idMesh: Graph-Based Disambiguation of Linked Data

Philippe Cudré-Mauroux -- MIT

joint work with
Parisa Haghani, Michael Jost, Karl Aberer (EPFL)
and Hermann de Meer (U. Passau)

April 24, 2009
World Wide Web Conference

Overview

- A Web of Resources
  - Distributed Naming Game
  - Entity Consolidation
- idMesh Constructs
- Link-Analysis Framework
- System Architecture
- Performance
- Conclusions & Future Work

CSAIL

A Web of Resources

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

CSAIL

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://lsirpeople.epfl.ch/pcudre/i

http://semanticweb.org/wiki/Special:ExportRDF/Philippe_Cudr%C3%A9-Mauroux

http://tw.rpi.edu/wiki/Special:ExportRDF/Philippe_Cudr%C3%A9-Mauroux

http://wiki.ontoworld.org/index.php/Special:ExportRDF/Philippe_Cudr%C3%A9-Mauroux

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

http://prauw.cs.vu.nl:8080/flink/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... .....

CSAIL

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[illegible]ple.epfl.ch/pcudre/i

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

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

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

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

http://prauw.cs.vu.nl:808[illegible]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/

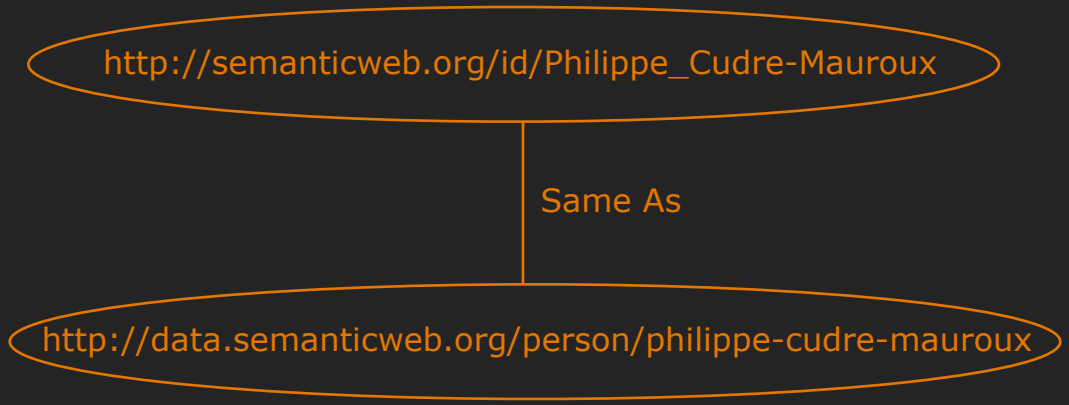
http://www.facebook.com/profile.php?id=1251943... .....

ID Jungle

CSAIL

Entity Consolidation (i)

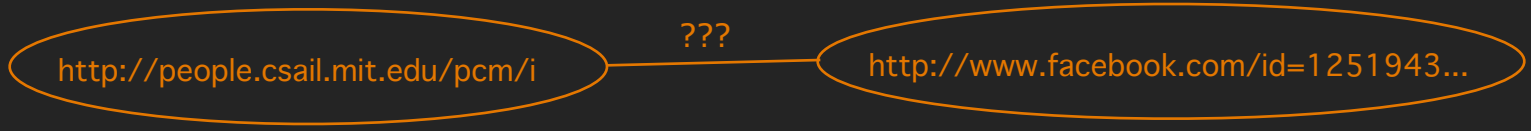
- A few constructs are increasingly used to consolidate Wed identifiers
  - OWL:SameAs, XFN rel:me, pipes, etc.



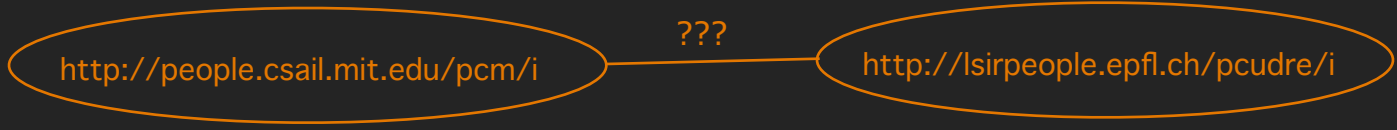
CSAIL

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)

CSAIL

New Twist on an Old Problem

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

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

CSAIL

The *idMesh* Approach

- *idMesh* suggests a radically different approach to online entity consolidation that is
  - *User-driven*
  - *Best-effort (probabilistic)*
  - *Decentralized*

- Link-analysis framework based on transitive closures of relationships
  - *Emergent semantics*
    - semantics of data derived through network
    - the sum is greater than the parts

CSAIL

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>
```

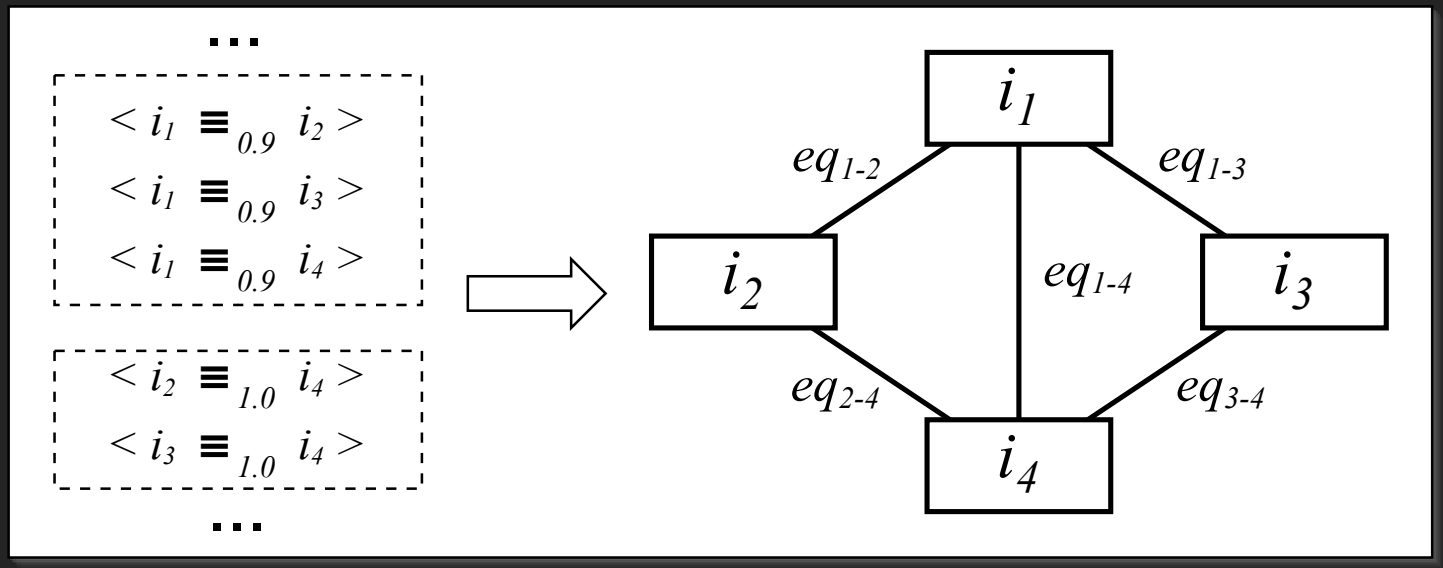
- Two levels of granularity
  - Entity disambiguation
  - Temporal discrimination
- Confidence values
- Can encompass previous constructs

```
<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* or *interrelated* identifiers
  - no associated contents, attributes, or properties...

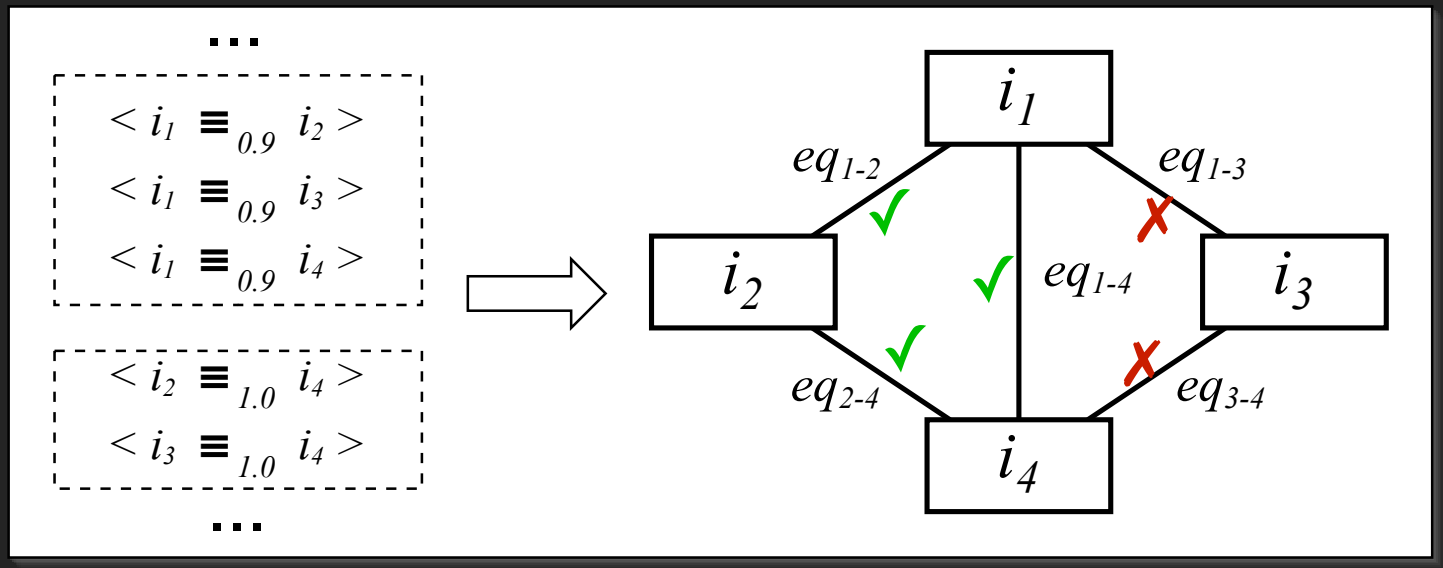


- Output: *clusters* of *equivalent* identifiers
  - probabilistic, *a posteriori* network equivalence
  - equivalence based on probabilistic threshold

CSAIL

Problem Definition

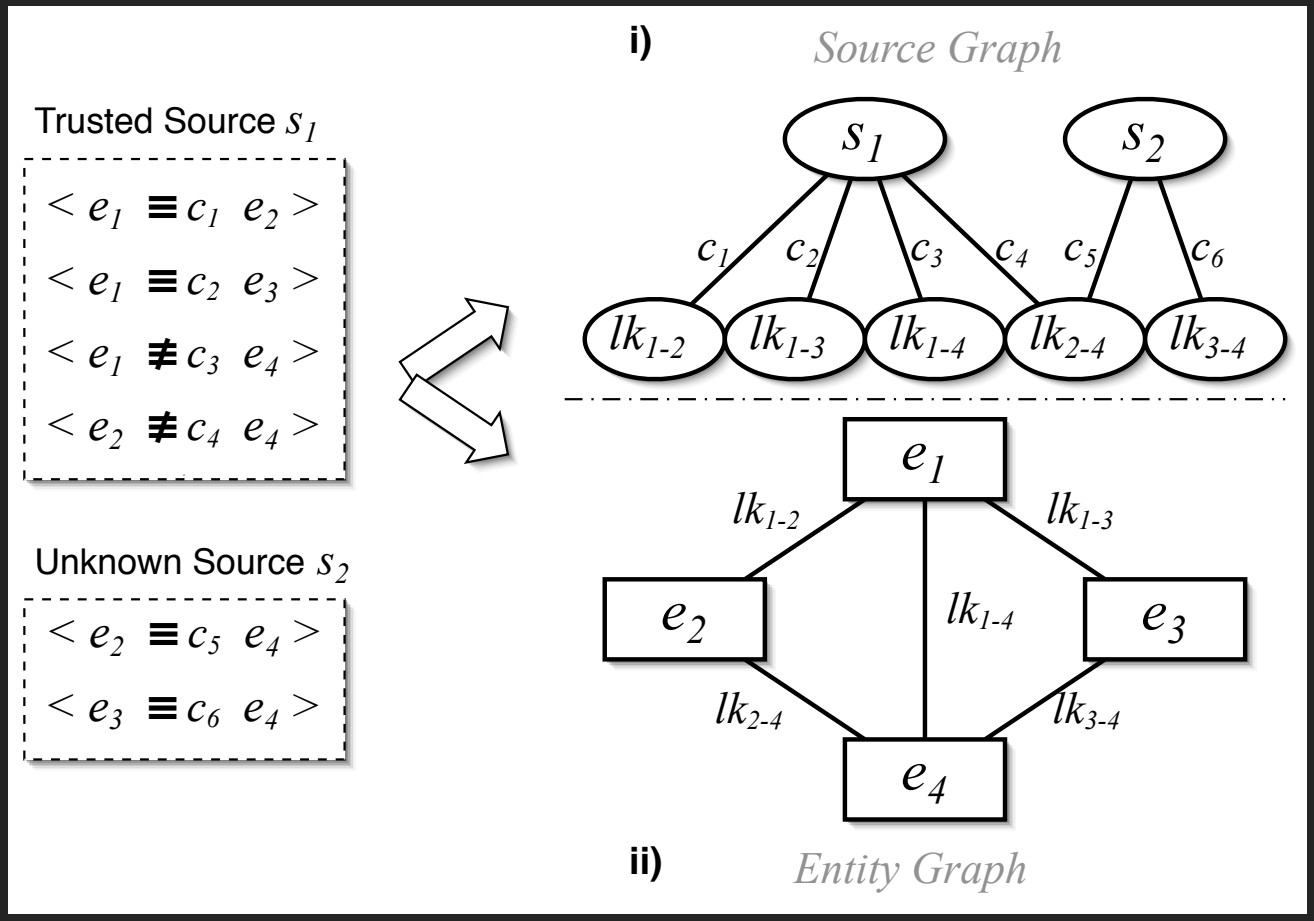
- Input: series of statements defining a *weighted graph* or *interrelated* identifiers
  - no associated contents, attributes, or properties...



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

CSAIL

Probabilistic Disambiguation (i)



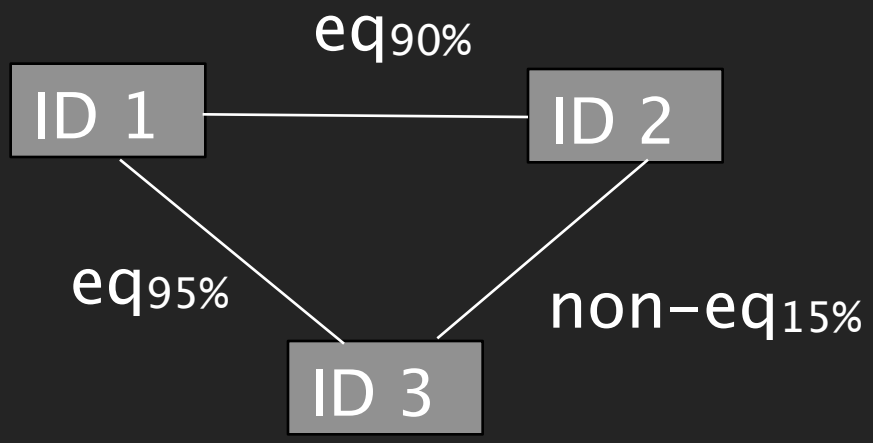
Definition of two graphs

CSAIL

Probabilistic Disambiguation (ii)

Definition of conditional probability functions relating links & sources

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

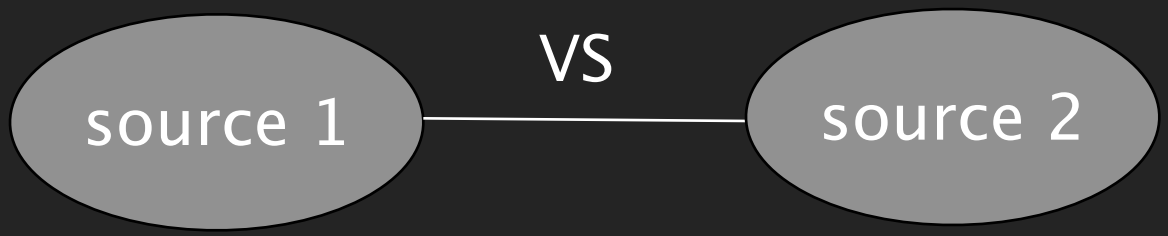


CSAIL

Probabilistic Disambiguation (iii)

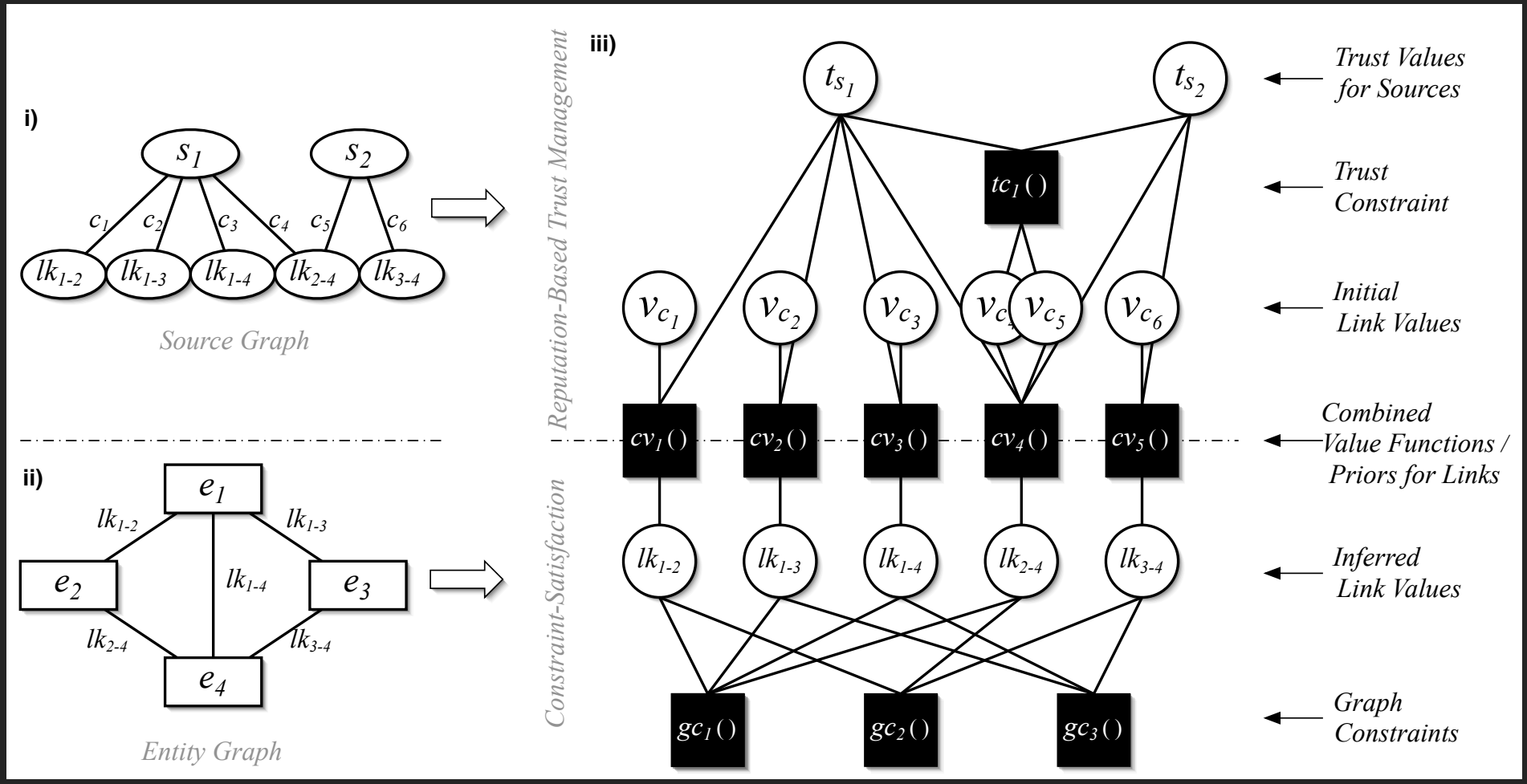
Definition of conditional probability functions relating links & sources

- 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 (iv)



*Probabilistic inference on *__combined__* graph*

CSAIL

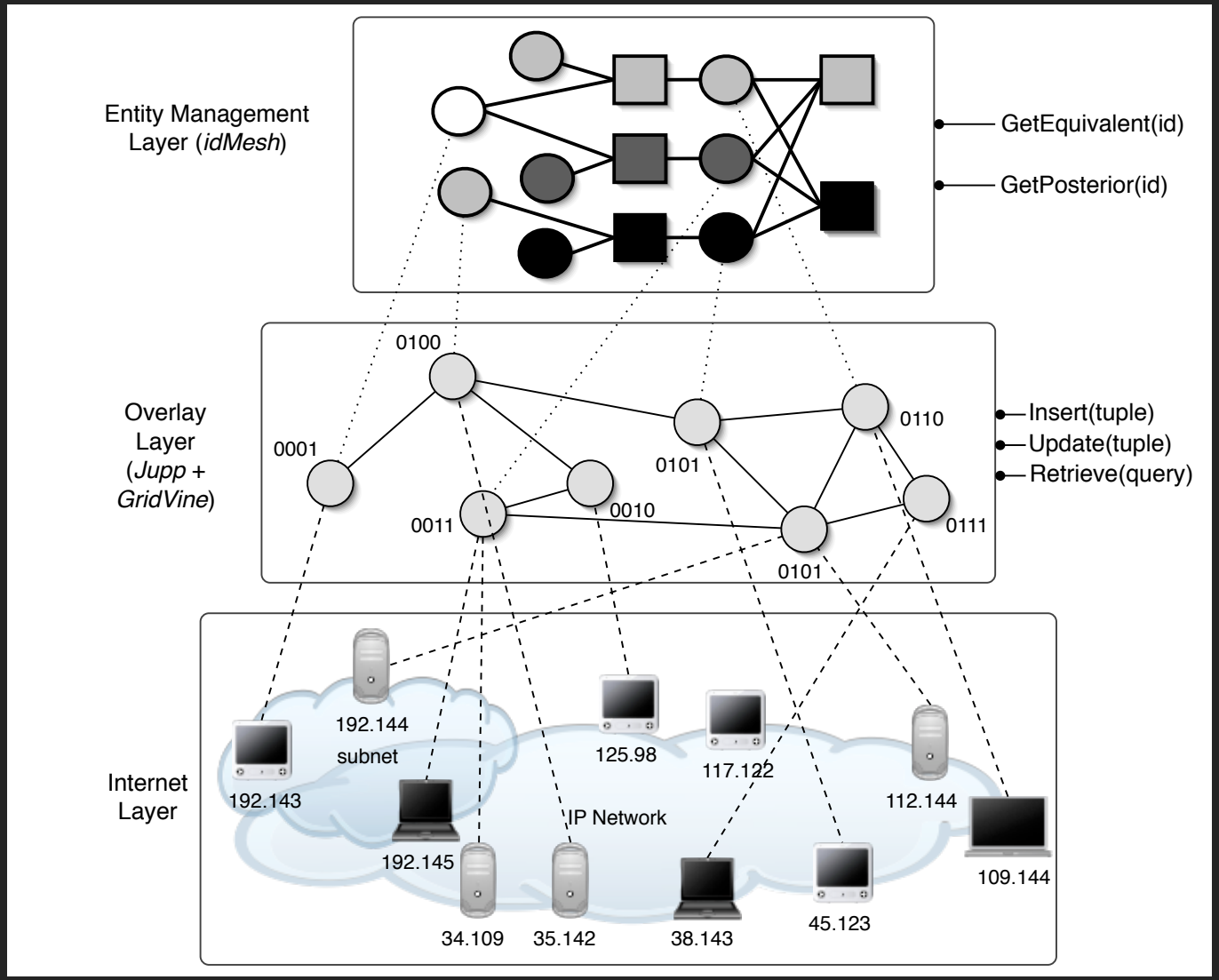
Scalability

- 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 should in itself be *decentralized*

CSAIL

Distributed, P2P Architecture



Message Passing

DHT

Internet

CSAIL

Summary of Results

- *Efficient*, *distributed* computations
  - Parallelized sums & products only
  - Quasi-instantaneous on a local machine
  - Naturally *scales up* in networked environments
    - 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

CSAIL

Conclusions & Future Work (i)

- *idMesh*: a ...
  - user-driven
  - probabilistic
  - decentralized

  ... link-analysis approach to disambiguate linked data.

- Can be combined with previous approaches
  - Previous constructs
  - Automated matching / content-based disambiguation
  - Reputation-based trust mechanisms

CSAIL

Conclusions & Future Work (ii)

- *Could* be extended to encompass further types of links
  - subsumption
  - relatedness
- *Should* be extended to support personalized disambiguation capabilities
  - context-sensitive

CSAIL

idMesh: Graph-Based Disambiguation of Linked Data

Philippe Cudré-Mauroux -- MIT

pcm@csail.mit.edu