Boundedness of Regular Path Queries in Data Integration Systems

Gösta Grahne, Alex Thomo

Regular Path Queries

Useful for expressing desired paths to follow in graph DB's.

E.g.

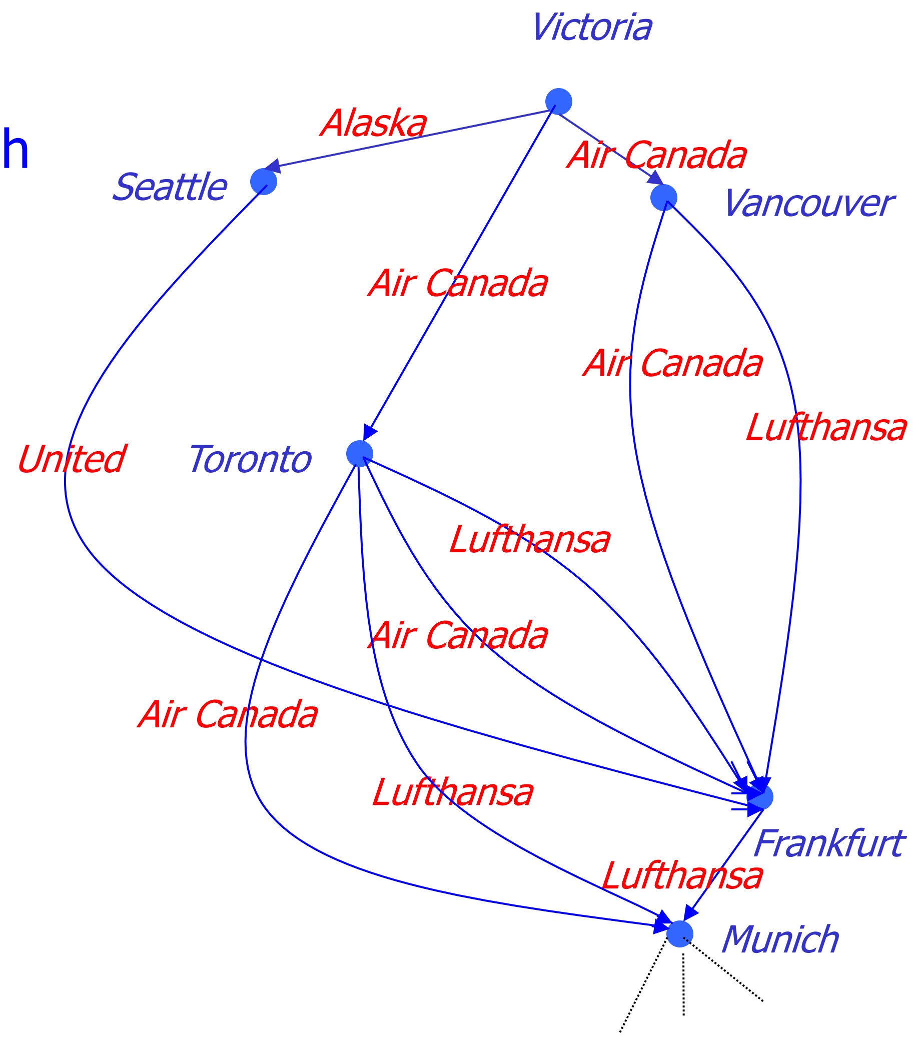
I want to go from Victoria to Munich taking Air Canada or Lufthansa or United.

Query:

(Air Canada+Lufthansa+United)*

Answer:

{ (Victoria,Vancouver),
(Victoria,Frankfurt),
(Victoria,Munich),
…
}



Data Sources

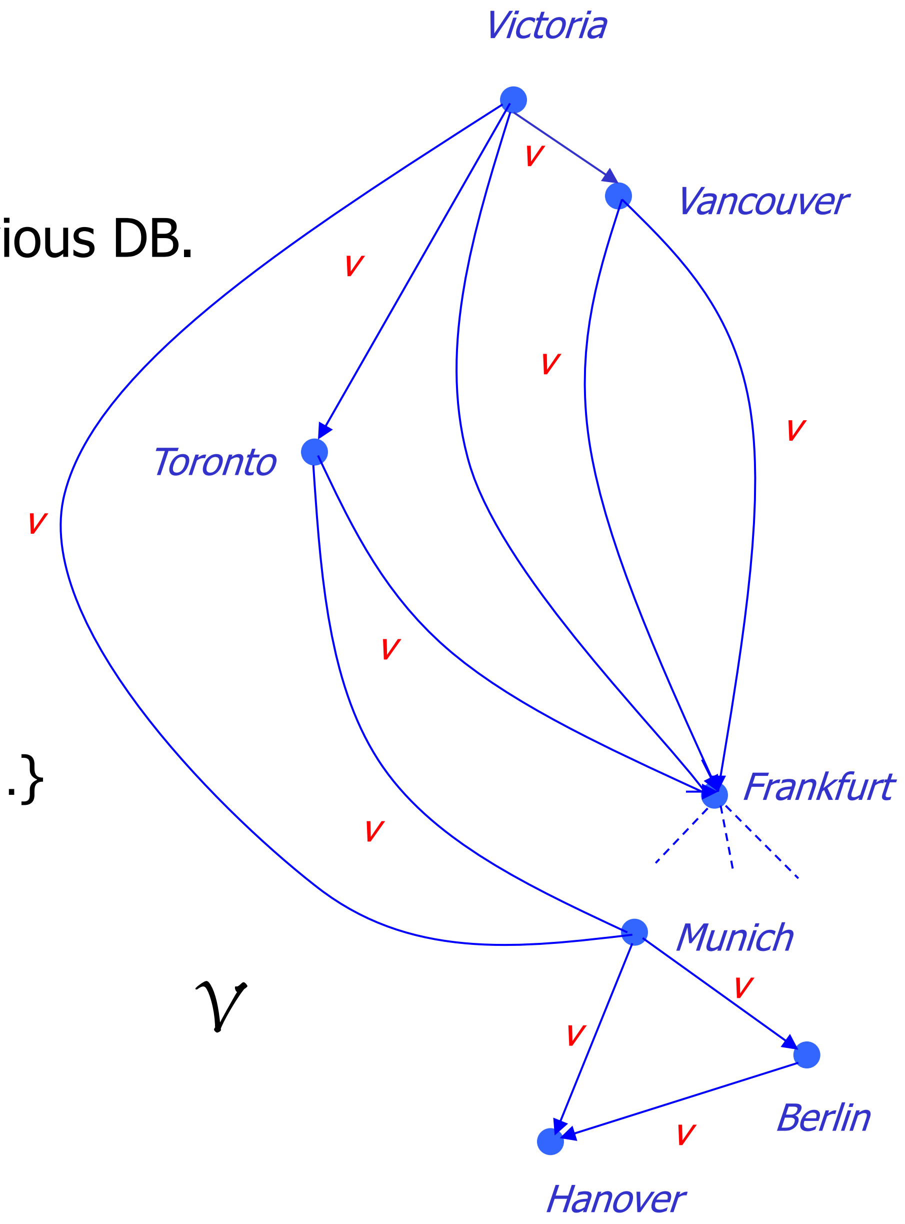
Suppose I have a not available the previous DB.

What I have is "data sources" (views)

V: (Air Canada+Lufthansa)*

Extension:
{(Victoria,Vancouver),
 (Victoria,Frankfurt),
 (Victoria,Munich), (Victoria,Hanover), ...}

LAV (local-as-view) data integration
Global Schema:
  Δ = {Air Canada, Lufthansa, United, BA, AA, Alaska,...}
Local Schema:
  Ω = {v, ...}
User posses queries on the global schema



Query Answering

Q: (Air Canada+Lufthansa+United)*

V: (Air Canada+Lufthansa)*

Two approaches for answering queries:

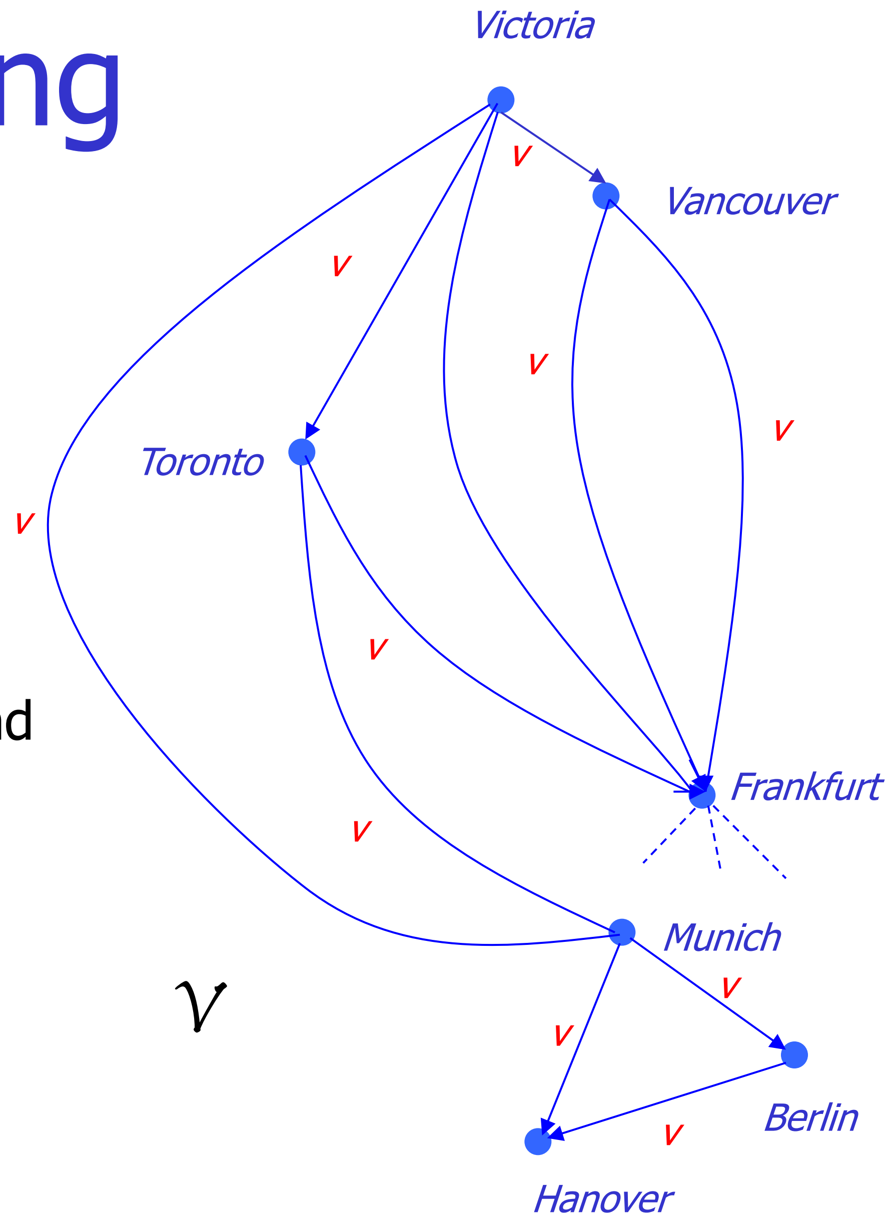
- Compute the **certain answer** (very expensive w.r.t to the data)
- Compute **view-based rewriting** and answer it on the view-graph (polynomial w.r.t. to data)
  Will go with this here.

View-Based Rewriting

[Calvanese, DeGiacomo, Lenzerini, Vardi PODS 1999]

Q′ = v* **:**

All words on Ω whose substitution is contained in Q.



Unnecessary Recursion

$Q' = v^*$

But why not just:

$Q'' = v$

Surely: $Q' \neq Q''$
…as languages on Ω.
However, they are equivalent should we "substitute" v by V, and have languages on Δ.

Hence, we should rather talk about Ω/Δ equivalence.

Unnecessary Recursion – Another Example

$Q=R^*R^k$

$V=R^+$

$Q'=(v^k)^+$ Recall, it's all words on Ω whose substitution is contained in Q

but…

$Q''=v^k$ which is clearly better.

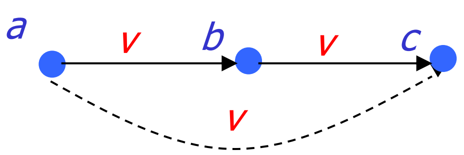
Possible Databases and Valid View-Graphs

- *poss* ($\mathcal{V}$) : Set of all databases from which a given view-graph $\mathcal{V}$ might have been generated.

- Valid $\mathcal{V}$: when *Poss* ($\mathcal{V}$) not empty

- Under **exact view assumption**, not all view graphs are valid.
  - E.g., consider $V=R^*$ and

$\mathcal{V}$



poss($\mathcal{V}$) = ∅.

because $\mathcal{V}$ "misses" a v-edge from *a* to *c*.

Characterization Theorem

Theorem. *Let Q_1 and Q_2 be queries on Ω. Under exact view assumption,*

$$Q_1 \equiv_{\Omega/\Delta} Q_2$$

iff

for each valid view graph $\mathcal{V}$

$$ans(Q_1, \mathcal{V}) = ans(Q_2, \mathcal{V}).$$

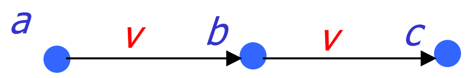
Corollary. Minimize as much as possible a query on Ω (i.e. a view-based rewriting) without loosing query-power as long as Ω /Δ -equivalence is preserved.

…and Ω /Δ -equivalence is algebraically weaker than Ω-equivalence.

Sound Views

- Previous theorem doesn't hold for sound views.
- **E.g.,** consider V=R*, which is Ω/Δ-equivalent with V*, and

$\mathcal{V}$



For $\mathcal{V}$, we have that $ans(v^*, \mathcal{V}) \neq ans(v, \mathcal{V})$.

- Clearly, the answer of V will be equal to the answer of V* on each database on Δ,

  …but because the view is assumed to be sound we cannot enforce $\mathcal{V}$ to have an additional v-edge from *a* to *c*.

Two Notions of Boundedness

- Q_k set of all Ω-words in Q, of length not more than k.

Definition

1. Q is k-bounded iff $Q_k \equiv_{\Omega/\Delta} Q$.
2. Q is finitely bounded iff $\exists\, k \in N$, such that Q is k-bounded.

Theorems

- *k-boundedness is PSPACE-complete w.r.t. the size of the query.*

- *Finite boundedness can be decided in EXPTIME w.r.t. the size of the query.*

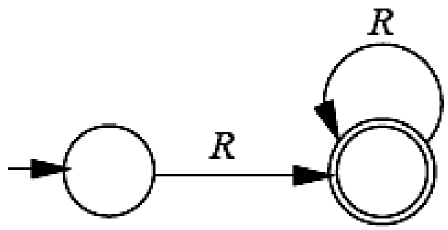
Limitedeness Problem in Distance Automata

- Let A be an ε-free **weighted** automaton (known as *distance automata*.)
  - $d_A(p,w,q)=$
    inf{weight(π) : π is a path spelling w, from p to q in A}
  - $d(A)$ =
    sup{$d_A(s,w,f)$: s start state, f final state}
  - A is limited in distance *iff* $d(A) < \infty$

- Limitedness Problem [Hashiguchi 82]:
  Is a given distance automaton A limited in distance?

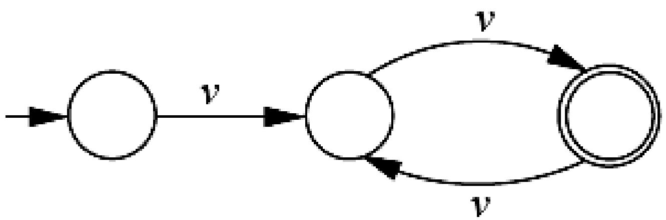
Reduction (I)

View definition

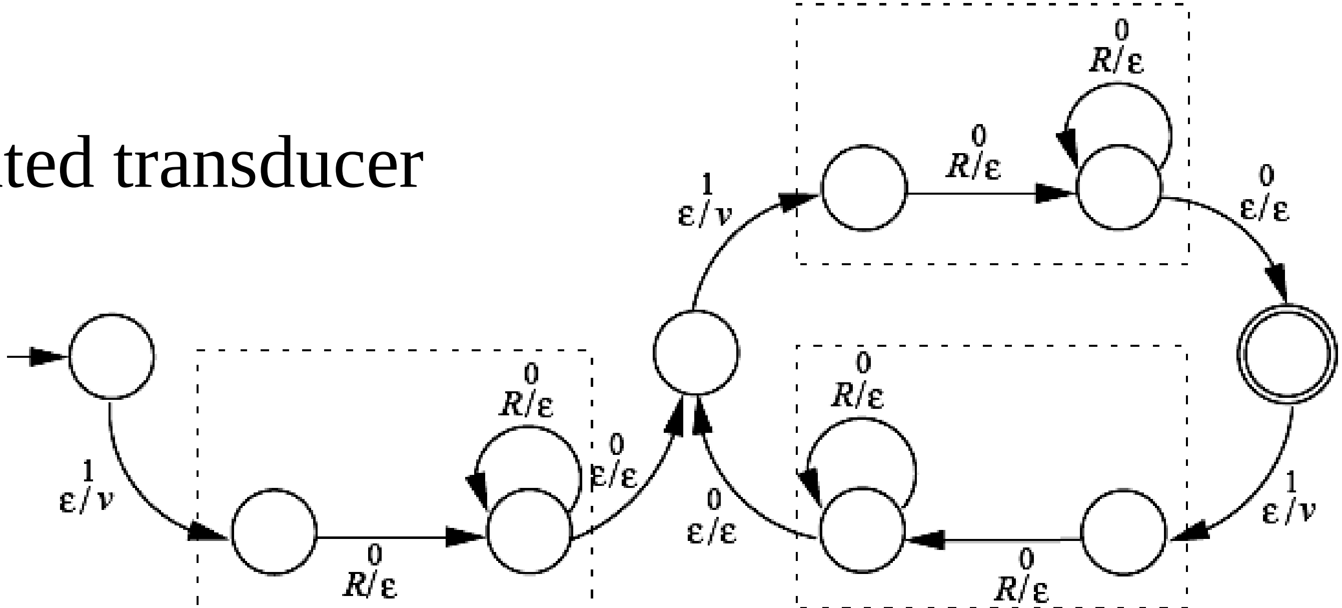


View-based

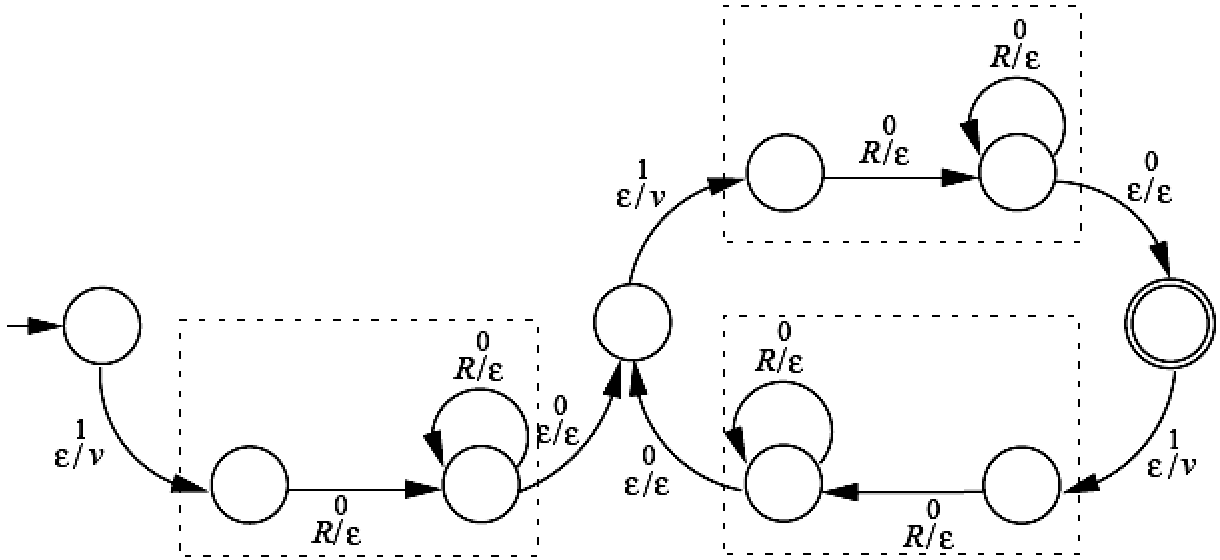
Rewriting



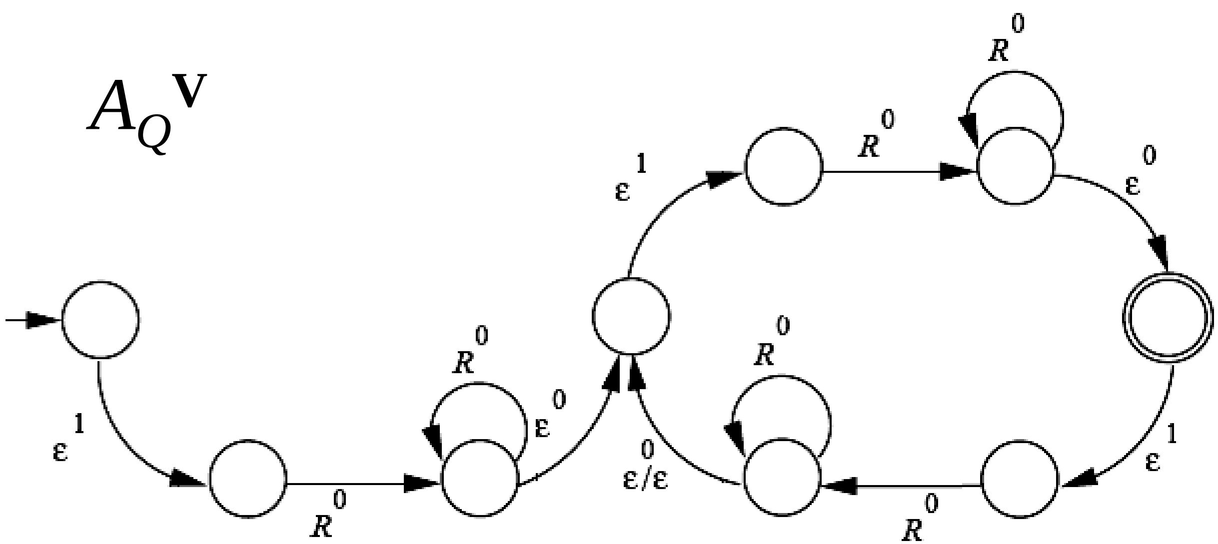
Weighted transducer



Reduction (I)



$A_Q^{\mathbf{v}}$



Drop output and obtain a weighted automaton.

Do epsilon removal.

Characterization

- **Our characterization**:

  Q is bounded iff $A_Q^{\mathbf{V}}$ is limited in distance.

References

- Gösta Grahne, Alex Thomo. Boundedness of Regular Path Queries in Data Integration Systems. IDEAS 2007: 85-92
- Gösta Grahne, Alex Thomo. Algebraic rewritings for optimizing regular path queries. Theoretical Computure Science 296(3): 453-471 (2003)
- Gösta Grahne, Alex Thomo: New Rewritings and Optimizations for Regular Path Queries. ICDT 2003: 242-258
- Gösta Grahne, Alex Thomo: Query containment and rewriting using views for regular path queries under constraints. PODS 2003: 111-122
- Gösta Grahne, Alex Thomo: Algebraic Rewritings for Optimizing Regular Path Queries. ICDT 2001: 301-315
- Gösta Grahne, Alex Thomo: Approximate Reasoning in Semistructured Data. KRDB 2001
- Gösta Grahne, Alex Thomo: An Optimization Technique for Answering Regular Path Queries. WebDB (Selected Papers) 2000: 215-225