

Hierarchical alignment of weighted directed acyclic graphs

Sean M. Falconer
University of Victoria
Victoria, BC, V8W 2Y2, Canada
seanf@uvic.ca

Dmitri Maslov
University of Waterloo
Waterloo, ON, N2L 3G1, Canada
dmitri.maslov@gmail.com

Abstract—In some applications of matching, the structural or hierarchical properties of the two graphs being aligned must be maintained. The hierarchical properties are induced by the direction of the edges in the two directed graphs. These structural relationships defined by the hierarchy in the graphs act as a constraint on the alignment. In this paper, we formalize this problem as the weighted alignment between two directed acyclic graphs. We prove that this problem is NP-complete, prove several upper bounds for approximating the solution, and finally introduce algorithms for several sub-classes of directed acyclic graphs.

I. THE PROBLEM

Matching or alignment problems are an important set of theoretical problems that appear in many different applications [3], [4], [8]. Depending on the structure of the problem, polynomial time algorithms may or may not exist. In this paper, we propose a new type matching problem called the *DAG (directed acyclic graph) alignment problem*. In this problem, we have two directed acyclic graphs and a set of possible matchings between vertices in both graphs. We wish to find the maximum weighted matching between the vertices in both graphs where the directed edges in both graphs act as *hierarchical constraints* on possible solutions to the matching. For example, if a vertex v_1 has a directed edge to a vertex v_2 , then any matched vertex to v_2 cannot be an ancestor of v_1 's matched vertex (see Figures 1 and 2).

We became interested in this problem through our interest in ontology alignment. An ontology is a conceptualization of a domain [11]. This conceptualization consists of a set of terms with certain semantics and relationships [23]. Generally, the terms are related by the *is_a* relationship. The relationships (edges) and terms (vertices) can be represented as a DAG. With ontology alignment, one wants to align terms from two different ontologies in order to merge, compare, or map the ontologies. Since the edges of the DAG represent an *is_a* relationship, this constrains the number of valid matchings, because we do not wish to violate this relationship in the matching.

This type of hierarchical or structural constraint is important in other applications as well. The domains of SVG (*Scalable Vector Graphics*) version comparison, source code comparison/merging, UML difference calculation, and file/folder merging, are all instances of hierarchical based matching. For example, an SVG document is rich with structure. The

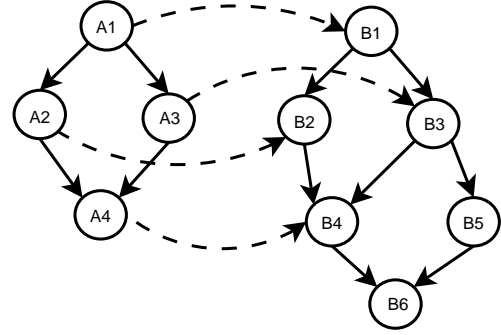


Fig. 1. Example of a valid matching between two graphs. The dashed lines represent valid assignments for the vertices A_1 , A_2 , A_3 , and A_4 .

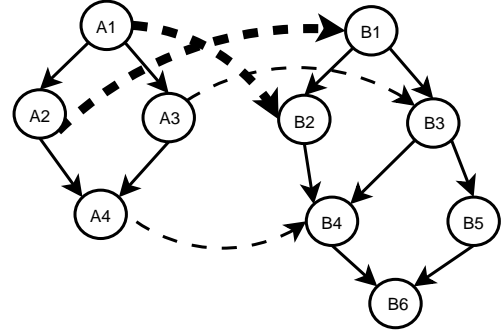


Fig. 2. Example of an invalid matching between two graphs. The dashed lines represent the assignments for the vertices A_1 , A_2 , A_3 , and A_4 . The two bold dashed lines represent an assignment violation because A_1 maps to a descendant of A_2 's mapped vertex B_1 .

document defines graphical objects, and how they relate, a form of the *is_a* relationship exists through the document graphic layers. In object-oriented programming, *is_a* relationships exist through the definitions of inheritance, and other relationships exist via class membership. Similarly, UML diagrams have structural relationships, and different versions of diagrams sometimes need to be merged or have their differences calculated for visual comparison [19]. Finally, in a file system, the folders represent an embedded hierarchy.

A. Related work

General graph matching is a well studied problem. Most graph matching problems can be divided into two categories, graph isomorphisms and weighted graph matching. In graph

isomorphism, the goal is to find a matching function f for two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. General graph isomorphism is still open, that is, it is not known whether the problem is NP-hard or can be solved in polynomial time [9]. Sub-graph isomorphism is known to be NP-complete [10]. With weighted graph matching, we are given a graph $G = (V, E)$, where the edges have associated weights and we wish to find a subset M of E , such that no two edges in M share a common end vertex and such that the sum of edge weights in M is maximum. For some classes of graphs, polynomial time algorithms are known, while some others are known to be NP-complete.

Both of these problems have many practical applications, in particular, graph isomorphism has received a lot of attention in the area of computer vision. Images or objects can be represented as a graph. A weighted graph can be used to formulate a structural description of an object [24]. There have been two main approaches to solving graph isomorphism: state-space construction with searching and nonlinear optimization. The first method consists of building the state-space, which can then be searched. This method has an exponential running time in the worst case scenario, but by employing heuristics, the search can be reduced to a low-order polynomial for many types of graphs [5], [25]. With the second approach (nonlinear optimization), the most successful approaches have been relaxation labeling [15], neural networks [18], linear programming [1], eigendecomposition [26], genetic algorithms [16], and Lagrangian relaxation [22].

As mentioned, we became interested in DAG alignment problem due to our interests in ontology alignment. Ontology alignment has recently received a lot of attention. An alignment between two ontologies can be formalized in terms of weighted graph matching, with certain constraints on the solution to any valid matching. Originally, alignments were performed by hand, and later, several researchers introduced semi-automatic alignment strategies, which make suggestions to the user about which terms to align [20], [21]. Since then, fully automatic alignment strategies have been explored. In [6], over twenty different tools/algorithms are discussed. Many of these approaches use heuristics to determine term similarities, by first comparing syntactic, semantic, and structural similarities, and then compute matches greedily or via some other local optimization technique.

In [7], graph matching is applied to conceptual system matching for translation. The work is very similar to ontology alignment, however, the authors formalize their problem in terms of any conceptual system rather than restricting the work specifically to an ontological formalization of a domain. They formalize conceptual systems as graphs, and introduce algorithms for matching both unweighted and weighted versions of these graphs.

B. Organization of the paper

The remainder of the paper is organized as follows. The next section introduces notations and definitions that will be used throughout the paper. The definitions include the formal description of the problem. Following this, we show that

the decision version of the problem is NP-complete via a reduction from 3SAT. Next, we prove two theorems, which yield upper bounds on approximating the DAG alignment problem. After this, we consider the alignment of different classes of graphs, such as chains, stars, and trees. We introduce polynomial time algorithms for the first two classes and a polynomial time algorithm for the trees with finite branching factor and logarithmic depth. We finish presentation with concluding remarks, a short discussion of open problems, and directions for future research.

II. NOTATIONS AND DEFINITIONS

A. Notations

Before formally defining the DAG alignment problem we must first introduce some objects. A DAG is a directed graph, $G = (V, E)$ that contains no oriented cycles, where V is a set of vertices and E is a set of edges. Let $anc(v)$ denote the set of ancestors for any $v \in V$, where an ancestor of v is any $a \in V$ such that there exists a directed path from a to v . Let $desc(v)$ denote the set of descendants for any $v \in V$, where a descendant of v is any $d \in V$ such that there exists a directed path from v to d .

B. Description of problem

In this section we formalize the problem of DAG alignment with hierarchy constraints. Without the hierarchy constraint, the problem reduces to weighted bipartite matching, since the edges that represent vertex relationships would be ignored. As was mentioned, in many practical applications these structural relationships cannot be ignored. Due to these relationships, many solutions that would be valid in weighted bipartite matching are invalid. In fact, we can think of any edge e as having a set of conflicting edges, where a conflict is any edge that would violate a matching solution that contained e . We formalize this in the following definition.

Definition 1 An **edge conflict** for edge $e = (a, b, w_e)$, $w_e \in [0, 1]$, is any edge $d = (f, g, w_d)$, $w_d \in [0, 1]$, and $d \neq e$, where one of the following conditions applies:

- 1) $a \in anc(f)$ and $b \notin anc(g)$.
- 2) $a \in desc(f)$ and $b \notin desc(g)$.
- 3) $a = f$.
- 4) $b = g$.

The set $conf(e)$ denotes the set of edges that have edge conflicts with edge e . We can now introduce the formal definition of the DAG alignment problem.

Definition 2 Given two DAGs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and a set of edges $\beta = \{(v_i, v_j, w_t)\}$ for all $v_i \in V_1$, all $v_j \in V_2$ and $w_t \in [0, 1]$, the **DAG alignment problem** is to find the maximum weight matching, $M \subseteq \beta$, such that each vertex in M appears only once and for any edge $e \in M$, $conf(e) \cap M = \emptyset$. We refer to this constraint on the matching as the *hierarchical constraint* for the remainder of this paper.

Our definition of the DAG alignment problem uses a complete bipartite graph of all possible matchings with the set of edges $\beta = \{(v_i, v_j, w_t)\}$ defined for all $v_i \in V_1$ and all

$v_j \in V_2$. This may appear to narrow the set of problems we are trying to solve, however, it does not. This is because a solution to the problem with an incomplete (some matchings may be inherently prohibitive) matching graph can be reduced to the problem with complete bipartite graph through the following consideration. Take a DAG alignment problem in which not every node of G_1 can potentially be mapped to any node of G_2 . Allow all the remaining matchings, but assign zero weights to them. Solve the DAG alignment problem with the complete set of possible matchings. Delete all zero weight matchings from the solution. The result is a solution for the DAG alignment problem with incomplete set of possible matchings.

III. INTRACTABILITY

The DAG alignment problem defined in the previous section is NP-complete. Before showing the proof of this, we begin by first defining the decision version of the problem.

Definition 3 We are given two DAGs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and a set of edges $\beta = \{(v_i, v_j, w_t)\}$ for all $v_i \in V_1$, $v_j \in V_2$ and $w_t \in [0, 1]$. Let $w(A)$, where $A \subseteq \beta$, be the sum of all weights w_t defined over all triples $(v_i, v_j, w_t) \in A$. Is there a matching $M \subseteq \beta$ with weight $w(M) \leq X$ and $|M| \leq Y$ such that each vertex in M appears only once and for any edge $e \in M$, $\text{conf}(e) \cap M = \emptyset$?

Theorem 1 DAG alignment, as introduced in Definition 3, is NP-complete.

Proof: It is easy to see that the decision version of DAG alignment is in NP, so this will be omitted.

We show a reduction of 3SAT to the decision version of the DAG alignment problem. In 3SAT we have a finite set of variables, $X = \{x_1, x_2, \dots, x_n\}$ and a finite set of clauses $C = \{c_1, c_2, \dots, c_m\}$, such that each clause is logic OR of 3 literals, where the literals over variable x_i are x_i^0 ($:= x_i$) and x_i^1 ($:= \bar{x}_i$). The problem is to find a truth assignment to variables in X such that the logic AND of all clauses in C is satisfied.

Let $\phi = (X, C)$ be an instance of 3SAT. We can define an instance of the DAG alignment problem as follows. We begin by defining the two DAGs used in the alignment. First, let us define $G_1 = (V_1, E_1)$ where V_1 is defined as follows.

$$V_1 = \bigcup_{c_i \in C} (x_j^{p1}, i) \cup (x_k^{p2}, i) \cup (x_l^{p3}, i), \text{ where}$$

$$c_i = (x_j^{p1}, x_k^{p2}, x_l^{p3}) \text{ and } p1, p2, p3 \in \{0, 1\} \text{ and } j, k, l \leq n.$$

We define the set of edges E_1 by creating directed edges over the vertices of V_1 as $((x_j^0, i), (x_j^1, t))$ for all $j \leq n$ and $i, t \leq m$.

Now, let us define a second DAG, $G_2 = (V_2, E_2)$. First, we define V_2 as

$$V_2 = \{\{y_1, z_1, y_2, z_2, \dots, y_n, z_n\} \times \{1, 2, \dots, m\}\} \\ \bigcup \{\{c_1, c_2, \dots, c_m\} \times \{1, 2\}\}.$$

Intuition behind this definition is y_i corresponds to x_i and z_i corresponds to \bar{x}_i .

We define E_2 by creating directed edges $((z_j, i), (y_j, t))$, $((y_j, t), (c_i, 1))$ and $((y_j, t), (c_i, 2))$ for all $j \leq n$ and $i, t \leq m$.

We now have two DAGs, G_1 and G_2 . We must define the set β , which describes the possible matches between the two DAGs, and the related weights. For every vertex, (x_j^0, i) or (x_j^1, i) , map this vertex to its corresponding vertex in V_2 with weight equal to one and add this to β . That is, $(x_j^0, i) \in V_1$ maps to $(y_j, i) \in V_2$ and $(x_k^1, t) \in V_1$ maps to $(z_k, t) \in V_2$, and so forth. Also, for each vertex $(x_j^p, i) \in V_1$, create mappings $((x_j^p, i), (c_i, 1))$ and $((x_j^p, i), (c_i, 2))$ both with weight equal to one and add this to β . Let the total weight and the total number of vertices for the matching be $3m$.

We now show that the DAG alignment problem, as described above, has a matching satisfying the hierarchical mapping constraint, if and only if ϕ is satisfiable.

(\Rightarrow) Assume ϕ is satisfiable. For each clause c_i , choose a single literal x_j^p . If variable $x_j \in X$ is true and $p = 0$ or $x_j \in X$ is false and $p = 1$, then include edge $((x_j^0, i), (y_j, i))$ in the matching M . Also, for any clause c_t with x_j^1 include edge $((x_j^1, t), (c_t, 1))$ if vertex $(c_t, 1)$ is not in the matching, otherwise include edge $((x_j^1, t), (c_t, 2))$. Similarly, if variable $x_j \in X$ is true and $p = 1$ or $x_j \in X$ is false and $p = 0$, then include $((x_j^1, i), (z_j, i))$ in the matching. Also, for any clause c_t with x_j^0 , include edge $((x_j^0, t), (c_t, 1))$ if vertex $(c_t, 1)$ is not in the matching, otherwise include edge $((x_j^0, t), (c_t, 2))$. Thus, M exactly maps all vertices in G_1 to vertices in G_2 . There are $3m$ vertices in V_1 , so $|M| = 3m$. Also, since the weight of each edge is one, $w(M) = 3m$. Finally, since both x_j^0 and x_j^1 cannot be true, both edges $((x_j^0, i), (y_j, i))$ and $((x_j^1, i), (z_j, i))$ cannot be in M , therefore the hierarchical constraint is satisfied.

(\Leftarrow) Let M be a solution to the DAG alignment problem. The truth value of any variable x_j is assigned as follows. If, for any clause c_i with literal x_j^0 , there exists an edge $((x_j^0, i), (y_j, i))$ from G_1 to G_2 , then let x_j be true. Similarly, if there exists an edge $((x_j^1, i), (z_j, i))$ from G_1 to G_2 , then let x_j be false. Since in G_1 , every vertex (x_j^0, i) has an edge to every (x_j^1, t) , and in G_2 every vertex (z_j, i) has an edge to every (y_j, t) , M cannot contain edges $((x_j^0, i), (y_j, i))$ and $((x_j^1, i), (z_j, i))$, otherwise the hierarchical constraint would be violated. Thus, x_j^0 or x_j^1 is true, but never both. Also, since any false literal in a clause c_i is mapped to a vertex $(c_i, 1)$ or $(c_i, 2)$, at most 2 vertices in any clause can be false. Thus, ϕ is satisfied. ■

IV. UPPER BOUNDS ON APPROXIMATING WEIGHTED DAG ALIGNMENT

Since weighted DAG alignment belongs to the class of NP-complete problems, it is unlikely that we will find a polynomial time solution to the problem. Thus, we must rely on an approximation scheme for computing alignments.

In this section, we introduce two polynomial time reductions of the DAG alignment problem to other known NP-complete problems and use these to provide upper bounds for approximating the weighted DAG alignment problem. The quality of the approximation is given as the ratio between the size of the maximum weighted DAG alignment and the

approximation found. The ratio in the worst-case scenario defines the *performance guarantee* of the algorithm.

We begin by reducing the DAG alignment problem to Weighted Independent Set (WIS). In the Independent Set problem, we are given a graph $G = (V, E)$, and we wish to find the largest subset $S \subseteq V$, such that no two vertices in S are connected by an edge in E . In the weighted version of this problem, each node, $v_i \in V$, has an associated weight w_i , and we wish to find the maximum weighted independent set.

Håstad [12] showed that Independent Set is hard to approximate within $n^{1-\epsilon}$, for $\epsilon > 0$, unless NP-hard problems have randomized polynomial time solutions. In [2], Boppana and Halldórsson introduced the Ramsey algorithm for solving WIS. The algorithm is an extension of the naive greedy approach, where in the greedy approach a vertex v is arbitrarily selected from the graph and added to the independent set, all adjacent vertices are removed, and this process is continued until all vertices are exhausted. The obvious problem with this solution is that the adjacencies are ignored. The first extension to this process is to consider not only the vertex v , but also the neighbors of v . The algorithm recurses by first considering v as part of the independent set, and then v not in the independent set, and selecting the better of the two results. This algorithm performs well provided the maximum *Clique* size is small. Boppana and Halldórsson further extended this algorithm by first removing the maximum set of disjoint k -cliques, and then apply the Ramsey algorithm to compute the independent set on this modified graph. From this, they were able to prove that the algorithm had a performance guarantee of $O(n/\log^2 n)$, where n is the number of vertices in the graph.

The following shows that any instance of the DAG alignment problem can be reduced, in polynomial time, to an instance of WIS. This reduction will allow us to use approximation strategies for Independent Set to find approximate solutions to the DAG alignment problem.

Theorem 2 The ontology alignment problem can be approximated within $O(m/\log^2 m)$ where $m = |\beta|$.

Proof: Consider an instance of the DAG alignment problem, defined by graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and the set of edges β . We define an instance of WIS, by constructing a graph $G = (V, E)$ as follows. For each edge $e = (a, b, w_e) \in \beta$, construct a corresponding $v \in V$, and let the weight of vertex v be $w := w_e$. Next, let $E = \{(v_i, v_j) | e_j \in \text{conf}(e_i) \text{ and } e_i, e_j \in \beta\}$.

Now, we claim that a solution to WIS, defined over graph G , corresponds to a solution to the DAG alignment problem. We construct this solution as follows. Let S be our solution to WIS. Then, for each $v_i \in S$, add the edge from β that corresponds to v_i , to our DAG alignment solution M . This precisely constructs a valid DAG alignment, since each $v_i \in S$ cannot be connected to any other $v_j \in S$, which implies that for edges $e_i, e_j \in M$, $e_i \notin \text{conf}(e_j)$. Since no edges in M conflict, this must be a valid solution.

WIS can be approximated within $O(n/\log^2 n)$, where n is the number of vertices in the graph. In our reduction, n corresponds to $|\beta|$, by letting $m = |\beta|$, we achieve an approximation of $O(m/\log^2 m)$. ■

Next, we improve this bound via a reduction to the Weighted Set Packing (WSP) problem. In WSP, we have a set S of m base elements, and a collection $\mathcal{U} = \{U_1, U_2, \dots, U_n\}$ of weighted subsets of S . We want to find a subcollection $\mathcal{U}' \subseteq \mathcal{U}$ of disjoint sets of maximum total weight.

In [13], an approximation guarantee of \sqrt{m} , where $m = |S|$ is given for WSP. The algorithm is based on a variant of the greedy algorithm for solving the non-weighted version introduced in [14]. In the following theorem, we show that any instance of the DAG alignment problem can be reduced to WSP in polynomial time, and that a solution to WSP corresponds to a solution of the DAG alignment problem.

Theorem 3 The DAG alignment problem can be approximated within \sqrt{m} where $m = |\beta|$.

Proof: Consider an instance of the DAG alignment problem, defined by graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and the set of edges β . We define an instance of WSP, by constructing S and the collection \mathcal{U} as follows.

We let our m base elements be the edges specified by β , thus our set $S = \beta$. We construct the collection \mathcal{U} , by defining subsets U_i for all $e_i \in \beta$ as $U_i = \{\{e_i\} \cup \text{conf}(e_i)\}$. Let the weight of U_i be equal to w_{e_i} . We now claim that any solution to WSP, \mathcal{U}' , corresponds to a solution the DAG alignment problem.

We can see this by considering any \mathcal{U}' . We construct a solution to the DAG alignment problem by taking each $U_i \in \mathcal{U}'$, and adding edge $e_i \in \beta$ to our ontology alignment solution M . This is a valid matching because every $U_i \in \mathcal{U}'$ is disjoint, which implies that for each $e_i \in M$ and $e_j \in M$, $e_i \notin \text{conf}(e_j)$, so no edges in M conflict.

Since a solution to WSP yields a solution to the DAG alignment problem, approximations of WSP correspond to approximations of the DAG alignment problem. Hence, we can approximate the DAG alignment problem within \sqrt{m} , where $m = |\beta|$. ■

V. ALGORITHMS

In this section we study certain types/classes of graphs with respect to their DAG alignment problem solution complexity. In particular, we show that the DAG alignment problems for chains, stars and some trees has a polynomial time solution. In this work, *chain* C_n is defined as a DAG with n vertices v_1, v_2, \dots, v_n and directed edges $(v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_n)$. We define a *star* as a finite set of chains $\{C_{n_1}, C_{n_2}, \dots, C_{n_k}\}$ that all share vertex v_1 . Finally, a complete binary tree, T , is a rooted tree where each internal non-root vertex has one parent and two children, the edge directions are from a parent to a child, and all leaves are at the same level of the tree. The root of T is a vertex with two children and no parent.

A. Polynomial-time algorithms

We first prove that any two chains can be aligned in polynomial time and then use this algorithm as a base for the star alignment problem. However, we first introduce an

object called a *conflict graph*—a concept serving as a main tool in construction of our alignment algorithms.

Definition 4 For a given DAG alignment problem (as outlined in Definition 3), its **conflict graph** is a non-oriented graph $G = (V, E)$ with vertex set $V := \{v_t = (v_i^1, v_j^2) \mid v_i^1 \in V_1, v_j^2 \in V_2\}$ and vertex weights $weight(v_t) := w_t$, where $w_t \in [0, 1]$. An edge $e \in E$ between the vertices $v_{t_1} = (v_{i_1}^1, v_{j_1}^2)$ and $v_{t_2} = (v_{i_2}^1, v_{j_2}^2)$ exists if and only if edges $(v_{i_1}^1, v_{j_1}^2, w_1)$ and $(v_{i_2}^1, v_{j_2}^2, w_2)$ of the DAG alignment problem form an edge conflict.

Given a DAG alignment problem, its conflict graph can be constructed in polynomial time with respect to the size of DAG alignment problem. Solution to the Weight Independent Set (WIS) problem over the conflict graph defines a solution to the DAG alignment problem which induced the given conflict graph (as shown in Theorem 2). Thus, a polynomial time solution to the WIS problem over the conflict graph defines a polynomial time solution to the DAG alignment problem. We use this observation in the proof of the following theorem.

Theorem 4 DAG alignment problem for two chains C_n and C_k can be solved in polynomial time.

Proof: We first construct the conflict graph for such DAG alignment problem and discuss some of its properties. The conflict graph has $n \times k$ vertices denoted (i, j) for $1 \leq i \leq n$ and $1 \leq j \leq k$. Weight $w(i, j) \in [0, 1]$ assigned to the vertex (i, j) of conflict graph is the weight associated with mapping i^{th} element of C_n to the j^{th} element of C_k . There are three types of edges in the conflict graph. Edge $((i_1, j), (i_2, j))$, where $1 \leq i_1, i_2 \leq n$ and $1 \leq j \leq k$ belongs to the conflict graph and represents impossibility of mapping two different nodes of C_n into one node of C_k . Analogously, edge $((i, j_1), (i, j_2))$ for $1 \leq i \leq n$ and $1 \leq j_1, j_2 \leq k$ belongs to the conflict graph since it represents impossibility of mapping a node in C_n into two different nodes of C_k . The last type of edges, $((i_1, j_1), (i_2, j_2))$, where $(i_1 - i_2) \times (j_1 - j_2) < 0$, is a manifest of the hierarchical constraint that must be preserved by the DAG alignment.

We draw the conflict graph with the vertices arranged in a lattice, where each string contains vertices (i, j) for a fixed i and j running from 1 to k (left to right) and each column contains vertices (i, j) for a fixed j and i running from 1 to n (top to bottom). A conflict graph for parameters $n = 3$ and $k = 3$ is illustrated in Figure 3.

Observe that once vertex (i, j) was added to the solution to WIS problem in a conflict graph for chains, any vertex (i', j') in the I or III quadrant or along the $x - y$ axes and assuming (i, j) is a center of coordinate system cannot be a part of this WIS solution. This is illustrated in Figure 4, where (i, j) is vertex V , $n = 9$ and $k = 7$. Shaded vertices cannot belong to the WIS solution that includes vertex V . A vertex that can be a part of the WIS solution must be in the II or IV quadrant with respect to vertex (i, j) and cannot be along the $x - y$ axes. This observation holds for any two vertices chosen as a solution for WIS meaning a set of vertices V_1, V_2, \dots, V_s of the WIS solution in this type graphs forms a polyline with the segments V_t, V_{t+1} being defined by a linear equation $y = kx + b$ with

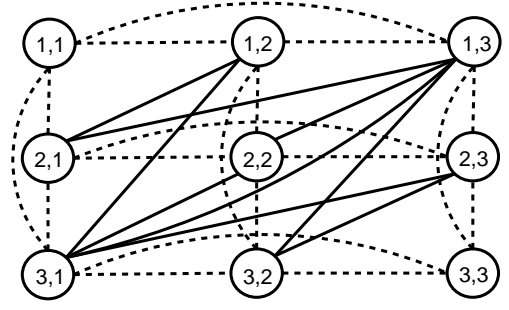


Fig. 3. Conflict graph for two chains with 3 vertices each. The (vertical) dashed lines represent edges $((i_1, j), (i_2, j))$, where $1 \leq i_1, i_2 \leq n$ and $1 \leq j \leq k$, and the (horizontal) dashed lines represent edges $((i, j_1), (i, j_2))$ for $1 \leq i \leq n$ and $1 \leq j_1, j_2 \leq k$. The regular lines stand for the edges $((i_1, j_1), (i_2, j_2))$, where $(i_1 - i_2) \times (j_1 - j_2) < 0$.

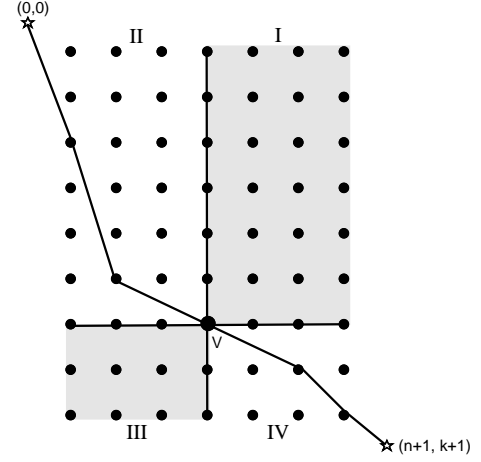


Fig. 4. Vertices eliminated from WIS solution to the conflict graph (shaded) for two chains once V is included into the solution, illustrated for parameters $n = 9$ and $k = 7$.

$k < 0$. Since, a solution to the WIS problem can be thought as a path from the artificially created vertex $(0, 0)$ with zero weight to the artificially created vertex $(n + 1, k + 1)$ with zero weight (Figure 4) with all segments having negative gradient.

The following dynamic programming algorithm solves the WIS problem for such conflict graphs. While running through all vertices (i, j) by increasing j while increasing i (or, symmetrically, increasing i while increasing j) update value $w(i, j)$ by adding $\max_{1 \leq s \leq i-1, 1 \leq t \leq j-1} w(s, t)$ to it. The weight value in node $(n + 1, k + 1)$ will give the answer to WIS problem and the vertices participating in the solution can be found through remembering the path that led to the $\max_{1 \leq s \leq i-1, 1 \leq t \leq j-1} w(s, t)$. Since all weights $w(i, j)$ are non-negative the algorithm can be improved through using formula $w(i, j) = \max\{w(0, j-1), w(1, j-1), \dots, w(i-1, j-1), w(i-1, j-2), \dots, w(i-1, 0)\}$. Thus, the total complexity of such algorithm is no more than the number of vertices visited (nk) times the maximal number of terms in the max operator, being equal to $(n + k - 3)$. Such algorithm is, thus, polynomial. ■

Theorem 5 DAG alignment problem for two stars S_1 and S_2 can be solved in polynomial time.

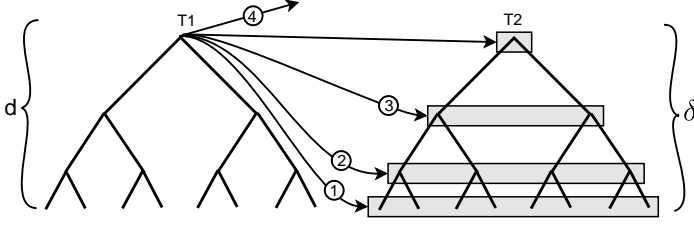


Fig. 5. Example of possible mappings for the root vertex of tree T_1 to the vertices of tree T_2 .

Proof: The proof of this theorem is based on the fact that the chain alignment problem has a polynomial time solution. When two stars are to be aligned, there are two cases to consider.

Case 1: the central vertex c_1 (vertex that has no parent) of a star S_1 is mapped into a non-central vertex v_2 of S_2 in the solution to graph alignment problem. Assuming the number of vertices in S_2 is k , there are $k - 1$ possible mappings. In each, we will need to align star S_1 with the chain starting with vertex v_2 . To find an optimal alignment it suffices to align each of the star's S_1 rays (chains starting with vertex c_1) to the chain starting with vertex v_2 . The number of rays in a star is never greater than the number of vertices in it, n . This means that finding the solution in case 1 is equivalent to aligning not more than nk chains of the sizes not more than n and not more than k correspondingly. This can be done in polynomial time.

Case 2: the central vertex c_1 of star S_1 is mapped into the central vertex c_2 of S_2 in the solution to the graph alignment problem. Assuming S_1 has r_1 rays ($r_1 < n$) and S_2 has r_2 rays ($r_2 < k$), create an $r_1 \times r_2$ matrix with the elements equal to the weights of aligning i^{th} ray of star S_1 with the j^{th} ray of star S_2 . The complexity of creating such matrix is no more than the complexity of aligning nk chains of sizes at most n and at most k correspondingly. This can be done in polynomial time. To find the maximum weight alignment of the two stars one needs to find the largest weight mapping of the set of matrix columns onto the set of matrix strings. This can be done in polynomial time [17]. ■

Lemma 1 The DAG alignment problem for two complete binary trees T_1 and T_2 can be solved with an $O(\frac{n^4}{\sqrt{\log n}})$ runtime algorithm.

Proof: Define $C(d, \delta)$ as the complexity for the DAG alignment problem for two complete binary trees of depths d and δ . In the solution to the DAG alignment problem, the root vertex of the first tree can be mapped into one of the 2^δ leaves (tree vertices at level δ) of the second tree (Figure 5, label ①). The best of such a mapping can be found with 2^δ simple weight comparison operations. The root vertex of tree T_1 can be mapped into one of the $2^{\delta-1}$ level- $(\delta-1)$ vertices of tree T_2 (Figure 5, label ②). In this case it suffices to align each of the two subtrees of the first tree with all possible leaves of the second tree. This calls for $2 * 2^\delta = 2^{\delta+1}$ alignments of a depth $d-1$ tree with a depth 0 tree, having a total complexity of $2^{\delta+1} C(d-1, 0)$.

There are $2^{\delta-2}$ possibilities to match the root of tree T_1 to the level- $(\delta-2)$ vertices of tree T_2 (Figure 5, label ③). In each case the two subtrees of depth $d-1$ of the first tree can be mapped to bottom subtrees of depth 1 or bottom subtrees of depth 0 of the second tree. Latter matchings were considered on the previous step, and the number of possible matchings of a tree depth $d-1$ (two such trees) to a tree depth 1 ($2^{\delta-1}$ such trees) is 2^δ . $2^{\delta-i+1}$ matchings of tree of depth $d-1$ with a tree of depth i (plus some smaller depth tree matchings considered at previous steps) must be considered when the root vertex of the first tree gets mapped to a level $\delta-i$ vertex of the second tree. Finally, if the root vertex of tree T_1 is not mapped anywhere (Figure 5, label ④), this calls for 2 mappings of depth $d-1$ tree to the depth δ tree (tree T_2) plus matchings of depth $d-1$ tree to some smaller subtrees of tree T_2 considered at previous steps.

When all numbers for optimal matchings were found through considering the possible matchings of the root vertex for the tree T_1 , one needs to find maximum of weight of the root mapped some place in tree T_2 plus maximal weight alignment of the left subtree depth $d-1$ and the right subtree of depth $d-1$ with subtrees of T_2 allowed by the hierarchy. This can be done with complexity $C_1 * 2^\delta$, where C_1 is a constant. To do that, one needs to create two copies of tree T_2 , T'_2 and T''_2 , and fill their node weights with the best values for aligning depth $d-1$ subtrees of T_1 with subtrees of T_2 having this node as a root. Term

$$\max_{node \in T_2} w(\text{root}_{T_1} \rightarrow \text{node}) + w(\text{node}_{T'_2}) + w(\text{node}_{T''_2})$$

can be found in linear (on 2^δ) time and it defines the weight of the best matching of T_1 with T_2 .

The result of this observation is the following formula:

$$C(d, \delta) \leq C_1 2^\delta + \sum_{i=0}^{\delta} 2^{i+1} C(d-1, \delta-i),$$

where C_1 is a constant. An upper bound for $C(d, d)$ can be constructed using the following recursion:

$$\begin{cases} D(0, 0) = 1 \\ D(d, \delta) = \sum_{i=0}^{\delta} 2^{i+1} D(d-1, \delta-i). \end{cases} \quad (1)$$

The solution for this recursion is $D(d, \delta) = 2^{d+\delta} \binom{d+\delta}{d}$. Substitution of $d = 0$ and $\delta = 0$ into this formula gives $D(0, 0) = 2^{0+0} \binom{0}{0} = 1$. We now need to show that

$$D(d, \delta) = 2D(d-1, \delta) + 2^2 D(d-1, \delta-1) + \dots + 2^{\delta+1} D(d-1, 0).$$

This will suffice for our proof since recurrence (1) defines function $D(d, \delta)$ uniquely. Substituting the proposed solution into the recurrence formula (1) gives

$$\begin{aligned} 2^{d+\delta} \binom{d+\delta}{d} &= 2^{1+d+\delta-1} \binom{d+\delta-1}{d-1} + \\ &+ 2^{2+d+\delta-2} \binom{d+\delta-2}{d-1} + \dots + 2^{\delta+1+d-1} \binom{d-1}{d-1}. \end{aligned}$$

Term $2^{d+\delta}$ can be deleted from the both sides of the equation leading to the well-known formula for binomial coefficients.

Assume constant $C_1 < 2$. In such case $C(d, \delta) \leq D(d, \delta + 1)$. This can be accurately proven by induction using the following set of inequalities

$$\begin{aligned}
 & D(d, \delta + 1) \\
 &= \sum_{i=0}^{\delta} 2^{i+1} D(d-1, \delta-i+1) + 2^{\delta+1} D(d-1, 0) \\
 &\geq \sum_{i=0}^{\delta} 2^{i+1} C(d-1, \delta-i) + 2^{\delta} C_1 \\
 &\geq C(d, \delta)
 \end{aligned}$$

For higher values of C_1 one might need to compare $C(d, \delta)$ to $D(d, \delta + \text{const})$. We finalize the proof of this Lemma by assuming $d = \delta$ (and $C_1 < 2$) and applying Stirling formula:

$$C(d, d) \leq D(d, d+1) = 2^{2d+1} \binom{2d+1}{d} = \Theta\left(\frac{2^{4d}}{\sqrt{d}}\right).$$

Recall that the problem size, n (number of vertices in the tree) is $2^{d+1} - 1$ in terms of the tree depth making $C(d, d) \preceq \frac{n^4}{\sqrt{\log n}}$. ■

Theorem 6 The DAG alignment problem for two trees with a branching factor bounded by a constant and of logarithmic depth can be solved in polynomial time.

Proof: We use the result of Lemma 1 to first show that any two binary trees of logarithmic depth can be aligned in polynomial time and then extend our solution to the trees with a bounded by constant branching factor and logarithmic depth.

First, aligning two binary trees of depth $c_1 \log n$ and $c_2 \log n$ is no more complex than aligning two complete binary trees of depths $c \log n$, where $c := \max\{c_1, c_2\}$. According to Lemma 1, the latter problem can be solved with a $\Theta\left(\frac{2^{4cd}}{\sqrt{cd}}\right)$ runtime algorithm, meaning at most degree $4c$ polynomial time solution algorithm.

The problem of aligning two depth $c_1 \log n$ and $c_2 \log n$ trees with finite branching factor b is no more complex than aligning two binary trees of depths $c_1 \log b \log n$ and $c_2 \log b \log n$. This is because each b -fork can be substituted by a binary tree of the maximal depth $\log b$. The tree matching algorithm from Lemma 1 would have to be updated as to not match the intermediate vertices introduced in each such substitution, but such an algorithm can only be faster. The overall solution to the problem is an at most degree $4c \log b$ polynomial runtime algorithm. ■

A consequence of this proof is the following result. Any two trees of logarithmic depth can be aligned with a quasi-polynomial ($O(n^{c \log n})$) time algorithm. To prove this upper bound it suffices to replace each b -fork in both graphs with at most $\log n$ depth binary trees.

VI. CONCLUSIONS

We introduced a new type of weighted matching problem called the DAG alignment problem. We formalized this problem, showed that it is NP-complete, proved several upper bounds for approximating solutions to the problem, and finally

introduced several algorithms for solving different classes of the problem. This problem developed through our research on the ontology alignment, however, it relates to many different applications, including, but not limited to, UML diagram comparison, SVG document comparison, and file/folder mapping.

In the future, we plan to find other classes of DAGs that can be aligned faster than with an exponential time algorithm, work on designing efficient heuristics, along with apply some of the ideas discussed to the problem of aligning ontologies.

With ontologies, the problem becomes even more complex because they can contain errors in their specification, meaning that in some circumstances the hierarchical constraint must be relaxed. Moreover, this is likely the case with most applications of the problem. Thus, it may also be an interesting problem to investigate approximate solutions that are allowed to contain a small number of edge conflicts, which will accommodate for some human error in an ontology specification.

ACKNOWLEDGEMENTS

Authors wish to acknowledge Prof. S. B. Gashkov from Lomonosov's Moscow State University and Prof. F. Ruskey from the University of Victoria for their help in proving polynomial time upper bound for the complete binary tree alignment problem (Lemma 1).

This work was supported in part by National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health, and by the PDF grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- [1] H. A. Almohamad and S. O. Duffuaa. A linear programming approach for the weighted graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):522–525, 1993.
- [2] R. Boppana and M. M. Halldórsson. Approximating maximum independent sets by excluding subgraphs. In J. R. Gilbert and R. Karlsson, editors, *SWAT 90 2nd Scandinavian Workshop on Algorithm Theory*, volume 447, pages 13–25, 1990.
- [3] S. Buss and P. Yianilos. A bipartite matching approach to approximate string comparison and search. Technical report, 1995.
- [4] Yves Caseau and Francois Laburthe. Solving various weighted matching problems with constraints. In *Principles and Practice of Constraint Programming*, pages 17–31, 1997.
- [5] M. A. Eshera and K. S. Fu. A graph distance measure for image analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 14(3):398–408, 1984.
- [6] J. Euzenat, T. Le Bach, J. Barrasa, P. Bouquet, J. DeBo, R. Dieng-Kuntz, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. Van Acker, and I. Zaihrayeu. State of the art on ontology. deliverable d2.2.3, 2004.
- [7] Y. Feng, R. L. Goldstone, and V. Menkov. A graph matching algorithm and its application to conceptual system translation. *International Journal on Artificial Intelligence Tools*, 14:77–100, 2005.
- [8] Z. Galil, S. Micali, and H. Gabow. An $o(ev \log v)$ algorithm for finding a maximal weighted matching in general graphs. *SIAM J. Comput.*, 15(1):120–130, 1986.
- [9] M. R. Garey and D. S. Johnson. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
- [10] S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388, 1996.
- [11] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):23–28, 1993.
- [12] J. Hästad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182:105–142, 1999.

- [13] M. M. Hallórssón. Approximation of weighted independent set and hereditary subset problems. In *Proceedings of COCOON'99*, 1999.
- [14] M. Hallórssón, J. Kratochvíl, and J. Telle. Independent sets with domination constraints. *Discrete Applied Mathematics*, 99:39–54, 1999.
- [15] R. A. Hummel and S. W. Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3), 1983.
- [16] M. Krcmar and A. Dhawan. Application of genetic algorithms in graph matching. In *International Conference on Neural Networks*, volume 6, pages 3872–3876, 1994.
- [17] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–95, 1955.
- [18] P. Kuner and B. Ueberreiter. Pattern recognition by graph matching combinatorial versus continuous optimization. *International Journal Pattern Recognition and Artificial Intelligence*, 2:527–542, 1988.
- [19] J. Niere. Visualizing differences of uml diagrams with fujaba. In *Proceedings of the Fujaba Days 2004*, 2004.
- [20] N. Noy and M. Musen. The prompt suite: Interactive tools for ontology merging and mapping. Technical report, 2002.
- [21] N. F. Noy and M. A. Musen. An algorithm for merging and aligning ontologies: Automation and tool support. In *Sixteenth National Conference on Artificial Intelligence (AAAI-99), Workshop on Ontology Management*, 1999.
- [22] A. Rangarajan and E. Mjolsness. A Lagrangian relaxation network for graph matching. In *International Conference on Neural Networks*, volume 7, pages 4629–4634. Inst. Electrical & Electronics Engineers, 1994.
- [23] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 1995.
- [24] L. Shapiro and R. Haralick. Structural descriptions and inexact matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:504–519, 1981.
- [25] W. H. Tsai and K. S. Fu. Subgraph error-correcting isomorphism for syntactic pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 13:48–62, 1983.
- [26] S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988.

