Tracking RDF Graph Provenance using RDF Molecules

Li Ding and **Tim Finin** and **Yun Peng** and **Anupam Joshi**
Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore MD

Paulo Pinheiro da Silva and **Deborah L. McGuinness**
Knowledge Systems Laboratory, Stanford University, Stanford CA

1 Introduction

The Semantic Web can be thought of as one large "universal" RDF graph distributed across many Web pages. Since the graph is an unwieldy view, we usually work with online RDF documents. This is natural and appropriate for most tasks but still too coarse for tracking the provenance of an RDF graph, which requires finding knowledge sources supporting the target graph. Supporting facts are typically partial – i.e., a source contains only a sub-graph of the target.

The graph $G1$ in Figure 1 is partially supported by two sources, i.e., graphs $G2$ (containing $t3, t4, t5$ together) and $G3$ (containing $t1$). Tracking its provenance at a sub-graph level yields better "recall" because there may be no single *RDF document* or *Named graph* [Carroll *et al.*, 2004] which derives $G1$. Triple level simply fails when the target graph has blank nodes. For example, $G4$ will be wrongly thought to support $G1$ by containing $t3$. This is because a triple only preserves the *existential semantics* but ignores the consequence of triples being bounded by virtue of sharing the same blank node. None of these approaches can both find all supporting sources like $G2$, $G3$, and reject irrelevant sources like $G4$.

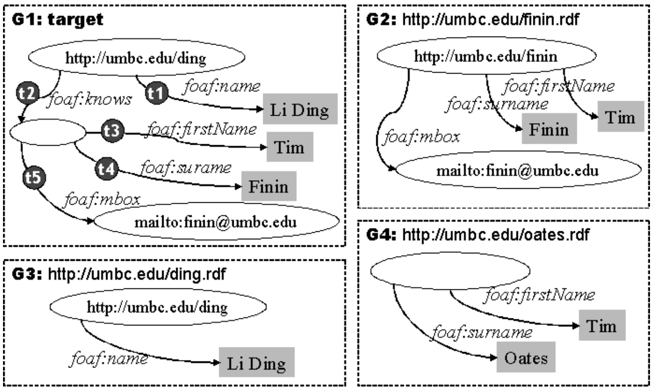

Figure 1: Target RDF graph $G1$ has five triples asserting that a thing with URI *http://umbc.edu/ding* and name *'Li Ding'* knows a thing with name *'Tim Finin'* and mbox *'mailto:finin@umbc.edu'*.

As shown in Figure 2, we define an intermediate decomposition for RDF graphs into sets of "molecules", each of which is a connected sub-graph of the original. The molecules are the "finest" in that they cannot be further decomposed without loss of information. The decomposition is "lossless" in that a graph's molecules can be recombined to yield the original graph (without introducing new triples) even if their blank nodes' IDs are "standardized apart".

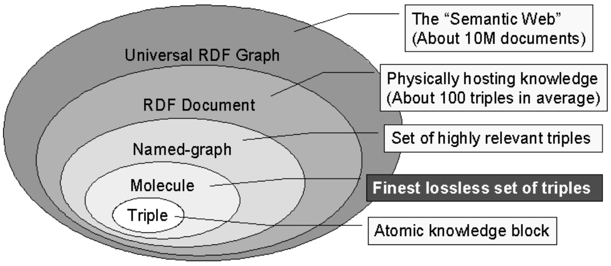
Universal RDF Graph
RDF Document
Named-graph
Molecule
Triple
The "Semantic Web" (About 10M documents)
Physically hosting knowledge (About 100 triples in average)
Set of highly relevant triples
Finest lossless set of triples
Atomic knowledge block

Figure 2: Various levels of granularity of the Semantic Web.

2 RDF Molecule and Lossless Decomposition

The semantics of blank nodes in RDF graphs has been studied in different application contexts, including F-logic inference [Yang and Kifer, 2003], signing entire graphs [Carroll, 2003] and minimal self-contained graphs [Tummarello *et al.*, 2005], tracking changes of graphs [Berners-Lee and Connolly, 2004] and definitions of resources [Stickler, 2005], and tracking knowledge provenance [Ding *et al.*, 2005]. Most approaches simply group triples which share the same blank node recursively, so as to preserve the existential and the binding semantics of blank node. Some also discusses the use of inverse functional properties and additional inference. We have formalized the definitions of a *lossless RDF graph decomposition* and an *RDF molecule* and have investigated three types of decomposition strategies. [1]

A **lossless RDF graph decomposition** has three elements (W, d, m): the background ontology W, the **decompose** operation $d(G, W)$ which breaks an RDF graph G into subgraphs $\hat{G} = \{G_1, G_2, ..., G_n\}$ using W, and the **merge** operation $m(\hat{G}, W)$ which combines all $\hat{G}$'s elements into a unified RDF graph G' using W. A lossless decomposition must

[1] Details is available in a technical report [Ding *et al.*, 2005].

satisfy that *for any RDF graph* G, $G = m(d(G, W), W)$. $\hat{G}$ is a **partition** of G if its elements are disjoint.

RDF molecules result from decomposing an RDF graph G into the finest, lossless sub-graphs according to a lossless decomposition (W, d, m). A sub-graph is *lossless* if it can be used to restore the original graph without introducing new triples, and it is the *finest* if it cannot be further decomposed into lossless sub-graphs.

A **Naive decomposition** decomposes an RDF graph without using any background ontologies. It is essentially computing connected components using only arcs connecting two blank nodes. It produces a partition with well-known time complexity – approximately O(V+E) for an RDF graph with V nodes and E arcs. This approach produces two molecules for G1 in Figure 1: (t1) and (t2,t3,t4,t5).

A **Functional decomposition** refines the result of a naive decomposition using functional dependencies asserted by the background ontologies. Inference which is supported by *owl:InverseFunctionalProperty* (IFP), *owl:FunctionalProperty* (FP), and OWL's same-as semantics can be used to label blank nodes with corresponding peers' URIs. Pre-inference in the background ontology can propagate them via *owl:inverseOf* and *rdfs:subPropertyOf*. For example, when *foaf:mbox* is declared as an IFP, this approach produces four molecules for G1 in Figure 1: (t1), (t2,t5), (t3,t5), and (t4,t5).

Heuristic decomposition studies blank nodes which can be uniquely identified by a set of properties acting like a 'key' in database literature. When *foaf:firstName* and *foaf:surname* together are used as a key according to the background ontologies, this approach produced three molecules for G1 in Figure 1: (t1), (t2,t3,t4), and (t3,t4,t5).

3 Current Status and Future Work

While the RDF molecule concept and the naive decomposition have been described independently by several researchers [Stickler, 2005; Tummarello *et al.*, 2005; Ding *et al.*, 2005], our formulation is more comprehensive. This work also differs from ontology partition [Grau *et al.*, 2005; Stuckenschmidt and Klein, 2004] in that it focus on finer decomposition dealing with the semantics of blank node but not the semantic dependencies among classes and properties.

We have implemented an RDF graph provenance service using the Swoogle [Ding *et al.*, 2004] search engine for tracking the provenance of integrated FOAF[2] profiles. It is motivated by the fact that provenance knowledge is usually needed before or after logical inference. Currently Swoogle has collected about 500K RDF documents from the Web and built a triple store with approximately 50M triples. For those RDF documents intended as ontologies, blank nodes are common due to the use of *owl:Restriction* and *owl:Union*. For example, the Inference Web ontology[3] contains 684 triples and decomposes into 349 one-triple molecules, and 78 molecules with four to eleven triples.

We also studied two specialized RDF collections (i.e. RSS files and FOAF files) that reveal interesting decomposition patterns. RSS files share a typical decomposition pattern – many one-triple molecules and only one multi-triple molecule, which is the instance of *rss:items* linking to a *rdf:sequence* of *rss:item* instances. FOAF files have various decomposition patterns since the FOAF ontology defined several IFPs. Some might worry about the complexity of enumerating all molecules; but it is necessary for 100% recall rate. Usually the number of generated molecules is less than the number of triples, and exceptions exist.

Our current work encompasses three areas: expanding the notion of decomposition to include heuristic grounding using Semantic Web compatible rule language like SWRL, exploring the utility of molecular decomposition for Semantic Web based hypothesis test, and integrating the molecular view into Inference Web [McGuinness and Pinheiro da Silva, 2004] to strengthen proofs using additional knowledge sources.

[2] see http://foaf-project.org

[3] See http://inferenceWeb.stanford.edu/2004/07/iw.owl.

References

[Berners-Lee and Connolly, 2004] Tim Berners-Lee and Dan Connolly. Delta: an ontology for the distribution of differences between rdf graphs. http://www.w3.org/DesignIssues/Diff, 2004.

[Carroll *et al.*, 2004] Jeremy J. Carroll, Christian Bizer, Patrick Hayes, and Patrick Stickler. Named graphs, provenance and trust. Technical Report HPL-2004-57, HP Lab, May 2004.

[Carroll, 2003] Jeremy J. Carroll. Signing RDF graphs. Technical Report HPL-2003-142, HP Lab, Jul 2003.

[Ding *et al.*, 2004] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C Doshi, , and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[Ding *et al.*, 2005] Li Ding, Tim Finin, Yun Peng, Paulo Pinheiro da Silva, and Deborah L. McGuinness. Tracking rdf graph provenance using rdf molecules. Technical Report TR CS-05-06, University of Maryland Baltimore County, April 2005.

[Grau *et al.*, 2005] Bernardo Cuenca Grau, Bijan Parsia, Evren Sirin, and Aditya Kalyanpur. Automatic partitioning of owl ontologies using e-connections. Technical report, University of Maryland Institute for Advanced Computer Studies, January 2005.

[McGuinness and Pinheiro da Silva, 2004] Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining answers from the semantic web: The inference web approach. *Journal of Web Semantics*, 1(4):397–413, October 2004.

[Stickler, 2005] Patrick Stickler. Cbd- concise bounded description. http://www.w3.org/Submission/CBD/, 2005.

[Stuckenschmidt and Klein, 2004] Heiner Stuckenschmidt and Michel C. A. Klein. Structure-based partitioning of large concept hierarchies. In *International Semantic Web Conference*, pages 289–303, 2004.

[Tummarello *et al.*, 2005] Giovanni Tummarello, Christian Morbidoni, Paolo Puliti, and Francesco Piazza. Signing individual fragments of an rdf graph. In *WWW (Special interest tracks and posters) 2005*, pages 1020–1021, 2005.

[Yang and Kifer, 2003] Guizhen Yang and Michael Kifer. Reasoning about anonymous resources and meta statements on the semantic web. *Journal on Data Semantics*, 1:69–97, 2003.