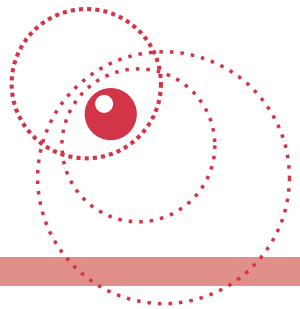


Uncertain Version Control in Open Collaborative Editing of Tree-Structured Documents

M. Lamine BA, Talel Abdessalem & Pierre Senellart

http://dbweb.enst.fr/

Licence de droits d'usage

Version Control of Uncertain Data

- **Data in large-scale, open and collaborative editing platforms, such as Wikipedia, are inherently uncertain**
  - contributors with different reliability, conflicts, malicous edits, . . .
  - need version control to maintain the quality of document versions
- Existing version control approaches are all deterministic
  - no room to uncertainty handling in the versioning process

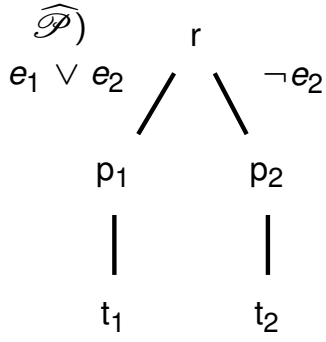
☛ A version control model aware of uncertain data may be helpful

Licence de droits d'usage

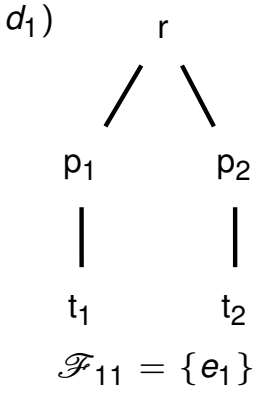
Uncertain Tree-Structured Data

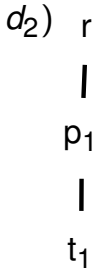
Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
  - annotations are **propositional formulas** of random Boolean variables



PrXML^{fie} p-Document





$\mathscr{F}_{11} = \{e_1\}$ $\mathscr{F}_{21} = \{e_2\}$ $\mathscr{F}_{22} = \{e_1, e_2\}$

$Pr(e_1) = 0.2$

$Pr(e_2) = 0.8$

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

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

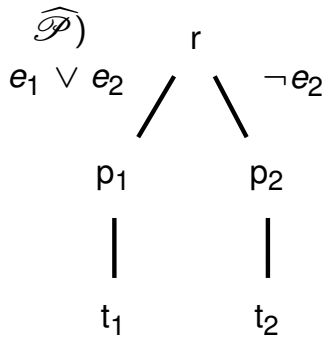
Possible worlds and their probabilities

Licence de droits d'usage

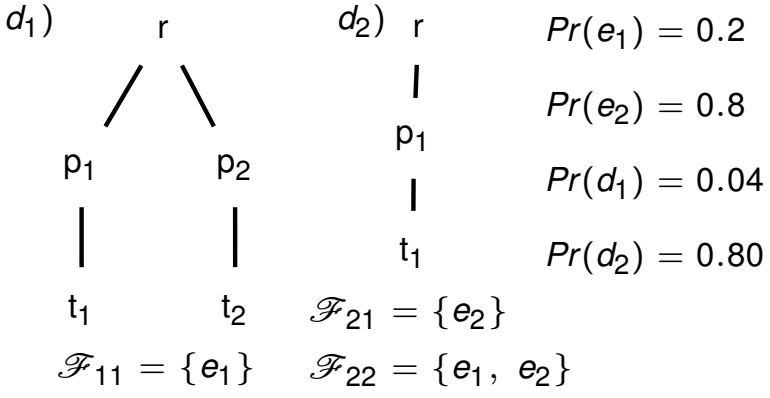
Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
  - annotations are **propositional formulas** of random Boolean variables



PrXML$^{\text{fie}}$ p-Document



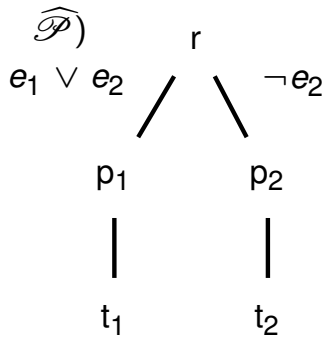
Possible worlds and their probabilities

Licence de droits d'usage

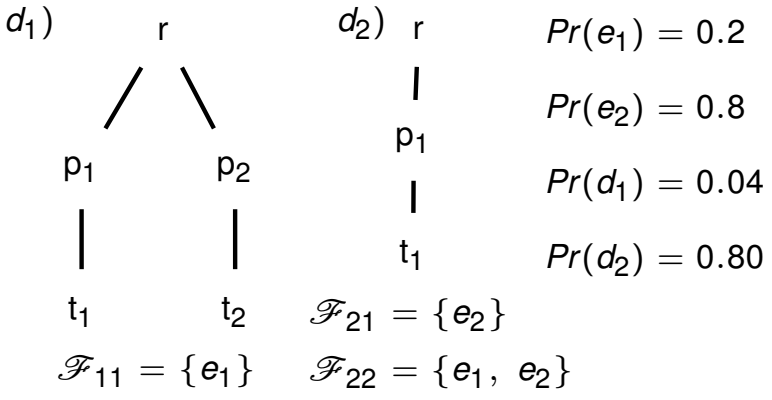
Uncertain Tree-Structured Data

Probabilistic XML [Kimelfeld & Senellart.(2013)]

- Unordered, unranked, and labeled **XML trees** with annotated edges
  - annotations are **propositional formulas** of random Boolean variables



PrXML$^{\text{fie}}$ p-Document



Possible worlds and their probabilities

- Enumerating all possible worlds and their probabilities
- Enable also to model uncertain updates on (uncertain) nodes [Kharlamov et al.(2010)]

☞ Integrate such a representation in a typical version control process

Licence de droits d'usage

Plan

Uncertain Multi-Version XML Document

- Uncertain Version Control Model
- Semantics of Updates

Evaluation of the model

- Performance Analysis
- Filtering capabilities

Licence de droits d'usage

Uncertain Multi-Version XML Document

Uncertain Version Control Model

Defines two equivalent views over any uncertain multi-version XML tree

- set $\mathscr{V}$ of random variables $e_0, e_1 \ldots e_n$ modeling the tree states
- infinite set $\mathscr{D}$ of all (unordered) XML trees including the versions

$(\mathscr{G}, \Omega)$: Logical View

- DAG $\mathscr{G}$ built on variables in $\mathscr{V}$
- Mapping $\Omega : 2^{\mathscr{V} \setminus \{e_0\}} \rightarrow \mathscr{D}$ which computes the possible versions according to sets of valid events

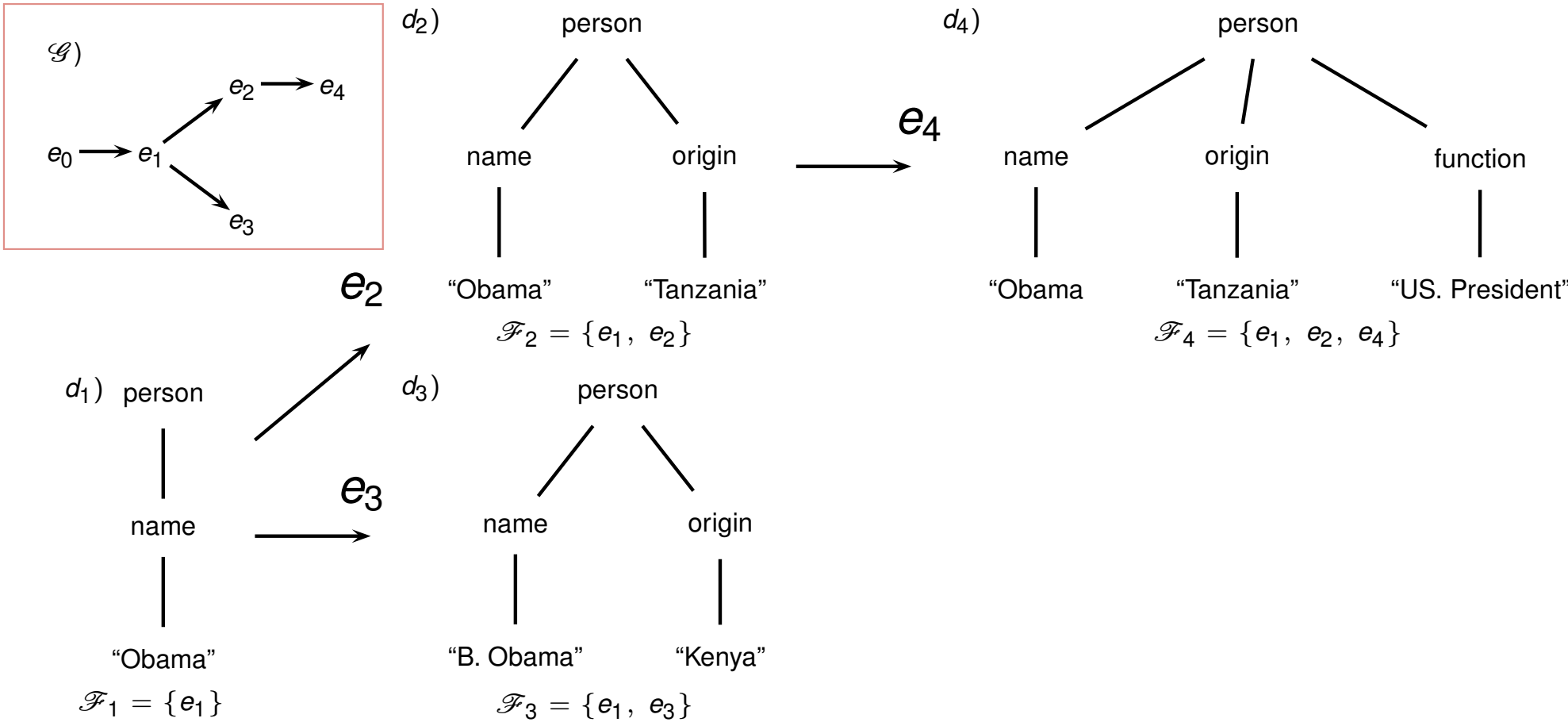
$(\mathscr{G}, \widehat{\mathscr{P}})$: Probabilistic XML Encoding

- Similar DAG $\mathscr{G}$ of random variables in $\mathscr{V}$
- Probabilistic XML tree $\widehat{\mathscr{P}}$ which defines the same probability distribution as Ω mapping

Licence de droits d'usage

Uncertain Multi-Version XML Document

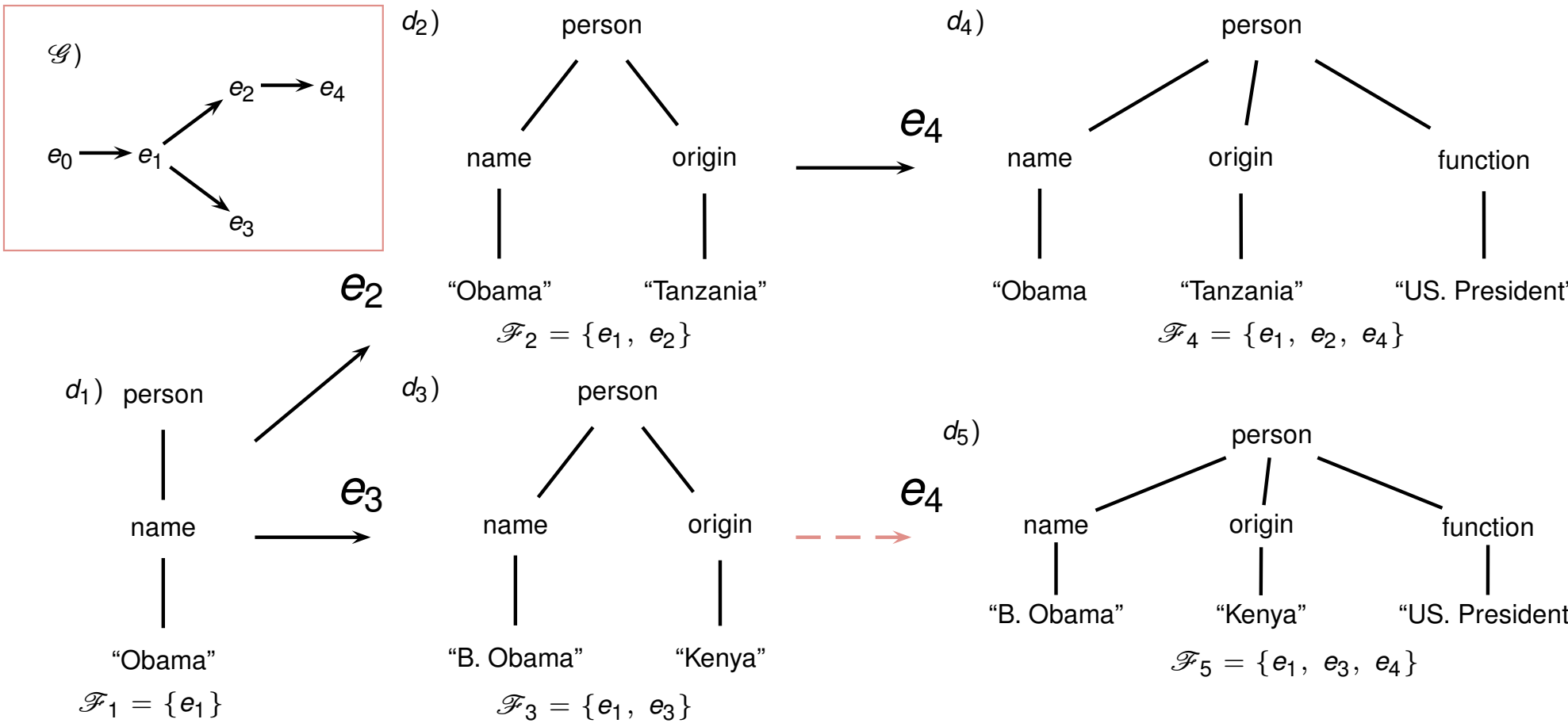
Uncertain Version Control Model (Example)



Licence de droits d'usage

Uncertain Multi-Version XML Document

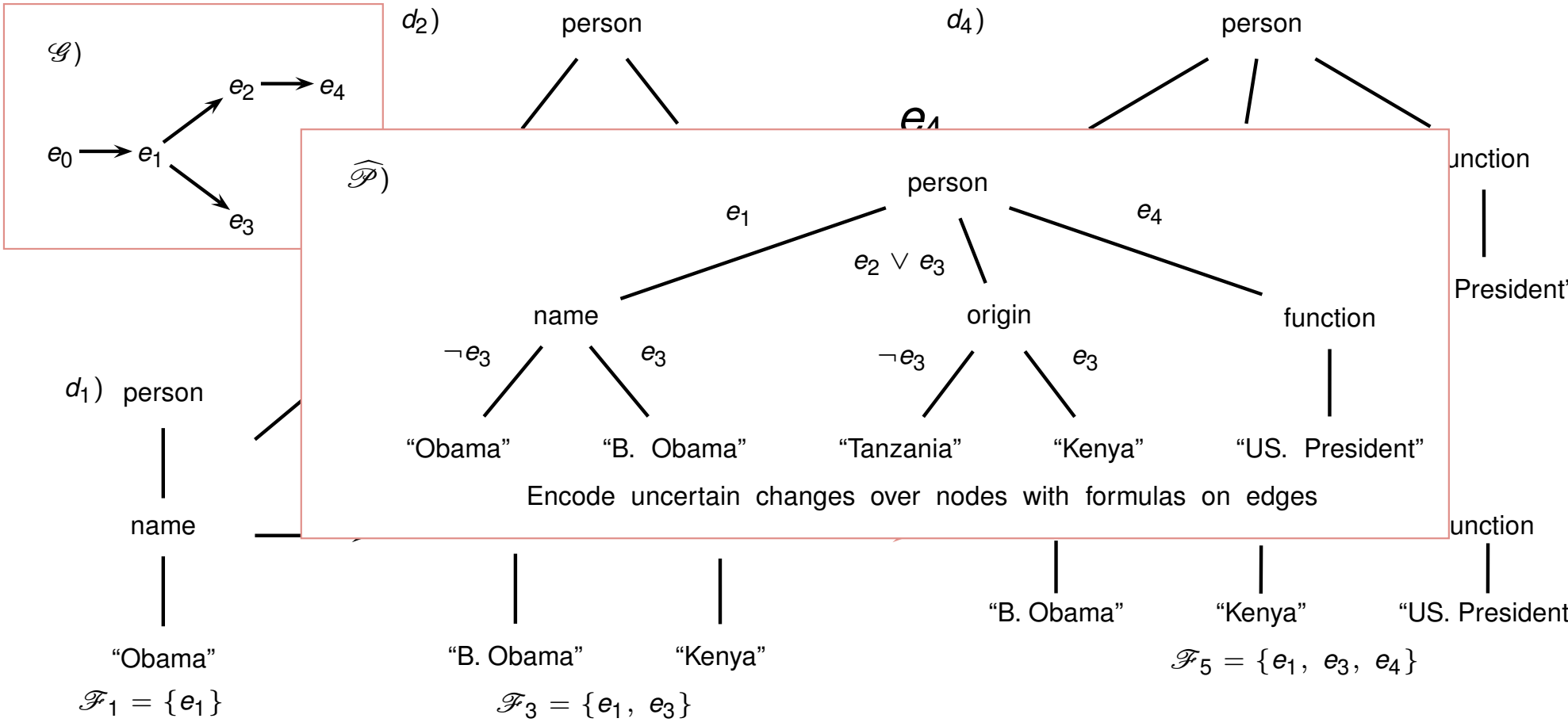
Uncertain Version Control Model (Example)



Licence de droits d'usage

Uncertain Multi-Version XML Document

Uncertain Version Control Model (Example)



Licence de droits d'usage

Uncertain Multi-Version XML Document

Semantics of Updates

- Assume an event e_i in $\mathscr{G}$ pointing to the edited version
- Come with a new event e_j *not in* $\mathscr{G}$ and an edit script Δ

Logical definition

Input: $(\mathscr{G},\ \Omega),\ e_i,\ e_j,\ \Delta$

- $\mathscr{G} := \mathscr{G} \cup (\{e_j\},\ \{(e_i,\ e_j)\})$
- Extension of Ω to a Ω' mapping

For each event set $\mathscr{F} \in 2^{(\mathscr{V} \setminus \{e_0\}) \cup \{e_j\}}$:

- $\Omega'(\mathscr{F}) = [\Omega(\mathscr{F} \setminus \{e_j\})]^{\Delta}$ if $e_j \in \mathscr{F}$
- $\Omega'(\mathscr{F}) = [\Omega(\mathscr{F})]$ if $e_j \notin \mathscr{F}$

$\Longleftrightarrow$

Probabilistic XML Update

Input: $(\mathscr{G},\ \widehat{\mathscr{P}}),\ e_i,\ e_j,\ \Delta$

- $\mathscr{G} := \mathscr{G} \cup (\{e_j\},\ \{(e_i,\ e_j)\})$
- Updating $\widehat{\mathscr{P}}$ with operations in Δ

For an insert of x and a delete of y:

- $fie(x) := fie(x) \vee (e_j)$ if $x \in \widehat{\mathscr{P}}$ or insert x in $\widehat{\mathscr{P}}$ with $fie(x) := (e_j)$
- $fie(y) := fie(y) \wedge \neg(e_j)$

Licence de droits d'usage

Plan

Uncertain Multi-Version XML Document

- Uncertain Version Control Model
- Semantics of Updates

Evaluation of the model

- Performance Analysis
- Filtering capabilities

Licence de droits d'usage

Evaluation of the model

Performance Analysis

- Estimation of two main metrics: commit time and checkout cost

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

- Set up our system (PrXML) in Java language

- Measures are obtained with all accesses in RAM Disk

Licence de droits d'usage

Evaluation of the model

Performance Analysis (Results)

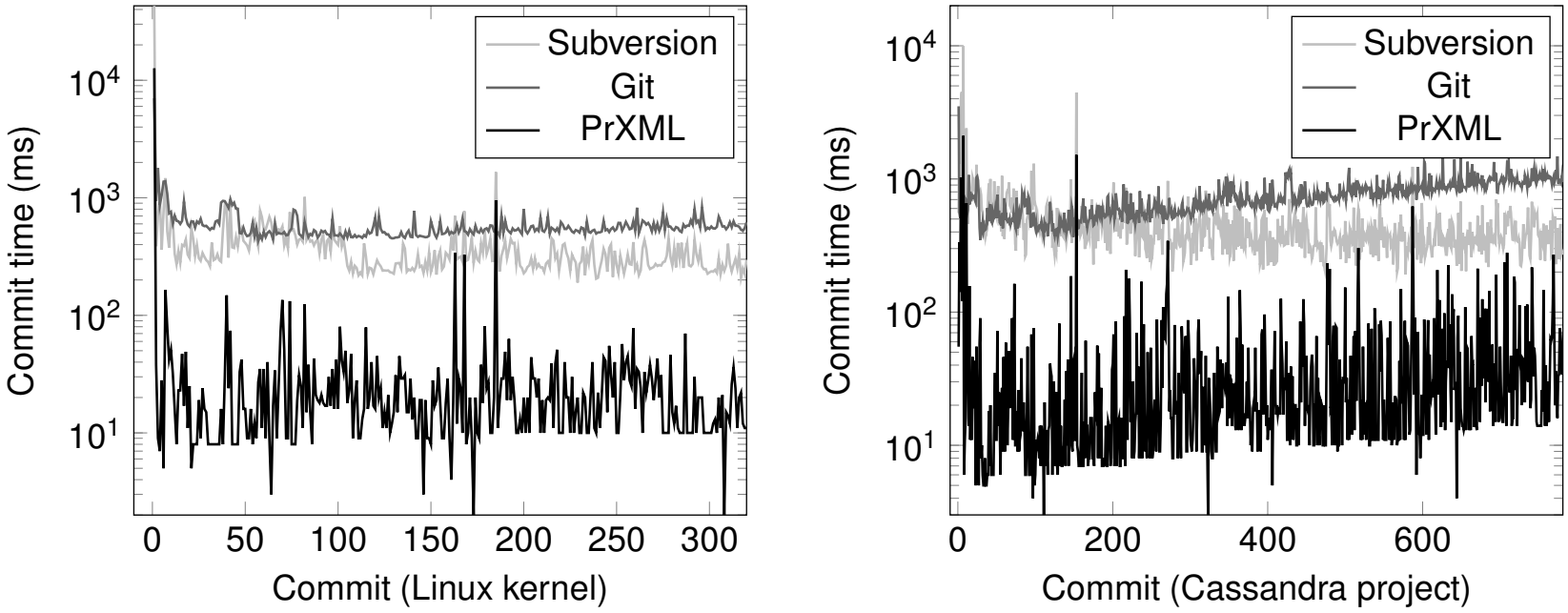


Figure : commit time over real-world datasets (logarithmic y-axis)

Licence de droits d'usage

Evaluation of the model

Performance Analysis (Results)

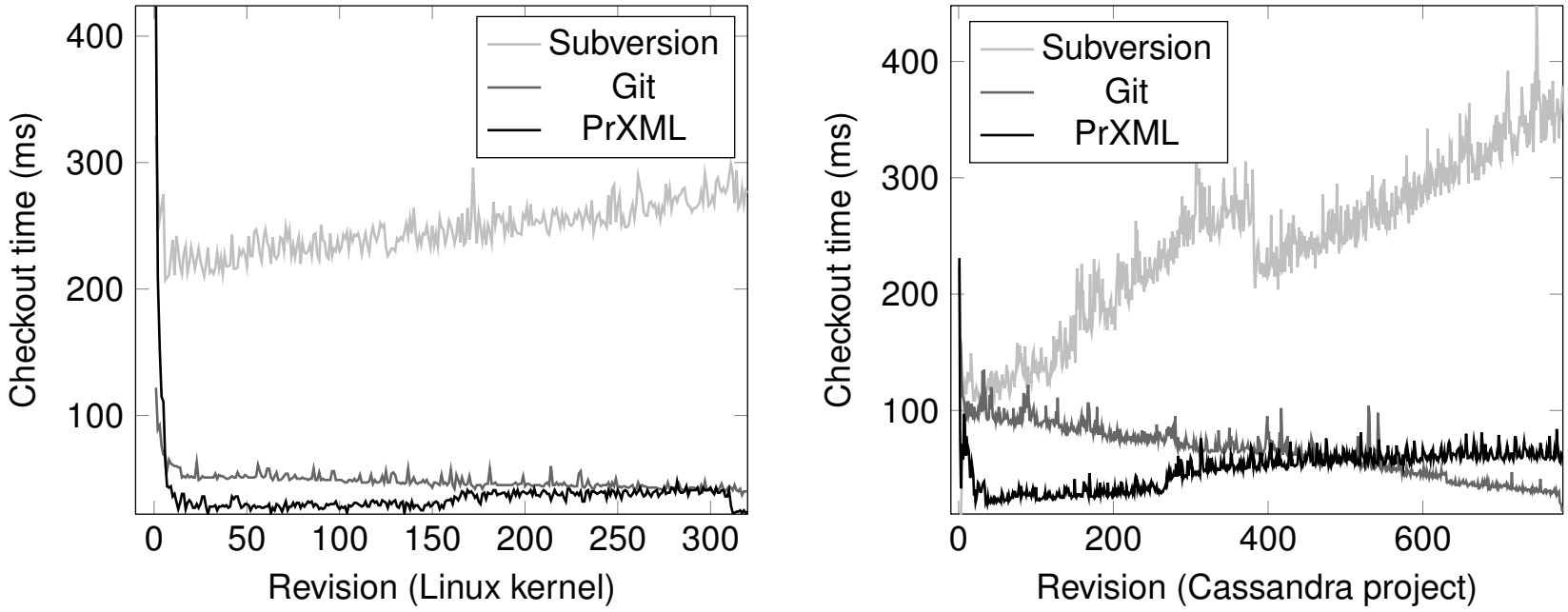


Figure : checkout time over real-world datasets (linear axes)

Licence de droits d'usage

Evaluation of the model

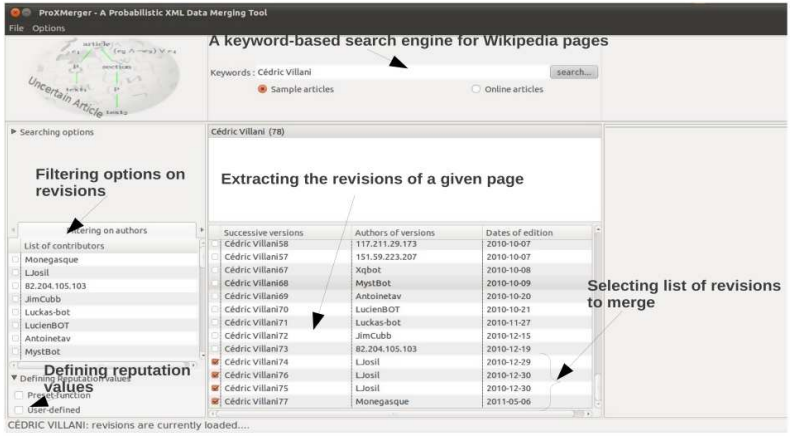
Filtering capabilities

- Tests are run over a sample of articles from the Wikipedia dump
  - Automatic filtering of unreliable content, e.g. spams, in versions of articles
  - Generate arbitrary versions that fit user preference
    - versions from trustworthy authors
  - Test more advanced operations over critical articles such as vandalized pages
    - e.g. study the impact of considering as reliable some versions affected by vandalism in the history of the edition of a given article
  - Detection of vandalism as well as Wikipedia robots do, automatically manage it while keeping all uncertain versions available for checkout.
  - etc.

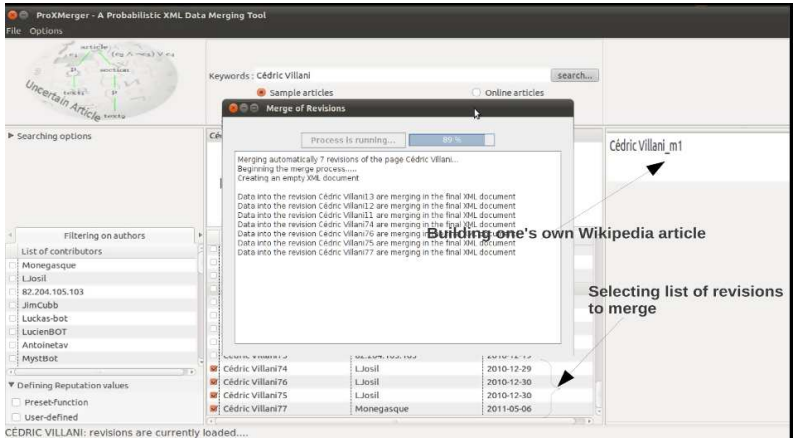
Licence de droits d'usage

Evaluation of the model

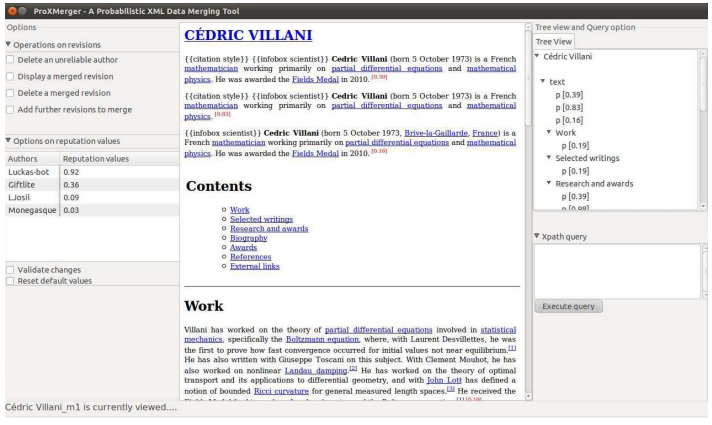
Filtering capabilities (Demo [Ba et al.(2011)])



(a) keyword-based search engine



(b) generation of arbitrary versions



(c)visualization features

Licence de droits d'usage

Thank for your attention !

Licence de droits d'usage

References

- Talel Abdessalem, M. Lamine Ba, and Pierre Senellart, *A probabilistic XML merging tool*, EDBT, 2011, Demonstration.
- M. Lamine Ba, Talel Abdessalem, and Pierre Senellart, *Towards a version control model with uncertain data*, PIKM, 2011.
- 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.