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

Palash Sashittal[1][0000−0001−6581−4839] and
Mohammed El-Kebir[2][0000−0002−1468−2407]

[1] Department of Aerospace Engineering, University of Illinois at Urbana-Champaign
[2] Department of Computer Science, University of Illinois at Urbana-Champaign
melkebir@illinois.edu http://www.el-kebir.net/

Abstract. Technological advances in genomic sequencing are facilitating the reconstruction of transmission histories during outbreaks in the fight against infectious diseases. However, accurate disease transmission inference using this data is hindered by a number of challenges due to within-host pathogen diversity and weak transmission bottlenecks, where multiple genetically-distinct pathogenic strains co-transmit. We formulate a combinatorial optimization problem for transmission network inference under a weak bottleneck from a given timed phylogeny and establish hardness results. We present SharpTNI, a method to approximately count and almost uniformly sample from the solution space. Using simulated data, we show that SharpTNI accurately quantifies and uniformly samples from the solution space of parsimonious transmission networks, scaling to large datasets. We demonstrate that SharpTNI identifies co-transmissions during the 2014 Ebola outbreak that are corroborated by epidemiological information collected by previous studies.

Keywords: Phylogenetics · Phylogeography · Migration · Transmission · Outbreak · Approximate counting · Almost-uniform sampling · Satisfiability.

Accurate inference of the transmission history of an infectious disease outbreak is pivotal for real-time outbreak management, public health policies and devising disease control strategies for future outbreaks. Traditional epidemiological approaches are fieldwork intensive and aim to uncover contact histories and exposure times of hosts to disease sources. With decreasing costs of genomic sequencing, molecular epidemiology has complemented these traditional approaches to effectively analyze and manage disease outbreaks.

Under a *weak transmission bottleneck* multiple genetically-distinct strains of the pathogen are simultaneously transmitted from a donor to a recipient through a non-negligibly small inoculum. Large inoculum sizes have been observed in a number of diseases. Here, we formulate the Transmission Network Inference (TNI) problem under a weak bottleneck for a given timed phylogeny (Fig. 1).

* supported by the National Science Foundation (CCF 18-50502).

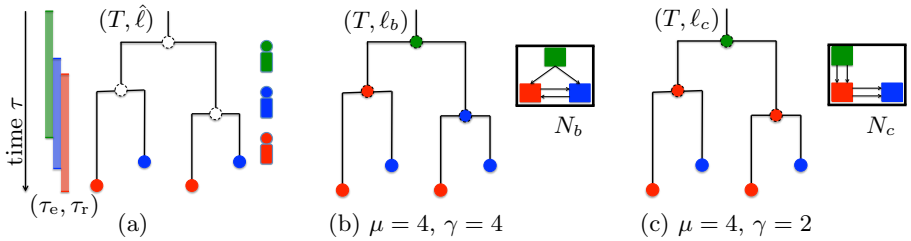

Fig. 1. Overview of the Transmission Network Inference (TNI) problem. (a)The evolutionary history of the pathogenic strains in an outbreak is described by a timed phylogeny T, assigning a time-stamp $\tau(v)$ to every vertex v. In addition, each leaf v is labeled by the host $\hat{\ell}(v)$ where the 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 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.

In this problem, we use the principle of parsimony to minimize the number of co-transmissions, which each may comprise of multiple transmitted strains. We prove hardness for the optimization and sampling versions of the problem. We introduce SharpTNI, a method to uniformly sample optimal solutions and quantify the size of the solution space. On simulated data, we show that SharpTNI accurately counts and samples parsimonious transmission networks, scaling to large datasets. We analyze a dataset from the 2014 Ebola outbreak [2], showing that SharpTNI outperforms SCOTTI [1] and recapitulates previously documented co-transmissions.

Accounting for weak transmission bottlenecks is crucial for accurate inference of transmission histories during outbreaks. SharpTNI is a parsimony-based method to reconstruct transmission networks for diseases with long incubation times and large inocula given timed phylogenies. The model and theoretical work of this paper pave the way for novel maximum likelihood methods to co-estimate timed phylogenies and transmission networks under a weak bottleneck.

References

1. De Maio, N., Wu, C.H., Wilson, D.J.: Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS computational biology **12**(9), e1005130 (2016)
2. Gire, S.K., Goba, A., Andersen, K.G., Sealfon, R.S., Park, D.J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al.: Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. science **345**(6202), 1369–1372 (2014)