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
A Web of Resources

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

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

foaf:made

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://semanticweb.org/wiki/Special:ExportRDF/Philippe_Cudr%C3%A9-Mauroux
http://tw.rpi.edu/wiki/Special:ExportRDF/Philippe_Cudre%3CA9-Mauroux
http://wiki.ontoworld.org/index.php/Special:ExportRDF/Philippe_Cudr%C3%A9-Mauroux
http://korrekt.org/index.php/Special:ExportRDF/Philippe_Cudre%3CA9-Mauroux
http://www.zoominfo.com/PersonID=402960578 http://www.flickr.com/photos/28735...@N00/
http://www.facebook.com/profile.php?id=1251943...
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://www.zoominfo.com/PersonID=402960578
http://www.flickr.com/photos/28735...@N00/
http://www.facebook.com/profile.php?id=1251943...
Entity Consolidation (i)

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

http://data.semanticweb.org/person/philippe-cudre-mauroux
http://semanticweb.org/id/Philippe_Cudre-Mauroux
Same As
Entity Consolidation (ii)

- Online entity consolidation is a complex game
  - Simple binary constructs are often insufficient

  - Social contexts (e.g., professional vs personal entities)
    - http://people.csail.mit.edu/pcm/i ➔ ??? ➔ http://www.facebook.com/id=1251943...

  - Granularity (e.g., out-of-date entities)

  - Uncertainty (e.g., automatically-generated entities)
New Twist on an Old Problem

- Well-known problem known 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
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
idMesh Constructs

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

• Input: series of statements defining a \textit{weighted graph} or \textit{interrelated} identifiers
  ■ no associated contents, attributes, or properties...

• Output: \textit{clusters} of \textit{equivalent} identifiers
  ■ probabilistic, \textit{a posteriori} network equivalence
  ■ equivalence based on probabilistic threshold
Problem Definition

• Input: series of statements defining a \textit{weighted graph} or \textit{interrelated} identifiers
  ■ no associated contents, attributes, or properties...

• Output: \textit{clusters} of \textit{equivalent} identifiers
  ■ probabilistic, \textit{a posteriori} network equivalence
  ■ equivalence based on probabilistic threshold
Probabilistic Disambiguation (i)

Definition of two graphs

Source Graph

Entity Graph

Trusted Source $s_1$

$< e_1 \equiv c_1 e_2 >$
$< e_1 \equiv c_2 e_3 >$
$< e_1 \equiv c_3 e_4 >$
$< e_2 \equiv c_4 e_4 >$

Unknown Source $s_2$

$< e_2 \equiv c_5 e_4 >$
$< e_3 \equiv c_6 e_4 >$
**Probabilistic Disambiguation (ii)**

*Definition of conditional probability functions relating links & sources*

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

```
<table>
<thead>
<tr>
<th></th>
<th>ID 1</th>
<th></th>
<th>ID 2</th>
<th></th>
<th>ID 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>eq</td>
<td>90%</td>
<td></td>
<td>eq95%</td>
<td></td>
<td>non-eq15%</td>
</tr>
</tbody>
</table>
```
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

source 1  VS  source 2
well-known, well-behaved  unknown, conflicting
Probabilistic Disambiguation (iv)

**Source Graph**

- $s_1$ connected by $c_1$, $c_2$, $c_3$, $c_4$, $c_5$, $c_6$
- $l_{k_{1-2}}$, $l_{k_{1-3}}$, $l_{k_{1-4}}$, $l_{k_{2-4}}$, $l_{k_{3-4}}$

**Entity Graph**

- $e_1$, $e_2$, $e_3$, $e_4$
- $l_{k_{1-2}}$, $l_{k_{1-3}}$, $l_{k_{1-4}}$, $l_{k_{2-4}}$, $l_{k_{3-4}}$

**Reputation-Based Trust Management**

- Initial Link Values
- Combined Value Functions / Priors for Links
- Inferred Link Values
- Graph Constraints
- Trust Values for Sources
- Trust Constraint

**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 should in itself be *decentralized*
Distributed, P2P Architecture

Entity Management Layer (idMesh)

Overlay Layer (Jupp + GridVine)

Internet Layer

Message Passing

DHT

Internet
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
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
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
idMesh: Graph-Based Disambiguation of Linked Data

Philippe Cudré-Mauroux -- MIT

pcm@csail.mit.edu