SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck

Palash Sashittal **Mohammed El-Kebir**

University of Illinois at Urbana-Champaign

RECOMB-CG 2019
Montepellier, France

Octotber 3^{rd}, 2019

Table of Contents

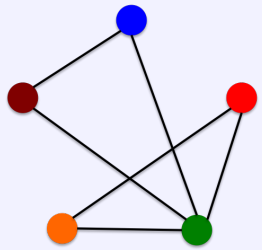
Background

Accurate inference of transmission network is pivotal for

- Real-time outbreak management
- Public health policies

Traditional approaches involved

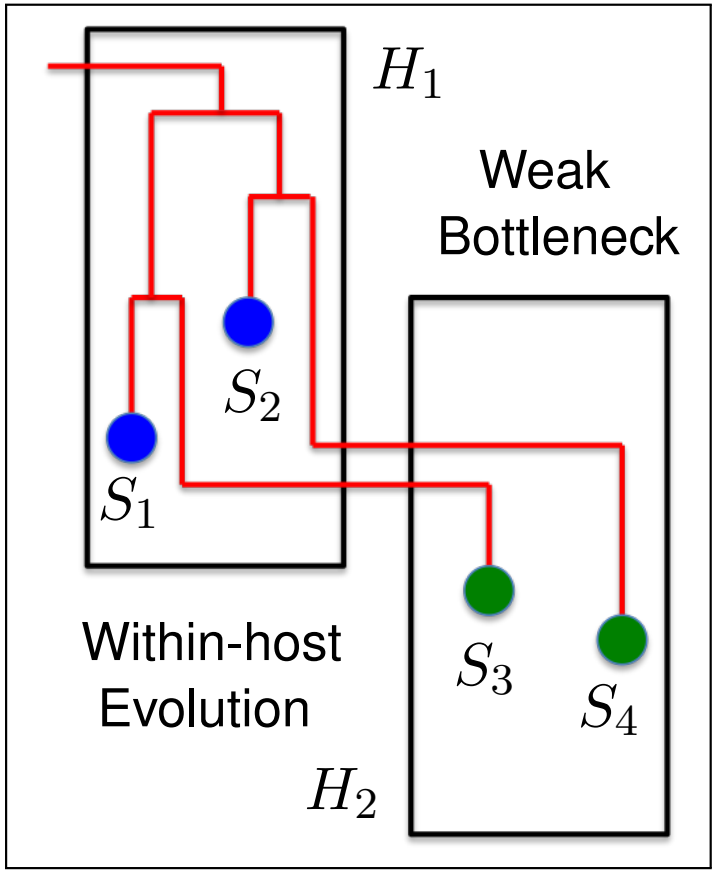
- fieldwork and interviews
- uncovering contact histories



With decreasing cost of genomic sequencing, molecular epidemiology has become indispensable in the analysis of disease outbreaks

Challenges:

- Pathogen evolutionary history does not necessarily match the transmission history of the outbreak[1]
- High mutation rates and long incubation times result in within-host diversity
- Further complication arises due to multi-strain infection or a weak transmission bottleneck



SharpTNI: infers parsimonious transmission networks under a weak bottleneck

[1] Ethan Romero-Severson et al. *Molecular biology and evolution* 31.9 (2014), pp. 2472–2482.

SharpTNI vs. Previous Work

Method	weak bottleneck	unsampled hosts	co-estimation of T and N	allows super-infection	allows co-transmissions
SharpTNI	✓	✓	✗	✓	✓
BadTrIP[2]	✓	✓	✗	✓	✓
SCOTTI[3]	partially	✓	✓	✓	✗
QUENTIN[4]	✗	✗	✗	✗	✗
Klinkenberg[5]	✗	✗	✓	✗	✗
Didelot[6]	✗	✓	✗	✗	✗
Hall[7]	✗	✗	✓	✗	✗
Didelot[8]	✗	✗	✗	✗	✗
Ypma[9]	✗	✗	✓	✗	✗

[2] Nicola De Maio et al. *PLoS computational biology* 14.4 (2018), e1006117.

[3] Nicola De Maio, Chieh-Hsi Wu, and Daniel J Wilson. *PLoS computational biology* 12.9 (2016), e1005130.

[4] Pavel Skums et al. *Bioinformatics* 34.1 (2017), pp. 163–170.

[5] Don Klinkenberg et al. *PLoS computational biology* 13.5 (2017), e1005495.

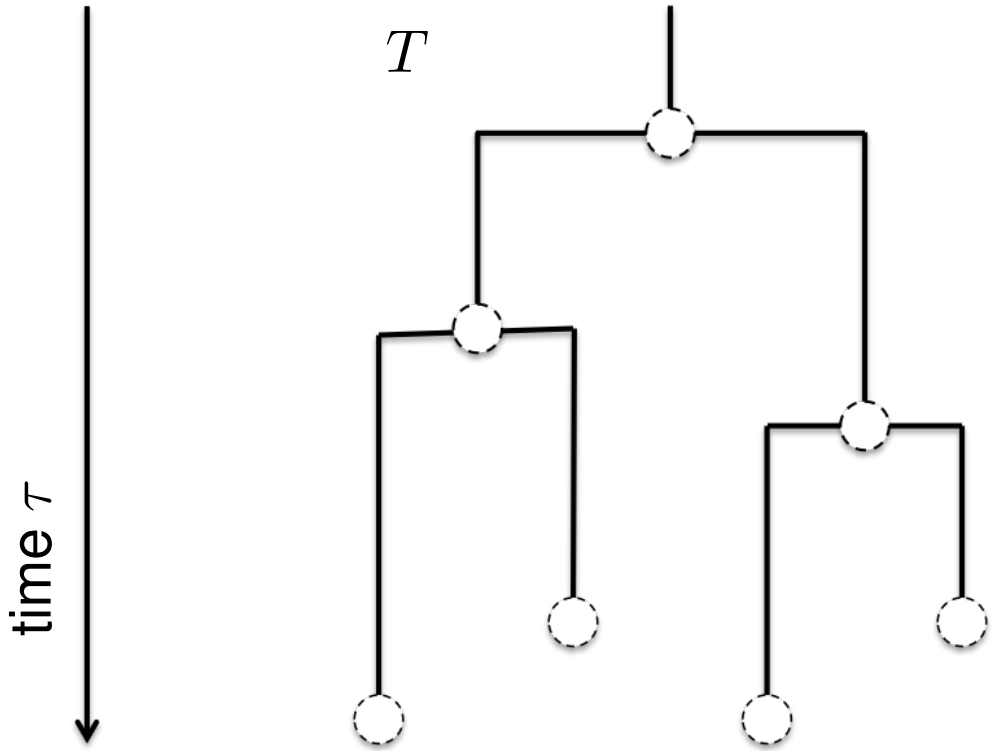
[6] Xavier Didelot et al. *Molecular biology and evolution* 34.4 (2017), pp. 997–1007.

[7] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. *PLoS computational biology* 11.12 (2015), e1004613.

[8] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. *Molecular biology and evolution* 31.7 (2014), pp. 1869–1879.

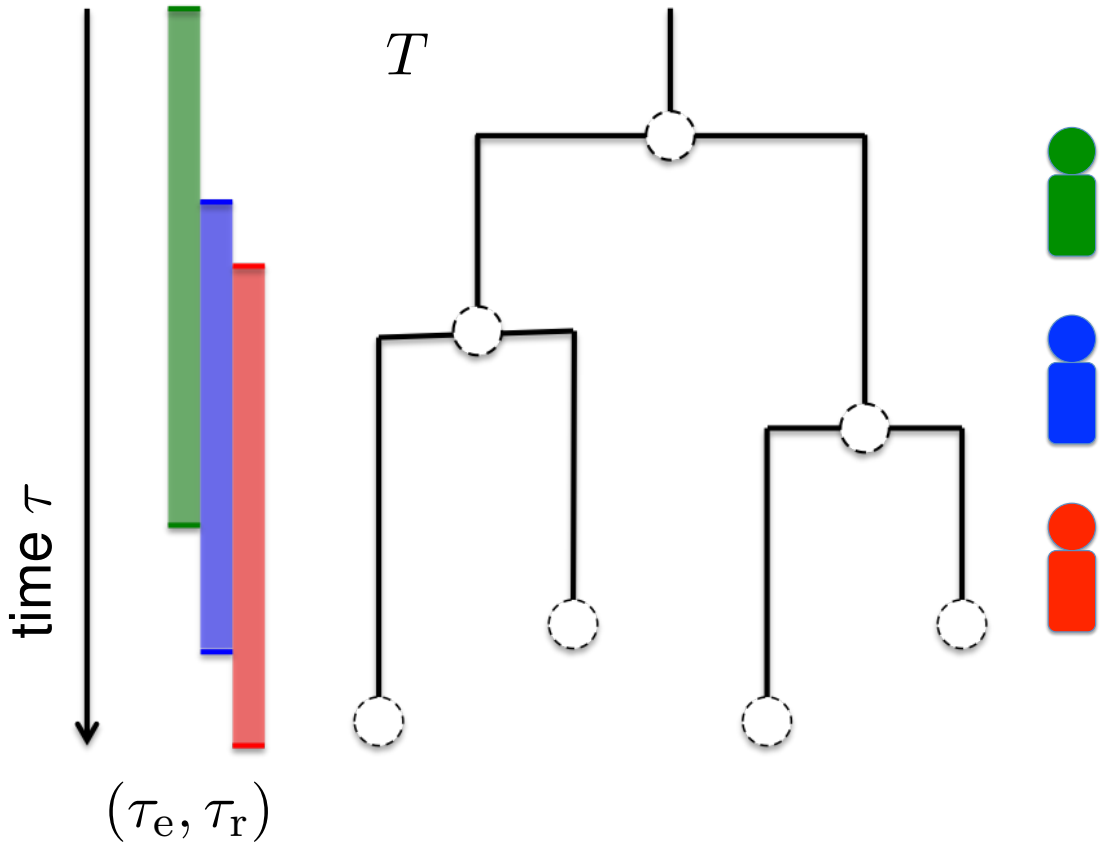
[9] Rolf JF Ypma, W Marijn van Ballegooijen, and Jacco Wallinga. *Genetics* 195.3 (2013), pp. 1055–1062.

TNI problem: Input



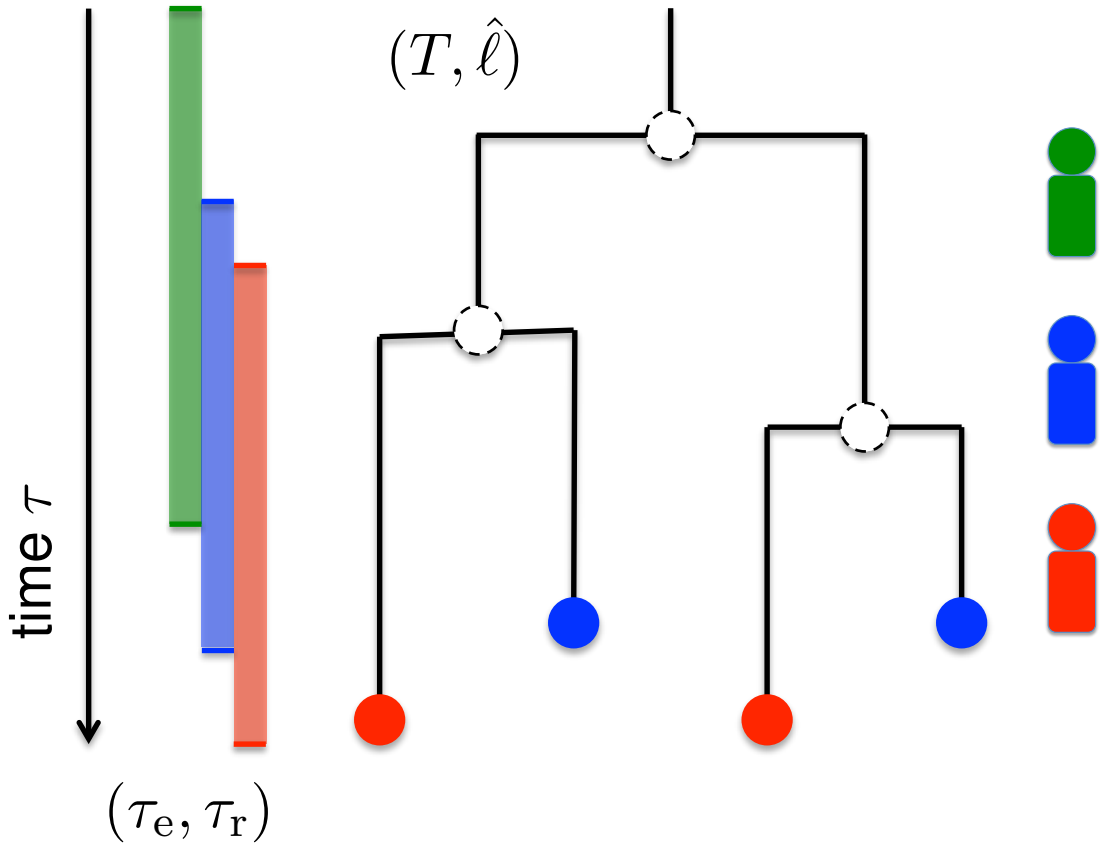
A timed phylogeny is a rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \to \mathbb{R}^{\geq 0}$ such that $\tau(u) < \tau(v)$ for all pairs u, v of vertices where $u \preceq_T v$.

TNI problem: Input



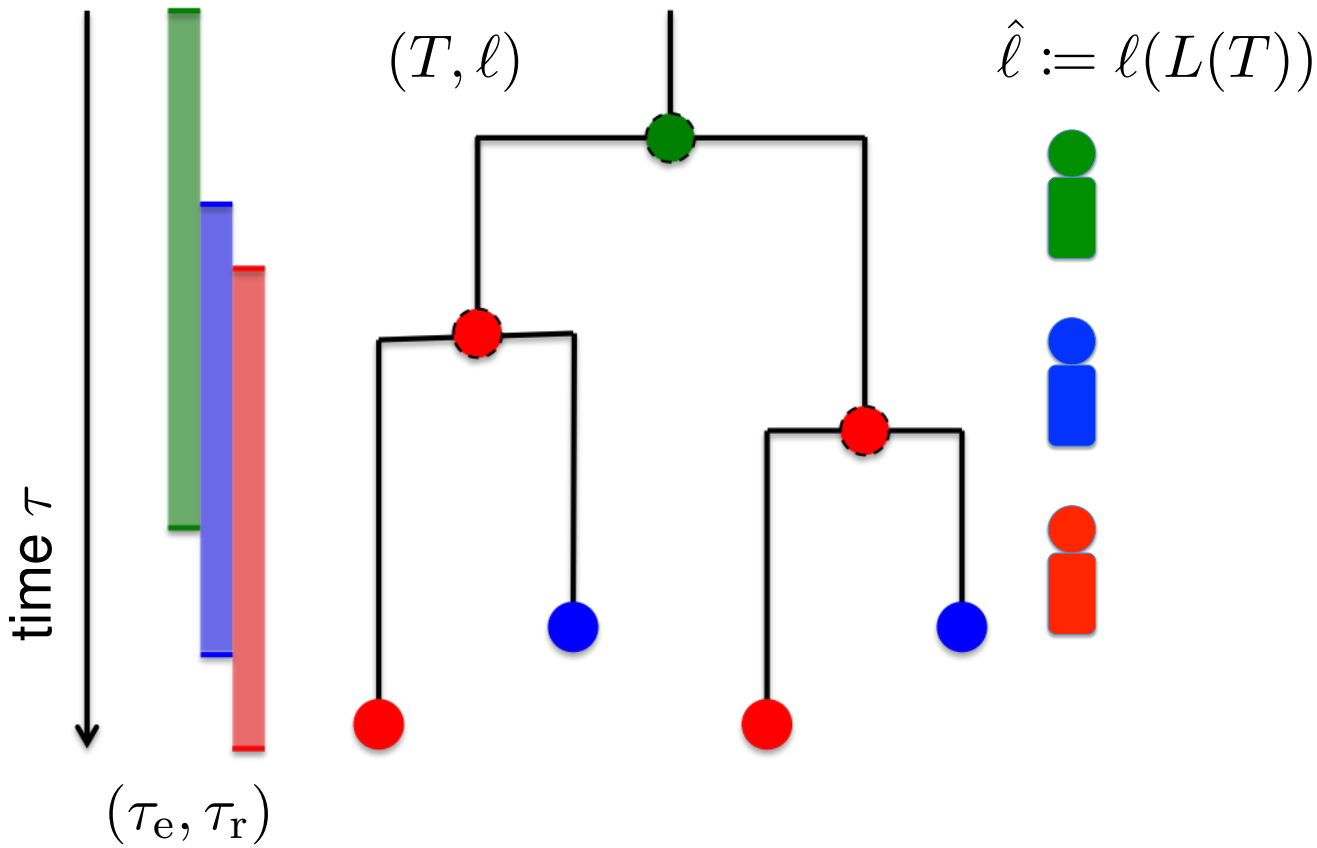
For each host $s \in \Sigma$, based on epidemiological data, we have an entrance time $\tau_{\mathrm{e}}(s)$ and the host is removed from the population at time $\tau_{\mathrm{r}}(s)$.

TNI problem: Input



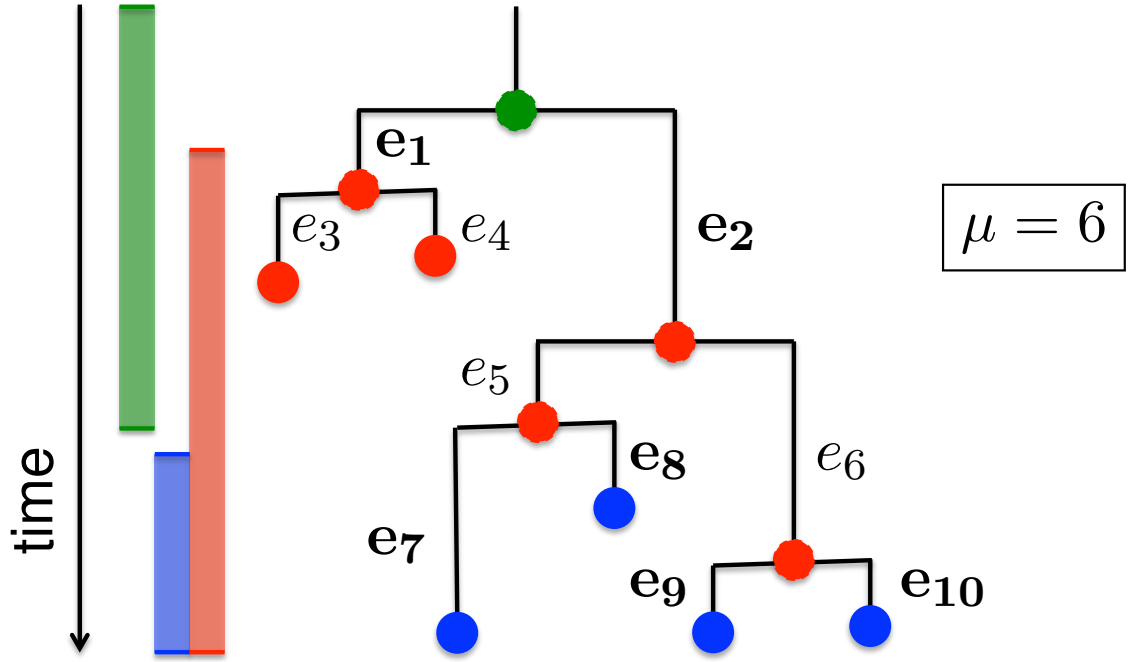
A leaf labeling of a timed phylogeny T is a function $\hat{\ell} : L(T) \to \Sigma$, assigning a host $\hat{\ell}(u)$ to each leaf u of T.

TNI problem: Output



A host labeling of a timed phylogeny T is a function $\ell : V(T) \to \Sigma$, assigning a host $\ell(u)$ to each vertex u of T.

TNI objective function

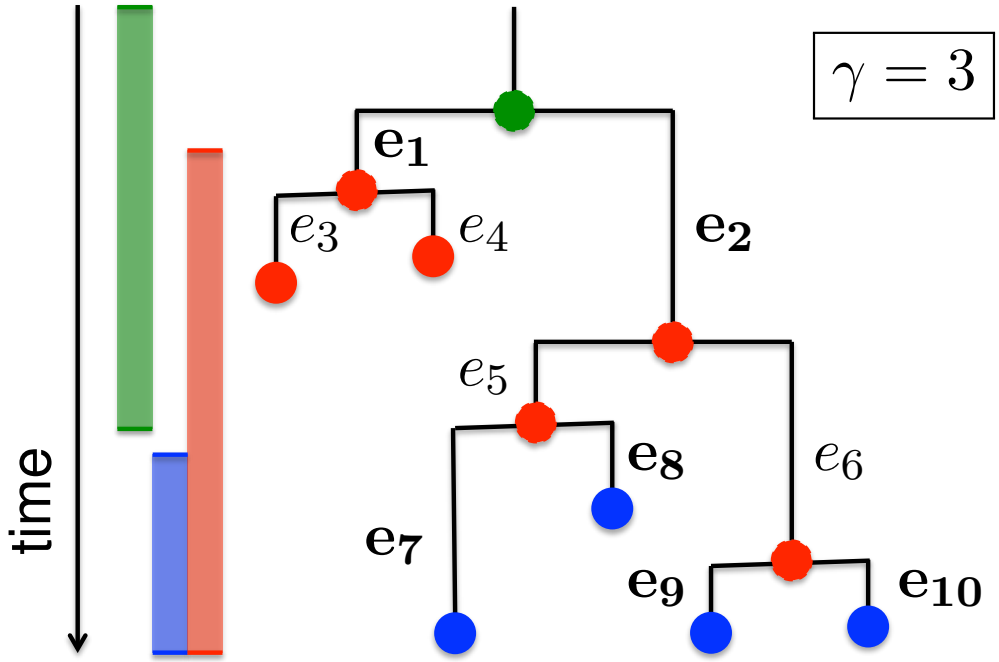


Transmission Edge

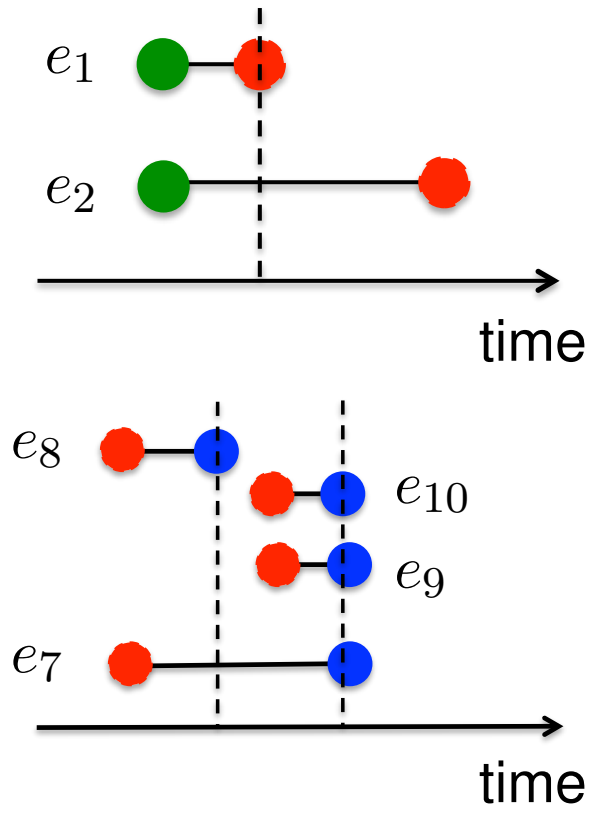
Given a timed phylogeny T and host labeling ℓ, an edge (u, v) of T is a transmission edge if $\ell(u) \neq \ell(v)$.

TNI objective function

$\Psi_1 = \{e_1, e_2\}$, $\Psi_2 = \{e_7, e_8\}$, $\Psi_3 = \{e_9, e_{10}\}$



$\gamma = 3$

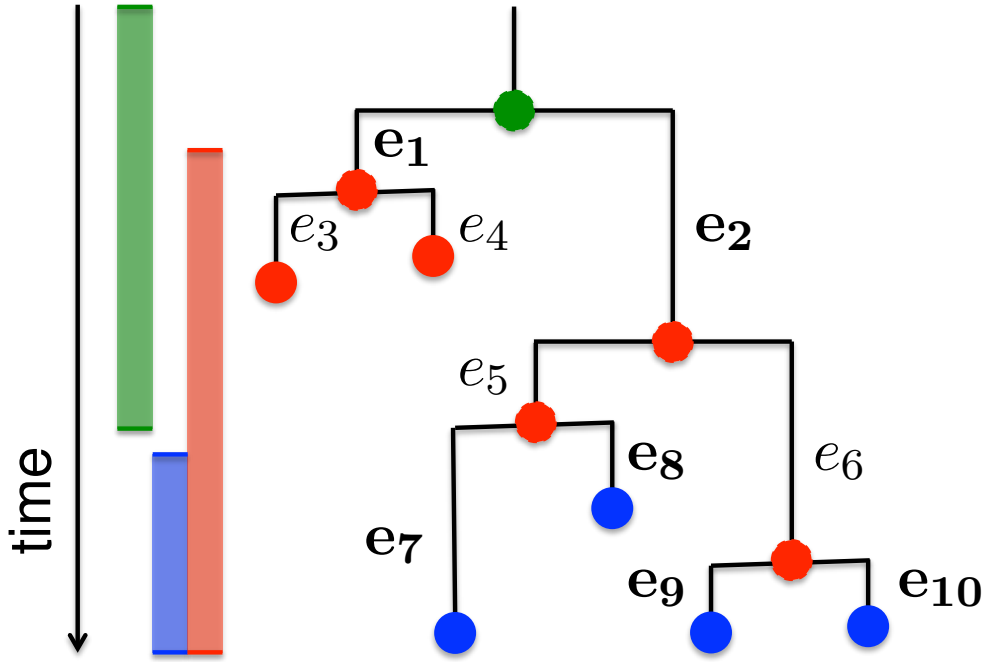


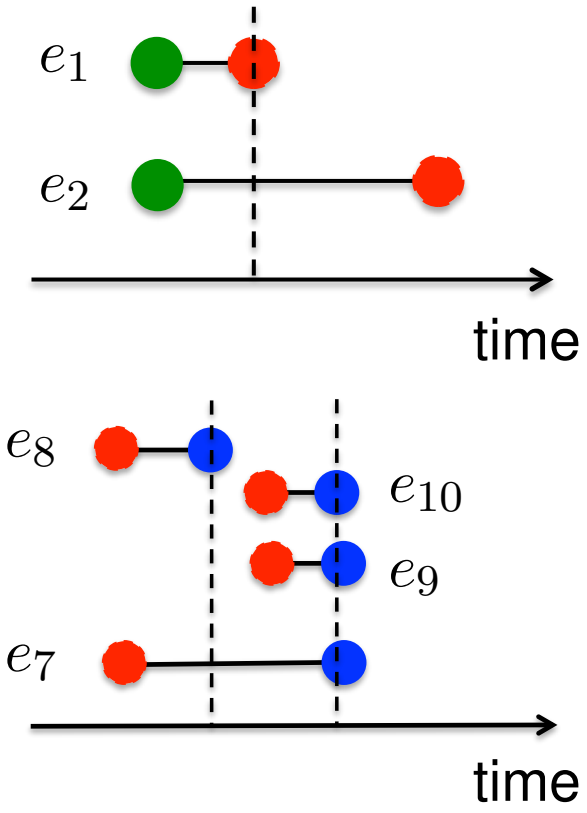
Transmission Event

Two transmission edges (u, v) and (u', v') can belong to the same event if (i) $\ell(u) = \ell(u'), \ell(v) = \ell(v')$ and (ii) $[\tau(u), \tau(v)] \cap [\tau(u'), \tau(v')] \neq \emptyset$.

TNI objective function

$$N = \{\{e_1, e_2\}, \{e_7, e_8\}, \{e_9, e_{10}\}\}$$



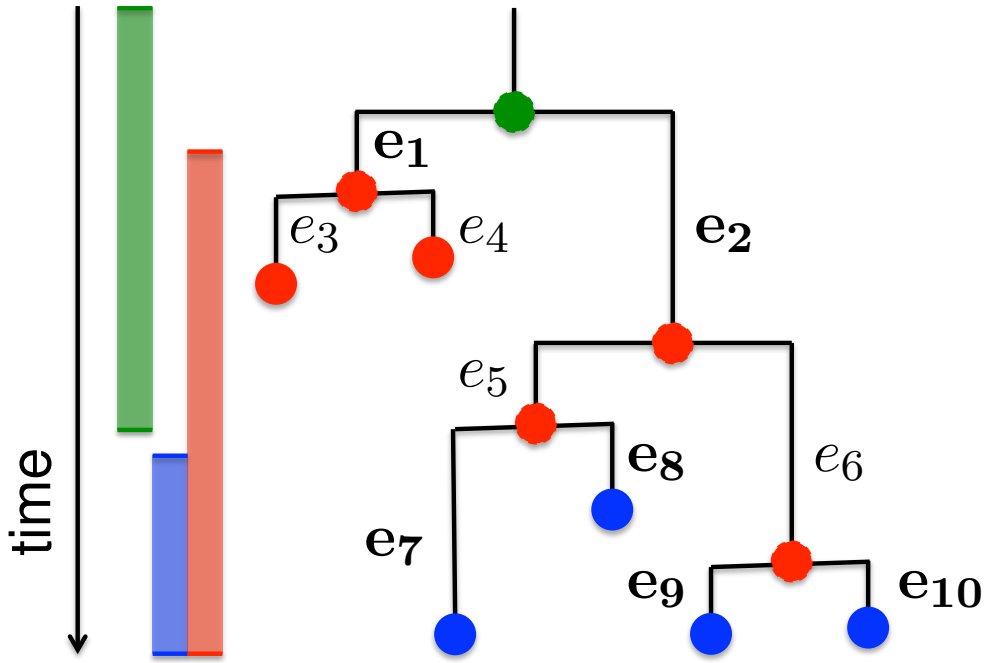


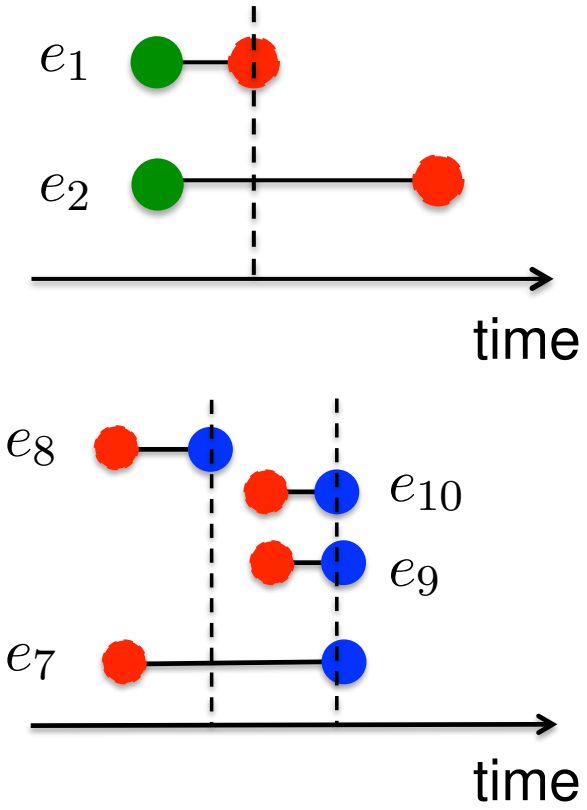
Transmission Network

A transmission network N is a partition of the transmission edges of (T, ℓ) into disjoint transmission events.

TNI objective function

$$N = \{\{e_1, e_2\}, \{e_7, e_8\}, \{e_9, e_{10}\}\}$$

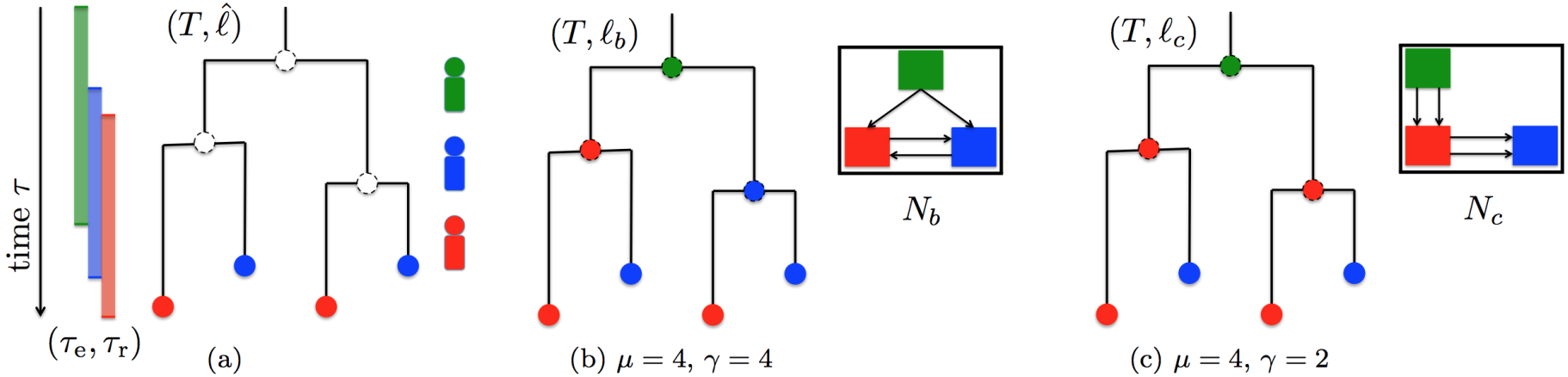




Objective Function

First minimize $\mu(N)$ and then minimize $\gamma(N)$

SharpTNI: Parsimonious Transmission Networks under Weak Bottleneck

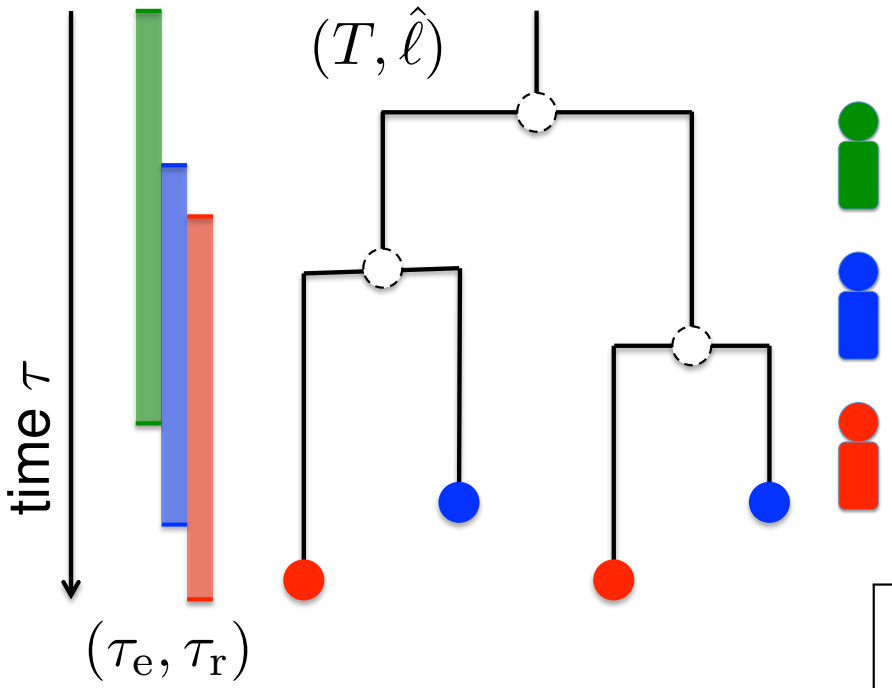


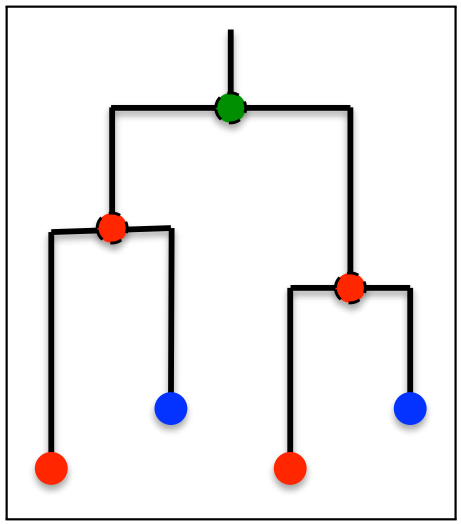
TNI Problem Statement

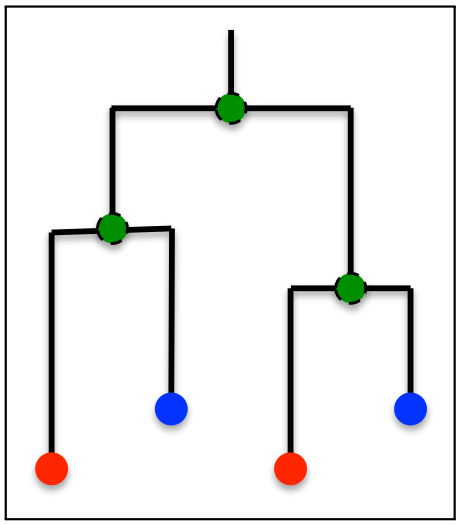
Given a timed phylogeny T with time-stamps τ, host-leaf labeling $\hat{\ell}$, entrance times τ_e and removal times τ_r, find a transmission network N and corresponding host labeling ℓ with minimum transmission number $\mu(N) = \mu^*$ and subsequently smallest co-transmission number $\gamma(N) = \gamma^*$ such that $\tau(u) \in [\tau_e(s), \tau_r(s)]$ for all hosts s and vertices u where $\ell(u) = s$.

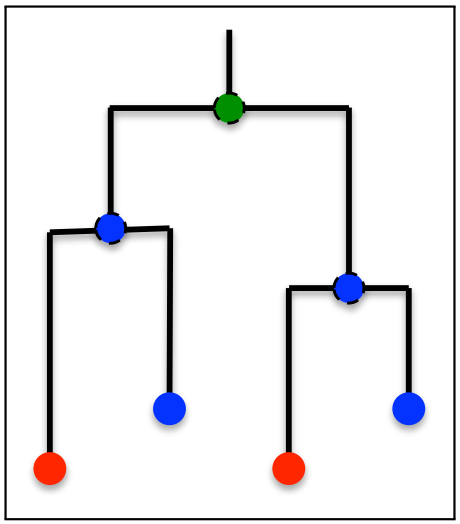
$$(T, \tau, \hat{\ell}, \tau_e, \tau_r) \rightarrow (\ell, \mu^*, \gamma^*, N)$$

Non-uniqueness of solutions



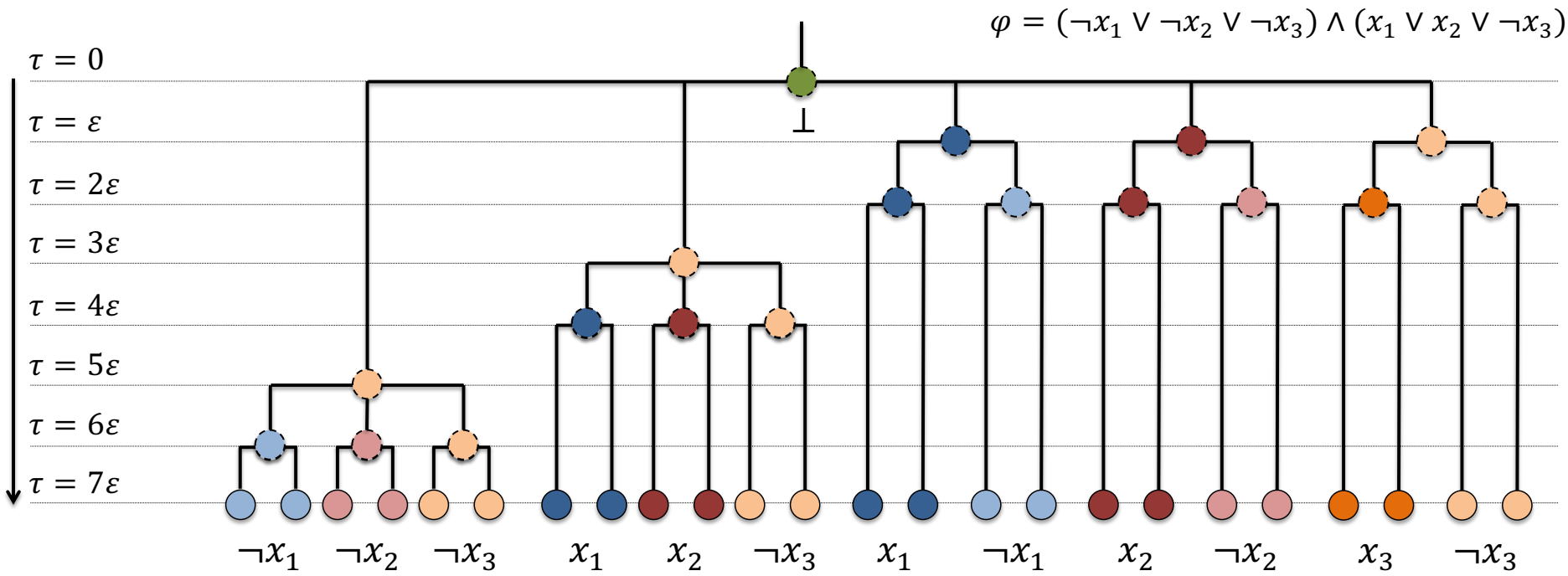






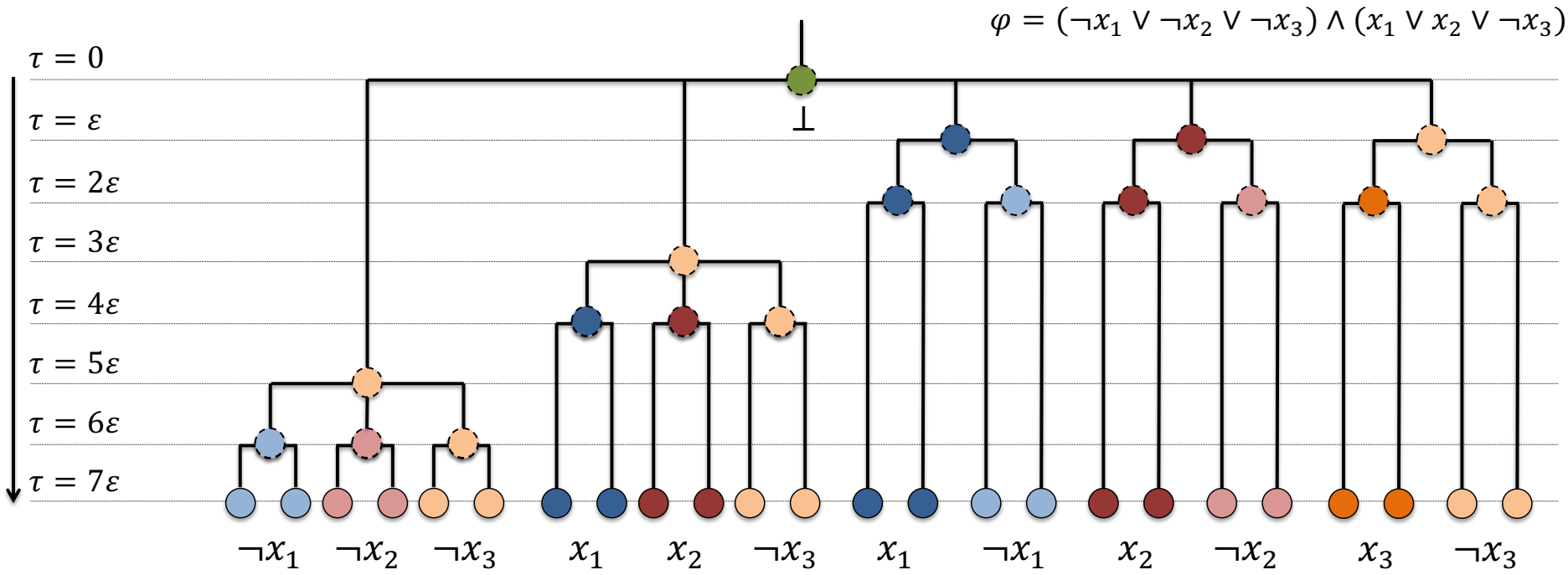
We have 3 possible solutions with $\mu^* = 4$ and $\gamma^* = 2$

Complexity



TNI is NP-hard (by reduction from 3-SAT)

Complexity



TNI is NP-hard (by reduction from 3-SAT)

There exists no FPAUS to sample solutions of TNI unless RP=NP

Methods Outline

- Polynomial time algorithm for a constrained TNI problem
- Relaxation of TNI
- Solving TNI via SAT

TNI problem constrained to a given host labeling

Problem Statement

Given a timed phylogeny T with time-stamps τ and host labeling $\ell : V(T) \to \Sigma$, find a transmission network N consistent with (T, ℓ) with minimum co-transmission number $\gamma(N)$.

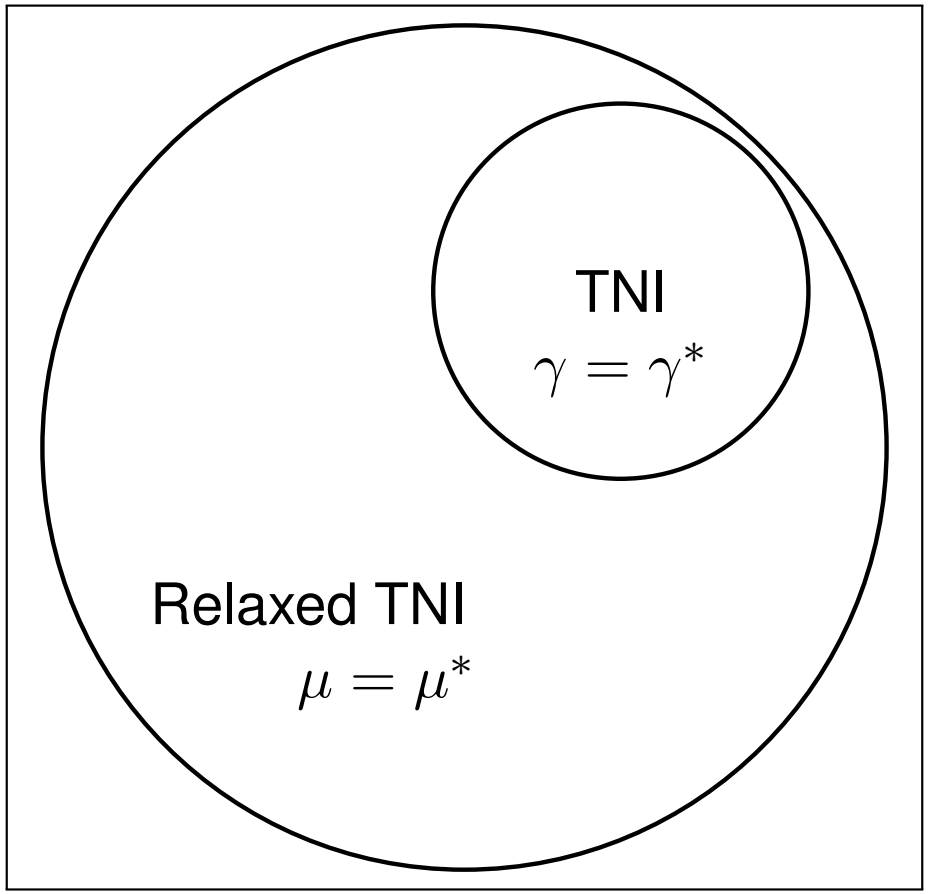
$$(T, \ell, \tau_{\mathrm{e}}, \tau_{\mathrm{r}}) \to (\gamma^*, N)$$

- This problem is equivalent to partitioning the vertices of an interval graph with the smallest number of cliques,
- Can be solved in polynomial time by a simple greedy algorithm[10]

[10] Gerd Finke et al. *Discrete Applied Mathematics* 156.5 (2008), pp. 556–568.

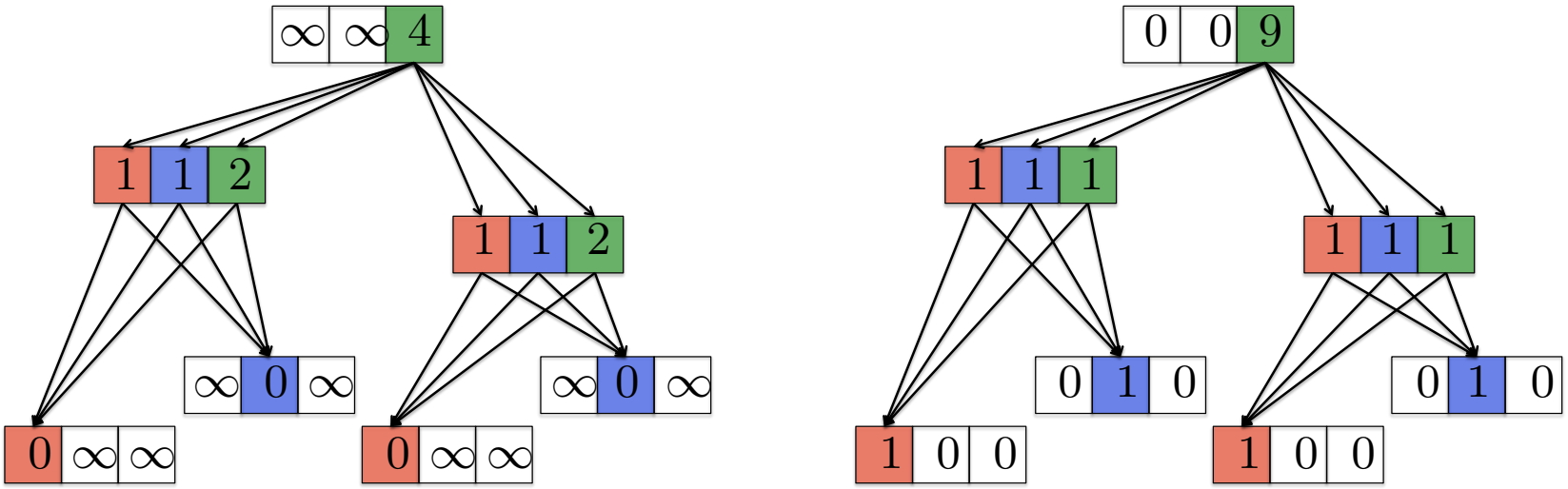
Relaxation of TNI

Consider host labelings ℓ that admit transmission network N with minimum transmission number $\mu(N) = \mu^*$ and any co-transmission number $\gamma(N)$



Relaxation of TNI

Consider host labelings ℓ that admit transmission network N with minimum transmission number $\mu(N) = \mu^*$ and any co-transmission number $\gamma(N)$



(i) minimum transmission number

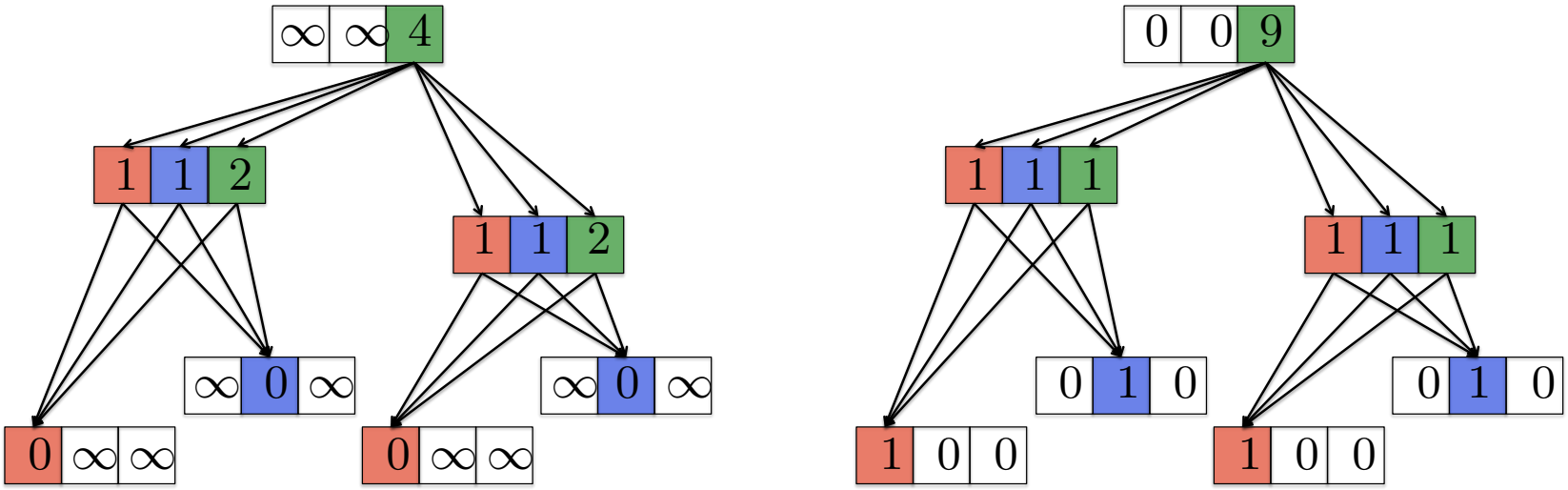
(j) minimum transmission labelings

Figure: **Dynamic programming for a relaxation of TNI.**

Can be solved in polynomial time using dynamic programming (equivalent to Sankoff solutions with some constraints)

Relaxation of TNI

Consider host labelings ℓ that admit transmission network N with minimum transmission number $\mu(N) = \mu^*$ and any co-transmission number $\gamma(N)$



(a) minimum transmission number

(b) minimum transmission labelings

Figure: **Dynamic programming for a relaxation of TNI.**

Can be solved in polynomial time using dynamic programming

Can be used for a naive rejection sampling algorithm

SAT formulation

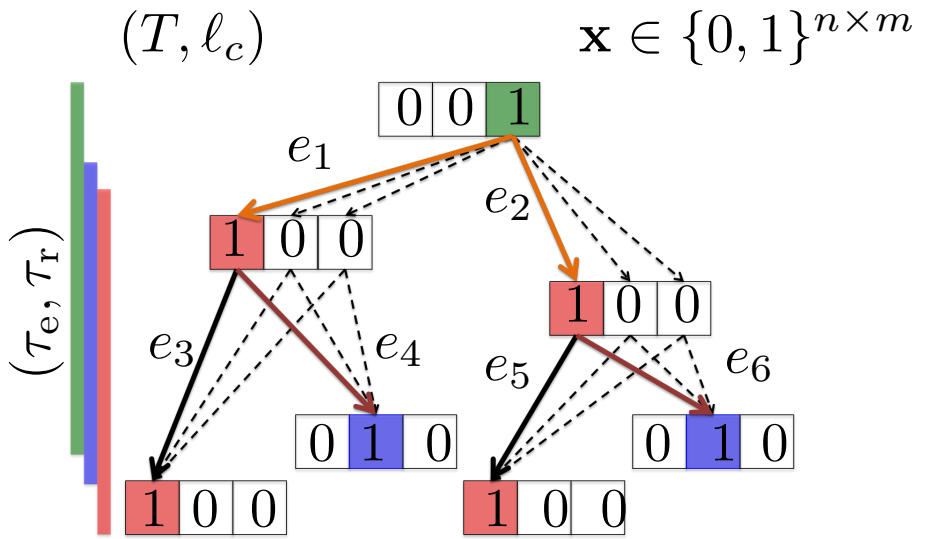
$\mathbf{x} \in \{0,1\}^{n \times m}$ encode a host labeling.

$$x_{i,s} = \begin{cases} 1, & \ell(v_i) = s, \\ 0, & \text{otherwise.} \end{cases}$$

$\mathbf{y} \in \{0,1\}^{(n-1)\times\alpha}$ encode the partition such that

$$y_{ij,p} = \begin{cases} 1, & \ell(v_i) \neq \ell(v_j), e_{ij} \in \Psi_p, \\ 0, & \text{otherwise.} \end{cases}$$

with $|V(T)| = n$, $|\Sigma| = m$ & $|N| = \alpha$



$N_c = \{\Psi_1, \Psi_2\}$ $\quad$ $\mathbf{y} \in \{0,1\}^{(n-1)\times\alpha}$

	e_1	e_2	e_3	e_4	e_5	e_6
Ψ_1	1	1	0	0	0	0
Ψ_2	0	0	0	1	0	1

Perform a sweep over $\alpha \in [m-1, \mu^*)$ for minimum co-transmission number

SAT formulation

$\mathbf{x} \in \{0,1\}^{n \times m}$ encode a host labeling.

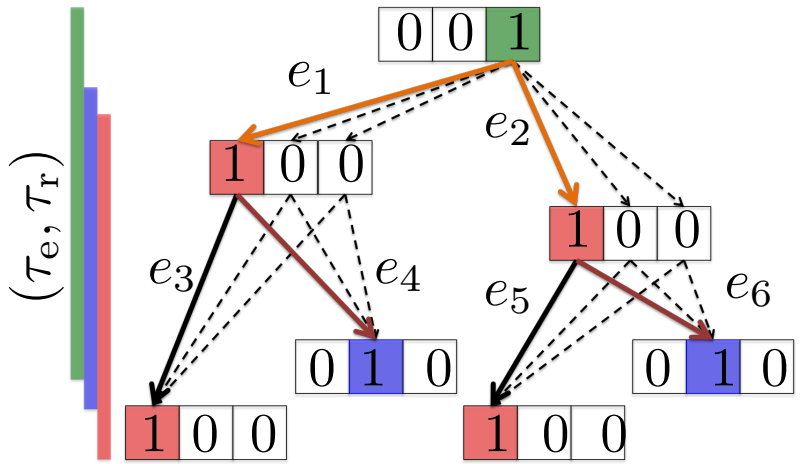
$$x_{i,s} = \begin{cases} 1, & \ell(v_i) = s, \\ 0, & \text{otherwise.} \end{cases}$$

$\mathbf{y} \in \{0,1\}^{(n-1)\times\alpha}$ encode the partition such that

$$y_{ij,p} = \begin{cases} 1, & \ell(v_i) \neq \ell(v_j), e_{ij} \in \Psi_p, \\ 0, & \text{otherwise.} \end{cases}$$

with $|V(T)| = n$, $|\Sigma| = m$ & $|N| = \alpha$

(T, ℓ_c) $\quad$ $\mathbf{x} \in \{0,1\}^{n \times m}$



$N_c = \{\Psi_1, \Psi_2\}$ $\quad$ $\mathbf{y} \in \{0,1\}^{(n-1)\times\alpha}$

	e_1	e_2	e_3	e_4	e_5	e_6
Ψ_1	1	1	0	0	0	0
Ψ_2	0	0	0	1	0	1

Perform a sweep over $\alpha \in [m-1, \mu^*)$ for minimum co-transmission number

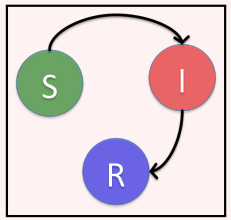
Results Outline

- Simulations of outbreaks with varying bottleneck and population sizes
- 2014 Ebola outbreak of Sierra Leone

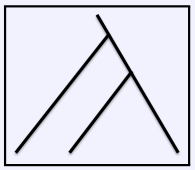
Outbreak Simulation

Epidemiological Model

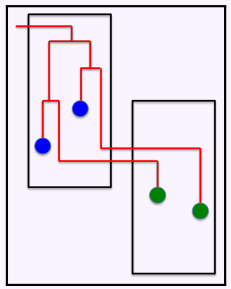
We use the compartmental SIR (Susceptible-Infectious-Recovered) model[11] with each infection event comprising of at most κ co-transmission strains



Evolution of pathogen modeled using coalescence model with constant population size and evolution rate



We perform complete sampling of the outbreak so that there is no incomplete lineage sorting. All pathogen strains transmitted from the donor to recipient are sampled from the donor.

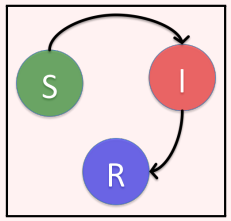


[11] Allen, L.J.: An introduction to stochastic epidemic models. In: Mathematical Epidemiology, pp. 81–130. Springer, (2008)

[12] Kingman, J.: b the coalescent. stoch. In: Proc. Appl, vol. 13, pp. 235–248 (1982)

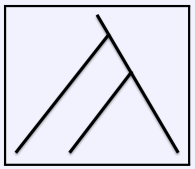
Outbreak Simulation

We use the compartmental SIR (Susceptible-Infectious-Recovered) model[11] with each infection event comprising of at most κ co-transmission strains

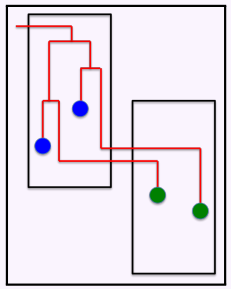


Within-Host Diversity

Evolution of pathogen modeled using coalescence model[12] with constant population size and evolution rate



We perform complete sampling of the outbreak so that there is no incomplete lineage sorting. All pathogen strains transmitted from the donor to recipient are sampled from the donor.

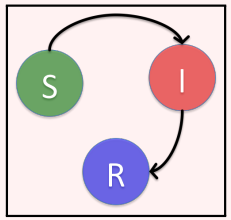


[11] Allen, L.J.: An introduction to stochastic epidemic models. In: Mathematical Epidemiology, pp. 81–130. Springer, (2008)

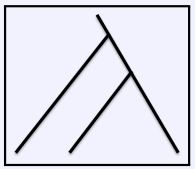
[12] Kingman, J.: b the coalescent. stoch. In: Proc. Appl, vol. 13, pp. 235–248 (1982)

Outbreak Simulation

We use the compartmental SIR (Susceptible-Infectious-Recovered) model[11] with each infection event comprising of at most κ co-transmission strains

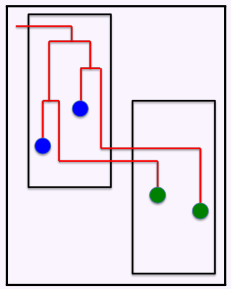


Evolution of pathogen modeled using coalescence model[12] with constant population size and evolution rate



Multiple Host Samples

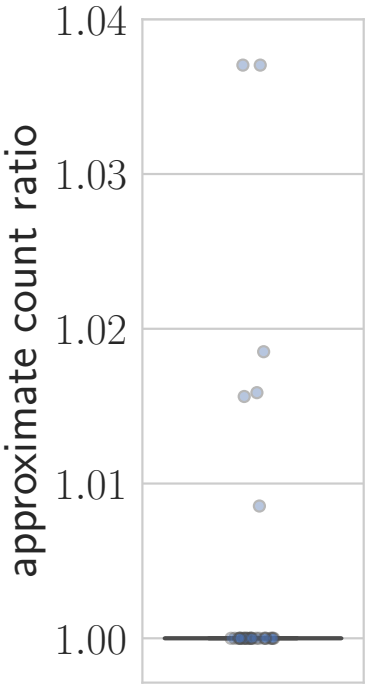
We perform complete sampling of the outbreak so that there is no incomplete lineage sorting. All pathogen strains transmitted from the donor to recipient are sampled from the donor.



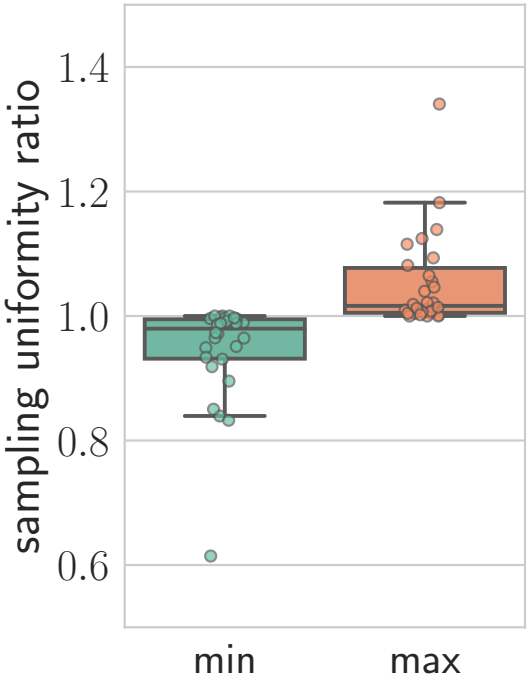
[11] Allen, L.J.: An introduction to stochastic epidemic models. In: Mathematical Epidemiology, pp. 81–130. Springer, (2008)

[12] Kingman, J.: b the coalescent. stoch. In: Proc. Appl, vol. 13, pp. 235–248 (1982)

Simulation Results



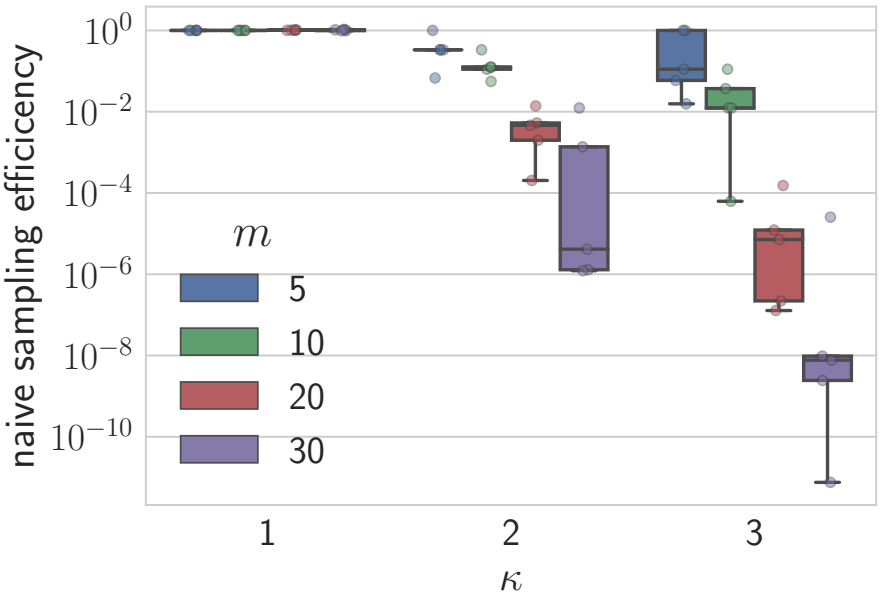
(a) count ratio



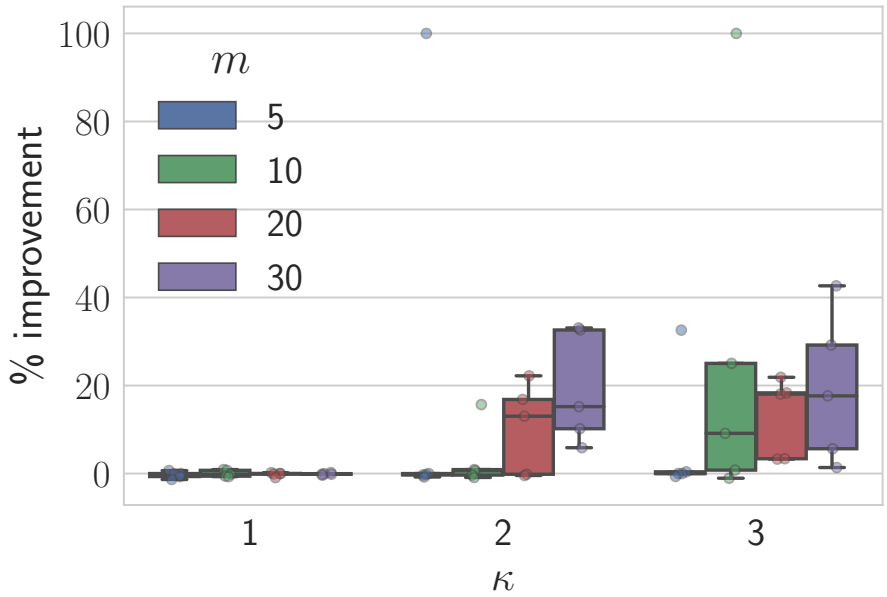
(b) sampling ratio

Simulations show that SharpTNI accurately counts and samples parsimonious transmission networks.

Simulation Results



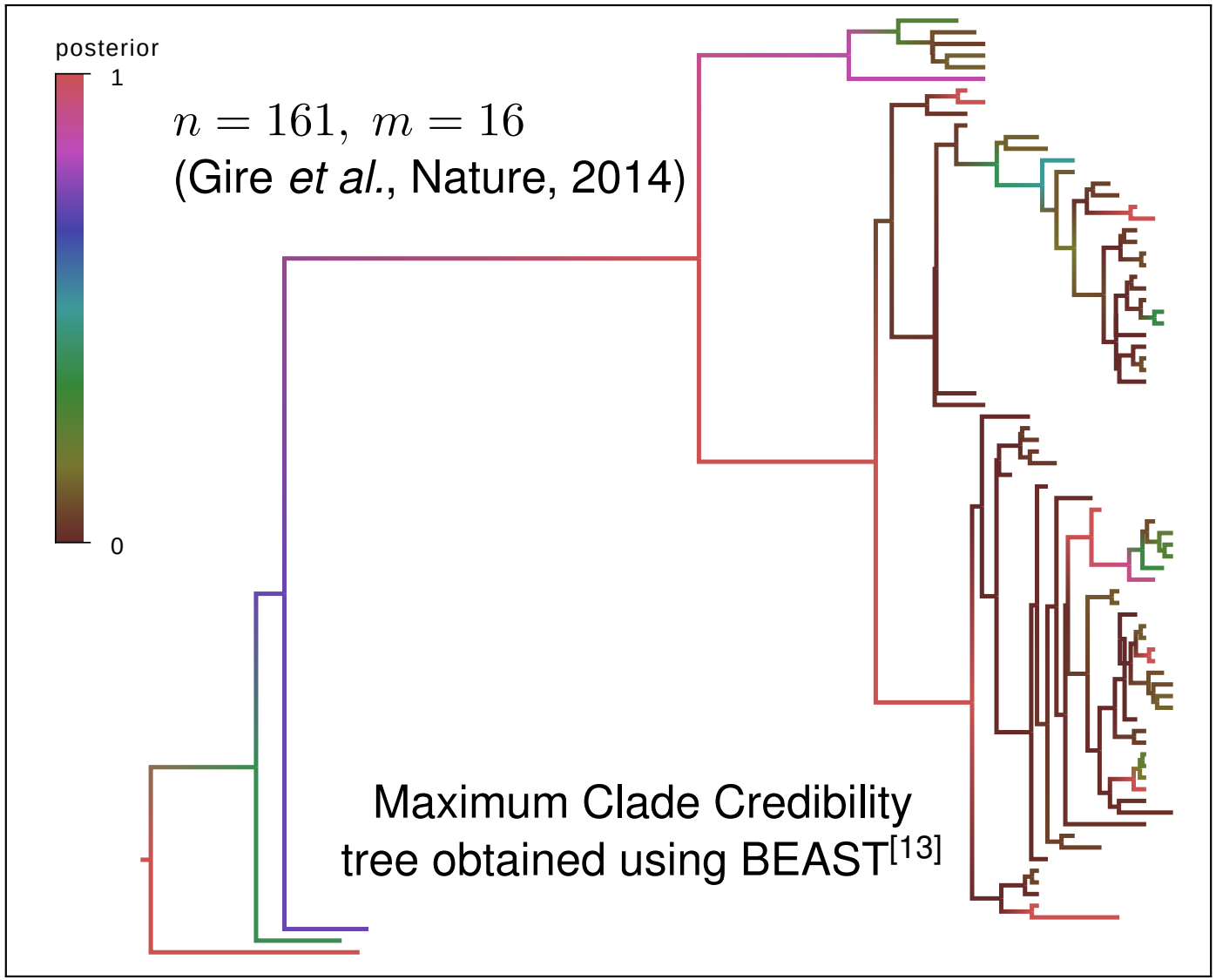
(c) rejection sampling efficiency



(d) SharpTNI performance improvement

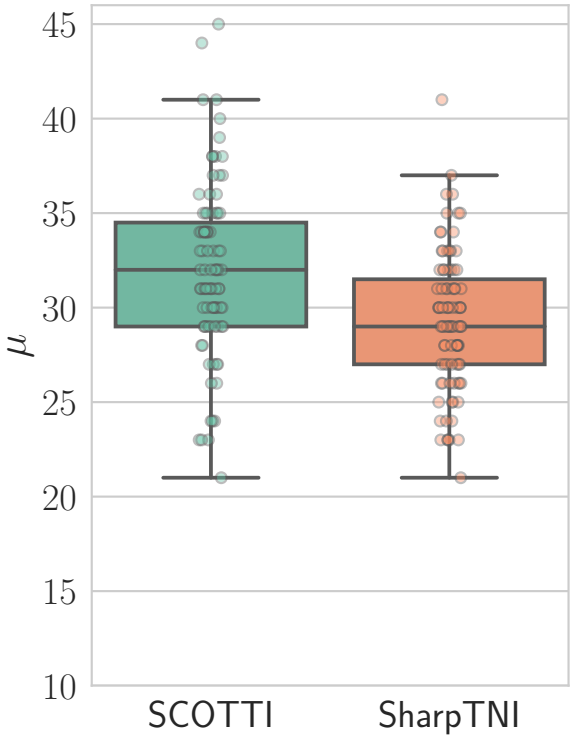
Simulations show that naive rejection sampling is not a practical approach for uniformly sampling parsimonious transmission networks.

Ebola 2014 Outbreak of Sierra Leone

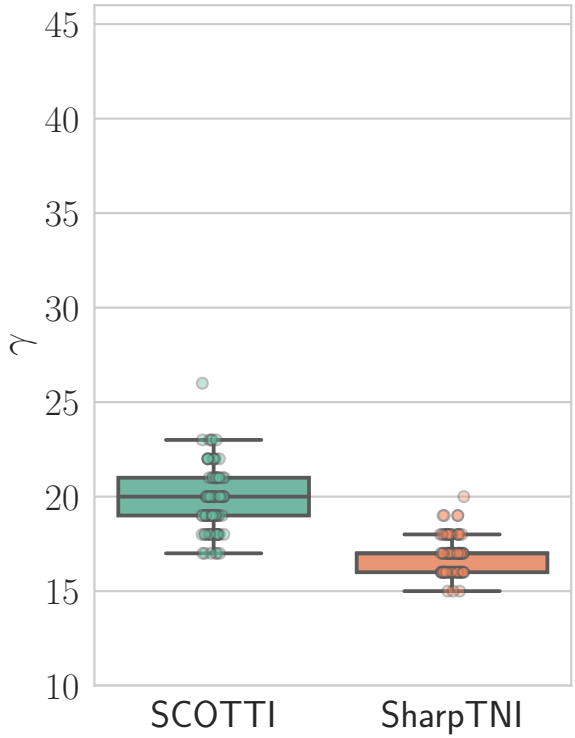


[13] Alexei J Drummond and Andrew Rambaut. *BMC Evolutionary Biology* 7.1 (2007), p. 214.

SharpTNI vs. SCOTTI



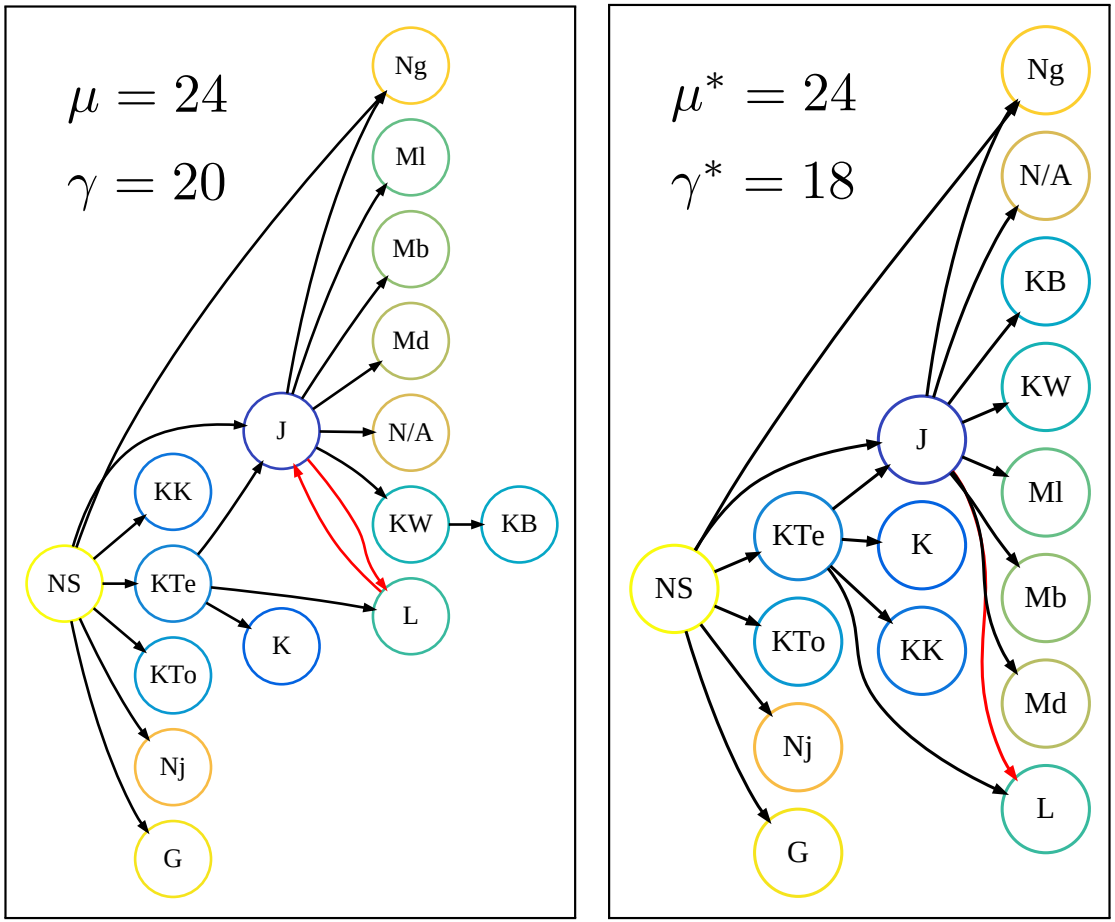
(e) transmission number



(f) co-transmission number

Transmission networks inferred by SharpTNI are more parsimonious compared to the transmission networks inferred by SCOTTI under a weak transmission bottleneck

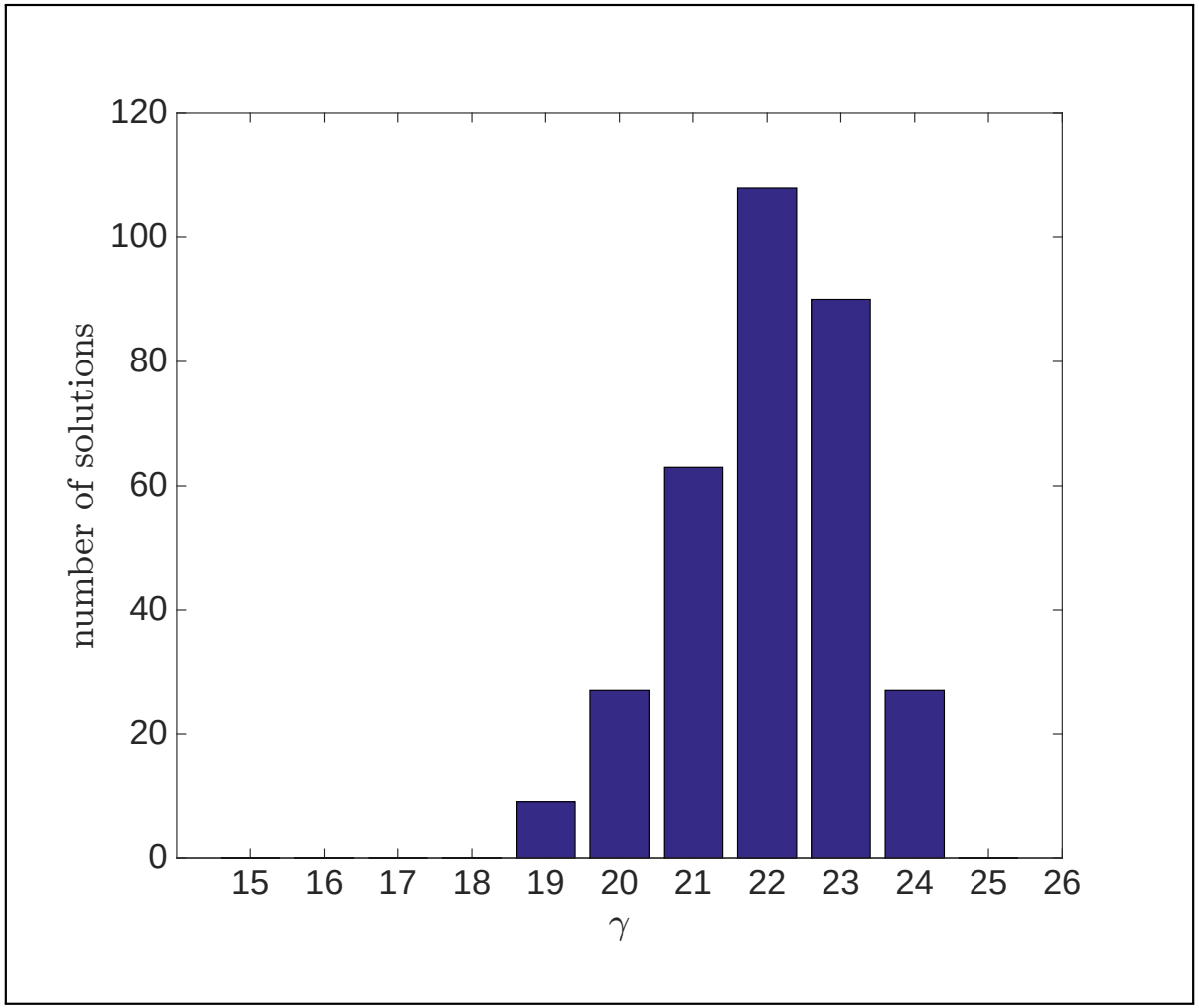
SharpTNI vs. SCOTTI



(g) SCOTTI (h) SharpTNI

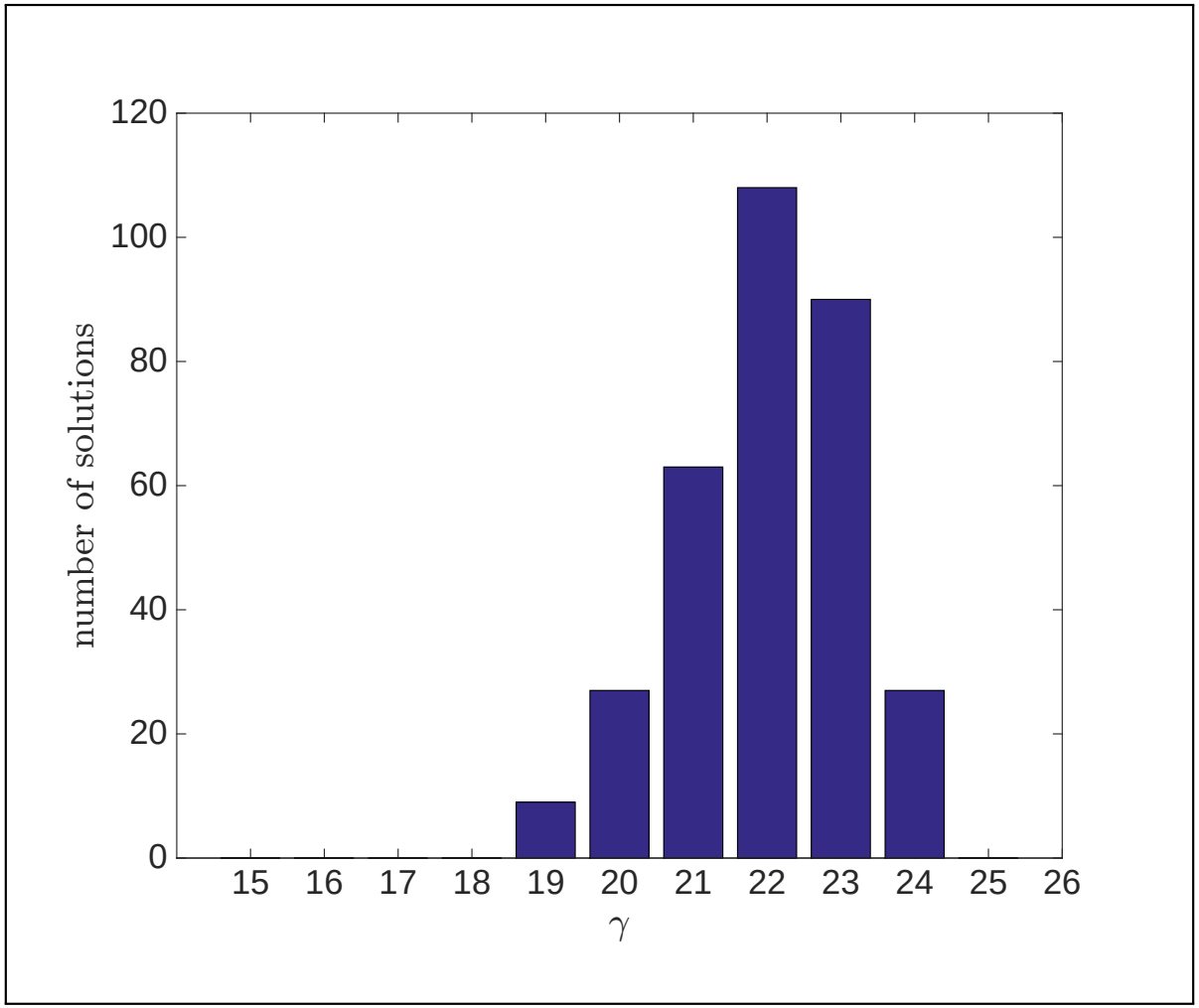
Influence of co-transmission number minimization on the network solution

(BEAST + SharpTNI) to re-analyse the Ebola outbreak



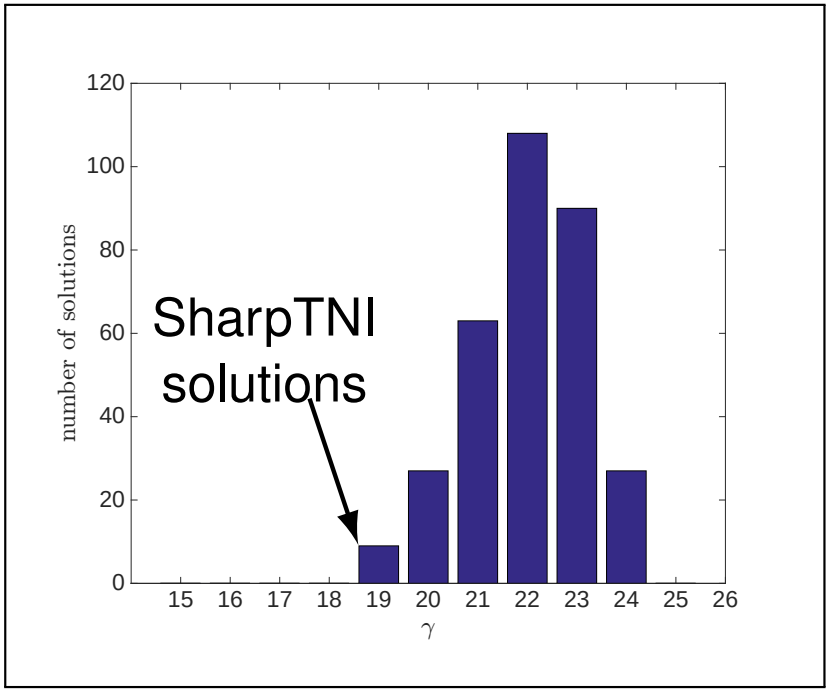
Distribution of the co-transmission number γ for all 324 minimum transmission ($\mu^* = 26$) host labelings of the BEAST MCC tree.

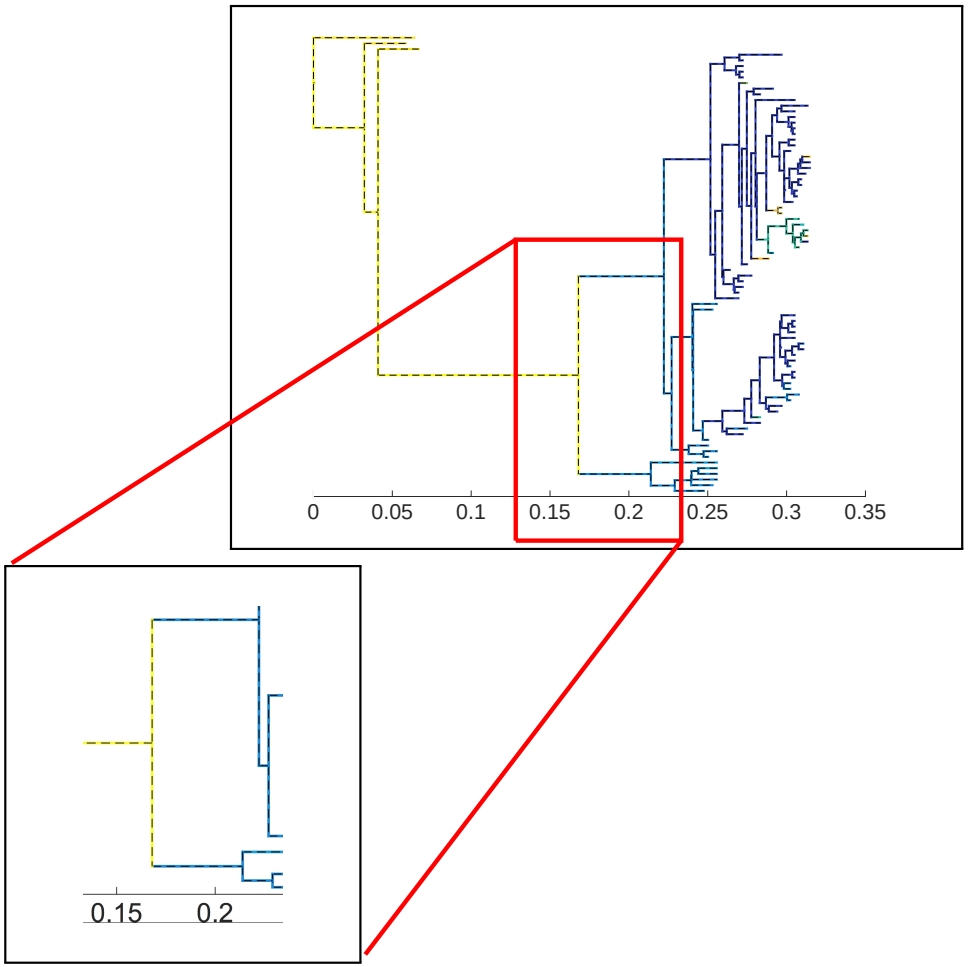
(BEAST + SharpTNI) to re-analyse the Ebola outbreak



216 solutions describe a migration of a single strain from Guinea to Kissi Teng which conflicts with fieldwork from Gire *et al.* 2014 paper.

(BEAST + SharpTNI) to re-analyse the Ebola outbreak

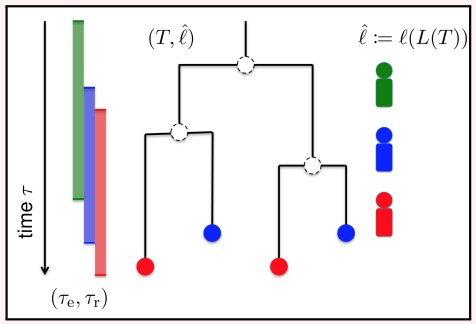




All 9 SharpTNI solutions contain co-transmission of two strains from Guinea to Kissi Teng which is corroborated by Gire *et al.* 2014 paper.

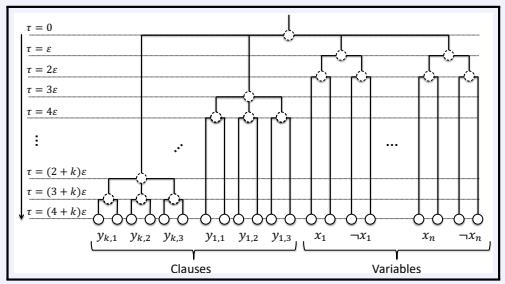
Discussion

We introduce the Transmission Network Inference (TNI) problem for estimating a parsimonious transmission network under a weak transmission bottleneck given a timed phylogeny.

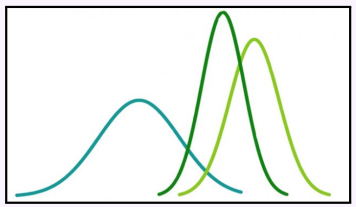


For optimization and sampling version of TNI,

- we establish hardness
- broader application of SAT
- novel method of counting/sampling



In the future, we plan to extend current framework to co-estimation of the timed phylogeny and the transmission network using MLE version of TNI.



Acknowledgements

- National Science Foundation, CCF 18-50502
- Experiments were NCSA's Blue Waters supercomputer

Questions?