Contrôle de version incertain dans l'édition collaborative ouverte de documents arborescents

M. Lamine BA, Talel Abdessalem & Pierre Senellart

BDA'13 – 22-25 October 2013 (Nantes, France)

http://dbweb.enst.fr/

Versioning on the Web is Uncertain (I)

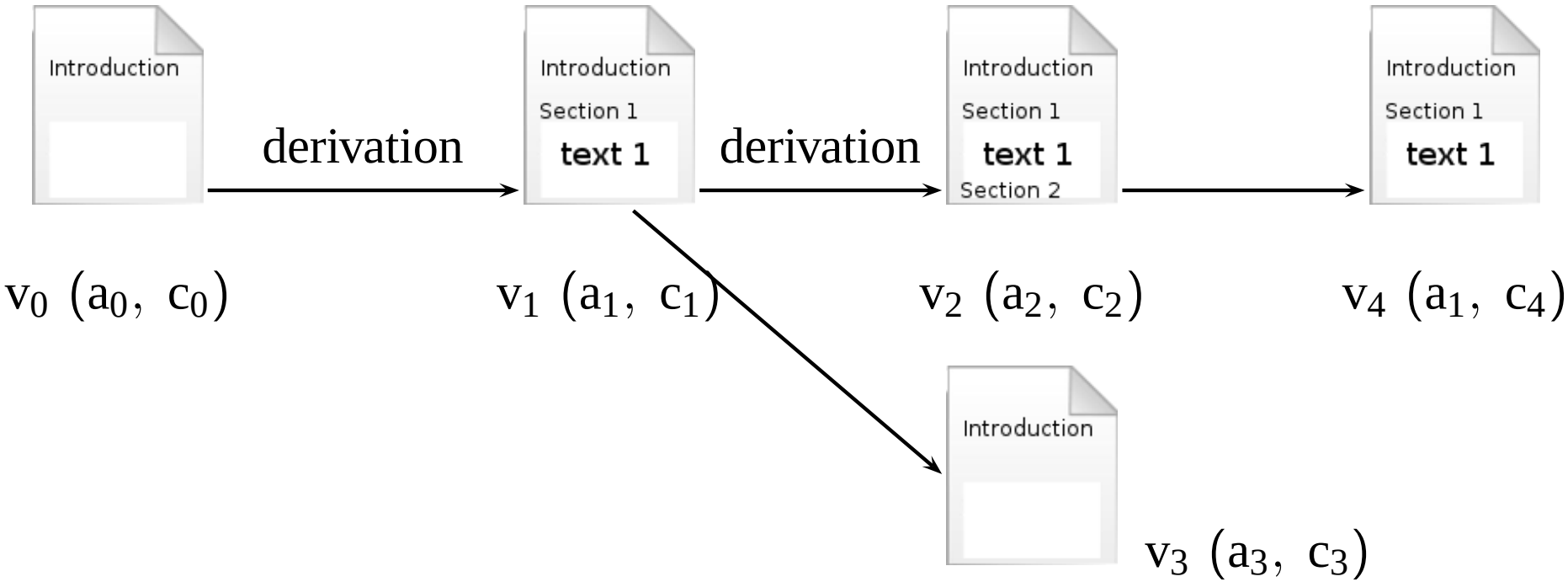
- Large-scale, open and collaborative editing platforms, e.g., wikipedia
  - **Unreliable contributors**, Novice vs. Experts, etc.
  - Malicious edits, Vandalism acts, **Contradictions**

Versioning on the Web is Uncertain (II)

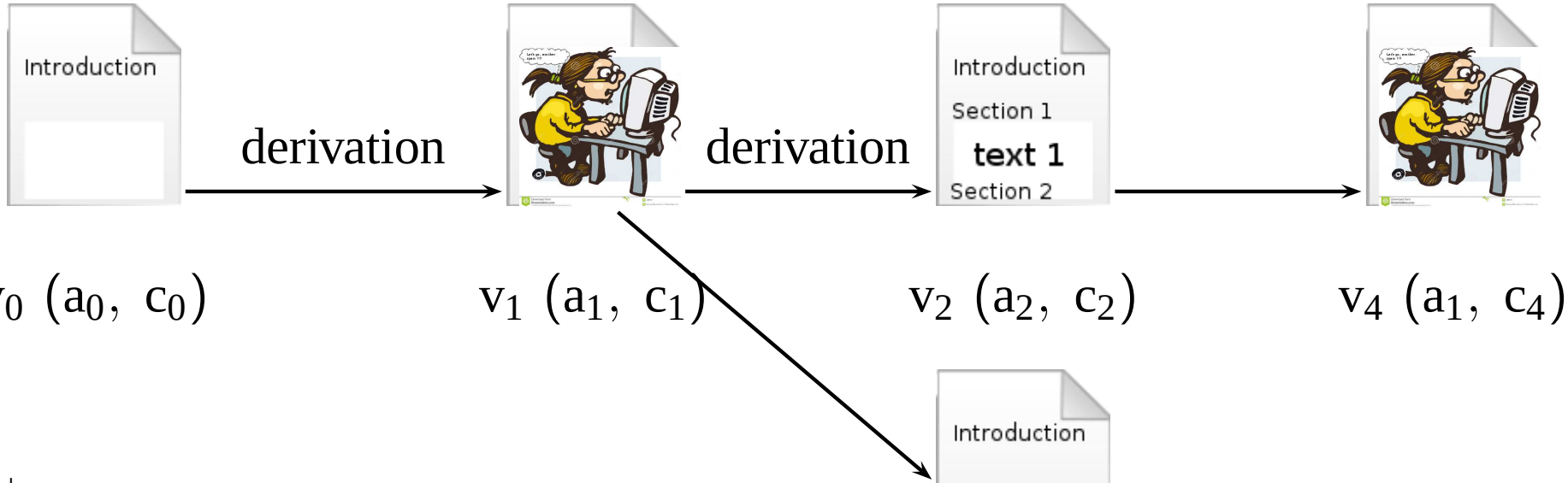
Version control is used in large-scale web collaborative editing platforms in order to **integrate** contributions from different sources and to support **fixing errors** with the possibility to query previous **data versions**

- No notion of more **relevant** versions or contributions which will fit to **user-preferences**, but just the concept of last **valid** revisions
- **Deterministic** version control models [Kerstin et al.(2009), Al Koc et al.(2011)] in the literature

Versioning on the Web is Uncertain (III)



Versioning on the Web is Uncertain (III)

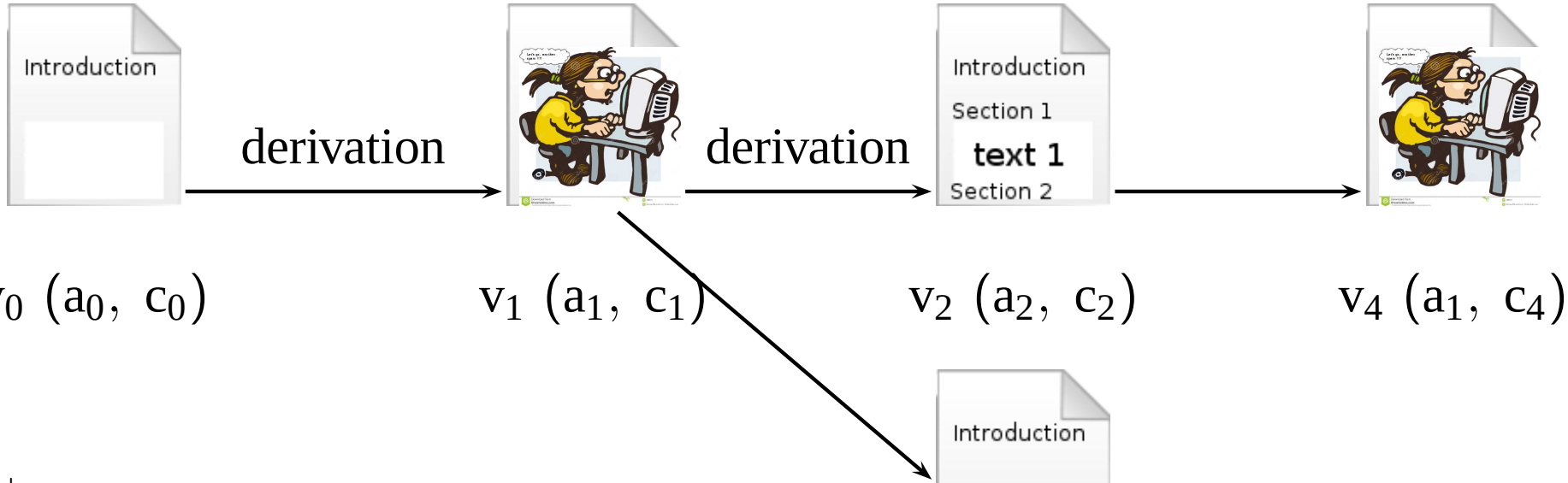


$Q_1 : v_2\ |\neg a_1$

$Q_2 : v_2\ |\neg c_1$

$Q_3 : \text{all } v_i |\ \Pr(v_i) \neq 0$

Versioning on the Web is Uncertain (III)



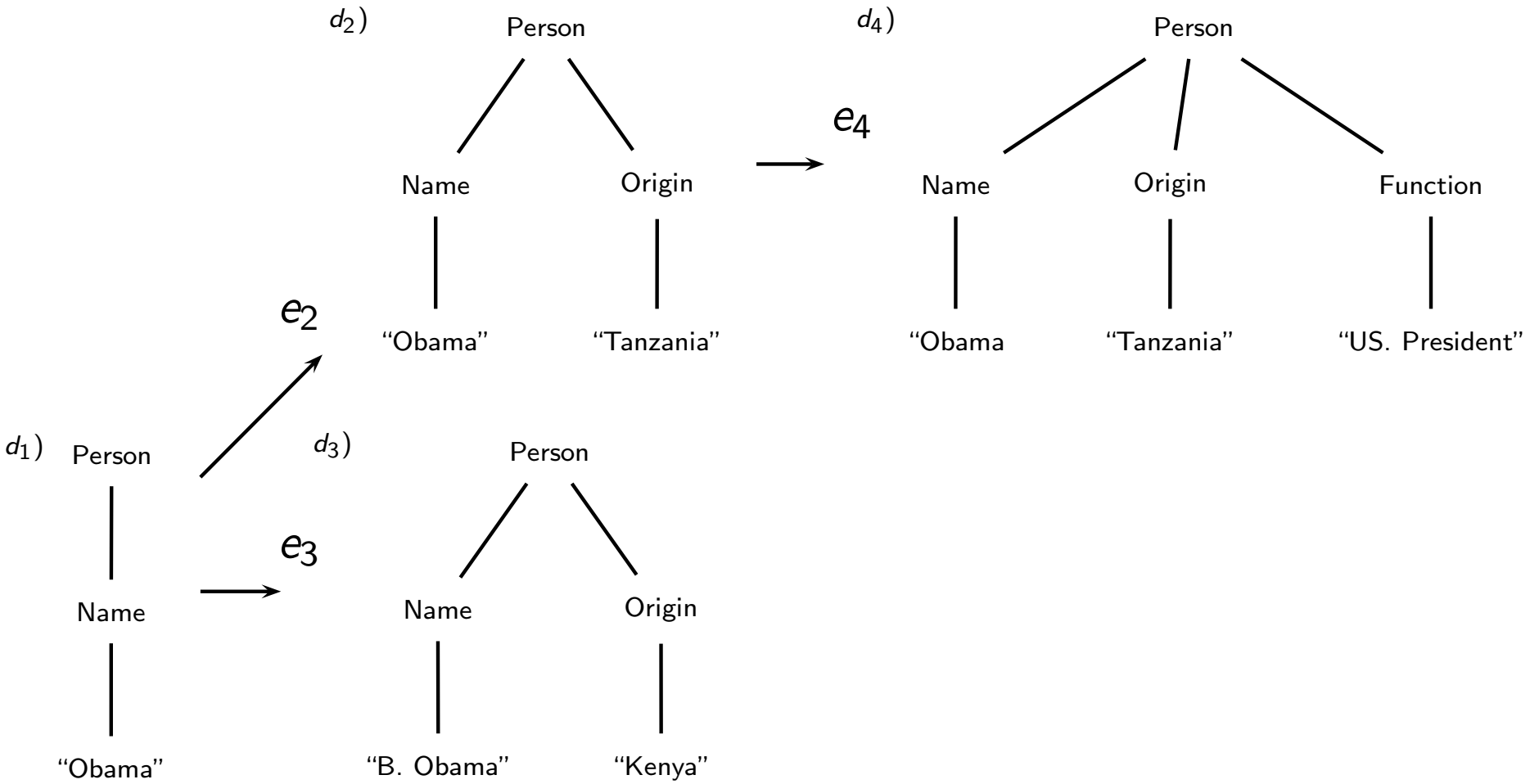
$Q_1 : v_2\ |\neg a_1$

$Q_2 : v_2\ |\neg c_1$

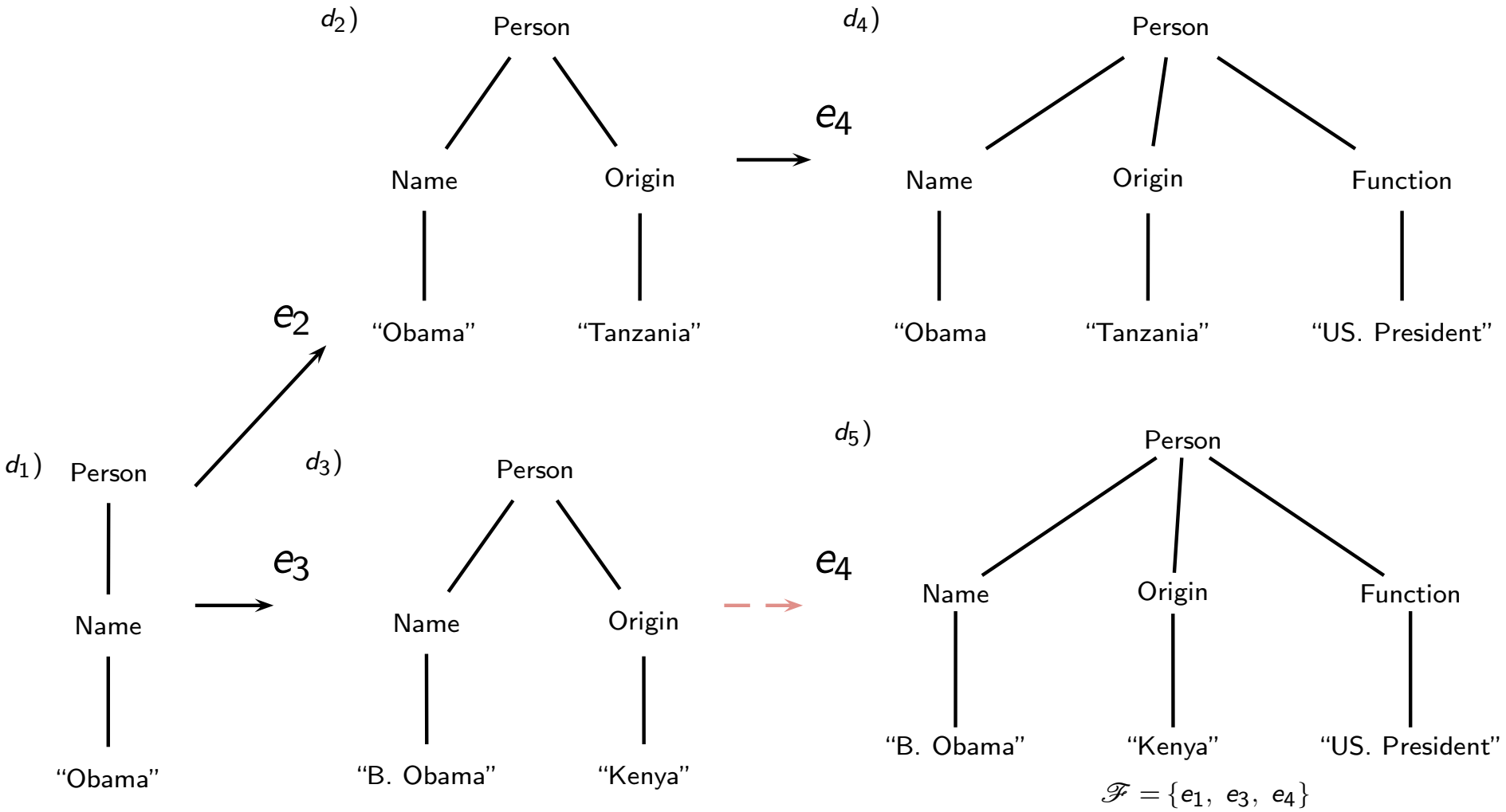
$Q_3 : \text{all } v_i |\ \Pr(v_i) \neq 0$

☛ **Uncertain version control in Open Collaborative Contexts**

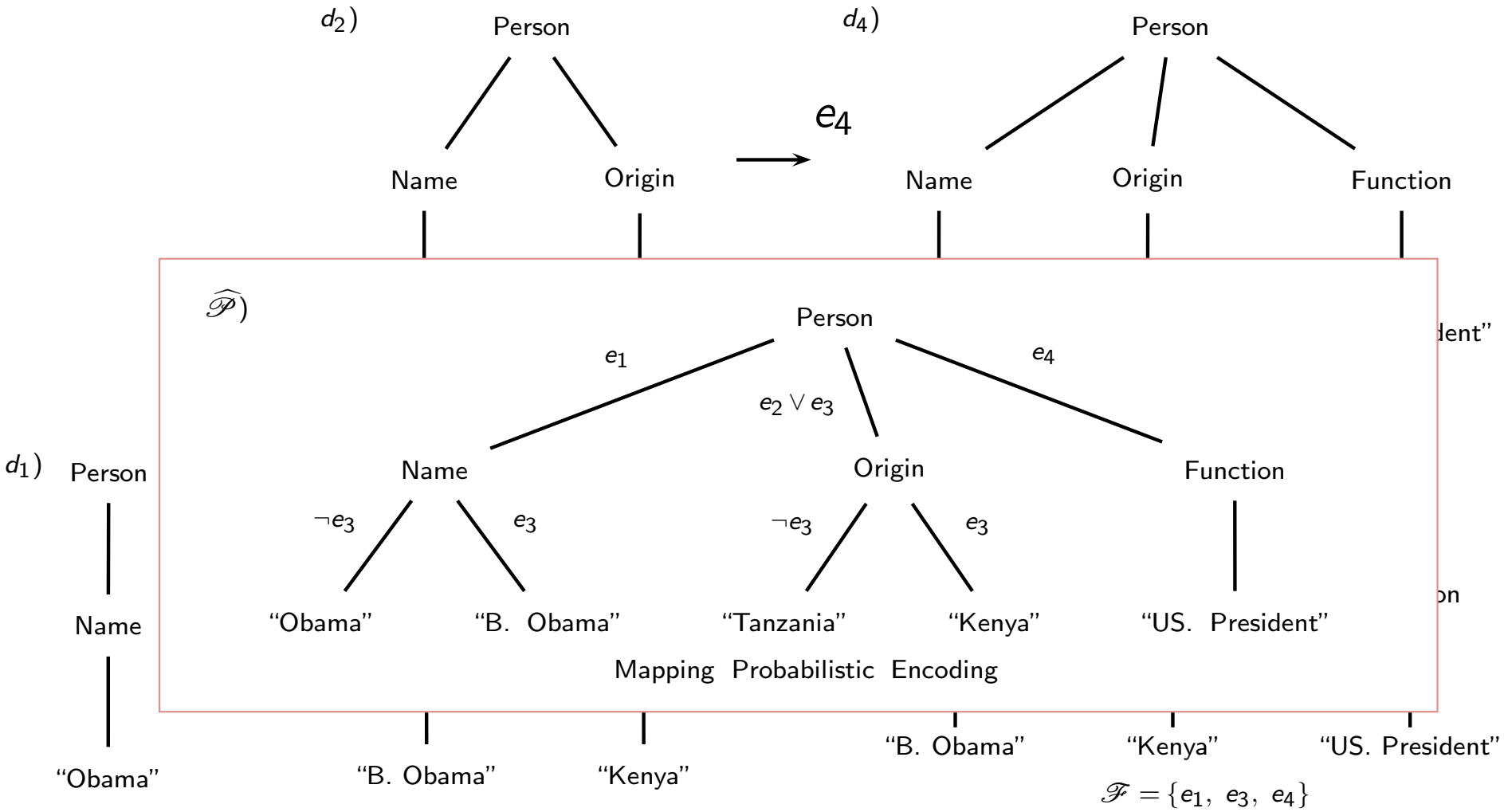
Versioning on the Web is Uncertain (IV)



Versioning on the Web is Uncertain (IV)



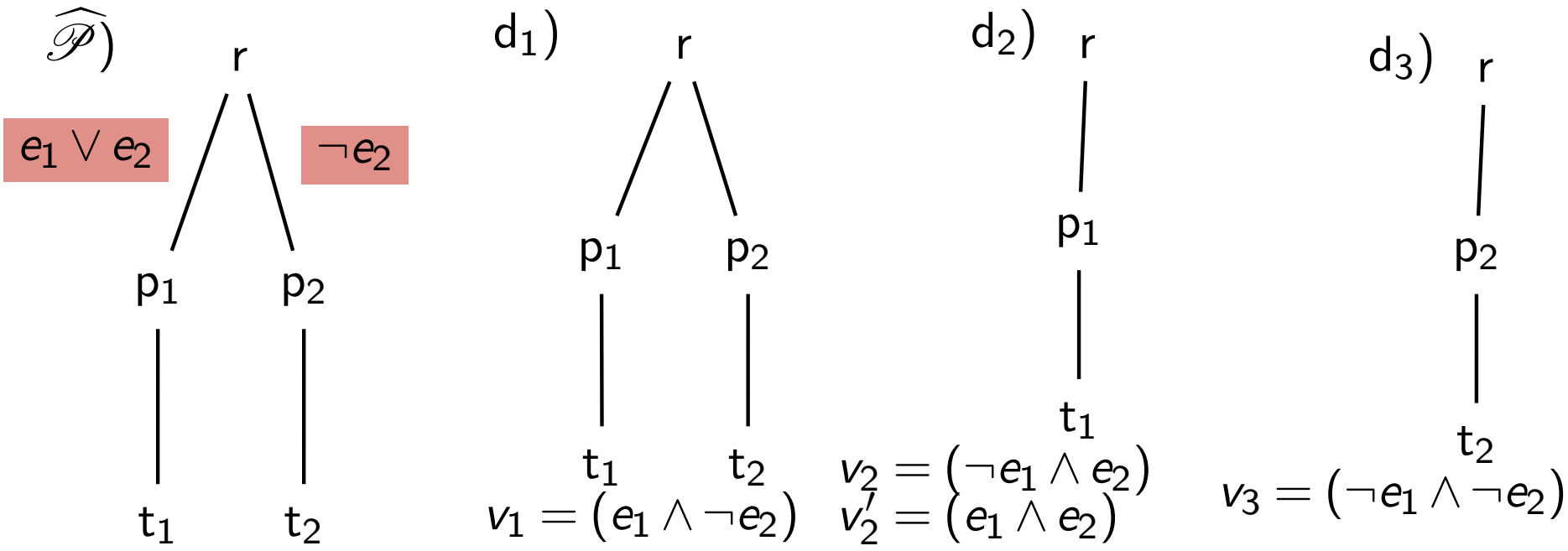
Versioning on the Web is Uncertain (IV)



Plan

Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]



$$\Pr(e_1) = 0.2;\ \Pr(e_2) = 0.8$$

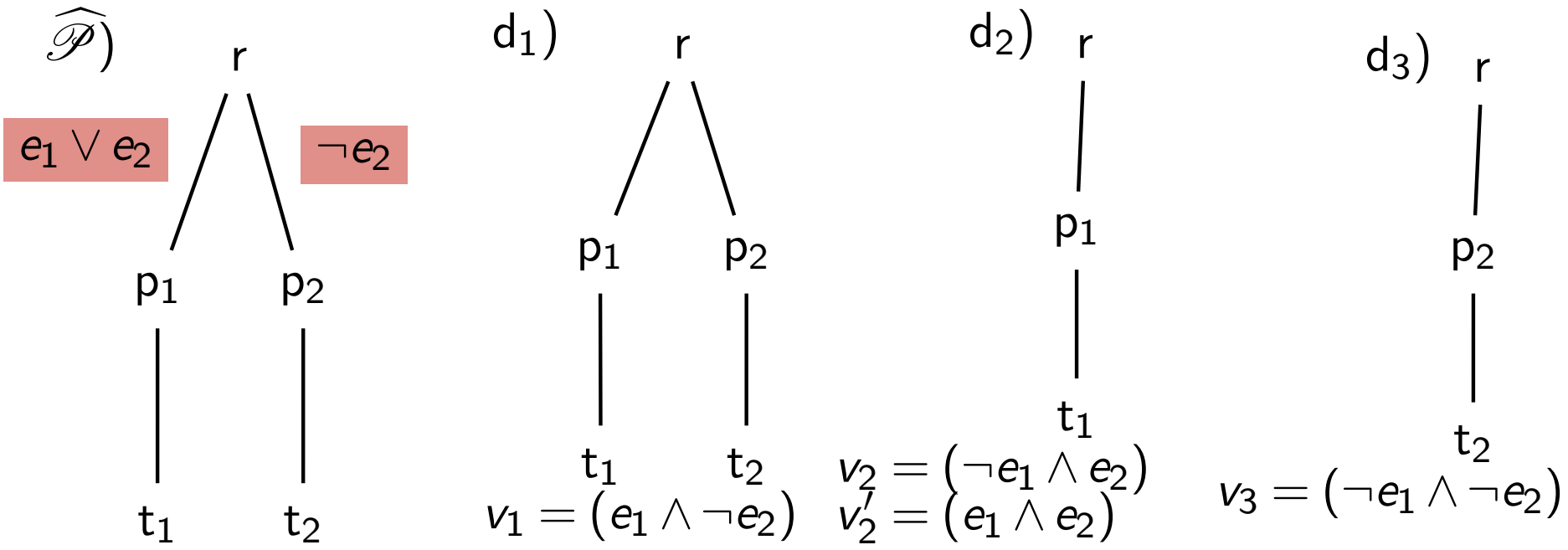
$$\Pr(d_1) = \Pr(e_1) \times \Pr(\neg e_2)$$

$$\Pr(d_2) = (\Pr(\neg e_1) \times \Pr(e_2)) + (\Pr(e_1) \times \Pr(e_2))$$

$$\Pr(d_3) = \Pr(\neg e_1) \times \Pr(\neg e_2)$$

Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]



$\Pr(e_1) = 0.2;\ \Pr(e_2) = 0.8$

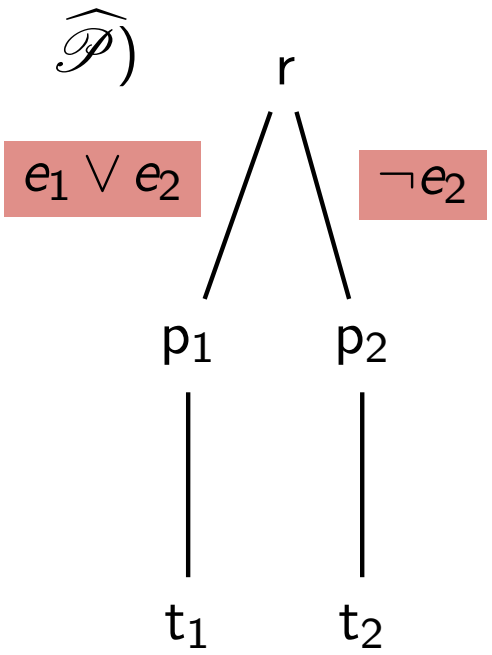
$\Pr(d_1) = 0.04$

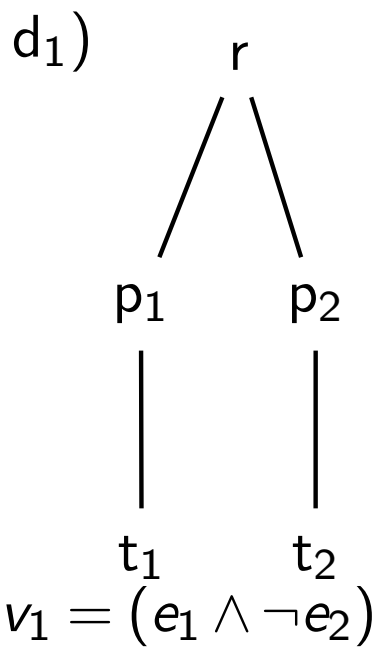
$\Pr(d_2) = 0.80$

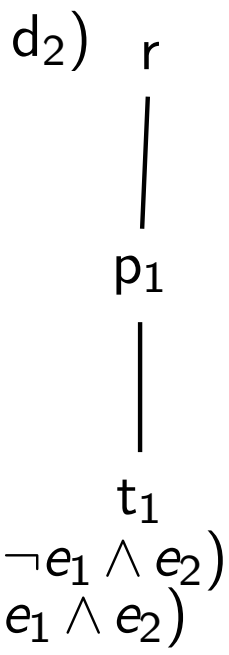
$\Pr(d_3) = 0.16$

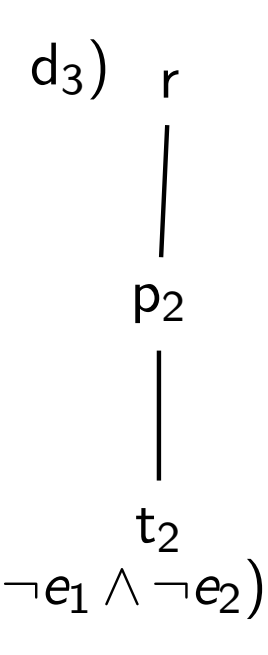
Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]









- $\widehat{\mathscr{P}} =< \mathscr{T}, C(E), fie, \Pr >$
- $[[\widehat{\mathscr{P}}]] =< \mathrm{D}, \Pr >;\ \Sigma\{\Pr(d) \mid d \in \mathrm{D}\} = 1$

Plan

Uncertain Multi-Version XML Document

Uncertain Version Control Model

- **Probability space** (PSV) over Uncertain versions of XML documents
- **Random** derivation graph (DG) over the document versions produced

Intuition: states in DG are complex variables e_i's called **version control events** based on simpler variables $b_1 \dots b_m$ managing uncertainty in data

$\mathscr{X} < \mathscr{G}, \Omega >$ defines a multi-version XML document with uncertain data

- $\mathscr{G}$ is DG over a set of versioning events $\mathscr{V} \cup e_0$ with $\mathscr{V} = \{e_1 \dots e_n\}$
- $\Omega : 2^{\mathscr{V} \setminus \{e_0\}} \rightarrow \mathscr{D}$ a mapping computing the PSV according to sets of valid events

Uncertain Multi-Version XML Document

Possible versions and Probabilities

Possible versions

- Version control events come with **edit scripts** updating content
- $\forall i, \forall \mathscr{F} \subseteq 2^{\mathscr{V} \setminus \{e_i\}}$, the possible version $\Omega(\{e_i\} \cup \mathscr{F}) = [\Omega(\mathscr{F})]^{\Delta_i}$

Probability of possible versions

- Assume a prior probability distribution over simple variables $b_1 \dots b_m$
- The probability of a given possible version $\Omega(\mathscr{F})$ is the probability of $\bigvee_{\mathscr{F} \subseteq \mathscr{V},\ \Omega(\mathscr{F})} \mathscr{F}$

Uncertain Multi-Version XML Document

Probabilistic XML Encoding

- **Compact** representation of all possible versions (PSV) in Ω mapping

Intuition: Represent PSV using **propositional formulas** of simple variables $b_1 \dots b_m$ attached to nodes in a global tree $\mathscr{T}$ containing all provided data

A probabilistic XML encoding of $<\mathscr{G}, \Omega>$ is a couple $<\mathscr{G}, \widehat{\mathscr{P}}>$ with

- $\widehat{\mathscr{P}} =< \mathscr{T}, C(\mathscr{V}), \mathit{fie}, \Pr >$ is a probabilistic XML document

Thm1: $[\![\widehat{\mathscr{P}}]\!]$ defines the same probability distribution over $\mathscr{D}$ as Ω, i.e.,

$$<\mathscr{G}, [\![\widehat{\mathscr{P}}]\!]>=<\mathscr{G}, \Omega>$$

Uncertain Multi-Version XML Document

Updating Uncertain Versions (I)

- **Uncertain** editions in Δ over nodes of a possible tree version

An update is an **uncertain** version control event defined based on a triple $< e_i, e_j, \Delta >$ ($e_i \in \mathscr{G}$ and $e_j \notin \mathscr{G}$)

$\texttt{updOP}(< e_i, e_j, \ \Delta >)[< \mathscr{G}, \Omega >]$

- $\mathscr{G} := \mathscr{G} \cup (\{e_j\}, \ \{(e_i, \ e_j)\})$
- Extension of Ω to a Ω' mapping with for each $\mathscr{F} \in 2^{(\mathscr{V} \setminus \{e_0\}) \cup \{e_j\}}$ $\Omega'(\mathscr{F}) = [\Omega(\mathscr{F} \setminus \{e_j\})]^{\Delta}$ if $e_j \in \mathscr{F}$ and $\Omega'(\mathscr{F}) = \Omega(\mathscr{F})$ otherwise

Uncertain Multi-Version XML Document

Updating Uncertain Versions (II)

$\texttt{updPrXML}(< e_i, e_j, \Delta >)[< \widehat{\mathscr{P}}, \Omega >]$

- $\mathscr{G} := \mathscr{G} \cup (\{e_j\},\ \{(e_i,\ e_j)\})$
- For all $\texttt{ins}(x,\ i)$ in Δ
  - $\texttt{Set}(x, fie(x) \vee (e_j))$ if x already in $\widehat{\mathscr{P}}$
  - Insert x in $\widehat{\mathscr{P}}$ and $\texttt{Set}(x, (e_j))$ otherwise
- For all $\texttt{del}(x)$ in Δ
  - $\texttt{Set}(x, fie(x) \wedge \neg(e_j))$

Thm2: $\texttt{updOP}(< e_i, e_j, \Delta >) \equiv \texttt{updPrXML}(< e_i, e_j, \Delta >)$

Thm3: updPrXML runs in $O(1)$ while $\widehat{\mathscr{P}}$ grows linearly according to $|\Delta|$

Uncertain Multi-version XML documents

Performance Analysis (Metrics, datasets and baseline)

- Estimation of the commit time and checkout cost of the model

Baseline Systems

☞ Versioning tools SubVersion and Git

– Use of their Java implementations based on the APIs SvnKit and JGit

Real Datasets

History of commits over two large file systems (shared tree-structured data)

☞ Linux kernel development

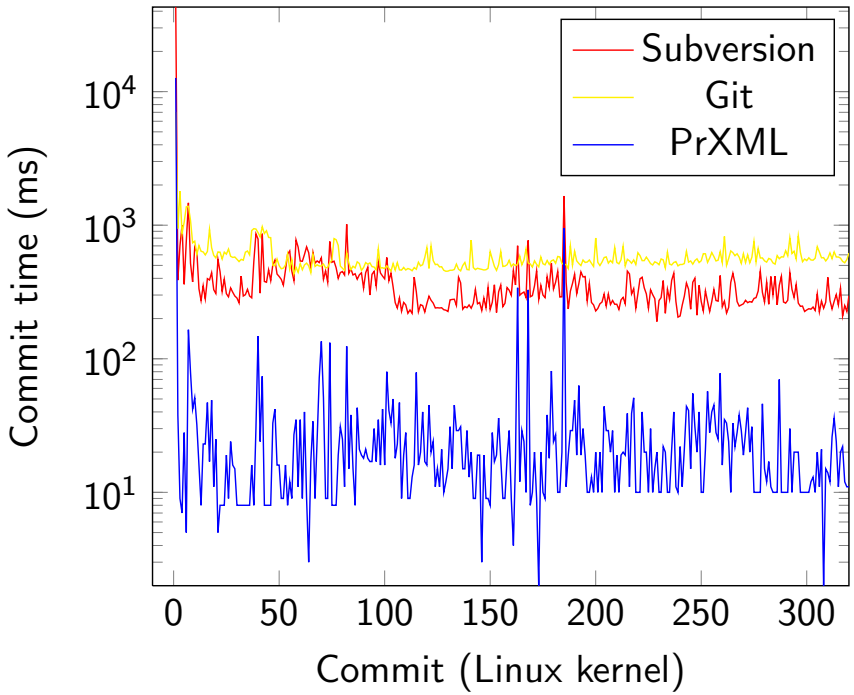
☞ Cassandra project

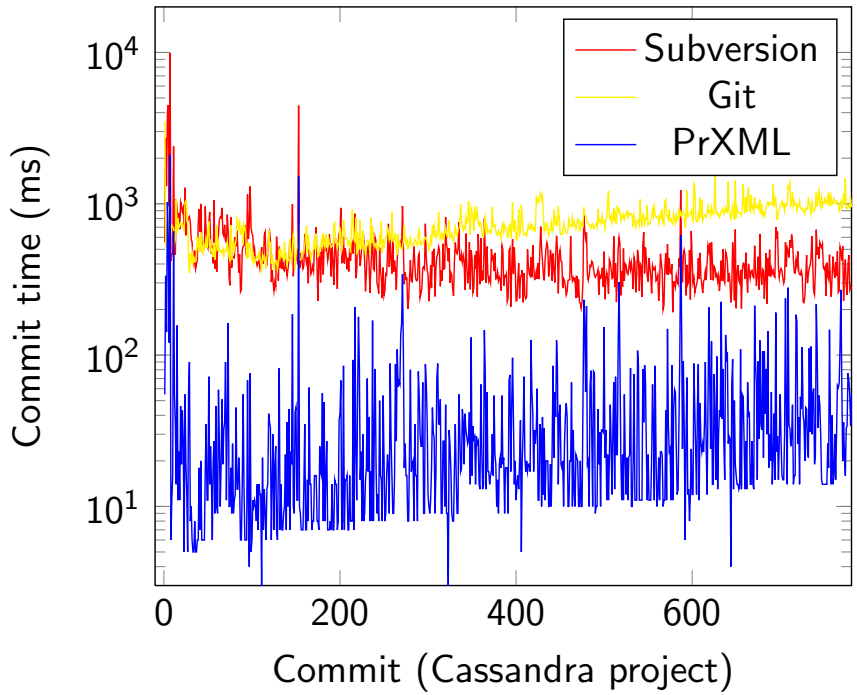
- Implementation of our system (PrXML) in Java

- Measures are obtained with all accesses in RAM Disk

Evaluation of the model

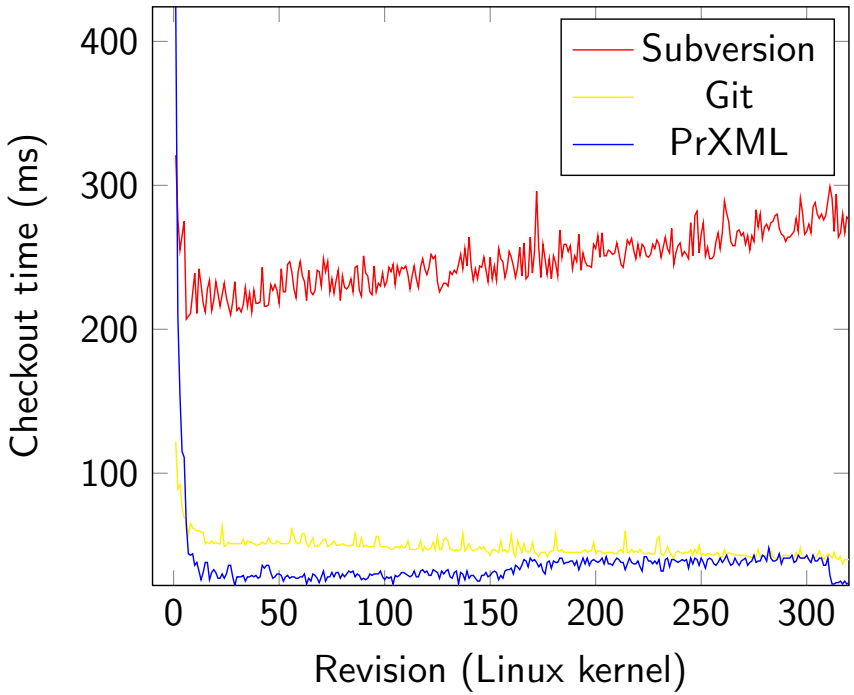
Performance Analysis (Commit time)

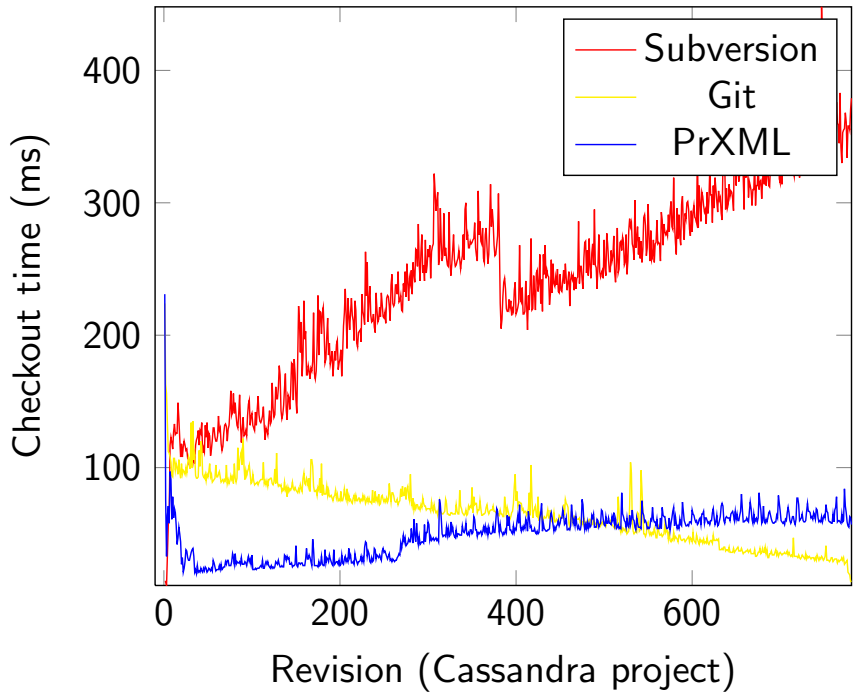




Evaluation of the model

Performance Analysis (Checkout time)





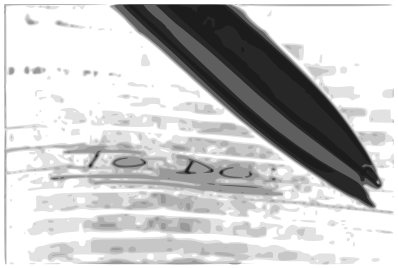
Plan

Conclusion and Further work (i)

- Design of a probabilistic version control approach for uncertain tree-structured documents
  - Both logical description and an efficient probabilistic compact XML encoding
  - Set-up of the most used version control operation, i.e., update and with an efficient mapping algorithms
  - Both theoretical and practical complexity analysis of the proposed model
- **Extension** of our model in [Ba et al.(DChanges, 2013)] with the **merging operation** over uncertain versions

Conclusion and Further work (ii)

- Support of more complex versioning operations such as copying, renaming etc.



- Study the impact of introducing some constraints over the order of nodes in uncertain version control

Merci!

References


Kerstin Altmanninger, Martina Seidl, and Manuel Wimmer, *A survey on model versioning approaches*, IJWIS **5** (2009).


M. Lamine Ba, Talel Abdessalem, and Pierre Senellart, *Towards a version control model with uncertain data*, PIKM, 2011.


_____, *Merging uncertain multi-version XML documents*, Proc. DChanges (Florence, Italy), sep 2013.


_____, *Uncertain version control in open collaborative editing of tree-structured documents*, Proc. DocEng (Florence, Italy), sep 2013.


Evgeny Kharlamov, Werner Nutt, and Pierre Senellart, *Updating Probabilistic XML*, Updates in XML, 2010.


Benny Kimelfeld and Pierre Senellart, *Probabilistic XML: Models and complexity*, Advances in Probabilistic Databases for Uncertain Information Management (Zongmin Ma and Li Yan, eds.), Springer-Verlag, 2013.


Al Koc and Abdullah Uz Tansel, *A survey of version control systems*, ICEME, 2011.