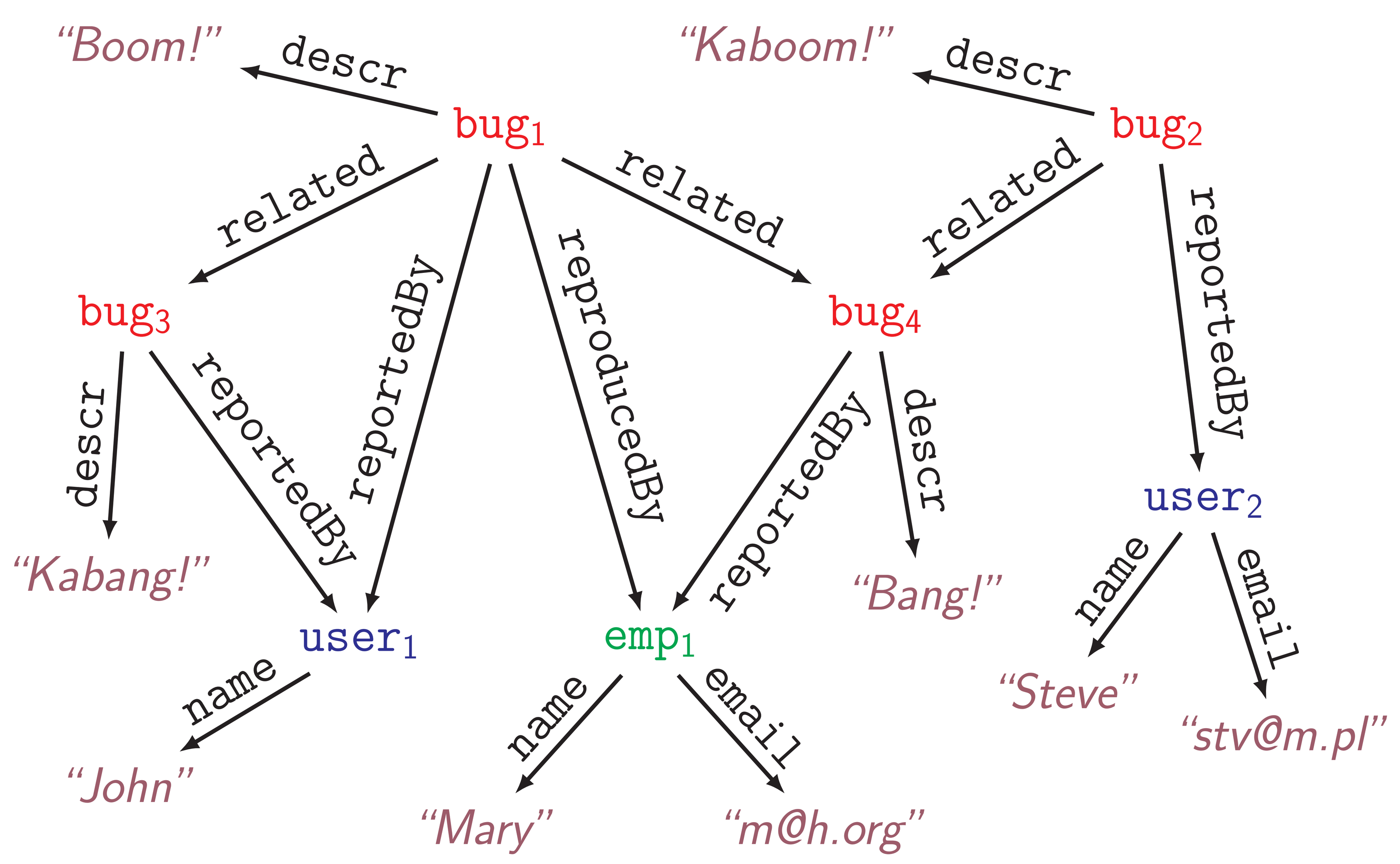
Containment of Shape Expression Schemas for RDF

Sławek Staworko, CRIStAL, INRIA LINKS, CNRS, University of Lille
Piotr Wieczorek, University of Wrocław

Université de Lille
Uniwersytet Wrocławski

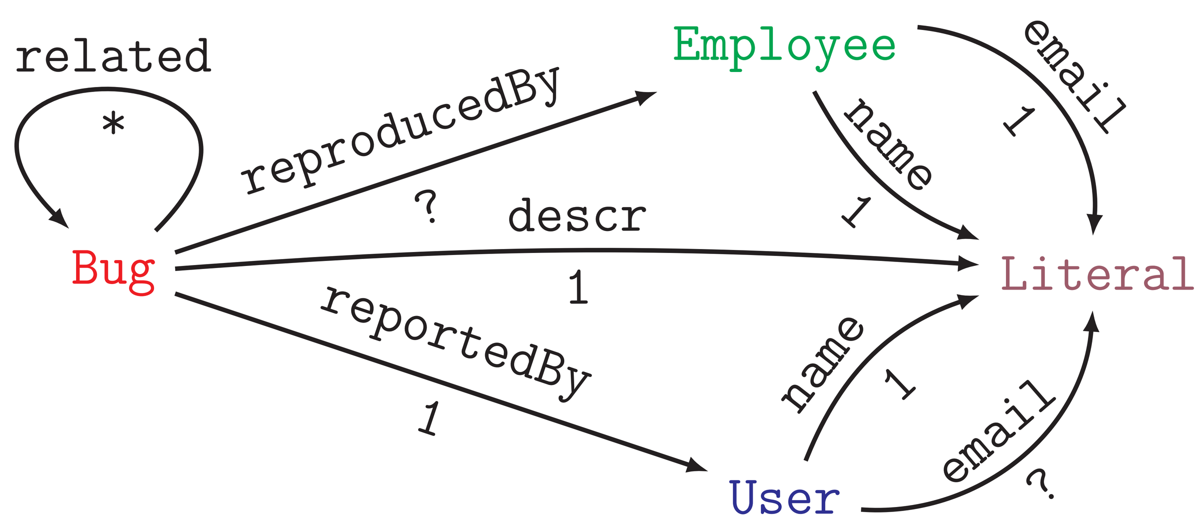
What is it all about?

Sample RDF Graph



Shape Graphs (a subclass of Shape Expression Schemas)

- A set of types is assigned for every node.
- For each type define the types and multiplicities (0, 1, ?, $*$) of outgoing neighbours.
- Every node has to satisfy the type definitions of all types assigned to the node and every node has to have at least one type.



- Simulation-style *embedding* guarantees containment for a pair of shape graphs.

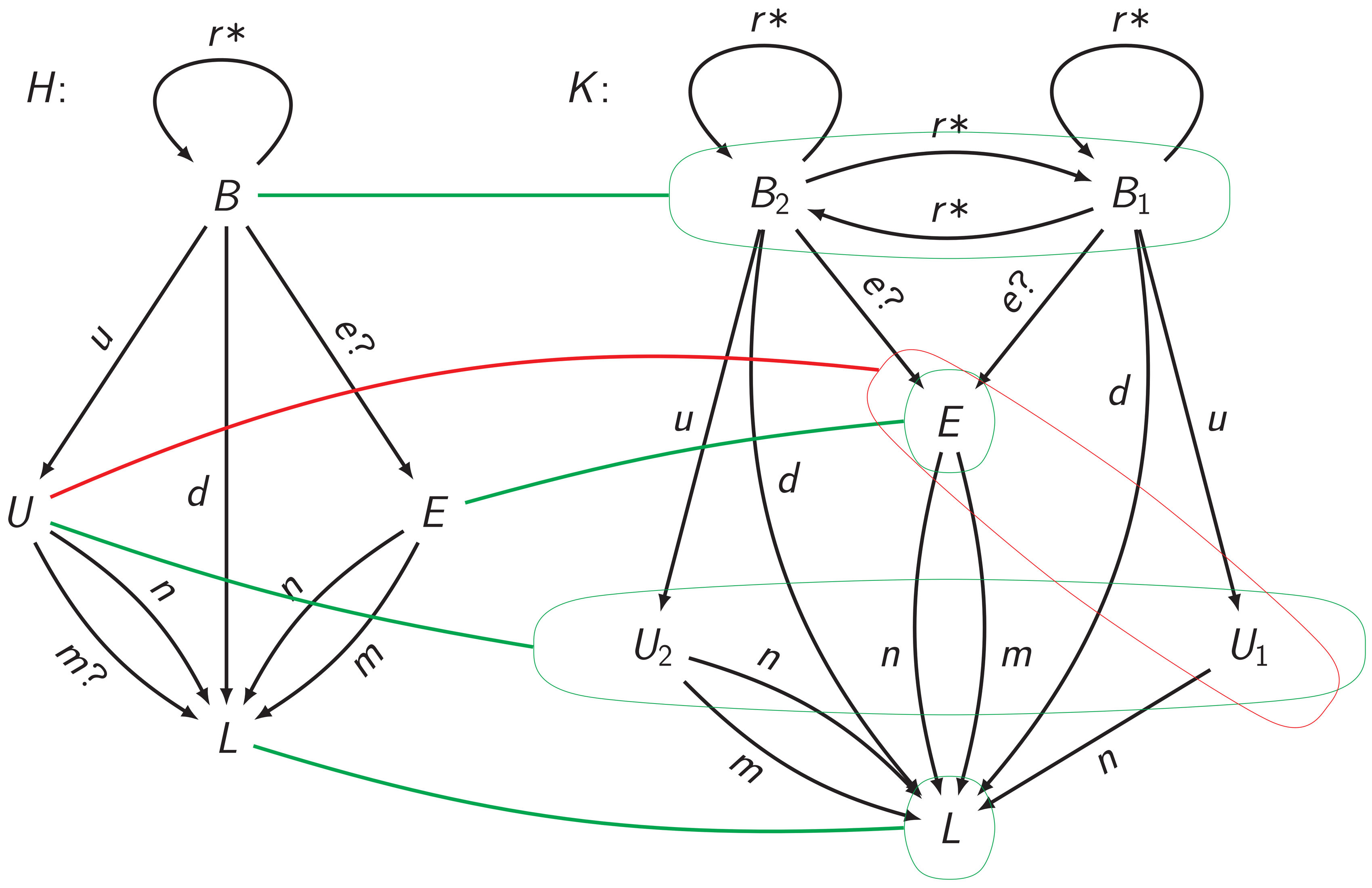
Theorem 1. Containment for Shape Graphs is EXPTIME-complete.

Deciding Containment of a Shape Graph H in a Shape Graph K.

- Containment is witnessed by the *covering*.
- Covering for every type of H identifies the set of types of K that capture the type of H e.g., on the example on the right, the type U is covered by $\{U_1, U_2\}$ and the type B by $\{B_1, B_2\}$
- Verify whether t of H is covered by a set S of types of K assuming some covering R, the *support*.
- Start with the full relation R of tuples of the from (a type of H, a set of types of K) and iteratively removes any (t, S) that is not supported by R.
- (t, S) is supported by R if the type definition of t can be unfolded into a disjunction contained in S with the use of unfolding operations based on some properties of RBE$_0$ e.g.,
  - $a{::}t^? \to \epsilon \mid a{::}t$ or $a{::}t^* \to \epsilon \mid (a{::}t \parallel a{::}t^*)$

  or based on R e.g., for every $(t, \{s_1, \ldots, s_m\})$
  - $a{::}t^? \to a{::}s_1^? \mid \ldots \mid a{::}s_m^?$
  - $a{::}t^* \to (a{::}s_1^* \parallel \ldots \parallel a{::}s_k^*) \mid (a{::}s_{k+1} \parallel a{::}t^*) \mid \ldots \mid (a{::}s_m \parallel a{::}t^*)$ $(0 \leq k \leq m)$

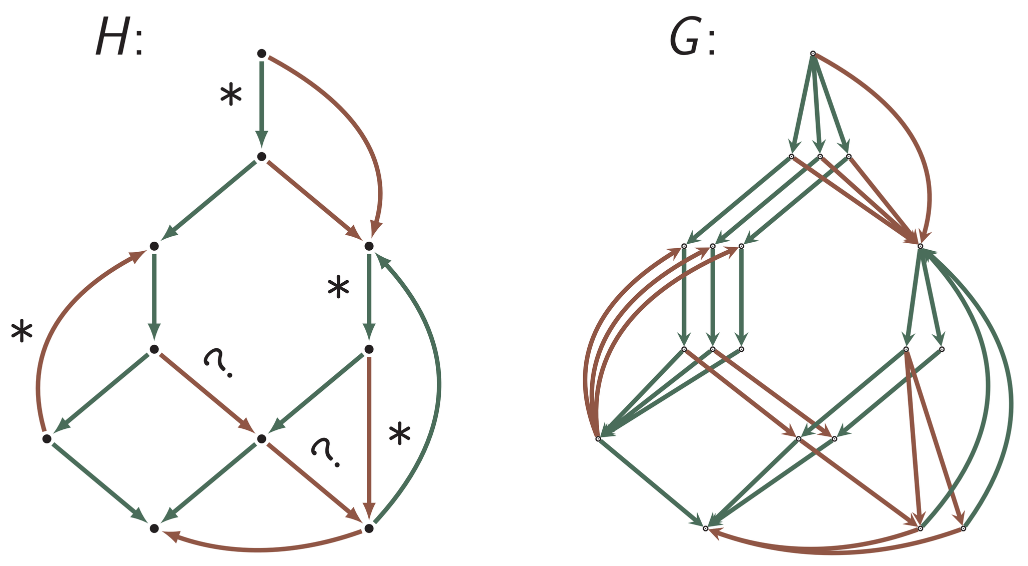
Covering



Theorem 2. Containment for Deterministic Shape Graphs is in PTIME.

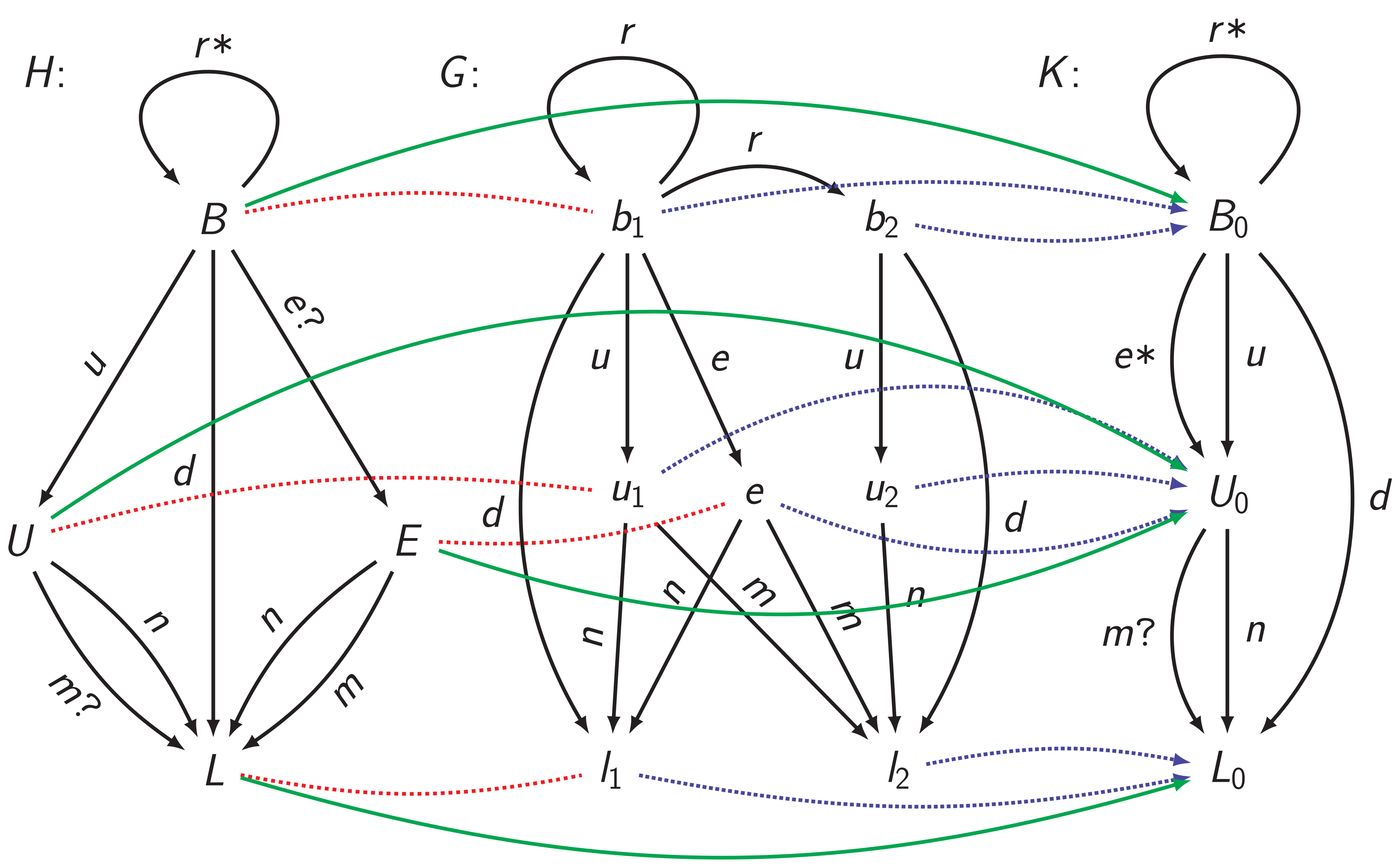
Deterministic Shape Expressions

- For every node n and every label a, n has at most one outgoing edge labeled with a; 1, $*$ are OK, ? restricted, no +.
- Embedding is a necessary and sufficient condition for containment.
- Main lemma: for each schema H there exists a *characteristic graph* $G \in L(H)$ which is a simple graph of polynomial size such that for any K we have that if there is an embedding from G to K then there is an embedding from H to K.
- Characteristic graph (CG). Different colors denote different labels.



- In essence, a number of nodes of type t serve the purpose of characterizing t.

Embedding of H in K based on the embedding of CG G in K.



Theorem 3. Containment for Shape Expressions is in co2NEXP$^{\text{NP}}$ and coNEXP-hard.

slawomir.staworko@inria.fr, piotr.wieczorek@cs.uni.wroc.pl