Cherries, trees, and cherries without trees

Giacomo Plazzotta

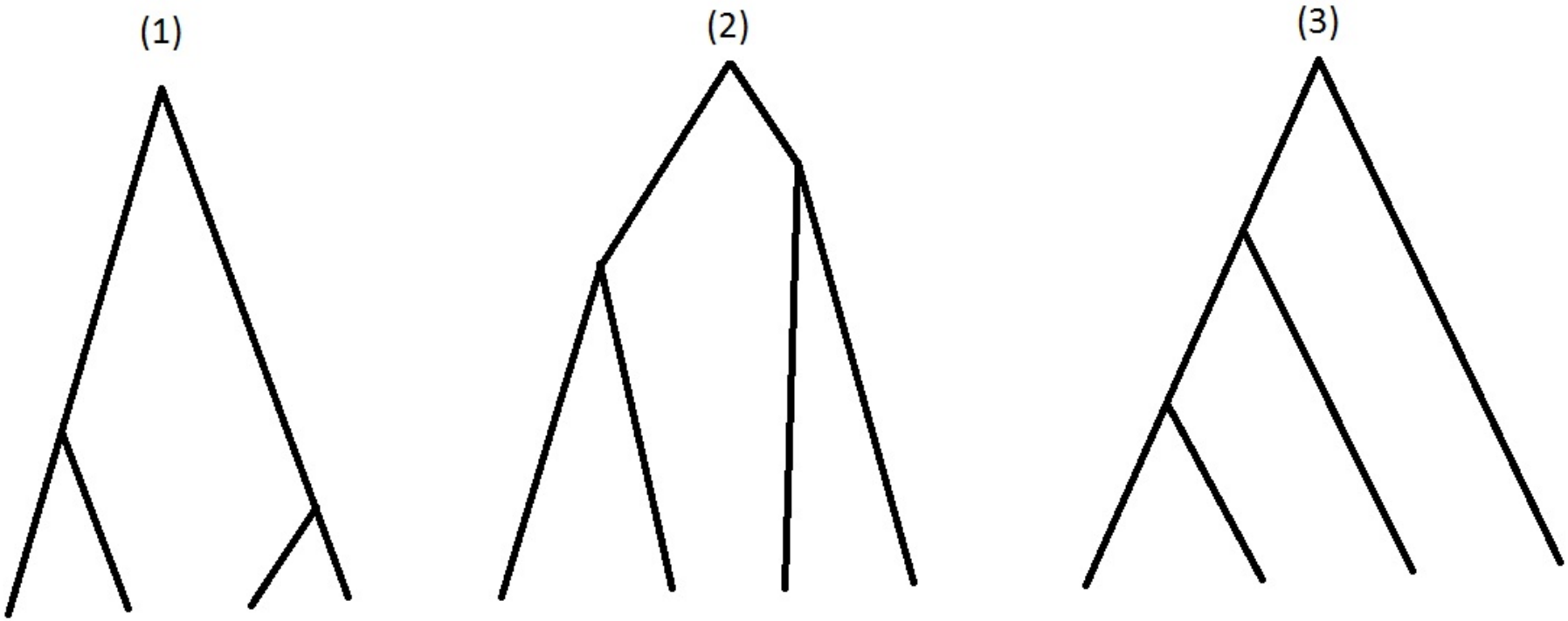
Banff International Research Station

14th February 2017

Imperial College London

Caroline Colijn

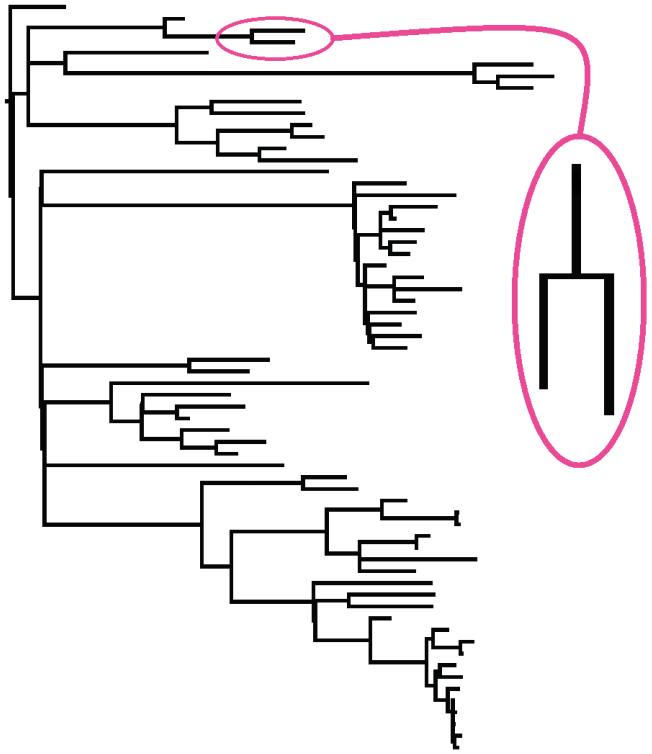
What is a shape?



Not rigorously, the shape $\mathcal{S}$ of a tree or subtree is the tree or subtree without the associated branch lengths

What is a shape frequency?

The frequency of a shape $\mathcal{S}$ in a tree is the ratio between the number of occurrences of $\mathcal{S}$ in the tree and the number of tips of the tree.



A cherry

Yule tree (McKenzie&Steel 2000)

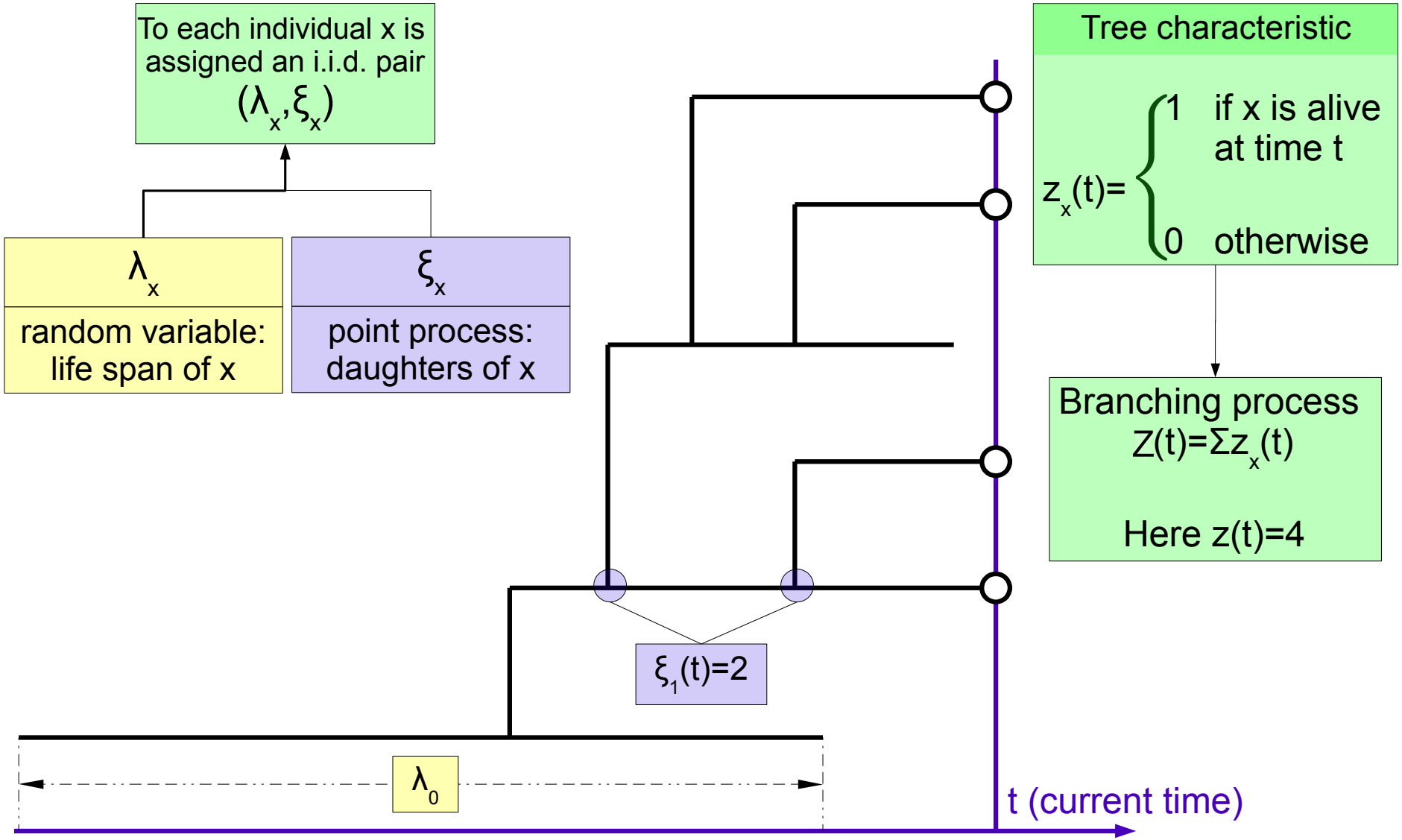
When the number of tips $n \geq 3$:

$$E[cherries] = \frac{n}{3}$$

$$var(cherries) = \frac{2n}{45}$$

Yule tree = a tree with no death and constant birth rate

General setting (Jagers, 1975)

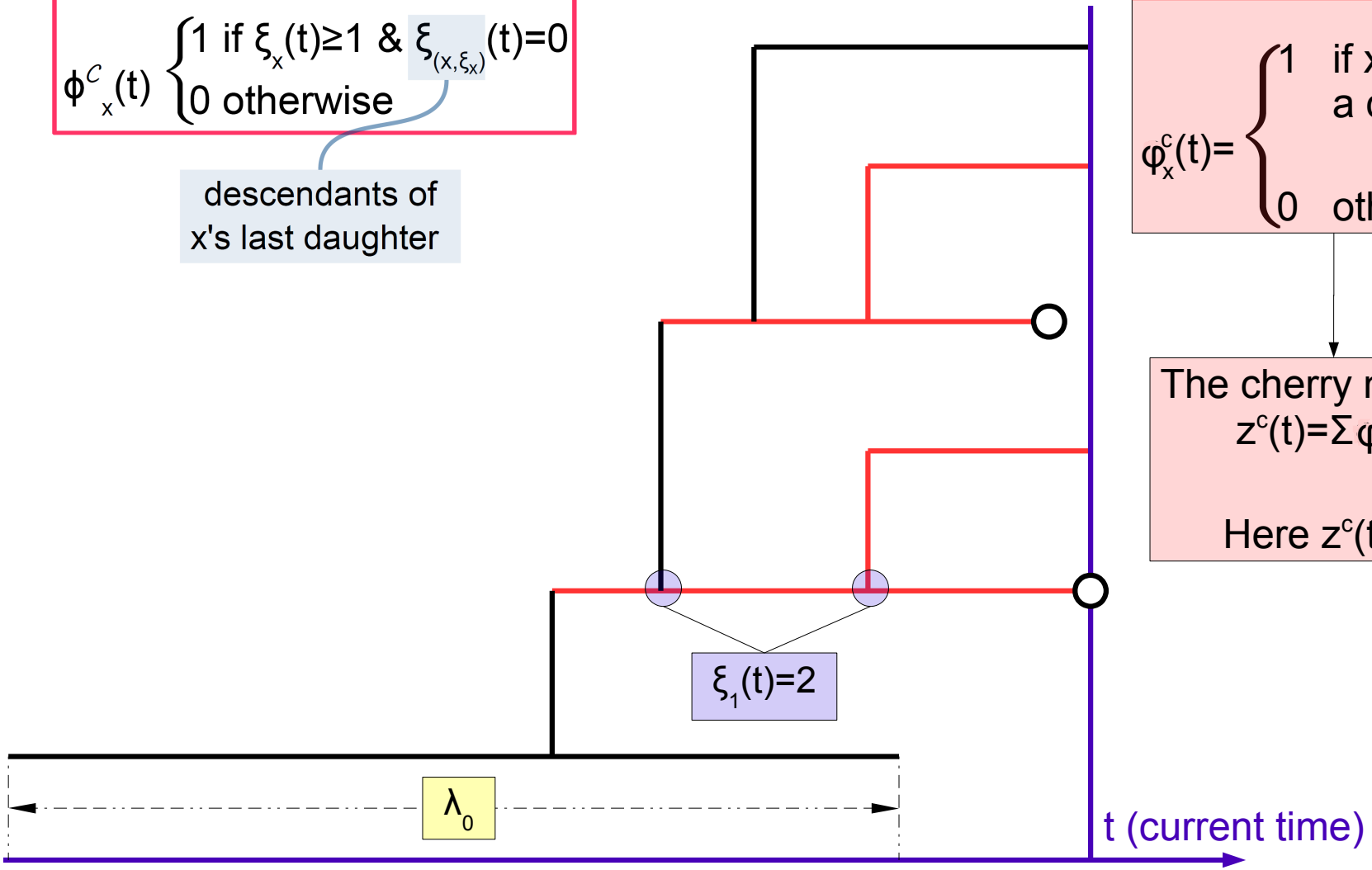


Counting shapes

Cherry characteristic:

$$\phi^{C}_{x}(t) \begin{cases} 1 \text{ if } \xi_x(t)\geq 1 \ \& \ \xi_{(x,\xi_x)}(t)=0 \\ 0 \text{ otherwise} \end{cases}$$

descendants of x's last daughter



Cherry characteristic

$$\varphi^{c}_{x}(t)= \begin{cases} 1 & \text{if x fathers a cherry} \\ 0 & \text{otherwise} \end{cases}$$

The cherry number

$z^c(t)=\Sigma\varphi^c_x(t)$

Here $z^c(t)=2$

Convergence of the shape frequency

Henceforth supercritical branching processes i.e. $E[\xi(\infty)] > 1$. Nerman (1981) proves a fundamental convergence:

(Nerman, 1981)

Given two characteristics ϕ^1, ϕ^2 (with some properties..) then:

$$\frac{Z^{\phi^1}(t)}{Z^{\phi^2}(t)} \xrightarrow{t\to\infty} \frac{E[Z^{\phi^1}(t)]}{E[Z^{\phi^2}(t)]} \xrightarrow{t\to\infty} \frac{\int_0^\infty \mathrm{e}^{-Mt} E[\phi^1(t)]dt}{\int_0^\infty \mathrm{e}^{-Mt} E[\phi^2(t)]dt},$$

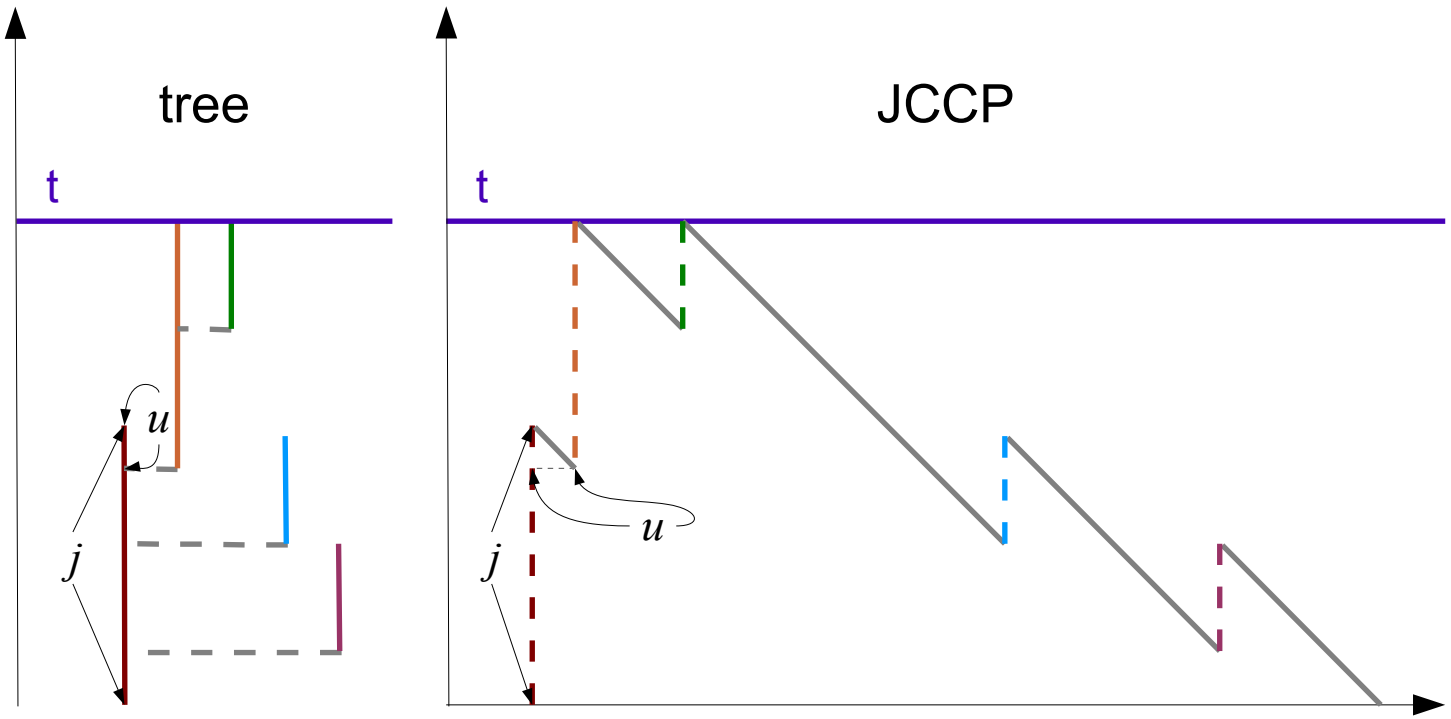
where M is the Malthusian parameter, i.e.: $\int_0^\infty \mathrm{e}^{-Mt}\mu(dt) = 1$.

In our case:

$$\lim_{t\to\infty} \frac{\text{Occurrences of } \mathcal{S}}{\text{Tips}} = M \int_0^\infty \mathrm{e}^{-Mt} E[\phi^{\mathcal{S}}(t)]dt.$$

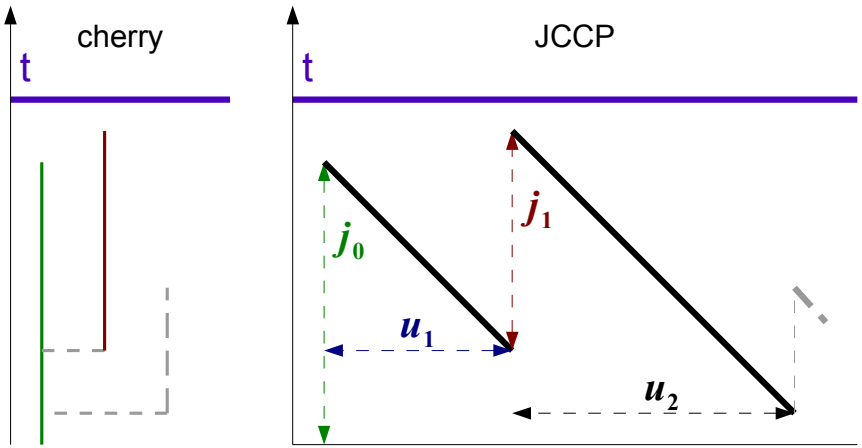
$E[\phi^{\mathcal{S}}(t)]$ is the probability that at t the ancestor has fathered a shape and is very hard to derive

Jumping Chronological Contour Process

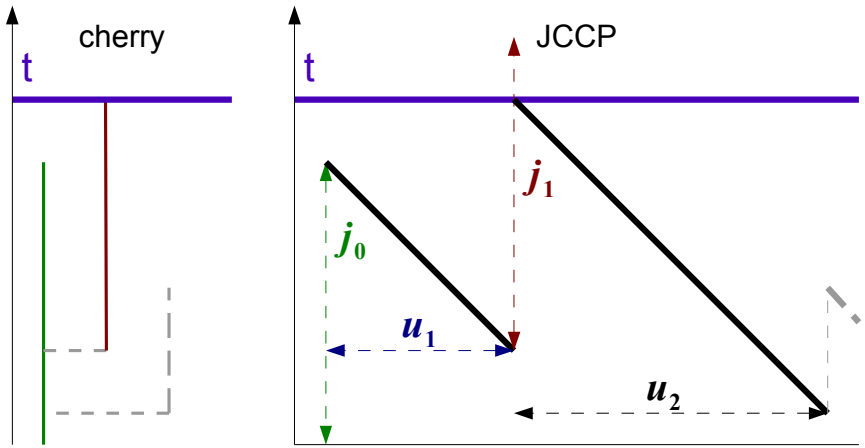


"This process can be seen as the path of a ball that follows an outline of the oriented tree, decreasing at unit speed along its edges and jumping instantaneously to the tip of the daughter edge when reaching a node." (Lambert, Alexander, Stadler, 2014)

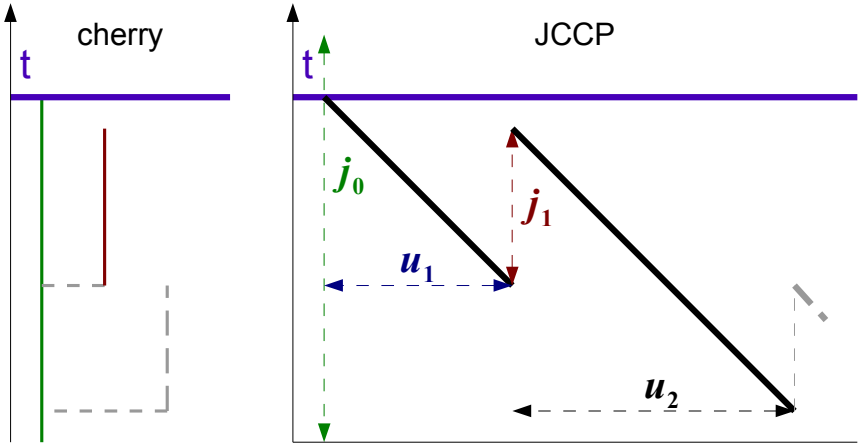
Evaluation of $E[\phi^{\mathcal{S}}(t)]$ in homogeneous trees



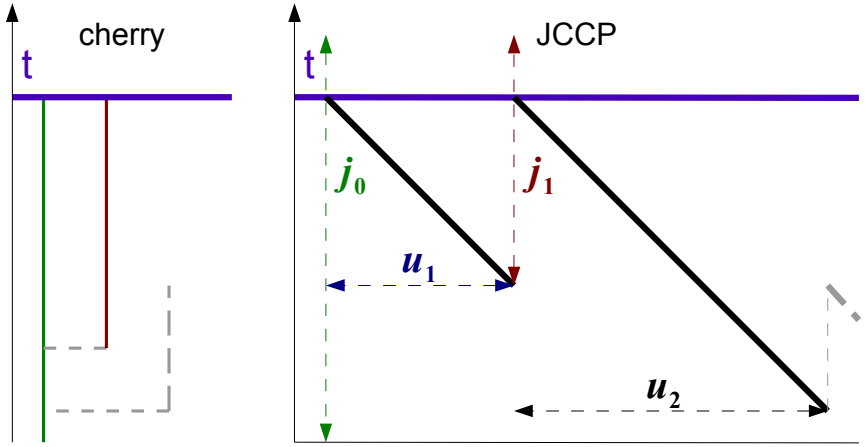
$$\int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{\infty} \left(\delta\, \mathrm{e}^{-\delta j_0}\right)\left(\beta\, \mathrm{e}^{-\beta u_1}\right)\left(\delta\, \mathrm{e}^{-\delta j_1}\right)\left(\beta\, \mathrm{e}^{-\beta u_2}\right) \mathrm{d}\, u_2\, \mathrm{d}\, j_1\, \mathrm{d}\, u_1\, \mathrm{d}\, j_0$$



$$\int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^{\infty} \left(\delta\, \mathrm{e}^{-\delta j_0}\right)\left(\beta\, \mathrm{e}^{-\beta u_1}\right)\left(\delta\, \mathrm{e}^{-\delta j_1}\right)\left(\beta\, \mathrm{e}^{-\beta u_2}\right) \mathrm{d}\, u_2\, \mathrm{d}\, j_1\, \mathrm{d}\, u_1\, \mathrm{d}\, j_0$$



$$\int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{\infty} \left(\delta\, \mathrm{e}^{-\delta j_0}\right)\left(\beta\, \mathrm{e}^{-\beta u_1}\right)\left(\delta\, \mathrm{e}^{-\delta j_1}\right)\left(\beta\, \mathrm{e}^{-\beta u_2}\right) \mathrm{d}\, u_2\, \mathrm{d}\, j_1\, \mathrm{d}\, u_1\, \mathrm{d}\, j_0$$



$$\int_t^{\infty} \int_0^t \int_{u_1}^{\infty} \int_t^{\infty} \left(\delta\, \mathrm{e}^{-\delta j_0}\right)\left(\beta\, \mathrm{e}^{-\beta u_1}\right)\left(\delta\, \mathrm{e}^{-\delta j_1}\right)\left(\beta\, \mathrm{e}^{-\beta u_2}\right) \mathrm{d}\, u_2\, \mathrm{d}\, j_1\, \mathrm{d}\, u_1\, \mathrm{d}\, j_0$$

Results: cherries to tips ratio in homogeneous models

A homogeneous tree has constant birth rate (β) and death rate (δ).

The cherries to tips ratio

$$\lim_{t\to\infty} \frac{\text{Cherries}}{\text{Tips}} = \frac{\beta}{3\beta+\delta} = \frac{R_0}{3R_0+1}$$

With $R_0 = \beta/\delta$.

As $R_0 \to \infty$ (Yule tree): $CTR \to \frac{1}{3}$ as in (McKenzie&Steel 2000).

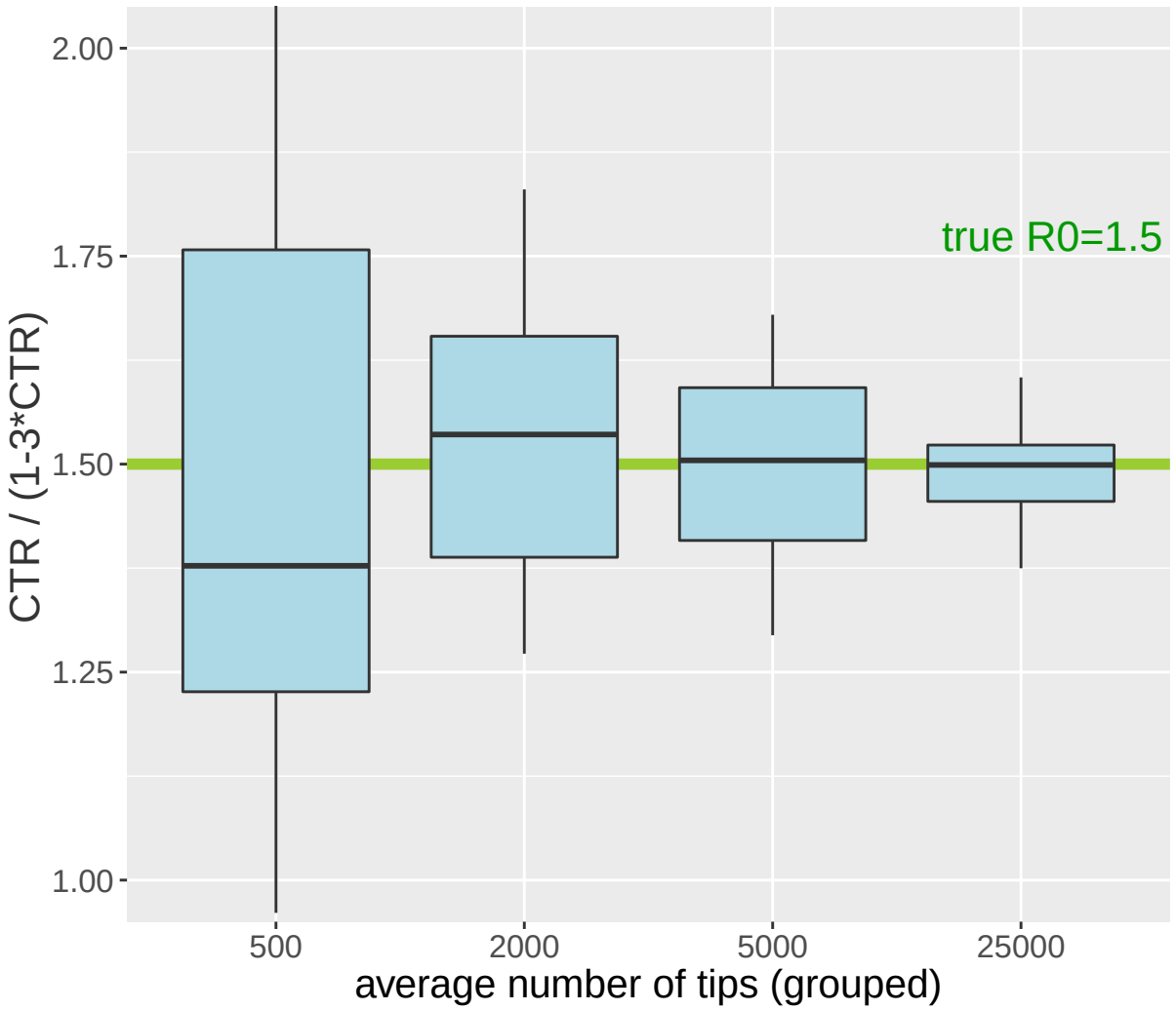
Remarks

Convergence is almost sure, tree is supercritical. For large trees:

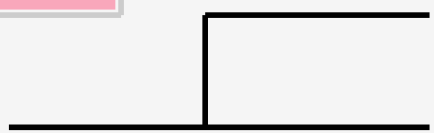
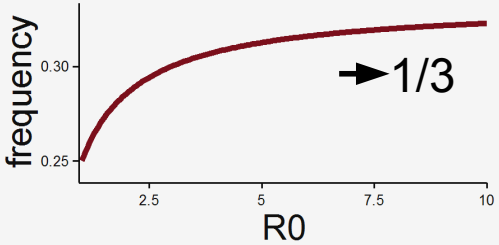
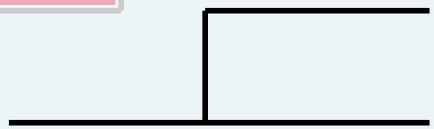
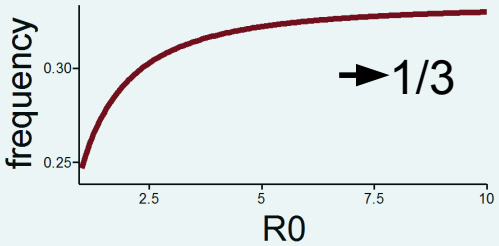
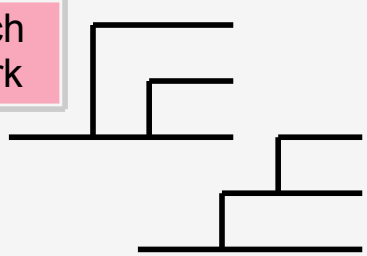
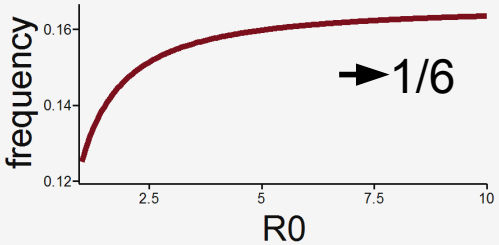
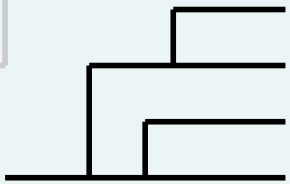
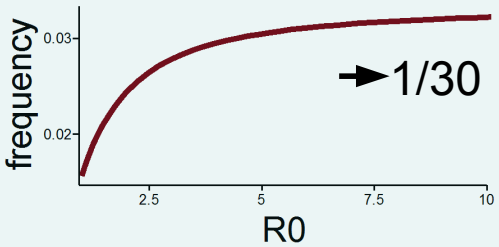
$$R_0 \approx \frac{\text{CTR}}{1-3\text{CTR}} \text{ when Tips} \gg 1$$

Implies a dynamic interpretation of R_0.

CTR/(1-3CTR) is close to R_0 in large trees



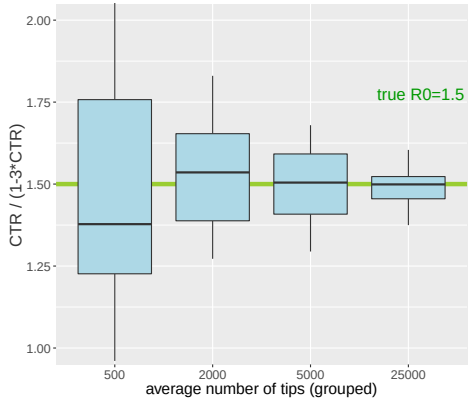
More results

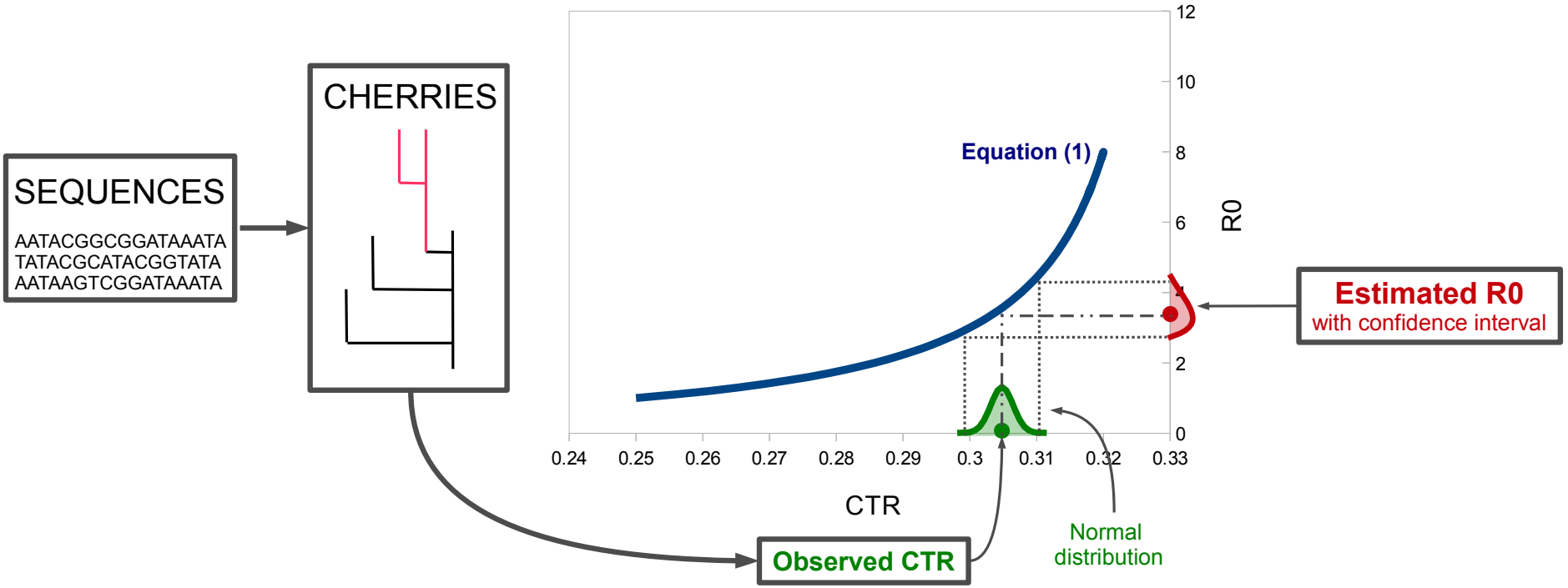
Configuration	Model	As. frequency	Plot and limit
cherry 	Homogeneous birth rate: β death rate: δ $R_0 = \beta/\delta$	$\frac{R_0}{3R_0+1}$	
cherry 	Non homogeneous constant birth rate β life span distribution: Gamma (rate=δ, shape=2) $R_0 = 2\beta/\delta$	$(512(243R_0^{12}+243R_0^{23/2}\sqrt{(R_0+8)}+5103R_0^{11}+ +4131R_0^{21/2}\sqrt{(R_0+8)}+40851R_0^{10}+26271R_0^{19/2}\sqrt{(R_0+8)}+ +160767R_0^{9}+80955R_0^{17/2}\sqrt{(R_0+8)}+338148R_0^{8}+ +131184R_0^{15/2}\sqrt{(R_0+8)}+387448R_0^{7}+112912R_0^{13/2}\sqrt{(R_0+8)}+ +235072R_0^{6}+49152R_0^{11/2}\sqrt{(R_0+8)}+68784R_0^{5}+ 9360R0^{9/2}\sqrt{(R_0+8)}+7616R_0^{4}+512R_0^{7/2}\sqrt{(R_0+8)}+128R_0^{3}))$ $((27R_0^{5}+27R_0^{9/2}\sqrt{(R_0+8)}+297R_0^{4}+189R_0^{7/2}\sqrt{(R_0+8)}+ +900R_0^{3}+360R_0^{5/2}\sqrt{(R_0+8)}+968R_0^{2}+240R_0^{(3/2)}\sqrt{(R_0+8)}+ +368R_0+48\mathrm{sqrt}(R_0+8)\sqrt{(R_0)}+32)^{-1}$ $(3R_0+\sqrt{(R_0+8)}\sqrt{(R_0)})^{-2}$ $(R_0+\sqrt{(R_0+8)}\sqrt{(R_0)})^{-2}(5R_0+\sqrt{(R_0+8)}\sqrt{(R_0)}+4)^{-3}$	
pitch -fork 	Homogeneous birth rate: β death rate: δ $R_0 = \beta/\delta$	$\frac{3R_0^2(R_0+1)}{(3R_0+1)^2(2R_0+1)}$	
double cherry 	Homogeneous birth rate: β death rate: δ $R_0 = \beta/\delta$	$1/4(2592R_0^9+11556R_0^8+18279R_0^7+ +13899R_0^6+4799R_0^5-65R_0^4- -546R_0^3-114R_0^2)$ $(19440R0^9+91044R0^8+ +187488R0^7+222741R0^6+ +168180R0^5+83666R0^4+ +27416R0^3+5705R0^2+ +684R0+36)^{-1}$	

Open questions

1. Do shape frequencies depend only on R_0?
2. Compute efficiently $E\left[\phi^{\mathcal{S}}(t)\right]$ in fully general processes?

R_0 inference from cherries - overview





R_0 estimate and confidence interval

- Lindeberg's CLT $\Rightarrow$ $CTR \approx$ Normal for large trees
- the variance of the CTR is bounded by $\frac{1}{4n}$

R_0 inference from CTR

$$R_0 \text{ estimate: } \frac{CTR}{1-3CTR}$$

$$R_0 \text{ CI: } \left[\frac{CTR-\frac{1}{2\sqrt{n}}}{1-3\left(CTR-\frac{1}{2\sqrt{n}}\right)}, \frac{CTR+\frac{1}{2\sqrt{n}}}{1-3\left(CTR+\frac{1}{2\sqrt{n}}\right)}\right]$$

This provides a theoretical 70% confidence (at least) but from simulations we found it is 95%.

How to derive the number of cherries?

- Count the cherries in a tree
- Reconstruct a tree from sequences and count the cherries
- We developed a tree-free cherry estimation CWT

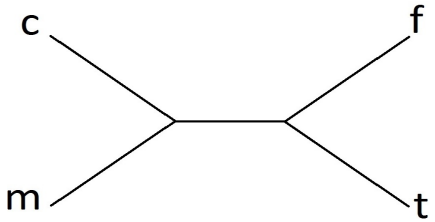
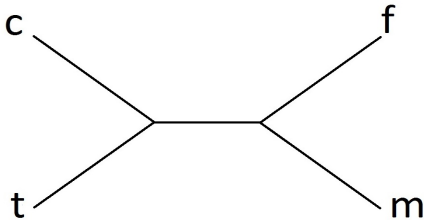
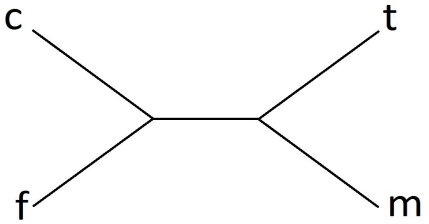
CWT overlook

- Consider tree unrooted (rooted and unrooted differ by 1 cherry)
- For each tip c, the algorithm looks among all other tips as candidates to form a cherry with c
- Current candidate m is tested against new candidate t in a quartet (c, f, m, t) where f for sure cannot form a cherry with c
- The quartet selection finds where is the split in (c, f, m, t) and at least one between m and t must be excluded

Quartet selection

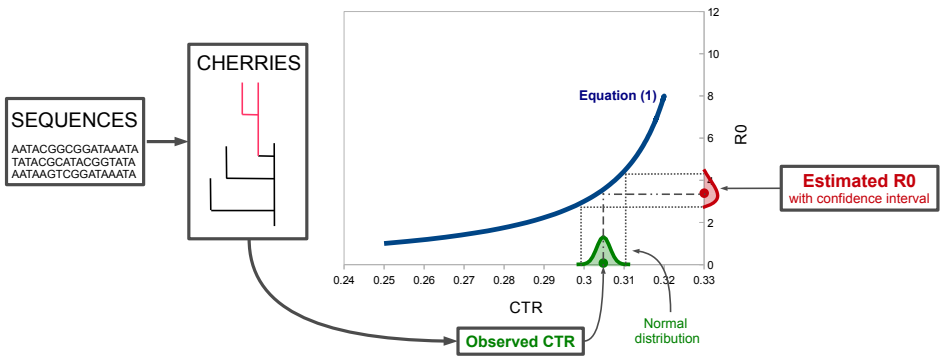
c to check, m current candidate

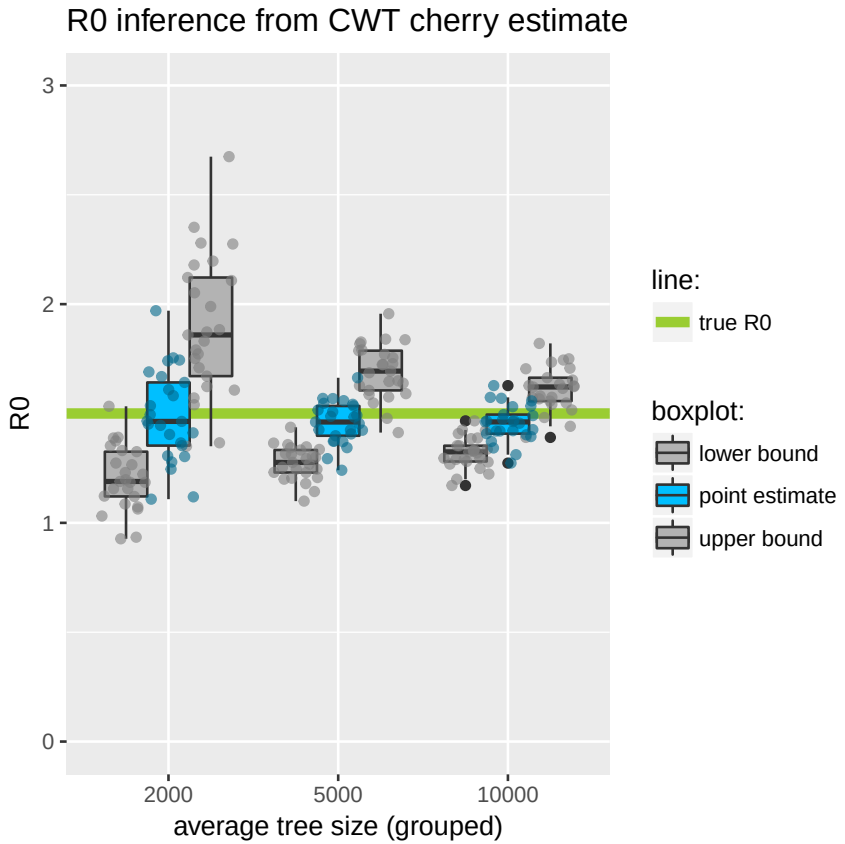
f surely not in a cherry, t is the test in this iteration

	characteristic	quartet	CWT update)
#1	$\left\lvert \overline{cf}+\overline{mt}-\overline{ct}-\overline{mf} \right\rvert=0$		t excluded $m \leftarrow m$, $f \leftarrow f$ $m.OK = m.OK$
#2	$\left\lvert \overline{cf}+\overline{mt}-\overline{cm}-\overline{ft} \right\rvert=0$		t is the new candidate $m \leftarrow t$, $f \leftarrow m$ $m.OK = 1$
#3	$\left\lvert \overline{ct}+\overline{fm}-\overline{cm}-\overline{ft} \right\rvert=0$		t excluded $m \leftarrow m$, $f \leftarrow f$ $m.OK = 0$

R_0 inference with CWT: simulation and real data

Simulations:





H1N1 2009 outbreak:

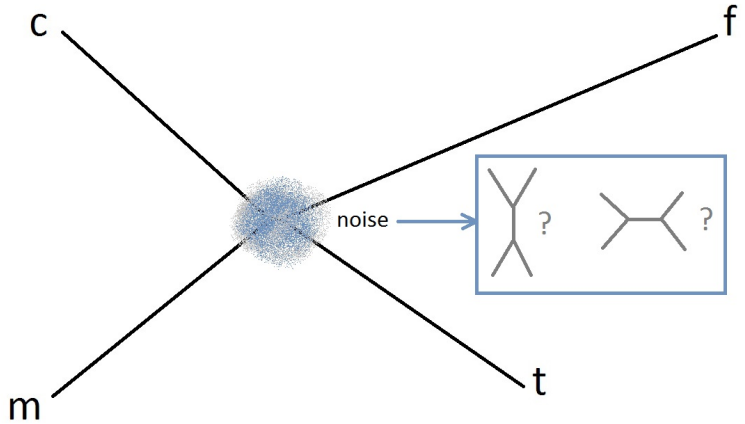
Method	# sequences	CTR estimate	R_0 estimate (CI)
CWT	2975	0.27042	1.43 (1.21 - 1.73)
FastTree	2975	0.28303	1.88 (1.53 - 2.36)

Benefits

- Can use any genetic distance
- No need of sequence alignment
- Tree-free
- Each step for a different c is independent, so CWT highly parallelisable
- Minimal memory required $O(l)$, time complexity from $O(ln^2)$ up to linear $O(ln)$ if fully parallelised
- Aimed at big data

Issues

- Requires a lot of sequences
- Long branch attraction:



- Sampling not considered (yet). Possibly not hard if sample rate is known

Open questions

3 Use CWT to reconstruct the tree?

Thank you.

1 Plazzotta Colijn *Asymptotic frequency of shapes in supercritical branching trees*, J App Prob 2016

2 Plazzotta Colijn *Phylodynamics without trees: estimating R0 directly from pathogen sequences*, BioRxiv 2017

CWT accuracy