T U M

Fakultät für Mathematik

Some Properties of Robinson Graphs

Oliver Bastert, Dan Rockmore, Peter F. Stadler,
Gottfried Tinhofer

TUM-M0101

1

Technische Universität München

Cataloging Data :
Oliver Bastert, Dan Rockmore, Peter F. Stadler, Gottfried Tinhofer: Some Properties of Robinson Graphs; Techn. Univ. München, Fak. f. Math, Report TUM M0101 (1)

Mathematics Subject Classification : 05C05 05C90

Editor :
H. Wähling (waehling@mathematik.tu-muenchen.de),
Fakultät für Mathematik der Technischen Universität München, D-80290 München, Germany.

For the Electronic Version see : http://www-lit.ma.tum.de/reports/

Some Properties of Robinson Graphs

Oliver Bastert[a]*, Dan Rockmore[b], Peter F. Stadler[c,d],
Gottfried Tinhofer[a]

[a]Zentrum Mathematik, Technische Universität Müchen
D-80290 München, Germany
bastert,gottin@mathematik.tu-muenchen.de

[b]Department of Computer Scienec, Dartmouth College
6211 Sudikoff Laboratory, Hanover, NH 03755-3510, USA
rockmore@cs.dartmouth.edu

[c]The Santa Fe Institute
1399 Hyde Park Road, Santa fe, NM 87501, USA

[d] Institut für Theoretische Chemie, Universität Wien
Währingerstr. 17, A-1090 Wien, Austria
studla@tbi.univie.ac.at

* This author has been mainly supported by the Deutsche Forschungsgemeinschaft through the graduate program Angewandte Algorithmische Mathematik, Technische Universität München, and partially by the German Israeli Foundation for Scientific Research and Development under contract # I-0333-263.06/93.

Abstract

Robinson graphs are configuration graphs over sets of phylogenetic trees. Their neighbor relation is given by all possible crossovers along inner edges of these trees. We show that Robinson graphs generate coherent algebras the cell partition of which equals the automorphism partition. Each cell consists of all phylogenetic trees having isomorphic inner trees (induced by the non-leaves).

Key words: phylogenetic tree; Robinson graphs; coherent algebra; cell partition;
AMS classification: 05C05; 05C90

1 Introduction

In biology evolutionary relationships between species or individual genes are customarily represented using *phylogenetic trees*. Finding the phylogenetic tree of a set of extant taxanomic units is known as the *phylogeny reconstruction problem*. In mathematical terms it may be stated as follows:

Given a set S of n elements, a *genetic tree* T on S is a tree having n vertices of degree 1, which are labeled with the numbers $\{1, 2, \ldots n\}$. All remaining vertices of T have degree 3. Furthermore, we are given a cost function f that allows one to determine "how well" a particular tree fits the genealogical relationship among the elements of S. The problem is now to find an *optimal* genetic tree with respect to f.

In this way, the above reconstruction problem is a combinatorial optimization problem, the basic variants of which are all known to be NP-complete [2, 4].

The set of all genetic trees for the set S may in turn be transformed into a configuration graph by defining an appropriate neighbor relation for genetic trees. The resulting graphs are called *Robinson graphs*. In Section 3 a particular neighbor relation is defined using crossover operations.

A *landscape* is a pair (G, f) of a configuration graph G and a function $f : V \longrightarrow \mathbb{R}$ defined on the vertex set V of G. Due to applications in biology, f is called *fitness function*, compare [10, 13].

General configuration graphs and their landscapes are widely used in combinatorial optimization theory. They provide useful mathematical models for studying functions on a discrete set V, the elements of which are structured objects. The configuration graph models a neighborhood relation on V, which defines how one is able to move within V. Such models have therefore a wide spectrum of applications.

Landscapes can be described by their autocorrelation functions which are defined in terms of random walks on G and can be investigated by either using the eigenvalues and eigenspaces of G or via equitable partitions derived from its coherent algebra, see [10, 11, 12] for details.

As outlined in [12] the eigenvalues of a graph G can be expressed in terms of the eigenvalues of graphs arising from pointed equitable partitions of G. Pointed equitable partitions can be derived from the coherent algebra of a graph very easily. Hence, the knowledge of the coherent algebra of G is of interest in this context as well.

In this paper we deal with the equitable partitions provided by the cell partitions of Robinson graphs. In particular, we proof that the cell partition of such a graph coincides with its automorphism partition.

2 Coherent algebras

A subalgebra $\mathcal{A}$ of the algebra Mat_V of all matrices the rows and columns of which are indexed with elements from V is called *coherent algebra* if it is invariant with respect to Hermitian conjugation, contains the unit matrix I and the all 1's matrix J, and is closed with respect to componentwise multiplication, i. e.

$$A, B \in \mathcal{A} \Longrightarrow A \circ B \in \mathcal{A}$$

where $(A \circ B)_{ij} = A_{ij} B_{ij}$. The set V is called the point set of $\mathcal{A}$. In our context $V = \{1, 2, \ldots, N\}$ for some natural number N.

Coherent algebras have been studied first in [14, 15], and independently in [5, 6, 7]. They play a fundamental role in algebraic combinatorics and have applications in mathematical chemistry. A friendly introduction to coherent algebras taking into account the interests of chemists is given in [8], while the collection [3] is written for mathematicians and covers the most important theoretical aspects.

We present here a list of facts about coherent algebras which are used in this paper. All proofs may be looked up in the cited literature.

The smallest coherent algebra containing the matrix A is called the coherent algebra *generated* by A and denoted by $\mathcal{A}(A)$. If A is the adjacency matrix of a graph G then we say also that $\mathcal{A}(A)$ is generated by G.

Every coherent algebra $\mathcal{A}$ of dimension d has a linear basis consisting of 0-1-matrices $A_1, \ldots, A_d$ and satisfying $A_i \circ A_j = A_i \delta_{ij}$. Hence, the supports

$$E_i = \{(u, v) | A_{uv} = 1\},\ 1 \leq i \leq d,$$

determine a partition

$$\mathcal{C} = \{E_1, \ldots, E_d\}$$

of $V \times V$. This partition is called a *coherent configuration* and is uniquely determined by the following properties:

(C1) $\displaystyle\bigcup_{i=1}^{d} E_i = V \times V,$

(C2) Using an appropriate numbering, there is a number $t < d$ such that

$$\bigcup_{i=1}^{t} E_i = \{(i,i)|i \in V\}(= \Delta, \text{the diagonal of } V \times V)$$

(C3) For each $i \in V$ there is an $i' \in V$ with $E_i = E_{i'}^T$.

(C4) There are non-negative integers p_{ij}^k such that

$$|\{w|(u,w) \in E_i \wedge (w,v) \in E_j\}| = p_{ij}^k$$

independent of $(u,v) \in E_k$.

With the help of the sets E_i, $i \leq t$, which are contained in the diagonal Δ, we define a partition of V, namely

$$C_i = \{u|(u,u) \in E_i\},\ 1 \leq i \leq t.$$

The sets C_i are called the *cells's*, and

$$\mathcal{C}_{\mathcal{A}} = \{C_1, \ldots, C_t\}$$

is called the *cell partition* of $\mathcal{A}$ (or of $\mathcal{C}$), or if $\mathcal{A}$ is generated by a graph G, the cell partition of G. For every basic set E_k there are cells C_i and C_j such that $E_k \subseteq C_i \times C_j$.

Let $a = (a_1, \ldots, a_l)$ be an arbitrary sequence of indices $a_i \in \{1, \ldots, d\}$. An a-walk from i to j is a sequence of arcs

$$(i_0, i_1), (i_1, i_2), \ldots, (i_{l-1}, i_l)$$

with the property

$$(i_s, i_{s+1}) \in E_{a_s},\ 0 \leq s < l.$$

In dealing with coherent configurations (like rather often in dealing with arbitrary partitions of objects) it has become standard language to express membership $(u,v) \in E_i$ by saying that (u,v) *has color* i. Analogously, we say that u *has color* k, if $u \in C_k$ holds.

The following lemma expresses a basic property of coherent configurations.

Lemma 1 *Let C and C' two arbitrary cells of $\mathcal{C}$ and assume $i, i' \in C$. Then for any a, the number of a-walks from i to a vertex in cell C' is the same as from i'. Note that a-walks counted here have only fixed starting point, but the final points vary over C'.*

Typically, this lemma is used when one has to demonstrate that two given items i and j cannot belong to the same cell. We shall make use of this lemma whenever it is appropriate and without quoting it at every occasion.

Graphs which generate the same coherent algebra are cospectral. Therefore the structure of the associated coherent algebra and of the corresponding coherent configuration, respectively, is an important structural property of graphs. In general, we are not able to compute the coherent configurations for Robinson graphs for larger n, those graphs have too huge a size. However, we are at least able to determine the cell partition.

3 Genetic Trees

A *leaf* of a tree T is a vertex v of degree 1. All other vertices are called *inner vertices*. Edges joining inner vertices are called *inner edges*. A tree is called *genetic* if all its inner vertices have degree 3 and its leaves are colored (labeled) with the colors $1, 2, \ldots, n$, whereas all inner vertices are uncolored.

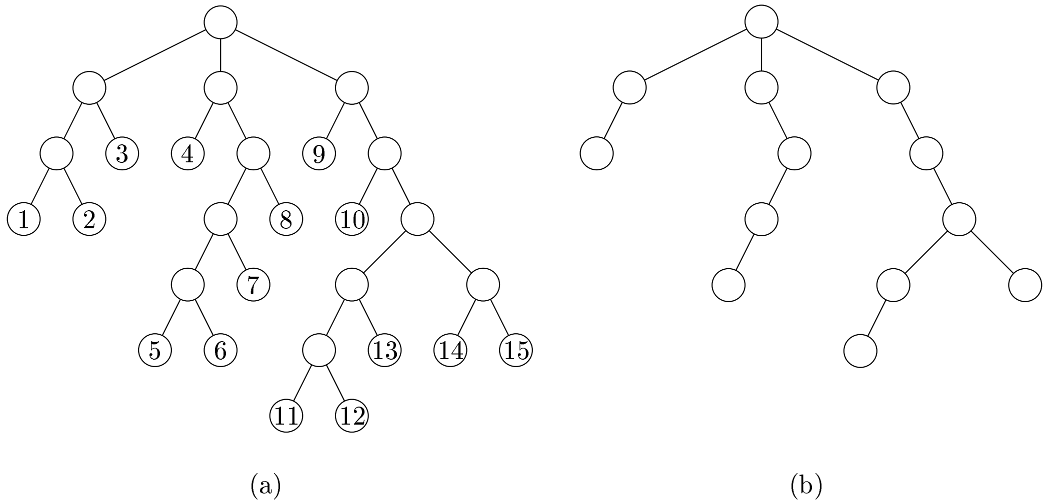
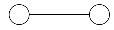


Figure 1: A genetic tree T (a) and its inner tree T° (b)

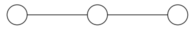
The set of genetic trees with n leaves will be denoted by $\mathcal{T}_n$. A member of $\mathcal{T}_n$ has $2n-3$ edges and $n-2$ inner vertices [9]. The *inner tree* T° of a genetic tree T is the subtree of T induced by the inner vertices of T. Two genetic trees are considered *equal* if and only if they are isomorphic as leaf colored trees. Observe that equal trees have isomorphic inner trees.

In **Figure 2** the inner trees in $\mathcal{T}_n$ for n up to 8 are depicted.

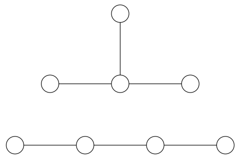
(a) $n = 3$



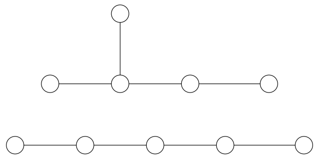
(b) $n = 4$



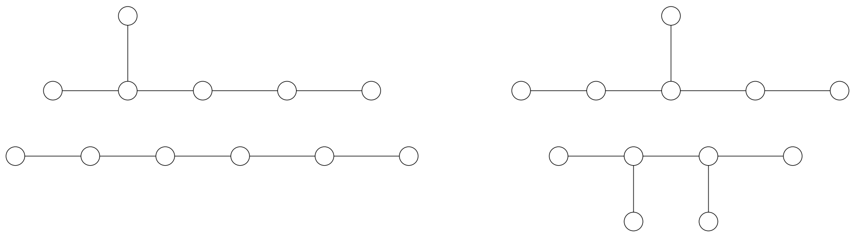
(c) $n = 5$



(d) $n = 6$



(e) $n = 7$



(f) $n = 8$

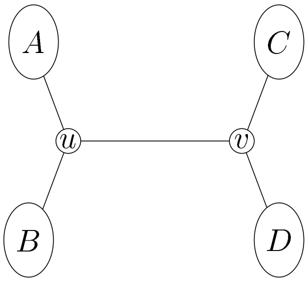
Figure 2: Inner trees of genetic trees

An inner vertex is called *s-vertex* if its degree with respect to the inner tree T° is s. The number of s-vertices will be denoted by n_s, $s = 1, 2, 3$. We call an edge of the inner tree an $(s : t)$-edge if the end vertices of the edge are an s- and a t-vertex.

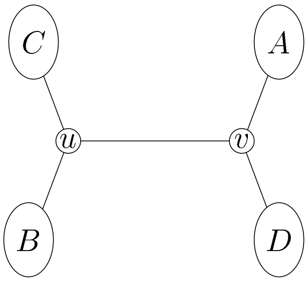
In many considerations, the coloring of the leaves is of no matter. In such cases, it is usually not mentioned.

With every inner edge $[u, v]$ of a genetic tree T, we associate four subtrees A, B, C, D as indicated in **Figure 3(a)**. The subtrees A, B, C, D are the four connected components which are obtained when deleting the edge $[u, v]$ and the vertices u and v from T. Note that each of these subgraphs may consist of a single vertex only.

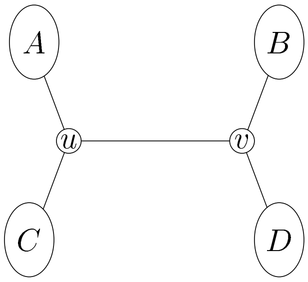
Definition 1 *The operations indicated in Figure 3(b) and (c) are called p(arallel)-crossover and d(iagonal)-crossover of* T *(on the inner edge* $[u, v]$*).*



(a) inner edge with subtrees



(b) p-crossover



(c) d-crossover

Figure 3: Crossovers

The type of the crossover is not a graph theoretical property. It depends on the drawing or the current ordering of the edges adjacent to some vertex. It is only introduced to simplify the argumentation at certain points.

A crossover on $[u, v]$ is called $(s : t)$-*crossover* if $[u, v]$ is an $(s : t)$-edge. We say that two trees are of the same *type* if and only if their inner trees are isomorphic.

Definition 2 *The configuration graph (Robinson graph) Γ_n has vertex set $\mathcal{T}_n$ and two trees T, T' are adjacent in Γ_n if and only if there exists an inner edge $e \in T$ such that T' results from T by a crossover on e. The vertices of Γ_n are called the trees of Γ_n.*

Observe that Γ_3 consists of 1 vertex only and that Γ_4 is the complete graph on 3 vertices.

Remark 2 The configuration graph Γ_n has $\prod_{i=0}^{n-3}(2i+1)$ vertices. It is $(2n-6)$-regular and the number of trees with distance two from a given tree equals

$2n^2 - 10n + 4n_1$, i.e., it depends only on n and the number of 1-vertices of the inner tree. Furthermore, the numbers n_2 and n_3 depend only on n and n_1 in a simple way, namely $n_2 = n_1 - 2$ and $n_3 = n - 2n_1$. Proofs for these observations can be found in [9].

4 The Cell Partition of Γ_n

Remember that by definition the coherent configuration $\mathcal{A}_{\Gamma_n}$ generated by the graph Γ_n is the coarsest partition of $\mathcal{T}_n \times \mathcal{T}_n$ satisfying (C1) – C(4). This partition contains a partition of $\mathcal{T}_n$, the cell partition $\mathcal{C}_{\Gamma_n}$ of Γ_n. In the following, we are going to determine the cell partition $\mathcal{C}_{\Gamma_n}$.

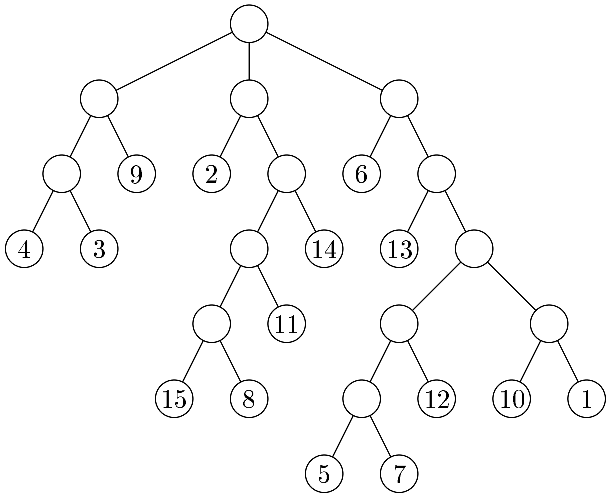


Figure 4: $\pi(T)$

Let $\{\tilde{T}_k \mid 1 \leq k \leq K\}$ denote the set of pairwise non-isomorphic inner trees with $n-2$ vertices. Let $[\tilde{T}_k]$ denote the set of genetic trees with inner tree isomorphic to $\tilde{T}_k$. Elements of $[\tilde{T}_k]$ differ only by the coloring of their leaves. Obviously, $\mathcal{C}_n := \{[\tilde{T}_k] \mid 1 \leq k \leq K\}$ is a partition of $\mathcal{T}_n$ as well.

Definition 3 *Let π be an arbitrary permutation of $\{1, 2, \ldots, n\}$. For a genetic tree $T \in \mathcal{T}_n$ define the tree $\pi(T)$ by replacing the colors $1, 2, \ldots, n$ on the leaves of T by the colors $\pi(1), \pi(2), \ldots, \pi(n)$, respectively.*

Let

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 4 & 3 & 9 & 2 & 15 & 8 & 11 & 14 & 6 & 13 & 5 & 7 & 12 & 10 & 1 \end{pmatrix}$$

and consider T as defined in **Figure 1**. The genetic tree $\pi(T)$ is depicted in **Figure 4**.

Lemma 3 *The partition $\mathcal{C}_n$ is at least as fine as the cell partition $\mathcal{C}_{\Gamma_n}$.*

Proof. Let π be an arbitrary permutation of $\{1, 2, \ldots, n\}$. This permutation preserves the partition $\mathcal{C}_n$, i.e., $\pi([\tilde{T}_k]) = [\tilde{T}_k]$, $1 \leq k \leq K$. Furthermore, if T' is a neighbor of T due to a crossover on an inner edge $[u, v]$, then $\pi(T')$ is a neighbor of $\pi(T)$ due to the same crossover. Hence, π induces an automorphism of Γ_n. Since for any two trees $T, T' \in [\tilde{T}_k]$, there is a π such that $\pi(T) = T'$, we obtain that $\mathcal{C}_n$ is at least as fine as the automorphism partition of Γ_n. Now, the claim follows by the fact that the cell partition of a graph is always coarser than or equal to its automorphism partition. $\square$

Lemma 3 shows that all genetic trees in Γ_n having isomorphic inner trees belong to the same cell of $\mathcal{C}_{\Gamma_n}$. In the following, we want to prove that the other direction is true as well. In other words: trees contained in the same cell have isomorphic inner trees, and thus, $\mathcal{C}_n$, the automorphism partition of Γ_n, and the cell partition of $\mathcal{A}_{\Gamma_n}$ coincide. To prove this, we show that two trees having non-isomorphic inner trees lie in different cells of $\mathcal{C}_{\Gamma_n}$.

Let U be a cellular set (a union of cells) of $\mathcal{A}_{\Gamma_n}$. In the discussion which follows, we use the fact that two vertices having a different number of neighbors in U, cannot belong to the same cell of $\mathcal{C}_{\Gamma_n}$ (see Lemma 1).

We will start by showing that the sets defined in Definition 4 are cellular sets of $\mathcal{A}_{\Gamma_n}$.

Definition 4 *Let $\mathcal{T}_n^{n_k}(i)$, $1 \leq k \leq 3$, be the subset of $\mathcal{T}_n$ in which each element has i k-vertices and $\mathcal{T}_n^{dm}(i)$ the subset of $\mathcal{T}_n$ in which each element has diameter equal to i.*

As mentioned above in Remark 2, the number of trees with distance 2 from a tree $T \in \mathcal{T}_n$ depends only on n and the number of 1-vertices. By Lemma 1, this implies the following lemma.

Lemma 4 *The trees in a given cell of $\mathcal{C}_{\Gamma_n}$ have the same number of 1-vertices (n_1 is constant on each cell).*

Immediately, we get:

Lemma 5 *The trees in a given cell of $\mathcal{C}_{\Gamma_n}$ have the same number of 2-vertices and the same number of 3-vertices.*

Proof. Since $n_2 = n - 2n_1$ and $n_3 = n_1 - 2$, the claim holds. $\square$

Thus, we get finally:

Lemma 6 *Each $\mathcal{T}_n^{n_k}(i)$, $1 \le k \le 3$, defines a cellular set.*

We now examine crossovers and their potential to change the diameter, i.e., we examine how the diameter of the resulting tree differs from the diameter of the tree we start with.

Consider the tree T of **Figure 3** again. Define l_A and l_B to be the length of a longest path from u to a leaf of A and B, respectively, and l_C and l_D to be the length of a longest path from v to a leaf of C and D, respectively.

Lemma 7 *The diameters of a tree T and a tree T' which is obtained by a crossover on some inner edge of T can differ by at most one.*

Proof. Assume without loss of generality that $l_A \ge l_C$, $l_A \ge l_B$, and $l_C \ge l_D$. This situation can always be met by properly renaming the different parts of T.

The diameter of T is

$$\max\{l_A + l_B, l_A + 1 + l_C\}.$$

Consider now the trees T_p and T_d which are the result of a p- and d-crossover, respectively, of T on $[u, v]$. The diameters have the following values:

$$\begin{aligned}\operatorname{diam}(T_p) &= \max\{l_A + 1 + l_B, l_A + 1 + l_C\} \text{ and}\\ \operatorname{diam}(T_d) &= \max\{l_A + 1 + l_B, l_A + 1 + l_D, l_A + l_C\}.\end{aligned}$$

The p-crossover leaves the diameter untouched or enlarges it by at most one. Since $l_A + l_D \le l_A + l_C$, $\operatorname{diam}(T_d)$ is at most $\operatorname{diam}(T) + 1$ and at least $l_A + l_C$ which is as least as large as $\operatorname{diam}(T) - 1$. $\square$

An edge is *incident* with a path P if exactly one of the end vertices of the edge lies on the path. A path $P = (v_1, v_2, \ldots v_k)$ in a tree T, consisting of inner vertices only, is a *longest inner path* if and only if $k = \operatorname{diam}(T) - 1$. As an immediate consequence of Lemma 7, we obtain the following lemma.

Lemma 8 *A tree with a larger diameter is obtained if and only if a crossover is performed on an edge incident with a longest inner path.*

Proof. Recall the situation in the proof above. Consider the case where $\mathrm{diam}(T_p) = \mathrm{diam}(T) + 1$. By simply analyzing the formulas for $\mathrm{diam}(T_p)$ and $\mathrm{diam}(T)$, we see that this happens if and only if there is a longest inner path in T starting in A and ending in B. $[u, v]$ is incident to this path.
Now, assume that $\mathrm{diam}(T_d) = \mathrm{diam}(T) + 1$. This is true if and only if there is a longest inner path in T starting in A and ending in B. Again, $[u, v]$ is incident to this path.

□

Lemma 9 *The only way to obtain a tree with a smaller diameter by a crossover is to perform the crossover on an edge which is part of all longest inner paths.*

Proof. Revisit the proof of Lemma 7 again. The only possibility for reducing the diameter by a crossover on $[u, v]$ is that all longest inner paths in T go from leaves in A to leaves in C. □

We are now able to prove that all trees with equal diameter define a cellular set of $\mathcal{A}_{\Gamma_n}$. First, we will have a closer look at the trees of Γ_n having largest diameter. The inner trees with the largest diameter, namely $n - 3$, are those isomorphic to the path on $n - 2$ vertices. Note that the diameter of the inner tree T° of a tree T is exactly $\mathrm{diam}(T) - 2$.

Lemma 10 $\mathcal{T}_n^{n_1}(2) = \mathcal{T}_n^{dm}(n-1)$ *is a cell of* $\mathcal{C}_{\Gamma_n}$.

Proof. $\mathcal{T}_n^{n_1}(2)$ is the set of trees whose inner trees are isomorphic to the path on $n - 2$ vertices. Due to Lemma 3 and Lemma 4, $\mathcal{T}_n^{n_1}(2)$ is a cell of $\mathcal{C}_{\Gamma_n}$. □

Lemma 11 *All trees in a cell of* $\mathcal{C}_{\Gamma_n}$ *have equal diameter. Thus,* $\mathcal{T}_n^{dm}(i)$ *is a cellular set for all* i.

Proof. The proof is by downward induction on the diameter of the trees.

We have shown already that $\mathcal{T}_n^{n_1}(2)$ is a cell of $\mathcal{C}_{\Gamma_n}$. Assume that two trees with different diameters greater than d lie in different cells of $\mathcal{C}_{\Gamma_n}$.

Let T be a tree with diameter $d < n - 1$. We will show that T has neighbors with diameter $d + 1$. Observe that due to Lemma 7, the diameter can increase by at most 1 after executing one crossover and thus, trees with diameter d are the only candidates for having neighbors with diameter $d + 1$.

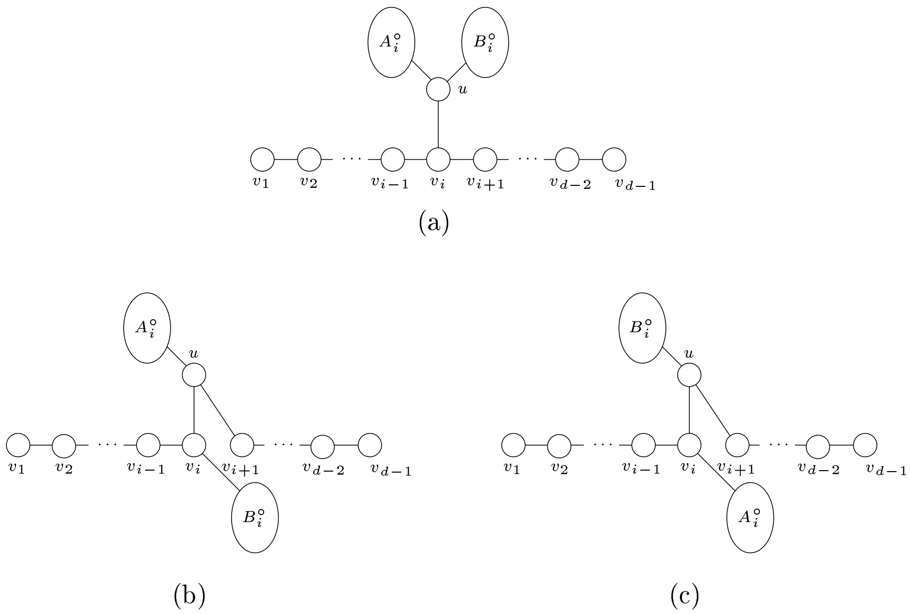


Figure 5: T° and the inner trees of the neighbors of T which result from crossovers on $[v_i, u]$

Since $T \notin \mathcal{T}_n^{n_1}(2)$, each longest inner path contains at least one 3-vertex. Let $(v_1, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_{d-1})$ be such a longest inner path, v_i a 3-vertex, and denote the third neighbor of v_i by u (see **Figure 5(a)**). Applying the two possible crossovers on the edge $[v_i, u]$ results in two trees with diameter $d+1$ (see **Figure 5(b),(c)**), realized by a new longest path $P' = (v_1, v_2, \ldots v_i, u, v_{i+1}, \ldots, v_{d-1})$.

Hence, T has neighbors with diameter $d+1$. This completes the proof. $\square$

Consider some longest inner path $P = (v_1, v_2, \ldots v_{d-1})$ in an inner tree T°. Assume that the 3-vertices on P are $\{v_{t_1}, v_{t_2}, \ldots, v_{t_k}\}$ with $\text{dist}(v_1, v_{t_i}) < \text{dist}(v_1, v_{t_j})$, $\forall i < j$, holds. Let u_i, $i \in \{1, 2, \ldots, k\}$, be the vertex not on P which is adjacent to v_{t_i}. If we perform both crossovers on $[v_{t_i}, u_i]$, a d-crossover and a p-crossover, we obtain two different trees. Observe that although the inner trees might be isomorphic, the resulting trees are different since they differ by the coloring of their leaves.

We will now, for each tree, identify "largest" neighbors among all neighbors with greater diameter. For this aim we are going to introduce an appropriate code for genetic trees. This task requires some preliminaries.

First, define the *code* $c_v(T)$ of a tree and an inner vertex v of this tree as the pair consisting of the length of a longest inner path from v to a leaf of T and of some complete invariant of T, for example a suitable code for T. We assume that codes can be compared lexicographically.

Next, define a function $c_T(P)$ on the set of inner paths P of a tree T. Let $P = (v_1, v_2, \ldots v_l)$ be an inner path between two 1-vertices v_1 and v_l. Assume that the vertices $\{v_{t_1}, v_{t_2}, \ldots, v_{t_k}\}$ are the 3-vertices on P. The subtree attached to v_{t_j} is denoted by Υ_j, and we assume $\mathrm{dist}(v_1, v_{t_j}) < \mathrm{dist}(v_1, v_{t_{j'}})$ if $j < j'$. Furthermore, each Υ_j consists of a node u_j adjacent to v_{t_j}, and the two subtrees A_j and B_j adjacent to u_j. The vertices in A_j and B_j adjacent to u_j are denoted by a_j and b_j, respectively. The situation is depicted in **Figure 6**.

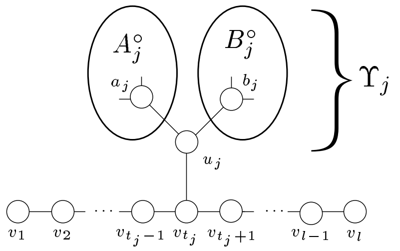


Figure 6: The path P in T° and the subtree Υ_j.

Let w.l.o.g. $c_{a_j}(A_j) \geq c_{b_j}(B_j)$. Define

$$c_T(P) := (l, ((\mathrm{dist}(v_{t_j}, v_k), c_{a_j}(A_j), c_{b_j}(B_j)) \mid j \in \{k, k-1, \ldots, 1\})).$$

Now we are able to introduce the following code for our trees. Define

$$c(T) = \max_{P \in \mathcal{P}(T)} \{c_P(T)\},$$

where

$$\mathcal{P}(T) := \{P \mid P \text{ is an inner path in } T\}$$

and where by "max" we mean the lexicographically largest value. We say that P is *responsible* for the code of T if $c(T) = c_T(P)$. Observe that if P is responsible for the code then it is a longest inner path in T. Obviously, given $c(T)$, we are able to reconstruct T in a unique way. A tree T is *larger* than another tree T' if $c(T)$ is lexicographically larger than $c(T')$.

Now we examine the situation with respect to the number of 3-vertices in more detail.

Lemma 12 *A* $(3:3)$*-crossover and a* $(3:2)$*-crossover leave the number of* 3*-vertices unchanged whereas a* $(3:1)$*-crossover reduces the number of* 3*-vertices by one.*

Proof. This is easy to verify. In **Figure 7(a)**, $[u,v]$ is a $(3:1)$-edge and the results of the two possible crossovers are shown in **Figure 7(b)** and **(c)**.

The edge $[u,v]$ becomes a $(2:2)$-edge. Note that the resulting inner trees are isomorphic.

In **Figure 8(a)**, $[u,v]$ is a $(3:2)$-edge and the results of the two possible crossovers are shown in **Figure 8(b)** and **(c)**. The edge $[u,v]$ remains a $(3:2)$-edge.

Obviously, a $(3:3)$-edge remains a $(3:3)$-edge. □

A neighbor of a tree is called *longer neighbor* if it has a larger diameter. A tree is called an *s-path* if its inner tree is a caterpillar with s legs, i.e., is composed of a longest inner path Q with s inner edges incident to it. Q is not necessarily unique. However, we assume that given an s-path, one of the possible longest inner paths is selected as Q.

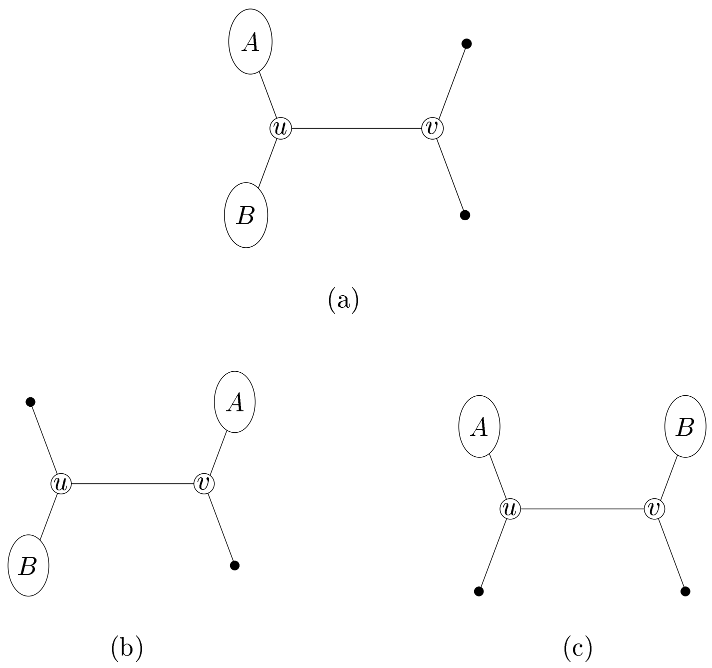


Figure 7: A tree and the two possible crossovers on a $(3:1)$-edge $[u,v]$

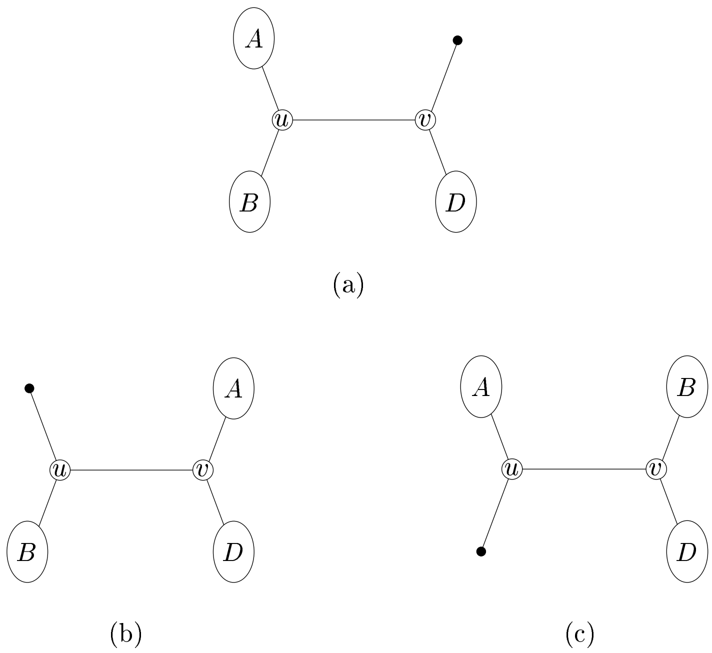


Figure 8: A tree and the two possible crossovers on a (3 : 2)-edge $[u, v]$

We first consider the set of all 1-paths and the set of all 2-paths, respectively, and show that they form cellular sets of $\mathcal{A}_{\Gamma_n}$. Afterwards, we turn to more general trees.

Let $Q = (v_1, v_2, \ldots, v_{d-1})$ be the selected longest inner path of an s-path, v_i the first and v_j the last 3-vertex on Q. Thus Q is the concatenation of three subpath Q_1, Q_2, Q_3, where Q_1 connects the 1-vertex v_1 to the first 3-vertex v_i, Q_3 connects the last 3-vertex v_j to the 1-vertex v_{d-1}. All other vertices of Q_1 and Q_3, if any, are 2-vertices. The subpaths Q_1 and Q_3 are called the *tails* of Q of length $i-1$ and $d-1-j$, respectively.

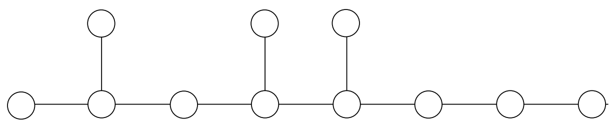
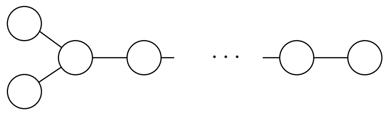


Figure 9: The inner tree of a 3-path with two choices for Q

Lemma 13 *The* 1*-paths build a cellular set and are distinguished in* $\mathcal{A}_{\Gamma_n}$ *if they have non-isomorphic inner trees.*

Proof. The 1-paths build a cellular set since they are the only trees with diameter $n-2$ and one 3-vertex and the intersection $\mathcal{T}_n^{n_3}(1) \cap \mathcal{T}_n^{dm}(n-2)$ of two cellular sets is obviously a cellular set.

The inner tree of a 1-path with only one tail having length greater than one is isomorphic to the graph depicted below.



Trees having this inner tree are distinguished from the other 1-paths since they are the only ones having four neighbors in $\mathcal{T}^{dm}(n-1)$. All other 1-trees have only two such neighbors.

Assume now that the trees in question have two tails of lengths l_1 and l_2, and that w.l.o.g. $l_1 \leq l_2$ and $l_1 \leq \frac{d-1}{2}$ holds.

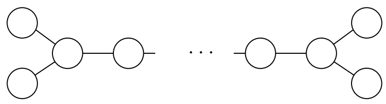
The proof is by induction on l_1. The proof for $l_1 = 1$ just has been given. Assume that the 1-paths with l_1 less than l are distinguished if they have non-isomorphic inner trees.

Consider now trees with $l_1 = l$. They are the only ones with $l_1 \geq l$ which have neighbors having a shortest tail of length $l-1$. This completes the proof. □

Lemma 14 *The 2-paths build a cellular set and will be distinguished in $\mathcal{A}_{\Gamma_n}$ if they have non-isomorphic inner trees.*

Proof. The 2-paths build a cellular set since they are the only trees with diameter $n-3$ and two 3-vertices, i.e., the only trees in $\mathcal{T}_n^{n_3}(2) \cap \mathcal{T}_n^{dm}(n-3)$. We define l_1 and l_2 as before.

The only 2-paths which have eight longer neighbors, which obviously are 1-paths, are trees the inner tree of which is isomorphic to the one depicted below ($l_1 = l_2 = 1$).



The length l_1 of the shortest tail of a 2-path T is determined by the code $c(T')$ of its largest neighbor T', which is a 1-path. By Lemma 13, 2-paths with different values of l_1 belong to different cells.

The remaining part of the proof is by induction on l_1+l_2. The case when $l_1+l_2=2$ has been considered already. The 2-paths with $l_1+l_2 \geq l$ having neighbors where the sum of the tails is shorter than l are graphs with $l_1+l_2=l$. $\square$

So far, we have proven that trees the inner trees of which are caterpillars with at most two legs, belong to the same cell of $\mathcal{C}_{\Gamma_n}$ if and only if their inner trees are isomorphic.

Now, we treat more general classes of trees. Let T have diameter d. Assume that there exist neighbors of T having greater diameter than T. Let T_l be a largest (with respect to the code) neighbor among those neighbors. Assume that the crossover on T to obtain T_l has been performed on $[v_i, u]$. Since the diameter of T_l is greater than the diameter of T, each longest path in T_l must contain the edge $[v_i, u]$. Let $P_l = (v_1, v_2, \ldots, v_{i-1}, u, v_i, v_{i+1}, \ldots, v_{d-1})$ be an inner path responsible for the code of T_l. Then $P = (v_1, v_2, \ldots, v_i, v_{i+1}, \ldots, v_{d-1})$ is a longest inner path in T. Obviously, v_i is the rightmost 3-vertex of P. Otherwise, P_l could not be a largest neighbor.

Lemma 15 *If the number of* 3*-vertices of* T *and* T_l *(as defined above) is equal, then* T° *is determined by* T_l°.

Proof. If T and T_l have the same number of 3-vertices, then the edge $[v_i, u]$ on which the crossover is performed is either a $(3:3)$- or a $(3:2)$-edge.

If $[v_i, u]$ is a $(3:3)$-edge, then subtrees isomorphic to A or B are attached to the rightmost 3-vertices (see Figure 10), namely u and v_i, on all paths in P_l' responsible for $c(T_l)$.

If the crossover has been performed on a $(3:2)$-edge, i.e., if in **Figure 10** B is a single vertex, then a subtree isomorphic to A is attached to the rightmost 3-vertex on all longest paths responsible for the code $c(T_l)$.

In both cases, the edge on which the crossover from T to T_l has been performed, is determined, namely the rightmost $(3:3)$-edge or $(2:3)$-edge, respectively, on a path responsible for the code $c(T_l)$.

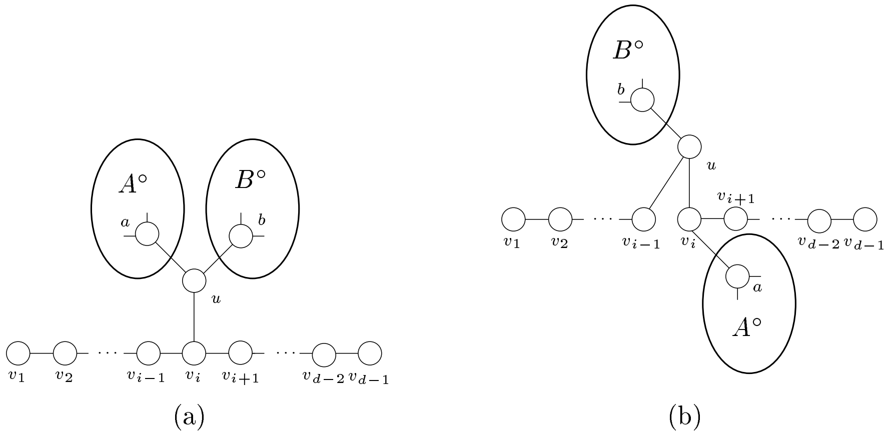


Figure 10:

Therefore, we are able to reconstruct T° by only considering T_l°. $\square$

Note that for finding a responsible path P, we need the whole tree rather than T° only, however, the coloring of the leaves in T is of no matter. Thus, what we need is the isomorphism class of T which is uniquely defined by T°.

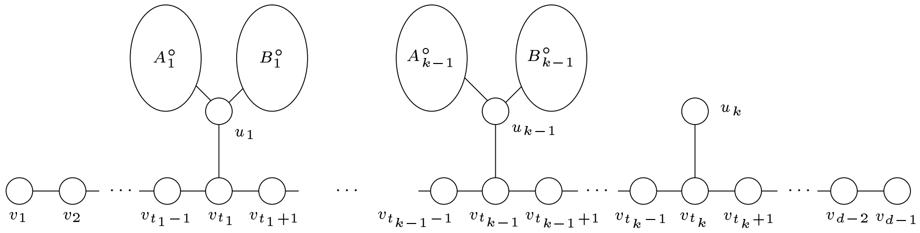
Now, let us consider the case where the largest neighbor T_l of T has fewer 3-vertices than T, i.e., the crossover leading from T to T_l is performed on a $(3:1)$-edge.

If this happens then clearly T° looks like in **Figure 11(a)** and T_l° like in **Figure 11(b)** where the path $P_l = (v_1, v_2, \ldots v_{t_k-1}, u_{k-1}, v_{t_k}, \ldots, v_{d-1})$ is responsible for the code of T_l. The 3-vertices of this path are $\{v_{t_1}, v_{t_2}, \ldots, v_{t_k-1}\}$.

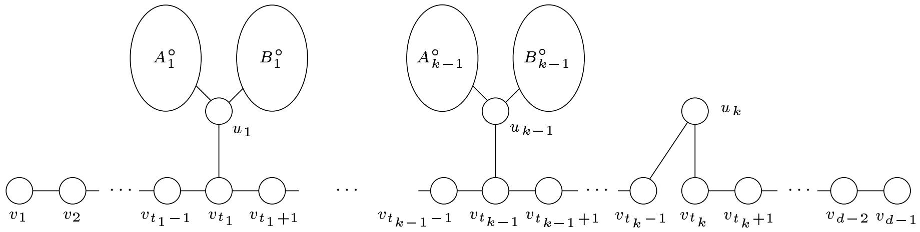
Consider now T_{xl}, the largest neighbor of T_l. It is clear that the crossover transforming T_l into T_{xl} has been made on the rightmost 3-vertex of P_l, namely $v_{t_{k-1}}$. This is because all paths in T_l which are responsible for the code contain the path from $v_{t_{k-1}}$ to v_{d-1}, since this is the only part of T_l where the length of a path has been increased with respect to T.

For the same reason, this path is the right tail of all largest paths in T_l. It cannot be the left tail, since $\text{dist}(v_{d-1}, v_{t_{k-1}}) > \text{dist}(v_1, v_{t_1})$.

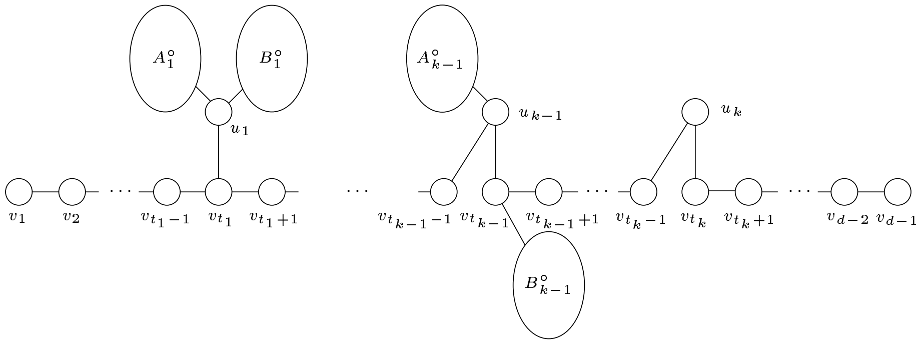
There is exactly one other path in Γ_n of length 2 from T to T_{xl} (by reversing the order of the two crossovers). Denote the tree on this path by T_x. Obviously, T_{xl} only exists if there are at least two 3-vertices on P. The situation is depicted in **Figure 11**.



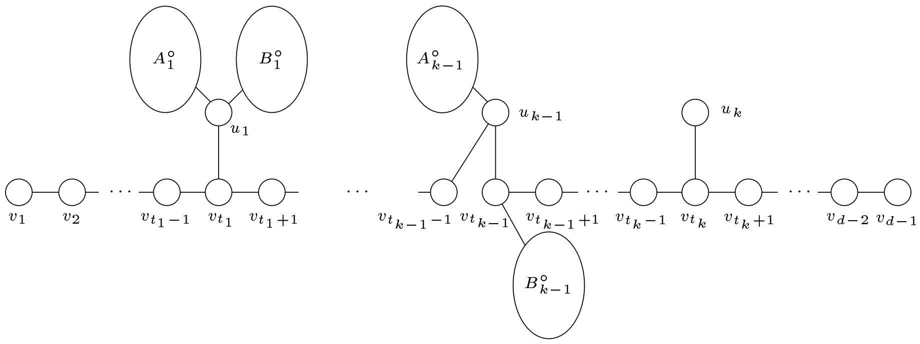
(a) T°



(b) T^{2°



(c) T^{4°



(d) T^{5°

Figure 11: T and larger neighbors

Lemma 16 *If all largest neighbors of T have less 3-vertices than T, T° is determined by T_l° and T_{xl}°.*

Proof. As we have seen, there is the unique tree T_x. Since T_l has less 3-vertices than T, Υ_k has only one inner vertex, namely u_k (see **Figure 6**). Observe that T° is determined up to the position of v_{t_k} by T_l°.

Assume now that there are either at least two 3-vertices on P_l or a subtree Υ_j, $j < k$, has more than one inner vertex. Otherwise, the tree T would be a caterpillar with at most 2 legs, for which the result is already clear due to Lemma 13 and Lemma 14.

Let P_x be a path in T_x responsible for the code $c(T_x)$. Since the distance from the beginning of P_x to the first 3-vertex in P_x and of the rightmost 3-vertex on P_x are exactly as in P, a path responsible for the code $c(T)$, the position of v_{t_k} is determined by T_x°. □

Theorem 17 *Each $[\tilde{T}_k]$ defines a cell of $\mathcal{C}_{\Gamma_n}$.*

Proof. The proof is done by a similar induction as in Lemma 11. From Lemma 10 we know that $\mathcal{T}^{dm}(n-1)$ defines a cell. In fact, we even have proved that the sets $[\tilde{T}_k]$ are cells if $\tilde{T}_k$ is a 1-path or a 2-path. This result has already been used in the proof of Lemma 16.

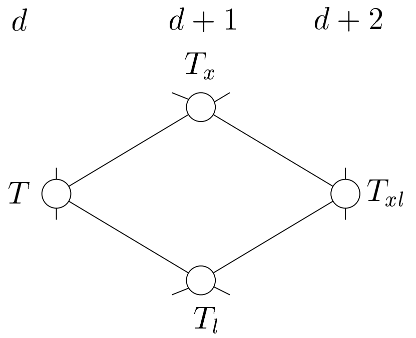


Figure 12: The crucial part of Γ_n

Assume that the trees with diameter larger than d lying in one cell have isomorphic inner trees. Consider a tree T with diameter d. As we have seen before, T° can be determined by considering only the inner trees of some longer trees with distance one or two of T. To be more precise, let us consider the situation in Γ_n as depicted in **Figure 12**.

If a largest neighbor of T has the same number of 3-vertices as T (and thus all largest neighbors have this property), then T° is determined only by T_l° (see Lemma 15). Since by induction hypothesis the set $[T_l^\circ]$ is a cell, trees having a largest neighbor not in $[T_l^\circ]$ are distinguished from T. Hence, trees having a largest neighbor with the same number of 3-vertices lie in different cells of $\mathcal{C}_{\Gamma_n}$ if and only if their inner trees are isomorphic.

The case where all largest neighbors of T have less 3-vertices than T is more involved. As proved before, we need to consider T_x° together with T_l° to determine T° (see Lemma 16). In $\mathcal{A}_{\Gamma_n}$, the color of the edge (T, T_{xl}) represents the set of colored paths from T to T_{xl} (compare Lemma 1 and remember what was said about the use of the term "color" and the end of Section 2). Hence, the color of (T, T_{xl}) depends on the colors of T_l and T_x as well. Thus, T° is determined by the color of the edge (T, T_{xl}) in $\mathcal{A}_{\Gamma_n}$.

Therefore, all trees with diameter d lie in the same cell only if they have isomorphic inner trees.
□

5 Summary

We have shown in this paper that the cell partition of Robinson graphs coincide with their automorphism partition. This result can be used in order to get information about the spectra of such graphs. Some results in this direction have been reported in [1].

References

[1] O. Bastert, D. Rockmore, P. F. Stadler and G. Tinhofer, *Landscapes on Spaces of Trees.* Submitted to *Applied Mathematics and Computations.*

[2] W. H. E. Day, D. S. Johnson and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.* **81** (1986), 33 - 42.

[3] I. A. Faradžev, M. H. Klin and M. ER. Muzychuk. Cellular rings and groups of automorphisms of graphs. In: I. A. Faradžev, A. A. Ivanonv, M. H. Klin and A. J. Woldar (eds.), Investigations in algebraic theory of combinatorial objects. *Mathematics and its applications* Soviet Series Vol. **84** (1994), Kluwer, Dordrecht.

[4] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Adv. Appl. Math.* **2** (1982), 43 - 49.

[5] D. Higman. Coherent configurations. I: Ordinary representation theory. *Geometria Dedicata* **4** (1975), 1 - 32.

[6] D. Higman. Coherent Configurations. Part II: Weights. *Geometria Dedicata* **5** (1976), 413 - 424.

[7] D. Higman. Coherent Algebras. *Lin. Alg. Appl.* **93** (1987), 209 - 239.

[8] M. Klin, C. Rücker, Ch. Rücker and G. Tinhofer. Algebraic combinatorics in mathematical chemistry. Methods and algorithms. I. Permutation groups and coherent (cellular) algebras. *MATXH* **40** (1999), 7 - 138.

[9] D. F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinbatorial Theory* **B11** (1971), 105 - 119.

[10] P. F. Stadler. Landscapes and their correlation functions. *Journal of Mathematical Chemistry* **20** (1996), 1 - 45.

[11] P. F. Stadler. Spectral landscape theory. In: J. P. Crutchfield and P. Schuster, eds., *Evolutionary dynamics – exploring the interplay of selection, neutrality and function.* Oxford University Press, 2000.

[12] P. F. Stadler and G. Tinhofer. Equitable partitions, coherent algebras and random walks. *Match* **40** (1999), 215 - 261.

[13] P. F. Stadler and G. P. Wagner. Algebraic theory of recombination spaces. *Evolutionary Computation* **5** (1998), 241 - 275.

[14] B. Y. Weisfeiler and A. A. Leman. Reduction of a graph to a canonical form and an algebra arising during this reduction. Naucho - Techn. Inf., Ser. 2, **9** (1968), 12 - 16.

[15] B. J. Weisfeiler (Ed.), *On construction and identification of graphs. Springer Lecture Notes* 558 (1976).