Extending Functional Dependency to Detect Abnormal Data in RDF Graphs

Yang Yu, Jeff Heflin
SWAT Lab
Department of Computer Science and Engineering
Lehigh University
PA, USA

Outline

- Motivation
- Functional Dependency (FD)
- Value-clustered Graph Functional Dependency (VGFD)
- System
  - Clustering
    - Pre-clustering
    - Optimal k-Means clustering
  - Handling multi-valued properties
  - Static pruning
  - Runtime pruning
- Conclusion and future work

Motivation

- Low data quality is a pressing problem for information integration
- Numerous problems could occur during the data generation process using a variety of tools
- Question answering needs to be more accurate and reliable
- Trust or other assessments require foundational input

Functional Dependency in DB

- It is by far the most common integrity constraints for databases.
- It is devised to specify missing semantics in mere syntactic definitions of database relations.
- Definition
  - **Given a relation *R*, a set of attributes *X*** in *R* is said to functionally determine another attribute *Y*, also in *R*, (written $X \rightarrow Y$) if, and only if, **each *X* value** is associated with precisely **one *Y* value**.
- E.g. {Course, Semester} → Instructor

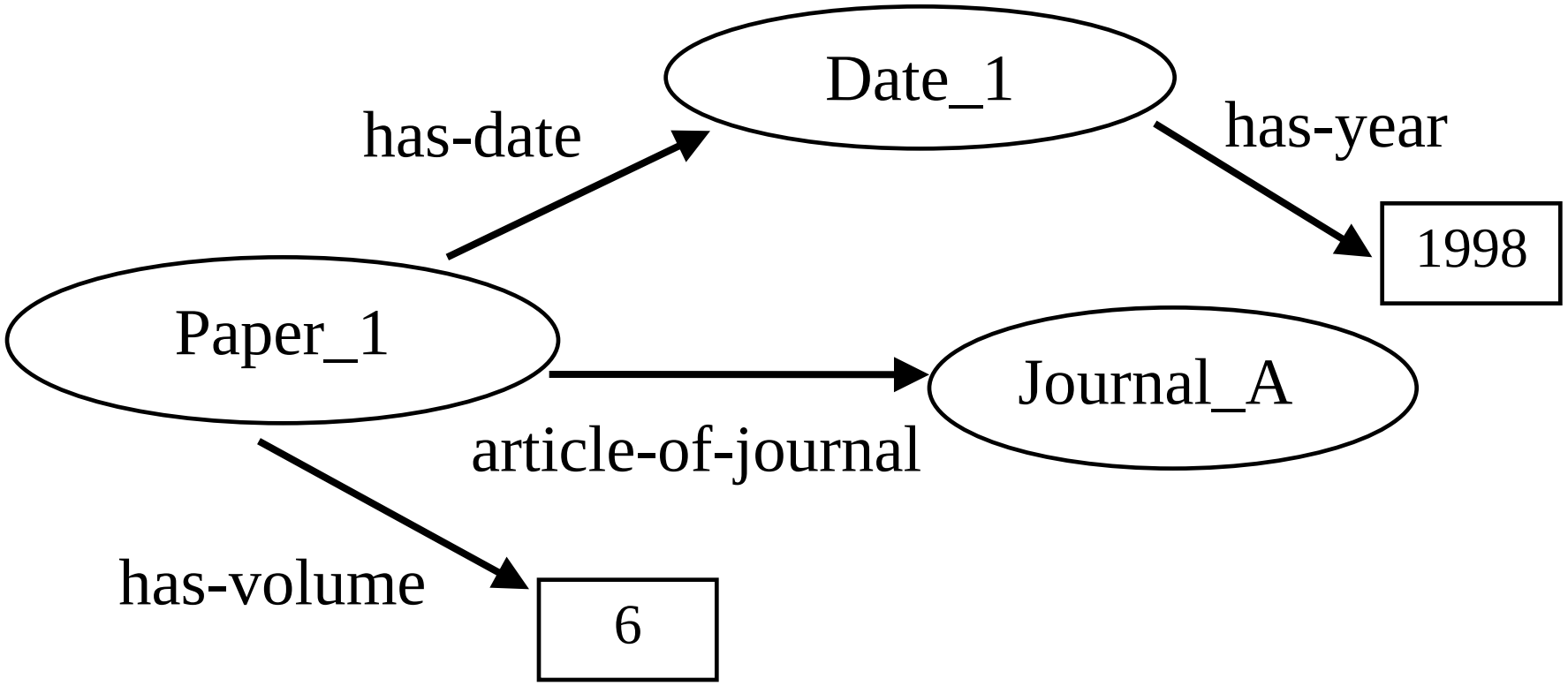
  {postalCode} → province

Challenges

- RDF data can be viewed as extremely decomposed relational tables
- More forms of combinations of properties convey latent semantics
  - E.g. *advisorOf* ∘ *studentOf* $^{-}$, *father* ∘ *brother*
- Multi-valued properties make it harder to determine value correlations
  - E.g. *city* and *province*
- Different syntactic values could be (almost) logically equivalent
  - E.g. 1.000001 vs. 1.0; “blue & red” vs. “red & blue”

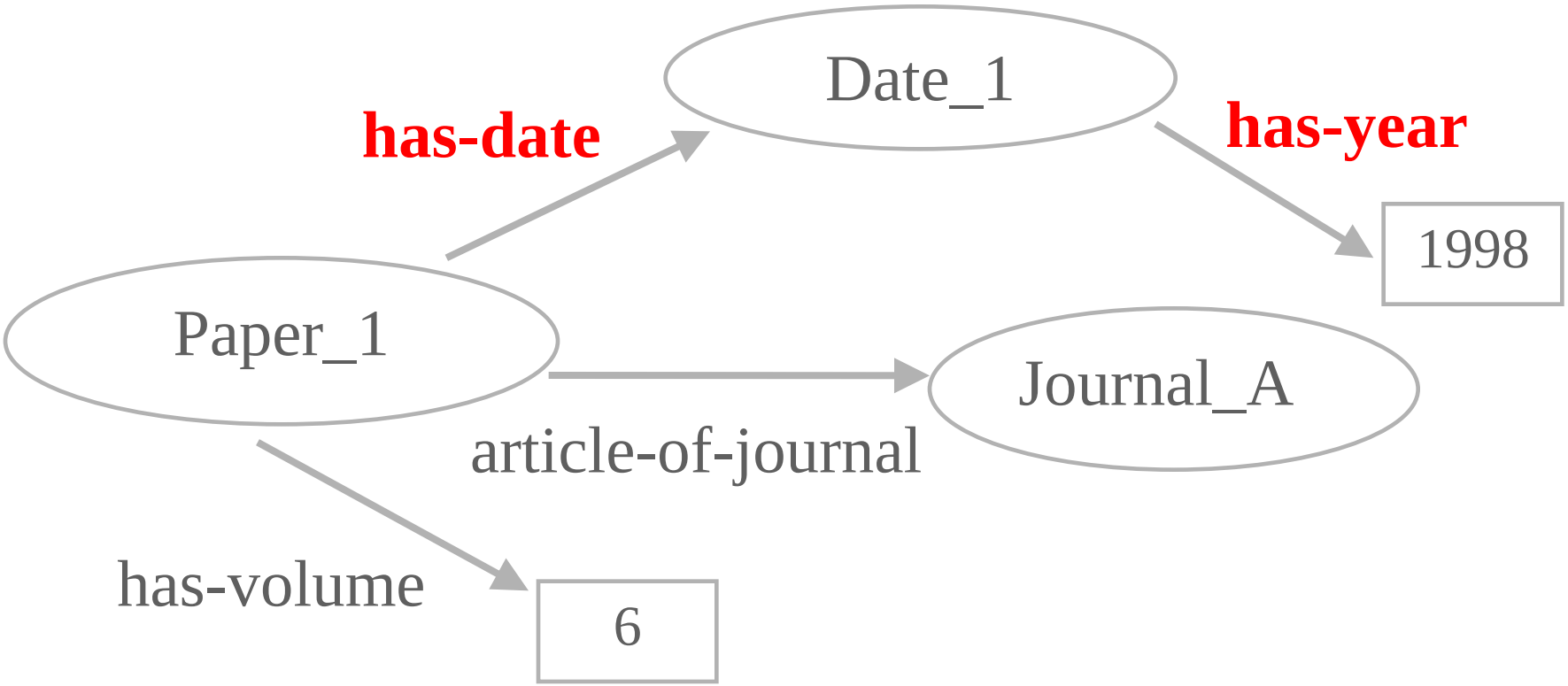
    upper ages of a certain type of school are in a small range

VGFD Definitions



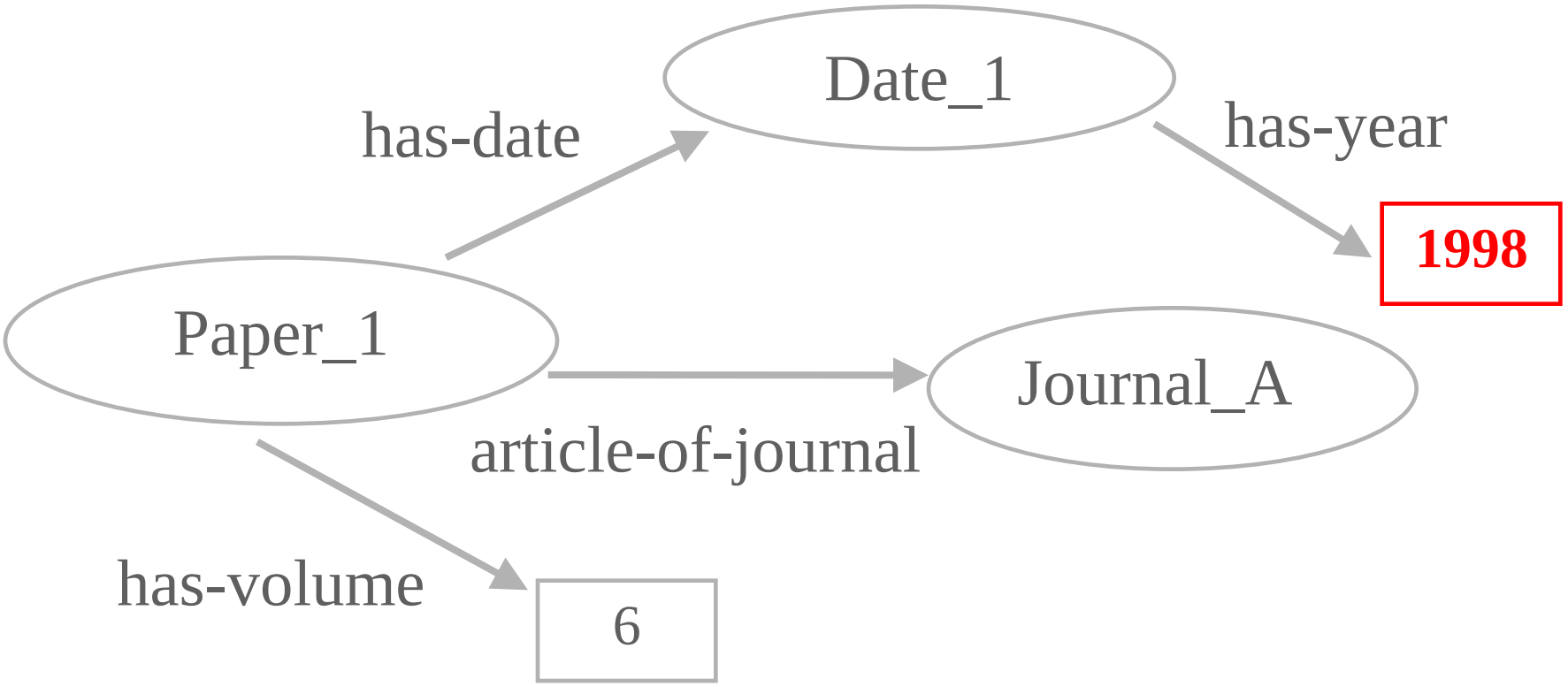
An RDF graph: directed, labeled graph

VGFD Definitions



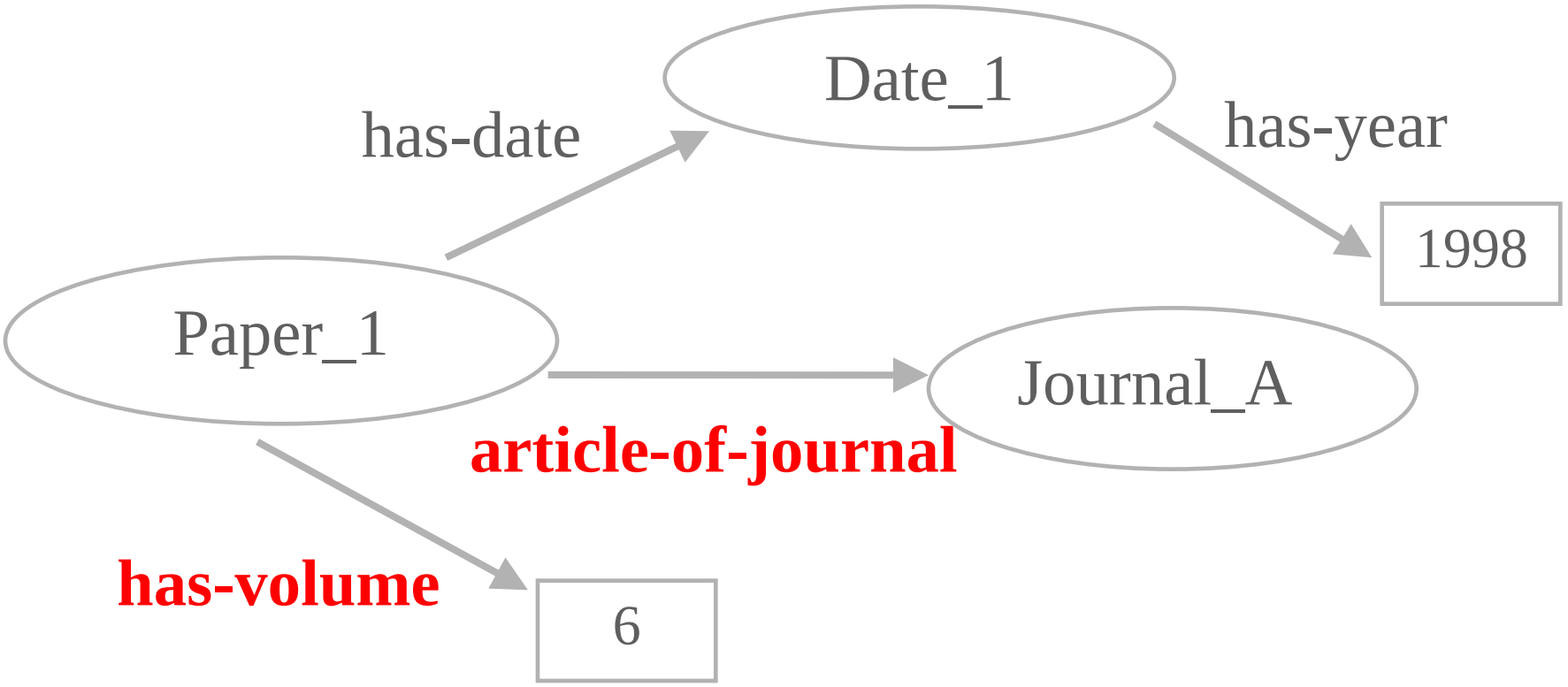
An example Composite Property: has-date ◦ has-year

VGFD Definitions



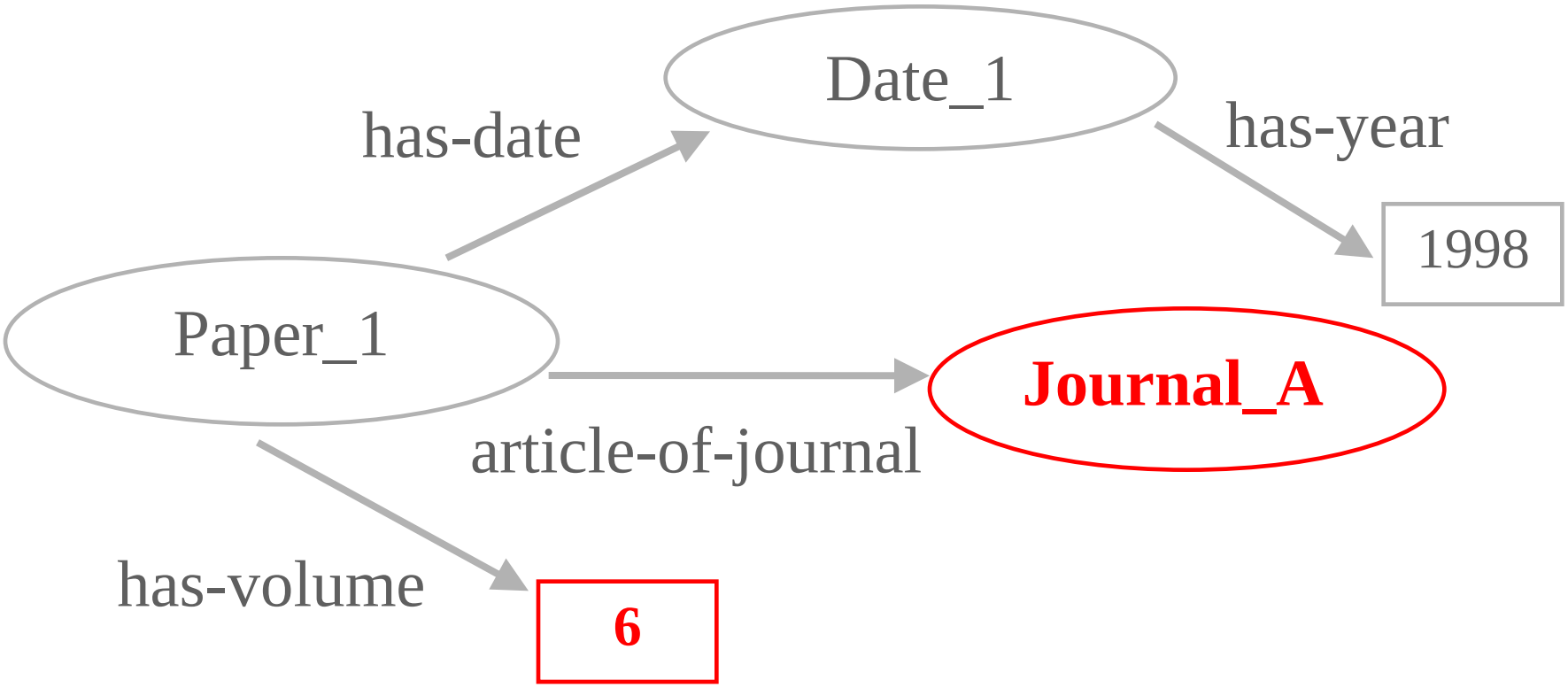
Value(Paper_1, has-date ◦ has-year) = 1998

VGFD Definitions



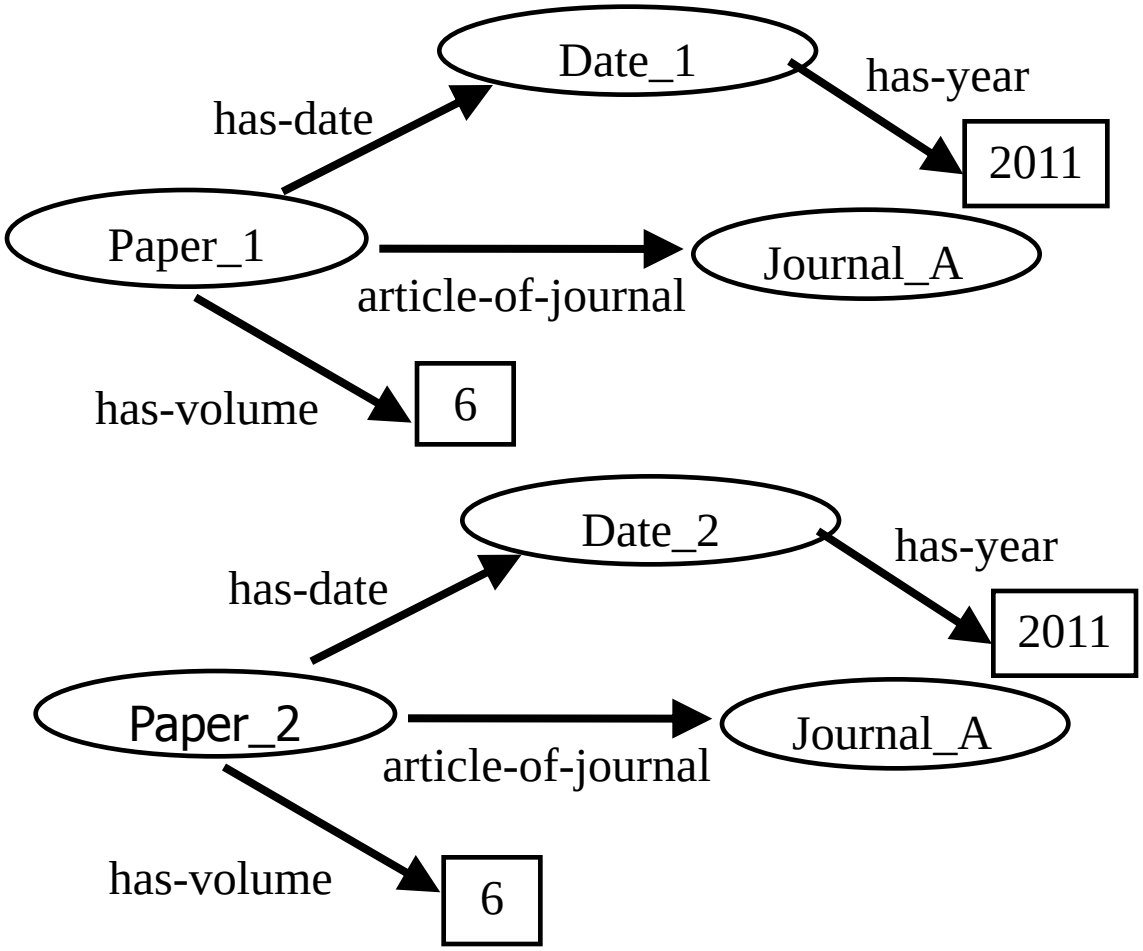
An example Conjunctive Property:
article-of-journal + has-volume

VGFD Definitions



Value(Paper_1, article-of-journal + has-volume) = (Journal_A, 6)

VGFD Definitions



A possible example VGFD:

article-of-journal + has-volume → has-date ◦ has-year

Handling Multi-valued Property

deptNo		deptName	
subject	object	subject	object
A	1	A	CS
A	2	A	EE
B	1	B	EE
C	2	C	CS
D	2	D	EE

Handling Multi-valued Property

deptNo		deptName	
subject	object	subject	object
A	1	A	CS
A	2	A	EE
B	1	B	EE
C	2	C	CS
D	2	D	EE

Candidate Value Mapping	Count
1→ EE	2
2→ EE	2
2→ CS	2
1→ CS	1

Handling Multi-valued Property

deptNo		deptName	
subject	object	subject	object
A	1	A	CS
A	2	A	EE
B	1	B	EE
C	2	C	CS
D	2	D	EE

Candidate Value Mapping	Count
1→ EE	2
2→ EE	2
2→ CS	2
1→ CS	1

Handling Multi-valued Property

deptNo		deptName	
subject	object	subject	object
A	1	A	CS
A	2	A	EE
B	1	B	EE
C	2	C	CS
D	2	D	EE

Candidate Value Mapping	Count
1→ EE	2
2→ EE	2
2→ CS	2
1→ CS	1

Handling Multi-valued Property

deptNo		deptName	
subject	object	subject	object
A	1	A	CS
A	2	A	EE
B	1	B	EE
C	2	C	CS
D	2	D	EE

Miss Match

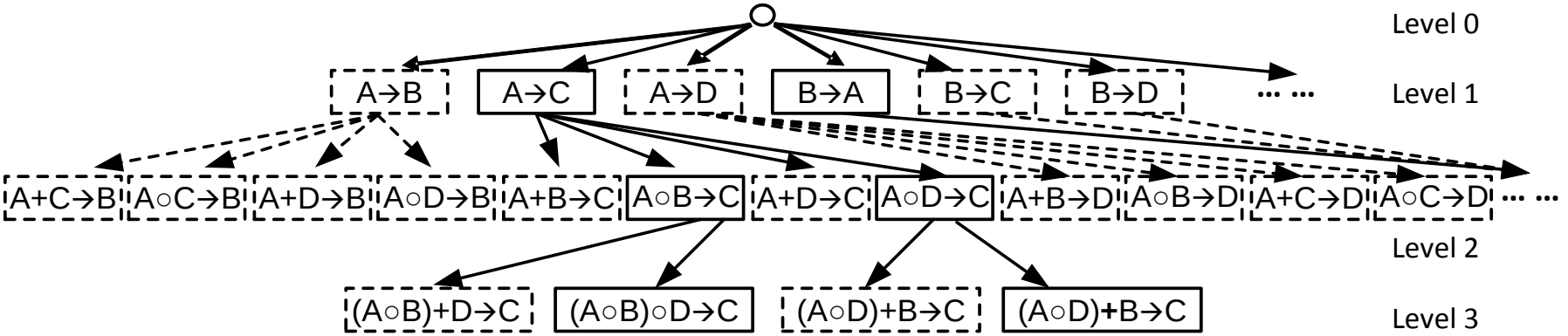
Candidate Value Mapping	Count
1→ EE	2
2→ EE	2
2→ CS	2
1→ CS	1

Confidence of VGFD deptNo → deptName = 4/5 = 0.8

Static Pruning Heuristics

- Insufficient subject/object overlap between properties to be composed or between the LHS and the RHS
- The discriminability of the LHS is less than that of the RHS
  - Discriminability of a property = # distinct values / the size of property extension
- The LHS or RHS has too high a discriminability
- The discriminability of a Composite Property (Conjunctive Property resp.) is no greater than (no less than resp.) that of each property involved.

Level-wise Discovering Process



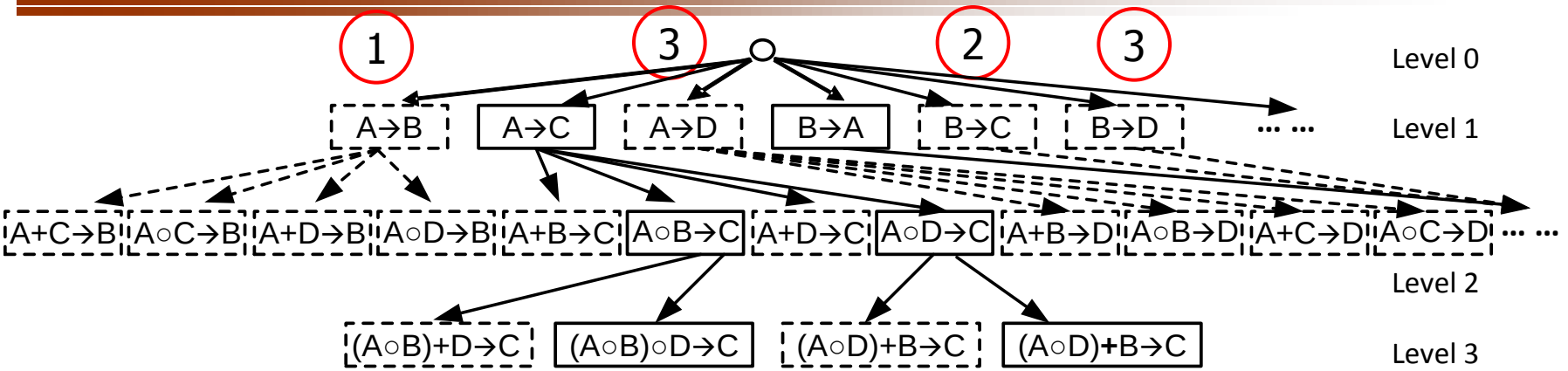
Static pruning heuristics:

1. insufficient subject/object overlap between properties to be composed or between the LHS and the RHS,
2. the discriminability of the LHS is less than that of the RHS,
3. the LHS or RHS has too high a discriminability,
4. the discriminability of a Composite Property (Conjunctive Property resp.) is no greater than (no less than resp.) that of each property involved

In this example, we suppose:

a. *A* and *B* have few common subjects
b. the discriminability of *B* is less then that of *C*
c. *D* has a high discriminability

Level-wise Discovering Process



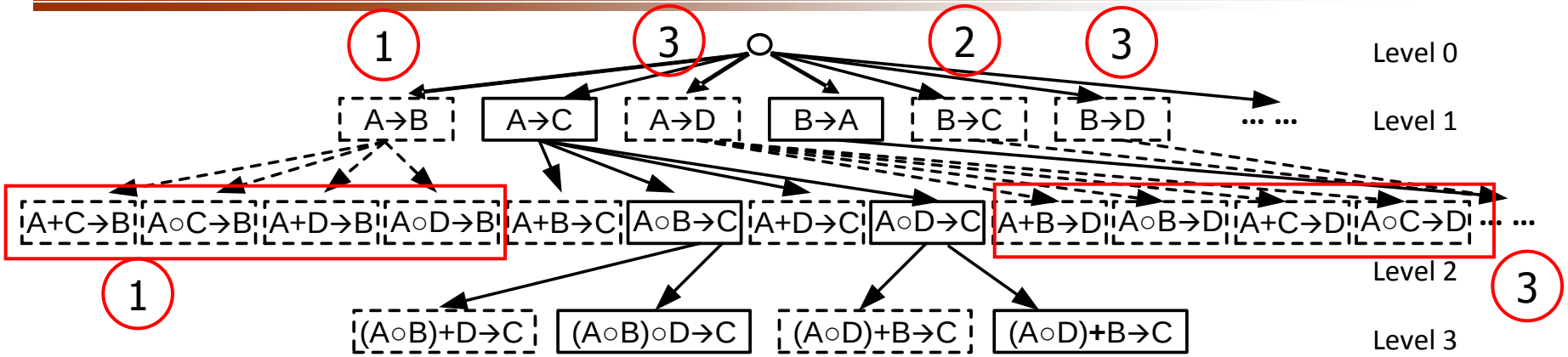
Static pruning heuristics:

1. insufficient subject/object overlap between properties to be composed or between the LHS and the RHS,
2. the discriminability of the LHS is less than that of the RHS,
3. the LHS or RHS has too high a discriminability,
4. the discriminability of a Composite Property (Conjunctive Property resp.) is no greater than (no less than resp.) that of each property involved

In this example, we suppose:

a. *A* and *B* have few common subjects
b. the discriminability of *B* is less then that of *C*
c. *D* has a high discriminability

Level-wise Discovering Process



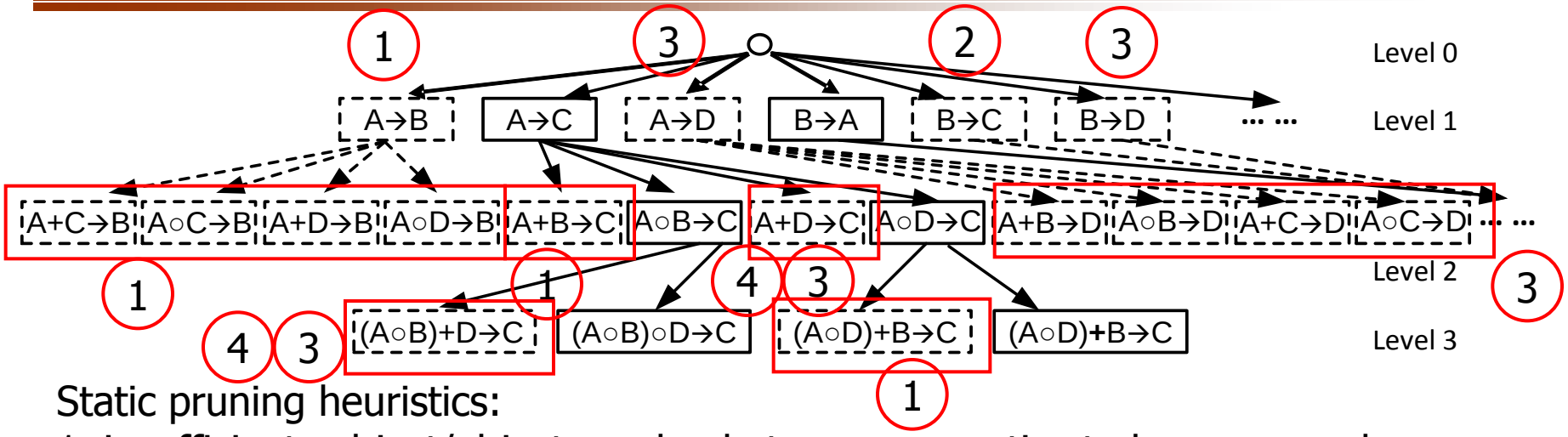
Static pruning heuristics:

1. insufficient subject/object overlap between properties to be composed or between the LHS and the RHS,
2. the discriminability of the LHS is less than that of the RHS,
3. the LHS or RHS has too high a discriminability,
4. the discriminability of a Composite Property (Conjunctive Property resp.) is no greater than (no less than resp.) that of each property involved

In this example, we suppose:

a. *A* and *B* have few common subjects
b. the discriminability of *B* is less then that of *C*
c. *D* has a high discriminability

Level-wise Discovering Process



Static pruning heuristics:

1. insufficient subject/object overlap between properties to be composed or between the LHS and the RHS,
2. the discriminability of the LHS is less than that of the RHS,
3. the LHS or RHS has too high a discriminability,
4. the discriminability of a Composite Property (Conjunctive Property resp.) is no greater than (no less than resp.) that of each property involved

In this example, we suppose:

a. *A* and *B* have few common subjects
b. the discriminability of *B* is less then that of *C*
c. *D* has a high discriminability

Runtime Pruning

- Mutual Information

$$I_{XY} = \sum_i \sum_j p_{i,j} \log (p_{i,j}/p_i p_j)$$

$$\begin{aligned} E_X + E_Y - E_{XY} &= \sum_i p_i \log 1/p_i + \sum_j p_j \log 1/p_j - \sum_i \sum_j p_{i,j} \log 1/p_{i,j} \\ &= \sum_i \sum_j p_{i,j} [\log p_{i,j} - \log p_i - \log p_j] = \sum_i \sum_j p_{i,j} \log (p_{i,j}/p_i p_j) \end{aligned}$$

- Entropy coefficient (EC)

$$EC(X|Y) = I_{XY}/E_Y$$

- Sampling a subset of instances for a VGFD to compute EC

Clustering

- Pre-clustering
  - Light-weight computation provides two benefits for finer clustering
    - The minimum number of clusters
    - Reserves expensive distance calculation for pairs of points within the same pre-cluster
  - Process
    - Pick a point that is closest to the center
    - Cluster points that are close to this center
    - Repeat the above two steps until all points are clustered

Clustering

- Optimal k-Means Clustering
  - Improved from Gap Statistic,
    - Estimating the number of clusters in a data set via the gap statistic, R. Tibshirani, G. Walther, T. Hastie, *Journal of the Royal Statistical Society*, Vol. 63, No. 2. (2001), pp. 411-423.
  - Without the input of k, automatically find the optimal number of clusters
  - Variation of normal k-Means clustering by restricting the distance calculation only on the pairs that within the same pre-cluster

Experimental Results

- Example VGFDs

VGFD	Description
genus→family	Organisms in the same genus also have the same family.
writer→genre	A work's writer determines the work's genre.
teamOwner→chairman	The teams with the same owner also have the same chairman.
composer→mediaType	The works by the same composer have the same media type.
militaryRank→title	The people of the same military rank also have the same title.
location→nearestCity	The things on the same location have the same nearest city.
topic→primaryTopic	The papers with the same topic have the same primary topic.
manufacturer+oilSystem →compressionRatio	The manufacturer and oil system determine the engine's compression ratio.
publisher ∘ country →language	The publisher's country determines the language of that published work.
article-of-journal+has-volume →has_date	The volume number of a journal where an article is published determines the published date of this article.
faculty→budget	The size of the faculty determines the budget range.
militaryRank→salary	The military rank determines the range of salary.
occupation→salary	The occupation determines the range of salary.
type→upperAge	A school's type determines the range of upper age.

Experimental Results

- Example Erroneous Data

1	<r:Shanghai_Jiao_Tong_University, o:university/undergrad, 194323445>
2	<r:Harrow_College, o:School/upperAge, 2009.0>
3	<r:Melbourne_Grammar_School, o:School/ranking, 2006.0>
4	<r:Google_Maps, o:Work/language, r:Coverage_details_of_Google_Maps>
5	<r:Wiktionary, o:Work/language, r:History_and_development>
6	<r:Dembela, o:Place/coordinates, coord|N|W>
7	<r:Hutt_Valley_High_School, o:EducationalInstitution/principal, r:2008>
8	<r:Wake_Island, o:Island/country, r:United_States_Air_Force>
9	<r:Albuquerque_Plaza, o:Building/floorCount, 2221>
10	<r:varedo, o:City/province, r:Province_of_Milan>
11	<i:796511, p:has-date, to-10-01>
12	<i:journals/jair/DarwicheP97, p:has-date, 1998>
13	<s:person/bastian-quilitz, s:ns/swc/ontology#affiliation, research assistant>
14	<s:person/ulf-leser, s:ns/swc/ontology#affiliation, professor>

- System performance

Dataset	SWRC	DBPedia	RKB
Precision	55%	86%	62%

Conclusion and Future Work

- Conclusion
  - Finding corroboration patterns in the form of Value-clustered Graph Functional Dependency
  - Detect abnormal data by checking conflicts with patterns
  - Experiments on different real world data sets that validate the system
- Future work
  - Further generalize the VGFDs
  - Apply system on knowledge acquisition areas

Thank you

Questions and comments?

Integrity Constraints

- Valuable in checking and enforcing data consistency
- Providing further semantics on data, and promoting semantic query optimization
- Open World Assumption (OWA) and the Non-Unique Name Assumption (NUNA) make them harder to be enforced
- How to automatically discover approximate integrity constraints?

System Process Overview

- 1. Find natural clusters of property values
- 2. List all possible combinations of properties on the LHS and RHS of VGFDs
- 3. Static pruning on all possible VGFDs
- 4. Runtime pruning of VGFDs
- 5. Compute the candidate VGFDs

Optimal k-Means Clustering

Algorithm 7 $Optimal\_kMeans(L, groups)$, L is a set of literal values; $groups$ is a set of pre-clustered groups of L.

1: $k = |groups|$
2: $Gap(k) = Gap\_Statistic(groups)$
3: $tmpC \leftarrow groups$
4: **repeat**
5: $k = k + 1$, $C \leftarrow tmpC$, $tmpC \leftarrow \emptyset$ //$tmpC$ is the set of k clusters
6: **for each** $i \leq k$ **do**
7: $Init(m_i), c_i \leftarrow c_i \cup m_i, tmpC \leftarrow tmpC \cup c_i$ //m_i is the center of each cluster
8: **repeat**
9: **for each** $x \in L$ **do**
10: $c_i \leftarrow c_i \cup \arg\min_{m_i \in Group(x)} Distance(x, m_i)$
11: **for each** $i \leq k$ **do**
12: $m_i = Mean(c_i)$
13: **until** $\forall i \leq k, m_i$ converges
14: $Gap(k) = Gap\_Statistic(tmpC)$
15: **until** $Gap(k) < Gap(k-1)$
16: **return** C

Initialize the estimated cluster center

Find the closest center within the same pre-cluster

Re-compute the center

Check if the gain of this clustering is slower than the expected rate

Pre-clustering

- Pre-clustering on strings
  - The weight of a word is based on its frequency in the values of the property
  - The string that has the largest sum of weights divided by the number of words in it is the center
  - Example pre-cluster results:

    {Coach, First Team Coach, Head coach, Senior Coach, Acting head coach, Reserve team coach, Chief Coach},

    {Director, Sports Director, Technical Director, Director general}