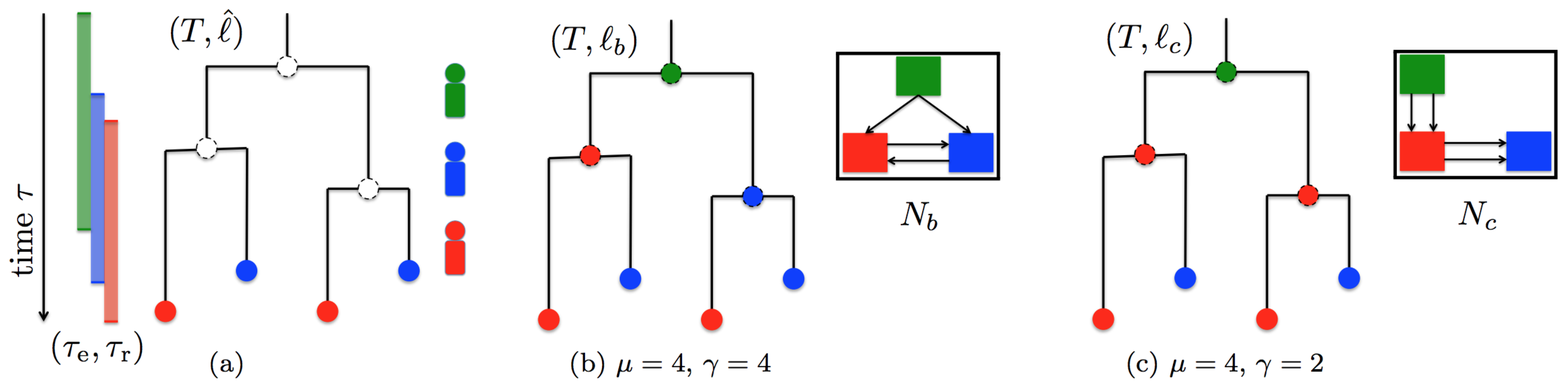
SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck

Palash Sashittal[1] and **Mohammed El-Kebir**[2]

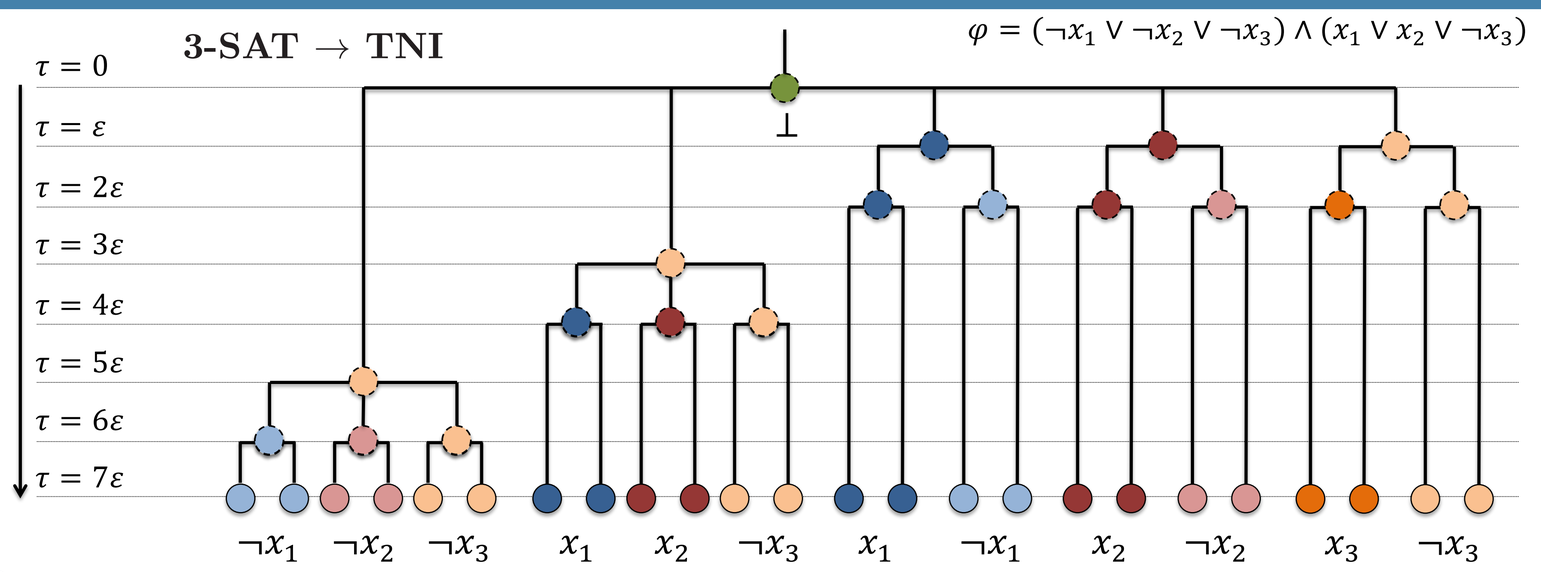
[1]Department of Aerospace Engineering, University of Illinois at Urbana-Champaign,
[2]Department of Computer Science, University of Illinois at Urbana-Champaign

1. Motivation and Problem Formulation

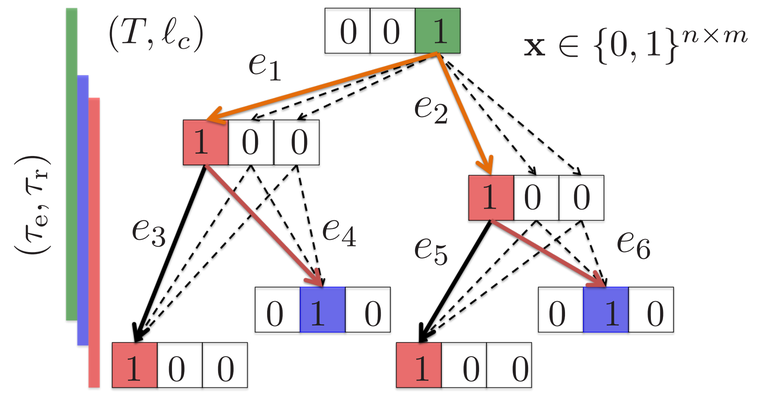


The evolutionary history of the pathogenic strains in an outbreak is described by a timed phylogeny T, assigning a time-stamp $\tau(v)$ to every vertex $v \in V(T)$, where $|V(T)| = n$. In addition, each leaf v is labeled by the host $\hat{\ell}(v)$ where the corresponding strain was observed (indicated by colors). Epidemiological data further constrain the entrance and removal time $[\tau_e(s), \tau_r(s)]$ of each host $s \in \Sigma$, where $|\Sigma| = m$. In the TNI problem, we seek a host labeling ℓ with minimum transmission number μ and subsequently smallest co-transmission number γ. (b) Host labeling ℓ_b with minimum transmission $\mu^* = 4$ but not the smallest co-transmission number $\gamma = 4$, resulting in a complex transmission network N_b. (c) Host labeling ℓ_c with minimum transmission $\mu^* = 4$ and smallest co-transmission number $\gamma^* = 2$, resulting in a parsimonious transmission network N_c. A time-invariant version of this problem has been applied to the analyses of migration in metastatic cancers [1].

2. Complexity - NP Hard



3. SAT formulation



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

	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

Definitions An edge (u, v) of T is a *transmission edge* if $\ell(u) \neq \ell(v)$. A *transmission event* Ψ is a subset of transmission edges between the same pair of hosts that have occurred simultaneously. A *transmission network* $N = \{\Psi_1, \ldots, \Psi_{|N|}\}$ is a partition of transmission edges into disjoint transmission events.

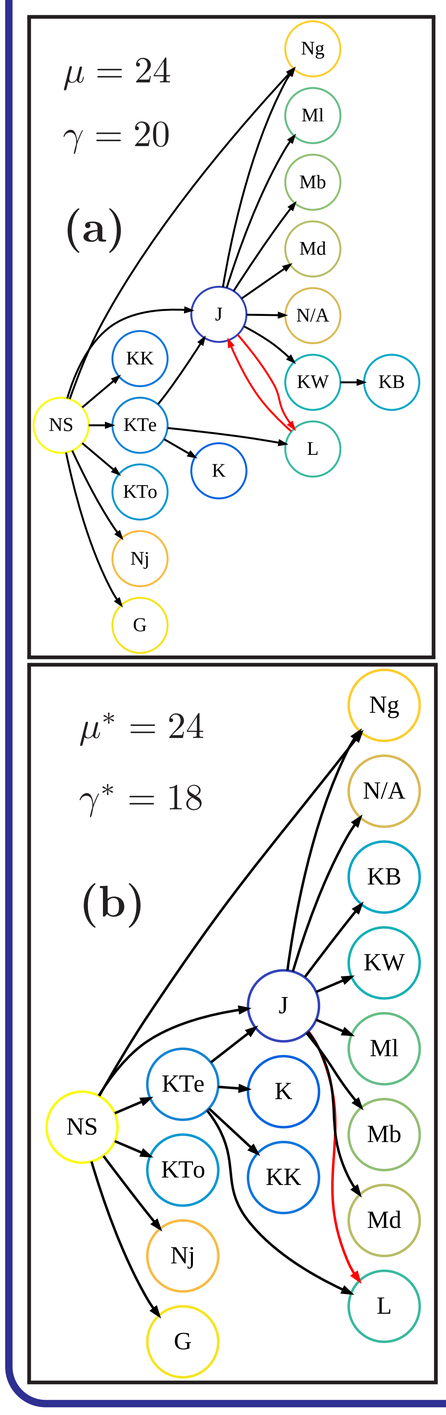
SAT Variables $\mathbf{x} \in \{0,1\}^{n\times m}$ encode a host labeling.

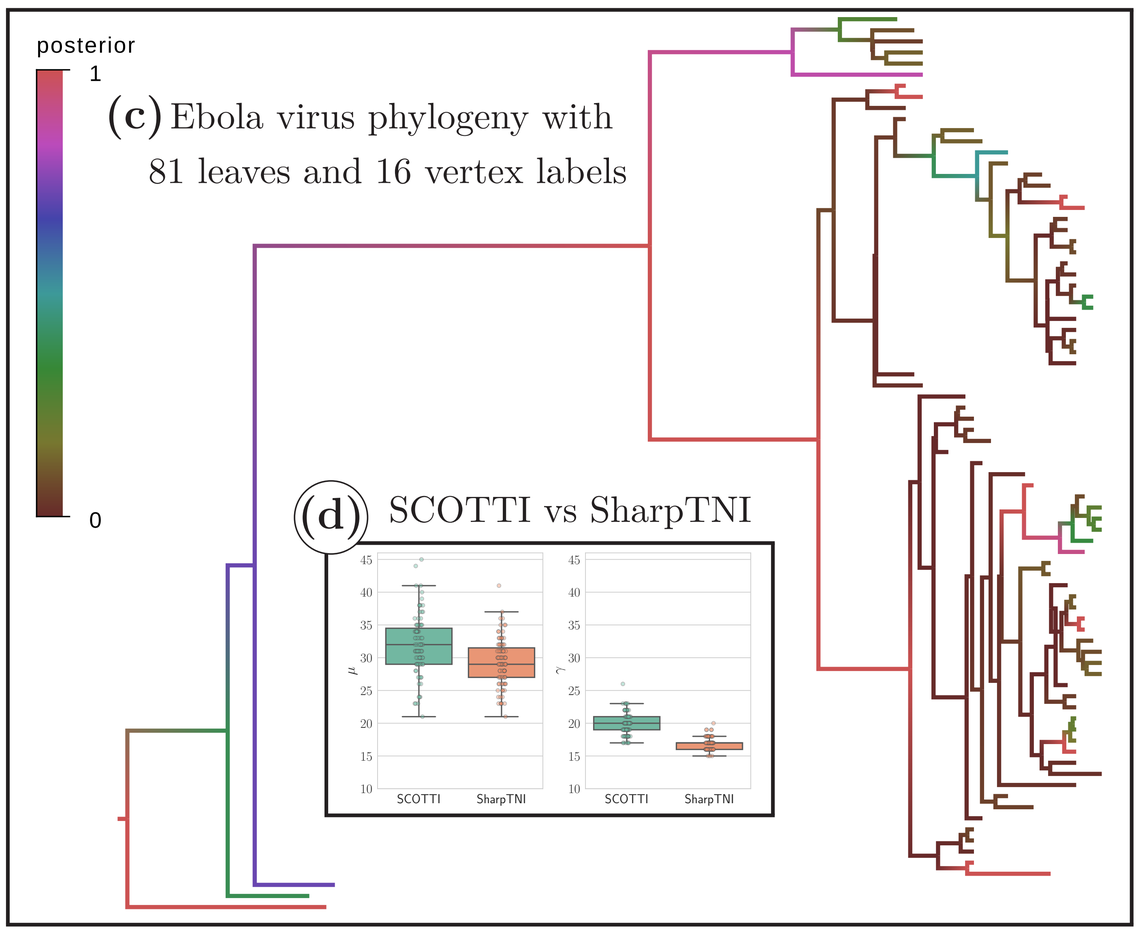
$$x_{i,s} = \begin{cases} 1, & \ell(v_i) = s, \\ 0, & \text{otherwise.} \end{cases}$$

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

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

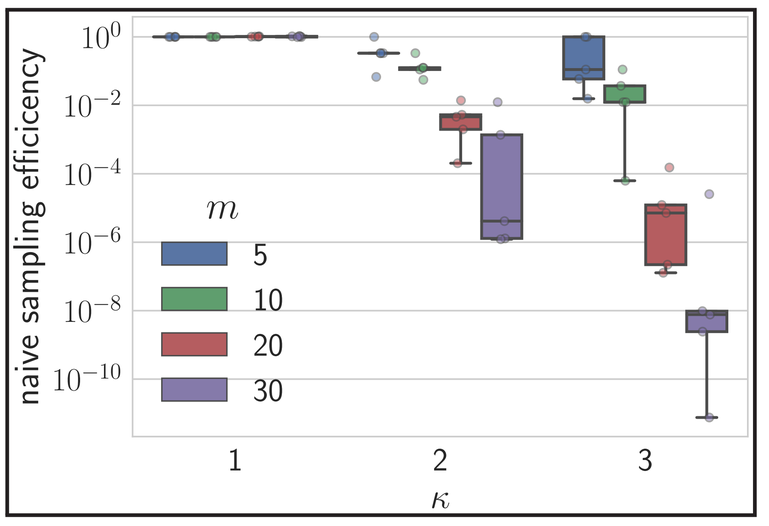
3. Ebola Outbreak of Sierra Leone in 2014





Transmission network inferred by **(a)** SCOTTI [2] with $\mu = 24$ and $\gamma = 20$ and **(b)** SharpTNI with $\mu^* = 24$ and $\gamma = 18$ for the same sample tree. The transmission network inferred by SCOTTI has a re-infection from Luawa to Jawie (highlighted in red) whereas in the transmission network inferred by SharpTNI there is no re-infection event. **(c)** The maximum clade credibility tree for the Ebola 2014 outbreak dataset obtained using BEAST [3]. **(d)** The transmission number μ (left) and smallest co-transmission number γ (right) inferred by SCOTTI (blue) and SharpTNI (orange) for 100 sample trees drawn from the posterior generated by SCOTTI. The γ of the host labeling inferred by SharpTNI is significantly smaller than the host labeling inferred by SCOTTI.

4. Simulation



Ratio between the approximate number of solutions to TNI (using UNIGEN [4]) and the number parsimonious Sankoff solutions for simulated outbreaks different number of hosts m and bottleneck sizes κ. This ratio corresponds to the success probability of the naive sampling algorithm to get parsimonious transmission network solutions.

6. References

[1] M. El-Kebir *et. al.*, *Nature Genetics*, vol. 50, pp. 718–726, May 2018.

[2] D. Maio *et. al.*, *PLoS computational biology*, vol. 12, no. 9, p. e1005130, 2016.

[3] A. J. Drummond *et. al.*, *BMC evolutionary biology*, vol. 7, no. 1, p. 214, 2007.

[4] S. Chakraborty *et. al.*, *Proc. of the 51st Annual Design Automation Conference*, pp. 1–6, 2014.