Codes Over Trees

Lev Yohananov , *Graduate Student Member, IEEE*, and Eitan Yaakobi , *Senior Member, IEEE*

Abstract—In graph theory, a *tree* is one of the more popular families of graphs with a wide range of applications in computer science as well as many other related fields. While there are several distance measures over the set of all trees, we consider here the one which defines the so-called *tree distance*, defined by the minimum number of edit operations, of removing and adding edges, in order to change one tree into another. From a coding theoretic perspective, codes over the tree distance are used for the correction of edge erasures and errors. However, studying this distance measure is important for many other applications that use trees and properties on their locality and the number of neighbor trees. Under this paradigm, the largest size of code over trees with a prescribed minimum tree distance is investigated. Upper bounds on these codes as well as code constructions are presented. A significant part of our study is dedicated to the problem of calculating the size of the ball of trees of a given radius. These balls are not regular and thus we show that while the *star tree* has asymptotically the smallest size of the ball, the maximum is achieved for the *path tree*.

Index Terms—Codes over graphs, tree distance, Prüfer sequences, Cayley's formula, tree edit distance.

I. INTRODUCTION

IN GRAPH theory, a *tree* is a special case of a connected graph, which comprises of n labeled nodes and $n-1$ edges. Studying trees and their properties has been beneficial in numerous applications. For example, in signal processing, trees are used for the representation of waveforms [5]. In programming languages, trees are used as structures to describe restrictions in the language. Trees also represent collections of hierarchical text which are used in information retrieval. In cybersecurity applications trees are used to represent fingerprint patterns [19]. One of the biology applications includes the tree-matching algorithm to compare between trees in order to analyze multiple RNA secondary structures [30]. Trees are also used in the subgraph isomorphism problem which, among its very applications, is used for chemical substructure searching [3].

An important feature when studying trees is defining an appropriate distance function. Several distance measures over trees have been proposed in the literature. Among the many examples are the tree edit distance [31], top-down distance [28], alignment distance [12], isolated-subtree

distance [32], and bottom-up distance [34]. These distance measures are mostly characterized by adding, removing, and relabeling nodes and edges as well as counting differences between trees with a different number of nodes. One of the more common and widely used distance, which will be referred in this work as the *tree distance* [10], [21], considers the number of edit edge operations in order to transform one tree to another. Namely, given two labeled trees over n nodes, the tree distance is defined to be half of the minimum number of edges that are required to be removed and added in order to change one tree to another. This value is also equivalent to the difference between $n-1$ and the number of edges that the two trees share in common. Despite the popularity of this distance function, the knowledge of its characteristics and properties is quite limited. The goal of this art is to close on these gaps and study trees under the tree distance from a coding theory perspective. To the best of our knowledge, this direction has not been explored rigorously so far.

Motivated by the coding theory approach, in this work we apply the tree distance, which is a metric, to study *codes over trees* with a prescribed minimum tree distance. This family of codes can be used for the correction of edge erasures. There are several applications in which such codes can be used. For example, in data structures, a tree is a widely used abstract data type that simulates a hierarchical tree structure [6]. Such tree data structures store the information in nodes and use edges as pointers between them. There are numerous examples for such tree data structures including *abstract syntax trees* (AST), *parsing trees* and *binary search trees* (BST) [6], [13]. AST represent the abstract syntactic structure of source code written in a programming language, while each node of the tree denotes a construct occurring in the source code. Parsing trees represent the syntactic structure of a string according to some context-free grammar. BST trees store in each node a value greater than all the values in the node's left subtree and less than those in its right subtree. These tree data structures can be implemented such that each node stores a list of pointers to other nodes in the tree. Theoretically, such pointers might have wrong addresses, which affects the reliability of the data structure. By adding redundancy edges and nodes, codes over trees may correct the unexpected pointer mismatches. Another family of applications include data structures such as *tries* and *suffix trees* [13] in which the information is stored on the edges rather than the nodes. Such data structures can be implemented by a list of $n-1$ edges which is a list of node pairs together with the information on every edge. Again, theoretically, such an edge list may have failures that can indeed be corrected using classical error-correction codes. However, these codes will not be cardinality optimal since

they do not take advantage of the structure of the tree. For the binary case, using classical error-correction codes, we show in the paper the construction of codes over trees of size $\Omega(n^{n-2d})$ where $d \leqslant n/2$ corresponds to the minimum tree distance of the code. Using codes over trees we show that it is possible to construct codes of cardinality $\Omega(n^2)$, while the minimum tree distance d approaches $\lfloor 3n/4 \rfloor$ and n is a prime number.

Another interesting problem for the tree distance is the study of the size of balls according to the tree distance. This investigation is useful not only for applying the sphere packing bound on codes over trees, but also for other applications. For example, in [7] it was claimed that recent research on nanotechnology discovered that structures of DNA molecules can be constructed into trees or lattices, and that future synthesis techniques may use physical constraints to enforce tree structures on the written base. The authors of [7] introduced the *tree trace reconstruction problem*, in which the goal is to reconstruct a tree from several of its copies while each copy can have node deletions. In this case, the size of the tree balls may be useful. Another approach deals with graph matching, i.e., the problem of finding a similarity between graphs [9], [16]. Graph matching is an important tool used for example in computer vision and pattern recognition. One of the problems under this setup is to find a *model graph*, which represents the prototype symbol, as a subgraph in an *input graph* that represents a diagram, which is also called *subgraph isomorphism problem* [16]. If the model graph cannot be found exactly in the input graph, then the goal is to find a subgraph that is close to the model graph while the similarity is determined by edit operations on the nodes and edges. This problem is also studied for trees called *subtree isomorphism problem* [29], and the size of balls of trees may be useful for this problem. Lastly, one of the classical problems in graph theory is finding a minimum spanning tree (MST) for a given graph. While the MST problem is solved in polynomial time [15], [22], it may become NP-hard under some specific constraints. For example, in the degree-constrained MST problem (d-MST) [14], [23], [24], [35], it is required that the degree of every vertex in the MST is not greater than some fixed value d. In another example, the goal is to look for an MST in which the length of the simple path between every two vertices is bounded from above by a given value $D \geqslant 4$ [25]. One of the common approaches for solving such problems uses *evolution algorithms* (*EA*). Under this setup, the goal is to find a feasible tree to the problem by iteratively searching for a candidate tree. This iterative procedure is invoked by using *mutation operations* over the current tree in order to produce a new candidate tree. These mutation operations typically involve the modification of edges in the tree and as such are highly related to the tree distance. Thus, in order to analyze the complexity of such algorithms, it is necessary to study the size of the balls according to the tree distance. In fact, in [10] the size of the radius-one ball was computed for all trees with at most 20 vertices. According to this computer search, it was observed that the smallest size of the ball is achieved when the tree is a star tree (i.e., the tree has one node connected to all other nodes), while the largest for a path tree (i.e., the tree has two leaves and the degree

of all other nodes is two). In this article, we establish this result for any number of nodes in the tree as well as for any radius. Furthermore, it is shown that the size of the radius-t ball ranges between $\Omega(n^{2t})$ (for a star tree) and $\mathcal{O}(n^{3t})$ (for a path tree), while the average size of all balls is $\Theta(n^{2.5t})$.

This article is organized as follows. In Section II, we formally define the tree distance and codes over trees as well as several more useful definitions and properties for balls of trees, that will be defined in the sequel. An *edge erasure* is the event in which one of the edges in the tree is erased and a forest is received with two connected components. This is also extended to the erasure of multiple edges. If t edges are erased, then a forest with $t+1$ connected components is received and the number of such forests is $\binom{n-1}{t}$. In Section III we summarize all main results of the paper. In Section IV, by using several known results on the number of forests with a fixed number of connected components we are able to derive a sphere packing bound for codes over trees. More specifically, the size of codes over trees of minimum tree distance d cannot be greater than $\mathcal{O}(n^{n-d-1})$. In Section V, we study balls of trees. The *tree ball of trees* of a given tree T consists of all trees such that their tree distance from T is at most some fixed radius t. These balls are not regular. In this section, these balls are studied for radius one. Balls with a general radius are studied in Section VI. In Section VII, the size of star, path tree ball is presented, respectively. Lastly, in Section VIII, for a fixed d we show a construction of codes over trees of size $\Omega(n^{n-2d})$. It is also shown that it is possible to construct codes of cardinality $\Omega(n^2)$, while the minimum distance d approaches $\lfloor 3n/4 \rfloor$ and n is a prime number. Finally, Section IX concludes the paper.

II. DEFINITIONS AND PRELIMINARIES

Let $G = (V_n, E)$ be a graph, where $V_n = \{v_0, v_1, \ldots, v_{n-1}\}$ is a set of $n \geqslant 1$ labeled *nodes*, also called *vertices*, and $E \subseteq V_n \times V_n$ is its *edge* set. In this article, we only study undirected trees and forests. By a slight abuse of notation, every undirected edge in the graph will be denoted by $\langle v_i, v_j \rangle$ where the order in this pair does not matter, i.e., the notation $\langle v_i, v_j \rangle$ is identical to the notation $\langle v_j, v_i \rangle$. Thus, there are $\binom{n}{2}$ possibilities for the edges and the edge set is defined by

$$E_n = \{\langle v_i, v_j \rangle \mid i, j \in [n]\}, \tag{1}$$

where $[n] \triangleq \{0, 1, \ldots, n-1\}$.

A finite undirected *tree* over n nodes is a connected undirected graph with $n-1$ edges. The *degree* of a node v_i is the number of edges that are incident to the node, and will be denoted by $\deg(v_i)$. Each node of degree 1 is called a *leaf*. The set of all trees over n nodes will be denoted by $\mathbf{T}(n)$. An undirected graph that consists of only disjoint union of trees is called a *forest*. The set of all forests over n nodes with exactly δ trees will be denoted by $\mathbf{F}(n, \delta)$. Denote by $F(n, \delta)$ the size of $\mathbf{F}(n, \delta)$. We sometimes use the notation $\{C_0, C_1, \ldots, C_{\delta-1}\} = F \in \mathbf{F}(n, \delta)$ to explicitly denote a forest with t connected components (or subtrees) of F. Note that $\mathbf{F}(n, 1) = \mathbf{T}(n)$.

By Cayley's formula [1] it holds that $|\mathbf{T}(n)| = n^{n-2}$. The proof works by showing a bijection $\mathcal{F} : \mathbf{T}(n) \to [n]^{n-2}$, where for every tree $T \in \mathbf{T}(n)$, the *prüfer sequence* of T is denoted by $\mathcal{F}(T) = \boldsymbol{w}_T$. An important property is that for each $T = (V_n, E)$, the number of appearances of node $v_i \in V_n$ in $\boldsymbol{w}_T$ is equal to $\deg(v_i) - 1$.

Definition 1: A *code over trees* $\mathcal{C}_T$, denoted by $\mathcal{T}\text{-}(n, M)$, is a set of M trees over n nodes. Each tree in the code $\mathcal{C}_T$ is called a *codeword-tree*. The *redundancy* r of the code $\mathcal{C}_T$ is defined by $r = (n-2)\log(n) - \log(M)$[1].

Every codeword-tree corresponds to unique information that is stored, sent, or read, i.e., the information is the structure of the codeword-tree. The storage of information depends mainly on the application that will be used. For example, in binary search trees [6] the information values and pointers that represent edges are stored in nodes. In tries or suffix trees [13] the symbols or strings are stored on edges. In order to deal with *erasures* and *errors* of edges of trees, we initiate the study of codes over trees as will be defined next.

Definition 2: An *erasure of ρ edges* in a tree $T \in \mathbf{T}(n)$ is the event in which ρ of the edges in T are erased and T is separated into a forest of $\rho + 1$ connected components over n nodes. An *error of ψ edges* in a tree $T \in \mathbf{T}(n)$ is the event in which ψ of the edges in T are replaced with other ψ edges such that we receive a new tree $T' \in \mathbf{T}(n)$.

The tree distance for trees is next defined.

Definition 3: The *tree distance* between two trees $T_1 = (V_n, E_1)$ and $T_2 = (V_n, E_2)$ will be denoted by $d_{\mathcal{T}}(T_1, T_2)$ and is defined to be,

$$d_{\mathcal{T}}(T_1, T_2) = n - 1 - |E_1 \cap E_2|.$$

It is clear that $d_{\mathcal{T}}(T_1, T_2) = |E_1 \setminus E_2| = |E_2 \setminus E_1|$. Every tree over n nodes can be represented by a binary vector of length $\binom{n}{2}$ called the *characteristic vector*. Such a vector is indexed by all possible $\binom{n}{2}$ edges that the tree can have and it has ones only in the indices of the tree's edges. Using this representation, the tree distance between any two trees is one half the Hamming distance between their characteristic vectors. Thus, the tree distance is a metric as was mentioned in [21] and is stated in the next lemma.

Lemma 4: The tree distance is a metric.

The *tree distance* of a code over trees $\mathcal{C}_T$ is denoted by $d_{\mathcal{T}}(\mathcal{C}_T)$, which is the minimum tree distance between any two distinct trees in $\mathcal{C}_T$, that is,

$$d_{\mathcal{T}}(\mathcal{C}_T) = \min_{\substack{T_1 \neq T_2 \\ T_1, T_2 \in \mathcal{C}_T}} \{d_{\mathcal{T}}(T_1, T_2)\}.$$

Definition 5: A *code over trees* $\mathcal{C}_T$ of tree distance d, denoted by $\mathcal{T}\text{-}(n, M, d)$, has M trees over n nodes and its tree distance is $d_{\mathcal{T}}(\mathcal{C}_T) = d$.

Since the tree distance is a metric the following theorem holds straightforwardly.

Theorem 6: A $\mathcal{T}\text{-}(n, M)$ code over trees $\mathcal{C}_T$ is of tree distance at least d if and only if it can correct any $d - 1$ edge erasures and if and only if it can correct any $\lfloor (d-1)/2 \rfloor$ edge errors.

[1]The base of all logarithms in the paper is assumed to be 2.
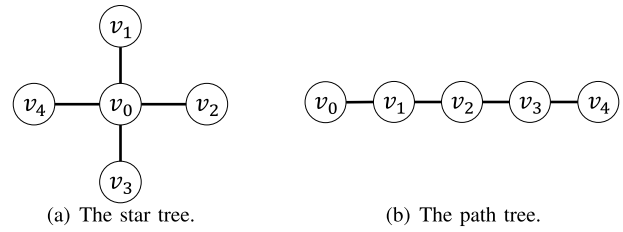

(a) The star tree. (b) The path tree.

Fig. 1. For $n = 5$ a star and a path trees are presented.

Next, we define the largest size of a code over trees with a prescribed tree distance.

Definition 7: The *largest size of a code over trees* with tree distance d is denoted by $A(n, d)$. The *minimum redundancy of a code over trees* will be defined by $r(n, d) = (n-2)\log(n) - \log(A(n, d))$.

A tree will be called a *star tree* (or a *star* in short) if it has a node $v_i, i \in [n]$ such that $\deg(v_i) = n - 1$, and all the other nodes $v_j, j \in [n], j \neq i$ satisfy $\deg(v_j) = 1$. A *path graph* or a *path tree* over n nodes is a graph whose nodes can be listed in the order $v_{i_0}, v_{i_1}, \ldots, v_{i_{n-1}}$, where $i_0, i_1, \ldots, i_{n-1} \in [n]$, such that its edges are $\langle v_{i_j}, v_{i_{j+1}} \rangle$ for all $j \in [n-1]$.

Definition 8: The *tree ball of a tree of radius t* in $\mathbf{T}(n)$ centered at $T \in \mathbf{T}(n)$ is defined to be

$$\mathcal{B}_T(n, t) = \{T' \in \mathbf{T}(n) \mid d_{\mathcal{T}}(T', T) \leqslant t\}.$$

The size of the tree ball of trees of T, $\mathcal{B}_T(n, t)$, is denoted by $V_T(n, t)$.

Note that $V_T(n, t)$ depends on the choice of its center T. For example, we will show that if T is a star then $V_T(n, 1) = (n-1)(n-2) + 1$ and if T is a path tree, then $V_T(n, 1) = (n-1)(n-2)(n+3)/6 + 1$. If T is a star, path tree the size of $V_T(n, t)$ is denoted by $V^\star(n, t), V^-(n, t)$, respectively. We define the *average ball size* of radius t to be the average value of all tree balls of trees of radius t, that is,

$$V(n, t) = \frac{\sum_{T \in \mathbf{T}(n)} V_T(n, t)}{n^{n-2}}.$$

Definition 9: The *sphere of radius $t \geqslant 0$ centered at $T \in \mathbf{T}(n)$* is defined to be

$$\mathcal{S}_T(n, t) = \mathcal{B}_T(n, t) \setminus \mathcal{B}_T(n, t-1),$$

where $\mathcal{S}_T(n, 0) = \mathcal{B}_T(n, 0) = \{T\}$, by definition. The size of the sphere of radius t is equal to the number of all trees in $\mathcal{S}_T(n, t)$ and is denoted by $S_T(n, t)$. If T is a star, path tree then we denote the sphere $\mathcal{S}_T(n, t)$ by $S^\star(n, t), S^-(n, t)$, respectively.

For each $T = (V_n, E) \in \mathbf{T}(n)$ and for each $E' \subseteq E, |E'| = t$, denote the forest $F_{T,E'} = (V_n, E \setminus E')$. Note that $F_{T,E'} \in \mathbf{F}(n, t+1)$.

Definition 10: The *forest ball of a tree $T = (V_n, E)$ of radius t* in $\mathbf{F}(n, t+1)$ is defined to be

$$\mathcal{P}_T(n, t) = \{F_{T,E'} \in \mathbf{F}(n, t+1) \mid E' \subseteq E, |E'| = t\}.$$

Given a tree $T = (V_n, E)$ and an edge-set $E' \in E, |E'| = t$, let $F_{T,E'} = (V_n, E \setminus E') \in \mathcal{P}_T(n, t)$ be the forest which is also denoted by $F_{T,E'} = \{C_0, C_1, \ldots, C_t\}$, such that $|C_0| \leqslant$

TABLE I

TABLE OF DEFINITIONS AND NOTATIONS

Notation	Meaning	Remarks
n	The number of nodes	Sec. II
$\mathbf{T}(n)$	The set of all labeled trees over n nodes	Sec. II
$\mathbf{F}(n,\delta)$	The set of all forests with δ connected components	Sec. II
$F(n,\delta)$	The size of $\mathbf{F}(n,\delta)$	Sec. II
d	The tree distance	Def. II
$\mathcal{T}\text{-}(n,M,d)$	A code over trees of size M	Def. 5
$A(n,d)$	The largest size of a $\mathcal{T}\text{-}(n,M,d)$ code	Def. 7
$r(n,d)$	The minimum redundancy of a $\mathcal{T}\text{-}(n,M,d)$ code	Def. 7
t	The radius of a ball	Def. 8
$\mathcal{B}_T(n,t)$	The tree ball of a tree of radius t centered at T	Def. 8
$V_T(n,t)$	The size of $\mathcal{B}_T(n,t)$	Def. 8
$V(n,t)$	The average ball size of radius t	Def. 8
$V^{\star}(n,t))$	The value of $V_T(n,t)$ if T is a star	Sec. II
$V^{-}(n,t)$	The value of $V_T(n,t)$ if T is a path tree	Sec. II
$\mathcal{S}_T(n,t)$	The sphere of a tree of radius t centered at T	Def. 9
$S_T(n,t)$	The size of $\mathcal{S}_T(n,t)$	Def. 9
$S^{\star}(n,t))$	The value of $S_T(n,t)$ if T is a star	Sec. II
$S^{-}(n,t)$	The value of $S_T(n,t)$ if T is a path tree	Sec. II
$\mathcal{P}_T(n,t)$	The forest ball of a tree of radius t centered at T	Def. 10
$P_T(n,t)$	The set of profiles of $\mathcal{P}_T(n,t)$	Def. 10
$\mathcal{B}_F(n,t)$	The tree ball of a forest of radius t centered at F	Def. 11
$V_F(n,t)$	The size of $\mathcal{B}_F(n,t)$	Def. 11

$|C_1| \leqslant \cdots \leqslant |C_t|$. The *profile vector of T and E'* is denoted by $\mathbf{P}_T(E') = (|C_0|, |C_1|, \ldots, |C_t|)$ and the multi-set $P_T(n,t)$ is given by

$$P_T(n,t) = \{\mathbf{P}_T(E') \mid E' \subseteq E, |E'| = t\}. \qquad (2)$$

It is can be verified that $|P_T(n,t)| = |\mathcal{P}_T(n,t)| = \binom{n-1}{t}$.

Definition 11: The *tree ball of a forest (or the forest's ball in short) of radius t centered at $F \in \mathbf{F}(n, t+1)$* is defined to be

$$\mathcal{B}_F(n,t) = \{T \in \mathbf{T}(n) \mid F \in \mathcal{P}_T(n,t)\}.$$

The size of the forest's ball of radius t is equal to the number of all trees in $\mathcal{B}_F(n,t)$ and is denoted by $V_F(n,t)$.

Notice that for every two distinct trees $T_1, T_2 \in \mathcal{B}_F(n,t)$ it holds that $d_{\mathcal{T}}(T_1, T_2) \leqslant t$. Note also that we have three different ball definitions, the forest ball of trees of Definition 10 denoted by $\mathcal{P}_T(n,t)$, the tree ball of trees of Definition 8, denoted by $\mathcal{B}_T(n,t)$, the forest's ball of Definition 11, denoted by $\mathcal{B}_F(n,t)$.

Furthermore, for the convenience of the reader, relevant notation and terminology referred to throughout the paper is summarized in Table I.

III. MAIN RESULTS

This section summarizes the main results in the paper. Theorem 12 states three main upper bounds which will be presented in Section IV. The first bound is a sphere packing bound that will be proved in Theorem 18. The second, third bound is an improved upper bound in case that $d = n-2, d = n-3$ that will be derived in Theorem 20, 23, respectively.

Theorem 12:

a) For all $n \geqslant 1$ and fixed d,

$$A(n,d) \leqslant F(n,d) / \binom{n-1}{d-1} = \mathcal{O}(n^{n-1-d}).$$

b) For all positive integers n, $A(n, n-2) \leqslant n$.

c) For all $n \geqslant 9$, $A(n, n-3) \leqslant n^2$.

While in Theorem 12 we obtained upper bounds on $A(n,d)$ using forest balls of trees, in Theorem 13 we show another approach to obtain both lower and upper bounds on codes over trees using tree balls of trees. For that, in Section V tree balls of trees of radius one are studied and the main results on these balls are summarized in the next theorem.

Theorem 13: a) For any $T \in \mathbf{T}(n)$,

$$V^{\star}(n,1) \leqslant V_T(n,1) \leqslant V^{-}(n,1).$$

b) For all $n \geqslant 1$, $V(n,1) \approx 0.5\sqrt{\frac{\pi}{2}} n^{2.5} = \Theta(n^{2.5})$.

c) For all $n \geqslant 1$, $V^{\star}(n,1) = \Theta(n^2)$, $V^{-}(n,1) = \Theta(n^3)$.

The first, second result of Theorem 13 is proved in Theorem 25, 28, respectively, while the third is deduced using Lemma 24. The reader can verify that the hardest part of this theorem is to prove, using inductive and recursive arguments, that $V_T(n,1) \leqslant V^{-}(n,1)$. In fact this is shown for arbitrary t in Section VII. By using the fact that $V^{\star}(n,1) = \Theta(n^2)$, the upper bound $A(n,3) = \mathcal{O}(n^{n-4})$ is concluded which is the same upper bound result as the sphere packing bound. Applying the *generalized Gilbert-Varshamov* bound[1] [33], while using the fact that $V(n,1) = \Theta(n^{2.5})$, it is then deduced that $A(n,2) = \Omega(n^{n-4.5})$. This bound is improved in Section VIII.

In Section VI similar results, summarized in the next theorem, of tree balls of trees with arbitrary radius are shown.

Theorem 14: For all $T \in \mathbf{T}(n)$ and fixed t, it holds that

a) $V_T(n,t) = \Omega(n^{2t})$, $V_T(n,t) = \mathcal{O}(n^{3t})$.

b) $V(n,t) = \Theta(n^{2.5t})$.

c) $V^{\star}(n,t) = \Theta(n^{2t})$ and $V^{-}(n,t) = \Theta(n^{3t})$.

The first result is shown in Theorem 35, the second is deduced in Corollary 38, and the third one is shown in Section VII as a result of Theorem 40 and Theorem 41. These results are derived from recursive formulas that calculate the size of the tree balls of trees of radius t.

Again, using the fact that $V^{\star}(n,t) = \Theta(n^{2t})$, it is deduced that for all $d = 2t+1$, $A(n,d) = \mathcal{O}(n^{n-1-d})$ which matches the upper bound results by the sphere packing bound. Applying the generalized Gilbert-Varshamov lower bound and using the fact that $V(n,t) = \Theta(n^{2.5t})$, it is also derived that for $d = t+1$, $A(n,d) = \Omega(n^{n-2-2.5(d-1)})$. This bound is also improved in Section VIII.

In Section VII, we study the sizes of tree balls of trees of stars and path trees for arbitrary radius. Our main contribution in this section is formulated in recursive formulas for the sizes of tree balls of trees for arbitrary trees. We then show upper and lower bounds on these formulas using the sizes of tree balls of trees of the star and path trees. We present these results in the following theorem.

Theorem 15: For all n and fixed t let

$$P = n^{t-1}\binom{n-1}{t}(n-t), \quad Q = n^{t-1}\binom{n+t}{2t+1}.$$

---

[1] Let X be a finite set with some distance function $d : X \times X \to \mathbb{N}$. Assume that the volume of every ball is $B_r(x) = \{y \in X | d(x,y) \leqslant r\}$. It was proved in [33] that if $\overline{\Delta}_r = \left(\sum_{x \in X} |B_r(x)|\right)/|X|$, then the generalized Gilbert-Varshamov bound asserts that there exists a code with minimum distance $r+1$ and of size at least $|X|/\overline{\Delta}_r$.

The following properties holds:

a)

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V^{\star}(n, t-i) = P.$$

b)

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V^{-}(n, t-i) = Q.$$

c) For all $T \in \mathbf{T}(n)$

$$P \leqslant \sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n, t-i) \leqslant Q.$$

The first result is deduced from Theorem 33 and is also shown in Equation (17). The second result is due to Theorem 41 and the last result can be found in the proof of Theorem 44. The reader will find out that the challenging part of this theorem is to prove that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n, t-i) \leqslant Q,$$

which is also used in order to prove that $V_T(n, 1) \leqslant V^{-}(n, 1)$. This section concludes with conjecturing that for fixed t and n large enough,

$$V^{\star}(n, t) \leqslant V_T(n, t) \leqslant V^{-}(n, t).$$

Lastly, in Section VIII we provide several constructions that improve upon the generalized Gilbert-Varshamov lower bounds. The results of these constructions are summarized in the following theorem.

Theorem 16: It holds that

a) There exists an $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ code.
b) There exists an $\mathcal{T}$-$(n, n, n-2)$ code.
c) For any positive integer $d \leqslant n/2$, there exists an $\mathcal{T}$-(n, M, d) code such that $M = \Omega(n^{n-2d})$.
d) For fixed m and prime n, there exists an $\mathcal{T}$-$(n, \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor, \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2)$ code.

The result in a) is proved in Theorem 46 using Construction 1, the result in b) is due to Theorem 47 and Construction 2, the result in c) holds according to Corollary 49 and Construction 3, and the result in d) follows from Construction 4 and is proved in Theorem 53. The result in d) assures that it is possible to construct codes of cardinality $\Omega(n^2)$, while the minimum distance d approaches $\lfloor 3n/4 \rfloor$ and n is a prime number. Comparing to Theorem 23 in which it was shown that $A(n, n-3) = \mathcal{O}(n^2)$, the result of Theorem 53 shows that $A(n, d) = \Omega(n^2)$, when d approaches $\lfloor 3n/4 \rfloor$ and n is prime. Thus, finding the range of values of d for which $A(n, d) = \Theta(n^2)$ is left for future work.

IV. UPPER BOUNDS ON CODES OVER TREES

In this section we show upper bounds for codes over trees. Remember that $F(n, \delta)$ is the size of $\mathbf{F}(n, \delta)$, i.e., the number of forests with n nodes and δ connected components. The value of $F(n, \delta)$ was shown in [18], to be

$$F(n, \delta) = \binom{n}{\delta} n^{n-\delta-1} \sum_{i=0}^{\delta} \left((-\frac{1}{2})^i \binom{\delta}{i} \frac{(\delta+i)(n-\delta)!}{n^i (n-\delta-i)!} \right)$$

or another representation of it in [4],

$$F(n, \delta) = n^{n-\delta} \left(\sum_{i=0}^{\delta} (-\frac{1}{2})^i \binom{\delta}{i} \binom{n-1}{\delta-1+i} \frac{(\delta+i)!}{n^i \delta!} \right).$$

The next corollary summarizes some of these known results.

Corollary 17: The following properties hold for all n.

a) $F(n, 1) = n^{n-2}$,
b) $F(n, 2) = \frac{1}{2} n^{n-4}(n-1)(n+6)$,
c) $F(n, 3) = \frac{1}{8} n^{n-6}(n-1)(n-2)(n^2 + 13n + 60)$,
d) $F(n, n-4) = \frac{1}{16} \binom{n}{4}(n^2 + 3n + 10)(n-4)(n+3)$,
e) $F(n, n-3) = \frac{1}{2} \binom{n}{4}(n^2 + 3n + 4)$,
f) $F(n, n-2) = 3\binom{n+1}{4}$,
g) $F(n, n-1) = \binom{n}{2}$,
h) $F(n, n) = 1$.

A. Sphere-Packing Bound

The following theorem proves the sphere packing bound for codes over trees.

Theorem 18: For all $n \geqslant 1$ and $1 \leqslant d \leqslant n$, it holds that $A(n, d) \leqslant F(n, d)/\binom{n-1}{d-1}$.

Proof: Let $\mathcal{C}_\mathcal{T}$ be a $\mathcal{T}$-(n, M, d) code such that $n \geqslant 1$ and $1 \leqslant d \leqslant n$. Using Theorem 6, it is deduced that given a codeword-tree T_1, each $d-1$ of its edge erasures can be corrected. Thus, every forest F in the forest ball of trees $\mathcal{P}_{T_1}(n, d-1)$ cannot appear in any other forest ball of trees $\mathcal{P}_{T_2}(n, d-1)$, for all $T_2 \in \mathcal{C}_\mathcal{T} \setminus \{T_1\}$. Thus, for every two distinct codeword-trees $T_1, T_2 \in \mathcal{C}_\mathcal{T}$ it holds that

$$\mathcal{P}_{T_1}(n, d-1) \cap \mathcal{P}_{T_2}(n, d-1) = \emptyset.$$

As already mentioned, for all $T = (V_n, E)$ it holds that $|\mathcal{P}_T(n, d-1)| = \binom{n-1}{d-1}$. Therefore,

$$M \cdot \binom{n-1}{d-1} = M \cdot |\mathcal{P}_T(n, d-1)| \leqslant F(n, d),$$

which leads to the fact that

$$A(n, d) \leqslant \frac{F(n, d)}{\binom{n-1}{d-1}}.$$

$\blacksquare$

It was also proved in [18] that for any fixed δ,

$$\lim_{n \to \infty} \frac{F(n, \delta)}{n^{n-2}} = \frac{1}{2^{\delta-1}(\delta-1)!},$$

which immediately implies the following corollary.

Corollary 19: For all $n \geqslant 1$ and fixed d, it holds that

$$A(n, d) \leqslant F(n, d)/\binom{n-1}{d-1} = \mathcal{O}(n^{n-1-d}),$$

and thus $r(n,d) = (d-1)\log(n) + \mathcal{O}(1)$.

Notice that by Corollary 17(g) it holds that

$$A(n, n-1) \leqslant \binom{n}{2}/(n-1) = n/2. \tag{3}$$

In Section VIII we will show that

$$A(n, n-1) = \lfloor n/2 \rfloor,$$

by showing a construction of a $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ code over trees for all $n \geqslant 1$. Similarly, by Corollary 17(f),

$$A(n, n-2) \leqslant 3\binom{n+1}{4}/\binom{n-1}{n-3} = \frac{1}{2}\binom{n+1}{2}, \tag{4}$$

however, we will next show how to improve this bound such that $A(n, n-2) \leqslant n$. In Section VIII, a construction of $\mathcal{T}$-$(n, n, n-2)$ codes over trees will be shown, leading to $A(n, n-2) = n$. Finally, by Corollary 17(e),

$$\begin{aligned} A(n, n-3) &\leqslant \frac{1}{2}\binom{n}{4}(n^2 + 3n + 4)/\binom{n-1}{n-4} \\ &= \frac{1}{8}n(n^2 + 3n + 4), \end{aligned} \tag{5}$$

where a better upper bound will be shown in the sequel, which improves this bound to be $A(n, n-3) \leqslant 1.5n^2$. Finding a construction for this case is left for future work.

Before we show the improved upper bound for $A(n, n-3)$, a few more definitions are presented. The girth of a graph is the length of a shortest cycle contained in the graph. If the graph does not contain any cycles (i.e. it is an acyclic graph), its girth is defined to be infinity. For a positive integer n, let E_n be the set of all $\binom{n}{2}$ edges as defined in (1). A graph $G = (U \cup V, \mathcal{E})$ is a *bipartite graph* with two sets of nodes U and V such that $U \cap V = \emptyset$ and every edge connects a vertex from U to a vertex from V, i.e., $\mathcal{E} \subseteq U \times V$. Reiman's inequality in [20] and [26] states that if $|V| \leqslant |U|$, then every bipartite graph $G = (U \cup V, \mathcal{E})$ with girth at least 6 satisfies

$$|\mathcal{E}|^2 - |U| \cdot |\mathcal{E}| - |V| \cdot |U| \cdot (|V| - 1) \leqslant 0. \tag{6}$$

B. An Improved Upper Bound for $A(n, n-2)$

According to Theorem 18, $A(n, n-2) \leqslant \frac{1}{2}\binom{n+1}{2}$ and in the next theorem this bound will be improved to be $A(n, n-2) \leqslant n$.

Theorem 20: For all positive integers n, $A(n, n-2) \leqslant n$.

Proof: Let $\mathcal{C}_{\mathcal{T}}$ be a $\mathcal{T}$-$(n, M, n-2)$ code. Let $G = (U \cup V, \mathcal{E})$ be a bipartite graph such that $V = \mathcal{C}_{\mathcal{T}}, U = E_n$ (defined in (1)) and $(T, e) \in \mathcal{E}$ if and only if the tree $T \in \mathcal{C}_{\mathcal{T}}$ has the edge $e \in E_n$. Clearly, $|V| = M, |U| = \binom{n}{2}$ and $|\mathcal{E}| = M(n-1)$.

Since $\mathcal{C}_{\mathcal{T}}$ is a $\mathcal{T}$-$(n, M, n-2)_{\mathcal{D}}$ code it holds that for all $T_1 = (V_n, E_1), T_2 = (V_n, E_2) \in \mathcal{C}_{\mathcal{T}}, |E_1 \cap E_2| \leqslant 1$. That is, there are no two codeword-trees in $\mathcal{C}_{\mathcal{T}}$ that share the same two edges. Hence, there does not exist a cycle of length four in G. If the girth of G is at least 6 (including the case in which G is acyclic by definition of the girth), by (4), for all $n \geqslant 3$, it holds that $|V| = M \leqslant \frac{1}{2}\binom{n+1}{2} \leqslant \binom{n}{2} = |U|$, so the inequality
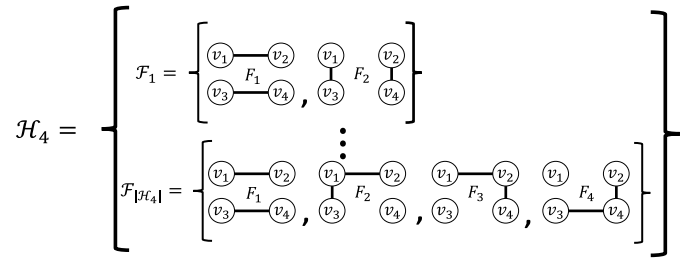

Fig. 2. An example of the $\mathcal{H}_4$ set. Given a forest-set $\mathcal{F} \in \mathcal{H}_4$, every two forests $F_1 = (V_n, E_1) \in \mathcal{F}, F_2 = (V_n, E_2) \in \mathcal{F}$ hold $|E_1 \cap E_2| \leqslant 1$.

stated in (6) will be used next. Since $|V| = M, |U| = \binom{n}{2}$ and $|\mathcal{E}| = M(n-1)$,

$$M^2(n-1)^2 - \binom{n}{2}M(n-1) - M\binom{n}{2}(M-1) \leqslant 0,$$

or equivalently

$$M(n-1) - \frac{n}{2}(M-1) \leqslant \binom{n}{2},$$

which is equivalent to

$$M(\frac{n}{2} - 1) \leqslant \binom{n}{2} - \frac{n}{2},$$

and since

$$\frac{\binom{n}{2} - \frac{n}{2}}{(\frac{n}{2} - 1)} = \frac{\frac{n}{2}(n-1) - \frac{n}{2}}{(\frac{n}{2} - 1)} = n\frac{(\frac{n}{2} - 1)}{(\frac{n}{2} - 1)} = n,$$

we deduce that $M \leqslant n$. ∎

As mentioned above, in Section VIII we will show that $A(n, n-2) = n$.

C. An Improved Upper Bound for $A(n, n-3)$

We showed in (5) that $A(n, n-3) \leqslant \frac{1}{8}n(n^2 + 3n + 4) = \mathcal{O}(n^3)$. In this section this bound will be improved by proving that $A(n, n-3) \leqslant n^2$.

Denote by $\mathcal{H}_n$ the set of forest-sets

$$\mathcal{H}_n = \left\{ \mathcal{F} \subseteq \mathbf{F}(n, 2) \, \middle| \, \begin{matrix} \forall F_1 = (V_n, E_1), F_2 = (V_n, E_2) \in \mathcal{F}, \\ |E_1 \cap E_2| \leqslant 1 \end{matrix} \right\}.$$

Example 1: For $n = 4$ we partially show an example of the forest-sets in $\mathcal{H}_4$.

We start with showing the following lemma.

Lemma 21: For $n \geqslant 9$ and for all $\mathcal{F} \in \mathcal{H}_n$ it holds that $|\mathcal{F}| \leqslant 2n$.

Proof: Let $\mathcal{F}$ be a forest-set in $\mathcal{H}_n$, and let $G = (U \cup V, \mathcal{E})$ be a bipartite graph such that $V = \mathcal{F}, U = E_n$ and $(F, e) \in \mathcal{E}$ if and only if the forest $F \in \mathcal{F}$ has the edge $e \in E_n$. Clearly $|V| = |\mathcal{F}|, |U| = \binom{n}{2}$, and $|\mathcal{E}| = |\mathcal{F}|(n-2)$. Note that G does not have girth 4 since for all $F_1 = (V_n, E_1), F_2 = (V_n, E_2) \in \mathcal{F}$ it holds that $|E_1 \cap E_2| \leqslant 1$.

Assume that the girth of G is at least 6. We consider the following two cases regarding the sizes of the V and U. In the first case, where $|V| \leqslant |U|$ we receive the bound stated in the lemma and we will show that the latter case cannot hold.

Case 1: Assume that $|V| = |\mathcal{F}| \leqslant \binom{n}{2} = |U|$. By (6)

$$|\mathcal{F}|^2(n-2)^2 - \binom{n}{2}|\mathcal{F}|(n-2) - |\mathcal{F}|\binom{n}{2}(|\mathcal{F}|-1) \leqslant 0,$$

or equivalently

$$|\mathcal{F}|(n-2)^2 - \binom{n}{2}(|\mathcal{F}|-1) \leqslant \binom{n}{2}(n-2),$$

which is equivalent to

$$|\mathcal{F}|\left((n-2)^2 - \binom{n}{2}\right) \leqslant \binom{n}{2}(n-2) - \binom{n}{2}.$$

Next it is deduced that

$$|\mathcal{F}|\left((n-2)^2 - \binom{n}{2}\right) \leqslant \binom{n}{2}(n-3),$$

which is equivalent to

$$|\mathcal{F}| \leqslant \frac{n^3 - 4n^2 + 3n}{n^2 - 7n + 8},$$

and therefore $|\mathcal{F}| \leqslant 2n$ for all $n \geqslant 9$.

Case 2: Assume that $|V| = |\mathcal{F}| > \binom{n}{2} = |U|$. Again, since the girth is at least six we have that

$$|\mathcal{F}|^2(n-2)^2 - |\mathcal{F}|^2(n-2) - \binom{n}{2}|\mathcal{F}|(\binom{n}{2}-1) \leqslant 0,$$

or equivalently

$$|\mathcal{F}|(n-2)^2 - |\mathcal{F}|(n-2) \leqslant \binom{n}{2}(\binom{n}{2}-1),$$

which is equivalent to

$$|\mathcal{F}|(n-3) \leqslant \frac{\binom{n}{2}(\binom{n}{2}-1)}{(n-2)}.$$

Hence for all $n \geqslant 9$

$$|\mathcal{F}| \leqslant \frac{n(n^2-1)}{4(n-3)} \leqslant \binom{n}{2},$$

which results with a contradiction. ∎

Let $\mathcal{C}_T$ be a $\mathcal{T}$-$(n, M, n-3)$ code. For all $e \in E_n$, denote by $c(\mathcal{C}_T, e)$ the number of codeword-trees of $\mathcal{C}_T$ having the edge e.

Lemma 22: Let $\mathcal{C}_T$ be a $\mathcal{T}$-$(n, M, n-3)$ code, where $n \geqslant 9$. Then, for all $e \in E_n$ it holds that $c(\mathcal{C}_T, e) \leqslant 2n$.

Proof: For $e \in E_n$, denote $k = c(\mathcal{C}_T, e)$ and let $T_0 = (V_n, E_0), T_1 = (V_n, E_1), \ldots, T_{k-1} = (V_n, E_{k-1}) \in \mathcal{C}_T$ be the k codeword-trees such that

$$e \in \bigcap_{i \in [k]} E_i. \qquad (7)$$

Denote by $\mathcal{F} \subseteq \mathbf{F}(n, 2)$ the set of k different forests received by removing the edge e from $T_0, T_1, \ldots, T_{k-1}$. Notice that since $\mathcal{C}_T$ is a $\mathcal{T}$-$(n, M, n-3)$ code it holds that $|E_i \cap E_j| \leqslant 2, i, j \in [k]$ and by (7) we deduce that for all distinct $F_i = (V_n, \mathcal{E}_i), F_j = (V_n, \mathcal{E}_j) \in \mathcal{F}, |\mathcal{E}_i \cap \mathcal{E}_j| \leqslant 1$. By Lemma 21, for all $n \geqslant 9, k = |\mathcal{F}| \leqslant 2n$ which leads to the fact that $c(\mathcal{C}_T, e) \leqslant 2n$. ∎

Lastly, the main result for this section is shown.

Theorem 23: For all $n \geqslant 9$, $A(n, n-3) \leqslant n^2$.

Proof: Let $n \geqslant 9$ and let $\mathcal{C}_T$ be a $\mathcal{T}$-$(n, M, n-3)$ code over trees. Since for all $e \in E_n$, $c(\mathcal{C}_T, e)$ is the number of codeword-trees of $\mathcal{C}_T$ having the edge e, we deduce that $\sum_{e \in E_n} c(\mathcal{C}_T, e) = M(n-1)$. By Lemma 22, for all $e \in E_n$, $c(\mathcal{C}_T, e) \leqslant 2n$. Therefore,

$$M(n-1) = \sum_{e \in E_n} c(\mathcal{C}_T, e) \leqslant \binom{n}{2} \cdot 2n = n^2(n-1),$$

and therefore, $M \leqslant n^2$. ∎

Lastly, we verified that for $4 \leqslant n \leqslant 8$, it holds that $A(n, n-3) \leqslant 1.5n^2$.

V. BALLS OF TREES OF RADIUS ONE

In previous section we introduced and studied the forest ball of a tree in order to derive a sphere packing bound on codes over trees with a prescribed minimum tree distance. In this section we study the size behavior of tree balls of trees. These results will also be used to apply the generalized Gilbert Varshamov bound [33] on codes over trees. We start from some definitions.

Our main goal in this section is to study the size of the radius-one tree ball of trees for all trees. This result is proved in the next lemma.

Lemma 24: For any $T \in \mathbf{T}(n)$ it holds that

$$V_T(n, 1) = \sum_{(i, n-i) \in P_T(n, 1)} \left(i(n-i) - 1\right) + 1. \qquad (8)$$

Proof: Let $T = (V_n, E) \in \mathbf{T}(n)$. For any tree $T' = (V_n, E') \in \mathcal{B}_T(n, 1) \setminus \{T\}$, if $e \in E \setminus E'$ and $e' \in E' \setminus E$, then T' is generated uniquely by removing an edge e from E, yielding two connected components (subtrees) $\{C_0, C_1\} \in P_T(n, 1), |C_0| \leqslant |C_1|$, and adding the edge $e' \neq e$ between C_0 and C_1. Thus,

$$|\mathcal{B}_T(n, 1) \setminus \{T\}| = \sum_{(|C_0|, |C_1|) \in P_T(n, 1)} \left(|C_0||C_1| - 1\right).$$

By denoting $|C_0| = i$ and $|C_1| = n-i$,

$$V_T(n, 1) = \sum_{(i, n-i) \in P_T(n, 1)} \left(i(n-i) - 1\right) + 1.$$

∎

Note that if T is a star, then

$$P_T(n, 1) = \left\{ \underbrace{(1, n-1), \ldots, (1, n-1)}_{n-1 \text{ times}} \right\}.$$

Therefore,

$$V^\star(n, 1) = \sum_{(1, n-1) \in P_T(n, 1)} \left(1 \cdot (n-1) - 1\right) + 1$$
$$= (n-1)(n-2) + 1.$$

If T is a path tree, for odd n,

$$P_T(n, 1) = \left\{ (i, n-i), (i, n-i) \mid 1 \leqslant i \leqslant \frac{n-1}{2} \right\},$$

and for even n,

$$P_T(n,1) = \left\{ (i, n-i), (i, n-i) \mid 1 \leqslant i \leqslant \frac{n-2}{2} \right\}$$
$$\cup \{(n/2, n/2)\}.$$

In both cases,

$$
\begin{aligned}
V^-(n,1) &= \sum_{i=1}^{n-1} \Big(i \cdot (n-i) - 1 \Big) + 1 \\
&= \sum_{i=1}^{n-1} \Big(i \cdot (n-i) \Big) - (n-1) + 1 \\
&\overset{(a)}{=} \binom{n+1}{3} - (n-1) + 1 \\
&= (n+1)n(n-1)/6 - 6(n-1)/6 + 1 \\
&= (n-1)(n^2 + n - 6)/6 + 1 \\
&= (n-1)(n-2)(n+3)/6 + 1,
\end{aligned}
$$

where (a) and its general case is shown in the proof of Theorem 41.

Our next goal is to show that for any $T \in \mathbf{T}(n)$ it holds that

$$V^\star(n,1) \leqslant V_T(n,1) \leqslant V^-(n,1).$$

The following claim is easily proved.

Claim 1: Given positive integers i, n such that $i \in [n]$, it holds that $n - 1 \leqslant i(n-i)$.

Next we state that for all $T \in \mathbf{T}(n)$,

$$\sum_{(i,n-i) \in P_T(n,1)} i(n-i) \leqslant \binom{n+1}{3}, \qquad (9)$$

while the proof will be shown in the general case in Lemma 43 where more than one edge is erased.

Theorem 25: For any $T \in \mathbf{T}(n)$ it holds that

$$V^\star(n,1) \leqslant V_T(n,1) \leqslant V^-(n,1).$$

Proof: First we prove the lower bound. For all $T \in \mathbf{T}(n)$

$$
\begin{aligned}
V_T(n,1) &= \sum_{(i,n-i) \in P_T(n,1)} \Big(i \cdot (n-i) - 1 \Big) + 1 \\
&\geqslant \sum_{(i,n-i) \in P_T(n,1)} \Big(1 \cdot (n-1) - 1 \Big) + 1 \\
&= (n-1)(n-2) + 1 = V^\star(n,1),
\end{aligned}
$$

where the inequality holds due to Claim 1. Next, due to (9),

$$
\begin{aligned}
V_T(n,1) &= \sum_{(i,n-i) \in P_T(n,1)} \Big(i \cdot (n-i) - 1 \Big) + 1 \\
&= \sum_{(i,n-i) \in P_T(n,1)} \Big(i \cdot (n-i) \Big) - (n-1) + 1 \\
&\leqslant \binom{n+1}{3} - (n-1) + 1 \\
&= (n-1)(n-2)(n+3)/6 + 1 = V^-(n,1),
\end{aligned}
$$

which leads to the fact that $V_T(n,1) \leqslant V^-(n,1)$. ∎

Our next goal is to show an approximation for the average ball of radius one, that is, the value $V(n,1)$. The first step in this calculation is established in the next lemma, where its proof can be found in Appendix A.

Lemma 26: For a positive integer n it holds that

$$\sum_{T \in \mathbf{T}(n)} V_T(n,1) = \sum_{F \in \mathbf{F}(n,2)} (V_F(n,1))^2 - (n-2)n^{n-2}.$$

In proof of Lemma 26 we use the equality

$$\sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} 1 = \sum_{F \in \mathbf{F}(n,2)} \sum_{T \in \mathcal{B}_F(n,1)} 1. \qquad (10)$$

which holds by changing the order of summation of all distinct couples of trees and forests. One can check that (10) is true also for $t > 1$, and we will use it in Lemma 36 which is in the next section. Notice also that from this equality it is deduced that

$$\sum_{F \in \mathbf{F}(n,t+1)} V_F(n,t) = \binom{n-1}{t} n^{n-2}.$$

Now, we are ready to show the following theorem.

Theorem 27: For all n,

$$\sum_{T \in \mathbf{T}(n)} V_T(n,1) = \frac{1}{2} n! \sum_{k=0}^{n-2} \frac{n^k}{k!} - (n-2)n^{n-2}.$$

Proof: It was shown in [18] that

$$F(n,2) = \frac{1}{2} \sum_{i=1}^{n-1} \binom{n}{i} i^{i-2}(n-i)^{n-i-2},$$

where i and $n - i$ represent the sizes of two connected components of each forest in $\mathbf{F}(n,2)$. Furthermore, since for all $\{C_0, C_1\} = F \in \mathbf{F}(n,2)$, if $|C_0| = i$ then $V_F(n,1) = i(n-i)$, it is deduced that,

$$
\begin{aligned}
\sum_{F \in \mathbf{F}(n,2)} (V_F(n,1))^2 &= \frac{1}{2} \sum_{i=1}^{n-1} \binom{n}{i} i^{i-2}(n-i)^{n-i-2} [i(n-i)]^2 \\
&= \frac{1}{2} \sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i} \overset{(a)}{=} \frac{1}{2} n! \sum_{k=0}^{n-2} \frac{n^k}{k!},
\end{aligned}
$$

where (a) holds according to Theorem 5.1 in [2]. Using Lemma 26 it is deduced that

$$\sum_{T \in \mathbf{T}(n)} V_T(n,1) = \frac{1}{2} n! \sum_{k=0}^{n-2} \frac{n^k}{k!} - (n-2)n^{n-2}.$$
∎

For two functions $f(n)$ and $g(n)$ we say that $f(n) \approx g(n)$ if $\lim_{n \to \infty} \frac{f(n)}{g(n)} = 1$. As a direct result of Theorem 27 the next corollary follows.

Corollary 28: It holds that,

$$V(n,1) \approx 0.5 \sqrt{\frac{\pi}{2}} n^{2.5}.$$

Proof: It was shown in [8] that

$$n! \sum_{k=0}^{n-2} \frac{n^k}{k!} \approx \sqrt{\frac{\pi}{2}} n^{n+0.5},$$

and therefore,

$$V(n,1) = \frac{\sum_{T \in \mathbf{T}(n)} V_T(n,1)}{n^{n-2}} \approx \frac{1}{2}\sqrt{\frac{\pi}{2}} n^{n+0.5-(n-2)}$$
$$= \frac{1}{2}\sqrt{\frac{\pi}{2}} n^{2.5}.$$

∎

To summarize the results of this section, we proved that for every $T \in \mathbf{T}(n)$ it holds that $V_T(n,1) = \Omega(n^2)$, $V_T(n,1) = \mathcal{O}(n^3)$ and the average ball size satisfies $V(n,1) = \Theta(n^{2.5})$. In order to apply the sphere packing bound for the tree balls of trees of radius one, we can only use the lower bound $V_T(n,1) = \Omega(n^2)$ and get that

$$A(n,3) \leqslant \frac{n^{n-2}}{\alpha n^2} = \frac{1}{\alpha} n^{n-4},$$

for some constant α. This bound is equivalent in its order to the one achieved in Corollary 19.

While we could not use the average ball size in applying the sphere packing bound, this can be done for the generalized Gilbert-Varshamov lower bound [33]. Intuitively, it is done by dividing the number of all trees over n nodes by the average ball size. Namely, according to [33], the following lower bound on $A(n,2)$ holds

$$A(n,2) = \Omega(n^{n-2-2.5}) = \Omega(n^{n-4.5}).$$

The reader can find Construction 3 in Section VIII for codes over trees with tree distance d and cardinality $\Omega(n^{n-2d})$. In case that $d=2$ the cardinality is

$$\Omega(n^{n-2d}) = \Omega(n^{n-4}),$$

which improves upon the generalized Gilbert-Varshamov lower bound of this case. In the next section, we show similar results of the ball $\mathcal{B}_T(n,t)$ for general radius t.

VI. BALLS OF TREES OF ARBITRARY RADIUS

The main goal of this section is to calculate for each $T \in \mathbf{T}(n)$ the size of its ball $\mathcal{B}_T(n,t)$ and sphere $\mathcal{S}_T(n,t)$ for general radius t. For that, in Subsection VI-A, it is first shown how to calculate the forest's ball. Using this result, in Subsection VI-B, a recursive formula for the tree ball of trees is given and finally in Subsection VI-C we study the average ball size of trees.

A. The Size of the Forest's Ball

In this subsection it is shown how to explicitly find the size of the forest's ball $\mathcal{B}_F(n,t)$. By using this result, we will be able to proceed to the next step, which is calculating the size of the tree ball of trees $\mathcal{B}_T(n,t)$. Throughout this section we use the notation $\deg_{\mathcal{T}}(v_i)$ for the degree of the node v_i in a tree $\mathcal{T}$ in order to emphasize over which tree the degree is referred to. We start with several definitions and claims.

Let $\mathcal{T} = (V_t, E) \in \mathbf{T}(t)$ be a tree, where $V_t = \{v_0, v_1, \ldots, v_{t-1}\}$, and let $F = \{C_0, C_1, \ldots, C_{t-1}\} \in \mathbf{F}(n,t)$
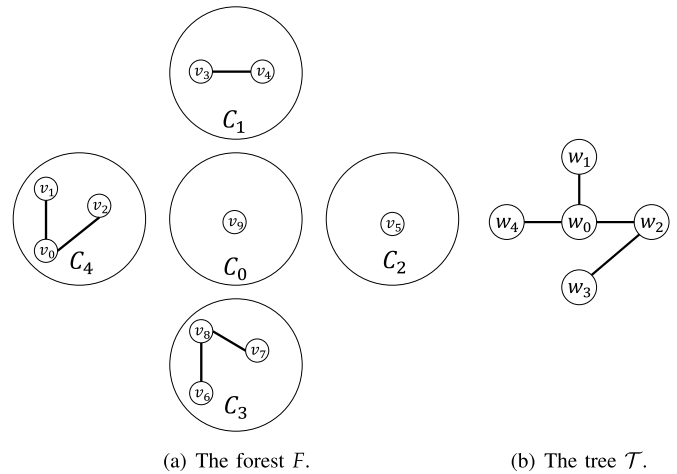

(a) The forest F. (b) The tree $\mathcal{T}$.

Fig. 3. For $n = 10$ and $t = 4$, a forest $F = \{C_0, C_1, C_2, C_3, C_4\} \in \mathbf{F}(10,5)$ over the set of nodes $\{v_i \mid i \in [10]\}$, and a tree $\mathcal{T} \in \mathbf{T}(4)$ over the set of nodes $\{w_i \mid i \in [5]\}$, are presented. Notice that $|C_0| = 1, |C_1| = 2, |C_2| = 1, |C_3| = 3, |C_4| = 3$, and thus, $\mathbf{P}_1(F, \mathcal{T}) = |C_0| \cdot |C_1| \cdot |C_0| \cdot |C_2| \cdot |C_0| \cdot |C_4| \cdot |C_2| \cdot |C_3| = 18$.

be a forest. Let $\mathbf{P}_1(F, \mathcal{T}) : \mathbf{F}(n,t) \times \mathbf{T}(t) \to \mathbb{N}$ be the following mapping. For all F and $\mathcal{T}$,

$$\mathbf{P}_1(F, \mathcal{T}) = \prod_{\langle v_i, v_j \rangle \in E} |C_i||C_j|.$$

The mapping $\mathbf{P}_1$ counts the number of options to complete a forest F with t connected components into a complete tree, according to a specific tree structure $\mathcal{T}$ with t nodes, corresponding to the t connected components of F. Since every $|C_i|$ appears in this multiplication exactly $\deg_{\mathcal{T}}(v_i)$ times (v_i is a node in $\mathcal{T}$), it is deduced that,

$$\mathbf{P}_1(F, \mathcal{T}) = \prod_{\langle v_i, v_j \rangle \in E} |C_i||C_j| = \prod_{C_i \in F} |C_i|^{\deg_{\mathcal{T}}(v_i)}. \quad (11)$$

Fig. 3 demonstrates the mapping $\mathbf{P}_1$.

Let $F = \{C_0, C_1, \ldots, C_{t-1}\} \in \mathbf{F}(n,t)$ be a forest and let E_F be its edge set. For all $T = (V_n, E_T) \in V_F(n,t)$ we denote its *component edge set* $E_{F,T}$ by

$$E_{F,T} = E_T \setminus E_F.$$

The component edge set is the set of edges that were added to the forest F in order to receive the tree T. We are ready to show the following claim.

Claim 2: For all $F \in \mathbf{F}(n, t+1)$ it holds that

$$V_F(n,t) = \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_1(F, \mathcal{T}).$$

Proof: Let $F = \{C_0, C_1, \ldots, C_t\}$ be a forest. Let H be a mapping $H : V_F(n,t) \to \mathbf{T}(t+1)$ that will be defined as follows. For each $T \in V_F(n,t)$ with a component edge set $E_{F,T}$, it holds that $H(T) = \mathcal{T}$ if for all $e \in E_{F,T}$ such that e connects between C_k and C_ℓ, the edge $\langle v_k, v_\ell \rangle$ exists in $\mathcal{T}$. Clearly, every $T \in V_F(n,t)$ is mapped and H is well defined. Moreover, for any $\mathcal{T} = (V_{t+1}, E) \in \mathbf{T}(t+1)$, H maps exactly

$$\prod_{\langle v_i, v_j \rangle \in E} |C_i||C_j|.$$

trees from $V_F(n, t)$ into $\mathcal{T}$, which is exactly the value of $\mathbf{P}_1(F, \mathcal{T})$. Thus,

$$V_F(n, t) = \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_1(F, \mathcal{T}).$$

■

Next, another mapping $\mathbf{P}_2(F, \mathcal{T}) : \mathbf{F}(n, t) \times \mathbf{T}(t) \to \mathbb{N}$ is defined. For every forest $F = \{C_0, C_1, \ldots, C_{t-1}\} \in \mathbf{F}(n, t)$ and a tree $\mathcal{T} \in \mathbf{T}(t)$ with a prüfer sequence

$$\boldsymbol{w}_{\mathcal{T}} = (i_0, i_1, \ldots, i_{t-3}) \in [t]^{t-2},$$

we let

$$\mathbf{P}_2(F, \mathcal{T}) = |C_{i_0}| \cdot |C_{i_1}| \cdots |C_{i_{t-3}}|.$$

Using the fact that each number i of node v_i appears in the prüfer sequence $\boldsymbol{w}_{\mathcal{T}}$ of $\mathcal{T}$ exactly $\deg_{\mathcal{T}}(v_i) - 1$ times, we deduce that

$$\mathbf{P}_2(F, \mathcal{T}) = \prod_{C_i \in F} |C_i|^{\deg_{\mathcal{T}}(v_i) - 1}. \tag{12}$$

Let $g_F(x)$ be the generating function of F, defined by

$$g_F(x) = \sum_{i=0}^{t-1} x^{|C_i|}.$$

This generating function will be used in the proof of the following claim.

Claim 3: Let F be a forest in $\mathbf{F}(n, t+1)$. Then,

$$\sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_2(F, \mathcal{T}) = \left(\sum_{C_i \in F} |C_i| \right)^{t-1} = n^{t-1}.$$

Proof: Let $F \in \mathbf{F}(n, t+1)$ be a forest and let $g_F(x)$ be its generating function. Let $G(x) = (g_F(x))^{t-1}$ and we deduce that

$$G(x) = (g_F(x))^{t-1} = \left(\sum_{i=0}^{t} x^{|C_i|} \right)^{t-1}$$
$$= \sum_{(i_0, i_1, \ldots, i_{t-2}) \in [t+1]^{t-1}} x^{|C_{i_0}| \cdot |C_{i_1}| \cdots |C_{i_{t-2}}|}.$$

Since each monomial of $G(x)$ is of the from $x^{|C_{i_0}| \cdot |C_{i_1}| \cdots |C_{i_{t-2}}|}$ for $(i_0, i_1, \ldots, i_{t-2}) \in [t+1]^{t-1}$, it holds that the sum of all the powers of x in $G(x)$ is

$$\sum_{(i_0, i_1, \ldots, i_{t-2}) \in [t+1]^{t-1}} |C_{i_0}| \cdot |C_{i_1}| \cdots |C_{i_{t-2}}|,$$

which is equal to the sum

$$\left(\sum_{C_i \in F} |C_i| \right)^{t-1}.$$

Furthermore, each vector $(i_0, i_1, \ldots, i_{t-2}) \in [t+1]^{t-1}$ is a prüfer sequence $\boldsymbol{w}_{\mathcal{T}}$ of some $\mathcal{T} \in \mathbf{T}(t+1)$. Thus we deduce that

$$G(x) = \sum_{\mathcal{T} \in \mathbf{T}(t+1)} x^{\mathbf{P}_2(F, \mathcal{T})},$$

and the powers sum of x is exactly

$$\sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_2(F, \mathcal{T}).$$

Therefore,

$$\sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_2(F, \mathcal{T}) = \left(\sum_{C_i \in F} |C_i| \right)^{t-1}.$$

Lastly, since $\sum_{C_i \in F} |C_i| = n$, it holds that

$$\left(\sum_{C_i \in F} |C_i| \right)^{t-1} = n^{t-1},$$

which concludes the proof.   ■

According to the last two claims, the next corollary is derived and provides an explicit expression to calculate the forest's ball size.

Corollary 29: For any $\{C_0, C_1, \ldots, C_t\} = F \in \mathbf{F}(n, t+1)$ it holds that

$$V_F(n, t) = n^{t-1} \prod_{C_i \in F} |C_i|.$$

Proof: The proof will hold by the following sequence of equations, that will be explained below,

$$V_F(n, t) \overset{(a)}{=} \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_1(F, \mathcal{T})$$

$$\overset{(b)}{=} \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \prod_{C_i \in F} |C_i|^{\deg_{\mathcal{T}}(v_i)}$$

$$\overset{(c)}{=} \prod_{C_i \in F} |C_i| \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \prod_{C_i \in F} |C_i|^{\deg_{\mathcal{T}}(v_i) - 1}$$

$$\overset{(d)}{=} \prod_{C_i \in F} |C_i| \sum_{\mathcal{T} \in \mathbf{T}(t+1)} \mathbf{P}_2(F, \mathcal{T})$$

$$\overset{(e)}{=} \prod_{C_i \in F} |C_i| \left(\sum_{C_i \in F} |C_i| \right)^{t-1}$$

$$\overset{(f)}{=} n^{t-1} \prod_{C_i \in F} |C_i|.$$

Equality (a) holds by Claim 2. Equality (b) holds due to (11). Equality (c) is a result of taking the common factor $\prod_{C_i \in F} |C_i|$ from the summation. Note also that for all $i \in [t+1]$, $\deg_{\mathcal{T}}(v_i) > 0$. Equality (d) holds due to (12). Equality (e) holds by Claim 3. Equality (f) holds since $(|C_0| + |C_1| + \cdots + |C_t|) = n$.   ■

B. The Size of the Tree Ball of Trees

In this subsection we present a recursive formula for the tree ball of trees $\mathcal{B}_T(n, t)$ and its sphere $\mathcal{S}_T(n, t)$, as well as asymptotic bounds on their sizes. First, according to Corollary 29, we immediately get the following corollary.

Corollary 30: For all $T \in \mathbf{T}(n)$ it holds that

$$\sum_{F \in \mathcal{P}_T(n, t)} V_F(n, t) = n^{t-1} \sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, t)} i_0 i_1 \cdots i_t.$$

Next, a recursive connection between the sizes of forest's balls and spheres (of trees) is shown.

Lemma 31: For all $T \in \mathbf{T}(n)$ it holds that

$$\sum_{F \in \mathcal{P}_T(n,t)} V_F(n,t) = \sum_{i=0}^{t} \binom{n-1-t+i}{i} S_T(n,t-i).$$

(13)

Proof: Let $T = (V,E) \in \mathbf{T}(n)$. First notice that for all $0 \leqslant i \leqslant t$,

$$\bigcup_{i=0}^{t} S_T(n,i) = \bigcup_{F \in \mathcal{P}_T(n,t)} V_F(n,t).$$

Therefore, our main goal in this proof is finding, for a given tree $T_i = (V, E_i) \in S_T(n,i)$, the number of forests in $\mathcal{P}_T(n,t)$ in which the tree belongs to their ball of trees. This number equals to the size of the intersection $\mathcal{P}_T(n,t) \cap \mathcal{P}_{T_i}(n,t)$ since all of these forests belong also to $\mathcal{P}_{T_i}(n,t)$. Thus, every forest $F \in \mathcal{P}_T(n,t) \cap \mathcal{P}_{T_i}(n,t)$ is received in two steps. First, remove from T_i the $t-i$ edges in $E_i \setminus E$. Then, i more edges from $E \cap E_i$ are chosen, where $|E \cap E_i| = n-1-(t-i)$. Note that indeed every forest in $\mathcal{P}_T(n,t) \cap \mathcal{P}_{T_i}(n,t)$ is generated by this procedure. Thus,

$$|\mathcal{P}_T(n,t) \cap \mathcal{P}_{T_i}(n,t)| = \binom{n-1-(t-i)}{i}.$$

Therefore, in (13) each tree $T_i \in S_T(n,t-i)$ belongs to the forest's balls of $\binom{n-1-(t-i)}{i}$ different forests in $\mathcal{P}_T(n,t)$. Since it is true for all $0 \leqslant i \leqslant t$ we conclude the lemma's statement. ∎

Combining Corollary 30 and Lemma 31, a recursive formula for the size of a sphere is presented.

Corollary 32: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n-1-t+i}{i} S_T(n,t-i) = n^{t-1} \sum_{(i_0,i_1,\ldots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t.$$

Using Corollary 32, a recursive formula for the tree ball of trees is immediately deduced, see Appendix B.

Theorem 33: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i) = n^{t-1} \sum_{(i_0,i_1,\ldots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t.$$

The proof of the following lemma can be found in Appendix C, and it is the last step before presenting the main result of this section.

Lemma 34: For any positive integer α, if

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i) = \Omega(n^{\alpha t}),$$

and $V_T(n,0) = 1$, then $V_T(n,t) = \Omega(n^{\alpha t})$.

Finally, the main result of this section is shown.

Theorem 35: For all $T \in \mathbf{T}(n)$ and fixed t, it holds that

$$V_T(n,t) = \Omega(n^{2t}), \quad V_T(n,t) = \mathcal{O}(n^{3t}).$$

Proof: First we will prove that $V_T(n,t) = \Omega(n^{2t})$. Given positive integers $i_0, i_1, \ldots, i_{t-1}, i_t, n$ such that $i_0 + i_1 + \cdots + i_{t-1} + i_t = n$, it holds that

$$n-t \overset{(a)}{\leqslant} i_0 i_1 \cdots i_t \overset{(b)}{\leqslant} \left(\frac{n}{t+1}\right)^{t+1}, \quad (14)$$

where (a) is well known and (b) holds by using the arithmetic-geometric mean inequality. Thus, for all $T \in \mathbf{T}(n)$

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i)$$

$$\overset{(a)}{=} n^{t-1} \sum_{(i_0,i_1,\ldots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t$$

$$\overset{(b)}{\geqslant} n^{t-1} \binom{n-1}{t}(n-t) = \Omega(n^{2t}),$$

and

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i)$$

$$\overset{(a)}{=} n^{t-1} \sum_{(i_0,i_1,\ldots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t$$

$$\overset{(b)}{\leqslant} n^{t-1} \binom{n-1}{t}\left(\frac{n}{t+1}\right)^{t+1} = \mathcal{O}(n^{3t}),$$

where in both cases (a) holds by Theorem 33 and inequality (b) holds according to (14). Therefore, it immediately deduced that $V_T(n,t) = \mathcal{O}(n^{3t})$. The result $V_T(n,t) = \Omega(n^{2t})$ is deduced according to Lemma 34. ∎

C. The Average Ball Size

In this section we study the asymptotic behavior of the average ball size (of trees). First, using Theorem 33 and Lemma 31 we deduce that for all $T \in \mathbf{T}(n)$

$$\sum_{F \in \mathcal{P}_T(n,t)} V_F(n,t) = \sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i).$$

(15)

The following recursive relation on the average ball size is presented.

Lemma 36: For all n and t, it holds that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V(n,t-i)$$

$$= n^{2t-n} \frac{1}{(t+1)!} \sum_{\substack{1 \leqslant i_0,i_1,\ldots,i_t \leqslant n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}.$$

Proof: The following holds,

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} \sum_{T \in \mathbf{T}(n)} V_T(n,t-i)$$

$$\overset{(a)}{=} \sum_{T \in \mathbf{T}(n)} \sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i)$$

$$\overset{(b)}{=} \sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,t)} V_F(n,t) \overset{(c)}{=} \sum_{F \in \mathbf{F}(n,t+1)} \sum_{T \in \mathcal{B}_F(n,t)} V_F(n,t)$$

$$= \sum_{F \in \mathbf{F}(n,t+1)} (V_F(n,t))^2 \overset{(d)}{=} \sum_{F \in \mathbf{F}(n,t+1)} \left(n^{t-1} \prod_{C_i \in F} |C_i|\right)^2$$

$$\overset{(e)}{=} \frac{n^{2t-2}}{(t+1)!} \sum_{\substack{0<i_0,\ldots,i_t<n \\ i_0+\cdots+i_t=n}} \binom{n}{i_0,\ldots,i_t} i_0^{i_0-2} \cdots i_t^{i_t-2}(i_0\cdots i_t)^2$$

$$= n^{2t-2}\frac{1}{(t+1)!} \sum_{\substack{0<i_0,i_1,\ldots,i_t<n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}.$$

Equality (a) holds by changing the summation order. Equality (b) holds due to (15) and Theorem 33. Equality (c) holds by changing the summation order of trees and forests as it was done in (10). Equality (d) holds by Corollary 29. We deduce equality (e) as follows. It was shown in [18] that

$$F(n,t+1) = \frac{1}{(t+1)!} \sum_{\substack{0<i_0,\ldots,i_t<n \\ i_0+\cdots+i_t=n}} \binom{n}{i_0,\ldots,i_t} i_0^{i_0-2} \cdots i_t^{i_t-2}.$$

For each $F \in \mathbf{F}(n,t+1)$ we denote $|C_j| = i_j$, $j \in [t+1]$. Thus,

$$\Big(\prod_{C_i \in F} |C_i|\Big)^2 = (i_0\cdots i_t)^2,$$

which verifies the equality in step (e). After dividing the last expression in the series of equations by n^{n-2}, the proof is concluded. ∎

Next, we seek to show the main result of this section, that is, the asymptotic size of the average ball. For that, we first show the following claim, where its proof is shown in Appendix D.

Claim 4: For a positive integer n and a fixed t it holds that

$$\sum_{i=1}^{n-1} \binom{n}{i} i^i(n-i)^{n-i}\Theta(i^{t/2}) = \Theta(n^{t/2})\sum_{i=1}^{n-1} \binom{n}{i} i^i(n-i)^{n-i}.$$

The following lemma is now presented.

Lemma 37:

$$\sum_{\substack{0<i_0,i_1,\ldots,i_t<n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t} = \Theta(n^{n+t/2}).$$

Proof: Consider the sequence of integers $1^1, 2^2, 3^3, \ldots$, that is $a_n = n^n$, for $n \geqslant 1$. Let $G(x)$ be its generating function, i.e.

$$G(x) = \sum_{n=1}^{\infty} a_n \frac{x^n}{n!} = \sum_{n=1}^{\infty} n^n \frac{x^n}{n!}.$$

Denote by $F_t(x)$ the function $F_t(x) = (G(x))^{t+1}$. Thus,

$$F_t(x) = \Big(\sum_{i_0=1}^{\infty} i_0^{i_0}\frac{x^{i_0}}{i_0!}\Big)\cdots\Big(\sum_{i_t=1}^{\infty} i_t^{i_t}\frac{x^{i_t}}{i_t!}\Big)$$

$$= \sum_{n=1}^{\infty} \Big(\sum_{\substack{0<i_0,i_1,\ldots,i_t<n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}\Big)\frac{x^n}{n!},$$

and the coefficient of $x^n/n!$ in $F_t(x)$ is exactly

$$\sum_{\substack{0<i_0,i_1,\ldots,i_t<n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}.$$

Next, it is shown by induction on t that the order of the coefficient of $x^n/n!$ in $F_t(x)$ is $\Theta(n^{n+t/2})$.
Base: Clearly the coefficient of $x^n/n!$ in $F_0(x)$ is n^n, since $F_0(x) = G(x)$.

Inductive Step: Assume that the coefficient of $x^n/n!$ in $F_t(x)$ is $\Theta(n^{n+t/2})$. Thus, the coefficient of $x^n/n!$ in $F_{t+1}(x)$ is exactly

$$F_{t+1}(x) = F_t(x)F_0(x)$$

$$= \Big(\sum_{i=1}^{\infty}\Big(\sum_{\substack{0<i_0,i_1,\ldots,i_t<i \\ i_0+i_1+\cdots+i_t=i}} \binom{i}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}\Big)\frac{x^i}{i!}\Big)$$

$$\cdot\Big(\sum_{j=1}^{\infty} j^j\frac{x^j}{j!}\Big) \overset{(a)}{=} \Big(\sum_{i=1}^{\infty}\Theta(i^{i+t/2})\frac{x^i}{i!}\Big)\cdot\Big(\sum_{j=1}^{\infty} j^j\frac{x^j}{j!}\Big)$$

$$\overset{(b)}{=} \sum_{i=1}^{n-1}\binom{n}{i} i^i(n-i)^{n-i}\Theta(i^{t/2})\frac{x^n}{n!}$$

$$\overset{(c)}{=} \Theta(n^{t/2})\sum_{i=1}^{n-1}\binom{n}{i} i^i(n-i)^{n-i}\frac{x^n}{n!}$$

$$\overset{(d)}{=} \Theta(n^{t/2})\sum_{i=1}^{n-1}\Theta(n^{n+0.5})\frac{x^n}{n!} = \sum_{i=1}^{n-1}\Theta(n^{n+(t+1)/2})\frac{x^n}{n!},$$

where equality (a) holds by the induction assumption, and equality (b) holds by denoting $i+j = n$. Equality (c) holds by Claim 4 and equality (d) holds due to Corollary 28, where we showed that the coefficient of $x^n/n!$ in $F_1(x)$ is $\Theta(n^{n+0.5})$. ∎

We are now ready to find the asymptotic size of the average ball.

Corollary 38: It holds that

$$V(n,t) = \Theta(n^{2.5t}).$$

Proof: It holds that

$$\sum_{i=0}^{t}\binom{n-2-t+i}{i}V(n,t-i)$$

$$\overset{(a)}{=} n^{2t-n}\frac{1}{(t+1)!}\sum_{\substack{1\leqslant i_0,i_1,\ldots,i_t\leqslant n \\ i_0+i_1+\cdots+i_t=n}} \binom{n}{i_0,i_1,\ldots,i_t} i_0^{i_0} i_1^{i_1} \cdots i_t^{i_t}$$

$$\overset{(b)}{=} \Theta(n^{2t-n})\Theta(n^{n+t/2}) = \Theta(n^{2.5t}),$$

where (a) holds by Lemma 36 and (b) holds using Lemma 37. Therefore it is deduced that $V(n,t) = \mathcal{O}(n^{2.5t})$. The result $V(n,t) = \Omega(n^{2.5t})$ is proved according to Lemma 34. ∎

In summary, we proved that for every $T \in \mathbf{T}(n)$ and fixed t it holds that $V_T(n,t) = \Omega(n^{2t})$, $V_T(n,t) = \mathcal{O}(n^{3t})$ and the average ball size satisfies $V(n,t) = \Theta(n^{2.5t})$. The sphere packing bound for smallest tree ball of trees size of radius t for $\mathcal{T}$-$(n,M,d=2t+1)$ codes over trees in this case shows that

$$A(n,d) \leqslant \frac{n^{n-2}}{\alpha n^{2t}} = \frac{1}{\alpha}n^{n-2-2t} = \frac{1}{\alpha}n^{n-1-d},$$

for some constant α. Thus, we derive a similar result as in Corollary 19.

By using the generalized Gilbert-Varshamov lower bound for the average ball size [33] for $\mathcal{T}$-$(n,M,d=t+1)$ codes over trees, we get,

$$A(n,d) = \Omega(n^{n-2-2.5(d-1)}) = \Omega(n^{n+0.5-2.5d}).$$

However, in Section VIII, based upon Construction 3, we will get that

$$A(n, d) = \Omega(n^{n-2d}).$$

In the next section similar results are shown for stars and path trees. While the exact size of the tree balls of trees is found for stars, for path trees we only find its asymptotic behavior and finding its exact expression is left for future work. It is also shown that for a fixed t the star tree has asymptoticly the smallest size of the tree of ball of trees, while the path tree achieves asymptoticly the largest size.

VII. THE TREE BALLS OF TREES FOR STARS AND PATH TREES

Several more interesting results on the size of the tree balls of trees and more specifically for stars and path trees are shown in this section. First we show an exact formula for $V^\star(n, t)$ and conclude that $V^\star(n, t) = \Theta(n^{2t})$. Then we simplify the recursive formula in Theorem 33 for path trees and we will show that $V^-(n, t) = \Theta(n^{3t})$. Finally, we will show the following explicit upper bound on the recursive formula in Theorem 33, that will not depend on the structure of the tree,

$$\sum_{i=0}^{t} \binom{n - 2 - t + i}{i} V_T(n, t - i) \leqslant n^{t-1} \binom{n + t}{2t + 1}.$$

This result will be shown in Theorem 44.

First, we derive some interesting properties from the recursive formula in Theorem 33, which proved that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n, t - i) = n^{t-1} \sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, t)} i_0 i_1 \cdots i_t.$$

Notice also that for all $T \in \mathbf{T}(n)$ and $t = n - 1$,

$$n^{n-2} = n^{(n-1)-1} \sum_{(1,1,\ldots,1) \in P_T(n, n-1)} 1$$

$$= \sum_{i=0}^{n-1} \binom{n - 2 - (n-1) + i}{i} V_T(n, n - 1 - i)$$

$$= \sum_{i=0}^{n-1} \binom{i - 1}{i} V_T(n, n - 1 - i) = V_T(n, n - 1),$$

where $\binom{-1}{0}$ is defined to be 1, and indeed $V_T(n, n - 1) = n^{n-2}$. Similarly, if $t = n - 2$ then

$$2(n - 1)n^{n-3} = n^{(n-2)-1} \sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, n-2)} 2$$

$$= \sum_{i=0}^{n-2} \binom{i}{i} V_T(n, n - 2 - i) = \sum_{i=0}^{n-2} V_T(n, n - 2 - i),$$

and thus,

$$\sum_{i=0}^{n-2} V_T(n, i) = 2(n - 1)n^{n-3}. \tag{16}$$

As for stars, applying Theorem 33, we simply draw the following formula

$$\sum_{i=0}^{t} \binom{n - 2 - t + i}{i} V^\star(n, t - i) = n^{t-1} \binom{n - 1}{t} (n - t). \tag{17}$$

Using this result and the proof of Theorem 33, the following interesting result holds.

Corollary 39: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n - 2 - t + i}{i} \left(V_T(n, t - i) - V^\star(n, t - i)\right) \geqslant 0.$$

Next an exact formula of the size of the tree ball of trees for stars is presented. The proof of this theorem is shown in Appendix E.

Theorem 40: The size of the sphere for a star satisfies

$$S^\star(n, t) = \binom{n - 1}{t} (n - 1)^{t-1}(n - t - 1),$$

and the size of the tree ball of trees for a star satisfies

$$V^\star(n, t) = \sum_{j=0}^{t} \binom{n - 1}{j} (n - 1)^{j-1}(n - j - 1).$$

Note that while in Theorem 35 it was shown that for all $T \in \mathbf{T}(n)$ it holds that $V_T(n, t) = \Omega(n^{2t})$, for stars it is deduced that $S^\star(n, t) = \Theta(n^{2t})$ and $V^\star(n, t) = \Theta(n^{2t})$, which verifies that stars have asymptotically the smallest size of the tree ball of trees.

We turn to study the size of the tree ball of trees for path trees. We first simplify the formula of Theorem 33 in the path tree case.

Theorem 41: The size of the tree ball of trees for a path tree satisfies

$$\sum_{i=0}^{t} \binom{n - 2 - t + i}{i} V^-(n, t - i) = n^{t-1} \binom{n + t}{2t + 1}.$$

Proof: Denote by A the set

$$A = \left\{ (j_0, j_1, \ldots, j_t) \,\middle|\, \begin{array}{c} 1 \leqslant j_0 \leqslant n - t \\ 1 \leqslant j_1 \leqslant n - (t - 1) - j_0 \\ \vdots \\ 1 \leqslant j_{t-1} \leqslant n - 1 - \sum_{s=0}^{t-2} j_s \\ j_t = n - \sum_{s=0}^{t-1} j_s \end{array} \right\}.$$

Let $T \in \mathbf{T}(n)$ be a path tree. The following equations hold.

$$\frac{1}{n^{t-1}} \sum_{i=0}^{t} \binom{n - 2 - t + i}{i} V^-(n, t - i)$$

$$\overset{(a)}{=} \sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, t)} i_0 i_1 \cdots i_t$$

$$\overset{(b)}{=} \sum_{(j_0, j_1, \ldots, j_t) \in A} j_0 j_1 \cdots j_t \overset{(c)}{=} \binom{n + t}{2t + 1}.$$

Equality (a) holds due to Theorem 33. As for equality (b), note that after an erasure of t edges of T, we get $t + 1$ connected components of T where each of them is a path tree. The value of j_i represents a path subtree as follows. The first path subtree will be of size j_0 which can be at least of size 1 and at most of size $n - t$. Similarly, the size j_1 of the second path subtree ranges by between 1 and $n - (t - 1) - j_0$, i.e. $1 \leqslant j_1 \leqslant n - (t - 1) - j_0$. Continuing with this analysis, the size j_t of the last path subtree satisfies $j_t = n - \sum_{s=0}^{t-1} j_s$. Hence, the set of all vectors $(j_0, j_1, \ldots, j_t)$ is exactly the

set A, which verifies equality (b). Equality (c) holds using combinatorial proof. Consider the problem of counting the number of options to choose $2t+1$ numbers from the set of numbers $[n+t]$. The right hand side is trivial. As for the left hand side, denote by $(x_1, x_2, \ldots, x_{2t+1})$ a vector such that $x_1 < x_2 < \cdots < x_{2t+1}$ representing an option of chosen $2t+1$ numbers. We choose these $2t+1$ in two steps. First we choose the values of $x_2, x_4, \ldots, x_{2t}$. We translate choosing these numbers to choosing the values of $j_0, j_1, \ldots, j_t$ such that

$$x_2 = j_0, x_4 = j_1 + j_0, \ldots, x_{2t} = \sum_{s=0}^{t-1} j_s = n - j_t.$$

In the next step we choose the values of $x_1, x_3, \ldots, x_{2t+1}$. Since $x_1 < x_2$, there are j_0 options to pick x_1. Similarly since $x_2 < x_3 < x_4$, there are $x_4 - x_2 = j_1$ options to pick x_3. Lastly, since $x_{2t} < x_{2t+1} < n$, there are j_t options to pick x_{2t+1}. Thus, every option of choosing $j_0, j_1, \ldots, j_t$ counts $j_0 j_1 \ldots j_{t-1} j_t$ solutions, and since all options of this problem are counted, the proof is concluded. ∎

Similarly to the case of stars, we showed in Theorem 35 that for all $T \in \mathbf{T}(n)$ it holds that $V_T(n, t) = \mathcal{O}(n^{3t})$, and it is also true for path trees as we can see in Theorem 41. According to Lemma 34 we also deduce that $V^-(n, t) = \Theta(n^{3t})$, that is, a path tree has asymptoticly the largest size of the tree ball of trees.

Although a path tree has asymptoticly the largest size of the tree ball of trees, it is not necessarily true that for every n and t its size is strictly the largest. We will show such an example at the end of this section.

Our last goal of this section is a stronger upper bound on the size of the tree ball of trees. According to Theorem 35, it was shown that for every tree $T \in \mathbf{T}(n)$ it holds that

$$\sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, t)} i_0 i_1 \cdots i_t \leqslant \binom{n-1}{t} \left(\frac{n}{t+1}\right)^{t+1},$$

while our goal is to improve this upper bound to be

$$\sum_{(i_0, i_1, \ldots, i_t) \in P_T(n, t)} i_0 i_1 \cdots i_t \leqslant \binom{n+t}{2t+1}. \tag{18}$$

While this result does not improve asymptotic upper bound of $\mathcal{O}(n^{3t})$, we believe that this upper bound is interesting and furthermore it verifies the statement of Theorem 44.

First the definition of (2) is slightly modified. Let $\boldsymbol{v}_\ell = (v_{i_0}, v_{i_1}, \ldots, v_{i_{\ell-1}})$ be a vector of ℓ not necessary distinct nodes of $T \in \mathbf{T}(n)$ where $1 \leqslant \ell \leqslant t+1$. Let $F_{T, E'}(\boldsymbol{v}_\ell) = (V_n, E \setminus E') \in \mathcal{P}_T(n, t)$ be a forest which is also denoted by $F_{T, E'}(\boldsymbol{v}_\ell) = \{C_0, C_1, \ldots, C_t\}$, such that $v_{i_j} \in C_j, j \in [\ell]$ and $|C_\ell| \leqslant |C_{\ell+1}| \leqslant \cdots \leqslant |C_t|$. In case there is more than one way to order to connected components $C_0, C_1, \ldots, C_t$, we choose one of them arbitrarily. For $1 \leqslant \ell \leqslant t+1$ denote the multi-set $P_T(n, t; \boldsymbol{v}_\ell)$

$$P_T(n, t; \boldsymbol{v}_\ell) =$$
$$\left\{ (|C_0|, \ldots, |C_t|) \mid \{C_0, \ldots, C_t\} = F_{T, E'}(\boldsymbol{v}_\ell) \in \mathcal{P}_T(n, t) \right\},$$

and for $\ell = 0$,
$$P_T(n, t; \boldsymbol{v}_\ell) = P_T(n, t).$$

Intuitively, this multi-set consist of profiles of forests such that all the nodes of the vector $\boldsymbol{v}_\ell$ are in different connected components of these forests. From this definition, in case that not all of the nodes in $\boldsymbol{v}_\ell$ are distinct, then $P_T(n, t; \boldsymbol{v}_\ell) = \emptyset$. Another property is that for all $\ell \in [t+2]$ it holds that

$$|P_T(n, t; \boldsymbol{v}_\ell)| \leqslant |P_T(n, t)| = \binom{n-1}{t}.$$

Next, for all $\ell \in [t+2]$, denote by $f_T(n, t; \boldsymbol{v}_\ell)$ the function

$$f_T(n, t; \boldsymbol{v}_\ell) = \sum_{(c_0, c_1, \ldots, c_t) \in P_T(n, t; \boldsymbol{v}_\ell)} c_\ell c_{\ell+1} \cdots c_t,$$

where in case that $\ell = t+1$, the function f_T is defined as

$$f_T(n, t, \boldsymbol{v}_{t+1}) = \sum_{(c_0, c_1, \ldots, c_t) \in P_T(n, t; \boldsymbol{v}_{t+1})} 1$$
$$= |P_T(n, t; \boldsymbol{v}_{t+1})| \leqslant \binom{n-1}{t}.$$

Again, if not all of the nodes in $\boldsymbol{v}_\ell$ are distinct, by definition $P_T(n, t; \boldsymbol{v}_\ell) = \emptyset$ and

$$f_T(n, t; \boldsymbol{v}_\ell) = 0.$$

Since in case that $t = n - 1$, each connected component is of size 1, the following property is immediately deduced,

$$f_T(n, n-1; \boldsymbol{v}_\ell) = \sum_{(c_0, c_1, \ldots, c_{n-1}) \in P_T(n, n-1; \boldsymbol{v}_\ell)} c_\ell c_{\ell+1} \cdots c_{n-1} = 1. \tag{19}$$

The main goal in this part is to show that

$$f_T(n, t; \boldsymbol{v}_\ell) \leqslant \binom{n+t-\ell}{2t+1-\ell}, \tag{20}$$

where in case that $\ell = 0$, the equality (18) is immediately deduced.

Let $T \in \mathbf{T}(n)$ for $n \geqslant 2$. For two integers t and ℓ such that $0 \leqslant \ell < t+1 < n$, let $\boldsymbol{v}_\ell = (v_{i_0}, \ldots, v_{i_{\ell-2}}, v_{i_{\ell-1}})$ be a vector of ℓ nodes in T. For $j \in [\ell+1]$ denote $\boldsymbol{v}_j = (v_{i_0}, \ldots, v_{i_{j-1}})$. For any node v_x in T denote $\boldsymbol{v}_{j+1}(v_x) = (v_{i_0}, \ldots, v_{i_{j-1}}, v_x)$. By a slight abuse of notation, given a vector $(c_0, \ldots, c_t) \in P_T(n, t; \boldsymbol{v}_\ell)$, the connected component C_j is referred to the value c_j, or in other words $(c_0, \ldots, c_t) = (|C_0|, \ldots, |C_t|)$. For any node v_x in T denote by $A_T(n, t; \boldsymbol{v}_\ell, v_x)$ the set

$$A_T(n, t; \boldsymbol{v}_\ell, v_x) = \left\{ (|C_0|, \ldots, |C_t|) \in P_T(n, t; \boldsymbol{v}_\ell) \mid v_x \in \bigcup_{i \in [\ell]} C_i \right\}.$$

Let v_x be a leaf connected to a node denoted by v_y in $T \in \mathbf{T}(n)$, and let $T_1 \in \mathbf{T}(n-1)$ be a tree generated by removing v_x from T. The definitions introduced above are used in the next claims and lemmas.

Claim 5: The following properties hold

a) It holds that

$$\sum_{(c_0, \ldots, c_t) \in P_T(n, t; \boldsymbol{v}_\ell)} c_\ell \cdots c_t$$
$$= \sum_{(c_0, \ldots, c_t) \in A_T(n, t; \boldsymbol{v}_\ell, v_y)} c_\ell \cdots c_t + \sum_{(c_0, \ldots, c_t) \in P_T(n, t; \boldsymbol{v}_{\ell+1}(v_y))} c_\ell \cdots c_t.$$

b) If v_x is not in $\boldsymbol{v}_\ell$ then,

$$\sum_{(c_0,\ldots,c_t)\in A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)} c_\ell\cdots c_t = \sum_{(c'_0,\ldots,c'_t)\in A_T(n,t;\boldsymbol{v}_\ell,v_x)} c'_\ell\cdots c'_t.$$

c) If v_x is not in $\boldsymbol{v}_\ell$ then,

$$\sum_{(c_0,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} c_\ell\cdots c_t$$
$$= \sum_{(c'_0,\ldots,c'_t)\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} (c'_\ell+1)c'_{\ell+1}\cdots c'_t$$
$$+ \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1\cdot c_{\ell+1}\cdots c_t.$$

d) If $v_x = v_{i_{\ell-1}}$ then

$$\sum_{(c_0,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t$$
$$= \sum_{(c'_0,\ldots,c'_t)\in P_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y))} c'_\ell\cdots c'_t$$
$$+ \sum_{(c_0,\ldots,c_{\ell-2},1,c_\ell,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t.$$

The proof of Claim 5 can be found in Appendix F. Next we show a recursive formula with respect to f_T.

Lemma 42: If v_x is not in $\boldsymbol{v}_\ell$ then,

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell)$$
$$+ f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell).$$

If v_x is in $\boldsymbol{v}_\ell$, and without loss of generality $v_x = v_{i_{\ell-1}}$, then

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1}).$$

The proof of Lemma 42 can be found in Appendix G. An example that illustrates this recursive formula is now presented.

Example 2: For $n = 10$, we illustrate in Fig. 4(a) a tree $T \in \mathbf{T}(10)$. In this example, $t = 4$ and $\ell = 1$, $v_x = v_5, v_y = v_6$ and $\boldsymbol{v}_\ell = (v_7)$. Let $T_1 \in \mathbf{T}(9)$ be a tree which is derived from T by removing the node v_5. After an erasure of 4 edges, the multiplication of the five connected components is counted in $f(n,t)$. Fig. 4(b), (c) and (d) represent the idea of the formula $f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell) + f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)$. The dashed edges in Fig. 4(b), (c) and (d) represent the erased edges from T, yielding a forest with five connected components C_0, C_1, C_2, C_3, and C_4. An example of possible erasure including the edge $\langle v_5, v_6\rangle$ is shown in Fig. 4(b). This example emphasizes the case which corresponds to the multiplication $|C_0| \cdot |C_1| \cdot |C_3| \cdot |C_4|$ that is also counted in $f_{T_1}(n-1,t;(v_7))$ since $|C_0| = 1$. Fig. 4(c) and (d) similarly emphasize the case in which an erasure of 4 edges does not include the edge $\langle v_5, v_6\rangle$. While Fig. 4(c) emphasizes the multiplication $|C_0| \cdot |C_1| \cdot \left(|C_3| - 1\right) \cdot |C_4|$, which is counted in $f_{T_1}(n-1,t-1;(v_7))$ (since v_5 is not in T_1), Fig. 4(d) emphasizes the multiplication $|C_0| \cdot |C_1| \cdot |C_4|$, which is counted in $f_{T_1}(n-1,t;(v_7,v_6))$. Hence, $|C_0| \cdot |C_1| \cdot |C_3| \cdot |C_4|$ is also counted in the case that the edge $\langle v_5, v_6\rangle$ is not erased.

Finally, the upper bound for $f_T(n,t;\boldsymbol{v}_\ell)$ is presented, while the proof is shown in Appendix H.

Lemma 43: For any tree $T \in \mathbf{T}(n), n \geqslant 1$ and a vector of $0 \leqslant \ell \leqslant t+1 \leqslant n$ nodes $\boldsymbol{v}_\ell = (v_{i_0}, v_{i_1}, \ldots, v_{i_{\ell-1}})$,

$$f_T(n,t;\boldsymbol{v}_\ell) \leqslant \binom{n+t-\ell}{2t+1-\ell}.$$

From Lemma 43 it is immediately deduced that for all $T \in \mathbf{T}(n)$,

$$\sum_{(i_0,i_1,\ldots,i_t)\in P_T(n,t)} i_0 i_1 \cdots i_t = f_T(n,t) \leqslant \binom{n+t}{2t+1}. \quad (21)$$

Using (21) the tighter upper bound for the recursive formula in Theorem 33 is shown in the following theorem.

Theorem 44: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i) \leqslant n^{t-1}\binom{n+t}{2t+1}.$$

From Theorem 44 and Theorem 41 we immediately deduce the following corollary.

Corollary 45: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i}\left(V_T(n,t-i) - V^-(n,t-i)\right) \leqslant 0.$$

Even though by Corollary 39,

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i}\left(V_T(n,t-i) - V^\star(n,t-i)\right) \geqslant 0,$$

and by Corollary 45,

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i}\left(V_T(n,t-i) - V^-(n,t-i)\right) \leqslant 0,$$

it does not imply that for all n and t, $V^\star(n,t) \leqslant V_T(n,t) \leqslant V^-(n,t)$. For example, if $t = n-2$, $V^\star(n,t) = n^{n-2}$ while $V^-(n,t) < n^{n-2}$, since one can check that there are two path trees $T_1, T_2 \in \mathbf{T}(n)$ such that $d_{\mathcal{T}}(T_1, T_2) = n - 1$. However, we conjecture that for fixed t and large enough n, it holds that $V^\star(n,t) \leqslant V_T(n,t) \leqslant V^-(n,t)$.

VIII. CONSTRUCTIONS OF CODES OVER TREES

In this section we show several constructions of codes over trees. The first is the construction of $\mathcal{T}\text{-}(n, \lfloor n/2\rfloor, n-1)$ codes, and the second is the construction of $\mathcal{T}\text{-}(n,n,n-2)$ codes. The third and our main result in this section is the construction of $\mathcal{T}\text{-}(n, M, d)$ codes for fixed d where $M = \Omega(n^{n-2d})$. For positive integers a and n we will use the notation $\langle a\rangle_n$ to denote the value of $(a \mod n)$.

A. A Construction of $\mathcal{T}\text{-}(n, \lfloor n/2\rfloor, n-1)$ Codes

A path tree $T = (V_n, E)$ with the edge set

$$E = \{(v_{i_j}, v_{i_{j+1}}) \mid j \in [n-1], i_j \in [n]\},$$

will be denoted by $T = (v_{i_0}, v_{i_1}, \ldots, v_{i_{n-1}})$, i.e., the nodes v_{i_0} and $v_{i_{n-1}}$ are leaves and the rest of the nodes have degree 2. Note that the number of path trees over n nodes is $n!/2$,
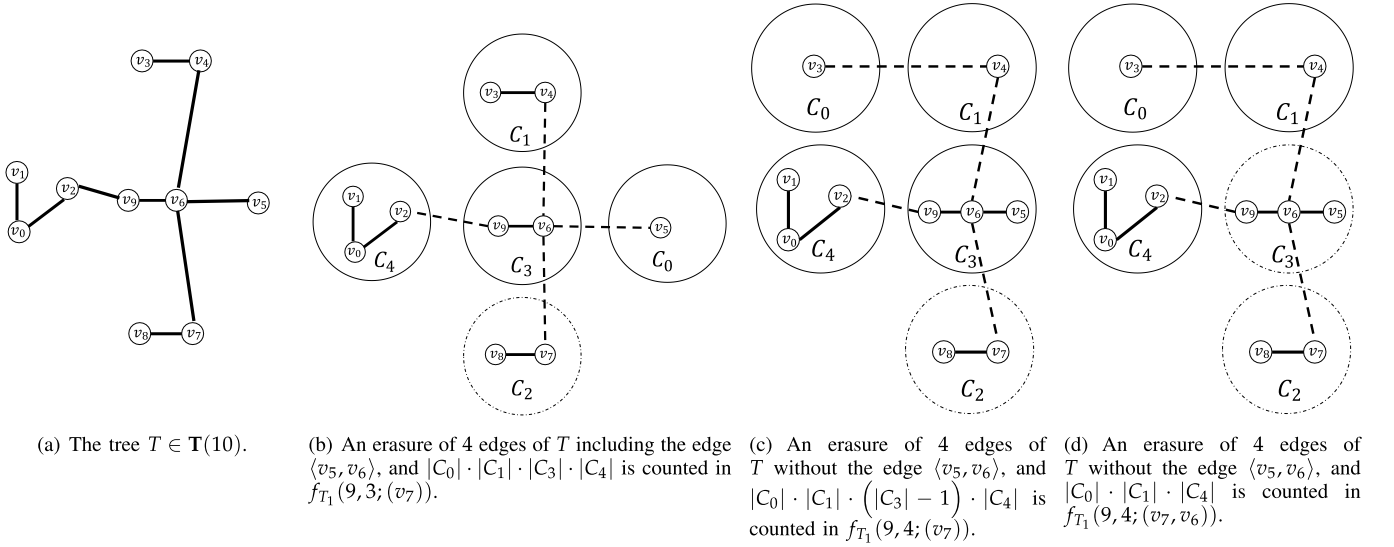
(a) The tree $T \in \mathbf{T}(10)$.

(b) An erasure of 4 edges of T including the edge $\langle v_5, v_6 \rangle$, and $|C_0| \cdot |C_1| \cdot |C_3| \cdot |C_4|$ is counted in $f_{T_1}(9, 3; (v_7))$.

(c) An erasure of 4 edges of T without the edge $\langle v_5, v_6 \rangle$, and $|C_0| \cdot |C_1| \cdot \left(|C_3| - 1\right) \cdot |C_4|$ is counted in $f_{T_1}(9, 4; (v_7))$.

(d) An erasure of 4 edges of T without the edge $\langle v_5, v_6 \rangle$, and $|C_0| \cdot |C_1| \cdot |C_4|$ is counted in $f_{T_1}(9, 4; (v_7, v_6))$.

Fig. 4. The idea of the recursive formula $f_T(10, 4; (v_7))) = f_{T_1}(9, 3; (v_7)) + f_{T_1}(9, 4; (v_7)) + f_{T_1}(9, 4; (v_7, v_6))$.
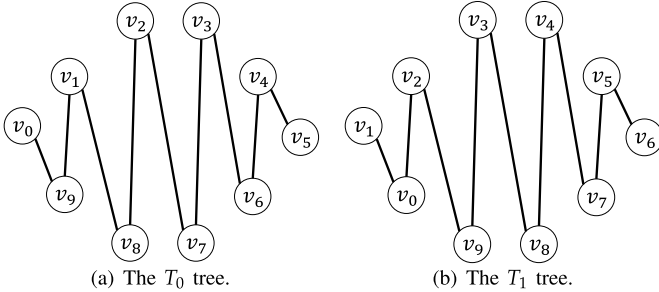

(a) The T_0 tree. (b) The T_1 tree.

Fig. 5. This code contains 5 trees, T_0, T_1, T_2, T_3, and T_4.

so every path tree has two representations in this form and we will use either one of them in the sequel. For $s \in [\lfloor n/2 \rfloor]$, denote by $T_s = (V_n, E)$ the path tree

$$T_s = \begin{cases} \left(v_{\langle s \rangle_n}, v_{\langle s-1 \rangle_n}, v_{\langle s+1 \rangle_n}, \ldots, v_{\langle s + \frac{n-1}{2} \rangle_n} \right) & : \text{if } n \text{ is odd,} \\ \left(v_{\langle s \rangle_n}, v_{\langle s-1 \rangle_n}, v_{\langle s+1 \rangle_n}, \ldots, v_{\langle s - \frac{n}{2} \rangle_n} \right) & : \text{if } n \text{ is even.} \end{cases}$$

Example 3: For $n = 10$ we show an example of the path tree T_0. By looking at the lower half of the circle in this figure, i.e. nodes $v_0, v_9, v_8, v_7, v_6, v_5$, there is a single edge connecting two vertices on this half circle. The path tree T_1 is received by rotating anticlockwise the nodes on this circle by one step. Note that all the edges in T_0 and T_1 are disjoint and this property holds also for the other path trees T_2, T_3, T_4.

The construction of a $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ code is given as follows. This construction is motivated by the factorization of the complete graph into mutually disjoint Hamiltonian paths; see [11], [17]. Even though this result is well known, for completeness we present it here along with its proof.

Construction 1: For all $n \geqslant 3$ let $\mathcal{C}_{\mathcal{T}_1}$ be the following code over trees

$$\mathcal{C}_{\mathcal{T}_1} = \{T_s = (V_n, E) | s \in [\lfloor n/2 \rfloor]\}.$$

Theorem 46: The code $\mathcal{C}_{\mathcal{T}_1}$ is a $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ code.

Proof: Clearly, since for all distinct $s_1, s_2 \in [\lfloor n/2 \rfloor]$ it holds that $s_1 \neq s_2 + \lfloor n/2 \rfloor$, it is deduced that $|\mathcal{C}_{\mathcal{T}_1}| = \lfloor n/2 \rfloor$. Next we prove that this code can correct $\rho = n - 2$ edge-erasures, by showing that $d_{\mathcal{T}}(\mathcal{C}_{\mathcal{T}_1}) > n - 2$.

Assume on the contrary that $d_{\mathcal{T}}(\mathcal{C}_{\mathcal{T}_1}) \leqslant n - 2$. Therefore, there are two distinct numbers $s_1, s_2 \in [\lfloor n/2 \rfloor]$ such that the trees $T_{s_1} = (V_n, E_1), T_{s_2} = (V_n, E_2) \in \mathcal{C}_{\mathcal{T}_1}$ hold $|E_1 \cap E_2| \geqslant 1$. Therefore, there exist two integers $t_1, t_2 \in [\lfloor n/2 \rfloor]$ such that one of the following cases hold:

a) $\langle v_{\langle s_1 + t_1 \rangle_n}, v_{\langle s_1 - (t_1 + 1) \rangle_n} \rangle = \langle v_{\langle s_2 + t_2 \rangle_n}, v_{\langle s_2 - (t_2 + 1) \rangle_n} \rangle$,

b) $\langle v_{\langle s_1 + t_1 \rangle_n}, v_{\langle s_1 - (t_1 + 1) \rangle_n} \rangle = \langle v_{\langle s_2 - t_2 \rangle_n}, v_{\langle s_2 + t_2 \rangle_n} \rangle$,

c) $\langle v_{\langle s_1 - t_1 \rangle_n}, v_{\langle s_1 + t_1 \rangle_n} \rangle = \langle v_{\langle s_2 - t_2 \rangle_n}, v_{\langle s_2 + t_2 \rangle_n} \rangle$.

We will eliminate all those options as follows.

a) If $\langle s_1 + t_1 \rangle_n = \langle s_2 + t_2 \rangle_n$ and $\langle s_1 - (t_1 + 1) \rangle_n = \langle s_2 - (t_2 + 1) \rangle_n$ then by summing those equations we deduce that $\langle 2s_1 - 1 \rangle_n = \langle 2s_2 - 1 \rangle_n$. Therefore, we deduce that $s_1 = s_2$ which is a contradiction. Similar proof shows that it is impossible to have $\langle s_1 + t_1 \rangle_n = \langle s_2 - (t_2 + 1) \rangle_n$ and $\langle s_2 + t_2 \rangle_n = \langle s_1 - (t_1 + 1) \rangle_n$.

b) If $\langle s_1 + t_1 \rangle_n = \langle s_2 - t_2 \rangle_n$ and $\langle s_1 - (t_1 + 1) \rangle_n = \langle s_2 + t_2 \rangle_n$ then by summing those equations we deduce that $\langle 2s_1 - 1 \rangle_n = \langle 2s_2 \rangle_n$. Since $s_1, s_2 \in [\lfloor n/2 \rfloor]$, if $s_1 \neq 0$ then $2s_1 - 1 < n - 1$ and $2s_2 < n - 1$. Clearly, $\langle 2s_1 - 1 \rangle_n$ is odd and $\langle s_2 \rangle_n$ is even, (since both of them smaller than n) so it is deduced that they are distinct. If $s_1 = 0$ then $\langle 2s_1 - 1 \rangle_n = n - 1$ but since $s_2 \in [\lfloor n/2 \rfloor]$ it holds that $2s_2 < n - 1$ and therefore we get again that $\langle 2s_1 - 1 \rangle_n \neq \langle 2s_2 \rangle_n$, which is a contradiction. Similar proof shows that it is impossible to have $\langle s_1 + t_1 \rangle_n = \langle s_2 + t_2 \rangle_n$ and $\langle s_1 - (t_1 + 1) \rangle_n = \langle s_2 - t_2 \rangle_n$.

c) If $\langle s_1 - t_1 \rangle_n = \langle s_2 - t_2 \rangle_n$ and $\langle s_1 + t_1 \rangle_n = \langle s_2 + t_2 \rangle_n$ then by summing those equations we deduce that $\langle 2s_1 \rangle_n = \langle 2s_2 \rangle_n$. Therefore, we deduce that $s_1 = s_2$ which is a contradiction. Similar proof shows that it is impossible to have $\langle s_1 - t_1 \rangle_n = \langle s_2 + t_2 \rangle_n$ and $\langle s_1 + t_1 \rangle_n = \langle s_2 - t_2 \rangle_n$. ∎

In this construction the result $A(n, n-1) \geqslant \lfloor n/2 \rfloor$ is shown, and since by (3), $A(n, n-1) \leqslant n/2$ it is deduced that $A(n, n-1) = \lfloor n/2 \rfloor$.

B. A Construction of $\mathcal{T}$-$(n, n, n-2)$ Codes

For convenience, a star T with a node v_i of degree $n-1$ will be denoted by T_{v_i}. The construction of a $\mathcal{T}$-$(n, n, n-2)$ code will be as follows.

Construction 2: For all $n \geqslant 4$ let $\mathcal{C}_{\mathcal{T}_2}$ be the following code

$$\mathcal{C}_{\mathcal{T}_2} = \{T_{v_i} = (V_n, E) | i \in [n]\}.$$

Clearly, the code $\mathcal{C}_{\mathcal{T}_2}$ is a set of all stars over n nodes. Next we prove that this code is a $\mathcal{T}$-$(n, n, n-2)$ code.

Theorem 47: The code $\mathcal{C}_{\mathcal{T}_2}$ is a $\mathcal{T}$-$(n, n, n-2)$ code.

Proof: Let $T_{v_i} = (V_n, E)$, $i \in [n]$, be a codeword-tree of $\mathcal{C}_{\mathcal{T}_2}$ with a node v_i of degree $n-1$. Since T_{v_i} is a star, after the erasure of $n-3$ edges from T_{v_i}, the node v_i will have degree 2 and all the nodes $v_j \in T_{v_i}$, $j \neq i$ will have degree of at most 1. Therefore the node v_i can be easily recognized and the codeword-tree T_{v_i} can be corrected. ∎

In this trivial construction we showed that $A(n, n-2) \geqslant n$ and since by Theorem 20, $A(n, n-2) \leqslant n$ it is deduced that $A(n, n-2) = n$.

C. A Construction of $\mathcal{T}$-$(n, \Omega(n^{n-2d}), d)$ Codes

In this section we show a construction of $\mathcal{T}$-$(n, \Omega(n^{n-2d}), d)$ codes for any positive integer $d \leqslant n/2$. Note that according to Corollary 19, for fixed d, $A(n, d) = \mathcal{O}(n^{n-1-d})$ and by Corollary 49 it will be deduced that $A(n, d) = \Omega(n^{n-2d})$.

For a vector $\boldsymbol{u} \in \mathbb{F}_2^m$ denote by $w_H(\boldsymbol{u})$ its Hamming weight, and for two vectors $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{F}_2^m$, $d_H(\boldsymbol{u}, \boldsymbol{w})$ is their Hamming distance. A binary code $\mathcal{C}$ of length m and size K over $\mathbb{F}_2$ will be denoted by (m, K) or (m, K, d), where d denotes its minimum Hamming distance. If $\mathcal{C}$ is also linear and k is its dimension, we denote the code by $[m, k]$ or $[m, k, d]$.

Let E_n be the set of all $\binom{n}{2}$ edges as defined in (1), with a fixed order. For any set $E \subseteq E_n$, let $\boldsymbol{v}_E$ be its characteristic vector of length $\binom{n}{2}$ which is indexed by the edge set E_n and every entry has value one if and only if the corresponding edge belongs to E. That is,

$$(\boldsymbol{v}_E)_e = \begin{cases} 1, & e \in E \\ 0, & otherwise \end{cases}.$$

The construction of $\mathcal{T}$-(n, M, d) code over trees will be as follows.

Construction 3: For all $n \geqslant 1$ let $\mathcal{C}$ be a binary code $(\binom{n}{2}, K, 2d-1)$. Then, the code $\mathcal{C}_{\mathcal{T}_3}$ is defined by

$$\mathcal{C}_{\mathcal{T}_3} = \{T \in \mathbf{T}(n) \mid \boldsymbol{v}_E \in \mathcal{C}\}.$$

Theorem 48: The code $\mathcal{C}_{\mathcal{T}_3}$ is a $\mathcal{T}$-(n, M, d) code over trees.

Proof: By Theorem 6, a code over trees $\mathcal{C}_{\mathcal{T}}$ with parameters $\mathcal{T}$-(n, M) has minimum distance d if and only if $\mathcal{C}_{\mathcal{T}}$ can correct any $d-1$ edge erasures. Notice also that since $\mathcal{C}$ is a code with Hamming distance $2d-1$, it can correct at most any $d-1$ substitutions.

Let $T = (V, E)$ be a codeword-tree of $\mathcal{C}_{\mathcal{T}_3}$ with its binary edge-vector $\boldsymbol{v}_E$. Suppose that T experienced at most $d-1$ edge erasures, generating a new forest F with the edge set E'. Since $E' \subseteq E$ and $|E'| \geqslant |E| - (d-1)$, it holds that $d_H(\boldsymbol{v}_{E'}, \boldsymbol{v}_E) \leqslant d-1$ and the vector $\boldsymbol{v}_E$ can be corrected using a decoder of $\mathcal{C}$. ∎

The next corollary summarizes the result of this construction.

Corollary 49: For positive integer n and fixed d, $A(n, d) = \Omega(n^{n-2d})$ and the redundancy is $r(n, d) \leqslant (d-1)\log(n) + \mathcal{O}(1)$.

Proof: Applying BCH codes (see Chapter 5.6 in [27]) in Construction 3 for all $n \geqslant 1$, linear codes $[\binom{n}{2}, k, 2d-1]$ are used with redundancy

$$r = (d-1)\log(\binom{n}{2}) + \mathcal{O}(1) = 2(d-1)\log(n) + \mathcal{O}(1)$$

redundancy bits. The 2^r cosets of the $\mathcal{C}$ codes are also binary $(\binom{n}{2}, 2^k, 2d-1)$ codes. Note that each tree T from $\mathbf{T}(n)$ can be mapped by Construction 3 to exactly one of these cosets. Thus, by the pigeonhole principle, there exists a code $\mathcal{C}_{\mathcal{T}_3}$ of cardinality at least

$$\frac{n^{n-2}}{2^{2(d-1)\log(n)+\mathcal{O}(1)}} = \frac{n^{n-2}}{\alpha n^{2d-2}} = \frac{1}{\alpha} n^{n-2d},$$

for some constant α. Thus, we also deduce that

$$r(n, d) \leqslant 2(d-1)\log(n) + \mathcal{O}(1).$$

∎

Remark 1: We note that the use of BCH codes can be changed to any linear codes. In fact, it is possible to use in Construction 3 a code correcting $d-1$ asymmetric errors. However, we chose to use symmetric error-correcting codes since the use of asymmetric error-correcting codes does not improve the asymptotic result and in order to derive the result in Corollary 49 we needed linear codes.

In this section we showed a family of codes with $\Omega(n^{n-2d})$ codeword-trees where $d \leqslant n/2$. Next we show a construction of codes over trees with $\Omega(n^2)$ codeword-trees where d is almost $3n/4$.

D. A Construction of $\mathcal{T}$-$(n, \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor, \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2)$ Codes

In this section, for a prime n, we show a construction of $\mathcal{T}$-$(n, \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor, \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2)$ codes, where m is a positive integer such that $3 \leqslant m \leqslant n-1$. By Corollary 49, $A(n, d) = \Omega(n^{n-2d})$ where $d \leqslant n/2$. Here we extend this result by showing that for d approaching $\lfloor 3n/4 \rfloor$, there exists a code with $\Omega(n^2)$ codeword-trees. First, several definitions are presented.

A *two-star tree* over n nodes is a tree who has exactly $n-2$ leaves. For a prime n and integers $s, t \in [n]$ where $t \neq 0$, denote the following two edge sets

$$E_{s,t}^{(+)} = \left\{ \langle v_s, v_{\langle s+it \rangle_n} \rangle | 1 \leqslant i \leqslant \frac{n+1}{2} \right\},$$

$$E_{s,t}^{(-)} = \left\{ \langle v_{\langle s+\frac{n+1}{2}t \rangle_n}, v_{\langle s+\frac{n+2j+1}{2}t \rangle_n} \rangle | 1 \leqslant j \leqslant \frac{n-3}{2} \right\}.$$
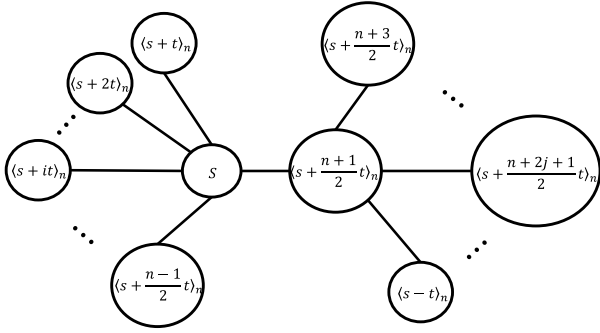
Fig. 6. A two-star tree. The nodes are marked by numbers $i \in [n]$ instead of nodes v_i. Note that by the definition of $E_{s,t}^{(-)}$, the node marked by $\langle s - t \rangle_n$ is exactly the node marked by $\langle s + \frac{n+2j+1}{2}t \rangle_n$, where $j = \frac{n-3}{2}$.

Denote by $T_{s,t} = (V_n, E_{s,t})$ the two-star tree with the edge set

$$E_{s,t} = E_{s,t}^{(+)} \cup E_{s,t}^{(-)}.$$

It is possible to verify that indeed according to this definition $T_{s,t}$ is well defined and is a two-star tree. Furthermore, It will be shown in Theorem 53 that each pair (s,t) defines a unique tree $T_{s,t}$. The nodes v_s and $v_{\langle s + \frac{n+1}{2}t \rangle_n}$ are called *the central nodes* of $T_{s,t}$. Also note that

$$\deg(v_s) = \frac{n+1}{2}, \quad \deg(v_{\langle s + \frac{n+1}{2}t \rangle_n}) = \frac{n-1}{2}. \quad (22)$$

In Fig. 6 we illustrate a two-star tree $T_{s,t}$.

For a prime n and an integer $1 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor$, where $3 \leqslant m \leqslant n-1$ and $\alpha \in \{\frac{n+1}{2}, \frac{n-1}{2}\}$, denote by $W(n,t,\alpha)$ the set

$$W(n,t,\alpha) = \{\langle t \rangle_n, \langle 2t \rangle_n, \langle 3t \rangle_n, \ldots, \langle \alpha t \rangle_n\}. \quad (23)$$

First we state the following claim.

Claim 6: For any two positive real numbers a, b such that $a < b$, the number of integers j such that $a < j \leqslant b$ is at most $\lceil b - a \rceil$.

The following lemma is now presented.

Lemma 50: Let n be a prime number, $\alpha = \frac{n+1}{2}$, and t_1, t_2 be two distinct integers $1 \leqslant t_1, t_2 \leqslant \lfloor \frac{n-1}{m} \rfloor$. Then

$$|W(n, t_1, \alpha) \cap W(n, t_2, \alpha)| < \left\lceil \frac{n}{4} \right\rceil + \left\lceil \frac{3n}{2m} \right\rceil + 1.$$

Proof: It is sufficient to prove this claim for $t_1 = 1$, since all the other cases are proved by relabeling t_1 to 1 and t_2 to $t_2 - t_1 + 1$. In this case,

$$W(n, 1, \alpha) = \left\{1, 2, \ldots, \frac{n+1}{2}\right\} = \left[\frac{n+3}{2}\right] \setminus \{0\}.$$

Thus, denote $t = t_2$ and since $0 \notin W(n, t, \alpha)$, it is sufficient to prove that for all $2 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor$,

$$\left|\left[\frac{n+3}{2}\right] \cap W(n, t, \alpha)\right| < \left\lceil \frac{n}{4} \right\rceil + \left\lceil \frac{3n}{2m} \right\rceil + 1.$$

For an integer k such that $1 \leqslant k \leqslant \frac{n+1}{2n}t$, let A_k be the set

$$A_k = \{\langle jt \rangle_n \mid (k-1)\frac{n}{t} < j \leqslant k\frac{n}{t}\}.$$

Note that $jt = (k-1)n + \langle jt \rangle_n$ and also

$$W(n, t, \alpha) = \bigcup_{k=1}^{\lfloor \frac{n+1}{2n}t \rfloor} A_k,$$

where all A_k's are mutually disjoint. Moreover, for all $1 \leqslant k \leqslant \frac{n+1}{2n}t$,

$$|A_k| \leqslant \left\lceil k\frac{n}{t} - (k-1)\frac{n}{t} \right\rceil = \lceil n/t \rceil,$$

which holds due to Claim 6. Hence,

$$\left|A_k \cap \left[\frac{n+3}{2}\right]\right| \leqslant \left|A_k \cap \left[\frac{n-1}{2}\right]\right| + 2$$

$$\overset{(a)}{\leqslant} |\{\langle jt \rangle_n \mid (k-1)\frac{n}{t} < j \leqslant (k-0.5)\frac{n}{t}\}| + 2$$

$$\overset{(b)}{\leqslant} \left\lceil (k-0.5)\frac{n}{t} - (k-1)\frac{n}{t} \right\rceil = \left\lceil \frac{n}{2t} \right\rceil + 2,$$

where (a) holds since for all $\langle jt \rangle_n \in A_k \cap [\frac{n-1}{2}]$ it holds that $0 < \langle jt \rangle_n < \frac{n-1}{2}$ and hence

$$(k-1)n < jt = (k-1)n + \langle jt \rangle_n$$

$$< (k-1)n + \frac{n-1}{2} < (k-0.5)n.$$

Equality (b) holds by Claim 6. Since $t \leqslant \lfloor \frac{n-1}{m} \rfloor$, it is deduced that

$$\left|W(n,t,\alpha) \cap \left[\frac{n+3}{2}\right]\right| \leqslant \left|\bigcup_{k=1}^{\lfloor \frac{n+1}{2n}t \rfloor} A_k \cap \left[\frac{n+3}{2}\right]\right|$$

$$\leqslant \frac{n+1}{2n}t\left(\left\lceil \frac{n}{2t} \right\rceil + 2\right)$$

$$\leqslant \frac{n+1}{2n}t\left(\frac{n}{2t} + 3\right)$$

$$= \frac{n+1}{4} + \frac{3}{2}\frac{n+1}{n}t$$

$$\leqslant \frac{n+1}{4} + \frac{3}{2}\frac{n+1}{n}\frac{n-1}{m}$$

$$= \frac{n+1}{4} + \frac{3}{2}\frac{n^2-1}{nm}$$

$$< \left\lceil \frac{n}{4} \right\rceil + \left\lceil \frac{3n}{2m} \right\rceil + 1.$$

∎

Note that this lemma holds also for $\alpha = \frac{n-1}{2}$. We state the following corollary which is derived directly from Lemma 50.

Corollary 51: Assume that W is a subset of one of the sets $W(n, t, \alpha)$, where $1 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor$ and $\alpha = \frac{n+1}{2}$. If $|W| \geqslant \lceil \frac{n}{4} \rceil + \lceil \frac{3n}{2m} \rceil + 1$, then the value of t can be uniquely determined.

This corollary holds also for $\alpha = \frac{n-1}{2}$. We proceed by introducing several more definitions. For all $1 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor$ denote the following set

$$B_t = \left\{\left\langle \frac{n+1}{2}it \right\rangle_n \mid i \in [n], i \text{ is odd}\right\}$$

and the set $A_{n,m}$ to be

$$A_{n,m} = \left\{(s,t) \mid s \in B_t, 1 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor \right\}.$$

Note that for every fixed $1 \leqslant t \leqslant \lfloor \frac{n-1}{m} \rfloor$, it holds that $|B_t| = \frac{n-1}{2}$. Thus,

$$|A_{n,m}| = \frac{n-1}{2} \cdot \left\lfloor \frac{n-1}{m} \right\rfloor. \tag{24}$$

Next, the following lemma is presented.

Lemma 52: For any $a \in [n]$, it holds that $\left(\langle a - \frac{n+1}{2} t \rangle_n, t \right) \in A_{n,m}$ if and only if $(a, t) \notin A_{n,m}$.

Proof: If $\left(\langle a - \frac{n+1}{2} t \rangle_n, t \right) \in A_{n,m}$ then $\langle a - \frac{n+1}{2} t \rangle_n \in B_t$. Therefore, there is an odd $i \in [n]$ such that

$$\left\langle a - \frac{n+1}{2} t \right\rangle_n = \left\langle \frac{n+1}{2} it \right\rangle_n.$$

Thus, $a = \langle \frac{n+1}{2}(i+1) t \rangle_n$ when $i+1$ is even. Therefore, $a \notin B_t$ which leads to $(a, t) \notin A_{n,m}$. The opposite direction is proved similarly. ∎

The construction of a $\mathcal{T}$-$(n, \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor, \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2)$ code will be as follows.

Construction 4: For a prime $n \geqslant 3$ let $\mathcal{C}_{\mathcal{T}_4}$ be the following code over trees

$$\mathcal{C}_{\mathcal{T}_4} = \{ T_{s,t} = (V_n, E_{s,t}) \mid (s, t) \in A_{n,m} \}.$$

Theorem 53: The code $\mathcal{C}_{\mathcal{T}_4}$ is a $\mathcal{T}$-$(n, \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor, \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2)$ code over trees.

Proof: First, it is deduced above in (24) that $|A_{n,m}| = \frac{n-1}{2} \cdot \lfloor \frac{n-1}{m} \rfloor$. We now prove that

$$|\mathcal{C}_{\mathcal{T}_4}| = |A_{n,m}| = \frac{n-1}{2} \cdot \left\lfloor \frac{n-1}{m} \right\rfloor.$$

It is clear that $|\mathcal{C}_{\mathcal{T}_4}| \leqslant |A_{n,m}|$ and assume in the contrary that $|\mathcal{C}_{\mathcal{T}_4}| < |A_{n,m}|$. Thus, there are two distinct pairs $(s, t), (s', t') \in A_{n,m}$ such that $T_{s,t} = T_{s',t'}$, which implies that the central nodes of $T_{s,t}$ and $T_{s',t'}$ are identical. Since $\deg(s) = \deg(s')$, the nodes v_s and $v_{s'}$ represent the same center node, so it is deduced that $s = s'$. From that, by the definition of the second central node, it is immediately implied that $t = t'$ which results with a contradiction.

Next, we show that $d = \lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 2$ by showing that it is possible to correct $\rho = d - 1$ edge erasures due to Theorem 6. Assume that ρ edges are erased in a tree $T_{s,t} \in \mathcal{C}_{\mathcal{T}_4}$. We separate the proof for two cases.

Case 1: after the erasure, both central nodes have degree of at least two, and will be denoted by v_a and v_b. If $a = s$ and $b = \langle s + \frac{n+1}{2} t \rangle_n$, then

$$\langle (a - b) \cdot 2 \rangle_n = \left\langle \left(s - (s + \frac{n+1}{2} t) \right) \cdot 2 \right\rangle_n = \langle -t \rangle_n.$$

Similarly, if $a = \langle s + \frac{n+1}{2} t \rangle_n$ and $b = s$, then

$$\langle (a - b) \cdot 2 \rangle_n = \left\langle \left((s + \frac{n+1}{2} t) - s \right) \cdot 2 \right\rangle_n = t.$$

Since $t \leqslant \lfloor \frac{n-1}{m} \rfloor$, it is deduced that $\lceil \frac{n-1}{m} \rceil < \langle -t \rangle_n \leqslant n - 1$, so only one of these options is valid and t is easily determined. Moreover, it is now determined which one of the values a or b is equal to s, and thus, $T_{s,t}$ is corrected.

Case 2: after the erasure, one of the central nodes has degree of at most one. Denote by v_a the central node with degree of at least two. Let α be a number such that if $a = s$ then

$\alpha = \frac{n+1}{2}$ and if $a = \langle s + \frac{n+1}{2} t \rangle_n$ then $\alpha = \frac{n-1}{2}$. Note that since ρ edges were erased, v_a has degree of at least

$$(n-1) - \rho - 1 = (n-1) - (\lfloor \frac{3n}{4} \rfloor - \lceil \frac{3n}{2m} \rceil - 3) - 1$$
$$= \lceil \frac{n}{4} \rceil + \lceil \frac{3n}{2m} \rceil + 1.$$

Thus, there are integers $i_1, i_2, \ldots, i_{(n-2)-\rho} \in [n]$ such that the edge set

$$E = \{ \langle v_a, v_{\langle a + i_j t \rangle_n} \rangle \mid 1 \leqslant j \leqslant (n-2) - \rho \}$$

consists of all the edges connected to v_a and were not erased. Let $W(n, t, \alpha)$ be the set defined in (23), and let W be the set

$$W = \left\{ \langle i_j t \rangle_n \in W(n, t, \alpha) \;\middle|\; \begin{array}{l} 1 \leqslant j \leqslant (n-2) - \rho, \\ \langle v_a, v_{\langle a + i_j t \rangle_n} \rangle \in E \end{array} \right\}.$$

Since $|W| = (n-2) - \rho = \lceil \frac{n}{4} \rceil + \lceil \frac{3n}{2m} \rceil + 1$, by Corollary 51, the value of t is uniquely determined. Therefore, the codeword-tree $T_{s,t}$ is either $T_{a,t}$ or $T_{\langle a - \frac{n+1}{2} t \rangle_n, t}$. By Lemma 52, it holds that $\left(\langle a - \frac{n+1}{2} t \rangle_n, t \right) \in A_{n,m}$ if and only if $(a, t) \notin A_{n,m}$. Thus, $T_{a,t} \in \mathcal{C}_{\mathcal{T}_4}$ if and only if $T_{\langle a - \frac{n+1}{2} t \rangle_n, t} \notin \mathcal{C}_{\mathcal{T}_4}$, and by finding either $T_{a,t}$ or $T_{\langle a - \frac{n+1}{2} t \rangle_n, t}$ in $\mathcal{C}_{\mathcal{T}_4}$ we find the codeword-tree $T_{s,t}$. ∎

Note that according to Theorem 53, it is possible to construct codes of cardinality $\Omega(n^2)$, while the minimum distance d approaches $\lfloor 3n/4 \rfloor$ and n is a prime number. In Theorem 23 we showed that $A(n, n-3) = \mathcal{O}(n^2)$, while from Theorem 53, $A(n, d) = \Omega(n^2)$, when d approaches $\lfloor 3n/4 \rfloor$ and n is prime. Thus, it is interesting to study the values of d for such that $A(n, d) = \Theta(n^2)$.

IX. CONCLUSION

In this article, we initiated the study of codes over trees over the tree distance. Upper bounds on such codes were presented together with specific code construction for several parameters of the number of nodes and minimum tree distance. For the tree ball of trees, it was shown that the star tree reaches the smallest size, while the maximum is achieved for the path tree. This guarantees that for a fixed value of t, the size of every ball of a tree is lower, upper-bounded from below, above by $\Omega(n^{2t})$, $\mathcal{O}(n^{3t})$, respectively. Furthermore, it was also shown that the average size of the ball is $\Theta(n^{2.5t})$. We also showed that optimal codes over trees ranged between $\mathcal{O}(n^{n-d-1})$ and $\Omega(n^{n-2d})$.

While the results in the paper provide a significant contribution in the area of codes over trees, there are still several interesting problems which are left open. Some of them are summarized as follows.

a) Improve the lower and upper bounds on the size of codes over trees, that is, the value of $A(n, d)$.

b) Find an optimal construction for $d = n - 3$.

c) Study codes over trees under different metrics such as the tree edit distance.

d) Study the problem of reconstructing trees based upon several forests in the forest ball of trees; for more details see [7].

APPENDIX A

Lemma 26.: For a positive integer n it holds that

$$\sum_{T \in \mathbf{T}(n)} V_T(n,1) = \sum_{F \in \mathbf{F}(n,2)} (V_F(n,1))^2 - (n-2)n^{n-2}.$$

Proof: The following sequence of equalities will be explained below,

$$\sum_{T \in \mathbf{T}(n)} V_T(n,1) =$$

$$\overset{(a)}{=} \sum_{T \in \mathbf{T}(n)} \left(\sum_{F \in \mathcal{P}_T(n,1)} (V_F(n,1)-1) + 1 \right)$$

$$= \sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} V_F(n,1) - \left(\sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} 1 \right) + n^{n-2}$$

$$\overset{(b)}{=} \sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} V_F(n,1) - (n-1)n^{n-2} + n^{n-2}$$

$$\overset{(c)}{=} \sum_{F \in \mathbf{F}(n,2)} \sum_{T \in \mathcal{B}_F(n,1)} V_F(n,1) - (n-1)n^{n-2} + n^{n-2}$$

$$= \sum_{F \in \mathbf{F}(n,2)} (V_F(n,1))^2 - (n-2)n^{n-2}.$$

In equality (a) we explain why

$$V_T(n,1) - 1 = |\mathcal{B}_T(n,1) \setminus \{T\}| = \sum_{F \in \mathcal{P}_T(n,1)} (V_F(n,1)-1).$$

Note that for all $T, T' \in \mathbf{T}(n)$ such that $d_T(T, T') = 1$, there exists exactly one forest $F \in \mathbf{F}(n,2)$ such that $F \in \mathcal{P}_T(n,1) \cap \mathcal{P}_{T'}(n,1)$. Thus, each $T' \in \mathcal{B}_T(n,1) \setminus \{T\}$ can be generated from T uniquely by removing and adding exactly one edge. Equivalently, each such a tree is counted by adding an edge to a forest $F \in \mathcal{P}_T(n,1)$. By doing so for all forests in $\mathcal{P}_T(n,1)$, while subtracting 1 for the tree T, equality (a) holds. In equality (b) it is deduced that

$$\sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} 1 = (n-1)n^{n-2}.$$

Lastly, in equality (c), by taking pairs of trees and forests, the order of summation is changed,

$$\sum_{T \in \mathbf{T}(n)} \sum_{F \in \mathcal{P}_T(n,1)} 1 = \sum_{F \in \mathbf{F}(n,2)} \sum_{T \in \mathcal{B}_F(n,1)} 1.$$

∎

APPENDIX B

Theorem 33.: For any $T \in \mathbf{T}(n)$ it holds that

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i) = n^{t-1} \sum_{(i_0,i_1,\ldots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t.$$

Proof: By definition, for $t \geq 1$, $S_T(n,t) = V_T(n,t) - V_T(n,t-1)$. Thus,

$$\sum_{i=0}^{t} \binom{n-1-t+i}{i} S_T(n,t-i) =$$

$$\sum_{i=0}^{t-1} \binom{n-1-t+i}{i} \left(V_T(n,t-i) - V_T(n,t-1-i) \right)$$

$$+ \binom{n-1}{t} V_T(n,0) = V_T(n,t)$$

$$+ \sum_{i=1}^{t} V_T(n,t-i) \left(\binom{n-1-t+i}{i} - \binom{n-2-t+i}{i-1} \right)$$

$$\overset{(a)}{=} V_T(n,t) + \sum_{i=1}^{t} V_T(n,t-i) \binom{n-2-t+i}{i}$$

$$= \sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i),$$

where (a) holds by the identity $\binom{n}{k} + \binom{n+1}{k} = \binom{n+1}{k+1}$. Using the result of Corollary 32, we conclude the proof. ∎

APPENDIX C

Lemma 34.: For any positive integer α, if

$$\sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i) = \Omega(n^{\alpha t}),$$

and $V_T(n,0) = 1$, then $V_T(n,t) = \Omega(n^{\alpha t})$.

Proof: This lemma is proved by induction on t.
Base: for $t = 0$, $V_T(n,0) = n^0 = 1$ which is true by the definition.
Inductive Step: suppose that the lemma holds for all $0 \leq t' \leq t-1$. Thus,

$$\Omega(n^{\alpha t}) = \sum_{i=0}^{t} \binom{n-2-t+i}{i} V_T(n,t-i)$$

$$= V_T(n,t) + \sum_{i=1}^{t} \binom{n-2-t+i}{i} V_T(n,t-i)$$

$$= V_T(n,t) + \sum_{i=1}^{t} \binom{n-2-t+i}{i} \Omega(n^{\alpha(t-i)})$$

$$= V_T(n,t) + \sum_{i=1}^{t} \Omega(n^i) \Omega(n^{\alpha(t-i)})$$

$$= V_T(n,t) + \Omega(n^{\alpha(t-1)+1}).$$

Therefore we deduce that

$$V_T(n,t) = \Omega(n^{\alpha t}) - \Omega(n^{\alpha(t-1)+1}) = \Omega(n^{\alpha t}).$$

∎

APPENDIX D

Claim 4.: For a positive integer n and a fixed t it holds that

$$\sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i} \Theta(i^{t/2}) = \Theta(n^{t/2}) \sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i}.$$

Proof: The upper bound is derived immediately,

$$\sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i} \Theta(i^{t/2}) = \mathcal{O}(n^{t/2}) \sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i}.$$

Next, the lower bound is proved by,

$$\sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i} \Omega(i^{t/2})$$

$$\geqslant \sum_{i=\lfloor \frac{n-1}{2} \rfloor}^{n-1} \binom{n}{i} i^i (n-i)^{n-i} \Omega(i^{t/2})$$

$$= \Omega(n^{t/2}) \sum_{i=\lfloor \frac{n-1}{2} \rfloor}^{n-1} \binom{n}{i} i^i (n-i)^{n-i}$$

$$= \Omega(n^{t/2}) \sum_{i=1}^{n-1} \binom{n}{i} i^i (n-i)^{n-i}.$$

■

APPENDIX E

Theorem 40.: The size of the sphere for a star satisfies

$$S^\star(n,t) = \binom{n-1}{t} (n-1)^{t-1}(n-t-1),$$

and the size of the tree ball of trees for a star satisfies

$$V^\star(n,t) = \sum_{j=0}^{t} \binom{n-1}{j} (n-1)^{j-1}(n-j-1).$$

Proof: Let $T \in \mathbf{T}(n)$ be a star tree, and denote the function

$$H(n,t) = \binom{n-1}{t} (n-1)^{t-1}(n-t-1).$$

We say that $\frac{d}{dn}(f(n))$ is the derivative of $f(n)$ with respect to n. Thus,

$$\sum_{i=0}^{t} \binom{n-1-t+i}{i} H(n,t-i)$$

$$= \sum_{i=0}^{t} \binom{n-1-t+i}{i} \binom{n-1}{t-i} (n-1)^{t-1-i}(n-t-1+i)$$

$$\stackrel{(a)}{=} \sum_{i=0}^{t} \binom{n-1}{t} \binom{t}{i} (n-1)^{t-1-i}(n-t-1+i)$$

$$= \binom{n-1}{t} \sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-1-i}(n-t-1+i)$$

$$= \binom{n-1}{t} \Big(\sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-1-i}(n-1)$$

$$\qquad - \sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-1-i}(t-i) \Big)$$

$$= \binom{n-1}{t} \Big(\sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-i} - \frac{d}{dn}\Big(\sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-i} \Big) \Big)$$

$$\stackrel{(b)}{=} \binom{n-1}{t} \Big(n^t - t n^{t-1} \Big) = \binom{n-1}{t} n^{t-1}(n-t)$$

$$\stackrel{(c)}{=} n^{t-1} \sum_{(i_0,i_1,\dots,i_t) \in P_T(n,t)} i_0 i_1 \cdots i_t,$$

where (a) holds by known formula

$$\binom{a-(b-c)}{c} \binom{a}{b-c} = \binom{a}{b} \binom{b}{c}$$

(i.e. $a = (n-1), b = t$, and $c = i$), and (b) holds by the binomial theorem, which is,

$$\sum_{i=0}^{t} \binom{t}{i} (n-1)^{t-i} = (n-1+1)^t = n^t.$$

Equality (c) holds due to (17). Thus, by Corollary 32, it is deduced that $S^\star(n,t) = H(n,t)$. Next,

$$V^\star(n,t) = \sum_{j=0}^{t} \binom{n-1}{j} (n-1)^{j-1}(n-j-1),$$

which is derived by the fact that for every $T \in \mathbf{T}(n)$,

$$V_T(n,t) = \sum_{i=0}^{t} S_T(n,i).$$

■

APPENDIX F

Claim 5.: The following properties hold

a) It holds that

$$\sum_{(c_0,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell)} c_\ell \cdots c_t$$

$$= \sum_{(c_0,\dots,c_t) \in A_T(n,t;\boldsymbol{v}_\ell,v_y)} c_\ell \cdots c_t + \sum_{(c_0,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_y))} c_\ell \cdots c_t.$$

b) If v_x is not in $\boldsymbol{v}_\ell$ then,

$$\sum_{(c_0,\dots,c_t) \in A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)} c_\ell \cdots c_t = \sum_{(c_0',\dots,c_t') \in A_T(n,t;\boldsymbol{v}_\ell,v_x)} c_\ell' \cdots c_t'.$$

c) If v_x is not in $\boldsymbol{v}_\ell$ then,

$$\sum_{(c_0,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} c_\ell \cdots c_t$$

$$= \sum_{(c_0',\dots,c_t') \in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} (c_\ell'+1) c_{\ell+1}' \cdots c_t'$$

$$+ \sum_{(c_0,\dots,c_{\ell-1},1,c_{\ell+1},\dots,c_t) \in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1 \cdot c_{\ell+1} \cdots c_t.$$

d) If $v_x = v_{i_{\ell-1}}$ then

$$\sum_{(c_0,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell \cdots c_t$$

$$= \sum_{(c_0',\dots,c_t') \in P_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y))} c_\ell' \cdots c_t'$$

$$+ \sum_{(c_0,\dots,c_{\ell-2},1,c_\ell,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell \cdots c_t.$$

Proof: a) By definition of the set $A_T(n,t;\boldsymbol{v}_\ell,v_y)$ it holds that $A_T(n,t;\boldsymbol{v}_\ell,v_y) \subseteq P_T(n,t;\boldsymbol{v}_\ell)$. Moreover,

$$\sum_{(c_0,c_1,\dots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell) \setminus A_T(n,t;\boldsymbol{v}_\ell,v_y)} c_\ell \cdots c_t$$

$$= \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_y))} c_\ell \cdots c_t,$$

and the proof is concluded.

b) Again, $A_T(n,t;\boldsymbol{v}_\ell,v_x) \subseteq P_T(n,t;\boldsymbol{v}_\ell)$. Since v_x is not in $\boldsymbol{v}_\ell$ it holds that v_x and v_y are always in the same connected component with respect to $A_T(n,t;\boldsymbol{v}_\ell,v_x)$, and thus, $|A_T(n,t;\boldsymbol{v}_\ell,v_x)| = |A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)|$. Moreover, since there is an index $j \in [\ell]$ such that $v_x,v_y \in C_j$ in T, it holds that $(c_0,\ldots,c_j,\ldots,c_t) \in A_T(n,t;\boldsymbol{v}_\ell,v_x)$ if and only if $(c_0,\ldots(c_j-1),\ldots,c_t) \in A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)$. Hence, this difference does not affect the equality, which concludes this proof.

c) Assume that v_x and v_y are in the same connected component C_ℓ with respect to T. In this case since $v_x,v_y \in C_\ell$, it holds that $(c_0,\ldots,c_\ell,\ldots,c_t) \in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))$ if and only if $(c_0,\ldots,c_\ell-1,\ldots,c_t) \in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))$. Thus, the following expression

$$\sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} (c_\ell'+1)c_{\ell+1}' \cdots c_t'$$

corresponds to all cases where the edge $\langle v_x,v_y \rangle$ was not removed, and

$$\sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1 \cdot c_{\ell+1} \cdots c_t,$$

corresponds to all cases where the edge $\langle v_x,v_y \rangle$ was removed. Hence, the sum of the two expressions equals to

$$\sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} c_\ell \cdots c_t.$$

d) Assume that v_x and v_y are in the same connected component $C_{\ell-1}$ with respect to T. Since $v_x,v_y \in C_{\ell-1}$, it holds that $(c_0,\ldots,c_{\ell-1},\ldots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell(v_x))$ if and only if $(c_0,\ldots,c_{\ell-1}-1,\ldots,c_t) \in P_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y))$. Thus, the following expression

$$\sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y))} c_\ell' \cdots c_t',$$

corresponds to all cases where the edge $\langle v_x,v_y \rangle$ was not removed, and

$$\sum_{(c_0,\ldots,c_{\ell-2},1,c_\ell,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell \cdots c_t.$$

corresponds to all cases where the edge $\langle v_x,v_y \rangle$ was removed. Again we get that the sum of the two expressions equals to

$$\sum_{(c_0,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell \cdots c_t.$$

∎

Appendix G

Lemma 42.: If v_x is not in $\boldsymbol{v}_\ell$ then,

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell)$$
$$+ f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell).$$

If v_x is in $\boldsymbol{v}_\ell$, and without loss of generality $v_x = v_{i_{\ell-1}}$, then

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1}).$$

Proof: In this proof, it is assumed that v_y is not in $\boldsymbol{v}_\ell$, although the proof is valid also for this case, where by the definition $f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y)) = 0$. First we prove the case where v_x is not in $\boldsymbol{v}_\ell$. In this case, we have that

$$f_{T_1}(n-1,t;\boldsymbol{v}_\ell) + f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))$$

$$\overset{(a)}{=} \sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_\ell)} c_\ell' \cdots c_t'$$

$$+ \sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} c_{\ell+1}' \cdots c_t'$$

$$\overset{(b)}{=} \sum_{(c_0',c_1',\ldots,c_t')\in A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)} c_\ell' \cdots c_t'$$

$$+ \sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} c_\ell' \cdots c_t'$$

$$+ \sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} c_{\ell+1}' \cdots c_t'$$

$$= \sum_{(c_0',c_1',\ldots,c_t')\in A_{T_1}(n-1,t;\boldsymbol{v}_\ell,v_y)} c_\ell' \cdots c_t'$$

$$+ \sum_{(c_0',c_1',\ldots,c_t')\in P_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))} (c_\ell'+1) \cdots c_t'$$

$$\overset{(c)}{=} \sum_{(c_0,c_1,\ldots,c_t)\in A_T(n,t;\boldsymbol{v}_\ell,v_x)} c_\ell \cdots c_t$$

$$+ \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} c_\ell \cdots c_t$$

$$- \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1 \cdot c_{\ell+1} \cdots c_t$$

$$\overset{(d)}{=} \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell)} c_\ell \cdots c_t$$

$$- \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1 \cdot c_{\ell+1} \cdots c_t.$$

Equality (a) holds by definition of the function f_{T_1}. Equality (b) holds due to Claim 5(a). Equality (c) holds by Claim 5(b) and (c). Equality (d) holds again due to Claim 5(a). Next we show that

$$f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)$$

$$= \sum_{(c_0',\ldots c_{\ell-1}',c_{\ell+1}',\ldots,c_t')\in P_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)} c_{\ell+1}' \cdots c_t'$$

$$\overset{(a)}{=} \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} c_{\ell+1} \cdots c_t,$$

where (a) holds since $(c_0,\ldots c_{\ell-1},c_{\ell+1},\ldots,c_t) \in P_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)$ if and only if $(c_0,\ldots c_{\ell-1},1,c_{\ell+1},\ldots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell)$. Thus,

$$f_{T_1}(n-1,t;\boldsymbol{v}_\ell) + f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y))$$
$$+ f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)$$

$$= \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell)} c_\ell \cdots c_t$$

$$- \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1\cdot c_{\ell+1}\cdots c_t$$

$$+ \sum_{(c_0,\ldots,c_{\ell-1},1,c_{\ell+1},\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_{\ell+1}(v_x))} 1\cdot c_{\ell+1}\cdots c_t$$

$$= \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell)} c_\ell\cdots c_t = f_T(n,t;\boldsymbol{v}_\ell).$$

Similarly, if v_x is in $\boldsymbol{v}_\ell$ and $v_x = v_{i_{\ell-1}}$,

$$f_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1})$$

$$\overset{(a)}{=} \sum_{(c'_0,c'_1,\ldots,c'_t)\in P_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y))} c'_\ell\cdots c'_t$$

$$+ \sum_{(c'_0,\ldots,c'_{\ell-2},c'_\ell,\ldots,c'_t)\in P_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1})} c'_\ell\cdots c'_t$$

$$\overset{(b)}{=} \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t$$

$$- \sum_{(c_0,\ldots,c_{\ell-2},1,c_\ell,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t$$

$$+ \sum_{(c_0,\ldots,c_{\ell-2},1,c_\ell,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t$$

$$= \sum_{(c_0,c_1,\ldots,c_t)\in P_T(n,t;\boldsymbol{v}_\ell(v_x))} c_\ell\cdots c_t$$

$$= f_T(n,t;\boldsymbol{v}_\ell(v_x)) = f_T(n,t;\boldsymbol{v}_\ell),$$

where equality (a) holds by the definition of f_{T_1}, equality (b) holds due to Claim 5(d), and since $(c_0,\ldots,c_{\ell-2},c_\ell,\ldots,c_t) \in P_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1})$ if and only if $(c_0,\ldots,c_{\ell-2},1,c_\ell,\ldots,c_t) \in P_T(n,t;\boldsymbol{v}_\ell(v_x))$. ∎

APPENDIX H

Lemma 43.: For any tree $T \in \mathbf{T}(n), n \geqslant 1$ and a vector of $0 \leqslant \ell \leqslant t+1 \leqslant n$ nodes $\boldsymbol{v}_\ell = (v_{i_0}, v_{i_1},\ldots,v_{i_{\ell-1}})$,

$$f_T(n,t;\boldsymbol{v}_\ell) \leqslant \binom{n+t-\ell}{2t+1-\ell}.$$

Proof: Note that if $\ell = t+1$ by the definition of f_T

$$f_T(n,t;\boldsymbol{v}_{t+1}) \leqslant \binom{n-1}{t} = \binom{n+t-(t+1)}{2t+1-(t+1)}.$$

As showed in (19), if $n = t+1$, then

$$f_T(n,n-1;\boldsymbol{v}_\ell) = 1 = \binom{n+(n-1)-\ell}{2(n-1)+1-\ell}, \quad (25)$$

so the lemma is correct for this two cases. Thus it is left to prove the cases where $0 \leqslant \ell < t+1 < n$, and it will be shown by the induction on $n \geqslant 1$.

Base: immediately derived from (25).

Inductive Step: assume that for any tree $T \in \mathbf{T}(n-1), n \geqslant 1$ and a vector of $1 \leqslant \ell \leqslant t+1 \leqslant n-1$ nodes $\boldsymbol{v}_\ell = (v_{i_0}, v_{i_1},\ldots,v_{i_{\ell-1}})$,

$$f_T(n-1,t;\boldsymbol{v}_\ell) \leqslant \binom{n-1+t-\ell}{2t+1-\ell}.$$

Let $T \in \mathbf{T}(n)$ and let v_x be a leaf connected to a node denoted by v_y. Assume that $T_1 \in \mathbf{T}(n-1)$ is the tree generated by

removing v_x from T. For two integers t and ℓ such that $0 \leqslant \ell < t+1 < n$, let $\boldsymbol{v}_\ell = (v_{i_0}, v_{i_1},\ldots,v_{i_{\ell-1}})$ be a vector of ℓ nodes in T. If v_x is not in $\boldsymbol{v}_\ell$, using Lemma 42 and the induction assumption, we deduce that

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell)$$
$$+ f_{T_1}(n-1,t;\boldsymbol{v}_{\ell+1}(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_\ell)$$
$$\leqslant \binom{n-1+t-\ell}{2t+1-\ell} + \binom{n-1+t-\ell-1}{2t+1-\ell-1}$$
$$+ \binom{n-1+t-1-\ell}{2t-1-\ell}$$
$$= \binom{n-1+t-\ell}{2t+1-\ell} + \binom{n-1+t-\ell}{2t-\ell}$$
$$= \binom{n+t-\ell}{2t+1-\ell},$$

where each equality holds by the identity $\binom{n}{k}+\binom{n+1}{k} = \binom{n+1}{k+1}$. Similarly, if $v_x \in \boldsymbol{v}_\ell$, and without loss of generality $v_x = v_{i_{\ell-1}}$, then

$$f_T(n,t;\boldsymbol{v}_\ell) = f_{T_1}(n-1,t;\boldsymbol{v}_\ell(v_y)) + f_{T_1}(n-1,t-1;\boldsymbol{v}_{\ell-1})$$
$$\leqslant \binom{n-1+t-\ell}{2t+1-\ell} + \binom{n-1+t-1-\ell+1}{2t-1-\ell+1}$$
$$= \binom{n+t-\ell}{2t+1-\ell}.$$

∎

REFERENCES

[1] M. Aigner and G. M. Ziegler, *Proofs From the Book*. New York, NY, USA: Springer-Verlag, 1998, pp. 141–146.

[2] T. Amdeberhan, D. Callan, and V. Moll, "Valuations and combinatorics of truncated exponential sums," in *Integers, Electronics Paper A21*. 2013. [Online]. Available: https://www.researchgate.net/publication/306129064_Valuations_and_combinatorics_of_truncated_exponential_sums/references

[3] J. M. Barnard, "Substructure searching methods: Old and new," *J. Chem. Inf. Comput. Sci.*, vol. 33, no. 4, pp. 532–538, Jul. 1993.

[4] B. Bollobas, *Graph Theory: An Introductory Course*. New York, NY, USA: Springer-Verlag, 1979.

[5] Y.-C. Cheng and S.-Y. Lu, "Waveform correlation by tree matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-7, no. 3, pp. 299–305, May 1985.

[6] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 1990.

[7] S. Davies, M. Z. Racz, and C. Rashtchian, "Reconstructing trees from traces," in *Proc. Conf. Learn. Theory (COLT)*, 2019, pp. 961–978.

[8] P. Flajolet, P. Grabner, P. Kirschenhofer, and H. Prodinger, "On Ramanujan's Q-function," *J. Comput. Appl. Math.*, vol. 58, pp. 103–116, Mar. 1995.

[9] L. Ganassali and L. Massoulié, "From tree matching to sparse graph alignment," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, Jul. 2020, pp. 1633–1665.

[10] J. Gottlieb, B. A. Julstrom, G. R. Raidl, and F. Rothlauf, "Prüfer numbers: A poor representation of spanning trees for evolutionary search," in *Proc. Genet. Evol. Comput. Conf.*, San Francisco, CA, USA, 2001, pp. 343–350.

[11] N. Hartsfield and G. Ringel, "Hamilton surfaces for the complete symmetric tripartite graph," *Archiv der Math.*, vol. 50, no. 5, pp. 470–473, May 1988.

[12] T. Jiang, L. Wang, and K. Zhang, "Alignment of trees—An alternative to tree edit," *Theor. Comput. Sci.*, vol. 143, no. 1, pp. 137–148, Jul. 1995.

[13] D. E. Knuth, *The Art of Computer Programming Volume 3: Sorting and Searching*. Reading, MA, USA: Addison-Wesley, 1973.

[14] M. Krishnamoorthy and A. T. Ernst, "Comparison of algorithms for the degree constrained minimum spanning tree," *J. Heuristics*, vol. 7, no. 6, pp. 587–611, 2001.

[15] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, no. 1, pp. 48–50, Feb. 1956.

[16] J. Llados, E. Marti, and J. J. Villanueva, "Symbol recognition by error-tolerant subgraph matching between region adjacency graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1137–1143, Oct. 2001.

[17] E. Lucas, "Les Rondes enfantines," in *Rècrèations Mathèmatiques*, vol. 2. Paris, France, 1894. [Online]. Available: https://link.springer.com/article/10.1007/BF02960735

[18] J. W. Moon, *Counting Labeled Trees*. 1970. [Online]. Available: https://books.google.co.il/books/about/Counting_Labelled_Trees_by_J_W_Moon.html?id=B6dDcgAACAAJ&redir_esc=y

[19] B. Moayer and K.-S. Fu, "A tree system approach for fingerprint pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 3, pp. 376–387, May 1986.

[20] I. Neuwirth, "The size of bipartite graphs with girth eight," 2001, *arXiv:math/0102210*. [Online]. Available: https://arxiv.org/abs/math/0102210

[21] T. Paulden and D. K. Smith, "Some novel locality results for the blob code spanning tree representation," in *Proc. 9th Annu. Conf. Genet. Evol. Comput. (GECCO)*, London, U.K., 2007, pp. 1320–1327.

[22] R. C. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. J.*, vol. 36, no. 6, pp. 1389–1401, Nov. 1957.

[23] G. R. Raidl, "An efficient evolutionary algorithm for the degree-constrained minimum spanning tree problem," in *Proc. Congr. Evol. Comput.*, Jun. 2000, pp. 104–111.

[24] G. R. Raidl and B. A. Julstrom, "Edge sets: An effective evolutionary coding of spanning trees," *IEEE Trans. Evol. Comput.*, vol. 7, no. 3, pp. 225–239, Jun. 2003.

[25] G. R. Raidl and B. A. Julstrom, "Greedy heuristics and an evolutionary algorithm for the bounded-diameter minimum spanning tree problem," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2003, pp. 747–752.

[26] I. Reiman, "Über ein problem von K. Zarankiewicz," *Acta Mathematica Hungarica*, vol. 9, nos. 3–4, pp. 269–273, Sep. 1958.

[27] R. M. Roth, *Introduction to Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[28] S. M. Selkow, "The tree-to-tree editing problem," *Inf. Process. Lett.*, vol. 6, no. 6, pp. 184–186, Dec. 1977.

[29] R. Shamir and D. Tsur, "Faster subtree isomorphism," *J. Algorithms*, vol. 33, no. 2, pp. 267–280, Nov. 1999.

[30] B. A. Shapiro and K. Zhang, "Comparing multiple RNA secondary structures using tree comparisons," *Comput. Appl. Biosci.*, vol. 6, no. 4, pp. 309–318, Oct. 1990.

[31] K.-C. Tai, "The tree-to-tree correction problem," *J. ACM*, vol. 26, no. 3, pp. 422–433, Jul. 1979.

[32] E. Tanaka and K. Tanaka, "The tree-to-tree editing problem," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 2, no. 2, pp. 221–240, Jun. 1988.

[33] L. M. G. M. Tolhuizen, "The generalized Gilbert-Varshamov bound is implied by Turan's theorem [code construction]," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1605–1606, Sep. 1997.

[34] G. Valiente, "An efficient bottom-up distance between trees," in *Proc. 8th Symp. String Process. Inf. Retr.*, Laguna de San Rafael, Chile, Nov. 2001, pp. 212–219.

[35] G. Zhou and M. Gen, "Approach to the degree-constrained minimum spanning tree problem using genetic algorithm," *Eng. Des. Autom.*, vol. 3, no. 2, pp. 157–165, 1997.

Lev Yohananov (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the Technion—Israel Institute of Technology, Haifa, Israel, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include algebraic error-correction codes, codes over graphs, and combinatorics.

Eitan Yaakobi (Senior Member, IEEE) received the B.A. degree in computer science and mathematics and the M.Sc. degree in computer science from the Technion—Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California at San Diego, San Diego, in 2011. From 2011 to 2013, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, California Institute of Technology, and the Center for Memory and Recording Research, University of California at San Diego. He is currently an Associate Professor with the Department of Computer Science, Technion—Israel Institute of Technology. His research interests include information and coding theory with applications to non-volatile memories, associative memories, DNA storage, data storage and retrieval, and private information retrieval. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship from 2010 to 2011.