Maximum Common Subelement Metrics and its Applications to Graphs

October 31, 2011

Abstract

In this paper we characterize a mathematical model called *Maximum Common Subelement (MCS) Model* and prove the existence of four different metrics on such model. We generalize metrics on graphs previously proposed in the literature and identify new ones by showing three different examples of MCS Models on graphs based on (1) subgraphs, (2) induced subgraphs and (3) an extended notion of subgraphs. This latter example can be used to model graphs with complex labels (e.g., graphs whose labels are other graphs), and hence to derive metrics on them. Furthermore, we also use (3) to show that graph edit distance, when a metric, is related to a maximum common subelement in a corresponding MCS Model.

1 Introduction

Graphs are a natural model for a number of concepts in many different domains such as molecules in chemistry, interaction networks in social studies and biochemistry, workflow descriptions in scientific computing, just to name a few. In each of these domains, when dealing with collections of such objects, it is usually important to have a precise notion of similarity/dissimilarity between them. An adequate and precise way to define similarity/dissimilarity between graphs is by means of a *metric* on (the set of) graphs.

Bunke and Shearer [2] showed that the function

$$d_B(g_1, g_2) = 1 - \frac{v_{12i}}{\max\{v_1, v_2\}}$$

is a metric on the set of graphs when g_1 and g_2 are graphs with, respectively, v_1 and v_2 vertices, and v_{12i} is the maximum number of vertices of a *common induced subgraph of* g_1 *and* g_2. Later, Wallis et al. [7] showed that, by rearranging the same terms, the function

$$d_W(g_1, g_2) = 1 - \frac{v_{12i}}{v_1 + v_2 - v_{12i}}$$

is also a metric on the set of graphs. We say that these two metrics are *based on induced subgraphs* because the term v_{12i} is related to a common induced subgraph of input graphs g_1 and g_2.

An initial motivation for this work was to identify metrics on the set of graphs *based on subgraphs* instead of induced subgraphs that would be analogous d_B and d_W. Note in Figure 1 that, for the same two graphs, the largest number of vertices of a common subgraph and of a common induced subgraph can be significatively different. This observation leads to the fact that, depending on the application, the graph similarity/dissimilarity notion is better modeled either by a function based on subgraphs or by one based on induced subgraphs. One application where a function based on subgraphs is a better fit is reported by [5]. In their paper, they argue that a common subgraph (not necessarily an induced one) that has the largest number of edges is a better model for the similarity of chemical graphs since, in their words, *"it is the bonded interactions between atoms in a molecule that are the most responsible for its perceived activity"*. In this application for chemical graphs, analogous versions of d_B and d_W based on subgraphs would be more adequate.

One metric on the set of graphs based on subgraphs was shown by Fernández and Valiente [4]. Their function is equivalent to the following definition:

$$d_F(g_1, g_2) = (v_1 + e_1) + (v_2 + e_2) - 2(v_{12s} + e_{12s}),$$

where the new terms e_1 and e_2 are the number of edges of g_1 and g_2, and v_{12s} and e_{12s} are the number of vertices and edges of a common subgraph of g_1 and g_2 that maximizes the sum of number of vertices and number of edges among all subgraphs of g_1 and g_2.

In this paper we characterize a mathematical structure called *Maximum Common Subelement (MCS) Model* (Section 3.2), that generalizes the one described by [3], and show that four metrics are valid in such model (Theorem 1), including general analogous versions of the functions d_B, d_W, and d_F. We then show three examples of MCS Models on graphs. The first two examples are based on the usual notions of subgraphs (Section 4.2) and induced subgraphs (Section 4.3), and the third example is based on a notion of *extended* subgraphs (Section 4.4). We refer to these three MCS Model on graphs as, respectively, *S-MCS Model*, *I-MCS Model*, and *E-MCS Model*. The importance of these MCS Models on graphs is that they enable us to reproduce previous metrics on graphs (e.g., d_B, d_W, d_F), extend them (weighting scheme), and derive new ones (e.g., analogous of d_B and d_W based on subgraphs, the metrics on the E-MCS Model).

One interesting aspect of the E-MCS Model is that the (vertex and edge) labels of its graphs are elements of other MCS Models. This permits an E-MCS Model to describe *rich structured objects* (e.g., graphs whose labels are other graphs) and similarity models on them (i.e., the general MCS Model metrics are readily available for these rich structured objects). In Section 5, we use E-MCS Models to show that for any graph edit distance that is a metric on graphs, we can derive a corresponding MCS Model where the edit distance of two graphs is related to the size of a maximum common subelement of the two graphs in this corresponding MCS Model.

2 Preliminaries

For the sake of completeness, in this section we state some standard concepts that are fundamental for the rest of the paper.

Definition 1. *(*METRIC, METRIC SPACE*) A metric d on a set X is a function $d{:}\,X\times X\to[0,\infty)$ that, for any $x_1, x_2, x_3 \in X$, the following conditions hold*

(M1) $d(x_1, x_1) = 0$;
(M2) $d(x_1, x_2) = d(x_2, x_1)$;
(M3) $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$;
(M4) if $d(x_1, x_2) = 0$ *then* $x_1 = x_2$.

In this case, the pair (X, d) is called a metric space. If X is finite then we also refer to (X, d) as a finite metric space.

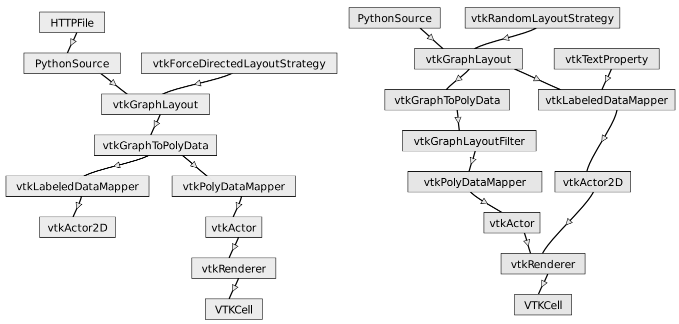
Definition 2. *(*PARTIAL ORDER*) Let $\preccurlyeq$ be a relation on a set X, i.e. $\preccurlyeq$ is a subset of $X \times X$. We use the notation $x_1 \preccurlyeq x_2$ to mean (x_1, x_2) is an element of $\preccurlyeq$. We say $\preccurlyeq$ is a partial order on X if the following conditions hold:*

(R1) $x \preccurlyeq x$ *(reflexivity)*
(R2) $x_1 \preccurlyeq x_2$ *and* $x_2 \preccurlyeq x_3$ *then* $x_1 \preccurlyeq x_3$ *(transitivity)*
(R3) $x_1 \preccurlyeq x_2$ *and* $x_2 \preccurlyeq x_1$ *then* $x_1 = x_2$ *(antisymmetry)*

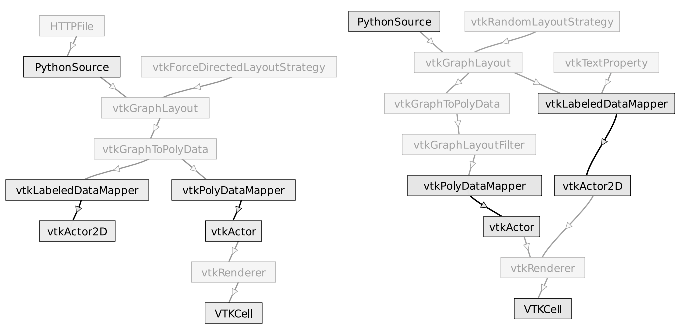
Furthermore, we use the notations $|A|$, $\mathcal{P}(A)$, and $[A]^k$ to mean, respectively, the number of elements in set A, the *power set of* A, and the set of all sets containin $k \geq 1$ elements of A.

3 Maximum Common Subelement (MCS) Model

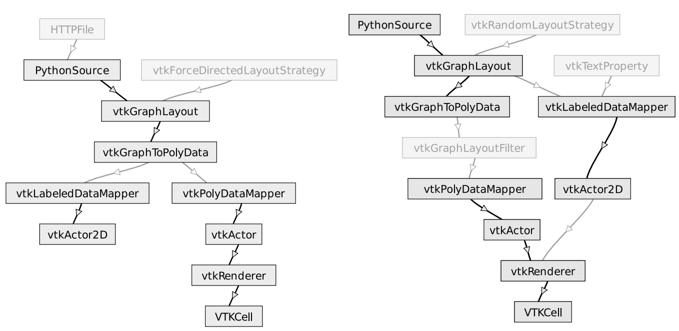
In general, a natural model to the similarity of two objects is given by a number reflecting *how much* do the two objects *overlap*. The *Maximum Common Subelement (MCS) Model* is a precise way of encoding this idea of similarity, framed in a general language that can fit many different scenarios (our focus application in the following sections are graphs). A MCS Model is composed of three parts. The first part is a *set* also called the *domain* of the model. The second part, used to to make the informal notion of *overlap* precise, is a *partial order* on the set or domain of the model. The third and last part, used to quantify *how much is an overlap*, is a *size function* which assigns a *size value* for each element of the domain. The following definitions fix some notation before we formally define a MCS Model.



(a) graphs g_1 (11 vertices) and g_2 (12 vertices)



(b) a common induced subgraph of g_1 and g_2 with maximum number of vertices (6 vertices)



(c) a common subgraph of g_1 and g_2 with maximum number of vertices (9 vertices)

Figure 1: Difference between the maximum number of vertices of a common induced subgraph and of a common subgraph of graphs g_1 and g_2 shown in (a). No common induced subgraph of g_1 and g_2 has more than 6 vertices (b), while there exist a common subgraph with 9 vertices (c). (these graphs represent scientific workflow descriptions generated using [6] (2008))

Definition 3. *(*SUBELEMENT, SUPERELEMENT, COMMON SUBELEMENTS*) Let $\preccurlyeq$ be a partial order on a set X. For $x_1, x_2 \in X$, if $x_1 \preccurlyeq x_2$, we say that x_1 is a subelement of x_2 and that x_2 is a superelement of x_1. We define the function common subelements, denoted by cs, as*

$$cs(X') = \{\, x \in X : x \preccurlyeq y, \forall y \in X'\}, \quad \textit{for } X' \subseteq X.$$

Definition 4. *(*SIZE FUNCTION*) Let $\preccurlyeq$ be a partial order on X. We say a function $s : X \to [0, \infty)$ is a* size function *on $(X, \preccurlyeq)$ if, for $x_1, x_2 \in X$, the following conditions hold*

(S1) if $x_1 \preccurlyeq x_2$ then $s(x_1) \leq s(x_2)$;

(S2) if $x_1 \preccurlyeq x_2$ and $s(x_1) = s(x_2)$ then $x_1 = x_2$.

The size function conditions (S1) and (S2) formalizes the idea that a subelement must have either a smaller size (*proper* subelement), or have the same size and be the same element (a *non-proper* subelement). Now we are ready to define a MCS Model.

Definition 5. *(*MAXIMUM COMMON SUBELEMENT MODEL*) A Maximum Common Subelement (MCS) Model on a set X is a triple*

$$(X, \preccurlyeq, s),$$

where $\preccurlyeq$ is a partial order on X, and s is a size function on $(X, \preccurlyeq)$ such that

(A1) Given $x_1, x_2 \in X$, $cs(\{x_1, x_2\}) \neq \emptyset$ and
$\{s(x) \mid x \in cs(\{x_1, x_2\})\}$ has a maximum;

(A2) Given $x_1, x_2, x \in X$ and $x_1, x_2 \preccurlyeq x$ there exists $x_{12} \in cs(\{x_1, x_2\})$ such that $s(x) \geq s(x_1) + s(x_2) - s(x_{12})$.

Condition (A1) on a MCS Model states that any two elements (not necessarily distinct) have at least one common subelement, and, among all common subelements, there is at least one (could be more than one) whose size is the largest possible. Condition (A2) is rooted on the idea that a superelement of any two elements must, some how, contain these two elements simultaneously, in other words, it contain a kind of *union* of these two elements. Imagine two finite sets S_1 and S_2, intuitively we expect that the number of elements of any superset S of sets S_1 and S_2 to have at least as much elements as their union: $|S| \geq |S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2|$, but never fewer elements than that.

The MCS Model is a generalization of the model proposed by [3], referred here as the *DR Model*. The motivation to define the DR Model in their paper was the same we had to define the MCS Model here: a template to fit applied situations into, and derive metrics. A terminology difference between the DR Model and the MCS Model is that the terms *pattern, generalization, specialization* in the former becomes, respectivelly, *element, subelement, superelement* in the latter. A more important difference is that, in our terminology, while the DR Model requires every two elements to have at least one subelement and one superelement, the MCS Model only requires the subelement to exist. Once the terminology between the two models is aligned, it is straightfoward to prove that MCS Model is in fact a generalization of the DR Model (e.g., *diamond inequality* there is equivalent to condition (A2)). Thus, all the examples given in that paper, namely weighted sets, strings and trees (with appropriate partial order relations and size functions) are also examples of MCS Models. [3] proved one metric function to be valid in any DR Model. In this paper (Theorem 1) we extend this list to four metric functions to be valid in an even more general model: the MCS Model.

When presenting examples of MCS Models in the following sections, instead of showing that property (A1) is valid, we show that the following more restrictive property (A1') is valid:

$$\text{(A1')} \quad 0 < |cs(\{x_1, x_2\})| < \infty;$$

Clearly, (A1') implies (A1), since the number of subelements of any two elements is finite.

Before going into some properties and the metrics of MCS Models we set more terminology

Definition 6. *(*Auxiliar Functions*) If $(X, \preccurlyeq, s)$ is a MCS Model and X' is a subset of X then the maximum common subelements size function, denoted by s', is defined by*

$$s'(X') = \max\{\ s(x) : x \in cs(X')\ \},$$

and the maximum common subelements function, denoted by mcs, is defined by

$$mcs(X') = \{\, x : s(x) = s'(X'), x \in cs(X')\,\}.$$

Note that, in general, s' and mcs might not be well defined (e.g., common subelements of three elements might be empty). By (A1), these functions are well defined when when $|X'| \leq 2$. In the rest of the paper we should use these functions only when they are well defined.

3.1 Some Properties of MCS Models

Proposition 1 (Uniqueness of minimum size element)**.** *Let $(X, \preccurlyeq, s)$ be a MCS Model and let $x_0 \in X$ be such that $s(x_0) = min(\{s(x)|x \in X\})$. Then, the following statements are true:*

(a) if $x \in X$ is such that $s(x) = s(x_0)$, then $x = x_0$.

(b) The element x_0 is a global subelement, i.e., $x_0 \in cs(X)$.

Proof. (a) Let $y \in cs(\{x_0, x\})$. It exists by $(A1)$. By (S1), we conclude that $s(y) = s(x_0) = s(x)$ and using (S2), we conclude that $x = y = x_0$. (b) Let $y \in X$ and, again, let $z \in cs(\{x_0, y\})$. By (S1), we conclude that $s(z) = s(x_0)$, and using (S2), we have that $x_0 = z \Rightarrow x_0 \preccurlyeq y$. $\square$

The following lemma provides a way to derive a MCS Model from a finite metric space. The interesting relation between this metric space and its derived MCS Model is that the metric is somehow preserved in the structure of the MCS Model. Figure 2 presents a finite metric space and a visual illustration of its derived MCS Model. We use this lemma in Section 5 to built a relation between Graph Edit Distance and MCS Models.

Lemma 1 (Metric Space to MCS Model)**.** *Let Σ be a finite set and the function $d : \Sigma \times \Sigma \to [0, \infty)$ be a metric on Σ. In this case, there is a MCS Model*

$$\mathcal{M}_X = (X, \preccurlyeq_X, s_X)$$

where $\Sigma \subseteq X$ and, for $\sigma_1, \sigma_2 \in \Sigma$,

$$d(\sigma_1, \sigma_2) = s_X(\sigma_1) + s_X(\sigma_2) - 2s'_X(\{\sigma_1, \sigma_2\}). \tag{1}$$

Proof. Let $n = |\Sigma|$ and $K_n = (\Sigma, [\Sigma]^2)$ be a complete (unlabeled simple) graph. Assume the natural interpretation: in K_n, an edge $\{\sigma_1, \sigma_2\} \in [\Sigma]^2$ has endpoints $\sigma_1, \sigma_2 \in \Sigma$. Furthermore, let Z be all non empty subsets of edges in K_n that induces a connected subgraph of K_n. We are now able to define the elements of our $\mathcal{M}_X$: the set X, the order relation $\preccurlyeq_X$, and the size function s_X. First, the elements of X are the vertices of K_n plus every subset of edges in K_n that induces a connected subgraph of K_n:

$$X = \Sigma \cup Z.$$

For $x_1, x_2 \in X$, let $x_1 \preccurlyeq_X x_2$ if

$(O1)$ $x_1 = x_2$,

$(O2)$ $x_1 = E \subseteq [\Sigma]^2, x_2 = \sigma \in \Sigma$,
and vertex σ is an endpoint of some edge in E.

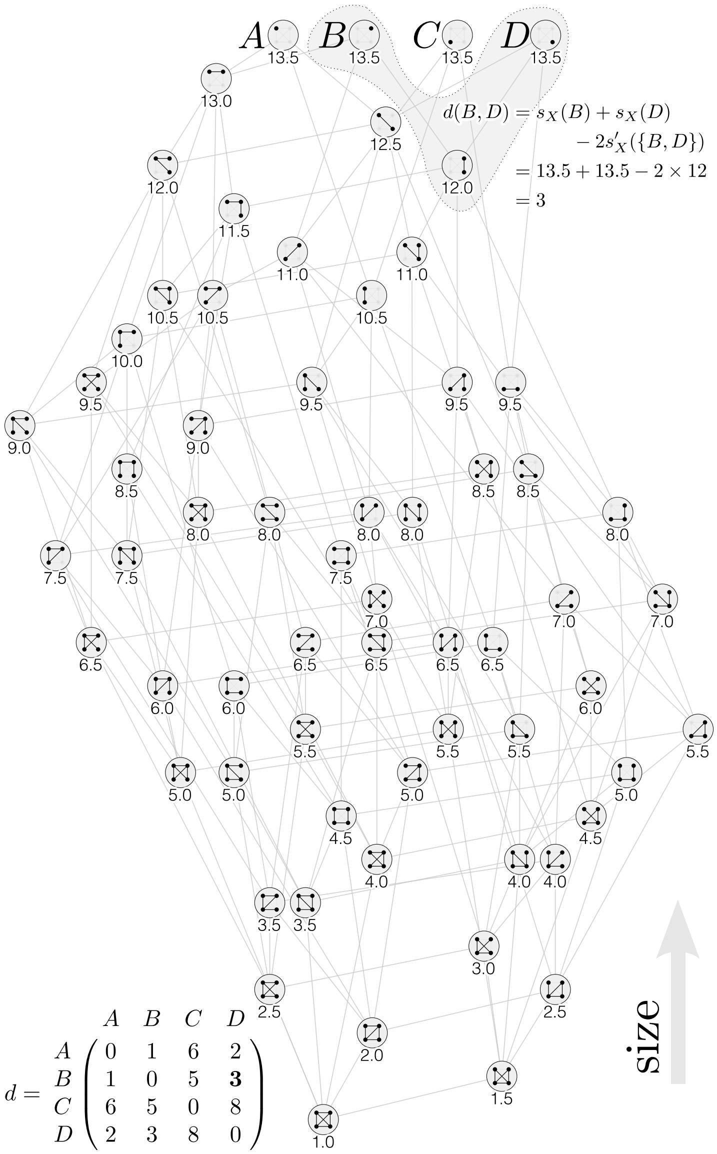
$(O3)$ $x_1, x_2 \subseteq [\Sigma]^2$ and $x_1 \supseteq x_2$.

Define R to be

$$R = \theta + \frac{1}{2} \sum_{\substack{\{\sigma_1, \sigma_2\} \\ \in [\Sigma]^2}} d(\sigma_1, \sigma_2),$$

for some $\theta > 0$. For $x \in X$ define

$$s_X(x) = \begin{cases} R, & \text{if } x \in \Sigma, \\ R - \frac{1}{2} \sum_{\substack{\{\sigma_1, \sigma_2\} \\ \in x}} d(\sigma_1, \sigma_2), & \text{for } x \subseteq [\Sigma]^2., \end{cases}$$



$$d = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 0 & 1 & 6 & 2 \\ 1 & 0 & 5 & \mathbf{3} \\ 6 & 5 & 0 & 8 \\ 2 & 3 & 8 & 0 \end{pmatrix} \end{array}$$

Figure 2: Metric d on the set $\Sigma = \{A, B, C, D\}$ and MCS Model $\mathcal{M}_X = (X, \preccurlyeq_X, s_X)$ on a set X where $\Sigma \subseteq X$. Each element of X is represented by a circle and A, B, C, D are the top four circles. Two elements x_1, x_2 in $\mathcal{M}_X$ are related by $x_1 \preccurlyeq_X x_2$ if there is an upward path from x_1 to x_2. The size function s_X grows bottom-up and its values are shown below each corresponding element. Model $\mathcal{M}_X$ is related to d by the fact that $d(\sigma_1, \sigma_2) = s_X(\sigma_1) + s_X(\sigma_2) - 2s'_X(\{\sigma_1, \sigma_2\})$ for any σ_1, σ_2 in Σ. By Lemma 1 for any other metric (on a finite set) there is a MCS Model satisfying the same properties as in this example.

Note that, with this definition of s_X, θ is the size of $[\Sigma]^2$ (which is an element in Z and is the smallest element in $\mathcal{M}_X$). We now prove that $\mathcal{M}_X = (X, \preccurlyeq_{\mathrm{X}}, s_X)$ is a MCS Model:

- ($\preccurlyeq_{\mathrm{X}}$ IS A PARTIAL ORDER)
  - (REFLEXIVE) By (O1), $\preccurlyeq_{\mathrm{X}}$ is reflexive.
  - (TRANSITIVE) Assume

    $$\text{(H1) } x_1 \preccurlyeq_{\mathrm{X}} x_2 \quad \text{and} \quad \text{(H2) } x_2 \preccurlyeq_{\mathrm{X}} x_3.$$

    If $x_1, x_2, x_3 \in Z$, then, by $(O3)$, $x_3 \supseteq x_2 \supseteq x_1$, therefore, $x_3 \supseteq x_1$ and, again by (O3), $x_1 \preccurlyeq x_3$. Note that, if $x \in \Sigma$, then it is maximal on $\preccurlyeq_{\mathrm{X}}$. Therefore, for $1 \leq i < j \leq 3$, if $x_i \in \Sigma$, then $x_j = x_i$. If $x_1 \in \Sigma$ then $x_1 = x_2 = x_3$ and, by (O1), $x_1 \preccurlyeq x_3$. If $x_1 \in Z$ and $x_2 \in \Sigma$, then $x_2 = x_3$ and (H1) is equivalent to $x_1 \preccurlyeq_{\mathrm{X}} x_3$. If $x_1, x_2 \in Z$ and $x_3 \in \Sigma$, then, by (H1), $x_1 \supseteq x_2$ and, by (H2), x_1 is an endpoint of some edge e in x_2. As edge e is also en edge in x_1 we can conclude $x_1 \preccurlyeq_{\mathrm{X}} x_3$.
  - (ANTISYMMETRIC) Assume

    $$\text{(H3) } x_1 \preccurlyeq_{\mathrm{X}} x_2 \quad \text{and} \quad \text{(H4) } x_2 \preccurlyeq_{\mathrm{X}} x_1.$$

    If $x_1 \in \Sigma$, then, by (H3), we must have $x_1 = x_2$. If $x_1, x_2 \in Z$, then (H3) and (H4) means $x_1 \supseteq x_1$ and $x_2 \supseteq x_2$ therefore $x_1 = x_2$.

- (s_X IS A SIZE FUNCTION) First, by the definitions of R and s_X it is easy to check that $s_X(x) \geq 0$ for all $x \in X$. If $x_1 \preccurlyeq_{\mathrm{X}} x_2$ then three cases can occur:

  $$\text{(C1) } x_1, x_2 \in Z, \ \ \text{(C2) } x_1 \in Z, x_2 \in \Sigma, \ \ \text{(C3) } x_1, x_2 \subseteq \Sigma.$$

  - (S1) We need to show that:

    $$\text{if } x_1 \preccurlyeq_{\mathrm{X}} x_2 \text{ then } s(x_1) \leq s(x_2).$$

    If (C1) occurs then $x_1 \supseteq x_1$ and, by definition, the expression for $s_X(x_1)$ will subtract from R at least the same edge terms $d(\sigma_1, \sigma_2)/2$ as the expression for $s_X(x_1)$, therefore, $s_X(x_1) \leq s_X(x_2)$. If case (2) occurs, x_1 has at least one edge e and the expression for $s_X(x_1)$ subtracts at least one positive value from R (e.g., the value relative to e). Since $x_2 = R$, then $s(x_1) \leq s(x_2)$. If case (3) occurs, then $x_1 = x_2$, therefore $s(x_1) \leq s(x_2)$.
  - (S2) We need to show that:

    $$\text{if } x_1 \preccurlyeq x_2 \text{ and } s(x_1) = s(x_2) \text{ then } x_1 = x_2.$$

    Assume $x_1 \preccurlyeq_{\mathrm{X}} x_2$ and $s(x_1) = s(x_2)$. If (C1) occurs, then $x_1 \supseteq x_2$ and, by definition, the expression for $s_X(x_1)$ subtracts from R at least the same edges as in $s_X(x_2)$. If we assume $x_1 \neq x_2$ then $s_X(x_1)$ would subtract at least one more positive term from R and therefore $s_X(x_1) < s_X(x_2)$, which contradicts the assumption $s_X(x_1) = s_X(x_2)$. Case (C2) cannot occur, because $x_1 \preccurlyeq_X x_2$, $x_1 \in Z$, and $x_2 \in \Sigma$ would imply $s_X(x_1) < s_X(x_2)$ since $s(x_2) = R$ and $s(x_1)$ definition subtracts at least one positive term from R. If (C3) occurs, then necessarily $x_1 = x_2$.

- (A1) We are going that (A1) holds, by showing that (A1') also holds. In order to do so, let $x_1, x_2 \in X$. By definition, the element $x = [\Sigma]^2 \in X$ is a subelement of all other objects in X, therefore $|cs(\{x_1, x_2\})| \geq 1$, for all $x_1, x_2 \in X$. As X is finite, then $|cs(\{x_1, x_2\})|$ must be finite.

- (A2) We have to show that

$$\begin{aligned}&\text{Given } x_1, x_2, x \in X \text{ and } x_1, x_2 \preccurlyeq x\\&\text{there exists } x_{12} \in cs(\{x_1, x_2\})\\&\text{such that } s(x) \geq s(x_1) + s(x_2) - s(x_{12}).\end{aligned}$$

Before going into this axiom, we first note that if $x_1, x_2 \in Z$, $x \in X$ and $x_1, x_2 \preccurlyeq_X x$, then $x_1 \cup x_2$ is also an element of Z. To see this fact, let $\sigma \in \Sigma$ be equal to x, $x \in \Sigma$, or an endpoint of one edge in x if $x \in Z$. Suppose σ_a is a vertex in x_1 and σ_b is a vertex in x_2 (both are in $x_1 \cup x_2$). There must be a path from σ_a to σ in x_1 (which is contained in $x_1 \cup x_2$) and there must be a path from σ_b to σ in x_2 (which is contained in $x_1 \cup x_2$). By joining these paths we have a path from σ_a to σ_b in $x_1 \cup x_2$. Therefore, $x_1 \cup x_2 \in Z$.

We will split this into three cases: (1) $x_1 \in \Sigma$; (2) $x \in Z$; (3) $x \in \Sigma$ and $x_1, x_2 \in Z$. By the symmetric roles that x_1 and x_2 take in this axiom, these cases are enough to cover all possibilities.

  - Case (1): As $x_1 \in \Sigma$, then we must have $x = x_1$ and making $x_{12} = x_2$ we have the axiom, since

$$\begin{aligned}s_X(x) \geq s_X(x) &= s_X(x_1)\\&= s_X(x_1) + s_X(x_2) + s_X(x_2)\\&= s_X(x_1) + s_X(x_2) + s_X(x_{12}).\end{aligned}$$

  - Case (2): As $x \in Z$, then $x_1 \supseteq x$, $x_2 \supseteq x$. Note that $x_1 \cap x_2 \supseteq x$ and x has at least one edge since it is an element in Z. Making $x_{12} = x_1 \cup x_2$ we have the axiom since:

$$\begin{aligned}&s_X(x)s_X(x_1 \cap x_2)\\&\qquad = s_X(x_1) + s_X(x_2) + s_X(x_{12}).\end{aligned}$$

    Note that $x_1 \cap x_2$ might not be a member of Z, but, since the formula for s_X is well defined any subset of $[\Sigma]^2$, we used it.

– Case (3): Note that the inequation

$$s_X(x) \geq s_X(x_1 \cap x_2)$$

is also true in this case. If $x_1 \cap x_2$ is the empty set we have that the only element in common between the graphs induced by x_1 and x_2 is the single vertex x. We can use the same development as in the previous case to estbilish the axiom in this case.

It now remains to show that Equation 1 is valid. The case where $\sigma_1 = \sigma_2$ is trivially true. Suppose $\sigma_1 \neq \sigma_2$. By the definition of $\preccurlyeq_X$ we know that $\{\{\sigma_1, \sigma_2\}\} \preccurlyeq_X \sigma_1$ and $\{\{\sigma_1, \sigma_2\}\} \preccurlyeq_X \sigma_2$. Furthermore, by the definition of s_X,

$$\begin{aligned} s_X(\{\{\sigma_1, \sigma_2\}\}) &= R - \frac{1}{2} d(\sigma_1, \sigma_2) \\ \equiv 2s_X(\{\{\sigma_1, \sigma_2\}\}) &= 2(R) - d(\sigma_1, \sigma_2) \\ \equiv d(\sigma_1, \sigma_2) &= 2(R - 2s_X(\{\{\sigma_1, \sigma_2\}\}) \\ \equiv d(\sigma_1, \sigma_2) &= s_X(x_1) + x_X(x_2) - 2s_X(\{\{\sigma_1, \sigma_2\}\}) \end{aligned}$$

which is in the form of Equation 1. If $\{\{\sigma_1, \sigma_2\}\} \in mcs(\{\sigma_1, \sigma_2\})$ then we have the result. Let's show that this is indeed true. Let x be a memeber of $mcs(\{\sigma_1, \sigma_2\})$. It then satisfies: $x \preccurlyeq_X \sigma_1, \sigma_2$. It also must be in Z since $\sigma_1 \neq \sigma_2$. There must be a path $\sigma_1, \beta_1, \beta_2, \ldots, \beta_k, \sigma_2$ in x otherwise $x \preccurlyeq_X \sigma_1$ and $x \preccurlyeq_X \sigma_2$ would not be true. Actually x must induce a path from σ_1 to σ_2 otherwise we could remove the extra (non-path) edges and still get a set of edges inducing a connected subgraph and with a larger s_X. Furthermore, for paths of the form $\sigma_1, \beta_1, \beta_2, \ldots, \beta_k, \sigma_2$, replacing edges $\{\sigma_1, \beta_1\}$ and $\{\beta_1, \beta_2\}$ by $\{\sigma_1, \beta_2\}$ we still have a path, and, by the fact that d is a metric, we have not increased the size s_X of our x. This way we can erase all intermediate graphs and get that the graph that induces the path σ_1, σ_2 must be a member of $mcs(\{\sigma_1, \sigma_2\})$. With this, the result is estabilished. □

Lemma 2 is a technical property used in Section 3.2 in the proof of Theorem 1.

Lemma 2. *Let* $(X, \preccurlyeq, s)$ *be a MCS Model. The inequality*

$$s'(\{x_1, x_2\}) + s'(\{x_2, x_3\}) \ \leq \ s(x_2) + s'(\{x_1, x_3\}). \tag{2}$$

holds for all $x_1, x_2, x_3 \in X$.

Proof. Let $x_{12} \in mcs(\{x_1, x_2\})$ and $x_{23} \in mcs(\{x_2, x_3\})$. As $x_{12}, x_{23} \preccurlyeq x_2$, we can use axiom (A2) to conclude that there exists $x_{123} \in cs(\{x_{12}, x_{23}\})$ such that $s(x_2) \geq s(x_{12}) + s(x_{23}) - s(x_{123})$. We then can write

$$\begin{aligned} s'(\{x_1, x_2\}) + s'(\{x_2, x_3\}) &= s(x_{12}) + s(x_{23}) \\ &\leq s(x_2) + s(x_{123}) \leq s(x_2) + s'(\{x_1, x_3\}). \end{aligned}$$

□

3.2 Metrics on MCS Models

The main result in this section is the Theorem 1 which states that four different functions are indeed metrics on MCS Models.

Theorem 1 (Metrics on MCS Models)**.** *Let $\mathcal{M} = (X, \preccurlyeq, s)$ be a MCS Model on X and let d_a, d_b, d_c, d_d be*

$$d_a(x_1, x_2) = s(x_1) + s(x_2) - 2s'(\{x_1, x_2\}), \tag{3}$$

$$d_b(x_1, x_2) = \max\{s(x_1), s(x_2)\} - s'(\{x_1, x_2\}), \tag{4}$$

$$d_c(x_1, x_2) = \begin{cases} 0\,, & \text{if } s(x_1) = s(x_2) = 0 \\ 1 - \dfrac{s'(\{x_1, x_2\})}{\max\{s(x_1), s(x_2)\}}\,, & \text{otherwise.} \end{cases} \tag{5}$$

$$d_d(x_1, x_2) = \begin{cases} 0\,, & \text{if } s(x_1) = s(x_2) = 0 \\ 1 - \dfrac{s'(\{x_1, x_2\})}{s(x_1) + s(x_2) - s'(\{x_1, x_2\})}\,, & \text{otherwise.} \end{cases} \tag{6}$$

Then, all of them are metrics on X.

Proof. Since $s'(\{x_1, x_1\}) = s(x_1)$, it is easy to check that (M1) is true for all formulas. Furthermore, as $s'(\{x_1, x_2\}) = s'(\{x_2, x_1\})$, it is also easy to see that (M2) is true for all formulas. Note that for any Maximum Common Subelement Model we have

$$\begin{aligned} x_1 \neq x_2 \;\Rightarrow\; s'(\{x_1, x_2\}) &< \max(\{s(x_1), s(x_2)\}) \\ &\leq s(x_1) + s(x_2) - s'(\{x_1, x_2\}). \end{aligned} \tag{7}$$

It is now easy to see that (M4) is true in all formulas (except for the case $s(x_1) = s(x_2) = 0$ on d_c and d_d) by using its contrapositive form

$$\text{if } x_1 \neq x_2 \text{ then } d(x_1, x_2) \neq 0.$$

Assume $x_1 \neq x_2$ and check that, using Equation 7 above, each of the four distance formulas will result in a positive number. The case when $s(x_1) = s(x_2) = 0$ on the formulas d_c and d_d is also true because, by Proposition 1, there can be only one element with size zero in a MCS Model. The proof of (M3) will be given separately for each formula. For all the following proofs let $x_1, x_2, x_3 \in X$.

- (M3) IS VALID FOR d_a: Using the Lemma 2, we can write

$$\begin{aligned} & 0 \leq s(x_2) + s'(\{x_1, x_3\}) - s'(\{x_1, x_2\}) - s'(\{x_2, x_3\}) \\ \Rightarrow\ & 0 \leq 2s(x_2) + 2s'(\{x_1, x_3\}) - 2s'(\{x_1, x_2\}) - 2s'(\{x_2, x_3\}) \\ \Rightarrow\ & s(x_1) + s(x_3) - 2s'(x_1, x_3) \leq s(x_1) + s(x_2) \\ & \quad - s'(x_1, x_2) + s(x_2) + s(x_3) - 2s'(x_2, x_3) \\ \Rightarrow\ & d_a(x_1, x_3) \leq d_a(x_1, x_2) + d_a(x_2, x_3). \end{aligned}$$

which proves (M3) for d_a.

- (M3) IS VALID FOR d_b: We split this proof in three cases. These are the only cases need to be considered, since the role played by x_1 and x_3 in (M3) are symmetric.

  – (Case 1) If $s(x_2) \leq s(x_1) \leq s(x_3)$ We can write:

$$\begin{aligned} \text{Lemma 2} \Rightarrow\ & 0 \leq s(x_1) + s'(\{x_1, x_3\}) - s'(\{x_1, x_2\}) - s'(\{x_2, x_3\}) \\ \Rightarrow\ & s(x_3) - s'(\{x_1, x_3\}) \leq s(x_1) \\ & \quad - s'(\{x_1, x_2\}) + s(x_3) - s'(\{x_2, x_3\}) \\ \Leftrightarrow\ & \max(\{s(x_1), s(x_3)\}) - s'(\{x_1, x_3\}) \leq \\ & \quad \max(\{s(x_1), s(x_2)\}) - s'(\{x_1, x_2\}) \\ & \quad + \max(\{s(x_2), s(x_3)\}) - s'(\{x_2, x_3\}) \\ \Rightarrow\ & d_b(x_1, x_3) \leq d_b(x_1, x_2) + d_b(x_2, x_3). \end{aligned}$$

  – (Case 2) If $s(x_1) \leq s(x_2) \leq s(x_3)$, adding $s(x_3)$ to both sides of (2) we have

$$\begin{aligned} & s(x_3) - s'(\{x_1, x_3\}) \leq s(x_2) \\ & \quad - s'(\{x_1, x_2\}) + s(x_3) - s'(\{x_2, x_3\}) \\ \Leftrightarrow\ & \max(\{s(x_1), s(x_3)\}) - s'(\{x_1, x_3\}) \leq \\ & \quad \max(\{s(x_1), s(x_2)\}) - s'(\{x_1, x_2\}) \\ & \quad + \max(\{s(x_2), s(x_3)\}) - s'(\{x_2, x_3\}) \\ \Rightarrow\ & d_b(x_1, x_3) \leq d_b(x_1, x_2) + d_b(x_2, x_3). \end{aligned}$$

  – (Case 3) If $s(x_1) \leq s(x_3) \leq s(x_2)$, adding $s(x_3)$ to the left hand side and $s(x_2)$ to the right hand side of (2) we have

$$\begin{aligned} & s(x_3) - s'(\{x_1, x_3\}) \leq s(x_2) \\ & \quad - s'(\{x_1, x_2\}) + s(x_2) - s'(\{x_2, x_3\}) \\ \Leftrightarrow\ & \max(\{s(x_1), s(x_3)\}) - s'(\{x_1, x_3\}) \leq \\ & \quad \max(\{s(x_1), s(x_2)\}) - s'(\{x_1, x_2\}) \\ & \quad + \max(\{s(x_2), s(x_3)\}) - s'(\{x_2, x_3\}) \\ \Rightarrow\ & d_b(x_1, x_3) \leq d_b(x_1, x_2) + d_b(x_2, x_3). \end{aligned}$$

- (M3) IS VALID FOR d_c: The triangle inequality (M3) requires expression $d_c(x_1, x_2) + d_c(x_2, x_3) - d_c(x_1, x_3)$ to be greater than or equal to zero. We split this into three cases:

  - (Case 1) If $s(x_1) = s(x_2) = 0$, then, by Property 1, $x_1 = x_2$ and the triangle inequality becomes $d_c(x_1, x_1) + d_c(x_1, x_3) - d_c(x_1, x_3) \geq 0$ which, by definition of d_c, can be reduced to $d_c(x_1, x_3) \geq d_c(x_1, x_3)$ which is obviously true. An analogous argument can be made to show that (M3) is valid in the cases where $s(x_1) = s(x_3) = 0$ and $s(x_2) = s(x_3) = 0$.

  If (Case 1) doesn't occur, then at least two elements in $\{x_1, x_2, x_3\}$ have size s greater than zero and the definition of d_c we need to use is the bottom one in Equation (5). In this case, the expression for (M3) becomes (8) ≥ 0, where (8) is

$$1 + \frac{s'\{x_1, x_3\}}{M_{13}} - \frac{s'\{x_1, x_2\}}{M_{12}} - \frac{s'\{x_2, x_3\}}{M_{23}} \tag{8}$$

  and M_{ij} is a short name for $\max\{s(x_i), s(x_j)\}$. The remaining cases that are sufficient to prove that (M3) is valid for d_c are:

  - (Case 2) If not (Case 1) and $s(x_2) \geq s(x_1), s(x_3)$ then

$$\begin{aligned}
& s(x_2) \times (8) \\
= \ & s(x_2) \times \left(1 + \frac{s'\{x_1, x_3\}}{M_{13}} - \frac{s'\{x_1, x_2\}}{M_{12}} - \frac{s'\{x_2, x_3\}}{M_{23}}\right) \\
\geq \ & s(x_2) + s'\{x_1, x_3\} - s'\{x_1, x_2\} - s'\{x_2, x_3\} \\
\geq \ & 0, \text{by Lemma 2.}
\end{aligned}$$

    Since $s(x_2) > 0$, this implies that (8) ≥ 0.

  - (Case 3) Similarly, if not (Case 1) and $s(x_1) \geq s(x_2), s(x_3)$ then

$$\begin{aligned}
& s(x_1) \times (8) \\
= \ & s(x_1) \times \left(1 + \frac{s'\{x_1, x_3\}}{M_{13}} - \frac{s'\{x_1, x_2\}}{M_{12}} - \frac{s'\{x_2, x_3\}}{M_{23}}\right) \\
= \ & s(x_1)\left(1 - \frac{s'\{x_2, x_3\}}{M_{23}}\right) + s'\{x_1, x_3\} - s'\{x_1, x_2\} \\
\geq \ & s(x_2)\left(1 - \frac{s'\{x_2, x_3\}}{M_{23}}\right) + s'\{x_1, x_3\} - s'\{x_1, x_2\} \\
= \ & s(x_2) - \frac{s(x_2)s'\{x_2, x_3\}}{M_{23}} + s'\{x_1, x_3\} - s'\{x_1, x_2\} \\
\geq \ & s(x_2) - s'\{x_2, x_3\} + s'\{x_1, x_3\} - s'\{x_1, x_2\} \\
\geq \ & 0, \text{by Lemma 2.}
\end{aligned}$$

Since $s(x_1) > 0$, this implies that $(8) \geq 0$.

- (M3) IS VALID FOR d_d: This proof follow the same lines as the one given for graphs in [7]. The triangle inequality (M3) requires expression $d_d(x_1, x_2) + d_d(x_2, x_3) - d_d(x_1, x_3)$ to be greater than or equal to zero. We split this into three cases:

  - (Case 1) If $s(x_1) = s(x_2) = 0$, then, by Property 1, $x_1 = x_2$ and the triangle inequality becomes $d_d(x_1, x_1) + d_d(x_1, x_3) - d_d(x_1, x_3) \geq 0$ which can be reduced to $d_d(x_1, x_3) \geq d_d(x_1, x_3)$ which is obviously true. An analogous argument can be made to show that (M3) is valid in the cases where $s(x_1) = s(x_3) = 0$ and $s(x_2) = s(x_3) = 0$.

  If (Case 1) doesn't occur, then at least two elements in $\{x_1, x_2, x_3\}$ have size s greater than zero and the definition of d_d we need to use is the bottom one in Equation (5). In this case, the expression for (M3) becomes $(9) \geq 0$, where (9) is

$$1 + \frac{s'\{x_1, x_3\}}{U_{13}} - \frac{s'\{x_1, x_2\}}{U_{12}} - \frac{s'\{x_2, x_3\}}{U_{23}} \tag{9}$$

  and $U_{ij} = s(x_i) + s(x_j) - s'\{x_i, x_j\}$. Let $x_{ij} \in mcs(\{x_i, x_j\})$. By (A2) of a Maximum Common Subelement Model there exists $x_{123} \preccurlyeq x_{12}, x_{23}$ such that $s(x_{123}) \leq s(x_{13}) = s'\{x_1, x_3\}$ and $s(x_{123}) \geq s(x_{12}) + s(x_{23}) - s(x_2)$. In this way we can write

$$\begin{aligned} (9) =&\ 1 + \frac{s(x_{13})}{U_{13}} - \frac{s(x_{12})}{U_{12}} - \frac{s(x_{23})}{U_{23}} \\ \geq&\ 1 + \frac{s(x_{123})}{s(x_1) + s(x_3) - s(x_{123})} - \frac{s(x_{12})}{U_{12}} - \frac{s(x_{23})}{U_{23}}. \end{aligned} \tag{10}$$

  Let non negative numbers $a_1, a_2, a_3, a_{12}, a_{23}, a_{123}$ be defined by $s(x_{123}) = a_{123}$; $s(x_{12}) = a_{12} + a_{123}$; $s(x_{23}) = a_{23} + a_{123}$; $s(x_1) = a_1 + a_{12} + a_{123}$; $s(x_2) = a_2 + a_{12} + a_{23} + a_{123}$; $s(x_3) = a_3 + a_{23} + a_{123}$;. And let $T = a_1 + a_2 + a_3 + a_{12} + a_{23} + a_{123}$. We now can write

$$(10) = 1 + \frac{a_{123}}{T - a_2} - \frac{a_{12} + a_{123}}{T - a_3} - \frac{a_{23} + a_{123}}{T - a_1} \tag{11}$$

  To show that $(11) \geq 0$ it is sufficient to show that (11) times a positive number is greater than or equal to zero. Let $(T - a_1)(T - a_2)(T - a_3)$ be this positive number, since (Case 1) is false.

$$\begin{aligned}
& (11) \times (T-a_1)(T-a_2)(T-a_3) \\
= \; & (T-a_1)(T-a_2)(T-a_3) + a_{123}(T-a_1)(T-a_3) \\
& - (a_{12}+a_{123})(T-a_1)(T-a_2) \\
& - (a_{23}+a_{123})(T-a_2)(T-a_3) \\
= \; & a_1a_2(T-a_3) + T\,(a_1a_3 + a_2a_3 + a_1a_{12} + a_2a_{12} \\
& \quad + a_2a_{123} + a_2a_{23} + a_3a_{23} + a_2a_{123}) \\
& + \; (a_1a_3a_{123} + a_1a_2a_{12} + a_1a_2a_{123} + a_2a_3a_{23} + a_2a_3a_{123}) \\
\geq \; & 0
\end{aligned}$$

The proof of Theorem 1 is complete. □

4 Graph MCS Models

In this section we present three examples of MCS Models on graphs and use Theorem 1 to derive different metrics on graphs for each of these examples. In particular, we are able to reproduce and generalize previous metrics on graphs based on subgraphs and induced subgraphs, and obtain new metrics on graphs based on an *extended* subgraph notion.

4.1 Graphs Terminology

Here is a series of graph related definitions we use in the rest of the paper. We chose undirected simple graphs as our default case, but the results we present in the following sections also work for directed graphs.

Definition 7. *(*Graph*) A graph is a 4-tuple $g = (V, E, \ell_v, \ell_e)$ where*

- *V is a finite set of vertices;*
- *$E \subseteq [V]^2$ is the set of edges;*
- *$\ell_V : V \to \Sigma_V$ is a function that assigns labels to vertices;*
- *$\ell_E : E \to \Sigma_E$ is a function that assigns labels to edges;*

If $V = \emptyset$ then g is called the empty graph.

Definition 8. *(*Subgraph*) A graph $g' = (V', E', \ell'_V, \ell'_E)$ is said to be a subgraph of $g = (V, E, \ell_V, \ell_E)$, if $V' \subseteq V$, $E' \subseteq E \cap [V']^2$, $\ell'_V(v) = \ell_V(v)$ for $v \in V'$, and $\ell'_E(e) = \ell_E(e)$ for $e \in E'$.*

Definition 9. *(*Induced Subgraph*) A graph $g' = (V', E', \ell'_V, \ell'_E)$ is said to be an induced subgraph of $g = (V, E, \ell_V, \ell_E)$ if g' is a subgraph of g and $E' = E \cap [V']^2$.*

Definition 10. *(*Isomorphism*) Let* $g_1 = (V_1, E_1, \ell_{V1}, \ell_{E1})$ *and* $g_2 = (V_2, E_2, \ell_{V2}, \ell_{E2})$ *be graphs. A bijection* $\phi : V_1 \to V_2$ *is an isomorphism between* g_1 *and* g_2 *if the conditions four conditions are valid: (1)* $E_2 = \{\{\phi(u), \phi(v)\} : \{u, v\} \in E_1\}$*; (2)* $\ell_{V1}(v) = \ell_{V2}(\phi(v))$*, for* $v \in V_1$*; (3)* $\ell_{E1}(\{u, v\}) = \ell_{E2}(\{\phi(u), \phi(v)\})$*, for* $\{u, v\} \in E_1$*. If there exists an isomorphism between two graphs we say they are isomorphic.*

Remark 1. *We use the notion* $\phi(e)$*, where* $e = \{u, v\} \in E_1$ *and* $u, v \in V_1$*, to mean the edge* $\{\phi(u), \phi(v)\} \in E_2$*.*

Definition 11. *(*Subgraph Isomorphic*) A graph* g *is subgraph isomorphic to a graph* g'*, denoted by* $g' \subseteq g$*, if there exists a subgraph of* g *that is isomorphic to* g'*.*

Definition 12. *(*Induced Subgraph Isomorphic*) A graph* g *is induced subgraph isomorphic to a graph* g'*, denoted by* $g' \subseteq_i g$*, if there exists an induced subgraph of* g *that is isomorphic to* g'*.*

Definition 13. *(*Graph n-Completion*) Let* $g = (V, E, \ell_V, \ell_E)$ *be a graph with vertex labels in* Σ_V *and edge labels in* Σ_E*. For* $n \geq |V|$*, a special vertex label* ε_V*, and a special edge label* ε_E*, we define the graph* n*-completion of* g *as*

$$\kappa_n^{\varepsilon_V, \varepsilon_E}(g) = (V', E', \ell'_V, \ell'_E)$$

where

- $V' = V \cup \{v_1, \dots, v_{n-|V|}\}$,
- $E' = [V']^2$,
- $\ell'(v) = \begin{cases} \ell(v), & \text{if } v \in V, \\ \varepsilon_V & \text{if } v \in V' \backslash V, \end{cases}$
- $\ell'(e) = \begin{cases} \ell(e), & \text{if } e \in E, \\ \varepsilon_E & \text{if } e \in E' \backslash E. \end{cases}$

When ε_V *and* ε_E *are clear in the context, we will denote the graph* n*-completion of* g *as* $\kappa_n(g)$*.*

4.2 Subgraph MCS Model

The first example of MCS Model on graphs is based on the subgraph relation $\subseteq$ (Definition 11).

Definition 14. *(*S-MCS Model*) A subgraph MCS Model or S-MCS Model is a triple*

$$(G, \subseteq, s_{GVE\alpha}),$$

where

- *G is the set of graphs (Definition 7) with vertex labels in Σ_V, and edge labels in Σ_E. Furthermore, we consider two graphs $g_1, g_2 \in G$ that are isomorphic to be the same graph: $g_1 = g_2$.*
- *$\subseteq$ is the subgraph isomorphic relation on G (Definition 11);*
- *$s_{GVE\alpha} : G \to [0, +\infty)$ is a function based on a label weighting function $\alpha : (\Sigma_V \cup \Sigma_E) \to (0, +\infty)$ and, for $g = (V, E, \ell_V, \ell_E)$, is defined by*

$$s_{GVE\alpha}(g) = \begin{cases} 0, & \text{if } V = \emptyset; \\ \displaystyle\sum_{v \in V} \alpha(\ell_V(v)) + \sum_{e \in E} \alpha(\ell_E(e)), & \text{otherwise.} \end{cases} \tag{12}$$

The following theorem shows a S-MCS Model is indeed a MCS Model.

Theorem 2. *The S-MCS Model is a MCS Model.*

Proof. It can be verified that $\subseteq$ is a partial order on G. Here we are only going to show (S1), (S2), (A1) and (A2).

(S1) Let $g_1 = (V_1, E_1, \ell_{V1}, \ell_{E1})$ and $g_2 = (V_2, E_2, \ell_{V2}, \ell_{E2})$ be graphs in G. If $g_1 \subseteq g_2$ then there is an isomorphism ϕ between g_1 and g_2', a subgraph of g_2. It should be clear that $s(g_1) = s(g_2')$ since for every vertex and edge in g_1 there is a ϕ-corresponding, equally labeled, vertex and edge in g_2' and vice-versa. As the vertices and edges of g_2' are subsets of V_2 and E_2, then $s(g_2') \leq s(g_2)$, since the vertex and edge sums in $s_{GVE\alpha}$ would run over these subsets. From this we can conclude $s_{GVE\alpha}(g_1) \leq s_{GVE\alpha}(g_2)$.

(S2) Consider the same setup as in (S1) above: $g_1 \subseteq g_2$ and g_1 isomorphic to subgraph g_2' of g_2. Add the extra hypothesis that $s_{GVE\alpha}(s_1) = s_{GVE\alpha}(s_2)$. This implies, as $s_{GVE\alpha}(g_1) = s_{GVE\alpha}(g_2')$, that $s_{GVE\alpha}(g_2') = s_{GVE\alpha}(g_2)$ which implies that the vertices and edges of g_2' are exactily V_2 and E_2. In other words, $g_2' = g_2$ and g_1 is isomorphic to g_2 which in our case is the same as $g_1 = g_2$.

(A1) We are going to show that (A1') holds. In fact, the empty graph is a subgraph of any other graph. This implies that, for any pair $g_1, g_2 \in G$, we have $\{\text{"empty graph"}\} \subseteq cs(\{g_1, g_2\})$ and, consequently, $0 < 1 \leq |cs(\{g_1, g_2\})|$. Also, by our definition, graphs have finite number of vertices and edges. This implies that all subgraphs of a graph is also finite (all possible subsets of the vertex and edge sets of a graph are finite). As we are consider isomorphic graphs to be equal, we know $cs(\{g_1, g_2\}) \subseteq cs(\{g_1\}) =$ "subgraphs of $g1$". As the right set in the previous chain is finite implies $cs(\{g_1, g_2\})$ is also finite.

(A2) Let $g_1, g_2 \subseteq h$. Let ϕ_1 be an isomorphim between g_1 and a subgraph $h_1 = (V_{h_1}, E_{h_1}, \ell_{V h_1}, \ell_{E h_1})$ of h and ϕ_2 be an isomorphism between g_2 and a subgraph $h_2 = (V_{h_2}, E_{h_2}, \ell_{V h_2}, \ell_{E h_2})$ of h. Define h_{12} to be another subgraph of h whose vertices and edges are, respectively, $V_{h_1} \cap V_{h_2}$ and $E_{h_1} \cap E_{h_2}$. The vertex and edges labels of h_{12} are chosen to match the ones in h. With this construction of h_{12} it can be verified that $h_{12} \subseteq g_1, g_2$ and that

$$s_{GVE\alpha}(h) \geq s_{GVE\alpha}(g_1) + s_{GVE\alpha}(g_2) - s_{GVE\alpha}(h_{12}).$$

To see this last inequation one should only notice that every vertex and edge of h that was counted twice in the sum $s_{GVE\alpha}(g_1) + s_{GVE\alpha}(g_2)$ is decreased once when we subtract $s_{GVE\alpha}(h_{12})$.

This completes the proof of Theorem 2. □

The Theorem 2 is true, if we use directed graph instead of undirected ones. Hence, we conclude, as a corollary of Theorem 2, that $(G, \subseteq, s_{GVE_{\alpha_1}})$, where α_1 denotes the constant function equal to one, is a MCS model and hence d_F is a metric on G (since, by Theorem 1, d_a is a metric on G), with this we reobtain the result in [4]. Furthermore, again by Theorem 1, d_c and d_d are metrics on G, which shows versions of d_B and d_W based on subgraphs (note that we need to use as size function the sum of the number of edges and the number of vertices).

It is worth noting that the Theorem 2 enables the use of different label weighting functions that makes possible to enconde application domain knowledge in the MCS Model definition and hence in the metrics in Theorem 1.

4.3 Induced Subgraph MCS Model

The second example of MCS Model on graphs is based on the induced subgraph relation $\subseteq_i$ (Definition 12).

Definition 15. *(*I-MCS MODEL*) An induced subgraph MCS Model or I-MCS Model is a triple*

$$(G, \subseteq_i, s_{GV\alpha}),$$

where

- *G is the set graphs (Definition 7) with vertex labels in Σ_V, and edge labels in Σ_E. Furthermore, we consider two graphs $g_1, g_2 \in G$ that are isomorphic to be the same graph: $g_1 = g_2$.*
- *$\subseteq_i$ is the induced subgraph relation on G (Definition 12);*
- *and $s_{GV\alpha} : G \to [0, +\infty)$ is a function based on a label weighting function $\alpha : \Sigma_V \to (0, +\infty)$ and, for $g = (V, E, \ell_V, \ell_E)$, is defined by*

$$s_{GV\alpha}(g) = \begin{cases} 0, & \text{if } V = \emptyset; \\ \displaystyle\sum_{v \in V} \alpha(\ell_V(v)), & \text{otherwise.} \end{cases} \tag{13}$$

The following theorem shows that I-MCS Models are indeed a MCS Model.

Theorem 3. *The I-MCS Model is a MCS Model.*

Proof. It can be verified that $\subseteq_i$ is a partial order on G. The arguments to show that (S1), (S2) and (A1) are valid here are essentially the same as in the proof of Theorem 2 if we replace the terms "subgraph" by "induced subgraph" and $s_{GVE\alpha}$ by $s_{GV\alpha}$. For (A2) it is sufficient to notice that the construction of h_{12} in the other proof replacing "subgraph" by "induced subgraph" yields an induced subgraph of g_1 and g_2 and the same argument used there to show the (A2) inequation was valid with h_{12} can also be used here. $\square$

Again, Theorem 3 also holds if we deal with directed graphs. Thus, we can get as corollary of Theorem 2 the fact that $(G, \subseteq_i, s_{GV_{\alpha_1}})$ is a MCS Model, where α_1 denotes the constant function equal to one. Thus, d_B and d_W are metrics on G (since, by Theorem 1, d_c and d_d are metrics on G), with this we reobtain the results by [2] and [7]. Furthermore, since d_a is a metric on G, we get a version of d_F based on the induced subgraph relation (using as size function the number of vertices). Also, as in the previous case, the use of different label weighting function allows application domain knowledge to be used in the definition of the metrics (similarity notion).

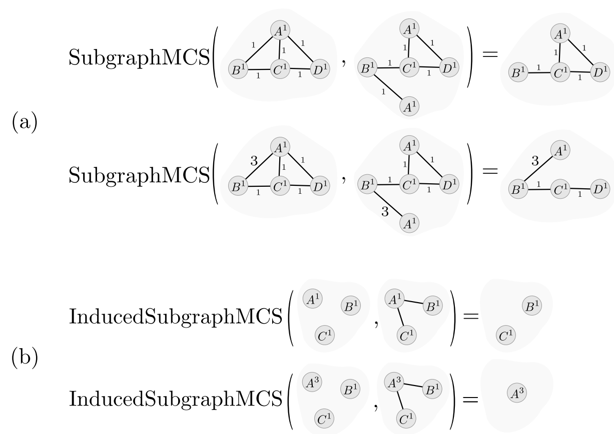


Figure 3: Effect of α (label weighting) on the notion of maximum common subelement of (a) S-MCS Model and (b) I-MCS Model. Vertex labels are letters superscripted with their α values. Edge labels match their α value in (a) and are ommited in (b), since they are not considered by $s_{GV\alpha}$. Note that depending on α, the maximum common subelement changes.

In Figure 3 we illustrate the effect on the notion of maximum common subelement (subgraphs and induced subgraphs) caused by using different label weighting functions α.

4.4 Extended Subgraph MCS Model

The previous two examples of MCS Models on graphs were based on two well known partial orders on graphs: subgraphs and induced subgraphs. We now define a third kind of partial order on graphs that we call *extended subgraphs.* The idea is a simple generalization of the subgraph partial order. Suppose that we fix partial orders on the vertex and edge label sets of a graph. Informally we say that a graph g_1 is an extended subgraph of a graph g_2 with respect to these label partial orders, if we can fit the structure of $g1$ into g_2 in a way that each aligned vertex and edge has a label in g_1 that is a subelement (by the label partial order) of the corresponding aligned element in g_2. This informal idea is defined precisely in the following two definitions.

Definition 16. *(*EXTENDED SUBGRAPH*) Let $g' = (V', E', \ell'_V, \ell'_E)$ and $g = (V, E, \ell_V, \ell_E)$ be graphs with vertex labels in Σ_V and edge labels in Σ_E. If $\preccurlyeq_{\Sigma_V}$ is a parital order on Σ_V and $\preccurlyeq_{\Sigma_E}$ is a partial order on Σ_E, we say g' is an extended subgraph of g with respect to $\preccurlyeq_{\Sigma_V}$ and $\preccurlyeq_{\Sigma_E}$ if*

$$
\begin{aligned}
&V' \subseteq V,\ E' \subseteq E \cap [V']^2, \\
&\ell'_V(v) \preccurlyeq_{\Sigma_V} \ell_V(v),\ \textit{for } v \in V', \\
&\ell'_E(e) \preccurlyeq_{\Sigma_E} \ell_E(e),\ \textit{for } e \in E'.
\end{aligned}
$$

When $\preccurlyeq_{\Sigma_V}$ and $\preccurlyeq_{\Sigma_E}$ are clear in the context we simply say that g' is an extended subgraph of g.

Definition 17. *(*EXTENDED SUBGRAPH ISOMORPHIC*) If g' is isomorphic to a graph that is an extended subgraph of g with respect to $\preccurlyeq_{\Sigma_V}$ and $\preccurlyeq_{\Sigma_E}$, we say that g is extended subgraph isomorphic to g', and denote this fact by $g' \subseteq_e g$.*

We are now able to define the third example of MCS Model on graphs based on the extended subgraph relation $\subseteq_e$ (Definition 17).

Definition 18. *(*E-MCS MODEL*) Let $\mathcal{M}_V$ and $\mathcal{M}_E$ be MCS Modelson Σ_V and Σ_E*

$$
\begin{aligned}
\mathcal{M}_V &= (\Sigma_V, \preccurlyeq_{\Sigma_V}, s_{\Sigma_V}), \\
\mathcal{M}_E &= (\Sigma_E, \preccurlyeq_{\Sigma_E}, s_{\Sigma_E}),
\end{aligned}
$$

with size functions being strictly positive: $s_{\Sigma_V} > 0$ and $s_{\Sigma_E} > 0$. An extended subgraph MCS or E-MCS Model with respect to $\mathcal{M}_V$ and $\mathcal{M}_E$ is a triple

$$(G, \subseteq_e, s_{GES})$$

where,

- *G is the set of graphs (Definition 7) with vertex labels in Σ_V, and edge labels in Σ_E; Furthermore, we consider two graphs $g_1, g_2 \in G$ that are isomorphic to be the same graph: $g_1 = g_2$.*

- $\subseteq_e$ *is the extended subgraph relation on* G *with respect to* $\preccurlyeq_{\Sigma_V}$ *and* $\preccurlyeq_{\Sigma_E}$ *(Definition 17);*

- *and for* $g = (V, E, \ell_V, \ell_E) \in G$,

$$s_{GES}(g) = \begin{cases} 0\,, & \text{if } V = \emptyset; \\ \displaystyle\sum_{v \in V} s_{\Sigma_V}(\ell_V(v)) + \sum_{e \in E} s_{\Sigma_E}(\ell_E(e)), & \text{otherwise.} \end{cases} \tag{14}$$

The following theorem shows that E-MCS Models are indeed MCS Models.

Theorem 4. *The E-MCS Model is a MCS Model.*

Proof. The proof goes as follows:

(S1) Let g_1, g_2 be graphs such that $g_1 \preccurlyeq^e g_2$. By definition of $\preccurlyeq^e$, there exists an extended subgraph g_2' of g_2 that is isomorphic to g_1, and, clearly, $s_{GES}(g_1) = s_{GES}(g_2')$. Since $s_{GES}(g_2')$ is a sum running over a subset of the vertices and edges of $s_{GES}(g_2')$, and each vertex or edge in the sum of $s_{GES}(g_2')$ yields a smaller or equal value than the one in the sum of $s_{GES}(g_2)$, we can conclude that (S1) is valid.

(S2) Let $g_1 \preccurlyeq^e g_2$ and $s(g_1) = s(g_2)$. By definition of $\preccurlyeq^e$, there exists an extended subgraph g_2' of g_2 that is isomorphic to g_1 and, clearly, $s_{GES}(g_1) = s_{GES}(g_2')$. This implies $s_{GES}(g_2') = s_{GES}(g_2)$. As g_2' is a subgraph of g_2 the only option to make $s_{GES}(g_2') = s_{GES}(g_2)$ is to have $g_2 = g_2'$.

(A1) Again, let $g_1 = (V_1, E_1, \ell_{V1}, \ell_{E1})$ and $g_2 = (V_2, E_2, \ell_{V2}, \ell_{E2})$ be two graphs. The fact that the empty graph is a subgraph of any graph implies that $\{\text{empty graph}\} \subseteq cs(\{g_1, g_2\})$ and, consequently, $0 < |cs(\{g_1, g_2\})|$. In order to prove that the set $\{s(g) | g \preccurlyeq^e g_1, g_2\}$ has a maximum, let $g_1^\emptyset = (V_1, E_1)$ and $g_2^\emptyset = (V_2, E_2)$ be the unlabelled copies (same structure) of g_1 and g_2 respectivelly. The set of common subgraphs (by the subgraph isomorphic relation) of $g_1^\emptyset$ and $g_2^\emptyset$ is finite, as in the proof of Theorem 2. Denote this set by $cs(g_1^\emptyset, g_2^\emptyset) = \{h_1, ..., h_n\}$. For each h_i, let $\Phi_i = \{\phi_i, ..., \phi_i^{k_i}\}$ be the set of all subgraph isomorphisms between h_i and g_1. Similarly, let $\Psi_i = \{\psi_i, ..., \psi_i^{l_i}\}$ be the set of all subgraph isomorphisms between h_i and g_2. Now, for each $s = 1, ..., k_i$ and $t = 1, ..., l_i$, the map $\phi_i^s \circ (\psi_i^t)^{-1}$ defines a isomorphism from a subgraph g_1^{st} of $g_1^\emptyset$ and a subgraph g_2^{st} of $g_2^\emptyset$. Finally, denote by g_{st} the extended subgraph of g_1 and g_2 that has the vertex and edge set the same as g_1^{st} and the labels are defined as follows: for each vertex v and edge e of g_{st} define its label as an element of $mcs(\ell_{V1}(v), \ell_{V2}(\phi_i^s \circ (\psi_i^t)^{-1}(v)))$ and $mcs(\ell_{E1}(e), \ell_{E2}(\phi_i^s \circ (\psi_i^t)^{-1}(e)))$, respectively. Now let s_0 and t_0 be such that $s_{GES}(g_{s_0 t_0}) = max(\{s_{GES}(g_{st}) | s = 1, ..., k_i \text{ and } t = 1, ..., l_i\})$. By construction, such $g_{s_0 t_0}$ is the extended subgraph of g_1 and g_2 with maximum size.

(A2) Let $g_1, g_2 \subseteq_e h$. Let ϕ_1 be an isomorphim between g_1 and an extended subgraph $h_1 = (V_{h_1}, E_{h_1}, \ell_{Vh_1}, \ell_{Eh_1})$ of h and ϕ_2 be an extended isomorphism between g_2 and a subgraph $h_2 = (V_{h_2}, E_{h_2}, \ell_{Vh_2}, \ell_{Eh_2})$ of h. Define h_{12} to be another extended subgraph of h whose vertices and edges are, respectively, $V_{h_1} \cap V_{h_2}$ and $E_{h_1} \cap E_{h_2}$. The label of vertex v of h_{12} is defined as follows: let $\sigma_h, \sigma_{h_1}, \sigma_{h_2}$ be the label of v in, respectively, h, h_1, h_2; by the fact that h_1, h_2 are extended subgraphs of h, we have that $\sigma_{h_1}, \sigma_{h_2} \preccurlyeq_{\Sigma_V} \sigma_h$; by axiom (A2) in $\mathcal{M}_V$ there exist $\sigma_{h_{12}} \in \Sigma_V$ such that $s_{\Sigma_V}(\sigma_h) \geq s_{\Sigma_V}(\sigma_{h_1}) + s_{\Sigma_V}(\sigma_{h_2}) - s_{\Sigma_V}(\sigma_{h_{12}})$; define the label of v in h_{12} to be σ_{12}. The label of an edge of h_{12} is defined in an analogous way. With this construction of h_{12} it can be verified that $h_{12} \subseteq g_1, g_2$ and that

$$s_{GES}(h) \geq s_{GES}(g_1) + s_{GES}(g_2) - s_{GES}(h_{12}).$$

The proof of Theorem 4 is complete. □

When modeling real world concepts using graphs, it is usually important to have flexibility when defining what information a vertex or an edge will carry. For example, in scientific workflow descriptions vertices represent parameterized modules that represent some kind of computation. Usually a single module is configured with a set of parameters and values which are not adequately represented by a single symbol, but, instead, by a more complicated object. The *nesting property* of E-MCS Modelsthat enables plugging other MCS Model elements as labels of vertices and edges, and be able to derive metrics for these objects that take into account all parts that form the final object is an interesting one.

To illustrate E-MCS Models, we will use them, in next section, to build a link between Graph Edit Distance and MCS Models. Before that we need an additional property of E-MCS Models that states that if we restrict the elements (graphs) of an E-MCS Model to complete graphs of n vertices, we still have a MCS Model. We will refer to this MCS model as a *n-restricted E-MCS Model.*

Proposition 2. *Let $\mathcal{M} = (G, \subseteq_e, s_{GES})$ be an E-MCS Model with respect to $\mathcal{M}_V = (\Sigma_V, \preccurlyeq_{\Sigma_V}, s_{\Sigma_V})$ and $\mathcal{M}_E = (\Sigma_E, \preccurlyeq_{\Sigma_E}, s_{\Sigma_E})$. Let K_n be the subset of G formed of complete graphs with n vertices. Let $\preccurlyeq_{K_n}$ and s_{K_n} be the restrictions of $\subseteq_e$ and s_{GES} to K_n. In this context, the triple $\mathcal{M}_{K_n} = (K_n, \preccurlyeq_{K_n}, s_{K_n})$ is also a MCS Model.*

Proof. Properties (R1),(R2) and (R3) clearly hold for $\preccurlyeq_{K_n}$, since they are valid for $\subseteq_e$. Similarly, the properties (S1) and (S2) hold for s_{K_n}, since they hold for s_{GES}.

- (A1) Let $g_1 = (V_1, E_1, \ell_{V1}, \ell_{E1}), g_2 = (V_2, E_2, \ell_{V2}, \ell_{E2}) \in K_n$. Let ϕ be a bijection between V_1 and V_2. Then, ϕ defines a one-to-one correspondence between any vertex and edge of g_1 and g_2. Then, we can define a graph $g_{12} \in K_n$, by using the same vertex and edge sets as in g_1 and defining the label for each vertex $v \in V_1$ as an element of $cs(\{\ell_{V1}(v), \ell_{V2}(\phi(v))\})$ and for each edge $e \in E_1$ as an element of $cs(\{\ell_{E1}(e), \ell_{E2}(\phi(e))\})$.

Now let $\{\phi_1, ..., \phi_{n!}\}$ be the set of all bijection between V_1 and V_2. For each bijection ϕ_k, we can define a extended subgraph g_{12}^k of g_1 and g_2 as before, but chosing as labels for each vertex an element of $mcs(\{\ell_{V1}(v), \ell_{V2}(\phi_k(v))\})$ and for each edge an element of $cs(\{\ell_{E1}(e), \ell_{E2}(\phi_k(e))\})$. Let k_0 be the index associated with the largest graph among g_{12}^k, *i.e.*, $s(g_{12}^{k_0}) = max(\{s_{K_n}(g_{12}^k)|k = 1, ..., n!)\})$. By construction, $s(g_{12}^{k_0}) = max(\{s_{K_n}(g)|g \in cs(\{g_1, g_2\})\})$.

- (A2) Let $g_1 = (V_1, E_1, \ell_{V1}, \ell_{E1}), g_2 = (V_2, E_2, \ell_{V2}, \ell_{E1}) \in K_n$ and let also $g = (V, E, \ell_V, \ell_E) \in K_n$ be such that $g_1, g_2 \preccurlyeq_{K_n} g$. We want to define a complete graph $g_{12} = (V_{12}, E_{12}, \ell_{V12}, \ell_{E12})$ such that $g_{12} \preccurlyeq_{K_n} g_1, g_2 \preccurlyeq_{K_n} g$ and also that $s(g) \geq s(g_1) + s(g_2) - s(g_{12})$. In order to do so, we fix one correspondences ϕ_{12} (bijection) between V_{12} and V_1 and other one ϕ_1 between V_1 and V_2. We are going to denote $v \in V_{12}, \phi_{12}(v)$ and $\phi_1(\phi_{12}(v))$, just by v. We define the label of g_{12} as follows: For each $v \in V_{12}$ we know that $\ell_{V1}(v), \ell_{V2}(v) \preccurlyeq \ell_V(v)$, then we can use the axiom (A2) for the MCS Model$(\Sigma_V, \preccurlyeq_{\Sigma_V}, s_{\Sigma_V})$ and conclude that there exists a label α_{12} such that $\alpha_{12} \preccurlyeq_{\Sigma_V} \ell_{V1}(v), \ell_{V2}(v) \preccurlyeq_{\Sigma_V} \ell_V(v)$ and $s_{\Sigma_V}(\ell_V(v)) \geq s_{\Sigma_V}(\ell_{V1}(v)) + s_{\Sigma_V}(\ell_{V2}(v)) - s_{\Sigma_V}(\alpha_{12})$. We define $\ell_{V12}(v) = \alpha_{12}$. With a similar construction, we can define the edge label function ℓ_{E12}. One can verify that g_{12} constructed this way satisfy the axiom (A2).

$\square$

5 Relation between Graph Edit Distance and MCS Models

In this section we show a relation between graph edit distance and MCS Models. Informally speaking, this connection states that if d_{GED} is a graph edit distance and a metric on G, then there is a corresponding MCS Model $\mathcal{M}$ such that

$$d_{\text{GED}}(g_1, g_2) = d_a(\theta(g_1), \theta(g_2)),$$

where $g_1, g_2 \in G$, θ takes the elements of G into their corresponding elements in $\mathcal{M}$, and d_a is the first of the four metrics in Theorem 1 valid in $\mathcal{M}$. Thus, the MCS Model $\mathcal{M}$ encodes d_{GED}. The problem of finding the graph edit distance between g_1 and g_2 becomes the problem of finiding a maximum common subelement between $\theta(g_1)$ and $\theta(g_2)$ in $\mathcal{M}$.

Before stating the main result of this section, we define precisely what we mean by graph edit distance. We use the notion of graph completion in this definition to facilitate the exposition: for any two graphs we can refer to bijections between the vertices of their completed versions instead of having to deal with functions between subsets of the vertices of the first graph into the vertices of the second graph. This definition of graph edit distance is equivalent to the common use of the term, where the cost of each vertex and edge operation (i.e., addition, deletion, and substitution) is based on labels and these operation costs are known a priori.

Definition 19. *(*Graph Edit Distance*) Let g_1 and g_2 be graphs with vertex labels in Σ_V and edge labels in Σ_E. Furthermore, let*

$$c_V : (\Sigma_V \cup \{\varepsilon_V\})^2 \to [0, \infty],$$
$$c_E : (\Sigma_E \cup \{\varepsilon_E\})^2 \to [0, \infty]$$

be, respectivelly, edit cost functions on vertex and edge labels, where ε_V and ε_E are special labels. Assume that

$$g_1' = \kappa^{\varepsilon_V, \varepsilon_E}_{|V_1|+|V_2|}(g_1) = (V_1', E_1', \ell_{V1}', \ell_{E1}'),$$
$$g_2' = \kappa^{\varepsilon_V, \varepsilon_E}_{|V_1|+|V_2|}(g_2) = (V_2', E_2', \ell_{V2}', \ell_{E2}').$$

Let $\mathcal{F}$ be the set of bijections from V_1' to V_2'. The cost $c(f)$ for $f \in \mathcal{F}$ is defined as

$$c(f) = \sum_{v \in V_1'} c_V(\ell_{V1}'(v), \ell_{V2}'(f(v))) \; + \sum_{e \in E_1'} c_E(\ell_{V1}'(e), \ell_{E2}'(f(e))). \tag{15}$$

In this context, we define the graph edit distance between g_1 and g_2 as

$$d_{GED}(g_1, g_2) = \min_{f \in \mathcal{F}} c(f).$$

Some uses of the term graph edit distance refer to a more general idea. For example, [1] shows a correspondence between the maximum number of vertices of a common induced subgraph and a specific graph edit distance notion where the edge operation cost depends on which operation was done in its end vertices.

Now we are able to state the main result of this section.

Theorem 5 (GED and MCS Model)**.** *Let G_n be the set of graphs with n or less vertices on finite label sets Σ_V and Σ_E. Let $c_V : (\Sigma_V \cup \varepsilon_V)^2 \to [0, \infty)$ and $c_E : (\Sigma_E \cup \varepsilon_E)^2 \to [0, \infty)$ be edit cost functions. Furthermore, let c_V and c_E be metrics on $\Sigma_V \cup \{\varepsilon_V\}$ and $\Sigma_E \cup \{\varepsilon_E\}$. Then, there exists a MCS Model*

$$\mathcal{M}_n = (X, \preccurlyeq_X, s_X)$$

and an injective function $\theta : G_n \to X$ such that

$$d_{GED}(g_1, g_2) = s_X(\theta(g_1)) + s_X(\theta(g_2)) - 2s_X'(\{\theta(g_1), \theta(g_2)\}).$$

Proof. Apply Lemma 1 on the finite metric spaces $(\Sigma_V \cup \{\varepsilon_V\}, c_V)$ and $(\Sigma_E \cup \{\varepsilon_E\}, c_E)$ to obtain corresponding MCS Models $\mathcal{M}_V = (\Sigma_V', \preccurlyeq_{\Sigma_V'}, s_{\Sigma_V'})$ and $\mathcal{M}_E = (\Sigma_E', \preccurlyeq_{\Sigma_E'}, s_{\Sigma_E'})$, where the size of the smallest element in these MCS Models are strictly positive. Let G' be the set of graphs with labels in Σ_V' and

Σ'_E. Observe that the triple $(G', \subseteq_e, s_{GES})$ with respect to $\mathcal{M}_V$ and $\mathcal{M}_E$ is an E-MCS Model. Define $X = K_{2n}$ to be the subset of G' consisting only of complete graphs with $2n$ vertices. By Proposition 2 we know that $\mathcal{M}_n = (X, \preccurlyeq_X, s_X)$ is a MCS Model, when $\preccurlyeq_X$ and s_X are restrictions of $\subseteq_e$ and s_{GES} to the set X. Assume $g_1, g_2 \in G_n$ and their vertex sets are, respectively, V_1 and V_2. By definition, the graph edit distance between g_1 and g_2 is the minimum value of function c (Equation 15) for a vertex bijection between $\kappa^{\varepsilon_V, \varepsilon_E}_{|V_1|+|V_2|}(g_1)$ and $\kappa^{\varepsilon_V, \varepsilon_E}_{|V_1|+|V_2|}(g_2)$. It can be checked that for our metric c_V and c_E this minimum value of function c is the same if we consider vertex bijections between $\kappa^{\varepsilon_V, \varepsilon_E}_{2n}(g_1)$ and $\kappa^{\varepsilon_V, \varepsilon_E}_{2n}(g_2)$. Define $\theta : G_n \to X$ to be the graph completion $\kappa^{\varepsilon_V, \varepsilon_E}_{2n}$. Make $x_1 = \theta(g_1) = (V'_1, E'_1, \ell'_{V_1}, \ell'_{E_1})$ and $x_2 = \theta(g_2) = (V'_2, E'_2, \ell'_{V_2}, \ell'_{E_2})$. Let f be a bijection between the vertices of x_1 and x_2. Define $x_f \in X$ in the following way: for every vertex v in x_1 there corresponds a vertex in x_f labeled with an element (any element) of $mcs(\{\ell'_{V_1}(v), \ell'_{V_2}(f(v))\})$, and for every edge e in x_1 there corresponds an edge in x_f labeled with an element (any element) of $mcs(\{\ell'_{E_1}(e), \ell'_{E_2}(f(e))\})$. Using this construction for x_f it is clear that $x_f \preccurlyeq_X x_1, x_2$ and it can be verified that $c(f) = s_X(x_1) + s_X(x_2) - 2s_X(x_f)$. Let f_0 be the bijection between vertices of x_1 and x_2 that yields the graph edit distance between g_1 and g_2. At this point we can write

$$d_{GED}(g_1, g_2) = c(f_0) = s_X(x_1) + s_X(x_2) - 2s_X(x_{f_0}).$$

To conclude the proof it remains showing that $s_X(x_{f_0}) = s'_X(\{x_1, x_2\})$. Assume $s_X(x_{f_0}) < s'_X(\{x_1, x_2\})$ and $x_{12} \in mcs(\{x_1, x_2\})$. Let ϕ_1 and ϕ_2 be isomorphisms between x_{12} and extended subgraphs of x_1 and x_2. Define $f_{12} = \phi_2 \circ \phi_1^{-1}$. Note that f_{12} is a bijection between vertices of x_1 and of x_2 and that $s_X(x_{12}) = s_X(x_{f_{12}})$. In this case, we can write $s_X(x_{f_0}) < s'_X(\{x_1, x_2\}) = s_X(x_{12}) = s_X(x_{f_{12}})$, for bijection f_{12}. This contradicts the hypothesis that $s_X(x_{f_0})$ is the maximum possible for a bijection between vertices of x_1 and x_2. The theorem is proven. □

An interesting aspect of this theorem is that it brings a different and precise materialization for the meaning of a metric graph edit distance between two graphs: we can see it encoded in an element of a corresponding MCS Model. We see as applications of this connection, the interpretation of natural notions in the MCS Model in terms of the original metric graph edit distance. For example, a maximum common subelement of three or more elements of the MCS Model could correspond to a natural generalization of the metric graph edit distance between three or more graphs.

6 Conclusions

In this paper we have introduced MCS Model which is a generalization of a model proposed by [3]. We then showed four metric functions to be valid in any MCS Model (three additional metrics to the one shown in [3]). The usefulness of the MCS Model is that it serves as a template to fit into applied scenarios and ease the derivation of metrics (precise similarity notions) in those scenarios. We show this usage of MCS Models by presenting three examples on graphs: the S-MCS (based on subgraphs), I-MCS (based on induced subgraphs), and E-MCS (based on a less common partial order that we name extended subgraphs). With these examples we are able to reproduce and extend previous reported metrics on graphs [2, 7, 4] as well as new ones (e.g., subgraph versions of d_B and d_W). The E-MCS Model has an interesting nesting property that allows one to derive distance metric for graphs with complex labels, which might be of important value when modeling real scenarios. A final contribution of this paper is an interpretation of the graph edit distance that is a metric on graphs as, essentially, a maximum common subelement on a corresponding MCS Model.

References

[1] Bunke, H., 1997. On a relation between graph edit distance and maximum common subgraph. Pattern Recognition Letters 18, 689–694.

[2] Bunke, H., Shearer, K., 1998. A graph distance metric based on the maximal common subgraph. Pattern Recognition Letters 19, 255–259.

[3] De Raedt, L., Ramon, J., 2009. Deriving distance metrics from generality relations. Pattern Recognition Letters 30, 187–191.

[4] Fernández, M., Valiente, G., 2001. A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recognition Letters 22, 753–758.

[5] Raymond, J., Willett, P., 2002. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. Journal of Computer-Aided Molecular Design 16, 521–533.

[6] VisTrails, 2008. The VisTrails Project. "http://www.vistrails.org".

[7] Wallis, W., Shoubridge, P., Kraetz, M., Ray, D., 2001. Graph distances using graph union. Pattern Recognition Letters 22, 701–704.