ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web

WWW2003 (Budapest, May 23, 2003)
Paper Presentation

Kemafor Anyanwu
Amit Sheth
Large Scale Distributed Information Systems Lab
University of Georgia

This material is based upon work supported by the National Science Foundation under Grant No. 0219649.

From …..

Finding things

to…..

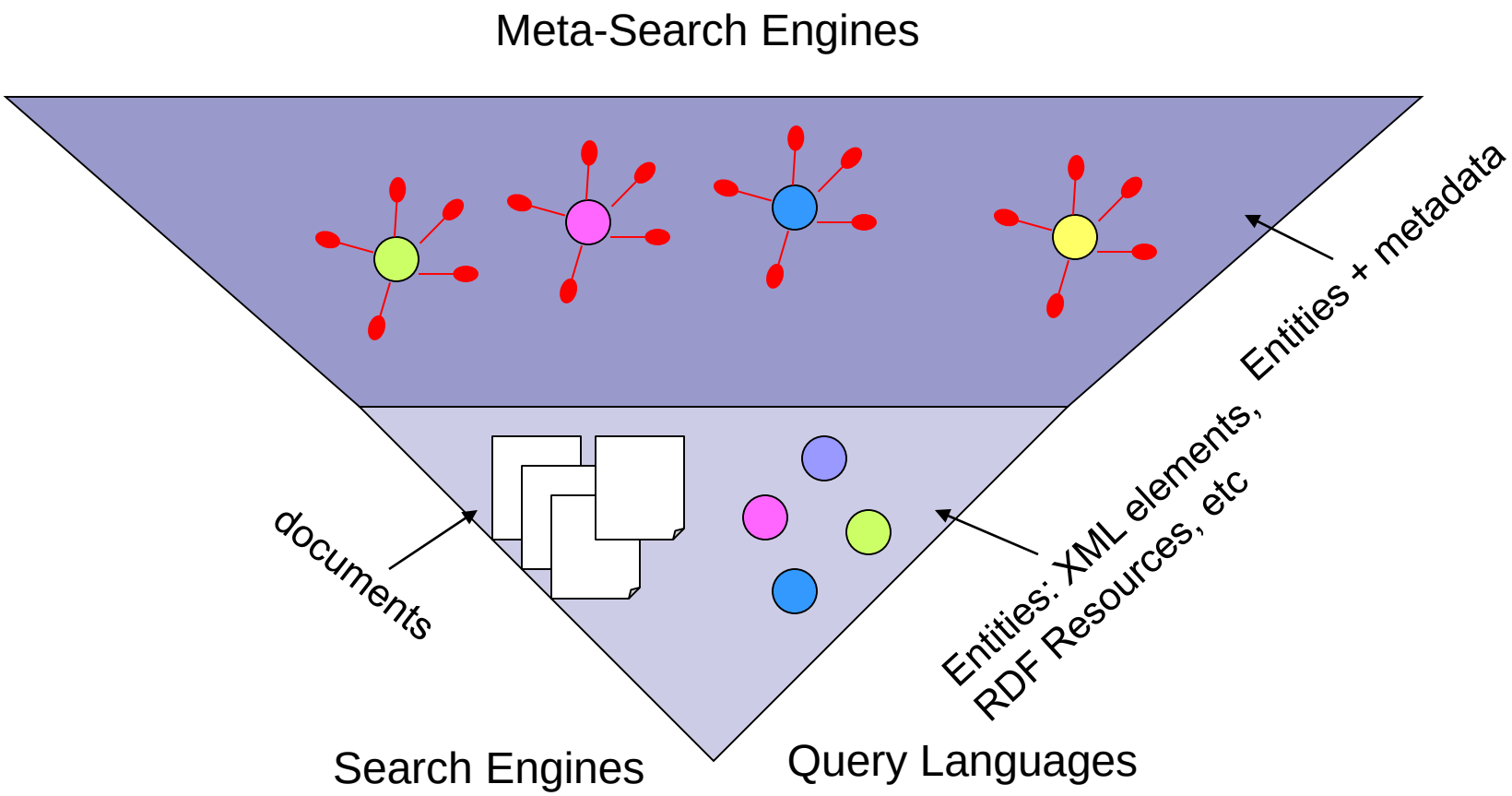
“Finding out about” [Belew00]

relationships!

Outline

- Semantic Associations: Introduction
- A Formal Framework for Semantic Associations on the Semantic Web
- ρ-Queries For Discovering Semantic Associations
  - Implementation Strategies & Issues
- Related Work
- Conclusion & Future Work

Web Search/Query Techniques are “Entity-Centric”



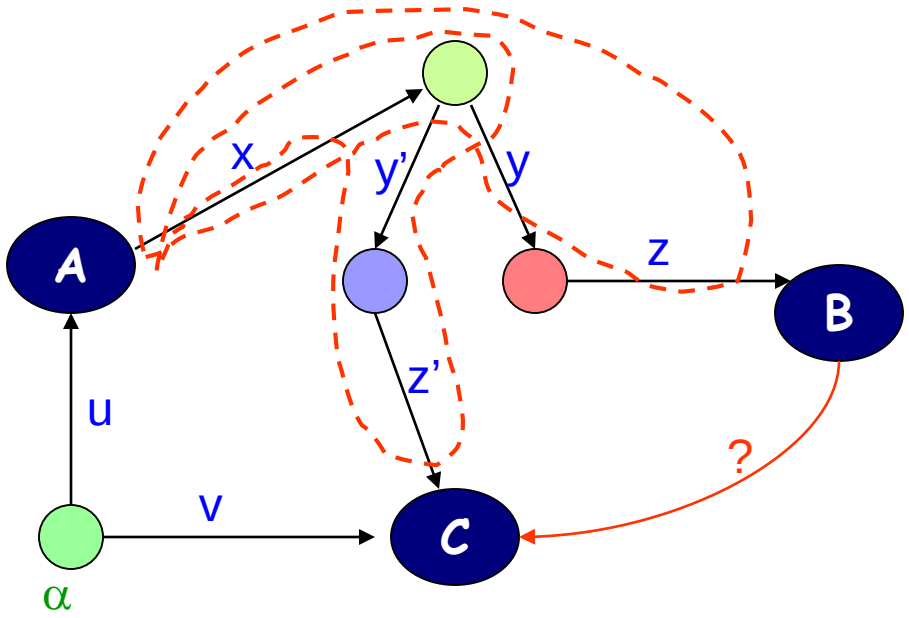
But…..

- “An object by itself is intensely uninteresting” .

  Grady Booch, Object Oriented Design with Applications, 1991

We need

- Mechanisms for querying about and retrieving **complex relationships** between entities.



1. A is related to B by x.y.z
2. A is related to C by
   i. x.y'.z'
   ii. u.v (*undirected path)*

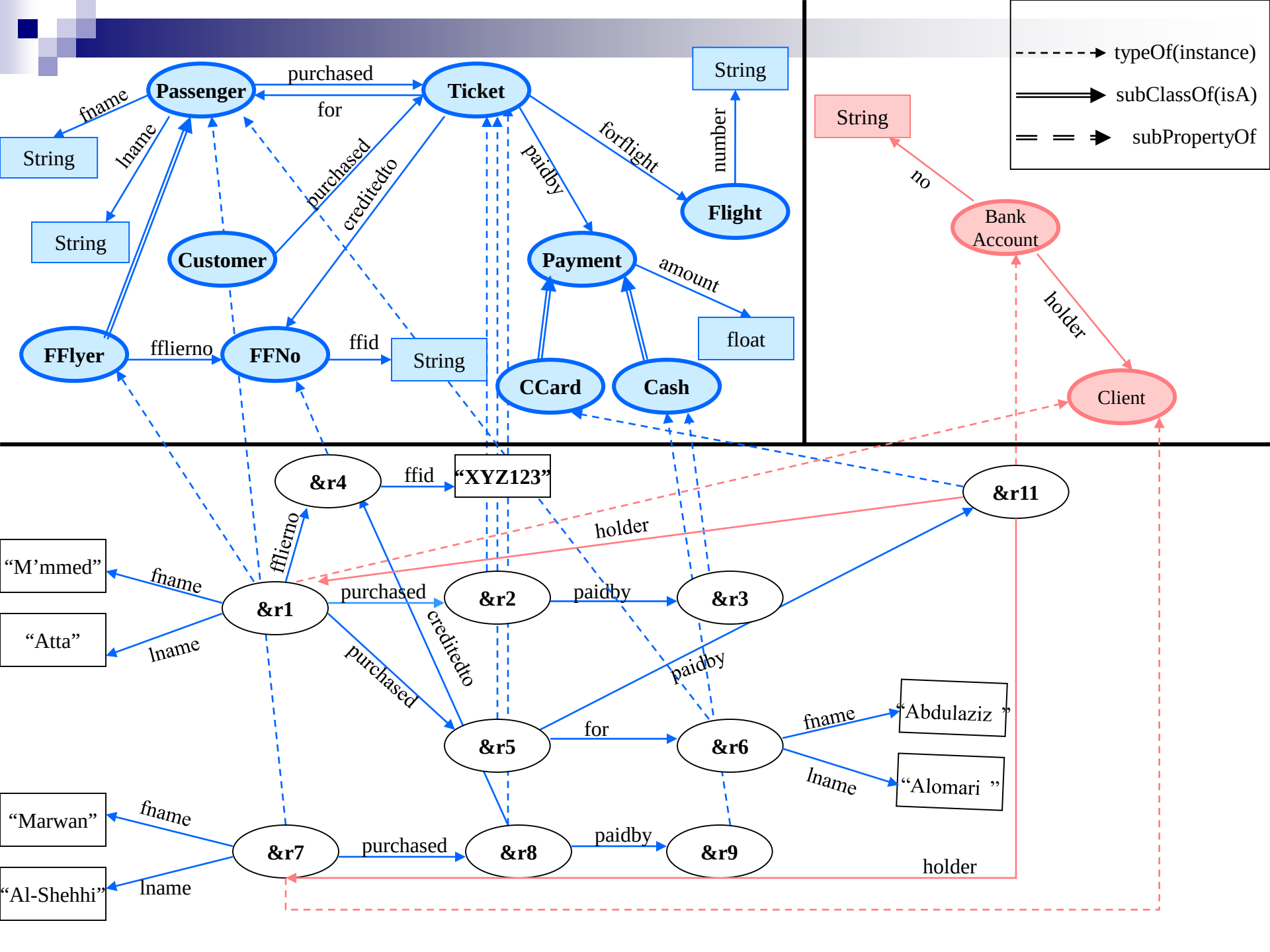
3. A is "related *similarly*" to B as it is to C
   (y' ⊆ y and z' ⊆ z → x.y.z ≅ x.y'.z')
   So are B and C related?

Why do we need this?

- Very useful in information analytics
  - national security
  - business intelligence
- Avoids the task of familiarizing oneself with schemas in order to formulate queries
  - especially when multiple schemas are involved !

Example in 9-11 context

- What are relationships between Khalid Al-Midhar and Majed Moqed ?
  - *Connections*
    - Bought tickets using same frequent flier number
  - *Similarities*
    - Both purchased tickets originating from Washington DC paidby cash and picked up their tickets at the Baltimore-Washington Int'l Airport
    - Both have seats in Row 12
- "What relationships exist (if any) between Osama bin Laden and the 9-11 attackers"


typeOf(instance)
subClassOf(isA)
subPropertyOf
Passenger
Ticket
purchased
for
fname
String
lname
String
String
number
forflight
Flight
paidby
purchased
creditedto
Customer
Payment
amount
float
FFlyer
fflierno
FFNo
ffid
String
CCard
Cash
String
no
Bank Account
holder
Client
&r4
ffid
"XYZ123"
&r11
holder
fflierno
"M'mmed"
fname
&r1
purchased
&r2
paidby
&r3
"Atta"
lname
creditedto
purchased
paidby
&r5
for
&r6
fname
"Abdulaziz "
lname
"Alomari "
"Marwan"
fname
&r7
purchased
&r8
paidby
&r9
holder
"Al-Shehhi"
lname

A Foundation for Semantic Associations on the Semantic Web

Complex Relationships?

- Traditional notions of relationships are captured by single n-ary relations
  - e.g. RDF:Property, UML Association, E-R:relationship, etc.
- Complex relationships can be viewed as specific compositions of multiple single n-ary relations
  - e.g. Sequence composition of binary relations allows us to capture paths
- Relation Sequences + certain operations allow us to detect very interesting relationships
  - Connectivity
  - Similarity

Semantic Web

- RDF is the current W3C standard for metadata representation on the Semantic Web
- Other proposals include OWL, DAML+OIL, UML, Topic Maps, etc.
- In RDF, the basic unit of relationship is a Property

Formal Data Model for RDF

- (Karvounarakis et al 2002) gives a formalization of RDF/RDFS which forms the basis for a typed RDF query language – RQL.
  - It provides a type system for RDF Schemas
  - For each type e.g. class type τ_c, property type τ_p, there is a mapping [[ ]] to its members
  - e.g. for a property type p, [[p]] is defined as {[v1, v2] | v1$\in$ [[p.domain]], v2$\in$ [[p.range]] } $\cup$ { [[p']] | ' $\subseteq$ p}

We add

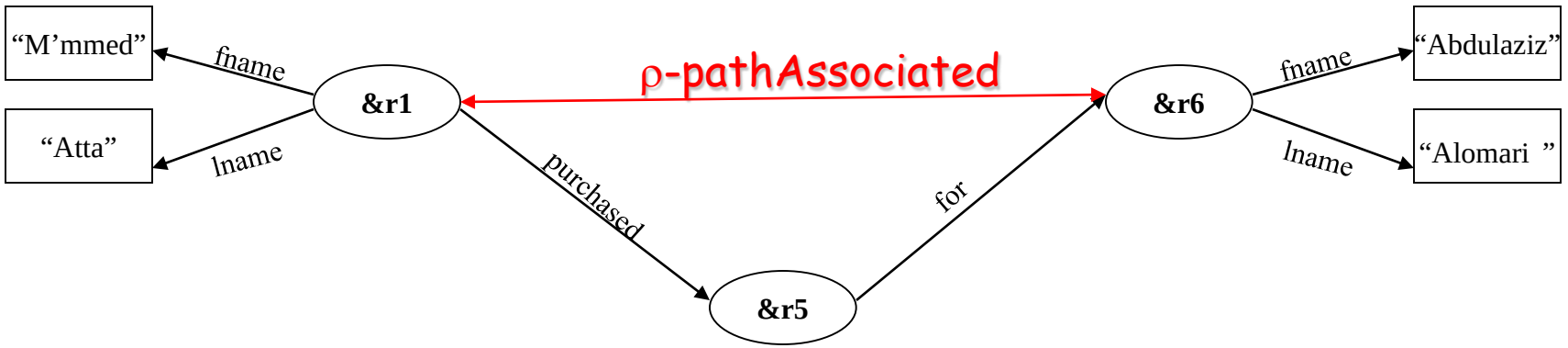
- The notion of an RDFS *Schema Set*. Basically, a union of a set of RDF Schemas supplying the *context* for a query
  - In the example, Flight + Banking Schemas
- The notion of a *Property Sequence*, which is the sequential composition of RDF Properties and define relations on Property Sequences
- A formalization for Semantic Associations based on Property Sequences and their relations

Property Sequence	Finite sequence of properties $PS = [P_1, P_2, P_3, \ldots P_n]$, P_i is a property defined in an RDF Schema RS_j of a schema set RSS. e.g. [*purchased*, *paidby*]. $[[PS]] \subseteq \prod_{i=1}^{n} [[P_i]]$ *such that* $ps \in [[PS]]$ implies i. $ps[i] \in [[P_i]]$ for $1 \leq i \leq n$ ii. $ps[i][1] = ps[i+1][0]$
Joined Property Sequences $(\bowtie_\rho)$	$PS_1 \bowtie_\rho PS_2 \leftarrow$ $\exists c \in (PS_1.NodesOfPS() \cap PS_2.NodesOfPS())$. c is called join node
ρ-Isomorphic Property Sequences $(\cong_\rho)$ A sequence such as "awarded.paidby" which means that a passenger was awarded a ticket, paid for by frequent miles is considered ρ-Isomorphic to "purchased.paidby".	$PS_1 \cong_\rho PS_2 \leftarrow$ i. $PS_1 = [P_1, P_2, P_3, \ldots P_m]$, $PS_2 = [Q_1, Q_2, Q_3, \ldots Q_m]$ ii. for all i, $1 \leq i \leq m$: $P_i = Q_i$ or $P_i \subseteq Q_i$ or $Q_i \subseteq P_i$ ($\subseteq$ means subpropertyOf) Note that the Property Sequences need not be exact to be ρ-Isomorphic, just similar.

Semantic Associations

ρ-pathAssociation

- Let PS be a Property Sequence and ps ∈ [[PS]].
- If x and y are the origin/terminus and terminus/origin of ps respectively,
  - ρ-pathAssociated (x, y)



ρ-joinAssociation

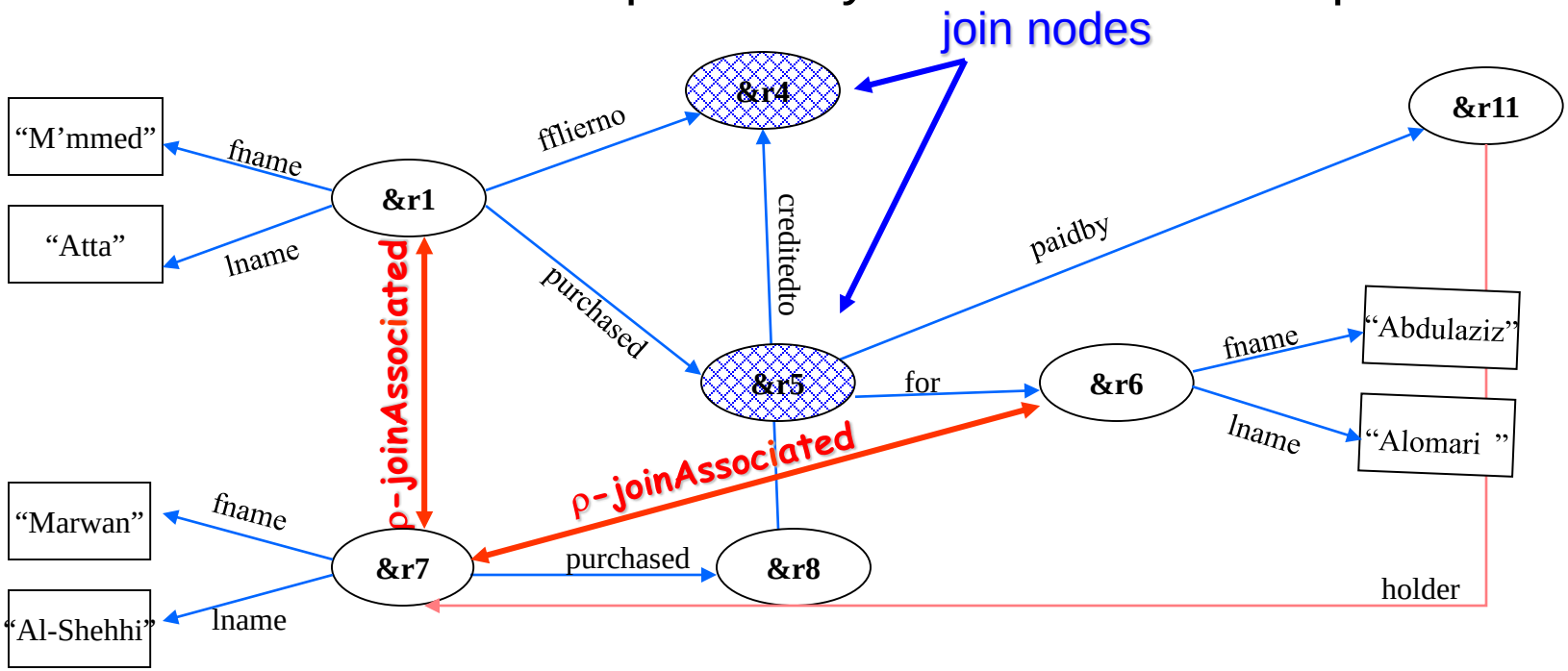
- ρ-joinAssociated (x, y) ←

a) ∃ PS_1,PS_2: PS_1 ⋈ρ PS_2

b) ∃ ps_1,ps_2: ps1 ∈ [[PS_1]], ps2 ∈ [[PS_2]]

   i. x is the origin of ps1 and y is the origin of ps2 or

   ii. x is the terminus of ps1 and y is the terminus of ps2.



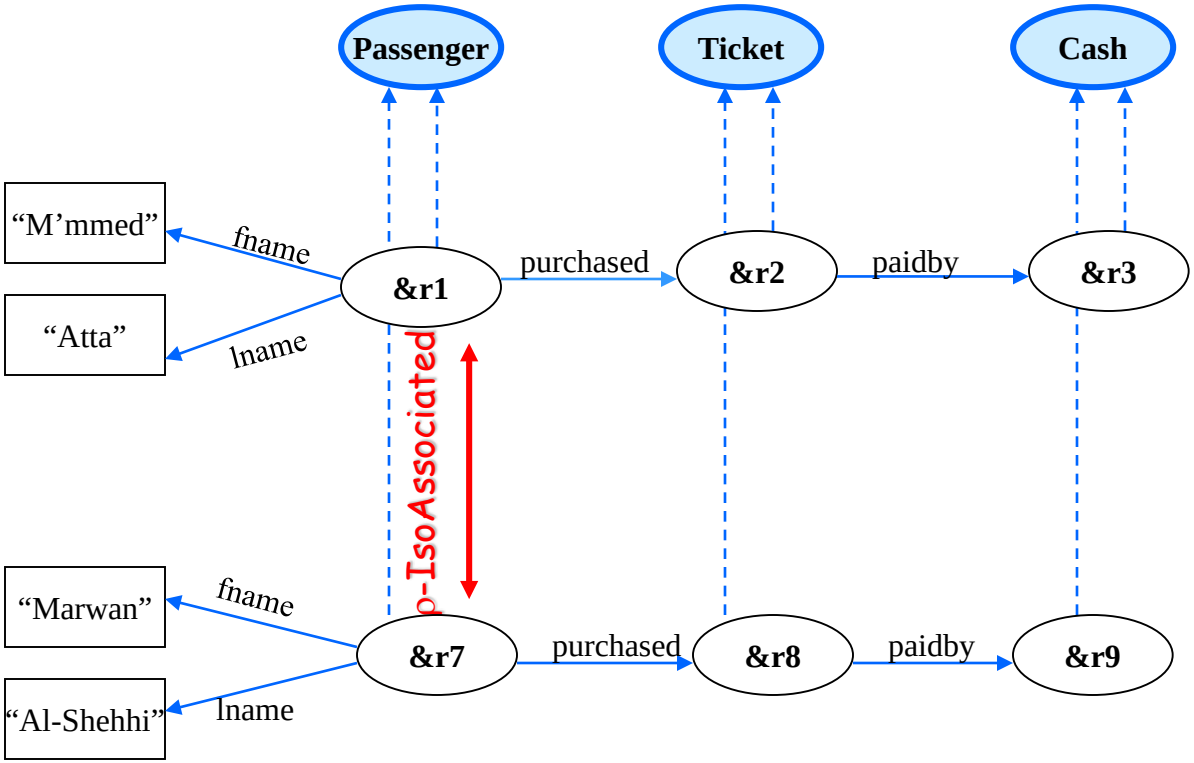
ρ-IsoAssociation

- ρ-IsoAssociated (x, y) ←

a) $\exists\ PS_1, PS_2 : PS_1 \cong_\rho PS_2$

b) $\exists\ ps_1, ps_2 : ps_1 \in [[PS_1]],\ ps_2 \in [[PS_2]]$

   i. x is the origin/terminus of ps_1 and y is the origin/terminus of ps_2.



ρ-Queries for Discovering Semantic Associations

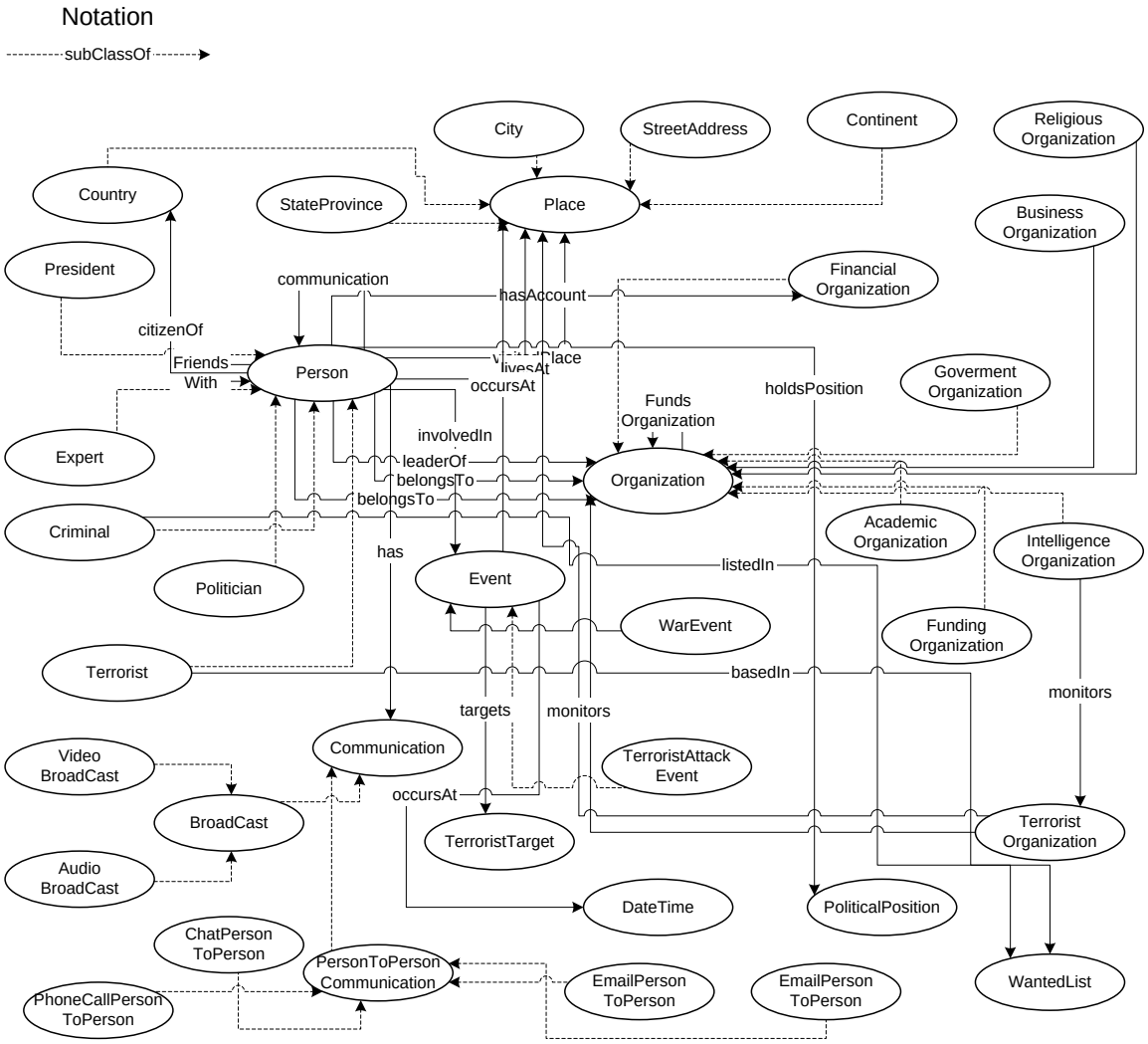
ρ-Queries

- Let $\tau_U^{(2)} = \{ \{x, y\} : x, y \in \tau_U \text{ and } x \neq y\}$,
  PS = {PS : PS is a Property Sequence},
  $\mathbf{PS}^{(2)} = \{\{PS_1, PS_2\} : PS_1, PS_2$ are Property Sequences$\}$
- A ρ-Query Q maps from a pair of keys to the **PS** and $\mathbf{PS}^{(2)}$ in the following manner:
  - $\rho: \tau_U^{(2)} \rightarrow 2^{\mathbf{PS}}$
  - $\rho^{\bowtie_\rho} : \tau_U^{(2)} \rightarrow 2^{\mathbf{PS}(2)}$
  - $\rho^{\cong\rho}: \tau_U^{(2)} \rightarrow 2^{\mathbf{PS}(2)}$

Implementation Approaches for ρ-Operators

- Exploit existing RDF storage & query infrastructure:
  - Persistent Stores → Translations to query expressions at data store layer, guided by index structures
  - Memory-Resident Stores → Employ graph traversal algorithms
- Alternative Representation with complimentary indexes and algorithms i.e. search-engine type Strategy

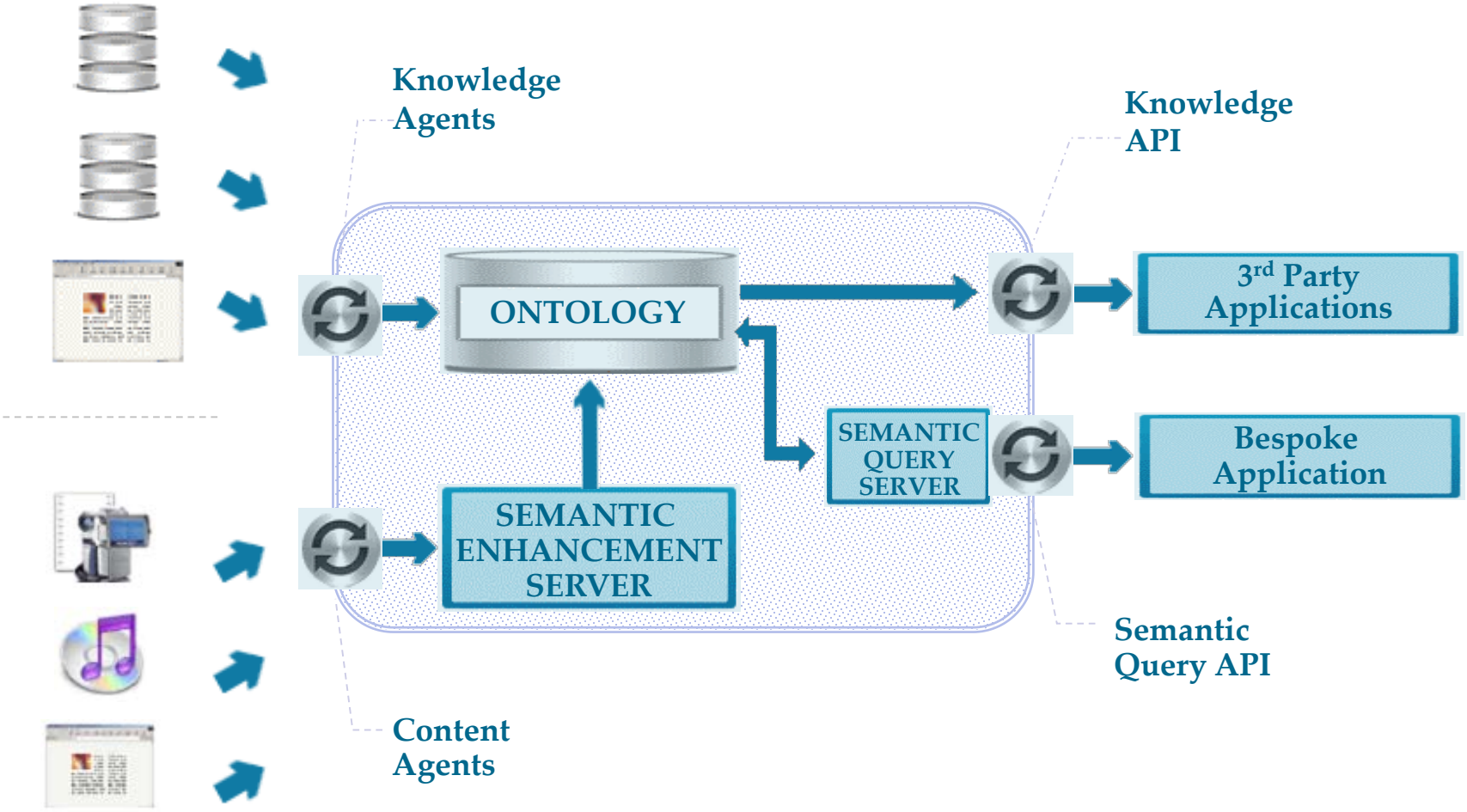
Evaluation Testbed Ontology



RDF Description Base *wrt* to this schema is populated from 30+ sources

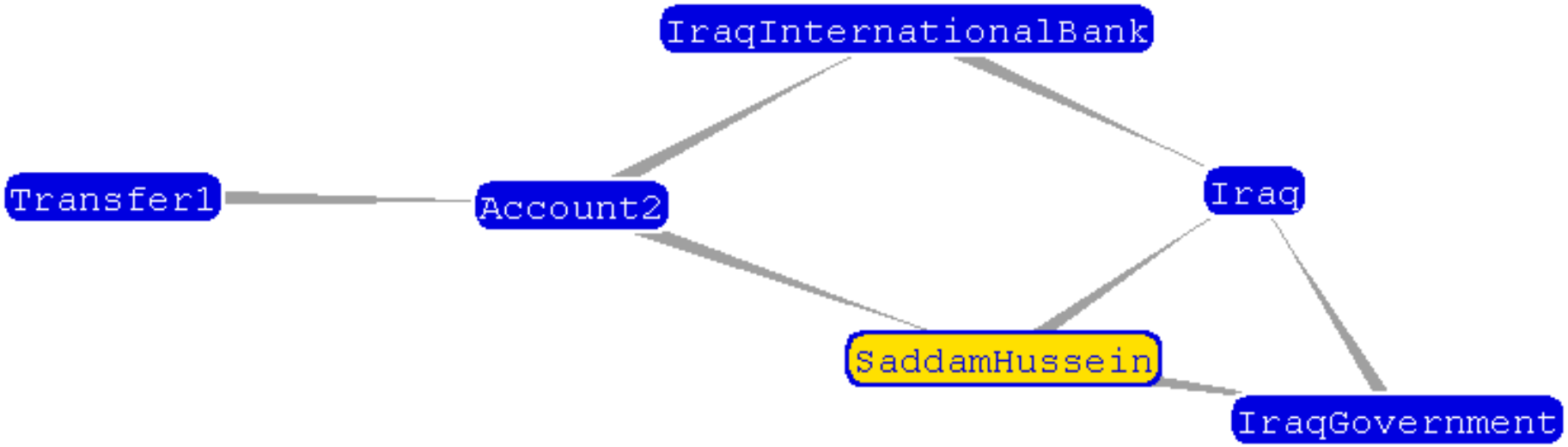
SEMAGIX

Use of Semagix Freedom for automatic ontology-driven metadata extraction to create large RDF description-base from many sources



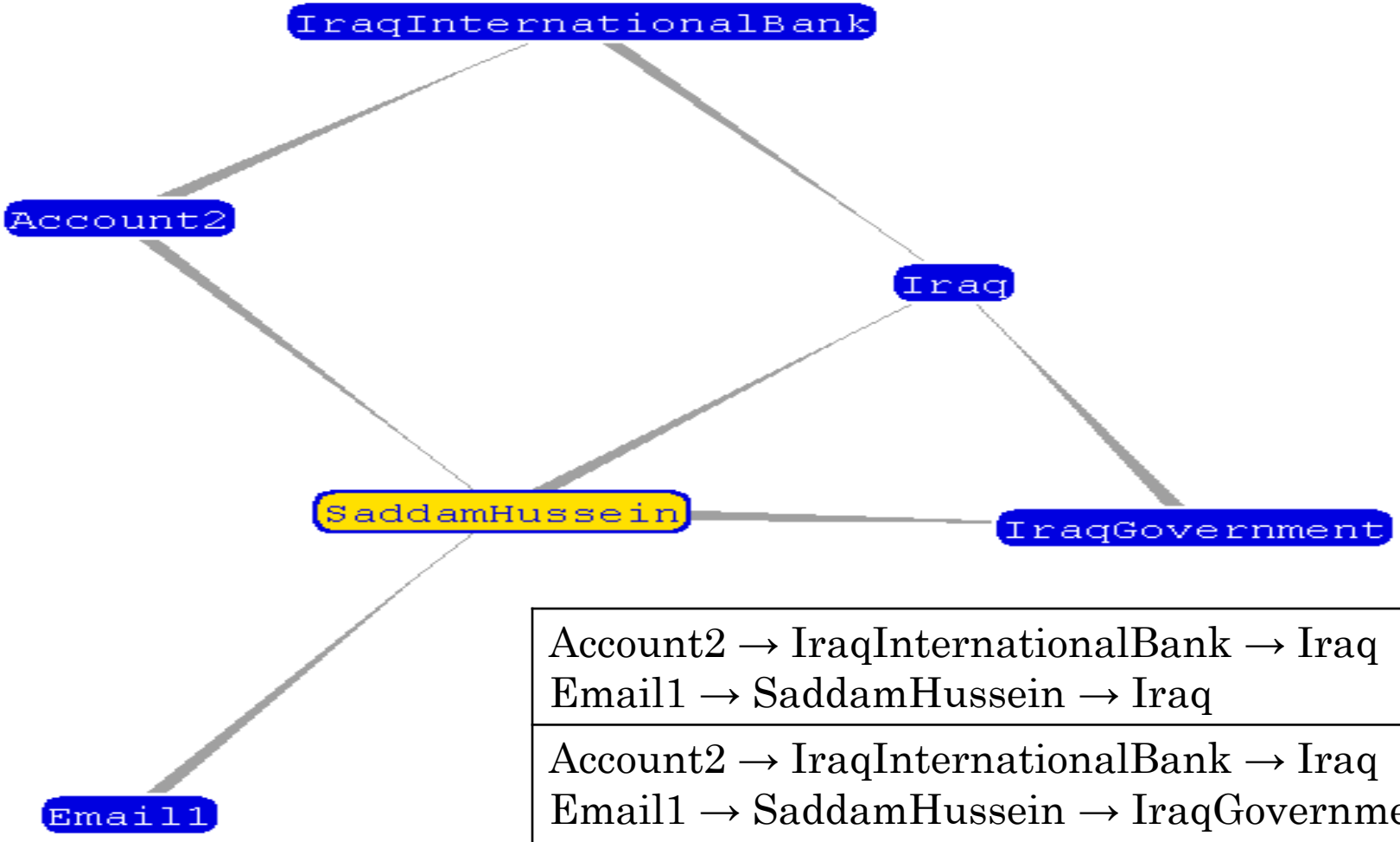
Semagix Freedom is based on prior research at the LSDIS Lab-> resulting SCORE technology.

ρ-PathAssociated(Transfer1, Iraq)



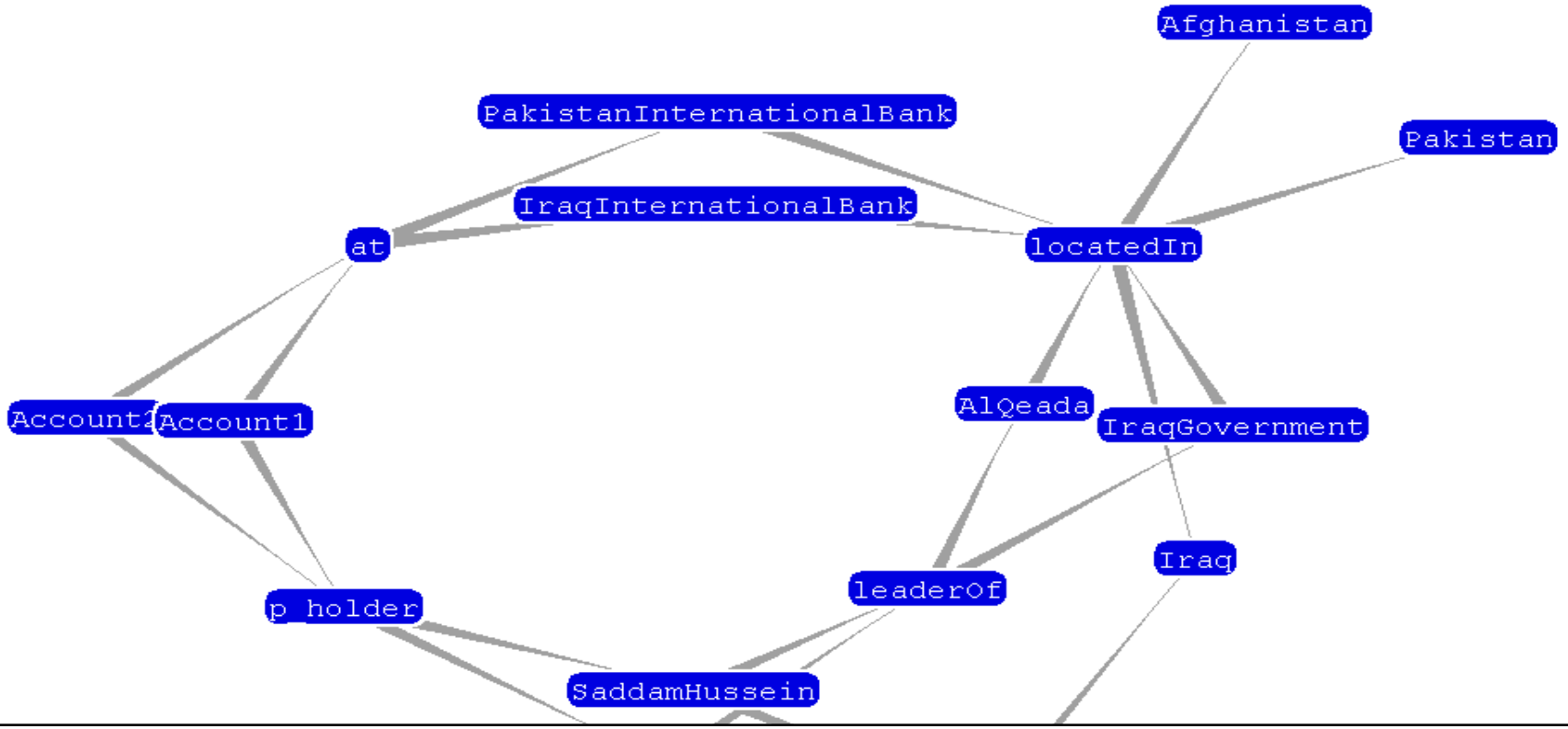
Transfer1 → Account2 → IraqInternationalBank → Iraq
Transfer1 → Account2 → SaddamHussein → Iraq
Transfer1 → Account2 → SaddamHussein → IraqGovernment → Iraq

ρ-joinAssociated(Account2, Email1)



Account2 → IraqInternationalBank → Iraq Email1 → SaddamHussein → Iraq
Account2 → IraqInternationalBank → Iraq Email1 → SaddamHussein → IraqGovernment → Iraq
Account2 → SaddamHussein Email1 → SaddamHussein

p-IsoAssociated(Account2, Account1)



Account2 → at → IraqInternationalBank → locatedIn → Iraq Account1 → at → PakistanInternationalBank → locatedIn → Pakistan
Account2 → p_holder → SaddamHussein → fromLocation → Iraq Account1 → p_holder → OsamaBinLaden → fromLocation → SaudiArabia
Account2 → p_holder→ SaddamHussein → leaderOf → IraqGovernment → locatedIn → Iraq Account1→ p_holder→ OsamaBinLaden → leaderOf → AlQeada → locatedIn →

Current & Future Work

- Data Preprocessing and Serialization
- Context
  - Specification & Representation
  - Streamline Query Processing
  - Ranking
- Query Processing Optimizations
  - Index structures
  - Heuristics
  - Complexity = $\Sigma_{(n-1)}^{(l=1)}$ (# paths of length l) (probability of keeping path of length l).
- Result Presentation
- Spatio-Temporal constraints

Related Work

- IR over XML, Relational Databases
  - [Hristidis et al 02,03], [Theobald et al 02],[Guha et al 03]
- Support for Path Expressions in Semi-Structured and Object-Oriented models
  - [Christophides et al 94], [Abiteboul et al 97], [Buneman et al 00], etc.
- Graph Databases
  - [Mendelzon, Wood 89]

More info.

- http://lsdis.cs.uga.edu/proj/SAI/
  - Project description, papers, presentations