Specifying and fitting a Dollo point-process model of binary trait-character evolution

Joint work with

Russell Gray, Psychology, Auckland

Thanks for the help

Quentin Atkinson, Psychology, Auckland
David Welch, Mathematics, Auckland

Jorge Luis Borges

Spanish - todas las palabras fueron, alguna vez, un neologismo
Catalan - totes les paraules van ser, en algun moment, un neologisme
French - chaque mot fut un jour un nologisme
Italian - tutte le parole sono state, un tempo, un neologismo
Portuguese - todas as palavras foram, um dia, um neologismo
Sicilian - tutti I paroli sunnu stati, un tempu un neologismu

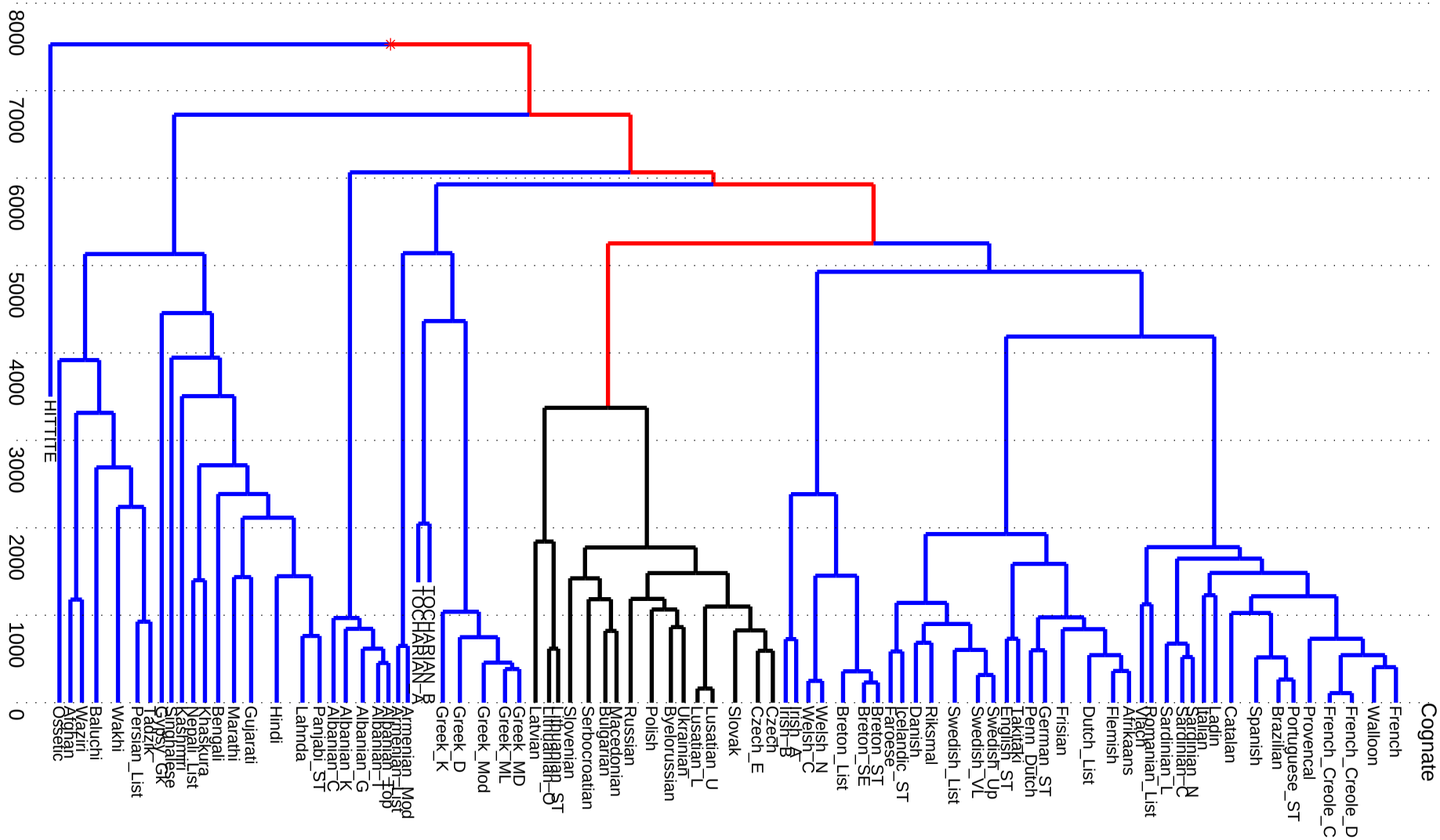
Danish - alle ord har engang vret en neologism
Dutch - alle woorden zijn, ooit, een neologisme geweest
English - every word was once a neologism
Flemish - alle woorden waren, eens, een neologisme
German - jedes Wort war einmal ein Neologismus
Swedish - alla ord har ngon gng varit en neologism

Basque - hitz guztiak izan ziren, noizbait ere, neologismo
Finnish - jokainen sana on joskus ollut neologismi

Verba Volant 28/7/04

N=87 rows/Indo-European languages
L=2398 columns/cognate classes/homologous traits
K=200 meaning classes

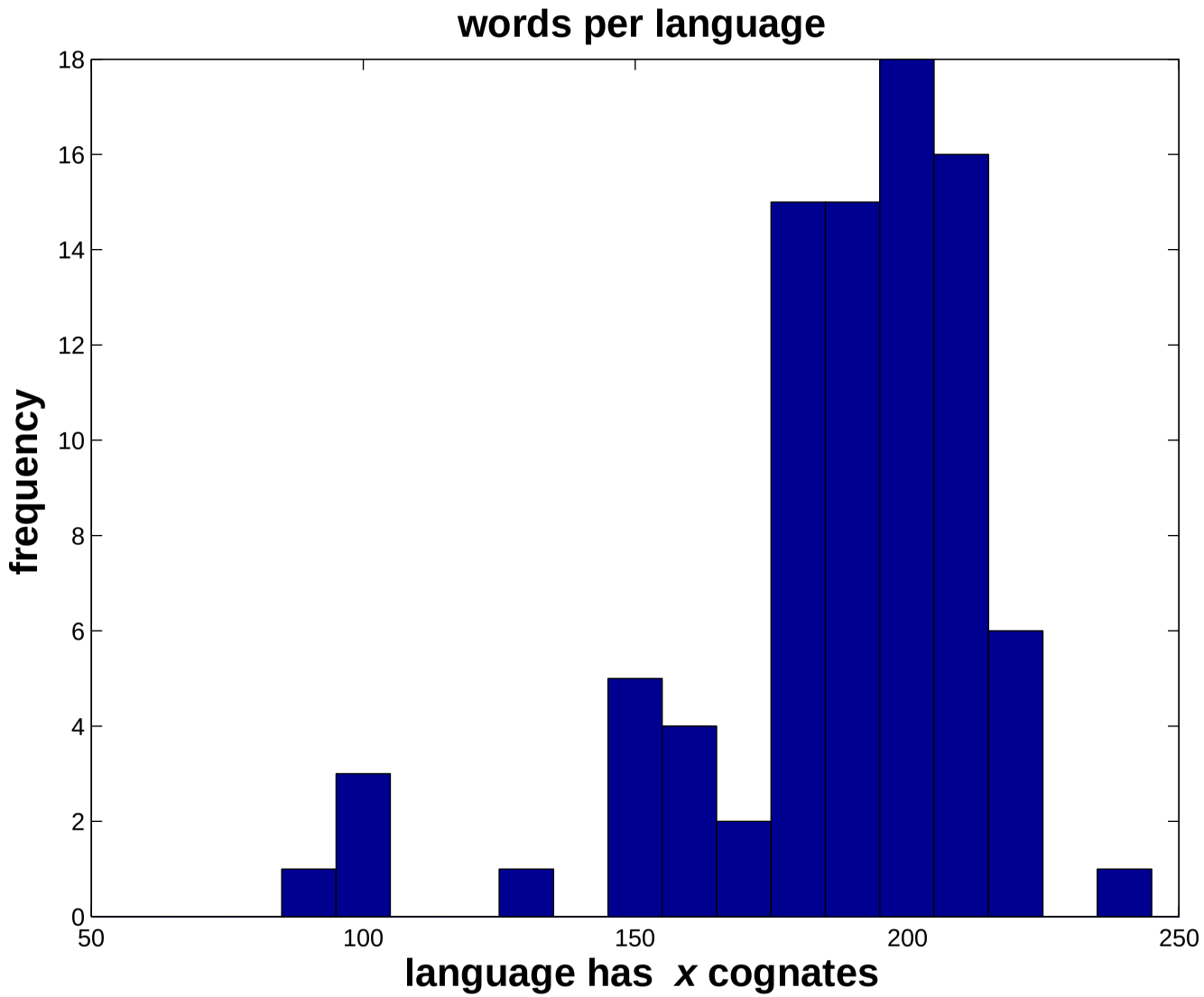
```
Danish              00100000010000000000000000010000 . . .
Riksmal             00100000010000000000000000010000
Icelandic_ST        00100000010000000000000000010000
Faroese             00100000010000000000000000010000
English_ST          00100000000000001000000000000000
French              01000000000000000000001000000000
Lithuanian_O        10000000000000000001000000000010 . . .
.                   .
.                   .
[1]  ALL_1    [2]  ALL_2    [3]  ALL_3    [4]  ALL_4    [5]  ALL_5
[6]  ALL_6    [7]  ALL_7    [8]  ALL_8    [9]  ALL_9    [10] AND_1
[11] AND_2    [12] AND_3    [13] AND_4    [14] AND_5    [15] AND_6
[16] AND_7    [17] AND_8    [18] AND_9    [19] AND_10   [20] AND_11
[21] AND_12   [22] AND_13   [23] AND_14   [24] AND_15   [25] ANIMAL_1
[26] ANIMAL_2 [27] ANIMAL_3 [28] ANIMAL_4 [29] ANIMAL_5 [30] ANIMAL_6 ...
```



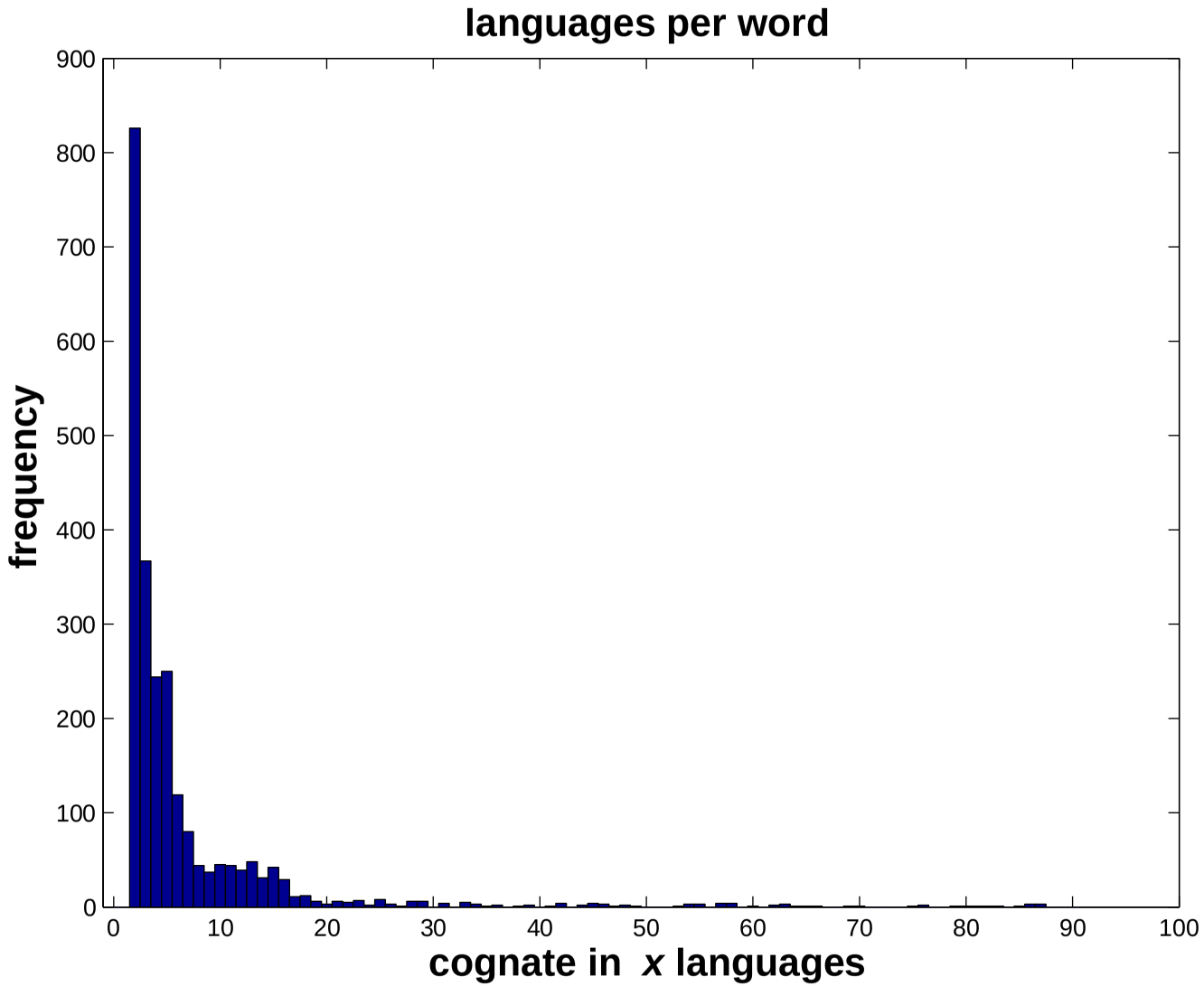
Clades: Celtic Brythonic Italic Iberian-French Germanic West-Germanic North-Germanic Balto-Slav Slav Indic Indo-Iranian Iranian Albanian Greek Armenian Tocharic

Should we fit a tree? Loan-words.

Above a single tree sampled from posterior: **not a single-tree "answer"**



Finite sites - L=2398 cognates? K=200 meaning classes?



Lost "zeros" and "ones"? About 2/3 of all word birth events.

Glottochronology c. 1970

Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts.
Proceedings of the American Philosophical Society, 96, 453-463 (1952).

Dyen, I., Kruskal, J. B. & Black
`http://www.ntu.edu.au/education/langs/ielex/PUBLICATIONS.html`

Two languages, $L = L_1 = L_2$ Swadesh-cognates, $L_{1,2}$ cognates in common.
Suppose $L_{1,2}(t;\mu) = L_1 \times \exp(-\mu t)$

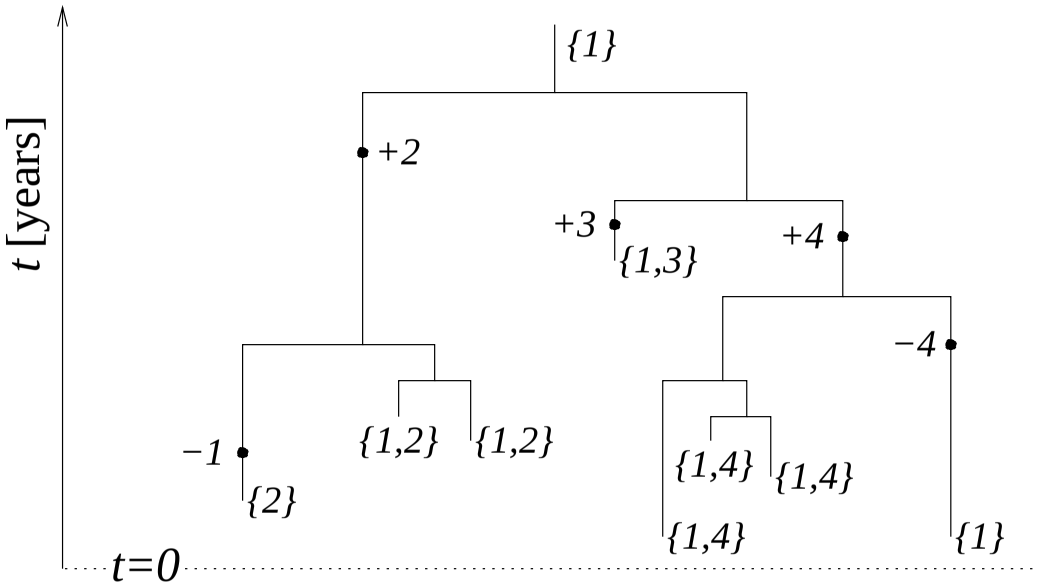
Estimate $\hat{\mu}$ using eg Latin/Italian [$\exp(-\mu \times 1000) \simeq 0.82$].
Estimate $\hat{t}$ for Persian/Danish using $\hat{\mu}$.

Uncertainty? Joint estimation? Constant-rate? L_1, L_2 random sizes?

Pagel (2004), Gray & Atkinson (2003), Warnow Evans Ringe & Nakhleh (2004)

Tree

- g: $g = (E, t) \quad t = (t_1, \ldots t_{2N-1})$
- $x_s = (\tau_s, i_s) \in [g], \quad x = (x_1, \ldots x_L)$



Rate Parameters

- θ: tree model parameter; $g \sim f_G(g|\theta)$
- λ: cognate birth rate [per year];
- μ: per capita death rate [per year]; $D \sim \Pr\{D|g, \lambda, \mu\}$

$$
\begin{aligned}
Y &\sim \Pi(\lambda; [g]), \\
d(z) &= \#\text{ leaf languages with cognate born at } z = (\tau, i) \\
\lambda(z) &= \lambda \Pr(d(z) \geq 2 | z, g, \mu) \\
X &\sim \Pi(\lambda(z); [g]), \\
\Pr(dx|g, \mu, \lambda) &= p_X(x|g, \mu, \lambda)dx \\
&= \exp\left(-\int_{[g]} \lambda(z)dz\right) \prod_{s=1}^{L} \lambda(x_s)dx_s
\end{aligned}
$$

$$
D = \begin{bmatrix}
(1) & (2) & (4) \\
0 & 1 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 0
\end{bmatrix}
$$

Statistical Inference:

$$P(x, g, \mu, \lambda, \theta|D) \propto \left(\prod_{s=1}^{L} P(D_{:,s}|x_s, g, \mu)\right) p_X(x|g, \mu, \lambda) f_G(g|\theta) p(\mu, \lambda, \theta)$$

$$f_G(g|\theta) \propto t_R^{-N+2}(\theta - t_R)^{-1}, \qquad 0 \leq t_R \leq \theta, \qquad \theta = 16000 \text{ Years BP}$$

$$f_G(g|\theta) \propto \mathbb{I}_{0 \leq t_R \leq \theta} \qquad \text{and Yule} \quad \theta^{N-3} \exp(-\theta|g|) \quad \text{and...}$$

$$P(g, \mu, \lambda, \theta|D) \propto \left(e^{-\int_{[g]} \lambda(z)dz} \int_{\Omega} \prod_{s=1}^{L} P(D_{:,s}|x_s, g, \mu)\lambda(x_s)dx_s\right) f_G(g|\theta) p(\mu, \lambda, \theta)$$

$$= \frac{\lambda^L}{L!} e^{-\int_{[g]} \lambda(z)dz} \prod_{s=1}^{L} \left(\int_{[g]} \Pr(D_{:,s}|y_s, g, \mu, d(y_s) \geq 2) \Pr(d(y_s) \geq 2|y_s, g, \mu)dy_s\right)$$

$$\times \qquad f_G(g|\theta) p(\mu, \lambda, \theta)$$

$$P(g, \mu, \lambda, \theta|D) \propto \frac{\lambda^L}{L!} e^{-\int_{[g]} \lambda(z)dz} \prod_{s=1}^{L} \left(\int_{[g]} \Pr(D_{:,s}|y_s, g, \mu)dy_s\right) f_G(g|\theta) p(\mu, \lambda, \theta)$$

$$
\begin{aligned}
\int_{[g]} \lambda(z)dz &= \lambda \int_{[g]} \Pr(d(z) \geq 2|z, g, \mu)dz \\
&= \lambda \sum_{\langle i,j \rangle \in E(g)} \int_{t_j}^{t_i} \Pr(d(\tau) \geq 2|(\tau, j), g, \mu)d\tau \\
&= \frac{\lambda}{\mu} \sum_{\langle i,j \rangle \in E(g)} \Pr(d(t_j, j) \geq 2|(t_j, j), g, \mu)(1 - e^{-\mu(t_i - t_j)})
\end{aligned}
$$

$$
\begin{aligned}
\int_{[g]} \lambda(z)dz &= \lambda \int_{[g]} \Pr(d(z) \geq 2|z, g, \mu)dz \\
&= \lambda \sum_{\langle i,j \rangle \in E(g)} \int_{t_j}^{t_i} \Pr(d(\tau) \geq 2|(t_j, j), g, \mu) e^{-\mu(\tau - t_j)} d\tau \\
&= \frac{\lambda}{\mu} \sum_{\langle i,j \rangle \in E(g)} \Pr(d(t_j, j) \geq 2|(t_j, j), g, \mu)(1 - e^{-\mu(t_i - t_j)})
\end{aligned}
$$

$$\begin{aligned}
\int_{[g]} \lambda(z)dz &= \lambda \int_{[g]} \Pr(d(z) \geq 2|z, g, \mu)dz \\
&= \lambda \sum_{\langle i,j \rangle \in E(g)} \int_{t_j}^{t_i} \Pr(d(\tau) \geq 2|(t_j, j), g, \mu) e^{-\mu(\tau - t_j)} d\tau \\
&= \frac{\lambda}{\mu} \sum_{\langle i,j \rangle \in E(g)} \Pr(d(t_j, j) \geq 2|(t_j, j), g, \mu)(1 - e^{-\mu(t_i - t_j)})
\end{aligned}$$

$\Pr\{d(t_i, i) \geq 2|(t_i, i), g, \mu\} = 1 - u_i^{(0)} - u_i^{(1)}$

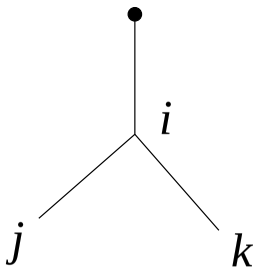
$u_i^{(0)} = \Pr\{d(t_i, i) = 0|(t_i, i), g, \mu\} \qquad u_i^{(1)} = \Pr\{d(t_i, i) = 1|(t_i, i), g, \mu\}$

$p_{i,j} = \exp(-\mu(t_i - t_j))$

$u_i^{(0)} = (1 - p_{i,j}(1 - u_j^{(0)})) \times (1 - p_{i,k}(1 - u_k^{(0)}))$

$u_i^{(1)} = (1 - p_{i,j}(1 - u_j^{(0)}))p_{i,k}u_k^{(1)} + (j \longleftrightarrow k)$

$u_i^{(0)} = 0$ and $u_i^{(1)} = 1$ if i is leaf.



$$P(g,\mu|D) \propto \int_0^\infty \int_0^\infty \left(\frac{\lambda}{\mu}\right)^L f_G(g|\theta)p(\mu,\lambda,\theta)$$
$$\exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E(g)} \Pr\{d(y)\geq 2|(t_j,j),g,\mu\}\left(1-e^{-\mu(t_i-t_j)}\right)\right) d\lambda d\theta$$
$$\prod_{s=1}^{L}\left(\sum_{\langle i,j\rangle\in E(g)} \Pr\{D_{:,s}|(t_j,j),g,\mu\}\left(1-e^{-\mu(t_i-t_j)}\right)\right)$$

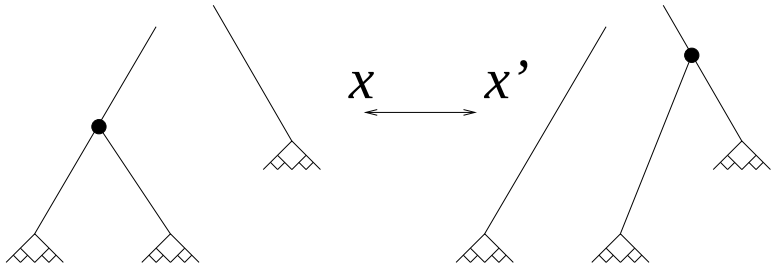
- Impose tree constraints - Italic/Brythonic/Germanic/Balto-Slav - fix μ
- For simple $p(\mu,\lambda,\theta)$ can integrate λ and θ by hand.
- Summarize $P(g,\mu|D)$ ($N-1$, 1, topology) using MCMC samples.

MCMC: Generate $(g, \mu) \sim P$ for MC inference (demo here)

`function` $(g^{(m+1)}, \mu^{(m+1)}) =$ `Markov`$(g^{(m)}, \mu^{(m)})$. Suppose $g^{(m)} = g,\ \mu^{(m)} = \mu$.

1. $a \sim U\{1, 2 \ldots A\}$
2. $u \sim q_a(u)$, $(g', \mu') = \phi_a(g, \mu, u)$
3. $\alpha((g, \mu) \to (g', \mu')) = \min\left\{1,\ \frac{P(g', \mu'|D)}{P(g, \mu|D)} \times \overbrace{\frac{q_a(u')}{q_a(u)}\left|\frac{\partial(g', \mu', u')}{\partial(g, \mu, u)}\right|}^{J}\right\}$
4. with prob. α set $(g^{(m+1)}, \mu^{(m+1)}) = (g', \mu')$ else $(g^{(m+1)}, \mu^{(m+1)}) = (g, \mu)$.

Topology



Scaling

$$\begin{aligned} x &= ((E, t_Y, t_I), \mu) \\ u &\sim 1/u\, \mathbb{I}_{1/2<u<2} \\ x' &= ((E, ut_Y, t_I), \mu/u) \\ J &= u^{N-2} \end{aligned}$$

and six other update types for irreducibility on trees + μ.

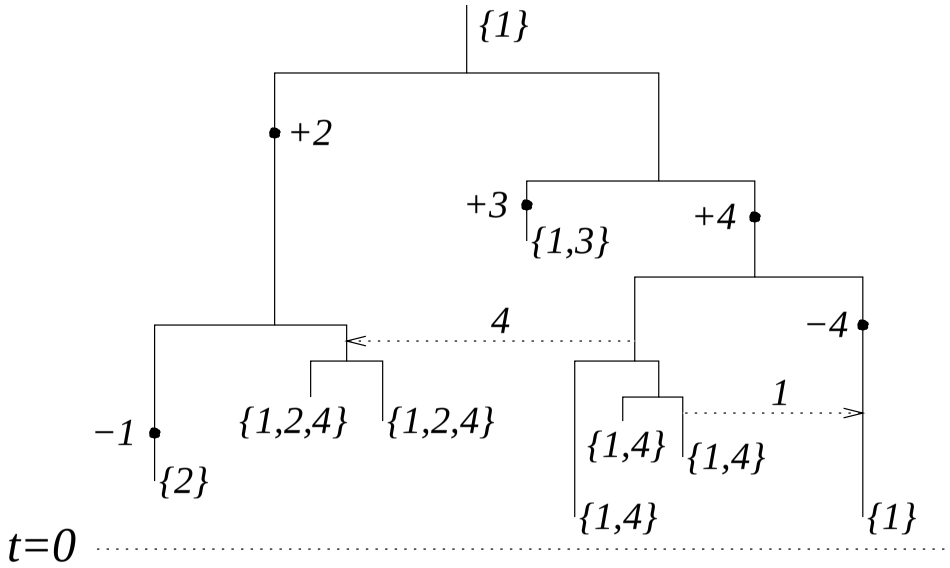
Results 1: compare language subsets

N taxa	$\overline{t_R}$	(σ_{t_R})	$\overline{\mu}$	(σ_μ)
87	7500	(210)	0.000263	(06)
30	8050	(360)	0.000204	(12)

Results 2: compare clade rates

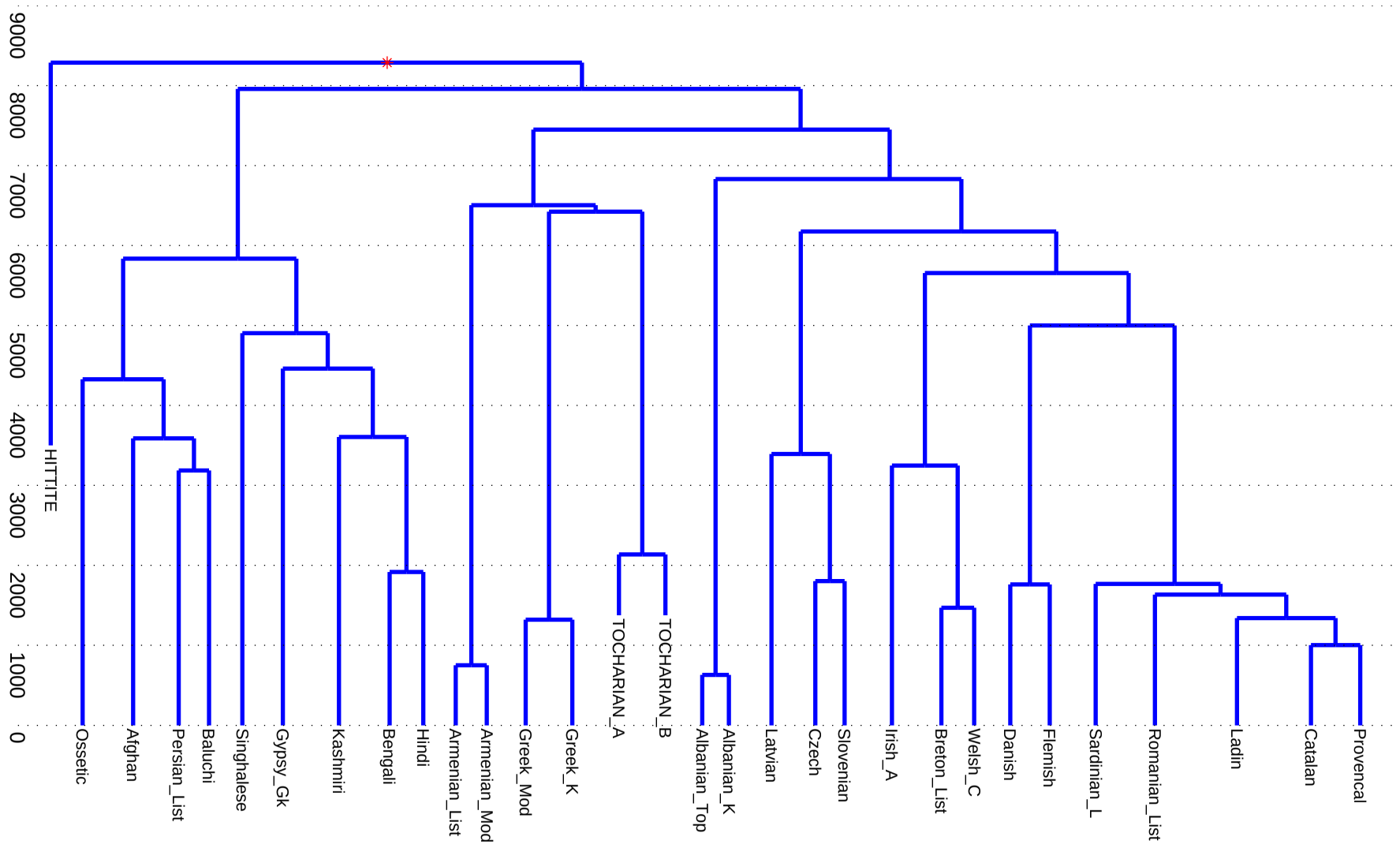
Clade	$\overline{\mu}$	(σ_μ)
Italic	0.000268	(11)
Iberian-French	0.000224	(21)
Germanic	0.000246	(13)
Balto-Slav	0.000260	(36)
Celtic	0.000176	(16)

Results 3: synthetic data with borrowing



At rate $b\mu$ each word copies itself into a randomly chosen language.

b	$\overline{t_R}$	(σ_{t_R})	true t_R
0	6720	(250)	|
0.1	6930	(230)	6900
0.2	6280	(180)	|



Single tree sampled from posterior: **not a single-tree "answer"**

Problems

- borrowing (model, get same misfit on synthetic data)
- word loss at branching events (model and fit)
- uneven representation of languages in data (model and fit)
- other rate heterogeneity (give up?)

Results

- better off without the lost "ones" - independence/observation model
- here is a stochastic model of Dollo parsimony (applications in anthropology/biology)
- here is a Monte Carlo toolbox to fit the model
- this model favors IE separation date around 7500 BP