The consequences of approximate conditional independence of subtrees on phylogeny estimation

Bret Larget
Departments of Botany
and of Statistics,
UW-Madison

Cats and Dogs Example

- Consider a data set from 12 species, closely related to domestic cats and dogs.
- The alignment is the 1545 base pair mitochondrial COX I gene.

Partial Alignment

```
Cat             ATGTTCATAAACCGGTGACTATTTTCAACTAATCACAAAGATATTGGTACTCTTTACCTT...
Cheetah         ATGTTCATAATCCGCTGATTATTTTCAACTAATCATAAAGATATCGGTACTCTTTACCTC...
Clouded Leopard ATGTTCATAAACCGCTGACTATTTTCAACTAACCATAAAGATATTGGAACTCTTTACCTT...
Snow Leopard    ATGTTCATAAACCGCTGACTATTTTCAACCAATCACAAAGATATTGGAACTCTTTACCTT...
Leopard         ATGTTCATAAACCGCTGACTATTTTCAACCAATCACAAAGATATTGGAACTCTTTACCTT...
Tiger           ATGTTCATAAACCGCTGACTATTTTCAACCAATCACAAGGATATTGGAACTCTTTACCTT...
Dog             ATGTTCATTAACCGATGATTGTTCTCCACTAATCACAAGGATATTGGTACTTTATACTTA...
Gray Wolf       ATGTTCATTAACCGATGATTGTTCTCCACTAATCACAAGGATATTGGTACTTTATACTTA...
Coyote          ATGTTCATTAACCGATGATTGTTCTCTACTAATCACAAAGATATTGGTACTTTATATCTA...
Dhole           ATGTTCATTAACCGATGGTTATTCTCTACTAATCACAAAGATATTGGGACTTTGTATCTA...
Red Fox         ATGTTCATTAATCGATGATTATTCTCTACTAACCACAAAGACATCGGTACTTTATATTTG...
Raccoon Dog     ATGTTCATTAACCGATGACTATTCTCTACTAACCACAAAGACATTGGCACTTTATATTTA...
```

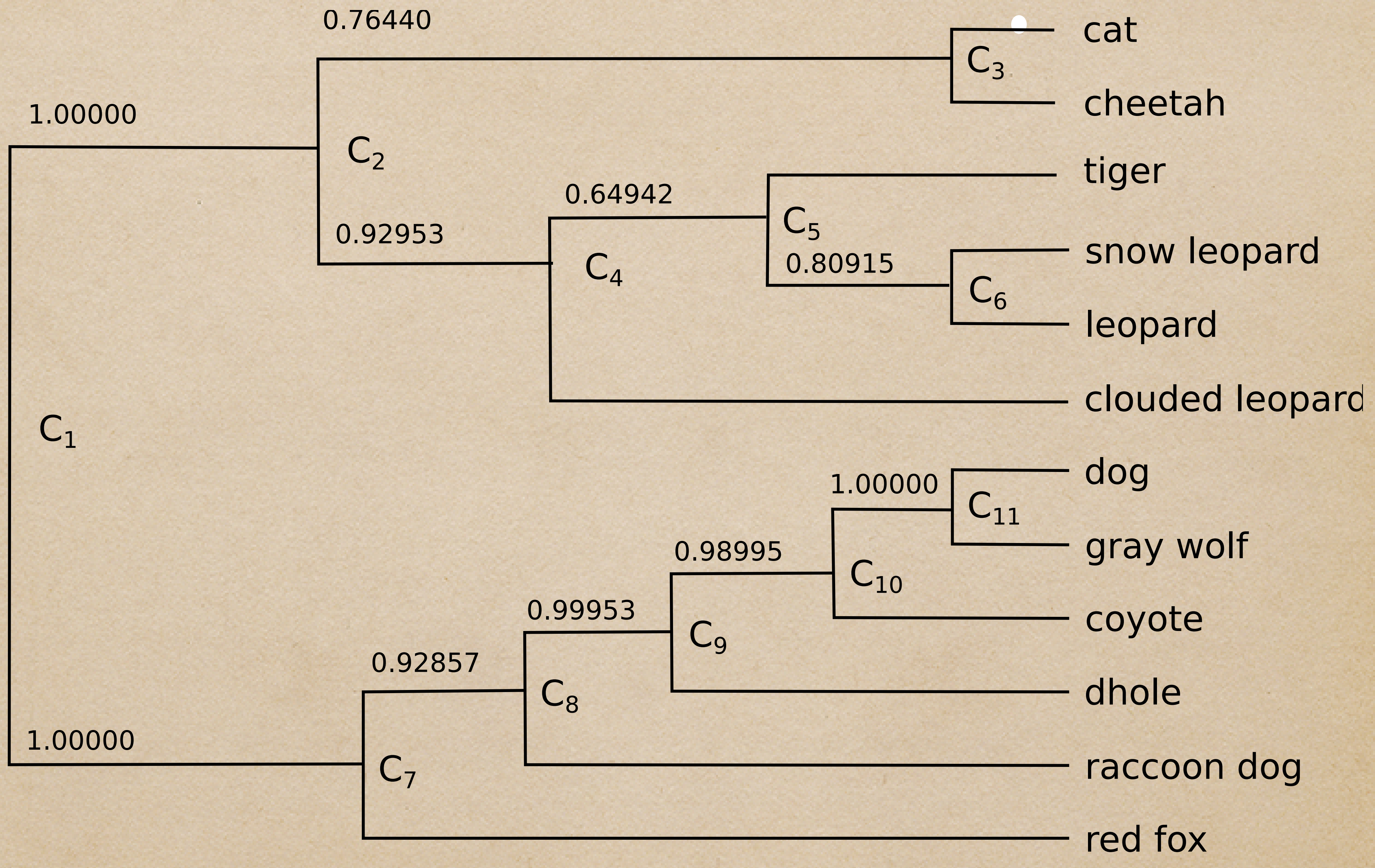
MCMC Samples

- Statisticians develop MCMC methods to sample trees from the posterior distribution.
- The posterior distribution of each topology T is estimated with its simple relative frequency in this sample.

Cat/Dog Example

- For a given likelihood model, one MCMC sample of 100,000 trees includes 229 unique trees.
- With a rooting that separates the dog-like and cat-like species, the following display shows the tree and the estimated probability of each of its clades.

A Probable Tree



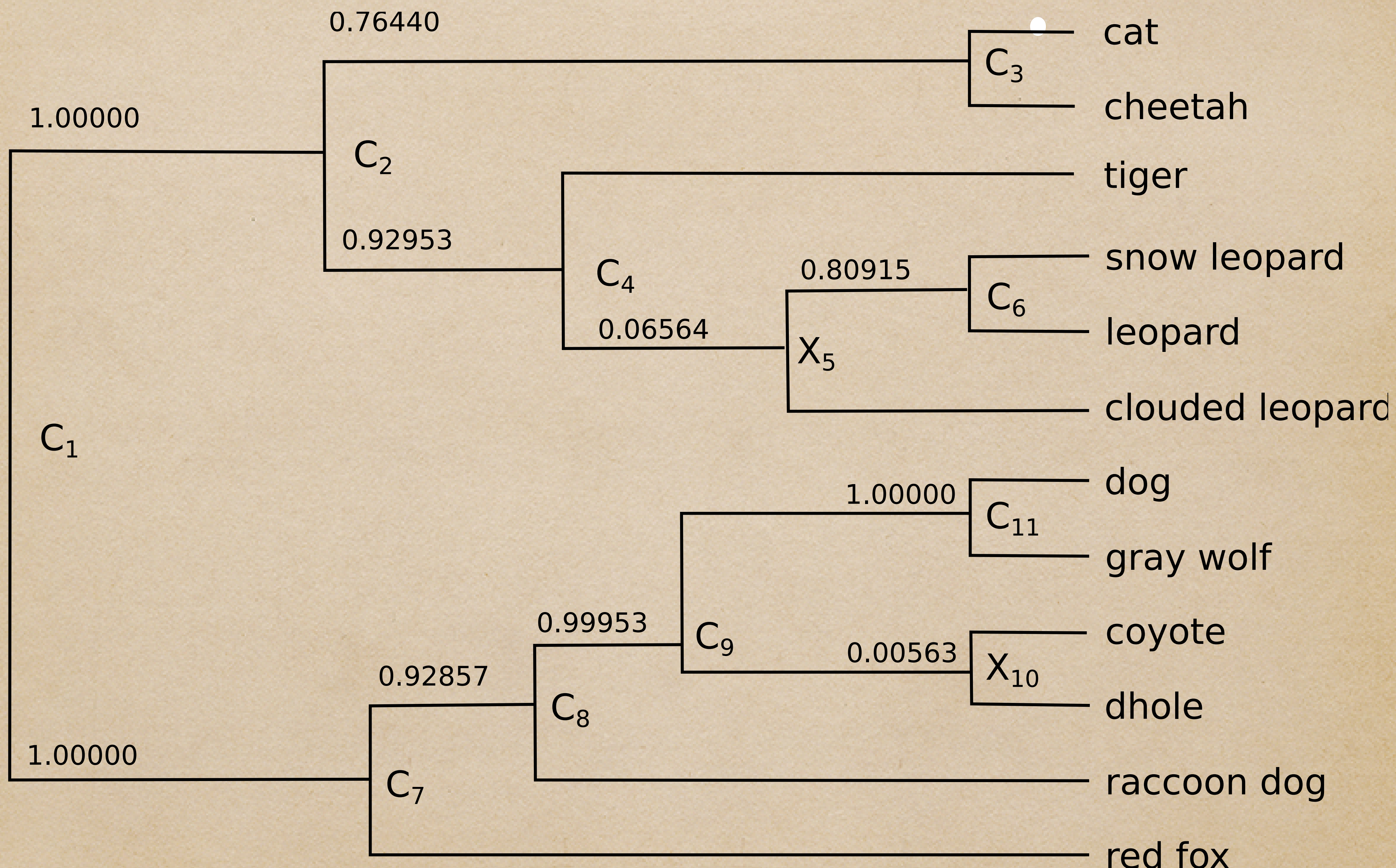
MCMC Summary

- There were 229 unique trees among the 100,000 sampled trees.
- 50 of these trees were sampled once.
- Different MCMC samples would contain slightly different sets of trees.
- Some unsampled trees are much more probable than others.

Trees as Intersections of Clades

- The tree is equivalent to its collection of clades.
- $P(T_1)= P(C_2 \cap C_3 \cap \cdots \cap C_{11})$
- However, the clades are not independent of each other.
- $P(T_1) \neq P(C_2) \times \cdots \times P(C_{11})$

A Less Probable Tree

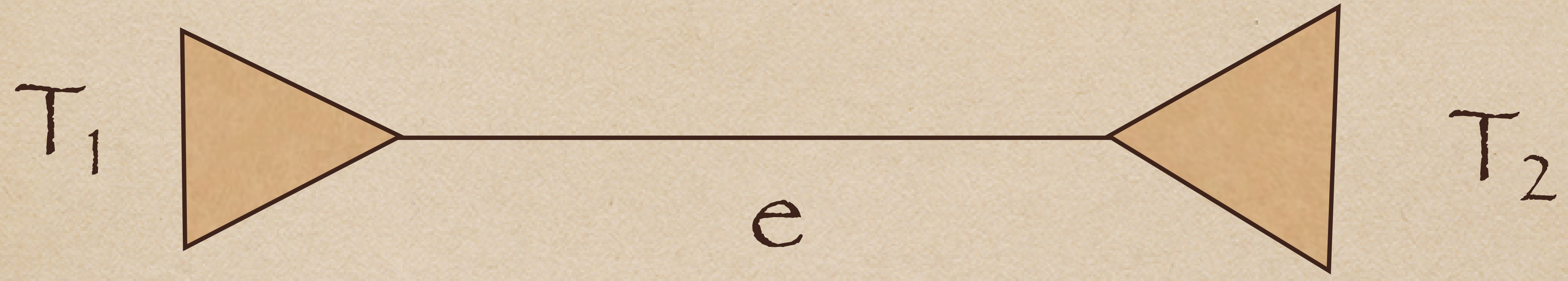


Simple Relative Frequencies

- The most probable tree had estimated probability 0.33925.
- The less probable tree had estimated probability 0, even though each of its clades had appeared multiple times in the sample.

Conditional Independence of Separated Subtrees

- A new alternative method to estimate the probabilities of trees from MCMC samples depends on the principle of the (approximate) conditional independence of separated subtrees.



Given e, the subtrees T_1 and T_2 are conditionally independent.

General Principle

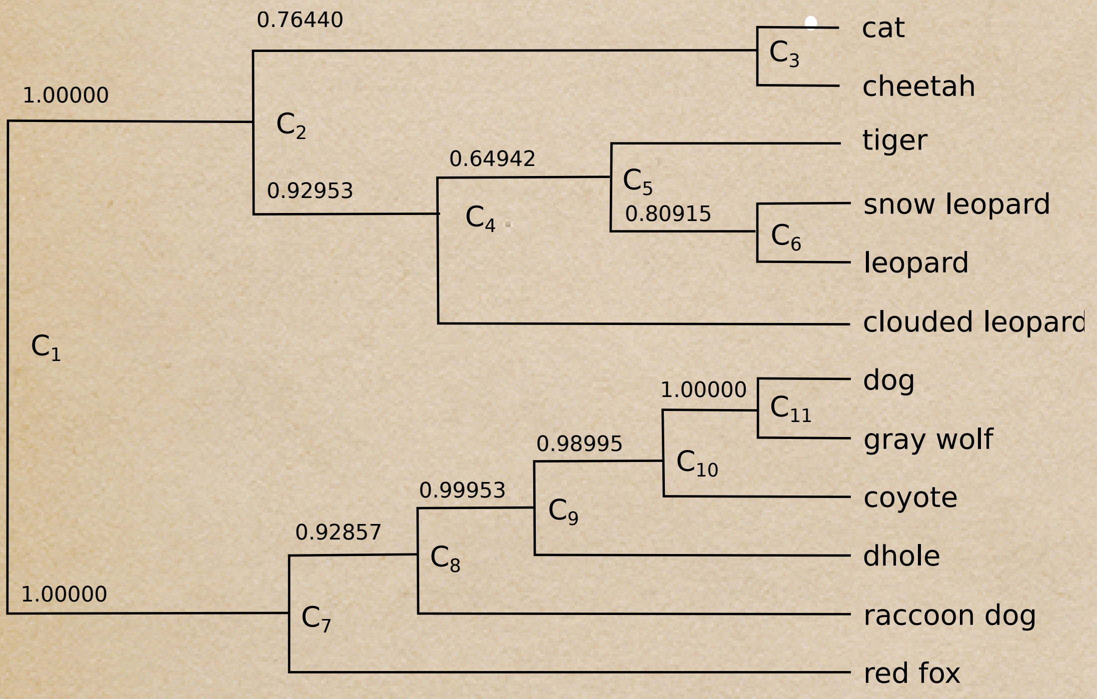
- Let split $s = A|B$ be a partition of the species set into two nonempty sets, A and B.
- Let $A_1,\ldots,A_m \subset A$ and $B_1,\ldots,B_n \subset B$ and let each subset also denote the existence of an edge in the tree that splits all species into itself and its complement.
- We say that s separates $\{A_i\}$ from $\{B_j\}$.
- Then, $P(A_1 \cap \cdots \cap A_m \cap B_1 \cap \cdots \cap B_n \mid s)$

$$\approx P(A_1 \cap \cdots \cap A_m \mid s) \times P(B_1 \cap \cdots \cap B_n \mid s)$$

Algorithm

- The principle of conditional independence of separated subtrees leads to an algorithm to approximate the probabilities of trees.
- This approximate probability is the product over nontrivial clades of the conditional probabilities of their subclades given the clade.

Example Calculation



$P(T1) \approx$

$P(C2 \cap C7 \mid C1)$

$\times\ P(C3 \cap C4 \mid C2)$

$\times\ P(C5 \cap \text{clouded leopard} \mid C4)$

$\times\ P(C6 \cap \text{tiger} \mid C5)$

$\times\ P(C8 \cap \text{red fox} \mid C7)$

$\times\ P(C9 \cap \text{raccoon dog} \mid C8)$

$\times\ P(C10 \cap \text{dhole} \mid C9)$

$\times\ P(C11 \cap \text{coyote} \mid C10)$

$$\frac{100000}{100000} \times \frac{69393}{100000} \times \frac{63655}{92953} \times \frac{50827}{64942} \times \frac{92857}{100000} \times \frac{92857}{92857} \times \frac{98948}{99953} \times \frac{98995}{98995} \doteq 0.3419$$

Compare to SRF estimate 0.3393

Calculation Details

- Each conditional probability takes the form:

  $$P(C_i \cap C_j \mid C_k) = P(C_i \cap C_j \cap C_k) / P(C_k)$$

- Note that $P(C_i \cap C_j \cap C_k)$ is the probability that the unrooted tree contains a node that partitions the species into three groups and $P(C_k)$ is the probability of an edge.

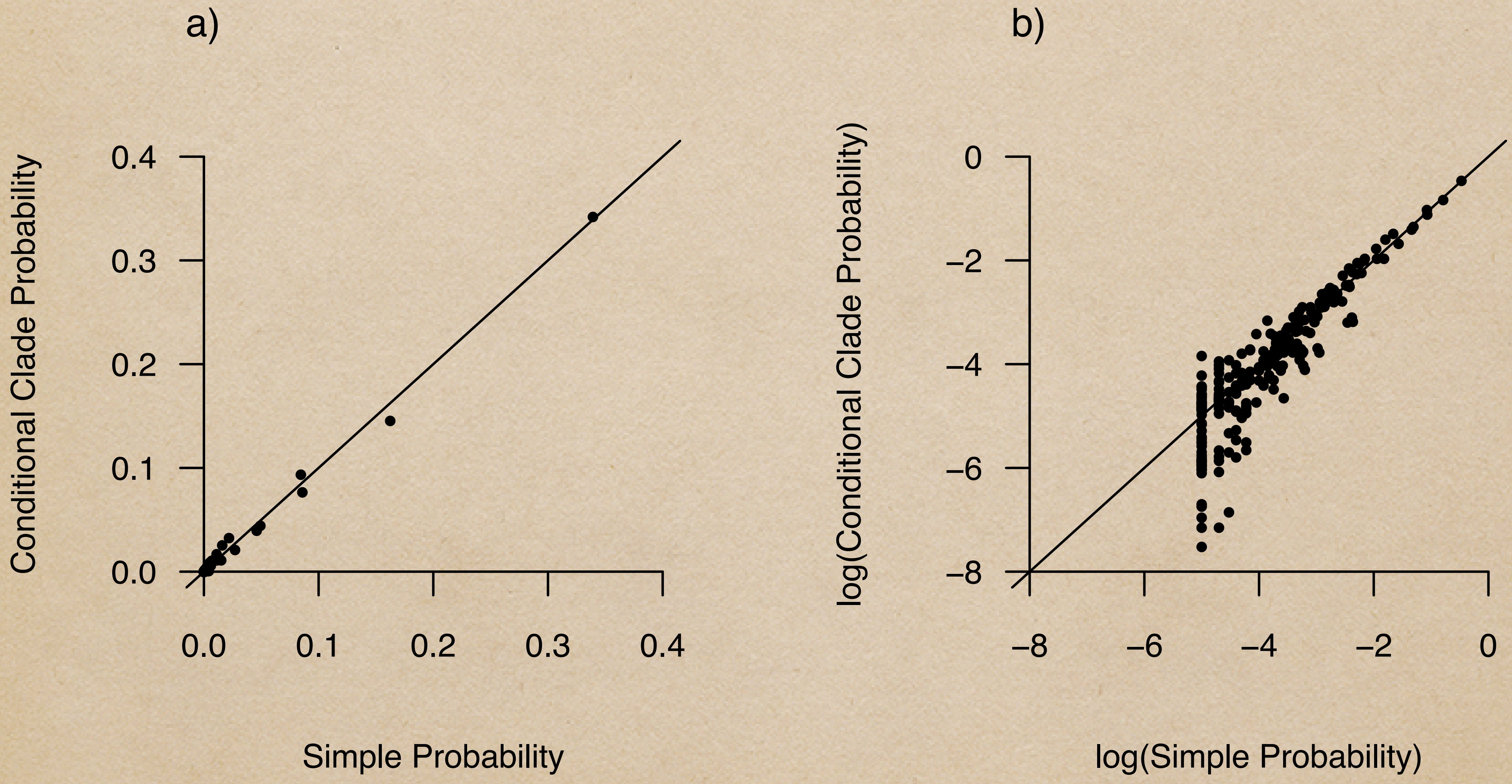
Unrooted Tree Version

- $$P(\tau) \approx \frac{\prod_{n \in \text{internal}(\tau)} \{ P(\text{triple split }(n)) \}}{\prod_{e \in \text{edges}(\tau)} \{ P(\text{edge } e) \}}$$

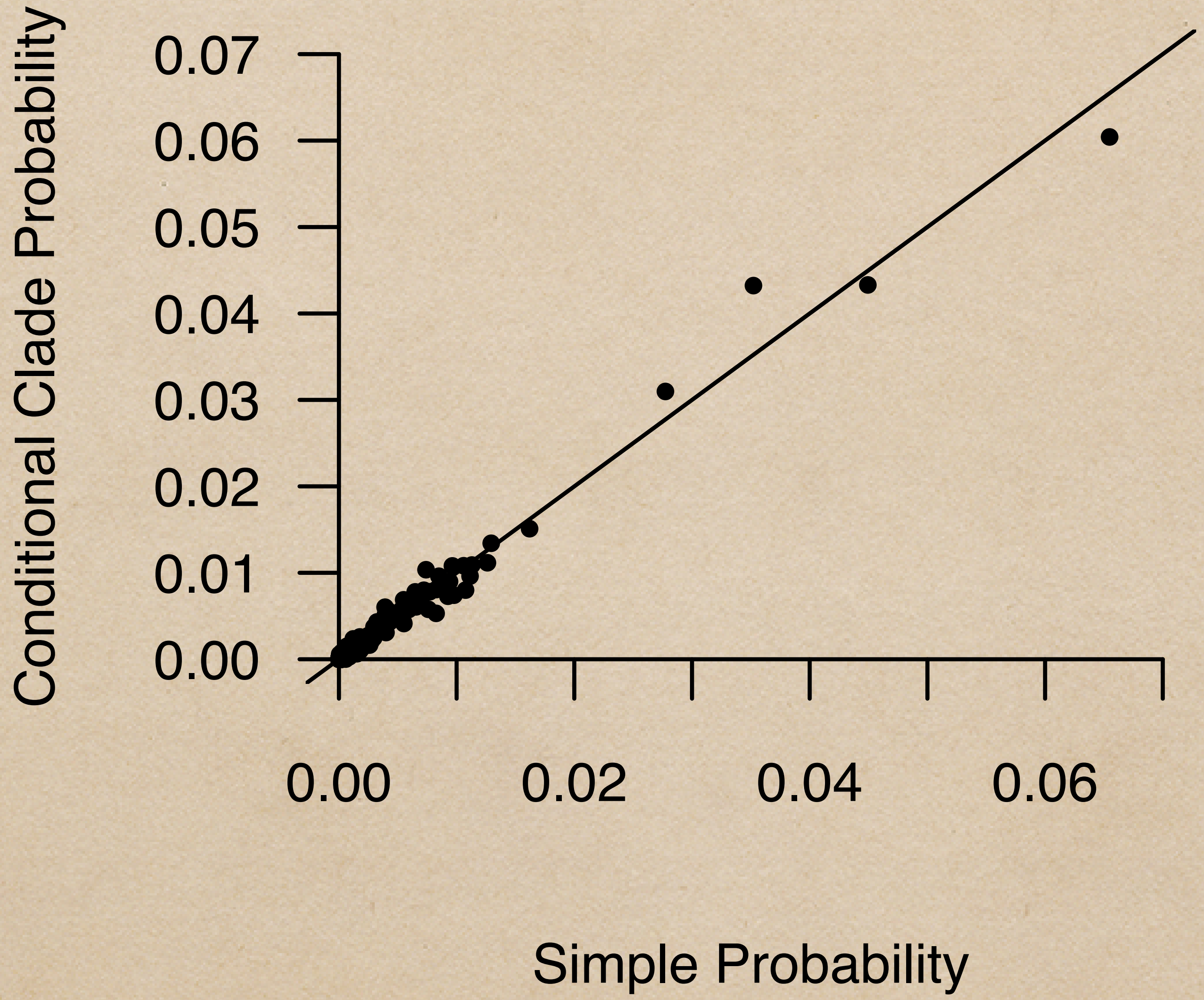
Estimating Conditional Probabilities

- Take an MCMC sample of trees.
- For each sampled tree, increment counters for each clade (edge partition) it contains and for each triple (internal node partition).
- Estimated conditional probabilities are ratios of node counts over edge counts.

Estimates for Cat/Dog Trees



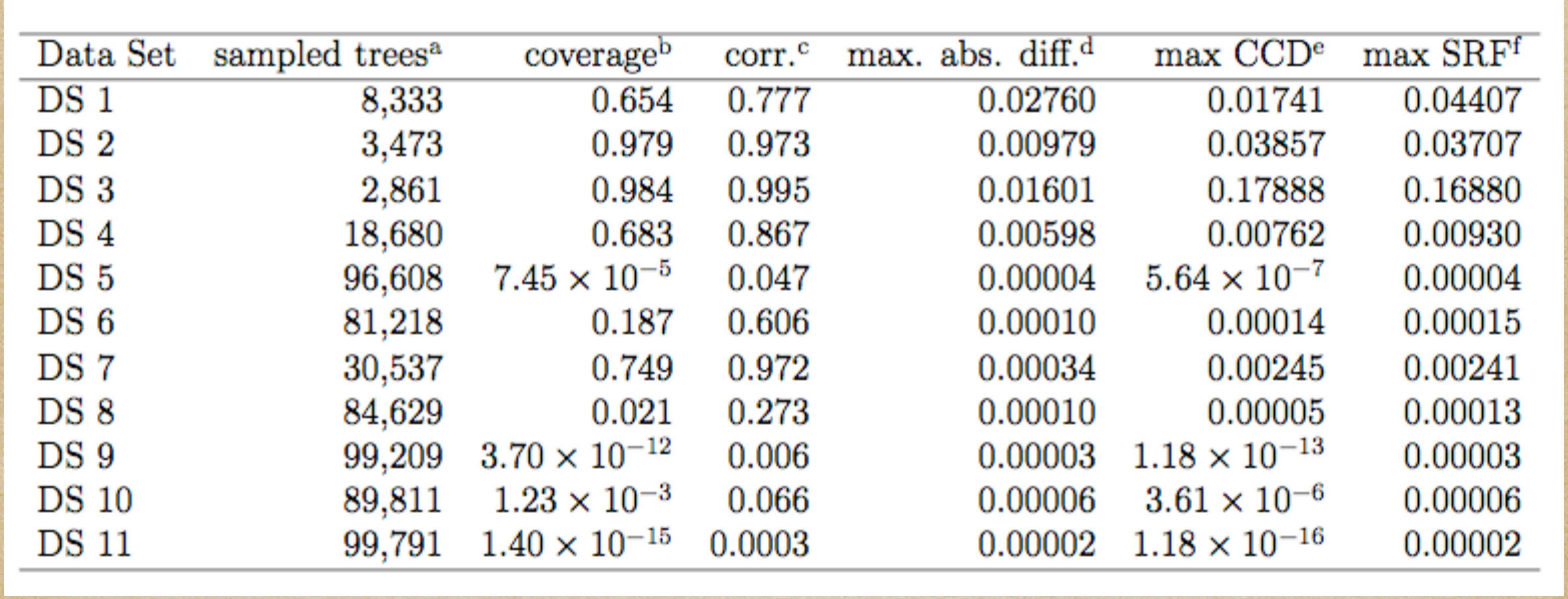
Larger 62-species Example



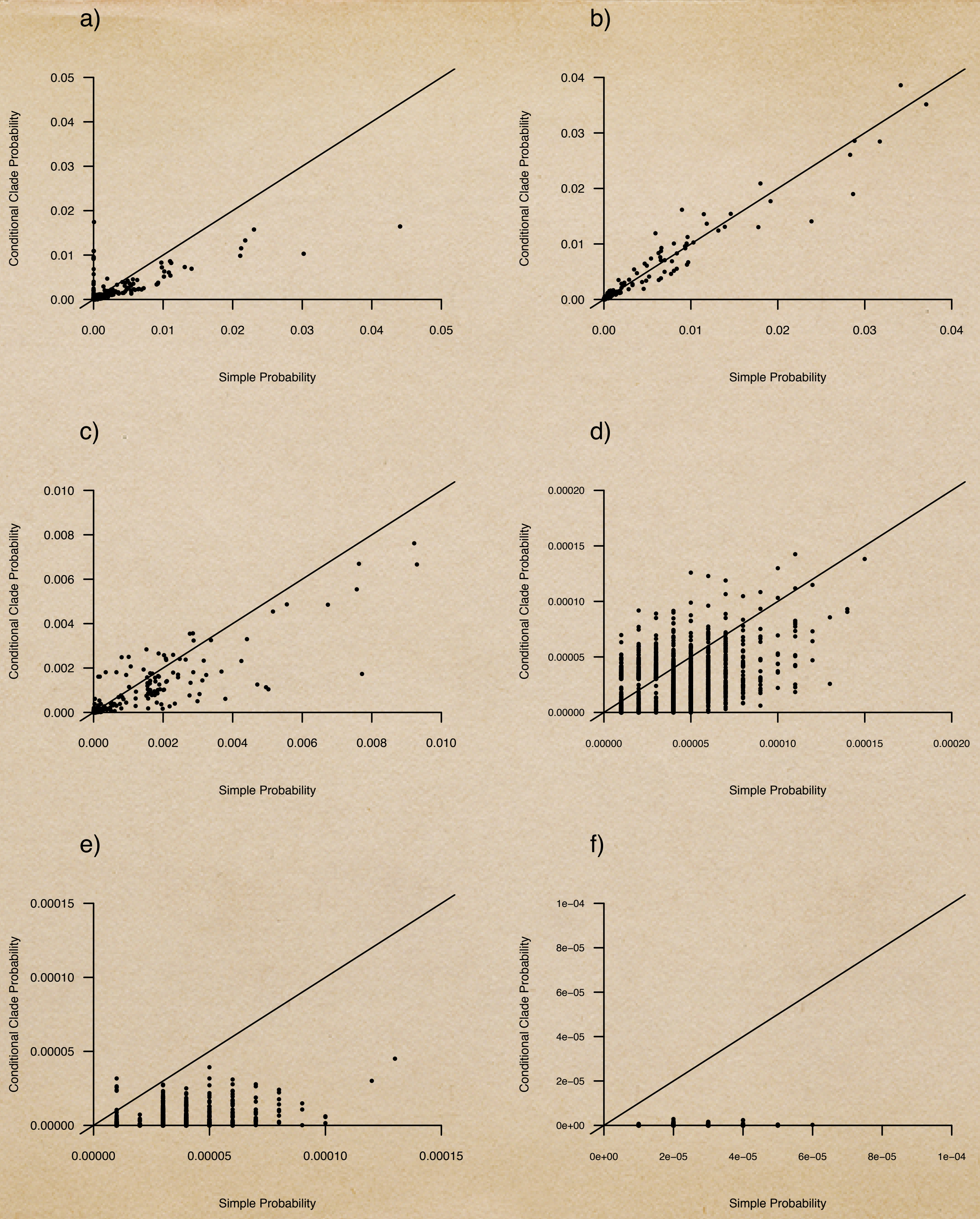
Coverage

- The sum of estimated probabilities taken over the entire MCMC sample is an estimate of the coverage of the sample.
- In the cat and dog example, the 229 trees in the sample are estimated to represent about 99.8 % of the total probability.

Coverage for 11 Data Sets



Data Set	sampled trees[a]	coverage[b]	corr.[c]	max. abs. diff.[d]	max CCD[e]	max SRF[f]
DS 1	8,333	0.654	0.777	0.02760	0.01741	0.04407
DS 2	3,473	0.979	0.973	0.00979	0.03857	0.03707
DS 3	2,861	0.984	0.995	0.01601	0.17888	0.16880
DS 4	18,680	0.683	0.867	0.00598	0.00762	0.00930
DS 5	96,608	7.45×10^{-5}	0.047	0.00004	5.64×10^{-7}	0.00004
DS 6	81,218	0.187	0.606	0.00010	0.00014	0.00015
DS 7	30,537	0.749	0.972	0.00034	0.00245	0.00241
DS 8	84,629	0.021	0.273	0.00010	0.00005	0.00013
DS 9	99,209	3.70×10^{-12}	0.006	0.00003	1.18×10^{-13}	0.00003
DS 10	89,811	1.23×10^{-3}	0.066	0.00006	3.61×10^{-6}	0.00006
DS 11	99,791	1.40×10^{-15}	0.0003	0.00002	1.18×10^{-16}	0.00002


a)
Conditional Clade Probability
0.05
0.04
0.03
0.02
0.01
0.00
0.00 0.01 0.02 0.03 0.04 0.05
Simple Probability
b)
Conditional Clade Probability
0.04
0.03
0.02
0.01
0.00
0.00 0.01 0.02 0.03 0.04
Simple Probability
c)
Conditional Clade Probability
0.010
0.008
0.006
0.004
0.002
0.000
0.000 0.002 0.004 0.006 0.008 0.010
Simple Probability
d)
Conditional Clade Probability
0.00020
0.00015
0.00010
0.00005
0.00000
0.00000 0.00005 0.00010 0.00015 0.00020
Simple Probability
e)
Conditional Clade Probability
0.00015
0.00010
0.00005
0.00000
0.00000 0.00005 0.00010 0.00015
Simple Probability
f)
Conditional Clade Probability
1e−04
8e−05
6e−05
4e−05
2e−05
0e+00
0e+00 2e−05 4e−05 6e−05 8e−05 1e−04
Simple Probability

Faster Sampling

- MCMC sampling produces highly dependent samples (millions of trees have the same information as a few hundred from an independent sample)
- The posterior approximation is one key step to a new sampling paradigm.

Importance Sampling

1. Determine conditional clade distributions.
2. Sample a tree topology.

   2.1. Sample branch lengths to complete the tree.

   2.2. Find the sampling probability density of the tree.

   2.3. Find the (unnormalized) posterior density of the tree.

   2.4. Give the sampled tree weight (posterior/ sampling density.

Importance Sampling (continued)

3. Take weighted average of sampled trees for estimates (of clade probabilities, branch lengths, et. cetera).

Importance Sampling (continued)

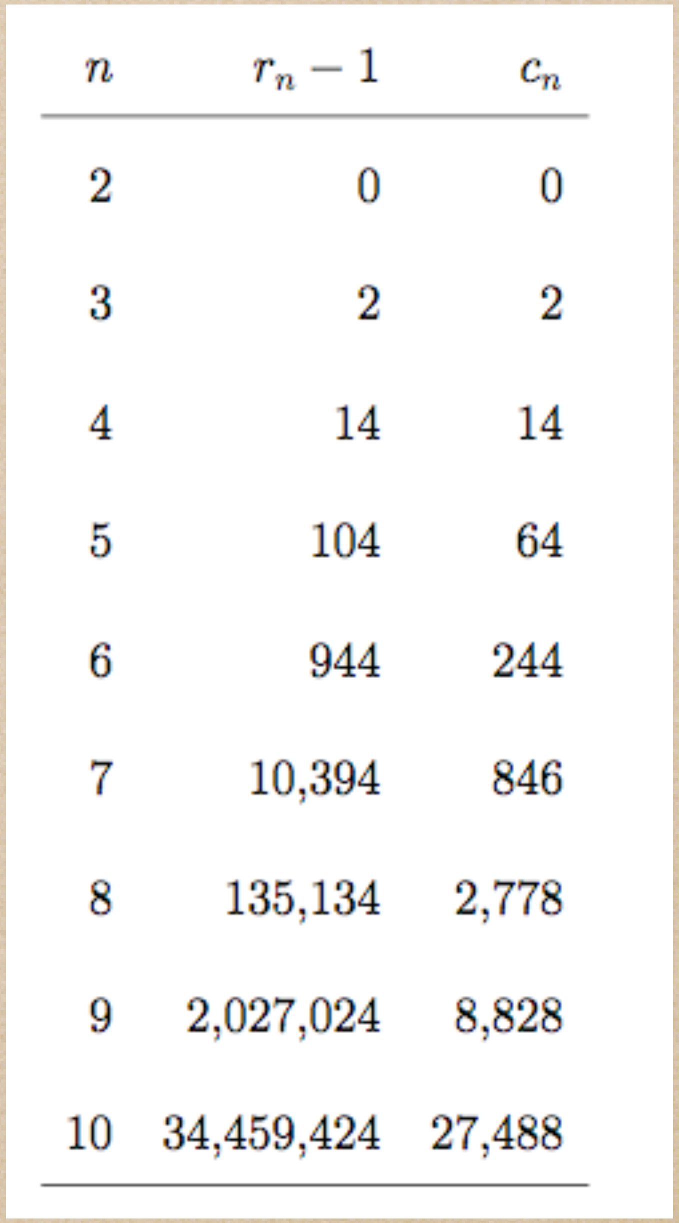
- If the incomplete steps can be finished and if the sampling distribution approximates well the actual posterior distribution, then independent sampling from a complicated high-dimensional posterior distribution will be possible.
- This may mean orders of magnitude less computational effort for Bayesian inference with large trees.

Approximation

- The true posterior distribution which sits in a huge parameter space is estimated to be close to one in a much smaller parameter space.

Number of Parameters

This table compares the number of free parameters for general tree distributions and distributions that satisfy conditional independence among separated subtrees.



n	$r_n - 1$	c_n
2	0	0
3	2	2
4	14	14
5	104	64
6	944	244
7	10,394	846
8	135,134	2,778
9	2,027,024	8,828
10	34,459,424	27,488

Summary

- The new method provides a means to estimate probabilities of rare trees much more accurately than simple sample averages.
- The new method is one key step in a new paradigm for computation in Bayesian phylogenetic inference.

Acknowledgments

- Thanks to Cécile Ané!
- Funding from NSF and NIH.