Abstract

We study anti-unification for possibly cyclic, unranked term-graps and develop an algorithm, which computes a minimal complete set of generalizations for them. For bisimilar graphs the algorithm computes the join in the lattice generated by a functional bisimulation. These results generalize anti-unification for ranked and unranked terms to the corresponding term-graps, and solve also anti-unification problems for rational terms and dags. Our results open a way to widen anti-unification based code clone detection techniques from a tree representation to a graph representation of the code.

1 Introduction

Term-graps are rooted, directed, labeled graphs, which may contain cycles. They can be used to represent functional expressions compactly and to process them efficiently with the help of graph transformations. Rewriting with term-graps has been studied quite intensively, see, e.g., [4, 6, 15, 19, 20, 27]. Term-graps can be represented in various ways, for instance, as constraints [6], hypergraphs [27], systems of recursion equations [4], or arrows in a category [15]. With cycles, term-graps can express infinite terms and can model regular infinite data structures. Some related (not necessarily equivalent) representations, that are widely used in computer science, include dags, \( \mu \)-terms, control flow graphs, abstract semantic graphs, program dependency graphs, certain kinds of flowcharts, process graphs, etc.
In this paper, we study the anti-unification problem for term-graphs: Given two such graphs $G_1$ and $G_2$ (maybe with cycles), our goal is to find a graph $G$, which is a least general common generalization of $G_1$ and $G_2$. It means, there should exist variable substitutions $\sigma_1$ and $\sigma_2$ such that the instances of $G$ with respect to them, i.e., the graphs $G_1 \sigma_1$ and $G_2 \sigma_2$, are equivalent to $G_1$ and $G_2$, respectively.

Our representation of term-graphs follows the approach from [4], based on recursion equations. The difference is that we are not restricted to ranked alphabets. Variadic function symbols are permitted and, to take the advantage of such variadicity, hedge variables are used together with individual variables. The latter stands for single graphs, while the former can be instantiated by hedges (finite sequences) of graphs. The equivalence relation is bisimilarity.

It has already been shown in [22] that anti-unification for unranked finite terms is finitary: There are, in general, finitely many least general generalizations (lggs). The same holds for unranked term-graphs, discussed in this paper. We develop an algorithm, which computes such lggs. Equivalence class of a term-graph with respect to bisimilarity is a complete lattice. For bisimilar terms, our algorithm computes the lgg, which is the join in this lattice.

The intuition behind lggs is that they should contain “maximal similarities” between the input graphs and should abstract differences between them by variables uniformly. While this might sound similar to the problem of computing maximal common subgraphs (mcs) between graphs [23,24], lggs, in general, might contain more edges than mcs’s and also give information about differences, which is usually neglected in mcs’s.

The results reported in this paper extend our previous results for unranked finite terms [8,22] to unranked cyclic graphs. In particular, we extend rigid anti-unification from terms to graphs. The rigid version is a more efficient variant of the unranked anti-unification algorithm, since it computes only certain kind of generalizations. It is guided by a rigidity function, which, essentially, decides which nodes of the input graphs should be retained in the generalization. Rigidity function is a parameter of the algorithm. Properties of the latter are proved for arbitrary values of this parameter. As special cases of our results, we obtain anti-unification for ranked term-graphs, rational trees, $\mu$-terms, and dags. To the best of our knowledge, generalization for these structures has not been addressed yet in the literature.

Anti-unification has a pretty wide scope of interesting applications. Originally, it was introduced for inductive reasoning [26]. As a method of computing generalizations, variants of anti-unification are important ingredients of techniques and tools that have found applications in various areas of artificial intelligence, machine learning, reasoning, linguistics, program synthesis, analysis, transformation, verification, etc. We can not give an exhaustive overview of all related work here. A couple of recent references (motivated by different applications) include, e.g., [1,2,7,9,10,13,18,21,25]. A particularly interesting motivation comes from software code clone detection, where anti-unification has been successfully incorporated at the level of abstract syntax trees [14,16,28]. Our results can serve as a starting point to extend these techniques for graph-based representation of code (e.g., abstract semantic graphs or program dependence graphs) or graph-based languages (e.g., for model transformation). Besides, term-graph anti-unification can be used to construct an index for sets of dags (e.g., substitution tree indexing), which can be useful in declarative programming and reasoning.

The paper is organized as follows: In Sect. 2, we introduce the notions related to unranked term-graphs. Sect. 3 briefly recalls results about term-graph bisimilarity. The notions related to substitutions and generalizations are introduced in Sect. 4. The term-graph generalization algorithm is described in Sect. 5. Conclusions and the future work are discussed in Sect. 6.

An experimental implementation of our anti-unification algorithm can be accessed online: http://www.risc.jku.at/projects/stout/software/tgau.php.
We start by defining unranked (possibly infinite) terms. A term without fixed arity), term variables $V_t$, and hedge variables $V_h$, an unranked term is a partial mapping $t : \mathbb{N}^* \rightarrow F \cup V_t \cup V_h$ such that
- the domain of $t$, denoted $\text{dom}(t)$, is non-empty and prefix-closed (i.e., if $p_1, p_2 \in \mathbb{N}^*$ and $p_1, p_2 \in \text{dom}(t)$, then $p_1 \in \text{dom}(t)$),
- for all $p \in \mathbb{N}^*$, if $t(p) \in F$, then there exists a natural number $n \geq 0$ such that $p.\hat{i} \in \text{dom}(t)$ for all $1 \leq i \leq n$ and $p.i \notin \text{dom}(t)$ for all $i > n$,
- for all $p \in \mathbb{N}^*$, if $t(p) \in V_t \cup V_h$, then for all $n$ we have $p.n \notin \text{dom}(t)$,
- $t(\epsilon) \notin V_t$.

A term $t$ is finite if $\text{dom}(t)$ is a finite set. Otherwise it is infinite. A term is rational if it has finitely many distinct subterms. Hedges are finite (possible empty) sequences of terms and hedge variables. The set of terms (respectively, hedges) over $F$, $V_t$, and $V_h$ is denoted by $T(F, V_t, V_h)$ (respectively, $H(F, V_t, V_h)$). We use the letters $f, g, h, a, b, c,$ and $d$ for function symbols, $x, y, z$, and $\nu$ for term variables, $X, Y, Z,$ and $U$ for hedge variables, $\chi, \nu, \upsilon$ and $\omega$ for a term variable or a hedge variable, and $t$ and $r$ for terms, $s$ and $q$ for a hedge variable or a term, and $\overline{s}$ and $\overline{q}$ for hedges. The empty hedge is denoted by $\epsilon$. Given a sequence $\overline{s}$, the $i$th element of $\overline{s}$ is denoted by $\overline{s}[i]$. Furthermore, $\overline{s}[i:j]$ denotes the subsequence between the positions $i$ and $j$ where $i < j$, that is, $\overline{s}[i+1], \ldots, \overline{s}[j-1]$. Unranked terms (resp. hedges) can be naturally represented as unranked trees (resp. forests).

Example 2. In Fig. 1 we visualize three examples of a finite, infinite non-rational, and infinite rational terms in form of trees. The triangles represent infinite subtrees:

For a term $t$, we denote by $V_t(t)$, $V_h(t)$, and $V(t)$ respectively the sets of term variables, hedge variables, and all variables occurring in $t$. The notation extends to hedges as well.

Now we define unranked cyclic term-graphs with the help of recursion equations. We start with a very general notion of a system of recursion equations and subsequently impose restrictions to get to the interesting concept.

Definition 3. A system of recursion equations over $F$, $V_t$, and $V_h$ is a set of equations $\{x_i \overset{t_i}{\rightarrow} x_j, x_i \overset{t_n}{\rightarrow} X_1 \overset{s_1}{\rightarrow} \ldots, X_m \overset{s_m}{\rightarrow} X_j\}$, where for all $i, j$, $1 \leq i < j \leq n$, $x_i \neq x_j$, all $t_i$’s are finite terms, for all $i, j$, $1 \leq i < j \leq m$, $X_i \neq X_j$, and $s_\nu$’s are hedges consisting of finite terms or hedge variables. The variables $x_1, \ldots, x_n, X_1, \ldots, X_m$ are called recursion variables. They are bound in the system. All other variables occurring in the system are free.
We will use different notation for free and bound variables in systems of recursion equations, writing the latter in bold font. One recursion variable (usually, the leading variable of the first equation) is a designated one and we call it the root of the system \( \Gamma \), denoted by \( \text{root}(\Gamma) \). It is always a term variable.

A recursion variable \( v \) is reachable from a recursion variable \( \chi \) in a system \( \Gamma \) if \( \Gamma \) contains an equation of the form \( \chi \doteq \hat{s} \in \Gamma \) and either \( v \in \mathcal{V}(\hat{s}) \), or \( v \) is reachable from some recursion variable \( \upsilon \in \mathcal{V}(\hat{s}) \). In particular, we say that a hedge variable \( Y \) is horizontally reachable from a hedge variable \( X \) in \( \Gamma \), if \( \Gamma \) contains an equation \( X \doteq \hat{\upsilon} \) such that either \( \hat{\upsilon} \) has the form \((\hat{s}_1, Y, \hat{s}_2)\) or it has the form \((\hat{s}_1, Z, \hat{s}_2)\) and \( Y \) is horizontally reachable from \( Z \). An equation is called useless in \( \Gamma \) if its leading recursion variable is not reachable from \( \text{root}(\Gamma) \).

A system \( \Gamma \) is called horizontally bounded if no hedge variable is reachable from itself in \( \Gamma \), i.e., \( \Gamma \) contains no horizontal cycle.\(^1\) For instance, \( \{ x \doteq f(x), X \doteq (x, Y) \} \) is a horizontally bounded system, while \( \{ x \doteq f(x), X \doteq (x, X) \} \) is not.

We do not distinguish between two systems of recursion equations if they differ from each other only by renaming of bound variables.

A system of recursion equations is called flat, if the equations have one of the following three possible forms: \( x \doteq f(X_1, \ldots, X_n) \), \( x \doteq u \) where \( u \) is a free or bound term variable, and \( X \doteq (v_1, \ldots, v_n) \) where \( n \geq 0 \) and each \( v_i \) is a free or bound term or hedge variable.

A system of recursion equations \( \Gamma \) is in canonical form if it does not contain useless equations and each equation in \( \Gamma \) has one of the following forms:

- \( x \doteq f(X_1, \ldots, X_n) \), where the \( X_i \)'s are (not necessarily distinct) recursion variables, or
- \( x \doteq y \), where \( y \) is a free variable, or
- \( X \doteq Y \), where \( Y \) is a free variable.

For instance, \( \{ x \doteq f(y, X, X), y \doteq g(x), X \doteq Y \} \) and \( \{ x \doteq f(y, z, z), y \doteq g(x), z \doteq a \} \) are in canonical form, while \( \{ x \doteq f(g(x), X), X \doteq Y \} \), \( \{ x \doteq f(y, X), y \doteq g(x) \} \), \( \{ x \doteq f(y, X), y \doteq g(x), X \doteq \epsilon \} \), and \( \{ x \doteq f(y, X), y \doteq g(x), X \doteq Y, Y \doteq Z \} \) are not.

Every canonical system is flat and horizontally bounded. On the other hand, each flat horizontally bounded system can be transformed to the canonical form by performing the following canonicalization steps as long as possible:

- Remove useless equations.
- Remove trivial equations of the form \( x \doteq y \) and replace all occurrences of \( x \) by \( y \). If \( y = x \), then replace the equation by \( x \doteq * \), where * is some predefined constant from \( \mathcal{F} \).
- Replace equations of the form \( X \doteq (v_1, \ldots, v_n) \), \( n > 1 \), where \( v_i \)'s are free or bound term or hedge variables, by \( n \) new equations \( Y_i \doteq v_i \), \( 1 \leq i \leq n \), where \( Y_i \)'s are fresh hedge variables, and replace each occurrence of \( X \) by \( (Y_1, \ldots, Y_n) \).
- Replace equations of the form \( X \doteq u \) by \( x \doteq u \) and replace each occurrence of \( X \) by \( x \), where \( u \) is a free or bound term variable and \( x \) is a fresh term variable.
- Remove trivial equations of the form \( X \doteq Y \) and replace all occurrences of \( X \) by \( Y \).
- Remove equations of the form \( X \doteq \epsilon \) and remove each occurrence of \( X \).

Essentially, this canonicalization extends the canonicalization from [4] by four steps dealing with hedge variables. These steps split each equation of the form \( X \doteq (s_1, \ldots, s_n) \) into \( n \) new equations (one for each \( s_i \)), and, eventually only those are retained for which \( s_i \) is a free variable. The bound \( s_i \)'s at the end replace their leading recursion variables.

\(^1\) Systems that are not horizontally bounded can be used to define cyclic term-graphs where cycles are formed both vertically and horizontally. Such term-graphs can model infinitely branching trees of infinite depth. These structures go beyond the scope of this paper.
Intuitively, canonical systems of recursion equations can be naturally represented by graphs: The nodes will be the recursion variables; a node $x$ will be connected to a node $\chi$ by an edge if the system contains an equation $x \doteq f(\ldots, \chi, \ldots)$; each node $x$ will have a label $f$ for an equation $x \doteq f(\ldots)$ or the label $y$ for an equation $x \doteq y$, each node $X$ will have a label $Y$ for an equation $X \doteq Y$. Every node is reachable from the root. Cycles and sharings are defined by the occurrences of recursion variables. This intuition justifies the definition:

- **Definition 4.** A term-graph is a system of recursion equations in canonical form.

- **Example 5.** We show some term-graphs and their defining recursion equations.

  1. Graph:

     \[ \begin{array}{c}
     x \\
     f \\
     \downarrow \\
     y \\
     \end{array} \]

     Equations: \( x \doteq f(y, x) \)

     \( y \doteq a \)

  2. Graph:

     \[ \begin{array}{c}
     x \\
     f \\
     \downarrow \\
     g \\
     y \\
     \end{array} \]

     Equations: \( x \doteq f(x, y) \)

     \( y \doteq g(y, x) \)

  3. Graph:

     \[ \begin{array}{c}
     y \\
     f \\
     \downarrow \\
     u \\
     \end{array} \]

     Equations: \( y \doteq f(u, z) \)

     \( z \doteq f(u, y) \)

     \( u \doteq x \)

A flat horizontally bounded system and its canonical form have the same (possibly infinite) term unwinding. In the rest of the paper we consider only canonical systems of recursion equations. The words “system of recursion equations” and “(term)-graphs” will be used interchangeably. The letter $\mathcal{G}$ will be used to denote term-graphs.

Given a term-graph $\mathcal{G}$ and an equation $x \doteq t$, the subgraph of $\mathcal{G}$ rooted at $x$ is the set $\text{subgraph}(\mathcal{G}, x) = \{ x \doteq t \} \cup \{ \chi \doteq s \mid \chi \in \mathcal{G} \text{ is reachable from } x \}$. Obviously, $\text{subgraph}(\mathcal{G}, \text{root}(\mathcal{G})) = \mathcal{G}$.

The set of nodes of a term-graph $\mathcal{G}$ is denoted by $\text{nodes}(\mathcal{G})$. If $x \in \text{nodes}(\mathcal{G})$ and $\mathcal{V}$ is its $i$th successor (i.e., $x \doteq t \in \mathcal{G}$ for some $t$ and $\mathcal{V}$ is $i$th argument of $t$), we will write $x \rightarrow_i \mathcal{V}$. An access path of $\mathcal{V} \in \text{nodes}(\mathcal{G})$ is a possibly empty finite sequence of positive natural numbers $(i_1, \ldots, i_j)$ such that there exist $x_1, \ldots, x_{j-1} \in \text{nodes}(\mathcal{G})$ with $\text{root}(\mathcal{G}) \rightarrow_{i_1} x_1 \rightarrow_{i_2} \cdots \rightarrow_{i_{j-1}} x_{j-1} \rightarrow_{i_j} \mathcal{V}$. A node may have several access paths. The set of all access paths of a node $\chi$ is denoted by $\text{acc}(\chi)$.

We will also consider term-graph hedges, defined analogously to hedges: they are finite, possibly empty sequences of term-graphs and hedge variables. We will use $\tilde{\mathcal{G}}$ to denote them.

### 3 Bisimilarity Relation

It is straightforward to adapt the notions of bisimulation and bisimilarity [4] to our graphs:

- **Definition 6.** Let $\Gamma_1 = \{ x_1 \doteq s_1, \ldots, x_n \doteq s_n \}$ and $\Gamma_2 = \{ v_1 \doteq q_1, \ldots, v_m \doteq q_m \}$ be two systems of recursion equations. Then $R$ is a bisimulation from $\Gamma_1$ to $\Gamma_2$ iff
  - $R$ is a binary relation with the domain $\{ x_1, \ldots, x_n \}$ and codomain $\{ v_1, \ldots, v_m \}$.
  - The roots of $\Gamma_1$ and $\Gamma_2$ are related: $x_1 \in R v_1$. 

---

**A. Baumgartner, T. Kutsia, J. Levy, and M. Villaret**

---

**FSCD 2018**
If \( \chi G_s, \chi \in l(x_1^1, \ldots, x_k^1) \in \Gamma_1, k_i \geq 0 \), and \( \nu_j \in l_2(\nu_k^2, \ldots, \nu_k^2) \in \Gamma_2, k_i \geq 0 \), then
\[
l_1 = l_2, k_i = k_j, \text{ and } \chi G_s(\nu), \text{ for all } 1 \leq u \leq k_i.
\]
(0.)
In short, bisimulation means that the roots are related, related nodes have the same label, and their successor nodes are again related.

**Definition 7.** Two graphs are bisimilar, if there exists a bisimulation from one to another.

Bisimilarity is an equivalence relation, see, e.g., [4]. We write \( \mathcal{G}_1 \simeq \mathcal{G}_2 \) if \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \) are bisimilar, and \( \mathcal{G}_1 \succeq \mathcal{G}_2 \) if there exists a functional bisimulation from \( \mathcal{G}_1 \) to \( \mathcal{G}_2 \) (i.e., a bisimulation which is a function).

Functional bisimulation collapses a graph into a smaller one. For the other way around, one says that the equivalence class of a term-graph \( \mathcal{G} \) with respect to bisimilarity is a complete lattice, partially ordered by functional bisimulation. The least upper bound in this lattice is a rational term, denoted by \( \triangledown \mathcal{G} \), and the greatest lower bound is a fully collapsed graph, denoted by \( \triangledown \mathcal{G} \).

**Example 8.** Let \( \mathcal{G} \) be the term graph \( \{ x \equiv f(y, z), y \equiv a, z = f(y, x) \} \). Then \( \triangledown \mathcal{G} \) is the finite rational term depicted in Fig. 1 in Example 2, and \( \triangledown \mathcal{G} \) is the first graph in Example 5.

Given a bisimulation relation \( R \) from a term-graph \( \mathcal{G}_1 \) to a term-graph \( \mathcal{G}_2 \), its associated graph \( \mathcal{G}_R^A \) is defined as follows: (i) nodes\( (\mathcal{G}_R^A) = R \), root\( (\mathcal{G}_R^A) = (\text{root}(\mathcal{G}_1), \text{root}(\mathcal{G}_2)) \), the label of each \((x_1, x_2) \in \text{nodes}(\mathcal{G}_R^A)\) is that of \( x_1 \) (which is the same as the label of \( x_2 \)); (ii) if \( x_1 \in \text{nodes}(\mathcal{G}_1), x_2 \in \text{nodes}(\mathcal{G}_2), (x_1, x_2) \in R, x_1 \rightarrow_1 x_1', \text{ and } x_2 \rightarrow_1 x_2' \), then in \( \mathcal{G}_R^A \) we have \((x_1, x_2) \rightarrow_1 (x_1', x_2')\).

### 4 Substitutions and Generalizations

The notions related to substitutions, formulated for finite unranked terms and hedges in [22], can be reused with a slight modification for (possibly) infinite terms and hedges.

A substitution is a mapping from term variables to terms and from hedge variables to hedges, which is the identity almost everywhere. We use the traditional finite set representation of substitutions, e.g., \( \{ x \mapsto f(a, f(a, \ldots)), X \mapsto e, Y \mapsto (X, g(Y, g(Y, \ldots, Y), Y) \} \), which stands for the substitution that maps every variable to itself except \( x, X, Y \) that are mapped respectively to \( f(a, f(a, \ldots)), e, \) and \( (X, g(Y, g(Y, \ldots, Y), Y)) \).

The lower case Greek letters are used to denote substitutions, with the exception of the identity substitution for which we write \( \text{Id} \). The domain and range of a substitution \( \sigma \) are defined in the usual way: \( \text{dom}(\sigma) = \{ x \in V \mid \sigma(x) \neq x \} \) and \( \text{ran}(\sigma) = \{ \sigma(x) \mid x \in \text{dom}(\sigma) \} \).

Substitutions can be applied to terms and hedges using the congruences \( \sigma(f(s_1, \ldots, s_n)) = f(\sigma(s_1), \ldots, \sigma(s_n)) \) and \( \sigma(s_1, \ldots, s_n) = (\sigma(s_1), \ldots, \sigma(s_n)) \). We call \( \sigma(s) \) and \( \sigma(\tilde{s}) \) the instances of respectively \( s \) and \( \tilde{s} \) and use postfix notation to denote them, writing \( s\sigma \) and \( \tilde{s}\sigma \).

We also say that \( \tilde{s} \) is more general than \( \tilde{q} \) if \( \tilde{q} \) is an instance of \( \tilde{s} \) and denote this fact by \( \tilde{s} \preceq \tilde{q} \). If \( \tilde{s} \preceq \tilde{q} \) and \( \tilde{q} \preceq \tilde{s} \), then we write \( \tilde{s} \sim \tilde{q} \). If \( \tilde{s} \preceq \tilde{q} \) and \( \tilde{q} \neq \tilde{q} \), then we say that \( \tilde{s} \) is strictly more general than \( \tilde{q} \) and write \( \tilde{s} \prec \tilde{q} \).

The composition of two substitutions \( \sigma \) and \( \vartheta \), written as \( \sigma \vartheta \), is defined as the composition of two mappings: We have \( s(\sigma \vartheta) = (s\sigma)\vartheta \) for all \( s \). A substitution \( \sigma_1 \) is more general than \( \sigma_2 \) with respect to a set of variables \( \mathcal{X} \subseteq \mathcal{Y} \), written \( \sigma_1 \preceq_{\mathcal{X}} \sigma_2 \), if there exists \( \vartheta \) such that \( \chi \sigma_1 \vartheta = \chi \sigma_2 \), for each \( \chi \in \mathcal{X} \). The relations \( \simeq \) and \( \sim \) are extended to substitutions: \( \sigma_1 \simeq_{\mathcal{X}} \sigma_2 \) means \( \sigma_1 \preceq_{\mathcal{X}} \sigma_2 \) and \( \sigma_2 \preceq_{\mathcal{X}} \sigma_1 \), and \( \sigma_1 \sim_{\mathcal{X}} \sigma_2 \) means \( \sigma_1 \preceq_{\mathcal{X}} \sigma_2 \) and \( \sigma_1 \not\preceq_{\mathcal{X}} \sigma_2 \).
Next we define substitutions directly for term-graphs, i.e., for systems of recursion equations (in canonical form). Instead of writing the whole systems of recursion equations in the range of substitutions, only the roots of the corresponding term-graphs appear there. Hedge variables in the image remain unchanged. For instance, assume the term-graphs $G_1$ and $G_2$ are given by the systems of recursion equations: $G_1 = \{ x \doteq f(y, x), y \doteq a \}$, and $G_2 = \{ x \doteq g(x, x, x), X \doteq Y \}$. Then the substitution $\{ x \mapsto f(a, f(a, \ldots)), X \mapsto \epsilon, Y \mapsto (X, g(Y, g(Y, \ldots, Y)), Y) \}$ we considered above can be written as $\{ x \mapsto \text{root}(G_1), X \mapsto \epsilon, Y \mapsto (X, \text{root}(G_2)) \}$. The bound variables in $G_1$ and $G_2$ should be appropriately renamed to guarantee that the names are distinct from each other and from free variables.

To define application of such a substitution to a term-graph, we assume that all term-graphs are in canonical form and the bound variables are appropriately renamed. Let $\sigma$ be a substitution and $G$ be a term-graph. Then the term-graph $\sigma(G)$, the instance of $G$ under $\sigma$, is obtained by canonicalizing the following flat horizontally bounded system of recursion equations: $\{ x \doteq \sigma(s) \mid x \doteq \bar{s} \in G \} \cup G_1 \cup \cdots \cup G_n$, where $G_1, \ldots, G_n$ are all term-graphs whose roots appear in $\text{ran}(\sigma)$. Substitution application naturally extends to term-graph hedges.

**Example 9.** Let $G$ be the term-graph:
\[
G = \{ x_0 \doteq f(x_0, X_1, x_1, X_1, x_2), X_1 \doteq X, x_1 \doteq g(X_2, x_2, X_2), X_2 \doteq Y, x_2 \doteq \epsilon \}.
\]
Let $\sigma = \{ x \mapsto \text{root}(G_1), X \mapsto (\text{root}(G_2), X), Y \mapsto \epsilon \}$, $G_1 = \{ y \doteq f(y) \}$, $G_2 = \{ z \doteq a \}$. Then
\[
\sigma(G) = \{ x_0 \doteq f(x_0, z, Z, x_1, z, Z, y), z \doteq a, Z \doteq X, x_1 \doteq g(y), y \doteq f(y) \}.
\]

The notion of more general term-graphs and term-graph hedges is defined modulo bisimilarity: $\bar{G}_1$ is more general than $\bar{G}_2$, if there is a substitution $\sigma$ such that $\bar{G}_1 \sigma \sim \bar{G}_2$.

We reuse the symbol $\simeq$ for this relation over term-graphs and term-graph hedges, and also write $\bar{G}_1 \simeq \bar{G}_2$ if $\bar{G}_1 \preceq \bar{G}_2$ and $\bar{G}_2 \preceq \bar{G}_1$. For the strict part of $\preceq$ we reuse $\prec$. Analogously for substitutions: A substitution over term-graphs $\sigma_1$ is more general than a substitution over term-graphs $\sigma_2$ with respect to a set of variables $X \subseteq V$, if there exists $\vartheta$ such that $\chi \sigma_1 \vartheta \prec \chi \sigma_2$ for each $\chi \in X$. Also in this case we reuse the $\preceq$ symbol and write $\sigma_1 \succeq_{x} \sigma_2$ (and similarly for the relations $\approx$ and $\prec$ for substitutions).

A term-graph hedge $\bar{G}$ is called a generalization of two-term-graph hedges $\bar{G}_1$ and $\bar{G}_2$ if $\bar{G} \preceq \bar{G}_1$ and $\bar{G} \preceq \bar{G}_2$. We say that a term-graph $\bar{G}$ is a least general generalization (lgg in short) of $\bar{G}_1$ and $\bar{G}_2$ if $\bar{G}$ is a generalization of $\bar{G}_1$ and $\bar{G}_2$ and there is no generalization $\bar{G}'$ of $\bar{G}_1$ and $\bar{G}_2$ that satisfies $\bar{G} \prec \bar{G}'$. That means, there are no generalizations of $\bar{G}_1$ and $\bar{G}_2$ that are strictly less general than their least general generalization.

An anti-unification triple, AUT in short, is written $\chi : \bar{G}_1 \nash \bar{G}_2$, where $\chi$ does not occur in $\bar{G}_1$ and $\bar{G}_2$. Intuitively, $\chi$ is a variable that stands for the most general generalization of $\bar{G}_1$ and $\bar{G}_2$. An anti-unifier of $\chi : \bar{G}_1 \nash \bar{G}_2$ is a substitution $\sigma$ such that $\text{dom}(\sigma) \subseteq \{ \chi \}$ and $\chi \sigma$ is a generalization of both $\bar{G}_1$ and $\bar{G}_2$. An anti-unifier $\sigma$ of an AUT $\chi : \bar{G}_1 \nash \bar{G}_2$ is least general (or most specific) if there is no anti-unifier $\vartheta$ of the same problem that satisfies $\chi \vartheta \preceq \chi \sigma$. If $\sigma$ is a least general anti-unifier of an AUT $\chi : \bar{G}_1 \nash \bar{G}_2$, then $\chi \sigma$ is an lgg of $\bar{G}_1$ and $\bar{G}_2$.

A complete set of generalizations of two-term-graph hedges $\bar{G}_1$ and $\bar{G}_2$ is a set of term-graph hedges that satisfies the properties:

**Soundness:** Each $\bar{G} \in G$ is a generalization of both $\bar{G}_1$ and $\bar{G}_2$.

**Completeness:** For each generalization $\bar{G}'$ of $\bar{G}_1$ and $\bar{G}_2$, there exists $\bar{G} \in G$ such that $\bar{G}' \preceq \bar{G}$.

$G$ is a minimal complete set of generalizations (mcsig) of $\bar{G}_1$ and $\bar{G}_2$ if, in addition to soundness and completeness, it satisfies also the following property:
Minimality: For each $\tilde{G}_1, \tilde{G}_2 \in G$, if $\tilde{G}_1 \not\leq \tilde{G}_2$ then $\tilde{G}_1 = \tilde{G}_2$.

Lemma 10. For any hedges $\tilde{s}$ and $\tilde{q}$ there exists their minimal complete set of generalizations. This set is finite and unique modulo $\simeq$.

Proof. Similar to the analogous lemma for hedges with finite terms, see [22].

Theorem 11. For any term-graph hedges $\tilde{G}_1$ and $\tilde{G}_2$ there exists their minimal complete set of generalizations. This set is finite and unique modulo $\simeq$ and $\sim$.

Proof. Note that $G \sim \delta G$ for all $G$. Let $\tilde{G}_1 = (G_1^1, \ldots, G_n^1)$ and $\tilde{G}_2 = (G_1^2, \ldots, G_n^2)$. By Lemma 10, the hedges $(\delta G_1^1, \ldots, \delta G_1^n)$ and $(\delta G_2^1, \ldots, \delta G_2^n)$ have a finite minimal complete set of generalizations, unique modulo $\simeq$.

Our goal is not to compute minimal complete sets of generalizations. We would rather focus on so called rigid generalizations, which we define below. The motivation comes from the experience with finite unranked term anti-unification, where unrestricted mcsg can grow too big and it makes sense to restrict consecutive hedge variables in the generalization. For the details, see [22].

Definition 12 (Alignment, Rigidity Function). Let $w_1$ and $w_2$ be strings of symbols. The sequence $a_1[i_1, j_1] \cdots a_n[i_n, j_n]$, for $n \geq 0$, is an alignment if $i$’s and $j$’s are positive integers such that $0 < i_1 < \cdots < i_n < |w_1|$ and $0 < j_1 < \cdots < j_n < |w_2|$, and $a_k = w_1[i_k] = w_2[j_k]$ for all $1 \leq k \leq n$. A rigidity function $R$ is a function that returns, for every pair of strings of symbols $w_1$ and $w_2$, a set of alignments of $w_1$ and $w_2$.

For instance, if $R$ computes the set of all longest common subsequences, then $R(abeda, bcd) = \{b[2, 1]c[3, 2]a[5, 3], b[2, 1]c[3, 2]d[4, 4]\}$.

The top symbol of a term is defined as $top(x) = x$ for any variable $x$, and $top(f(\tilde{s})) = f$ for any term $f(\tilde{s})$. The notion is extended to hedges: $top(X) = X$ and $top(s_1, \ldots, s_n) = (top(s_1), \ldots, top(s_n))$. If $\{x_1 = s_1, \ldots, x_l = s_l\} \subseteq G$, $n > 0$, then we define $top(x_1, \ldots, x_n, G)$ as $top(s_1, \ldots, s_n)$. Moreover, we define $top(\tilde{G}) = top(root(\tilde{G}), G)$.

Definition 13 ($R$-Generalization). Given two term-graphs $G_1$ and $G_2$ (without common free and bound variables) and the rigidity function $R$, we say that a term-graph $\tilde{G}$ that generalizes both $G_1$ and $G_2$ is their generalization with respect to $R$, or, shortly, an $R$-generalization, if either

\[ R(top(\tilde{G}_1), top(\tilde{G}_2)) \in \{\emptyset, \{t\}\} \] and $G = \{x \equiv y\}$, where $x$ is a new bound term variable and $y$ is a new free term variable, or

\[ f[1, 1] \in R(top(\tilde{G}_1), top(\tilde{G}_2)) \] for some $f$ and $G = \{root(\tilde{G}) \equiv f(\tilde{X})\} \cup \tilde{Y} \cup \tilde{G}_1' \cup \cdots \cup \tilde{G}_n'$, where $\tilde{X}$ does not contain pairs of consecutive hedge recursion variables.

The sequence $\tilde{X}$, the set $\tilde{Y}$, and the graphs $\tilde{G}_1', \ldots, \tilde{G}_n'$ are defined as follows:

For $i = 1, 2$, the original graph $\tilde{G}_i$ contains an equation $root(\tilde{G}_i) \equiv f(\tilde{Y}_i)$ and there exists an alignment $g_1[i_1, j_1] \cdots g_n[i_n, j_n] \in R(top(\tilde{Y}_1, \tilde{G}_1), top(\tilde{Y}_2, \tilde{G}_2))$, satisfying the following conditions:

1. If we remove all hedge recursion variables that occur as elements of $\tilde{X}$, we get a sequence of term recursion variables $(x_1, \ldots, x_n)$, such that $x_k = root(\tilde{G}_k')$ and each $\tilde{G}_k'$ contains an equation of the form $x_k \equiv g_k(\tilde{V}_k)$ for all $1 \leq k \leq n$, and

Note that unrestricted unranked term anti-unification (i.e., without a rigidity function) can be also modeled as associative anti-unification with the unit element. The latter problem has been studied, e.g., in [2, 3].
2. For every \(1 \leq k \leq n\), there exists a pair of term recursion variables \(y_k^1\) and \(y_k^2\) such that \(\overline{v}_1|_{l_k} = y_k^1\), \(\overline{v}_2|_{l_k} = y_k^2\), and \(G_k^*\) is an \(R\)-generalization of subgraph \((G_1, y_1^1)\) and subgraph \((G_2, y_1^2)\).

3. \(\mathcal{Y} = \{Y_1 \doteq Z_1, \ldots, Y_m \doteq Z_m\}\), where \(Y_1, \ldots, Y_m\) are all hedge recursion variables in \(\chi\) and \(Z_1, \ldots, Z_m\) are new free hedge variables.

Example 14. Let \(R\) compute the set of all longest common subsequences and let \(G_1 = \{x_0 \doteq f(x_1, x_2), x_1 \doteq g(x_2, x_2), x_2 \doteq a\}\) and \(G_2 = \{y_0 \doteq f(y_1, y_0, y_0), y_1 \doteq g, y_2 \doteq a\}\). The term graph \(\{z_0 \doteq f(z_1, Z_1, Z_2, Z_1), z_1 \doteq g(Z_2), Z_2 \doteq Z_2\}\) is an \(R\)-generalization of \(G_1\) and \(G_2\) while \(\{z_0 \doteq f(z_1, Z_1, Z_2, Z_1), z_1 \doteq g(Z_2, Z_2), Z_2 \doteq Z_2\}\) is not.

5 The Algorithm

We present our anti-unification algorithm as a rule-based algorithm that works on quadruples \(A; S; T; G\), called configurations. Here \(A\), \(S\), and \(T\) are sets of anti-unification triples and \(G\) is a term-graph. The rules transform configurations into configurations. Intuitively, the problem set \(A\) contains AUTs that have not been solved yet, the store \(S\) contains the already solved AUTs, the trail \(T\) keeps track of the names of recursion variables, and \(G\) is the generalization which becomes more and more specific as the algorithm progresses by applying the rules.

To keep the notation short, in anti-unification triples we only use variables from the graphs to be generalized. Those graphs are not explicitly present in the configurations, but are global parameters, denoted by \(G_1\) and \(G_2\). For simplicity, we assume that \(G_1\) and \(G_2\) do not contain free variables. This is not a restriction, because we can replace free variables by new constants, use the algorithm defined below, and in the generalization replace those constants back with variables. (In case of hedge variables, we might need to replace their corresponding generalization term variables by generalization hedge variables.) The rigidity function \(R\) is yet another global parameter. In the rules below, generalization term-graphs are assumed to be implicitly transformed into the canonical form.

Step: Simplification Step

\(\{x : y \doteq z\} \cup A; S; T; G \implies A_0 \cup A; S; T \cup \{u : y \doteq z\}; G\{x \mapsto u\} \cup \{u \doteq t\}\),

where \(y \doteq l(\overline{v}) \in G_1, z \doteq l(\overline{w}) \in G_2, l[1, l[1, l \in R(top(y, G_1), top(z, G_2))\), \(T\) does not contain an AUT of the form \(\_ : y \doteq z\), and \(u\) is a fresh recursion term variable. If \(\overline{v} = \overline{u} = \epsilon\) then \(t = l\) and \(A_0 = 0\), otherwise \(t = l(X)\) and \(A_0 = \{X : \overline{v} \doteq \overline{u}\}\) where \(X\) is a fresh hedge variable.

Dec-S: Decomposition and Solving

\(\{X : \overline{v} \doteq \overline{u}\} \cup A; S; T; G \implies\)

\(A \cup \{y_k : \overline{v}|_{l_k} \doteq \overline{u}|_{l_k} | 1 \leq k \leq n\};\)

\(S \cup \{y_0 : \overline{v}|_{l_0} \doteq \overline{u}|_{l_0}\} \cup \{y_k : \overline{v}|_{l_k} \doteq \overline{u}|_{l_k} | 1 \leq k \leq n-1\} \cup \{y_n : \overline{v}|_{l_n} \doteq \overline{u}|_{l_n}\};\)

\(T; G | \sigma \cup \{Z_0 \doteq Y_0, \ldots, Z_n \doteq Y_n\},\)

if \(R(top(\overline{v}, G_1), top(\overline{u}, G_2))\) contains a sequence \(l_1[l_1, j_1] \cdots l_n[l_n, j_n], n > 0\). The \(y\)'s are fresh term variables, the \(Y\)'s are fresh hedge variables, the \(Z\)'s are fresh recursion hedge variables, and the substitution is \(\sigma = \{X \mapsto \{Z_0, y_1, Z_1, \ldots, Z_{n-1}, y_n, Z_n\}\}\). For each \(1 \leq i \leq n\), if the new AUT has the form \(Y_i : \epsilon \doteq \epsilon\), then it is not added to \(S\) and \(Z_i\) does not appear in \(\sigma\).
Term-Graph Anti-Unification

Solve: Solving
\{χ: \tilde{v} \triangleq G\} \cup A; S; T; G \implies A; S \cup \{χ: \tilde{v} \triangleq \tilde{u}\}; T; G[χ \mapsto \omega] \cup \{\omega \equiv \chi\}.

if \mathcal{R}(\text{top}(\tilde{v}, G_1), \text{top}(\tilde{u}, G_2)) = \emptyset or \mathcal{R}(\text{top}(\tilde{v}, G_1), \text{top}(\tilde{u}, G_2)) = \{\epsilon\}. The variable \omega is a fresh recursion variable. If \chi \in V_1, then \omega \in V_1 and if \chi \in V_2, then \omega \in V_2.

Share: Sharing
\{x: y \triangleq z\} \cup A; S; \{u: y \triangleq z\} \cup T; G \implies A; S; \{u: y \triangleq z\} \cup T; G(x \mapsto u).

Merging Nodes in the Store
\emptyset; \{X_1: \tilde{v} \triangleq \tilde{u}, X_2: \tilde{v} \triangleq \tilde{u}\} \cup S; T; \{\omega_1 \equiv X_1, \omega_2 \equiv X_2\} \cup G \implies
\emptyset; S \cup \{X_1: \tilde{v} \triangleq \tilde{u}\}; T; G(\omega_2 \mapsto \omega_1) \cup \{\omega_1 \equiv X_1\},

where \chi_1, \chi_2 \in V_1 \cup V_2 such that if \chi_1 \in V_3, then \chi_2 \notin V_3.

The rules never generate the AUTs of the form \(X: \epsilon \triangleq \epsilon\). To compute \mathcal{R}\text{-generalizations of } G_1 and G_2, we start with \{x: \text{root}(G_1) \triangleq \text{root}(G_2)\}, \emptyset, \emptyset, \{x \equiv x\} and apply the rules on the selected AUTs in all possible ways. The obtained procedure is denoted by Gen(\mathcal{R}).

The notation \(\implies^*\) abbreviates finite (possibly empty) sequence of rule applications. If we want to make it clear which rule is used to transform a configuration, we will write the rule name as the index at the arrow like, e.g., \(A: S; T; G \implies^* A'; S': T'; G'\) for the transformation with the rule Simplification Step.

Example 15. Let \mathcal{R} be the longest common subsequence. Then the term-graphs \(G_1\) and \(G_2\) below have a unique \mathcal{R}\text{-}lpg \(G\):

\[
G_1 = \{x_0 \equiv f(x_1, x_2, x_3, x_0, x_3, x_2, x_3), x_1 \equiv g(x_1, x_2), x_2 \equiv b, x_3 \equiv a\}.
\]

\[
G_2 = \{y_0 \equiv f(y_1, y_0, y_3), y_1 \equiv g(y_1, y_2), y_2 \equiv b, y_3 \equiv a\}.
\]

\[
G = \{z_0 \equiv f(z_1, Z_1, z_0, z_3, Z_1), z_1 \equiv g(z_1, z_2), Z \equiv U, z_3 \equiv a, z_2 \equiv b\}.
\]

Graphically:

The algorithm Gen(\mathcal{R}) computes \(G\), e.g., in the following way:

\[
\{u_0: x_0 \equiv y_0\}; \emptyset; \emptyset; \{z_0 \equiv u_0\} \implies^* \{U_0: (x_1, x_2, x_3, x_0, x_3, x_2, x_3) \equiv (y_1, y_0, y_3)\}; \emptyset; \{z_0: x_0 \equiv y_0\};
\]

\[
\{z_0 \equiv f(Z_0), Z_0 \equiv U_0\} \implies^\text{Dec-S}
\]

(Choosing the common subsequence: \(g[1,1][f[4,2][a[5,3]]\), corresponding to the node pairs \(x_1\) and \(y_1\), \(x_0\) and \(y_0\), the second occurrence of \(x_3\) and \(y_3\).

\[
\{u_1: x_1 \equiv y_1, u_2: x_0 \equiv y_0, u_3: x_2 \equiv y_3\}; \{U_1: (x_2, x_3) \equiv \epsilon, U_2: (x_2, x_3) \equiv \epsilon\};
\]

\[
\{z_0: x_0 \equiv y_0\};
\]

\[
\{z_0 \equiv f(u_1, Z_1, u_2, u_3, Z_2), u_1 \equiv u_1, Z_1 \equiv U_1, u_2 \equiv u_2, u_3 \equiv u_3, Z_2 \equiv U_2\} \implies^\text{Step}
\]

\[
\{u_2: x_0 \equiv y_0, u_3: x_1 \equiv y_3, U_3: (x_1, x_2) \equiv (y_1, y_2)\};
\]

\[
\{U_1: (x_2, x_3) \equiv \epsilon, U_2: (x_2, x_3) \equiv \epsilon\}; \{z_0: x_0 \equiv y_0, z_1: x_1 \equiv y_1\};
\]
\[
\begin{align*}
\{z_0 &\triangleq f(z_1, Z_1, u_2, u_3, Z_2), \\
z_1 &\triangleq g(Z_3), Z_1 \triangleq U_1, u_2 \triangleq u_2, u_3 \triangleq u_3, Z_2 \triangleq U_2, Z_3 \triangleq U_3\} \implies \text{Share} \\
\{u_3 : x_3 \triangleq y_3, U_3 : (x_1, x_2) \triangleq (y_1, y_2)\}; \\
\{U_1 : (x_2, x_3) \triangleq \epsilon, U_2 : (x_2, x_3) \triangleq \epsilon\}; \\
\{z_0 : x_0 \triangleq y_0, z_1 : x_1 \triangleq y_1\}; \\
\{z_0 \triangleq f(z_1, Z_1, z_0, z_3, Z_2), \\
z_1 \triangleq g(Z_3), Z_1 \triangleq U_1, u_3 \triangleq u_3, Z_2 \triangleq U_2, Z_3 \triangleq U_3\} \implies \text{Step} \\
\{U_3 : (x_1, x_2) \triangleq (y_1, y_2)\}; \\
\{U_1 : (x_2, x_3) \triangleq \epsilon, U_2 : (x_2, x_3) \triangleq \epsilon\}; \\
\{z_0 : x_0 \triangleq y_0, z_1 : x_1 \triangleq y_1, z_3 : x_3 \triangleq y_3\}; \\
\{z_0 \triangleq f(z_1, Z_1, z_0, z_3, Z_2), \\
z_1 \triangleq g(u_4, u_5), Z_1 \triangleq U_1, z_3 \triangleq a, Z_2 \triangleq U_2, u_4 \triangleq u_4, u_5 \triangleq u_5\} \implies \text{Share} \\
\{u_5 : x_2 \triangleq y_2\}; \\
\{U_1 : (x_2, x_3) \triangleq \epsilon, U_2 : (x_2, x_3) \triangleq \epsilon\}; \\
\{z_0 : x_0 \triangleq y_0, z_1 : x_1 \triangleq y_1, z_3 : x_3 \triangleq y_3\}; \\
\{z_0 \triangleq f(z_1, Z_1, z_0, z_1, Z_2), \\
z_1 \triangleq g(z_1, u_5), Z_1 \triangleq U_1, z_3 \triangleq a, Z_2 \triangleq U_2, u_5 \triangleq u_5\} \implies \text{Step} \\
\emptyset; \\
\{U_1 : (x_2, x_3) \triangleq \epsilon\}; \\
\{z_0 : x_0 \triangleq y_0, z_1 : x_1 \triangleq y_1, z_3 : x_3 \triangleq y_3, z_2 : x_2 \triangleq y_2\}; \\
\{z_0 \triangleq f(z_1, Z_1, z_0, z_1, Z_2), \\
z_1 \triangleq g(z_1, Z_2), Z_1 \triangleq U_1, z_3 \triangleq a, Z_2 \triangleq U_2, z_2 \triangleq b\} \implies \text{Merge} \\
\emptyset; \\
\{U_1 : (x_2, x_3) \triangleq \epsilon\}; \\
\{z_0 : x_0 \triangleq y_0, z_1 : x_1 \triangleq y_1, z_3 : x_3 \triangleq y_3, z_2 : x_2 \triangleq y_2\}; \\
\{z_0 \triangleq f(z_1, Z_1, z_0, z_1, Z_1), z_1 \triangleq g(z_1, z_2), Z_1 \triangleq U_1, z_3 \triangleq a, z_2 \triangleq b\}.
\end{align*}
\]

The obtained generalization is equal to \(G\) modulo renaming variables. The store and the trail suggest how to obtain the original term-graphs from the computed generalization. For instance, to obtain \(G_1\) from \(G\), we just apply the substitution \(\{U_1 \mapsto (x_2, x_3)\}\) to \(G\). In the obtained term-graph we will have \(x_2 \triangleq b\) and \(x_3 \triangleq a\) alongside to \(z_2 \triangleq b\) and \(z_3 \triangleq a\), but it will be bisimilar to \(G_1\).

**Example 16.** Let \(G_1 = \{x_0 \triangleq f(x_1, x_2), x_1 \triangleq g(x_0, x_3), x_2 \triangleq a, x_3 \triangleq b\}\) and \(G_2 = \{y_0 \triangleq f(y_1, y_2), y_1 \triangleq h(y_0, y_3), y_2 \triangleq a, y_3 \triangleq b\}\). Then the algorithm ends with \(\emptyset, \{z : x_1 \triangleq y_1\}, \{z_0 : x_0 \triangleq y_0, z_1 : x_2 \triangleq y_2\}, \{z_0 \triangleq f(z_1, z_2), z_1 \triangleq z, z_2 \triangleq a\}\). Obtaining \(G_1\) from the computed generalization can be illustrated as \(\{z_0 \triangleq f(z_1, z_2), z_1 \triangleq z, z_2 \triangleq a\}\) \(\{z \mapsto x_1\} = \{z_0 \triangleq f(x_1, z_2), z_2 \triangleq a, x_1 \triangleq g(x_0, x_3), x_0 \triangleq f(x_1, x_2), x_2 \triangleq a, x_3 \triangleq b\} \sim G_1\).

**Theorem 17 (Termination).** The procedure \(\text{Gen}(R)\) terminates on any input and produces a configuration \(\emptyset; S; T; G\), where \(S\) is irreducible with respect to the merging rule.

**Proof.** Let the size of a hedge, \(\text{size}(\tilde{h})\), be the number of symbols in it. the size of an AUT \(x : t_1 \triangleq t_2\) be \(\text{size}(t_1) + \text{size}(t_2) + 1\), and the size of \(X : \tilde{s}_1 \triangleq \tilde{s}_2\) be \(\text{size}(\tilde{s}_1) + \text{size}(\tilde{s}_2) + 2\).
The size of a set of AUTs is the multiset of the sizes of its elements. Then the only rule that increases the size of $A$ is Step. However, this step can be applied only finitely many times, since each time it strictly decreases the number of unvisited node pairs $(x_1, x_2)$, where $x_1 \in G_1$ and $x_2 \in G_2$. Any other rule strictly decreases the size of $A$ or, in case of Merge, the size of $S$. Moreover, Merge does not change the size of $A$. The rule Dec-$S$ can introduce only finite branching. Therefore, the algorithm terminates.

**Definition 18.** Given a set $A$ of AUTs where all the generalization variables are pairwise distinct. We define two substitutions that can be obtained from $A$:

$$
\sigma_L(A) := \{ x \mapsto \tilde{v} \mid x : \tilde{v} \triangleq \tilde{u} \in A \} \quad \sigma_R(A) := \{ x \mapsto \tilde{v} \mid x : \tilde{v} \triangleq \tilde{u} \in A \}
$$

**Lemma 19 (Transformation Invariant).** Let $G_1$, $G_2$ be the two term graphs to be generalized and let $A; S; T; G$ be a configuration such that all the generalization variables from $A, S, T$ are unique among all the other variables from $A, S, T$, including those occurring in graphs or hedges. Furthermore, let $G \sigma_L(T) \sigma_L(S) \sigma_L(A) = G_1$ and $G \sigma_R(T) \sigma_R(S) \sigma_R(A) = G_2$, and let $G$ be a rigid generalization of $G_1$ and $G_2$, where $i \in \{1, 2\}$.

If $A; S; T; G \Rightarrow A'; S'; T'; G'$ is a transformation step applying one of the defined rules then all the generalization variables from $A', S', T'$ are unique among all the other variables from $A', S', T'$.

Moreover, $G' \sigma_L(T') \sigma_L(S') \sigma_L(A') = G_1$ and $G' \sigma_R(T') \sigma_R(S') \sigma_R(A') = G_2$, and $G'$ is a rigid generalization of $G'$ and $G_2$, where $i \in \{1, 2\}$.

**Proof.** We prove that each rule preserves those properties. We can omit the proof for $G' \sigma_L(T') \sigma_R(S') \sigma_R(A') = G_2$, since it is equivalent to proving $G' \sigma_L(T') \sigma_L(S') \sigma_L(A') = G_1$.

For the same reason, we omit the proof that $G'$ is a rigid generalization of $G'$ and $G_2$.

In Step we have two cases, namely (i) $\tilde{v} = \tilde{v} = \epsilon$, and (ii) $\tilde{v} \neq \epsilon$ or $\tilde{v} \neq \epsilon$. We only illustrate the more general case (ii) since the two proofs are largely identical. Therefore, we have $A = \{ x : y \triangleq z \} \cup (A' \setminus \{ X : \tilde{v} \triangleq \tilde{u} \})$, $S = S'$, $T \cup \{ u : y \triangleq z \} = T'$, and $G\{ x \mapsto u \} \cup \{ u : l(X) \} = G'$, where $y = l(\tilde{v}) \in G_1$, $z \triangleq l(\tilde{u}) \in G_2$ and $u, X$ are fresh. Since $u, X$ are fresh, all the generalization variables from $A', S', T'$ are still unique among all the other variables from $A', S', T'$. Obviously, $G \sigma_L(T) \sigma_L(S) \sigma_L(A) = G \sigma_L(T) \sigma_L(S) \sigma_L(A') = G_1$.

Since $G'$ can't lead to consecutive hedge variables and $l[1, 1] \in R(\text{top}(y, G_1), \text{top}(z, G_2))$, it follows that $G'$ is a rigid generalization of $G'$ and $G_1$.

Now we analyze Dec-$S$, which is a bit more involved. We have $A = \{ X : \tilde{v} \triangleq \tilde{u} \} \cup (A' \setminus \{ y_k : \tilde{v}_{i_k} \triangleq \tilde{u}_{j_k} \mid 1 \leq k \leq n \})$, $S \cup \{ Y_0 : \tilde{v}_{i_0} \triangleq \tilde{u}_{i_0} \} \cup \{ Y_k : \tilde{v}_{i_k}^{(k+1)} \triangleq \tilde{v}_{j_k}^{(k+1)} \mid 1 \leq k \leq n - 1 \} \cup \{ Y_n : \tilde{v}_{i_n}^{(n+1)} \triangleq \tilde{u}_{j_n}^{(n+1)} \} = S'$, $T = T'$, and $G \sigma \cup \{ Z_0 = Y_0, \ldots, Z_n = Y_n \} = G'$, where $R(\text{top}(\tilde{v}, G_1), \text{top}(\tilde{u}, G_2))$ contains a sequence $l_{i_1, j_1} \cdots l_{i_n, j_n}$, $n > 0$, the $y$'s, $Y$'s, and $Z$'s are fresh, and $\sigma = \{ X \mapsto (Z_0, Y_1, Z_1, \ldots, Z_{n-1}, Y_n, Z_n) \}$. Since all the variables introduced by the transformation are fresh, all the generalization variables from $A', S', T'$ are still unique. We get $G \sigma_L(T) \sigma_L(S) \sigma_L(A) = G \sigma_L(T) \sigma_L(S') \sigma_L(A') \Rightarrow (\{ Y_0 : \tilde{v}_{i_0} \triangleq \tilde{u}_{i_0} \}) \cup \{ Y_k : \tilde{v}_{i_k}^{(k+1)} \triangleq \tilde{u}_{j_k}^{(k+1)} \mid 1 \leq k \leq n - 1 \} \cup \{ Y_n : \tilde{v}_{i_n}^{(n+1)} \triangleq \tilde{u}_{j_n}^{(n+1)} \}) = G_1$. By uniqueness of $X$, follows $G_1 = G \{ X \mapsto \tilde{v} \} \sigma_L(T) \sigma_L(S') \sigma_L(A') \Rightarrow (\{ Y_0 : \tilde{v}_{i_0} \triangleq \tilde{u}_{i_0} \}) \cup \{ Y_k : \tilde{v}_{i_k}^{(k+1)} \triangleq \tilde{u}_{j_k}^{(k+1)} \mid 1 \leq k \leq n \}).$
The assumptions of Lemma 19 hold for the initial configuration $\{x : \hat{\nu}(G_1) \triangleq \hat{\nu}(G_2)\}; \emptyset; \{x \equiv x\}$ is a derivation in $Gen(\mathcal{R})$, then $G$ is a derivation in $Gen(\mathcal{R})$, then $G$ is an $\mathcal{R}$-generalization of $G_1$ and $G_2$.

**Proof.** The assumptions of Lemma 19 hold for the initialization $\{x : \hat{\nu}(G_1) \triangleq \hat{\nu}(G_2)\}; \emptyset; \{x \equiv x\}$. Since $Gen(\mathcal{R})$ terminates on any input (Theorem 17), it follows that all the generalization variables from $S$ and $T$ are unique among all the other variables from $S$ and $T$. Moreover, $G\sigma_L(T)\sigma_L(S) = G_1$ and $G\sigma_R(T)\sigma_R(S) = G_2$, and $G$ is a rigid generalization of $G_1$ and $G_2$. To prove that $G$ is an $\mathcal{R}$-generalization, it remains to show that the recursion from Definition 13 item 2 has been applied exhaustively. This follows from the fact that the store is complete, i.e., $G\sigma_L(T)\sigma_L(S) = G_1$ and $G\sigma_R(T)\sigma_R(S) = G_2$, and from the condition of the rule Solve that $R(top(\bar{v}, G_1), top(\bar{v}, G_2))$ is either $\emptyset$ or $\{x\}$.

**Theorem 21 (Soundness of the Store).** If $\{x : \hat{\nu}(G_1) \triangleq \hat{\nu}(G_2)\}; \emptyset; \{x \equiv x\}$ is a derivation in $Gen(\mathcal{R})$, then $G\sigma_L(T)\sigma_L(S) = G_1$ and $G\sigma_R(T)\sigma_R(S) = G_2$.

Notice that $Gen(\mathcal{R})$ computes generalizations that do not have free term variables. Therefore, they are not considered in the completeness theorem. However, we show in [12] that this restriction can be lifted by adding an additional transformation rule.

**Theorem 22 (Completeness).** Let $G$ be an $\mathcal{R}$-generalization of $G_1$ and $G_2$. Then $Gen(\mathcal{R})$ computes an $\mathcal{R}$-generalization $G'$ of $G_1$ and $G_2$ such that $G \preceq G'$. 
The computed generalization


Example 24. Let \( G_1 = \{ x_0 \triangleright f(x_1), x_1 \triangleright f(x_2), x_2 \triangleright f(x_3), x_3 \triangleright f(x_4), x_4 \triangleright f(x_5), x_5 \triangleright f(x_6) \} \) and \( G_2 = \{ y_0 \triangleright f(y_1), y_1 \triangleright f(y_2), y_2 \triangleright f(y_3), y_3 \triangleright f(y_4), y_4 \triangleright f(y_5), y_5 \triangleright f(y_6), y_6 \triangleright f(y_7), y_7 \triangleright f(y_8) \} \). They are bisimilar. The algorithm computes their lgg \( G = \{ z_0 \triangleright f(z_1), z_1 \triangleright f(z_2), z_2 \triangleright f(z_3), z_3 \triangleright f(z_4), z_4 \triangleright f(z_5), z_5 \triangleright f(z_6), z_6 \triangleright f(z_7), z_7 \triangleright f(z_8), z_8 \triangleright f(z_9) \} \). It is the join in the lattice of the bisimilarity class of \( G_1 \) and \( G_2 \) [4].
6 Conclusion

We have presented an anti-unification algorithm for (unranked) term-graphs, which are given as systems of recursion equations. The algorithm is sound, complete, and terminating, and uses a parameter, called rigidity function. The function selects common edges outgoing from the pair of nodes to be generalized. While longest common subsequence is the most intuitive instance of the rigidity function, the properties of the algorithm hold for any concrete rigid instance of the parameter. As a future work, extending simply typed lambda term anti-unification [11] to cyclic lambda terms [5] would provide a generalization of our results from a first-order language to a higher-order one.

References


2 María Alpuente, Santiago Escobar, Javier Espert, and José Meseguer. ACUOS: A system for modular ACU generalization with subtyping and inheritance. In Fermé and Leite [17], pages 573–581. doi:10.1007/978-3-319-11558-0.


