Schemaless Semistructured Data Revisited

—Reinventing Buneman's Deterministic Semistructured Data Model—

Keishi Tajima

School of Informatics, Kyoto University

PBF **at Edinburgh, Scotland**
28 Oct. 2013

School of Informatics
Kyoto University

Background

Why Semistructured Data in PBF 2013?

One of the topics where Peter Buneman has made big contributions.

I was:

- **visiting U. Penn DB group from 2000 to 2001, and**
- **joined their research on efficient archiving of version history of semistructured data.**

Background

What is Deterministic Semistructured Data Model?

The basic idea of the research was partly based on:

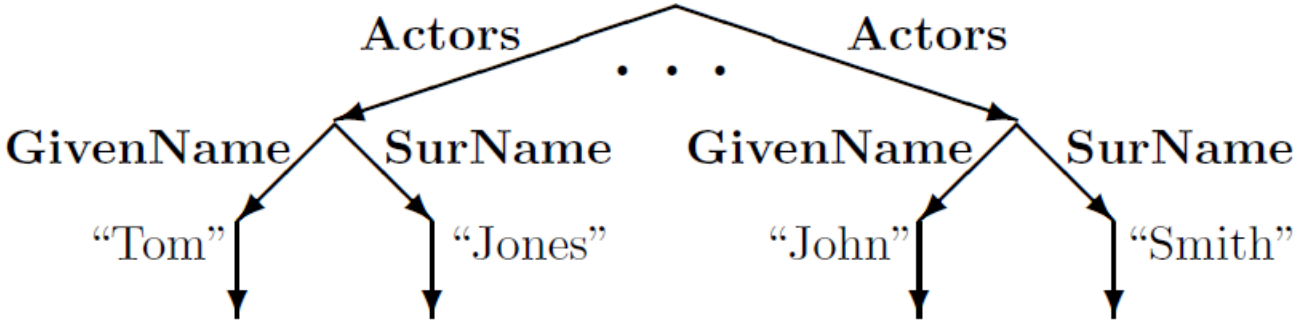
"A Deterministic Model for Semistructured Data"
by P. Buneman, A. Deutsch, W.-C. Tan
Workshop in conj. with ICDT'99, pp. 14–19, 1999

Their Deterministic Semistructured Data Model:

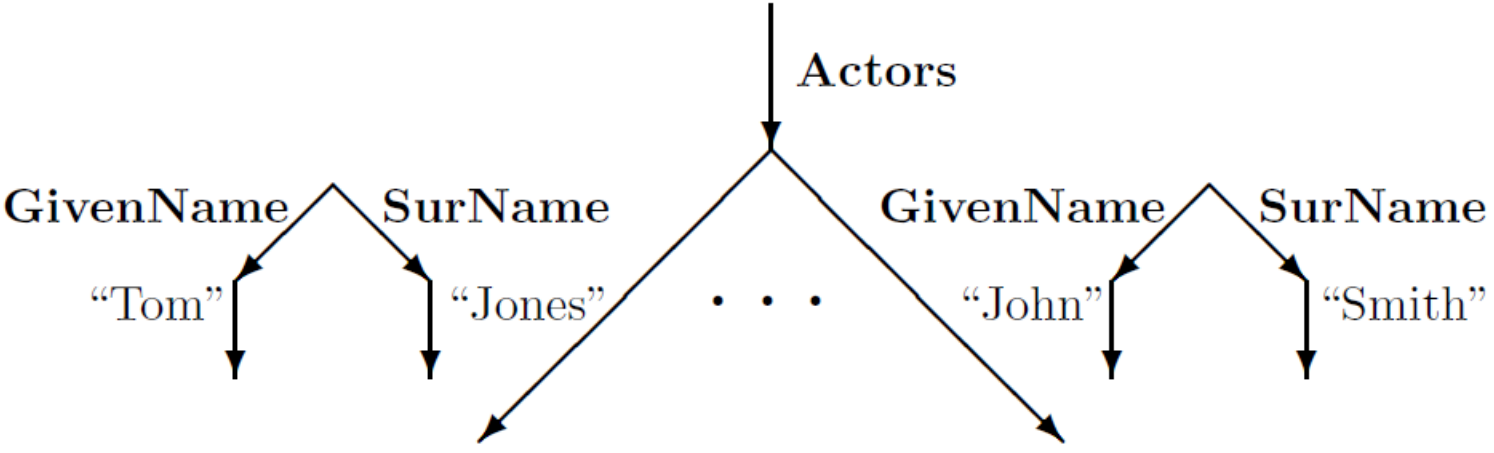
- Edge-labeled graph.
- Edges outgoing from a node have unique labels.
- Edge labels are also graphs.

Standard Model and Their Model

Standard Model



Their Model



School of Informatics
Kyoto University

Goal of this Research

Reinventing the data model through another way.

Contribution:

- **Providing additional rationale of the design of the model.**

School of Informatics
Kyoto University

Goal of this Research

the same?

Reinventing the data model through ~~another~~ way.

Contribution:

- ~~**Providing additional rationale of the design of the model.**~~
- **Providing an archive of their discussion (provenance).**

School of Informatics
Kyoto University

Roadmap

1. **What is important for models for semistructured data?**

   ⇒ **uniform treatment of data and metadata**

2. **What are most important metadata?**

   ⇒ **attribute names and key values**

3. **We extend a standard model following the discussion.**

   Start: **edge-labeled graphs**

   3.1. Use key values as edge labels.

   3.2. Why unique edge labels?

   3.3. Why any graphs as edge labels?

What is important in models for semistructured data?

What is semistructured data?

- nested irregular structure
- no predefined schema defining the structure (schemaless)

What is schema?

- metadata annotating data
  - for systems to parse data
  - for users to know the meaning of data

Standard approach

- embed metadata inside data (self-describing)

Uniform treatment of data and metadata is important.

What are data and metadata?

Let's see the most classic way of organizing large data:

tables

Two types of tables has been popularly used:

- relational tables
  - a set of tuples indexed by column names
- multidimensional tables (or matrices if two-dimensional)
  - a multidimensional array indexed by arbitrary values

School of Informatics
Kyoto University

Relational Tables

A relational table with column names

Nutrition Facts

Item	Cal	Sugar	Fat	···	Allergen		
					Milk	Wheat	Egg
Hamburger	250	5.5	9	···	-	√	√
Cheeseburger	300	6.5	12	···	√	√	√
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gigaburger	540	8.8	29	···	-	√	√

- column names = attribute names = metadata
- Some attribute names have hierarchical structure. (e.g., Allergen.Milk)

Multidimensional Tables

A multidimensional table with row/column names

Schuylkill River Trail Mileage Chart

	Philadelphia	Manayunk	···	Tamaqua
Philadelphia	—	7	···	114.5
Manayunk	7	—	···	107.5
⋮	⋮	⋮	⋮	⋮
Tamaqua	114.5	107.5	···	—

- Both column names and row names are metadata

Relational Table or Multidimensional Table?

No clear distinction between them.

Car Sales by Month and State

	NY	NJ	PA	···	CA
Jan 2010	233	149	183	···	258
Feb 2010	358	187	170	···	286
Mar 2010	285	174	191	···	225
⋮	⋮	⋮	⋮	⋮	⋮
Dec 2010	169	89	115	···	188

- A typical multidimensional table in OLAP.
- Can also be interpreted as a relation.

Relational Table or Multidimensional Table?

No clear distinction between them.

Car Sales by Month and State

Month	NY	NJ	PA	···	CA
Jan 2010	233	149	183	···	258
Feb 2010	358	187	170	···	286
Mar 2010	285	174	191	···	225
⋮	⋮	⋮	⋮	⋮	⋮
Dec 2010	169	89	115	···	188

- A typical multidimensional table in OLAP.
- Can also be interpreted as a relation.

Relational Table or Multidimensional Table?

No clear distinction between them.

Car Sales by Month and State

State	NY	NJ	PA	···	CA
Jan 2010	233	149	183	···	258
Feb 2010	358	187	170	···	286
Mar 2010	285	174	191	···	225
⋮	⋮	⋮	⋮	⋮	⋮
Dec 2010	169	89	115	···	188

- A typical multidimensional table in OLAP.
- Can also be interpreted as a relation.

Interpreting Relations as Multidimensional Tables

Item	Cal	Sugar	Fat	···	Milk	Wheat	Egg
Hamburger	100	5.5	9	···	-	√	√
Cheeseburger	200	6.5	12	···	√	√	√
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gigaburger	540	8.8	29	···	-	√	√

- attribute names = column names
- key values = row names

Relations can be interpreted as multidimensional tables if we interpret key values as row names.

Interpreting Relations as Multidimensional Tables

Item	Cal	Sugar	Fat	···	Milk	Wheat	Egg
Hamburger	100	5.5	9	···	-	√	√
Cheeseburger	200	6.5	12	···	√	√	√
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gigaburger	540	8.8	29	···	-	√	√

Benefit:

- Symmetric treatment of rows and columns.
- More intuitive for human.
  "cal of Hamburger is 100, sugar of Hamburger is 5.5, ..., cal of Cheeseburger is 200, ..."

What are metadata in tables?

Attribute names and key values most often and naturally play roles of metadata.

Why are they distinguished in RDB?

- **Columns are static, while rows are dynamic.**
- **Cells in a column store the same type, while cells in a row may not.**

These two are not important in semistructured data.

What should we do with semistructured data?

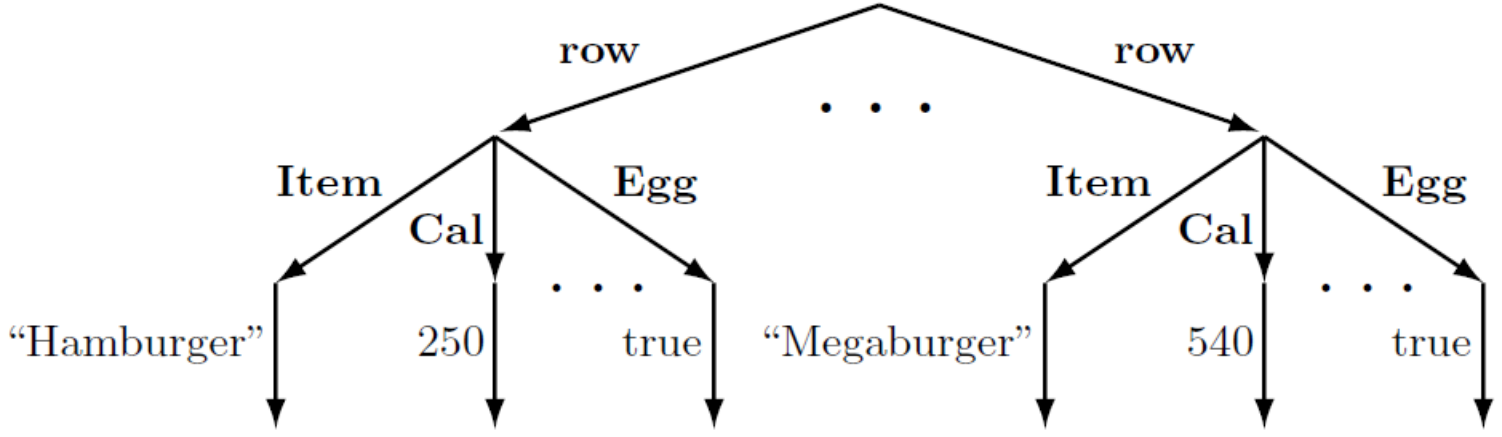
Let's use attribute names and key values as metadata.

Benefit of treating key values as metadata:

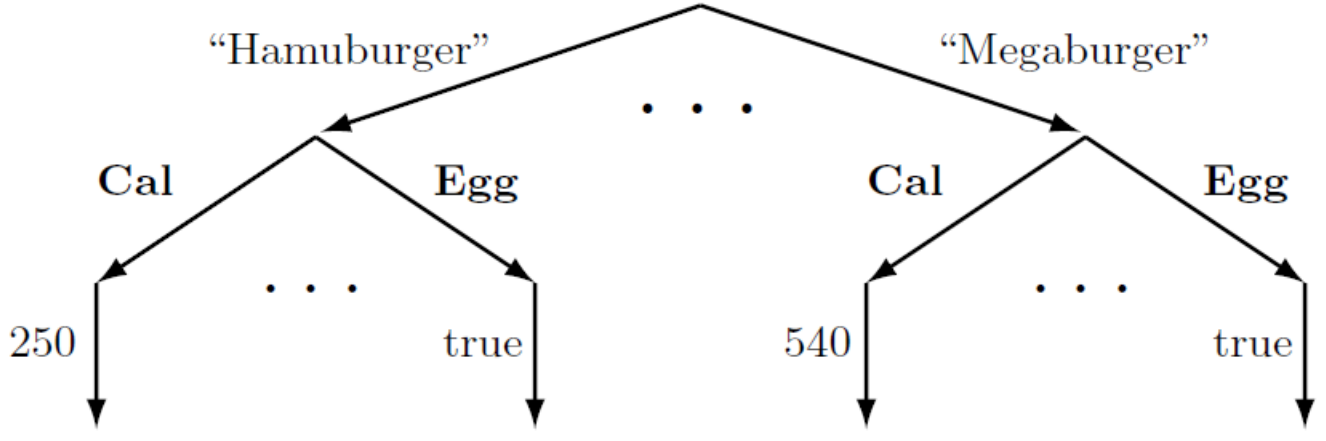
- **Natural way of organizing data.**
- **We often specify data items by specifying key values, so why not?**

Using Key Values as Labels in Semistructured Data

Representation of tables in ordinary model:

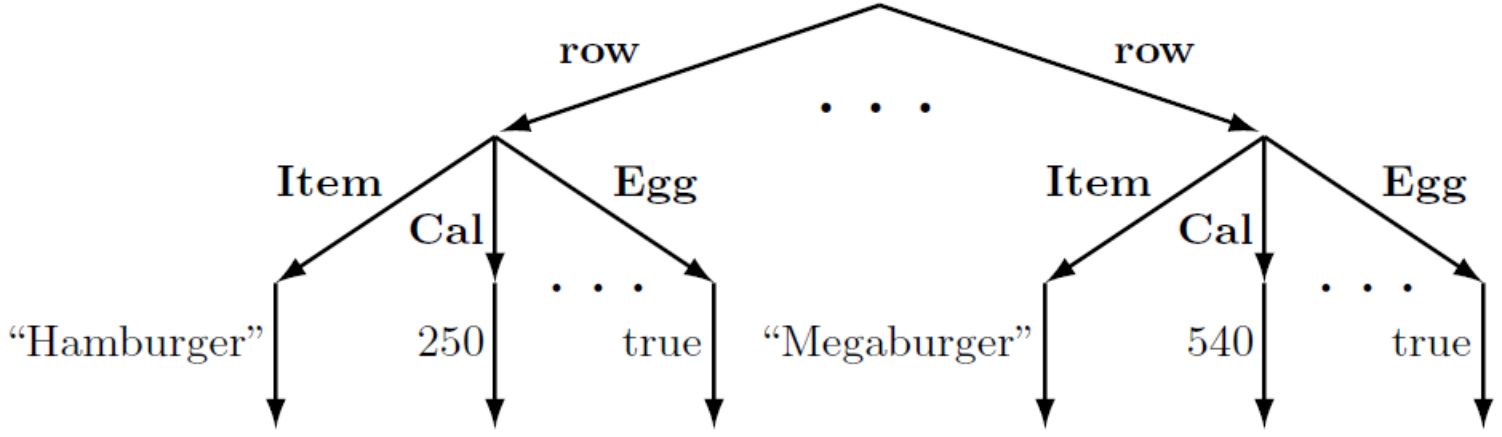


With key values as metadata (row-wise):

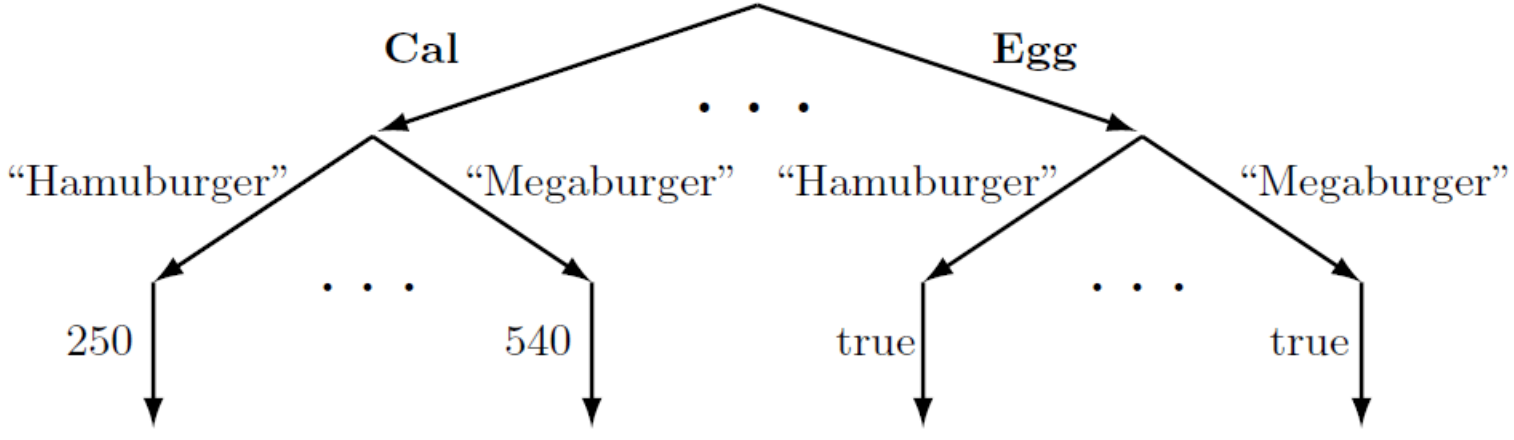


Using Key Values as Labels in Semistructured Data

Representation of tables in ordinary model:



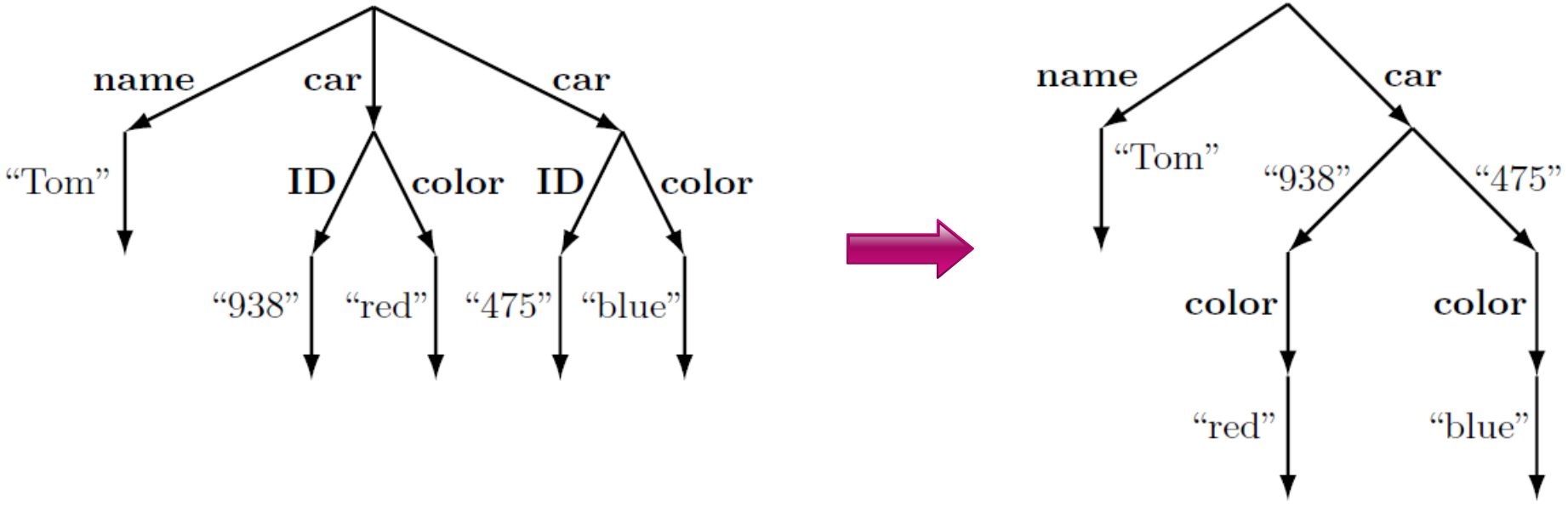
With key values as metadata (column-wise):



Deterministic?

Do we need multiple outgoing edges with the same label?

No, if we use key values as edge labels



Benefit?

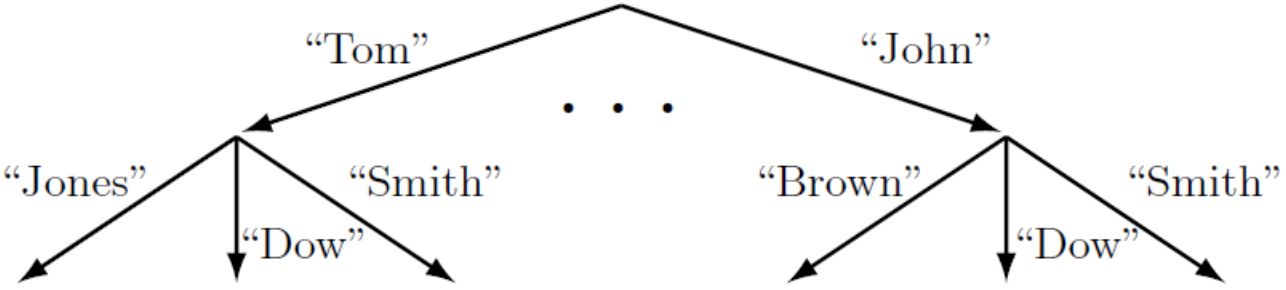
Uniform treatment of set-value and single-value

In the ordinary models, it was achieved by carefully designed QLs.

Only atomic values as edges?

We have composite keys.

Hierarchical representation of composite keys?



Problems:

- **Produces meaningless nodes.**
- **Forces some order unnecessarily. (data independence!)**
- **Especially bad when we have keys that are set-valued.**

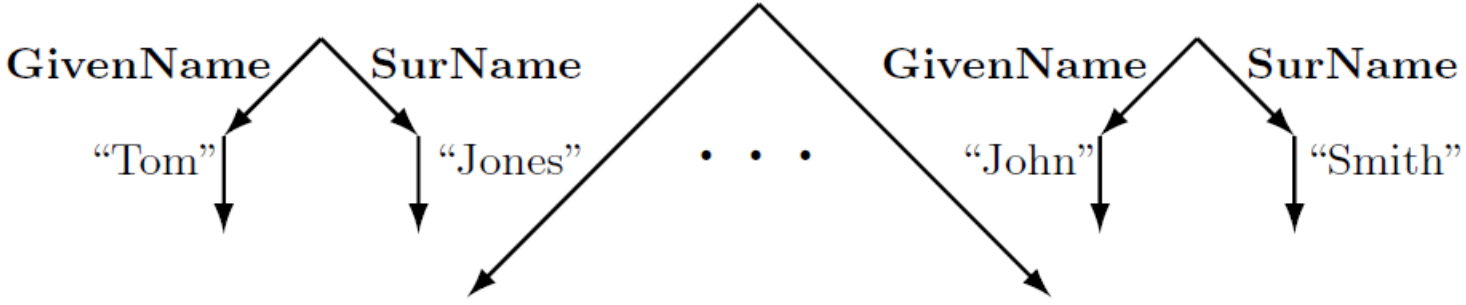
School of Informatics
Kyoto University

How we represent composite or set-valued keys?

We represent composite or set values by graphs.

⇓

We use graphs as edge labels.



Now we have reinvented the data model.

Summary

1. **Uniform treatment of data and metadata is important in semistructured data**
2. **attribute names and key values are most important metadata**
3. **We extend a standard model in the following way**
   - **3.1. Use key values as edge labels.**
   - **3.2. Unique edge labels because:**
     - **uniform treatment of single-value/set-value.**
   - **3.3. Allow any graph values as edge labels because:**
     - **we have composite or set-valued keys.**