



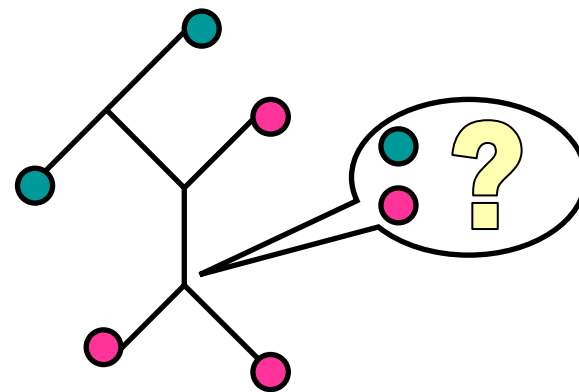
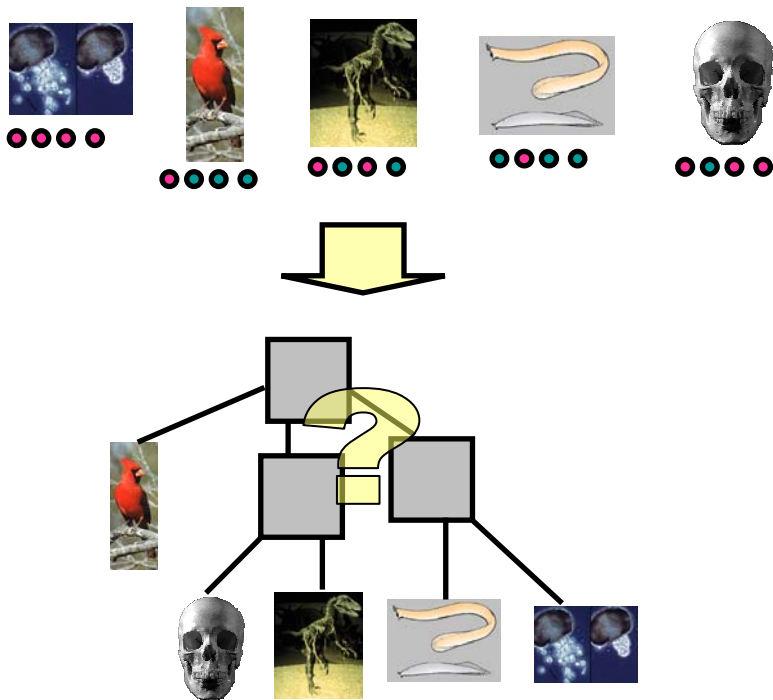
Optimal Phylogenetic Reconstruction

Joint Work with C. Daskalakis and E. Mossel
Available at <http://arxiv.org/abs/math/0509575>

Sébastien Roch
Department of Statistics
UC Berkeley

Seattle, WA, May 21, 2006

Overview: Result in a Nutshell



Tree reconstruction can be solved from very short sequences



There exists a good estimator for root reconstruction

Defn I: Markov Model on a Tree

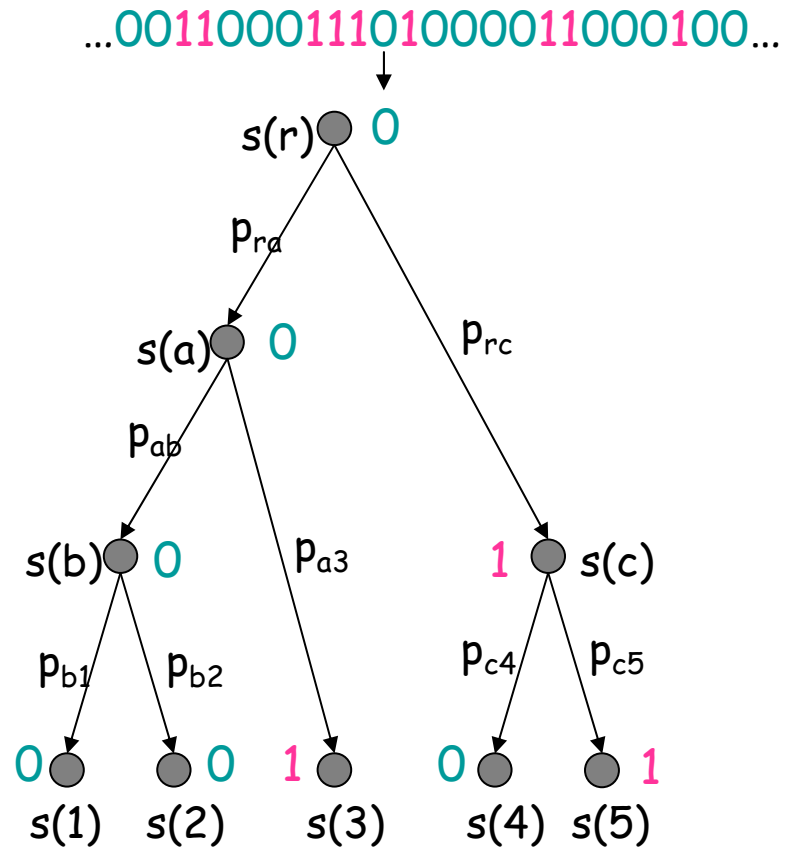
■ Ising/BSC/CFN Model:

- Tree: $T = (V, E)$
- Node states:

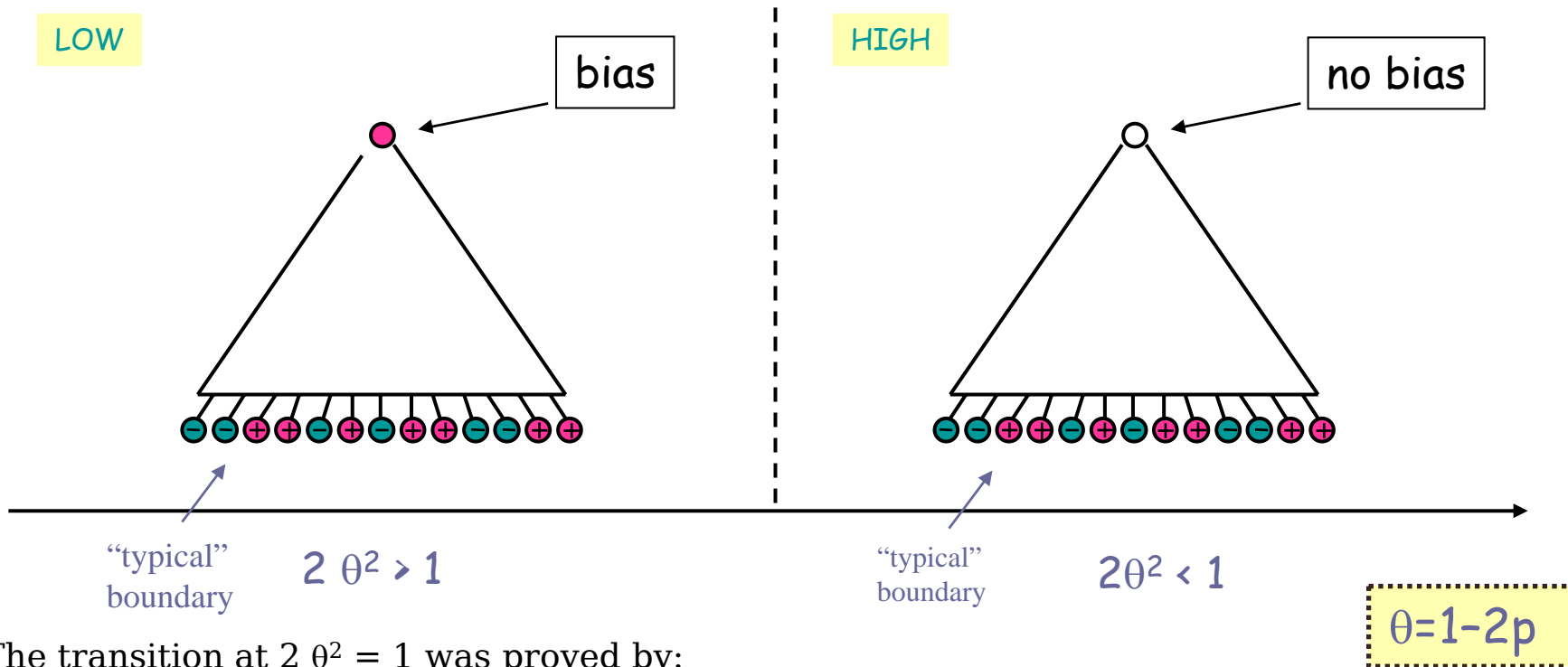
$$\{s(v) \in \{0,1\} : v \in V\}$$
- Mutation probabilities:

$$\{0 < p_e < 1/2 : e \in E\}$$
- Number of leaves: n

0: Purines (A,G)
1: Pyrimidines (C,T)



Defn II: “Reconstruction Problem”



The transition at $2\theta^2 = 1$ was proved by:
 [Bleher-Ruiz-Zagrebnov’95], [Ioffe’96],[Evans-Kenyon-Peres-Schulman’00],
 [Kenyon-Mossel-Peres’01],[Martinelli-Sinclair-Weitz’04], [Borgs-Chayes-Mossel-R’06].
 Also, “spin-glass” case studied by [Chayes-Chayes-Sethna-Thouless’86]. Solvability for
 $2\theta^2 > 1$ was first proved by [Higuchi’77] (and [Kesten-Stigum’66]).

Defn III: Phylogenetic Reconstruction Problem

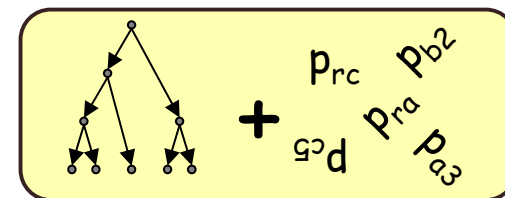
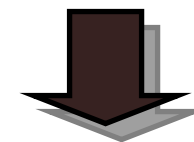
■ Reconstruction:

- *Given:* i.i.d. samples at the leaves
- *Task:* fully reconstruct the model, i.e. find **tree and mutation probabilities** (and, *if possible*, do so **efficiently**)

■ Previous Work:

- *Biology:* [Felsenstein'04]
- *TCS (Learning):* [Ambainis-Desper-Farach-Kannan'97], [Farach-Kannan'96], [Cryan-Goldberg-Goldberg'02], [Mossel-R'05]
- *Combinatorial Phylogeny:* [Erdos-Steel-Szekely-Warnow'97, '98], [Mossel'04a]

$s(1)$	$s(2)$	$s(3)$	$s(4)$	$s(5)$
0	0	1	1	1
0	0	0	1	1
1	1	0	0	1
0	0	1	1	1
1	0	0	1	1



Resolution of Steel-Mossel Conjecture

Statistical physics

Phylogeny

[DMR'05]	Low Temp		$k = O(\log n)$
[Mossel'04b]	High Temp		$k = \text{poly}(n)$
[Mossel-Steel'03]	Percolation		Random Cluster
[Mossel'04b]	Ising model		CFN Balanced Tree

Main Result: “Optimal” Reconstruction

- **Th** [Daskalakis-Mossel-R’06]: If \mathcal{T} is a tree on n leaves s.t.
 - For all e , $\theta_{\min} < \theta(e) < \theta_{\max}$ and $2\theta_{\min}^2 > 1$, $\theta_{\max} < 1$.
 - **Then** $k = O(\log n - \log \delta)$ characters suffice to reconstruct the topology with probability $1 - \delta$ (where the constant depends on $\theta_{\min}, \theta_{\max}$).
- Easy counting argument: $O(\log n)$ is necessary.
- By [Mossel’04b], the result is tight: when $2\theta^2 < 1$ polynomially samples are necessary in general.
- Our proof has same basic structure as in [Mossel’04b].

Ingredients: Tree Metrics and Recursive Majority

1) Tree metrics:

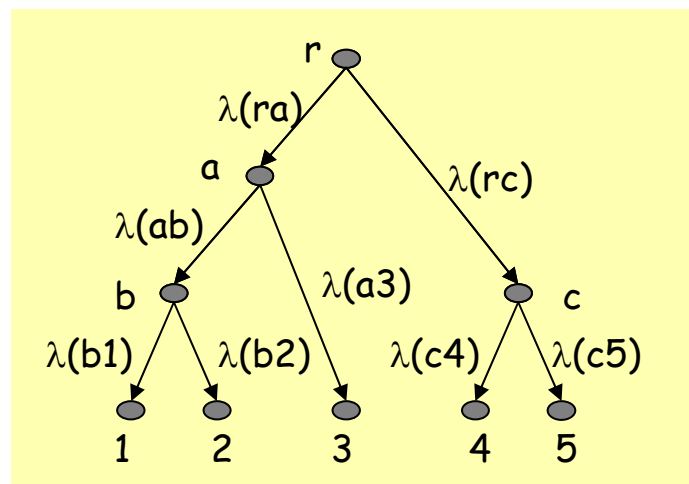
- Associate to each edge a **weight**:

$$\lambda(e) = -\frac{1}{2} \log(1 - 2p_e)$$

- Defines a **tree metric**:

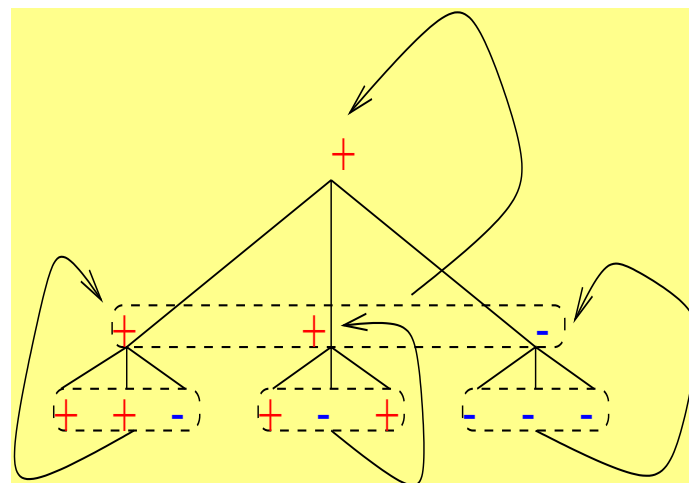
$$d_{ab} = \sum \{ \lambda(e) : e \in P_{ab} \}$$

- Reconstruction**: Deduce topology from estimated distance at leaves



2) Recursive-Majority [Mossel'98,'04b]:

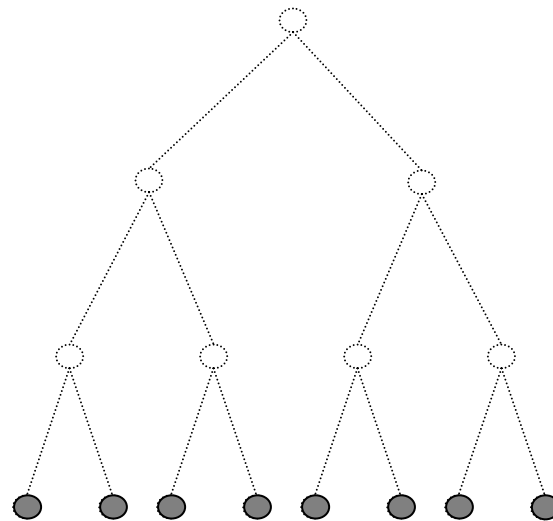
- Suppose $2\theta_{\min}^2 > 1$.
- There exists $L, \eta > 0$ s. t. if:
 - T is the binary tree with L levels;
 - $\theta(e) \geq \theta_{\min}$, for all e ;
 - Noise level $< \eta$ at leaves.
- Then, correlation between majority on leaves and root is $\geq \eta$.



Basic Algorithm: Balanced Trees

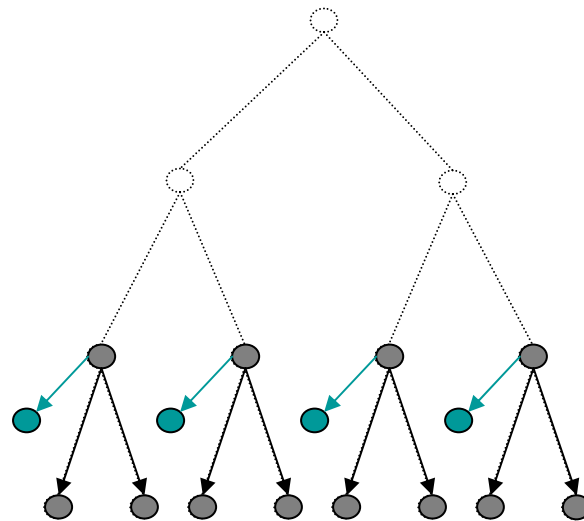
■ Two-Step Algorithm [Mossel'04b]:

- 1) Reconstruct one (or a few) level(s)
- 2) Infer sequences at roots
- 3) Start over



Basic Algorithm: Balanced Trees

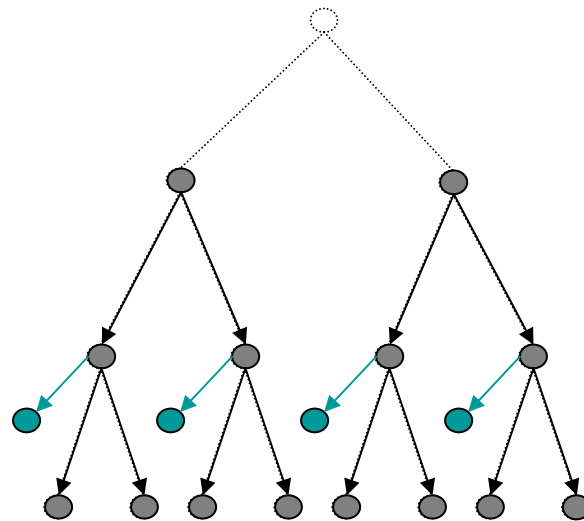
- **Two-Step Algorithm [Mossel'04b]:**
 - 1) Reconstruct one (or a few) level(s)
 - 2) Infer sequences at roots
 - 3) Start over



Basic Algorithm: Balanced Trees

■ Two-Step Algorithm [Mossel'04b]:

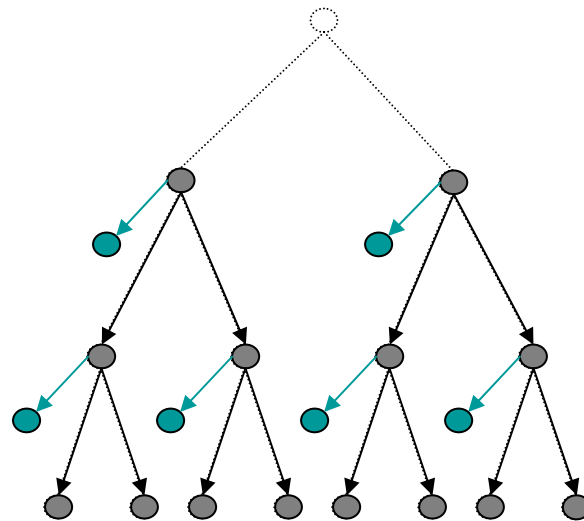
- 1) Reconstruct one (or a few) level(s)
- 2) Infer sequences at roots
- 3) Start over



Basic Algorithm: Balanced Trees

■ Two-Step Algorithm [Mossel'04b]:

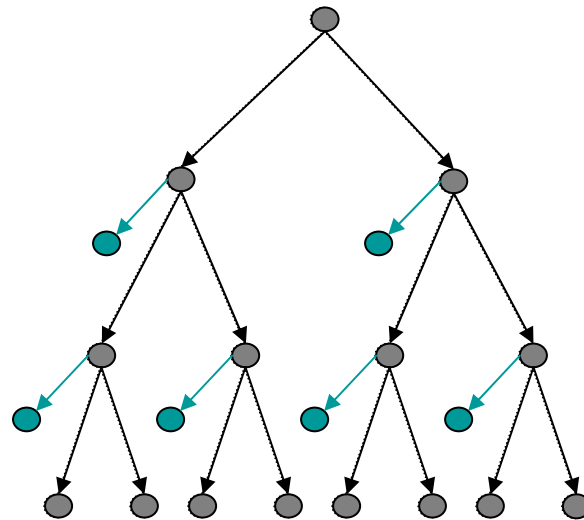
- 1) Reconstruct one (or a few) level(s)
- 2) Infer sequences at roots
- 3) Start over



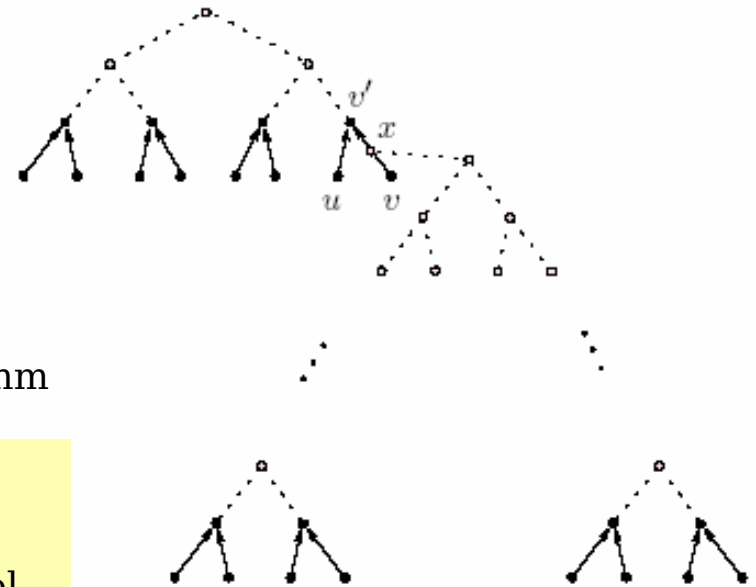
Basic Algorithm: Balanced Trees

■ Two-Step Algorithm [Mossel'04b]:

- 1) Reconstruct one (or a few) level(s)
- 2) Infer sequences at roots
- 3) Start over



General Trees: Blindfolded Cherry Picking

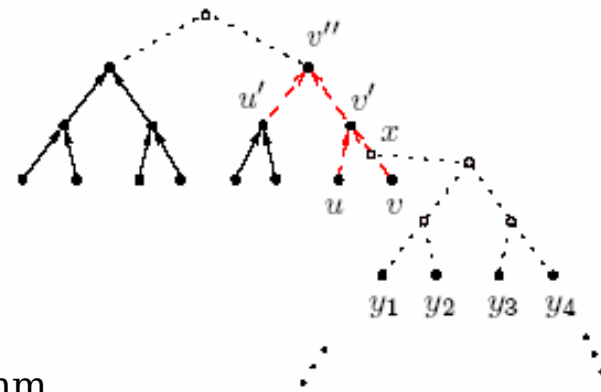


- Need “only” **one extra step** in the algorithm

■ **Main Loop:**

- 1) Distance estimation
- 2) Identify cherries from the next level
- 3) Sequence reconstruction
- 4) Detect “fake cherries”

General Trees: Blindfolded Cherry Picking



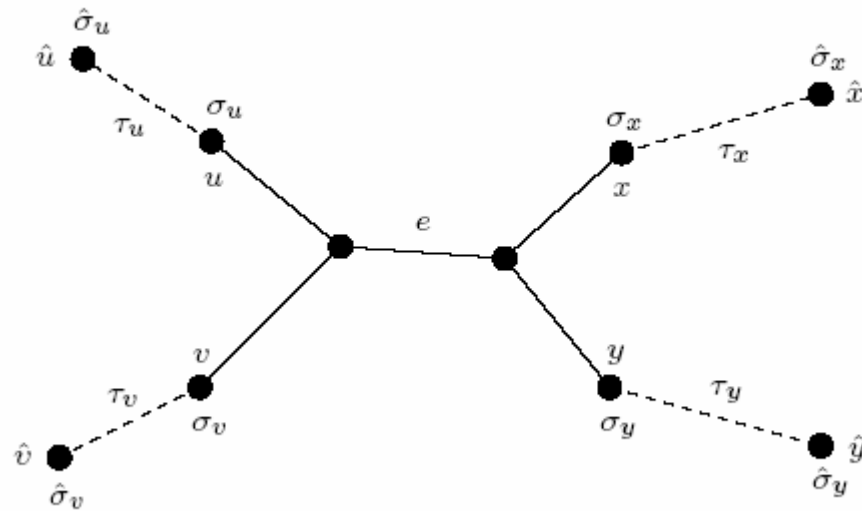
- Need “only” **one extra step** in the algorithm

■ **Main Loop:**

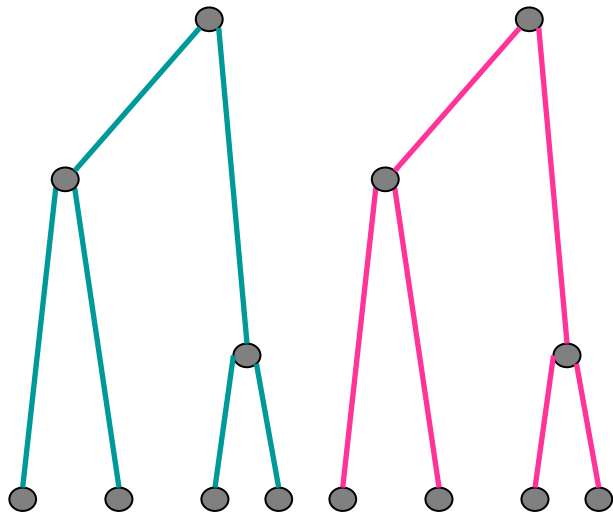
- 1) Distance estimation
- 2) Identify cherries from the next level
- 3) Sequence reconstruction
- 4) Detect “fake cherries”



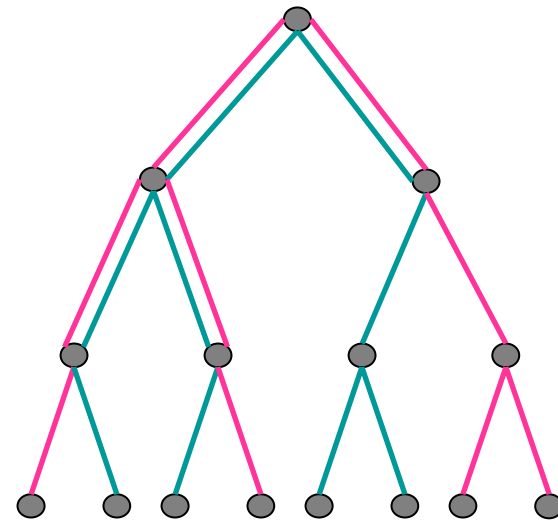
Blindfolded Cherry Picking I: Weight Estimation



Blindfolded Cherry Picking II: Edge Disjointness

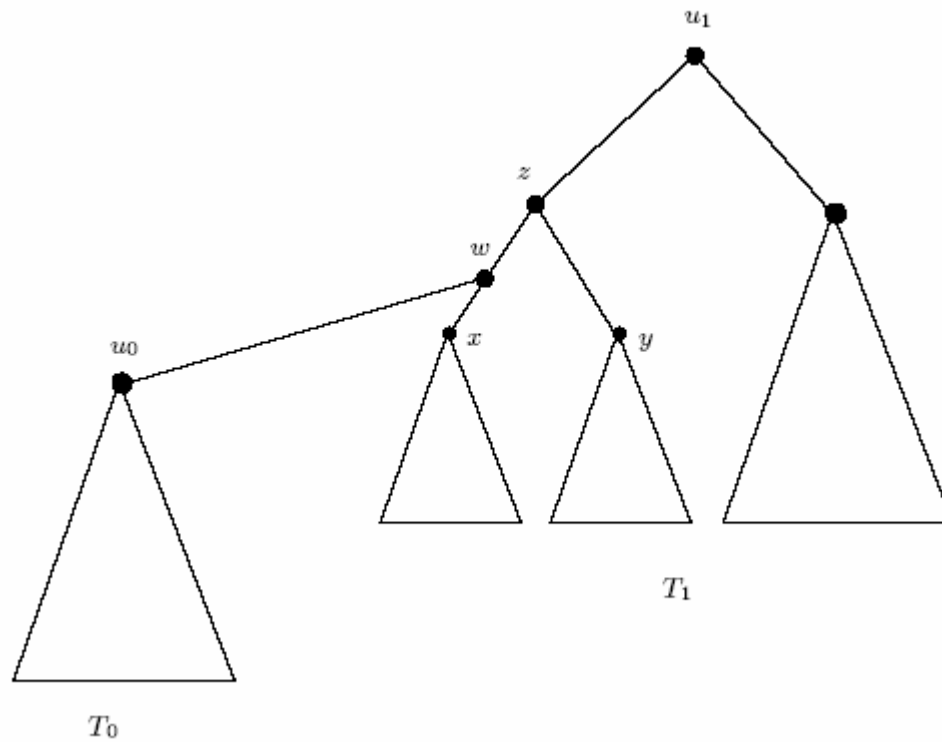


Non Edge-Disjoint Reconstruction



True Tree

Blindfolded Cherry Picking III: Collisions



Open Problems

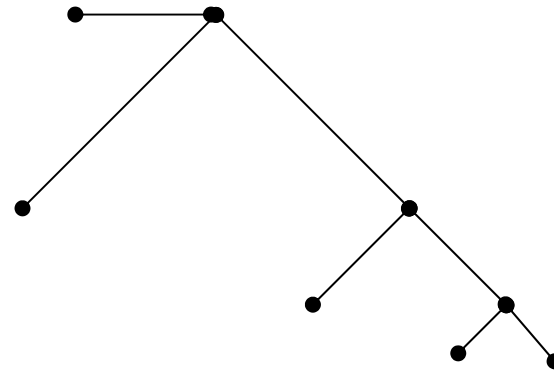
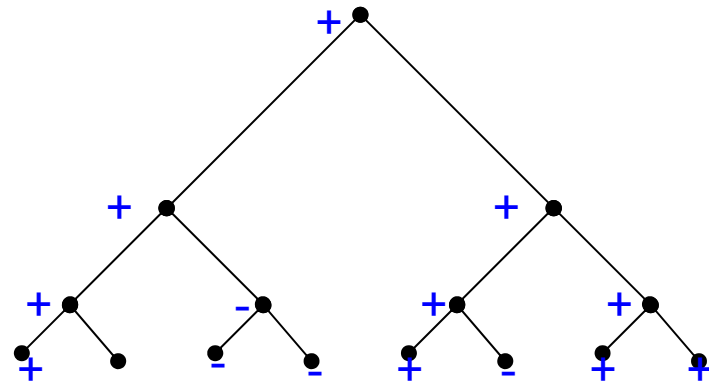
- Can this result be extended to more general models of evolution?
- Can it actually be useful in practice?



Thanks!

Trees

- (3-)regular trees.
- Binary -- All internal degrees are 3 (bifurcating speciation; results valid if degrees are ≥ 3 , or $\geq b+1$).
- General trees.



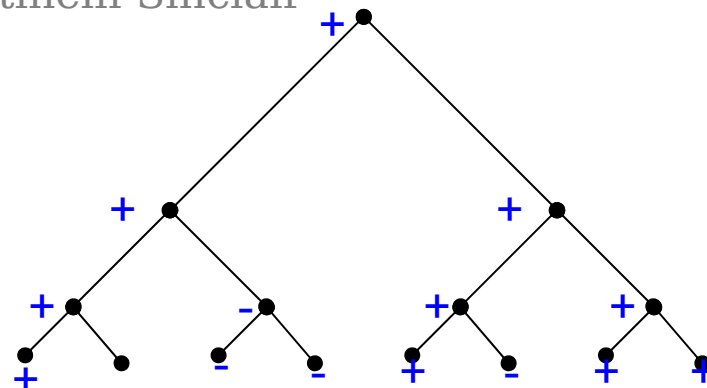
Markov Models on Trees

- Markov Chain on a Tree:
 - Finite set A of information values.
 - Tree $T=(V,E)$ rooted at r .
 - Vertex v in V , has information σ_v in A .
 - Edge $e=(v, u)$, where v is the parent of u , has a mutation matrix M^e of size $|A| \times |A|$:
 - $M_{i,j}^{(v,u)} = P[\sigma_u = j \mid \sigma_v = i]$
- For each character σ , we are given $\sigma_{\partial T} = (\sigma_v)_{v \text{ in } \partial T}$, where ∂T is the **boundary** of the tree.
- We will focus on the Ising-CFN model:

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

Statistical Physics on Trees

- The **Ising** model on the binary tree can be defined:
 - Set σ_r , the root spin, to be $+/-$ with probability $\frac{1}{2}$.
 - For all pairs of (parent, child) = (v, w) , set $\sigma_w = \sigma_v$, with probability θ , otherwise $\sigma_w = +/-$ with probability $\frac{1}{2}$.
- This is exactly the **CFN** model.
- Studied in statistical physics [Spitzer 75, Higuchi 77, Bleher-Ruiz-Zagrebnoy 95, Evans-Kenyon-Peres-Schulman 2000, Ioffe 99, M 98, Haggstrom-M 2000, Kenyon-M-Peres 2001, Martinelli-Sinclair Weitz 2003, Martine 2003]



Reconstruction Solvability

- Let \mathcal{T} be an infinite rooted tree and \mathcal{T}_n denote the first n levels of \mathcal{T} .
- We say that the **reconstruction problem is solvable** if *one* of the following equivalent conditions hold:
 - Exists π s.t. (for all non-degenerate π) $\lim_n I(X_0, X_n) > 0$, where $I(X_0, X_n) = H(X_0) + H(X_n) - H(X_0, X_n)$; H is the **entropy operator**, $H(X) = -\sum_x P[X = x] \log_2 P[X = x]$.
 - Exists i, j s.t. $\lim_n |P_n^i - P_n^j| > 0$, where P_n^j denotes the distribution of X_n conditional on $X_0 = j$.
 - If X_0 has the uniform distribution then, $\liminf_n \Delta_n > 1/m$, where Δ_n is the probability of correct reconstruction of X_0 given X_n .
 - Exists π (for all non-degenerate π) $\liminf_n \text{Var}[E[X_0|X_n]] > 0$.

Reconstruction by Majority

- Let \mathcal{M} be the Ising (BSC) model on a b -ary tree \mathcal{T} .
- Is σ_0 correlated with $f(\sigma_n) = \text{sign}(\sum\{\sigma(v) : v \text{ in } L_n\})$?
- **Theorem** [Higuchi 77]:
 - $\lim_n P[\sigma_0 = f(\sigma_n)] > \frac{1}{2}$ if $b\theta^2 > 1$.
- Generalization:
 - Let \mathcal{M} be any chain and \mathcal{T} the b -ary tree
 - Let λ be the 2nd eigenvalue of \mathcal{M} in absolute value.
 - **Theorem**[Kesten-Stigum66] $b |\lambda|^2 > 1$ reconstruction.

A Diagram

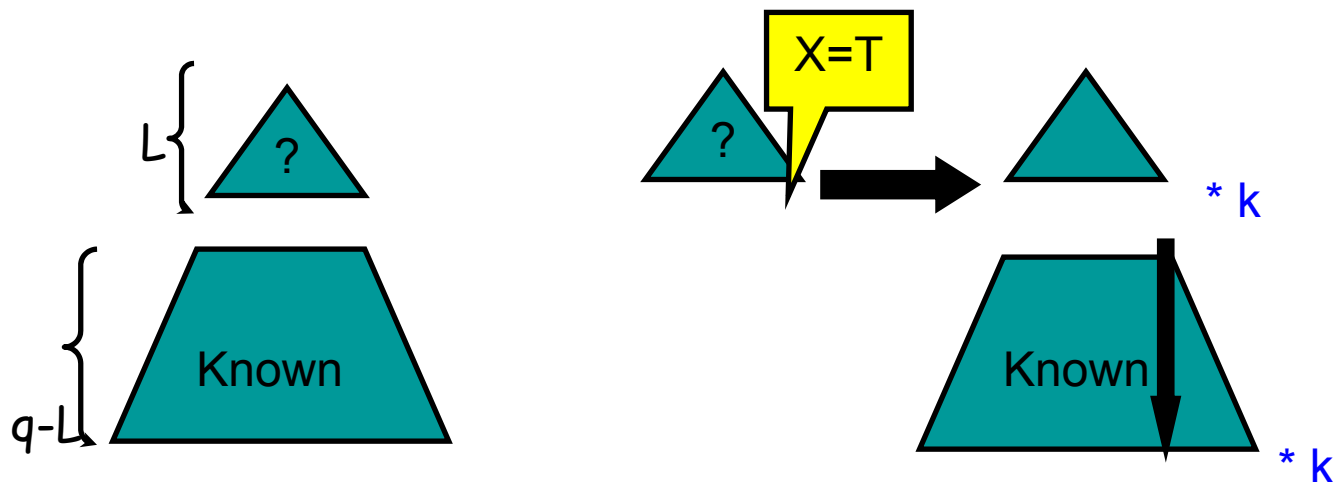
$$\begin{array}{ccc} \text{Trees} \otimes [\theta_{\min}, \theta_{\max}] & \longrightarrow & \sigma \\ \uparrow \psi & & \downarrow \otimes^k \\ (\sigma_{\partial}^t)_{t=1}^k & \longleftarrow & (\sigma^t)_{t=1}^k \end{array}$$

Polynomial Lower Bound at High Mutation Rates I

- **Th1**[Mossel 2004]: Suppose that $n=3 \times 2^q$ and
 - \mathbb{T} is a uniformly chosen $(q+1)$ -level 3-regular tree.
 - For all e , $\theta(e) < \theta$, and $2\theta^2 < 1$.
 - **Then** in order to *reconstruct* the topology with probability $\delta > 0.1$, at least $k = \Omega(n^{(-2\log_2(\theta) - 1)})$ characters are needed.
- **In Fact:**
 - **Thm [ESSW97]** Suppose that for all e , $1 - \varepsilon > \theta(e) > \varepsilon > 0$. Then given k characters of the process at n leaves, it is possible to reconstruct the underlying topology with probability $1 - \delta$, if $k = n^{O(-\log \varepsilon)}$.

Polynomial Lower Bound at High Mutation Rates II

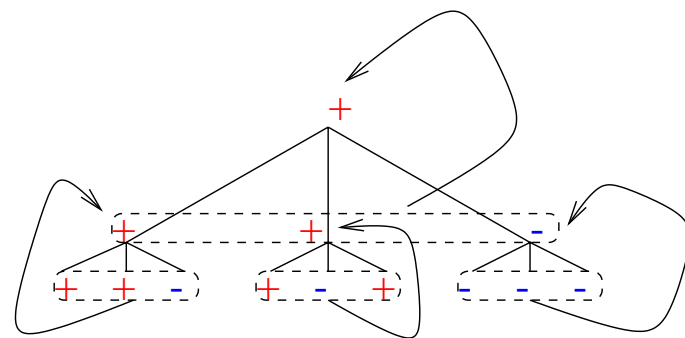
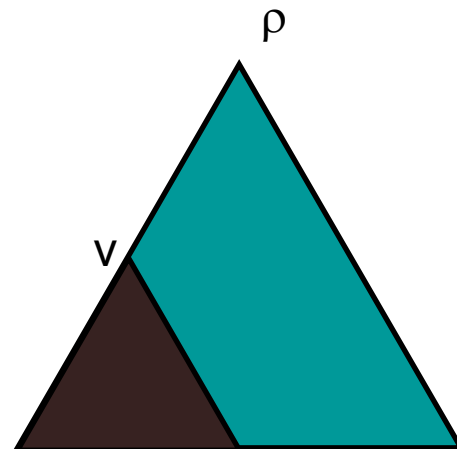
- **Proof:**



- **Mutual Information:** $I(X,Y) = H(X) - H(X | Y)$
- **Data Processing Lemma:** If X and Z are cond. indep. given Y then $I(X,Y) \geq I(X,Z)$

Proof Ia: Recursive Reconstruction for Ising Models

- **One (of many) proof** for reconstruction for $\theta^2 > 1$ [Mossel98]
- **Advantage:** Works also when we have **lower bound** on θ . Majority doesn't.
 - Blue edges have θ^1 , black θ^2 , $\theta^1 < \theta^2 \sim 1$.
 - $\text{Maj}(\sigma_\rho) \sim \text{Maj}$ of black tree.
 - Maj of black tree $\sim \sigma_v$.
 - σ_v and σ_ρ have exp. small correlation.
 - **Phylogeny:** reconstruction given bounds.
- Instead we will use **recursive-majority**.



Proof: General Case

Proposition 2 (Properties of $\widehat{\mathcal{F}}_i$) *The following properties hold at the beginning of BCP's i -th iteration, $\forall i \geq 1$:*

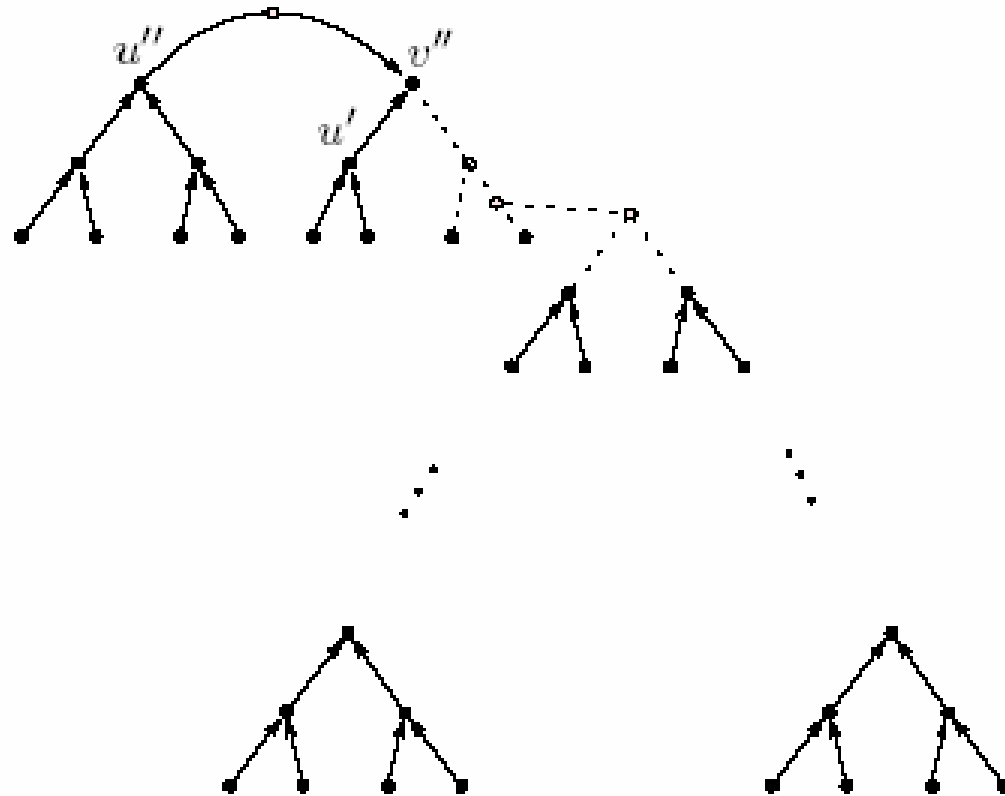
1. [Edge Disjointness] $\widehat{\mathcal{F}}_i = \{T_{\leq u}^{\text{Child}} : u \in \widehat{L}_i\}$ is an edge disjoint subforest of T .
2. [Edge Lengths] $\forall u \in \widehat{L}_i, T_{\leq u}^{\text{Child}}$ is a rooted full binary tree with edge lengths at most g .
3. [Weight Estimation] *The estimated lengths of the edges in $\widehat{\mathcal{F}}_i$ are within ε_2 from their right values.*
4. [Collisions] *There is no collision at distance $20g$.*

Proposition 3 (Progress) *Let*

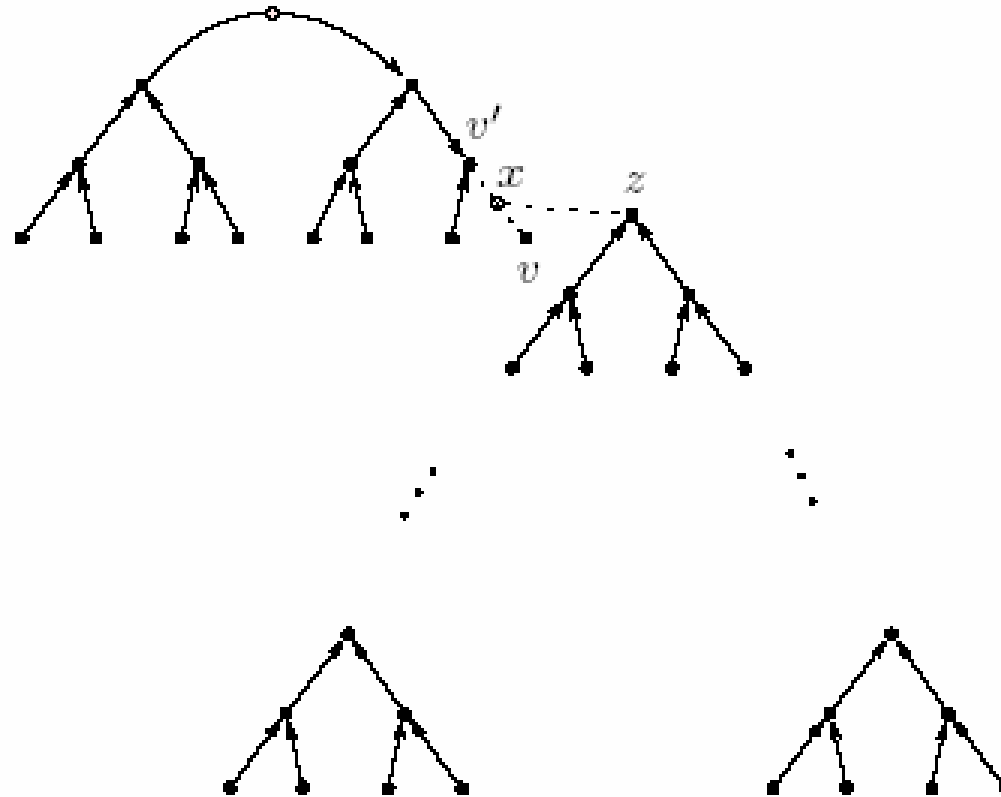
$$\widehat{\mathcal{F}}_i = \{T_{\leq u}^{\text{Child}} : u \in \widehat{L}_i\}$$

(where \widehat{L}_i is taken at the beginning of iteration i) for all $i \geq 0$ with corresponding maximal fixed subforest $\widehat{\mathcal{F}}_i^$. Then for all $i \geq 0$ (before the termination step), $\widehat{\mathcal{F}}_i^* \subseteq \widehat{\mathcal{F}}_{i+1}^*$ and $|\mathcal{V}(\widehat{\mathcal{F}}_{i+1}^*)| > |\mathcal{V}(\widehat{\mathcal{F}}_i^*)|$.*

General Trees [Daskalakis, Mossel, R, 2005]



General Trees [Daskalakis, Mossel, R, 2005]



Reconstruction Solvability

- Let \mathcal{T} be an infinite rooted tree and \mathcal{T}_n denote the first n levels of \mathcal{T} .
- We say that the **reconstruction problem is solvable** if *one* of the following equivalent conditions hold:
 - $\lim_n |P_n^0 - P_n^1| > 0$, where P_n^j denotes the distribution of X_n conditional on $X_0 = j$.
 - If X_0 has the uniform distribution then, $\liminf_n \Delta_n > 1/2$, where Δ_n is the probability of correct reconstruction of X_0 given X_n .
- **Theorem** [Higuchi'77]: Let $f(\sigma_n) = \text{sign}(\sum\{\sigma(v) : v \text{ in } L_n\})$. Then $\lim_n P[\sigma_0 = f(\sigma_n)] > \frac{1}{2}$ if $2\theta^2 > 1$.
- An alternative proof for reconstruction for $2\theta^2 > 1$ [Mossel'98]: **recursive-majority**.

$$\theta = 1 - 2p$$

