On The Similarity Metric

Elie Alhajjar*[1] and Clément Lefèvre†[2]

[1]Department of Mathematical Sciences, United States Military Academy
[2]Department of Mathematics, Ecole Saint Cyr

May 24, 2019

Abstract

In mathematics, and more specifically in topology, the notion of distance metric is well known since the nineteenth century. It is used to measure the "difference" between two objects. When it comes to characterizing the similarity between two objects, a similarity metric is needed. Although widely used in computer science, such a metric is not clearly defined mathematically. We fill in the existing gap in the current literature concerning similarity metrics, connecting them to the well-known notion of partial metrics in general topology.

1 Introduction

Comparing two objects is a top priority task for a lot of scientists nowadays. In biology for instance, one may wonder how similar two DNA sequences are. In cybersecurity, while analyzing the traffic of a network, one wants to compare the packets which are traveling on the network with packets of reference in order to tell if they represent a threat [2]. A way to achieve such a comparison is to use similarity metrics.

Intuitively, the similarity between two objects represents the common information shared by these two objects. The term "metric" refers to the notion of distance metric in mathematics. It has been studied since 1906 when it was first introduced by Maurice Fréchet [10], a French Mathematician who established metric spaces.

The main point is that distance metrics - such as the Euclidean distance, the Gaussian Kernel, the Hausdorff metric, the Minkowski distance, etc. - are well known in mathematics, but they are not similarity metrics as defined in the next section. Indeed, metrics like the Jaccard index, the Keyword distance and the

*elie.alhajjar@westpoint.edu
†clement.lefevre@st-cyr.terre-net.defense.gouv.fr

Overlap coefficient are used in computer science [9]. However, it seems that there is a lack of foundation from a mathematical point of view.

The definition of a similarity metric, on which the current work is based, was given by Chen, Ma and Zhang in their 2009 paper *On the Similarity Metric and the Distance Metric* [4], where they describe this notion as derived from distance metric. Here, the originality of our work lies in the fact that we no longer consider this analogy, but rather the derivation from partial metrics. Put in simple terms, a partial metric is a distance metric in which self-distances are allowed to be nonzero.

The main goal of this paper is to draw the connection between two notions previously thought of as unrelated, namely similarity metrics and partial metrics. Section 2 serves as an overview for the concepts and definitions that will be used later. In section 3, we make concrete the relationship between similarity and partial metrics. In Section 4, we introduce a new concept, the "ϵ-similarity", and discuss its application within our framework. Finally, we end the paper with a short section for remarks and future directions.

2 Preliminaries

In this section we gather all the material that will later be useful in the paper, such as the basic definitions, key notions and previous results in the literature. First of all, we recall the definition of a distance metric, and the one of a similarity metric as well.

Definition 1. *(Distance metric [10])*
Let X be a non-empty set. A metric, or a distance metric, is a function $d : X \times X \to \mathbb{R}$, such that for all x, y, z in X, d satisfies the following conditions:

1. $d(x, y) \geq 0$ *(non-negativity)*
2. $d(x, y) = d(y, x)$ *(symmetry)*
3. $d(x, y) = 0$ *if and only if* $x = y$ *(coincidence)*
4. $d(x, z) \leq d(x, y) + d(y, z)$ *(triangle inequality).*

The first example that comes to mind is the geometric distance between two points in Euclidean space, namely the *Euclidean distance*. It is a folklore exercise to show that it is indeed a distance metric, so we omit the proof (see for example [1]).

The next definition is that of a similarity metric. It is adopted from [4] and will be the basis of our work in the remainder of the paper.

Definition 2. *(Similarity metric [4,7])*
Let X be a non-empty set and x, y, z be in X. Let s: $X \times X \to \mathbb{R}$ be a function. We say that s is a similarity metric if it satisfies the following conditions:

1. $s(x, y) = s(y, x)$
2. $s(x, x) \geq 0$

3. $s(x, y) \leq s(x, x)$

4. $s(x, x) = s(y, y) = s(x, y)$ *if and only if* $x = y$

5. $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$.

If we take a close look at the conditions given for a function to be a similarity metric, we see that the first three are quite intuitive. First, the symmetry is obvious: the similarity between x and y is the same as the similarity between y and x. Second, even if the 0 as a lower bound is not mandatory, we can understand why the self-similarity should be positive. If we compare an object with itself, it seems intuitive that this similarity should be bigger than 0. Finally, the third condition expresses the fact that an object is always more similar to itself than to any other object.

Concerning the last two properties, we will see below that there is a connection with partial metrics. We have *almost* the same triangular inequality, and the fourth assertion is also a characteristic property of partial metrics.

A very well-known example to illustrate the latter definition is the *Jaccard index*.

Example 1. *(The Jaccard index [2])*
The Jaccard index is a function $J : \{sets\} \to [0, 1]$ *used to compare two given sets* A *and* B*. It is defined as*

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A *and* B *are not allowed to be both empty. Clearly, if* A *has nothing in common with* B*, then* $A \cap B = \varnothing$ *and* $J(A, B) = 0$*, while if they are the same set, then* $A \cap B = A \cup B = A = B$ *and* $J(A, B) = 1$*. For all other cases, it is not hard to see that* $J(A, B) \in (0, 1)$*.*

Note that $J(A, B) = J(B, A)$*,* $J(A, A) = 1$*, and that* $J(A, B) = 1$ *if and only if* $A = B$*. This takes care of conditions 1 through 4. It remains to show that the Jaccard index satisfies the last condition of a similarity metric. We record this result separately.*

Theorem 1. *The Jaccard index is a similarity metric.*

Proof. Let A, B and C be three given sets. It remains to show that

$$J(A, B) + J(B, C) \leq J(A, C) + 1.$$

The following drawing will be useful in the proof. Note that $a, b, c, d, e, f, g \geq 0$.

We want to show that

$$\frac{|A \cap B|}{|A \cup B|} + \frac{|B \cap C|}{|B \cup C|} - \frac{|A \cap C|}{|A \cup C|} - 1 \leq 0$$

$$\Leftrightarrow \frac{a+b}{a+b+c+d+e+f} + \frac{a+d}{a+b+c+d+e+g} - \frac{a+c}{a+b+c+d+f+g} - 1 \leq 0$$

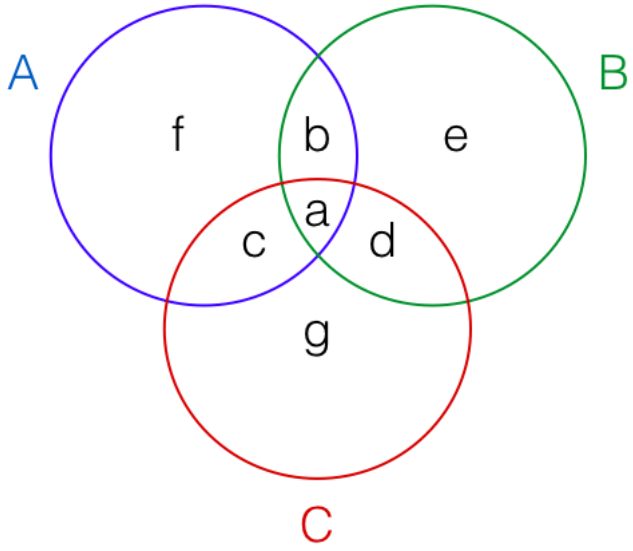

Figure 1: Venn Diagram

$$\Leftrightarrow \frac{N}{(a+b+c+d+e+f)(a+b+c+d+e+g)(a+b+c+d+f+g)} \leq 0,$$

$$N = -(2a^2c + 2a^2e + 4abc + 3abe + abf + 4ac^2 + 4acd + 6ace + 3acf + 3acg + 3ade + adg + 2ae^2 + 2aef + 2aeg + 2afg + 2b^2c + b^2e + b^2f + 4bc^2 + 4bcd + 5bce + 4bcf + 3bcg + 2bde + bdf + bdg + be^2 + 2bef + 2beg + bf^2 + 2bfg + 2c^3 + 4c^2d + 4c^2e + 3c^2f + 3c^2g + 2cd^2 + 5cde + 3cdf + 4cdg + 2ce^2 + 4cef + 4ceg + cf^2 + 4cfg + cg^2 + d^2e + d^2g + de^2 + 2def + 2deg + 2dfg + dg^2 + e^2f + e^2g + ef^2 + 2efg + eg^2 + f^2g + fg^2).$$

Since $a, b, c, d, e, f, g \geq 0$, then $N \leq 0$. This completes the proof. □

We now recall the definition of normalized metrics and s-metrics and we then point out the existing link between them.

Definition 3. *(Normalized Distance Metric [4])*
Let X be a non-empty set and d be a distance metric on X. Then d is a normalized distance metric if $d(x, y) \leq 1$ for all $x, y \in X$.

Definition 4. *(Normalized Similarity Metric[4])*
Let X be a non-empty set and s be a similarity metric on X. Then s is a normalized s-metric if $|s(x, y)| \leq 1$ for all $x, y \in X$.

The next proposition is taken from [4]. It simply states that the greater the distance between two objects is, the less similar they are. We provide a detailed proof for the sake of completeness.

Proposition 1. *If d is a normalized distance metric, then $1-d$ is a normalized similarity metric. If s is a normalized similarity metric with $s(x,y) \geq 0$ and $s(x,x) = 1$, then $1 - s$ is a normalized distance metric.*

Proof. Let d be a normalized distance metric. Set $s := 1 - d$, we show that s is a normalized similarity metric.

The first three axioms are trivial. For the fourth, assume $s(x,x) = s(y,y) = s(x,y)$, then $1 = 1 - d(x,y)$ which means $d(x,y) = 0$, which implies that $x = y$. The converse is trivial.

Finally, we have: $s(x,z) + s(y,y) = 1 - d(x,z) + 1 - d(y,y) = 2 - d(x,z)$, and $s(x,y) + s(y,z) = 2 - d(x,y) - d(y,z)$. By the triangle inequality for the distance metric d, we get $s(x,y) + s(y,z) - (s(x,z) + s(y,y)) = d(x,z) - d(x,y) - d(y,z) \leq 0$. This implies that $s(x,y) + s(y,z) \leq s(x,z) + s(y,y)$ and the first statement is deduced by noticing that $s(x,y) = 1 - d(x,y) \leq 1$ since d is a metric.

If we let s satisfy the hypothesis of the second statement and set $d := 1 - s$, then positivity and symmetry can be checked immediately. The triangle inequality for d follows from condition 5 for s. It remains to show that $d(x,y) = 0$ if and only if $x = y$.

Assume $d(x,y) = 0$, then $s(x,y) = 1$ which implies that $s(x,y) = s(x,x)$. Using condition 4 of a similarity metric, we get that $x = y$. The converse is trivial. Moreover, since s is normalized and $s(x,y) \geq 0$, d is indeed a normalized distance metric. □

Example 2. *(Graph Distance [4])*
The distance between two graphs G_1 and G_2 is defined as

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

where $|G_i|$ is the cardinality (number of vertices) of G_i. It is known that d is a distance metric and it is clear that d is normalized. Therefore, by Proposition 1, we obtain that

$$s(G_1, G_2) := \frac{|G_1 \cap G_2|}{\max\{|G_1|, |G_2|\}}$$

is a normalized similarity metric.

We now mention some other properties from [4] linking s-metrics to distance metrics. Unless otherwise stated, we consider X to be a non-empty set, and x, y, z three elements of this set.

Lemma 1. *Let s be a s-metric and define d as*

$$d(x,y) := \frac{s(x,x) + s(y,y)}{2} - s(x,y).$$

Then d is a distance metric.

The previous lemma shows how to switch from a s-metric to a distance metric. The next one illustrates how to go the other way around.

Lemma 2. *Let d be a distance metric. Then for any $k \geq 1$ and any fixed point ϕ in X,*

$$s_k(x, y) := \frac{d(x, \phi) + d(y, \phi)}{k} - d(x, y)$$

is a similarity metric.

3 Similarity metrics as partial metrics

In this section, we will first define partial metrics, then introduce the relationship between these metrics and similarity metrics. Partial metrics are considered a type of generalization of distance metrics as they fail to satisfy some of the metric axioms mentioned in the beginning of the paper.

Definition 5. *(Partial metric [6])*
Let X be a non-empty set. A partial metric, or a p-metric for short, is a function $p : X \times X \to \mathbb{R}$, such that for all x, y, z in X we have:

1. $p(x, y) = p(y, x)$ *(symmetry)*
2. $p(x, x) = p(x, y) = p(y, y)$ *if and only if* $x = y$
3. $p(x, x) \leq p(x, y)$ *(small self-distances)*
4. $p(x, z) + p(y, y) \leq p(x, y) + p(y, z)$ *(triangle inequality for partial metrics).*

Here we can see that the main difference between distance metrics and partial metrics lies in the fact that the self-distance is no longer required to be equal to zero in partial metrics. That is to say that previously, with a given distance metric d and a point x of a non-empty set X, we had $d(x, x) = 0$. Now with partial metrics $p(x, x)$ is not necessarily equal to zero.

Note that in this article, we decided not to get too deep into the theory behind partial metrics. It is worth mentioning though that some authors have required that a partial metric takes its values in $[0, \infty]$, while others allow the values to be taken in $\mathbb{R}$. We state the definition in the most generality, keeping the specific details for case by case examples. Anything that might interest the reader about this topic can be found in the survey papers *Partial Metric Topology* by S.G. Matthews [8] and *Two Topologies Are Better Than One* by S.J. O'neill [5].

Here is an example of a partial metric over the real line $\mathbb{R}$.

Example 3. *We define $p : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by:*

$$p(x, y) := -\min\{x, y\}, \forall x, y \in \mathbb{R}$$

Then it is not hard to show that p is a partial metric on $\mathbb{R}$. In his paper [5], O'Neill calls p the usual p-metric.

Remark 1. *If p is a partial metric such that $p(x,x) = 0$ for all $x \in X$, then p is a distance metric. Every distance metric is clearly a partial metric, however the converse is not necessarily true, which justifies the fact that the latter generalizes the former.*

The following two theorems make clear the connection between partial metrics and similarity metrics. In other words, they give a way of creating one from the other via simple transformations.

Theorem 2. *Consider $p : X \times X \to [0,1]$, and define $s := 1 - p$. Then p is a partial metric if and only if s is a similarity metric.*

Proof. First of all, suppose that p is a partial metric. We show that s is a similarity metric. Let x, y, z be in X, we have:

1. $s(x,y) = 1 - p(x,y) = 1 - p(y,x) = s(y,x)$, thus s is symmetric.
2. $s(x,x) = 1 - p(x,x) \geq 0$, since we know that $p(x,x) \leq 1 \Leftrightarrow$ $1 - p(x,x) \geq 0$.
3. We know that for all x,y in X, $p(x,x) \leq p(x,y)$. Hence, $1-p(x,x) \geq 1-p(x,y)$. Thus $s(x,x) \geq s(x,y)$.
4. If $x = y$, then trivially $s(x,x) = s(y,y) = s(x,y)$.
   Conversely, suppose $s(x,x) = s(y,y) = s(x,y)$, let us show that $x = y$.
   The hypothesis gives us that:

   $$1 - p(x,x) = 1 - p(y,y) = 1 - p(x,y)$$

   Then obviously we obtain the result:

   $$p(x,x) = p(y,y) = p(x,y),$$

   which implies that

   $$x = y.$$

5. We have:

   $$\begin{aligned} & p(x,z) + p(y,y) \leq p(x,y) + p(y,z) \\ \Leftrightarrow & -(p(x,z) + p(y,y)) \geq -(p(x,y) + p(y,z)) \\ \Leftrightarrow & 1 - p(x,z) + 1 - p(y,y) \geq 1 - p(x,y) + 1 - p(y,z) \\ \Leftrightarrow & s(x,z) + s(y,y) \geq s(x,y) + s(y,z) \end{aligned}$$

Therefore, p is a partial metric implies that s is a similarity metric.

Conversely, suppose that s is a s-metric and let us show that p is a p-metric.

1. Obvious.
2. If $x = y$, then it is trivial. Suppose $p(x,x) = p(y,y) = p(x,y)$, then we have $s(x,x) = s(y,y) = s(x,y)$ which implies that $x = y$.

3. We know that $p(x,x) - p(x,y) = s(x,y) - s(x,x) \leq 0$. Hence, $p(x,x) \leq p(x,y)$.

4. We have:

$$\begin{aligned} & s(x,z) + s(y,y) \geq s(x,y) + s(y,z) \\ \Leftrightarrow & -(s(x,z) + s(y,y)) \leq -(s(x,y) + s(y,z)) \\ \Leftrightarrow & 1 - s(x,z) + 1 - s(y,y) \leq 1 - s(x,y) + 1 - s(y,z) \\ \Leftrightarrow & p(x,z) + p(y,y) \geq p(x,y) + p(y,z) \end{aligned}$$

Therefore, s is a similarity metric implies that p is a partial metric. □

This first theorem shows that it is possible in some cases to treat similarity metrics as partial metrics, which justifies that the link between topology and s-metrics is much deeper than what one might think.

Theorem 3. *Consider $p : X \times X \to \mathbb{R}$ such that $p(x,x) \leq 0$ for all $x \in X$, and $s := -p$. Then p is a partial metric if and only if s is a similarity metric.*

Proof. Let x, y, z be in X. We suppose first that p is a p-metric, and show that s is a s-metric.

1. s is symmetric because p is symmetric.

2. $s(x,x) = -p(x,x) \in \mathbb{R}_+$.

3. $s(x,x) = -p(x,x) \geq -p(x,y) = s(x,y)$.

4. If $x = y$, then obviously $s(x,x) = s(y,y) = s(x,y)$.
   Conversely, suppose $s(x,x) = s(y,y) = s(x,y)$. Then $-p(x,x) = -p(y,y) = -p(x,y)$, which implies that $p(x,x) = p(y,y) = p(x,y)$, which then leads to $x = y$.

5. $s(x,y) + s(y,z) = -(p(x,y) + p(y,z)) \leq -(p(x,z) + p(y,y)) = s(x,z) + s(y,y)$.

Thus, p is a p-metric implies that s is a s-metric.
The converse holds by the same reasoning. Hence, s is a s-metric implies that p is a p-metric. □

Theorem 3 is very powerful, because it is no longer mandatory to consider normalized p-metrics. The only condition on p is for its diagonal to take values in $\mathbb{R}_-$, therefore we can consider any p-metric, and map its diagonal to $\mathbb{R}_-$ instead of $\mathbb{R}_+$. Due to the symmetry of the real line about the origin, this consideration behaves in the same way as the original definition. This allows us to recover the results mentioned in section 2, by generalizing them using p-metrics. In the remainder of the section, we will consider s to be a s-metric defined by $s := -p$, where p is a p-metric whose diagonal values are restricted to $\mathbb{R}_-$.

3.1 Recovering of Lemma 1

Let p be a partial metric, X a non-empty set, and d the induced metric as defined in [6]:

$$d(x, y) := 2p(x, y) - p(x, x) - p(y, y),$$

for all $x, y \in X$. If we define $d^* := \frac{d}{2}$, then d^* is still a distance metric and we recover Lemma 1 by setting $s := -p$ as follows:

$$d^*(x, y) = \frac{s(x, x) + s(y, y)}{2} - s(x, y).$$

3.2 Recovering of Lemma 2

Let $d : X \times X \to \mathbb{R}_+$ be a distance metric, $k \geq 1$, and ϕ a fixed point in X. Then for any $x, y \in X$,

$$p(x, y) := d(x, y) - \frac{d(x, \phi) + d(y, \phi)}{k}$$

is a p-metric. Setting $s := -p$, we thus recover Lemma 2.

Proof. Let x, y, z be in X. We show that p is a p-metric.

1. p is symmetric because d is symmetric.

2. Suppose $p(x, y) = p(y, y) = p(x, x)$. We want to prove that $x = y$. We have:

$$d(x, y) - \frac{d(x, \phi) + d(y, \phi)}{k} = d(x, x) - \frac{2d(x, \phi)}{k} = d(y, y) - \frac{2d(y, \phi)}{k}$$

   Knowing that $d(x, x) = 0$ for any x, we obtain:

$$d(x, \phi) = d(y, \phi) \text{ and } d(x, y) = \frac{d(x, \phi) - d(y, \phi)}{k}$$

   Which means that $d(x, y) = 0 \Rightarrow x = y$. The converse is trivial.

3. We have:

$$\begin{aligned} p(x, x) - p(x, y) &= \frac{d(y, \phi) - d(x, \phi)}{k} - d(x, y) \\ &\leq \frac{d(x, y)}{k} - d(x, y) \leq 0 \end{aligned}$$

   By the triangle inequality, because we know that: $d(y, \phi) \leq d(y, x) + d(x, \phi)$, hence the fact that $d(y, \phi) - d(x, \phi) \leq d(x, y)$.

4. Finally, we take a look at the following:

$$p(x,z) + p(y,y) - p(x,y) - p(y,z)$$

$$= d(x,z) - \frac{d(x,\phi) + d(z,\phi)}{k} - \frac{2d(y,\phi)}{k} - d(x,y) + \frac{d(x,\phi) + d(y,\phi)}{k} - d(y,z) + \frac{d(y,\phi) + d(z,\phi)}{k}$$

$$= d(x,z) - d(x,y) - d(y,z) \leq 0,$$

by the triangle inequality for d.
This implies that $p(x,z) + p(y,y) \leq p(x,y) - p(y,z)$.

Therefore, p is indeed a partial metric.
Moreover, for $x \in X$, we have:

$$p(x,x) := d(x,x) - \frac{d(x,\phi) + d(x,\phi)}{k} = -\frac{2d(x,\phi)}{k} \leq 0$$

□

The above two results highlight an important conclusion: similarity metrics are partial metrics in disguise. The line of effort we put in this work is mainly towards re-conciliating the discrepancy between the metric notions in the fields of computer science and mathematics.

4 A Variation On Similarity

Now that we can consider a similarity metric as $s := -p$, where p is a partial metric, it is natural to ask the following question: what topological properties can be "translated" from p to s? In this section, we aim to provide a partial answer to this open ended question.

In order to do so, we first give the definition of boundedness in terms of s-metrics, then introduce the notion of **ϵ-Similarity**. Throughout the section, we consider a s-metric $s := -p$, where $p : X \times X \to \mathbb{R}_-$ is a p-metric.

4.1 Boundedness

Definition 6. *A similarity metric s is said to be **bounded** if there exists $M \in \mathbb{R}_+$ such that*

$$s(x,y) \leq M, \forall x, y \in X.$$

This definition is intuitive, since it is the natural way to say that a function is bounded. The following theorem shows the importance of bounded s-metrics.

Theorem 4. *Given a similarity metric s, the function $s^* := \frac{s}{s+1}$ is a bounded similarity metric.*

Proof. Let s be a similarity metric and $s^* := \frac{s}{s+1}$. We show that s^* is a s-metric, and that it is indeed bounded.

- s^* is bounded because for all $x, y \in X$ we have :

$$s^*(x,y) = \frac{s(x,y)}{s(x,y)+1} \leq \frac{s(x,y)}{s(x,y)} = 1$$

- Let us now show that s^* is a s-metric. For any $x, y, z \in X$ we have:

1. s^* is symmetric because s is symmetric.

2. $s^*(x,x) \geq 0$, because $s(x,x) \geq 0$ for all x.

3. We can write:

$$\begin{aligned} s^*(x,x) - s^*(x,y) &= \frac{s(x,x)}{s(x,x)+1} - \frac{s(x,y)}{s(x,y)+1} \\ &= \frac{s(x,x)[s(x,y)+1] - s(x,y)[s(x,x)+1]}{(s(x,x)+1)(s(x,y)+1)} \\ &= \frac{s(x,x) - s(x,y)}{(s(x,x)+1)(s(x,y)+1)} \geq 0. \end{aligned}$$

   This implies that $s^*(x,x) - s^*(x,y) \geq 0$ and $s^*(x,x) \geq s^*(x,y)$.

4. Since $s(x,x) = s(y,y) = s(x,y)$ if and only if $x = y$, then it is straightforward to see that $s^*(x,x) = s^*(y,y) = s^*(x,y)$ if and only if $x = y$.

5. We want to show that $s^*(x,y) + s^*(y,z) - s^*(x,z) - s^*(y,y) \leq 0$.
   In order to do so, we focus first on $s^*(x,y) - s^*(x,z)$. We have :

$$\begin{aligned} (1) : s^*(x,y) - s^*(x,z) &= \frac{s(x,y)}{s(x,y)+1} - \frac{s(x,z)}{s(x,z)+1} \\ &= \frac{s(x,y)(s(x,z)+1) - s(x,z)(s(x,y)+1)}{(s(x,z)+1)(s(x,y)+1)} \\ &= \frac{s(x,y) - s(x,z)}{(s(x,z)+1)(s(x,y)+1)} \leq s(x,y) - s(x,z). \end{aligned}$$

   The same way we have :

$$\begin{aligned} (2) : s^*(y,z) - s^*(y,y) &= \frac{s(y,z)}{s(y,z)+1} - \frac{s(y,y)}{s(y,y)+1} \\ &= \frac{s(y,z)(s(y,y)+1) - s(y,y)(s(y,z)+1)}{(s(y,y)+1)(s(y,z)+1)} \\ &= \frac{s(y,z) - s(y,y)}{(s(y,z)+1)(s(y,y)+1)} \leq s(y,z) - s(y,y). \end{aligned}$$

   Then finally :

$$(1) + (2) \leq s(x,y) + s(y,z) - s(x,z) - s(y,y) \leq 0.$$

Therefore, s^* is a bounded s-metric and the result follows. □

4.2 ϵ-similarity

Below we recall the definition of open balls in the partial metric context. It will be crucial in introducing the new notion of ϵ-similarity.

Definition 7. *(Open balls with p-metrics [5])*
Let p be a p-metric and $\epsilon > 0$. Then

$$B_\epsilon(x) := \{y \in X |\ p(x,y) < p(x,x) + \epsilon\}$$

is the open ball centered at x with radius ϵ.

Remark 2. *If we compare the above definition with the usual open ball definition in metric spaces, we see once again that the main difference between p-metrics and distance metrics lies in the fact that in general $p(x,x) \neq 0$. More precisely, if $p(x,x) = 0$, then Definition 7 is exactly the definition of open balls for a distance metric.*

What we define as ϵ-similarity is adapted from the definition of open balls for p-metrics. Indeed, it is a reformulation of the above definition via the substitution $s := -p$. For practical reasons, as shown in the examples below, we do not require strict inequality. This can be justified by the fact that the parameter ϵ can always be perturbed for that purpose.

Definition 8. *(ϵ-similarity)*
Let $x, y \in X$ and $\epsilon > 0$. Then y is said to be ϵ-similar to x if

$$s(x,x) - \epsilon \leq s(x,y) \leq s(x,x).$$

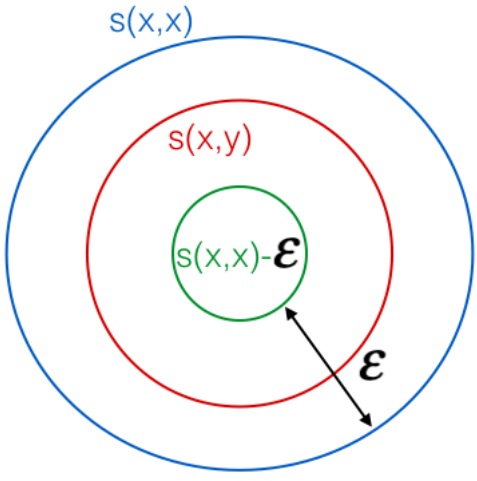

Figure 2: ϵ-Similarity

Note that the second inequality follows from the definition of similarity metrics, while the first one is inspired by open balls in the partial metric topology. Figure 2 serves as a naive illustration of the meaning behind ϵ-similarity. With this drawing in mind, it is clear that the smaller ϵ is, the more y is similar to x. The green circle gets bigger, and then forces the red one to become closer to the blue one.

For application purposes, it is sometimes more convenient to substitute the strict inequality by $s(x,x) - \epsilon \leq s(x,y)$. This in turn means that we are dealing now with closed balls instead of open balls in the partial metric topology. It is trivial to say that if y is ϵ-similar to x, then y is δ-similar to x for all $\delta \geq \epsilon$.

The next example is a good application of this new notion. It deals with a variation of the Jaccard index discussed in Section 2.

Example 4. *(ϵ-Jaccard index)*
Recall the definition of the Jaccard index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B are not allowed to be both empty. We say that a set B is ϵ-similar to a set A if for some $\epsilon > 0$,

$$J(A,A) - \epsilon \leq J(A,B) \leq J(A,A)$$

$$\Leftrightarrow 1 - \epsilon \leq J(A,B) \leq 1$$

$$\Leftrightarrow J(A,B) \in [1-\epsilon, 1]$$

It is clear that the smaller ϵ is, the more B is similar to A. We can even see this in terms of percentage, for example if $\epsilon = 0.05$, then A and B are said to be 95% similar.

For instance, if we consider different sets A, B, C, D, E, F, G, where

$$A = \{0, ..., 99\}$$

$$B = \{1, ..., 95\}$$

$$C = \{1, ..., 75\}$$

$$D = \{\ i \in [0, 99] \mid i \text{ is an odd number}\}$$

$$E = \{1, ..., 25\}$$

$$F = \{1, ..., 10\}$$

$$G = \{101\},$$

and fix A as a reference, then we can compare the other sets to A, using the Jaccard index. The results are as follow:

$$J(A,B) = 0.95$$

$$J(A,C) = 0.75$$

$$J(A,D) = 0.5$$

$$J(A,E) = 0.25$$

$$J(A, F) = 0.1$$

$$J(A, G) = 0.$$

Now we can pick different values for ϵ and see which sets are ϵ-similar to A. For $\epsilon = 0.05$, we obtain that B is 0.05-similar to A, which means that B and A are 95% similar. Likewise,

$$\epsilon = 0.25 \Rightarrow B \text{ and } C \text{ are } \epsilon\text{-similar to } A$$

$$\epsilon = 0.5 \Rightarrow B,\ C,\ D \text{ are } \epsilon\text{-similar to } A$$

$$\epsilon = 0.75 \Rightarrow B,\ C,\ D \text{ and } E \text{ are } \epsilon\text{-similar to } A$$

$$\epsilon = 0.9 \Rightarrow B,\ C,\ D,\ E \text{ and } F \text{ are } \epsilon\text{-similar to } A$$

Therefore, the ϵ-similarity is a good indicator of how similar two objects are.

5 Concluding Remarks

To summarize, the originality of this article is centered around the connection between two concepts: similarity metrics as defined in computer science and partial metrics as defined in general topology. To the best of our knowledge, this connection did not previously exist in the literature and we hope that we filled such a gap therein.

Beyond expanding the basis of the relationship between s-metrics and p-metrics, there remains a lot of work to do. This paper only scratches the surface of what we believe to be an impactful field of research, namely the study of similarity from a topological point of view. In an upcoming paper, the authors are planning to explore the properties of a partial metric space and their consequences on the similarity metrics. These properties include: convergence, Cauchy sequences, completeness, and many others.

6 Acknowledgment

The authors would like to thank the referee for their valuable comments that helped drastically improve the exposition of the paper and the clarity of the proofs. The second author is supported by a grant from the French Government. This work started during his stay at the United States Military Academy, West Point NY and continued during the first author's visit to Saint Cyr, France. Special thanks to Mr. Guy Chassé and to the head of the department of mathematical sciences at USMA, Col Tina Hartley, whose hospitality and support will be forever remembered.

References

[1] J. L. Kelley, *General Topology*, University Series In Mathematics, D. van Nostrand, 1955.

[2] Leigh Metcalf, William CASEY, *Cybersecurity and Applied Mathematics*, 1st edition, 2016.

[3] Suzhen Han, Jiangfeng Wu, Dong Zhang, *Properties and Principles on Partial Metric Spaces*, Topology and Its Application, 230 (2017) 77-98.

[4] Shihyen Chen, Bin Ma, Kaizhong Zhang, *On the Similarity Metric and the Distance Metric*, Theoretical Computer Science, 410 (2009) 2365-2376.

[5] S. J. O'neill, *Two Topologies are Better Than One*, 1995.

[6] Michael Bukatin, Ralph Kopperman, Steve Matthews, Homeira Pajoohesh *Partial Metric Spaces*, The American Mathematical Monthly, Vol. 116, No. 8 (Oct., 2009) 708-718.

[7] Ming Li, Xin Chen, Xin Li, Bin Ma, Paul M. B. Vintayin, *The Similarity Metric*, IEEE Transactions On Information Theory, Vol. 50, No. 12 (Dec., 2004) 3250-3264.

[8] S. G. Matthews, *Partial Metric Topology*, in, Papers on General Topology and Its Applications, Proc. 8^th Summer Conf., Queen's College, 1992, in, Ann. N.Y Acad. Sci., Vol. 728, (1994) 183-197.

[9] Saimadhu Polamuri, *Five Most Popular Similarity Measures Implementation In Python*, dataaspirant.com, http://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/ April, 2015.

[10] Maurice Fréchet, *Sur Quelques Points du Calcul Fonctionnel*, Thèse, Paris, 1905, Rendiconti Circolo Mat. Palermo, Vol. 22, (1906) 1–74.