Compacting XML Data

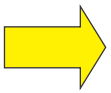
Shuohao Zhang, Curtis Dyreson and Zhe Dang
School of E.E. and Computer Science
Washington State University
Pullman, Washington USA

Book Data

```
<bib>
  <publisher>Doubleday
    <book>
      <title>The Da Vinci Code</title>
      <author>
        <name>Dan Brown</name>
      </author>
    </book>
  </publisher>
  <publisher>Pocket Star
    <book>
      <title>Angels & Demons</title>
      <author>
        <name>Dan Brown</name>
      </author>
    </book>
  </publisher>
</bib>
```

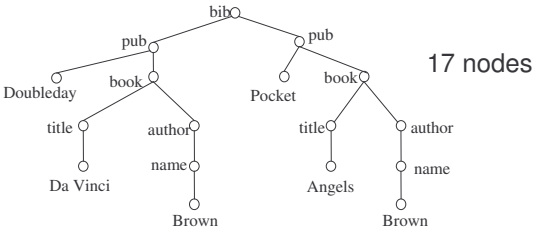

Same Data, Different Structure

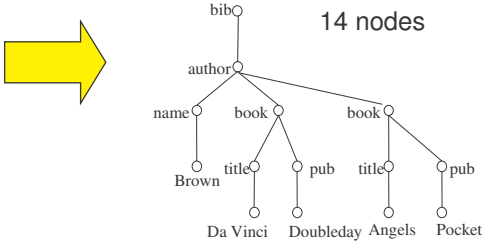
```
<bib>
  <publisher>Doubleday
    <book>
      <title>The Da Vinci Code</title>
      <author>
        <name>Dan Brown</name>
      </author>
    </book>
  </publisher>
  <publisher>Pocket Star
    <book>
      <title>Angels & Demons</title>
      <author>
        <name>Dan Brown</name>
      </author>
    </book>
  </publisher>
</bib>
```



```
<bib>
 <author><name>Dan Brown</name>
    <book>
      <title>The Da Vinci Code
      </title>
      <publisher>Doubleday
      </publisher>
    </book>
    <book>
      <title>Angels & Demons
      </title>
      <publisher>Pocket Star
      </publisher>
    </book>
  </author>
<bib>
```

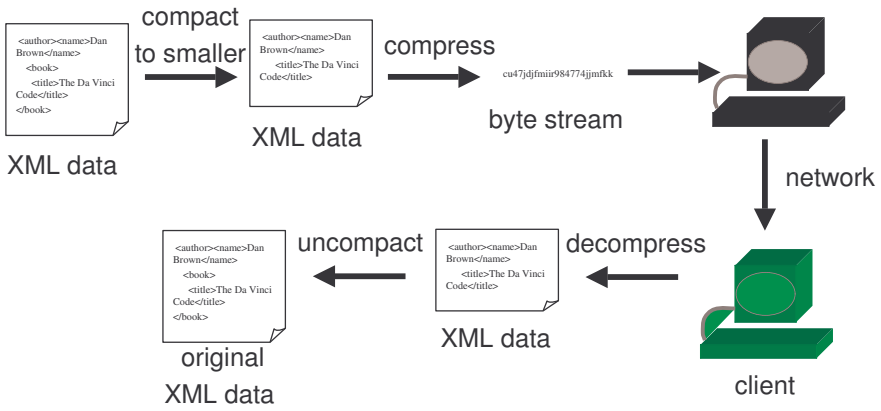

Structure Comparison





Target Application

- Compression – make smaller
- Compaction – make smaller, output is XML



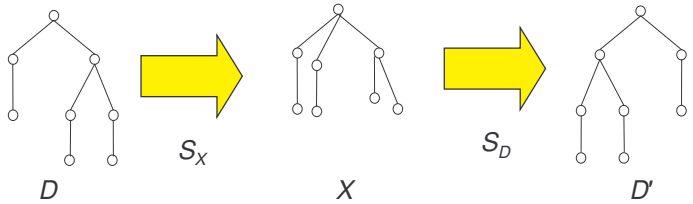
Related Work

- Compression
  - Huffman (*IRE* 1952)
  - Lempel/Ziv (*Trans. on Inf. Theory* 1977)
  - **XMill** – Liefke and Suciu (*SIGMOD* 2000)
  - Homomorphic XML Compression
    - **XPress** – Jun-Ki Min, Myung-Jae Park, Chin-Wan Chung (*SIGMOD* 2003)
    - **XGrind** – P. M. Tolani and J. R. Haritsa (*ICDE* 2002)
  - Compaction – us
- Restructuring
  - Zhang and Dyreson (*IIWeb* 2006)
  - LCA queries
    - Schmidt, Kersten, Windhouwer (*ICDE* 2001)
    - Cohen, Mamou, Kanza, Sagiv (*VLDB* 2003)
    - Li, Yu, Jagadish (*VLDB* 2004)
    - Zhang and Dyreson (*WWW* 2006)

Restructuring

- Process that transforms data structure
- Takes a *structural specification* (*signature*)

  $Trans(D,S) = X$



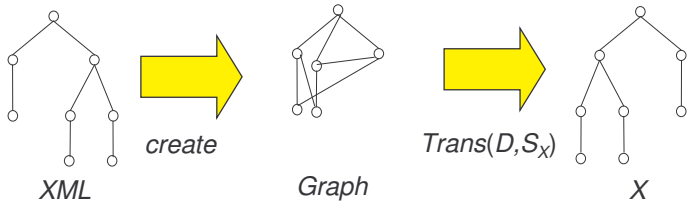
- Three outcomes of $Trans(Trans(D, S_X),S_D) \equiv D'$
  1. D' has at least the same data as D (inclusive)
  2. D' has no more data than D (non-additive)
  3. D' has the same data as D (*reversible*)

Same Data, Different Document

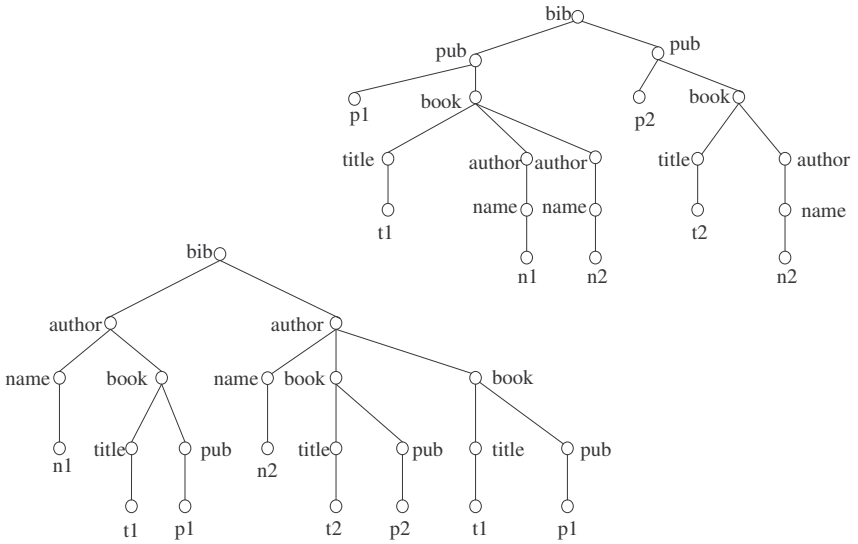
- *Same* data modulo
  - sibling order
  - duplicate siblings
  - white space

Poly-Transform

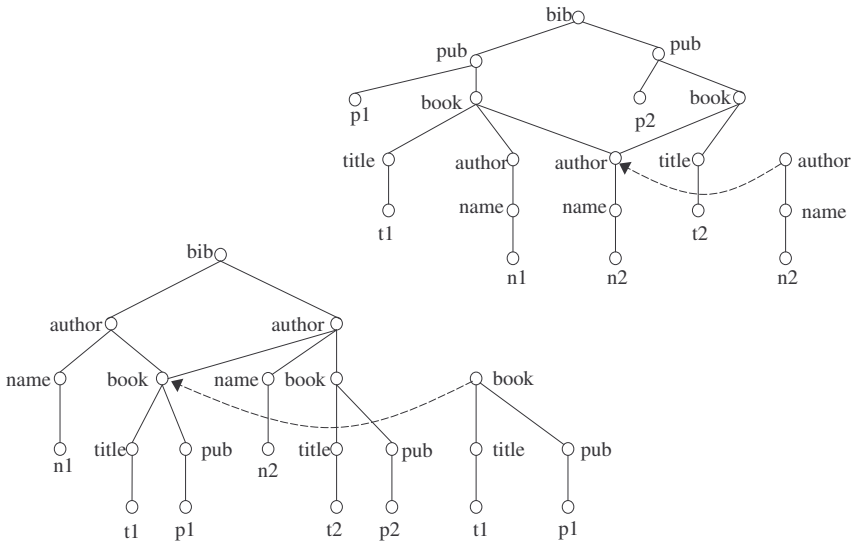
- Reversible transformation
- Construct *canonical* graph



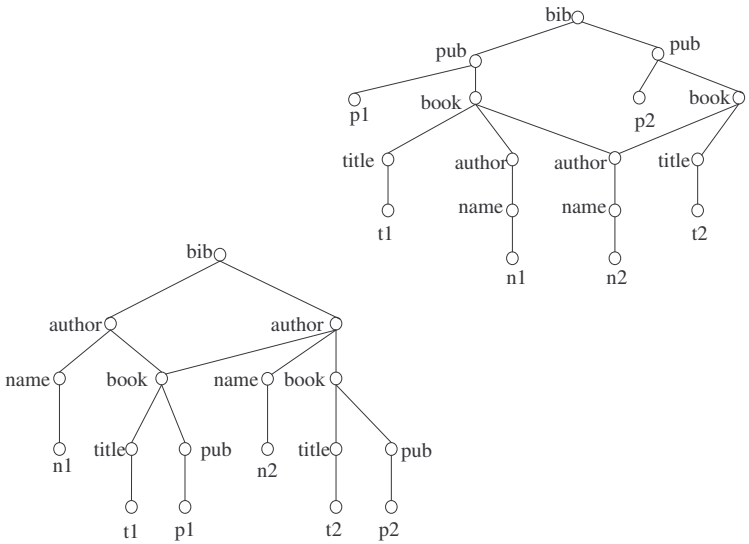
Same Data?



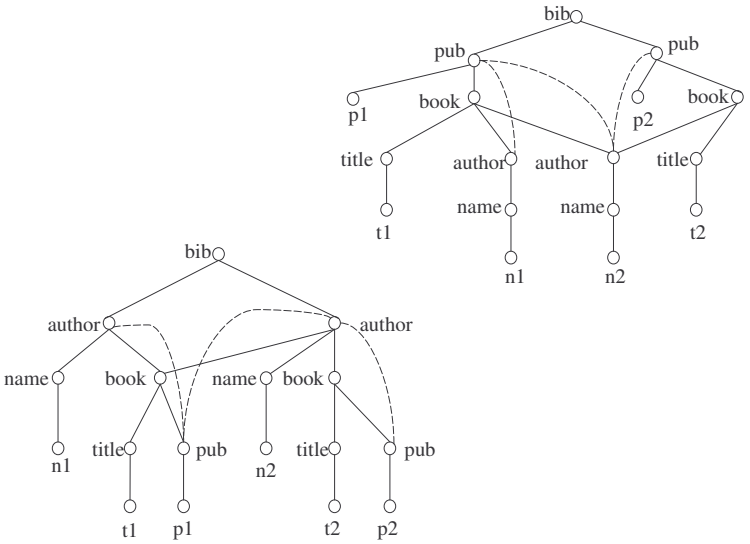
Group Identical Subtrees



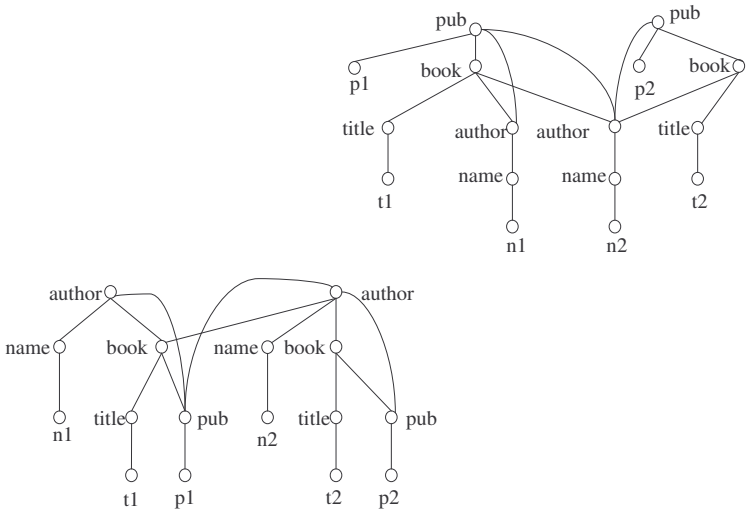
After Grouping



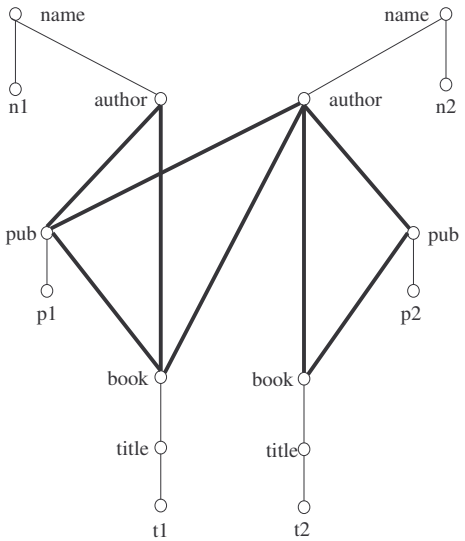
Associate Closest Nodes



After Association

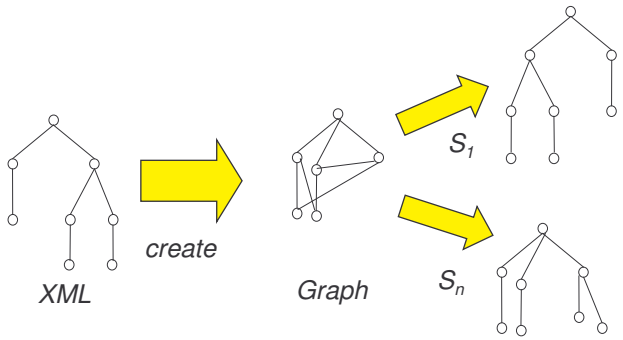


Graphs are Isomorphic



Compaction

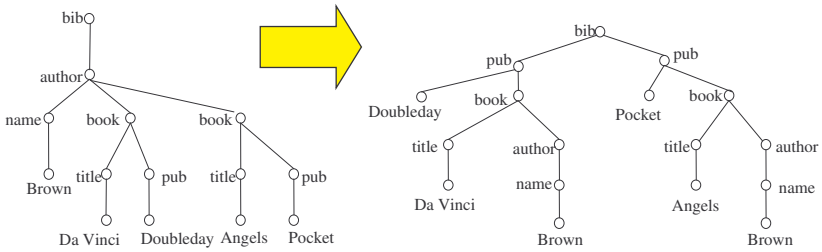
- Many different reversible signatures, $S_1 \ldots S_n$



- Different signatures produce different-sized documents
- Choosing *smallest* is NP-hard

Use Heuristic

- How many parents to each child?
  - 1-1 – either can be parent
  - 1-many – make 1 side the parent
  - Many-many – make smallest side the parent
- Example:
  - author-name is 1-1, keep as is
  - publisher-book is 1-many so move it above book
  - book-author is many-many (1000-2000), move book up



Conclusion

- Experiment
  - DBLP article data, 309KB, 7312 elements

```
<article>…
  <year>2005</year>…
  <journal>TODS</journal>…
 </article>
```

  - Compacted, 252 KB, 5441 elements

```
<year>2005
  <journal>TODS…
    <article>…</article>…
  </journal>…
</year>
```

  - Compaction reduced file size by 18%, # of elements by 25%
  - Amount of compaction achieved depends on the data
- Can use a *reversible* restructuring to compact