Approximate Data Exchange

Michel de Rougemont
University Paris II & LRI

Adrien Vieilleribière
University Paris-Sud & LRI

ICDT 2007

Motivation

1. Data from different imperfect sources. Framework for Data-Exchange and Data-Integration

2. Logic and Approximation
   - Definability and Complexity (scaling)
   - Robustness

3. Statistics based computations

Plan

1. Classical Data Exchange on words and trees

2. Approximation based on Property Testing.

   Tester for regular words and regular trees

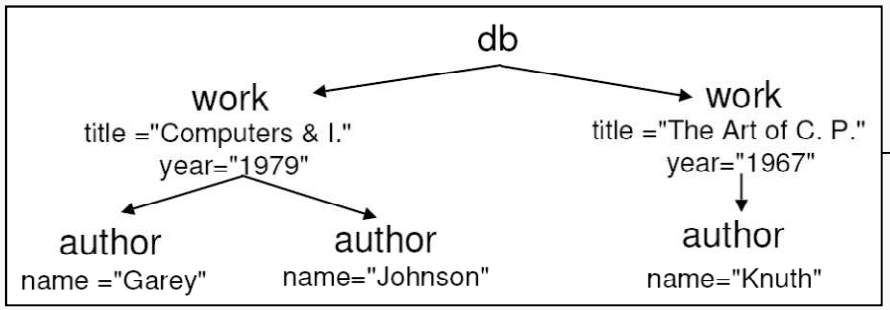
   (Edit Distance with Moves)

   - Property testing for regular tree languages (ICALP 2004)
   - Approximate Satisfiability and Equivalence (LICS 06)

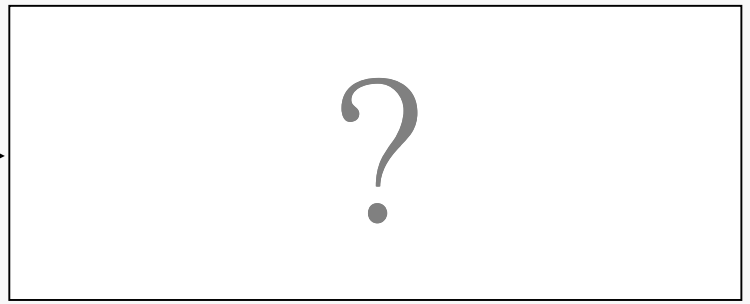
3. Approximate Data Exchange

1. Data Exchange on Trees

Source



Targets



<!ELEMENT db (work*)>
<!ELEMENT work (author*)>
<!ATTLIST work title CDATA #REQUIRED year CDATA>
<!ELEMENT author (EMPTY)>
<!ATTLIST author name CDATA #REQUIRED>

<!ELEMENT bib (livre*)>
<!ELEMENT livre (auteur+, titre , annee)>
<!ELEMENT auteur #PCDATA>
<!ELEMENT titre #PCDATA>
<!ELEMENT annee #PCDATA>

Classical Data-Exchange

Data Exchange setting: (K_S, τ, K_T)

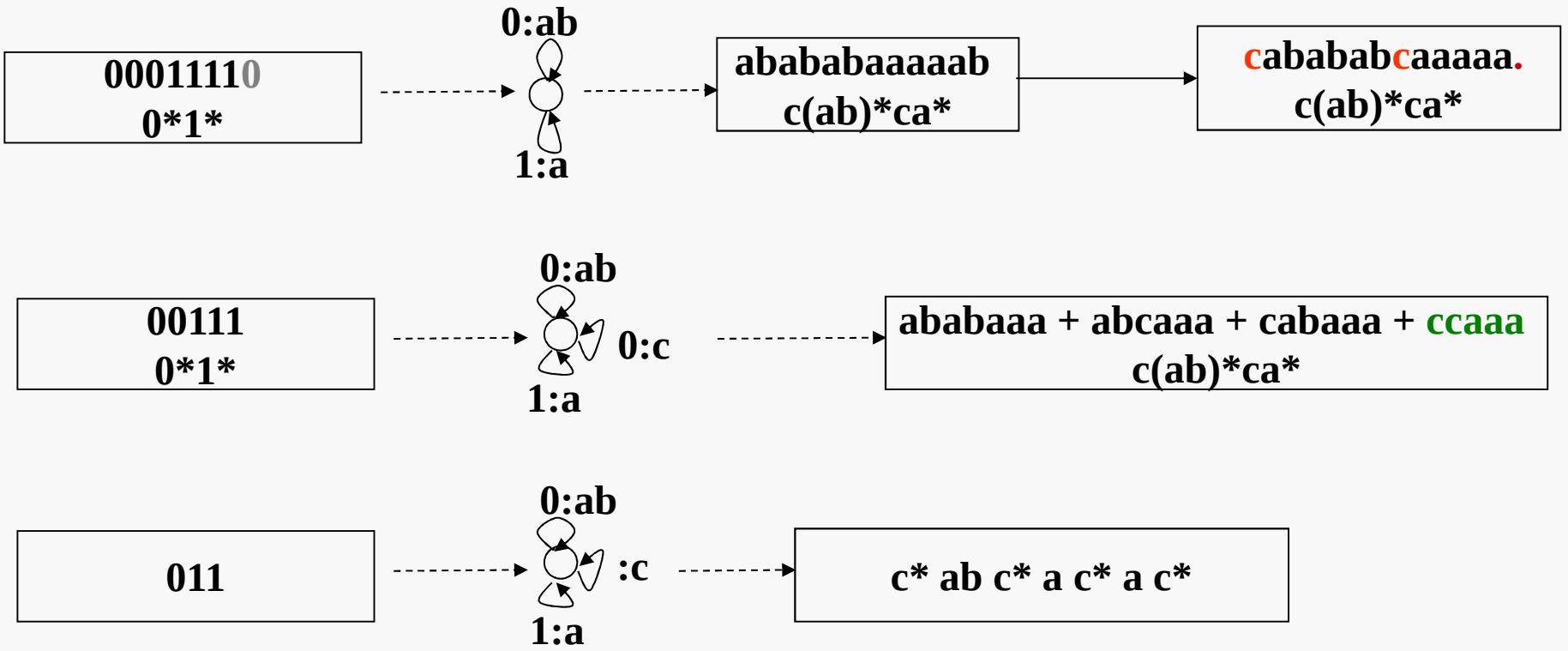
- Fagin et al. 2002: τ defined by Source-Target-Dependencies on relations
- Arenas, Libkin 2005: τ defined by Tree-Pattern-Formulas on trees

- **Source-Consistency:** Given a source structure I in K_S, is there a target J in K_T s.t. (I,J) in τ ?
- **Typechecking:** Decide if for all I in K_S and all J s.t. (I,J) in τ, J is in K_T.
- **Composition** of settings ?
- **Query Answering:** Given a source structure I in K_S, decide if for all J s.t. (I,J) in τ, J is in K_Q.

Class τ defined by Transducers

Deterministic Transducer on unranked trees with attributes. In practice, XSLT program.

Generalization to non-deterministic Transducers..



Approximate Data Exchange

(K_S,τ,K_T) is a setting, where τ is a transducer:

- **ε-Source-Consistency:** Given a source structure *I*, is there a source $I' \in K_S$, ε-close to *I* s.t. $\tau(I')$ *is* ε-close to K_T ?
- **ε-Typechecking:** Decide if for all *I* in K_S, $\tau(I)$ *is* ε-close to K_T.
- **ε-Composition** of settings.

General transducer τ :

- **ε-Query Answering:** Given a source structure *I*, is there a source *I'* ε-close to *I* s.t. any *J* [s.t. (*I'*,*J*) is in τ] is ε-close to K_Q ?.

2. Property Testing

Let F be a property on a class K of structures U

An **ε-tester** for F is a probabilistic algorithm A such that:

- If U |= F, A accepts
- If U is ε-far from F, A rejects with high probability

A property F is **testable** if there exists a probabilistic algorithm A s.t.

- For all ε it is an **ε-tester** for F
- Time(A) independent of n=|U|.

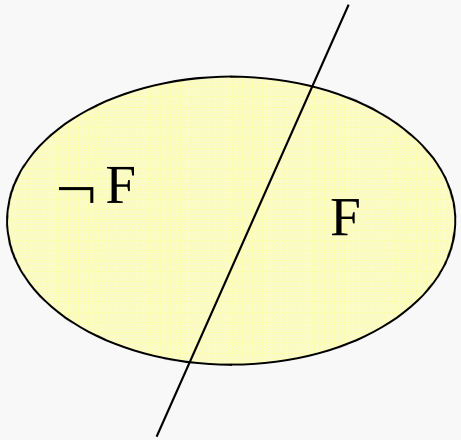
R. Rubinfeld, M. Sudan, Robust characterizations of polynomials, 1994
O. Goldreich, S. Goldwasser and D. Ron, Property Testing and its connection to Learning and Approximation, 1996.

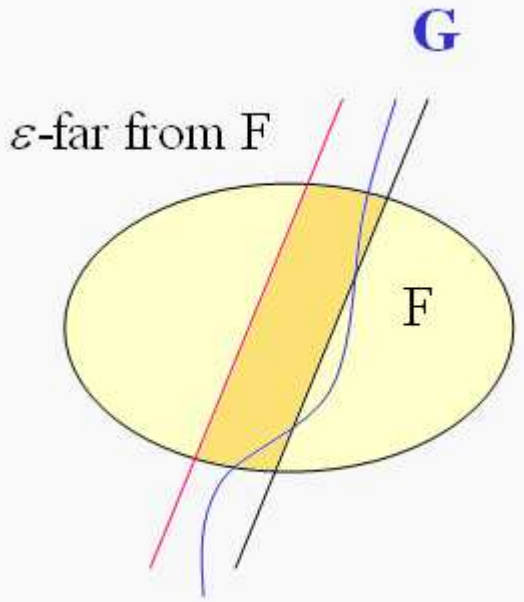
Tester usually implies a linear time corrector. $(\varepsilon_1, \varepsilon_2)$-Tolerant Tester.

Approximate Satisfiability and Equivalence

1. Satisfiability: $T \models F$
2. Approximate Satisfiability: $T \models_{\varepsilon} F$
3. Approximate Equivalence: $F \equiv_{\varepsilon} G$

Image on a class K of trees





Edit Distances with Moves

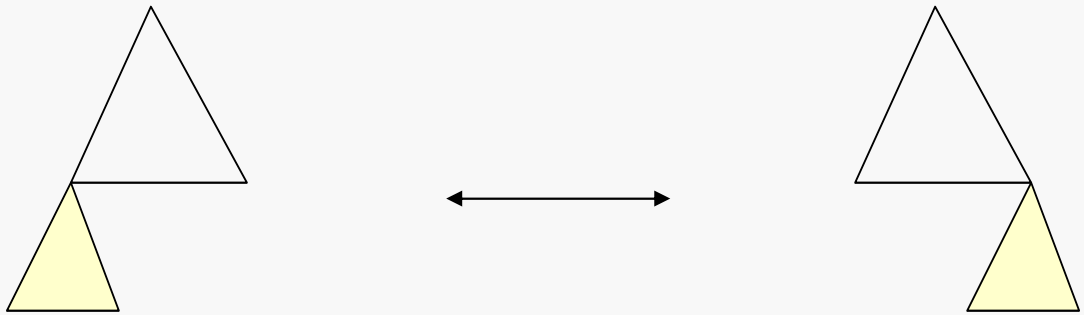
1. Classical Edit Distance: *Insertions, Deletions, Modifications*
2. Edit Distance with moves .

01110000111100 11001

01110 111100 00011001

$$dist(W, W') \;;\; dist(W, L) = Min_{W' \in L} \{dist(W, W')\}$$

3. Edit Distance with Moves generalizes to Ordered Trees



Uniform Statistics: k=1/ε

$$u.stat(W) = \begin{pmatrix} \#n_1 \\ .. \\ .. \\ \#n_{2^k} \end{pmatrix} . \frac{1}{n-k+1}$$

$\#n_1$ number of "00...0"
$\#n_2$ number of "00...1"
.. ..
.. ..
$\#n_{2^k}$ number of "11...1"

***W*=00101010111 0** length n,
n-k+1 blocks of length k
For k=2, n=12, 11 blocks

$$u.stat(W) = \begin{pmatrix} 1 \\ 4 \\ 4 \\ 2 \end{pmatrix} . \frac{1}{11} \approx Y(W) + \varepsilon$$

Fact 1: dist(W,W') $\approx |u.stat(W)\text{-}u.stat(W')|_1$ for words of similar length

Fact 2: $|u.stat(W)\text{-}Y(W)|_1 \leq \varepsilon$ for $Y(W)$ the *u.stat* vector on N samples

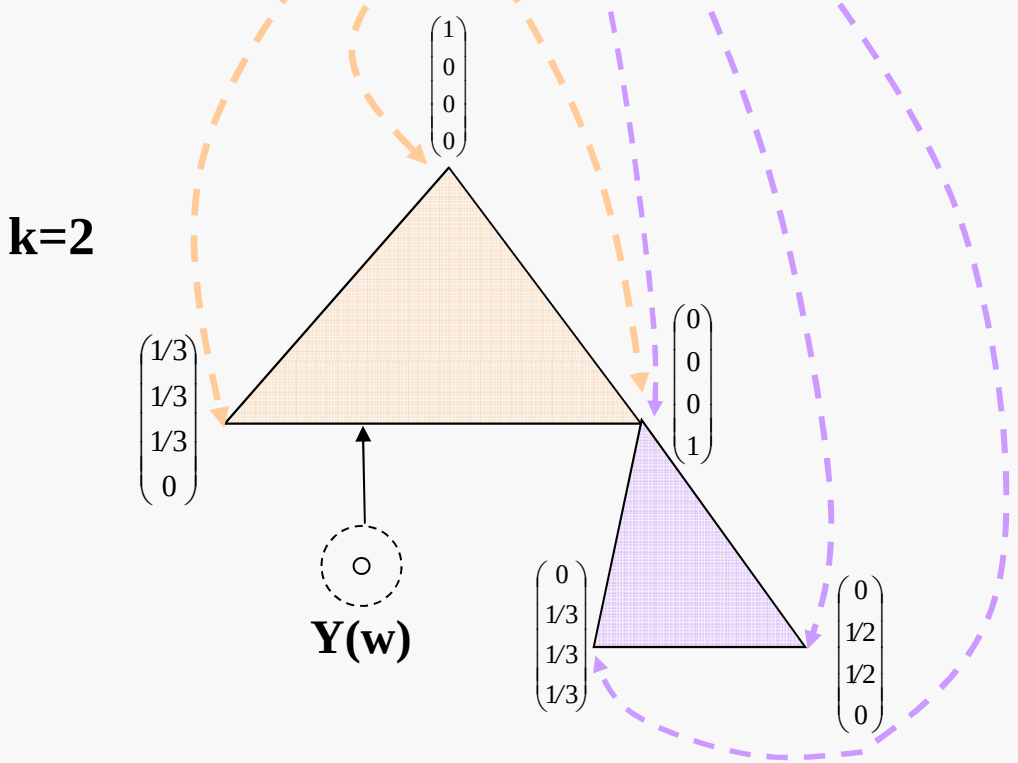
Distance between words (NP-complete)

- Testable, O(1): Sample N subwords of length k: $Y(W)$ and $Y(W')$

  If $|Y(w)\text{-}Y(w')|_1 < \varepsilon$ accept, else reject

Statistics on Regular Expressions

r = (010)*0*1* + 1*(01)*(110)*



H={*u.stat(w) : w* in *r* } is a union of polytopes.

2 polytopes for r.

.

Membership Tester:

Compute *Y(w)*. Accept if d(*Y(w)*,H) ≤ ε, else reject

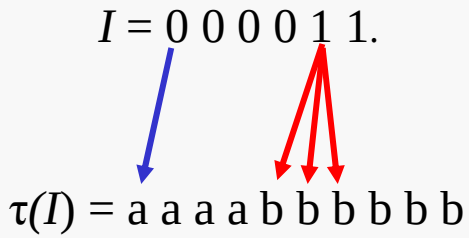
3. Approximate Data Exchange

ε-Source-Consistency: Given a source structure I, is there a source $I' \in K_S$ ε-close to I s.t. $\tau(I')$ is ε-close to K_T ?

Complexity parameter: n=|I|

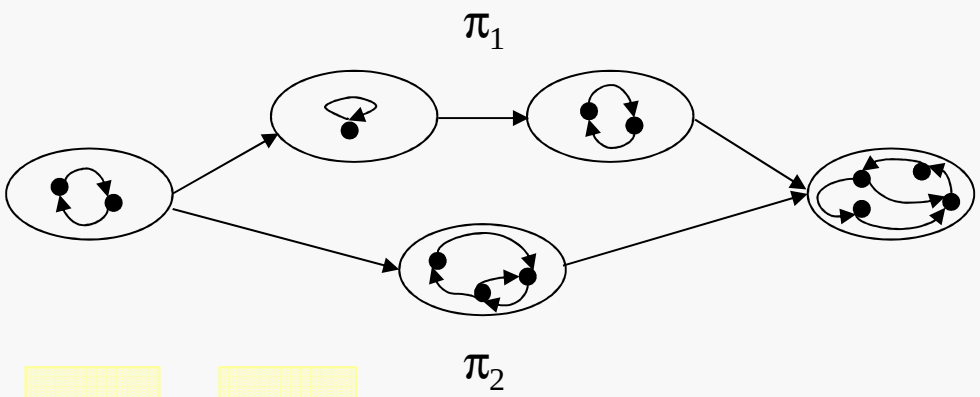
Case of 1-state on words: how to k-sample uniformly in $\tau(I)$ *?*

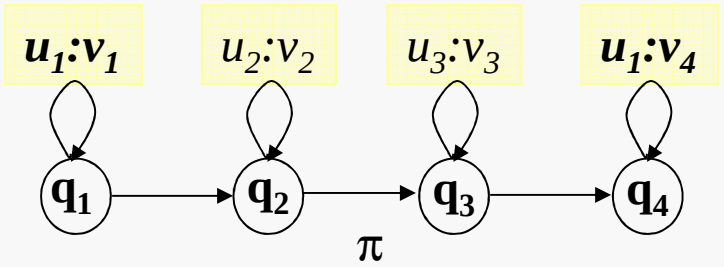
Suppose τ(0)=a, τ(1)=bbb. Adjust the probabilities:
If s=0…, 1 possible block from τ(0), adjust with 1/3
If s=1…, 3 possible blocks from τ(1), choose a shift in {0,1,2} uniformly



Approximate $u.stat(\tau(I))$.

Analysis of τ for ε-Source-consistency:

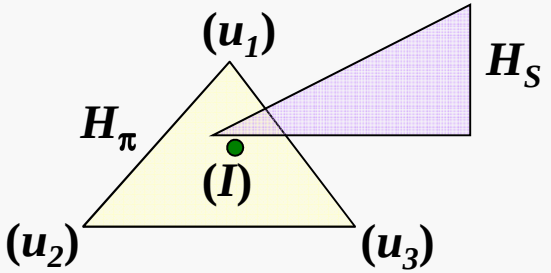




$$u.stat(I) \approx \lambda_1(u_1) + \lambda_2(u_2) + \lambda_3(u_3)$$

$$\sum_{i=1}^{3} \lambda_i = 1$$

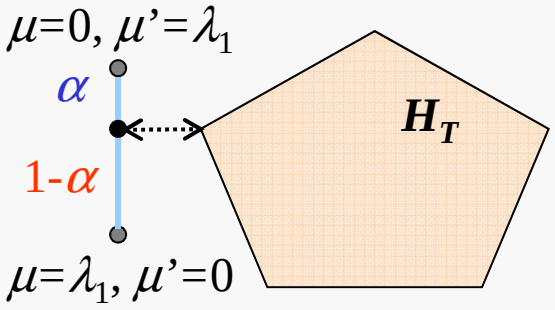
$$H_S \leftrightarrow u.stat(K_S)$$
$$H_\pi \leftrightarrow u.stat(\tau_\pi)$$
$$H_T \leftrightarrow u.stat(K_T)$$



$$u.stat(\tau_\pi(I)) = \mu(v_1) + \mu'(v_4) + \lambda_2(v_2) + \lambda_3(v_3)$$
$$with\ \mu + \mu' = \lambda_1.$$

Tester for ε-Source-consistency:

$u.stat(\tau_\pi(I)) = \mu(v_1)+\mu'(v_4)+\lambda_2(v_2)+\lambda_3(v_3)$
with $\mu+\mu'=\lambda_1$.



Tester:

- *u.stat*(*I*) is ε-far from H_S: reject [*I* is far from K_S] → **Tester for K_S.**
- Generate $\Pi=\{\pi \mid u.stat(I)$ is ε-close from being decomposable over $H_\pi\}$ → **Testers for K_π**
- While ($\Pi\neq\varnothing$) {
  - take a π in Π, approximate $u.stat(\tau_\pi(I))$ and $x=d(u.stat(\tau_\pi(I)), H_T)$
  - If $x\leq\varepsilon$, then accept and stop
    else remove π from Π }
- Reject

Find *I'*: If the test accepts, split λ_1 with the α proportions :

$I = u_2\, u_1u_1u_1\, u_1u_1u_1u_1u_1u_1\, u_3u_3$

$I' = u_1u_1u_1\, u_2\, u_3u_3\, u_1u_1u_1u_1u_1u_1$

Approximate ε-Source-Consistency:

Lemma: If I is s.t. $\tau(I) \in K_T$, then A accepts because there is a π with $\text{dist}(\tau_\pi(I), K_T)=0$

Lemma: If I is ε-far from being Source-Consistent, then the tester reject with high probabilities.

Theorem: For every $\varepsilon > 0$, there is an ε-tester for the ε-Source-Consistency on words.

Corollary: If I is ε-Source-Consistent, the procedure leads to an I' s.t. $\tau(I')$ is ε-close to K_T.

ε-Source-Consistency

Given a source structure *I*, is there a source *I'* ε-close to *I* s.t. *τ(I')* is ε-close to K_T ?

Case of 1-state: how to k-sample uniformly in *τ(I)* ?

Suppose τ(0)=ab, τ(1)=a, τ(2)=ccc. Adjust the probabilities:

If *s*=0…,	2 possible blocks from τ(0),	adjust with 1/3
If *s*=1…,	1 possible block from τ(1),	adjust with 1/6
If *s*=2…,	3 possible blocks from τ(2),	adjust with 1/2

Approximate *ustat(τ(I))*.

0 0 2 1 1 1 1

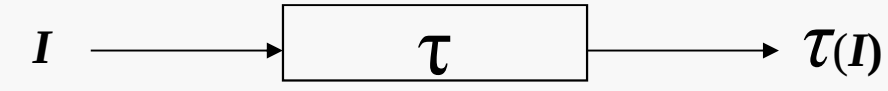
a b a b c c c a a a a

Outputs:
aa : 4/7.1/6.3/4=1/14
ab : 2/7.1/3.1/2=1/21
bc : 1/7.1/3.1/2=1/42
ca : 1/7.1/3.1/2=1/42
cc : 1/7.1/2.2/3=1/21

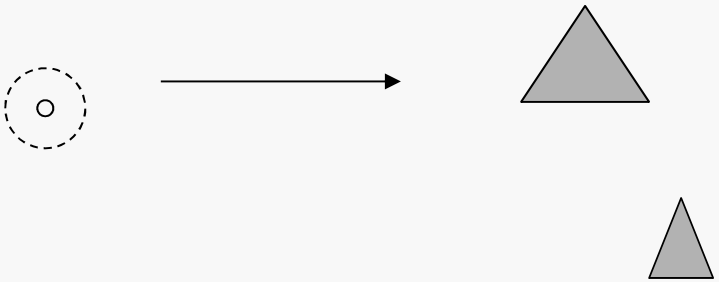
$\rightarrow \begin{pmatrix} 0.34 \\ 0.23 \\ 0.1 \\ 0.1 \\ 0.2 \end{pmatrix}$

ustat(τ(w))= $\begin{pmatrix} 0.33 \\ 0.22 \\ 0.1 \\ 0.1 \\ 0.22 \end{pmatrix}$

Image of the statistics by a general transducer



$$u.stat(I) = \begin{pmatrix} 1/11 \\ 4/11 \\ 4/11 \\ 2/11 \end{pmatrix}$$

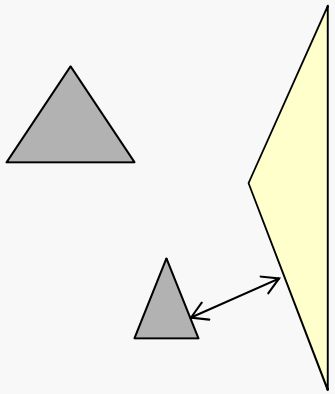


Union of polytopes

Applications:

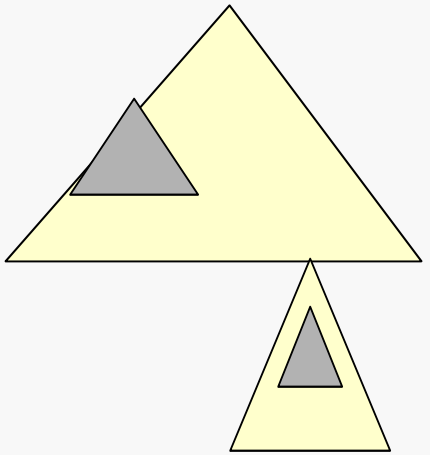
ε-Source-Consistency:

$d(u.stat[\tau(I)], H_T) \leq \varepsilon$?



ε-Query Answering:

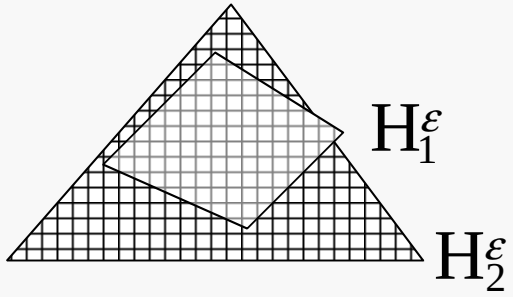
$u.stat[\tau(I)] \subseteq_\varepsilon H_Q$?



Inclusion Tester for regular properties

Tester for inclusion : $r_1 \subseteq r_2$

$H_1^{\varepsilon} \subseteq H_2^{\varepsilon}$?

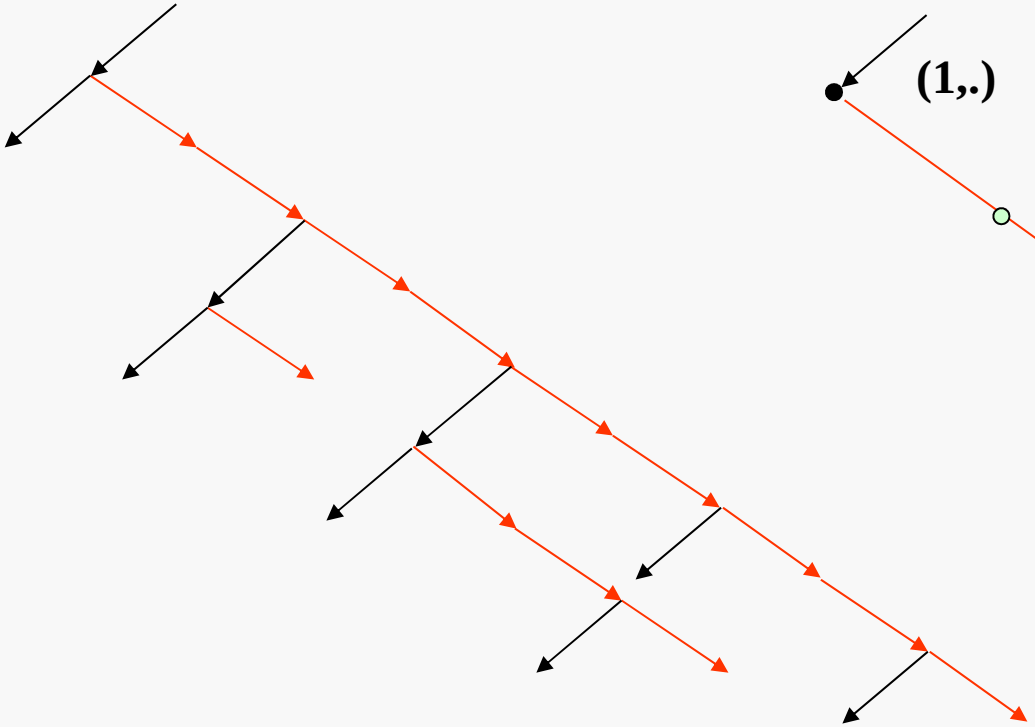


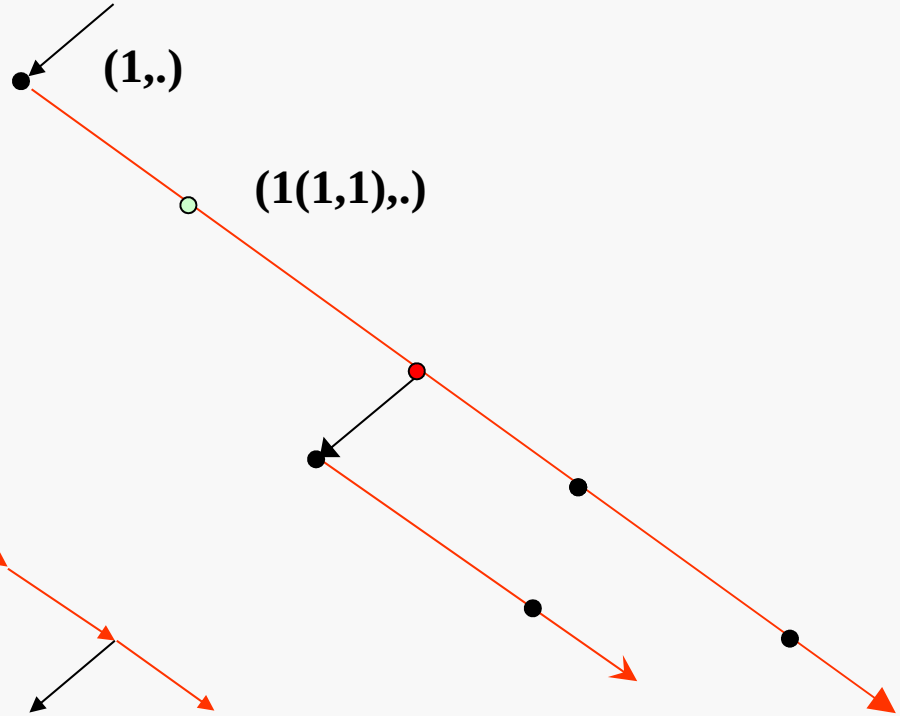
Application: **ε-Typechecking:** Decide if J is ε-close to K_T [for all I in K_S and all (I,J) in τ] .

Solution: Inclusion Tester for $\tau(K_S) \subseteq_{\varepsilon} K_T$.

Time polynomial in m=Max($|r_1|$,$|r_2|$): $m^{|\Sigma|^{O(k)}}$

Statistics on Trees





***T*: Ordered (extended) Tree of rank 2.**

***T'*: squeleton**

W: word with labels. Apply *u.stat* on *W* and define *u.stat*(*T*).

Extension to trees

Statistics on DTDs:

H={stat(t) : t in DTD} is still a union of polytopes (harder analysis to construct it)

Transducer τ with attributes:

- $\delta : \Sigma_S \times Q \rightarrow Hedge_{\Sigma_T, A_T}[Q]$
- $h : \Sigma_S \times Q \times A_S \rightarrow \{1\} \cup Var$ extended to $\Sigma_S \times Q \times Str \rightarrow Str \cup Var$
- $\mu : \Sigma_S \times Q \times A_T \times D_T \rightarrow \{1,\ldots,k\}$ where D_T is the hedge defined by δ.

τ is decomposable in a finite number of paths in the graph of the strongly connected components.

Lemma: The image of a statistical vector through a path is a union of polytopes.

ε-Source-Consistency on trees

Test: If there is a π (allowing a decomposition of t on H_π) s.t. $u.stat(\tau_\pi(t))$ is ε-close to H_T then accept, else reject

Lemma: If $\tau(t) \in K_T$, then there is a π with $dist(\tau_\pi(t),K_T)=0$.

Lemma: If t is ε-far from being ε-Source-Consistent, then we reject with high probabilities.

→ Testers for K_S, K_π;
→ x: approximation of $u.stat(\tau_\pi(t))$,
$d(x,H_T) \leq \varepsilon$?

Theorem: For every $\varepsilon > 0$, there is an ε-tester for the ε-Source-Consistency on trees.

Corollary: If t is ε-Source-Consistent, the procedure leads to an t' s.t. $\tau(t')$ is ε-close to K_T

Composition of close settings

An *ε-corrector* for a class $K_0 \in K$ is a algorithm A which takes as input a structure *I* which is ε-close to K_0 and outputs a structure $I_0 \in K_0$, such that I_0 is ε-close to *I*.

Ex : If an XML file *F* is ε-close from a DTD, find a valid *F'* ε-close to *F*:
http://www.lri.fr/~mdr/xml/

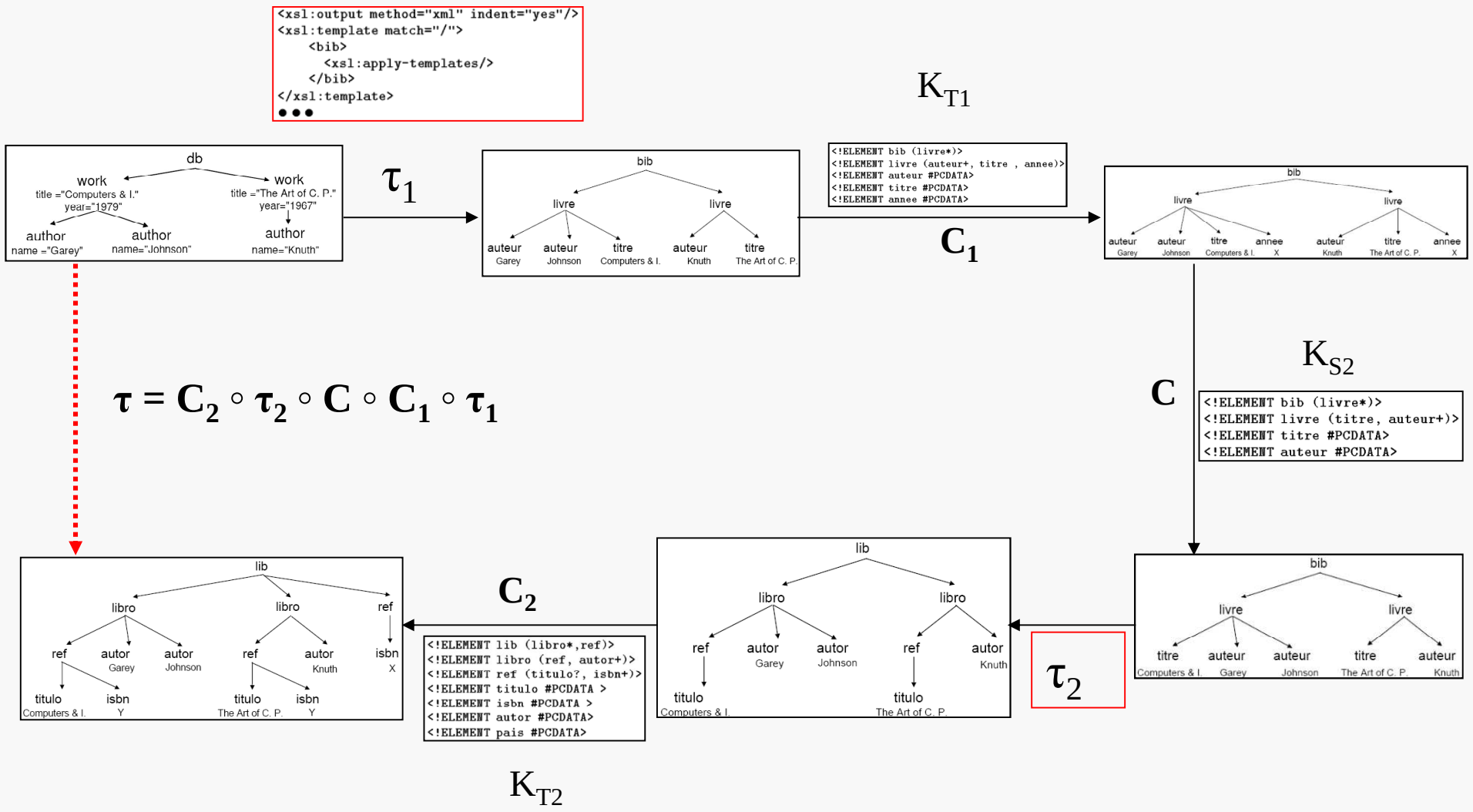
Data Exchange settings: (K_{S1}, τ_1, K_{T1}), (K_{S2}, τ_2, K_{T2}):
Solution if they are *ε-composable*

- K_{T1} and K_{S2} are ε-close.
- the settings satisfy ε-typechecking

Composition: Apply correctors at every stage to define the new τ.

$\rightarrow (K_{S1}, \tau, K_{T2})$ satisfies 3ε-typechecking.

Composition



Conclusion

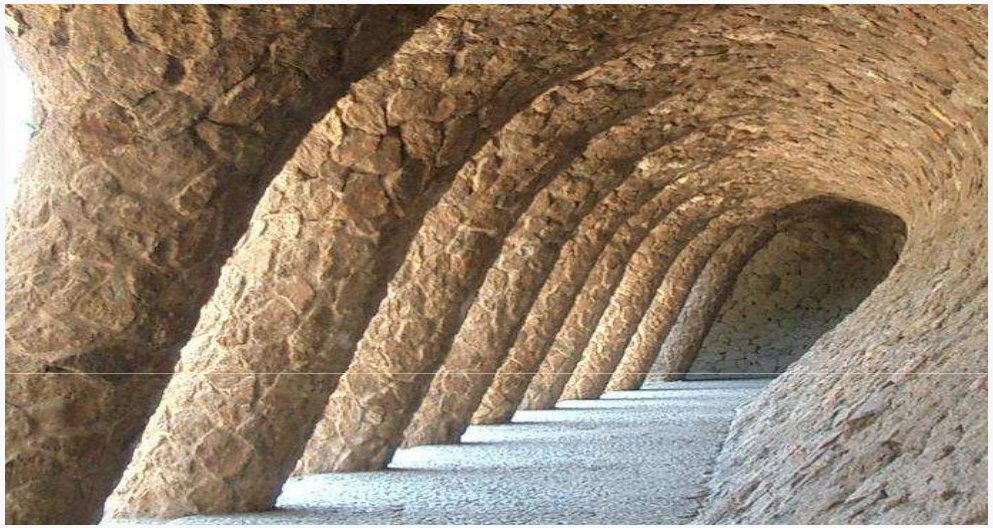
1. Data Exchange:
   - Source-Consistency,
   - Typechecking,
   - Query-Answering.

2. Approximate Data Exchange:

   Property Testing based Approximation

   - ε-Source-Consistency,
   - ε-Typechecking,
   - ε-Query-Answering,
   - ε-Composition.

Questions ?



Adrien Vieilleribière: vieille@lri.fr
Michel de Rougemont: mdr@lri.fr