam Smith: (1965-)

- Definition (***Agreement decay***)
  - Agreement decay of attribute A over time Δt is the probability that different entities share the same A-value within time Δt.

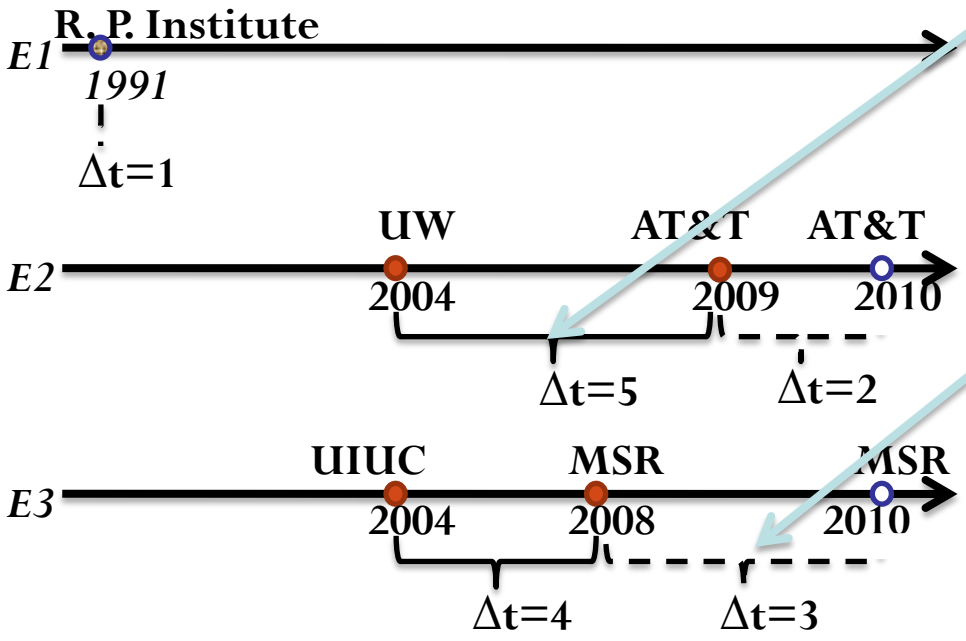
Applying Decay

$$sim(r, r') = \frac{\sum_{A \in \mathbf{A}} w_A(\Delta t) \cdot sim(r.A, r'.A)}{\sum_{A \in \mathbf{A}} w_A(\Delta t)}$$

$$w_A(\Delta t) = 1 - d(A, \Delta t)$$

- E.g.
  - r1 <Xin Dong, Uni. of Washington, 2004>
  - r2 <Xin Dong, AT&T Labs-Research, 2009>
- **Decayed similarity**
  - w(name, Δt=5)=1-d_{agree}(name , Δt=5)=.95,
  - w(affi., Δt=5)=1-$d_{disagree}$(affi. , Δt=5)=.1
  - sim(r1, r2)=(.95*1+.1*0)/(.95+.1)=**.9** **Match**
- **No decayed similarity:**
  - w(name)=w(affi.)=.5
  - sim(r1, r2)=.5*1+.5*0=**.5** **Un-match**

Learning Disagreement Decay



Change & last time point Change point

Last time point Full life span Partial life span

1. Full life span: [t, t_{next})

A value exists from t to t_{next}, for time (t_{next}-t)

2. Partial life span: [t, t_{end}+1)*

A value exists since t, for at least time (t_{end}-t+1)

$$d^{\neq}(A, \Delta t) = \frac{|\{l \in \bar{L}_f | l \leq \Delta t\}|}{|\bar{L}_f| + |\{l \in \bar{L}_p | l \geq \Delta t\}|}$$

L_p={1, 2, 3}, L_f={4, 5}

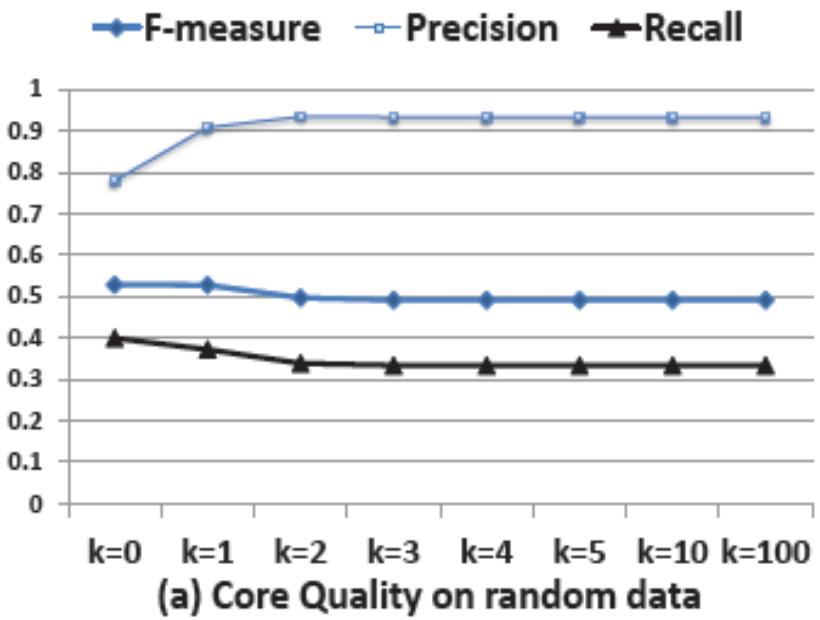
d(Δt=1)=0/(2+3)=0

d(Δt=4)=1/(2+0)=0.5

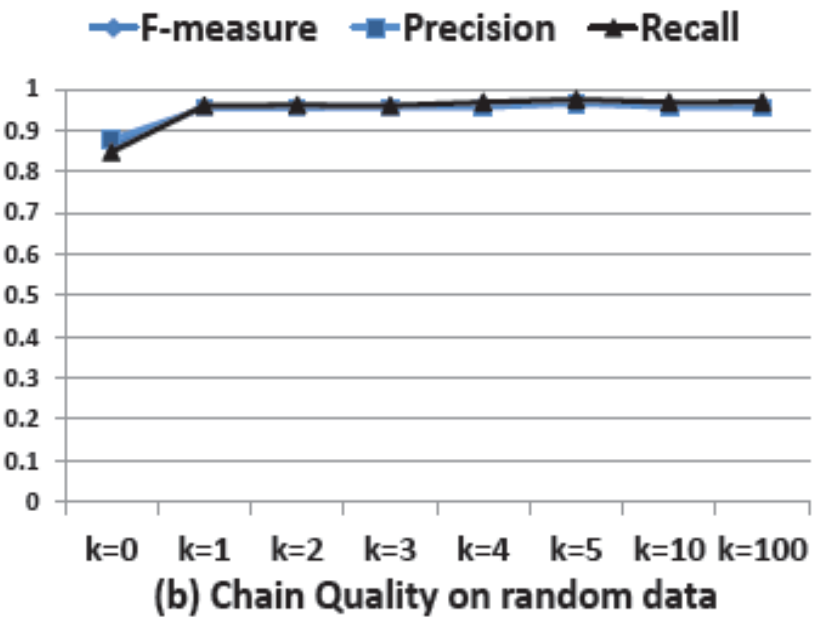
d(Δt=5)=2/(2+0)=1

Experimental Evaluation

- **K-robustness of cores:**



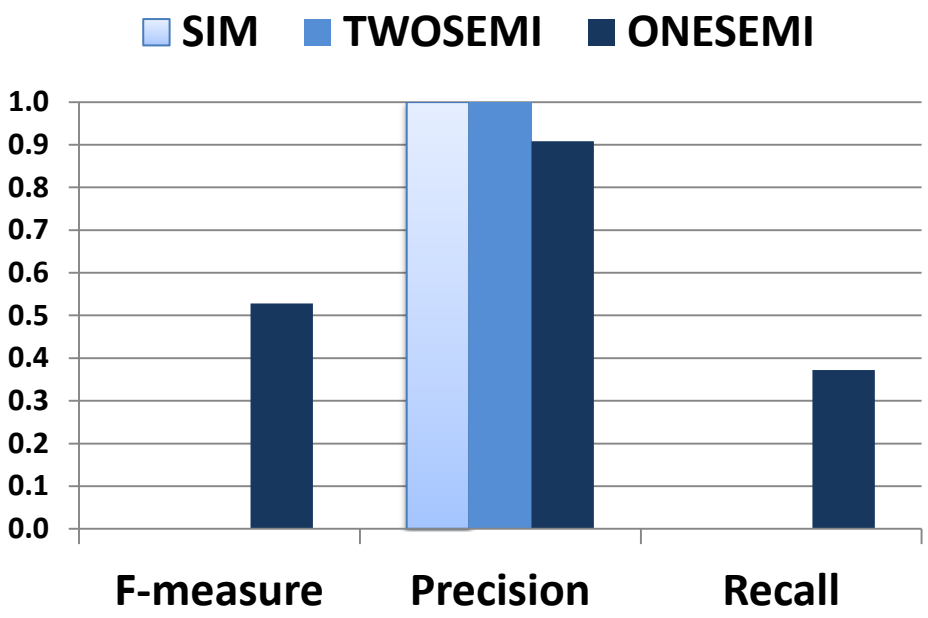
(a) Core Quality on random data



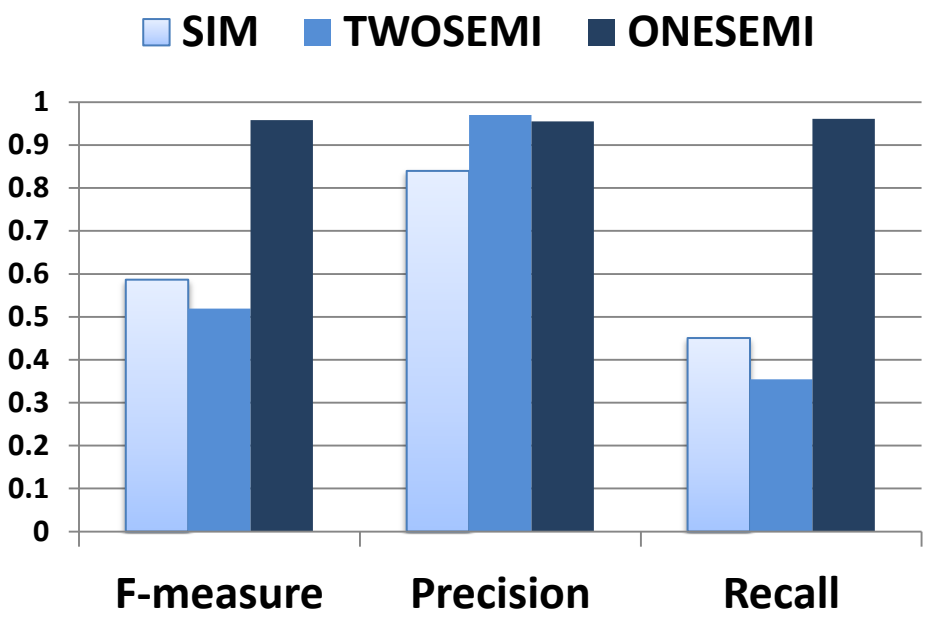
(b) Chain Quality on random data

Experimental Evaluation

- **Different strategies of cores:**



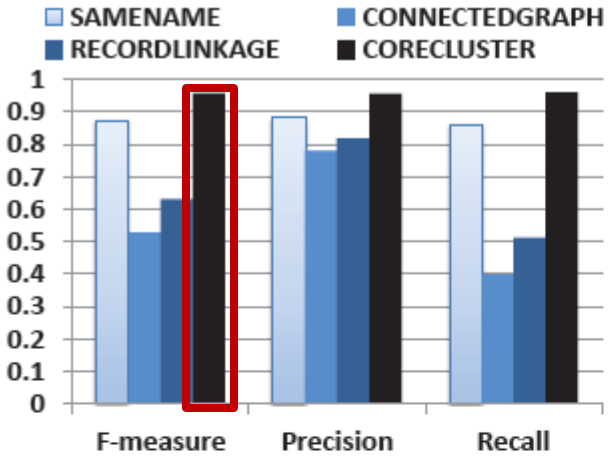
(a) Core quality on random data



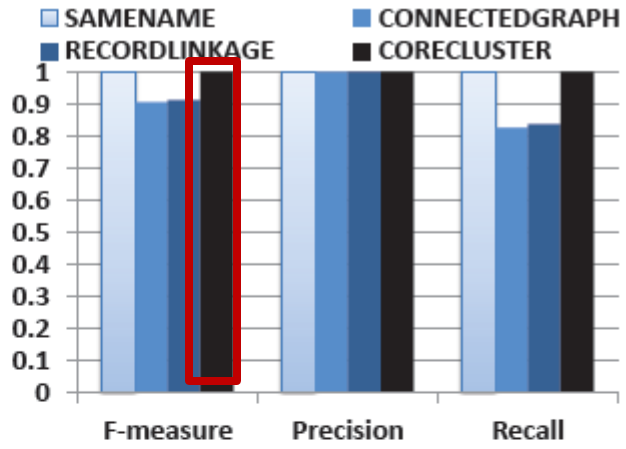
(b) Chain quality on random data

Experimental Evaluation

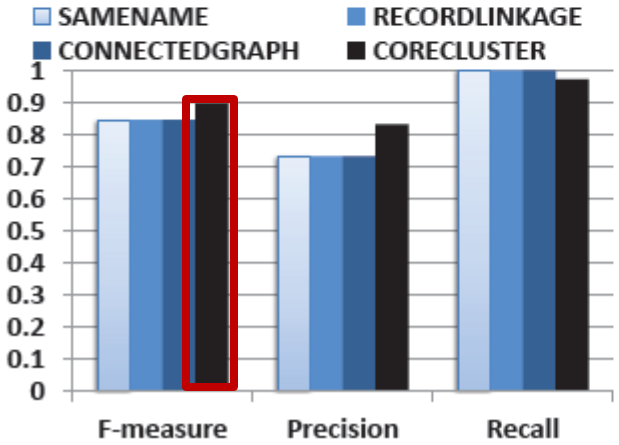
- **Overall results:**



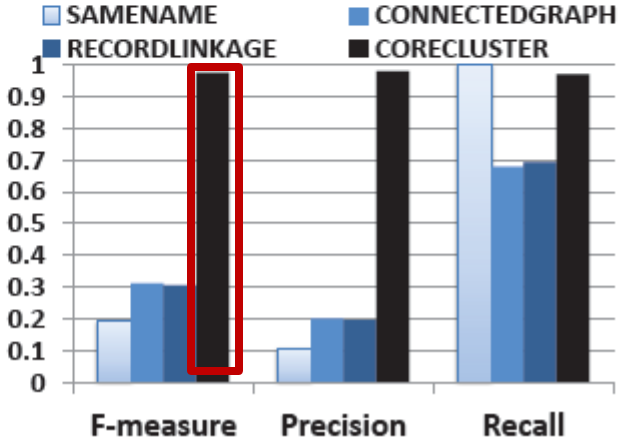
(a) Random data



(b) AI data



(c) UB data



(d) Perturbed FBIns data