



# Jaccard index

From Wikipedia, the free encyclopedia



This article includes a [list of references](#), but **its sources remain unclear because it has insufficient [inline citations](#)**. Please help to [improve](#) this article by [introducing](#) more precise citations. *(March 2011)* ([Learn how and when to remove this template message](#))

The **Jaccard index**, also known as **Intersection over Union** and the **Jaccard similarity coefficient** (originally given the French name *coefficient de communauté* by [Paul Jaccard](#)), is a [statistic](#) used for gauging the [similarity](#) and [diversity](#) of [sample](#) sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the [intersection](#) divided by the size of the [union](#) of the sample sets:

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

(If *A* and *B* are both empty, define  $J(A, B) = 1$ .)

$$0 \leq J(A, B) \leq 1.$$

The **Jaccard distance**, which measures *dissimilarity* between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

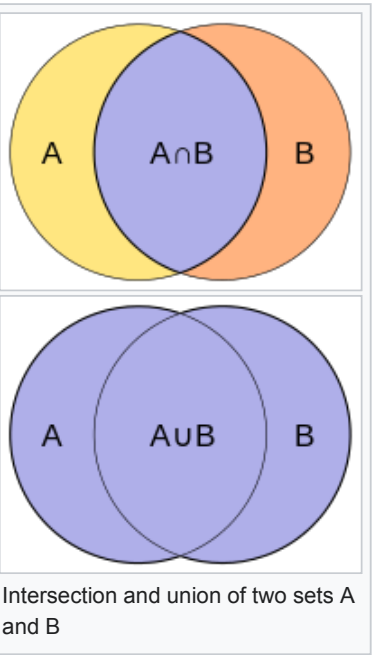
$$d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

An alternative interpretation of the Jaccard distance is as the ratio of the size of the [symmetric difference](#)  $A \Delta B = (A \cup B) - (A \cap B)$  to the union. Jaccard distance is commonly used to calculate an  $n \times n$  matrix for [clustering](#) and [multidimensional scaling](#) of  $n$  sample sets.

This distance is a [metric](#) on the collection of all finite sets.<sup>[1][2][3]</sup>

There is also a version of the Jaccard distance for [measures](#), including [probability measures](#). If  $\mu$  is a measure on a [measurable space](#)  $X$ , then we define the Jaccard coefficient by

$$J_\mu(A, B) = \frac{\mu(A \cap B)}{\mu(A \cup B)},$$



and the Jaccard distance by

$$d_{\mu}(A,B) = 1 - J_{\mu}(A,B) = \frac{\mu(A\Delta B)}{\mu(A\cup B)}.$$

Care must be taken if  $\mu(A\cup B) = 0$  or  $\infty$ , since these formulas are not well defined in these cases.

The [MinHash](#) min-wise independent permutations [locality sensitive hashing](#) scheme may be used to efficiently compute an accurate estimate of the Jaccard similarity coefficient of pairs of sets, where each set is represented by a constant-sized signature derived from the minimum values of a [hash function](#).

Contents [\[hide\]](#)

- 1 Similarity of asymmetric binary attributes
- 1.1 Difference with the simple matching coefficient (SMC)
- 2 Weighted Jaccard similarity and distance
- 3 Probability Jaccard similarity and distance
- 3.1 Optimality of the Probability Jaccard Index
- 4 Tanimoto similarity and distance
- 4.1 Tanimoto's definitions of similarity and distance
- 4.2 Other definitions of Tanimoto distance
- 5 See also
- 6 Notes
- 7 References
- 8 External links

## Similarity of asymmetric binary attributes [\[edit\]](#)

Given two objects, *A* and *B*, each with *n* [binary](#) attributes, the Jaccard coefficient is a useful measure of the overlap that *A* and *B* share with their attributes. Each attribute of *A* and *B* can either be 0 or 1. The total number of each combination of attributes for both *A* and *B* are specified as follows:

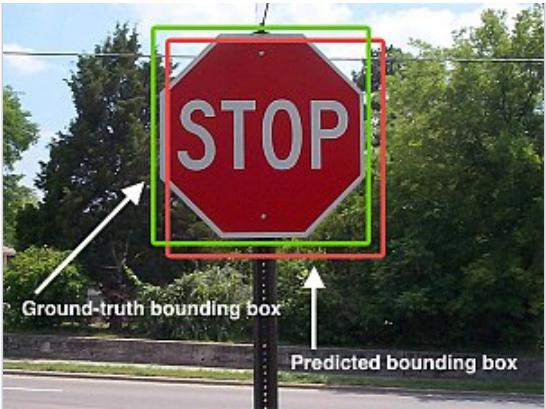
- $M_{11}$  represents the total number of attributes where *A* and *B* both have a value of 1.
- $M_{01}$  represents the total number of attributes where the attribute of *A* is 0 and the attribute of *B* is 1.
- $M_{10}$  represents the total number of attributes where the attribute of *A* is 1 and the attribute of *B* is 0.
- $M_{00}$  represents the total number of attributes where *A* and *B* both have a value of 0.

Each attribute must fall into one of these four categories, meaning that

$$M_{11} + M_{01} + M_{10} + M_{00} = n.$$

The Jaccard similarity coefficient, *J*, is given as

A	
0	1



$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Intersection over Union as a similarity measure for [object detection](#) on images - an important task in [computer vision](#).

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

The Jaccard distance,  $d_J$ , is given as

$$d_J = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}} = 1 - J.$$

<b>B</b>	0	$M_{00}$	$M_{10}$
	1	$M_{01}$	$M_{11}$

### Difference with the simple matching coefficient (SMC) [\[ edit \]](#)

When used for binary attributes, the Jaccard index is very similar to the [simple matching coefficient](#). The main difference is that the SMC has the term  $M_{00}$  in its numerator and denominator, whereas the Jaccard index does not. Thus, the SMC counts both mutual presences (when an attribute is present in both sets) and mutual absence (when an attribute is absent in both sets) as matches and compares it to the total number of attributes in the universe, whereas the Jaccard index only counts mutual presence as matches and compares it to the number of attributes that have been chosen by at least one of the two sets.

In [market basket analysis](#), for example, the basket of two consumers who we wish to compare might only contain a small fraction of all the available products in the store, so the SMC will usually return very high values of similarities even when the baskets bear very little resemblance, thus making the Jaccard index a more appropriate measure of similarity in that context. For example, consider a supermarket with 1000 products and two customers. The basket of the first customer contains salt and pepper and the basket of the second contains salt and sugar. In this scenario, the similarity between the two baskets as measured by the Jaccard index would be 1/3, but the similarity becomes 0.998 using the SMC.

In other contexts, where 0 and 1 carry equivalent information (symmetry), the SMC is a better measure of similarity. For example, vectors of demographic variables stored in [dummy variables](#), such as gender, would be better compared with the SMC than with the Jaccard index since the impact of gender on similarity should be equal, independently of whether male is defined as a 0 and female as a 1 or the other way around. However, when we have symmetric dummy variables, one could replicate the behaviour of the SMC by splitting the dummies into two binary attributes (in this case, male and female), thus transforming them into asymmetric attributes, allowing the use of the Jaccard index without introducing any bias. The SMC remains, however, more computationally efficient in the case of symmetric dummy variables since it does not require adding extra dimensions.

### Weighted Jaccard similarity and distance [\[ edit \]](#)

If  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  and  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  are two vectors with all real  $x_i, y_i \geq 0$ , then their Jaccard similarity coefficient (also known then as Ruzicka similarity) is defined as

$$J_{\mathcal{W}}(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)},$$

and Jaccard distance (also known then as Soergel distance)

$$d_{J\mathcal{W}}(\mathbf{x}, \mathbf{y}) = 1 - J_{\mathcal{W}}(\mathbf{x}, \mathbf{y}).$$

With even more generality, if  $f$  and  $g$  are two non-negative measurable functions on a measurable space  $X$  with measure  $\mu$ , then we can define

$$J_{\mathcal{W}}(f, g) = \frac{\int \min(f, g) d\mu}{\int \max(f, g) d\mu},$$

where **max** and **min** are pointwise operators. Then Jaccard distance is

$$d_{\mathcal{W}}(f, g) = 1 - J_{\mathcal{W}}(f, g).$$

Then, for example, for two measurable sets  $A, B \subseteq X$ , we have  $J_{\mu}(A, B) = J(\chi_A, \chi_B)$ , where  $\chi_A$  and  $\chi_B$  are the characteristic functions of the corresponding set.

## Probability Jaccard similarity and distance [\[edit\]](#)

The weighted Jaccard similarity described above generalizes the Jaccard Index to positive vectors, where a set corresponds to a binary vector given by the [indicator function](#), i.e.  $\mathbf{x}_i \in \{0, 1\}$ . However, it does not generalize the Jaccard Index to probability distributions, where a set corresponds to a uniform probability distribution, i.e.

$$\mathbf{x}_i = \begin{cases} \frac{1}{|X|} & i \in X \\ 0 & \text{otherwise} \end{cases}$$

It is always less if the sets differ in size. If  $|X| > |Y|$ , and  $\mathbf{x}_i = \mathbf{1}_X(i)/|X|, \mathbf{y}_i = \mathbf{1}_Y(i)/|Y|$  then

$$J_{\mathcal{W}}(\mathbf{x}, \mathbf{y}) = \frac{|X \cap Y|}{|X \setminus Y| + |X|} < J(X, Y).$$

Instead, a generalization that is continuous between probability distributions and their corresponding support sets is

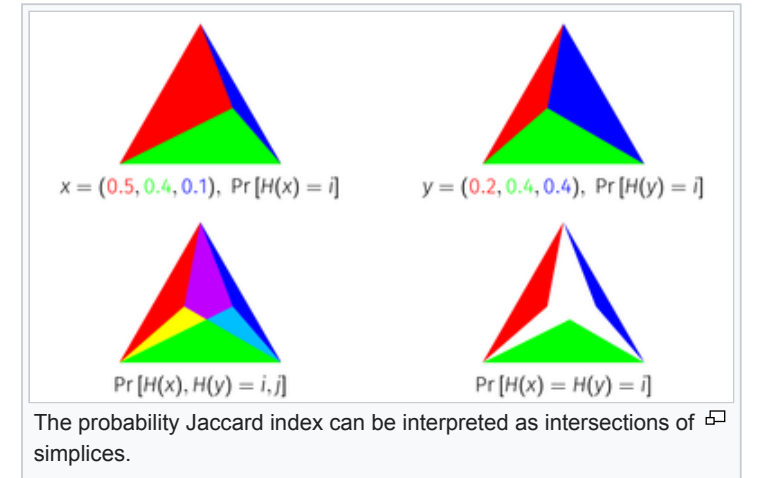
$$J_{\mathcal{P}}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}_i \neq 0, \mathbf{y}_i \neq 0} \frac{1}{\sum_j \max\left(\frac{x_j}{x_i}, \frac{y_j}{y_i}\right)}$$

which is called the "Probability" Jaccard.<sup>[4]</sup> It has the following bounds against the Weighted Jaccard on probability vectors.

$$J_{\mathcal{W}}(\mathbf{x}, \mathbf{y}) \leq J_{\mathcal{P}}(\mathbf{x}, \mathbf{y}) \leq \frac{2J_{\mathcal{W}}(\mathbf{x}, \mathbf{y})}{1 + J_{\mathcal{W}}(\mathbf{x}, \mathbf{y})}$$

Here the upper bound is the (weighted) [Sørensen–Dice coefficient](#). The corresponding distance,  $1 - J_{\mathcal{P}}(\mathbf{x}, \mathbf{y})$ , is a metric over probability distributions, and a [pseudo-metric](#) over non-negative vectors.

The Probability Jaccard Index has a geometric interpretation as the area of an intersection of [simplices](#). Every point on a unit  $k$ -simplex corresponds to a probability distribution on  $k + 1$  elements, because the unit  $k$ -simplex is the set of points in  $k + 1$  dimensions that sum to 1. To derive the Probability Jaccard Index geometrically, represent a probability distribution as the unit simplex divided into sub simplices according to the mass of each item. If you overlay two distributions represented in this way on top of each other, and intersect the simplices corresponding to each item, the area that remains is equal to the Probability Jaccard Index of the distributions.



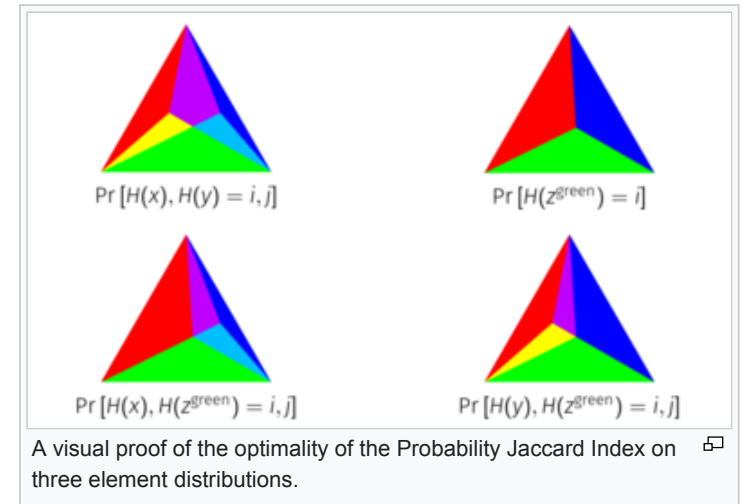
## Optimality of the Probability Jaccard Index [\[ edit \]](#)

Consider the problem of constructing random variables such that they collide with each other as much as possible. That is, if  $X \sim x$  and  $Y \sim y$ , we would like to construct  $X$  and  $Y$  to maximize  $\Pr[X = Y]$ . If we look at just two distributions  $x, y$  in isolation, the highest  $\Pr[X = Y]$  we can achieve is given by  $1 - \text{TV}(x, y)$  where **TV** is the [Total Variation distance](#). However, suppose we weren't just concerned with maximizing that particular pair, suppose we would like to maximize the collision probability of any arbitrary pair. One could construct an infinite number of random variables one for each distribution  $x$ , and seek to maximize  $\Pr[X = Y]$  for all pairs  $x, y$ . In a fairly strong sense described below, the Probability Jaccard Index is an optimal way to align these random variables.

For any sampling method  $G$  and discrete distributions  $x, y$ , if  $\Pr[G(x) = G(y)] > J_{\mathcal{P}}(x, y)$  then for some  $z$  where  $J_{\mathcal{P}}(x, z) > J_{\mathcal{P}}(x, y)$  and  $J_{\mathcal{P}}(y, z) > J_{\mathcal{P}}(x, y)$ , either  $\Pr[G(x) = G(z)] < J_{\mathcal{P}}(x, z)$  or  $\Pr[G(y) = G(z)] < J_{\mathcal{P}}(y, z)$ .<sup>[4]</sup>

That is, no sampling method can achieve more collisions than  $J_{\mathcal{P}}$  on one pair without achieving fewer collisions than  $J_{\mathcal{P}}$  on another pair, where the reduced pair is more similar under  $J_{\mathcal{P}}$  than the increased pair. This theorem is true for the Jaccard Index of sets (if interpreted as uniform distributions) and the probability Jaccard, but not of the weighted Jaccard. (The theorem uses the word "sampling method" to describe a joint distribution over all distributions on a space, because it derives from the use of [weighted minhashing algorithms](#) that achieve this as their collision probability.)

This theorem has a visual proof on three element distributions using the simplex representation.



## Tanimoto similarity and distance [\[ edit \]](#)

Various forms of functions described as Tanimoto similarity and Tanimoto distance occur in the literature and on the Internet. Most of these are synonyms for Jaccard similarity and Jaccard distance, but some are mathematically different. Many sources<sup>[5]</sup> cite an IBM Technical Report<sup>[6]</sup> as the seminal reference. The report is available from [several libraries](#)<sup>[7]</sup>.

In "A Computer Program for Classifying Plants", published in October 1960,<sup>[7]</sup> a method of classification based on a similarity ratio, and a derived distance function, is given. It seems that this is the most authoritative source for the meaning of the terms "Tanimoto similarity" and "Tanimoto Distance". The similarity ratio is equivalent to Jaccard similarity, but the distance function is *not* the same as Jaccard distance.

## Tanimoto's definitions of similarity and distance [\[ edit \]](#)

In that paper, a "similarity ratio" is given over [bitmaps](#), where each bit of a fixed-size array represents the presence or absence of a characteristic in the plant being modelled. The definition of the ratio is the number of common bits, divided by the number of bits set (*i.e.* nonzero) in either sample.

Presented in mathematical terms, if samples  $X$  and  $Y$  are bitmaps,  $X_i$  is the  $i$ th bit of  $X$ , and  $\wedge, \vee$  are [bitwise and, or](#) operators respectively, then the similarity ratio  $T_s$  is

$$T_s(X, Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

If each sample is modelled instead as a set of attributes, this value is equal to the Jaccard coefficient of the two sets. Jaccard is not cited in the paper, and it seems likely that the authors were not aware of it.

Tanimoto goes on to define a "distance coefficient" based on this ratio, defined for bitmaps with non-zero similarity:

$$T_d(X, Y) = -\log_2(T_s(X, Y))$$

This coefficient is, deliberately, not a distance metric. It is chosen to allow the possibility of two specimens, which are quite different from each other, to both be similar to a third. It is easy to construct an example which disproves the property of [triangle inequality](#).

### Other definitions of Tanimoto distance [\[ edit \]](#)

Tanimoto distance is often referred to, erroneously, as a synonym for Jaccard distance  $1 - T_s$ . This function is a proper distance metric. "Tanimoto Distance" is often stated as being a proper distance metric, probably because of its confusion with Jaccard distance.

If Jaccard or Tanimoto similarity is expressed over a bit vector, then it can be written as

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

where the same calculation is expressed in terms of vector scalar product and magnitude. This representation relies on the fact that, for a bit vector (where the value of each dimension is either 0 or 1) then

$$A \cdot B = \sum_i A_i B_i = \sum_i (A_i \wedge B_i)$$

and

$$|A|^2 = \sum_i A_i^2 = \sum_i A_i.$$

This is a potentially confusing representation, because the function as expressed over vectors is more general, unless its domain is explicitly restricted. Properties of  $T_s$  do not necessarily extend to  $f$ . In particular, the difference function  $1 - f$  does not preserve [triangle inequality](#), and is not therefore a proper distance metric, whereas  $1 - T_s$  is.

There is a real danger that the combination of "Tanimoto Distance" being defined using this formula, along with the statement "Tanimoto Distance is a proper distance metric" will lead to the false conclusion that the function  $1 - f$  is in fact a distance metric over vectors or [multisets](#) in general, whereas its use in similarity search or clustering algorithms may fail to produce correct results.

Lipkus<sup>[2]</sup> uses a definition of Tanimoto similarity which is equivalent to  $f$ , and refers to Tanimoto distance as the function  $1 - f$ . It is, however, made clear within the paper that the context is restricted by the use of a (positive) weighting vector  $W$  such that, for any vector  $A$  being considered,  $A_i \in \{0, W_i\}$ . Under these circumstances, the function is a proper distance metric, and so a set of vectors governed by such a weighting vector forms a [metric space](#) under this function.

## See also [ edit ]

- [Overlap coefficient](#)
- [Simple matching coefficient](#)
- [Most frequent k characters](#)
- [Hamming distance](#)
- [Sørensen–Dice coefficient](#), which is equivalent:  $J = S/(2 - S)$  and  $S = 2J/(1 + J)$  ( $J$ : Jaccard index,  $S$ : Sørensen–Dice coefficient)
- [Tversky index](#)
- [Correlation](#)
- [Mutual information](#), a normalized [metricated](#) variant of which is an entropic Jaccard distance.


## Notes [ edit ]

- <sup>^</sup> Kosub, Sven; "A note on the triangle inequality for the Jaccard distance" [arXiv:1612.02696](#)
- <sup>^</sup> <sup>[a](#)</sup> <sup>[b](#)</sup> Lipkus, Alan H. (1999), "A proof of the triangle inequality for the Tanimoto distance", *Journal of Mathematical Chemistry*, **26** (1–3): 263–265, doi:[10.1023/A:1019154432472](#)
- <sup>^</sup> Levandowsky, Michael; Winter, David (1971), "Distance between sets", *Nature*, **234** (5): 34–35, doi:[10.1038/234034a0](#)
- <sup>^</sup> <sup>[a](#)</sup> <sup>[b](#)</sup> Moulton, Ryan; Jiang, Yunjiang (2018), "Maximally Consistent Sampling and the Jaccard Index of Probability Distributions", *International Conference on Data Mining, Workshop on High Dimensional Data Mining*: 347–356, [arXiv:1809.04052](#), doi:[10.1109/ICDM.2018.00050](#), ISBN 978-1-5386-9159-5
- <sup>^</sup> For example Qian, Huihuan; Wu, Xinyu; Xu, Yangsheng (2011). *Intelligent Surveillance Systems*. Springer. p. 161. ISBN 978-94-007-1137-2.
- <sup>^</sup> Tanimoto, Taffee T. (17 Nov 1958). "An Elementary Mathematical theory of Classification and Prediction". *Internal IBM Technical Report*. **1957** (?).
- <sup>^</sup> Rogers, David J.; Tanimoto, Taffee T. (1960). "A Computer Program for Classifying Plants". *Science*. **132** (3434): 1115–1118. doi:[10.1126/science.132.3434.1115](#). PMID 17790723.

## References [ edit ]

- Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2005), *Introduction to Data Mining*, ISBN 0-321-32136-7
- Jaccard, Paul (1901), "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin de la Société vaudoise des sciences naturelles*, **37**: 547–579
- Jaccard, Paul (1912), "The Distribution of the flora in the alpine zone", *New Phytologist*, **11** (2): 37–50, doi:[10.1111/j.1469-8137.1912.tb05611.x](#)

## External links [ edit ]

- [Introduction to Data Mining lecture notes from Tan, Steinbach, Kumar](#) 
- [SimMetrics](#) a sourceforge implementation of Jaccard index and many other similarity metrics
- [A web-based calculator for finding the Jaccard Coefficient](#)
- [Tutorial on how to calculate different similarities](#)
- [Intersection over Union \(IoU\) for object detection](#)
- [Kaggle Dstl Satellite Imagery Feature Detection - Evaluation](#)



This page was last edited on 7 June 2020, at 11:24 (UTC).

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#).  
Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.

[Privacy policy](#) [About Wikipedia](#) [Disclaimers](#) [Contact Wikipedia](#) [Developers](#) [Statistics](#) [Cookie statement](#) [Mobile view](#)