Name Disambiguation in AMiner:

Jing Zhang*, Jie Tang+

*Information School, Renmin University
+Computer Science Department, Tsinghua University

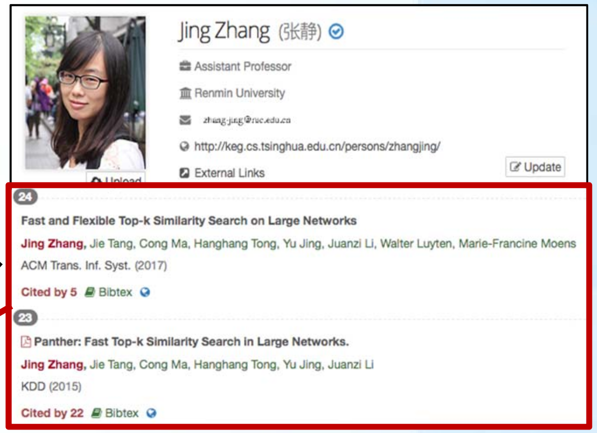
AMiner

清華大學
Tsinghua University

中国人民大學
RENMIN UNIVERSITY OF CHINA

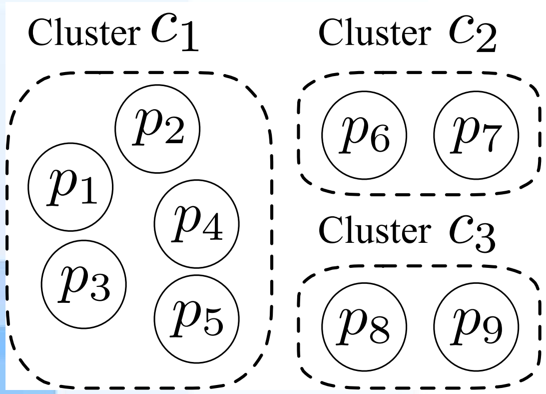
An Example in AMiner

All the researchers named "Jing Zhang"



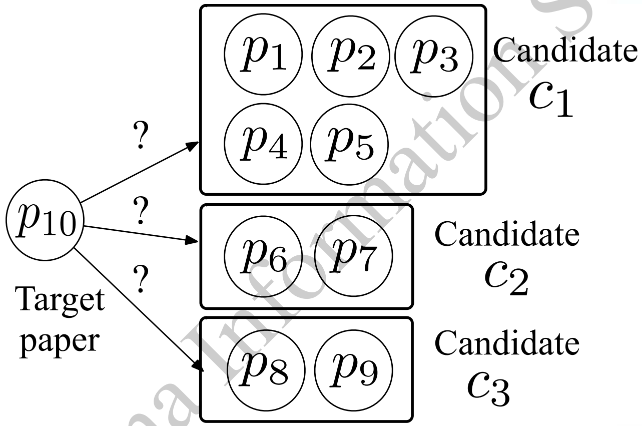
Assign papers with author name "Jing Zhang" to right authors

Three Scenarios



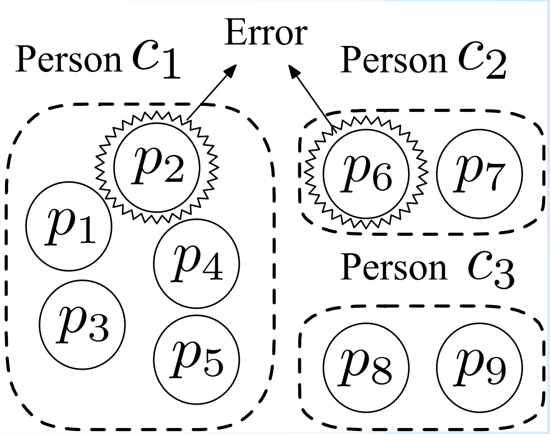
Full ND

- Name disambiguation when the system is built from scratch



Continuous ND

- Name disambiguation when persons' profilesare continuously updated

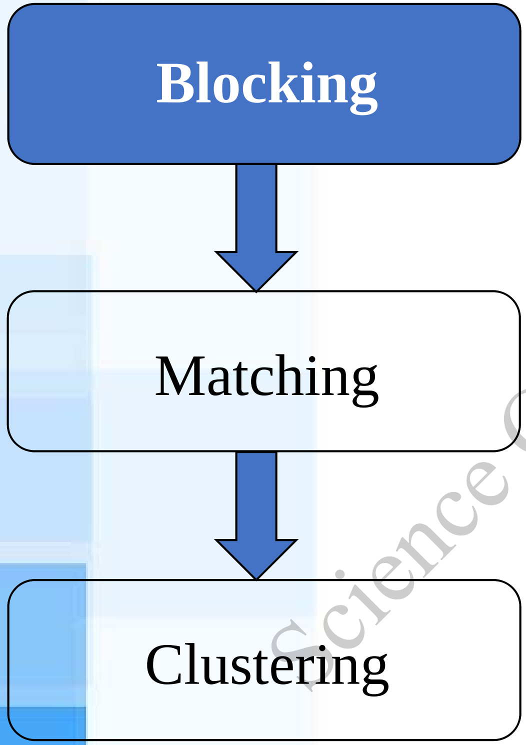


Error

- Error detection upon existing persons' profiles

1st Scenario: Full ND

- Stage One: Constructing candidate sets of potential matchings.



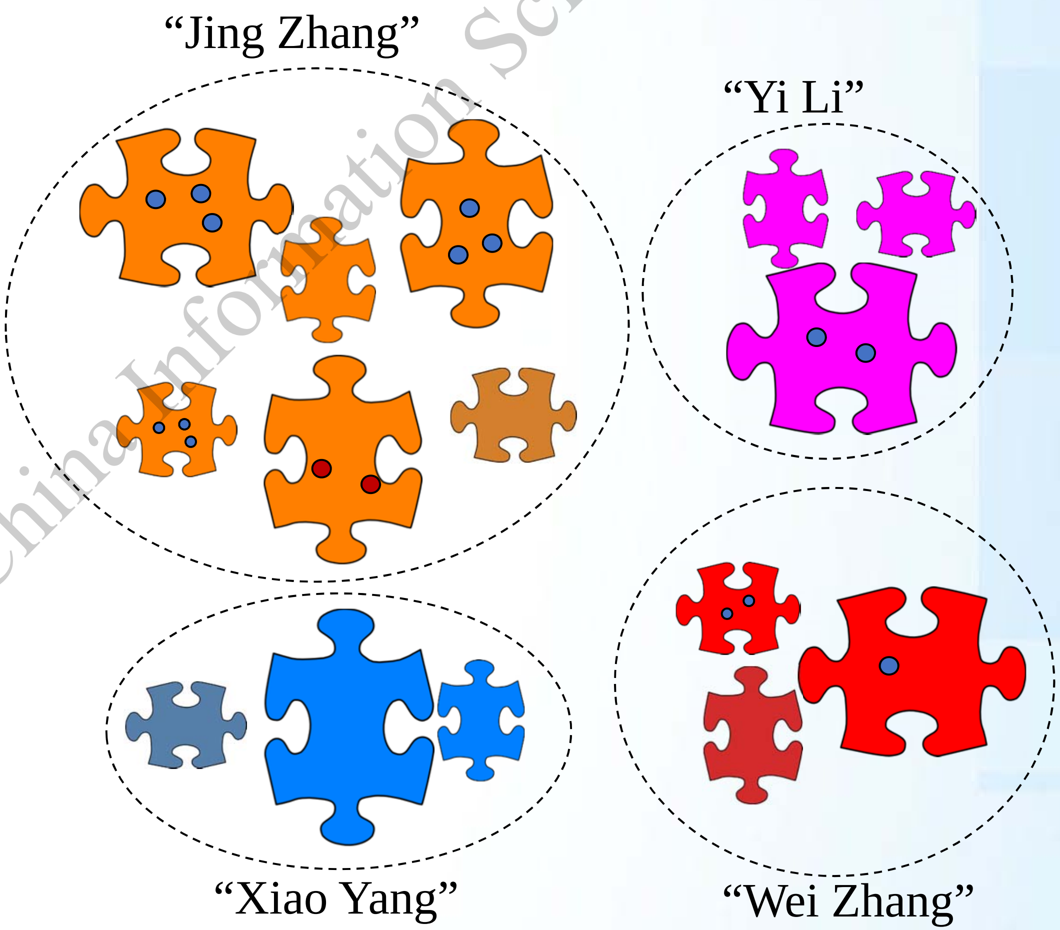


Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

Stage Two: Similarity Matching

- Constructing similarity matrix for each candidate set.

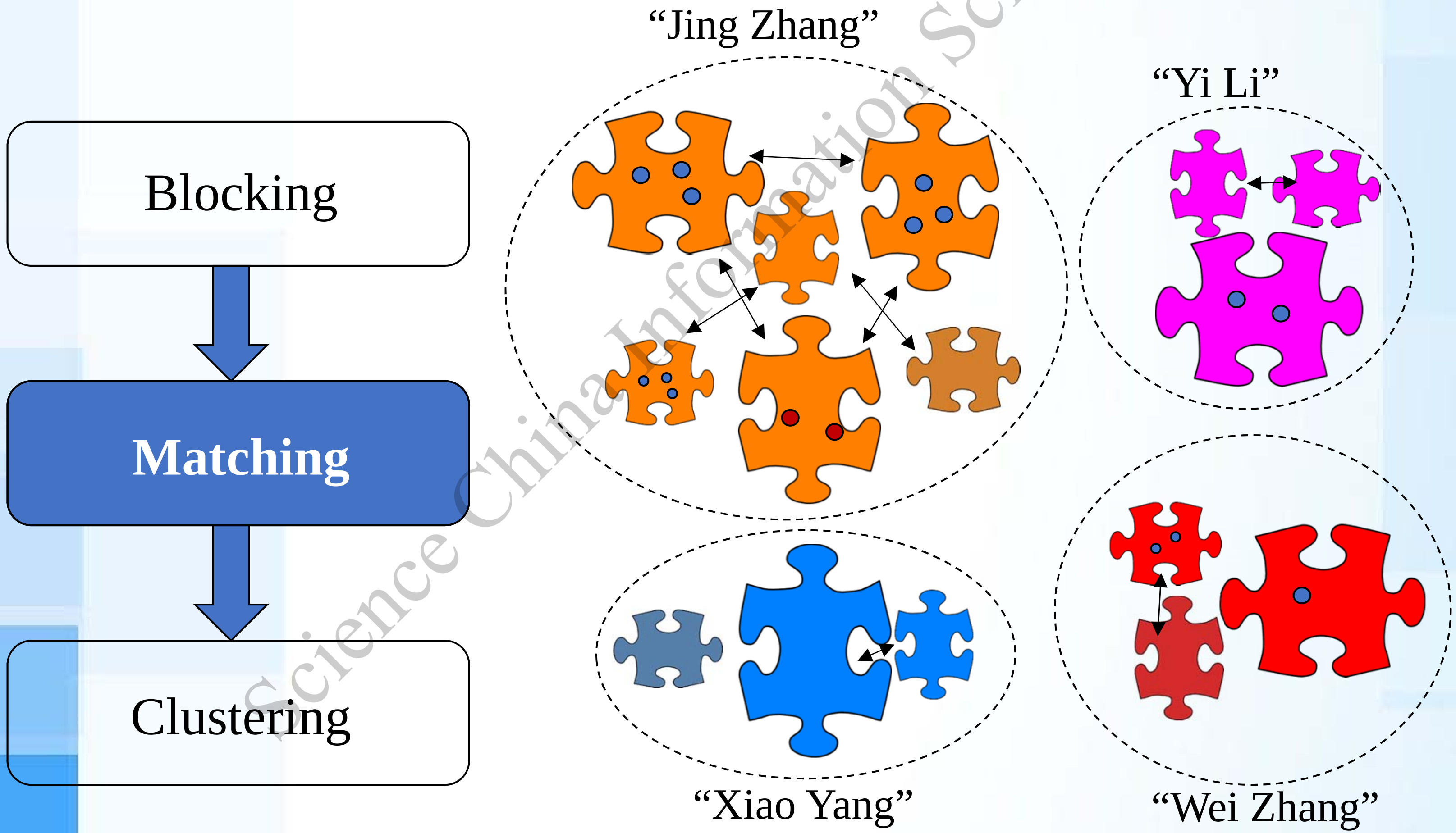
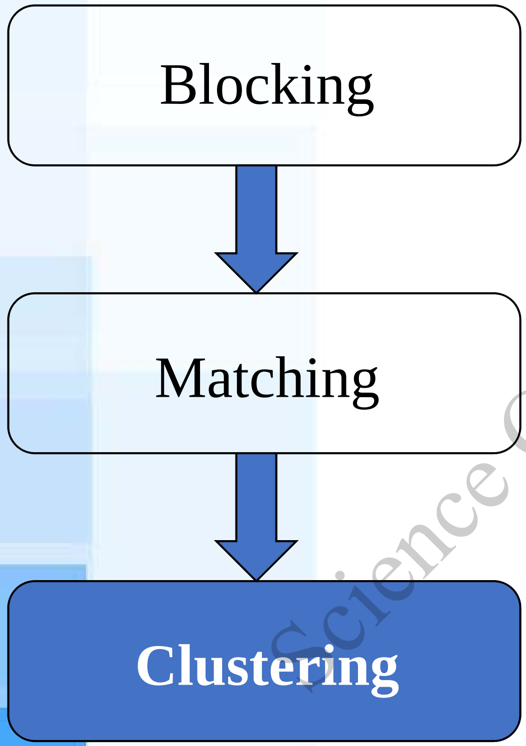


Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

Stage Three: Clustering

- Partitioning each candidate set based on similarity.



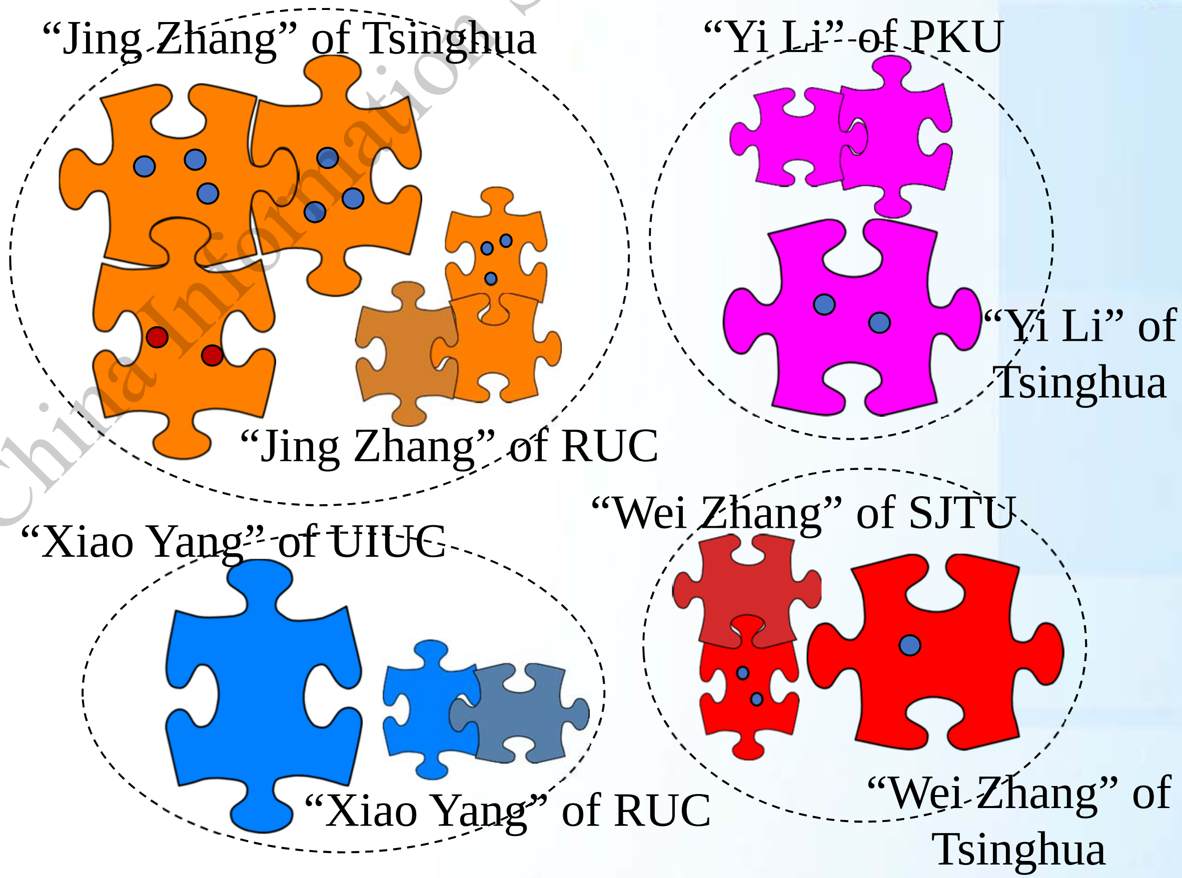


Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

Feature-based Matching

TABLE 1
Attributes of Each Publication p_i

Attribute	Description
p_i.title	title of p_i
p_i.pubvenue	published conference/journal of p_i
p_i.year	published year of p_i
p_i.abstract	abstract of p_i
p_i.authors	authors name set of p_i $\{a_i^{(0)}, a_i^{(1)}, \ldots a_i^{(u)}\}$
p_i.references	references of p_i

TABLE 2
Relationships between Papers

R	W	Relation Name	Description
r_1	w_1	CoPubVenue	$p_i.pubvenue = p_j.pubvenue$
r_2	w_2	CoAuthor	$\exists r, s>0, a_i^{(r)}=a_j^{(s)}$
r_3	w_3	Citation	p_i cites p_j or p_j cites p_i
r_4	w_4	Constraint	feedback supplied by users
r_5	w_5	τ-CoAuthor	τ-extension co-authorship (τ>1)

Local features

- Similarity between a paper and its cluster centroid according to papers' attributes

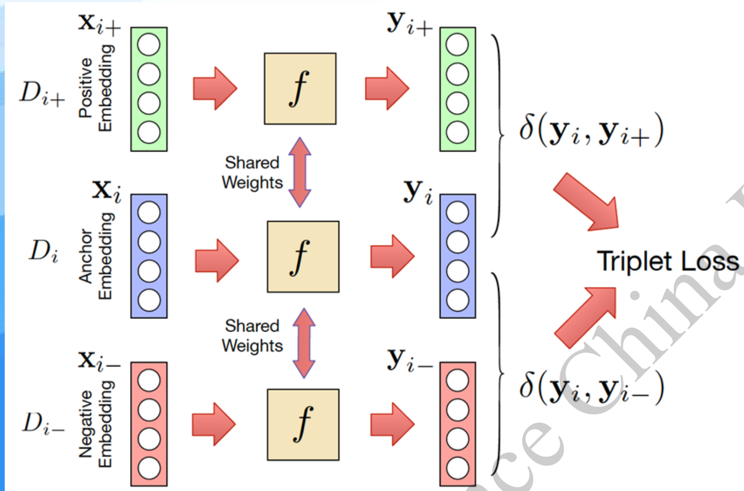
Correlation features

- Similarities between two papers according to their relationships.

Embedding-based Matching

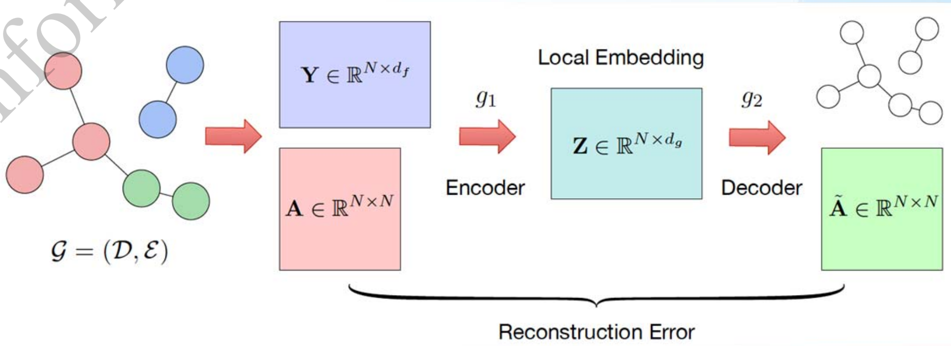
- Calculating similarities between embeddings.

Global embeddings



- Map every paper to a **unified representation space.**
- **Share supervision** across different candidate sets.

Local embeddings



- For each candidate set: build a graph by linking each two similar papers.
- Train a **graph auto-encoder** to learn **separate representation space.**

2nd Scenario: Continuous ND

- Papers come in a streaming fashion (**500,000/month**).
- Assigning new papers to right persons continuously.
- Stage One: Constructing candidates related to the author of the target paper.

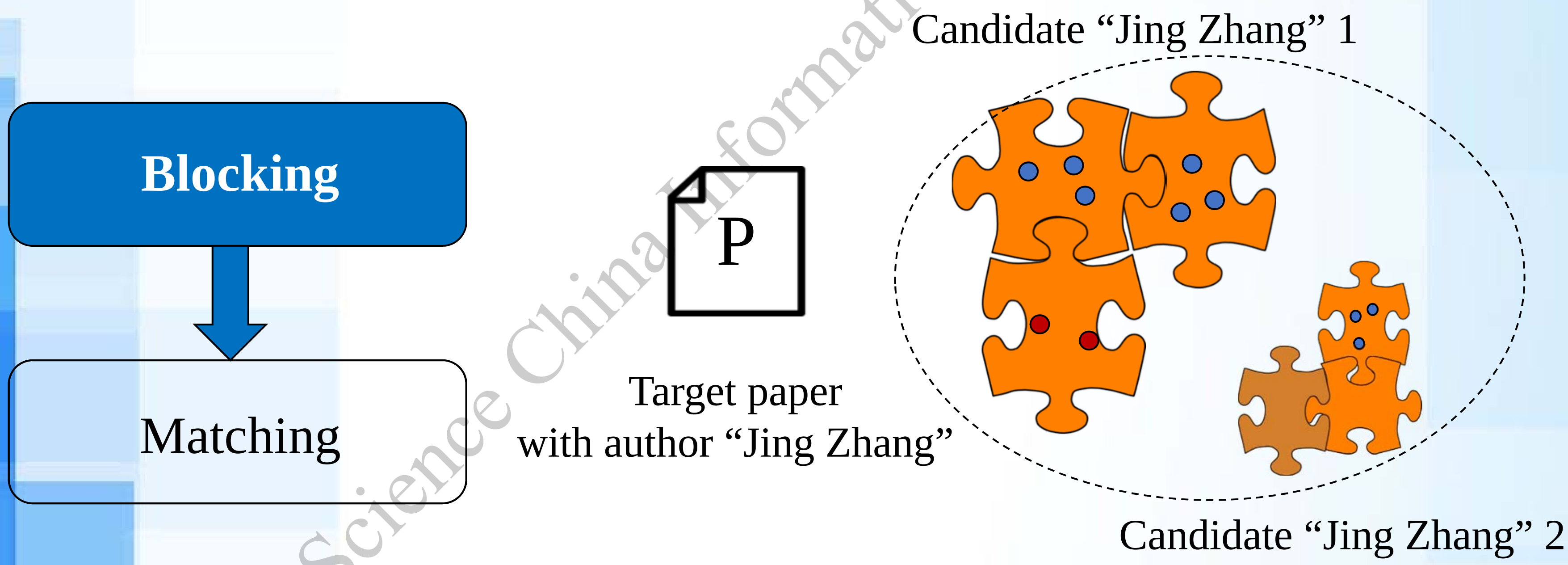


Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

Stage Two: Similarity Matching

- Matching the target paper and each candidate.
- Assigning the paper to the candidate with the largest matching score.

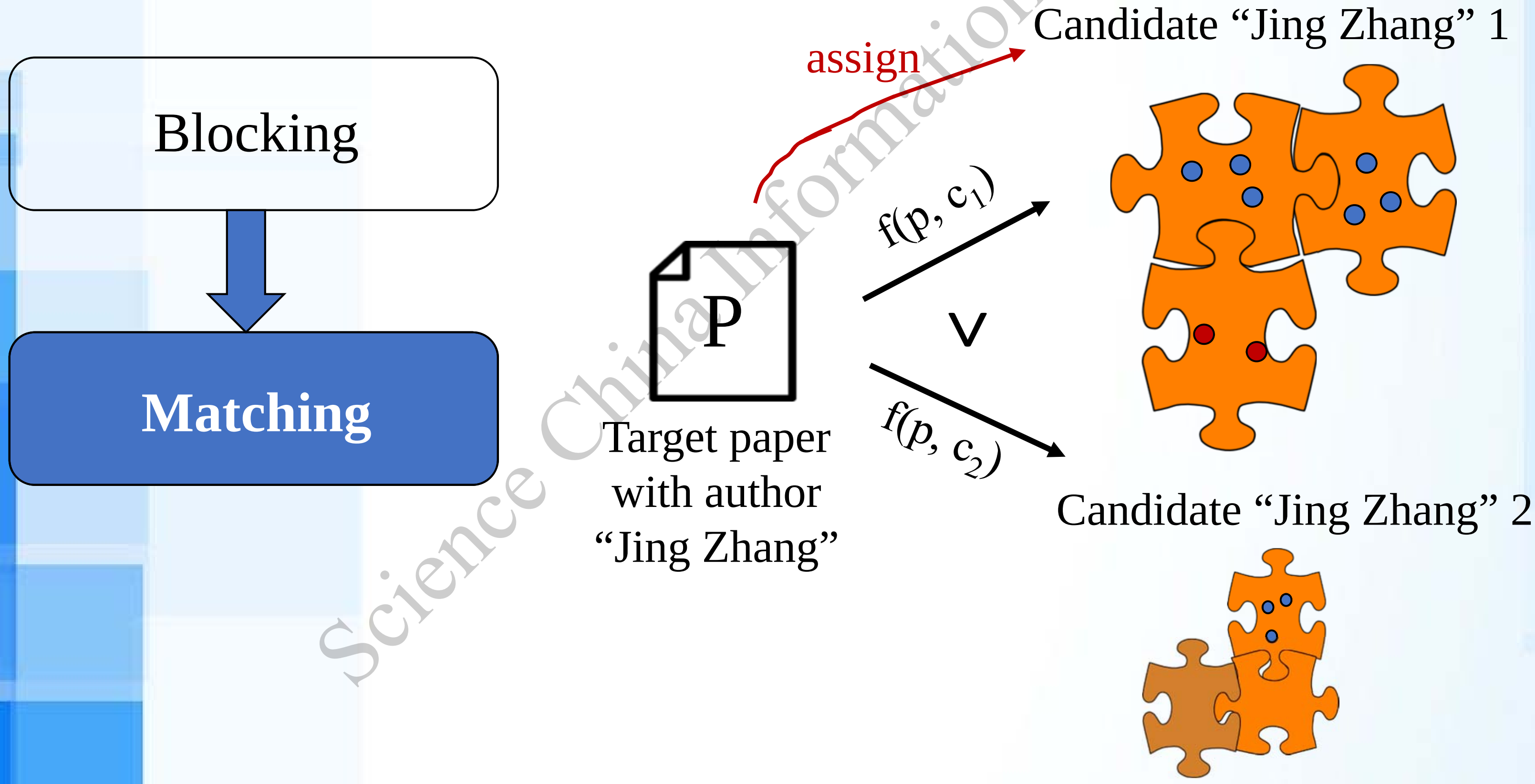
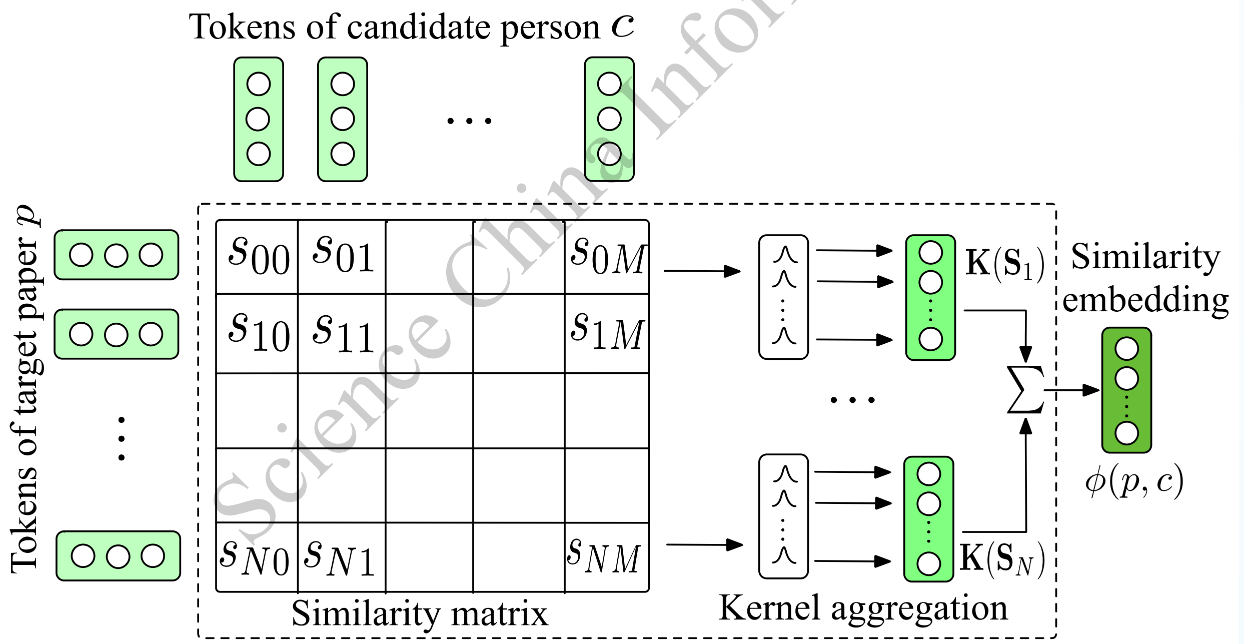


Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

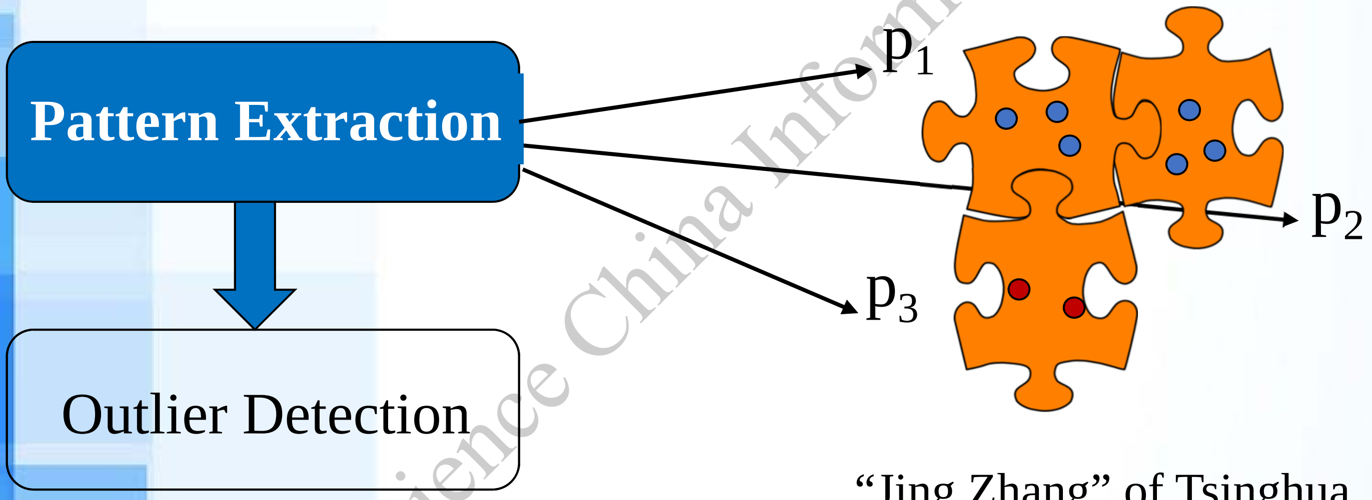
Interaction-based matching

- Calculating the similarities between the embeddings of each pairs of tokens in p and c.
- Capturing both the exact and the soft matches.



3rd Scenario: Error Detection

- The accuracy of ND algorithms cannot be 100%.
- An additional error detection function is needed.
- Stage One: Extract patterns that can distinguish the right (normal) and wrong (abnormal) assigned papers.



"Jing Zhang" of Tsinghua

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in *ICDE'13, VLDB'13*

Pattern Extraction

- Construct a multi-relation egonet for each candidate c.
  - Co-author, co-venue, citation relationships
- For each relation, extract:
  - Number of neighbors of ego c
  - Number of edges in c's egonet
  - Total weight of c's egonet
  - Principal eigenvalue of the weighted adjacency matrix of c's egonet

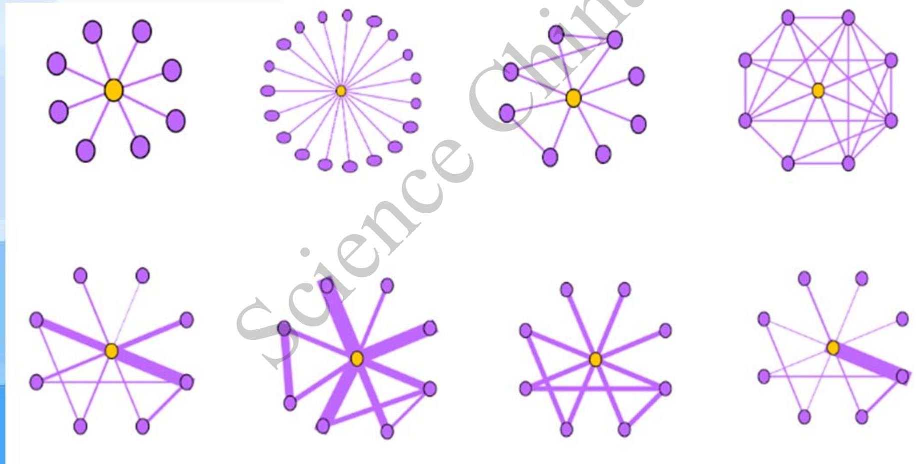
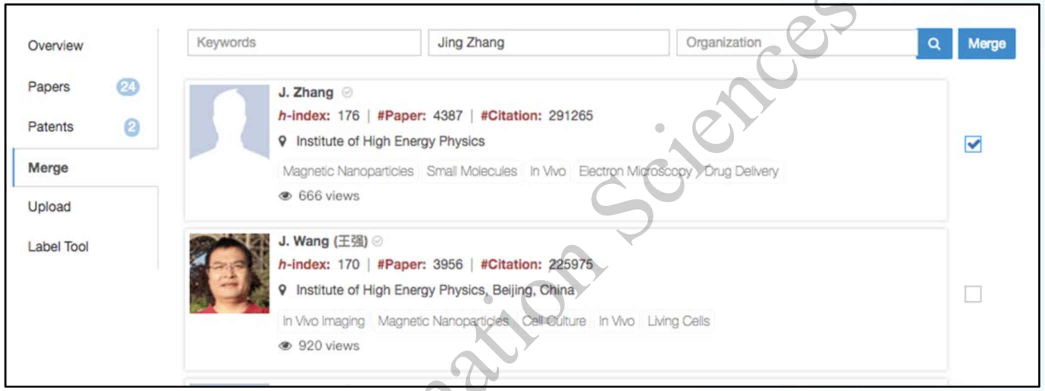


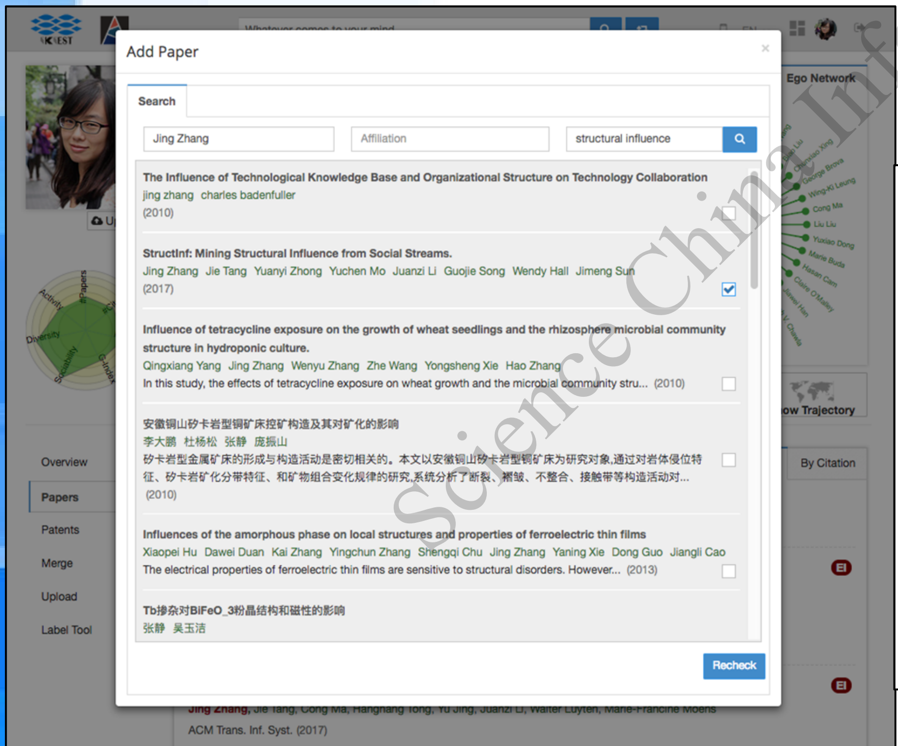
Image Reference: Xin Luna Dong, Christos Faloutsos, et al. Fact Checking: Theory and Practice. Tutorial in *KDD'18*

User Feedbacks

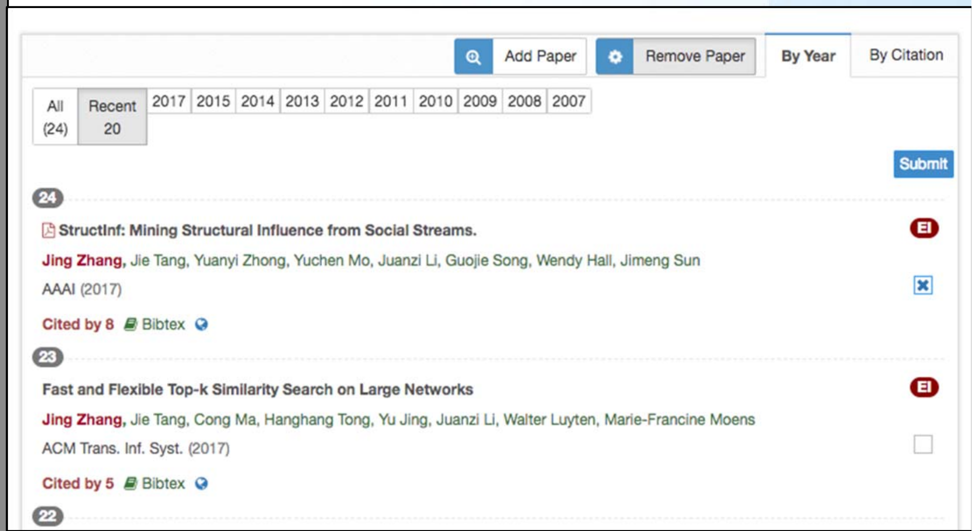
Merge Function:
Merge person profiles



Add function:
Assign new papers to persons



Remove function:
Remove the wrongly assigned papers



Name Disambiguation in AMiner:

Jing Zhang*, Jie Tang+

*Information School, Renmin University
+Computer Science Department, Tsinghua University

Dataset: https://www.aminer.cn/na-data
AMiner: https://www.aminer.cn/

AMiner 清華大學 Tsinghua University 中國人民大學 RENMIN UNIVERSITY OF CHINA