Schema Based Deduplication

Pei Li, Andrea Maurino

SEQUOIAS group – SErvice and Quality Oriented InformAtion Systems
DISCo - Dipartimento di Informatica Sistemistica e Comunicazione
Università di Milano-Bicocca

Contents

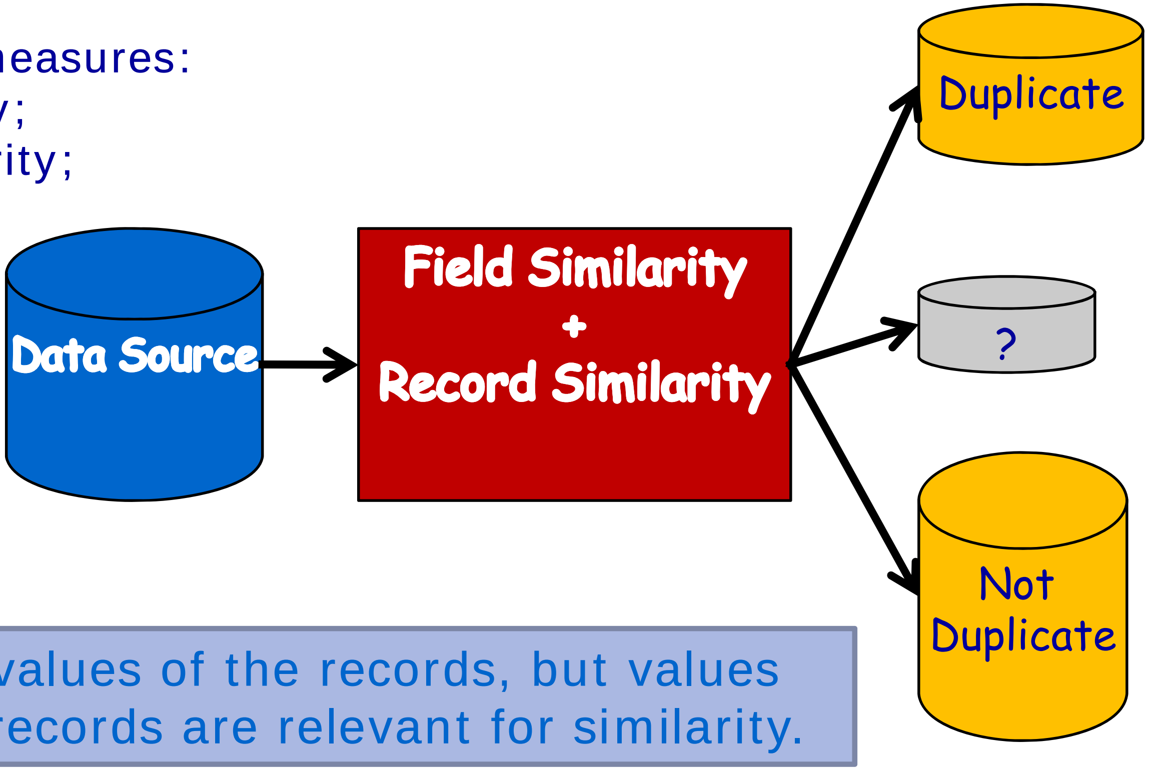
- Duplicate detection
- Motivating Examples
- Schema-Based Deduplication
- Example
- Conclusion

Duplicate detection

Duplicate detection is the discovery of multiple representations of the same real-world object

Deduplication is the discovery of multiple representation of same real-world object on the same table

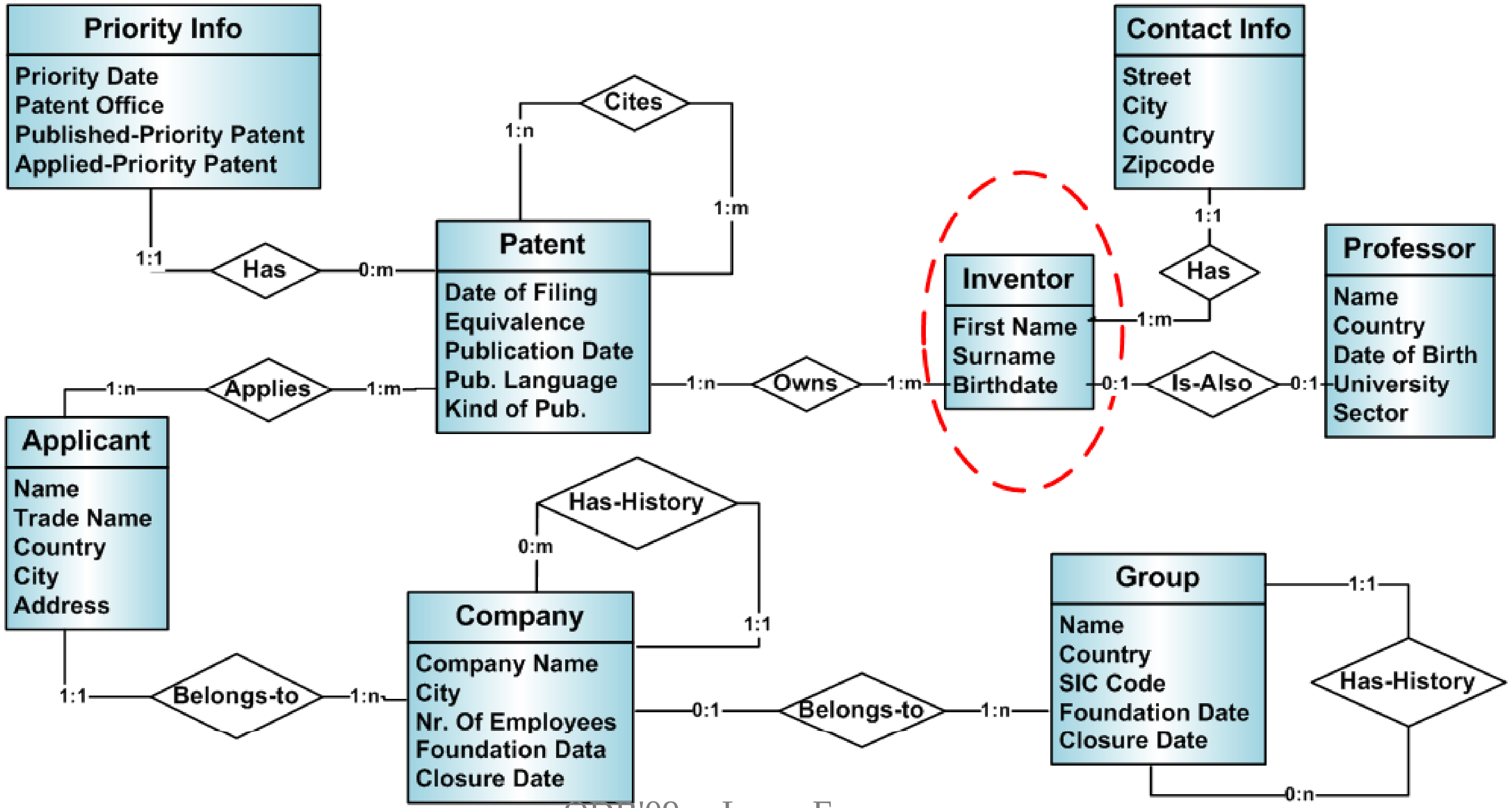
Two similarity measures:
- Field Similarity;
- Record Similarity;



Not only values of the records, but values of set of records are relevant for similarity.

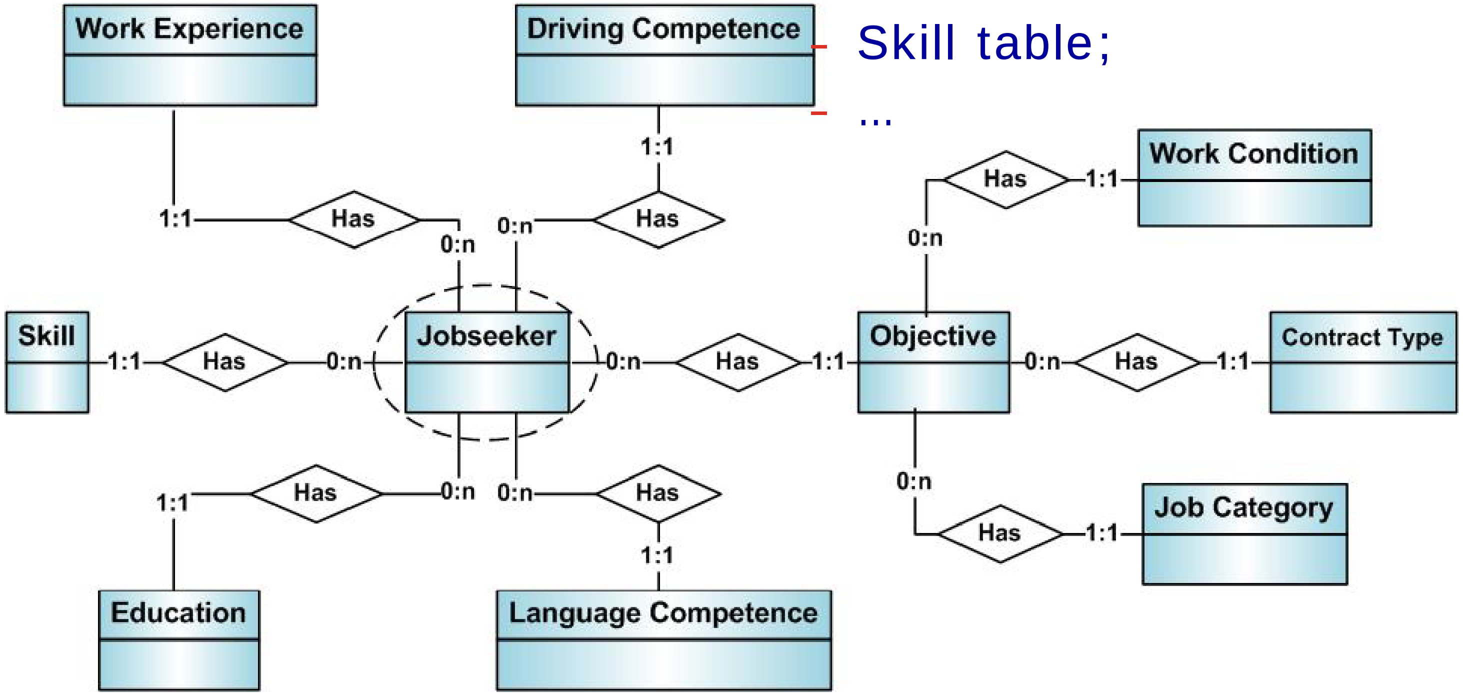
Motivating Examples

- European Paten Office (ESF project APE-INV)

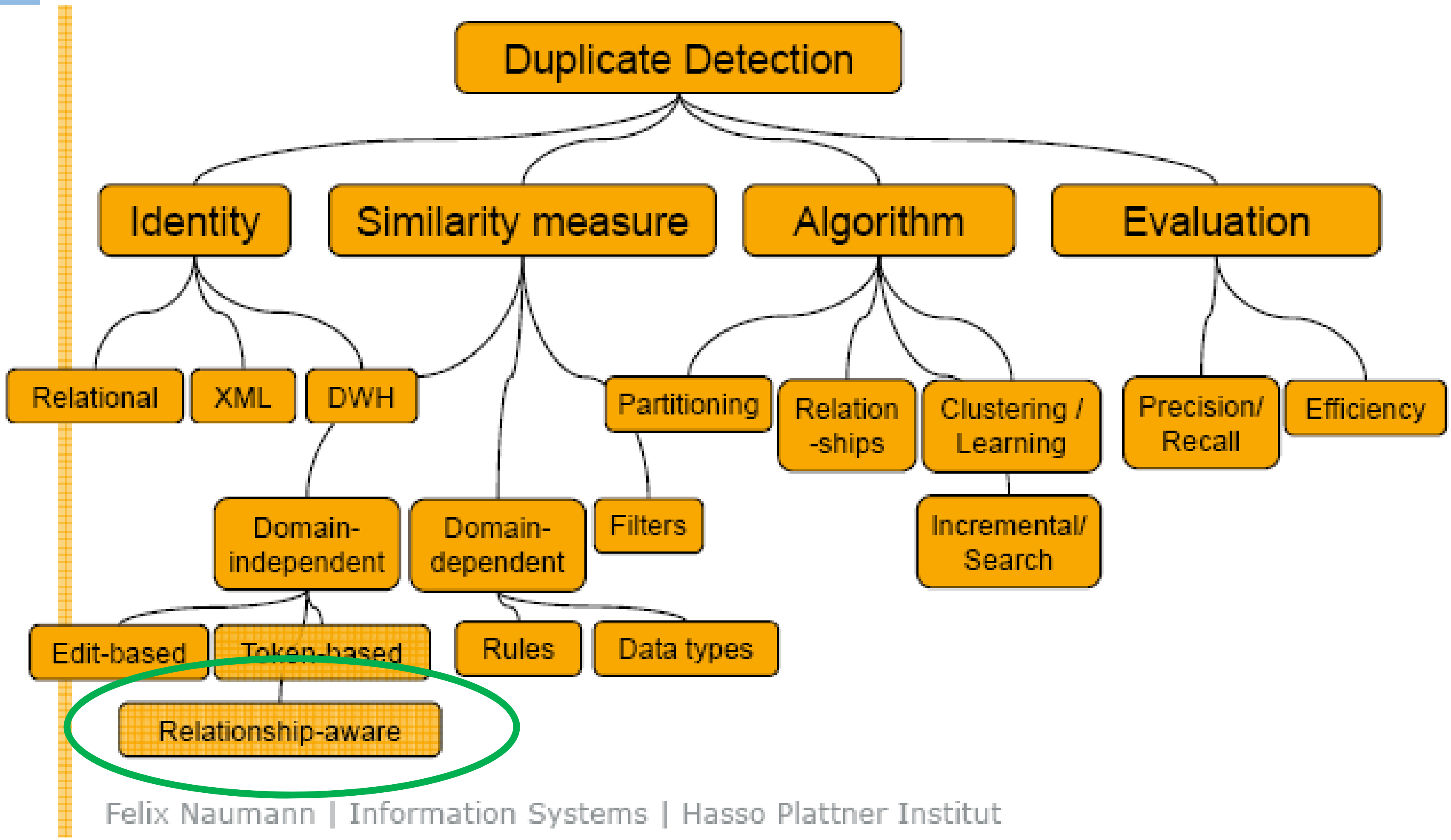


Motivating Examples

- ❖ Job placement Database (FP6 SEEMP project)
  - ✓ Non-anonymous:
    - – Jobseeker table;
  - ✓ Anonymous:
    - – Work Experience table;
    - – Education table;
    - – Skill table;
    - – ...



Where we started



Duplicate Detection

- Identity
  - Relational
  - XML
  - DWH
- Similarity measure
  - Domain-independent
    - Edit-based
    - Token-based
      - Relationship-aware
  - Domain-dependent
    - Rules
    - Data types
  - Filters
- Algorithm
  - Partitioning
  - Relation-ships
  - Clustering / Learning
    - Incremental/ Search
- Evaluation
  - Precision/ Recall
  - Efficiency

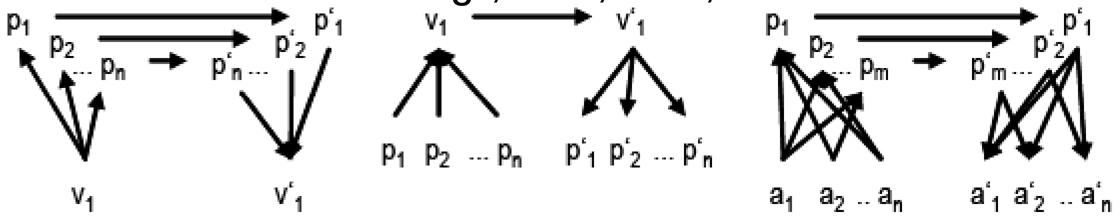
Felix Naumann | Information Systems | Hasso Plattner Institut

Our Contribution

- Our approach are similar to...
  - ✓ Group Linkage (a.k.a. Group ER)
  - ✓ Inter-relationship Deduplication
- But we aim at providing a better solution which is...
  - ✓ General purpose
  - ✓ Exploiting context information via schema analysis
  - ✓ Covering multiple types of record linkage:
    - Dispersed record linkage problem (scattered information)
    - Ambiguous record linkage problem (dirty data)

Knowledge Network

To improve record linkage based on schemas where objects are mapped into each other as e.g., 1:n; n:1; n:m.

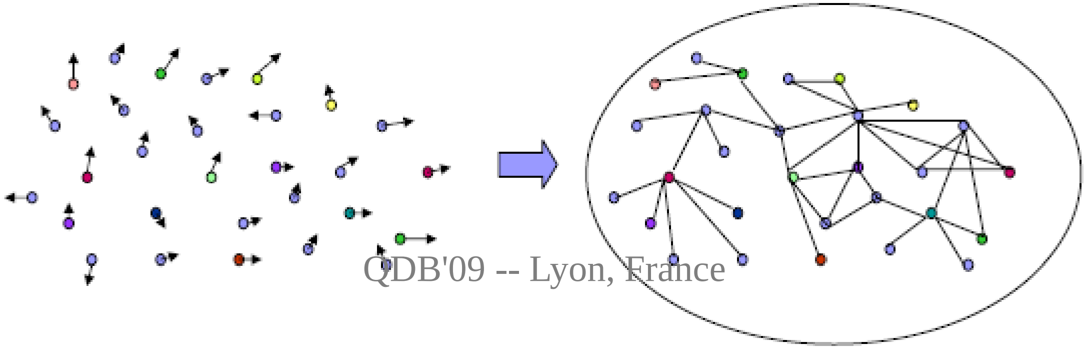


(a) 1:n (b) n:1 (c) n:m

(Venue-Publication) (Publication-Venue) (Author-Publication)

Every object is represented as a knowledge network based on the above schema structure; (scattered information)

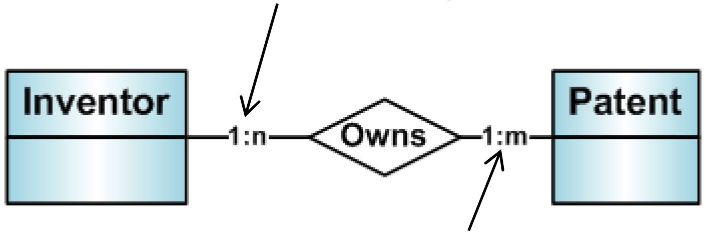
Every tuple, either being a dispersed or ambiguous reference to the object, reflects partially, or an extension of existing knowledge network



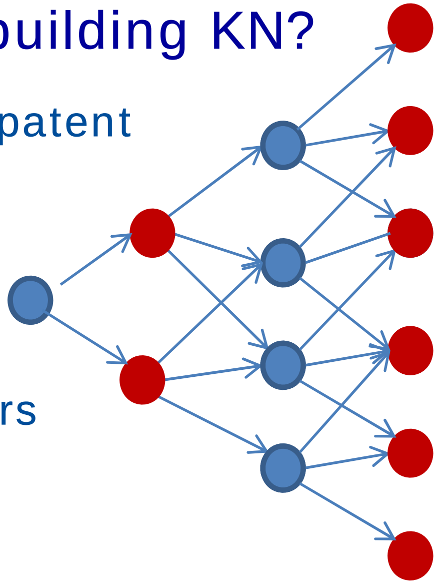
Schema-Based Deduplication

- Why ER cardinalities are considered in building KN?

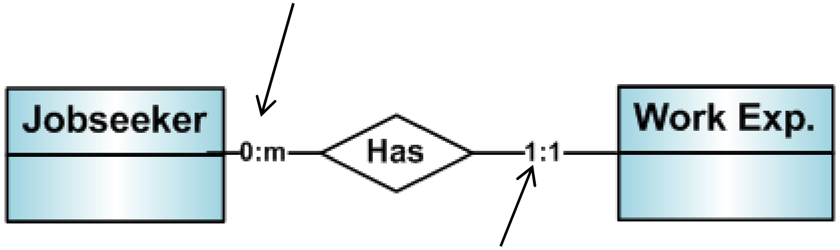
an inventor definitely owns one or more than one patent



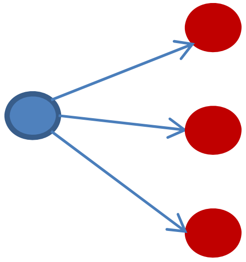
a patent definitely belongs to one or more inventors



a jobseeker could have no work experiences



each work experience is related to exactly one jobseeker



Schema-Based Deduplication

- Similarity Functions:
  - ✓ Many to Many Relationship
  - ✓ Optional Many to One Relationship
  - ✓ Many to One Relationship
  - ✓ One to Many Relationship
  - ✓ (Optional)One to (Optional) One relationship

Schema-Based Deduplication

- Many to Many Relationship
  - ✓ E.g.: *Inventor-Patent, Paper/Patent-Citation*
  - ✓ Similarity Metric: KN G*raph G <V,E> and its subgraph KN Gi<Vi,Ei>*
    - *τ: relevance of v w.r.t. G and Gi,*

$$\tau(v, G_i, G) = \frac{(|\, I_i(v) \,| + |\, O_i(v) \,|) \subseteq G_i}{(|\, I(v) \,| + |\, O(v) \,|) \subseteq G}$$

    - *ρ: sum of relevances w.r.t. all nodes in Gi*

$$\rho(G_i, G) = \sum_{v \subseteq V_i} \tau(v, G_i, G)$$

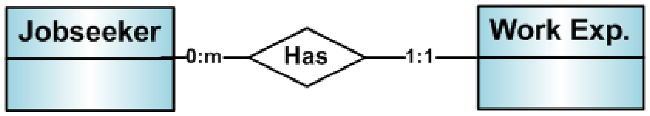
    - *δ: relevance of a set of common subgraphs w.r.t. G and G'*

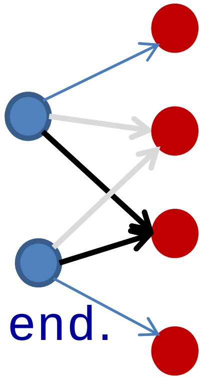
$$\delta(\Gamma) = \frac{\sum(\rho(G_i, G)\rho(G_i, G'))}{|\, V \,||\, V' \,|}$$

s:maximal proportion of all common subgraphs

Schema-Based Deduplication

- Optional Many to One Relationship
  - E.g., *Jobseeker-Work Experience*
  - Similarity metric:
    - SimRank
    - Average similarity score of out-neighbor nodes between to objects
    - Shortest Path in a graph:
    - walk from (a, b) which touches a singleton node at the end and only at the end.





$$s^{n1}(G^a,G^b) = \frac{C_1}{|\,O(G^a)\,||\,O(G^b)\,|} \sum_{i=1}^{|O(G^a)|}\sum_{j=1}^{|O(G^b)|} s^{n1}(O_i(G^a),O_j(G^b))$$

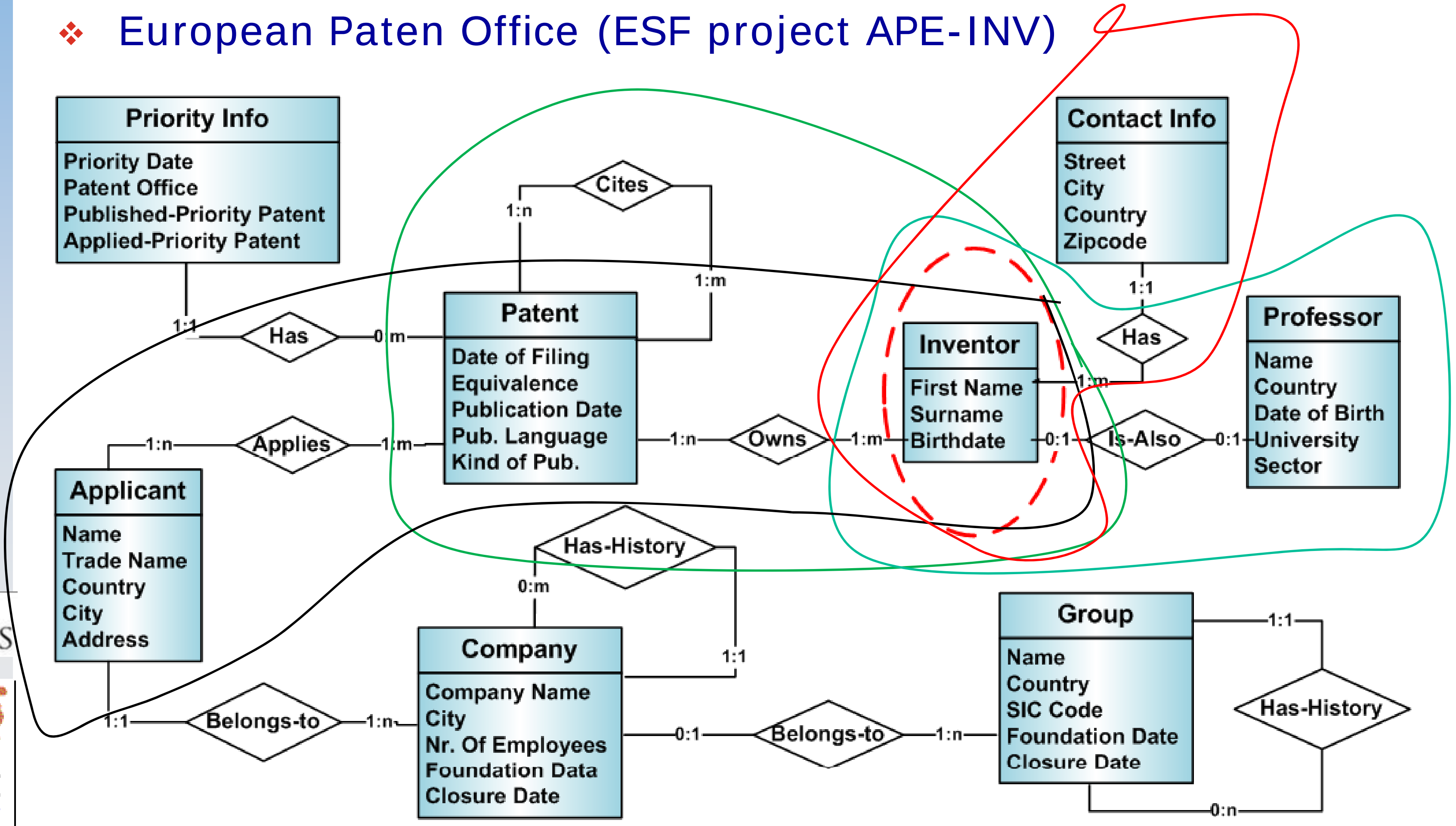
$$s^{n1}(O_i(G^a),O_j(G^b)) = \frac{C_2}{|\,I(O_i(G^a))\,||\,I(O_j(G^b))\,|} \cdot \sum_{n=1}^{|I(O_i(G^a))|}\sum_{m=1}^{|I(O_j(G^b))|} s^{n1}(I_n(O_i(G^a)),I_m(O_j(G^b)))$$

Schema-Based Deduplication

- Many to One Relationship
  - ✓ E.g., *Country-Region-Province-City*
  - ✓ *Similarity Metric:*
    - *Hierarchy Graph*
- One to Many Relationship
  - ✓ E.g., *Kid-Mather (1:1 – 1:n)*
  - ✓ *Similarity Metric:*
    - *many-to-many relationship*
- (Optional)One to (Optional) One relationship
  - ✓ E.g., Inventor-Professor
  - ✓ Similarity Metric:
    - *No similarity metric, merge entities*

Example

- European Paten Office (ESF project APE-INV)



Priority Info
Priority Date
Patent Office
Published-Priority Patent
Applied-Priority Patent

Patent
Date of Filing
Equivalence
Publication Date
Pub. Language
Kind of Pub.

Contact Info
Street
City
Country
Zipcode

Inventor
First Name
Surname
Birthdate

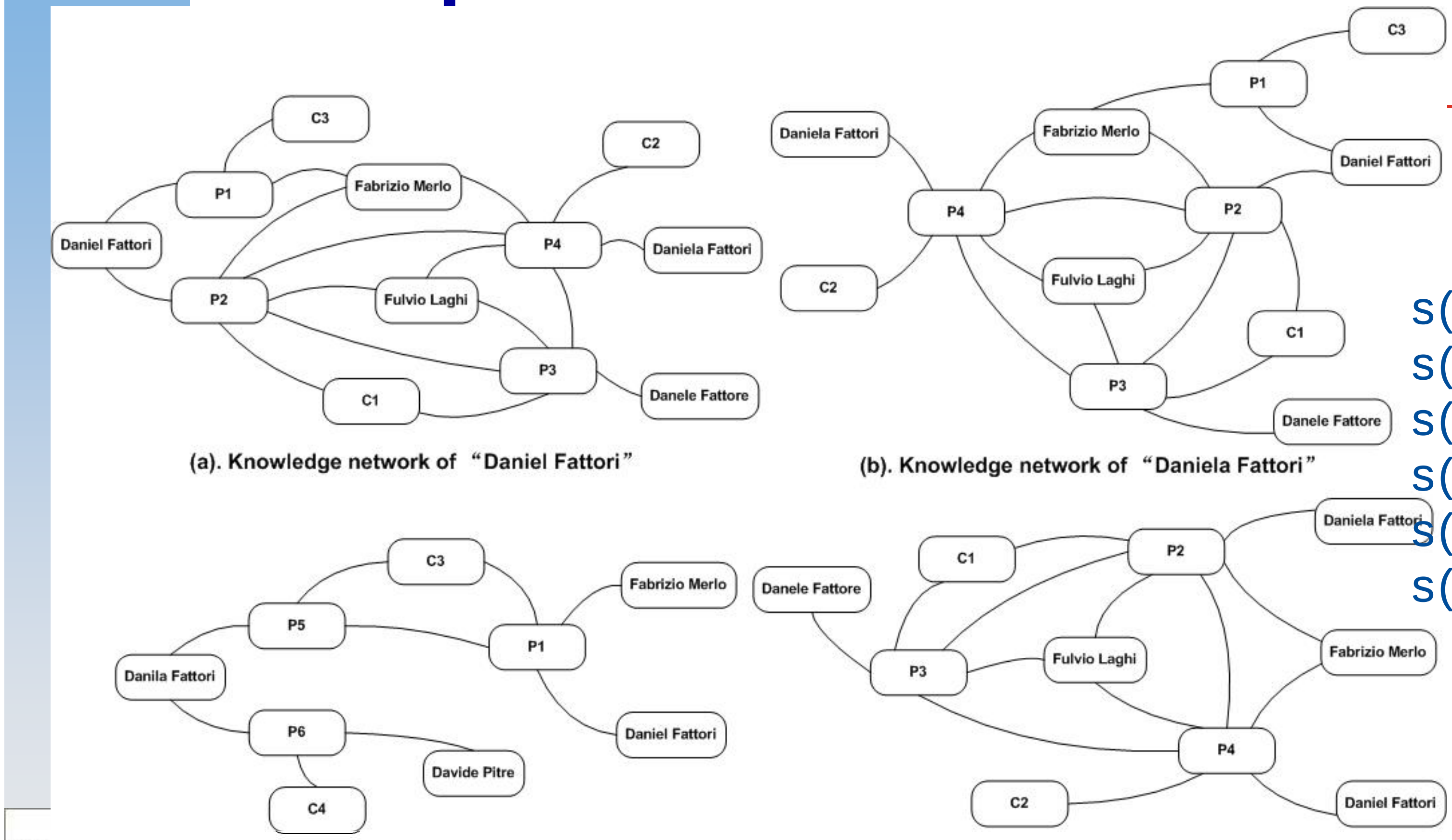
Professor
Name
Country
Date of Birth
University
Sector

Applicant
Name
Trade Name
Country
City
Address

Company
Company Name
City
Nr. Of Employees
Foundation Data
Closure Date

Group
Name
Country
SIC Code
Foundation Date
Closure Date

Example



(a). Knowledge network of "Daniel Fattori"

(b). Knowledge network of "Daniela Fattori"

(c). Knowledge network of "Danila Fattori"

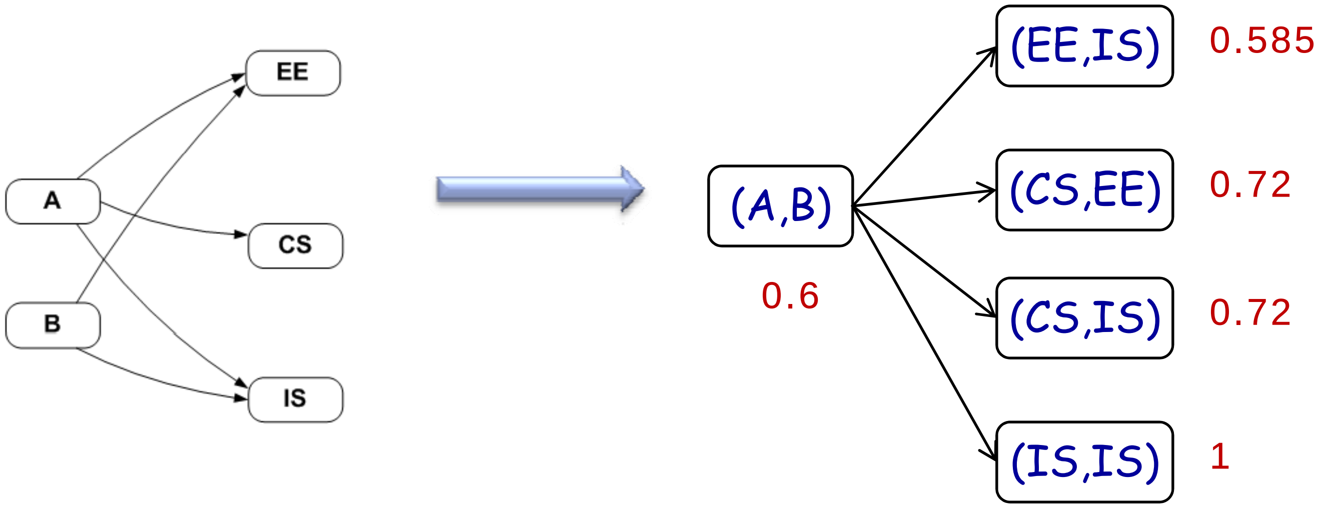
(d). Knowledge network of "Danele Fattore"

s(a,b)= 1
s(a,c)=0.085
s(a,d)= 0.764
s(b,c)= 0.085
s(b,d)=0.764
s(d,c)=0

"Daniel Fattori", "Daniela Fattori" and "Danele Fattore" refer to the same inventor,
"Danila Fattori" probably represents another inventor.

Preliminary Result

- SMEEP Database
  - ✓ the similarity of two jobseekers is improved by the similarity of their education information as an computation example



Discussion and Conclusion

- Some Discussions
  - ✓ Size of KN (small) (remember 7 persons distance)
  - ✓ Computing Efficiency
  - ✓ False Positive

- Future Work
  - ✓ Deeper analysis of all kinds of relationships;
  - ✓ Optimization techniques for prerequisite blocking;
  - ✓ Investigating the performance of different subgraph detection algorithms;
  - ✓ Evaluation of efficiency and effectiveness.

Thank you! Questions?