Structures arborescentes complexes :
analyse combinatoire, génération aléatoire et applications

Alexis Darrasse

26 janvier 2010

Complex tree-like structures

combinatorial analysis

generating functions

singularity analysis

Boltzmann model

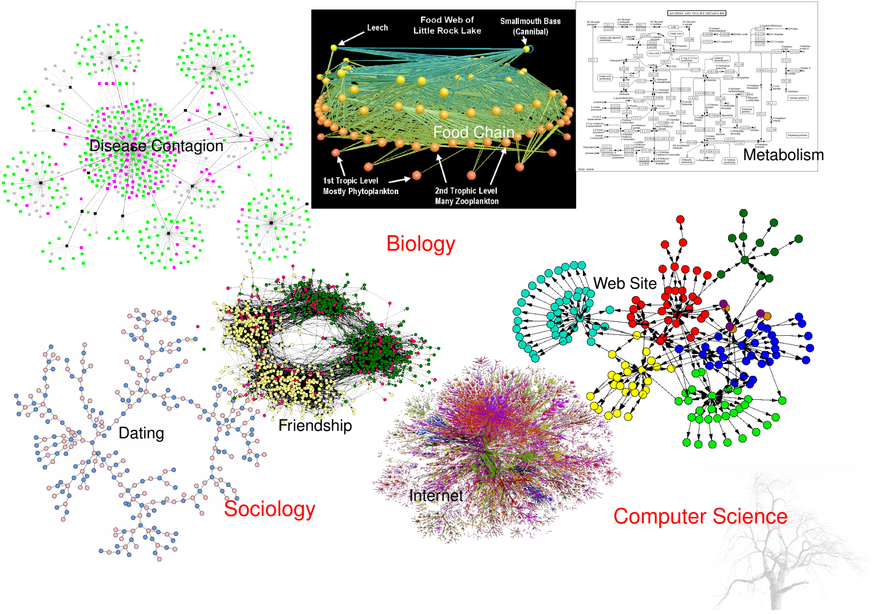
"real world" graphs

random sampling — applications

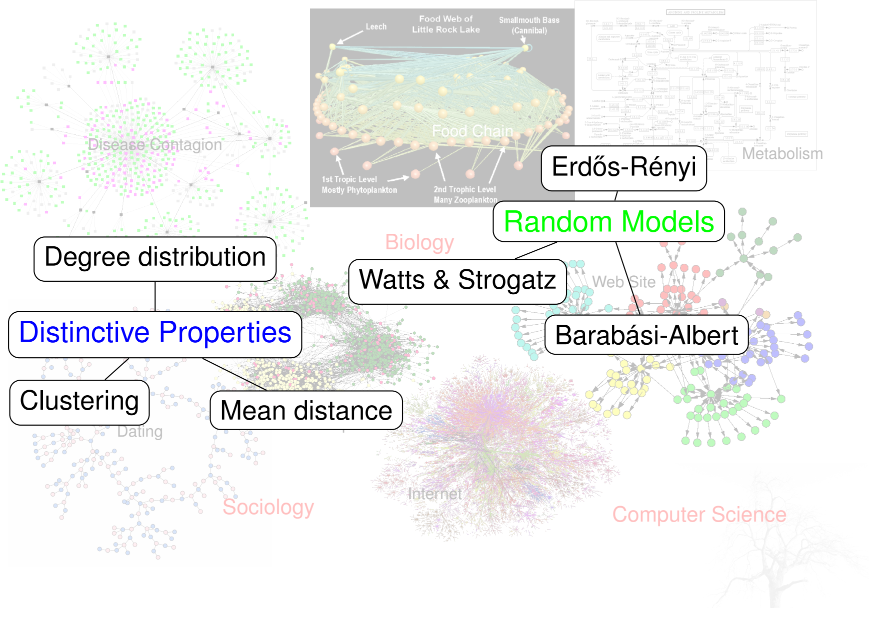
uniform & efficient

tree-like data struct.

“Real world” graphs



“Real world” graphs

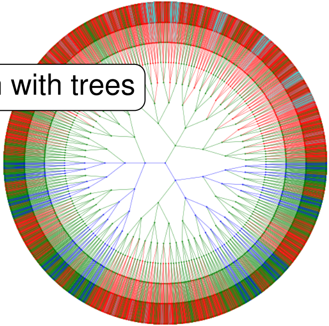


“Real world” graphs

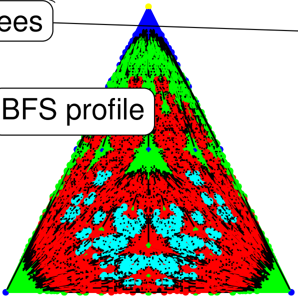
Apollonian networks

k-trees

Bijection with trees



BFS profile

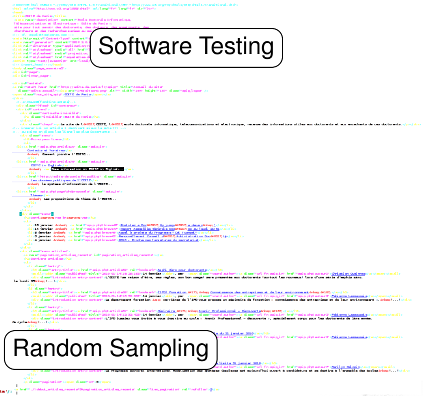


Analytic Combinatorics

Combinatorial Analysis

Tree-like data structures

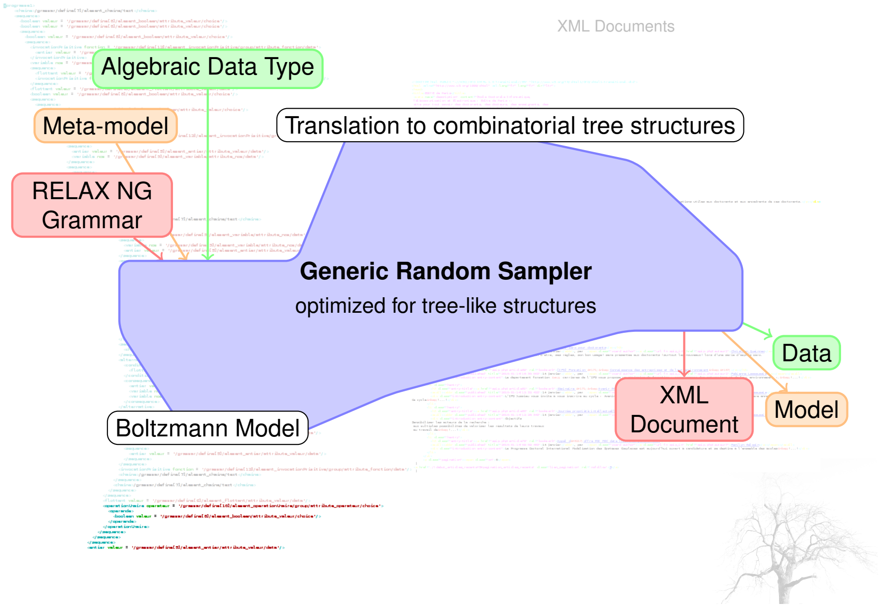
XML Documents



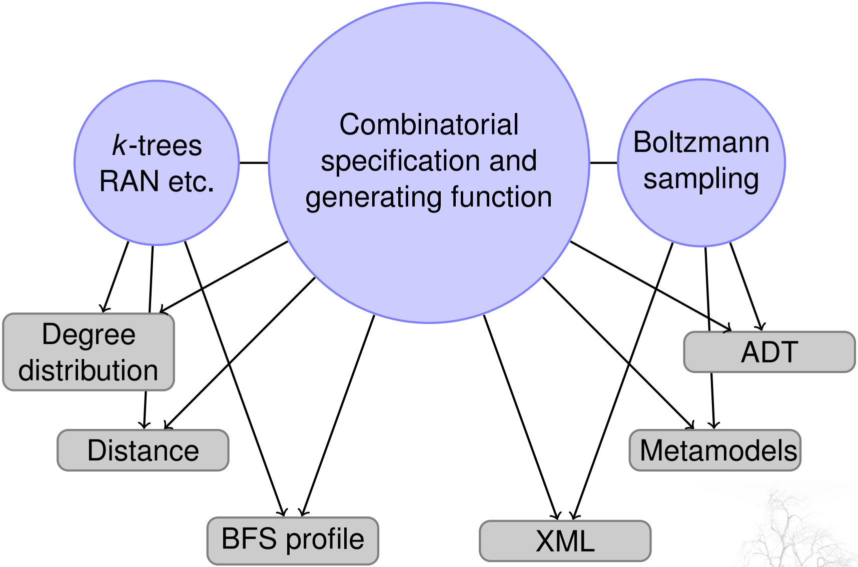
Software Testing

Random Sampling

Tree-like data structures



Plan



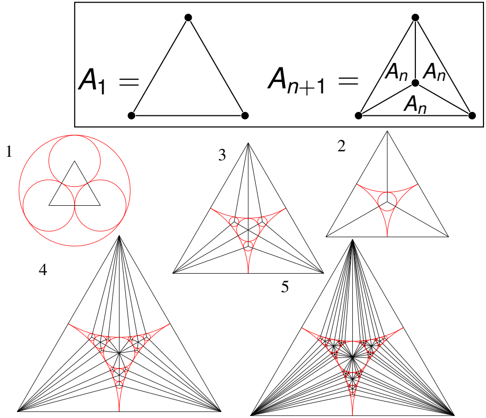
Thanks to H.-K. Hwang for this figure

First part

Analysis of k-trees

bijection between planar 3-trees and ternary trees
estimating distances in planar 2-trees
BFS-profile of general k-trees

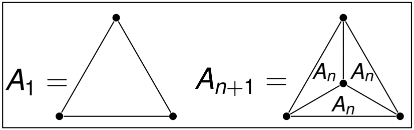
Apollonian networks (deterministic)

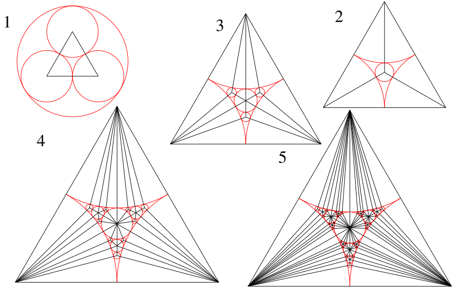


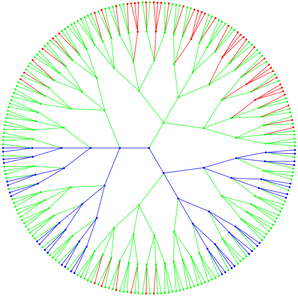
[*Andrade et al.* 05]

- Properties similar to “real world” graphs
- Inspired from the apollonian packings

Apollonian networks (deterministic)

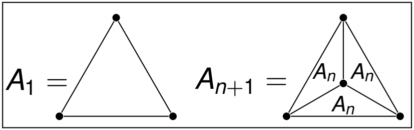


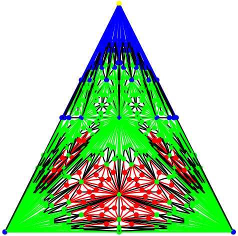


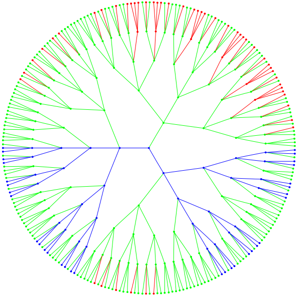


The (trivial) bijection with ternary trees can be used to study distances to an external vertex.

Apollonian networks (deterministic)

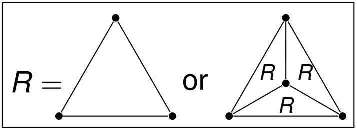




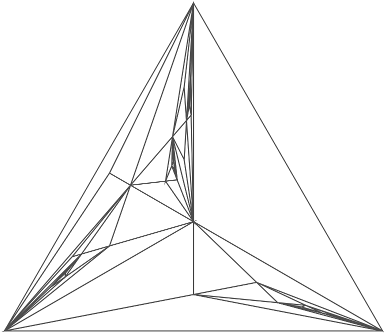


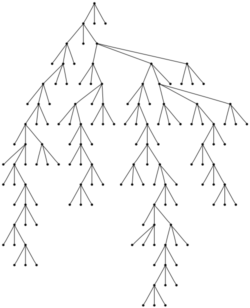
The (trivial) bijection with ternary trees can be used to study distances to an external vertex.

Random Apollonian networks (Planar 3-trees, Stack triangulations)

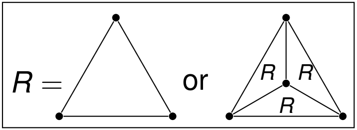


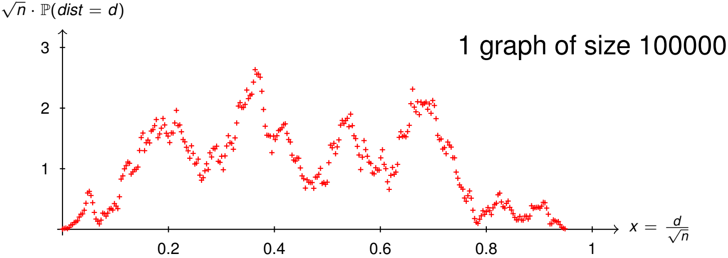
[*Zhou et al.* 05]



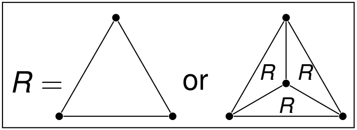


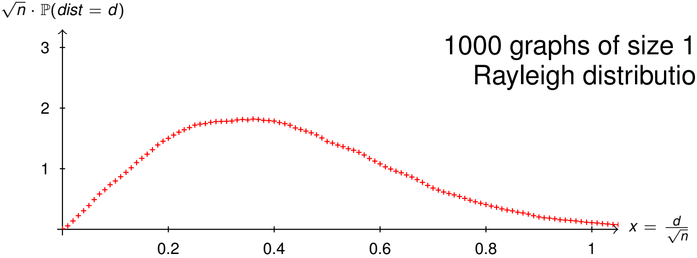
Random Apollonian networks (Planar 3-trees, Stack triangulations)





Random Apollonian networks (Planar 3-trees, Stack triangulations)

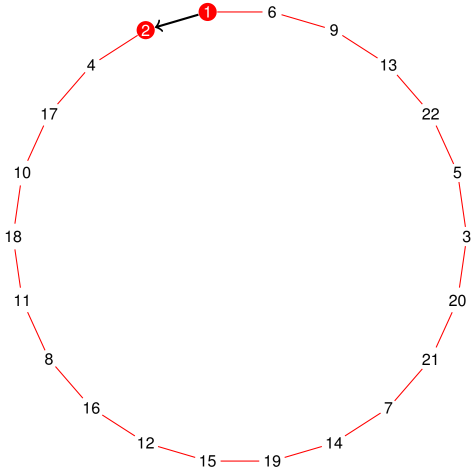




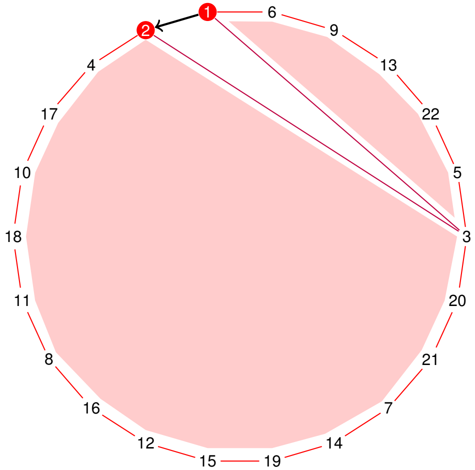
[A.D., *Soria* 07] [*Bodini*, A.D., *Soria* 08]

[*Albenque*, *Marckert* 08]

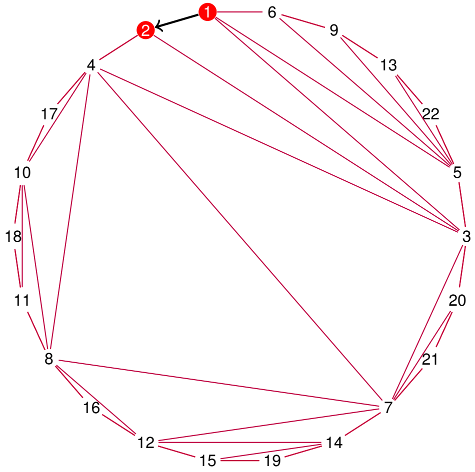
Planar 2-trees ≡ triangulations of a regular polygon



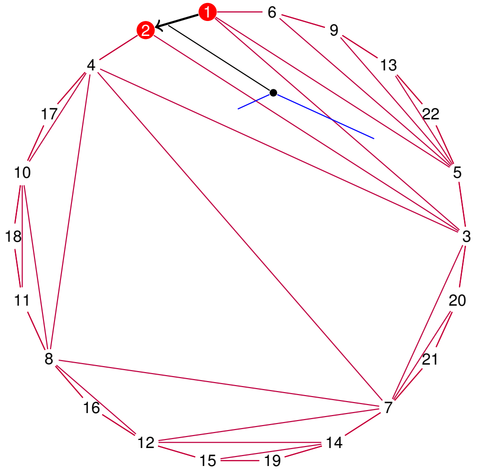
Planar 2-trees $\equiv$ triangulations of a regular polygon

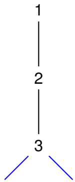


Planar 2-trees $\equiv$ triangulations of a regular polygon

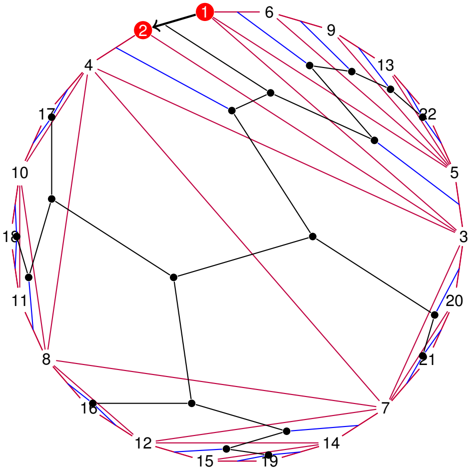


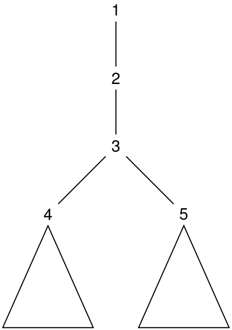
Planar 2-trees $\equiv$ triangulations of a regular polygon



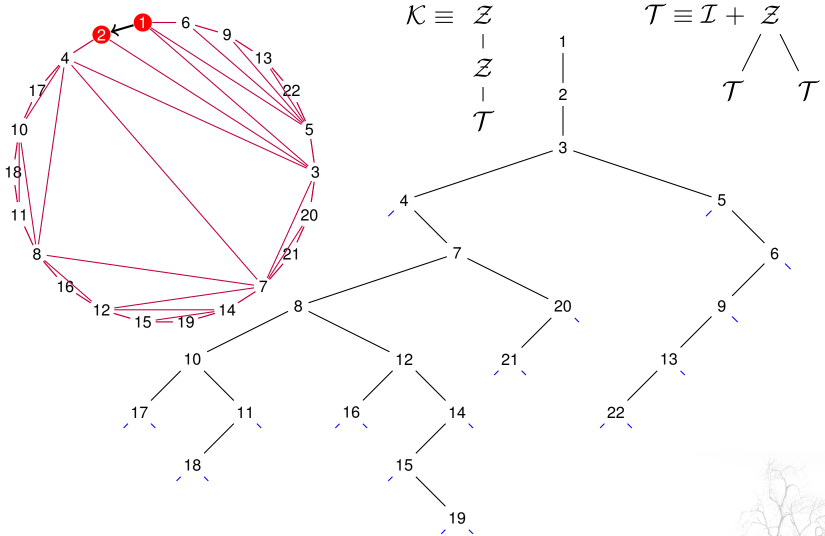


Planar 2-trees ≡ triangulations of a regular polygon

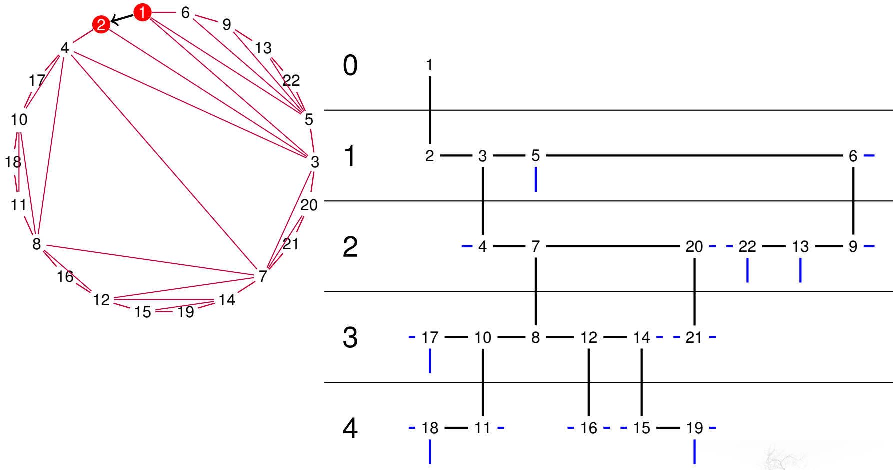




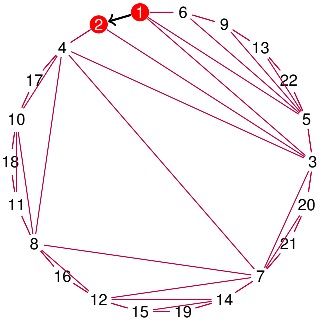
Planar 2-trees $\equiv$ triangulations of a regular polygon

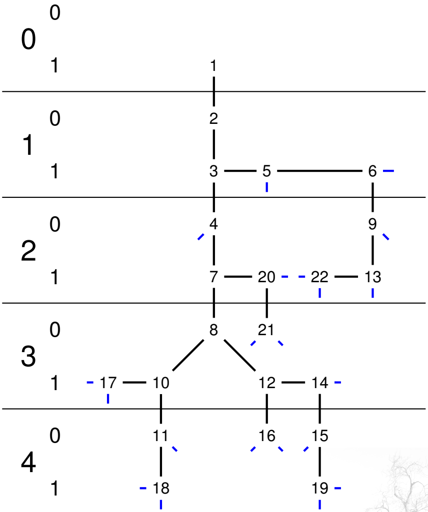


Planar 2-trees ≡ triangulations of a regular polygon



Planar 2-trees $\equiv$ triangulations of a regular polygon





$$\mathcal{K} \equiv \begin{array}{c}\mathcal{Z}\\ | \\ \mathcal{Z} \\ | \\ \mathcal{T}_{1,1}\end{array}$$

$$\mathcal{T}_{d,0} \equiv \mathcal{I} + \begin{array}{c}\mathcal{Z}_d \\ \diagup \quad \diagdown \\ \mathcal{T}_{d,1} \qquad \mathcal{T}_{d,1}\end{array}$$

$$\mathcal{T}_{d,1} \equiv \mathcal{I} + \begin{array}{l}\mathcal{Z}_d - \mathcal{T}_{d,1} \\ \ | \\ \mathcal{T}_{d+1,0}\end{array}$$

Planar 2-trees $\equiv$ triangulations of a regular polygon

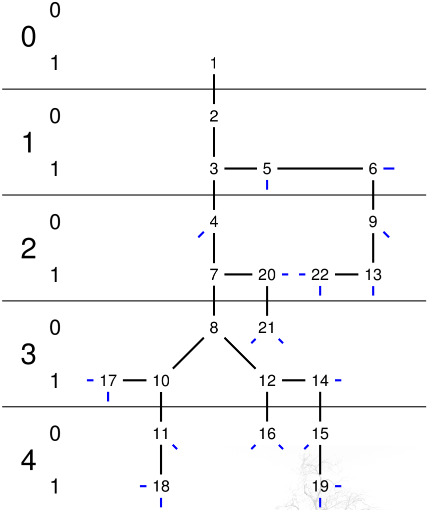
$$K_d(z,u) = \sum_{n,m} r_{n,m} u^m z^n$$

$r_{n,m}$: # of rooted k-trees with n total vertices and m vertices at distance d

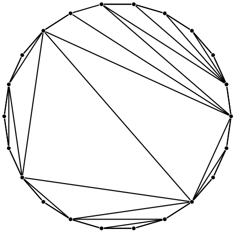
$$\frac{\partial}{\partial u} K_d(z,u)\Big|_{u=1} = \sum_n r_n z^n$$

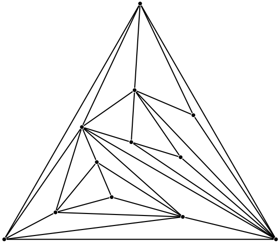
r_n: # of vertices at distance d in all rooted k-trees with n total vertices

$$K_d(z,u) = z^2 T_{d-1,1}(z,u)$$
$$T_{d,1}(z,u) = 1 + zT_{d,1}(z,u)T_{d-1,0}(z,u)$$
$$T_{d,0}(z,u) = 1 + zT^2_{d,1}(z,u)$$
$$T_{0,1}(z,u) = 1 + zT_{0,1}(z,u)T(z)$$
$$T_{0,0}(z,u) = 1 + zT^2_{1,1}(z,u)$$



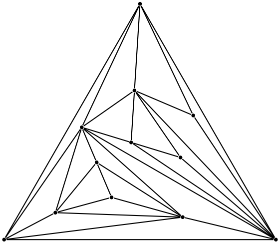
Generalization to planar k-trees



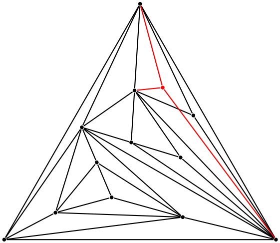


$k = 2$	$k = 3$	any k
binary trees	ternary trees	k-ary trees
two sub-levels	three sub-levels	k sub-levels

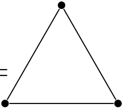
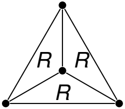
Generalization to (non-planar) k-trees



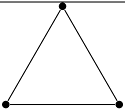
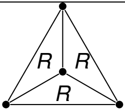
planar

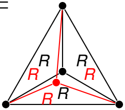


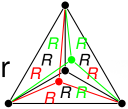
general

$R =$  or 

$$\mathcal{T} = \mathcal{I} + \mathcal{Z} \star \mathcal{T}^k$$

$R =$  or 

or 

or 

or …

$$\mathcal{T} = \textsc{Set}(\mathcal{Z} \star \mathcal{T}^k)$$

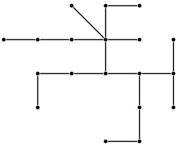
k-trees

Definition [*Beineke, Pippert* 69] ($k = 2$ [*Harary, Palmer* 68])

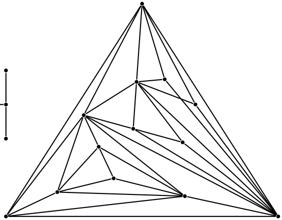
A k-tree is:

- either a k-clique,
- or a k-tree with one of its k-cliques connected to a new vertex.

$k = 1$



$k = 3$



Some graph theory NP-hard problems have linear time algorithms on partial k-trees [*Arnborg, Proskurowksi* 89]

Generating function for distances

$K(z) = \sum_n K_n z^n$

$$K(z) = z^k T(z)$$
$$T(z) = \exp(zT^k(z))$$

$K_d(z,u) = \sum_{n,m} r_{n,m} u^m z^n$

$$\begin{aligned} K_d(z,u) &= z^k T_{d,1}(z,u) \\ T_{d,i}(z,u) &= \exp(zT^{k-i}_{d,i}(z,u)T^i_{d,i+1}(z,u)) \\ T_{d,d}(z,u) &= T_{d-1,0}(z,u) \end{aligned}$$

$$\begin{aligned} T_{0,i}(z,u) &= \exp(uzT^{k-i}_{0,i}(z,u)T^i_{0,i+1}(z,u)) \\ T_{0,d-1}(z,u) &= \exp(uzT^{k-1}_{0,d-1}(z,u)T(z)) \end{aligned}$$

$\frac{1}{nK_n}[z^n] \; \frac{\partial}{\partial u} K_d(z,u)\Big|_{u=1} =$

mean # of vertices at distance d

Generating function for distances

$K(z) = \sum_n K_n z^n$

$$K(z) = z^k T(z)$$
$$T(z) = \exp(zT^k(z))$$

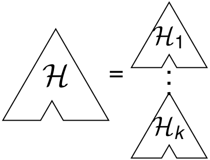
$K_d(z,u) = \sum_{n,m} r_{n,m} u^m z^n$

$$K_d(z,u) = z^k T_{d,1}(z,u)$$
$$T_{d,i}(z,u) = \exp(zT_{d,i}^{k-i}(z,u)T_{d,i+1}^{i}(z,u))$$
$$T_{d,d}(z,u) = T_{d-1,0}(z,u)$$

$$T_{0,i}(z,u) = \exp(uzT_{0,i}^{k-i}(z,u)T_{0,i+1}^{i}(z,u))$$
$$T_{0,d-1}(z,u) = \exp(uzT_{0,d-1}^{k-1}(z,u)T(z))$$

$\frac{1}{nK_n}[z^n]\ \frac{\partial}{\partial u}K_d(z,u)\Big|_{u=1} =$
mean # of vertices at distance d

$$\frac{\partial}{\partial u}K_d(z,u)\Big|_{u=1} = H^{d-2}(z)\ \frac{\partial}{\partial u}K_2(z,u)\Big|_{u=1}$$



combinatorial interpretation

$$\begin{aligned} H(z) &= k!(zT^k(z))^k \prod_{i=1}^{k-1} \frac{1}{1 - izT^k(z)} \\ &= 1 - c_k\sqrt{2(1-kez)} + O(1-kez). \end{aligned}$$

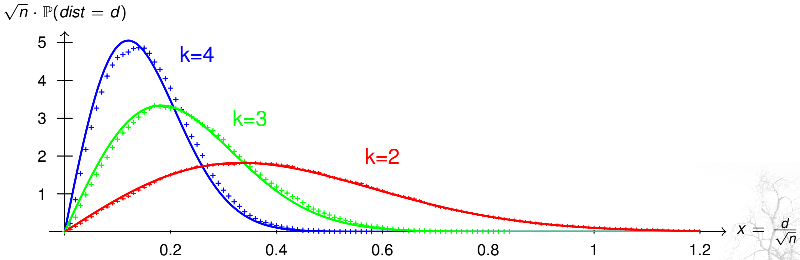
Semi-large power theorem for calculating $[z^n]H^{d-2}(z)$ in the range $x\sqrt{n}$.

Main Result

Theorem [A.D., *Soria* 09] (RAN [*Bodini*, A.D., *Soria* 08])

Given a rand. k-tree G over n vert., the distance between the root vertex r and a random vertex v of G has asymptotic mean value of order $\sqrt{n}$ and is Rayleigh distributed in the range $x\sqrt{n}$:

$$\sqrt{n} \cdot \lim_{n\to\infty} \mathbb{P}(D(r,v) = \lfloor x\sqrt{n} \rfloor) = c_k^2 x e^{-\frac{(c_k x)^2}{2}}, \text{ with } c_k = k \sum_{i=1}^{k} \frac{1}{i}$$



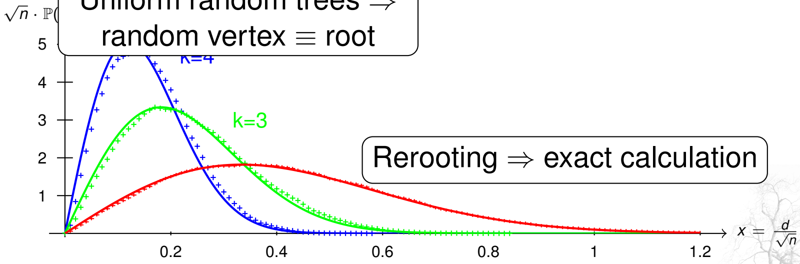
Main Result

Theorem [A.D., *Soria* 09] (RAN [*Bodini*, A.D., *Soria* 08])

Given a rand. k-tree … ance between the root vertex r and a r… as asymptotic mean value of order $\sqrt{n}$ ar… d in the range $x\sqrt{n}$:

What about the distance between 2 random vertices?

$$\sqrt{n} \cdot \lim_{n \to \infty} \mathbb{P}(D(r, v) = \lfloor x\sqrt{n} \rfloor) = c_k^2 x e^{-\frac{(c_k x)^2}{2}}, \text{ with } c_k = k \sum_{i=1}^{k} \frac{1}{i}$$



Uniform random trees $\Rightarrow$ random vertex $\equiv$ root

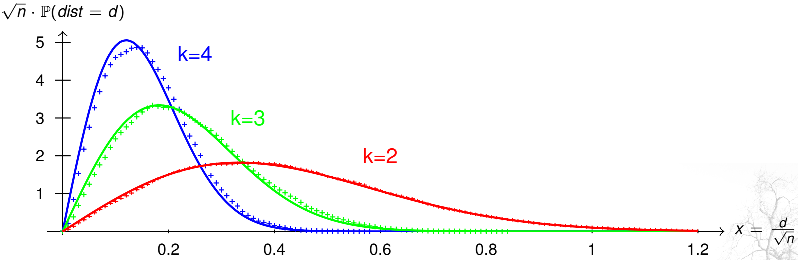
Rerooting $\Rightarrow$ exact calculation

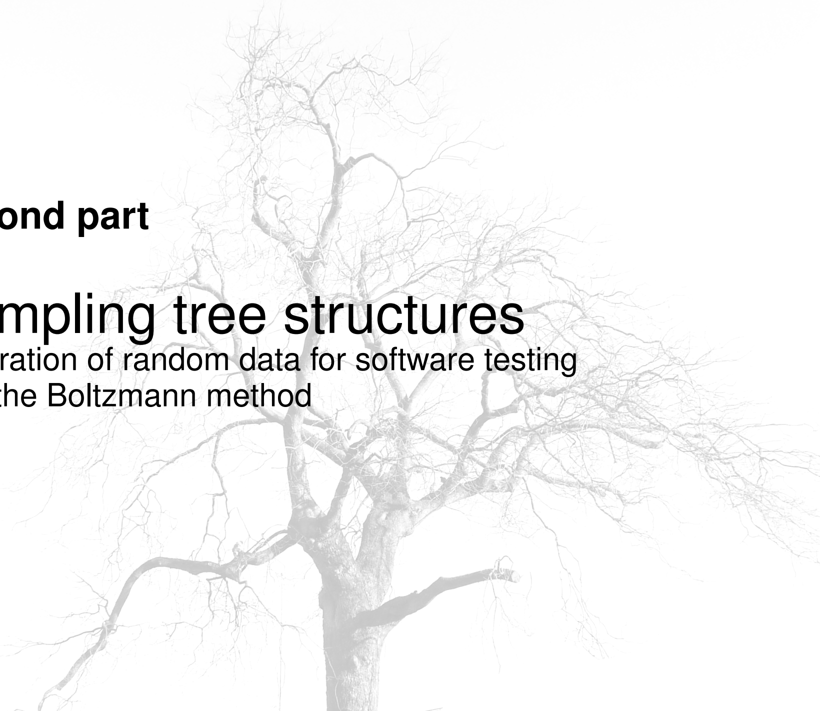
Main Result

Theorem [A.D., *Soria* 09]

Given a rand. k-tree G over n vert., the distance between the two random vertices v, w of G has asymptotic mean value of order $\sqrt{n}$ and is Rayleigh distributed in the range $x\sqrt{n}$:

$$\sqrt{n} \cdot \lim_{n\to\infty} \mathbb{P}(D(v,w) = \lfloor x\sqrt{n} \rfloor) = c_k^2 x e^{-\frac{(c_k x)^2}{2}}, \text{ with } c_k = k \sum_{i=1}^{k} \frac{1}{i}$$



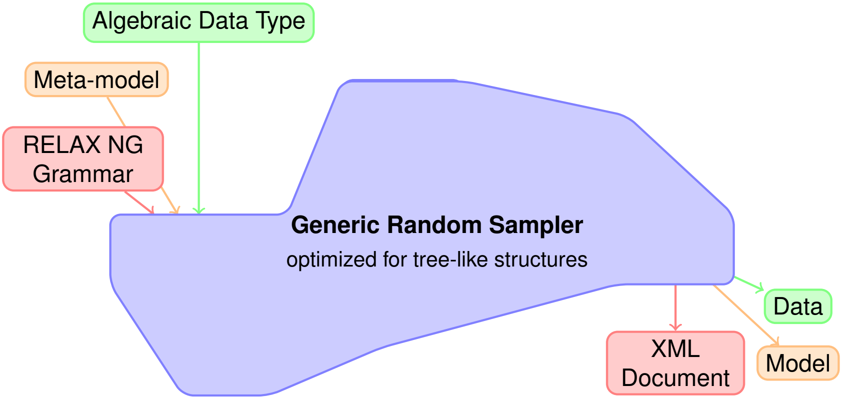


Second part

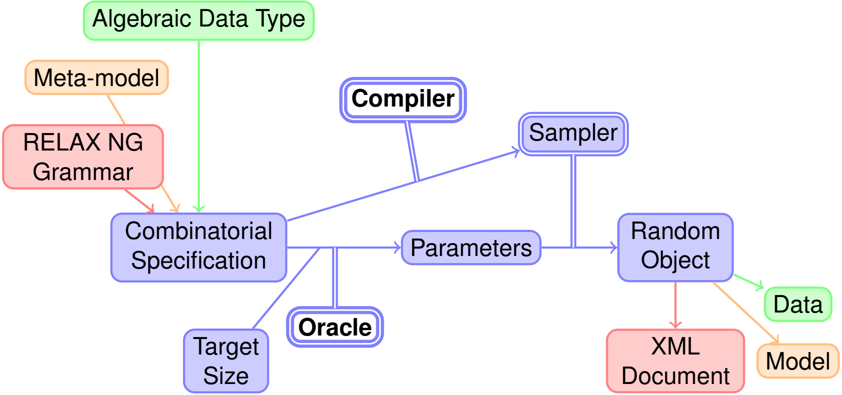
Sampling tree structures

generation of random data for software testing
with the Boltzmann method

Tree-like data structures



Tree-like data structures

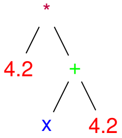


Generic oracle in Maple by Pivoteau

Optimized oracle for tree structures in C

Algebraic Data Types

```
type expression =
    Const of float
  | Var of string
  | Sum of expression * expression
  | Prod of expression * expression
```



$$\mathcal{T} = \mathcal{Z} + \mathcal{Z} + \mathcal{T} \star \mathcal{T} + \mathcal{T} \star \mathcal{T}$$

Boltzmann sampler ($a = 0.25, b = 0.5, c = 0.75$)

```
let rec rand_exp = function (a,b,c) as v ->
  let r = Random.float 1.0 in
  if      r < a then Const 4.2
  else if r < b then Var "x"
  else if r < c then Sum (rand_exp v, rand_exp v)
  else               Prod (rand_exp v, rand_exp v)
;;
```

Boltzmann sampling

[*Duchon, Flajolet, Louchard, Schaeffer* 04]

Simple translation from specification to probabilistic algorithm

construction	Boltzmann sampler with param. x
$\mathcal{C} = \mathcal{I}$	$\Gamma C(x) := \varepsilon$
$\mathcal{C} = \mathcal{Z}$	$\Gamma C(x) := z$
$\mathcal{C} = \mathcal{A} + \mathcal{B}$	$\Gamma C(x) :=$ Bern $\frac{A(x)}{C(x)} \longrightarrow \Gamma A(x) \mid \Gamma B(x)$
$\mathcal{C} = \mathcal{A} \times \mathcal{B}$	$\Gamma C(x) := \langle\, \Gamma A(x)\ ;\ \Gamma B(x)\, \rangle$

Boltzmann sampling

[*Duchon, Flajolet, Louchard, Schaeffer* 04]

$\mathcal{C}$ set of objects γ with size function $|\cdot| : \mathcal{C} \to \mathbb{N}$

Boltzmann model, parameter x

$$\mathbb{P}_x(\gamma) = \frac{x^{|\gamma|}}{C(x)} \quad \text{where} \quad C(x) = \sum_{\gamma \in \mathcal{C}} x^{|\gamma|}$$

"Oracle" for computing $C(x)$ for $x < \rho_{\mathcal{C}}$
[*Pivoteau, Salvy, Soria* 08]

Properties

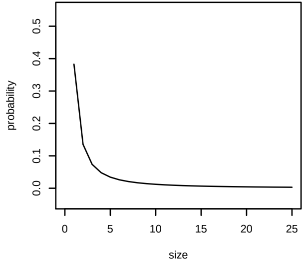
- Guarantees uniformity (between objects of the same size)
- Output size is random, but can be tuned with parameter x
- Linear complexity of sampling (for approximate size) $\to$ ability to generate huge objects

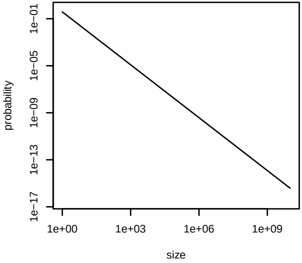
Boltzmann sampling of trees

[*Duchon, Flajolet, Louchard, Schaeffer* 04]

Singular sampler $\Rightarrow$ power law

$$\mathbb{P}(|\gamma| = n) \sim cn^{-3/2}$$



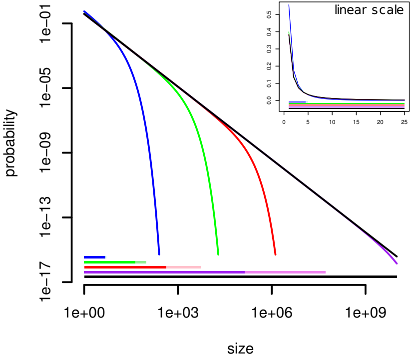


Ceiled rejection $\Rightarrow$ still linear

Boltzmann sampling of trees

Approximation of $\rho \Rightarrow$ power law with exp. cutoff

$$\mathbb{P}(|\gamma| = n) \sim c' \left(\frac{x}{\rho}\right)^n n^{-3/2}$$



Still linear for small enough n

$x = (1 - 10^{-10})\rho \Rightarrow$ trees of size 10^8

Algebraic Data Types [*Canou*, A.D. 09]

Very natural translation from type definition to combinatorial specification.

$$\begin{aligned}
\phi(\textsf{int}), \phi(\textsf{bool}), \ldots &= \mathcal{Z} \\
\phi(t) &= \mathcal{X}_t \\
\phi(t\ \textsf{list}) &= \mathcal{Z} \star \textsc{Seq}(\mathcal{Z} \star \phi(t)) \\
\phi(t_1 * \ldots * t_n) &= \mathcal{Z} \star \phi(t_1) \star \ldots \star \phi(t_n) \\
\phi(\textsf{A1 of } t_1 \mid \ldots \mid \textsf{An of } t_n) &= \mathcal{Z} \star (\phi(t_1) + \ldots + \phi(t_n))
\end{aligned}$$

XML documents following a RELAX NG grammar

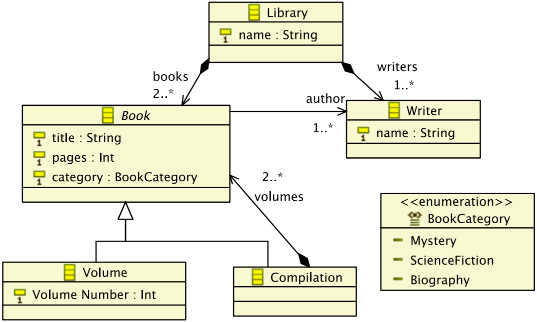
Very large specifications, interesting for testing and benchmarking the oracle.

Grammar	nb eqs	max deg	nb sols	Generic[1] $C(x)$	Tree optim.[2] ρ and $C(\rho)$
rss	10	5	2	0.02s	0.02s
PNML	22	4	4	0.05s	0.26s
xslt	40	3	10	0.4s	2.5s
relaxng	34	4	32	0.4s	1.8s
xhtml-basic	53	3	13	1.2s	4s
mathml2	182	2	18	3.7s	18s
docbook	407	11		67.7s	66s

[1] Evaluation of $C(x)$ using Pivoteau's Maple oracle, made by Salvy
[2] Estimation of ρ and evaluation of $C(\rho)$ using optimized oracle in C

Meta-models [*Mougenot*, A.D., *Blanc, Soria* 09]

Initiated and developed by people in the application's domain.



Model size (10% margin)	Av. simulation time	Building time
50 000	0.501s	9.87s
100 000	0.934s	26.0s
250 000	2.48s	63.2s
1 000 000	8.86s	N/A

Perspectives

k-trees

Another random model: increasing trees.

(joint work with Bodini, Hwang, Soria)

Extension to chordal graphs.

Random Sampling

Go from proof of concept to finished product.

Multi-parameter sampling. (using work by Bodini, Ponty)

Add random generating feature to the Encyclopedia of Combinatorial Structures. (ECS by INRIA Algorithms)