University of Washington
Department of Computer Science & Engineering

# Managing XML and Semistructured Data

Lecture 3: Bisimulation

Prof. Dan Suciu

Spring 2001

## In this lecture

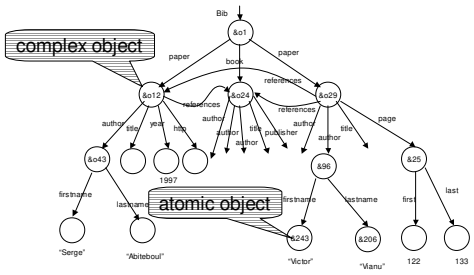
- Semistructured data model
- Graph Simulation and Bisimulation
- Computing (bi)simulation

Resources

Adding structure to semistructured data by Buneman, Davidson, Fernandez, Suciu, in ICDT 97

**Data on the Web** Abiteboul, Buneman, Suciu : *section 6.4*

## The Semistructured Data Model




Object Exchange Model (OEM)

## Syntax for Semistructured Data

May omit oid's:

```
{ paper: { author: "Abiteboul",
           author: { firstname: "Victor",
                     lastname: "Vianu"},
           title: "Regular path queries …",
           page: { first: 122, last: 133 }
         }
}
```

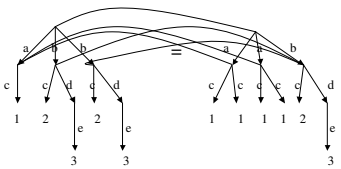
## Set Semantics for Trees

Want to say that {a, a, b} = {a, b}
Define equality for trees first, then for graphs

**Definition** Two trees t, t' are equal, t=t', if:

1. They are both atomic values with same value
2. $t = \{t_1, ..., t_m\}$, $t' = \{t_1', ..., t_n'\}$ and:
   - $\forall i=1,...,m, \exists j=1,...,n$ s.t. $t_i = t_j'$
   - $\forall j=1,...,n, \exists i=1,...,m$ s.t. $t_i = t_j'$

## Set Semantics: Example



## Set Semantics for Graphs

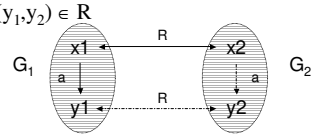
- Previous definition does not apply directly to graphs with cycles
- Need to adapt it ➔ ***bisimulation***
- First, we will define a ***simulation***

## Graph Simulation

**Definition** Two edge-labeled graphs $G_1$, $G_2$

A *simulation* is a relation R between nodes:

- if $(x_1, x_2) \in R$, and $(x_1,a,y_1) \in G_1$, then exists $(x_2,a,y_2) \in G_2$ (same label) s.t. $(y_1,y_2) \in R$




Note: if we insist that R be a function ➔ graph homeomorphism

## Graph Bisimulation

**Definition** Two edge-labeled graphs G1, G2

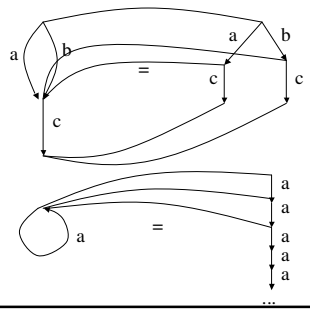
A *bisimulation* is a relation R between nodes s.t. both R and $R^{-1}$ are simulations

## Set Semantics for Semistructured Data

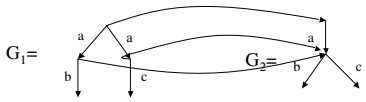
**Definition** Two rooted graphs $G_1$, $G_2$ are equal if there exists a bisimulation R from $G_1$ to $G_2$ such that $(root(G_1), root(G_2)) \in R$

- Notation: $G_1 \approx G_2$
- For trees, this is precisely our earlier definition

## Examples of Bisimilar Graphs




## Examples of non-Bisimilar Graphs

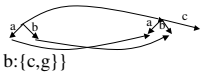



- This is a *simulation* but not a *bisimulation* – Why ?
- Notice: $G_1$, $G_2$ have the same sets of *paths*

## Examples of Simulation

- Simulation acts like "subset"

  {a, b} ⊆ {a, b, c}



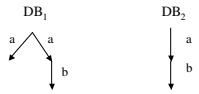

  {a, b:{c}} ⊆ {d, a:{e,f}, b:{c,g}}

- Question:
- if $DB_1 \subseteq DB_2$ and $DB_2 \subseteq DB_1$ then $DB_1 \approx DB_2$ ?

## Answer

if $DB_1 \subseteq DB_2$ and $DB_2 \subseteq DB_1$ then $DB_1 \approx DB_2$ ?

No. Here is a counter example:




$DB_1 \subseteq DB_2$ and $DB_2 \subseteq DB_1$ but NOT $DB_1 \approx DB_2$

## Facts About a (Bi)Simulation

- The empty set is always a (bi)simulation
- If R, R' are (bi)simulations, so is R U R'
- Hence, there always exists a *maximal* (bi)simulation:
  - Checking if $DB_1$=$DB_2$: compute the maximal bisimulation R, then test (root($DB_1$),root($DB_2$)) in R

## Computing a (Bi)Simulation

- Computing the maximal (bi)simulation:
  - Start with R = nodes($G_1$) x nodes($G_2$)
  - While exists ($x_1$, $x_2$) ∈ R that violates the definition, remove ($x_1$, $x_2$) from R
- This runs in polynomial time ! Better:
  - O((m+n)log(m+n)) for bisimulation
  - O(m n) for simulation
  - Compare to finding a graph homeomorphism !

NP Complete