JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 10, 2021
Mary Ann Liebert, Inc.
Pp. 945–960
DOI: 10.1089/cmb.2020.0500

Research Articles

Scale-Free Spanning Trees and Their Application in Genomic Epidemiology

YURY ORLOVICH,[1] KIRILL KUKHARENKO,[2,i] VOLKER KAIBEL,[2] and PAVEL SKUMS[3]

ABSTRACT

We study the algorithmic problem of finding the most "scale-free-like" spanning tree of a connected graph. This problem is motivated by the fundamental problem of genomic epidemiology: given viral genomes sampled from infected individuals, reconstruct the transmission network ("who infected whom"). We use two possible objective functions for this problem and introduce the corresponding algorithmic problems termed *m*-SF (-scale free) and *s*-SF Spanning Tree problems. We prove that those problems are APX- and NP-hard, respectively, even in the classes of cubic and bipartite graphs. We propose two integer linear programming (ILP) formulations for the *s*-SF Spanning Tree problem, and experimentally assess its performance using simulated and experimental data. In particular, we demonstrate that the ILP-based approach allows for accurate reconstruction of transmission histories of several hepatitis C outbreaks.

Keywords: computational complexity, genomic epidemiology, integer linear programming, scale-free network, transmission network.

1. INTRODUCTION

VIRAL OUTBREAKS CONTINUE to be major causes of morbidity and mortality. The ongoing pandemic of the coronavirus SARS-CoV-2 (Huang et al., 2020) is a vivid example, but long-standing epidemics of HIV, hepatitis B virus, and hepatitis C virus (HCV) are hardly less damaging (Kilmarx, 2009; Hajarizadeh et al., 2013). Viral epidemics are complex processes defined by evolutionary dynamics of pathogens and social dynamics of susceptible populations (e.g., individual behaviors, social interactions, and mobility patterns).

Recent advances in sequencing technologies invigorated the field of genomic epidemiology (Armstrong et al., 2019; Knyazev et al., 2020) that aims to use viral genomic data to understand the epidemiological dynamics of pathogens. The fundamental algorithmic problem of genomic epidemiology could be formulated as follows:

- Given viral genomes sampled from n infected individuals, infer a transmission network indicating who of them infected whom (Knyazev et al., 2020). If each individual is supposed to be infected only once, then a transmission network is a tree called a *transmission tree*.

[1]Faculty of Applied Mathematics and Computer Science, Belarusian State University, Minsk, Belarus.
[2]Institute for Mathematical Optimization, Otto von Guericke University Magdeburg, Magdeburg, Germany.
[3]Department of Computer Science, Georgia State University, Atlanta, Georgia, USA.
[i]ORCID ID (https://orcid.org/0000-0003-2959-6816).

This problem has been approached by a variety of methods (Jombart et al., 2011, 2014; Sledzieski et al., 2019; Wertheim et al., 2014; Campo et al., 2016; De Maio et al., 2016; Klinkenberg et al., 2017; Skums et al., 2018). One family of methods is based on the so-called network approach. It is particularly popular among researchers of HIV and HCV and has been adopted as a standard methodology for outbreak investigations carried out by the CDC (Wertheim et al., 2014; Campo et al., 2016; Campbell et al., 2017; Kosakovsky Pond et al., 2018; Ramachandran et al., 2018; Ragonnet-Cronin et al., 2019). This approach usually consists of two stages. First, a weighted *relatedness graph* G_R is constructed. Its vertices represent infected hosts, and edges connect the hosts whose viral populations are close to each other according to a selected population genetics measure. Often G_R itself supplies enough information for epidemiologists and provides a *fast and scalable alternative* to phylogenetic trees when applied to next-generation sequencing (NGS) data (Wertheim et al., 2014; Campo et al., 2016; Ragonnet-Cronin et al., 2019). However, usually it contains many edges that do not represent actual transmissions. Thus, at the second stage, the transmission tree is inferred as the spanning tree of G_R.

Under the maximum parsimony criterion, the most likely transmission network is a minimum spanning tree of G_R (Jombart et al., 2011). However, experiments demonstrated that this approach is not accurate (Jombart et al., 2014). Furthermore, genomic data alone often do not allow to resolve ambiguities in transmission tree inference, and incorporation of additional evidence is necessary (Jombart et al., 2014; Villandre et al., 2016; Jha et al., 2017). Such evidence usually comes in the form of epidemiological information, such as sample collection times and exposure intervals. However, HIV, HCV, and many other infections tend to be initially asymptomatic, and consequently, sampling times may not accurately reflect the infection times. In addition, in outbreaks with high transmission rates (e.g., HIV/HCV among injection drug users), susceptible hosts are almost constantly exposed to the virus, which makes exposure intervals useless. Another important drawback of many existing methods is their implicit assumption that transmission tree edges are independent. In reality, it is not the case, as, for example, certain hosts (so-called superspreaders) infect more people than an average person (Galvani and May, 2005).

Skums et al. (2018) proposed an alternative approach. It is known that for viruses, whose transmissions are associated with behavioral risk factors, their transmission trees have properties of so-called scale-free graphs (Leigh Brown et al., 2011; Wertheim et al., 2014). Those graphs have specific features, including power-law degree distribution, small diameter, and the presence of high-degree vertices (hubs). This observation gives rise to the following informally defined algorithmic problem (*scale-free spanning tree problem*): find the most ''scale-free-like'' spanning tree T of the graph G_R. In addition, constraints on the weight of T could be imposed. This approach was the basis of the Bayesian framework and the Markov Chain Monte Carlo algorithm for the transmission network inference described by Skums et al. (2018) and implemented as a tool called QUENTIN. Although QUENTIN is efficient in practice, it is a heuristic, and the questions about computational complexity and possibility of the exact solution of the problem were left open.

In this article, we present the first detailed study of the scale-free spanning tree problem. Our major contributions are as follows.

(1) We propose two rigorous formulations of the scale-free spanning tree problem further referred to as m-SF Spanning Tree and s-SF Spanning Tree problems. They are based on two related objective functions and, to the best of our knowledge, have not been previously studied.
(2) We establish the computational complexity of both problems by demonstrating that they are NP-hard or APX-hard, even when restricted to cubic graphs and bipartite graphs.
(3) We propose two integer linear programming (ILP) formulations for the problems, and perform computational experiments to assess their performance using simulated data. Then we apply an ILP approach to real genomic data from several epidemiologically curated HCV outbreaks investigated by the CDC (Campo et al., 2016; Skums et al., 2018) and demonstrate that it allows for accurate inference of transmission trees.

2. PRELIMINARIES

2.1. Problem formulations

We consider only finite undirected simple graphs and use standard graph-theoretic terminologies, see, for example, Chartrand et al. (2016). Let $G=(V, E)$ be a connected graph. For a vertex $x \in V(G)$, the *neighborhood* $N_G(x)$ of x is the set of all vertices that are adjacent to x in G. The *degree* of x is defined as

$\deg_G x = |N_G(x)|$. Several definitions of scale-free graphs of different degrees of mathematical rigor are known in a literature. We utilize the rigorous combinatorial characterization that has been introduced by Li et al. (2005) using the so-called *s*-metric of a graph. This graph invariant is defined as follows:

$$s(G) = \sum_{uv \in E(G)} \deg_G u \deg_G v. \tag{1}$$

The same parameter is known in mathematical chemistry as *second Zagreb index* (Das and Gutman, 2004; Borovicanin et al., 2017). Li et al. (2005) demonstrated that a higher *s*-metric indicates with high probability the presence of most of the expected properties of scale-free graphs. The intuition behind these results is that in a graph with a high *s*-metric, a large number of edges should be incident to high-degree vertices, thus forcing them to resemble preferential attachment graphs—a standard Barabási and Albert (1999) model for scale-free networks. Therefore, another mathematical chemistry parameter called the *first Zagreb index* (Borovicanin et al., 2017) or *m-metric* also can serve as a measure of ''scale-freeness'' of a graph:

$$m(G) = \sum_{uv \in E(G)} (\deg_G u + \deg_G v) = \sum_{u \in V(G)} (\deg_G u)^2. \tag{2}$$

Thus, we can formulate *m*-SF Spanning Tree and *s*-SF Spanning Tree problems: given a connected graph G, find the spanning tree T of G such that $m(T)$ (respectively, $s(T)$) is maximal. The respective maximum values of $m(T)$ and $s(T)$ are called *first* and *second SF-dimensions* of G and denoted by $\tau_1(G)$ and $\tau_2(G)$. By T^{sopt} and T^{mopt}, we denote an *s-optimal tree* and an *m-optimal tree* of G, respectively.

A somehow related problem has been studied by Kincaid et al. (2016): find a spanning subgraph with *prescribed vertex degrees* such that its *s*-metric is maximum. This problem is polynomially solvable in general, but becomes NP-hard, when the output spanning subgraph is required to be connected.

2.2. Mathematical preliminaries

2.2.1. Subgraph counting. Here we establish the characterizations for the *m*-metric and *s*-metric in terms of numbers of small subgraphs in a graph. This technique is used to establish complexity results in Section 3 and ILP formulations in Section 4.

Proposition 1. *For any graph G,*

$$m(G) = 2\gamma_2(G) + 2\gamma_1(G), \qquad s(G) = 3\gamma_\Delta(G) + \gamma_3(G) + 2\gamma_2(G) + \gamma_1(G),$$

where $\gamma_\Delta(G)$ is the number of triangles and $\gamma_t(G)$ is the number of paths of length t in G, respectively.

Proof. We prove only the second equality, the first one can be proved similarly. Let $A = [a_{ij}]$ be the adjacency matrix of G and $\boldsymbol{d}$ be its degree vector. We have $s(G) = \frac{1}{2}\boldsymbol{d}^T \cdot A \cdot \boldsymbol{d}$ and $\boldsymbol{d} = A \cdot \mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$. Therefore

$$s(G) = \frac{1}{2}\mathbf{1}^T \cdot A^3 \cdot \mathbf{1} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^{(3)},$$

where $a_{ij}^{(3)}$ denotes (i, j)-entry in the matrix A^3.

It is known that $a_{ij}^{(3)}$ equals the number of walks of length 3 between vertices i and j. Thus, $s(G)$ is equal to one-half of the total number of three-walks in G. An edge v_1v_2 produces exactly two such walks: $W_{11} = (v_1, v_2, v_1, v_2)$ and $W_{12} = (v_2, v_1, v_2, v_1)$. Each 2-path $\{v_1v_2, v_2v_3\}$ produces four 3-walks: $W_{21} = (v_1, v_2, v_3, v_2)$, $W_{22} = (v_2, v_3, v_2, v_1)$, $W_{23} = (v_2, v_1, v_2, v_3)$, and $W_{24} = (v_3, v_2, v_1, v_2)$. Each 3-path $\{v_1v_2, v_2v_3, v_3v_4\}$ produces two 3-walks: $W_{31} = (v_1, v_2, v_3, v_4)$ and $W_{32} = (v_4, v_3, v_2, v_1)$. Finally, each triangle with vertex set $\{v_1, v_2, v_3\}$ produces six 3-walks: $W_{\Delta 1} = (v_1, v_2, v_3, v_1)$, $W_{\Delta 2} = (v_1, v_3, v_2, v_1)$, $W_{\Delta 3} = (v_2, v_3, v_1, v_2)$, $W_{\Delta 4} = (v_2, v_1, v_3, v_2)$, $W_{\Delta 5} = (v_3, v_1, v_2, v_3)$, and $W_{\Delta 6} = (v_3, v_2, v_1, v_3)$. As every three-walk of G has one of these forms, the statement of the lemma follows. □

2.2.2. Neighbor switching. This is a tree rearrangement technique that is used for obtaining structural and complexity results. Let T be a tree and (u, v) be a pair of distinct vertices $u, v \in V(T)$, where $\deg_T u = p \geq 2$ and $\deg_T v = t \geq 2$. We denote the unique $u - v$ path in T by $P_T(u, v)$, and neighbors of u and

v laying on $P_T(u, v)$ by u^+ and v^-, respectively. In case u and v are not adjacent, the neighbor of u^+ distinct from u and laying on $P_T(u, v)$ is denoted by u^{++}. Let $A = N_T(u)\backslash\{u^+\} = \{a_1, \ldots, a_{p-1}\}$, and let the set $N_T(v)\backslash\{v^-\}$ be partitioned into two subsets $B = \{b_1, \ldots, b_q\}$ and $C = \{c_1, \ldots, c_r\}$, where $B \neq \emptyset$. Furthermore, let $\deg_T u^+ = \alpha$ and $\deg_T v^- = \beta$. Define numbers D_A, D_B, and D_C as follows:

$$D_A = \sum_{i=1}^{p-1} \deg_T a_i, \qquad D_B = \sum_{j=1}^{q} \deg_T b_j, \qquad D_C = \sum_{k=1}^{r} \deg_T c_k. \tag{3}$$

Given the pair (u, v), the *neighbor switch* $\mathcal{S}^B_{v\to u}$ is a transformation producing a new tree $\widetilde{T}$ from T by replacing the edges $vb_1, \ldots, vb_q$ with new edges $ub_1, \ldots, ub_q$ (Fig. 1). This operation changes only degrees of the vertices u and v, namely $\deg_{\widetilde{T}} u = p + q$, $\deg_{\widetilde{T}} v = r + 1$.

Lemma 2. *Suppose that $\mathcal{S}^B_{v\to u}(T) = \widetilde{T}$. If $p \geq r + 1$, $D_A > D_C$ and, in case u and v are not adjacent, additionally $\alpha \geq \beta$, then $s(\widetilde{T}) > s(T)$.*

Proof. We prove lemma when u and v are not adjacent, that is, $u \neq v^-$ and $v \neq u^+$ (the proof for the other case is similar). Define by X (resp., Y) the set of edges of T (resp., $\widetilde{T}$) incident to u or v. Let us denote by $\lambda(X)$ (resp., $\widetilde{\lambda}(Y)$) the contribution to $s(T)$ (resp., $s(\widetilde{T})$) from the edges of X (resp., Y). Then

$$s(\widetilde{T}) - s(T) = \widetilde{\lambda}(Y) - \lambda(X). \tag{4}$$

Using Equation (3) one can easily calculate

$$\lambda(X) = \deg_T u \deg_T u^+ + \deg_T v^- \deg_T v + \sum_{i=1}^{p-1} \deg_T u \deg_T a_i + \sum_{j=1}^{q} \deg_T v \deg_T b_j$$
$$+ \sum_{k=1}^{r} \deg_T v \deg_T c_k = p\alpha + \beta t + pD_A + tD_B + tD_C.$$

After substituting $t = q + r + 1$, we obtain

$$\lambda(X) = p\alpha + \beta q + \beta(r + 1) + pD_A + qD_B + (r + 1)D_B + qD_C + (r + 1)D_C. \tag{5}$$

Similarly,

$$\widetilde{\lambda}(Y) = p\alpha + q\alpha + \beta(r + 1) + pD_A + qD_A + pD_B + qD_B + (r + 1)D_C. \tag{6}$$

Using equalities (4)–(6) we obtain

$$\begin{aligned} s(\widetilde{T}) - s(T) = \widetilde{\lambda}(Y) - \lambda(X) &= q\alpha + qD_A + pD_B - \beta q - (r + 1)D_B - qD_C \\ &= q(\alpha - \beta) + D_B(p - r - 1) + q(D_A - D_C). \end{aligned} \tag{7}$$

Since $\alpha \geq \beta$ and $p \geq r + 1$, it follows that $q(\alpha - \beta) + D_B(p - r - 1) \geq 0$. On the contrary, since $q \geq 1$ and $D_A > D_C$, we have $q(D_A - D_C) > 0$ and therefore $s(\widetilde{T}) - s(T) > 0$. □

If $B = N_T(v)\backslash\{v^-\}$, then the neighbor switch produces a tree $\widetilde{T}$ with v being a leaf. In this case $\mathcal{S}^B_{v\to u}$ is a *total neighbor switch*. For our goals it suffices to prove the following corollary.

Corollary 3. *If $\widetilde{T}$ is obtained from T by a total neighbor switch $\mathcal{S}^B_{v\to u}$ and, in case of u and v not being adjacent, additionally $\alpha \geq \beta$ or $p \geq \beta$, then $s(\widetilde{T}) > s(T)$.*

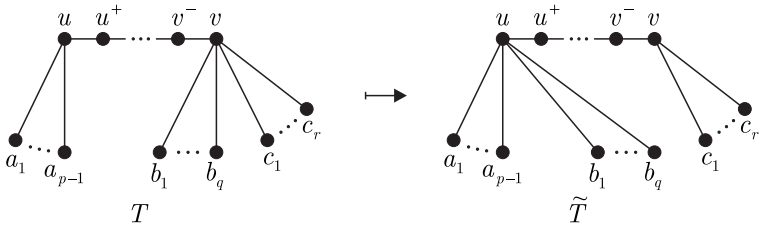

FIG. 1. Neighbor switch.

Proof. We check that all conditions of Lemma 2 are satisfied for the total neighbor switch. Indeed, since $D_A \geq p-1 \geq 1$ (recall $\deg_T u = p \geq 2$) and $D_C = r = 0$, we have $D_A > D_C$ and $p \geq r+1$. If u and v are not adjacent, we still require that $\alpha \geq \beta$, as in Lemma 2. However, this condition can be replaced if we rewrite Equation (7) as follows:

$$s(\widetilde{T}) - s(T) = q(\alpha - \beta) + D_B(p-1) + qD_A = q(\alpha + D_A - \beta) + D_B(p-1).$$

Note that the latter expression is positive in case of $p \geq \beta$, since $\alpha \geq 2$ and $D_A \geq p-1 \geq 1$.

In the same way we can compare the trees T and $\widetilde{T} = \mathcal{S}^B_{v \to u}(T)$ in terms of their m-metrics.

Lemma 4. *Suppose that $\mathcal{S}^B_{v \to u}(T) = \widetilde{T}$ and $p > r+1$. Then $m(\widetilde{T}) > m(T)$.*

Proof. The idea is similar to the proof of Lemma 2. Since the neighbor switch changes only degrees of vertices u and v, $m(\widetilde{T}) - m(T) = \deg^2_{\widetilde{T}} u + \deg^2_{\widetilde{T}} v - \deg^2_T u - \deg^2_T v = 2q(p-r-1)$, which proves the lemma, since $q \geq 1$. □

For further results we need weaker modifications of Lemmas 2 and 4 for the case $\deg_T u = p \geq 1$ (and therefore $D_A \geq 0$). Recall $\deg_T v = t \geq 2$ since we still require at least one vertex to switch.

Lemma 5. *Suppose $\widetilde{T}$ is obtained from T by a total neighbor switch $\mathcal{S}^B_{v \to u}$, then the following propositions hold:*

(a) $m(\widetilde{T}) \geq m(T)$;
(b) $s(\widetilde{T}) \geq s(T)$ (unless u and v are not adjacent with $\alpha < \beta$).

Now we consider a special case, when u is a vertex of maximum degree in T and all vertices in $N_T(v) \backslash \{v^-\}$ are leaves. In addition, let $B, C \neq \emptyset$. We introduce a *double neighbor switch* $\mathcal{S}^{B,C}_{v \to u,\, u^+}(T) = \mathcal{S}^C_{v \to u^+}(\mathcal{S}^B_{v \to u}(T))$. The reason to treat this two-step switch as a single operation is that the first switch itself might cause the descend of s-metric, however, the decrease would be compensated by the second switch.

Lemma 6. *If $\widehat{T}$ is obtained from T by a double neighbor switch $\mathcal{S}^{B,C}_{v \to u,\, u^+}(T)$, then $s(\widehat{T}) > s(T)$.*

Proof. In case u and v are adjacent, that is, $u^+ = v$, a double switch $\mathcal{S}^{B,C}_{v \to u,\, u^+}(T)$ gets reduced to the first neighbor switch $\mathcal{S}^B_{v \to u}(T)$, which produces a tree with a higher s-metric due to Lemma 2. Therefore, assume u and v are not adjacent. Consider the first switch and let $\widetilde{T} = \mathcal{S}^B_{v \to u}(T)$. From Equation (7), since $D_B = q$ and $D_C = r$, we obtain

$$s(\widetilde{T}) - s(T) = q(\alpha + D_A - \beta) + q(p-r-1) - qr. \tag{8}$$

Next let $\widehat{T} = \mathcal{S}^C_{v \to u^+}(\widetilde{T})$ be obtained by the total neighbor switch. To avoid reassigning of notations we denote $D_E = \sum_{w \in N_{\widetilde{T}}(u^+) \backslash \{u^{++}\}} \deg_{\widetilde{T}} w$ and $\gamma = \deg_{\widetilde{T}} u^{++}$. Other notations stay the same from the first switch. Again from Equation (7) we get

$$s(\widehat{T}) - s(\widetilde{T}) = r(\gamma - \beta) + r(\gamma - 1) + rD_E. \tag{9}$$

Summation of Equations (8) and (9) gives

$$s(\widehat{T}) - s(T) = q(\alpha + D_A - \beta) + q(p-r-1) + r(\gamma + D_E - \beta - q) + r(\gamma - 1),$$

where $D_A \geq p-1$, $\alpha \geq 2$, $\gamma \geq 2$, and $D_E \geq \deg_{\widetilde{T}} u = p+q$. Furthermore, since u is a vertex of maximum degree in T, $p \geq \beta$ and $p \geq q+r+1 > r+1$ (recall $q, r > 0$), which proves the lemma. □

2.3. Bounds in terms of the maximum degree

There exist bounds for both SF-dimensions of a graph in terms of its order only (de Caen, 1998; Das, 2003; Das and Gutman, 2004). However, they are not particularly efficient, when used as ILP cuts. Here we provide the adjusted upper bounds that turned out to be more useful for that purpose. Let $\Delta(G)$ denote the maximum vertex degree of G and $S_{m,k}$ denote a *double star*, that is, a tree obtained from two disjoint stars $K_{1,m}$ and $K_{1,k}$ with m and k leaves, respectively, by adding an edge joining their central vertices.

Theorem 7. *For any graph G of order $n \geq 2$,*

$$\begin{aligned} \tau_1(G) &\leq m(S_{\Delta(G)-1,\, n-\Delta(G)-1}) = 2\Delta^2(G) + n^2 - 2n\Delta(G) + n - 2, \\ \tau_2(G) &\leq s(S_{\Delta(G)-1,\, n-\Delta(G)-1}) = n(n-\Delta(G)-1) + \Delta^2(G). \end{aligned}$$

Proof. We provide the proof for the second SF-dimension only (the other proof is similar). Suppose T^{sopt} is an s-optimal tree of G and $T^{sopt} \neq S_{\Delta(G)-1,\, n-\Delta(G)-1}$. We prove the statement by performing a sequence of neighbor switches on T^{sopt}, with each of them increasing s-metric, so that the resulting tree is $S_{\Delta(G)-1,\, n-\Delta(G)-1}$.

Let u be a vertex of maximum degree in T^{sopt}. Then for every v in T^{sopt} follows $\deg_{T^{sopt}} v \leq \deg_{T^{sopt}} u \leq \deg_G u \leq \Delta(G)$. Let $T := T^{sopt}$. We divide the sequence of neighbor switches into three stages.

STAGE 1: *For each vertex v with all vertices in $N_T(v) \backslash \{v^-\}$ (where $v^- \in P_T(u, v)$) being leaves, we either perform the total neighbor switch $T := \mathcal{S}^B_{v \to u}(T)$ or double neighbor switch $T := \mathcal{S}^{B,C}_{v \to u,\, u^+}(T)$ until the degree of u is not equal to $\Delta(G)$.*

One can observe that a double neighbor switch is needed to ensure that $\deg_T u$ can be increased exactly to $\Delta(G)$. Since $\deg_T u$ increases after each switch, only the finite number of switches is required. In case the tree T obtained after the first stage differs from $S_{\Delta(G)-1,\, n-\Delta(G)-1}$, we perform the second stage if there exist at least two vertices w_1 and w_2 in $N_T(u)$ with $\deg_T w_1 \geq \deg_T w_2 \geq 2$ or jump directly to Stage 3 otherwise.

STAGE 2: *For each distinct w_1 and w_2 in $N_T(u)$ with $\deg_T w_1 \geq \deg_T w_2 \geq 2$ perform a total neighbor switch $T := \mathcal{S}^B_{w_2 \to w_1}(T)$.*

After each iteration, the number of vertices in $N_T(u)$ with degree at least two decreases by one. Thus, Stage 2 terminates after a finite number of switches leaving at most one vertex $w \in N_T(u)$ with degree at least two. Finally if T still differs from $S_{\Delta(G)-1,\, n-\Delta(G)-1}$, the third stage is required.

STAGE 3: *While there exists vertex v in $N_T(w) \backslash \{u\}$ with $\deg_T v \geq 2$ perform a total neighbor switch $T := \mathcal{S}^B_{v \to w}(T)$.*

Since the number of neighbors of w with degrees at least two decreases after each switch, Stage 3 terminates after finite number of steps with all neighbors of w, except for u, being leaves, that is, $T = S_{\Delta(G)-1,\, n-\Delta(G)-1}$. Note that each iteration of Stages 1–3 produces a tree with a higher s-metric due to Lemmas 2, 6 and Corollary 3. □

3. HARDNESS RESULTS

In this section, we study the computational complexity of both the m-SF and the s-SF SPANNING TREE problem. The following known fact is used:

Theorem 8 (Kleitman and West, 1991). *Any connected graph of order n with minimum vertex degree at least 3 has a spanning tree with at least $n/4 + 2$ leaves.*

We start by investigating the complexity of our problems for cubic graphs.

Theorem 9. *The m-SF SPANNING TREE problem is APX-hard for cubic graphs.*

Proof. Let G be a cubic graph on n vertices and T be a spanning tree with $\ell = \ell(T)$ leaves and $n_i = n_i(T)$ vertices of degree i, $i \in \{2, 3\}$. Then

$$m(T) = \ell + 4n_2 + 9n_3, \tag{10}$$

with the numbers n_i satisfying the equalities $\ell + n_2 + n_3 = n$ and $\ell + 2n_2 + 3n_3 = 2(n-1)$. Deriving n_2 and n_3 from these equalities gives us

$$n_2 = n + 2 - 2\ell, \qquad n_3 = \ell - 2. \tag{11}$$

After substituting these expressions into Equation (10), we get

$$m(T) = 2\ell + 4n - 10. \tag{12}$$

Thus, finding a spanning tree with maximum m-metric in this case is polynomially equivalent to finding a spanning tree with maximum number of leaves (MAXLEAF problem). For cubic graphs, the latter problem

was shown to be APX-hard by Bonsma (2012). Thus, we prove the APX-hardness of the m-SF Spanning Tree problem by providing an L-reduction (Papadimitriou and Yannakakis, 1991) from MaxLeaf.

Given an optimization problem P and an instance I of this problem, we use $opt_P(I)$ to denote the optimum value of I, and $val_P(I, S)$ to denote the value of a feasible solution S of instance I. Let A and B be two optimization problems. The problem A is said to be L-reducible to B if there exist polynomial-time computable functions f, g and constants $\alpha, \beta > 0$ such that

(L1) f maps any instance I of A to an instance $f(I)$ of B such that $opt_B(f(I)) \leq \alpha \cdot opt_A(I)$;

(L2) for any instance I of A and a solution S' of the instance $f(I)$, g maps S' to a solution S for I such that $|val_A(I, S) - opt_A(I)| \leq \beta \cdot |val_B(f(I), S') - opt_B(f(I))|$.

Let T^{mopt} be an m-optimal spanning tree of G and ℓ^* be the maximum number of leaves in spanning trees of G. Note that $\ell^* \geq n/4 + 2$ by Theorem 8, and therefore, $n \leq 4\ell^* - 8$. Then using Equation (12) we get

$$\tau_1(G) = m(T^{mopt}) = 2\ell(T^{mopt}) + 4n - 10 \leq 2\ell^* + 16\ell^* - 32 \leq 18\ell^*.$$

Moreover, for every spanning tree T of G we have $\frac{1}{2}|m(T) - m(T^{mopt})| = |\ell(T) - \ell^*|$. As a result, Equation (12) implies an L-reduction with identity mappings f and g and constants $\alpha = 18$ and $\beta = \frac{1}{2}$, thus proving the theorem.

Theorem 10. *The* s-SF Spanning Tree *problem is* NP*-hard for cubic graphs.*

Proof. For the reduction, we use the following problem proved to be NP-complete by Lemke (1988):

Instance: A connected cubic graph G of order n.

Question: Is there a spanning tree of G without vertices of degree 2?

According to Equation (11), $n_2 = n_2(T) = n + 2 - 2\ell(T)$. Thus, the answer for the problem's question is negative if n is odd. Hence, we concentrate only on the case when $n \geq 4$ is even, thus n_2 is even as well. We show that among all trees T of order n with $\Delta(T) \leq 3$, the trees without vertices of degree 2 have the highest s-metric. Indeed, the following claim holds:

Claim 11. *If* $\Delta(T) \leq 3$ *and* $n \geq 4$ *are even, then* $s(T) \leq 6n - 15$. *The equality holds if and only if* T *has no vertices of degree* 2.

Proof. If T has no vertices of degree 2, then Equation (11) implies $\ell = \ell(T) = \frac{n+2}{2}$. Furthermore, $s(T) = 3m_1 + 9m_3$, where m_1 is the number of edges incident to a leaf and m_3 is the number of edges with both ends of degree 3. Obviously, $m_1 = \ell$ and $m_3 = n - 1 - \ell$, thus yielding $s(T) = 6n - 15$.

Now suppose that T has $n_2 \geq 2$ vertices of degree 2. Let u and v be two vertices of degree 2 lying on a path $P_T(u, v)$ and $\deg_T u^+ \geq \deg_T v^-$. Iteratively applying a total neighbor switch $\mathcal{S}^B_{v \to u}$ for all pairs of vertices u and v of degree 2, we obtain a tree with higher s-metric (due to Corollary 3) and without vertices of degree 2. This proves the claim. □

Thus, $\tau_2(G) = 6n - 15$ if and only if G has a spanning tree without vertices of degree 2. This concludes the proof. □

Next, we consider bipartite graphs.

Theorem 12. *The* m-SF Spanning Tree *and* s-SF Spanning Tree *problems are* NP*-hard for bipartite graphs.*

Proof. We present a polynomial-time reduction from the NP-complete 3-Dimensional Matching (3-DM) problem (Garey and Johnson, 1979):

Instance: Pairwise disjoint sets X, Y, Z of cardinality n, and a collection $\mathcal{M}$ of m three-element sets, where each $M \in \mathcal{M}$ includes exactly one element from each of X, Y, and Z.

Question: Is there a set of pairwise disjoint members of $\mathcal{M}$ (a *perfect* 3*-dimensional matching*), whose union is $X \cup Y \cup Z$?

Let $Q = (X, Y, Z, \mathcal{M})$ be an instance of 3-DM. We construct a graph $G = G_Q$ on $3n + m + 1$ vertices as follows. The vertex set of G is the disjoint union $\{r\} \cup A \cup B$, where $A = \mathcal{M}$, $B = X \cup Y \cup Z$, and r is the special root vertex. The edge set includes all edges ra, $a \in A$, as well as the edges Mx, My, and Mz for each $M = \{x, y, z\} \in A$ (Fig. 2). We may assume that G is connected. Note also that G is a bipartite graph with the parts A and $\{r\} \cup B$.

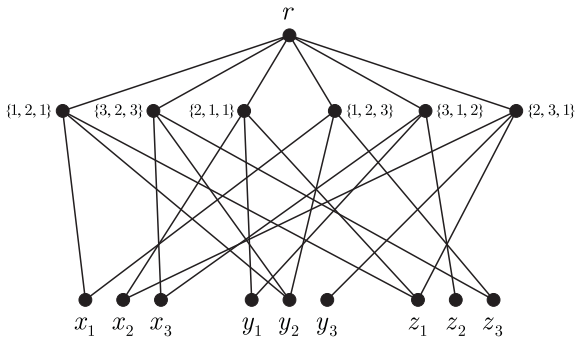

FIG. 2. An example of the graph G for $n=3$, $X=\{x_1, x_2, x_3\}$, $Y=\{y_1, y_2, y_3\}$, $Z=\{z_1, z_2, z_3\}$, and $\mathcal{M}=\{\{x_1, y_2, z_1\}, \{x_3, y_2, z_3\}, \{x_2, y_1, z_1\}, \{x_1, y_2, z_3\}, \{x_3, y_1, z_2\}, \{x_2, y_3, z_1\}\}$. Here each vertex labeled $\{p, q, r\}$ represents a set $\{x_p, y_q, z_r\}$.

For a vertex v of G and a subset $W \subseteq V(G)$ let us denote by $(v : W)$ the set of edges connecting v to vertices in W.

Lemma 13. *There are spanning trees T_1 and T_2 in G, both containing all edges of $(r : A)$, with $m(T_1)=\tau_1(G)$ and $s(T_2)=\tau_2(G)$.*

Proof. We provide the proof for the s-metric, the proof for the m-metric is similar. Among the optimal spanning trees of G, let T_2 be the one with the maximum number of edges from $(r : A)$. We claim that T_2 contains all these edges.

Suppose for a contradiction that the set $C \subseteq A$ of all vertices that are adjacent to r in T_2 differs from A. Then there must be a vertex $b \in B$ with $P_{T_2}(r, b)$ having two edges, such that set $D=N_{T_2}(b) \cap \{A \backslash C\}$ is nonempty. By Lemma 5, since $\deg_{T_2} r^+ = \deg_{T_2} b^-$ and $\deg_{T_2} r \geq 1$, we can apply total neighbor switch $\mathcal{S}^D_{b \to r}$ to construct a spanning tree T'_2 from T_2 with $s(T'_2) \geq s(T_2)$, and the root r having more neighbors in T'_2 than it has in T_2. □

Any spanning tree T of G containing all edges of $(r : A)$ has $m+3n$ edges, $3n(m-1)$ paths of length three (each of the $3n$ edges of the tree connecting A and B induces exactly $m-1$ such paths), and $m(m-1)/2+3n$ paths of length two that are not formed by a pair of edges between A and B. There are $3\delta_4+\delta_3$ remaining paths of length two, where δ_i is the number of vertices in A that have degree i in the tree. Indeed, a vertex $v \in A$ with $j \in \{0, 1, 2, 3\}$ neighbors from B in the tree contributes no such path in case of $j \in \{0, 1\}$, one such path in case of $j=2$, and three such paths in case of $j=3$. Thus by Proposition 1

$$m(T)=m^2+m+12n+6\delta_4+2\delta_3, \qquad s(T)=m^2+3mn+6n+6\delta_4+2\delta_3.$$

Since $|B|=3n$, we have $3\delta_4+2\delta_3 \leq 3n$ and $6\delta_4+2\delta_3 \leq 6\delta_4+4\delta_3 \leq 6n$. Hence, $6\delta_4+2\delta_3 \leq 6n$ with equality holding if and only if $\delta_3=0$ and $\delta_4=n$.

A perfect 3-DM $\mathcal{M}^*=\{M_1, \ldots, M_n\}$ induces the spanning tree $T_{\mathcal{M}^*}$ that contains all edges from $(r : A)$ and edges ax, ay, az for each $a=\{x, y, z\} \in \mathcal{M}^*$. For this tree we have $\delta_4=n$ and

$$m(T)=m^2+m+18n := t_1(n, m), \qquad s(T)=m^2+3mn+12n = : t_2(n, m).$$

Conversely, every spanning tree T that contains all edges from $(r : A)$ and $m(T)=t_1(n, m)$ or $s(T)=t_2(n, m)$ (and thus $\delta_4=n$) arises from a perfect 3-DM.

By Lemma 13, the graph G satisfies $\tau_1(G) \geq t_1(n, m)$ (resp., $\tau_2(G) \geq t_2(n, m)$) if and only if there is a spanning tree T of G that contains all edges from $(r : A)$ and whose m-metric (resp., s-metric) is equal to $t_1(n, m)$ (resp., $t_2(n, m)$). The latter is true if and only if Q has a perfect 3-DM. □

4. ILP FORMULATIONS

Here we describe two ILP models for the s-SF Spanning Tree problem (for the m-SF Spanning Tree problem the approach is similar). For a given spanning tree T of a graph $G=(V, E)$ of order n, consider the indicator variables $(x_e)_{e \in E}$:

$$x_e = \begin{cases} 1, & e \in E(T); \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Using Proposition 1, we can represent $s(T)$ as

$$s(T) = \sum_{\{e_i, e_j, e_k\} \in \Gamma_3(G)} x_{e_i} x_{e_j} x_{e_k} + 2 \sum_{\{e_i, e_j\} \in \Gamma_2(G)} x_{e_i} x_{e_j} + \sum_{e \in E(G)} x_e, \tag{14}$$

where $\Gamma_i(G)$ denotes the set of all paths of length i in G. To linearize (14), we introduce Boolean variables y_{ijk} and y_{ij} and the following constraints:

$$\begin{aligned} &y_{ijk} \le x_{e_i}, && y_{ij} \le x_{e_i}, \\ &y_{ijk} \le x_{e_j}, && y_{ij} \le x_{e_j}, \\ &y_{ijk} \le x_{e_k}, && y_{ij} \ge x_{e_i} + x_{e_j} - 1, \\ &y_{ijk} \ge x_{e_i} + x_{e_j} + x_{e_k} - 2, \end{aligned} \tag{15}$$

for every $\{e_i, e_j, e_k\} \in \Gamma_3(G)$ and $\{e_i, e_j\} \in \Gamma_2(G)$, which are equivalent to $y_{ijk} = x_{e_i} x_{e_j} x_{e_k}$ and $y_{ij} = x_{e_i} x_{e_j}$. Thus the objective function (14) can be rewritten as

$$s(T) = \sum_{\{e_i, e_j, e_k\} \in \Gamma_3(G)} y_{ijk} + 2 \sum_{\{e_i, e_j\} \in \Gamma_2(G)} y_{ij} + \sum_{e \in E(G)} x_e. \tag{16}$$

We use two types of constraints to describe the spanning trees. The first type is the extended formulation of Martin (1991), which uses auxiliary variables

$$z^r_{(v, w)}, z^r_{(w, v)} \ge 0 \quad \text{for every } r \in V(G),\ vw \in E(G), \tag{17}$$

where $z^r_{(v, r)} = 0$ for every $r \in V(G)$ and $vr \in E(G)$. A 0/1-vector x describes a spanning tree of G if and only if these variables satisfy the constraints

$$\begin{aligned} x_{vw} - z^r_{(v, w)} - z^r_{(w, v)} &= 0, \quad r \in V(G),\ vw \in E(G), \\ \sum_{vw \in E(G)} z^r_{(v, w)} &= 1, \quad r, w \in V(G),\ r \ne w, \\ \sum_{vr \in E(G)} z^r_{(v, r)} &= 0, \quad r \in V(G). \end{aligned} \tag{18}$$

The second type exploits the Miller–Tucker–Zemlin (MTZ) constraints (Miller et al., 1960). We introduce the auxiliary variables

$$\begin{aligned} &z_{(v, w)}, z_{(w, v)} \in \{0, 1\} \quad \text{for every } vw \in E(G), \\ &t_v \in [0, n-1] \quad \text{for every } v \in V(G), \end{aligned} \tag{19}$$

and constraints

$$\begin{aligned} x_{vw} - z_{(v, w)} - z_{(w, v)} &= 0, \quad vw \in E(G), \\ \sum_{vw \in E(G)} z_{(v, w)} &= 1, \quad w \in V(G) \backslash \{r\}, \\ \sum_{vr \in E(G)} z_{(v, r)} &= 0, \\ t_v - t_w + n z_{(v, w)} &\le n - 1, \quad v, w \in V(G),\ vw \in E(G), \end{aligned} \tag{20}$$

where $r \in V(G)$ is some fixed vertex. Finally we add the additional constraint

$$s(T) = \sum_{\{e_i, e_j, e_k\} \in \Gamma_3(G)} y_{ijk} + 2 \sum_{\{e_i, e_j\} \in \Gamma_2(G)} y_{ij} + \sum_{e \in E(G)} x_e \le n(n - \Delta(G) - 1) + \Delta^2(G), \tag{21}$$

defined by Theorem 7, which turns out to significantly improve the algorithm running times. Maximization of the objective (16) subject to the constraints (15), (18), (21) is further referred to as Martin formulation, while maximization of Equation (16) subject to Equations (15), (20), (21) as MTZ formulation.

5. EXPERIMENTAL RESULTS

In this section, we investigate the practical aspects of scale-free spanning tree problems by conducting computational experiments for various simulated and experimental data sets to evaluate the performance of the ILP models. All computations below were performed on a standard laptop with 2.0 GHz dual core processor and 16 GB of RAM, and ILP problems were solved using Gurobi 8.1.

5.1. Synthetic data

5.1.1. Synthetic graphs. We used graphs from the following synthetic data sets:

Erdős-Rényi graphs constructed by adding each possible edge uniformly and independently with the probability $p = 4.25/n$. The number of nodes n in our experiments varied from 10 to 40 (corresponding to the sizes of HCV outbreaks analyzed later).

$n \times m$ grid graphs (Cartesian products of paths P_n and P_m) with $n, m = 4, \ldots, 7$.

Scale-free graphs of two types generated using NetworkX library (Hagberg et al., 2008): those based on the classical Barabási and Albert (1999) model and those constructed with NetworkX default parameters. The latter graphs are usually denser.

For all synthetic data sets except for grid graphs, we generated 10 graphs per node number. Figures 3 and 4 illustrate the running times of the ILP solver on both the MTZ formulation and the Martin formulation compared with the published tool QUENTIN (Skums et al., 2018) runtimes for all four simulated graph classes.[1] The results demonstrate that for those graph classes, the ILP algorithms in average perform much better than in the worst case and are able to produce optimal results in a reasonable amount of time. Moreover, for considered graph sizes, they outperform QUENTIN. For Erdős–Rényi graphs and grids (Fig. 3), which are characterized by relatively large sets of feasible solutions, the Martin formulation was superior to MTZ and QUENTIN, while for Barabási–Albert scale-free graphs (Fig. 4a), the MTZ formulation was leading to the faster algorithm. In general, the ILP approach allows to solve the problem within minutes or few hours for small-to-medium-sized problems (up to several dozens of vertices) on Erdős–Rényi graphs and grids, and for medium-sized problems (several hundred vertices) on scale-free graphs.

5.1.2. Simulated outbreaks. We simulated outbreaks over scale-free Barabási–Albert contact networks of $n = 10 - 30$ nodes using the following model. The infection spreads over each network according to the susceptible infected (SI) model (Newman, 2010) with the transmission rate $\rho = 10^{-2}$. Each infected individual is assumed to carry a viral sequence of length $m = 13200$, and at each transmission event, the source's sequence is transmitted to the recipient. Sequence evolution is described by a skyline model with the piecewise constant decreasing mutation rate, that is, viral sequences mutate at the basic rate of $\mu = 10^{-5}$ changes/position/time unit, and the mutation rate is decreasing by 30% every $\tau = 100$ time units. This model captures the decrease of the speed of intrahost evolution as the infection progresses from an acute to a persistent stage (De Maio et al., 2016; Icer et al., 2020).

For each simulated outbreak, we compared the performance of the ILP algorithm for the Martin formulation, with the standard approach based on the phylogenetic trait inference (Sagulenko et al., 2018). First, we constructed a maximum likelihood phylogeny using MEGA (Kumar et al., 2018). Each patient was encoded by a discrete trait, and the marginal likelihood ancestral traits were reconstructed using the Felsenstein pruning algorithm (Felsenstein, 2004) with the pairwise between-trait transition rates equal to ρ. Inferred transmission links then correspond to trait changes along the phylogeny branches. The genetic relatedness network G_R used as an input for the ILP was constructed using a threshold-based approach suggested by Kosakovsky Pond et al. (2018). A pair of vertices of G_R are adjacent, if the Hamming distance between the corresponding sequences does not exceed a threshold t that was estimated as the minimal integer such that the graph G_R is connected. The obtained graph was further sparsed out by applying the same procedure to each of its biconnected component.

[1]Running times for MTZ formulation on grids and Martin formulation on Barabási–Albert scale-free graphs are plotted only for smaller n, since for large values they are significantly higher than for the other formulation. In particular, Martin formulation on Barabási–Albert scale-free graphs works $\sim$ 150 seconds for 1000 vertices, $\sim$ 480 seconds for 1500 vertices, and exceeds 1800 seconds for 2000 and more vertices. QUENTIN running times are not plotted in Figure 4, since they already exceed timeout of 3600 seconds for 50 vertices for both scale-free graphs.

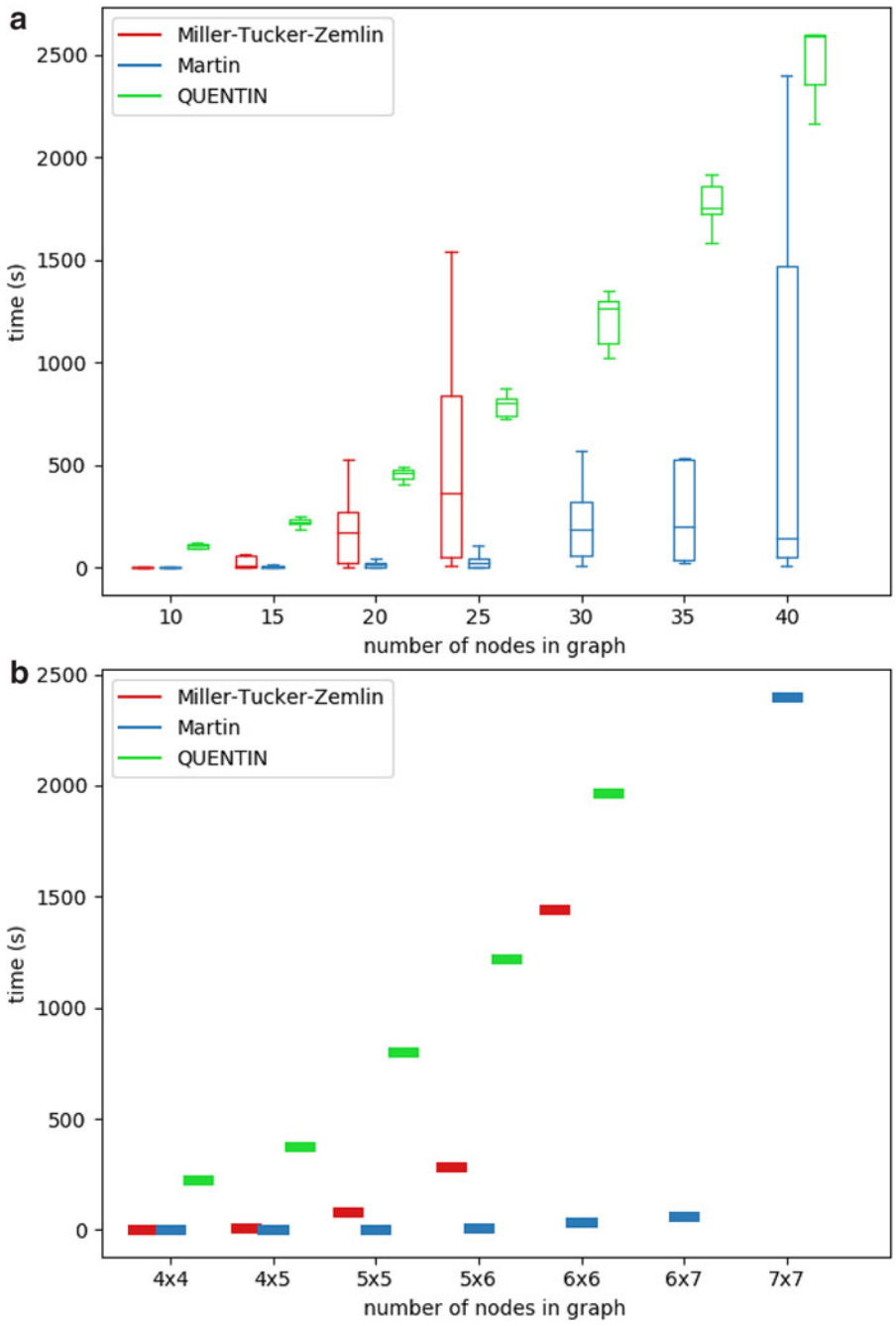

FIG. 3. Running times of ILP solver and QUENTIN on Erdős–Rényi graphs **(a)** and grids **(b).** ILP, integer linear programming.

The results of algorithms' comparison are shown in Figure 5. We measured algorithm accuracy by the proportion of correctly inferred transmission links and transmission ancestries (i.e., pairs, ancestor/descendant). *s*-SF-based ILP clearly outperformed the phylogenetic approach: the average transmission link detection accuracy was 82.44% for the former and 72.61% for the latter, while the average transmission ancestry detection accuracies were 97.48% and 73.96%, respectively.

5.2. Data from hepatitis C outbreaks

We applied the concept of scale-free spanning trees to the graphs arising from the benchmark data set consisting of several epidemiologically curated HCV outbreaks investigated by the CDC (Campo et al.,

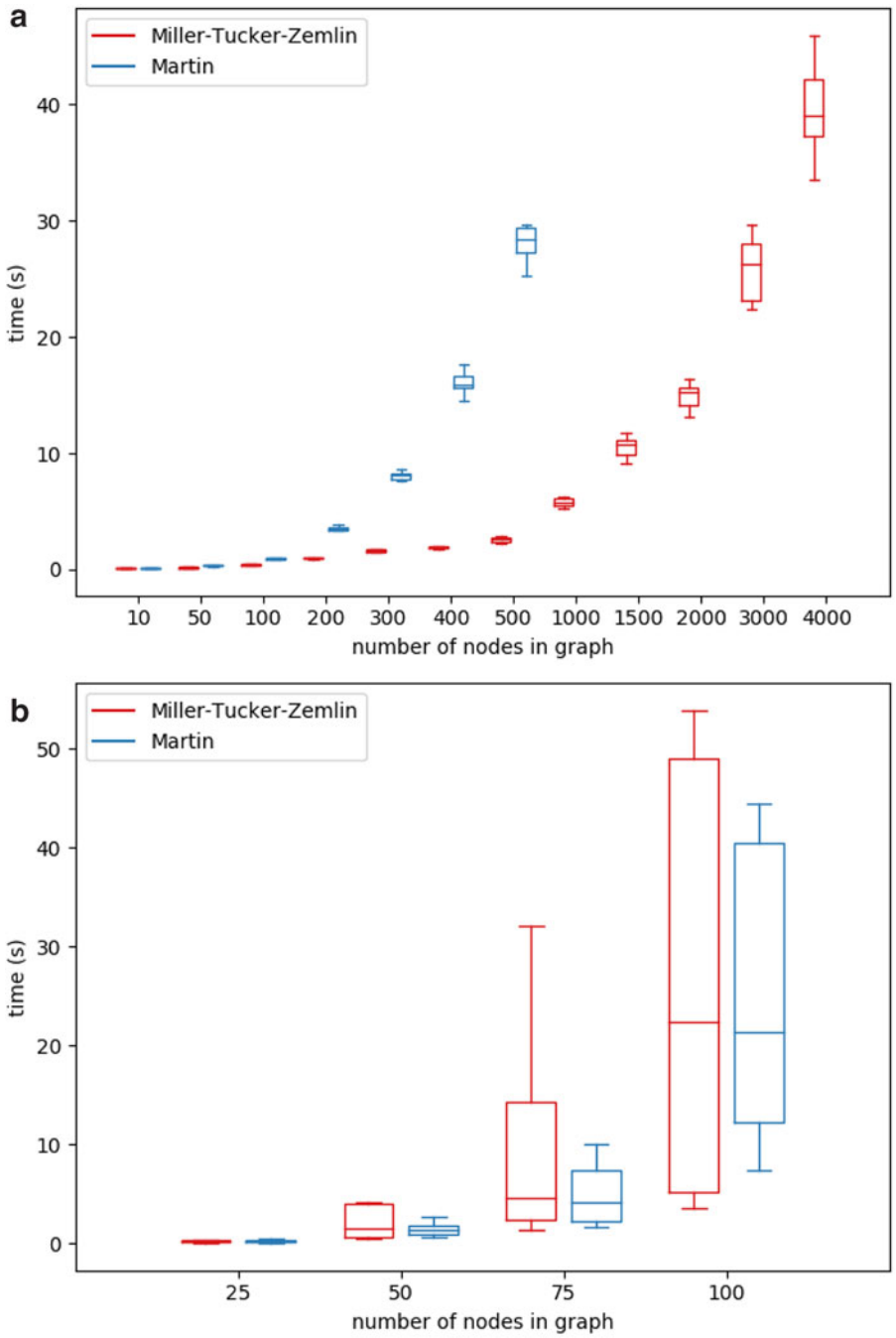

FIG. 4. Running times of ILP solver on Barabási–Albert **(a)** and NetworkX **(b)** scale-free graphs.

2016; Glebova et al., 2017; Skums et al., 2018). This data set comprises HCV quasispecies populations sampled from 81 infected individuals involved in 10 viral outbreaks. Each population consists of RNA sequences of HCV hypervariable region 1 (HVR1) of length 264 bp. Transmission histories of the outbreaks (''who infected whom'') are known as a result of epidemiological investigations. In this case, we are dealing with intrahost viral populations rather than single sequences, and therefore, we compared the proposed approach with QUENTIN, which has been specifically designed to handle such data (Skums et al., 2018).

For each outbreak, the genetic relatedness network G_R was constructed using the threshold-based approach suggested by Campo et al. (2016). The vertices of G_R are adjacent, if the *minimal* Hamming distance

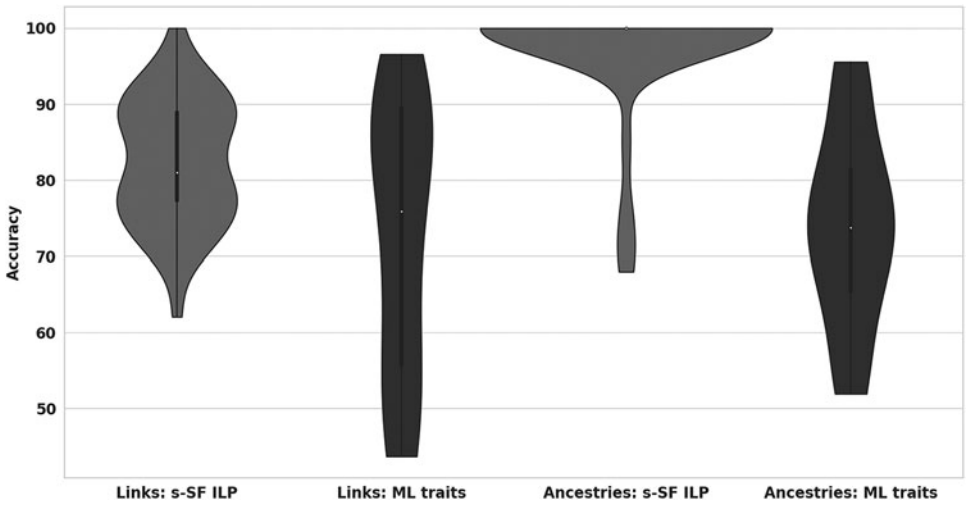

FIG. 5. Accuracy of *s*-SF ILP model compared with the phylogenetic trait inference algorithm.

between the sets of sequences sampled from these patients does not exceed the threshold t. The threshold value was estimated as described in Subsection 5.1.2. Next, the ILP algorithm for the Martin formulation has been applied to G_R. For all outbreaks, the ILP problem has been solved to optimality.

We tested the accuracy of inference of transmission links and identification of the superspreaders (the sources of majority of infections). The results are reported in Table 1. The superspreaders correspond to vertices of highest degrees in *s*-optimal and *m*-optimal trees for 9 out of 10 outbreaks. It should be noted that all algorithms incorrectly identified a superspreader for the same outbreak. It is the only outbreak where the virus was transmitted via a nonsocial interaction (namely, through blood transfusions), while all other outbreaks were associated with unsafe injection practices or sexual contacts. For those outbreaks, both ILP approaches correctly recovered 92% of transmission links and all ancestor/descendant pairs, thus outperforming QUENTIN.

6. DISCUSSION

In genomic epidemiology, reconstruction of viral transmission histories from genomic data is fundamental for the investigation of outbreaks and understanding of epidemic spread. Genomic analysis has become one of the major tools for the investigation of outbreaks and surveillance of transmission dynamics (Armstrong et al., 2019; Knyazev et al., 2020). Naturally, graphs are the primary models used in such studies (Wertheim et al., 2014; Campo et al., 2016; Ragonnet-Cronin et al., 2019). In many settings, graph-based methods have been shown to be more efficient to ascertain transmission links compared with methods based on binary phylogenies (Wertheim et al., 2014), as phylogenetic clades are not easily resolvable into transmission clusters and pairs (Lewis et al., 2008; Hughes et al., 2009; Kouyos et al., 2010), while the statistical support for a clade does not necessarily indicate the statistical support for a

Table 1. Results on Experimental Data with Different Models

	Evaluation metric		
Methods	*(A)*	*(B)*	*(C)*
QUENTIN	0.9	0.78	0.98
s-SF	0.9	0.92	1.0
m-SF	0.9	0.92	1.0

(A) Superspreader inference accuracy, (B) accuracy of transmission link inference, and (C) accuracy of transmission ancestry inference.

relationship between individual genomes inside a clade (Volz et al., 2012; Wertheim et al., 2014). However, in many cases, transmission links cannot be inferred using the genomic data alone (Jombart et al., 2014; Villandre et al., 2016). It leads to the need to introduce additional constraints on the reconstructed transmission networks or utilize more complicated objectives.

As a result, the associated algorithmic problems become harder. In this article, we studied one such problem—scale-free spanning tree problem—that arises in epidemiological studies of viruses whose spread is highly influenced by social networks of contacts between susceptible individuals. This includes HIV, HCV, and other pathogens transmitted through sexual contact or needle sharing. We demonstrated that this problem in its two possible algorithmic formulations is NP-hard, even if restricted to relatively simple graph classes. However, it admits an ILP formulation allowing to efficiently solve the problem for small-to-medium networks. It is often enough for the vast majority of outbreaks of HIV and HCV that involve dozens of infected individuals.

However, some outbreaks involve hundreds or even thousands of hosts, and in such cases, more scalable algorithmic solutions are needed. Thus, an important open problem is to establish whether constant or logarithmic approximation exists for the *m*-SF Spanning Tree and *s*-SF Spanning Tree problems. In this context, it would be interesting to explore the relationships between scale-free spanning tree problems and max-leaf spanning tree problems. The latter is a well-studied combinatorial problem (Griggs et al., 1989; Galbiati et al., 1994), which seems to be the closest to our problem. Indeed, both problems aim to find a ''star-like'' spanning tree; furthermore, several reduction schemes for the proof of NP-completeness used by us exploit this relationship. Importantly, Lu and Ravi (1998) and Reich (2016) showed that the max-leaf spanning tree problem is approximable within a constant factor. Although the problems are far from being equivalent, it may seem reasonable for future studies to try to adopt algorithmic machinery developed for the max-leaf spanning tree problem to the scale-free spanning tree problem.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

Y.O. was partially supported by the BRFFR grant (Project F20UKA-005). The work of V.K. and K.K. was supported by the German National Science Foundation via DFG-Research Training Group 2297 (Mathematical Complexity Reduction—MathCoRe). P.S. was supported by the National Institutes of Health grant 1R01EB025022 and by the National Science Foundation grant 2047828.

REFERENCES

Armstrong, G.L., MacCannell, D.R., Taylor, J., et al. 2019. Pathogen genomics in public health. *N. Engl. J. Med.* 381, 2569–2580.

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–512.

Bonsma, P. 2012. Max-leaves spanning tree is APX-hard for cubic graphs. *J. Discrete Algorithms.* 12, 14–23.

Borovicanin, B., Das, K.C., Furtula, B., et al. 2017. Bounds for Zagreb indices. *MATCH Commun. Math. Comput. Chem.* 78, 17–100.

Campbell, E.M., Jia, H., Shankar, A., et al. 2017. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. *J. Infect. Dis.* 216, 1053–1062.

Campo, D.S., Xia, G.-L., Dimitrova, Z., et al. 2016. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. *J. Infect. Dis.* 213, 957–965.

Chartrand, G., Lesniak, L., and Zhang, P. 2016. *Graphs & Digraphs*. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Das, K.C. 2003. Sharp bounds for the sum of the squares of the degrees of a graph. *Kragujev. J. Math.* 25, 19–41.

Das, K.C., and Gutman, I. 2004. Some properties of the second Zagreb index. *MATCH Commun. Math. Comput. Chem.* 52, 103–112.

de Caen, D. 1998. An upper bound on the sum of squares of degrees in a graph. *Discrete Math.* 185, 245–248.

De Maio, N., Wu, C.-H., and Wilson, D.J. 2016. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.* 12, e1005130.

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.

Galbiati, G., Maffioli, F., and Morzenti, A. 1994. A short note on the approximability of the maximum leaves spanning tree problem. *Inform. Process. Lett.* 52, 45–49.

Galvani, A.P., and May, R.M. 2005. Dimensions of superspreading. *Nature* 438, 293–295.

Garey, M.R., and Johnson, D.S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman & Co., New York, NY.

Glebova, O., Knyazev, S., Melnyk, A., et al. 2017. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genomics.* 18, 918.

Griggs, J.R., Kleitman, D.J., and Shastri, A. 1989. Spanning trees with many leaves in cubic graphs. *J Graph Theory* 13, 669–695.

Hagberg, A., Swart, P., and Chult, D. 2008. Exploring network structure, dynamics, and function using NetworkX. *In Proceedings of the 7th Python in Science Conference*. Online publication, Pasadena, California; pp. 11–16.

Hajarizadeh, B., Grebely, J., and Dore, G.J. 2013. Epidemiology and natural history of HCV infection. *Nat. Rev. Gastroenterol. Hepatol.* 10, 553–562.

Huang, C., Wang, Y., Li, X., et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506.

Hughes, G.J., Fearnhill, E., Dunn, D., et al. 2009. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathogens* 5, e1000590.

Icer Baykal, P., Lara, J., Khudyakov, Y., et al. 2020. Quantitative differences between intra-host HCV populations from persons with recently established and persistent infections. *Virus Evol*. 7, veaa103.

Jha, D., Skums, P., Zelikovsky, A., et al. 2017. Modeling the spread of HIV and HCV infections based on identification and characterization of high-risk communities using social media, 425–430. *In International Symposium on Bioinformatics Research and Applications*. Springer, Cham.

Jombart, T., Cori, A., Didelot, X., et al. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* 10, e1003457.

Jombart, T., Eggo, R., Dodd, P., et al. 2011. Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity* 106, 383–390.

Kilmarx, P.H. 2009. Global epidemiology of HIV. *Curr. Opin. HIV AIDS.* 4, 240–246.

Kincaid, R.K., Kunkler, S.J., Lamar, M.D., et al. 2016. Algorithms and complexity results for finding graphs with extremal Randić index. *Networks* 67, 338–347.

Kleitman, D.J., and West, D.B. 1991. Spanning trees with many leaves. *SIAM J Discrete Math.* 4, 99–106.

Klinkenberg, D., Backer, J.A., Didelot, X., et al. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* 13, e1005495.

Knyazev, S., Hughes, L., Skums, P., et al. 2020. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Brief. Bioinformatics* 22, 96–108.

Kosakovsky Pond, S.L., Weaver, S., Leigh Brown, A.J., et al. 2018. HIV-TRACE (TRansmission Cluster Engine): A tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Mol. Biol. Evol.* 35, 1812–1819.

Kouyos, R.D., Von Wyl, V., Yerly, S., et al. 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* 201, 1488–1497.

Kumar, S., Stecher, G., Li, M., et al. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549.

Leigh Brown, A.J., Lycett, S.J., Weinert, L., et al. 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* 204, 1463–1469.

Lemke, P. 1988. The maximum leaf spanning tree problem for cubic graphs is NP-complete. IMA Preprint Series No. 428.

Lewis, F., Hughes, G.J., Rambaut, A., et al. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5, e50.

Li, L., Alderson, D., Doyle, J. C., et al. 2005. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.* 2, 431–523.

Lu, H.I., and Ravi, R. 1998. Approximating maximum leaf spanning trees in almost linear time. *J. Algorithms.* 29, 132–141.

Martin, R.K. 1991. Using separation algorithms to generate mixed integer model reformulations. *Oper. Res. Lett.* 10, 119–128.

Miller, C.E., Tucker, A.W., and Zemlin, R.A. 1960. Integer programming formulation of traveling salesman problems. *J. ACM* 7, 326–329.

Newman, M. 2010. *Networks*. Oxford University Press, New York, NY.

Papadimitriou, C., and Yannakakis, M. 1991. Optimization, approximation, and complexity classes. *J. Comput. Syst. Sci.* 43, 425–440.

Ragonnet-Cronin, M., Hu, Y.W., Morris, S.R., et al. 2019. HIV transmission networks among transgender women in Los Angeles County, CA, USA: A phylogenetic analysis of surveillance data. *Lancet HIV.* 6, e164–e172.

Ramachandran, S., Thai, H., Forbi, J.C., et al. 2018. A large HCV transmission network enabled a fast-growing HIV outbreak in rural Indiana, 2015. *EBioMedicine* 37, 374–381.

Reich, A. 2016. Complexity of the maximum leaf spanning tree problem on planar and regular graphs. *Theoret.Comput. Sci.* 626, 134–143.

Sagulenko, P., Puller, V., and Neher, R.A. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* 4, vex042.

Skums, P., Zelikovsky, A., Singh, R., et al. 2018. QUENTIN: Reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics* 34, 163–170.

Sledzieski S., Zhang C., Mandoiu I., et al. 2019. TreeFix-TP: Phylogenetic error-correction for infectious disease transmission network inference. *bioRxiv*. 1:813931.

Villandre, L., Stephens, D.A., Labbe, A., et al. 2016. Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: Applications to HIV-1. *PLoS One* 11, e0148459.

Volz, E.M., Koopman, J.S., Ward, M.J., et al. 2012. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Comput.Biol.* 8, e1002552.

Wertheim, J.O., Leigh Brown, A.J., Hepler, N.L., et al. 2014. The global transmission network of HIV-1. *J. Infect. Dis.* 209, 304–313.

Address correspondence to:
Dr. Pavel Skums
Department of Computer Science
Georgia State University
1 Park Place
Atlanta, GA 30303
USA

E-mail: pskums@gsu.edu