Extending the Relational Data Model for Disjunctive Genealogical Data

Lars Olson
DEG Lab
BYU Computer Science Dept.

Supported by National Science Foundation
Grant #0083127

Introduction

- Multiple sources for genealogical data
  - Census records
  - Birth / death records
  - Journals / diaries
  - Etc.
- Some data doesn’t fit
  - Multiple data sources → conflicting data
  - Uncertain or imprecise data
  - Constraint violations
- Not always possible to resolve

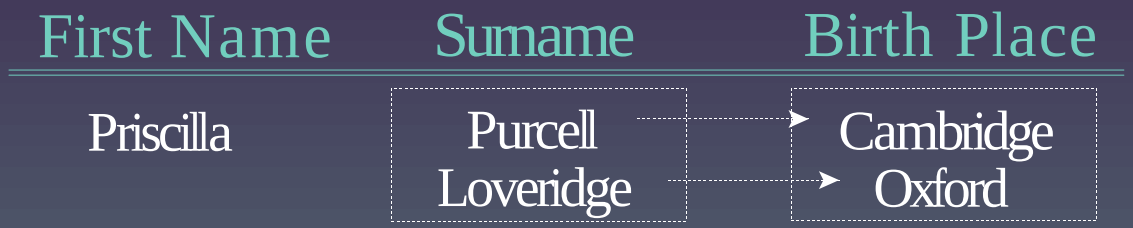
Disjunctive Databases

"OR-tables," Imielinski and Vadaparty, 1989

Name	Birth Date	Marriage Date	Death Date
James I	Dec. 1394	2 Feb. 1423 2 Feb. 1424	21 Feb. 1436 21 Feb. 1437
Joseph Harrison	26 Jan. 1781 26 Jan. 1782 26 Jul. 1782	19 Dec. 1811	5 Apr. 1861
⋮	⋮	⋮	⋮

Shortcomings of "OR-tables"

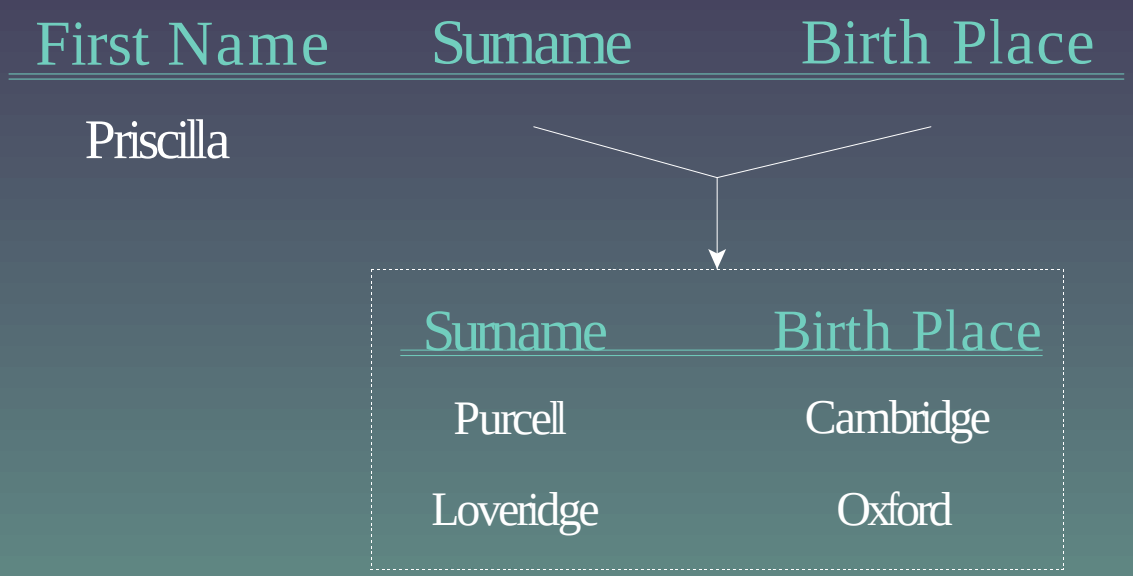
- Can't correlate between possible values



- Answering queries in general is CoNP-complete (Imielinski & Vadaparty)

Sub-relation Data Construct

- Solution: store the correlated data in its own relation



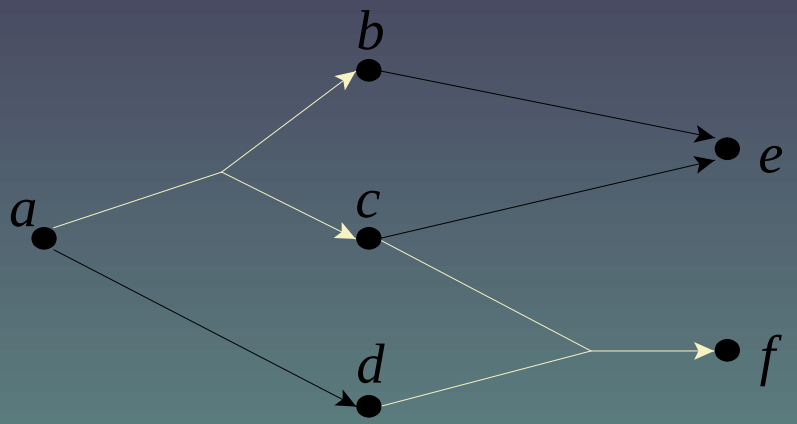
Disjunctive Database Problems

- How do we avoid the CoNP-completeness problem and answer queries efficiently?
- If more than one value is possible, which one is the most likely?
- Other questions to be solved as part of the thesis:
  - Where are the constraint violations?
  - How do we map sub-relations to physical storage?
  - How do we efficiently update the database?

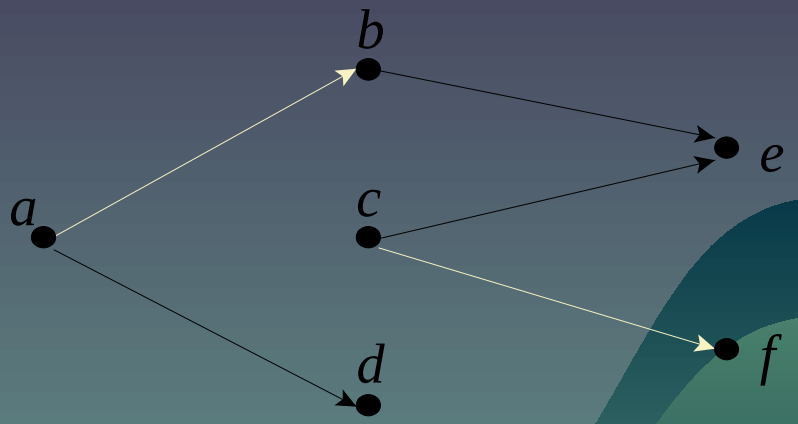
Transitive Closure of Disjunctive Graphs

Solving the CoNP completeness problem [LYY95]

Disjunctive graph



Possible model



Transitive closure of a: $\{a, d, e\}$

Using Disjunctive Graphs to Answer Queries

Table *Person:*

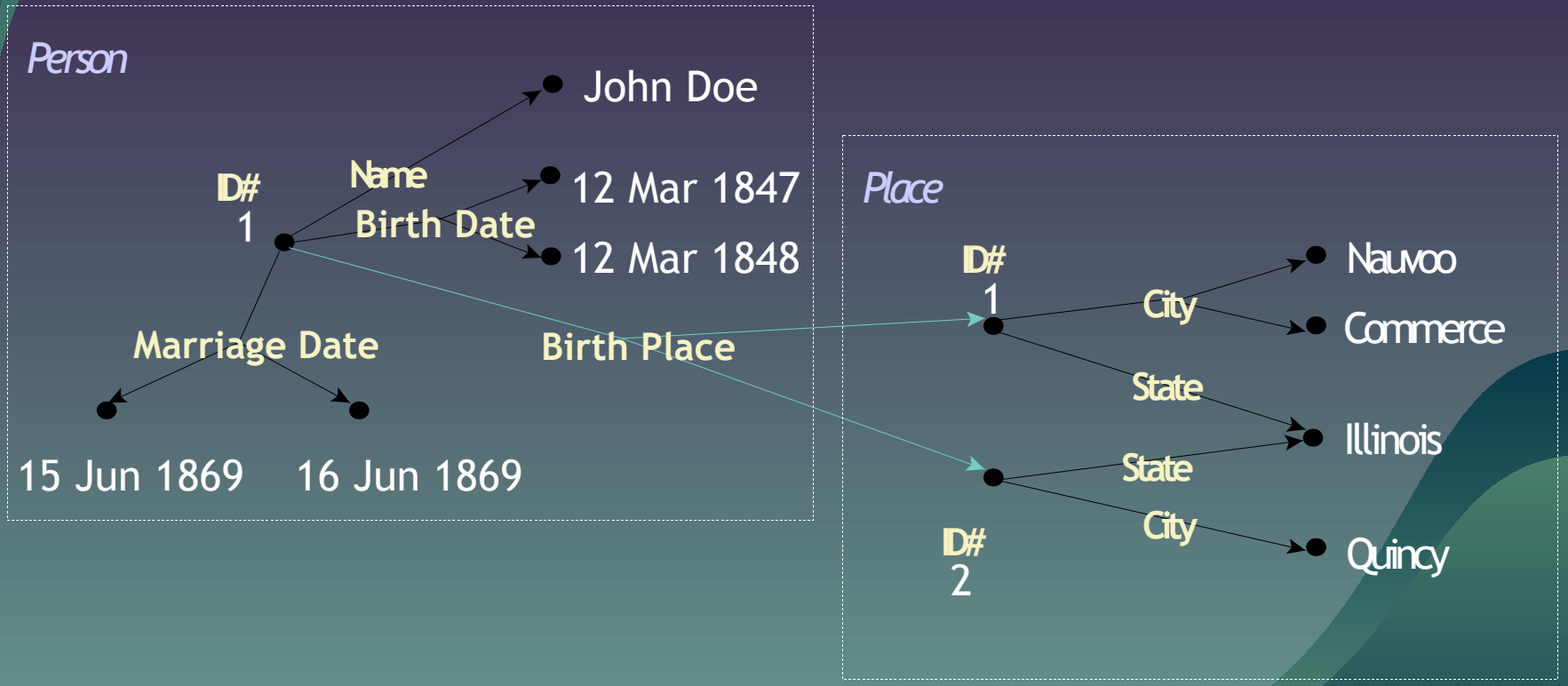
ID#	Name	Birth Date	Birth Place ID# (references Table *Place*)	Marriage Date
1	John Doe	12 Mar. 1840 or 12 Mar. 1841	1 or 2	15 Jun. 1869 or 16 Jun. 1869
⋮	⋮	⋮	⋮	⋮

Table *Place:*

ID#	City	State
1	Commerce or Nauvoo	Illinois
2	Quincy	Illinois
⋮	⋮	⋮

Using Disjunctive Graphs to Answer Queries

$$\pi_{State}(\sigma_{ID=1} Person \bowtie Place)$$

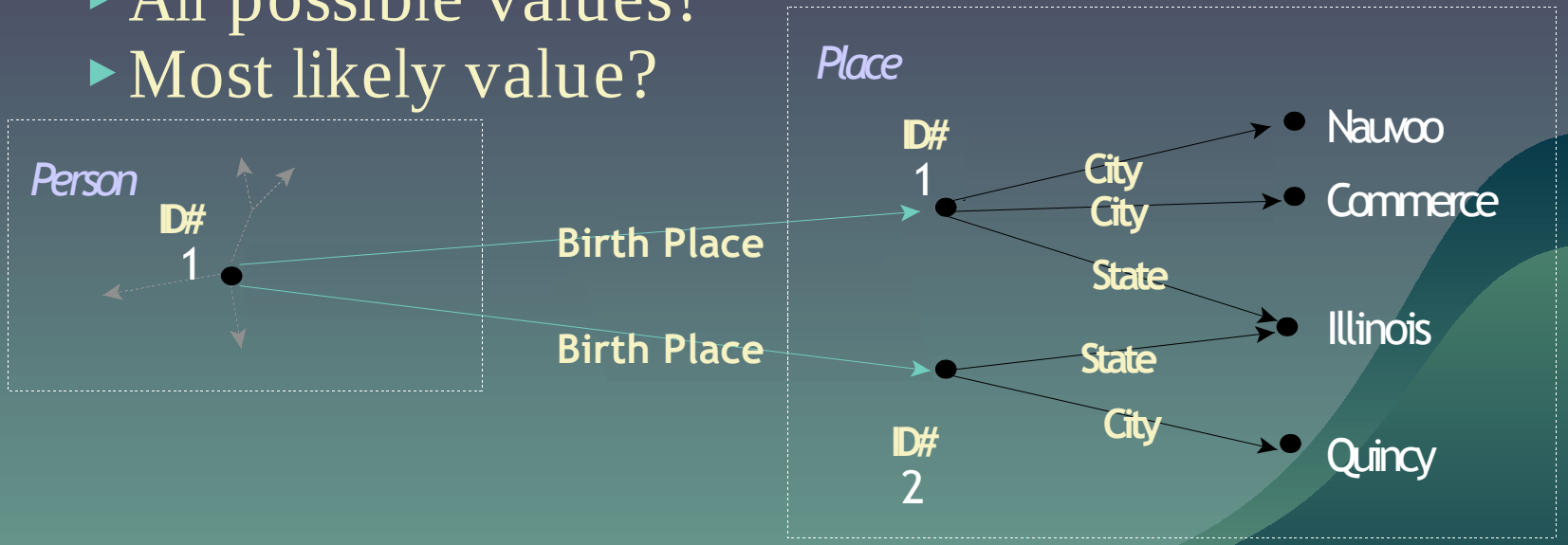


Using Disjunctive Graphs to Answer Queries

$$\pi_{City,State}(\sigma_{ID=1} Person \bowtie Place)$$

...meaning what?

- Definitely known?
- All possible values?
- Most likely value?

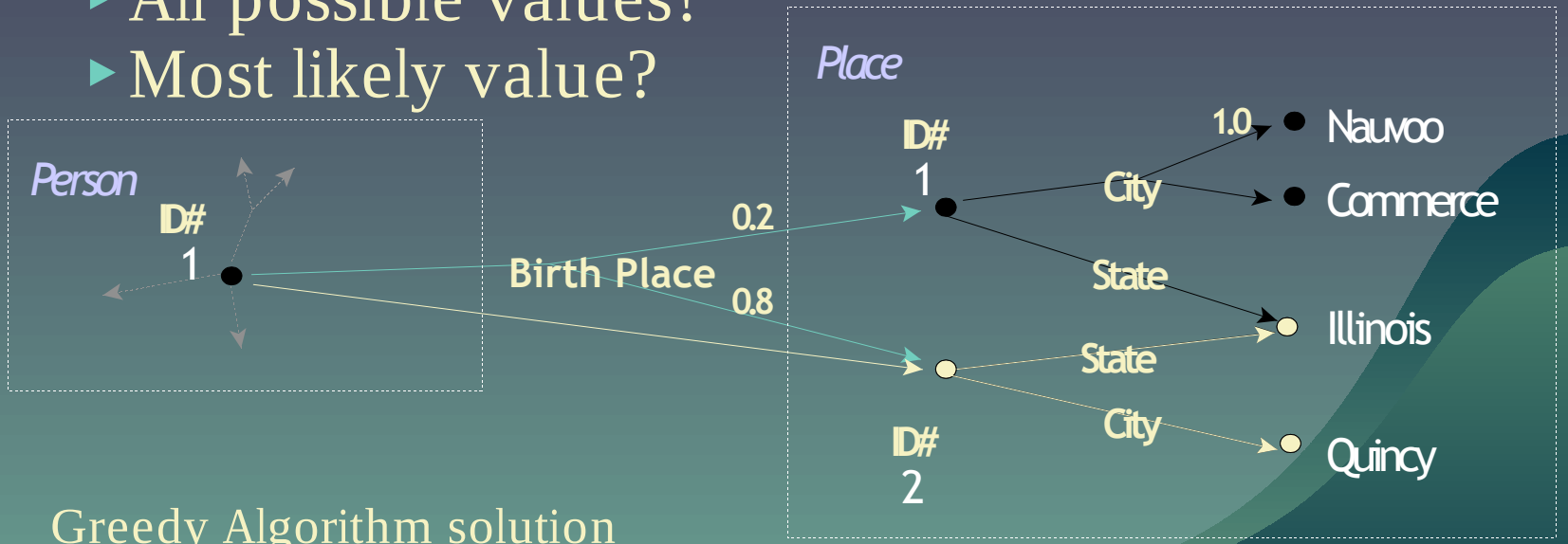


Using Disjunctive Graphs to Answer Queries

$$\pi_{City,State}(\sigma_{ID=1} Person \bowtie Place)$$

...meaning what?

- Definitely known?
- All possible values?
- Most likely value?

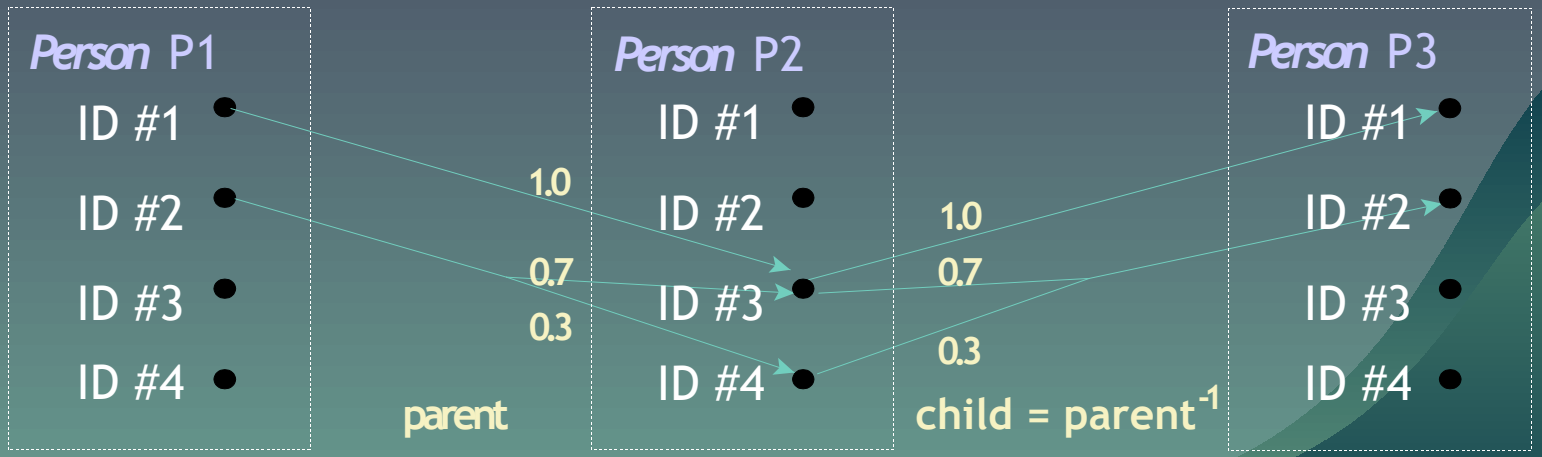


Greedy Algorithm solution

Finding a Most-Likely Interpretation

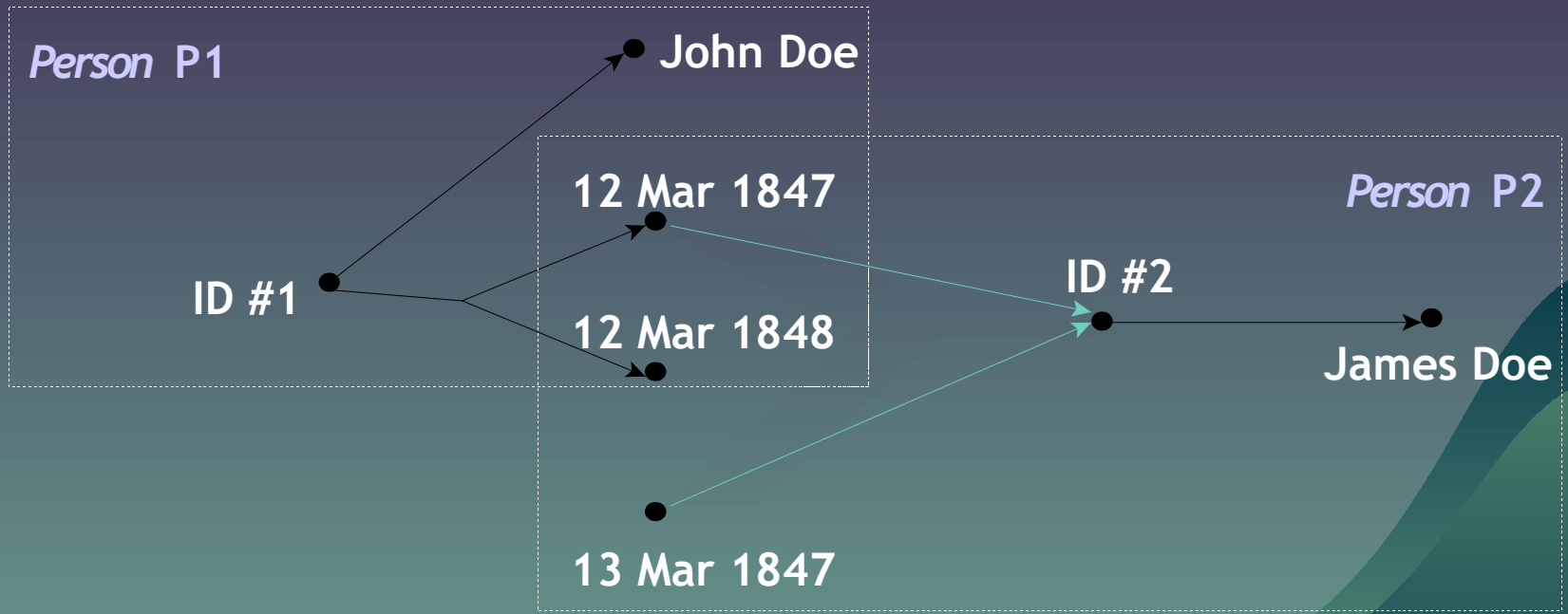
Example: Unknown birth dates among family members

- Example constraints: parents must be born before their children, and each child must be born at least 9 months apart (except perhaps twins)
- Build a relation containing all family members



Using Disjunctive Graphs to Answer Queries

$$\pi_{P1.Name,\ P2.Name}(Person\ P1 \bowtie_{P1.BirthDate = P2.BirthDate} Person\ P2)$$



Conclusions

- Genealogical data can be stored in a disjunctive database format.
- Many common queries can be computed in polynomial time.
- We can detect intractable queries and limit the search space required, usually enough to get polynomial time.