Provenance-Aware Entity Resolution

Leveraging Provenance to Improve Quality

Qing Wang[1], Klaus-Dieter Schewe[2] and Woods Wang[3]

[1]Research School of Computer Science, Australian National University, Australia
[2]Software Competence Center Hagenberg and Johannes-Kepler-University Linz, Austria
[3]Alcatel-Lucent Beijing, China

Introduction – Entity Resolution

- *Entity resolution* (ER) is to determine whether or not different entity representations (e.g., records) correspond to the same real-world entity.

Introduction – Entity Resolution

- *Entity resolution* (ER) is to determine whether or not different entity representations (e.g., records) correspond to the same real-world entity.
- Consider the following relation AUTHORS:

aid	name	affiliation	email
1	Qing Wang		qw@gmail.com
2	Mike Lee	Curtin University	
3	Qinqin Wang	Curtin University	
4	Jan Smith		jan@gmail.com
5	Q. Wang	University of Otago	qw@gmail.com
6	Jan V. Smith	RMIT	jan@gmail.com
7	Q. Q. Wang		
8	Wang, Qing	University of Otago	

– Are Qing Wang (1) and Q. Wang (5) the same person?
– Are Qinqin Wang (3) and Q. Wang (5) not the same person?
– ...

State of the Art

- State-of-the-art approaches to entity resolution favor similarity-based methods.
- Numerous techniques have been developed under a variety of perspectives:
  a. **Rule-based**: positive rules, negative rules, soft rules, hard rules, etc.
  b. **Supervised**: decision tree, SVM, etc.
  c. **Active learning**: uncertainty sampling, query by committee, etc.
  d. **Clustering-based**: k-means, correlation clustering, etc.
  e. **Collective**: take into account relational similarity
  f. ...
- The central idea is

> "The more similar two entity representations are, the more likely they are referring to the same real-world entity."

Entity Resolution – Example

- Nevertheless, ER results in real-world applications are still largely *imprecise*.

aid	name	affiliation	email
1	Qing Wang		qw@gmail.com
2	Mike Lee	Curtin University	
3	Qinqin Wang	Curtin University	
4	Jan Smith		jan@gmail.com
5	Q. Wang	University of Otago	qw@gmail.com
6	Jan V. Smith	RMIT	jan@gmail.com
7	Q. Q. Wang		
8	Wang, Qing	University of Otago	

eid	aids
e_1	$\langle 1, 3, 5, 7, 8 \rangle$
e_2	$\langle 2 \rangle$
e_3	$\langle 4, 6 \rangle$

An ER result (incorrect)

eid	aids
e_1	$\langle 1, 5, 8 \rangle$
e_2	$\langle 2 \rangle$
e_3	$\langle 4, 6 \rangle$
e_4	$\langle 3, 7 \rangle$

An ER result (correct)

Goals of the Paper

- To answer some research questions:

  Can we leverage provenance information to improve the quality of ER?

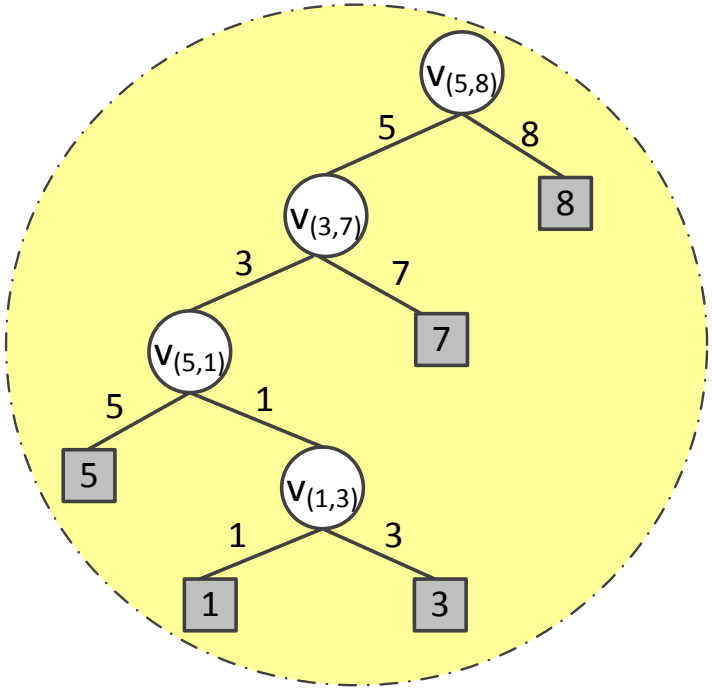
  What kind of provenance information should we capture for ER?

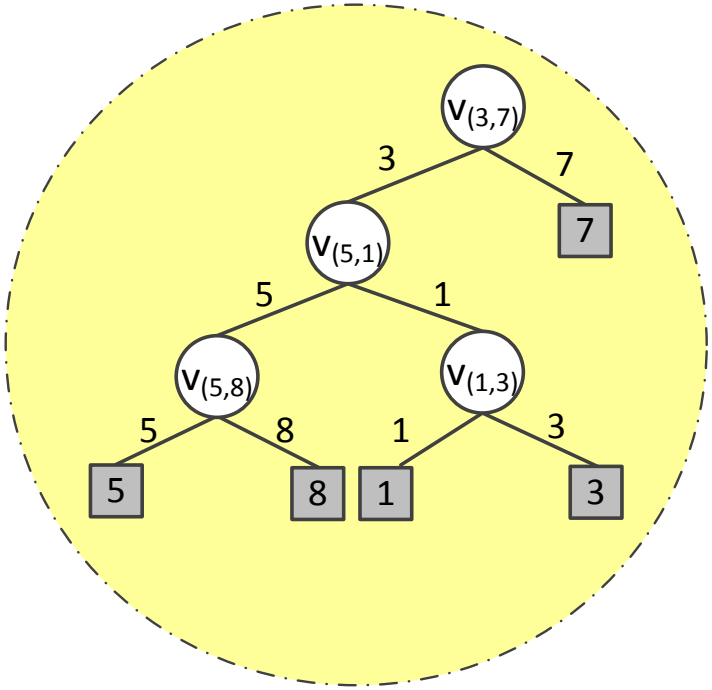
  How can such provenance information be used efficiently and effectively?

- To develop techniques for **provenance-aware entity resolution**.

Entity Resolution Index (ERI)

- ERI is a data structure that maps each entity e to an ER tree t_e, in which each index entry has the form: (entity e, ER tree t_e).
- An ER tree t_e keeps track of the matches that are relevant to the entity e.





- The ER tree t_e of e is determined by the matches among records in e, depending on a chosen matching algorithm.

Inconsistent ER Results

- We focus on inconsistent ER results governed by *two kinds of ER constraints*:

  (1) **Must-link constraints** $k_1 \backsimeq k_2$

  i.e., records k_1 and k_2 must be matched to the same entity.

  (2) **Cannot-link constraints** $k_1 \not\backsimeq k_2$

  i.e., records k_1 and k_2 cannot be matched to the same entity.

  - Such constraints can be gathered from user feedback, domain experts, background knowledge, integrity constraints, etc.

  - **Example:** Evaluating the following constraint over a relation AUTHOR(aid, name, affiliation, email) would yield a set of must-link constraints:

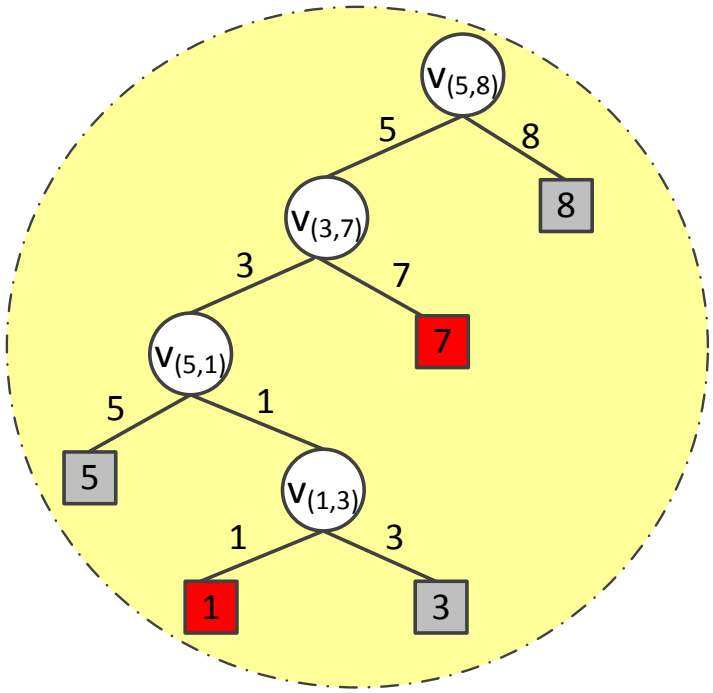
$$x \backsimeq x' \Leftarrow \text{AUTHOR}(x, y, z, w) \wedge \text{AUTHOR}(x', y', z', w).$$

Repairing ER Results

- How can we use an ERI index to support repairs of inconsistent ER results?

- **Basic operations:**
  - Must-link constraints $k_1 \backsimeq k_2$
    $\hookrightarrow$ Merging records k_1 and k_2 into the same entity using (k_1, k_2).
  - Cannot-link constraints $k_1 \not\backsimeq k_2$
    $\hookrightarrow$ Splitting records k_1 and k_2 into two different entities (this is hard).

- **Example:** If $1 \not\backsimeq 7$, how should we split the entity $e = \langle 1, 3, 5, 7, 8 \rangle$ with the following ER tree t_e?



Algorithms for Splitting

- Given an ER tree t_e that violates a cannot-link constraint σ, do the following:

(1) **Start** with finding the guard node in t_e.

- The *guard node* "guards" the search for critical nodes and cut nodes in the next steps. *Exactly one guard node* exists in t_e w.r.t. σ.

(2) **Traverse downward** to find all critical nodes in t_e.

- *Critical nodes* correspond to matches that are relevant to the violation of σ, but do not necessarily correspond to erroneous matches.

(3) **Human feedback** to classify all critical records relating to critical nodes, and then cut nodes in t_e are automatically generated based on human feedback.

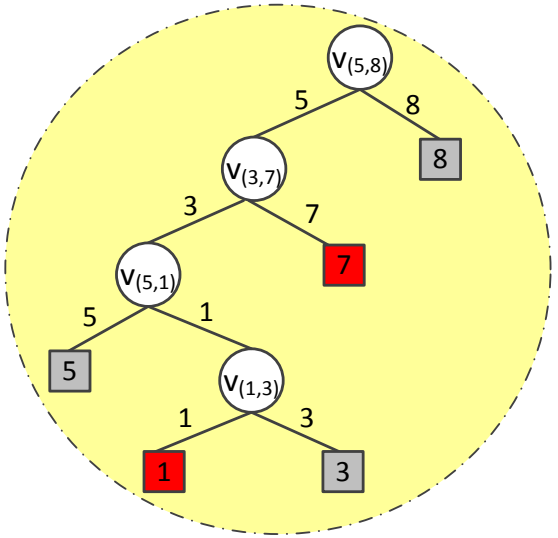
- *Cut nodes* correspond to erroneous matches.

(4) **Traverse upward** to remove all cut nodes and propagated erroneous nodes, and split t_e into subtrees.

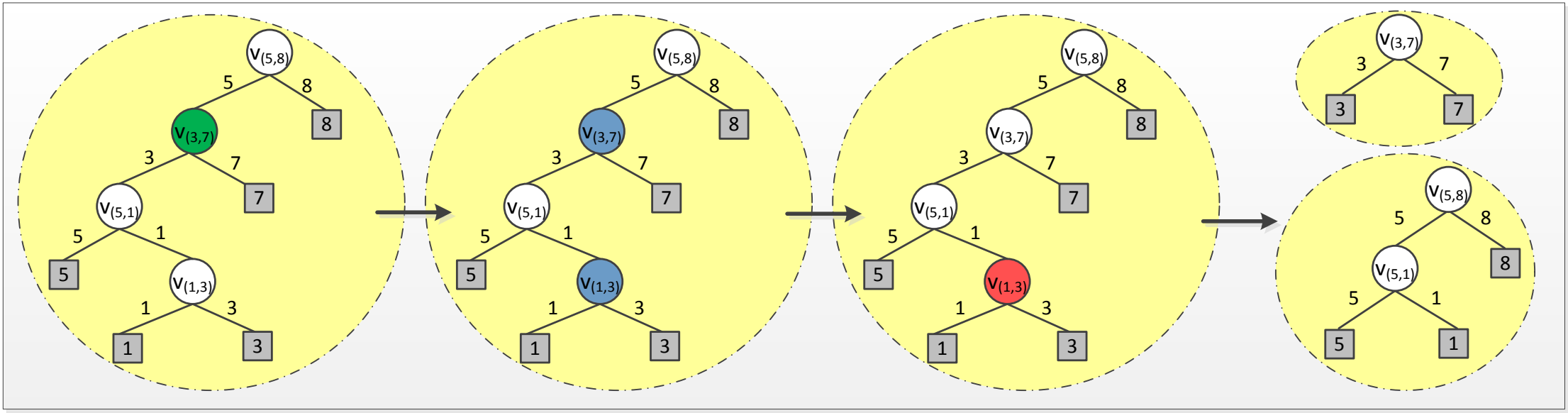
Algorithms for Splitting

- **Example:**

If $1 \not\leq 7$, how should we split the entity $e = \langle 1, 3, 5, 7, 8 \rangle$ with the following ER tree t_e?



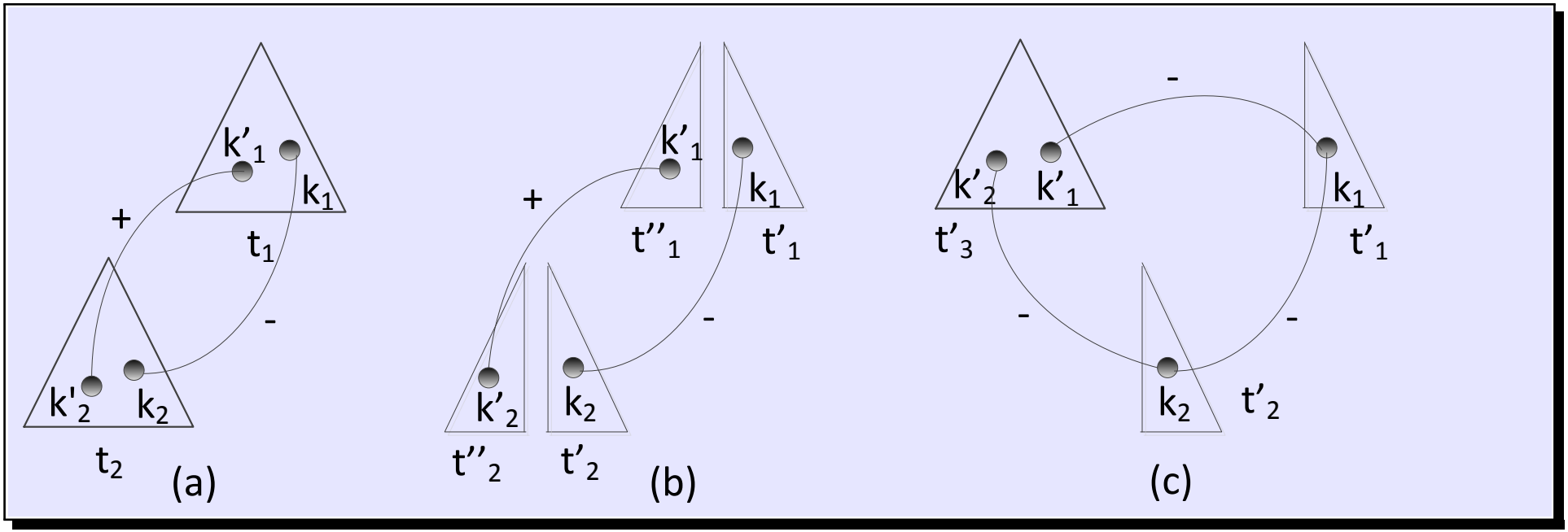
- Applying the splitting algorithm on t_e w.r.t. $1 \not\leq 7$:



green=guard node blue=critical nodes red=cut nodes splitting result

Repairing ER Results - Strategies

- A naive strategy is to apply a merge operation whenever a must-link constraint is violated, and apply a split operation whenever a cannot-link constraint is violated.
- However, this naive strategy does not work well when must-link and cannot-link constraints coexist and interweave.
- We propose a strategy, called *coordinate-split-merge* (CSM), to repair ER results by taking into account the interaction between must-link and cannot-link constraints.



Experiments

- Three sets of experiments:

  (1) Time and space requirements for building an ERI index

  (2) Human effort: how efficient can ER results be repaired with human feedback?

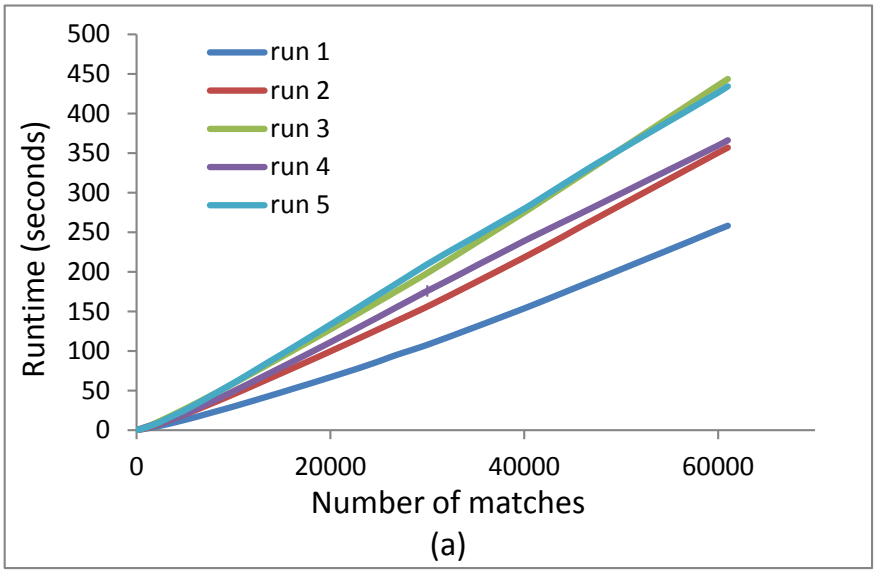
  (3) ER quality: how effective can ER results be repaired?

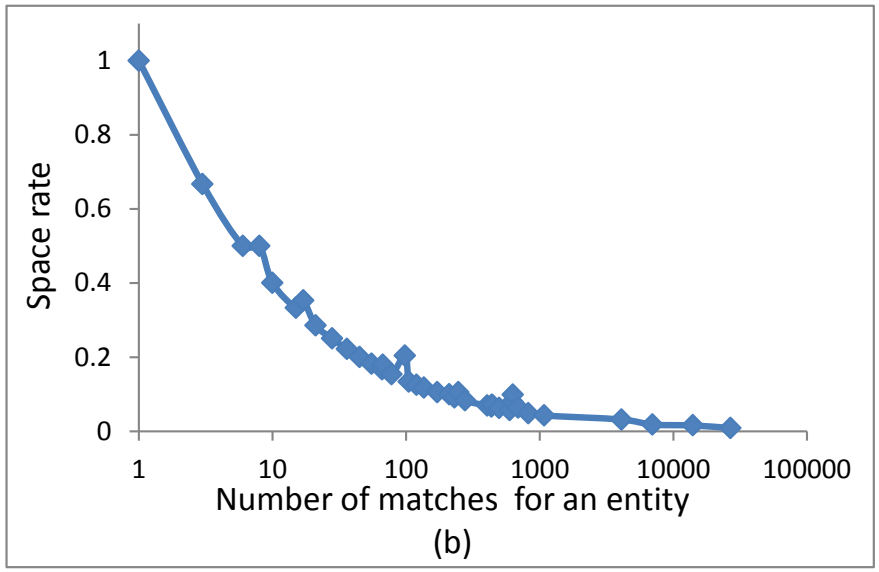
- Two data sets:
  - **Cora:** contains 1,879 publication records, and the "gold standard" is available.
  - **Scopus:** contains 10,784 publication records. A "gold standard" for 4,865 publication and 19,527 author records was established by domain experts.

Results	Cora	Scopus
matches	61,453	75,447
non-singleton ER trees	117	2,969
singleton ER trees	1,674	36,003

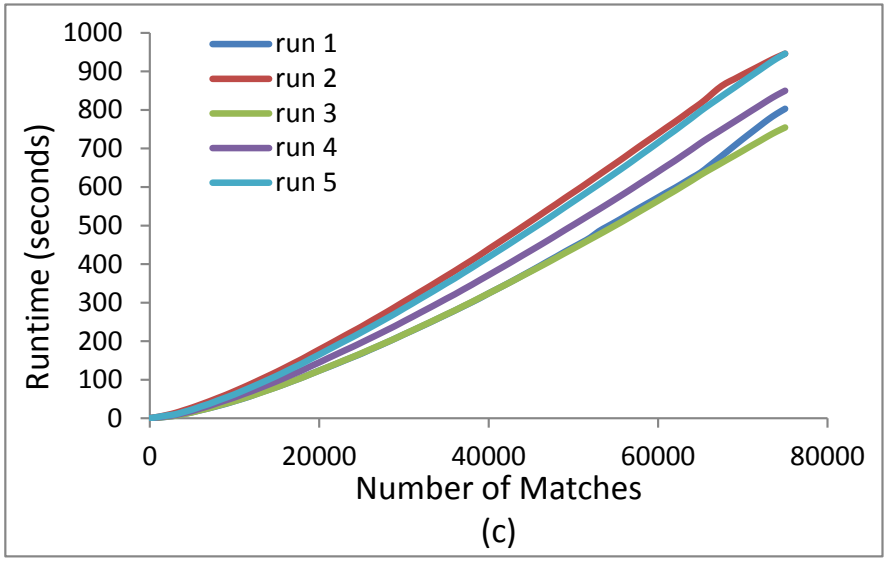
Experiments (1) – Time and Space Requirements

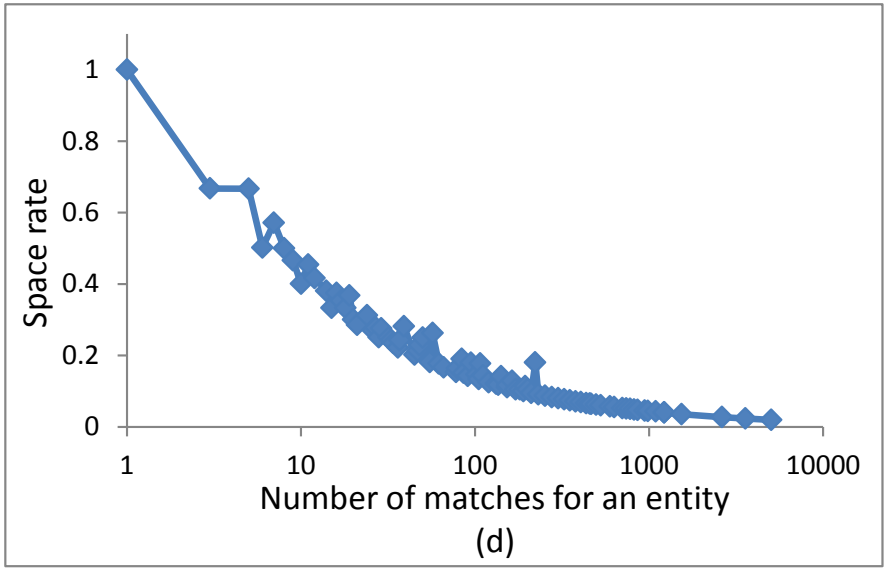
Cora





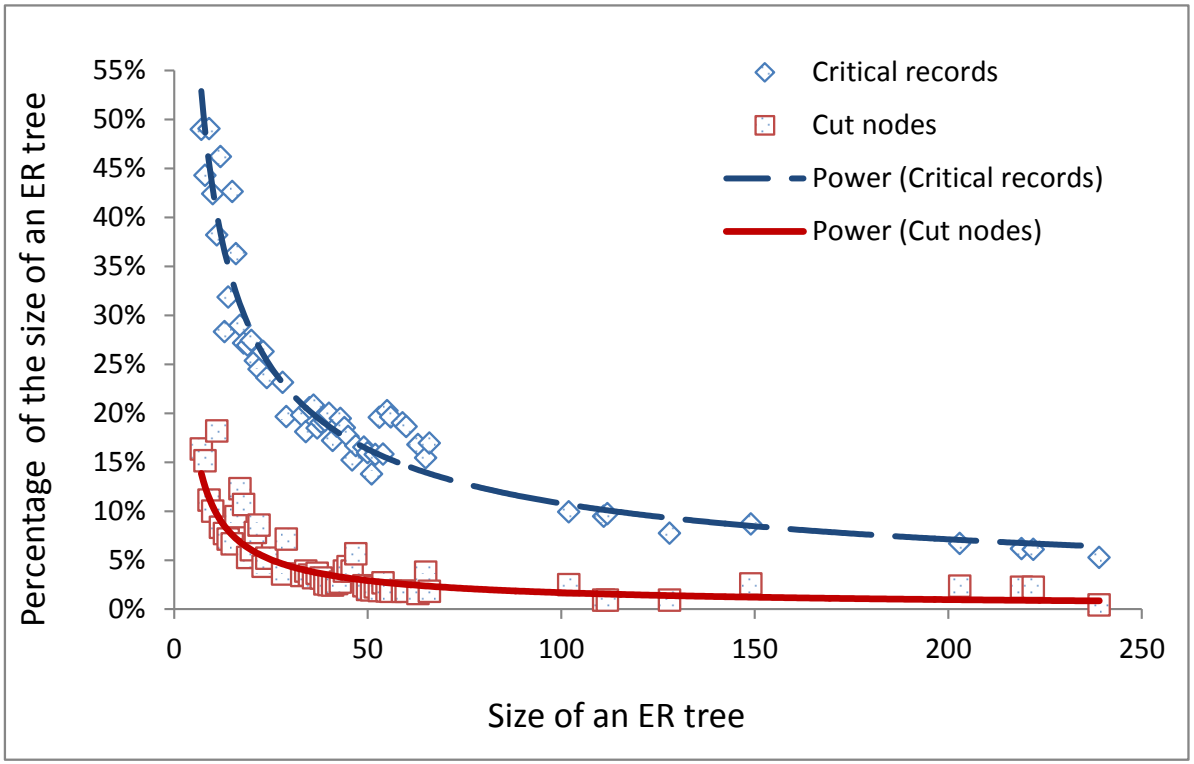
Scopus





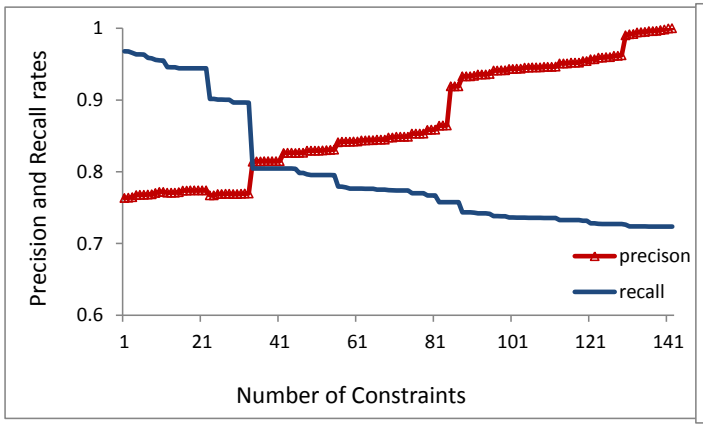
Experiments (2) – Human Effort

- **Some observations** (on Cora):
  - About 30% to 50% of records are critical when the size of a cluster is small (i.e., less than 10), whereas 5% to 9% of records are critical when the size of a cluster is over 100.
  - The percentage of the number of cut nodes in terms of the size of a cluster remains quite stable (between 1% - 2%) when the size of a cluster is greater than 50.

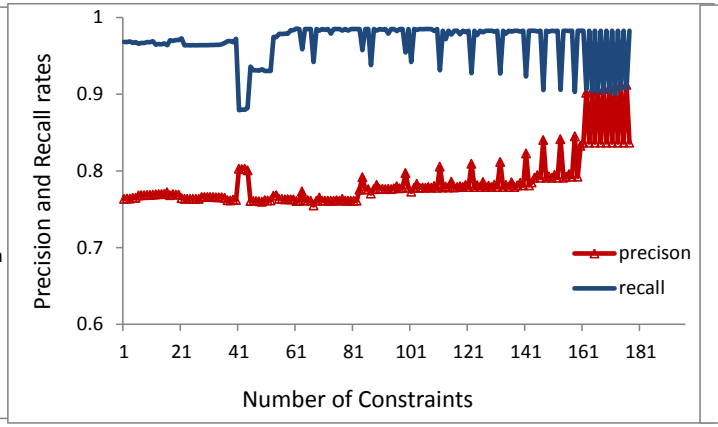


Experiments (3) – ER Quality

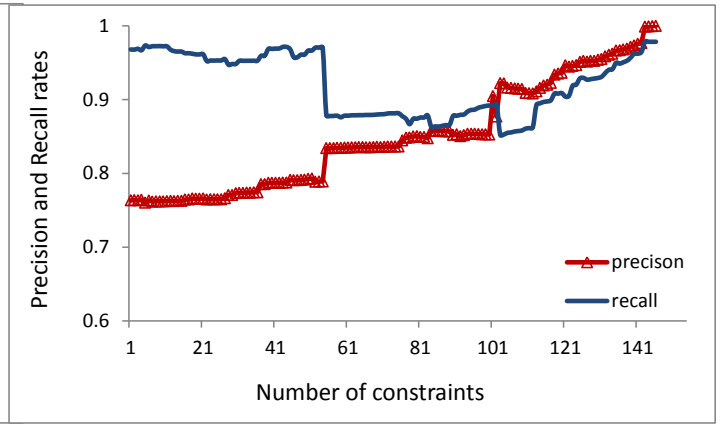
- **Baseline:** an ER result for Cora with precision=76.35% and recall=96.80%.
- **Some results:**
  - Using cannot-link constraints leads to increased precision but decreased recall.
  - The naive strategy shows an inverse correlation between precision and recall.
  - The CSM strategy can control the negative effects of merging operations on precision so that the precision can be consistency improved over time.



(a) Cannot-link constraints



(b) The naive strategy



(c) The CSM strategy

Conclusions

- We have studied the entity resolution problem in terms of provenance, which is largely unexplored in the literature.
- A tree-based indexing method was proposed, which can efficiently manage the provenance information of the ER process.
- The ERI indexing enables us to repair inconsistent ER results that violate must-link and cannot-link constraints.
- Our experimental results confirmed that the ERI indexing not only exhibits good scalability properties for building a provenance structure, but also supports repairing erroneous ER matches with reduced human efforts.

For questions, please email qing.wang@anu.edu.au.