Supporting Queries Spanning Across Phases of Evolving Artifacts using Steiner Forests

Siarhei Bykau
bykau@disi.unitn.eu

joint work with John Mylopoulos, Flavio Rizzolo and Yannis Velegrakis
University of Trento via Sommarive 18, 38050 Povo, Trento – Italy

The Problem of Evolution

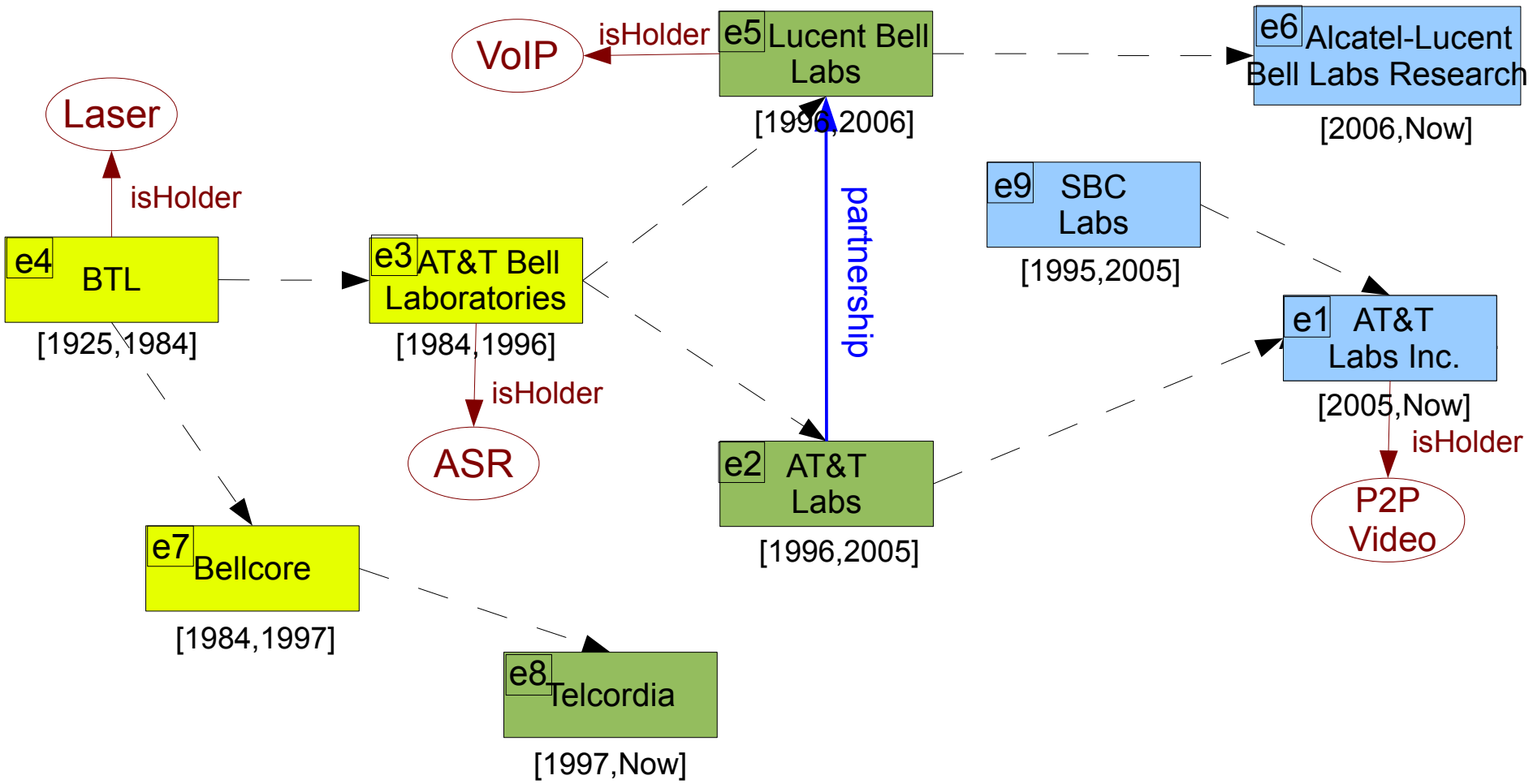
- Schema evolution [McBrien et al. 2002]
- Data evolution [Chawathe et al. 1998]
- Data transformation [Velegrakis et al. 2005]
- Temporal databases [Buneman et al. 2002]
- ...

The Problem of Evolution

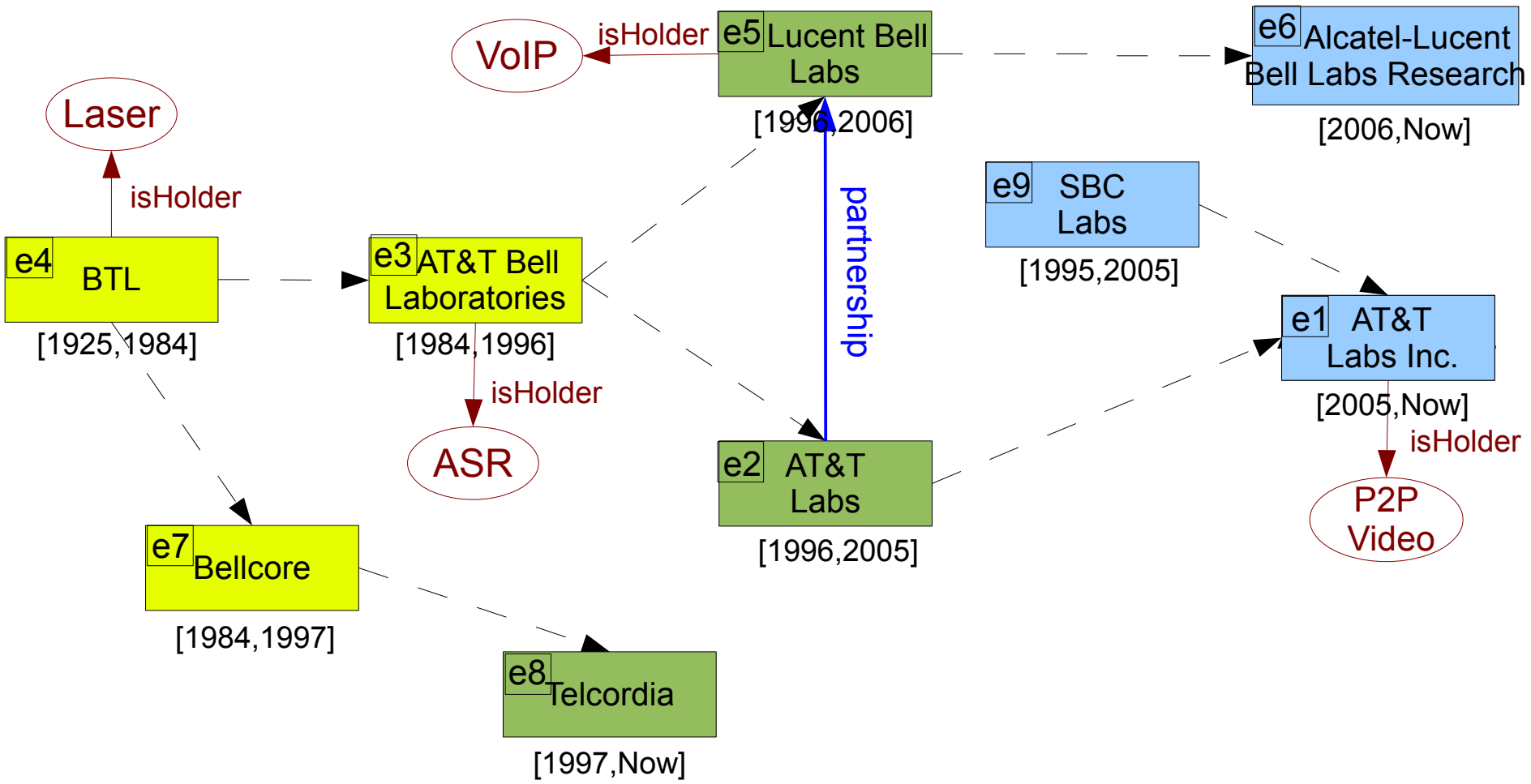
- Schema evolution [McBrien et al. 2002]
- Data evolution [Chawathe et al. 1998]
- Data transformation [Velegrakis et al. 2005]
- Temporal databases [Buneman et al. 2002]
- ...

Evolution doesn't span different concepts

AT&T Labs History



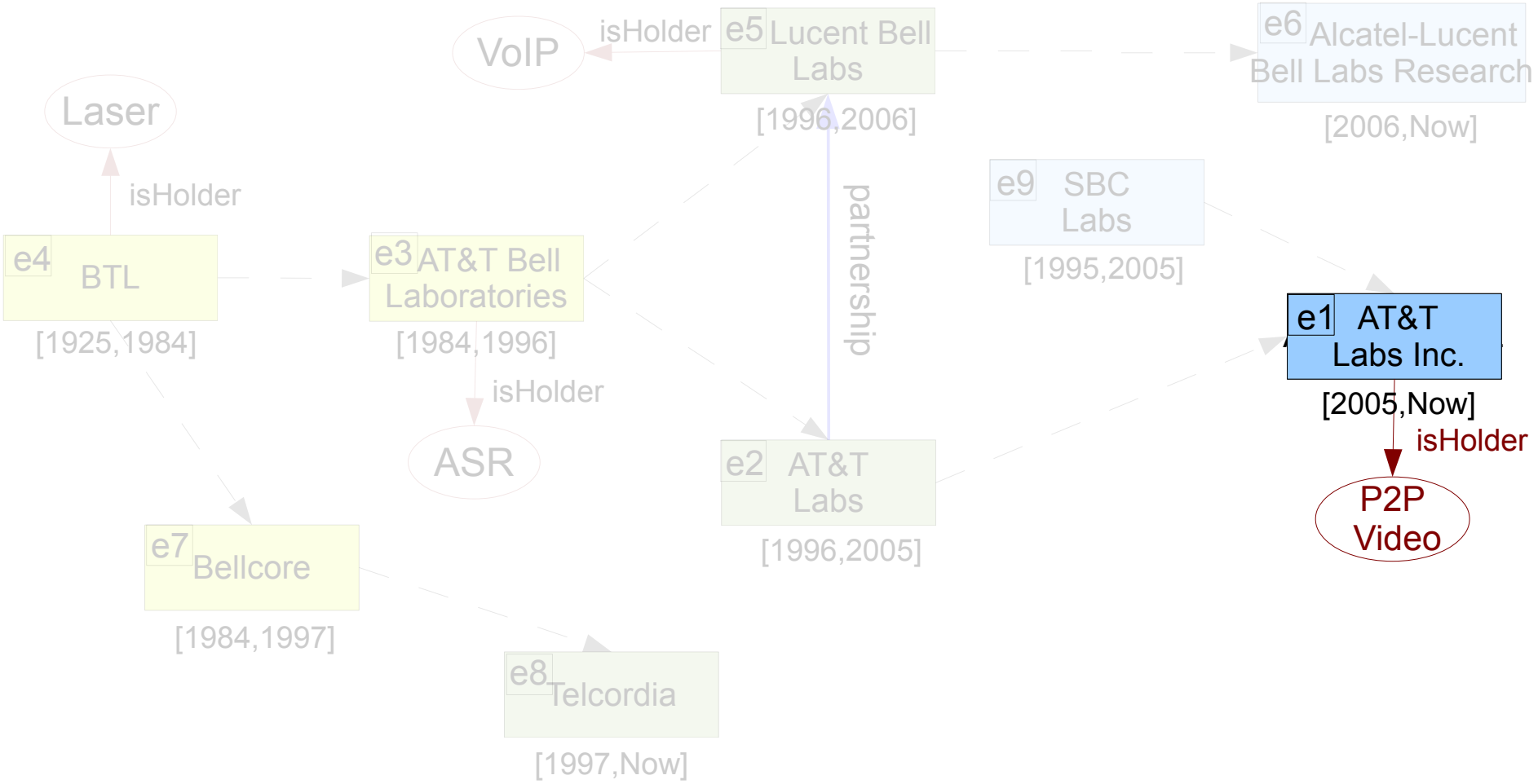
Motivating Example



Find the patents of AT&T Labs Inc.:

$x(isHolder:$y):-$x(name:AT&T Labs Inc; isHolder:$y)

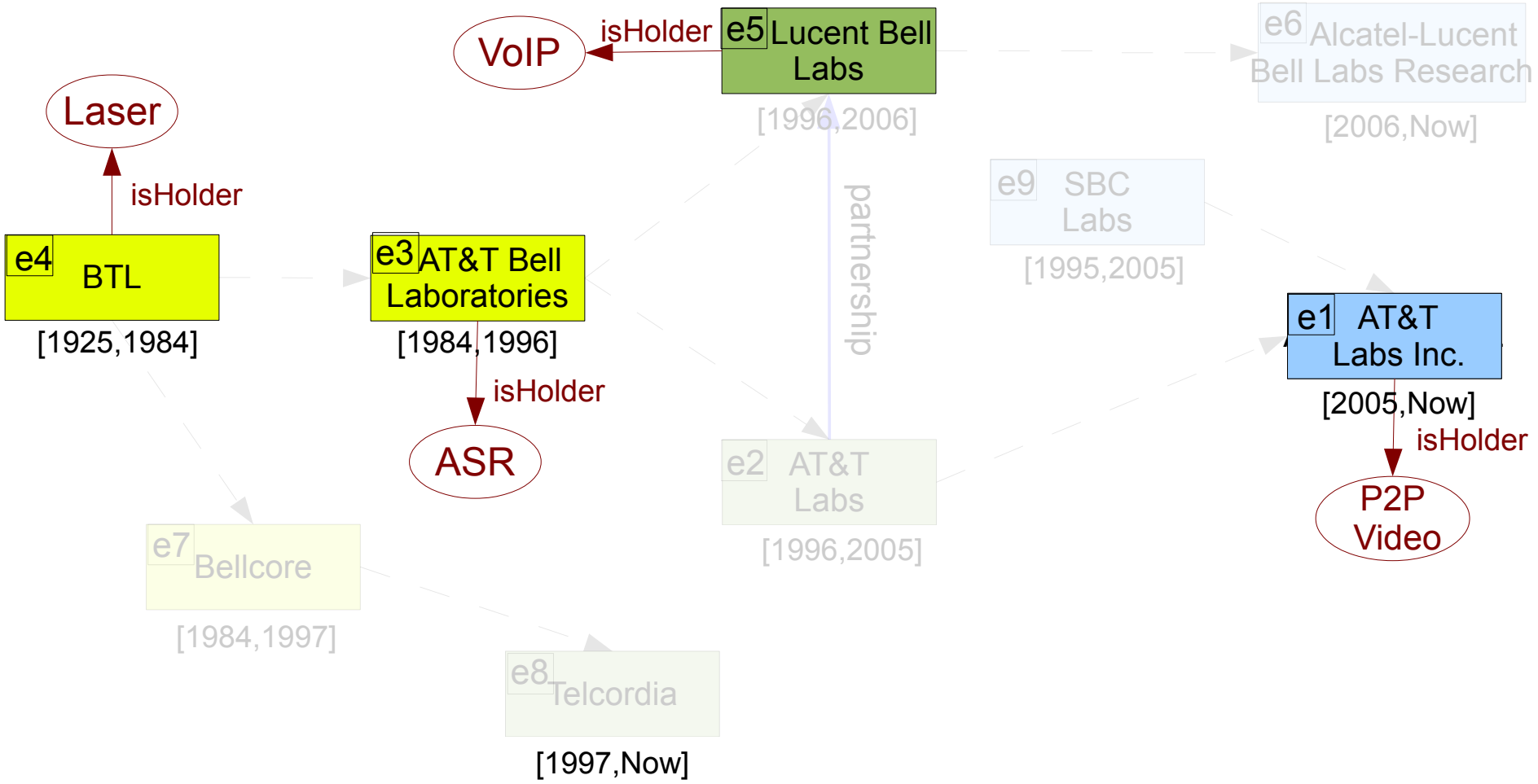
Motivating Example



Query: find the patents of AT&T Labs Inc.:

$x(isHolder:$y):-$x(name:AT&T Labs Inc; isHolder:$y)

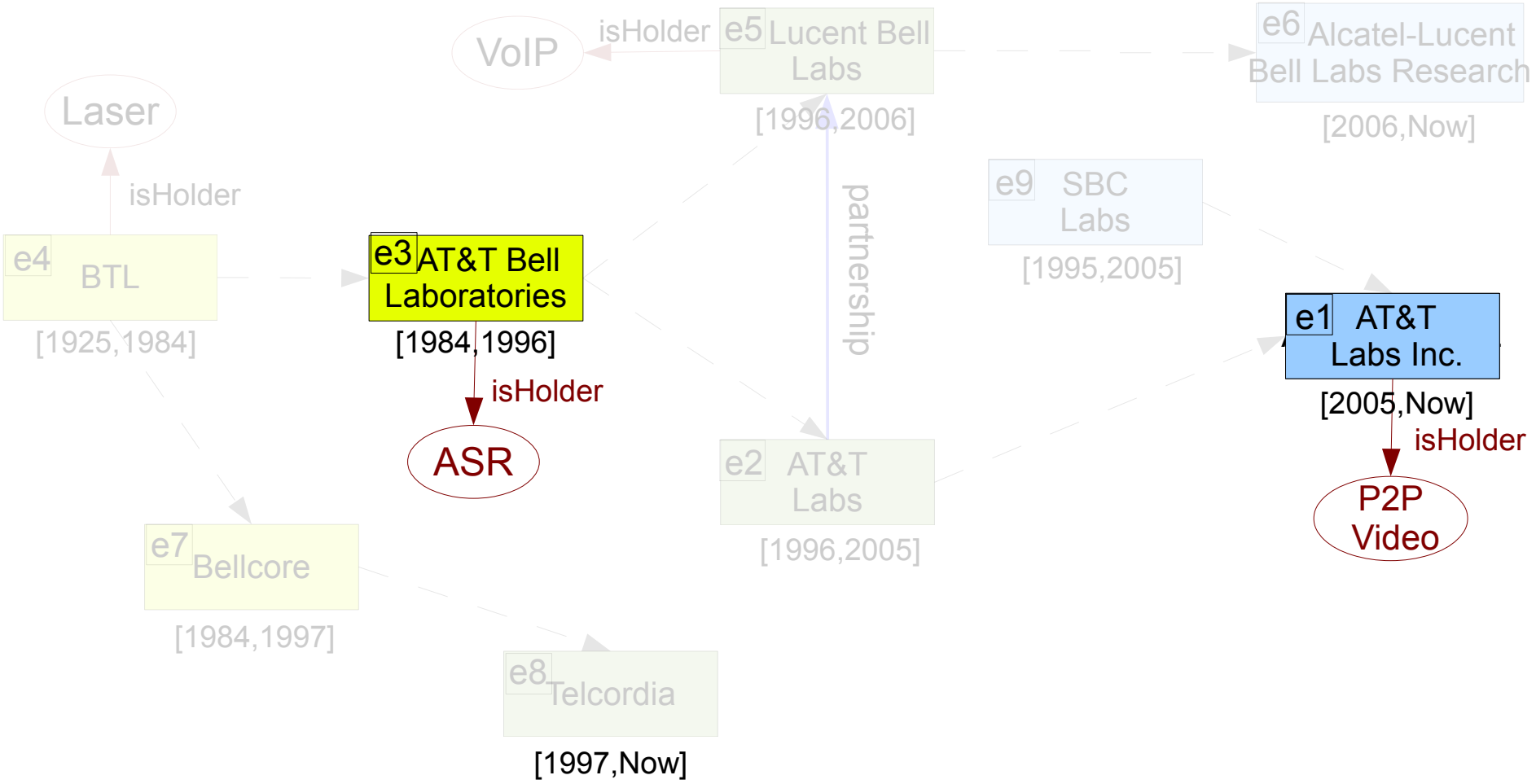
Motivating Example



Query: find the patents of AT&T Labs Inc.:

$x(isHolder:$y):-$x(name:AT&T Labs Inc; isHolder:$y)

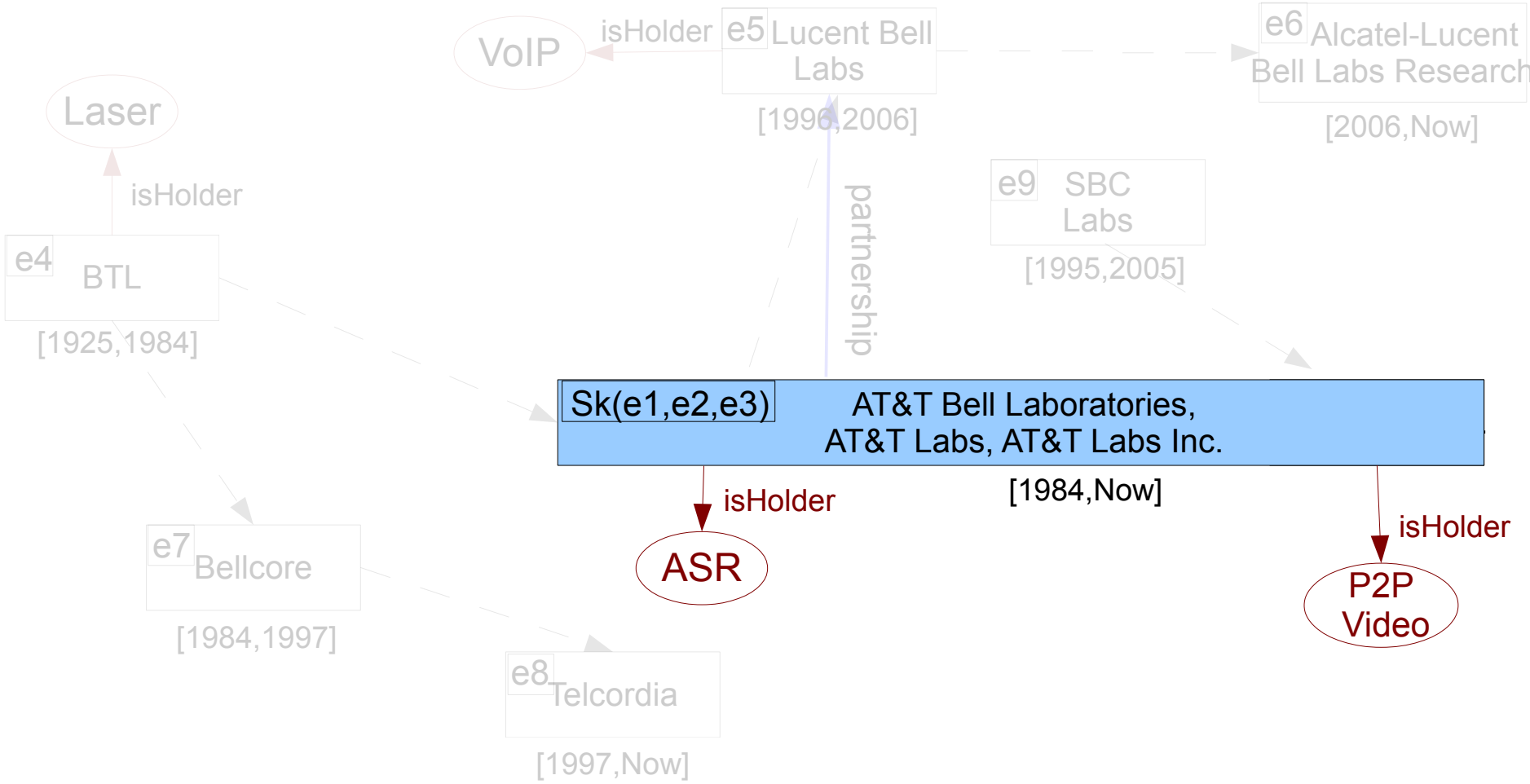
Entity Coalescence



Query: find the patents of AT&T Labs Inc.:

$x(isHolder:$y):-$x(name:AT&T Labs Inc; isHolder:$y)

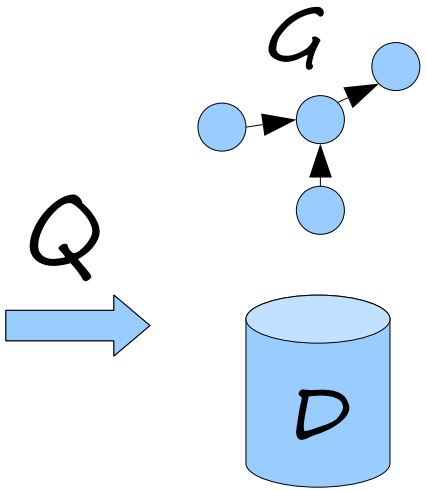
Entity Coalescence



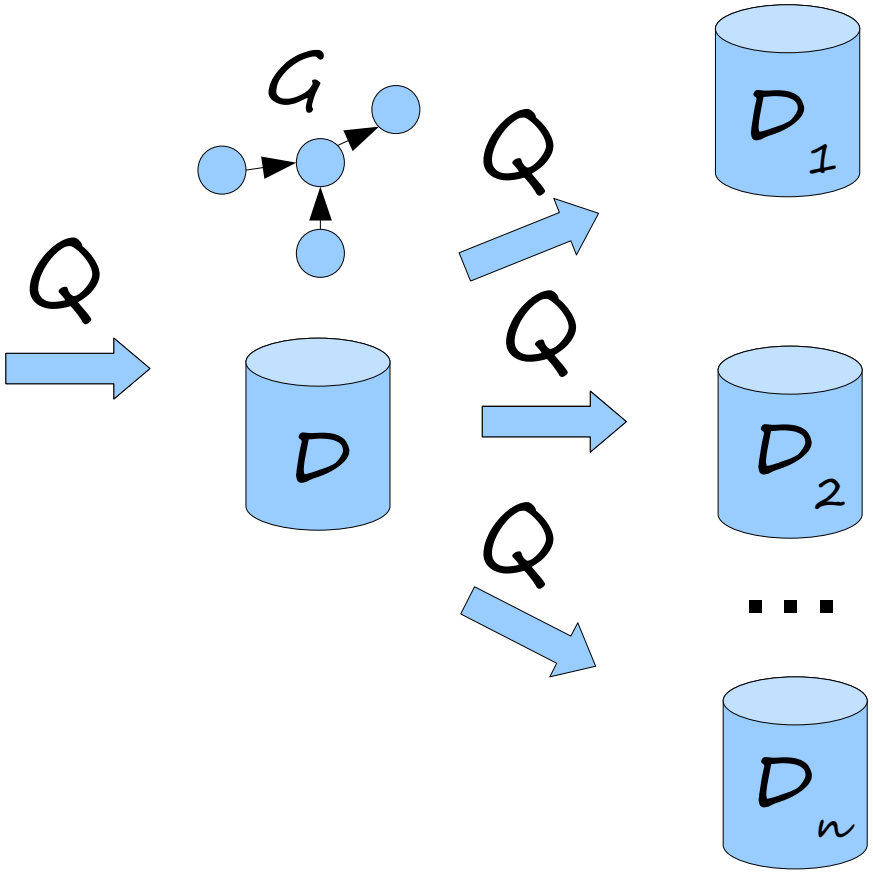
Query: find the patents of AT&T Labs Inc.:

$x(isHolder:$y):-$x(name:AT&T Labs Inc; isHolder:$y)

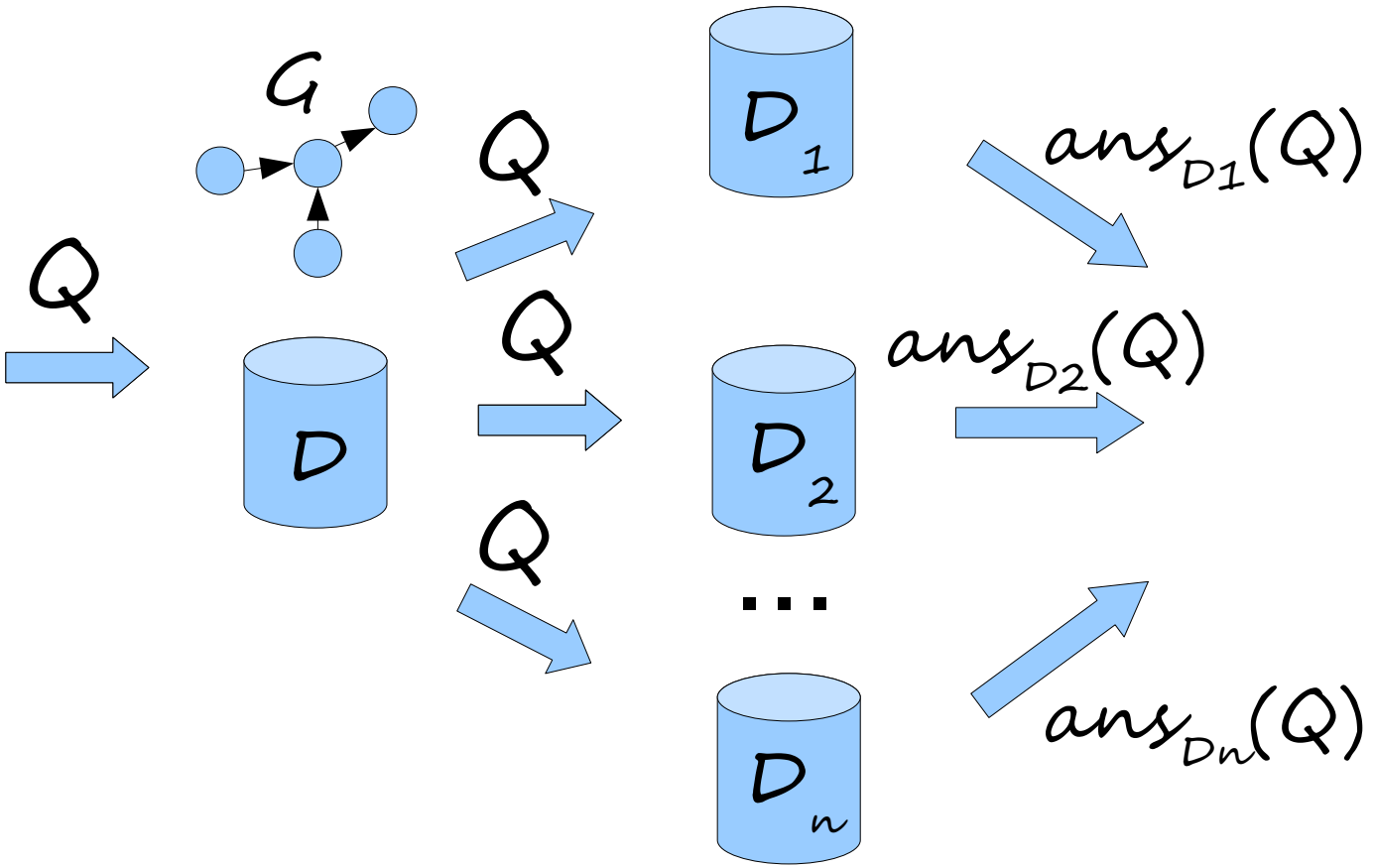
Query Semantics



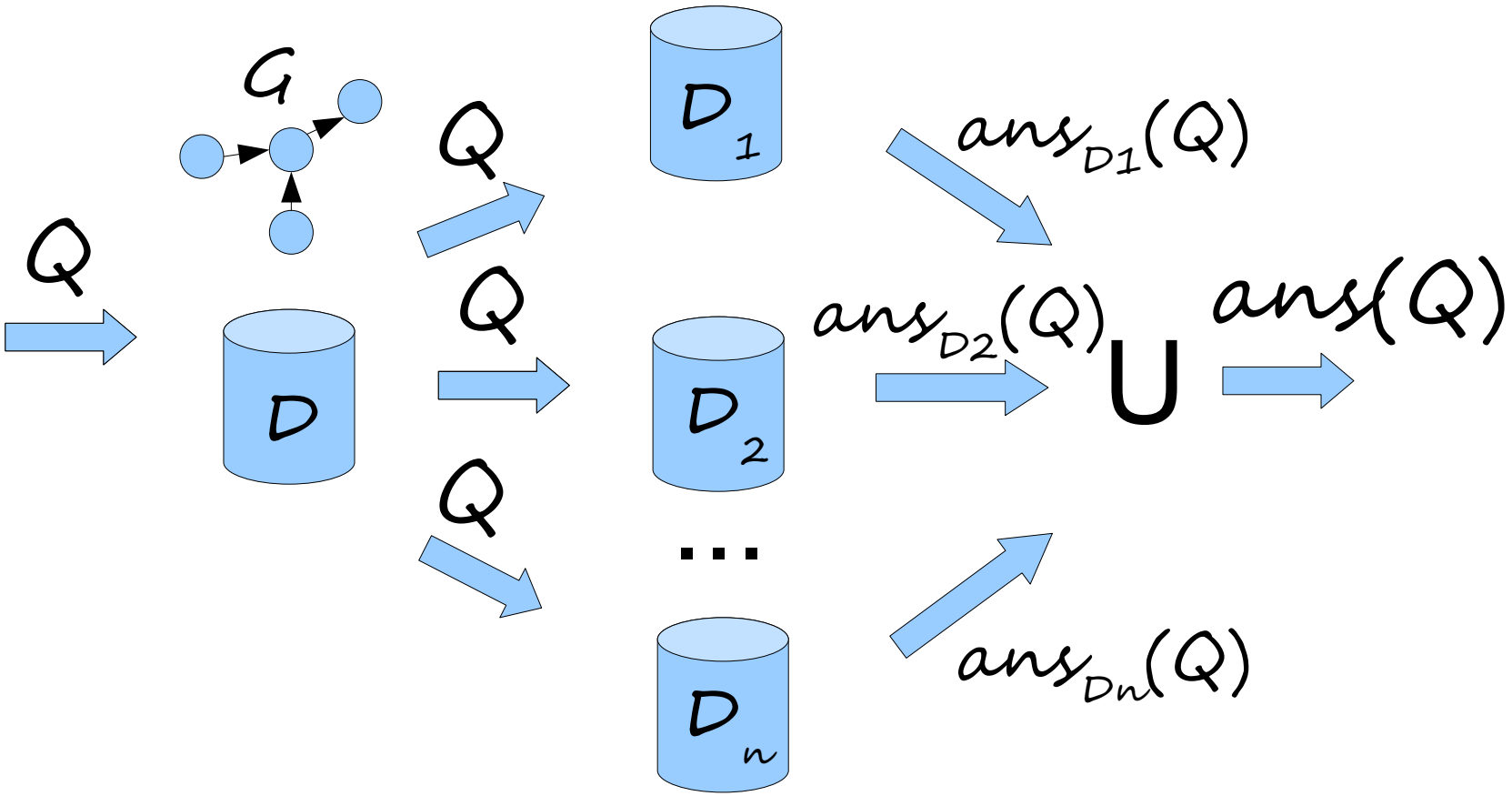
Query Semantics



Query Semantics



Query Semantics



Variable Assignments

$x(name:AT&T Labs Inc; isHolder:$y)

$x	$y	Possible World	Answer	Cost
e1	P2P Video	0	e1(isHolder:"P2P Video")	0
Sk(e1,e2)	**P2P Video**	**e1,e2**	**Not generated**	**1**
Sk(e1,e2,e3)	**P2P Video**	**e1,e2,e3**	**Not generated**	**2**
Sk(e1,e2,e3)	ASR	e1,e2,e3	Sk(e1,e2,e3)(isHolder:"ASR")	2
Sk(e1,e2,e3,e4)	Laser	e1,e2,e3,e4	Sk(e1,e2,e3,e4)(isHolder:"Laser")	3
...	...	...	...	...

Query Evaluation Techniques

- Brute force: generate all possible worlds

Query Evaluation Techniques

- Brute force: generate all possible worlds

computationally expensive

Query Evaluation Techniques

- Brute force: generate all possible worlds

  computationally expensive

- Materialize possible worlds

Query Evaluation Techniques

- Brute force: generate all possible worlds

  computationally expensive

- Materialize possible worlds

  needs too much space

Query Evaluation Techniques

- Brute force: generate all possible worlds

  computationally expensive

- Materialize possible worlds

  needs too much space

- Materialize only the maximum possible world

Query Evaluation Techniques

- Brute force: generate all possible worlds

  computationally expensive

- Materialize possible worlds

  needs too much space

- Materialize only the maximum possible world

  redundant coalescences

  doesn't distinguish different evolution phases of an entity

On-the-fly coalescence computations

On-the-fly coalescence computations

Step1: $x(name:”AT&T Labs Inc.”),$x(isHolder:$y)

Step2:

{$x}	{$x,$y}
{e1}	{e1, ”P2P Video”}
	{e3, ASR}
	{e4, Laser}
	{e5, VoIP}

On-the-fly coalescence computations

Step1: $x(name:”AT&T Labs Inc.”),$x(isHolder:$y)

Step2:

{$x}	**{$x,$y}**
{e1}	{e1, ”P2P Video”}
	{e3, ASR}
	{e4, Laser}
	{e5, VoIP}

Step3:

{$x,$x,$y}
{e1,e1, ”P2P Video”}
{e1,e3, ASR}
{e1,e4, Laser}
{e1,e5, VoIP}

On-the-fly coalescence computations

Step1: $x(name:”AT&T Labs Inc.”),$x(isHolder:$y)

Step2:

{$x}	{$x,$y}
{e1}	{e1, ”P2P Video”}
	{e3, ASR}
	{e4, Laser}
	{e5, VoIP}

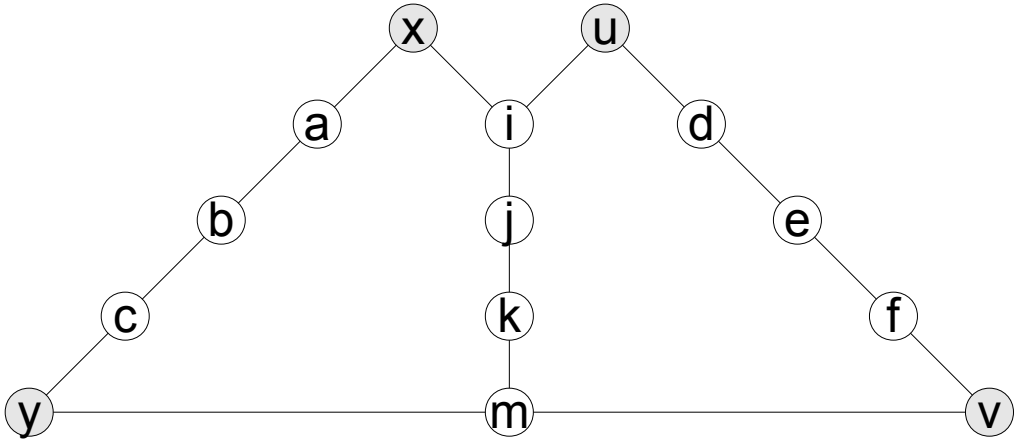
Step3:

{$x,$x,$y}
{e1,e1, ”P2P Video”}
{e1,e3, ASR}
{e1,e4, Laser}
{e1,e5, VoIP}

Step5:

V1={e1,e3}
V2={e1,e4}
V3={e1,e5}

Steiner Forest Problem



V1=(x,y) V2=(u,v)

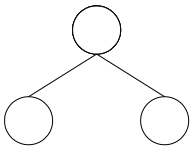
Steiner forest problem is NP-hard [Gassner 2010]

Steiner Tree Algorithm

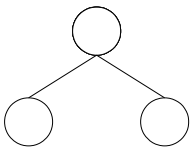
Optimal dynamic programming algorithm for Steiner trees [Ding et al. 2007]:

$$O(3^{\sum l_i} n + 2^{\sum l_i}((\sum l_i + \log n)\, n + m))$$

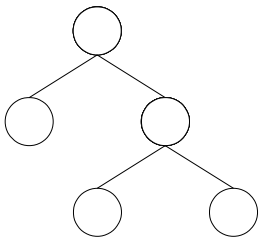
ST(x,y)



ST(u,v)



ST(x,y,u,v)

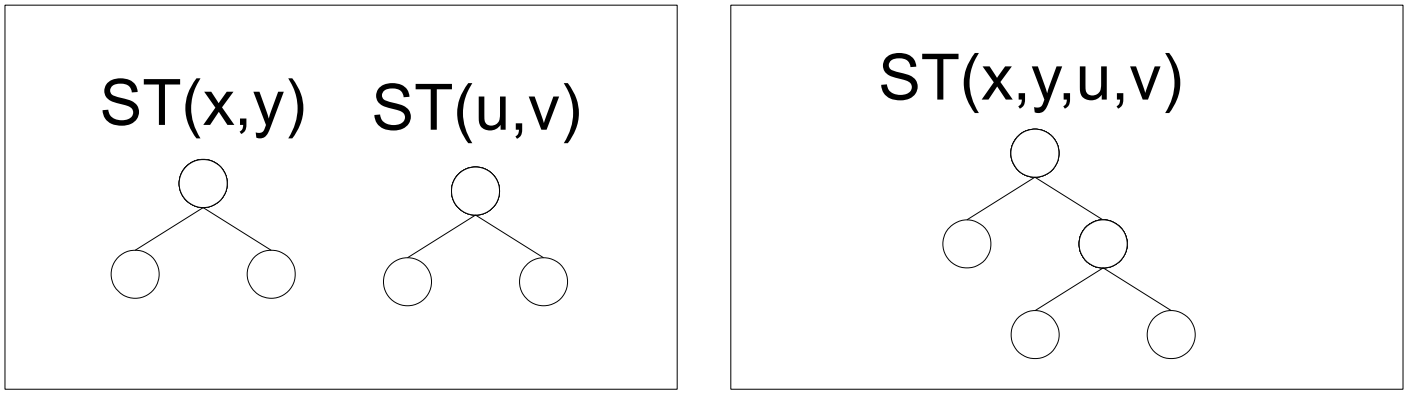


Steiner Forest Algorithm

$$SF(\boldsymbol{V}) = ST(V_i)$$

$$SF(\boldsymbol{V}) = min_{\boldsymbol{H} \subseteq \boldsymbol{V}} (ST(maxflat(\boldsymbol{H})) + SF(\boldsymbol{V} \setminus \boldsymbol{H}))$$

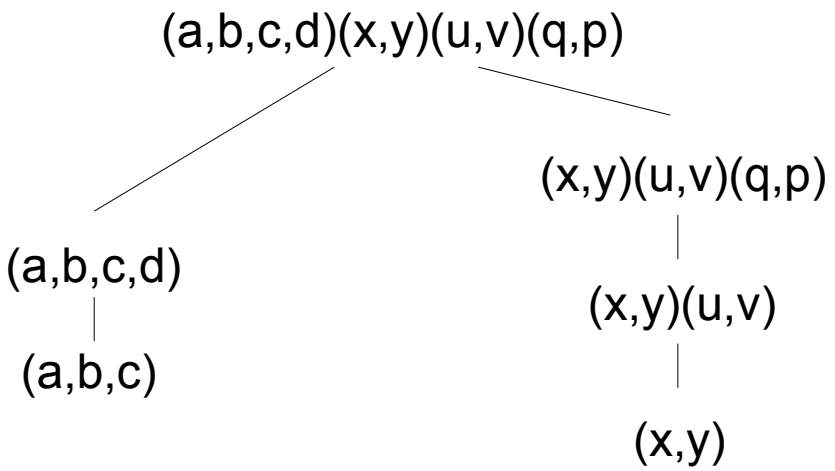
SF((x,y),(u,v)):



$$O(3^L - 2^L(L/2-1) - 1)$$

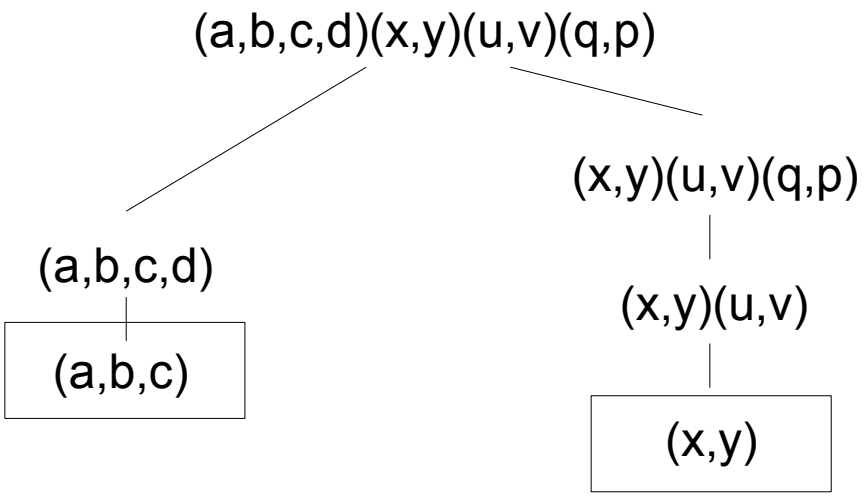
Top-K Query Evaluation Optimization

Lemma: for two sets of sets of nodes V' and V'' on graph G if $V' \subseteq V''$ then $cost(SF(V')) \leqslant cost(SF(V''))$



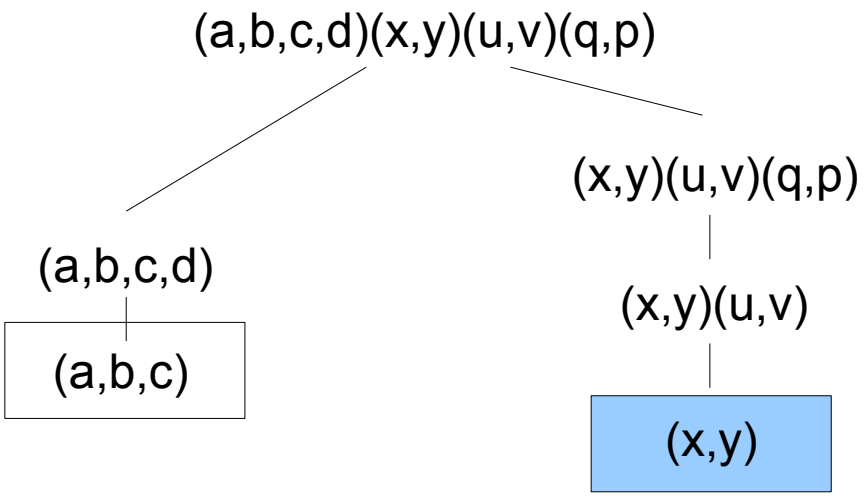
Top-K Query Evaluation Optimization

Lemma: for two sets of sets of nodes V' and V'' on graph G if $V' \subseteq V''$ then $cost(SF(V')) \leqslant cost(SF(V''))$



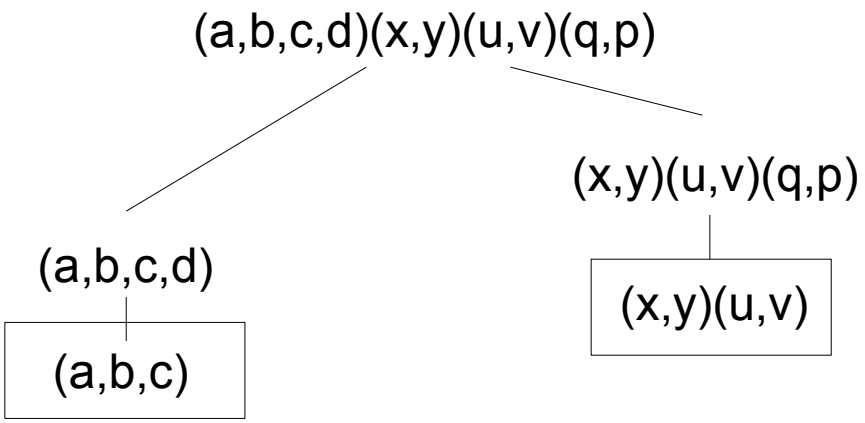
Top-K Query Evaluation Optimization

Lemma: for two sets of sets of nodes V' and V'' on graph G if $V' \subseteq V''$ then $cost(SF(V')) \leqslant cost(SF(V''))$



Top-K Query Evaluation Optimization

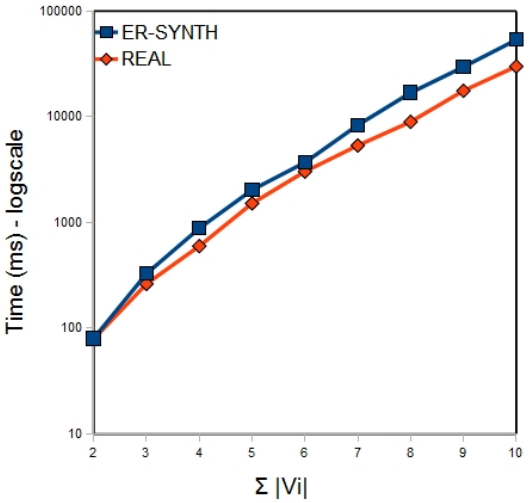
Lemma: for two sets of sets of nodes V' and V'' on graph G if $V' \subseteq V''$ then $cost(SF(V')) \leqslant cost(SF(V''))$



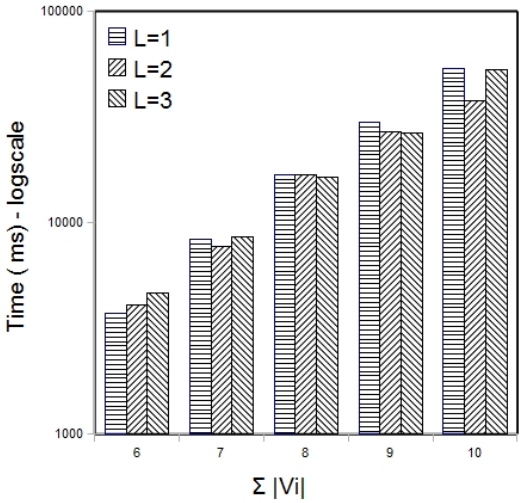
Experiments

- Synthetic Data:
  - Evolution Graph: Erdös-Rényi generator
  - Attribute/Association Data: Zipfian distribution
- Real Data:
  - Evolution Graph: Extracted from the US Trademark dataset
    - REAL_CHAIN
    - REAL_STAR

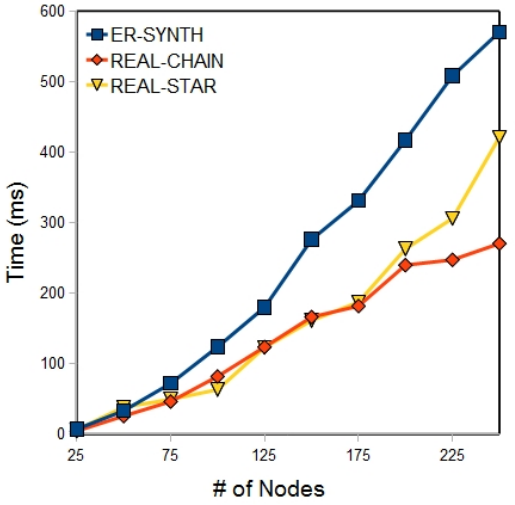
Steiner Forest Experiments



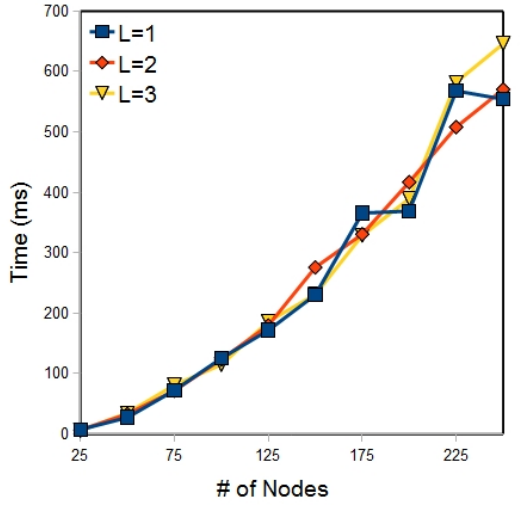
(a)



(b)

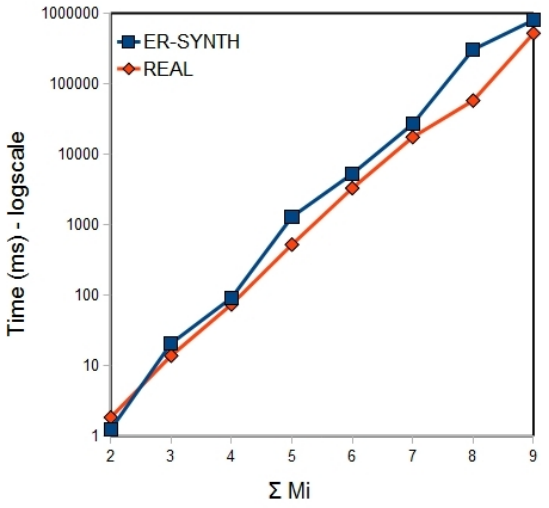


(c)

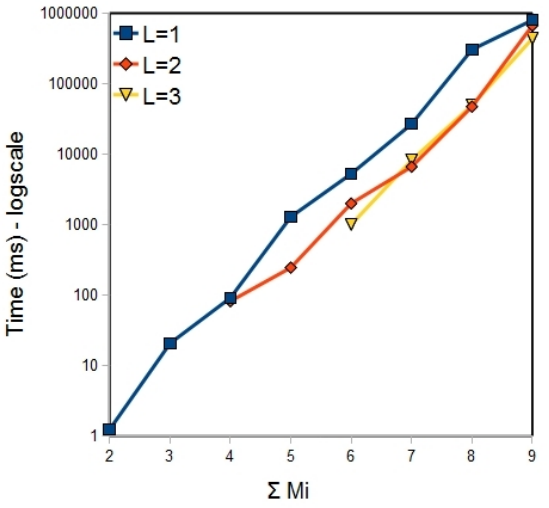


(d)

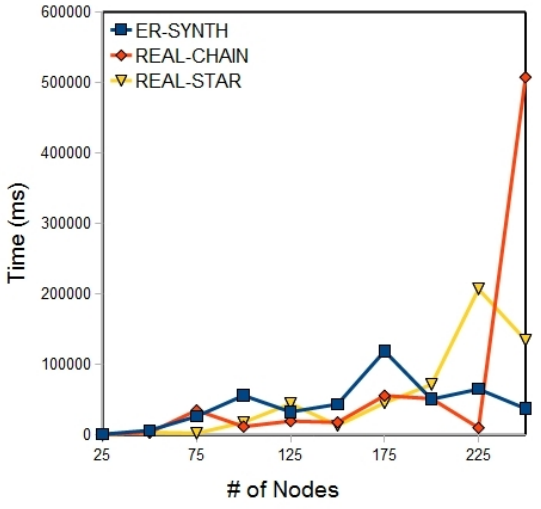
Query Evaluation Experiments



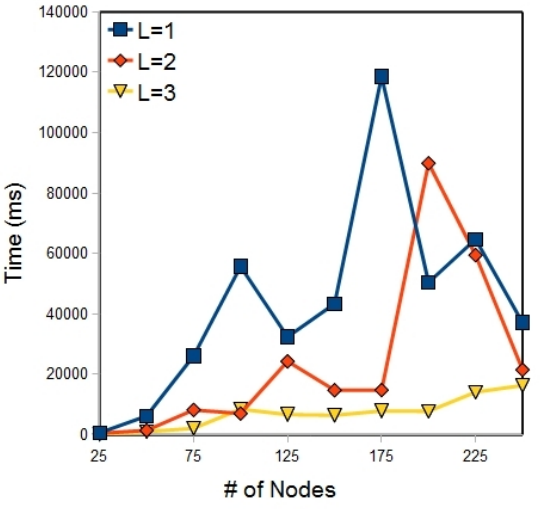
(a)



(b)



(c)

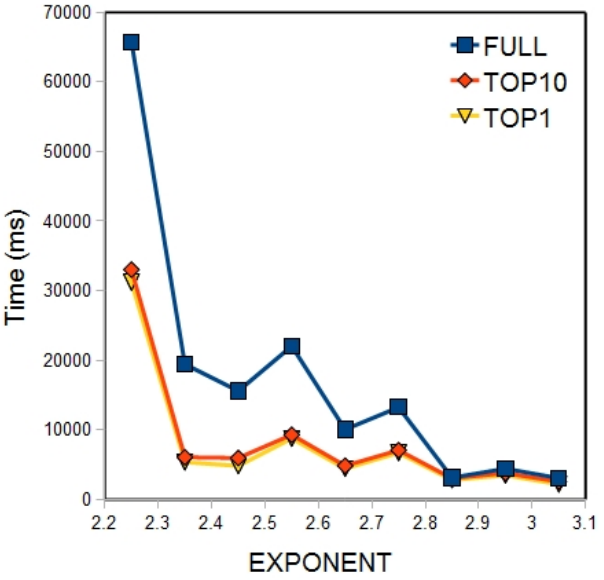


(d)

of connected components and data distribution experiments

# of Branches	s=2.5	s=3.0	s=3.0
1	93,845	44,098	35,382
2	2,400	1,414	4,686
3	374	637	485
4	485	20	91
5	124	260	16

(a)



(b)

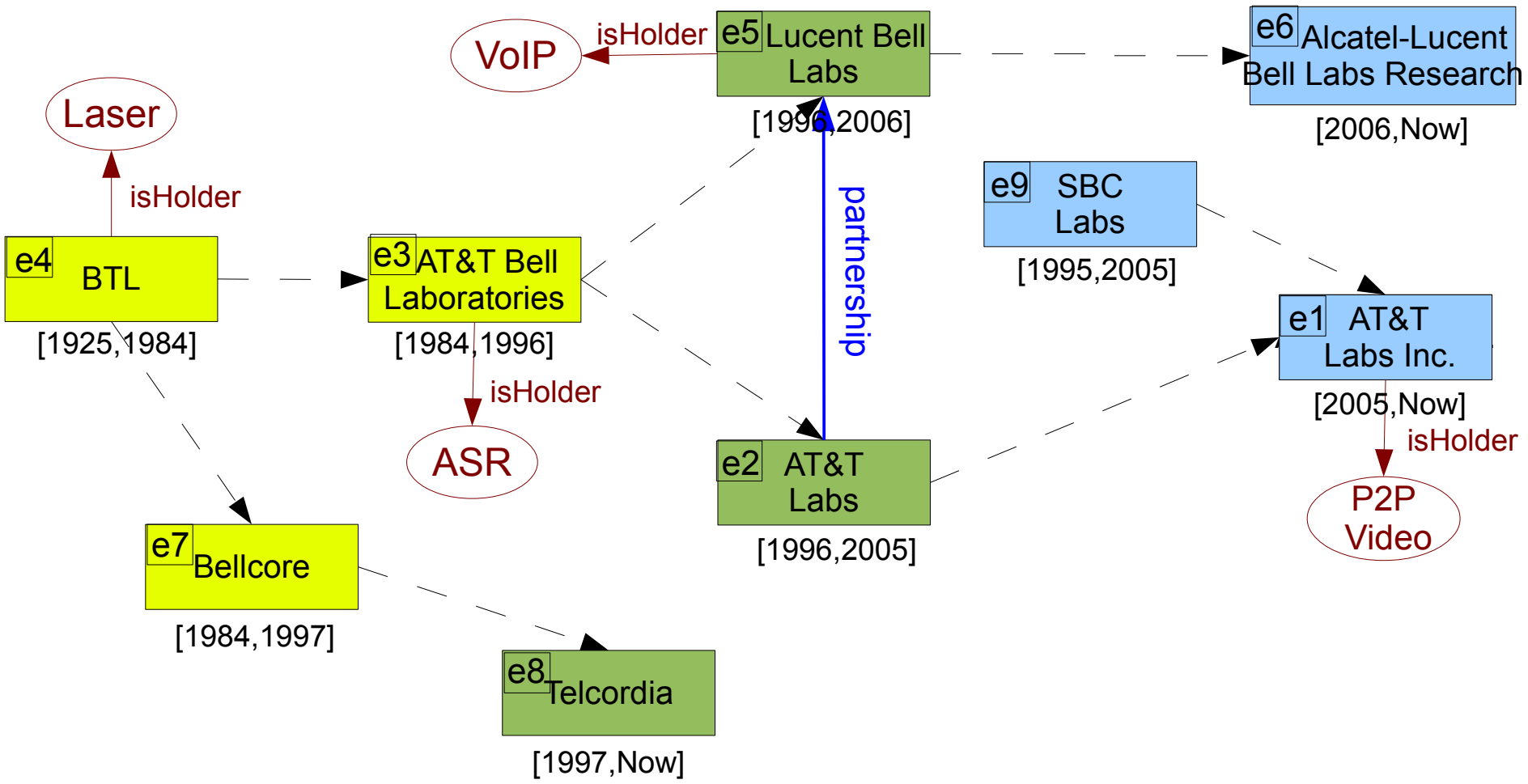
Conclusion

- A new insight into evolution: evolution spans different concepts
- Possible worlds are used to define the semantics of evolution
- A strategy to evaluate a query over possible worlds
- A new optimal algorithm for the Steiner forest problem
- Top-k optimization technique
- Experimental evaluation

Thank you for you attention!

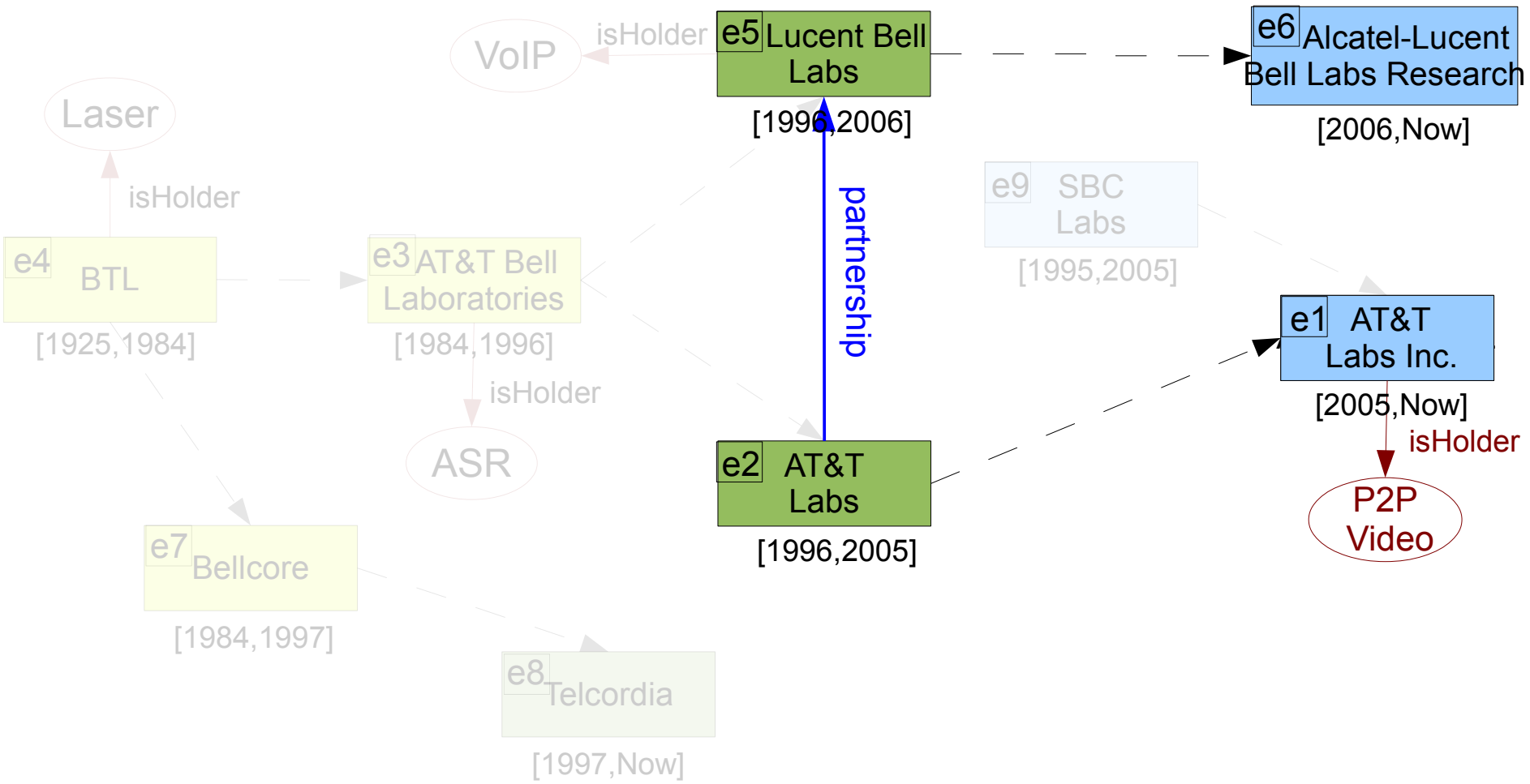
Backup slides

Motivating Query with Associations



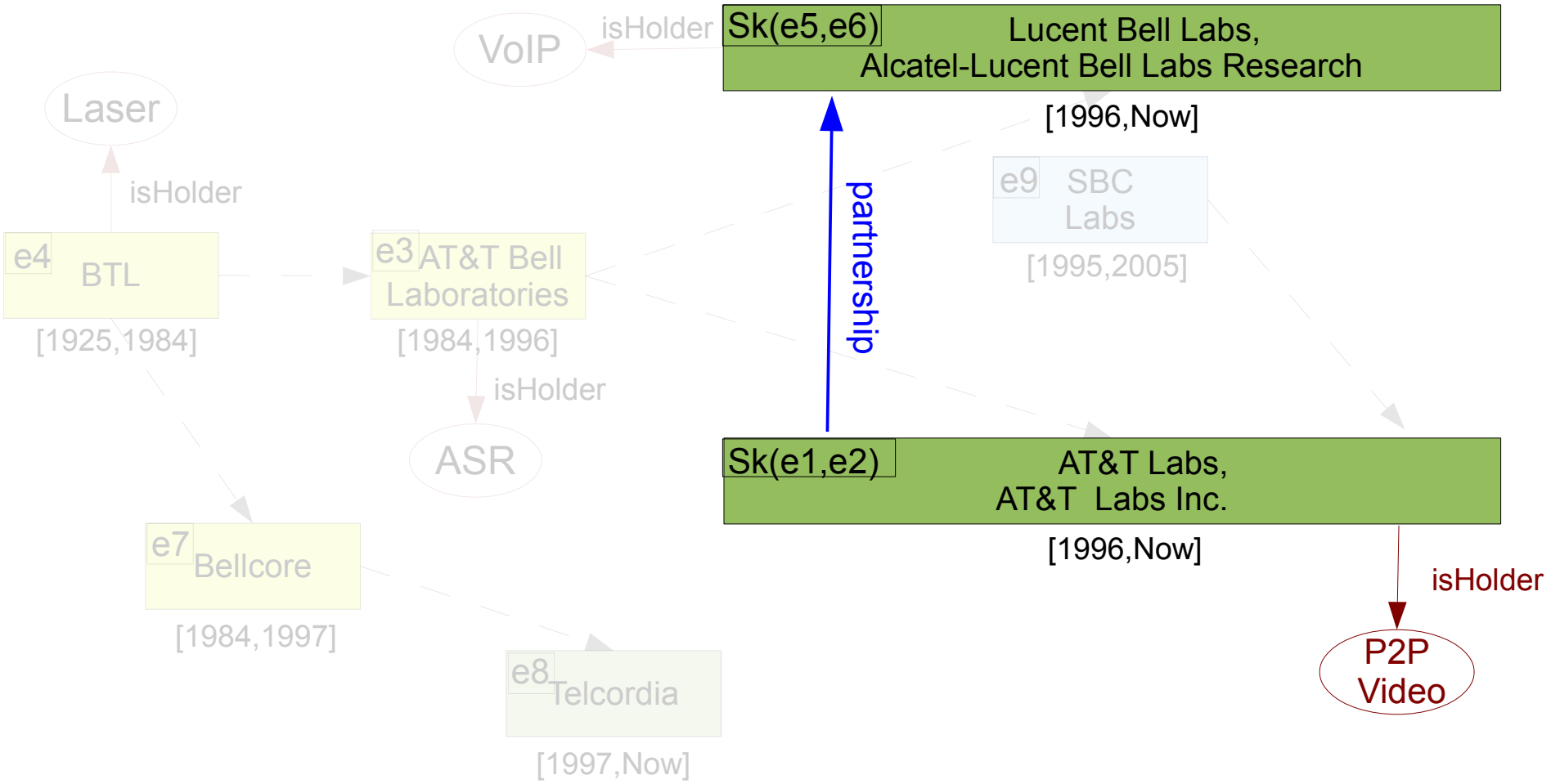
Query: find all the partners of AT&T Labs Inc

Motivating Query with Associations



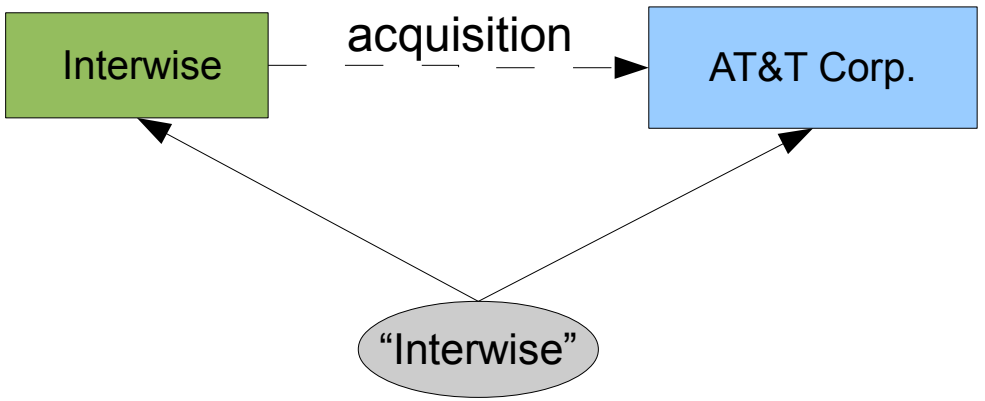
Query: find all the partners of AT&T Labs Inc

Motivating Query with Associations



Query: find all the partners of AT&T Labs Inc

Trademark Dataset*



- 16K unique companies
- 200K attributes
- evolution graph with 573 components of sizes between 5 and 373

*http://www.uspto.gov/

References

1. P. Buneman, S. Khanna, K. Tajima, and W. Tan. Archiving scientific data. In SIGMOD, pages 1–12, 2002.
2. S. Chawathe, S. Abiteboul, and J. Widom. Representing and Querying Changes in Semistructured Data. In ICDE, pages 4–19, 1998.
3. Bolin Ding, J Xu Yu, Shan Wang, Lu Qin, Xiao Zhang, and Xuemin Lin. Finding Top-k Min-Cost Connected Trees in Databases. ICDE, pages 836–845, 2007.
4. Elisabeth Gassner. The Steiner Forest Problem revisited. Journal of Discrete Algorithms, 8(2):154–163, June 2010.
5. P. McBrien and A. Poulovassilis. Schema Evolution in Heterogeneous Database Architectures, A Schema Transformation Approach. In CAiSE, pages 484–499, 2002.
6. Y. Velegrakis, R. J. Miller, and J. Mylopoulos. Representing and Querying Data Transformations. In ICDE, pages 81–92, 2005.