Emergent Semantics: Rethinking Interoperability for Large Scale Decentralized Information Systems

Philippe Cudré-Mauroux

Computer Science & Artificial Intelligence Lab
Massachusetts Institute of Technology

Self-Organizing Knowledge Systems Symposium, VU Amsterdam, April 29, 2010

Outline

- 1. Introduction
  - 1.1. Semantic Interoperability in the Internet Era
  - 1.2. Peer Data Management Systems (PDMSs)
  - 1.3. Syntactic Semantics
- 2. Methods
  - 2.1. Semantic Gossiping
  - 2.2. Graph-Theoretic Semantic Interoperability
- 3. Systems
  - 3.1. GridVine: A P2P Semantic Overlay Network
  - 3.2. idMesh: Disambiguation of Linked Data
- 4. Conclusions

➡ breadth rather than depth

CSAIL

Part 1

Introduction

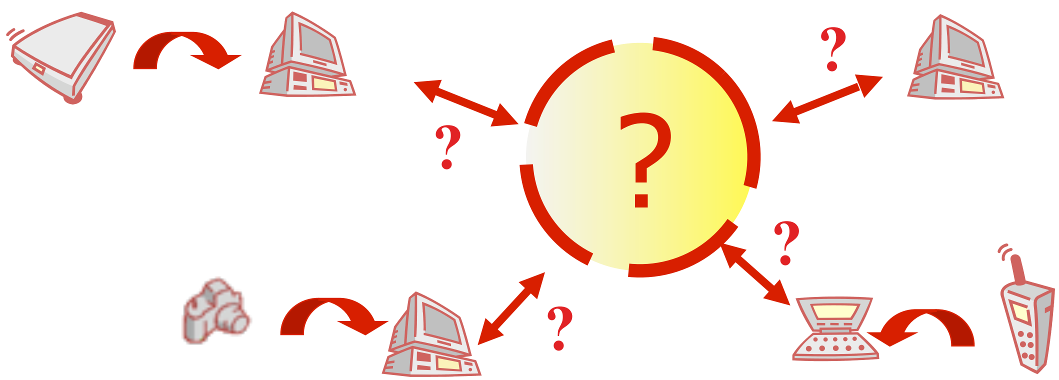
Interoperability in the Internet Era

Searching semantically richer objects in large scale heterogeneous networks

date?

<xap:CreateDate>**2001-12-19T18:49:03Z**</xap:CreateDate>
<xap:ModifyDate>**2001-12-19T20:09:28Z**</xap:ModifyDate>

<es:DofCreation> **05/08/2004** </es:DofCreation>



<myRDF:Date> **Jan 1, 2005** </myRDF:Date>

➠ Lack of semantic interoperability

On Information Heterogeneity

- Syntactic discrepancies

ImageGUID	cDate
A0657B25	05.08.04

VS

```
<es:cDate> 05/08/2004 </es:cDate>
```

- Semantic heterogeneity
  - All the aforementioned standards are extensible

```
<rdf:Property rdf:ID="width">
  <rdfs:label>Width</rdfs:label>
  <rdfs:subPropertyOf rdf:resource="#length"/>
</rdf:Property>
```

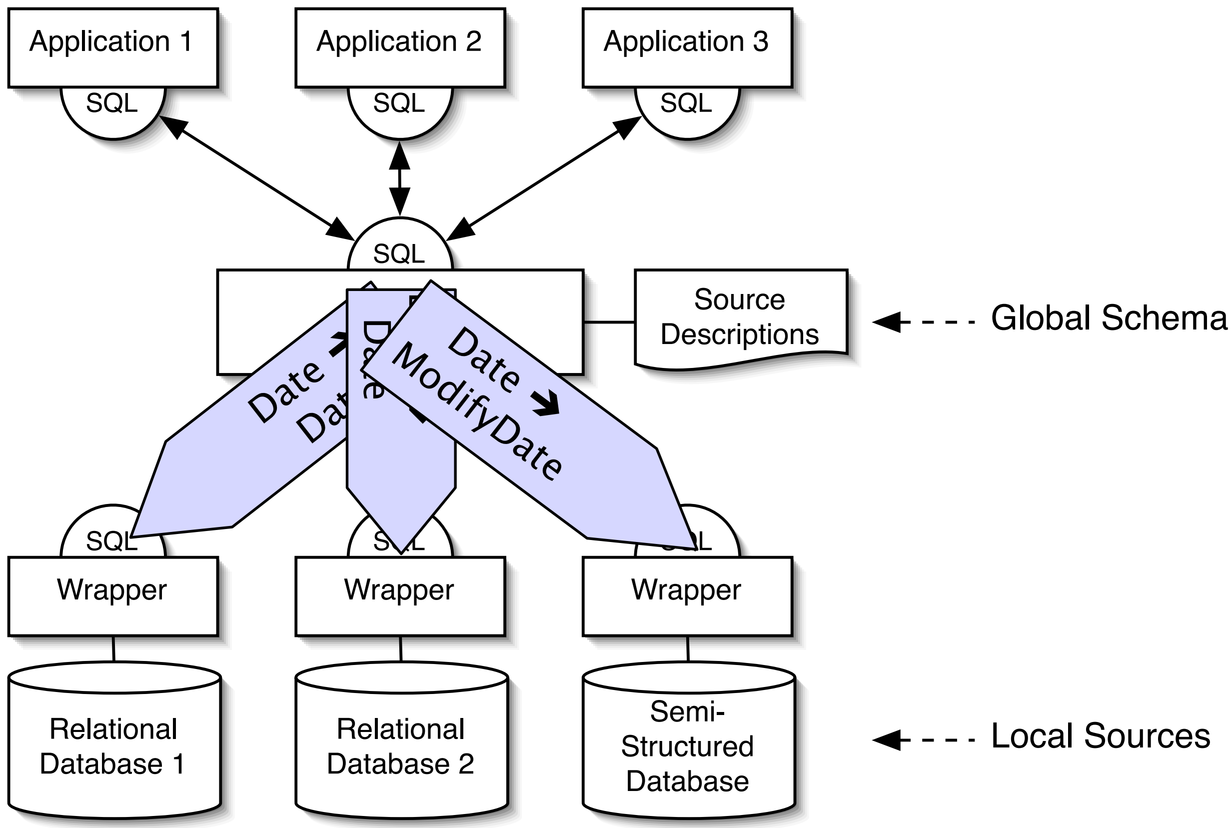
VS

```
<rdf:Property rdf:ID="Length-Y">
  <rdfs:label>Length-Y</rdfs:label>
  <rdfs:subPropertyOf rdf:resource="#length"/>
</rdf:Property>
```

➠ Shared representation is *not* enough

Integrating Data in Distributed Databases

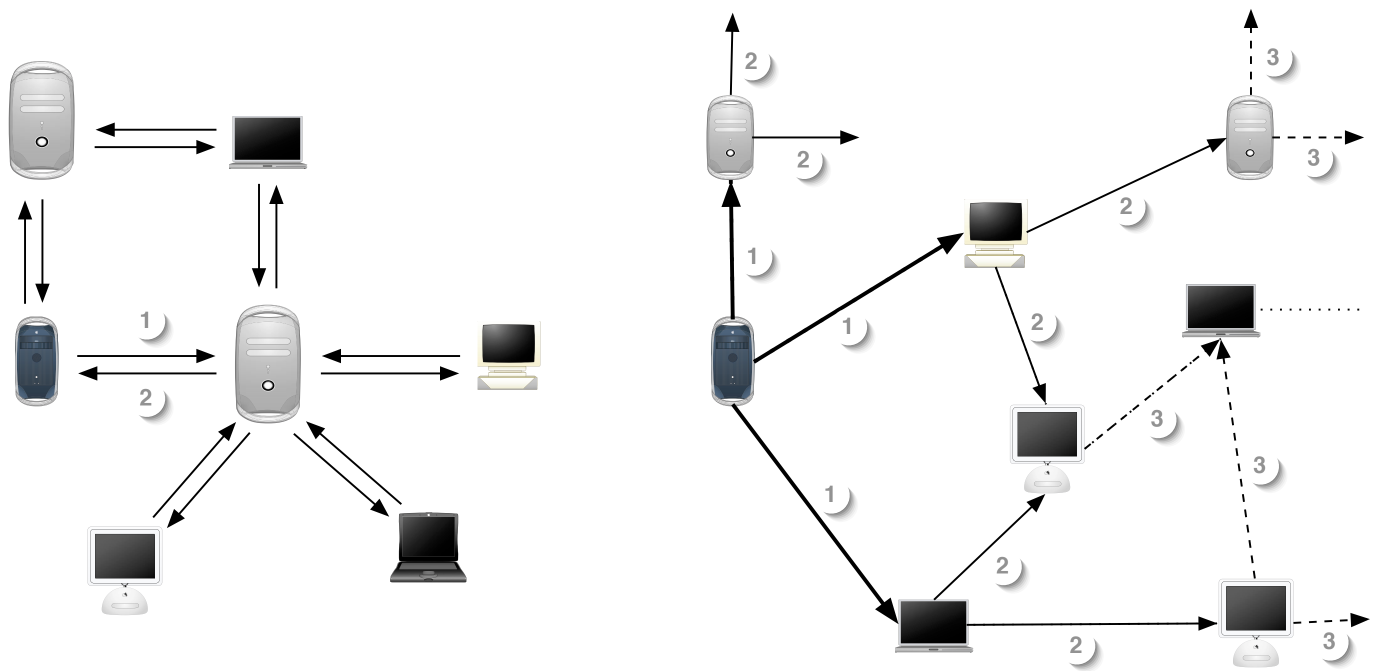
- The Wrapper-Mediator architecture



Integrating Data in the new Web Ecology

	Distributed Databases	Large Scale Information Systems (e.g., WWW))
Scale	Number of sources < 100	Number of sources > 1000
Uncertainty	Consistent Data - Coordination - Manually curated data Schemas created by administrators	Uncertain Data - Autonomy - Semi-automatic creation of data Schemas created by end users
Dynamicity	Relatively stable set of sources - stable mediator Sources known a priori	Network churn - node failures Unknown sources
Expressivity	Relational Data Structured Schemas - Integrity constraints Structured Queries	Semi-structured data Schematas - No integrity constraints Simple S-P Queries

Opportunity: P2P Architectures



i) Client-Server

ii) Peer-to-Peer

- Scalability (decentralized architectures)
- Autonomy (self-organization)
- Robustness (adaptivity, no single point of failure)

Decentralized Interoperability

```
Q1=
<GUID>$p/GUID</GUID>
FOR $p IN /Photoshop_Image
WHERE $p/Creator LIKE "%Robi%"
```

```
Q2=
<GUID>$p/GUID</GUID>
FOR $p IN T12
WHERE $p/Creator LIKE "%Robi%"
```

Photoshop (own schema)

```
<Photoshop_Image>
 <GUID>178A8CD8865</GUID>
  <Creator>Robinson</Creator>
  <Subject>
   <Bag>
    <Item>
     Tunbridge Wells
    </Item>
    <Item>Royal Council</I
   </Bag>
  </Subject>
  …
</Photoshop_Image>
```

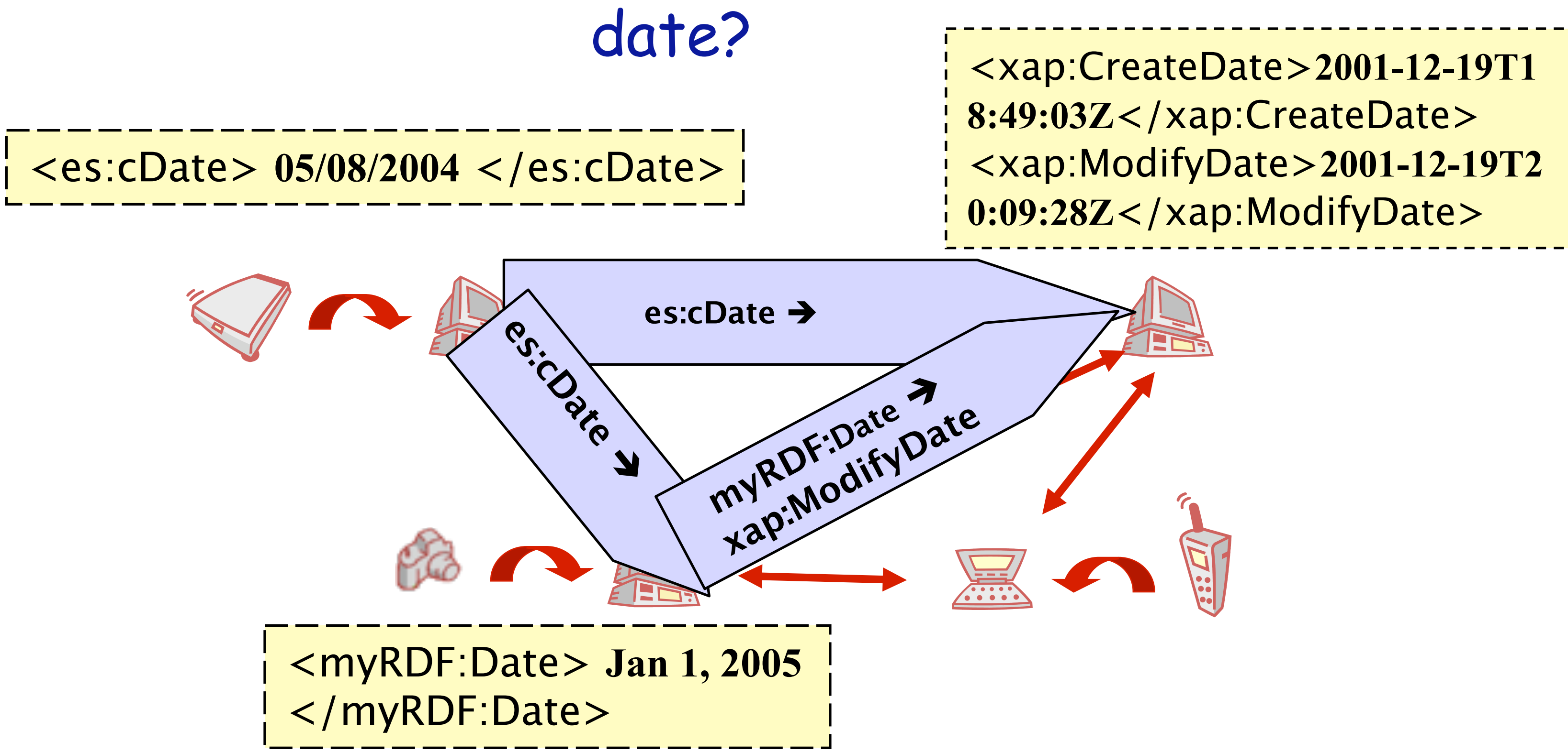
```
T12 =
<Photoshop_Image>
 <GUID>$fs/GUID</GUID>
 <Creator>
  $fs/Author/DisplayName
 </Creator>
</Photoshop_Image>
FOR $fs IN /WinFSImage
```

WinFS (known schema)

```
<WinFSImage>
 <GUID>178A8CD8866</GUID>
 <Author>
  <DisplayName>
   Henry Peach Robinson
  <DisplayName>
  <Role>Photographer</Role>
 <Author>
 <Keyword>
  Tunbridge
 </Keyword>
 <Keyword>Council</Keyword>
 …
</WinFSImage>
```

➠ Extending integration techniques to decentralized settings

Peer Data Management Systems



- Pairwise mappings
  - Peer Data Management Systems (PDMS)
- Local mappings overcome global heterogeneity
  - Iterative query reformulation

Emergent Semantics (1)

- Contrary to the wrapper-mediator architecture, no definite, global semantics defined *a priori*
  - What is the resulting semantics of the overall system?
- Long-standing debate:”What is semantics?”
  - Standard response: “Mapping of a syntactic structure into a semantic domain”

Semantic Grounding

- The meaning of symbols can be explained by its semantic correspondences to other symbols alone [“Understanding understanding” Rapaport 93]
  - Type 1 semantics: understanding in terms of something else
    - Problem: how to ground semantics?
  - Type 2 semantics: understanding something in terms of itself
    - “syntactic semantics”: grounding through recursive understanding

Emergent Semantics (2)

- Emergent Semantics:
  - Semantics as *a posteriori* agreements on conceptualizations
  - Semantics of symbols as recursive correspondences to other symbols
    - Analyzing transitive closures of mappings
  - Self-organizing, bottom-up approach
    - Global semantics (stable states) emerging from multiple local interactions
  - Syntactic semantics
    - Studying semantics from a syntactic perspective

Problems (1/2): Precision / Recall

- Semantic Query routing
  - To whom shall I forward a query posed against my local schema?
- Some (most) mappings will be (partially) faulty
  - Low expressive power of mappings
    - samePropertyAs / sameClassAs / subclassOf
    - ... or event worse (Microformats)
  - Automatic schema alignment techniques
  - Different views on conceptualizations
- Local query resolution
  - Low recall
- Flooding
  - Low precision

➠ Standard deductive integration is not sufficient

➠ Uncertainty on mappings and conceptualizations

CSAIL

Problems (2/2): Global Interoperability

- What is the global impact of local actions?
  - Issuing a query locally
    - Diffusion on the global scale
    - cf. precision/recall
  - Creating local mappings
    - Mapping scarcity
      - Semantic partitions
    - Mapping abundance
      - Mapping Quality
      - Computational overhead
      - Network overhead

➠ Model encompassing interoperability at global scale.

CSAIL

Part 2

Methods

Semantic Gossiping

- Local, selective and query-specific forwarding paradigm
  - Mapping completeness
    - Capability of reformulating arbitrary queries
      - Lost predicates
    - Syntactic analysis
  - Mapping soundness
    - Capability of reformulating queries in semantically correct ways
      - Agreements on conceptualizations
    - Semantic analyses

➠ Self-organization of query diffusion

  ➠ Precision/Recall tradeoff

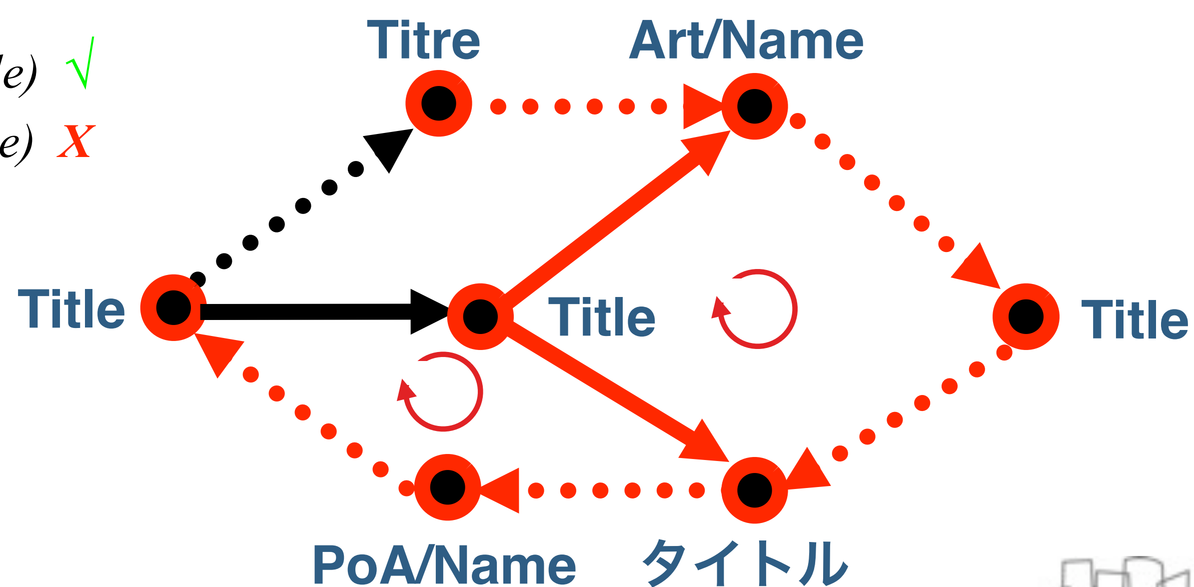
Syntactic Analysis

- Measure the syntactic losses in successive query reformulations (mapping completeness)
  - attributes lost in the projections
    - $\pi_{Title,\ Format,\ Length} \longrightarrow \pi_{Format,\ Length} \longrightarrow \pi_{Length} \longrightarrow \ldots$
  - predicates lost in the selections
    - $\sigma_{Title=\text{"The Vitruvian Man"},\ Year < 1600} \longrightarrow \sigma_{Year < 1600} \longrightarrow \ldots$
- Losses can have various impacts
  - Selectivity of the selection predicates
  - Query-dependent weights of the attributes
- Losses aggregated in two similarity values
  - $0 \leq SIM_{\pi|\sigma}(q, (\mu_n \circ \ldots \circ \mu_1)(q)) \leq 1$

Semantic Analyses (1/2)

- Measure the semantic losses in successive query reformulations (mapping soundness)
- Cycle analysis: agreement on conceptualizations derived through transitive closure of mapping operations

- $(\mu_n \circ \ldots \circ \mu_1)\,(Title) \equiv (Title)$ √
- $(\mu_n \circ \ldots \circ \mu_1)\,(Title) \not\equiv (Title)$ X
- $(\mu_n \circ \ldots \circ \mu_1)\,(Title) = \emptyset$

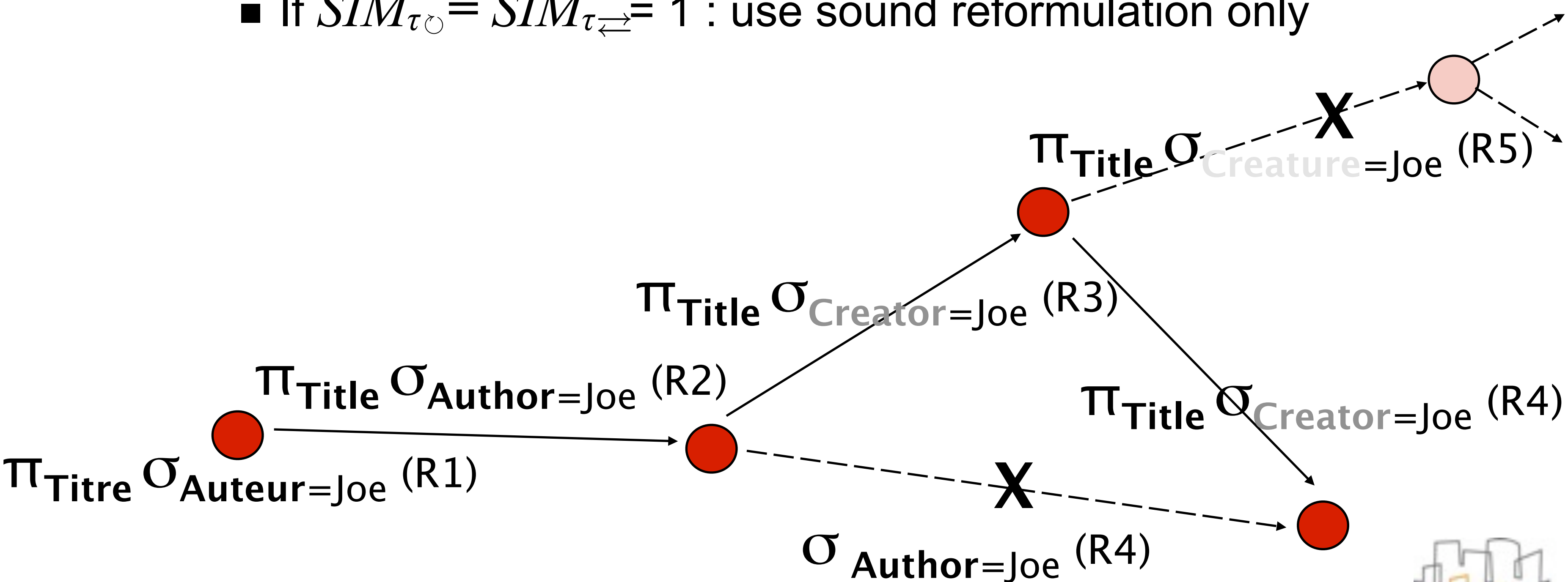


Semantic Analyses (2/2)

- Derive likelihood on mapping soundness from multiple feedback cycles
  - $P(f^{+}_{\circlearrowright}|m=1) = (1-\epsilon_{cyc})^{\|f_{\circlearrowright}\|-1} + (1-(1-\epsilon_{cyc})^{\|f_{\circlearrowright}\|-1})\delta_{cyc})$
  - $P(m=1|\boldsymbol{f}_{\circlearrowright}) = K\ P(m=1)$
    $$\prod_{f^{+}_{\circlearrowright}\in\boldsymbol{f}^{+}_{\circlearrowright}} P(f^{+}_{\circlearrowright})^{-1}P(f^{+}_{\circlearrowright}|m=1) \prod_{f^{-}_{\circlearrowright}\in\boldsymbol{f}^{-}_{\circlearrowright}} P(f^{-}_{\circlearrowright})^{-1}P(f^{-}_{\circlearrowright}|m=1)$$
- Similar analysis for returned results
  - Agreements on document classification
- Iteratively update a semantic similarity value along with the reformulations
  - $0 \leq SIM_{\circlearrowright|\rightleftarrows}(q,(\mu_n \circ \ldots \circ \mu_1)(q)) \leq 1$

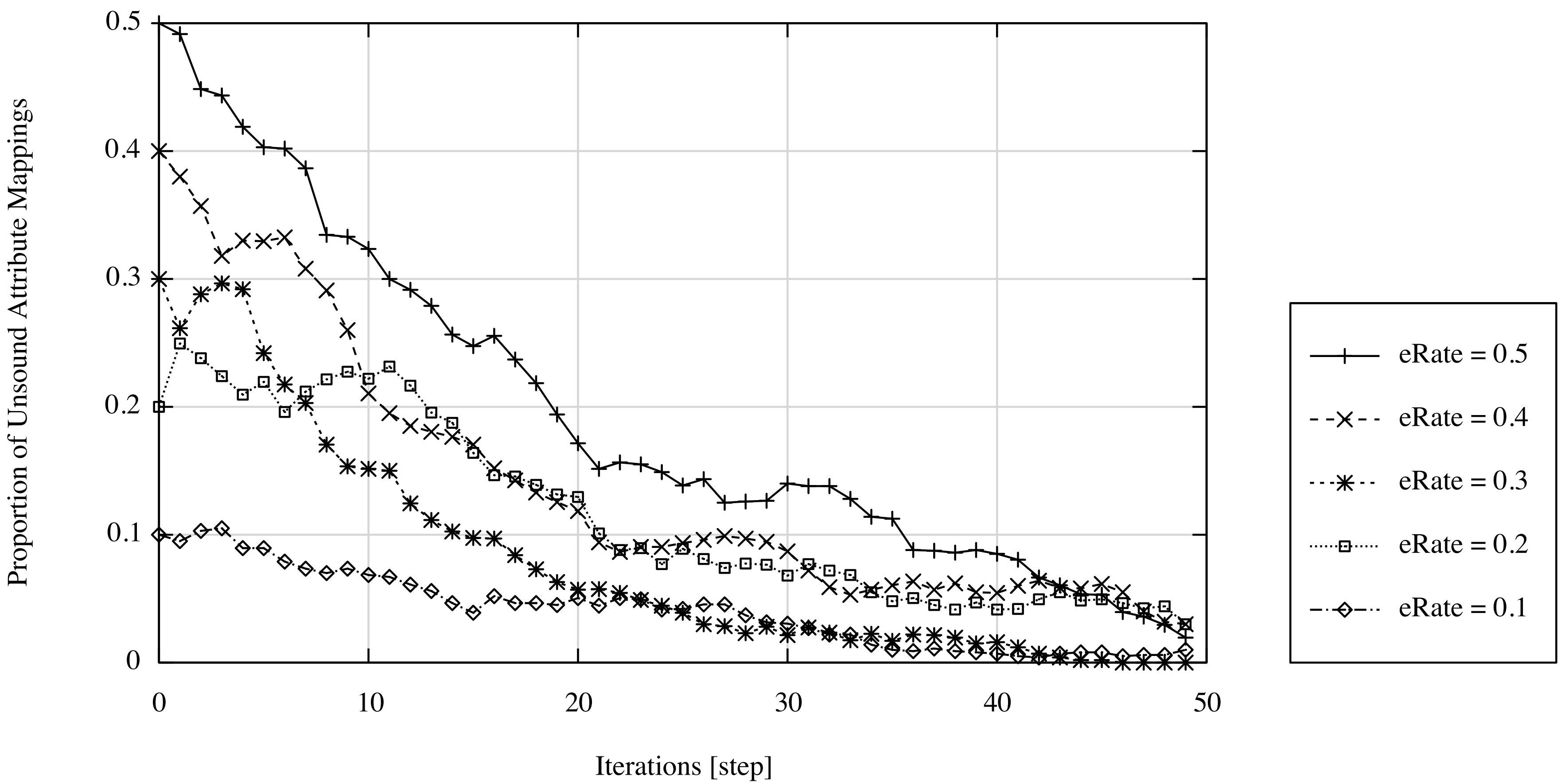
Semantic Gossiping: Per-Hop Forwarding

- Query specific thresholds on similarities $\boldsymbol{SIM_\tau}$
  - User / System generated
  - Reformulate query through mapping if $\boldsymbol{SIM_{q'} \geq SIM_\tau}$
    - If $SIM_{\tau\pi} = SIM_{\tau\sigma} = 1$: use complete reformulations only
    - If $SIM_{\tau\circlearrowright} = SIM_{\tau\rightleftarrows} = 1$: use sound reformulation only



CSAIL

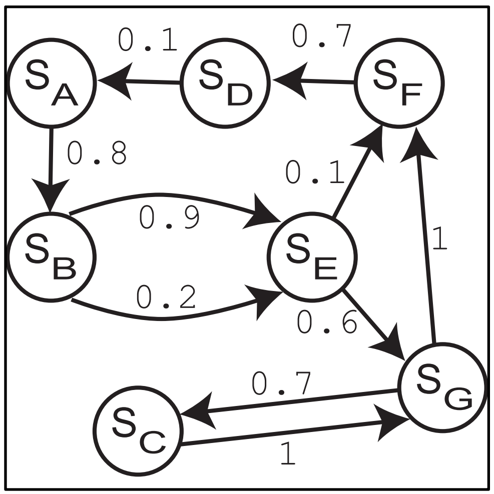
Self-Healing Semantic Networks



Combined Analysis (random graph, 4 att., 25 schemas, TTL=6 (cycle)/3(results), 10 consecutive runs)

Graph-Theoretic Semantic Interoperability

- What about interoperability at a global scale?
- Modeling semantic interoperability:



Schema-to-Schema Graph

- Logical model
- Directed
- Weighted
- Redundant

- The semantic connectivity graph
  - Idea: as for physical network analyses, define a connectivity layer
  - Unweighted, non-redundant version of the Schema-to-Schema graph

Semantic Interoperability in the Large

- Definition

  Peers in a set P_s are semantically interoperable iff S_s is strongly connected, with $S_s \equiv \{s \mid \exists p \in P_s, p \leftrightarrow s\}$

- Observation 1

  A set of peers P_s cannot be semantically interoperable if $|E_s| < |V_s|$

- Observation 2

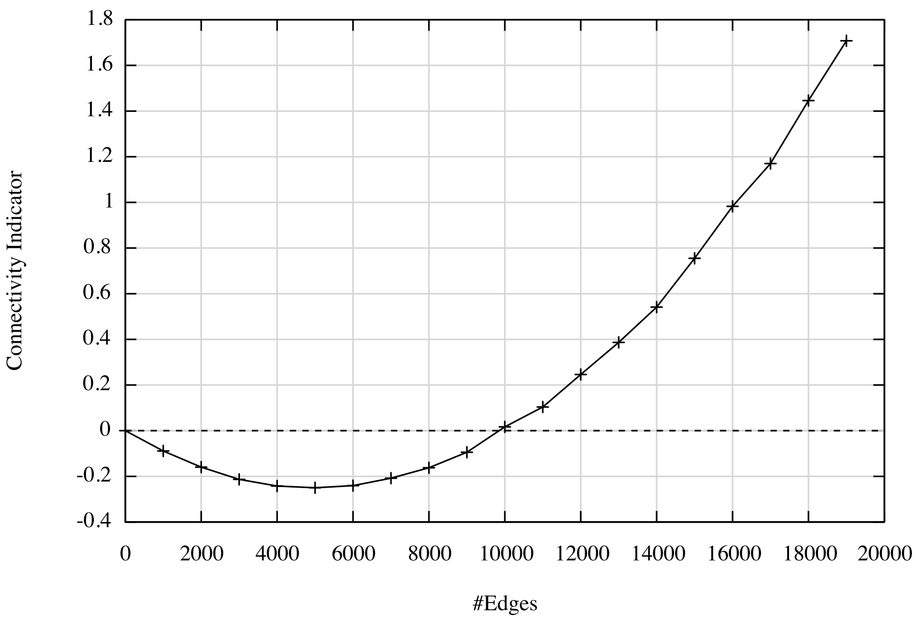
  A set of peers P_s is semantically interoperable if $|E_s| > |V_s| (|V_s|-1) - (|V_s|-1)$

- What happens between those two bounds?
  - What is the proportion of interoperable systems?

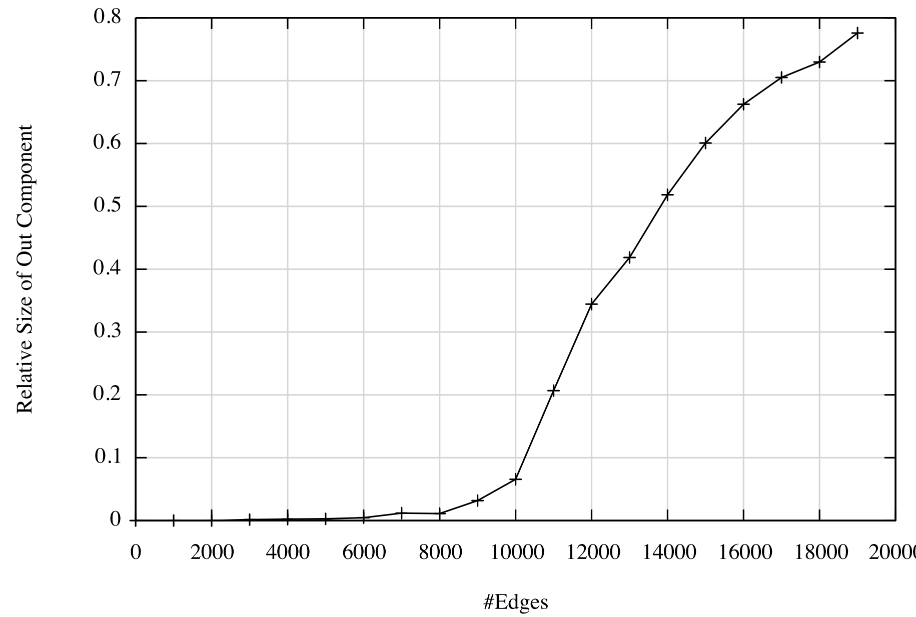
A Necessary Condition for Semantic Interoperability in the Large

- Analyzing semantic interoperability in large-scale, decentralized networks
  - Percolation theory for directed graphs
  - Based on a recent graph-theoretic framework
  - Graphs with specific degree distributions p_{jk}, clustering coefficients cc and bidirectionality coefficient bc
- Based on generatingfunctionality $\mathcal{G}(x,y) = \sum_{j,k} p_{jk} x^j y^k$
- Connectivity indicator: $ci = \sum_{j,k} (jk - j(bc+cc) - k)\, p_{jk}$
  - Necessary condition for semantic interoperability in the large: $ci \geq 0$
- Also: approximations of the size of semantically interoperable clusters

Example: Directed Graph



a)



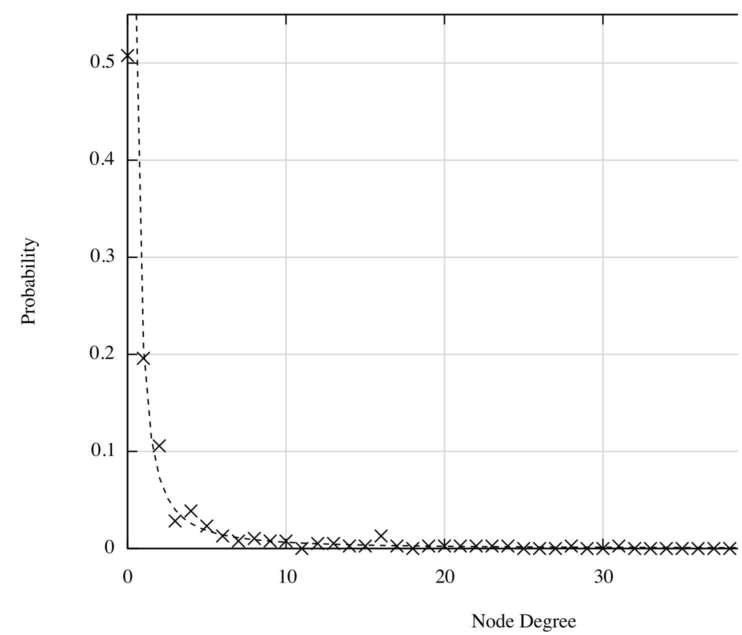
b)

Connectivity Indicator (a) and maximal connected cluster size (b)
Random network of 10000 vertices and a varying number of edges.

Analysis of a bioinformatic system

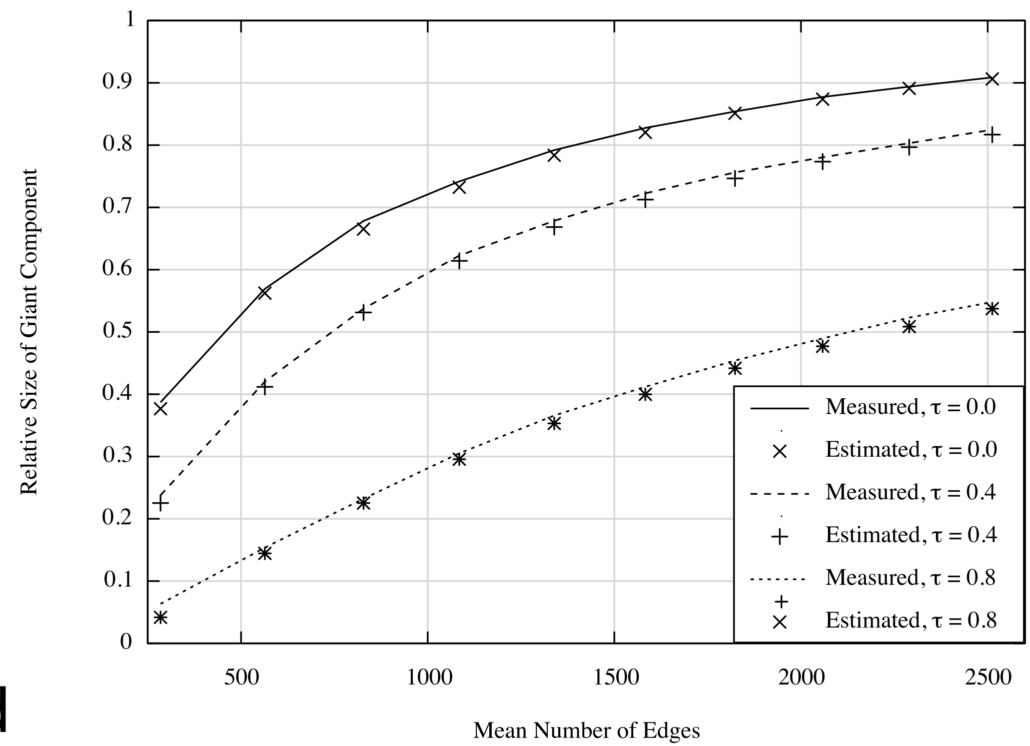
- Analysis of the Sequence Retrieval System (SRS)
  - Commercial information indexing and retrieval system for bioinformatic libraries
  - Schemas described in a custom language (Icarus)
  - Mappings (foreign keys) from one database to others
- Crawling the EBI repository
  - 388 databanks
  - 518 (undirected) links
  - Power-law distribution of node degrees

    $y(x) = \alpha x^{-\gamma}$ with $\alpha = 0.21$ and $\gamma = 1.51$
  - Clustering coefficient = 0.32
  - Diameter = 9
- Connectivity indicator ci = 25.4
  - Super-critical state
- Size of the giant component
  - **0.47** (derived) VS **0.48** (observed)

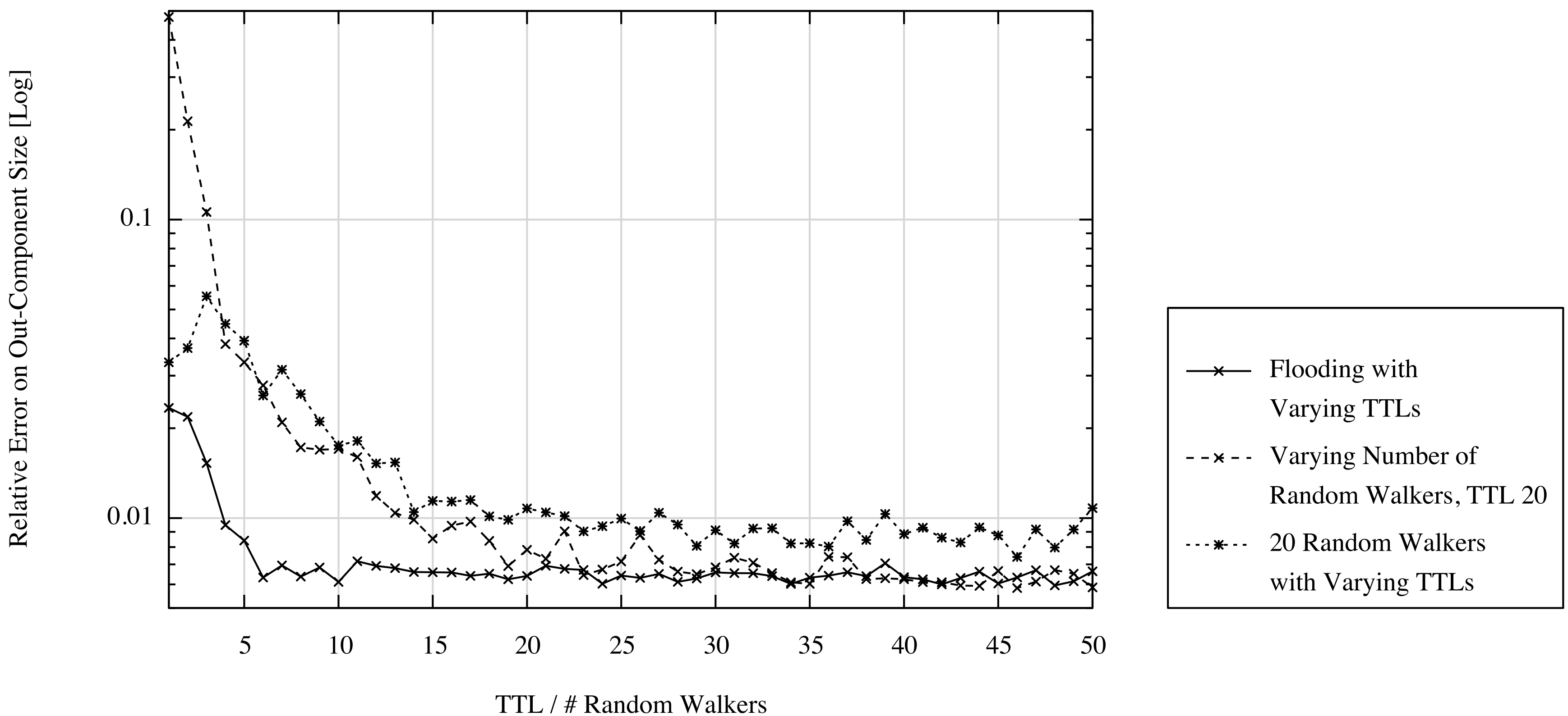


Query Dissemination in Weighted Networks

- Per-hop forwarding behaviors
- Only forward if $w_i >= \tau$
  - $\tau = 0$: flooding
  - $\tau = 1$: exact answers
- Degree distribution taken from the SRS system
- Uniformly distributed weights between 0 and 1



Local View on Global Properties



(Random graph, 1000 vertices, 4000 edges)

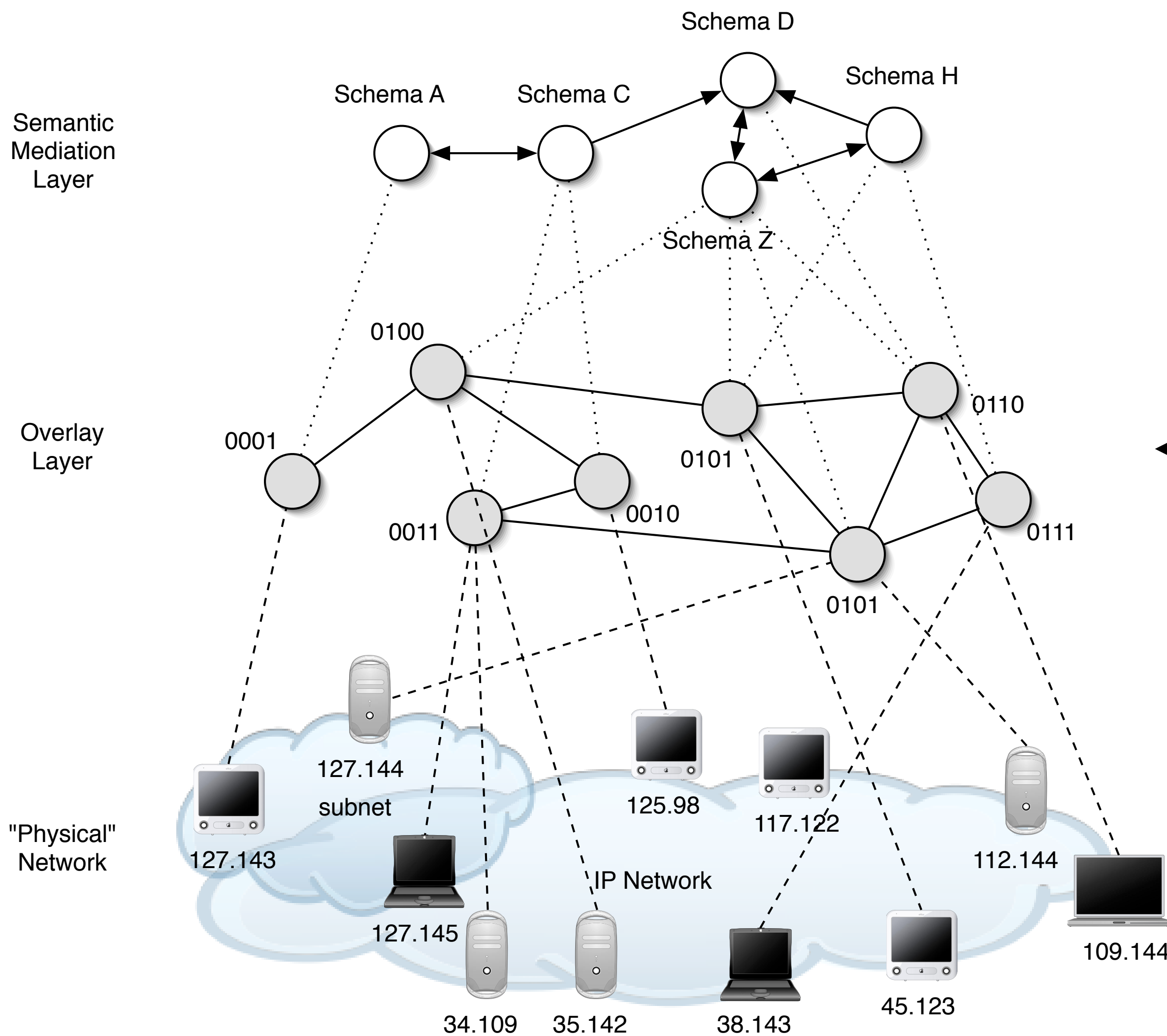
➠ Local View on Global Semantic Properties

CSAIL

Part 3

Systems

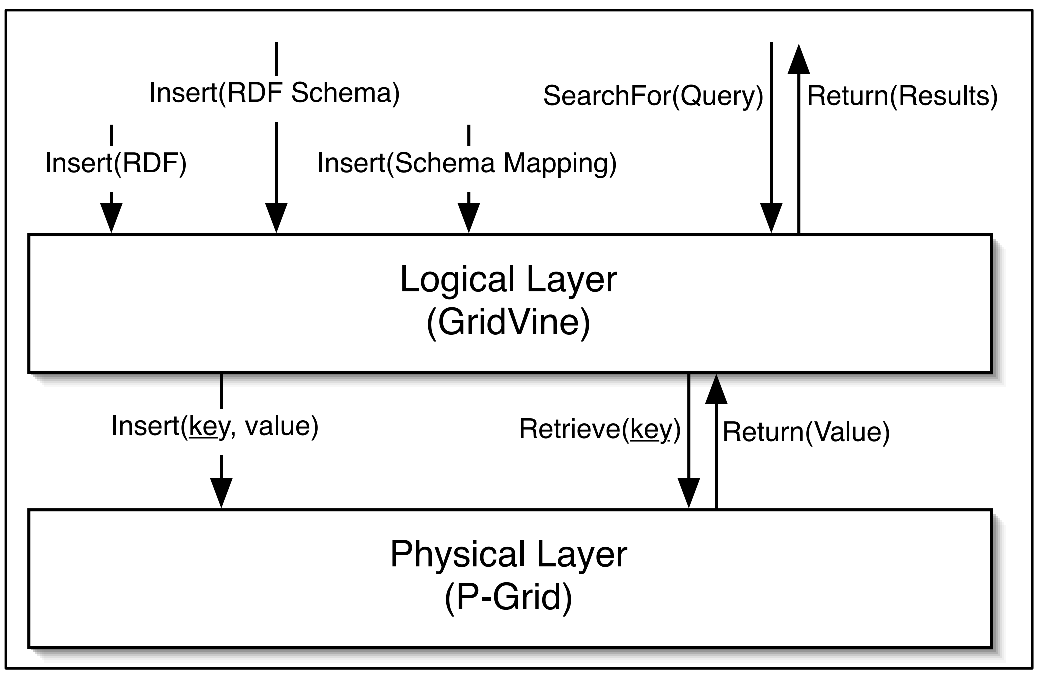
GridVine: a P2P Semantic Overlay Network



← P-Grid

GridVine: Data Independence

- Building large-scale semantic systems
  - Self-organizing semantic overlay network
- Principle of data independence
  - Scalable physical layer
  - Semantic logical layer



Indexing semi-structure data in GridVine

- Triple t = <lsir:GridVine> <dc:creator> <lsir:pcm>

Put(Hash(**lsir:GridVine**), t)

Put(Hash(**dc:creator**), t)

Put(Hash(**lsir:pcm**), t)

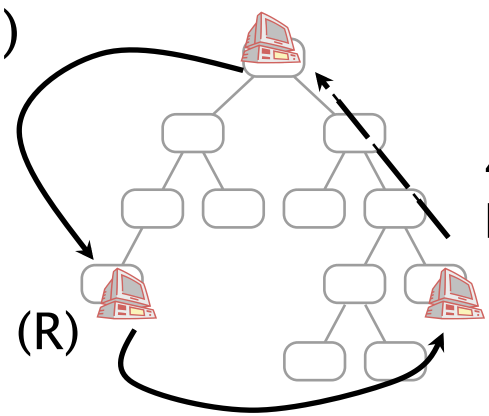
- Insertion of schemas and mappings

➠ Decentralized conjunctive query resolution based on iterative look-ups

Query Resolution

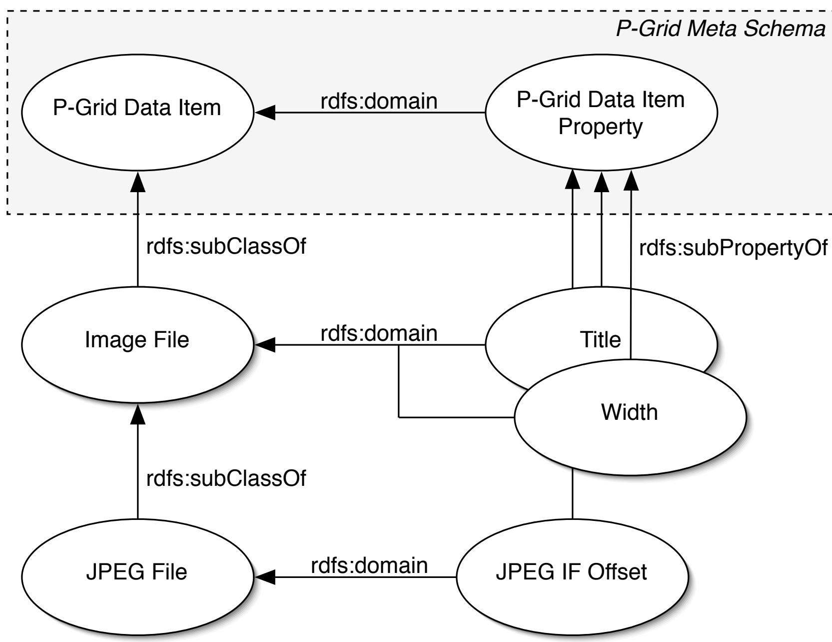
- Triple pattern queries {(?s, ?p, ?o)}
  - path queries, conjunctive queries
  - Iterative, distributed table lookup

(?x, <rdf:type>, <foaf:Person>)
(?x, <foaf:name>, "John")



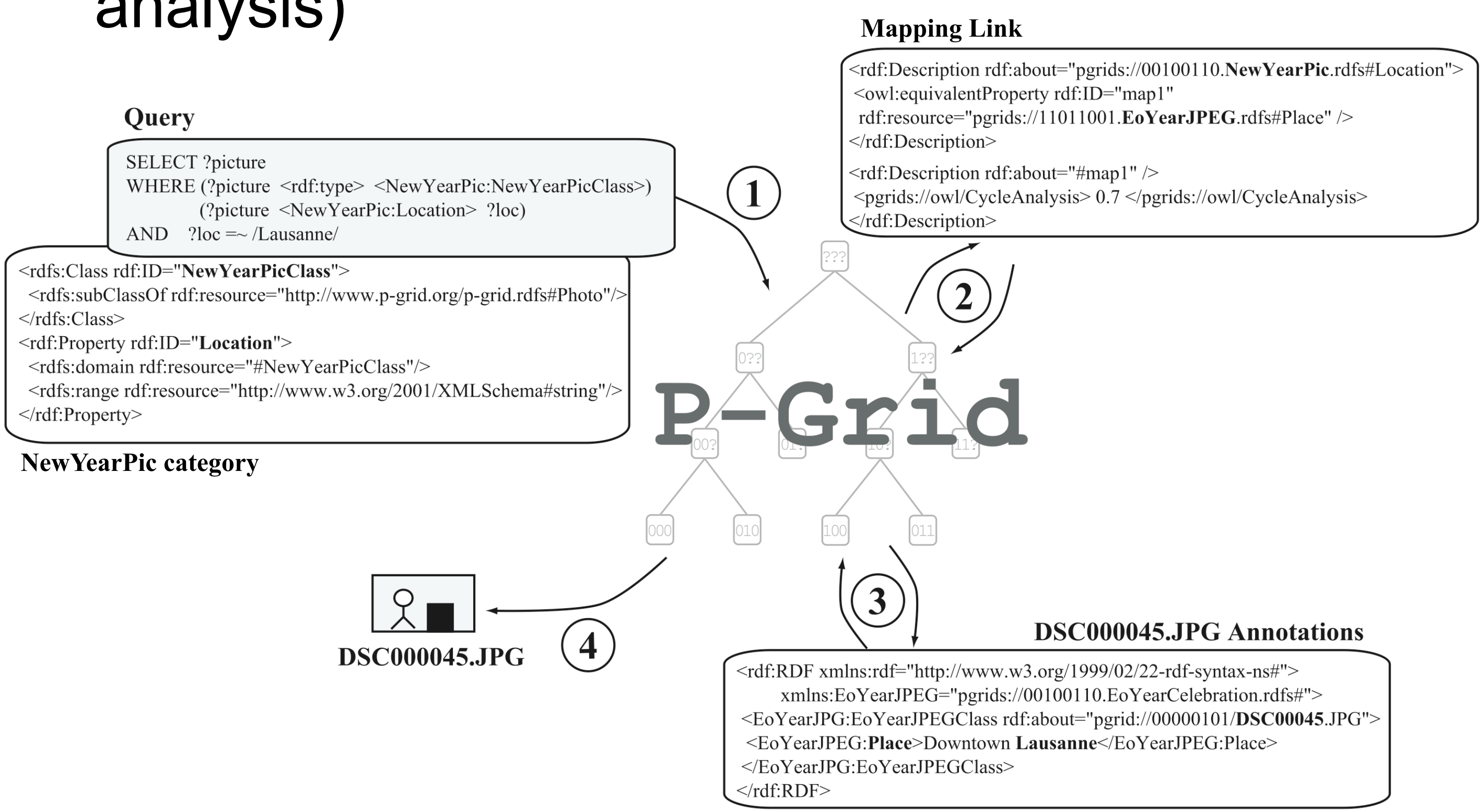
Semantic Integration in GridVine

- Vertical integration: hierarchy of classes
  - Fostering semantic interoperability through reuse of conceptualizations
  - Popular base classes bootstrapping interoperability through monotonic inheritance of properties
  - RDFS entailment can be materialized



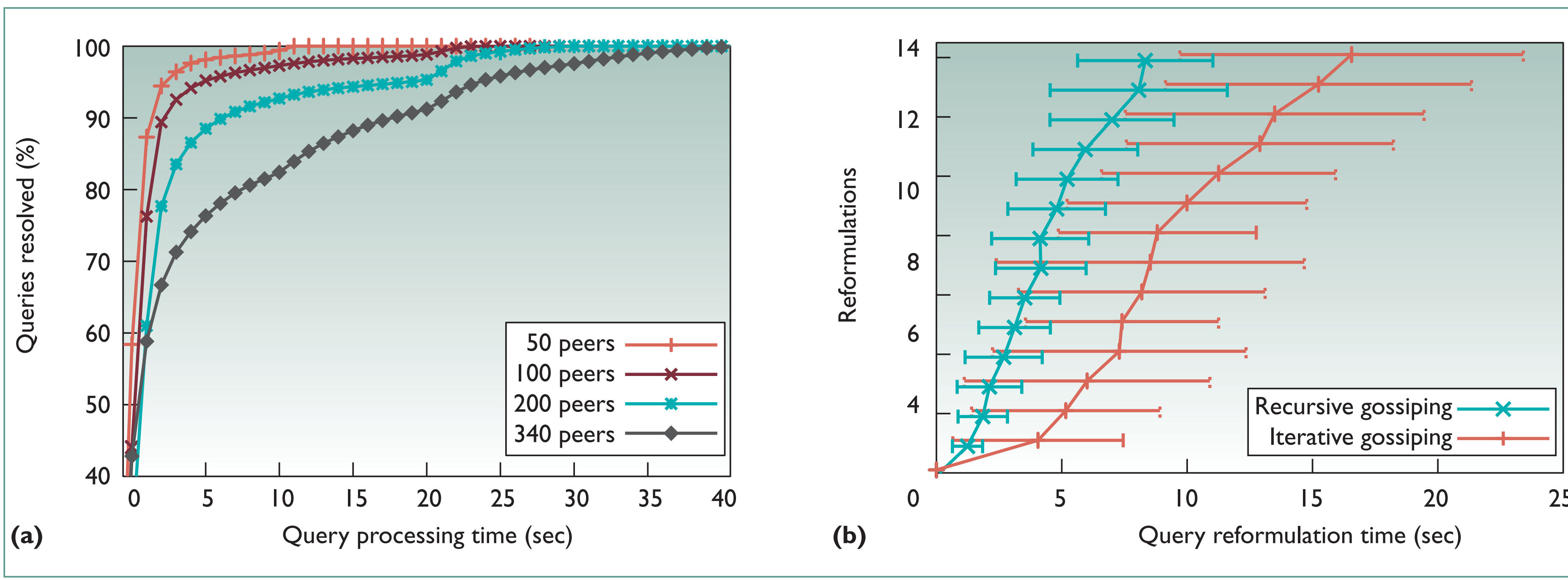
Semantic integration in GridVine

- Horizontal integration: mappings
  - Message passing + feedback analyses to get probabilistic guarantees on mapping soundness
  - Generation of new mappings if necessary (graph analysis)



Semantic Gossiping in GridVine

- Decoupling of the indexing and mediation layers
  - No more constraints on gossiping
- Different query forwarding paradigms
  - Iterative forwarding
  - Recursive forwarding



idMesh: Disambiguation of Linked Data

- Increasingly, the world is modeled as a collection of (interlinked) identifiers
  - Linked Data
  - Semantic Web
  - RESTful services
  - ...

Naming & Decentralization

- The great thing about *unique identifiers* is that there are *so many* to choose from
  - Decentralized naming game
  - Soaring dimensions in Web 2.0 / 3.0 contexts
    - Social websites
    - Exported (linked) data
    - Automated mash-ups

http://semanticweb.org/id/Philippe_Cudre-Mauroux

http://data.semanticweb.org/person/philippe-cudre-mauroux

http://people.csail.mit.edu/pcm/i http://lsi ple.epfl.ch/pcudre/i

http://semanticweb.org/wiki/Special:E ppe_Cudr%C3%A9-Mauroux

http://tw.rpi.edu/wiki/Special:E r%C3%A9-Mauroux

http://wiki.ontoworld.o xportRDF/Philippe_Cudr%C3%A9-Mauroux

http://korrekt.org/index. ExportRDF/Philippe_Cudr%C3%A9-Mauroux

http://prauw.cs.vu.nl:808 nk/graph?profile=http%3A%2F%2Fwww.cs.vu.nl%2F%7Epmika%2Fsocionet%23Philippe%2BCudre-Mauroux

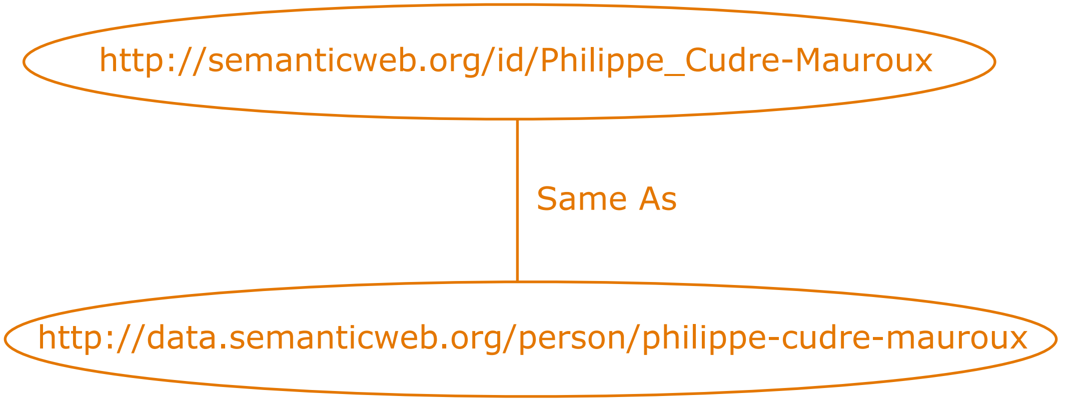
http://www.zoominfo.com/PersonID=402960578 http://www.flickr.com/photos/28735...@N00/

http://www.facebook.com/profile.php?id=1251943... ....

ID Jungle

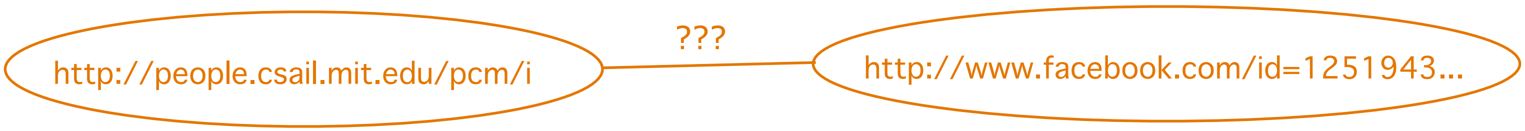
Entity Consolidation (i)

- A few constructs are increasingly used to consolidate Wed identifiers
  - OWL:SameAs, XFN rel:me, pipes, etc.

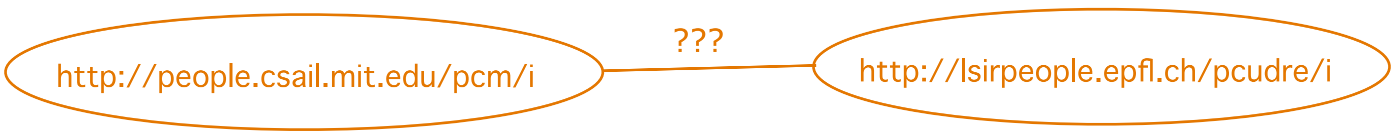


Entity Consolidation (ii)

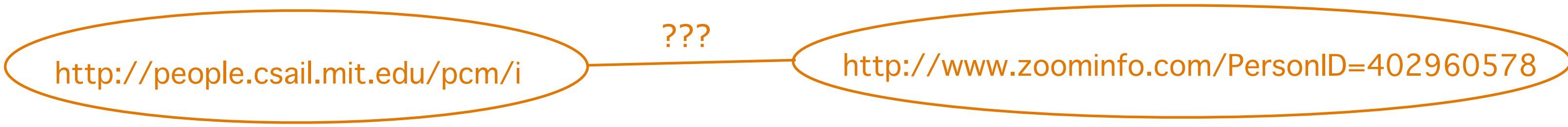
- Online entity consolidation is a *complex* game
  - Simple binary constructs are often insufficient
    - Social contexts (e.g., professional vs personal entities)



    - Granularity (e.g., out-of-date entities)



    - Uncertainty (e.g., automatically-generated entities)



New Twist on an Old Problem

- Well-known problem know as *Entity Disambiguation* or *Resolution*
  - Large body of related work
- *New context*
  - Unprecedented scale
  - Networked game
  - Social dimension

➡ *central* problem impeding all automated, large-scale online data processing endeavors

➡ new approach based on graph analysis only

idMesh Constructs

```
...
<rdfs:Class rdf:ID="Entity"/>
<rdf:Property rdf:ID="idMeshProperty">
      <rdfs:domain rdf:resource="#Entity" />
      <rdfs:range rdf:resource="#Entity" />
</rdf:Property>
<rdf:Property rdf:ID="LinkConfidence">
      <rdfs:domain rdf:Statement />
      <rdfs:range rdf:datatype="&xsd;decimal" />
</rdf:Property>
<rdf:Property rdf:ID="EquivalentTo">
      <rdfs:subPropertyOf rdf:resource="#idMeshProperty" />
</rdf:Property>
<rdf:Property rdf:ID="NotEquivalentTo">
      <rdfs:subPropertyOf rdf:resource="#idMeshProperty" />
</rdf:Property>
<rdf:Property rdf:ID="Predates">
      <rdfs:subPropertyOf rdf:resource="#EquivalentTo" />
</rdf:Property>
<rdf:Property rdf:ID="Postdates">
      <rdfs:subPropertyOf rdf:resource="#EquivalentTo" />
</rdf:Property>
<rdf:Property rdf:ID="Equidates">
      <rdfs:subPropertyOf rdf:resource="#EquivalentTo" />
</rdf:Property>
```

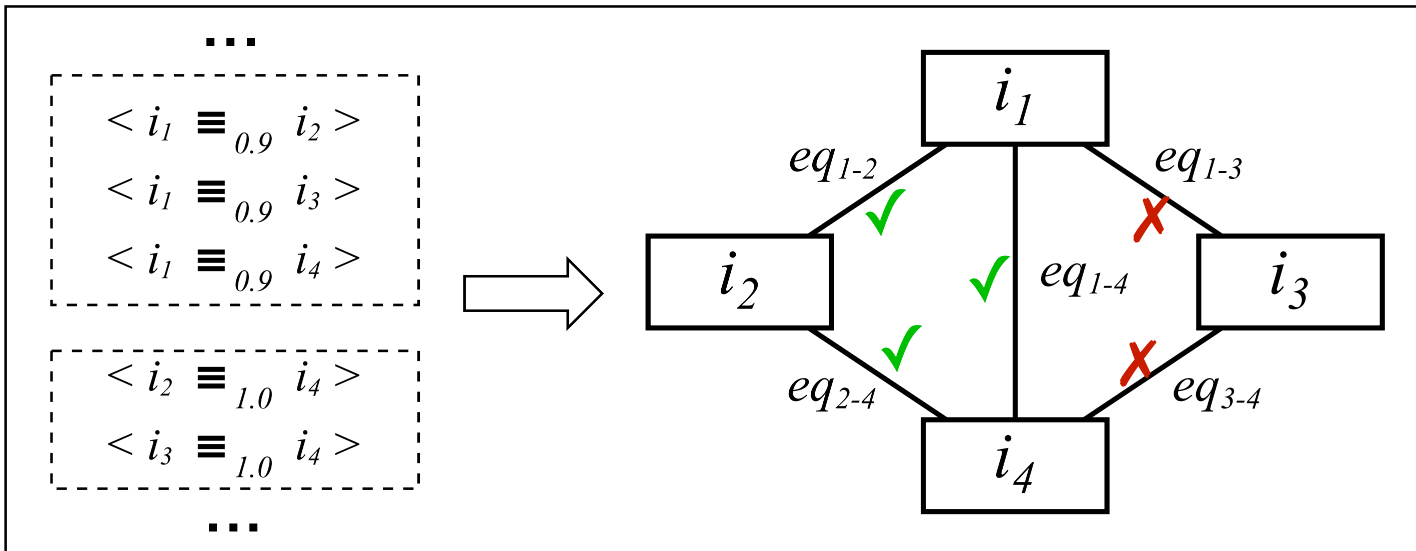
- Two levels of granularity
  - Entity disambiguation
  - Temporal discrimination
- Confidence values
- Can encompass previous constructs

```
<rdf:Description rdf:about="http://www.epfl.ch/">
      <idMesh: NotEquivalentTo rdf:ID="link0001"
        rdf:resource="http://www.ethz.ch"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.epfl.ch/">
      <idMesh:EquivalentTo rdf:ID="link0002"
        rdf:resource="http://en.wikipedia.org/wiki/EPFL"/>
</rdf:Description>
<rdf:Description rdf:about="#link0002">
      <idMesh:LinkConfidence
        rdf:datatype="&xsd;decimal"> 0.9 </idMesh:LinkConfidence>
</rdf:Description>
```

CSAIL

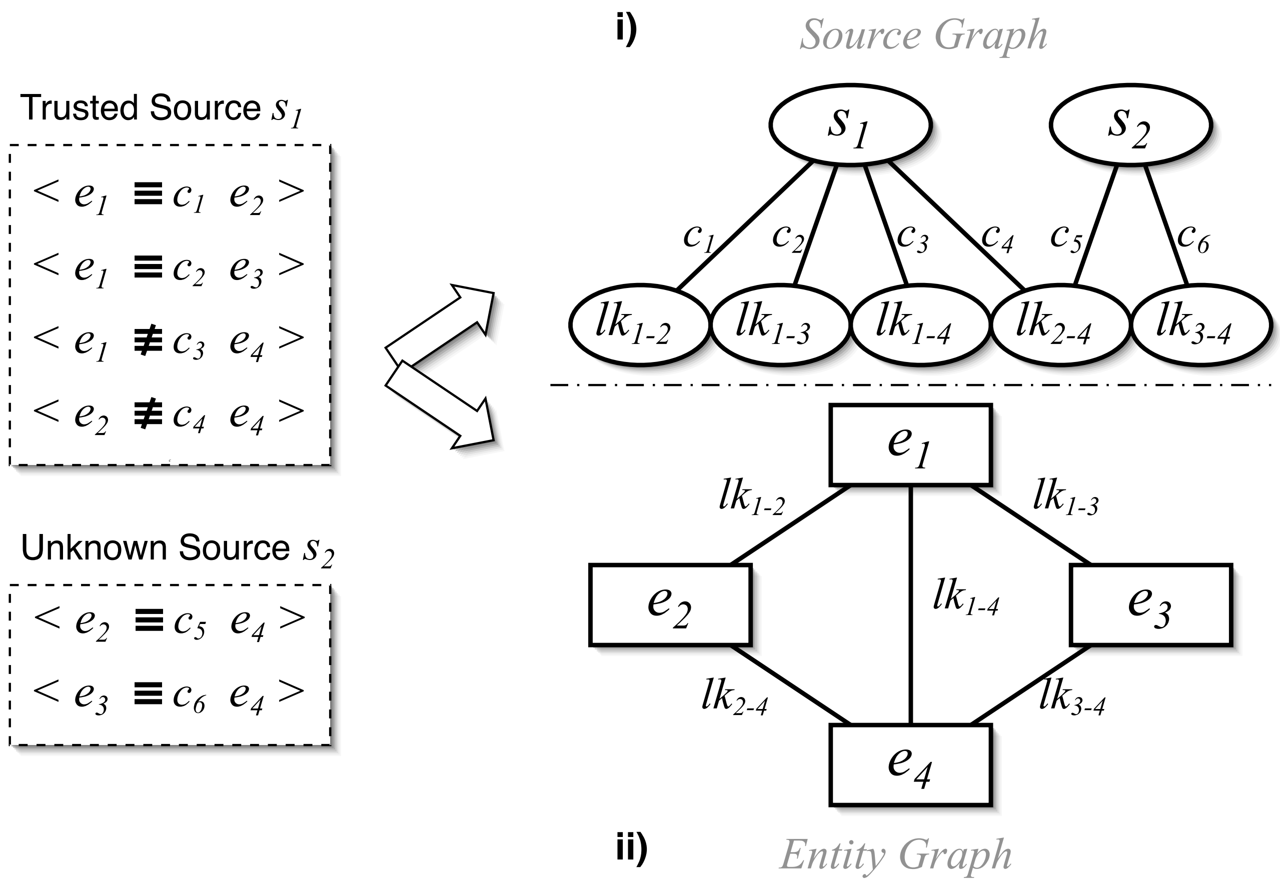
Problem Definition

- Input: series of statements defining a *weighted graph* of *interrelated* identifiers
  - no associated contents, attributes, or properties...



- Output: *clusters* of *equivalent* identifiers
  - probabilistic, *a posteriori* network equivalence
  - equivalence based on probabilistic threshold

Probabilistic Disambiguation

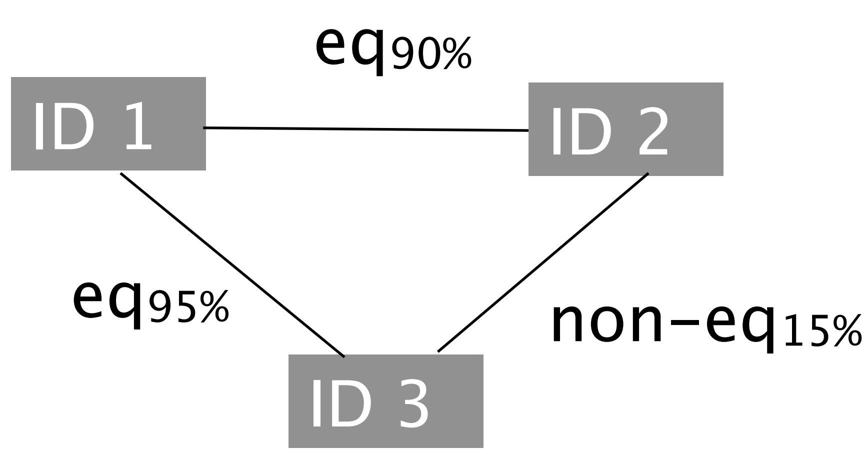


Definition of two graphs

Probabilistic Disambiguation (ii)

Definition of conditional probability functions relating links & sources

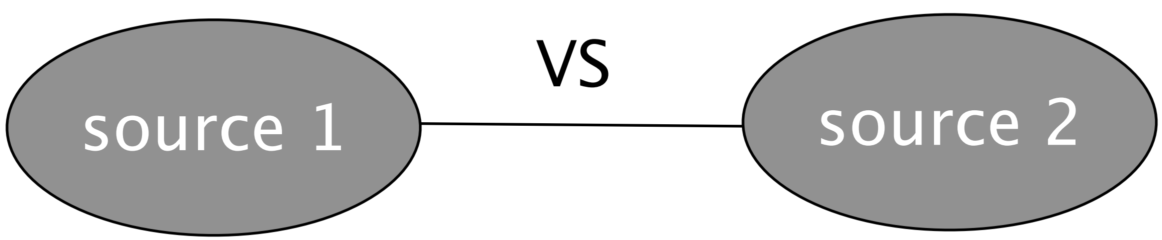
- Transitive closures of link properties (*entity graph*)
  - *ID Equivalence* is
    - *symmetric*
    - *transitive*



Probabilistic Disambiguation (iii)

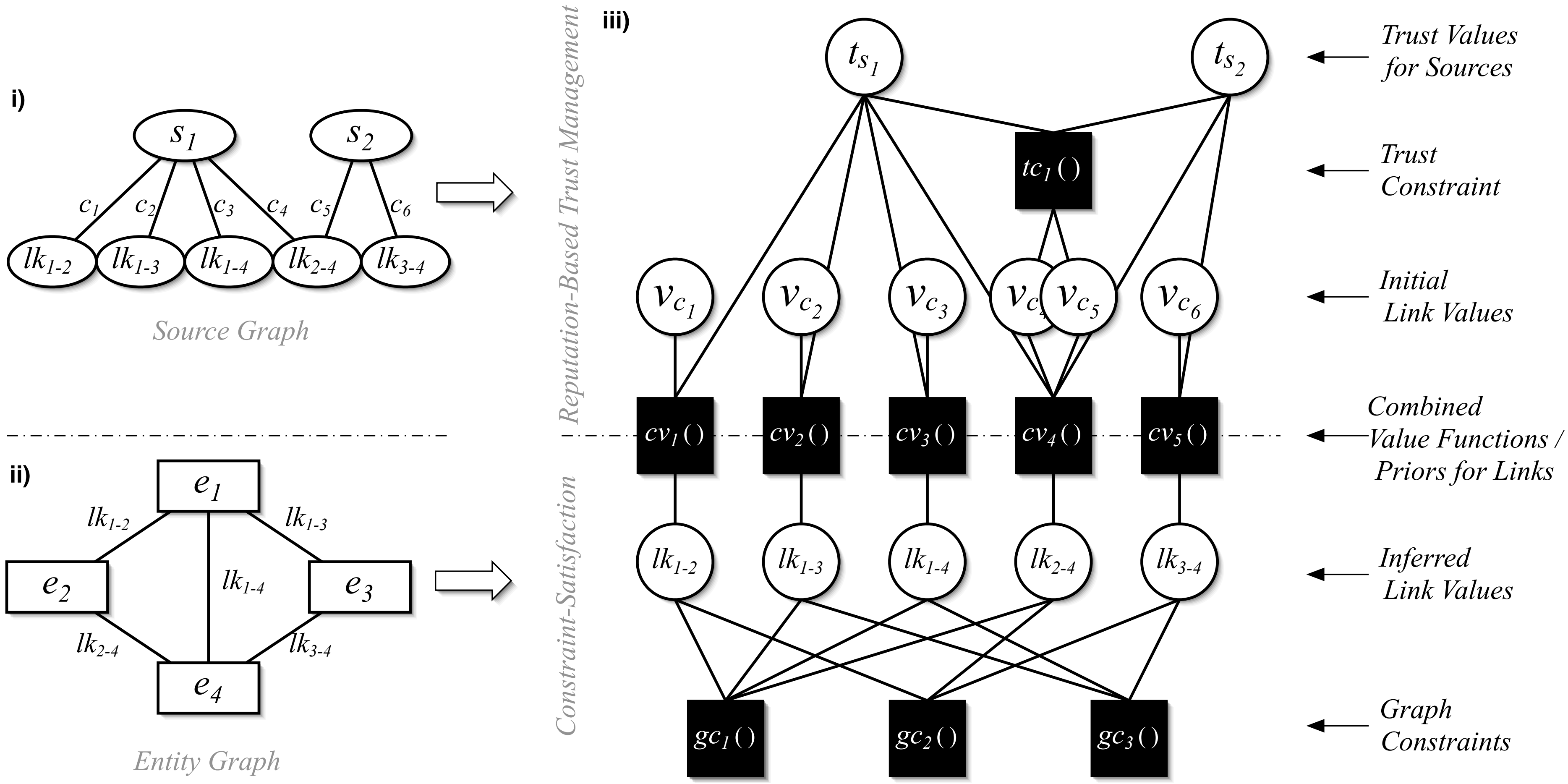
Definition of conditional probability functions relating links & sources

- Source discrimination (*source graph*)
  - Through internet domains / authentication mechanisms
    - openid, foaf-ssl, etc.
  - High confidence values for well-known + well-behaved sources



CSAIL

Probabilistic Disambiguation



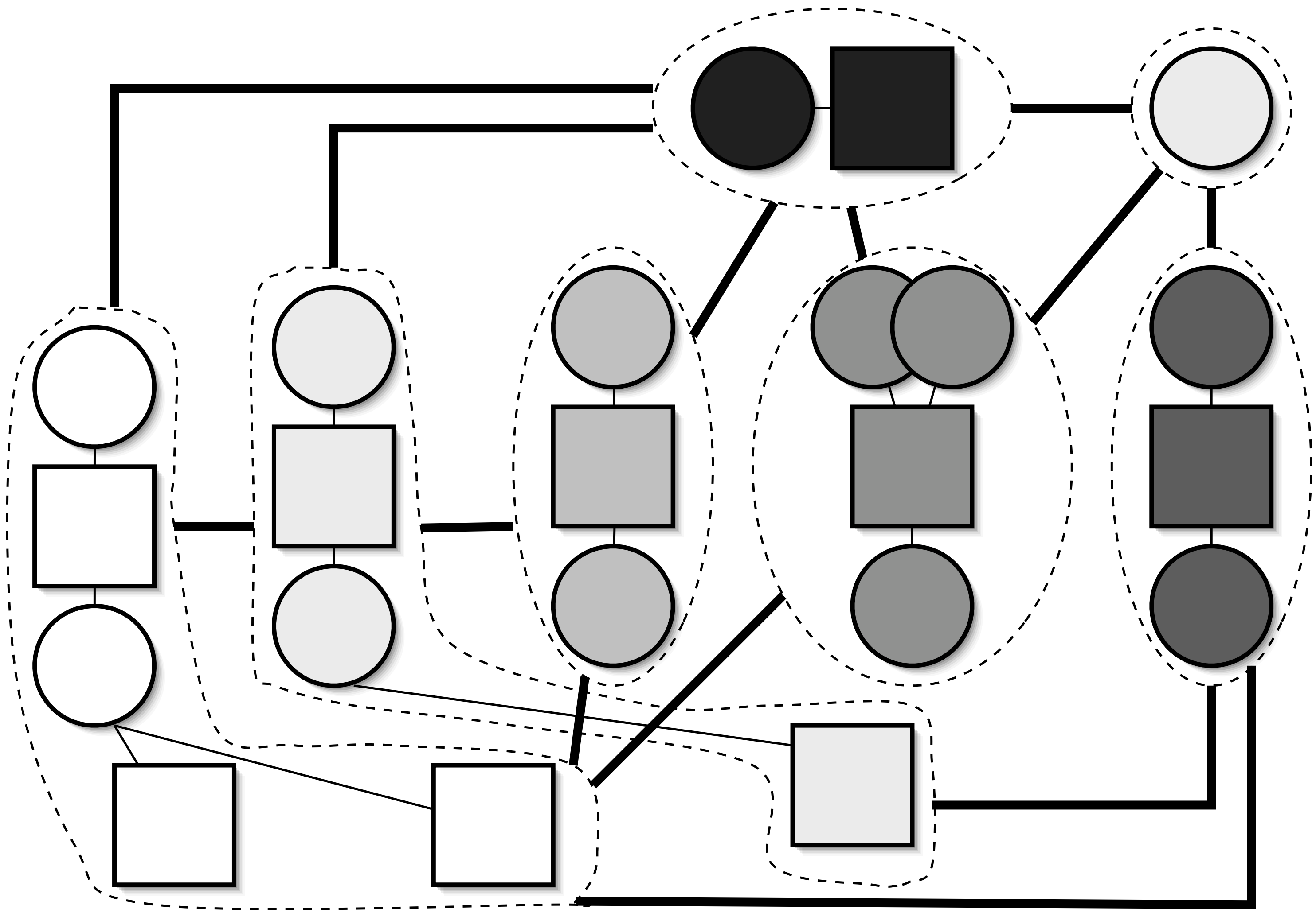
*Probabilistic inference on *combined* graph*

Scalability

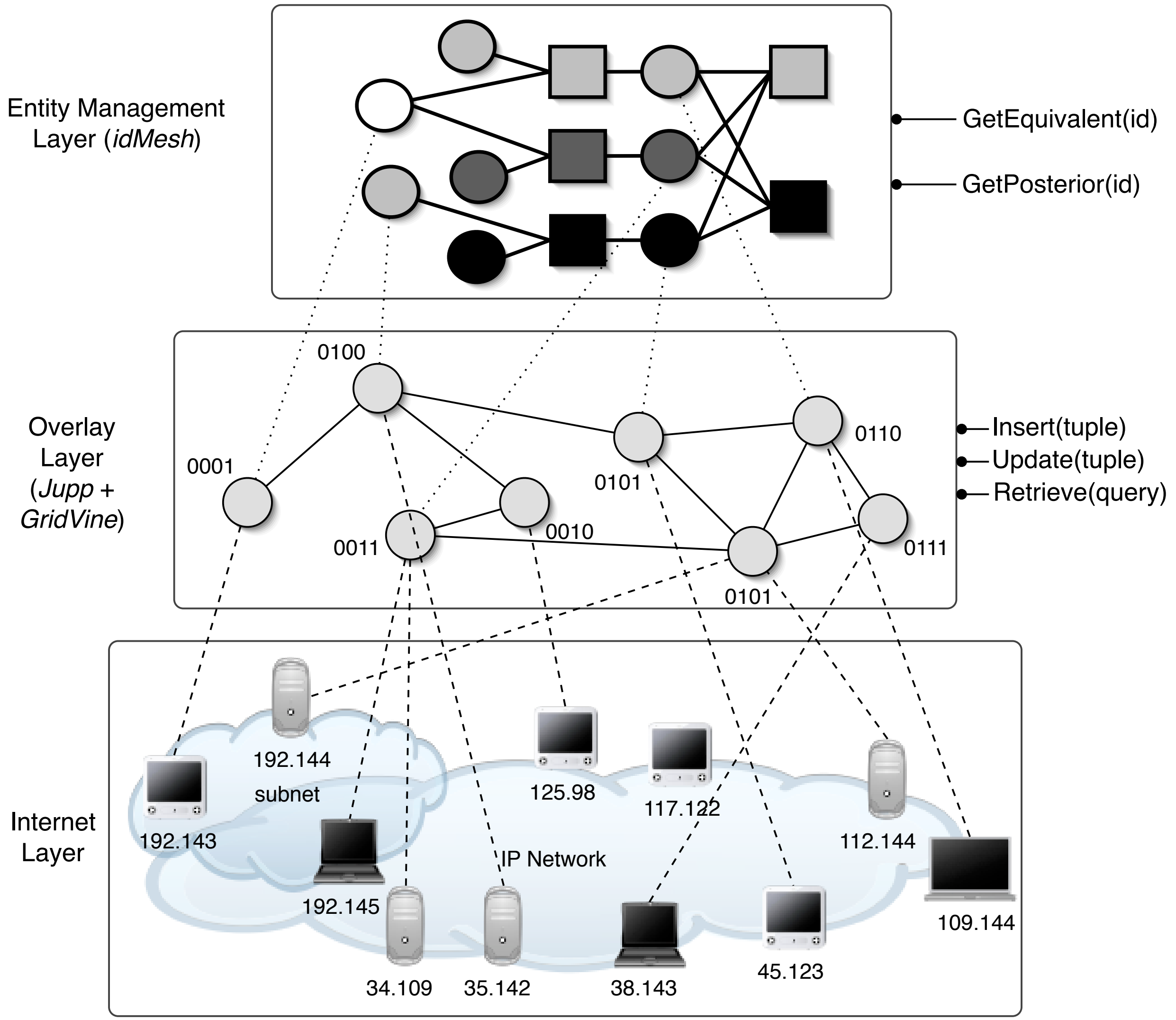
- Problem: both source / entity graphs can become *very large* in practice
  - Unbounded number of sources
    - peer production
  - Cheap production of (uncertain) links
    - automated matching algorithms

➡ inference in itself should be *decentralized*

Distributing the Probabilistic Graph



Distributed, P2P Architecture



Message Passing

DHT

Internet

idMesh: summary of Results

- *Efficient*, *distributed* computations
  - Parallelized sums & products only
  - Quasi-instantaneous on a local machine
  - Naturally *scales out* in networked environments
    - A couple of seconds to disambiguate 8'000 entities interlinked by 24'000 links on 400 machines

- High *discriminative power* in practice
  - $90\%^{+}$ accuracy with well-behaved but uncertain sources
  - 75% accuracy with 90% malign sources

Conclusions

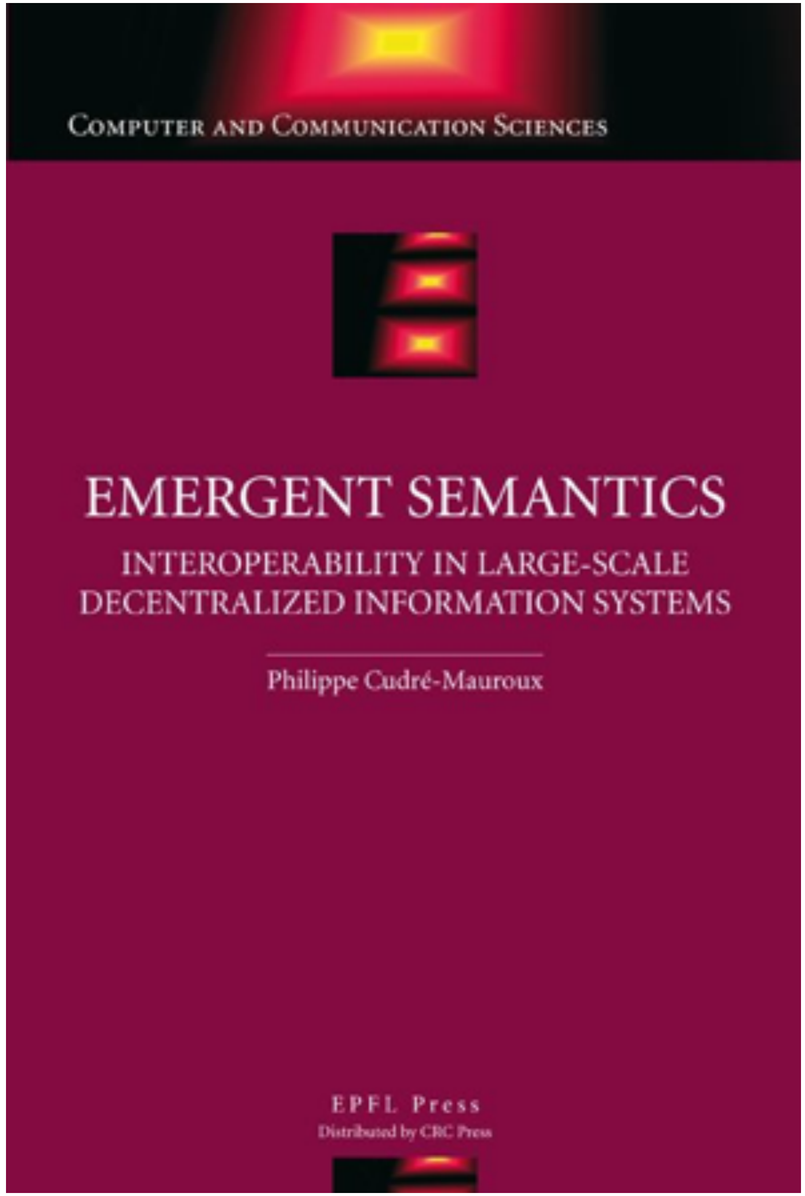
- More and more machine-processable (semi-structured) data available
  - Sensing Technologies
  - Peer Production
  - Human Computation

- Top-down efforts to align data have failed largely

- Emergent Semantics
  - Bottom-up
  - Dynamic, self-organizing
  - *Best-Effort*

➠ Only resort to foster interoperability in the large scale decentralized data spaces currently emerging

CSAIL


COMPUTER AND COMMUNICATION SCIENCES
EMERGENT SEMANTICS
INTEROPERABILITY IN LARGE-SCALE
DECENTRALIZED INFORMATION SYSTEMS
Philippe Cudré-Mauroux
EPFL Press
Distributed by CRC Press

Emergent Semantics:
Rethinking Interoperability for Large Scale Decentralized Information Systems

references:

http://people.csail.mit.edu/pcm/

Self-Organizing Knowledge Systems Symposium, VU Amsterdam, April 29, 2010