# CHARACTERIZATION OF THE JACCARD DISSIMILARITY METRIC AND A GENERALIZATION

GEORGIOS GERASIMOU

Abstract. The Jaccard dissimilarity metric identifies the distance between two finite sets by the number of their unique elements as a proportion of their joint cardinality. This note gives an elementary characterization of the Jaccard metric by means of three simple axioms. Relaxing the most substantial one allows for a general family of new dissimilarity quasi-metrics to emerge that encompasses Jaccard's metric.

## 1. Introduction

For two finite sets $A$ and $B$ that are not both empty, the Jaccard metric [2, 3, 4, 7, 11, 12] defines the distance between them by

$$J(A,B) := \frac{|A\cup B|-|A\cap B|}{|A\cup B|} = 1-\frac{|A\cap B|}{|A\cup B|}. \tag{1}$$

This simple and intuitive formula is often interpreted as reflecting the dissimilarity between $A$ and $B$. It is frequently applied in such diverse fields as ecology, genetics, meteorology, information theory, operations research and recommender systems [1, 6, 9, 10, 13, 14, 15].

While axiomatic characterizations of other dissimilarity measures in different domains are known [5, 8, 16], it appears that this is not the case for (1). We provide an elementary such characterization by means of three simple axioms. These do not include the triangle inequality although, of course, they imply it.

Our analysis highlights the following two properties that are implicit in (1): (i) dissimilarity increases by a constant factor when any element belonging to two distinct sets is removed from one but not both sets (**A3**; Section 2); (ii) the cumulative dissimilarity between each of two disjoint sets and their union is the maximum possible (**A6**; Section 3). Relaxing these two requirements allows for a general and, as we show, novel family of dissimilarity quasi-metrics to emerge, which includes Jaccard's metric as a special case. Our analysis, finally, helps shed some new light on three distinct and pre-existing such quasi-metrics.

## 2. Axioms for the Jaccard Metric

We assume an unstructured finite set $X$ and consider a mapping $R : 2^X \times 2^X \to \mathbb{R}$ that is well-defined everywhere except at $(\emptyset,\emptyset)$. **A1**-**A3** below are some properties

2020 *Mathematics Subject Classification.* 68R99, 68T99, 68U01.

I thank a referee for constructive suggestions that strengthened the characterization and shortened its proof.

that $R$ could have if $R(A,B)$ were to be interpreted as reflecting the dissimilarity between sets $A$ and $B$ in this domain:

**A1.** $R(A,B) = R(B,A)$.

**A2.** $R(A,B) = 0 \iff A = B$.

**A3.** $A \not\ni x \in B \implies R(A,B) - R(A \cup \{x\}, B) = \dfrac{1}{|A \cup B|}$.

**A1** and **A2** are the standard metric symmetry and identity properties. **A3** requires constant marginal sensitivity of $R$ to the gradual removal from one of the two sets of elements that belong to both sets. In addition, it requires such constant marginal sensitivity to be reciprocally dependent on the joint cardinality of the two sets, which of course remains fixed during this gradual removal process.

**Proposition 1.** *$R$ satisfies* **A1**, **A2**, **A3** *if and only if $R \equiv J$.*

*Proof.* It is immediate that these axioms are implied by $J$. For the converse implication, let $R : 2^X \times 2^X \to \mathbb{R}$ satisfy **A1**, **A2**, **A3**. Consider $A, B \subseteq X$, with $A, B$ not both empty. It will be shown that, regardless of whether $A \subseteq B$, $B \subseteq A$, or otherwise, we have

$$\begin{aligned} R(A,B) &= \frac{|A \setminus B|}{|A \cup B|} + \frac{|B \setminus A|}{|A \cup B|} \\ &\equiv \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \\ &\equiv J(A,B) \end{aligned} \tag{2}$$

Suppose first that there is $x \in B \setminus A$. By **A3**, $R(A,B) = R(A \cup \{x\}, B) + \frac{1}{|A \cup B|}$. Applying **A3** iteratively over all elements of $B \setminus A$ yields

$$R(A,B) = R(A \cup B, B) + \frac{|B \setminus A|}{|A \cup B|}. \tag{3}$$

Now, either there is $x \in A \setminus B$ or $A \subseteq B$. Consider the former case first. By **A3**, $R(B, A \cup B) = R(B \cup \{x\}, A \cup B) + \frac{1}{|A \cup B|}$ for such $x$. Applying **A3** iteratively over all elements of $A \setminus B$ gives

$$\begin{aligned} R(A \cup B, B) &= R(B, A \cup B) \\ &= R(A \cup B, A \cup B) + \frac{|A \setminus B|}{|A \cup B|} \\ &= \frac{|A \setminus B|}{|A \cup B|}, \end{aligned} \tag{4}$$

with the first step following from **A1** and the last from **A2**. If $A \subseteq B$ holds instead, then

$$\begin{aligned} R(A \cup B, B) &= R(B, B) \\ &= 0, \end{aligned} \tag{5}$$

by **A2**.

Now suppose there is no $x \in B \setminus A$, so that $B \subseteq A$. This implies

$$R(A,B) = R(A \cup B, B). \tag{6}$$

If $B \subset A$, then (4) and (6) imply

$$R(A,B) = \frac{|A \setminus B|}{|A \cup B|}. \tag{7}$$

If $A = B$ instead, then (5) and (6) imply $R(A,B) = 0$. Thus, (2) holds in all cases.

To verify that **A1**, **A2** and **A3** are independent we show that there is a set $X$ such that, for every $i, j \in \{1,2,3\}$, there exists $R_j : 2^X \times 2^X \to \mathbb{R}$ that satisfies each $\mathbf{A}i$ except $\mathbf{A}j$. To this end, it suffices to consider the domains $X := \{a,b\}$ and $2^X \equiv \big\{\emptyset, \{a\}, \{b\}, \{a,b\}\big\}$.

$R_1$:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $R_1(\{a\},\{a\})$ | = | 0 | $R_1(\{b\},\{b\})$ | = | 0 | $R_1(\{a,b\},\{a,b\})$ | = | 0 |
| $R_1(\{a\},\{a,b\})$ | = | $\frac{1}{2}$ | $R_1(\{b\},\{a,b\})$ | = | $\frac{1}{2}$ | $R_1(\emptyset,\{a\})$ | = | 1 |
| $R_1(\emptyset,\{b\})$ | = | 1 | $R_1(\emptyset,\{a,b\})$ | = | 1 | $R_1(\{a\},\{b\})$ | = | 1 |
| $R_1(\{a\},\emptyset)$ | = | $\frac{1}{3}$ | $R_1(\{b\},\emptyset)$ | = | $\frac{1}{3}$ | $R_1(\{a,b\},\emptyset)$ | = | $\frac{1}{3}$ |
| $R_1(\{b\},\{a\})$ | = | 1 | $R_1(\{a,b\},\{a\})$ | = | $\frac{1}{4}$ | $R_1(\{a,b\},\{b\})$ | = | $\frac{1}{4}$ |

$R_2$:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $R_2(\{a\},\{a\})$ | = | 0 | $R_2(\{b\},\{b\})$ | = | 0 | $R_2(\{a,b\},\{a,b\})$ | = | 0 |
| $R_2(\{a\},\{a,b\})$ | = | $\frac{1}{2}$ | $R_2(\{b\},\{a,b\})$ | = | $\frac{1}{2}$ | $R_2(\emptyset,\{a\})$ | = | 1 |
| $R_2(\emptyset,\{b\})$ | = | 1 | $R_2(\emptyset,\{a,b\})$ | = | 1 | $R_2(\{a\},\{b\})$ | = | 0 |
| $R_2(\{a\},\emptyset)$ | = | 1 | $R_2(\{b\},\emptyset)$ | = | 1 | $R_2(\{a,b\},\emptyset)$ | = | 1 |
| $R_2(\{b\},\{a\})$ | = | 0 | $R_2(\{a,b\},\{a\})$ | = | $\frac{1}{2}$ | $R_2(\{a,b\},\{b\})$ | = | $\frac{1}{2}$ |

$R_3$:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $R_3(\{a\},\{a\})$ | = | 0 | $R_3(\{b\},\{b\})$ | = | 0 | $R_3(\{a,b\},\{a,b\})$ | = | 0 |
| $R_3(\{a\},\{a,b\})$ | = | $\frac{1}{3}$ | $R_3(\{b\},\{a,b\})$ | = | $\frac{1}{3}$ | $R_3(\emptyset,\{a\})$ | = | 1 |
| $R_3(\emptyset,\{b\})$ | = | 1 | $R_3(\emptyset,\{a,b\})$ | = | 1 | $R_3(\{a\},\{b\})$ | = | 1 |
| $R_3(\{a\},\emptyset)$ | = | 1 | $R_3(\{b\},\emptyset)$ | = | 1 | $R_3(\{a,b\},\emptyset)$ | = | 1 |
| $R_3(\{b\},\{a\})$ | = | 1 | $R_3(\{a,b\},\{a\})$ | = | $\frac{1}{3}$ | $R_3(\{a,b\},\{b\})$ | = | $\frac{1}{3}$ |

□

## 3. A General Class of Dissimilarity Quasi-Metrics

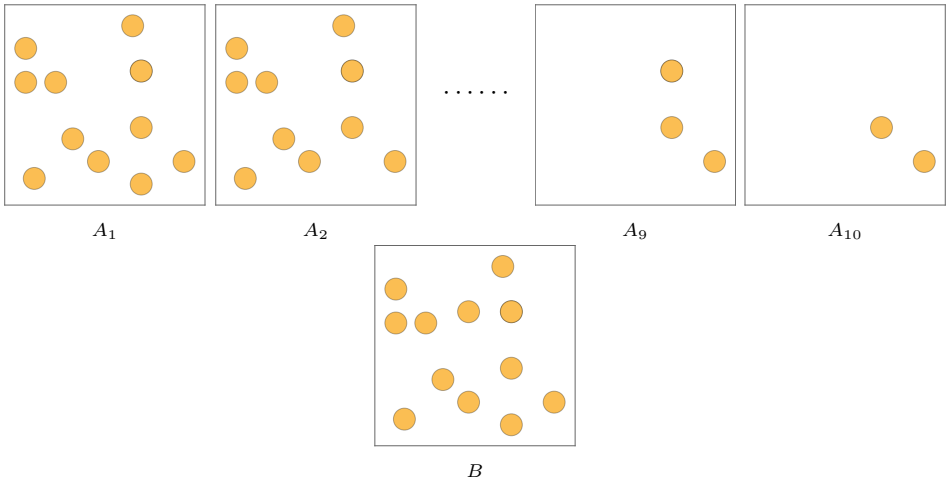
The constant marginal sensitivity axiom **A3** may be unappealing in dissimilarity judgments within collections of pairs of sets such as the one depicted in Figure 1. There, $B \supset A_1 \supset A_2 \supset \cdots \supset A_9 \supset A_{10}$; $|B| - |A_1| = |A_j| - |A_{j+1}| = 1$ for $j \leq 9$; and $J(A_1,B) - J(A_2,B) = J(A_9,B) - J(A_{10},B) = \frac{1}{12}$. But since $A_1$ and $A_2$ are less Jaccard-dissimilar than $A_9$ and $A_{10}$, one might have expected the dissimilarity difference between $A_1, B$ and $A_2, B$ to be perceived as being smaller than that between $A_9, B$ and $A_{10}, B$.

A simple generalization of (1) that allows for such potentially *increasing* marginal sensitivity to the gradual removal of common elements is achievable by letting the mapping $J^w : 2^X \times 2^X \to \mathbb{R}$ be defined by

$$\begin{aligned} J^w(A,B) &:= \frac{|A \cup B|^\alpha - |A \cap B|^\beta}{|A \cup B|^\alpha} \\ &= 1 - \frac{|A \cap B|^\beta}{|A \cup B|^\alpha}, \end{aligned} \tag{8}$$

for some $1 \geq \alpha \geq \beta > 0$, with $\alpha \geq \beta$ sufficing for $J^w$ to be non-negative.

FIGURE 1. $J$ shows constant marginal sensitivity to the removal of jointly owned elements. Here, $J(A_1, B) - J(A_2, B) = J(A_9, B) - J(A_{10}, B)$.




Moreover, it is easily seen that, whenever $\alpha = \beta \leq 1$, $J^w$ obeys the first two of the next three additional conditions that are satisfied by $J$ (and hence are implied by **A1**-**A3**):

**A4.** $R(A, B) \in [0, 1]$.

**A5.** $R(A, B) = 1 \iff A \cap B = \emptyset$.

**A6.** $A \cap B = \emptyset \implies R(A, A \cup B) + R(B, A \cup B) = 1$.

**A4** is a boundedness property with an obvious (dis)similarity meaning. In conjunction with the normalization imposed by **A4**, **A5** can be intuitively thought of as requiring that any two sets be maximally dissimilar if and only if they have nothing in common, and **A6** as demanding that there be maximum cumulative dissimilarity between each of two disjoint sets and their union.

Both **A3** and **A6** are violated by $J^w$ whenever $\alpha \neq 1 \neq \beta$. However, $J^w$ obeys the following *mid-point convexity* generalization of **A3** and *sub-additivity* generalization of **A6** if $\alpha = \beta \leq 1$:

**A3′.** $A \not\ni x, y \in B \implies R(A, B) - R(A \cup \{x\}, B) \geq R(A \cup \{x\}, B) - R(A \cup \{x, y\}, B)$.

**A6′.** $A \cap B = \emptyset \implies R(A, A \cup B) + R(B, A \cup B) \leq 1$.

**Proposition 2.** *$J^w \equiv J \iff \alpha = \beta = 1$. Moreover, if $\alpha = \beta \leq 1$, then $J^w$ implies* **A1**, **A2**, **A3′**, **A4**, **A5**, **A6′**. *The converse of this statement is false.*

We note that $R_3$ in the proof of Proposition 1 satisfies the six axioms of Proposition 2 and coincides with the specific $J^w$ where $\alpha = \beta = \frac{\ln(3) - \ln(2)}{\ln(2)} \approx 0.5849$. On the slightly richer set $X := \{a, b, c\}$, however, we provide below an example of

some $R$ that satisfies these axioms but is not a $J^w$ function, thereby also proving the non-obvious part of Proposition 2.

Indeed, consider $X := \{a, b, c\}$, $2^X \equiv \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$ and, excluding again the pair $(\emptyset, \emptyset)$, let $\widehat{R} : 2^X \times 2^X \to \mathbb{R}$ be defined as follows:

$$\begin{array}{rcll} \widehat{R}(A,B) & = & 0 & \iff \quad A = B, \\ \widehat{R}(A,B) & = & 1 & \iff \quad A \cap B = \emptyset, \\ \widehat{R}(A,B) & = & \widehat{R}(B,A) & \quad \forall\, A, B \in 2^X, \\ \widehat{R}(\{x,y\},\{y,z\}) & = & \dfrac{1}{2} & \quad \forall\, x, y, z \in X : x \neq y \neq z \neq x, \\ \widehat{R}(\{x\},\{x,y\}) & = & \dfrac{1}{3} & \quad \forall\, x, y \in X : x \neq y, \\ \widehat{R}(\{x\}, X) & = & \dfrac{1}{4} & \quad \forall\, x \in X : \\ \widehat{R}(\{x,y\}, X) & = & \dfrac{1}{8} & \quad \forall\, x, y \in X : x \neq y. \end{array}$$

We leave it as an exercise for the reader to verify that $\widehat{R}$ violates **A3** and **A6** but satisfies all axioms appearing in the statement of Proposition 2. That there is no $\alpha \in (0, 1)$ such that $\widehat{R} \equiv J^w$ under $\alpha$ can be seen, for example, by observing that $\widehat{R}(\{a\}, \{a, b\}) = \frac{1}{3}$ is uniquely compatible with $J^w$ under the $\alpha$ associated with function $R_3$ above, whereas $\widehat{R}(\{a, b\}, \{b, c\}) = \frac{1}{2}$ is so under the distinct $\alpha' = \frac{\ln(2)}{\ln(3)}$. Yet, $J^w$ requires the value of this scalar to be invariant across all pairs.

One may wonder whether, in addition to **A3** and **A6**, $J^w$ generally violates the triangle inequality too when $\alpha = \beta < 1$:

**A7.** $R(A, B) + R(B, C) \geq R(A, C)$.

The positive answer to this question can be confirmed, for example, at $\alpha = \beta := 0.5$, $A := \{1, 2, 3, 4\}$, $B := \{3, 4, 5, 6, 7\}$ and $C := \{5, 6, 7, 8\}$. This fact and Proposition 2 now imply that $J^w$ with $\alpha = \beta < 1$ becomes a *quasi-metric* that features increasing marginal sensitivity to the gradual removal of common elements. We can use the axioms introduced above to compare the structure of this class of dissimilarity quasi-metrics to other such functions in the existing literature.

To this end, we recall the Sørensen-Dice ($SD$), Salton cosine ($SC$) and *Overlap* dissimilarity functions (see [15] and references therein), which are defined by

$$SD(A,B) := 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{9}$$

$$SC(A,B) := 1 - \frac{|A \cap B|}{\sqrt{|A| \times |B|}} \tag{10}$$

$$Overlap(A,B) := 1 - \frac{|A \cap B|}{\min\{|A|, |B|\}} \tag{11}$$

As can be easily checked, all three satisfy **A1**, **A2**, **A4**, **A5** and violate **A3**, **A6**, **A7**. Furthermore, $SD$ and $SC$ satisfy **A3′**, whereas *Overlap* violates this axiom (e.g. at $A = \{a, b\}$ and $B = \{b, c, d\}$ with $x = c$, $y = d$). Finally, they all satisfy **A6′**, and *Overlap* does so trivially (the relevant sum is either 0 or 1). This information is summarized in Table 1.

TABLE 1. Axioms satisfied or violated by the different dissimilarity functions.

| | **A1** | **A2** | **A3** | **A3′** | **A4** | **A5** | **A6** | **A6′** | **A7** |
|---|---|---|---|---|---|---|---|---|---|
| $J$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $J^w$ | ✓ | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | × |
| $SD$ | ✓ | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | × |
| $SC$ | ✓ | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | × |
| $Overlap$ | ✓ | ✓ | × | × | ✓ | ✓ | × | ✓ | × |

Despite $SD$, $SC$ and $J^w$ sharing the same features as far as these 9 axioms are concerned, it is straightforward to see that the values of different pairs of sets $(A, B)$, $(A', B')$ under each of the terms that appear with a negative sign on the right hand sides of (9) and (10) are generally distinct from those under the corresponding term in the single-parameter specification of (8) with a fixed $\alpha \in (0, 1)$. Thus, (8) and (9)-(11) are non-nested classes of dissimilarity quasi-metrics, with the one defined by (8) when $\alpha = \beta \leq 1$ apparently being a novel addition to that category. A complete characterization of this general class is left as an open problem.

## REFERENCES


1. Gilbert, G. F. (1884) Finley's Tornado Predictions. *Amer. Meteor. J.* 1(5):166-172.
2. Jaccard, P. (1901) Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37(142):547–579.
3. Rogers, D. J. and Tanimoto, T. T. (1960) A Computer Program for Classifying Plants. *Science.* 132(3434):1115-1118.
4. Levandowsky, M. and Winter, D. (1971) Distance Between Sets. *Nature.* 234(5):34–35.
5. Tversky, A. (1977) Features of Similarity. *Psychol. Rev.* 84(4):327-352.
6. Schaefer, J. T. (1990) The Critical Success Index as an Indicator of Warning Skill. *Weather Forecast.* 5(4):570-575.
7. Lipkus, A. H. (1999). A Proof of the Triangle Inequality for the Tanimoto Distance. *J. Math. Chem.* 26:263–265.
8. Bertoluzza, C., Di Bacco, M. and Doldi, V. (2004) An Axiomatic Characterization of the Measures of Similarity. *Sankhyã.* 66(3):474-486.
9. Azaele, S., Muneepeerakul, R. Maritan, A. and Rodriguez-Iturbe, I. (2009) Predicting Spatial Similarity of Freshwater Fish Biodiversity. *Proc. Natl. Acad. Sci. U.S.A.* 106(17):7058-7062.
10. Vorontsov, I. E., Kulakovskiy, I. V. and Makeev, V. J. (2013) Jaccard Index Based Similarity Measure to Compare Transcription Factor Binding Site Models. *Algorithms Mol. Biol.* 8:23.
11. Grygorian, A. and Iacob, I. E. (2018) A Concise Proof of the Triangle Inequality for the Jaccard Distance. *Coll. Math. J.*. 49(5):363-365.
12. Kosub, S. (2019) A Note on the Triangle Inequality for the Jaccard Distance. *Pattern Recognit. Lett.* 120:36-38.
13. Bag, S., Kumar, S. K. and Tiwari, M. K. (2019) An Efficient Recommendation Generation Using Relevant Jaccard similarity. *Inf. Sci.* 483:53-64.
14. Besta, M., Kanakagiri, R., Mustafa, H., Karasikov, M., Rätsch, G., Hoefler, T. and Solomonik, E. (2020) Communication-Efficient Jaccard similarity for High-Performance Distributed Genome Comparisons. *IEEE Trans. Parallel Distrib. Syst.* 2020:1122-1132.
15. Verma, V. and Aggarwal, R. K. (2020) A Comparative Analysis of Similarity Measures Akin to the Jaccard Index in Collaborative Recommendations: Empirical and Theoretical Perspective. *Soc. Netw. Anal. Min.* 10:43.
16. Bouyssou, D., Marchant, T. and Pirlot, M. (2022) Axiomatic Characterization of the $\chi^2$ Dissimilarity Measure. *Aequat. Math.* 96:307–323.



UNIVERSITY OF GLASGOW
*Email address*: Georgios.Gerasimou@glasgow.ac.uk