Provenance for Explaining Taxonomy Alignments

Mingmin Chen[1], Shizhuo Yu[1], Parisa Kianmajd[1], Nico Franz[2], Shawn Bowers[3], Bertram Ludäscher[1]

[1]University of California at Davis, [2] Arizona State University, [3]Gonzaga University

Motivation

- Logical derivations are a form of provenance.
- System-generated proofs are often overly complex or detailed; making sense of them is non-trivial.
- No provenance isn't an option either.

Approach

We study provenance in the application of Euler/X , a logic-based toolkit for aligning multiple biological taxonomies.
We propose a combination of approaches to explain

- logical inconsistencies in the input alignment,
- the derivation of new facts in the output taxonomies.

Running Example

Suppose we have two taxonomies, the first one has three concepts **a**, **b**, **c** with intra-taxonomy constraints:

r_1: b *isa* a
r_2: c *isa* a
r_3: a = b ∪ c
r_4: b ∩ c = ∅

The other taxonomy has three concepts **d**, **e**, **f** with intra-taxonomy constraints:

r_5: e *isa* d
r_6: f *isa* d
r_7: d = e ∪ f
r_8: e ∩ f = ∅

We have the following **articulations** (inter-taxonomy constraints):

A_1: a = d
A_2: b ⊊ e
A_3: c ⊊ f
A_4: b ⊊ d

We would like to apply all of the relations between concepts across two taxonomies, and output a merged taxonomy.

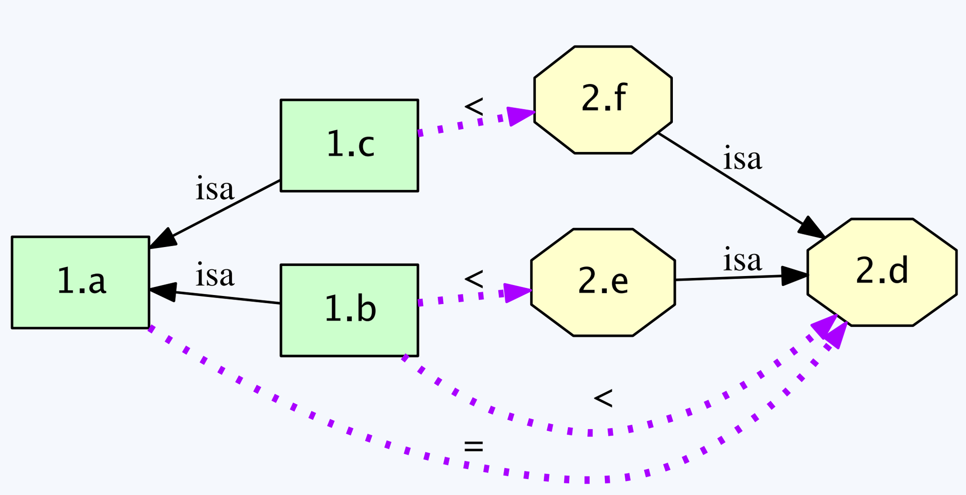

Fig1. Input Alignment

After running Euler/X on the input in Fig.1 we find the constraints to be unsatisfiable.

We thus perform a model-based diagnosis.

Inconsistency Explanation

The result lattice (Fig. 2) highlights minimal inconsistent subsets (MIS) and maximal consistent subsets (MCS).

The MIS {A1, A2, A3} indicates which articulations are inconsistent with T1 , T2 .

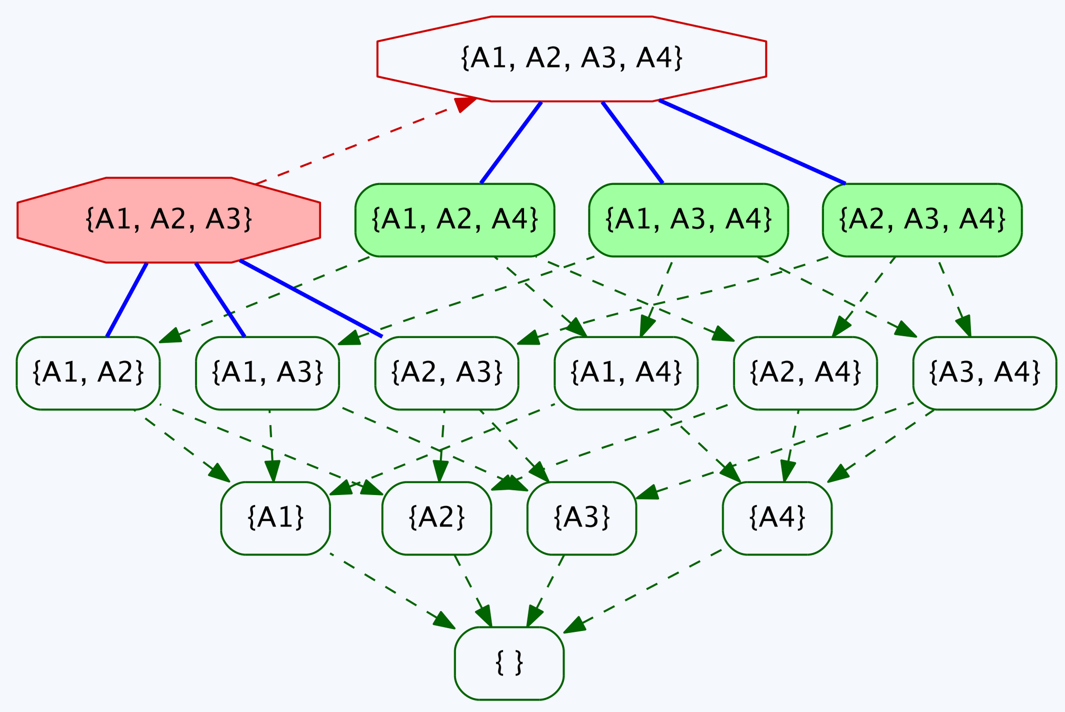

Fig.2 Diagnosis for A = {A1, . . . , A4}: solid red octagons and solid green boxes denote MIS and MCS, respectively. The (in)consistency of all other combinations are implied.

To further explore the inconsistency, the system-derived MCS can be employed: Fig. 3 shows the merged taxonomies (a.k.a. "possible worlds") obtained from the MCS.

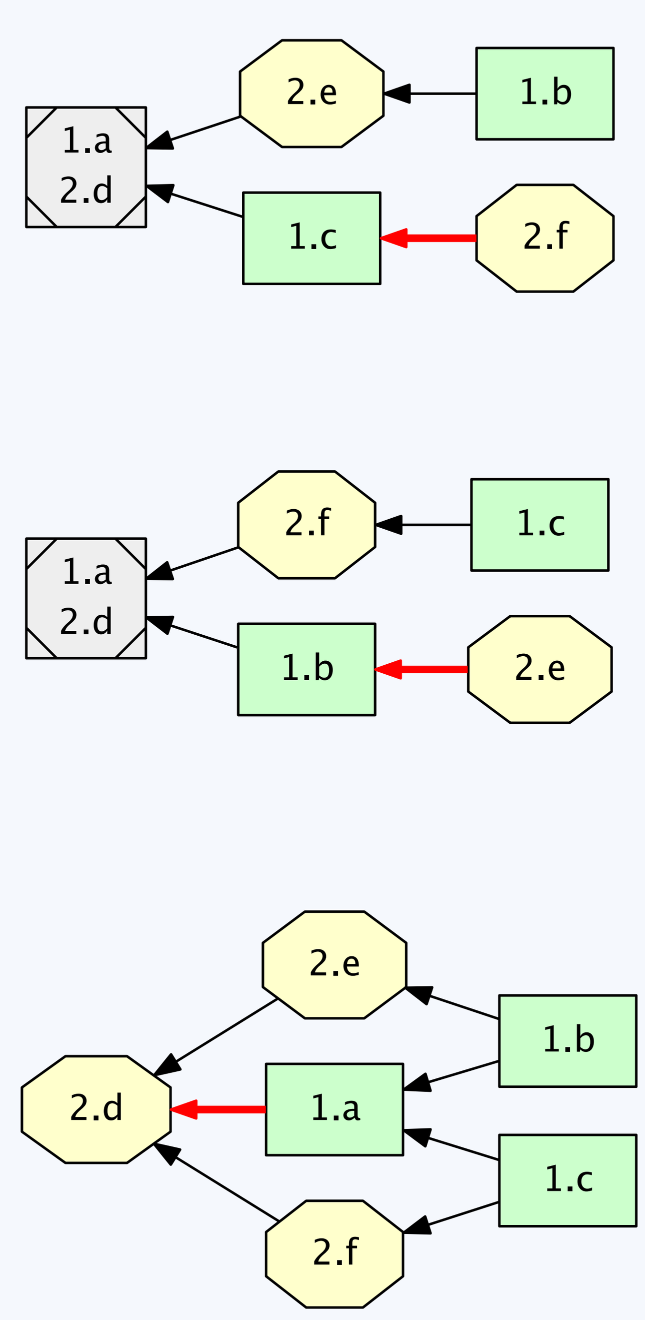

Fig. 3. Merged taxonomies (possible worlds) for MCS {A1,A2,A4}, {A1,A3,A4}, and {A2, A3, A4}. Grey boxes are fused concepts; bold, red edges represent inferred relations.

Acknowledgements

Work supported in part by NSF awards IIS-1118088 and DBI-1147273.

Resolving Inconsistency

- Using expert knowledge or further constraints, a preferred merge result can be selected to further analyze and then repair the inconsistency.
- Assume the user picks {A1,A2,A4}. From {A1,A2} it follows that f ⊊ c, but A3 states that c ⊊ f which is a contradiction.
- Now we need to explain why f ⊊ c is inferred.

Derivation Explanation

- To understand how f ⊊ c is inferred, we inspect its abstract logical derivation.
- We obtain this provenance in Euler/X by keeping track of the rules r1 , . . . , r8 and input alignments A1 , . . . , A4 used by the reasoner.

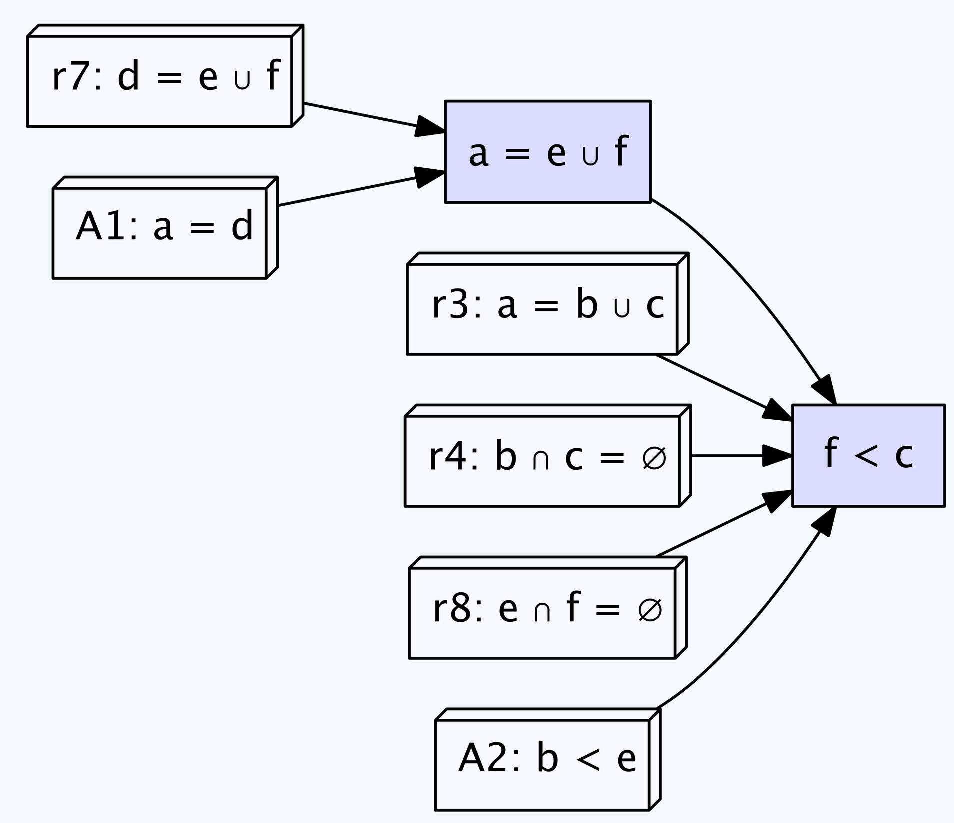

Fig. 4. Derivation Explanation

Related Work

Data provenance is an actively researched area and is closely related to proofs and derivations in logical reasoning.
Our inconsistency explanation is based on Reiter's model-based diagnosis, which has been studied extensively and applied to many areas, e.g., type error debugging, circuit diagnosis, OWL debugging, etc.
We have adapted the HST algorithm to compute all MIS and MCS for inconsistency explanation. The problem was shown to be TRANS-ENUM-complete by Eiter and Gottlob. Inspired by the ideas of a provenance semirings and Datalog debugging, our approach explains the derivation of the inferred relations.

Source Code

Scan to go to the Euler/X source code page: