Phylogenetic[1] Correlations in Mutation Processes

Eli Ben-Naim

Theoretical Division, Los Alamos National Lab

I Motivation: Evolutionary Trees

II The Mutation-Duplication Model

III Pair & Higher Order Correlations

IV Asymptotic Analysis

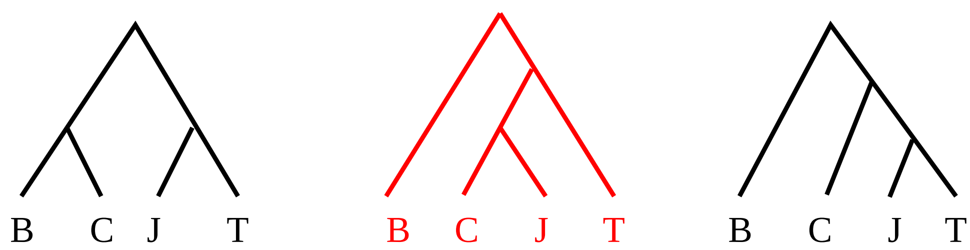
V Generalizations

---

[1] Phylogeny = Family Tree

Evolutionary Tree Reconstruction

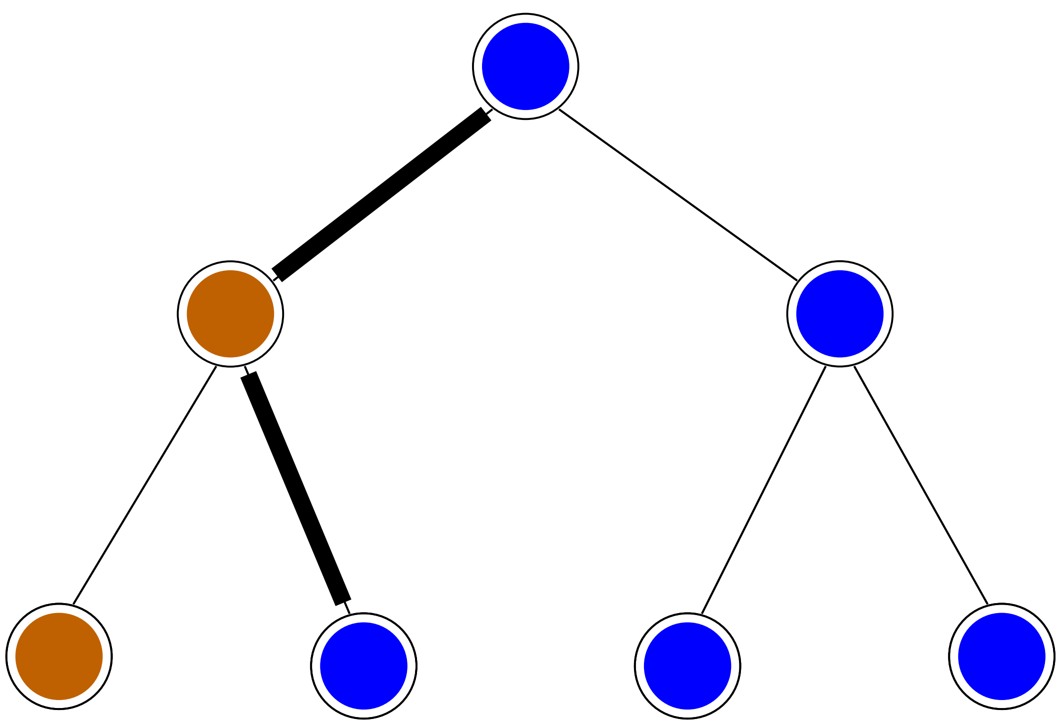
- **Genetic Sequences:** Evolution's "fingerprints"
- DNA/RNA, amino acid sequences: like words taken from alphabet of size 4, 20
- Example: Ostrich (ABCD), Turkey(A**CB**D), Jaguar(A**CD**D), Tiger (ABC**A**)
- Tree Reconstruction:
  1. Assume evolution/mutation model
  2. Enumerate all possible evolutionary trees
  3. Choose "most likely" tree



Role of tree morphology?

The Mutation-Duplication Model

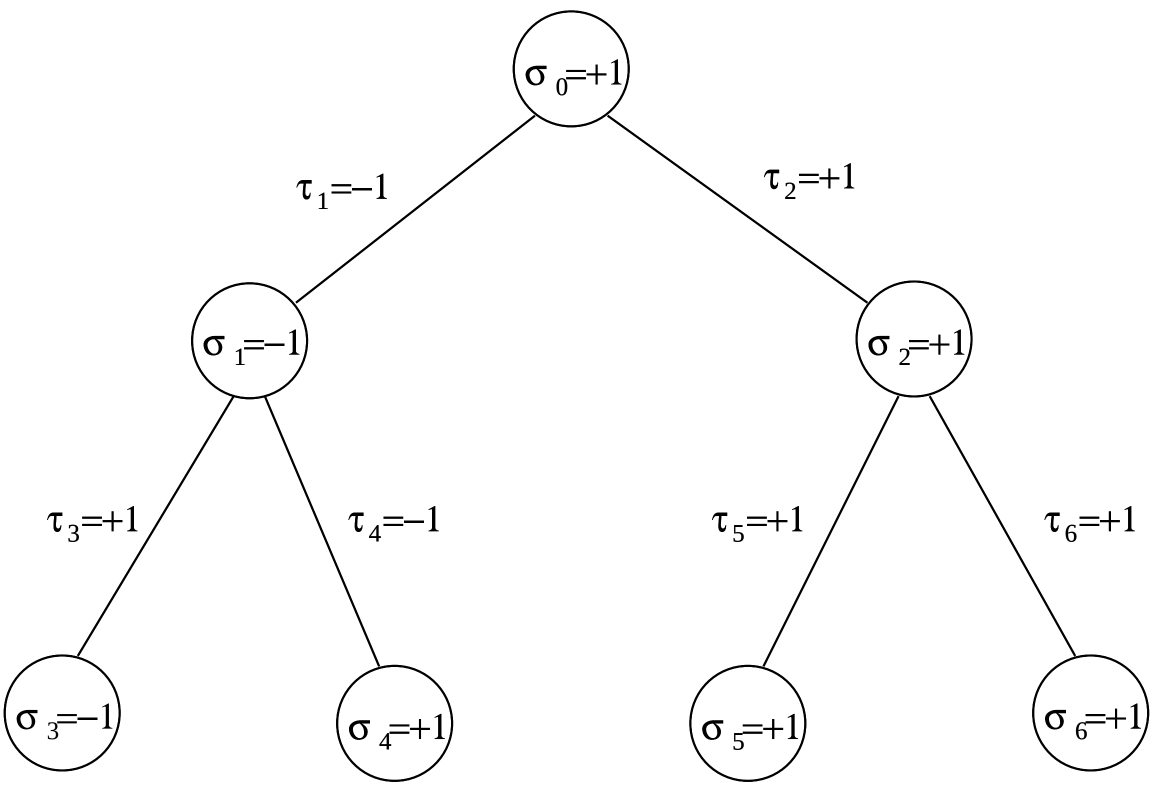
- **Simplified two-state sequences:**
  Alphabet size=2, Length=1.
- **Random (Poisson) mutation process**:
  Mutation occurs with probability p.
- **Binary tree phylogeny:**
  Every parent has 2 children.



Set-Up

- **Numeric representation:** $\sigma = \pm 1$
- **Invariant:** under $\sigma \to -\sigma$, $p \to 1-p$
  Restrict attention to $0 \le p \le 1/2$
- **Multiplicative variables:** $\sigma_i = \sigma_j \tau_i$

$$\langle \tau \rangle \equiv \langle \tau_i \rangle = (1-p) \times (+1) + p \times (-1) = 1 - 2p$$



Calculating Average Correlations

- **Pair Correlation:** $\langle \sigma_i \sigma_j \rangle$
- Example: $\langle \sigma_3 \sigma_4 \rangle$
- Method: trace history for every node

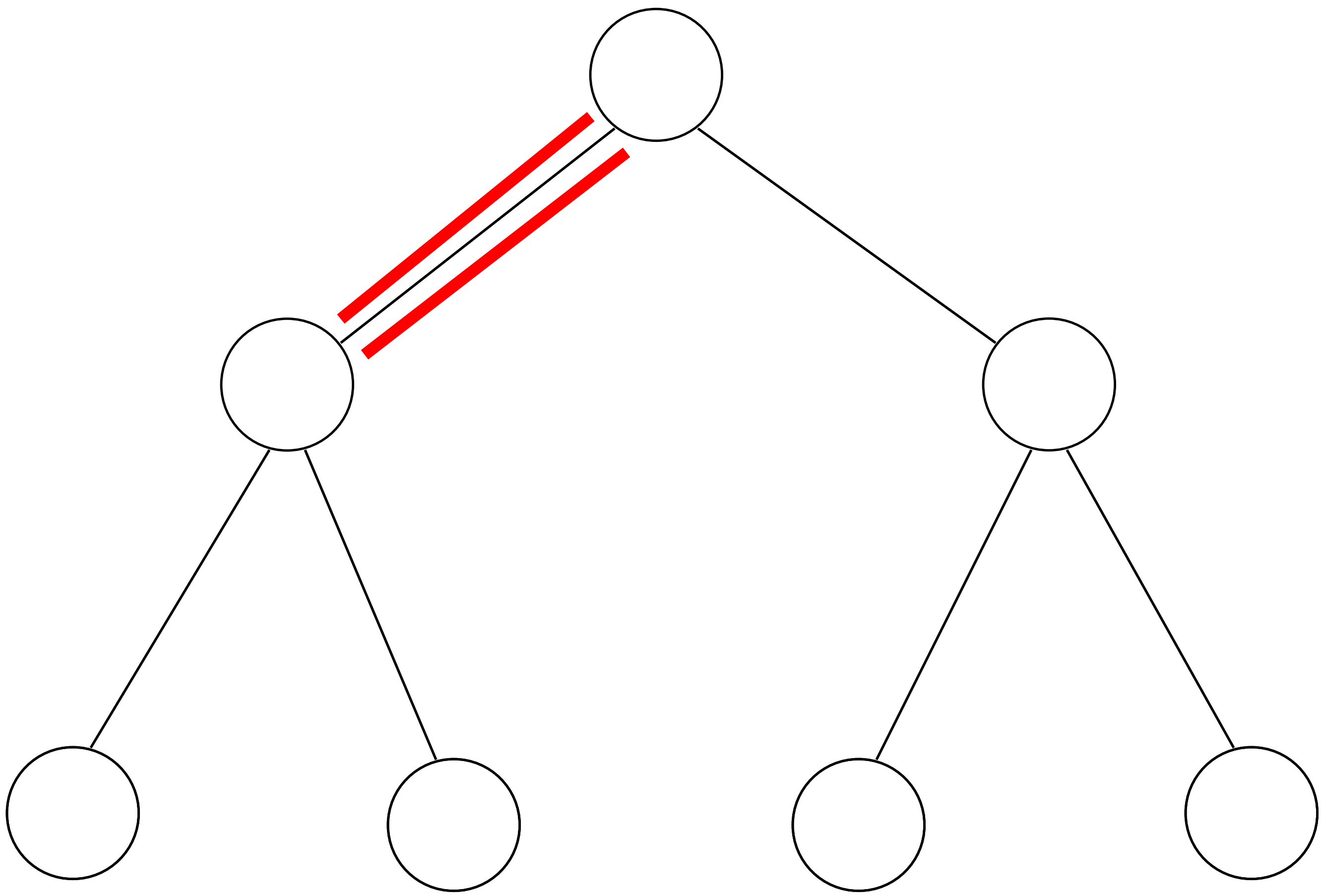
$$\sigma_3 = \sigma_0 \tau_1 \tau_3 \qquad \sigma_4 = \sigma_0 \tau_1 \tau_4$$

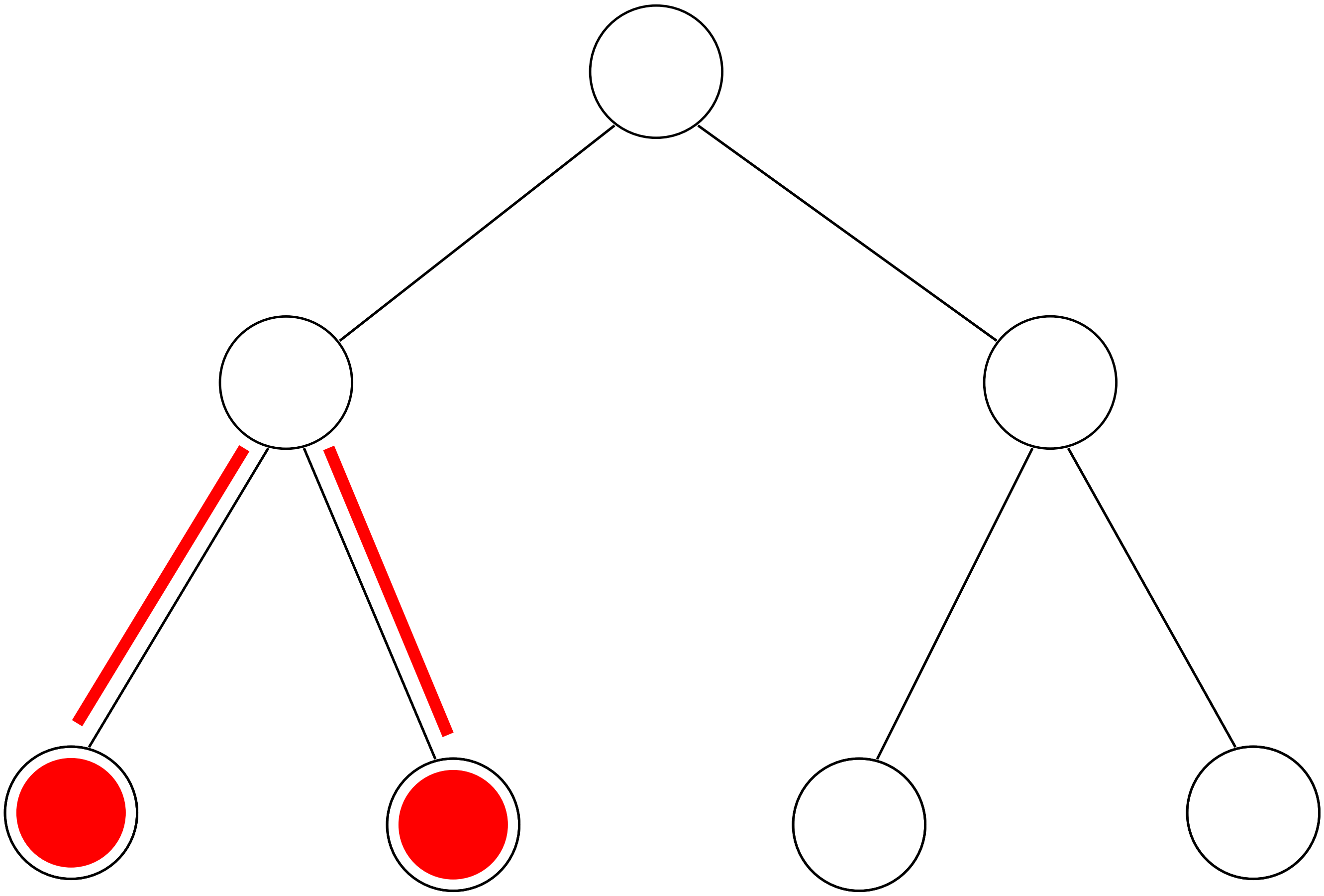
- Use (i) $\sigma^2 = \tau^2 = 1$ (ii) τ_i are i.i.d

$$\langle \sigma_3 \sigma_4 \rangle = \langle \sigma_0^2 \tau_1^2 \tau_3 \tau_4 \rangle = \langle \tau \rangle^2$$

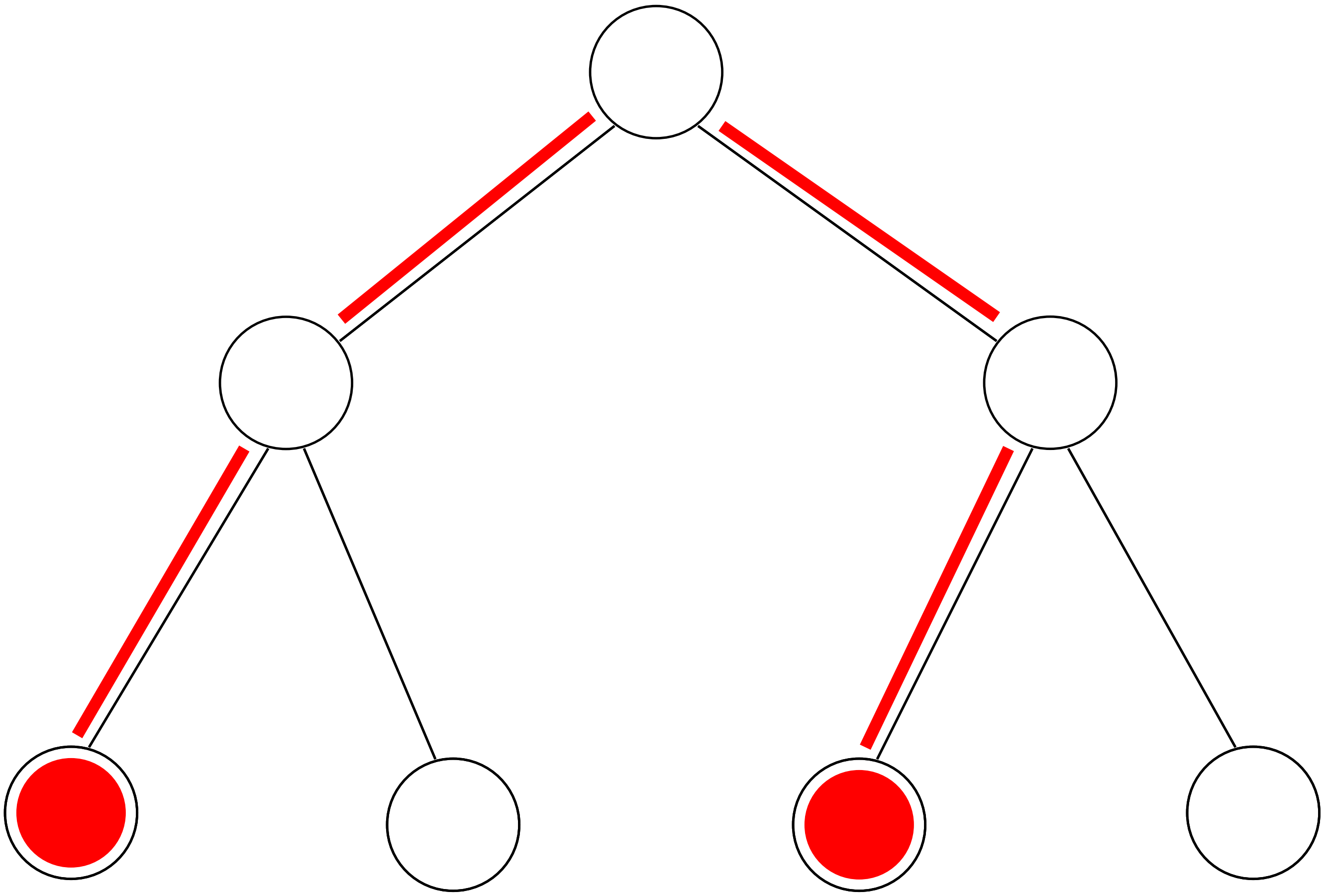
- **Recipe**
  1. Trace path to origin for each node.
  2. Cancel doubly counted bonds.
  3. Average correlation $= \langle \tau \rangle^{\text{number of remaining bonds}}$

Common bonds cancel in pairs

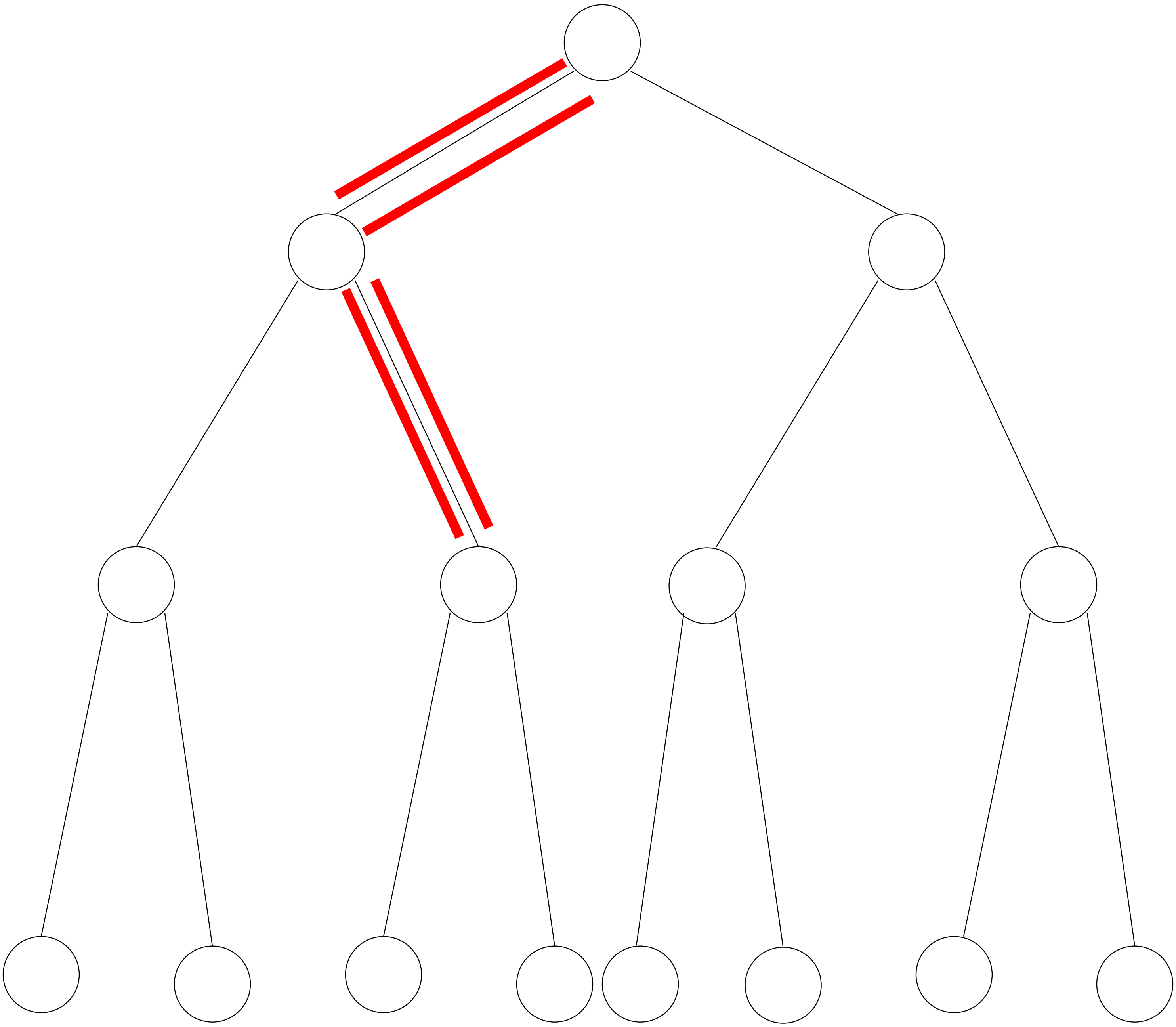


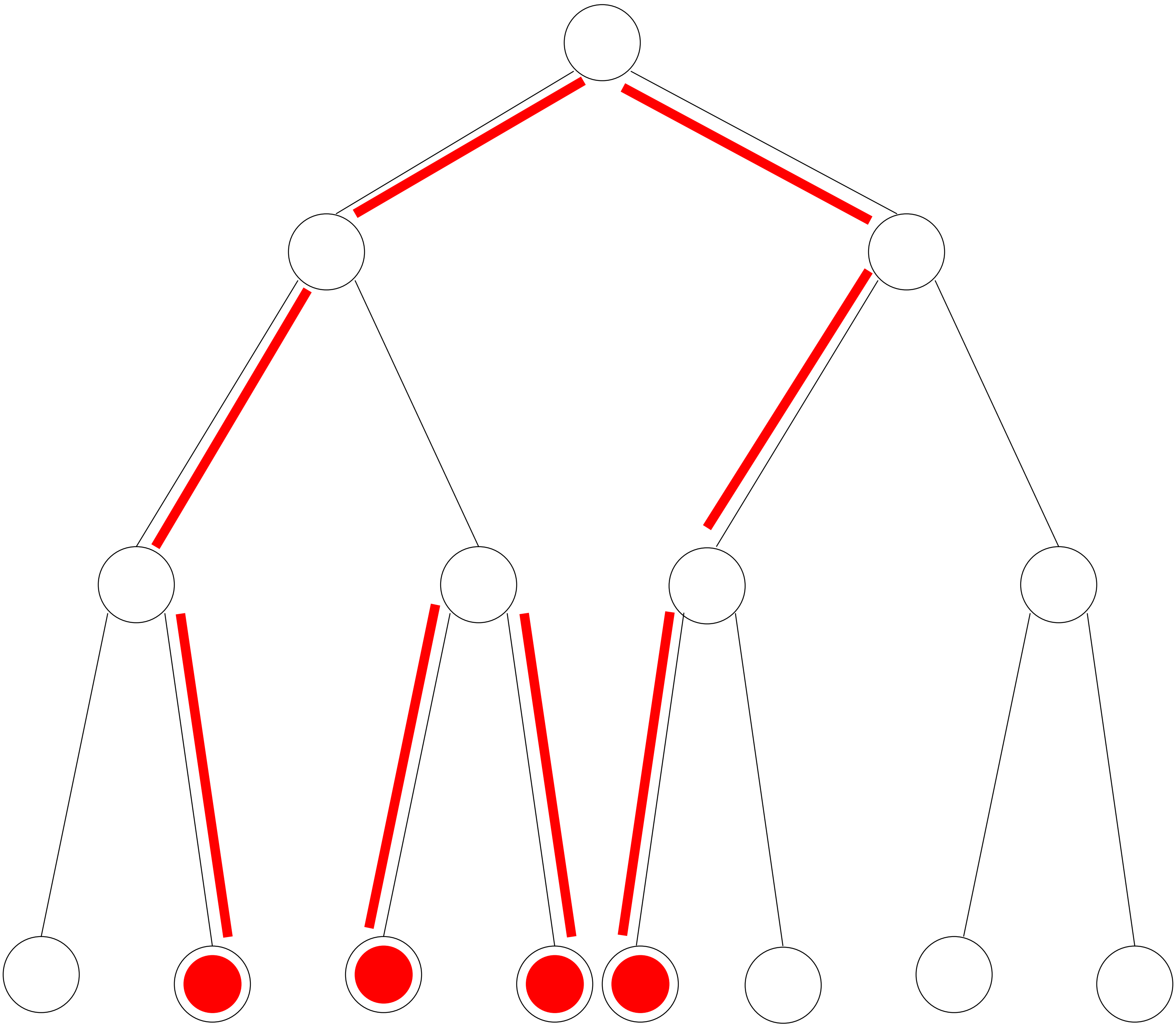


$$\langle \sigma_3 \sigma_4 \rangle = \langle \tau \rangle^2$$



$$\langle \sigma_3 \sigma_5 \rangle = \langle \tau \rangle^4$$





$$\langle \sigma_8 \sigma_9 \sigma_{10} \sigma_{11} \rangle = \langle \tau \rangle^8$$

Law for Correlations

- **Genetic Distance:** $d_{i,j} =$ minimal number of bonds connecting i and j

- Two-point correlation:

$$\langle \sigma_i \sigma_j \rangle = \langle \tau \rangle^{d_{i,j}}$$

- Similarly, four-point correlation:

$$\langle \sigma_i \sigma_j \sigma_k \sigma_l \rangle = \langle \tau \rangle^{d_{i,j,k,l}}$$

$$d_{i,j,k,l} = \min\{d_{i,j} + d_{k,l}, d_{i,k} + d_{j,l}, d_{i,l} + d_{j,k}\}.$$

- n-point genetic distance $d_n =$ minimal # of bonds connecting n nodes **in pairs**

$$\boxed{n\textbf{-point correlations} = \langle \tau \rangle^{d_n}}$$

Average Pair Correlations

- **Average pair correlation** at kth gen
  average over: (i) realizations (ii) pairs

$$G_2(k) = \langle\langle \sigma_i \sigma_j \rangle\rangle$$

- **Geometric Series:**

$$G_2(k) = \frac{\langle\tau\rangle^2 + 2\langle\tau\rangle^4 + \cdots + 2^{k-1}\langle\tau\rangle^{2k}}{2^k - 1}$$

- **Asymptotic Behavior:** $k \to \infty$

$$G_2(k) \sim \begin{cases} \langle\tau\rangle^{2k} & p < p_c; \\ 2^{-k} & p > p_c. \end{cases} \qquad p_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$$

- **Trivial "star" phylogeny**: $G_2^*(k) = \langle\tau\rangle^{2k}$

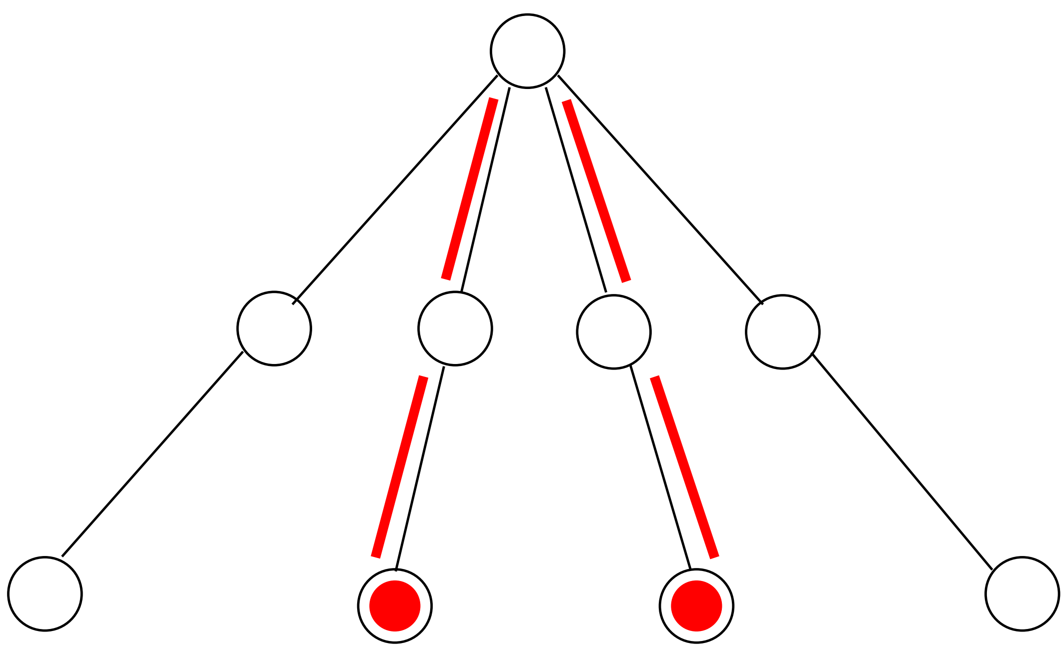
$$\frac{G_2(k)}{G_2^*(k)} \to \begin{cases} \text{const.} & p < p_c; \\ \infty & p > p_c. \end{cases}$$

$p > p_c$: **Phylogeny causes strong correlations**

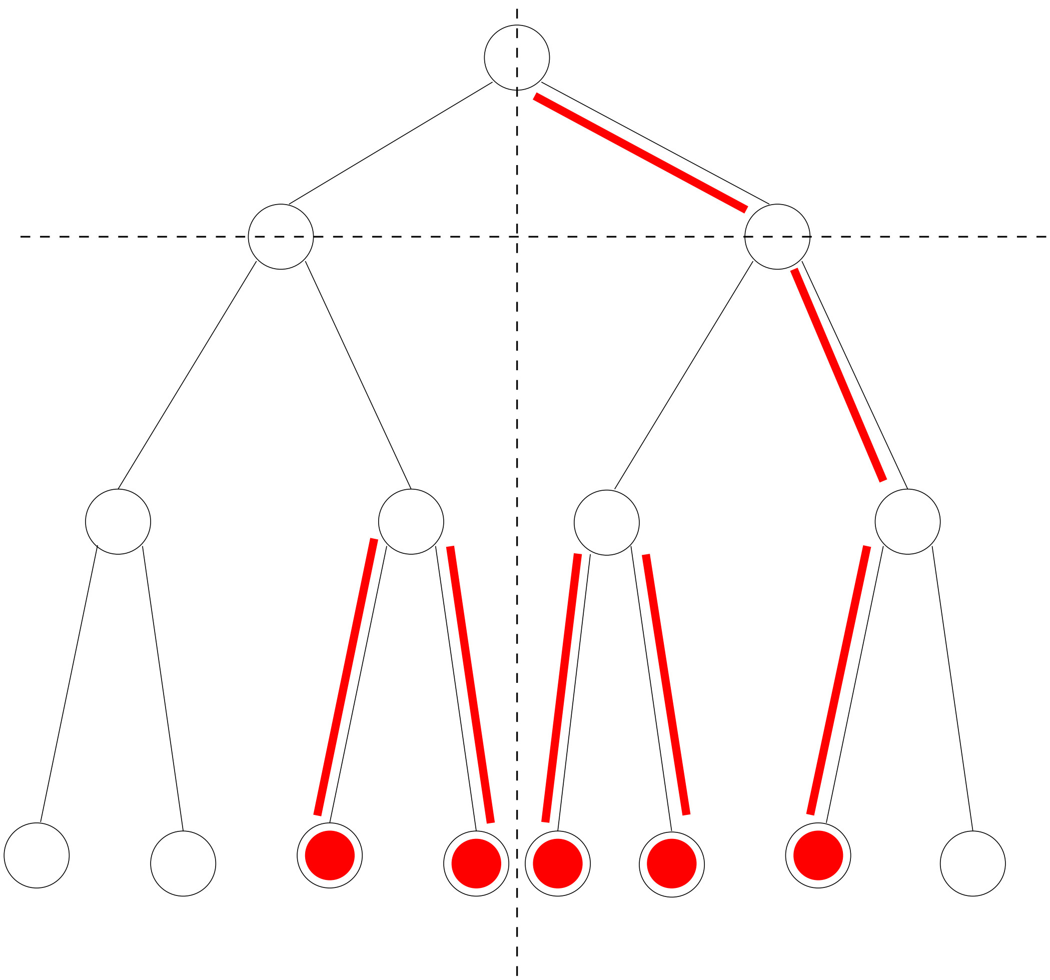
The Star Phylogeny

- **Trivial structure:** serves as reference
- All paths pass through root
- **Genetic distances:** $d_n = nk$

$$G_n(k) = \langle \tau \rangle^{nk}$$



Recursive Calculation



$$G_5 = G_2G_3 + G_1G_4 + \cdots$$

Reduce to 2 trees of 1 less generation

Higher Order Correlators

- **Average n-point correlation** at kth gen

$$G_n(k) = \langle\langle \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_n}\rangle\rangle$$

- Obtain from $G_n(k) = F_n(k)\Big/\binom{2^k}{n}$

$$F_n(k) = \sum_{1\le i_1<i_2<\cdots<i_n\le 2^k} \langle \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_n}\rangle.$$

- $F_n(k)$ obey **recursion relations**

$$F_n(k)=\sum_{m=0}^{n} F_m(k-1)F_{n-m}(k-1)\langle\tau\rangle^{(m \bmod 2)+(n-m \bmod 2)}$$

- **Generating functions** analysis for $k\to\infty$

Tree morphology allows recursive calculation

Leading Asymptotic Behavior

- **Low Mutation rates:** $p < p_c$
  $G_2(k) \simeq g_2 \langle \tau \rangle^{2k}$ generalizes
  Correlations are marginally larger, $g_n > 1$

$$G_n(k) \simeq g_n \langle \tau \rangle^{nk}$$

- **High Mutation rates:** $p > p_c$
  Decay same as for $G_2(k) \simeq \left(1/\sqrt{2}\right)^{2k}$

$$G_{2r}(k) \simeq (2r+1)\frac{G_{2r+1}(k)}{G_1(k)} \simeq f_{2r} \left(\frac{1}{\sqrt{2}}\right)^{2rk}$$

- Prefactors diverge near critical point

$$g_{2r} \simeq f_{2r} = \frac{(2r)!}{r!}\beta^r |p_c - p|^r \quad p \to p_c$$

$$\boxed{\textbf{Same critical behavior underlies all correlators}}$$

Heuristic Picture

- **Least correlated** nodes dominate at $p < p_c$

  $G_2 \propto \langle \tau \rangle^{2k}$ likelihood $\propto 1$

- **Most correlated** nodes dominate at $p > p_c$

  $G_2 \propto 1$ likelihood $\propto 2^{-k}$

- **Critical point** found by comparing the two

$$2\langle \tau \rangle^2 = 1 \quad \Rightarrow \quad p_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$$

- **Stochastic tree morphologies:** average number of children $\langle N \rangle$ relevant parameter

$$\langle N \rangle \langle \tau \rangle^2 = 1 \quad \Rightarrow \quad p_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{\langle N \rangle}}\right)$$

Multiplicity & correlation degree compete

Generalizations

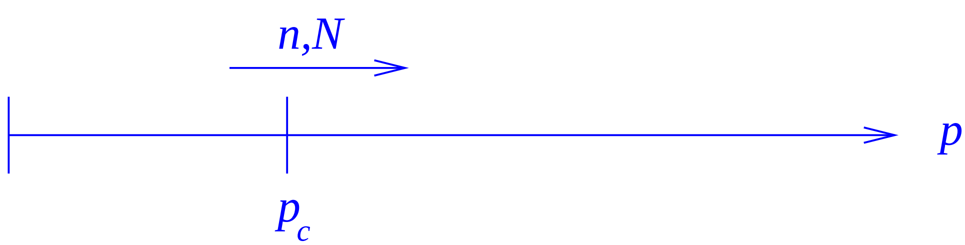
- **Continuous time formulation**
  mutation rate$=\theta$, birth rate$=1$

$$\theta_c = \frac{1}{4}$$

- **Multiple states** $\sigma^n = 1$, $\sigma \to \sigma \exp(2\pi i/n)$

$$\theta_c = \frac{1}{2(1-\cos\frac{2\pi}{n})}$$

- **Role of Phylogeny** decreases with increasing $\#$ of states n, children N



$$\boxed{\textbf{Nature of transition remains the same}}$$

Conclusions

- **Correlations decay exponentially with time, genetic distance.**
- **Phylogeny matters only when the mutation rates is high.**
- **Transition is critical in nature: all correlations behave similarly.**
- **Results apply to a large class of mutation/duplication processes.**
- **Role of phylogeny decreases as alphabet size, tree size increases.**

cond-mat/9812182, *Phys. Rev. E* (submitted)