International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 3 (2004) 299–313

World Scientific
www.worldscientific.com

METRICS FOR ATTRIBUTED GRAPHS BASED ON THE MAXIMAL SIMILARITY COMMON SUBGRAPH

DŽENA HIDOVIĆ
School of Computer Science, University of Birmingham,
Edgbaston, Birmingham, B15 2TT, UK
D.Hidovic@cs.bham.ac.uk

MARCELLO PELILLO
Dipartimento di Informatica, Università Ca' Foscari di Venezia,
Via Torino 155, 30172 Venezia Mestre, Italy
pelillo@dsi.unive.it

Two distance measures for attributed graphs are presented that are based on the maximal similarity common subgraph of two graphs. They are generalizations of two existing distance measures based on the maximal common subgraph. The new measures are superior to the well-known measures based on elementary edit transformations in that no particular edit operations (together with their costs) need to be defined. Moreover, they can deal not only with structural distortions, but also with perturbations of attributes. It is shown that the new distance measures are metrics.

1. Introduction

Graphs have long been an important tool in the computer vision and pattern recognition fields, especially because of their representational power and flexibility, and recently there has been a renewed and growing interest towards formulating abstract, internal representations of objects and scene in terms of graphs, as well as explicitly formulating computer vision problems as graph problems.[6] As a representational device, graphs are often used to describe objects by representing, for example, their parts by vertices and the relations between them by edges,[1] and once objects are abstracted in terms of graphs, object recognition becomes the problem of matching graphs. Graph matching is therefore a fundamental problem in computer vision and pattern recognition, and a great deal of effort has been devoted over the past decades to devise efficient and robust algorithms for it (see Ref. 3 for an update on recent developments).

A crucial aspect of matching and recognition problems involves determining a suitable similarity measure between "objects". In many applications it is required that such a measure possesses certain properties. In particular, it is often desired that a distance measure d fulfills the following properties:

- $d(A, B) \geq 0$ (nonnegativity)
- $d(A, A) = 0$ (identity)
- $d(A, B) = 0 \iff A = B$ (uniqueness)[a]
- $d(A, B) = d(B, A)$ (symmetry)
- $d(A, B) + d(B, C) \geq d(A, C)$ (triangle inequality).

A distance function satisfying these five properties is called a *metric*. It establishes a partial order over the objects in consideration, and is particulary important for searching and indexing in large databases, or whenever some numerical comparisons between distances have to be done.

A classical approach to comparing graphs is based on the idea of computing their edit-distance, namely, the minimum cost to transform one graph into another by elementary edit operations. This idea is attractive especially when the structures being matched are subject to significant structural distortions. Unfortunately, it turns out that computing the edit-distance on arbitrary graphs is NP-complete, which implies that all exact algorithms have a worst-case time complexity that very likely is exponential in the number of vertices in the graph. Moreover, determining the set of elementary edit operations and the associated costs depends heavily on the application domain and can be problematic (see, e.g. Refs. 8 and 9 for some examples of edit operations motivated by shape matching problems). This choice is in fact crucial as two graphs that are similar under one cost function may be quite dissimilar using another, and the optimal node correspondences may vary considerably.

Recently, another approach has emerged to measuring the distance between graphs, namely substructure-based methods. Within this framework, one looks for a common substructure that satisfies some properties, which typically is maximum cardinality. Specifically, Bunke and Shearer[4] developed a graph distance metric based on maximal common subgraph of two graphs. A variant of their distance is the union-based graph distance introduced by Wallis *et al.*[11] Finally, Fernandez and Valiente[7] measured the distance between graphs by measuring the missing structural information expressed as the difference between minimal common supergraph and maximal common subgraph. The approach can naturally deal with several types of noise and distorsions such as the addition or deletion of nodes in both graphs and is particularly advantageous as it does not require the use of any cost function, thereby avoiding the major drawback of edit-distance-based approaches. It is also worth mentioning that Bunke[5] has shown that on generic graphs, under certain assumptions concerning the edit-costs, determining the maximum common subgraph is equivalent to computing the graph edit-distance.

In many computer vision and pattern recognition applications, however, abstract representations of complex objects and patterns are often endowed with

[a]In the case of graphs, we replace the property of being "equal" with that of being "isomorphic" (see below for a formal definition).

information regarding geometric properties. Hence, the graphs being matched are typically equipped with symbolic and/or numeric attributes, which encode geometric information. All substructure-based measures developed so far deal only with structural distortions, and are therefore not applicable to attributed relational structures. In this paper, we present a generalization of Bunke and Shearer's work to this kind of structures. In an attempt to also take into account the noise and errors in attributes that are very likely to rise in real-world applications, we use the concept of maximal similarity, instead of maximal cardinality, and propose two new attributed graph distance measures based on the maximal similarity common subgraph of two graphs, thereby generalizing previous works by Bunke and Shearer,[4] and Wallis *et al.*[11] The main contribution of this paper is the definition of these new distance measures and the formal proof that they fulfill the metric properties.

2. Preliminaries

Let $G = (V, E)$ be a graph, where $V = \{1, \cdots, n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. Two nodes $u, v \in V$ are said to be *adjacent* if they are connected by an edge. Given a subset of nodes $C \subseteq V$, the *induced subgraph* $G[C]$ is the graph having C as its node set, and two nodes are adjacent in $G[C]$ if and only if they are adjacent in G.

Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs. Any bijection $f : H_1 \to H_2$, with $H_1 \subseteq V_1$ and $H_2 \subseteq V_2$, that preserves the adjacency relationships between the matched nodes is called a *subgraph isomorphism*. Formally, this amounts to stating that, given $u, v \in H_1$, we have $(u, v) \in E_1$ if and only if $(f(u), f(v)) \in E_2$. If $H_1 = V_1$ and $H_2 = V_2$, then we say that G_1 and G_2 are isomorphic, in which case we write $G_1 \cong G_2$. A subgraph isomorphism is *maximal* if there is no other subgraph isomorphism $f' : H_1' \to H_2'$ with H_1 a strict subset of H_1', and *maximum* if H_1 has largest cardinality. The maximal (maximum) subgraph isomorphism problem is to find a maximal (maximum) subgraph isomorphism between two rooted trees.

A word of caution about terminology is in order here. Despite name similarity, we are not addressing the standard subgraph isomorphism problem, which consists of determining whether a given graph is isomorphic to a subgraph of a larger one. In fact, we are dealing with a generalization thereof, the maximum common subgraph problem, which consists of determining the largest isomorphic subgraphs of two given graphs. We shall continue to use our own terminology, however, as it emphasizes the role of the isomorphism f.

Formally, an *attributed graph* is a triple $G = (V, E, \alpha)$, where (V, E) is the "underlying" graph and α is a function which assigns an attribute vector $\alpha(u)$ to each node $u \in V$. It is clear that in matching two attributed graphs, our objective is to find an isomorphism which pairs nodes having "similar" attributes. To this end, let σ be any similarity measure on the attribute space, i.e. any (symmetric) function which assigns a positive number to any pair of attribute vectors. If $f : H_1 \to H_2$ is a subgraph isomorphism between two attributed graphs $G_1 = (V_1, E_1, \alpha_1)$ and

$G_2 = (V_2, E_2, \alpha_2)$, the overall similarity between the induced subgraphs $G_1[H_1]$ and $G_2[H_2]$ can be defined as follows:

$$S(f) = \sum_{u \in H_1} \sigma(\alpha_1(u), \alpha_2(f(u))) \ . \tag{1}$$

The isomorphism f is called a *maximal similarity subgraph isomorphism* if there is no other subgraph isomorphism $f' : H_1' \to H_2'$ such that H_1 is a strict subset of H_1' and $S(f) < S(f')$. It is called a *maximum similarity subgraph isomorphism* if $S(f)$ is largest among all subgraph isomorphisms between G_1 and G_2.

In what follows, we shall assume that the similarity function is bounded from above, namely that it is a function of the form $\sigma : A \times A \to [0, M]$, where A is the space of attribute vectors, and $M \in \mathbb{R}_+$ is the upper bound. Moreover, we shall assume throughout that σ is "derived" from a metric, i.e. it is of the form

$$\sigma(a, b) = M - \delta(a, b) \qquad \forall a, b \in A \tag{2}$$

where δ is a metric. Note that for $M = 1$, δ is a normalized metric. The metric properties of δ imply that the function σ fulfills the following properties:

$$\begin{array}{lll} 1. & 0 \leq \sigma(a, b) \leq M & \\ 2. & \sigma(a, b) = M \quad \text{iff} \quad & a = b \\ 3. & \sigma(a, b) = \sigma(b, a) & \\ 4. & \sigma(a, c) \geq \sigma(a, b) + \sigma(b, c) - M & \end{array} \tag{3}$$

for all $a, b, c \in A$. These properties will be instrumental to show that our distance measures are indeed metrics.

3. Distance Based on the Maximal Similarity Common Subgraph

Let $G_1 = (V_1, E_1, \alpha_1)$ and $G_2 = (V_2, E_2, \alpha_2)$ be two nonempty attributed graphs, and let $f : H_1 \to H_2$ be a maximum similarity subgraph isomorphism between G_1 and G_2 ($H_1 \subseteq V_1$, $H_2 \subseteq V_2$). Let also G_{12} be the maximum similarity common subgraph induced by f. The first graph distance we propose in this paper is defined as:

$$d(G_1, G_2) = 1 - \frac{W(G_{12})}{M \max(|G_1|, |G_2|)} \tag{4}$$

where

$$W(G_{12}) = S(f) \tag{5}$$

is the overall similarity of nodes paired by f, as defined in (1).

Note that the choice of the similarity measure depends on the particular set of attributes assigned to the nodes in the graph. Therefore, (4) describes a whole family of distance measures between attributed graphs and not just one.

Theorem 1. *Let G_1, G_2, G_3 be three nonempty attributed graphs. If the similarity σ is derived from a metric, as in* (2)*, then the following properties hold true*:

1. $0 \leq d(G_1, G_2) \leq 1$
2. $d(G_1, G_2) = 0 \quad$ *iff* $\quad G_1 \cong G_2$
3. $d(G_1, G_2) = d(G_2, G_1)$
4. $d(G_1, G_2) + d(G_2, G_3) \geq d(G_1, G_3)\,.$

In other words, d is a normalized metric.

Proof. 1. Nonnegativity: $\quad 0 \leq d(G_1, G_2) \leq 1$

$$G_{12} \subseteq G_1\,,\ G_{12} \subseteq G_2 \quad \Rightarrow \quad |V_{12}| \leq \min(|V_1|, |V_2|) \leq \max(|V_1|, |V_2|)$$

$$\begin{aligned} (3) \quad \Rightarrow \quad W(G_{12}) &= \textstyle\sum_{v_1 \in H_1} \sigma(\alpha_1(v_1), \alpha_2(f(v_1))) \\ &\leq M|H_1| = M|V_{12}| \leq M \max(|V_1|, |V_2|) \\ \Rightarrow \quad 0 \leq d(G_1, G_2) &\leq 1\,. \end{aligned}$$

2. Uniqueness and identity: $\quad d(G_1, G_2) = 0 \quad$ iff $\quad G_1 \cong G_2$

$$\begin{aligned} &\text{Let} \quad d(G_1, G_2) = 0 \quad \Leftrightarrow \quad W(G_{12}) = M \max(|V_1|, |V_2|) \\ &\quad G_{12} \subseteq G_1, G_2 \quad \Rightarrow \quad |V_{12}| \leq \min(|V_1|, |V_2|) \\ (5) \quad &\Rightarrow \quad W(G_{12}) \leq M|V_{12}| \quad \Rightarrow \quad \max(|V_1|, |V_2|) \leq |V_{12}| \\ &\Rightarrow \quad \max(|V_1|, |V_2|) = |V_{12}| = \min(|V_1|, |V_2|) \\ &\Rightarrow \quad |V_{12}| = |V_1| = |V_2| \quad \Rightarrow \quad W(G_{12}) = M|V_{12}| \\ &\Rightarrow \quad \sigma(\alpha_1(v_i), \alpha_2(f(v_i))) = M \quad \forall v_i \in V_1 \\ &\Rightarrow \quad \alpha_1(v_i) = \alpha_2(f(v_i)) \quad \forall v_i \in V_1 \quad \text{[due to (3)]} \end{aligned}$$

$$(|V_1| = |V_2| \quad \wedge \quad \alpha_1(v_i) = \alpha_2(f(v_i)) \quad \forall v_i \in V_1) \quad \Rightarrow \quad G_1 \cong G_2\,.$$

On the other hand,

$$\begin{aligned} G_1 \cong G_2 \quad &\Leftrightarrow \quad |V_{12}| = |V_1| = |V_2| \quad \text{and} \quad \alpha_1(v_i) = \alpha_2(f(v_i)) \quad \forall\ v_i \in V_1 \\ \alpha_1(v_i) = \alpha_2(f(v_i)) \quad &\wedge \quad (3) \\ &\Rightarrow \quad \sigma(\alpha_1(v_i), \alpha_2(f(v_i))) = M \\ &\Rightarrow \quad W(G_{12}) = M|V_{12}| = M \max(|V_1|, |V_2|) \\ &\Rightarrow \quad d(G_1, G_2) = 0\,. \end{aligned}$$

3. Symmetry: $\quad d(G_1, G_2) = d(G_2, G_1)$

It follows directly by the symmetry of the function σ (3) and that of the maximum function: $\max(|V_1|, |V_2|) = \max(|V_2|, |V_1|)$.

4. Triangle inequality:

It has to be shown:

$$1 - \frac{W(G_{12})}{M\max(|G_1|,|G_2|)} + 1 - \frac{W(G_{23})}{M\max(|G_2|,|G_3|)} \geq 1 - \frac{W(G_{13})}{M\max(|G_1|,|G_3|)}\,. \quad (6)$$

Let:

- $f : V_1^{'} \to V_2^{'}$ a maximal similarity subgraph isomorphism between G_1 and G_2 that introduces the maximum similarity common subgraph $G_{12} = (V_{12}, E_{12}) \Rightarrow |V_1^{'}| = |V_2^{'}| = |V_{12}|$.
- $g : V_2^{''} \to V_3^{''}$ a maximal similarity subgraph isomorphism between G_2 and G_3 that introduces the maximum similarity common subgraph $G_{23} = (V_{23}, E_{23}) \Rightarrow |V_2^{''}| = |V_3^{''}| = |V_{23}|$.
- $h : V_1^{'''} \to V_3^{'''}$ a maximal similarity subgraph isomorphism between G_1 and G_3 that introduces the maximum similarity common subgraph $G_{13} = (V_{13}, E_{13}) \Rightarrow |V_1^{'''}| = |V_3^{'''}| = |V_{13}|$.

Let also $V_{123} = V_2^{'} \bigcap V_2^{''}$, i.e. the set of nodes belonging to both G_{12} and G_{23} and hence, common to all three graphs. It induces the common subgraph of G_1 and G_3 with the weight:

$$W_{13}(G_{123}) = \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), g(v_2))\,. \quad (7)$$

G_{123} is not necessarily of maximum similarity when considered as subgraph of G_1 and G_3, but it obviously holds true that $W_{13}(G_{123}) \leq W(G_{13})$ where $W(G_{13})$ is the weight of the maximum similarity common subgraph of G_1 and G_3. The inequality $W_{13}(G_{123}) > W(G_{13})$ being in contradiction with the fact that G_{13} is of the maximum similarity, does not hold true.

Therefore, to prove the triangle inequality, it suffices to show

$$1 - \frac{W(G_{12})}{M\max(|G_1|,|G_2|)} + 1 - \frac{W(G_{23})}{M\max(|G_2|,|G_3|)} \geq 1 - \frac{W_{13}(G_{123})}{M\max(|G_1|,|G_3|)} \quad (8)$$

which, after some algebra, is equivalent to:

$$\begin{aligned} M\max(|G_1|,|G_2|)\max(|G_2|,|G_3|)\max(|G_1|,|G_3|) \\ \geq\ W(G_{12})\max(|G_2|,|G_3|)\max(|G_1|,|G_3|) \\ + W(G_{23})\max(|G_1|,|G_2|)\max(|G_1|,|G_3|) \\ - W_{13}(G_{123})\max(|G_1|,|G_2|)\max(|G_2|,|G_3|)\,. \end{aligned} \quad (9)$$

Before beginning the formal proof of the last inequality, some additional useful relations are to be noted:

$$\begin{aligned} W(G_{12}) &= \sum_{v_2 \in V_2^{'} \setminus V_{123}} \sigma(f^{-1}(v_2), v_2) + \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) \\ &\leq M(|V_2^{'}| - |V_{123}|) + \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2)\,. \end{aligned}$$

The analogue relation holds true for $W(G_{23})$:

$$W(G_{23}) \leq M(|V_{23}| - |V_{123}|) + \sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2))$$

$$W(G_{13}) \geq \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), g(v_2))\,.$$

We also have:

$$\begin{aligned} M|V_2| &\geq M(|V_{12}| + |V_{23}| - |V_{123}|) \\ &= M(|V_{12}| - |V_{123}|) + M(|V_{23}| - |V_{123}|) + M|V_{123}| \\ &\geq W(G_{12}) - \textstyle\sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) \\ &\quad + W(G_{23}) - \textstyle\sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2)) + M|V_{123}| \end{aligned}$$

i.e.

$$M|V_2| \geq W(G_{12}) + W(G_{23}) - \sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M)\,.$$

Two cases are to be distinguished now:

1. $\displaystyle\sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M) < 0$

$$\Rightarrow \quad M|V_2| \geq W(G_{12}) + W(G_{23}) \geq W(G_{12}) + W(G_{23}) - W_{13}(G_{123})$$
$$\text{due to} \quad W_{13}(G_{123}) \geq 0\,.$$

2. $\displaystyle\sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M) \geq 0\,.$

From property 4 of (3), it follows:

$$\sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), g(v_2)) \geq \sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M) \qquad (10)$$

that implies:

$$\begin{aligned} M|V_2| &\geq W(G_{12}) + W(G_{23}) - \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), g(v_2)) \\ &= W(G_{12}) + W(G_{23}) - W_{13}(G_{123})\,. \end{aligned}$$

Therefore, we can conclude:

$$M|V_2| \geq W(G_{12}) + W(G_{23}) - W_{13}(G_{123})\,. \qquad (11)$$

Another note regarding the notation that will be used during the proof:

$$W'(G_{12}) = \sum_{v_2 \in V_2' \setminus V_{123}} \sigma(f^{-1}(v_2), v_2), \quad W'(G_{23}) = \sum_{v_2 \in V_2'' \setminus V_{123}} \sigma(v_2, g(v_2))\,.$$

There are six possible cases to be distinguished and proven:

- Case 1: $|V_1| \geq |V_2| \geq |V_3|$.

In this case, inequality (9) is equivalent to the following one:

$$M|V_1||V_2| \geq |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123})\,.$$

Consider the inequality: $|V_2| \geq |V_{12}| + |V_{23}| - |V_{123}|$

$$\begin{aligned}
&M|V_1||V_2| \\
&\geq M|V_1||V_{12}| + M|V_1||V_{23}| - M|V_1||V_{123}| \\
&= |V_1|\ M(|V_{12}| - |V_{123}|) + |V_1|M(|V_{23}| - |V_{123}|) + M|V_1||V_{123}| \\
&\geq |V_1|W^{'}(G_{12}) + |V_1|W^{'}(G_{23}) + M|V_1||V_{123}| \\
&= |V_1|W^{'}(G_{12}) + |V_1|W^{'}(G_{23}) + M|V_2||V_{123}| + (|V_1| - |V_2|)M|V_{123}| \\
&\geq |V_2|W^{'}(G_{12}) + |V_1|W^{'}(G_{23}) + M|V_2||V_{123}| + (|V_1| - |V_2|) \sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2)) \\
&= |V_2|W^{'}(G_{12}) + |V_1|W^{'}(G_{23}) + (|V_1| - |V_2|) \sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2)) \\
&\quad + M|V_2||V_{123}| + |V_2| \left[\sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) - \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) \right] \\
&= |V_2| \left[W^{'}(G_{12}) + \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) \right] + |V_1| \left[W^{'}(G_{23}) + \sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2)) \right] \\
&\quad - |V_2| \left[\sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), v_2) + \sum_{v_2 \in V_{123}} \sigma(v_2, g(v_2)) - M|V_{123}| \right] \\
&= |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2| \sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M)\,.
\end{aligned}$$

If $\sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M) < 0$ then the following holds true:

$$\begin{aligned}
M|V_1||V_2| &\geq |V_2|W(G_{12}) + |V_1|W(G_{23}) \\
&\geq |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123})\,.
\end{aligned}$$

If, on the other hand, $\sum_{v_2 \in V_{123}} (\sigma(f^{-1}(v_2), v_2) + \sigma(v_2, g(v_2)) - M) \geq 0$, (10) implies:

$$\begin{aligned}
M|V_1||V_2| &\geq |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2| \sum_{v_2 \in V_{123}} \sigma(f^{-1}(v_2), g(v_2)) \\
&= |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123})\,.
\end{aligned}$$

In both cases we have

$$M|V_1||V_2| \geq |V_2|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123})\,.$$

- Case 2: $|V_1| \geq |V_3| \geq |V_2|$.

In this case, inequality (9) is equivalent to:

$$M|V_1||V_3| \geq |V_3|W(G_{12}) + |V_1|W(G_{23}) - |V_3|W_{13}(G_{123})\,.$$

The proof is analogous to Case 1.

- Case 3: $|V_2| \geq |V_1| \geq |V_3|$.

In this case, inequality (9) is equivalent to:

$$M|V_1||V_2| \geq |V_1|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123})\,.$$

From inequality (11) it follows:

$$\begin{aligned} M|V_1||V_2| &\geq |V_1|W(G_{12}) + |V_1|W(G_{23}) - |V_1|W_{13}(G_{123}) \\ &\geq |V_1|W(G_{12}) + |V_1|W(G_{23}) - |V_2|W_{13}(G_{123}) \end{aligned}$$

due to $|V_1| \leq |V_2| \quad \Rightarrow \quad -|V_1| \geq -|V_2|$.

- Case 4: $|V_2| \geq |V_3| \geq |V_1|$.

In this case, inequality (9) is equivalent to:

$$M|V_3||V_2| \geq |V_3|W(G_{12}) + |V_3|W(G_{23}) - |V_2|W_{13}(G_{123})\,.$$

The proof is analogous to Case 3.

- Case 5: $|V_3| \geq |V_1| \geq |V_2|$.

In this case, inequality (9) is equivalent to:

$$M|V_1||V_3| \geq |V_3|W(G_{12}) + |V_1|W(G_{23}) - |V_1|W_{13}(G_{123})\,.$$

The proof is analogous to Case 1.

- Case 6: $|V_3| \geq |V_2| \geq |V_1|$.

In this case, inequality (9) is equivalent to:

$$M|V_2||V_3| \geq |V_3|W(G_{12}) + |V_2|W(G_{23}) - |V_2|W_{13}(G_{123})\,.$$

The proof is analogous to Case 1.

It has been proven that in all six possible cases, inequality (9) holds true. Its equivalence to the triangle inequality also implies that the latter one is always satisfied. This concludes the proof of the metric properties of the new distance measure based on the maximal similarity common subgraph we proposed. □

If we restrict ourselves to mappings that preserve the attributes assigned to nodes, i.e. match nodes with *identical* attribute vectors, then $W(G_{12}) = M|V_{12}|$. The same relation also holds when the graphs at hand are unlabeled, by assuming w.l.o.g. that all nodes carry the same label, or attribute vector. In these cases, our distance measure becomes:

$$d(G_1, G_2) = 1 - \frac{|G_{12}|}{\max(|G_1|, |G_2|)} \tag{12}$$

which coincides with the distance based on the maximum common subgraph defined by Bunke and Shearer in Ref. 4, and is therefore a proper generalization thereof.

4. A Distance Based on the Graph Union

Another distance measure between graphs that follows the idea of finding the maximum common subgraph was introduced by Wallis *et al.* in Ref. 11. It is very similar to the distance proposed by Bunke and Shearer[4] since both can be expressed in the same form:

$$d(G_1, G_2) = 1 - \frac{m(G_1, G_2)}{M(G_1, G_2)}$$

with $m(G_1, G_2)$ representing the similarity of graphs while $M(G_1, G_2)$ represents the size of the problem. In both distances $m(G_1, G_2) = |G_{12}|$, while they differ in the definition of $M(G_1, G_2)$. Bunke and Shearer used the size of the larger of two graphs, while Wallis *et al.* used the size of the graph union to model the size of the problem. The authors explained the last choice by observing that it can capture the variance in the size of the smaller of the two graphs, which is important in some applications.

In Ref. 11, the union of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is defined as a graph $G = (V_1 \cup V_2, E_1 \cup E_2)$, where the common nodes, i.e. nodes belonging to the intersection of sets of nodes of two graphs, are nodes made correspondent by the maximum subgraph isomorphism between G_1 and G_2. That is one possible interpretation of the graph union when graphs being isomorphic, but not equal are considered as equivalent. The distance measure based on the graph union is then defined as follows[11]:

$$d(G_1, G_2) = 1 - \frac{|G_{12}|}{|G_1| + |G_2| - |G_{12}|} \tag{13}$$

where $|G|$ represents the cardinality of the node set of the graph G. The measure of distance above defined is a normalized metric as was formally proven in Ref. 11. We now generalize this measure on attributed graphs.

The generalization of the union-based graph distance on attributed graphs uses the analogous idea of pairing the nodes having assigned similar, but not exactly equal, set of attributes. Once again the subgraph isomorphism we are interested in is characterized by the maximal similarity instead of maximal cardinality.

A straightforward generalization of the union-based distance measure to attribute graphs would be the following:

$$d(G_1, G_2) = 1 - \frac{W(G_{12})}{M(|G_1| + |G_2| - |G_{12}|)} \tag{14}$$

where again, G_{12} is a maximal similarity common subgraph derived from a maximal similarity subgraph isomorphism f, and $W(G_{12}) = S(f)$. Note, however, that since the function σ fulfills the property of nonnegativity, a maximal similarity common subgraph has also maximal cardinality. G_{12} has been derived from the maximal similarity subgraph isomorphism, but there could exist another subgraph G'_{12} such that

$$W(G_{12}) = W(G'_{12}) \quad \wedge \quad |G_{12}| < |G'_{12}| \,.$$

In this case we should have:

$$\begin{aligned} d(G_1, G_2) &= 1 - \frac{W(G_{12})}{M(|G_1| + |G_2| - |G_{12}|)} \\ &> 1 - \frac{W(G'_{12})}{M(|G_1| + |G_2| - |G'_{12}|)} = d'(G_1, G_2) \end{aligned}$$

which obviously makes no sense. Hence, to correctly define the graph distance based on graph union, it is necessary that the maximal similarity common subgraph used to measure the distance has the largest cardinality among all maximal similarity common subgraphs. This leads to our "correct" union-based distance.

Let $G_1 = (V_1, E_1, \alpha_1)$ and $G_2 = (V_2, E_2, \alpha_2)$ be two nonempty attributed graphs and let $f : H_1 \to H_2$ ($H_1 \subseteq V_1$, $H_2 \subseteq V_2$) be a subgraph isomorphism between $G_1[H_1]$ and $G_2[H_2]$. The distance between graphs G_1 and G_2 based on the union is defined as:

$$d(G_1, G_2) = 1 - \max_{f: H_1 \to H_2} \frac{S(f)}{M(|G_1| + |G_2| - |H_1|)} \,. \tag{15}$$

Theorem 2. *Let G_1, G_2, G_3 be three nonempty attributed graphs. If the similarity σ is derived from a metric, as in* (2)*, then the distance function defined in* (15) *is a normalized metric.*

Proof. In order to simplify the notation we write the distance in the form given in (14), remembering that G_{12} has the largest cardinality between all the maximal similarity common subgraphs of G_1 and G_2.

To prove the theorem it is necessary to show that the distance function satisfies the properties of nonnegativity, uniqueness, identity, symmetry and triangle inequality.

1. Nonnegativity: $0 \le d(G_1, G_2) \le 1$

$$G_{12} \subseteq G_1, \quad G_{12} \subseteq G_2 \quad \Rightarrow \quad |V_{12}| \le \min(|V_1|, |V_2|) \,.$$

$$(3) \quad \Rightarrow \quad W(G_{12}) = \sum_{v_1 \in H_1} \sigma(\alpha_1(v_1), \alpha_2(f(v_1)))$$
$$\leq M|H_1| = M|V_{12}| \leq M(|V_1| + |V_2| - |V_{12}|)$$
$$\Rightarrow \quad 0 \leq d(G_1, G_2) \leq 1\,.$$

2. Uniqueness and identity: $d(G_1, G_2) = 0$ iff $G_1 \cong G_2$

Let $d(G_1, G_2) = 0 \quad \Leftrightarrow \quad W(G_{12}) = M(|V_1| + |V_2| - |V_{12}|)$
$G_{12} \subseteq G_1, G_2 \quad \Rightarrow \quad |V_{12}| \leq \min(|V_1|, |V_2|).$

$$\begin{aligned}
(3) \quad &\Rightarrow \quad W(G_{12}) \leq M|V_{12}| \\
&\Rightarrow \quad |V_1| + |V_2| - |V_{12}| \leq |V_{12}| \\
&\Rightarrow \quad |V_{12}| = |V_1| = |V_2| \qquad \Rightarrow \quad W(G_{12}) = M|V_{12}| \\
&\Rightarrow \quad \sigma(\alpha_1(v_1), \alpha_2(f(v_1))) = M \qquad \forall v_1 \in V_1 \\
&\Rightarrow \quad \alpha_1(v_1) = \alpha_2(f(v_1)) \qquad \forall v_1 \in V_1 \qquad \text{[due to (3)]}
\end{aligned}$$

$$(\ |V_1| = |V_2| \quad \wedge \quad \alpha_1(v_1) = \alpha_2(f(v_1)) \quad \forall v_1 \in V_1) \quad \Rightarrow \quad G_1 \cong G_2\,.$$

For the opposite direction, we have:
$G_1 \cong G_2 \quad \Leftrightarrow \quad |V_{12}| = |V_1| = |V_2| \quad e \quad \alpha_1(v_1) = \alpha_2(f(v_1)) \quad \forall v_1 \in V_1$

$$\begin{aligned}
\alpha_1(v_1) = \alpha_2(f(v_1)) \quad &\wedge \quad (3) \\
&\Rightarrow \quad \sigma(\alpha_1(v_1), \alpha_2(f(v_1))) = M \quad \forall v_1 \in V_1 \\
&\Rightarrow \quad W(G_{12}) = M|V_{12}| = M(|V_1| + |V_2| - |V_{12}|) \\
&\Rightarrow \quad d(G_1, G_2) = 0\,.
\end{aligned}$$

3. Symmetry: $d(G_1, G_2) = d(G_2, G_1)$

It follows directly from the symmetry of function σ (3).

4. Triangle inequality:

$$1 - \frac{W(G_{12})}{M(|G_1| + |G_2| - |G_{12}|)} + 1 - \frac{W(G_{23})}{M(|G_2| + |G_3| - |G_{23}|)}$$
$$\geq 1 - \frac{W(G_{13})}{M(|G_1| + |G_3| - |G_{13}|)}\,. \qquad (16)$$

Analogous to what was proved for the labeled or unlabeled graphs, the maximal similarity common subgraph is considered like the intersection of two graphs.

$$G_1 \cap G_2 = G_{12} \quad \Rightarrow \quad |G_1 \cup G_2| = |G_1| + |G_2| - |G_{12}|\,.$$

Therefore, the triangle inequality (16) can be rewritten as:

$$1 - \frac{W(G_{12})}{M|G_1 \cup G_2|} + 1 - \frac{W(G_{23})}{M|G_2 \cup G_3|} \geq 1 - \frac{W(G_{13})}{M|G_1 \cup G_3|}\,.$$

Let us denote:

- x_1 the number of nodes of G_1 that do not belong to $G_2 \cup G_3$
- x_{12} the number of nodes of $G_1 \cap G_2$ that do not belong to G_3
- $x_{123} = |G_{123}|$ the number of nodes common to all three graphs
- X the total number of nodes in: $G_1 \cup G_2 \cup G_3$.

Analogously, we can define x_2, x_3, x_{13}, x_{23}. Using this notation the previous inequality can be rewritten as:

$$1 - \frac{W(G_{12})}{M(X - x_3)} + 1 - \frac{W(G_{23})}{M(X - x_1)} \geq 1 - \frac{W(G_{13})}{M(X - x_2)} \tag{17}$$

$$\begin{aligned} \Leftrightarrow \quad & M(X - x_1)(X - x_2)(X - x_3) - W(G_{12})(X - x_1)(X - x_2) \\ & - W(G_{23})(X - x_2)(X - x_3) \\ & + W(G_{13})(X - x_1)(X - x_3) \geq 0\,. \end{aligned}$$

$$\begin{aligned} \Leftrightarrow \quad \text{Dis} = {} & X^2(MX - Mx_1 - Mx_2 - Mx_3 - W(G_{12}) - W(G_{23})) \\ & + X[W(G_{12})(x_1 + x_2) + W(G_{23})(x_2 + x_3) + Mx_1x_3] \\ & + x_1x_2(MX - W(G_{12}) - Mx_3) + x_2x_3(MX - W(G_{23})) \\ & + x_1x_3W(G_{13}) + XW(G_{13})(X - x_1 - x_3) \geq 0\,. \end{aligned}$$

Recall that

$$\begin{aligned} W(G_{12}) &\leq M|G_{12}| = M(x_{12} + x_{123}) \\ W(G_{23}) &\leq M|G_{23}| = M(x_{23} + x_{123})\,. \end{aligned}$$

Using last two relations and the implication:

$\forall a, b \geq 0 \quad a \leq b \quad \Rightarrow \quad -a \geq -b$ we obtain:

$$\begin{aligned} \text{Dis} \geq {} & X^2M(X - x_1 - x_2 - x_3 - x_{12} - x_{23} - x_{123} - x_{123}) \\ & + X[W(G_{12})(x_1 + x_2) + W(G_{23})(x_2 + x_3) + Mx_1x_3] \\ & + x_1x_2M(X - x_{12} - x_{123} - x_3) + x_2x_3M(X - x_{23} - x_{123})) \\ & + x_1x_3W(G_{13}) + XW(G_{13})(X - x_1 - x_3)\,. \end{aligned}$$

$$\begin{aligned} \text{Dis} \geq {} & X^2M(X - x_{13} - x_{123}) + x_1x_3W(G_{13}) + XW(G_{13})(X - x_1 - x_3) \\ & + x_1x_2M(X - x_{12} - x_{123} - x_3) + x_2x_3M(X - x_{23} - x_{123}) \\ & + X[W(G_{12})(x_1 + x_2) + W(G_{23})(x_2 + x_3) + Mx_1x_3] \geq 0\,. \end{aligned}$$

This implies that the inequalities (17) and (16) are always true that concludes the proof of the metric properties of the distance function given in (15). □

5. Conclusions

We have proposed two new attributed graph distance measures that are based on the notion of a maximal similarity common subgraph. Their main advantage is the fact that they do not depend on the definition of elementary edit operations and their costs, which represents a difficult task in traditional edit-distance measures. We have proven that the new distance measures are metrics. The proposed distance measures are general and therefore applicable in a variety of domains. Whenever abstract representations of complex objects and patterns are given by attributed graphs, our framework can be applied to deal not only with structural perturbations but also with geometric distorsions.

The main drawback of our distance measures, shared by all other metrics based on common substructures, is the associated computational complexity. This coincides with the complexity of determining the maximum similarity common subgraph, which is an NP-complete problem. A recently introduced approach to attack this problem consists of transforming it into the equivalent problem of finding a maximum weight clique (i.e. a complete subgraph having largest weight) in an auxiliary structure called the (weighted) association graph.[10] Although the maximum weight clique problem is NP-complete, many powerful heuristics have been developed which efficiently find good approximate solutions.[2]

References

1. D. Ballard and C. M. Brown, *Computer Vision* (Prentice-Hall, Englewood Cliffs, NJ, 1982).
2. I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, The maximum clique problem, *Handbook of Combinatorial Optimization* (*Suppl. Vol. A*), eds. D.-Z. Du and P. M. Pardalos (Kluwer, Boston, MA, 1999), pp. 1–74.
3. H. Bunke, Recent developments in graph matching, in *Proc. 15th Int. Conf. Pattern Recognition*, Vol. 2 (IEEE Computer Society Press, 2000), pp. 117–124.
4. H. Bunke and K. Shearer, A graph distance metric based on the maximal common subgraph, *Patt. Recogn. Lett.* **19** (1998) 255–259.
5. H. Bunke, Error correcting graph matching: on the influence of the underlying cost function, *IEEE Trans. Patt. Anal. Mach. Intell.* **21** (1999) 917–922.
6. S. Dickinson, M. Pelillo and R. Zabih (eds.), Special section on graph algorithms in computer vision, *IEEE Trans. Patt. Anal. Mach. Intell.* **23** (2001) 1049–1052.
7. M. L. Fernandez and G. Valiente, A graph distance metric combining maximum common subgraph and minimum common supergraph, *Patt. Recogn. Lett.* **22** (2001) 753–758.
8. P. Klein, S. Tirthapura, D. Sharvit and B. Kimia, A tree-edit-distance algorithm for comparing simple, closed shapes, in *Proc. 10th Ann. ACM–SIAM Symp. Discr. Algorithms* (*SODA*) San Francisco, CA, 2000, pp. 696–704.
9. T.-L. Liu and D. Geiger, Approximate tree matching and shape similarity, in *Proc. ICCV'99 — 7th Int. Conf. Computer Vision*, 1999, Kerkyra, Greece, pp. 456–462.
10. M. Pelillo, K. Siddiqi and S. W. Zucker, Matching hierarchical structures using association graphs, *IEEE Trans. Patt. Anal. Mach. Intell.* **21** (1999) 1105–1120.
11. W. D. Wallis, P. Shoubridge, M. Kraetz and D. Ray, Graph distances using graph union, *Patt. Recogn. Lett.* **22** (2001) 701–704.

Marcello Pelillo received the "Laurea" degree with honors in computer science from the University of Bari, Italy, in 1989. From 1988 to 1989 he was at the IBM Scientific Center in Rome, where he was involved in studies on natural language and speech processing.

In 1991 he joined the Faculty of the University of Bari, Italy, as an Assistant Professor of computer science. Since 1995, he has been with the Univesity of Venice, Italy, where he is currently an Associate Professor of computer science. He held visiting research positions at Yale University (USA), University College London (England), McGill University (Canada), the University of Vienna (Austria), and York University (England).

He has organized a number of scientific events, including the Neural Information Processing Systems (NIPS) 1999 Workshop on "Complexity and Neural Computation: The Average and the Worst Case" (Breckenridge, Colorado, December 1999). In 1997 he established a new series of international workshops devoted to energy minimization methods in computer vision and pattern recognition (EMMCVPR). He also co-edited two journal special issues on this theme: one, in 2000, for *Pattern Recognition*, and the other, in 2003/2004, for the *IEEE Trans. Pattern Analysis and Machine Intelligence*. In 2001 he was a guest co-editor of a special issue of the *IEEE Trans. Pattern Analysis and Machine Intelligence* devoted to "Graph algorithms in computer vision." He has been on the program committees of several international conferences and workshops and serves on the editorial board for the journal *Pattern Recognition*. Professor Pelillo is a member of the IEEE, the International Association for Pattern Recognition and the Pattern Recognition Society.

His research interests are in the area of computer vision, pattern recognition and neural computation, where he has published more than 90 papers in refereed journals, handbooks, and conference proceedings.

Džena Hidović received the "Laurea" degree with honours in computer science from Ca' Foscari University of Venice, Italy, in 2002. Currently, she is doing a Ph.D. in computer science at the University of Birmingham, UK.

Her research interests include tree and graph metrics for image analysis and pattern recognition, as well as medical image understanding and physics-based models of tissue optics.