

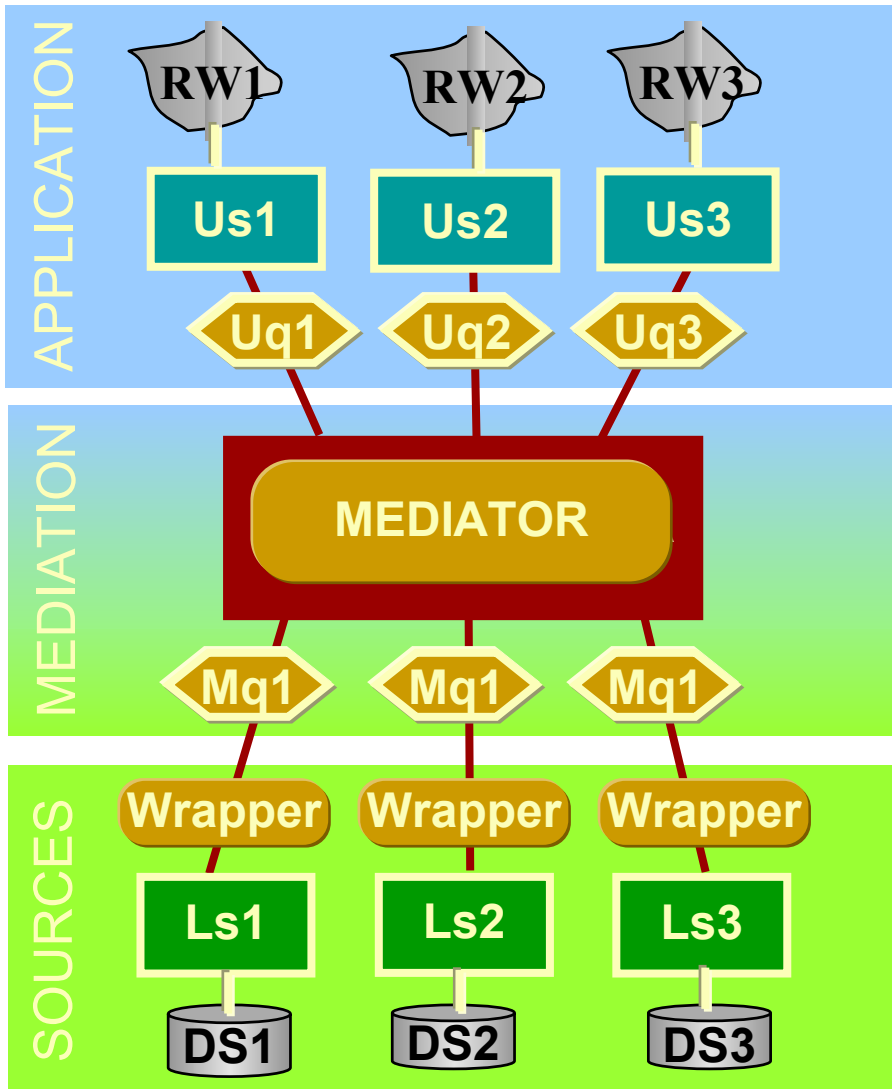
Discovering Mediation Queries from Metadata

Zoubida KEDAD

Laboratoire PRiSM

Université de Versailles St Quentin en Yvelines

Research Problems During Design



Mediation Schema design

Mediation query generation (GAV)

Data cleaning & reconciliation

Mediat. Sch/Query Evolution (GAV)

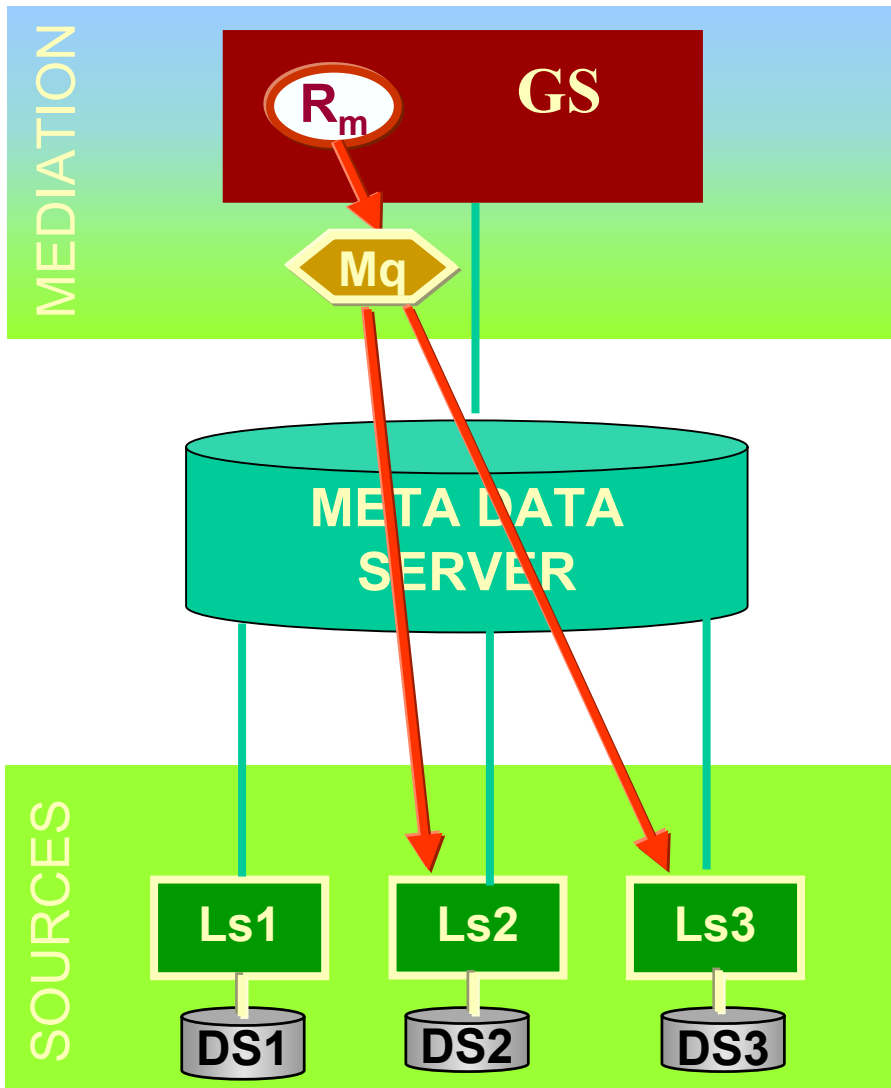
User query rewriting (LAV)

Query optimization (LAV & GAV)

Data materialisat. & refreshment

Data quality

Objective



- Assuming the existence of a meta data server which provides
 - Sources descriptions (schemas)
 - Structural and linguistic mappings between the GS and LS
 - [possibly some inter-sources mappings]
- Given a relation R_m in the global schema, defined by
 - Its schema
 - Its keys and foreign keys,
 - A set of functional dependencies
- Pb: how to generate the mediation query M_q over a set of heterogeneous sources ?

Description of the metadata

➤ Description of data sources

✓ Relational schemas

✓ Assertions :

▪ Intra-schema assertions :

- *Functional Dependencies* ($R_1.K \rightarrow R_1.A$)
- *Referential Constraints* ($R_1.A \Rightarrow R_2.A$)
- *Value Constraints* ($R_1.A < R_1.B$)

▪ Inter-schema assertions :

- *Semantic Equivalence* ($R_1.A \equiv R_2.A$)
- *Semantic Equivalence of key instances*
 $I(R_1) = I(R_2), I(R_1) \cap I(R_2) = \emptyset \dots$
- *Structures comparison :*
 $\underline{R}_1 = \underline{R}_2, \underline{R}_1 \cap \underline{R}_2 \neq \emptyset \dots \dots$

➤ Description of mediation relation

✓ Relational schemas

- $R_m(\underline{K}, A_1, \dots, A_n)$

✓ Constraints :

- **Keys**
- **References**
- **Functional Dependencies**

Intuitive Approach

$R_m(\underline{K}, A, B, C, D, E)$

Mediation queries / m-relation

$$q_1: \pi_{K,A,B,C,D,E}(T1 \bowtie T2)$$

$$q_2: \pi_{K,A,B,C,D,E}(T1 \bowtie T3)$$

$$q_3: \pi_{K,A,B,C,D,E}(T1 \bowtie (T2 \cup T3))$$

$T1(\underline{K}, A, B, K')$

$T2(\underline{K}', C, D, E)$

Actual mediation queries after rewritings

$$q_1: \pi_{K,A,B,C,D,E}(\pi_{K,A,B,K'} S1 \bowtie \pi_{K',C,D,E} S2)$$

$$q_2: \pi_{K,A,B,C,D,E}(\pi_{K,A,B,K'} S1 \bowtie \pi_{K',C,D,E} S3)$$

$$q_3: \pi_{K,A,B,C,D,E}(\pi_{K,A,B,K'} S1 \bowtie (\pi_{K',C,D,E} S2 \cup \pi_{K',C,D,E} S3))$$

Issues in the definition of mediation queries

➤ **For each mediation relation R :**

- ✓ **Q1 : which source relations are relevant for computing R ?**
 - *Determined using the assertions*

- ✓ **Q2 : which operations are relevant between these source relations ?**
 - *Determined using the assertions and a set of integration rules*

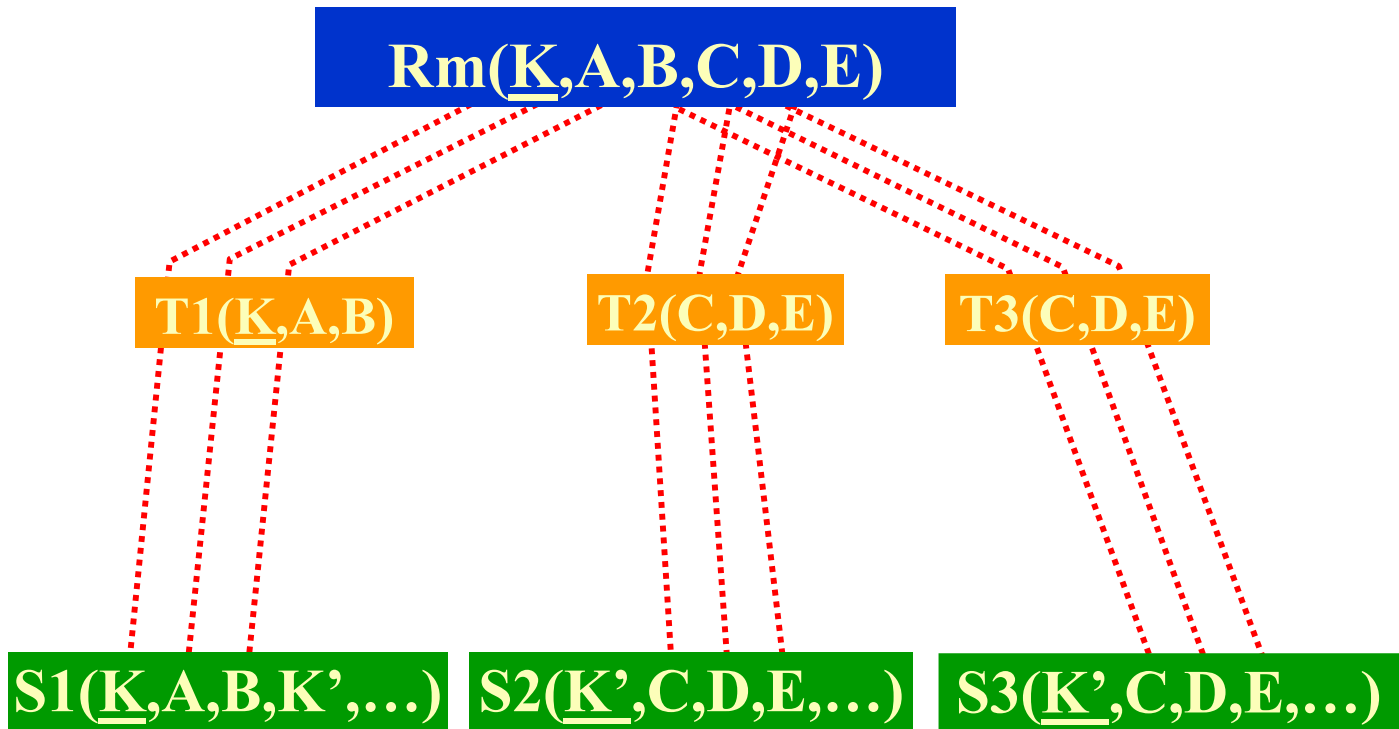
- ✓ **Q3 : which query can be generated from these operations ?**
 - *Generated using computation paths*

Mapping relations

➤ **Definition 1 - Mapping Relation (m-relation for short):** Let $R(Y)$ be a relation of the mediation schema and $S(Z)$ a source relation, $T(X)$ is a mapping relation between R and S if $X \subseteq Y$ et $X \subseteq Z$.

✓ The mapping relation T can be computed as a projection of S on X :

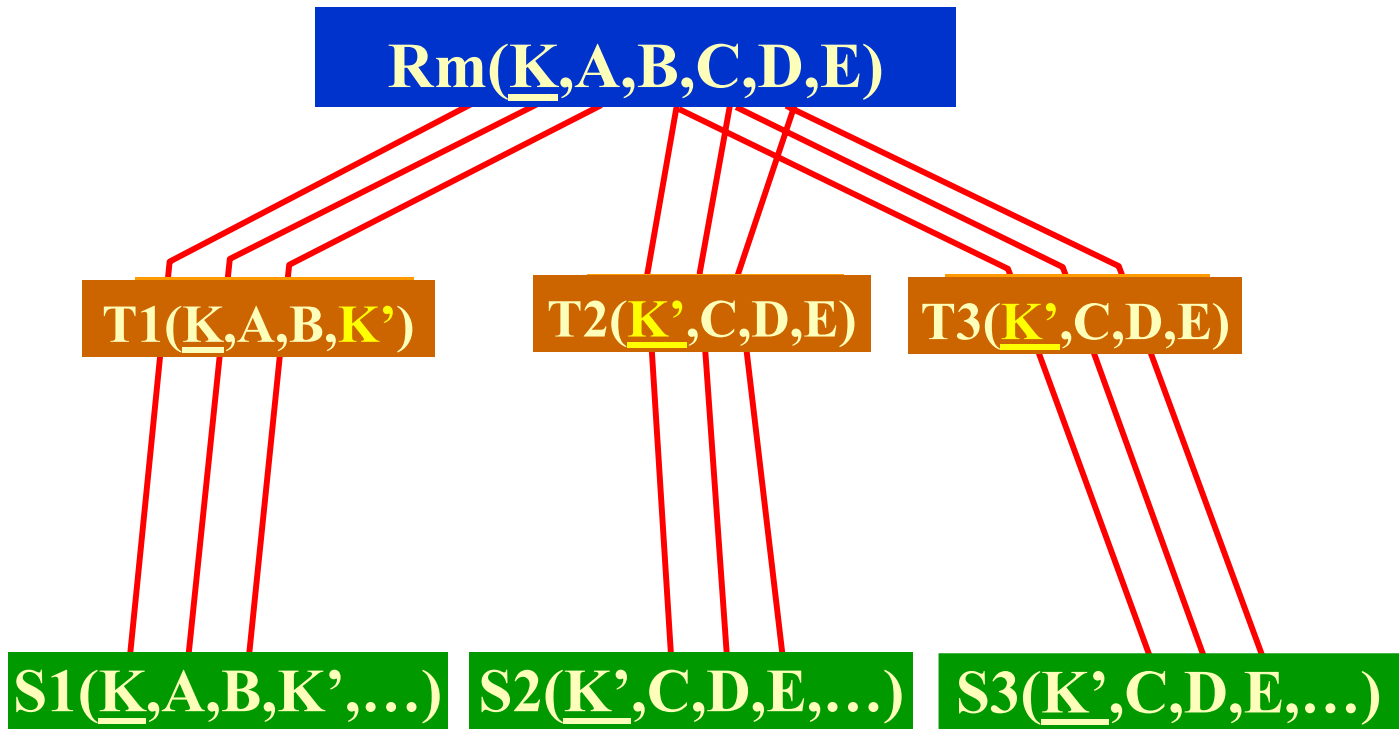
$$T(X) = \pi_X(S).$$



(Extended) Mapping relations

➤ *Definition 2 – Extended mapping relation (xm-relations) : An extended mapping relation T between R and S is a mapping relation which includes both keys and foreign keys of its source relation S .*

- ✓ A mediated relation R has as many xm-relations as sources having common attributes with R .

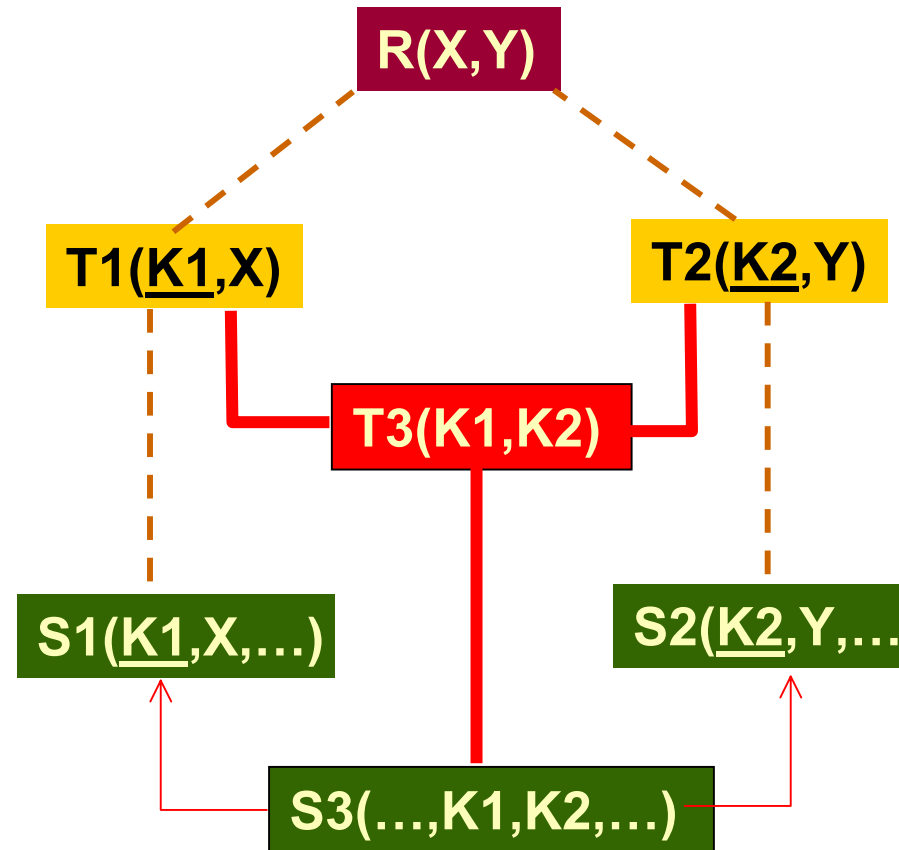


Transition relations

➤ **Definition 3** – **Transition relation:** A relation $T3(K1,K2)$ is a transition relation between $T1(\underline{K1},X)$ and $T2(\underline{K2},Y)$, with respect to $R(X,Y)$, if

- ✓ $T1$ and $T2$ are xm-relations of R
- ✓ there exist a source relation $S3$ such that $T3 = \pi_{K1,K2} S3$
- ✓ $S3.K1 \rightarrow S1.K1$ and $S3.K2 \rightarrow S2.K2$ (foreign keys)

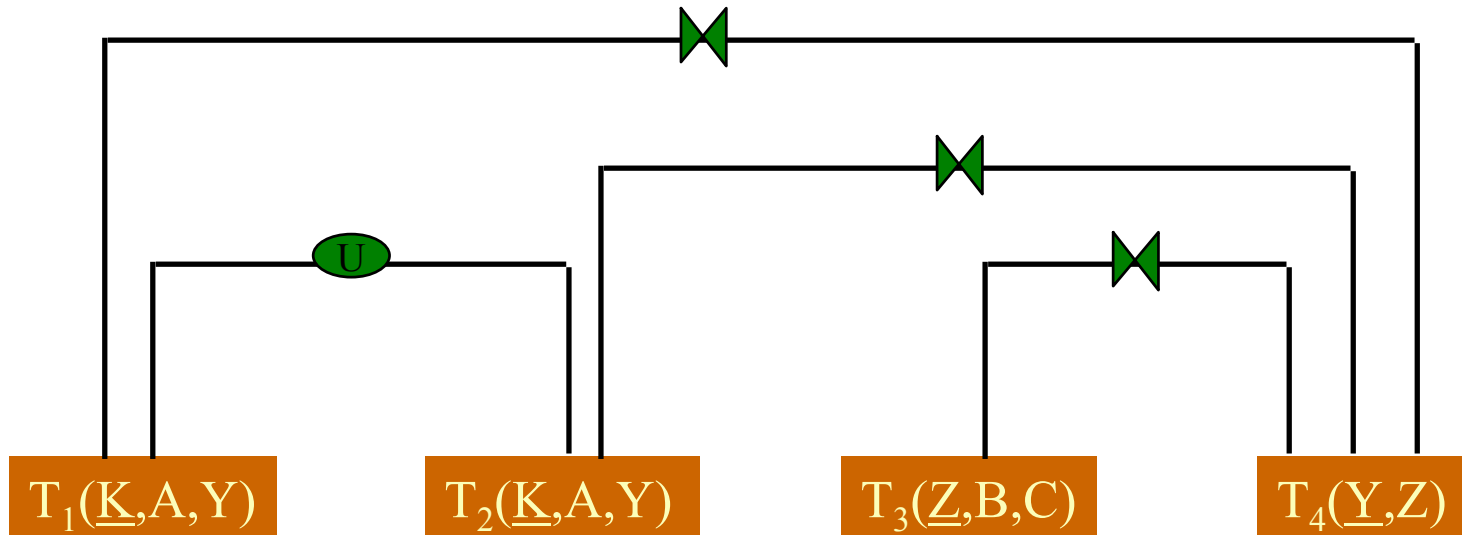
➤ xm-relations and tm-relations inherit FD and RefC from their corresponding sources



Operations graph

➤ *Definition 4 – Operations graph: An operations graph is the set of all possible operations defined over a set of xm-relation / tm-relations associated to a given mediated relation.*

➤ Potential operations are defined using ‘integration rules’



Rules to generate operations graph (integration rules)

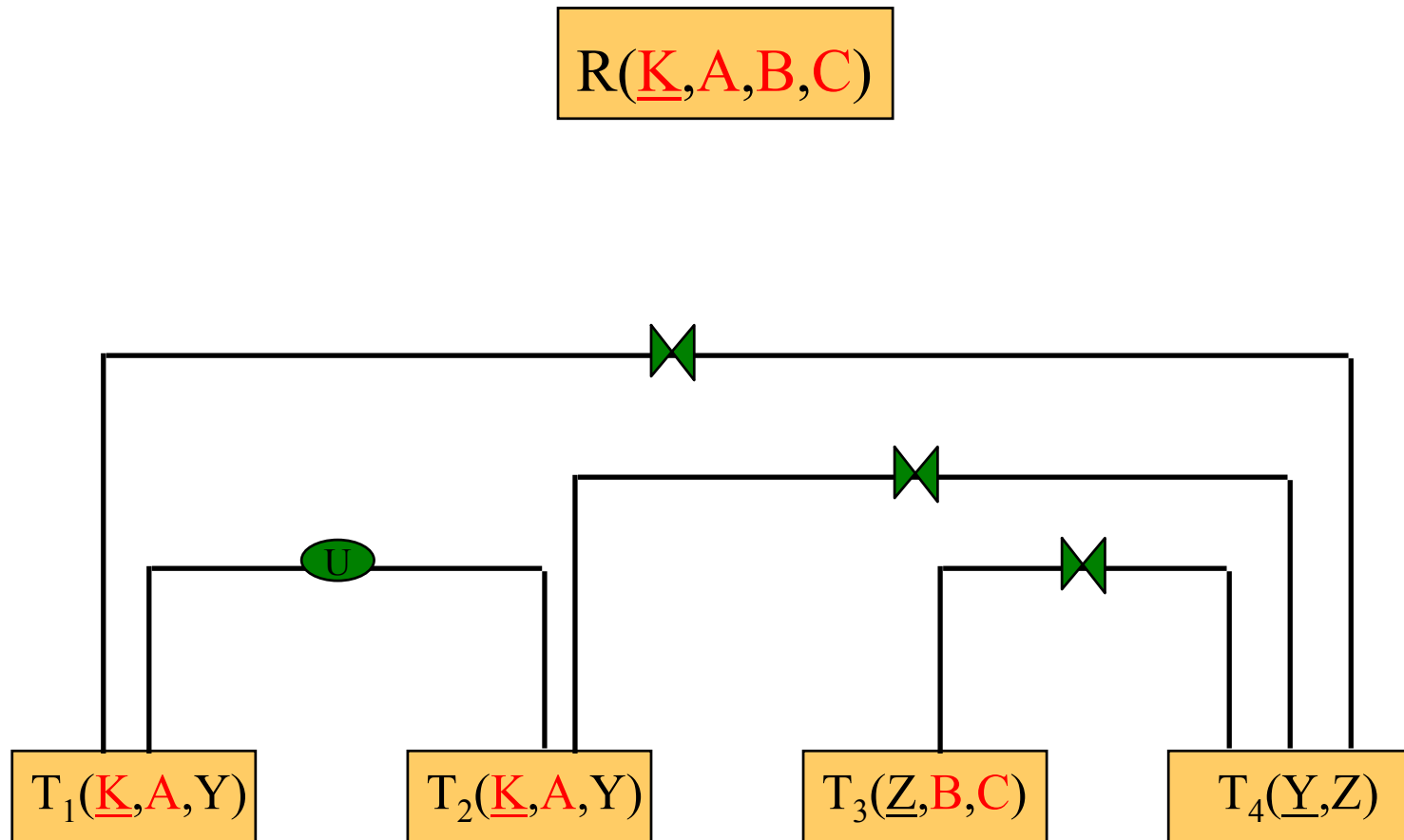
Rule 1

IF T1 = T2 and T1.K1 ↔ T2.K2
THEN T = T1 ∪ T2 or
T = T1 ∩ T2 or
T = T1 - T2 or
T = T2 - T1

Rule 2

IF T1 < > T2 and T2.K → T1.K
THEN T = join(T1, T2)

Query generation: Deriving computation paths from operation graph



Computation path

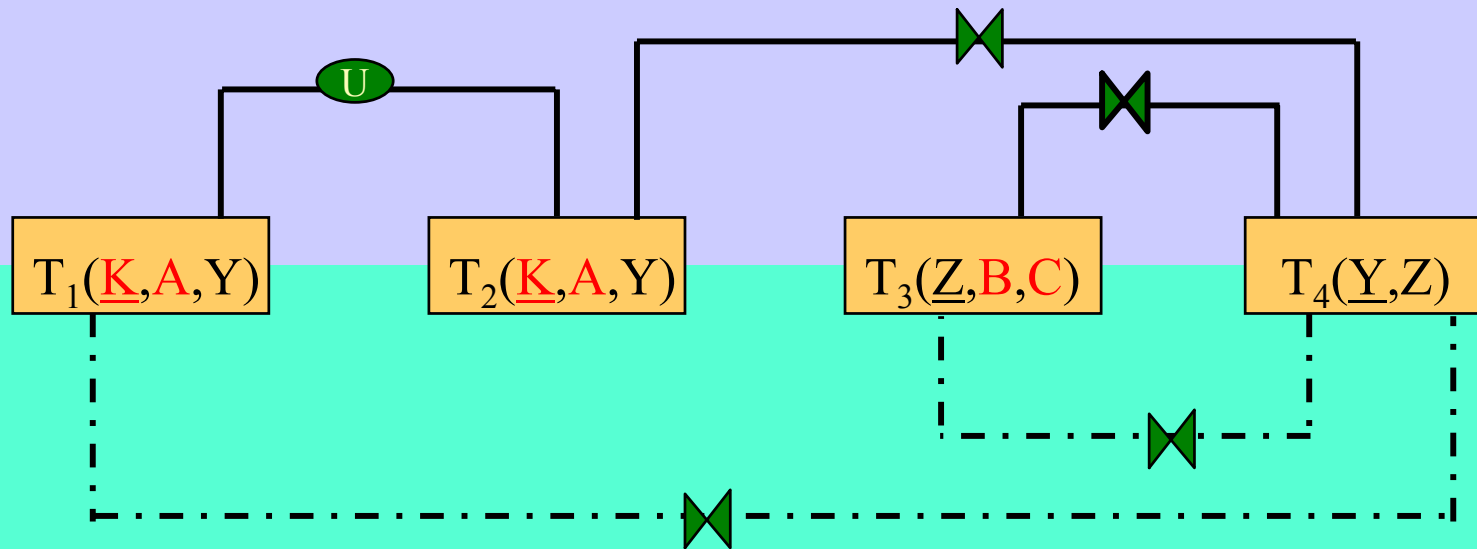
➤ **A computation path associated with a given mediated relation is defined as either :**

- ✓ **A mapping relation which involves all the attributes of the mediated relation**
- or**
- ✓ **Any acyclic and connected subgraph in the operations graph which involves all the attributes of the mediated relation.**

Examples of computation paths

$$Q_1 = (T_1 \cup T_2) \bowtie T_4 \bowtie T_3$$

Solution 1



Solution 2

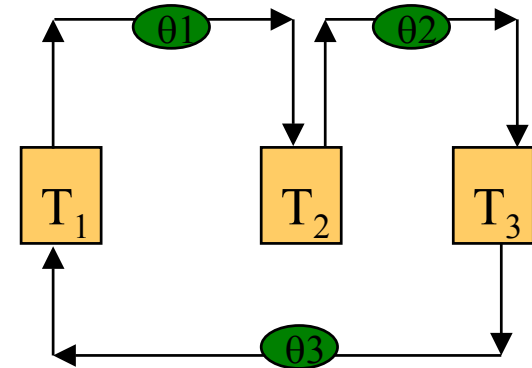
$$Q_2 = T_1 \bowtie T_4 \bowtie T_3$$

Restrictions on computation paths

➤ Problem of cycles :

- $T_1 \theta_1 T_2 \theta_2 T_3$
- $T_1 \theta_1 T_2 \theta_2 T_3 \theta_3 T_1$
- $T_1 \theta_1 T_2 \theta_2 T_3 \theta_3 T_1 \theta_1 T_2$
-

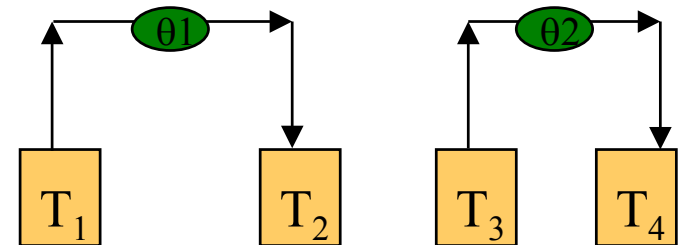
⇒ Infinite number of queries



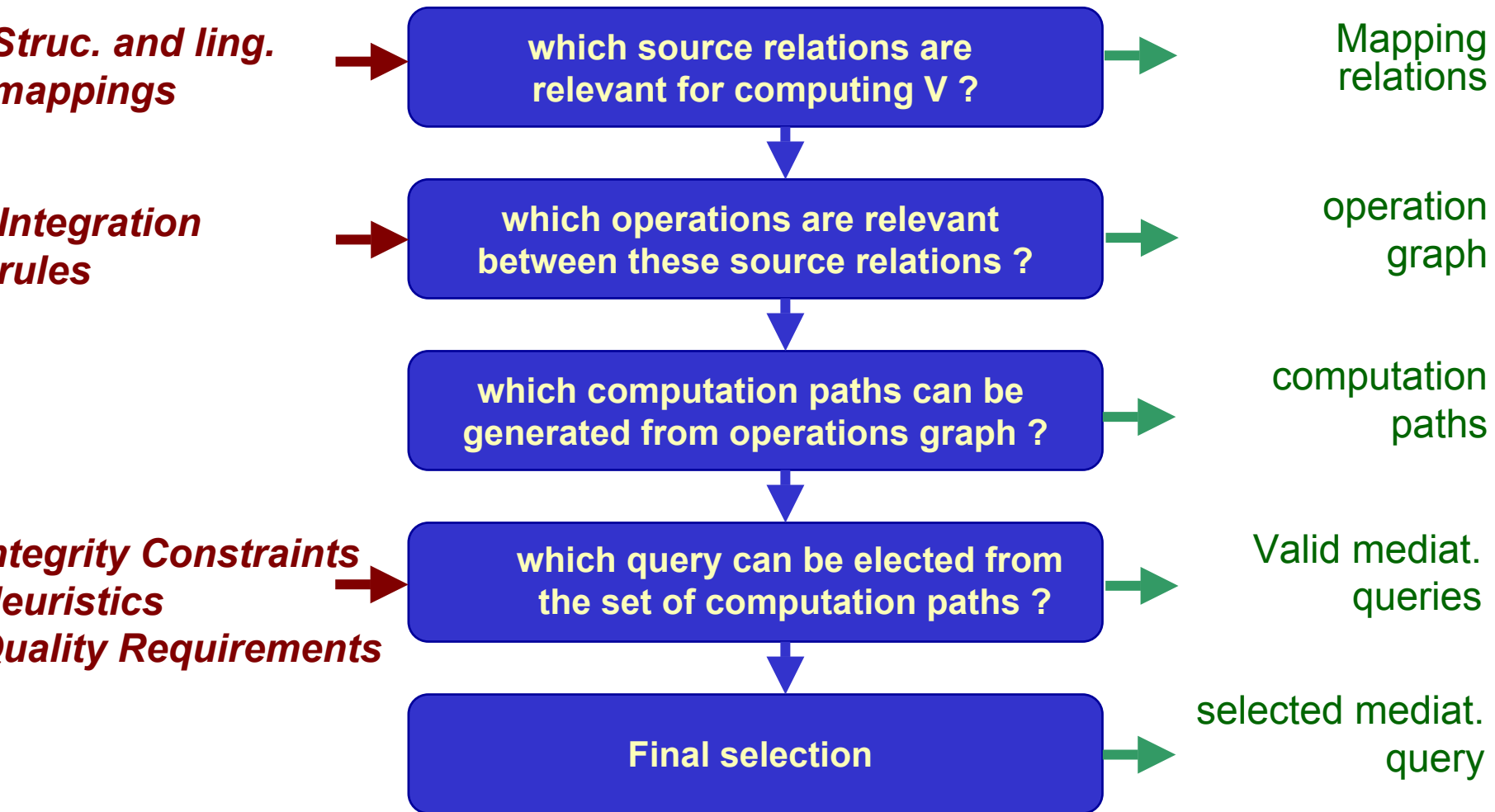
➤ Problem of non-connected sub-graphs

- $A \in \underline{T_i}$ and $A \in \underline{V}$
- $A \notin \underline{T_j}$ ($i \neq j$)

⇒ No possible query



Main steps of the query discovery process



Improvements

➤ Take care of data cleaning

- ✓ Data sources are not homogeneous
- ✓ The generated queries cannot be executed without data transformation

➤ Quality of the mediation schema and queries

- ✓ The solution space is still too much important, explore more heuristics to select a best solution
 - Data sources quality
 - User requirements in terms of data quality

➤ Evolution of the mediation schema and queries

- ✓ Data sources are evolving systems
- ✓ User requirements may evolve