

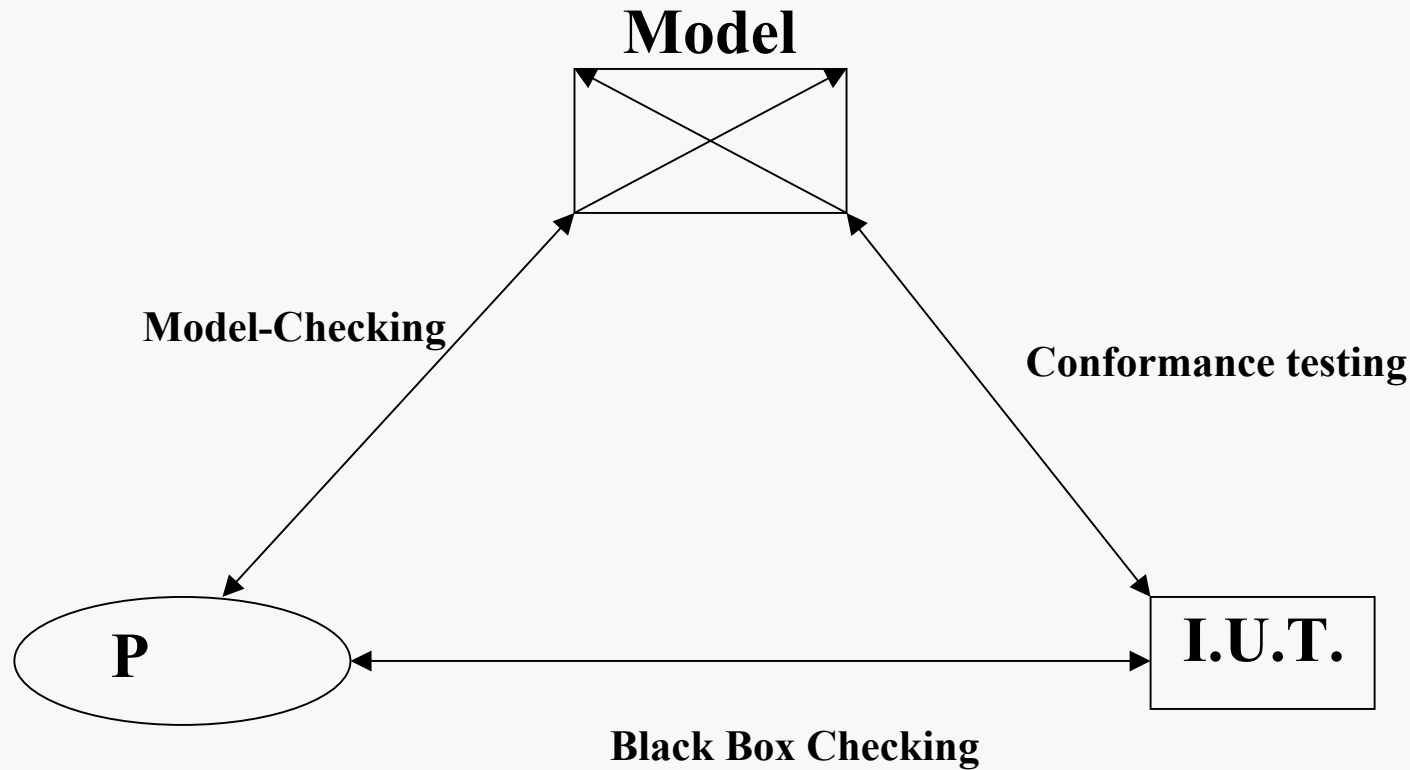
# Approximate verification (part 1)

Michel de Rougemont, LRI , *University Paris II*

Joint work with E. Fischer, *Technion*, F. Magniez, *LRI*



# Model Checking and Test



**Example:** finite automata,  $P: 0^*1^*$

# Approximate Verification

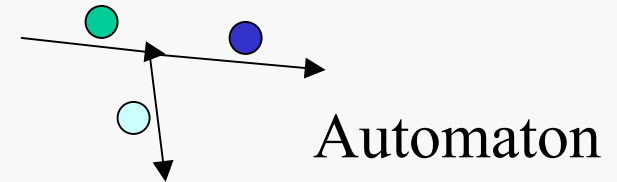
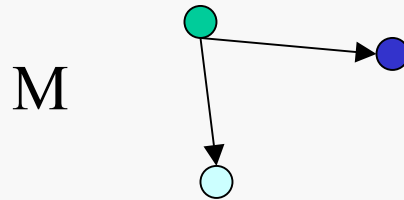
1. Probabilistic verification  
Interactive Proofs, PCP, Property testing
2. Approximate equality (constant time)
3. Test if a word  $w$  is in a regular language (constant time)  
Test if a tree  $T$  is in a regular Tree language (constant time)
4. Equivalence Testing: Compare two languages  
Finite automata (regular expressions): Polynomial time approximation (exact version is PSPACE complete)  
Buchi Automata: Polynomial time approximation  
Extensions to CF, Regular Expression with squaring and negation (Meyer, Stockmeyer 1972)

# Verification

1. Logic  $M \models \Theta$  ( $M$  satisfies  $\Theta$ )

$M = (S, R, P_1, \dots, P_k, s)$  Kripke Structure,  $R \subseteq S.S$  and  $P_i \subseteq S, s \in S$

$\Theta$ :  $E(p_1 U p_2)$  CTL formula



2. Complexity:  $f(x)=y$

Given  $x, y$  check if  $f(x)=y$

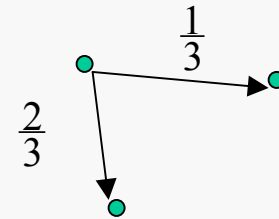
Example:  $Permanent(A)=y$

# Probabilistic verification

1. Logic  $M \models \text{Prob}[\Theta] > \frac{1}{2}$

$M = (S, R, P_1, \dots, P_k, s)$  where  $R \subseteq S \times S$ ,  $[0, 1]$  and  $P_i \subseteq S$ ,  $s \in S$

$\Theta$ :  $(p_1 U p_2)$  LTL formula



2. Complexity

Given  $x, y$  verify with a probabilistic protocol if  $f(x) = y$

IP : Interactive proof

PCP: Probabilistic Checkable Proof

Property Testing

# Approximations in MC and testing

1. Model-Checking: Automata, OBDD, SAT

**Problem:** combinatorial explosion

**Approximations** in Model-Checking:

BMC: Bounded Model-Checking

SAT solvers

Abstraction methods

Probabilistic MC

$\text{Prob}_{\Omega} [ (p_1 \cup p_2) ]$

2. Testing

PAC Learning

Statistical tests

# 1. New approximation for MC

## Testers and correctors

Property is  $\epsilon$  testable if there exists an efficient algorithm which distinguishes between true instances and  $\epsilon$  far instances.

**PAC Learning** : discover  $f$  from samples labelled by  $f$ .

## Approximate satisfiability:

$U \models F$  generalized to  $U \models_{\epsilon} F$

There exists  $U'$  s. t.  $\text{dist}(U, U') < \epsilon$  and  $U' \models F$

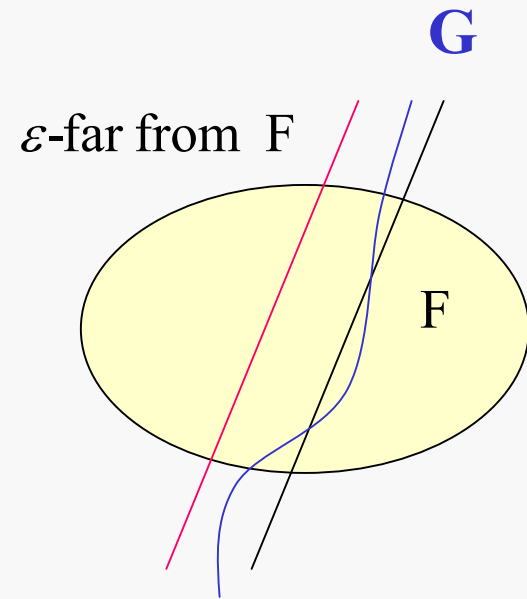
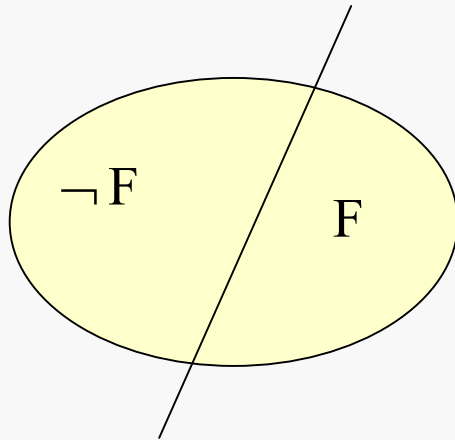
**Application to MC:**  $M \approx_{\epsilon} \Theta$  |

OBDD can still be exponential (LICS 02)

# Approximate Satisfiability and Equivalence

1. Satisfiability :  $\text{Tree} \models F$
2. Approximate satisfiability  $\text{Tree} \models_{\varepsilon} F$
3. Approximate equivalence  $F \equiv_{\varepsilon} G$

Image on a class K of trees



# Testers on a class $K$

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

An  $\varepsilon$ -tester for  $F$  is a probabilistic algorithm  $A$  such that:

- If  $U \models F$ ,  $A$  accepts
- If  $U$  is  $\varepsilon$  far from  $F$ ,  $A$  rejects with high probability
- $\text{Time}(A)$  independent of  $n$ .

(Goldreich, Golwasser, Ron 1996, Rubinfeld, Sudan 1994)

Tester usually implies a linear time corrector.

# History of Testers

Self-testers and correctors for Linear Algebra ,Blum & Kanan 1989

Robust characterizations of polynomials, R. Rubinfeld, M. Sudan, 1994

Testers for graph properties : k-colorability, Goldreich and al. 1996

$\Sigma_2$  graph properties have testers, Alon and al. 1999

Regular languages have testers, Alon and al. 2000s

**Testers for Regular tree languages** , Mdr and Magniez, ICALP 2004


## 2. Equality tester

### 1. Classical Edit Distance:

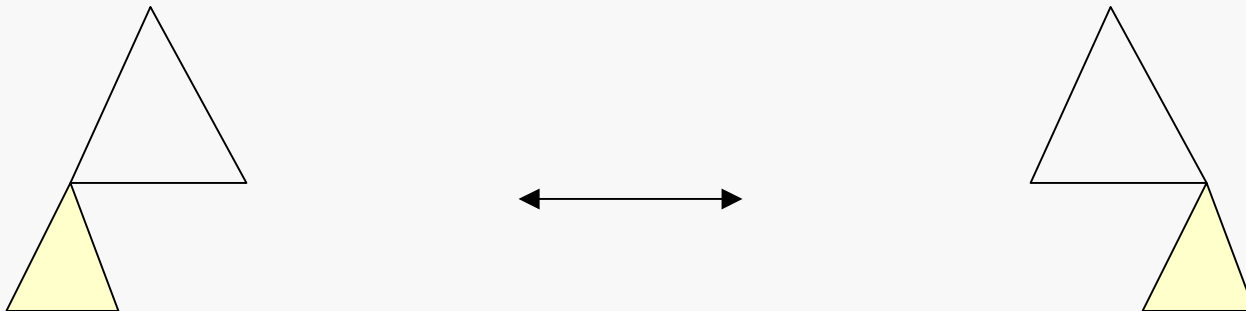
Insertions, Deletions, Modifications

### 2. Edit Distance with moves

**0111000011110011001**  
**0111011110000011001**



### 3. Edit Distance with Moves generalizes to Trees



# Statistics on words

**Block statistics: b.stat**

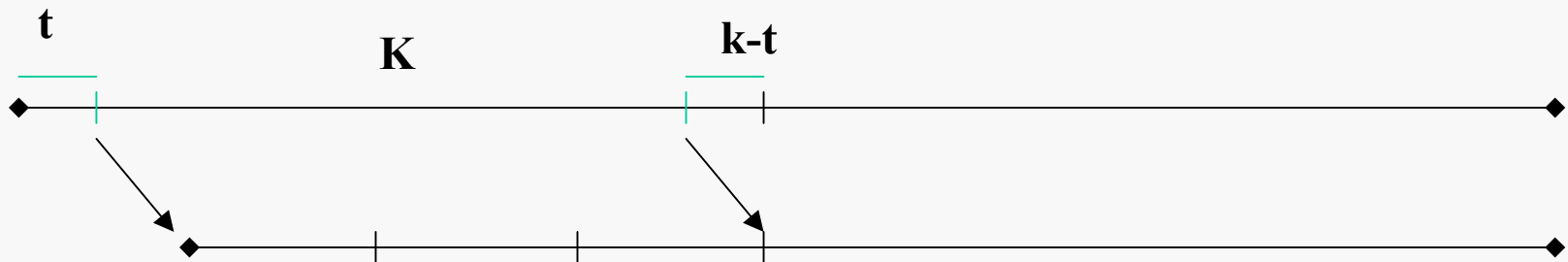
$$k = \frac{1}{\varepsilon}$$



**Uniform statistics: u.stat**



**Block Uniform statistics: bu.stat**



# Block statistics

$W=001010101110\dots$  length  $n$ , subword of length  $k$ ,  $n/k$  blocks

$$b.stat(W) = \begin{pmatrix} \#n_1 \\ \dots \\ \dots \\ \#n_{2^k} \end{pmatrix} \cdot \frac{1}{n/k}$$

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

$$b.stat(W) = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 1 \end{pmatrix} \cdot \frac{1}{6}$$

For  $k=2$ ,  $n/k=6$

# Uniform statistics

**W=001010101110**

**length n, subword of length k, n-k+1 blocks**

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

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

$$u.stat(W) = \begin{pmatrix} 1 \\ 4 \\ 4 \\ 2 \end{pmatrix} \cdot \frac{1}{11}$$

# Statistics and distance

$$W = 00\mathbf{1010101110} \quad (n=12, k=2) \quad W' = 110100001111$$

$$w_1 = 00\mathbf{1101001110} \quad w_2 = \mathbf{110100001110} \quad w_3 = \mathbf{110100001111}$$

$$\text{dist}(W, W') = 3$$

$$\text{dist}(W, W') / 12 = 0.25$$

$$b.stat(W) = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 1 \end{pmatrix} \cdot \frac{1}{6}$$

$$d_0 = \frac{8}{6} = 1.33$$

$$b.stat(W') = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 3 \end{pmatrix} \cdot \frac{1}{6}$$

$$u.stat(W) = \begin{pmatrix} 1 \\ 4 \\ 4 \\ 2 \end{pmatrix} \cdot \frac{1}{11}$$

$$d_1 = \frac{8}{11} = 0.73$$

$$u.stat(W') = \begin{pmatrix} 3 \\ 2 \\ 2 \\ 4 \end{pmatrix} \cdot \frac{1}{11}$$

# Goal: $d_1$ approximates the distance

Let  $\varepsilon = 1/k$  : For  $n > n_0$   $\mathbf{dist} - \varepsilon.n < d_1 < \mathbf{dist} + \varepsilon.n$

**Practical application:**  $\varepsilon = 10^{-2}$  hence  $k = 100$ , stat dimension  $2^{100}$   
Words of length  $n = 10^9$ ,  $d_1$  is approximated by  $N$  samples and a good approximation after  $N = O(1/\varepsilon^3)$  trials.

## Remarks:

1. Distance with Moves.

$$\begin{array}{l} W = 000 \dots 000 1111 \dots 111 \\ W' = 1111 \dots 111 000 \dots 000 \end{array} \quad b.stat(W) \approx \begin{pmatrix} 0.5 - \varepsilon/2 \\ \ddots \\ \varepsilon \\ \ddots \\ 0.5 + \varepsilon/2 \end{pmatrix} \approx b.stat(W')$$

2. Robustness to noise

If  $W, W'$  are noisy inputs (but  $\varepsilon$ -close), the method still works.

3. Random words are close with the moves, far without.

# Classical complexity

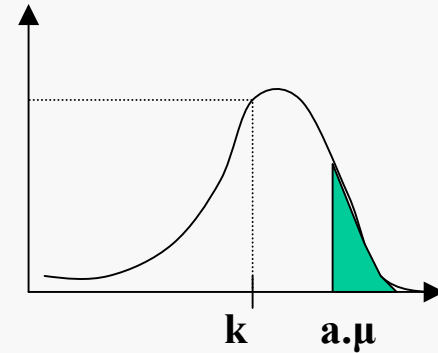
## Edit distances:

1. P problem on words without the moves.
  - Approximation?
  - Sublinear algorithm?
2. NP-complete problem on words with the moves.
  - $O(1)$ -approximable
3. P problem on ordered trees without the moves
4. NP-complete problem on unordered trees and trees with the moves.

# Basic tool: Chernoff bound

**Random variables:**

Prob[X=k]



**Markov bound**

**Chebyshev bound**

**Chernoff bound:** sum of independent variables  $X_i$ , whose average is  $\mu$

$$Y = \frac{1}{N} \sum_{i=1 \dots N} X_i$$

$$\Pr[|X - Y| \geq a \cdot \mu] \leq e^{-8 \cdot N \cdot a^2}$$

$$X_i = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \ddots \\ 0 \end{pmatrix}$$

**Hoeffding bound**

# Soundness and Robustness

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

$F$  is Equality

**Soundness:** close structures have close statistics

$$\text{dist}(w, w') \leq \varepsilon \cdot n$$

**Robustness:** far structures have far statistics

$$\text{dist}(w, w') \geq \varepsilon \cdot n$$

# Robustness of b.stat

**Robustness of b-stat:**  $dist(w, w') \leq (\frac{1}{2} \cdot |b.stat(w) - b.stat(w')| + \epsilon) \cdot n$

if  $b.stat(w) = b.stat(w')$  then  $dist(w, w') \leq \epsilon \cdot n$

if  $b.stat(w) \neq b.stat(w')$  then construct  $w''$  s. t.  $b.stat(w'') = b.stat(w')$   
after at most  $\frac{n}{2} \cdot |b.stat(w) - b.stat(w')|$  substitutions on  $w$ . Example:

$$b.stat(W) = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 1 \end{pmatrix} \cdot \frac{1}{6} \qquad b.stat(W') = \begin{pmatrix} 2 \\ 0 \\ 3 \\ 1 \end{pmatrix} \cdot \frac{1}{6}$$

#"00"=1 in  $W$  and 2 in  $W'$  but #"10"=4 in  $W$  and 3 in  $W'$

$W''$ : take one block of "00" in  $W$  and change it into "10"

# Soundness of u.stat

**Soundness of u-stat:**  $dist(w, w') \leq \varepsilon^2 \cdot n \implies |u.stat(w) - u.stat(w')| \leq 6 \cdot \varepsilon$

Simple edit:  $|u.stat(w) - u.stat(w')| \leq \frac{2k}{n-k+1} \leq \frac{2}{n \cdot \varepsilon}$

Move  $w = A.B.C.D$ ,  $w' = A.C.B.D$ :  $|u.stat(w) - u.stat(w')| \leq \frac{2 \cdot 3(k-1)}{n-k+1} \leq \frac{6}{n \cdot \varepsilon}$

Hence, for  $\varepsilon^2 \cdot n$  operations,  $|u.stat(w) - u.stat(w')| \leq 6 \cdot \varepsilon$

Problem: robustness of **u.stat** ?

Harder! You need an auxiliary distribution and two key lemmas.

# Block Uniform Statistics

**Lemma 1:**  $\forall w \exists v |bu.stat(w) - b.stat(v)| \leq \frac{\epsilon}{2}$

$$bu.stat(w) = \frac{K}{n} \sum_{i=1, \dots, n/K} E_{t_i}(b.stat(v_i)) = E_v(b.stat(v))$$

$$X_i = b.stat(v_i), \quad X_i[u] = b.stat(v_i)[u], \quad 0 \leq X_i[u] \leq 1$$

Each  $X_i[u]$  is independent. Average is  $bu.stat(w)[u]$

Chernoff Bound:  $\Pr[|b.stat(v)[u] - bu.stat(w)[u]| \geq t \times bu.stat(w)[u]] \leq e^{-\frac{8n}{K}t^2}$

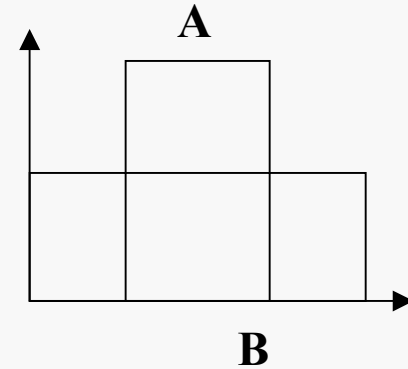
Union Bound:  $\Pr[|b.stat(v) - bu.stat(w)| \geq t \times bu.stat(w)] \leq |\Sigma|^k \cdot e^{-\frac{8n}{K}t^2}$

$$t = \frac{\epsilon}{2} \Rightarrow \Pr[|b.stat(v) - bu.stat(w)| \leq \frac{\epsilon}{2}] \geq 0$$

# Uniform Statistics

Lemma : Let  $A \subseteq B$  and two uniform distributions  $\mu_A, \mu_B$ .

$$\text{Then } |\mu_A - \mu_B| = 2 \cdot \frac{|B| - |A|}{|A|}$$



**Lemma 2:**  $\forall w \left| bu.stat(w) - u.stat(w) \right| \leq \frac{\log|\Sigma|}{\varepsilon^4 \cdot n}$

#subwords of length  $k$  missed by bu :  $(k-1) \cdot \left(\frac{n}{K} - 1\right)$

Apply the previous lemma with  $|B| = n - k + 1$ ,  $K = \frac{\varepsilon^3 \cdot n}{\log|\Sigma|}$

$$\left| u.stat(v) - bu.stat(w) \right| = O\left(\frac{\log|\Sigma|}{\varepsilon^4 \cdot n}\right)$$

# Robustness of the uniform Statistics

**Robustness of u-stat:**  $dist(w, w') \geq 5\varepsilon .n \Rightarrow |u.stat(w) - u.stat(w')| \geq 6,5.\varepsilon$

By Lemma 1:  $\forall w \exists v |bu.stat(w) - b.stat(v)| \leq \frac{\varepsilon}{2}$

Get  $v, v'$  from  $w, w'$

By Lemma 3:  $\forall w |bu.stat(w) - u.stat(w)| \leq \frac{\log|\Sigma|}{\varepsilon^4.n}$

Robustness of b-stat implies robustness of u-stat.

# Tester for the distance with moves

NP-complete problem, but  $O(1)$ -approximable.

## Approximate u.stat:

Sample  $N$  subwords of length  $k$ , compute  $Y$ :

$$Y = \frac{1}{N} \sum_{i=1 \dots N} X_i$$

$$X_i = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$Y$  is a good approximation of u.stat (Chernoff),

$$\Pr[|u.stat(W) - Y| \geq a \cdot \mu] \leq e^{-8 \cdot N \cdot a^2}$$

Uniform statistics is a good approximation of the distance by soundness and robustness.

**Tester:** If  $Y < \epsilon \cdot n$  accept, else reject.