Codes over Trees

Lev Yohananov
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 3200003, Israel
Email: levyohananov@campus.technion.ac.il

Eitan Yaakobi
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 3200003, Israel
Email: yaakobi@cs.technion.ac.il

***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 *line tree*.**

I. Introduction

In graph theory, a *tree* is a special code 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. In signal processing trees are used for the representation of waveforms [4]. 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 cyber applications trees are used to represent fingerprint patterns [14]. One of the biology applications includes the tree-matching algorithm to compare between trees in order to analyze multiple RNA secondary structures [22]. Trees are used in the subgraph isomorphism problem which, among its very applications, is used for chemical substructure searching [2].

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 [23], top-down distance [21], alignment distance [9], isolated-subtree distance [24], and bottom-up distance [26]. 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 to this work as the *tree distance*, 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 into 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 paper 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.

One of the classical problems in graph theory is finding a minimal spanning tree (MST) for a given graph. While the MST problem is solved in polynomial time [11], [16], it may become NP-hard under some constraints. For example, in the degree-constrained MST problem (d-MST) [10], [17], [18], [28], 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 $D \geqslant 4$ [19]. One of the common approaches for solving such problems uses *evolution algorithms* (*EA*). 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 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 [7] 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, while the largest for a line. In this paper, we establish this result for any number of nodes in the tree. 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 line tree), while the average size of the ball is $\Theta(n^{2.5t})$.

Motivated by the coding theory approach, in this work we apply the tree distance, which is a metric, in order to study *codes over trees* with a prescribed minimum tree distance. This family of codes can be used for the correction of edge erasures and errors. In Section II, we formally define the tree distance and codes over trees as well as several more useful definitions and properties. An *edge erasure* is the event in which one of the edges in the tree is erased and a forest is received with two sub-trees. This is also extended to the erasure of multiple edges. If t edges are erased, then a forest with $t+1$ sub-trees is received and the number of such forests is $\binom{n-1}{t}$. In Section III, by using several known results on the number of forests with a fixed number of sub-trees, 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 IV, we study balls of trees. The *tree ball of trees* of a given tree T consists of all trees for which their tree distance from T is at most some fixed radius t. It is observed that these balls are not regular and in Section IV, it is shown that the size of the star, line tree ball is $\Theta(n^{2t}), \Theta(n^{3t})$, respectively, while the average size of the ball is $\Theta(n^{2.5t})$. Lastly, in Section V, for a fixed d we show a construction of codes over trees of size $\Omega(n^{n-2d})$. Finally, Section VI concludes the paper. Due to the lack of space, some of the proofs in the paper are omitted. These proofs and several more results can be found in the full version of this work in [27].

II. Definitions and Preliminaries

Let $G = (V_n, E)$ be a graph, where $V_n = \{v_0, \ldots, v_{n-1}\}$ 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 paper, 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 $\binom{n}{2}$ 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 $\mathbb{T}(n)$. By Cayley's formula [1] it holds that $|\mathbb{T}(n)| = n^{n-2}$. 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 t trees will be denoted by $\mathbb{F}(n,t)$. We sometimes use the notation $\{C_0, C_1, \ldots, C_{t-1}\} = F \in \mathbb{F}(n,t)$ to explicitly denote a forest with t sub-trees (or connected components) of F. Note that $\mathbb{F}(n,1) = \mathbb{T}(n)$.

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

The definition of the redundancy follows from the observation that $\mathbb{T}(n) = n^{n-2}$. For each codeword-tree, a topology and an arrangement of the nodes is unique information that we want to store or send and read, even in the presence of errors. For this purpose, *erasures* and *errors* in trees are defined.

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

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 $d_{\mathcal{T}}(T_1, T_2) = |E_1 \setminus E_2| = |E_2 \setminus E_1|$. The following lemma is easily proved.

Lemma 4. *The tree distance is a metric.*

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

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

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

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

Theorem 6. *A $\mathcal{T}$-(n, M) code over trees $\mathcal{C}_{\mathcal{T}}$ is of tree distance 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.*

Lastly, 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 $r(n,d) = (n-2)\log(n) - \log(A(n,d))$.*

III. Bounds on Codes over Trees

In this section we show upper bounds for codes over trees. We start with several definitions. Denote by $F(n,d)$ the size of $\mathbb{F}(n,d)$, i.e., the number of forests with n nodes and d sub-trees. The value of $F(n,d)$ was shown in [13], to be

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

or another representation of it in [3],

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

The next corollary summarizes some of these known results.

Corollary 8. *The following properties hold for all n.*

1) $F(n,1) = n^{n-2}$,
2) $F(n,2) = \frac{1}{2} n^{n-4}(n-1)(n+6)$,
3) $F(n,3) = \frac{1}{8} n^{n-6}(n-1)(n-2)(n^2+13n+60)$,
4) $F(n,n-4) = \frac{1}{16}\binom{n}{4}(n^2+3n+10)(n-4)(n+3)$,
5) $F(n,n-3) = \frac{1}{2}\binom{n}{4}(n^2+3n+4)$,
6) $F(n,n-2) = 3\binom{n+1}{4}$,
7) $F(n,n-1) = \binom{n}{2}$,
8) $F(n,n) = 1$.

For each $T = (V_n, E) \in \mathbb{T}(n)$, and $0 \leqslant t \leqslant n-1$, denote by $\mathsf{E}_T(n,t)$ the set

$$\mathsf{E}_T(n,t) = \{E' \mid |E'| = t, E' \subseteq E\}, \tag{2}$$

where clearly $|\mathsf{E}_T(n,t)| = \binom{n-1}{t}$. For each $T \in \mathbb{T}(n), E' \in \mathsf{E}_T(n,t)$, denote the forest $F_{T,E'} = (V_n, E \setminus E')$, where $F_{T,E'} \in \mathbb{F}(n,t+1)$.

Definition 9. *The forest ball of a tree $T = (V_n, E)$ of radius t in $\mathbb{F}(n,t+1)$ is defined to be*

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

A. Sphere-Packing Bound

The following theorem proves the sphere packing bound for codes over trees, while the proof can be found in [27].

Theorem 10. *For all $n \geqslant 1$ and $1 \leqslant d \leqslant n$, it holds that $A(n,d) \leqslant F(n,d)/\binom{n-1}{d-1}$.*

It was also proved in [13] that for any fixed d,

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

which immediately implies the following corollary.

Corollary 11. *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 8(7) it holds that

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

In Section V 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 8(6),

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

however, we will next show how to improve this bound such that $A(n,n-2) \leqslant n$. In Section V, 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 8(5),

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

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.

B. An Improved Upper Bound for $A(n,n-2)$ and $A(n,n-3)$

Before we show the improved upper bounds for $A(n,n-2)$ and $A(n,n-3)$, a few more definitions are presented. 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 node sets U and V if $U \cap V = \emptyset$ 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 [15] and [20] states that if $|V| \leqslant |U|$, then every bipartite graph $G = (U \cup V, \mathcal{E})$ with girth at least 6 satisfies

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

According to Theorem 10, $A(n,n-2) \leqslant \frac{1}{2}\binom{n+1}{2}$ and in the next theorem this bound will be improved to be $A(n,n-2) \leqslant n$.

Theorem 12. *For all positive integer 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, i.e., the girth of G is at least six. 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 stated in (6) will be used next. Since $|V| = M, |U| = \binom{n}{2}$ and $|\mathcal{E}| = M(n-1)$,

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

and it can be verified that it holds if and only if $M \leqslant n$. ■

As mentioned above, in Section V we will show that $A(n,n-2) = n$. Next, we showed in (5) that $A(n,n-3) \leqslant \frac{1}{8}n(n^2+3n+4) = \mathcal{O}(n^3)$. Using similar techniques, we state the following theorem while its proof is shown in [27].

Theorem 13. *For all n, $A(n,n-3) \leqslant 1.5n^2$.*

IV. Tree Balls

In Section III, 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 introduce several more ball definitions and study their size behavior. These results will also be used to apply the generalized Gilbert Varshamov bound [25] on codes over trees. We start from some definitions.

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 *line tree* (or a *line* in short) 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 14. *The tree ball of a tree of radius t in $\mathbb{T}(n)$ centered at $T \in \mathbb{T}(n)$ is defined to be*

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

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

We define the *average ball size* to be the average value of all tree ball of trees, that is,

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

Let us remind here the definition of the forest ball of a tree $\mathcal{P}_T(n,t)$ from Definition 9 and the set $\mathsf{E}_T(n,t)$ as defined in (2). Given a tree T and an edge-set $E' \in \mathsf{E}_T(n,t)$, 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 $|C_0| \leqslant |C_1| \leqslant \cdots \leqslant |C_t|$. 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

$$P_T(n,t) = \{\mathbf{P}_T(E') \mid E' \in \mathsf{E}_T(n,t)\}.$$

Notice that $|P_T(n,t)| = |\mathcal{P}_T(n,t)| = |\mathsf{E}_T(n,t)| = \binom{n-1}{t}$.

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 stated in the next lemma.

Lemma 15. *For any $T \in \mathbb{T}(n)$ it holds that*

$$V_T(n,1) = \sum_{(i,n-i) \in P_T(n,1)} \Big(i(n-i)-1\Big) + 1. \tag{7}$$

Note that $V_T(n,t)$ depends on the choice of its center T. For example, it can be shown that if T is a star then $V_T(n,1) = (n-1)(n-2)+1$ and if T is a line tree, then $V_T(n,1) = (n-1)(n-2)(n+3)/6+1$. If T is a star, line the size of $V_T(n,t)$ is denoted by $V^\star(n,t), V^-(n,t)$, respectively.

Our next goal is to show that for any $T \in \mathbb{T}(n)$ 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 the following lemma is stated.

Lemma 16. *For all $T \in \mathbb{T}(n)$,*

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

Next, we show the following important result.

Theorem 17. *For any $T \in \mathbb{T}(n)$ 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 $T \in \mathbb{T}(n)$

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

where the inequality holds due to Claim 1. Next, due to Lemma 16,

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

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

An approximation for the average ball of radius one, or the value $V(n,1)$, is shown next. First, the following theorem is shown, while its proof is shown in [27].

Theorem 18. *For all n,*

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

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 18 the next corollary follows.

Corollary 19. *It holds that,*

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

Proof: It was shown in [6] that $n!\sum_{k=0}^{n-2}\frac{n^k}{k!} \approx \sqrt{\frac{\pi}{2}}n^{n+0.5}$, and therefore,

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

∎

To summarize the results of this section, we proved that for every $T \in \mathbb{T}(n)$ it holds that $V^{\star}(n,1) = \Theta(n^2)$, $V^{-}(n,1) = \Theta(n^3)$, $V_T(n,1) = \Omega(n^2)$, $V_T(n,1) = \mathcal{O}(n^3)$ and the average ball size satisfies $V(n,1) = \Theta(n^{2.5})$. In the extended version of this work [27], it is shown that for every fixed radius t, $V^{\star}(n,t) = \Theta(n^{2t})$, $V^{-}(n,t) = \Theta(n^{3t})$, $V_T(n,t) = \Omega(n^{2t})$, $V_T(n,t) = \mathcal{O}(n^{3t})$, and $V(n,t) = \Theta(n^{2.5t})$. 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 11.

By using the Gilbert-Varshamov lower bound for the average ball size [25] 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 V, based upon Construction 3, we will get that

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

V. Constructions of Codes over Trees

In this section we show several constructions of codes over trees. The first is the construction of $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ codes, and the second is the construction of $\mathcal{T}$-$(n, n, n-2)$ codes. The third and our main result in this section is the construction of $\mathcal{T}$-(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 $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ Codes

A line 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 line trees over n nodes is $n!/2$, so every line 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 line tree

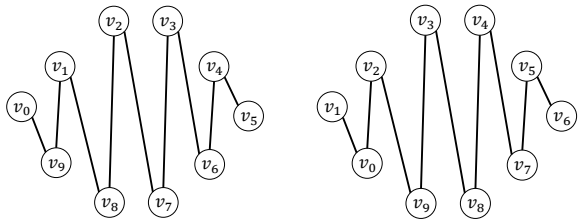
$$T_s = \begin{cases} \Big(v_{\langle s\rangle_n}, v_{\langle s-1\rangle_n}, v_{\langle s+1\rangle_n}, \ldots, v_{\langle s+\frac{n-1}{2}\rangle_n}\Big) : \text{if } n \text{ is odd}, \\ \Big(v_{\langle s\rangle_n}, v_{\langle s-1\rangle_n}, v_{\langle s+1\rangle_n}, \ldots, v_{\langle s-\frac{n}{2}\rangle_n}\Big)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

Example 1. For $n = 10$ an example of the line-tree T_0 is shown. 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 line 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 tree lines 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 [8], [12].

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

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



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

Fig. 1. This code contains 5 trees, T_0, T_1, T_2, T_3, and T_4.

Theorem 20. *The code $\mathcal{C}_{\mathcal{T}_1}$ is a $\mathcal{T}$-$(n, \lfloor n/2 \rfloor, n-1)$ code.*

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 state that this code is a $\mathcal{T}$-$(n, n, n-2)$ code, while its proof is written in [27].

Theorem 21. *The code $\mathcal{C}_{\mathcal{T}_2}$ is a $\mathcal{T}$-$(n, n, n-2)$ code.*

In this trivial construction it is shown that $A(n, n-2) \geqslant n$ and since by Theorem 12, $A(n, n-2) \leqslant n$ it is deduced that $A(n, n-2) = n$.

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

In this section we show a construction of $\mathcal{T}$-(n, M, d) codes for any positive integer d. Note that according to Corollary 11, $A(n, d) = \mathcal{O}(n^{n-1-d})$ and by Corollary 23 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

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

Theorem 22. *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 characteristic 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 construction result.

Corollary 23. *For positive integer n and fixed d, $A(n, d) = \Omega(n^{n-2d})$ and the redundancy is $r(n, d) \leqslant 2(d-1)\log(n) + \mathcal{O}(1)$.*

Proof: Applying BCH codes 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 $\mathbb{T}(n)$ can be mapped by Construction 22 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. Note that in Construction 3 we could use a code correcting $(d-1)$ asymmetric errors. However, we chose to use symmetric error-correcting codes since this does not improve the asymptotic result and in order to derive the result in Corollary 23 we needed linear codes.

VI. Conclusion

In this paper, we initiated the study of codes over trees over the tree distance. Upper bounds on such codes were presented together with specific code constructions 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 line tree. This guarantees that for fixed values 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 shown that the ball average size 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.

1) Improve the lower and upper bounds on the size of codes over trees, that is, the value of $A(n, d)$, when d is fixed and also for $n/2 \leqslant d \leqslant n-3$.
2) Find an optimal construction for $d = n-3$.
3) Study codes over trees under different metrics such as the tree edit distance.
4) Study the problem of reconstructing trees based upon several forests in the forest ball of trees; for more details see [5].

Acknowledgment

This work was partially supported by the United States-Israel BSF grant 2018048.

REFERENCES

[1] M. Aigner and G. M. Ziegler, *Proofs from THE BOOK*, pp. 141–146, Springer-Verlag, New York, 1998.

[2] J. M. Barnard, "Substructure searching methods: old and new," *Journal of chemical information and computer sciences*, vol. 3, issue 33, pp 532–538, Jul. 1993.

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

[4] Y. C. Cheng and S. Y. Lu, "Waveform correlation by tree matching," *IEEE transactions on pattern analysis and machine intelligence*, vol. PAMI-7, issue 3, pp. 299–305, May 1985.

[5] S. Davies, M. Z. Racz, and C. Rashtchian, "Reconstructing trees from traces," *Conference On Learning Theory (COLT)*, 2019.

[6] P. Flajolet, P. Grabner, P. Kirschenhofer, and H. Prodinger, "On ramanujan's qfunction," *Journal of computational and applied mathematics*, vol. 58, pp. 103–116, 1995.

[7] J. Gottlieb ,B. A. Julstrom,G. R. Raidl, and F. Rothlauf, "Prüfer numbers: a poor representation of spanning trees for evolutionary search," *Proceedings of the genetic and evolutionary computation conference*, pp. 343–350, San Francisco, CA, 2001.

[8] N. Hartsfield, G. Ringel, "Hamilton surfaces for the complete symmetric tripartite graph," *Archiv der Mathematik*, Springer-Verlag, vol. 50, pp. 470–473, 1988.

[9] T. Jiang, L. Wang, and K. Zhang, "Alignment of trees – an alternative to tree edit," *Theoretical computer science*, vol. 143, issue 1, pp. 137–148, 1995.

[10] M. Krishnamoorthy and A. T. Ernst, "Comparison of algorithms for the degree constrained minimum spanning tree," *Journal of Heuristics*, vol. 7, pp. 587–611, 2001.

[11] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the american mathematics society*, vol 7, issue 1, pp. 48– 50, 1956.

[12] E. Lucas, "Les rondes enfantines," *Rècrèations mathèmatiques*, vol. 2, Paris, 1894.

[13] J. W. Moon, *Counting labeled trees*, 1970.

[14] B. Moayer and K. S. Fu, "A tree system approach for fingerprint pattern recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. PAMI-8, issue 3, pp. 376–387, May 1986.

[15] I. Neuwirth, "The size of bipartite graphs with girth eight," *arXiv:math.CO/0102210*, 2001.

[16] R. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, pp. 1389–1401, 1957.

[17] G. R. Raidl, "An efficient evolutionary algorithm for the degree-constrained minimum spanning tree problem," *Proceedings of the IEEE congress on evolutionary computation*, pp. 104–111, 2000.

[18] G.R. Raidl, B.A. Julstrom, "Edge sets: An effective evolutionary coding of spanning trees," *IEEE transactions on evolutionary computation*, vol. 7, pp. 225–239, 2003.

[19] G.R. Raidl, B.A. Julstrom, "Greedy heuristics and an evolutionary algorithm for the bounded-diameter minimum spanning tree problem," *In proceedings of the ACM symposium on applied computing*, pp. 747–752, 2003.

[20] I. Reiman, "Über ein Problem von K. Zarankiewicz," *Acta mathematica hungarica*, vol. 9, issue 3–4, pp. 269–273, Hungary, Budapest, Sep. 1958.

[21] S. M. Selkow, "The tree-to-tree editing problem," *Inform. Process. Lett.*, vol. 6, issue 6, pp. 184–186, 1977.

[22] B. A. Shapiro and K. Zhang, "Comparing multiple RNA secondary structures using tree comparisons," *Computer applications in the biosciences*, vol. 6, issue 4, pp. 309–318, Oct. 1990.

[23] K.C. Tai, "The tree-to-tree correction problem," *Journal of the ACM*, vol. 26, issue 3, pp. 422–433, 1979.

[24] E. Tanaka and K. Tanaka, "The tree-to-tree editing problem," *International journal of pattern recognition and artificial intelligence*, vol. 2, issue 2, pp. 221–240, 1988.

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

[26] G. Valiente, "An efficient bottom-up distance between trees," *Proceedings eighth symposium on string processing and information retrieval*, Chile, Nov. 2001.

[27] L. Yohananov and E. Yaakobi, "Codes over trees," *arXiv:2001.01791*, Jan. 2020.

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