MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **74** (2015) 223-246

ISSN 0340 - 6253

On the Hausdorff Distance between Some Families of Chemical Graphs

Aleksander Kelenc[1], **Andrej Taranenko**[1,2]

[1]*Faculty of Natural Sciences and Mathematics, University of Maribor, Slovenia*

[2]*Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia*

aleksander.kelenc@um.si, andrej.taranenko@um.si

(Received January 14, 2015)

Abstract

In chemistry, the problem known as similarity searching involves finding a set of molecules that are similar to a given sample molecule. The problem is tackled using graph theory and related algorithmic approaches involving some kind of measure of similarity of two graphs. In this paper we study a recently introduced Hausdorff distance between some families of graphs that often appear in chemical graph theory. Next to a few results for general graphs, we determine formulae for the distance between paths and cycles. For trees some bounds are proved. Also, an exact (exponential time) algorithm for determining the distance between two arbitrary trees is presented. We give examples emphasizing the difference between this new measure and some other known approaches, also reasons why some well-known algorithms for trees may not suffice to determine the Hausdorff distance of two trees are shown.

1 Introduction

In this paper we study a new measure of similarities of graphs (introduced in [3]) on common families of chemical graphs, namely paths, cycles and trees. Determining the distance between two graphs is closely related to the study of similarity of molecular structures and related algorithmic problems [11, 14].

The subgraph isomorphism problem is defined as follows. Given two graphs G and H, does there exist a subgraph of G isomorphic to H. The subgraph isomorphism problem is known to be NP-complete [7].

The so-called *structure searching* mostly uses a graph isomorphism algorithm to determine whether two molecular compounds are identical; *substructure searching* involves the subgraph isomorphism problem and involves determining whether any of the sample structures (usually saved in a database) contain a sample structure. Closely related to the topic of this paper is the problem in chemistry known as *similarity searching*: given a molecule of interest find in a database its nearest neighbours - those molecules which are most similar to the given sample using some measure of inter-molecular similarity [8].

Generally, in graph theory the distance between two graphs has been defined in various ways, for examples see [4–6, 9]. A common way is to define the distance as the minimum number of some operations (on vertices or edges) one needs to transform one graph into the other. Under the assumption that the graphs compared are of the same order and size, the operations defined were edge move [4], edge rotation [6] and edge slide [4, 9], among others.

A graph G is said to be a common subgraph of the graphs G_1 and G_2 if it holds that $G \subseteq G_1$ and $G \subseteq G_2$. We say that a common subgraph G of G_1 and G_2 is a maximal common subgraph if there does not exist a common subgraph H with $|V(H)| > |V(G)|$ and $G \subseteq H$. In [5], the authors use the notion of the maximal common subgraph to define the distance between two non-empty graphs, where the metric they define uses only the order of a maximal common subgraph and the order of the graphs compared. A measure of similarity of graphs based on a maximal (maximum) common subgraph is often used chemical graph theory to search for molecules that are measured to be close to each other.

In [3], Banič and Taranenko define the concept of so-called Hausdorff graphs. Together with amalgams (cf. [2, 10]) these graphs are used to define the Hausdorff distance on the class of all connected simple graphs as a new measure of similarity of two such graphs. We use this measure to determine the distance between paths, cycles and trees. The notion of the Hausdorff distance considers a special kind of a common subgraph of the graphs compared, as do many such measures, as well as the structural properties outside of the common subgraphs, which is, to our knowledge, new.

We proceed as follows. In the next section we present basic definitions and results needed for proper understanding of the results of this paper. We also prove some general results related to the measure itself. In Section 3 we give formulas for exact values of the Hausdorff distance between paths and cycles. Section 4 deals with the Hausdorff distance between two trees. We present some bounds and an exact exponential time algorithm that is used to determine the Hausdorff distance of two trees. Moreover, we give examples that show why some already known polynomial time algorithms for trees do not suffice in the case of the Hausdorff distance. We conclude the paper with two open problems.

2 Definitions and Notations

A *graph* $G = (V(G), E(G))$ is determined by a non-empty *vertex set* $V(G)$ and a set $E(G)$ of unordered pairs of vertices $\{u, v\}$, called the set of *edges*. We will use the short notation uv for edge $\{u, v\}$. We say that a vertex u is adjacent to a vertex v if $uv \in E(G)$.

Let $G = (V(G), E(G))$ and $H = (V(H), E(H))$ be any graphs. If $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$, then we say that H is a subgraph of G and write $H \subseteq G$.

All graphs considered in the paper are simple graphs, i.e. the graphs without multiple edges and without loops ($uu \notin E(G)$ for any $u \in V(G)$).

Let G be a graph and let $S \subseteq V(G)$. By $\langle S\rangle$ we denote the subgraph of G induced by the set S, i.e. for all $u, v \in S$, $uv \in E(\langle S\rangle)$ if and only if $uv \in E(G)$.

Two graphs are *isomorphic*, if there is a bijection between their vertex sets that preserves adjacency and non-adjacency of the vertices.

A path P from a vertex x to a vertex y in G is a sequence $x = v_0v_1v_2 \ldots v_{k-1}v_k = y$ of pairwise different vertices of G, where $v_iv_{i+1} \in E(G)$, for each $i \in \{0, \ldots, k-1\}$. The vertices x and y are called the *endpoints* of the path. The *length* of a path P, denoted by $\ell(P)$, is the number of edges in P.

The *distance* between vertices x and y, denoted by $d_G(x, y)$, is the length of a shortest path between x and y in G.

A graph G is connected if for each $u, v \in V(G)$ there is a path in G from u to v.

A connected subgraph H of a graph G is convex in G if for any $u, v \in V(H)$, $P \subseteq H$ for any shortest path P from u to v in G.

Let G be a graph and v be a vertex of G. The *eccentricity* of the vertex v, denoted $\mathrm{e}(v)$ is the maximum distance from v to any vertex of G. That is, $\mathrm{e}(v) = \max\{d_G(v, u) | u \in V(G)\}$. The *radius* of G, denoted $\mathrm{rad}(G)$, is the minimum eccentricity among the vertices of G. Therefore, $\mathrm{rad}(G) = \min\{\mathrm{e}(v) | v \in V(G)\}$. The *diameter* of G, denoted $\mathrm{diam}(G)$, is the maximum eccentricity among the vertices of G. Thus, $\mathrm{diam}(G) = \max\{\mathrm{e}(v) | v \in V(G)\}$. The *center* of G is the set of vertices with eccentricity equal to the radius. Hence, center$(G) = \{v \in V(G)|\mathrm{e}(v) = \mathrm{rad}(G)\}$. A vertex $v \in$ center(G) is called a *central vertex* of G. It is well known that for an arbitrary graph G it holds that $\mathrm{rad}(G) \leq \mathrm{diam}(G) \leq 2 \cdot \mathrm{rad}(G)$.

We will also need the following definitions form [3].

Definition 2.1. [3] Let G be an arbitrary graph. The Hausdorff graph of the graph G, denoted by 2^G, has for the vertex set $V(2^G)$ the set of all non-empty subgraphs of G. The adjacency of vertices in 2^G is defined as follows. For all $H_1, H_2 \in V(2^G)$, $H_1 \neq H_2$, it holds that $H_1H_2 \in E(2^G)$ if and only if

1. for each $v \in V(H_1)$ there exists $v' \in V(H_2)$ such that $d_G(v, v') \leq 1$ and
2. for each $v' \in V(H_2)$ there exists $v \in V(H_1)$ such that $d_G(v', v) \leq 1$.

The Hausdorff metric h_G between two subgraphs of a graph G is defined in the following definition. It will tell us how much two subgraphs of G coincide.

Definition 2.2. [3] Let G be an arbitrary graph. The *distance between two subgraphs H_1 and H_2 of G*, denoted by $h_G(H_1, H_2)$, is the distance between their corresponding vertices in 2^G. In other words,

$$h_G(H_1, H_2) := d_{2^G}(H_1, H_2).$$

We call h_G the *Hausdorff metric on* 2^G.

Note that for two different isomorphic subgraphs H_1 and H_2 of a graph G, the value $h_G(H_1, H_2)$ may be arbitrarily large. Moreover, the following corollary is also proven in [3].

Corollary 2.3. [3] If G is connected, then h_G is a metric on $V(2^G)$.

Definition 2.4. Let H_1 be a (convex) subgraph of G_1 and H_2 a (convex) subgraph of G_2. If H_1 and H_2 are isomorphic graphs, then a (convex) amalgam of G_1 and G_2 is any graph A obtained from G_1 and G_2 by identifying their subgraphs H_1 and H_2. We call the isomorphic copies of G_1 and G_2 in A the covers of the amalgam A and denote them by G_1^A and G_2^A, respectively. See Figure 1 for reference.

Denote by $\mathcal{A}(G_1, G_2)$ and $\mathcal{X}(G_1, G_2)$ the sets of all amalgams and all convex amalgams of the graphs G_1 and G_2, respectively.

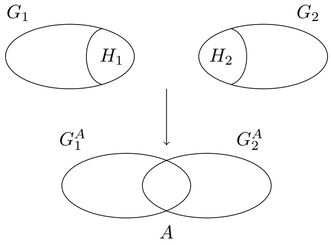

Figure 1: An amalgam A of G_1 and G_2.

Remark 2.5. Let A be an amalgam of G_1 and G_2 obtained from G_1 and G_2 by identifying their subgraphs H_1 and H_2. Then $G_1^A \cap G_2^A = H_1^A = H_2^A$ is isomorphic to H_1 and H_2.

Let $\mathcal{G}$ be the family of all simple connected graphs.

Theorem 2.6. *[3, Theorem 4.10] Let $G_1, G_2 \in \mathcal{G}$. Let d be a non-negative integer and A an amalgam of G_1 and G_2. Then $h_A(G_1^A, G_2^A) = d$ if and only if*

(i) for each $u \in V(G_1^A)$ there is a vertex $v \in V(G_2^A)$ such that $d_A(u,v) \leq d$,

(ii) for each $u \in V(G_2^A)$ there is a vertex $v \in V(G_1^A)$ such that $d_A(u,v) \leq d$, and

(iii) there is $u \in V(G_1^A)$ such that for each vertex $v \in V(G_1^A \cap G_2^A)$ the distance $d_A(u,v) \geq d$ or

there is $u \in V(G_2^A)$ such that for each vertex $v \in V(G_1^A \cap G_2^A)$ the distance $d_A(u,v) \geq d$.

From Theorem 2.6 we get the following Corollary.

Corollary 2.7. Let $G_1, G_2 \in \mathcal{G}$. Let A be an amalgam of G_1 and G_2. Then

$$h_A(G_1^A, G_2^A) = \max_{u \in V(A)} \{d_A(u, G_1^A \cap G_2^A\} \ .$$

Proof. Let $d := \max_{u \in V(A)}\{d_A(u, G_1^A \cap G_2^A\}$ and $u \in V(G_i^A)$, for some $i \in \{1,2\}$, such that $d_A(u, G_1^A \cap G_2^A) = d$. Then for every vertex $v \in V(G_1^A \cap G_2^A)$ it holds that $d_A(u,v) \geq d$. Therefore, the condition (iii) of Theorem 2.6 holds true.

Choose a vertex $u_1 \in V(G_1^A)$. Let $v_1 \in V(G_1^A \cap G_2^A)$ be such that $d_A(u_1, v_1) = d_A(u_1, G_1^A \cap G_2^A)$. Then $d_A(u_1, v_1) \leq \max_{u \in V(G_1^A)}\{d_A(u, G_1^A \cap G_2^A\} \leq d$. It follows that the condition (i) of Theorem 2.6 holds true.

Following the same line of thought one can prove that the condition (ii) of Theorem 2.6 is also fulfilled.

Since all of the conditions of Theorem 2.6 hold true, the assertion follows immediately. ■

Given $G_1, G_2 \in \mathcal{G}$ and an amalgam A of G_1 and G_2, Corollary 2.7 says that to determine $h_A(G_1^A, G_2^A)$ it suffices to find a vertex $v \in V(A)$ with the maximum distance to $G_1^A \cap G_2^A$, since $h_A(G_1^A, G_2^A) = d_A(v, G_1^A \cap G_2^A)$. This idea will be used a lot in our proofs.

Finally, the Hausdorff distance $\mathcal{H} : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ on $\mathcal{G}$ can be defined as follows:

Definition 2.8. [3] For any graphs $G_1, G_2 \in \mathcal{G}$, we define

$$\mathcal{H}(G_1, G_2) = \begin{cases} \min\left\{h_A(G_1^A, G_2^A) \mid A \in \mathcal{X}(G_1, G_2)\right\}, & \text{if } G_1 \ncong G_2 \\ 0, & \text{if } G_1 \cong G2 \end{cases}.$$

We call $\mathcal{H}$ *the Hausdorff distanc*e on $\mathcal{G}$.

Note, Definition 2.8 is equivalent to definition of the Hausdorff distance in [3, Definition 4.18]. Moreover, it is proven in [3] that $\mathcal{H}$ is a metric on the class of all simple connected pairwise non-isomorphic graphs. A convex amalgam A of two simple connected graphs G_1 and G_2, for which $h_A(G_1^A, G_2^A) = \mathcal{H}(G_1, G_2)$ is called an *optimal amalgam.*

As noted in [3], for fixed isomorphic subgraphs H_1 and H_2 of G_1 and G_2, respectively, there may be many isomorphisms from H_1 onto H_2. Therefore there may be more than just one amalgam A of G_1 and G_2, which is obtained by identifying H_1 and H_2 (see Example 2.9).

Example 2.9. Let G_1 and G_2 be the graphs depicted in Figure 2, and H_1 and H_2 their subgraphs, respectively, both isomorphic to P_2. Let f_1 and f_2 be two isomorphisms from H_1 onto H_2. In Figure 2 they are depicted by dotted and dashed arrows, respectively. Next, let A_i be the amalgam of G_1 and G_2 obtained by identifying H_1 and H_2 according to the isomorphism f_i, $i \in \{1, 2\}$. Obviously, A_1 and A_2 are not isomorphic, although they were both obtained by identifying the same subgraphs.

In the next theorem we prove that distance between the covers of a convex amalgam (therefore also the Hausdorff distance between two graphs) is not dependant on the choice of the isomorphism between the subgraphs.

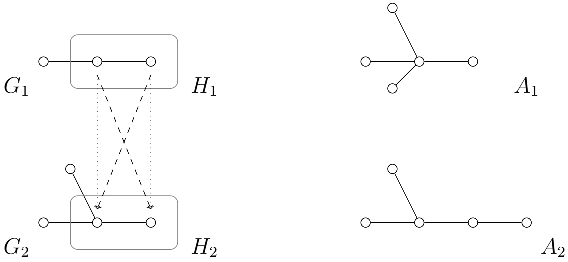

Figure 2: The amalgams A_1 and A_2 from Example 2.9.

Theorem 2.10. *Let $G_1, G_2 \in \mathcal{G}$ and let H_1 and H_2 be fixed isomorphic convex subgraphs of G_1 and G_2, respectively. Also, let f_1 and f_2 be any two isomorphisms between H_1 and H_2, and A_1 and A_2 the two convex amalgams of G_1 and G_2 obtained by identifying H_1 and H_2 with respect to isomorphisms f_1 and f_2, respectively. Then $h_{A_1}(G_1^{A_1}, G_2^{A_1}) = h_{A_2}(G_1^{A_2}, G_2^{A_2})$.*

Proof. Let $h_k = h_{A_k}(G_1^{A_k}, G_2^{A_k})$, for each $k \in \{1,2\}$. Towards contradiction, suppose $h_{A_1}(G_1^{A_1}, G_2^{A_1}) < h_{A_2}(G_1^{A_2}, G_2^{A_2})$. Then by Corollary 2.7 there exists a vertex $u \in G_i^{A_1}$, for some $i \in \{1,2\}$, with $d_{A_1}(u, G_1^{A_1} \cap G_2^{A_1}) = h_1$. Let $v \in V(G_1^{A_1} \cap G_2^{A_1})$, with $d_{A_1}(u,v) = h_1$. Similarly, there exists a vertex $x \in G_i^{A_2}$, for some $i \in \{1,2\}$, with $d_{A_2}(x, G_1^{A_2} \cap G_2^{A_2}) = h_2$. Let $y \in V(G_1^{A_2} \cap G_2^{A_2})$ with $d_{A_2}(x,y) = h_2$. Denote by x' the vertex in a cover of A_1 corresponding to x, and by $y' \in V(G_1^{A_1} \cap G_2^{A_1})$ corresponding to y.

Obviously, $d_{A_1}(x', G_1^{A_1} \cap G_2^{A_1}) = d_{A_1}(x', y')$. Moreover, $d_{A_1}(x',y') = d_{A_2}(x,y) = h_2$. By Corollary 2.7 for all $w \in V(A_1)$ holds that $d_{A_1}(u,v) \geq d_{A_1}(w, G_1^{A_1} \cap G_2^{A_1})$. Therefore $h_1 = d_{A_1}(u,v) \geq d_{A_1}(x', G_1^{A_1} \cap G_2^{A_1}) = d_{A_1}(x',y') = h_2$, so $h_1 \geq h_2$, a contradiction with our assumption.

Similarly one can disprove the case that $h_{A_2}(G_1^{A_2}, G_2^{A_2}) < h_{A_1}(G_1^{A_1}, G_2^{A_1})$. Therefore the assertion follows. ■

Let G be a graph and H its convex subgraph. The distance between H and G is defined as $\max_{v \in V(G)}\{d_G(v,H)\}$. Note, that G can be looked at as an amalgam of G and H', where H' is isomorphic to H, and the amalgam of G and H' is obtained by identifying H and H'. Therefore by Corollary 2.7, $\max_{v \in V(G)}\{d_G(v,H)\} = h_G(G^G, H^G)$.

Proposition 2.11. *Let $G_1, G_2 \in \mathcal{G}$. Let H_1 and H_2 be two isomorphic convex subgraphs of G_1 with $d_1 \leq d_2$, where d_1 and d_2 are the distances between H_1 and G_1, and H_2 and*

G_1, respectively. Let H_3 be a convex subgraph of G_2 isomorphic to H_1 (and H_2). Let A_1 be a convex amalgam of G_1 and G_2 obtained by identifying H_1 and H_3, and A_2 be a convex amalgam of G_1 and G_2 obtained by identifying H_2 and H_3. Then $h_{A_1}(G_1^{A_1}, G_2^{A_1}) \leq h_{A_2}(G_1^{A_2}, G_2^{A_2})$ holds true.

Proof. Let d_3 be the distance between H_3 and G_2. From Corollary 2.7 it follows that $h_{A_1}(G_1^{A_1}, G_2^{A_1}) = \max\{d_1, d_3\}$ and $h_{A_2}(G_1^{A_2}, G_2^{A_2}) = \max\{d_2, d_3\}$. Since $d_1 \leq d_2$ it follows that $\max\{d_1, d_3\} \leq \max\{d_2, d_3\}$ which implies $h_{A_1}(G_1^{A_1}, G_2^{A_1}) \leq h_{A_2}(G_1^{A_2}, G_2^{A_2})$. ■

For two arbitrary simple connected graphs, the upper bound for the Hausdorff distance can be expressed using the radius of the graphs.

Theorem 2.12. *Let G_1 and G_2 be two arbitrary simple, connected graphs. Then*

$$\mathcal{H}(G_1, G_2) \leq \max\left\{\mathrm{rad}(G_1), \mathrm{rad}(G_2)\right\}.$$

Proof. Let c_1 be a central vertex of G_1 and c_2 be a central vertex of G_2. Let A be an amalgam which is created by identifying c_1 and c_2. Since there is exactly one vertex in $G_1^A \cap G_2^A$, A is a convex amalgam. In G_1 it holds that for each $v \in V(G_1)$ the distance $d_{G_1}(v, c_1) \leq \mathrm{rad}(G_1)$. Similarly, in G_2 it holds that for each $v \in V(G_2)$ the distance $d_{G_2}(v, c_2) \leq \mathrm{rad}(G_2)$. Since A is a convex amalgam, the same holds for the corresponding vertices of G_1^A and G_2^A in A. Using Corollary 2.7, it follows that $h_A(G_1^A, G_2^A) = \max\left\{\mathrm{rad}(G_1), \mathrm{rad}(G_2)\right\}$ and $\mathcal{H}(T_1, T_2) \leq \max\left\{\mathrm{rad}(G_1), \mathrm{rad}(G_2)\right\}$. ■

Note, this bound is sharp if one of the graphs is trivial (a vertex graph).

3 Results on some simple families of graphs

In this section we give some results about the Hausdorff distance between two graphs of some simple families of graphs, that often appear in chemical graph theory.

First, consider the following Remarks which can be easily verified.

Remark 3.1. We will often use the following implication. If a and b are two arbitrary positive integers with $a < b$, then $2a < 2b - 1$. Clearly, if $b \geq a + 1$, then $2b \geq 2a + 2 > 2a + 1$.

Remark 3.2. For an arbitrary positive integer m the following equality holds:

$$\left\lceil \frac{\left\lfloor \frac{m}{2} \right\rfloor}{2} \right\rceil = \left\lceil \frac{m-1}{4} \right\rceil.$$

Note, for a path every connected subgraphs is also a convex subgraph. Now we give formulae for the Hausdorff distance between some simple families of graphs. In all cases we construct a convex amalgam and thus obtain an upper bound. Then we show there can be no amalgam, that would give a better upper bound.

Proposition 3.3. *Let P_n and P_m be two paths on n and m vertices, respectively, with $n \geq m \geq 1$. Then $\mathcal{H}(P_n, P_m) = \left\lceil \frac{n-m}{2} \right\rceil$.*

Proof. Denote the vertices of P_n with $u_1, \ldots, u_n$, where $u_i u_{i+1} \in E(P_n)$, for each $i \in \{1, \ldots, n-1\}$, and the vertices of P_m with $v_1, \ldots, v_m$, where $v_i v_{i+1} \in E(P_m)$, for each $i \in \{1, \ldots, m-1\}$.

Let A be an amalgam which is created by identifying pairs of vertices $u_{\lceil \frac{n-m}{2} \rceil + i}$ and v_i for each $1 \leq i \leq m$. A is clearly a convex amalgam. Using Corollary 2.7 we immediately get that $h_A(P_n^A, P_m^A) = \left\lceil \frac{n-m}{2} \right\rceil$ and therefore $\mathcal{H}(P_n, P_m) \leq \left\lceil \frac{n-m}{2} \right\rceil$.

Suppose now, that there exists an amalgam $A' \in \mathcal{X}(P_n, P_m)$ such that $k := h_{A'}(P_n^{A'}, P_m^{A'}) < \left\lceil \frac{n-m}{2} \right\rceil$. Due to Corollary 2.7, for each $w \in V(A')$ it holds that $k \geq d_{A'}(w, P_n^{A'} \cap P_m^{A'})$. The graph $P_n^{A'} \cap P_m^{A'}$ is isomorphic to a path with at most m vertices. Then, for every path P in A' it follows that the length $\ell(P) \leq m - 1 + 2k$. It holds that $\ell(P) \leq m - 1 + 2k < m - 1 + 2\left\lceil \frac{n-m}{2} \right\rceil - 1 \leq m - 1 + 2\frac{n-m+1}{2} - 1 = n - 1$. So, for every path P in A' it holds that $\ell(P) < n - 1$. But $P_n^{A'} \subseteq A'$ and $\ell(P_n^{A'}) = n - 1$; this is a contradiction with the assumption that such an amalgam A' exists. ■

Let C_n be a cycle on n vertices, with $n \geq 3$. Then the largest convex subgraph of C_n is a path on $\left\lceil \frac{n}{2} \right\rceil$ vertices.

Proposition 3.4. *Let P_n and C_m be a path and a cycle on n and m vertices, respectively, with $n \geq 1$ and $m \geq 3$. Then*

$$\mathcal{H}(P_n, C_m) = \begin{cases} \left\lceil \frac{m-n}{2} \right\rceil, & n \leq \frac{m}{2} \\ \left\lceil \frac{m-1}{4} \right\rceil, & \frac{m}{2} < n \leq m \\ \left\lceil \frac{n - \left\lceil \frac{m}{2} \right\rceil}{2} \right\rceil, & n > m \end{cases}$$

Proof. Denote vertices of P_n with $u_1, \ldots, u_n$ where $u_i u_{i+1} \in E(P_n)$, for each $i \in \{1, \ldots, n-1\}$, and vertices of C_m with $v_0, v_1, v_2, \ldots, v_{m-1}$ where $v_i v_{i+1} \in E(C_m)$, for each $i \in \{0, \ldots, m-1\}$. All indices in C_m are computed modulo m.

Let $n \leq \frac{m}{2}$. Let A be an amalgam which is created by identifying pairs of vertices u_i and v_i, for each $1 \leq i \leq n$. Since every subgraph of C_m isomorphic to a path on n vertices is a convex subgraph of C_m, A is a convex amalgam. Clearly, $\max_{u \in V(A)}\{d_A(u, P_n^A \cap C_m^A)\} = \left\lceil \frac{m-n}{2} \right\rceil$. Using Corollary 2.7 it follows that $h_A(P_n^A, C_m^A) = \left\lceil \frac{m-n}{2} \right\rceil$ and $\mathcal{H}(P_n, C_m) \leq \left\lceil \frac{m-n}{2} \right\rceil$.

Suppose there exists a convex amalgam A' with $h_{A'}(P_n^{A'}, C_m^{A'}) < \left\lceil \frac{m-n}{2} \right\rceil$. Define $h := \left\lceil \frac{m-n}{2} \right\rceil$. Due to convexity, $P_n^{A'} \cap C_m^{A'}$ is isomorphic to a path on k vertices, $1 \leq k \leq n$. Say the vertices in $P_n^{A'} \cap C_m^{A'}$ are $v_i^{A'}, v_{i+1}^{A'}, \ldots, v_{i+k-1}^{A'}$, with an edge between two consecutive vertices. We now consider the vertex $v_{i-h}^{A'}$. Clearly, $d_{A'}(v_i^{A'}, v_{i-h}^{A'}) = h = \left\lceil \frac{m-n}{2} \right\rceil$. On the other hand,

$$\begin{aligned} d_{A'}(v_{i+k-1}^{A'}, v_{i-h}^{A'}) &= \\ \ell(C_m) - h - (k-1) &= \\ m - h - k + 1 &\geq \\ m - h - n + 1 &= \\ 2\frac{m-n+1}{2} - h &\geq \\ 2\left\lceil \frac{m-n}{2} \right\rceil - h &= h. \end{aligned}$$

It follows that $d_{A'}(v_{i-h}^{A'}, P_n^{A'} \cap C_m^{A'}) = \left\lceil \frac{m-n}{2} \right\rceil > h_{A'}(P_n^{A'}, C_m^{A'})$. A contradiction with Corollary 2.7.

Let $\frac{m}{2} < n \leq m$. Set $l := \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$. Let A be an amalgam which is created by identifying pairs of vertices u_{i+l+1} and v_i for each $0 \leq i < \lceil \frac{m}{2} \rceil$, see Figure 3 for reference. It is easy to verify that A is a convex amalgam. Due to Corollary 2.7, to determine the value of $h_A(P_n^A, C_m^A)$ it suffices to find the vertex in A with the maximum distance to $P_n^A \cap C_m^A$. Clearly, the candidates are the two endpoints of the path P_n^A that are outside of $P_n^A \cap C_m^A$ (vertices u_1^A and u_n^A) and a vertex of $V(C_m^A) \setminus V(P_n^A \cap C_m^A)$ with the maximum distance to $P_n^A \cap C_m^A$ (the vertex $v_{0-\lceil \frac{\lfloor m/2 \rfloor}{2} \rceil}^A$).

Note, that $d_A(u_1^A, P_n^A \cap C_m^A) = d_A(u_1^A, u_{l+1}^A) = l$ and $d_A(u_n^A, P_n^A \cap C_m^A) = d_A(u_n^A, u_{l+\lceil \frac{m}{2} \rceil}^A)$. The distance between the vertices u_n^A and $u_{l+\lceil \frac{m}{2} \rceil}^A$ can be expressed as the difference between the length of the path P_n and the length of the path between u_1^A and $u_{l+\lceil \frac{m}{2} \rceil}^A$.

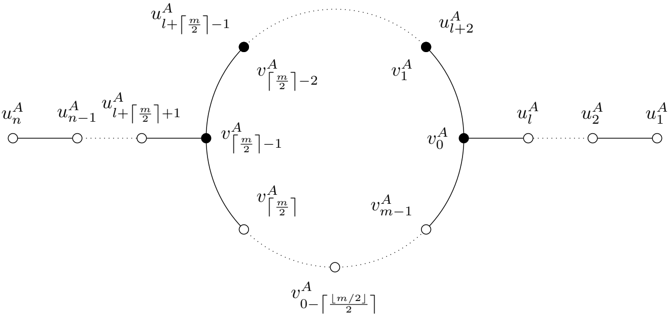
$u^A_{l+\lceil \frac{m}{2}\rceil-1}$
u^A_{l+2}
$v^A_{\lceil \frac{m}{2}\rceil-2}$
v^A_1
u^A_n
u^A_{n-1}
$u^A_{l+\lceil \frac{m}{2}\rceil+1}$
$v^A_{\lceil \frac{m}{2}\rceil-1}$
v^A_0
u^A_l
u^A_2
u^A_1
$v^A_{\lceil \frac{m}{2}\rceil}$
v^A_{m-1}
$v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}$

Figure 3: An amalgam A of path P_n (vertices u_i) and cycle C_m (vertices v_j).

Therefore,

$$\begin{aligned} d_A(u^A_n, u^A_{l+\lceil \frac{m}{2}\rceil}) &= \\ n-1-(l+\left\lceil \frac{m}{2}\right\rceil -1) &= \\ n-1-l-\left\lceil \frac{m}{2}\right\rceil +1 &= \\ 2\frac{n-\lceil \frac{m}{2}\rceil}{2} - l &\leq \\ 2\left\lceil \frac{n-\lceil \frac{m}{2}\rceil}{2}\right\rceil - l &= \\ 2l-l = l. \end{aligned}$$

The distance $d_A(v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}, P^A_n \cap C^A_m) = \min\{\left\lceil \frac{\lfloor m/2\rfloor}{2}\right\rceil, d_A(v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}, v^A_{\lceil \frac{m}{2}\rceil-1})\}$. It holds that

$$\begin{aligned} d_A(v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}, v^A_{\lceil \frac{m}{2}\rceil-1}) &= \\ m - d_A(v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}, v^A_0) - d_A(v^A_0, v^A_{\lceil \frac{m}{2}\rceil-1}) &= \\ m - \left\lceil \frac{m}{2}\right\rceil + 1 - \left\lceil \frac{\lfloor \frac{m}{2}\rfloor}{2}\right\rceil &= \\ \left\lfloor \frac{m}{2}\right\rfloor - \left\lceil \frac{\lfloor \frac{m}{2}\rfloor}{2}\right\rceil + 1 &= \\ \left\lfloor \frac{\lfloor \frac{m}{2}\rfloor}{2}\right\rfloor + 1 \geq \left\lceil \frac{\lfloor \frac{m}{2}\rfloor}{2}\right\rceil. \end{aligned}$$

Therefore, $d_A(v^A_{0-\lceil \frac{\lfloor m/2\rfloor}{2}\rceil}, P^A_n \cap C^A_m) = \left\lceil \frac{\lfloor m/2\rfloor}{2}\right\rceil$. Since $l \leq \left\lceil \frac{\lfloor m/2\rfloor}{2}\right\rceil$, by Corollary 2.7, $h_A(P^A_n, C^A_m) = \left\lceil \frac{\lfloor m/2\rfloor}{2}\right\rceil$. It follows that $\mathcal{H}(P_n, C_m) \leq \left\lceil \frac{\lfloor m/2\rfloor}{2}\right\rceil$. See Figure 3 for reference.

Suppose there exists a convex amalgam A' with $h_{A'}(P_n^{A'}, C_m^{A'}) < \left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil$. Define $h := \left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil$. Again, due to convexity, $P_n^{A'} \cap C_m^{A'}$ is isomorphic to a path on k vertices, $1 \leq k \leq \lceil \frac{m}{2} \rceil$. Say the vertices in $P_n^{A'} \cap C_m^{A'}$ are $v_i^{A'}, v_{i+1}^{A'}, \ldots, v_{i+k-1}^{A'}$. We consider the vertex $v_{i-h}^{A'}$. Since $d_{A'}(v_i^{A'}, v_{i-h}^{A'}) = h$ and

$$\begin{aligned}
d_{A'}(v_{i+k-1}^{A'}, v_{i-h}^{A'}) &= \\
m - d_{A'}(v_i^{A'}, v_{i+k-1}^{A'}) - d_{A'}(v_i^{A'}, v_{i-h}^{A'}) &= \\
m - (k-1) - h &= \\
m - k + 1 - h &\geq \\
m - \left\lceil \frac{m}{2} \right\rceil + 1 - h &= \\
\left\lfloor \frac{m}{2} \right\rfloor + 1 - \left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil &= \\
\left\lfloor \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rfloor + 1 &\geq \\
\left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil &= h,
\end{aligned}$$

it follows that $d_{A'}(v_{i-h}^{A'}, P_n^{A'} \cap C_m^{A'}) = h = \left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil > h_{A'}(P_n^{A'}, C_m^{A'})$. A contradiction with Corollary 2.7. Using Remark 3.2 the assertion follows.

Let $n > m$. Set $l := \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$. Let A be an amalgam which is created by identifying pairs of vertices u_{i+l+1} and v_i for each $0 \leq i < \lceil \frac{m}{2} \rceil$. It is easy to verify that A is a convex amalgam. As in the previous case, the value of $h_A(P_n^A, C_m^A)$ can be determined by finding a vertex of A with the maximum distance to $P_n^A \cap C_m^A$; the same candidate vertices have to be considered (vertices u_1^A, u_n^A and $v^A_{0 - \lceil \frac{\lfloor m/2 \rfloor}{2} \rceil}$). Following the same line of thought as in the previous case and taking into account that $l \geq \left\lceil \frac{\lfloor \frac{m}{2} \rfloor}{2} \right\rceil$, it follows that $h_A(P_n^A, C_m^A) = \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$ and $\mathcal{H}(P_n, C_m) \leq \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$.

Suppose there exists a convex amalgam A' with $h_{A'}(P_n^{A'}, C_m^{A'}) < \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$. Due to convexity, $P_n^{A'} \cap C_m^{A'}$ is isomorphic to a path on k vertices, $1 \leq k \leq \lceil \frac{m}{2} \rceil$. Say the vertices in $P_n^{A'} \cap C_m^{A'}$ are $v_i^{A'}, v_{i+1}^{A'}, \ldots, v_{i+k-1}^{A'}$. The length of path $P_n^{A'}$ is clearly $n-1$ and equals $d_{A'}(u_1^{A'}, v_i^{A'}) + d_{A'}(v_i^{A'}, v_{i+k-1}^{A'}) + d_{A'}(v_{i+k-1}^{A'}, u_n^{A'})$. On the other hand, by Corollary 2.7, it holds that $d_{A'}(u_1^{A'}, v_i^{A'}) \leq h_{A'}(P_n^{A'}, C_m^{A'})$ and $d_{A'}(v_{i+k-1}^{A'}, u_n^{A'}) \leq h_{A'}(P_n^{A'}, C_m^{A'})$. Putting this together we get that

$$\begin{aligned}\ell(P_n^{A'}) = d_{A'}(u_1^{A'}, v_i^{A'}) + d_{A'}(v_i^{A'}, v_{i+k-1}^{A'}) + d_{A'}(v_{i+k-1}^{A'}, u_n^{A'}) &\le \\ h_{A'}(P_n^{A'}, C_m^{A'}) + k - 1 + h_{A'}(P_n^{A'}, C_m^{A'}) &< \\ 2\left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil - 1 + k - 1 &\le \\ 2\frac{n - \lceil \frac{m}{2} \rceil + 1}{2} - 1 + \left\lceil \frac{m}{2} \right\rceil - 1 &= n - 1.\end{aligned}$$

So, $n - 1 = \ell(P_n^{A'}) < n - 1$, a contradiction. ∎

Now, we derive a formula for the Hausdorff distance between two cycles. If the cycles are isomorphic, the Hausdorff distance equals 0 by definition. For non-isomorphic cycles we get the following proposition.

Proposition 3.5. *Let C_n and C_m be two cycles of length n and m, respectively, with $n > m \geq 3$. Then $\mathcal{H}(C_n, C_m) = \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$.*

Proof. Denote vertices of C_n with $u_0, \ldots, u_{n-1}$, where $u_i u_{i+1} \in E(C_n)$, for each $i \in \{0, \ldots, n-1\}$, and vertices of C_m with $v_0, \ldots, v_{m-1}$, where $v_i v_{i+1} \in E(C_m)$, for each $i \in \{0, \ldots, m-1\}$. All indices are computed modulo of the length of the corresponding cycle.

Let A be an amalgam which is created by identifying pairs of vertices u_i and v_i, for each $1 \leq i \leq \lceil \frac{m}{2} \rceil$. Since every subgraph of C_m (C_n) isomorphic to a path on $\lceil \frac{m}{2} \rceil$ vertices is a convex subgraph of C_m (and also C_n), A is a convex amalgam. Then by Corollary 2.7 $h_A(C_n^A, C_m^A) = \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$ and $\mathcal{H}(C_n, C_m) \leq \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$.

Suppose there exists a convex amalgam A' with $h_{A'}(C_n^{A'}, C_m^{A'}) < \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$. Therefore $C_n^{A'} \cap C_m^{A'}$ is isomorphic to a path on k vertices, $1 \leq k \leq \lceil \frac{m}{2} \rceil$. Say the vertices in $C_n^{A'} \cap C_m^{A'}$ are $u_i^{A'}, u_{i+1}^{A'}, \ldots, u_{i+k-1}^{A'}$. We now choose the vertex $u^{A'}_{i - \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil}$. Since $d_{A'}(u_i^{A'}, u^{A'}_{i - \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil}) = \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$ and $d_{A'}(u_{i+k-1}^{A'}, u^{A'}_{i - \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil}) \geq \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil$, it follows that $d_{A'}(u^{A'}_{i - \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil}, C_n^{A'} \cap C_m^{A'}) = \left\lceil \frac{n - \lceil \frac{m}{2} \rceil}{2} \right\rceil > h_{A'}(C_n^{A'}, C_m^{A'})$. A contradiction with Corollary 2.7. ∎

4 Trees and the Hausdorff distance

Trees often appear in chemical graph theory, since many organic molecules have a graph representation that is a tree (e.g. saturated hydrocarbons). Isomers, for example, have the same chemical formula but different molecular structures. One of the problems arisen with respect to the chemical structure is to determine whether two chemical structures are the same or how similar they are. Say that the chemical structures can be presented as trees. This means we have to determine whether two trees are isomorphic; this is a simple problem and can be done in linear time [1]. Also, as a measure of similarity of two non-isomorphic trees one can use a maximum common subtree of the two trees compared. The problem of finding a maximum common subtree of two arbitrary trees can be done in non-linear polynomial time [13].

On the other hand, to determine the Hausdorff distance between two trees, using a maximum common subtree to form a convex amalgam of two arbitrary trees may not produce an optimal amalgam (see Example 4.8). Therefore the mentioned algorithms may not suffice to determine the Hausdorff distance of two arbitrary trees.

In the next section we present some bounds for the Hausdorff distance between two trees, some formulae for special cases and in Section 4.2 an exact (exponential time) algorithm for computing the Hausdorff distance between two trees.

4.1 Hausdorff distance between trees

It is well known that any tree has either exactly one central vertex or exactly two central vertices that are adjacent. We say that a tree T is *central*, if $|\text{center}(T)| = 1$, otherwise it is *bicentral*. Also, for an arbitrary tree T it holds that $\text{diam}(T) = 2\text{rad}(T) - 1$, if T is bicentral, and $\text{diam}(T) = 2\text{rad}(T)$, if T is central. This fact, together with Theorem 2.12, immediately implies the following corollary.

Corollary 4.1. Let T_1 and T_2 be two arbitrary trees. Then

$$\mathcal{H}(T_1, T_2) \leq \max\left\{ \left\lceil \frac{\text{diam}(T_1)}{2} \right\rceil, \left\lceil \frac{\text{diam}(T_2)}{2} \right\rceil \right\}.$$

Clearly, if one of the trees is trivial, one obtains an optimal amalgam of the two trees by identifying the only vertex of the trivial tree with a central vertex of the other tree. For this reason in the following results we restrict ourselves to non-trivial trees.

Proposition 4.2. *Let T_1 and T_2 be two non-trivial trees with* $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$*. If T_1 is bicentral then $\mathcal{H}(T_1, T_2) < \operatorname{rad}(T_1)$.*

Proof. Let $\operatorname{center}(T_1) = \{c_1, c_2\}$. Let c be a central vertex of T_2 and c' its arbitrary neighbour, if T_2 is central, otherwise let c' be the other central vertex of T_2. Let H_1 be the subgraph of T_1 induced on the set $\operatorname{center}(T_1)$, and H_2 the subgraph of T_2 induced on the set $\{c, c'\}$. Let A be a convex amalgam of T_1 and T_2 obtained by identifying the graphs H_1 and H_2.

For any vertex $u \in V(T_1^A)$ it holds that $d_A(u, T_1^A \cap T_2^A) < \operatorname{rad}(T_1)$, since both central vertices are in $T_1^A \cap T_2^A$. Let $v \in V(T_2^A)$. If T_2 is bicentral (both its central vertices are also in $T_1^A \cap T_2^A$), then $d_A(v, T_1^A \cap T_2^A) < \operatorname{rad}(T_2) \leq \operatorname{rad}(T_1)$. If T_2 is central, then $\operatorname{rad}(T_2) < \operatorname{rad}(T_1)$. Since $c^A \in V(T_1^A \cap T_2^A)$ it holds that $d_A(v, T_1^A \cap T_2^A) \leq \operatorname{rad}(T_2) < \operatorname{rad}(T_1)$. Using Corollary 2.7 the assertion follows immediately. ■

Next, we study some properties of optimal amalgams of trees. Remember, a convex amalgam of two graphs is called optimal, if it gives rise to the Hausdorff distance between the two graphs.

Theorem 4.3. *Let T_1 and T_2 be two arbitrary non-trivial trees, with* $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$*. Let $c \in \operatorname{center}(T_1)$. Then for every optimal amalgam $A \in \mathcal{X}(T_1, T_2)$ it holds that $\{c^A\} \subseteq V(T_1^A \cap T_2^A)$.*

Proof. Assume there exists $A \in \mathcal{X}(T_1, T_2)$ with $h_A(T_1^A, T_2^A) = \mathcal{H}(T_1, T_2)$ such that at least one central vertex of T_1, say v, is not in $T_1^A \cap T_2^A$. Then it holds that $d_A(v, T_1^A \cap T_2^A) \geq 1$.

Suppose T_1 is central. Since $T_1^A \cap T_2^A$ is convex in A, then there exists a vertex $u \in V(T_1^A) \backslash V(T_1^A \cap T_2^A)$ with $d_A(v, u) = \left\lceil \frac{\operatorname{diam}(T_1^A)}{2} \right\rceil$. But then, $d_A(u, T_1^A \cap T_2^A) \geq \left\lceil \frac{\operatorname{diam}(T_1^A)}{2} \right\rceil + 1$. This is a contradiction with Corollary 2.7 and Corollary 4.1 together with the assumption $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$.

Suppose T_1 is bicentral. Since $T_1^A \cap T_2^A$ is convex in A, then there exists a vertex $u \in V(T_1^A) \backslash V(T_1^A \cap T_2^A)$ with $d_A(v, u) = \operatorname{rad}(T_1) - 1$. But then, $d_A(u, T_1^A \cap T_2^A) \geq \operatorname{rad}(T_1)$. This is a contradiction with Corollary 2.7 and Proposition 4.2. ■

Let G be a graph and H its subgraph with a property P. We say H is minimal subgraph with the property P if there exists no proper subgraph of H with the property P.

Theorem 4.4. *Let T_1 and T_2 be two arbitrary non-trivial trees, with* $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$*. Let $0 \leq k \leq rad(T_1)$ be a fixed integer. Let H be a minimal subtree of T_1 containing a central vertex of T_1, such that* $\max_{u \in V(T_1) \setminus V(H)}\{d_{T_1}(u, H)\} \leq k$*. If T_2 does not contain a subgraph isomorphic to H then $\mathcal{H}(T_1, T_2) > k$.*

Proof. Suppose, $\mathcal{H}(T_1, T_2) \leq k$, for some fixed integer $0 \leq k \leq rad(T_1)$. Then there exists a convex amalgam A of T_1 and T_2 such that $h_A(T_1^A, T_2^A) = k$. Let H' be the subgraph of T_1 corresponding to $T_1^A \cap T_2^A$. By Theorem 4.3 the graph H' contains a central vertex of T_1. By Corollary 2.7 it holds true that $\max_{u \in V(T_1) \setminus V(H')}\{d_{T_1}(u, H')\} \leq k$. Now let H be a minimal subtree of H' such that $\max_{u \in V(T_1) \setminus V(H)}\{d_{T_1}(u, H)\} \leq k$ is still true. Clearly, H is a (convex) subgraph of H', therefore H^A is a convex subgraph of $T_1^A \cap T_2^A$. Then T_2 clearly contains a subgraph isomorphic to H. ■

The minimal subgraph H of a tree T, with the properites as required by Theorem 4.4 can be easily found as follows. Set $S := \operatorname{center}(T)$. Say k is a fixed integer as in Theorem 4.4. Choose a central vertex c of the tree T. Now, for each leaf u of the tree consider the path P_u from the leaf to the central vertex c. If $\ell(P_u) \leq k$, then do nothing. Otherwise, let $v_u \in V(P_u)$ be the vertex with $d_T(u, v_u) = k$. Let R_u be the path from v_u to c. Add the vertices of R_u to S. Clearly, the graph induced on the vertices in S is the subgraph we are constructing, i. e. $H = \langle S \rangle$.

Theorem 4.3 says that the center of the tree with larger diameter is always in the intersection of an optimal amalgam. On the other hand, there exist trees T_1 and T_2, with $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$, such that no central vertex of T_2 is in $T_1^A \cap T_2^A$ for any optimal amalgam A of T_1 and T_2, as the following example demonstrates.

Example 4.5. In Figure 4 we have two non-isomorphic trees T_1 and T_2. The (connected) subgraphs induced on the sets of blacks vertices in each tree are clearly isomorphic. Moreover, since they are connected, they are also convex in the corresponding graphs. Therefore by identifying these two subgraphs we obtain a convex amalgam A such that, by Corollary 2.7, $h_A(T_1^A, T_2^A) = 4$. Therefore, $\mathcal{H}(T_1, T_2) \leq 4$.

To see that $\mathcal{H}(T_1, T_2) \geq 4$, suppose that there exists an amalgam $A' \in \mathcal{X}(T_1, T_2)$ for which it holds that $h_{A'}(T_1^{A'}, T_2^{A'}) \leq 3$. Using Theorem 4.4, a minimal subtree H of T_1 containig the center of T_1 and satisfying the condition $\max_{u \in V(T_1) \setminus V(H)}\{d_{T_1}(u, H)\} \leq 3$ is the subgraph induced on the set of vertices $\{v_1, v_2, \ldots, v_{11}\}$. Clearly, T_2 contains no subgraph isomorphic to H, therefore $\mathcal{H}(T_1, T_2) > 3$. It follows that $\mathcal{H}(T_1, T_2) = 4$.

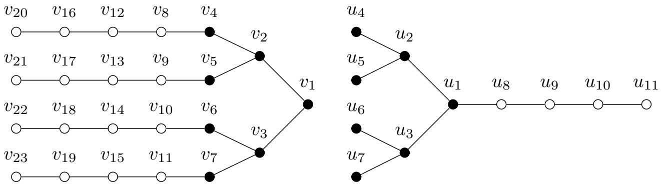
v_{20} v_{16} v_{12} v_8 v_4
v_2
v_{21} v_{17} v_{13} v_9 v_5
v_1
v_{22} v_{18} v_{14} v_{10} v_6
v_3
v_{23} v_{19} v_{15} v_{11} v_7
u_4
u_2
u_5
u_1 u_8 u_9 u_{10} u_{11}
u_6
u_3
u_7

Figure 4: Trees T_1 (left) and T_2 (right).

Now, we show that no central vertex of T_2 is in some optimal amalgam of T_1 and T_2. Note, that u_8 is the only central vertex of T_2. Suppose that there exist an amalgam A' such that $h_{A'}(T_1^{A'}, T_2^{A'}) = 4$ and $u_8^{A'} \in V(T_1^{A'} \cap T_2^{A'})$. From the same reason as above the set of vertices $\{v_1^{A'}, v_2^{A'}, \ldots, v_7^{A'}\}$ is a subset of $V(T_1^{A'} \cap T_2^{A'})$. Since the subgraph of T_2 induced on the set of black vertices in the Figure 4 is the only subgraph of T_2 which is isomophic to the subgraph of T_1 induced on the set of (black) vertices $\{v_1, v_2, \ldots, v_7\}$, and it does not contain u_8, it follows that no such amalgam A' exists.

Proposition 4.6. *Let T_1 and T_2 be two arbitrary non-trivial trees, with* $\operatorname{diam}(T_1) \geq \operatorname{diam}(T_2)$. *Let $A \in \mathcal{X}(T_1, T_2)$ be an optimal amalgam of T_1 and T_2. Then there exist $c_1 \in \operatorname{center}(T_1)$ and $c_2 \in \operatorname{center}(T_2)$ such that $d_A(c_1^A, c_2^A) \leq \mathcal{H}(T_1, T_2)$.*

Proof. Choose vertices $c_1 \in \operatorname{center}(T_1)$ and $c_2 \in \operatorname{center}(T_2)$ such that distance $d_A(c_1^A, c_2^A)$ is the smallest possible. Choose the vertex u for which it holds that $\operatorname{rad}(T_1) = d_A(c_1^A, u) \leq d_A(c_2^A, u)$. Such a vertex $u \in V(T_1^A)$ exists because $c_1 \in \operatorname{center}(T_1)$. Note, if T_1 is bicentral, the second central vertex is on the shortest path between c_1 and u. Choose a vertex v for which it holds that $v \in V(T_1^A \cap T_2^A)$ and $d_A(u, v)$ is the smallest possible. Then

$$\begin{aligned}
\mathcal{H}(T_1, T_2) \geq& \\
d_A(u, v) =& \\
d_A(c_1^A, u) - d_A(c_1^A, v) =& \\
d_A(c_1^A, u) - \big(d_A(c_2^A, v) - d_A(c_1^A, c_2^A)\big) \geq& \\
\operatorname{rad}(T_1) - \big(\operatorname{rad}(T_2) - d_A(c_1^A, c_2^A)\big) =& \\
d_A(c_1^A, c_2^A) + (\operatorname{rad}(T_1) - \operatorname{rad}(T_2)) \geq& \quad d_A(c_1^A, c_2^A).
\end{aligned}$$

■

The following proposition shows that the bound from Proposition 4.6 is sharp.

Proposition 4.7. *For an arbitrary non-negative integer k there exist trees T_1 and T_2, with* $\mathrm{diam}(T_1) \geq \mathrm{diam}(T_2)$ *and* $\mathcal{H}(T_1, T_2) = k$, *such that for every optimal amalgam A of T_1 and T_2 it holds that* $d_A(c_1^A, c_2^A) = \mathcal{H}(T_1, T_2)$, *where* $c_1 \in \mathrm{center}(T_1)$ *and* $c_2 \in \mathrm{center}(T_2)$.

Proof. Let k be a fixed non-negative integer. We will construct two non-isomorphic trees T_1 and T_2, such that the Hausdorff distance between T_1 and T_2 is $\mathcal{H}(T_1, T_2) = k$ and the distance between the vertices c_1^A and c_2^A corresponding to the centers of T_1 and T_2 in every optimal convex amalgam A is $d_A(c_1^A, c_2^A) = k$.

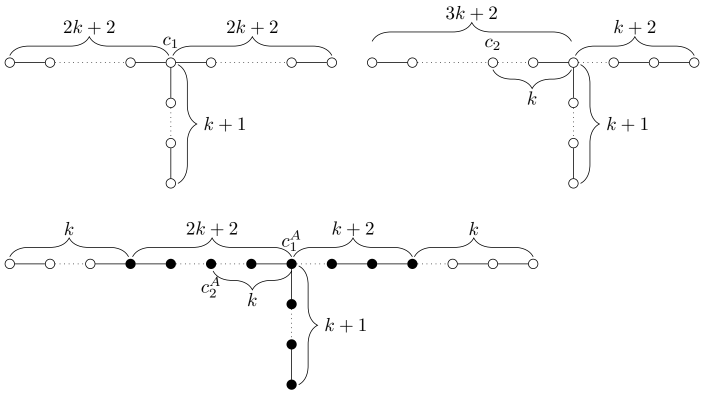

Figure 5: Trees T_1, T_2 and an optimal amalgam A of T_1 and T_2.

Let T_1 be the tree constructed from a path of length $4k+4$ and a path of length $k+1$, where we identify one end-vertex of the shorter path with the central vertex of the longer path; see the top left-hand tree in Figure 5 for reference. So T_1 is a star-like tree with three rays, two of length $2k+2$ and one of length $k+1$. Clearly, c_1 is the only central vertex of T_1.

Next, let T_2 be the tree constructed from a path of length $4k+4$ and a path of length $k+1$, where we identify one end-vertex of the shorter path with a vertex at distance k from the central vertex of the longer path; see the top right-hand tree in Figure 5 for reference. By construction, also T_2 is a star-like tree with three rays, one of length $3k+2$, one of length $k+2$ and one of length $k+1$, with exactly one central vertex, namely c_2.

Now, we construct an amalgam A of T_1 and T_2 as shown in the bottom tree in Figure 5. Clearly, A is a convex amalgam of T_1 and T_2, the distance between vertices corresponding to the centers of the trees is $d_A(c_1^A, c_2^A) = k$. From the construction and Corollary 2.7 it is also obvious that $\mathcal{H}(T_1, T_2) \leq h_A(T_1^A, T_2^A) = k$. Using Theorem 4.4, it can be easily checked that $\mathcal{H}(T_1, T_2) > k - 1$.

All that is left, is to show that in every optimal amalgam of trees T_1 and T_2 the distance between vertices corresponding to the central vertices of covers is k. Let A be an arbitrary optimal amalgam of T_1 and T_2. Note, $\mathrm{diam}(T_1) = \mathrm{diam}(T_2) = 4k + 4$. By Theorem 4.3 to follows that $c_1^A \in V(T_1^A \cap T_2^A)$. Moreover, we claim that the vertices corresponding to all neighbours of c_1 are also in $T_1^A \cap T_2^A$. Towards contradiction, let $v \in V(T_1)$ be a neighbour of c_1 such that $v^A \notin V(T_1^A \cap T_2^A)$. Also, let w denote the leaf of T_1 such that the path $P_{v,w}$ from v to w does not contain c_1. Since $T_1^A \cap T_2^A$ is convex (and therefore connected) no vertex of $P_{v,w}$ can be in $T_1^A \cap T_2^A$. But then $d_A(w^A, T_1^A \cap T_2^A) = k+1 > k$, a contradiction with Theorem 2.6 and the fact that $\mathcal{H}(T_1, T_2) = k$. It follows that c_1 and all its three neighbours are in $T_1^A \cap T_2^A$. Since T_2 contains exactly one vertex, say u, of degree three and A is a convex amalgam of T_1 and T_2, this vertex and its neighbours must also be in $T_1^A \cap T_2^A$, moreover c_1 is mapped with an isomorphism to u. Since A was arbitrarily chosen, the distance between the vertices c_1^A and c_2^A is the same in all optimal amalgams. Clearly, $d_A(c_1^A, c_2^A) = k$. ■

4.2 Algorithm

In this section we present an exact (exponential time) algorithm for computing the Hausdorff distance between two trees. As the following example demonstrates, using a maximum common subtree of two arbitrary trees to form a convex amalgam may not always produce an optimal amalgam for the Hausdorff distance.

Example 4.8. In Figure 6 we have two non-isomorphic trees T_1 (left hand side) and T_2 (right hand side) with central vertices c_1 and c_2, respectively. A maximum common subtree of T_1 and T_2 is clearly isomorphic to T_2.

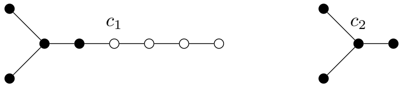

Figure 6: Maximum common subtree does not suffice.

Let A_1 be a convex amalgam obtained from T_1 and T_2 by identifying the subgraphs induced by the sets of black vertices (using maximum common subtree). In this case

$h_{A_1}(T_1^{A_1}, T_2^{A_1}) = 4$. On the other hand, one can form a convex amalgam A_2 by identifying the central vertices of the two trees for which $h_{A_2}(T_1^{A_2}, T_2^{A_2}) = 3$ and therefore $\mathcal{H}(T_1, T_2) \leq 3$. Therefore A_1 is not an optimal amalgam. It can be easily verified that there is no optimal convex amalgam of T_1 and T_2 that would be obtained by identifying T_2 with any other subgraph of T_1 isomorphic to T_2.

Now, we present three algorithms we use to compute the Hausdorff distance between two arbitrary trees. The first two are used as a subroutine of the Algorithm 3, which returns the actual Hausdorff distance and an optimal amalgam.

Algorithm 1: ConnectedSubgraphsRooted

input : An arbitrary rooted tree T and result passed by reference.
output: An array of lists of connected subgraphs of input tree T which include root vertex, saved in result.

```
ts ← firstSubgraph(T)
while ts ≠ null do
    ts ← nextSubgraph(T, ts)
    if ts = null then
        break
    end
    ti ← null
    foreach subgraph s in result[ts.size()] do
        if Isomorphism(s, ts) then
            ti ← s
            break
        end
    end
    if ti ≠ null and ts.dist < ti.dist then
        Remove ti from result[ts.size()]
        Insert ts in result[ts.size()]
    else if ti = null then
        Insert ts in result[ts.size()]
    end
end
```

Given a rooted tree as the input, Algorithm 1 computes those connected subgraphs (called r-subgraphs) of the input tree which include the root vertex. If two or more isomorphic r-subgraphs are found then, by Proposition 2.11, it suffices to save in the result only one of them, namely the one that has the smallest distance to the input tree. The result of the algorithm is an array of lists of subgraphs, where the list at index i contains all pairwise non-isomorphic connected r-subgraphs on i vertices.

The function `firstSubgraph()` returns the (r-)subgraph induced on the root vertex. The function `nextSubgraph()` returns the next r-subgraph which has not been created yet. Details about this procedure are described in [12]. In this procedure we also compute the distance of the created subgraph to the original tree. The order in which this procedure generates r-subgraphs is not important in our algorithm.

When we get the result from `nextSubgraph()` (one r-subgraph) we check if there already exists an isomorphic r-subgraph on i vertices in the array of saved r-subgraphs. If no such r-subgraph is found then we save our r-subgraph in the array. Otherwise, due to Proposition 2.11, we keep in the array only the one of the two r-subgraphs that has the smallest distance to the input tree.

The time complexity of the outer while loop is proportional to number of all r-subgraphs, that is $O(2^{n-1})$ in the worst case. In line 3 we compute one r-subgraph. The most time consuming in this procedure is computing the distance of the subgraph to original tree which gives the routine the time complexity $O(n^2)$. Lines 8-13 have the time complexity $O(n \cdot 2^{n-1})$, where $O(n)$ is the complexity of the method `Isomorphism()` and $O(2^{n-1})$ (number of all r-subgraphs) is the upper bound for the number of r-subgraphs on fixed number of vertices. Lines 14-19 have a constant time complexity. Therefore, the total time complexity for Algorithm 1 is $O(2^{n-1} \cdot n \cdot 2^{n-1}) = O(n \cdot 4^n)$.

Algorithm 2: ConnectedSubgraphs

```
input  : An arbitrary tree T and result passed by reference.
output: An array of lists of connected subgraphs of input tree T, saved in result.
foreach vertex $v \in V(T)$ do
    Root tree T in v
    ConnectedSubgraphsRooted(T, result)
end
```

Algorithm 2 is similar to Algorithm 1. It computes all connected pairwise non-isomorphic subgraphs of the input tree, whilst Algorithm 1 computes only such r-subgraphs. Note, from all pairwise isomorphic subgraphs we only keep the one with the smallest distance to the input tree. Because of similarity, we can use Algorithm 1 inside Algorithm 2. We root the input tree in each vertex and call Algorithm 1 to add the current r-subgraphs to the result. The time complexity of Algorithm 2 is $O(n^2 \cdot 4^n)$.

However, we can slightly improve the complexity by changing the function `nextSubgraph()`. When rooting the tree in two different vertices and computing their r-subgraphs, the same subgraph of the input tree can appear in the result of both rooted trees. This means that we iterate through the same subgraphs multiple times. This can be avoided by adding a condition that prevents computing r-subgraphs which include a vertex that has already been a root in some previous iteration. Adding this condition does not worsen the time complexity of the function `nextSubgraph()`. With this improvement, the time complexity of Algorithm 2 is $O(n \cdot 4^n)$.

Algorithm 3: Hausdorff distance

```
input  : Two arbitrary trees T1 and T2, with diam (T1) $\geq$ diam (T2)
output: Isomorphic subgraphs r1 and r2 of T1 and T2 which give rise to
        $T1^A \cap T2^A$, where $A$ is an optimal amalgam and Hausdorff distance hd
1  Calculate the set center(T1) and then root the tree T1 from an arbitrary
   $c \in$ center(T1)
2  sub1 $\leftarrow$ ConnectedSubgraphsRooted (T1)
3  sub2 $\leftarrow$ ConnectedSubgraphs (T2)
4  hd $\leftarrow$ rad (T1)
5  for i $\leftarrow$ 1 to min(sub1.size(), sub2.size()) do
6      foreach element s1 in sub1[i] do
7          foreach element s2 in sub2[i] do
8              if Isomorphism(s1,s2) and max(s1.dist,s2.dist) $<$ hd then
9                  hd $\leftarrow$ max(s1.dist,s2.dist)
10                 r1 $\leftarrow$ s1
11                 r2 $\leftarrow$ s2
12             end
13         end
14     end
15 end
```

Algorithm 3 computes the Hausdorff distance of two input trees T1 and T2. It returns the distance and two subgraphs of the input trees which give rise to an optimal amalgam.

The first part of the algorithm computes all suitable r-subgraphs (for T1, since by Theorem 4.3 we know that a central vertex of T1 is in the intersection of the amalgam) and subgraphs of the input trees using Algorithm 1 and Algorithm 2, respectively. Then the algorithm iterates (lines 5-15) through all pairs of subgraphs of the same order and looks for isomorphisms between them. If we find an isomorphism between two subgraphs we can construct amalgam A of the input trees with respect to this isomorphism. Using Corollary 2.7, the distance $h_A(\mathsf{T1}^A, \mathsf{T2}^A)$ for such an amalgam equals the maximum of the distances of the subgraphs to their corresponding input tree. Since we check this for all possible pairs of subgraphs, the algorithm computes the Hausdorff distance (and the optimal amalgam).

The time complexity for lines 1-4 is $O(n \cdot 4^n)$. Time complexity for lines 5-15 is also $O(n \cdot 4^n)$, because in the worst case we check all possible pairs of subgraphs and there are at most 2^{n-1} subgraphs in `sub1` and at most 2^n subgraphs in `sub2`. For every pair of subgraphs we check if they are isomorphic which takes $O(n)$ time per pair, other operations inside the loops can be done in constant time. Therefore the total time complexity for Algorithm 3 is $O(n \cdot 4^n)$.

5 Open problems

We conclude the paper with two open problems. It is clear from the complexity analysis of Algorithm 3 that it is polynomial for trees with a polynomial number of connected subgraphs, and exponential in the general case. So the next question arises naturally.

Problem 5.1. Is there a polynomial algorithm that determines the Haudsorff distance between two arbitrary trees?

In chemical graphs the vertices present atoms and edges present bonds. So when determining the similarity of two (chemical) graphs it would make sense to restrict which vertices can map to each other when making an amalgam.

Problem 5.2. Define a measure of similarity of two graphs based on the Hausdorff distance for labelled graphs with an additional restriction to what labels are allowed to map to each other.

References

[1] A. V. Aho, J. E. Hopcroft, J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison–Wesley, Boston, 1974.

[2] H.-J. Bandelt, H. M. Mulder, E. Wilkeit, Quasi–median graphs and algebras, *J. Graph Theory* **18** (1994) 681–703.

[3] I. Banič, A. Taranenko, Measuring closeness of graphs – the Hausdorff distance, *Bull. Malays. Math. Sci. Soc.*, in press.

[4] G. Benadé, W. Goddard, T. A. McKee, P. A. Winter, On distances between isomorphism classes of graphs, *Math. Bohem.* **116** (1991) 160–169.

[5] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, *Patt. Recog. Lett.* **19** (1998) 255–259.

[6] G. Chartrand, F. Saba, H. B. Zou, Edge rotations and distance between graphs, *Čas. Pěst. Mat.* **110** (1985) 87–91.

[7] S. A. Cook, The complexity of theorem–proving procedures, M. A. Harrison, R. B. Banerji, J. D. Ullman (Eds.), *STOC '71 Proceedings of the Third Annual ACM Symposium on Theory of Computing*, ACM, New York, 1971, pp. 151–158.

[8] G. M. Downs, P. Willett, Similarity searching in databases of chemical structures, in: K. B. Lipkowitz, D. B. Boyd (Eds.), *Reviews in Computational Chemistry*, Wiley, Hoboken, 2007, pp. 1–66.

[9] M. Johnson, An ordering of some metrics defined on the space of graphs, *Czech. Math. J.* **37** (1987) 75–85.

[10] S. Klavžar, Wiener index under gated amalgamations, *MATCH Commun. Math. Comput. Chem.* **53** (2005) 181–194.

[11] J. W. Raymond, P. Willett, Maximum common subgraph isomorphism algorithms for the matching of chemical structures, *J. Comput. Aid. Mol. Des.* **16** (2002) 521–533.

[12] F. Ruskey, Listing and counting subtrees of a tree, *SIAM J. Comput.* **10** (1981) 141–150.

[13] G. Valiente, *Algorithms on Trees and Graphs*, Springer, Berlin, 2010.

[14] P. Willett, Matching of chemical and biological structures using subgraph and maximal common subgraph isomorphism algorithms, in: D. G. Truhlar, W.J. Howe, A. J. Hopfinger, J. Blaney, R. A. Dammkoehler (Eds.), *Rational Drug Design*, Springer, New York, 1999, pp. 11–38.