A 1-parameter family of metrics connecting the Jaccard distance and the Normalized Information Distance

Bjørn Kjos-Hanssen
Saroj Niraula, Soowhan Yoon
Sabrina Hardisty, Guanhong Li, Jacqueline Millard
bjoern.kjos-hanssen@hawaii.edu
{sniraula|syoon2}@hawaii.edu
{shardist|guanhong|jmillard}@hawaii.edu
University of Hawai'i at Mānoa
Honolulu, USA

ABSTRACT

Jiménez, Becerra, and Gelbukh (2013) defined a family of symmetric Tversky ratio models S parametrized by $0 \le \alpha \le 1$ and $\beta > 0$. Letting $D = 1 - S$ we have a semimetric which we show is a metric if and only if $0 \le \alpha \le \frac{1}{2}$ and $\beta \ge 1/(1-\alpha)$.

For $\beta = 1/(1-\alpha)$, the two endpoints $\alpha = 0, \frac{1}{2}$ correspond to the normalized information distance (NID) and the Jaccard distance, respectively.

This result, realizing Jaccard distance and NID as the endpoints of a family of metrics, is formally verified in the Lean proof assistant.

CCS CONCEPTS

• **Theory of computation** → *Computational geometry*; • **Applied computing** → *Bioinformatics*.

KEYWORDS

Jaccard distance, normalized information distance, proof assistants,alignment-free methods

ACM Reference Format:
Bjørn Kjos-Hanssen, Saroj Niraula, Soowhan Yoon, and Sabrina Hardisty, Guanhong Li, Jacqueline Millard. 2018. A 1-parameter family of metrics connecting the Jaccard distance and the Normalized Information Distance. In *KDD '21: Knowledge Discovery and Data Mining, 2021, Online.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Distance metrics are used in a wide variety of scientific contexts. In bioinformatics, M. Li, Badger, Chen, Kwong, and Kearney [10] introduced an information-based sequence distance. In an information-theoretical setting, M. Li, Chen, X. Li, Ma and Vitányi [11] rejected the distance of [10] in favor of a *normalized information distance* (NID). The Encyclopedia of Distances [1] describes the NID on page 205 out of 583, as

$$\frac{\max\{K(x \mid y^*), K(y \mid x^*)\}}{\max\{K(x), K(y)\}}$$

where $K(x \mid y^*)$ is the Kolmogorov complexity of x given a shortest program y^* to compute y. It is equivalent to be given y itself in hard-coded form:

$$\frac{\max\{K(x \mid y), K(y \mid x)\}}{\max\{K(x), K(y)\}}$$

Another formulation sometimes used is

$$\frac{K(x,y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}.$$

The fact that the NID is in a sense a normalized metric is proved in [11]. Then in 2017, while studying malware detection, Raff and Nicholas [12] suggested Lempel–Ziv Jaccard distance (LZJD) as a practical alternative to NID. As we shall see, this is a metric. In a way this constitutes a full circle: the distance in [10] is itself essentially a Jaccard distance, and the LZJD is related to it as Lempel–Ziv complexity is to Kolmogorov complexity. In the present paper we aim to shed light on this back-and-forth by showing that the NID and Jaccard distances constitute the endpoints of a parametrized family of metrics.

For comparison, the Jaccard distance between two sets X and Y, and our analogue of the NID, are

$$\frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|} = 1 - \frac{|X \cap Y|}{|X \cup Y|}, \quad \text{and} \tag{1.0.1}$$

$$\frac{\max\{|X \setminus Y|, |Y \setminus X|\}}{\max\{|X|, |Y|\}}, \tag{1.0.2}$$

respectively. Our main result Theorem 10 shows which interpolations between these two are metrics.

The way we arrived at (1.0.2) as an analogue of NID is via Lempel–Ziv complexity. While there are several variants [8, 16, 17], the LZ 1978 complexity [17] of a sequence is the cardinality of a certain set, the dictionary. It will not be used in our paper except in conferring the spirit of Kolmogorov complexity onto set distances by suggesting the following notation.

Definition 1. *Let* $\mathrm{LZSet}(A)$ *be the Lempel–Ziv dictionary for a sequence* A*. We define LZ–Jaccard distance* LZJD *by*

$$\mathrm{LZJD}(A,B) = 1 - \frac{|\mathrm{LZSet}(A) \cap \mathrm{LZSet}(B)|}{|\mathrm{LZSet}(A) \cup \mathrm{LZSet}(B)|}.$$

KDD '21, 2021, Online

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Reference	Jaccard notation	NID notation
[10]	d	
[11]	d_s	d
[7]	D	D'
[12]	LZJD	NCD

Table 1: Overview of notation used in the literature. (It seems that authors use simple names for their favored notions.)

It is shown in [10, Theorem 1] that the triangle inequality holds for a function which they call an information-based sequence distance. Later papers give it the notation d_s in [11, Definition V.1], and call their normalized information distance d. Raff and Nicholas [12] introduced the LZJD and did not discuss the appearance of d_s in [11, Definition V.1], even though they do cite [11] (but not [10]).

Kraskov et al. [6][7] use D and D' for continuous analogues of d_s and d in [11] (which they cite). The *Encyclopedia* calls it the normalized information metric,

$$\frac{H(X \mid Y) + H(X \mid Y)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)}$$

or Rajski distance [13].

This d_s was called d by [10] — see Table 1. Conversely, [11, near Definition V.1] mentions mutual information.

A more general setting is that of STRM (Symmetric Tversky Ratio Models), Definition 9. These are variants of the Tversky index (Definition 3) proposed in [4].

Definition 2. *A* semimetric *on $\mathcal{X}$ is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the first three axioms of a metric space, but not necessarily the triangle inequality: $d(x,y) \geq 0$, $d(x,y) = 0$ if and only if $x = y$, and $d(x,y) = d(y,x)$ for all $x, y \in \mathcal{X}$.*

Definition 3. *For sets X and Y the Tversky index with parameters $\alpha, \beta \geq 0$ is a number between 0 and 1 given by*

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X \setminus Y| + \beta|Y \setminus X|}.$$

We also define the corresponding Tversky dissimilarity $d^T_{\alpha,\beta}$ by

$$d^T_{\alpha,\beta}(X,Y) = \begin{cases} 1 - S(X,Y) & \text{if } X \cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

To motivate Definition 2, we include the following lemma without proof.

Lemma 4. *Suppose d is a metric on a collection of nonempty sets $\mathcal{X}$, with $d(X,Y) \leq 2$ for all $X, Y \in \mathcal{X}$. Let $\hat{\mathcal{X}} = \mathcal{X} \cup \{\emptyset\}$ and define $\hat{d} : \hat{\mathcal{X}} \times \hat{\mathcal{X}} \to \mathbb{R}$ by stipulating that for $X, Y \in \mathcal{X}$,*

$$\hat{d}(X,Y) = d(X,Y); \quad d(X,\emptyset) = 1 = d(\emptyset,X); \quad d(\emptyset,\emptyset) = 0.$$

Then $\hat{d}$ is a metric on $\hat{\mathcal{X}}$.

Theorem 5 (Gragera and Suppakitpaisarn [2, 3]). *The optimal constant ρ such that $d^T_{\alpha,\beta}(X,Y) \leq \rho(d^T_{\alpha,\beta}(X,Y) + d^T_{\alpha,\beta}(Y,Z))$ for all X, Y, Z is*

$$\frac{1}{2}\left(1 + \sqrt{\frac{1}{\alpha\beta}}\right).$$

Corollary 6. *$d^T_{\alpha,\beta}$ is a metric only if $\alpha = \beta \geq 1$.*

Proof. Clearly, $\alpha = \beta$ is necessary to ensure $d^T_{\alpha,\beta}(X,Y) = d^T_{\alpha,\beta}(Y,X)$. Moreover $\rho \leq 1$ is necessary, so Theorem 5 gives $\alpha\beta \geq 1$. □

Definition 7. *The Szymkiewicz—Simpson coefficient is defined by*

$$\text{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

We may note that $\text{overlap}(X,Y) = 1$ whenever $X \subseteq Y$ or $Y \subseteq X$, so that $1 - \text{overlap}$ is not a metric.

Definition 8. *The Sørensen–Dice coefficient is defined by*

$$\frac{2|X \cap Y|}{|X| + |Y|}.$$

Definition 9 ([4, Section 2]). *Let $\mathcal{X}$ be a collection of finite sets. We define $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows. For sets $X, Y \in \mathcal{X}$ we define $m(X,Y) = \min\{|X \setminus Y|, |Y \setminus X|\}$ and $M(X,Y) = \max\{|X \setminus Y|, |Y \setminus X|\}$. The symmetric TRM is defined by*

$$S(X,Y) = \frac{|X \cap Y| + \text{bias}}{|X \cap Y| + \text{bias} + \beta\,(\alpha m + (1-\alpha)M)}$$

The unbiased symmetric TRM is the case where bias $= 0$*, which is the case we shall assume we are in for the rest of this paper. The Tversky semimetric $D'_{\alpha,\beta}$ is defined by $D'_{\alpha,\beta}(X,Y) = 1 - S(X,Y)$, or more precisely*

$$D'_{\alpha,\beta} = \begin{cases} \beta\frac{\alpha m + (1-\alpha)M}{|X \cap Y| + \beta(\alpha m + (1-\alpha)M)}, & \text{if } X \cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

Note that for $\alpha = 1/2$, $\beta = 1$, the STRM is equivalent to the Sørensen–Dice coefficient. Similarly, for $\alpha = 1/2$, $\beta = 2$, it is equivalent to Jaccard's coefficient.

Our main result is (see Figure 1):

Theorem 10. *Let $0 \leq \alpha \leq 1$ and $\beta > 0$. Then $D'_{\alpha,\beta}$ is a metric if and only if $0 \leq \alpha \leq 1/2$ and $\beta \geq 1/(1-\alpha)$.*

Theorem 10 gives the converse to the Gragera and Suppakitpaisarn inspired Corollary 6:

Corollary 11. *The Tversky dissimilarity $d^T_{\alpha,\beta}$ is a metric iff $\alpha = \beta \geq 1$.*

Proof. Suppose the Tversky dissimilarity $d^T_{\alpha,\beta}$ is a semimetric. Let X, Y be sets with $|X \cap Y| = |X \setminus Y| = 1$ and $|Y \setminus X| = 0$. Then

$$1 - \frac{1}{1+\beta} = d^T_{\alpha,\beta}(Y,X) = d^T_{\alpha,\beta}(X,Y) = 1 - \frac{1}{1+\alpha},$$

hence $\alpha = \beta$. Let $\gamma = \alpha = \beta$.

Now, $d^T_{\gamma,\gamma} = D'_{\alpha_0,\beta_0}$ where $\alpha_0 = 1/2$ and $\beta_0 = 2\gamma$. Indeed, let $m = \min\{|X \setminus Y|, |Y \setminus X|\}$ and $M = \max\{|X \setminus Y|, |Y \setminus X|\}$. Since

$$D'_{\alpha_0,\beta_0} = \beta_0 \frac{\alpha_0 m + (1-\alpha_0)M}{|X \cap Y| + \beta_0\,[\alpha_0 m + (1-\alpha_0)M]},$$

$$D'_{\frac{1}{2},2\gamma} = 2\gamma \frac{\frac{1}{2}m + (1-\frac{1}{2})M}{|X \cap Y| + 2\gamma\left[\frac{1}{2}m + (1-\frac{1}{2})M\right]}$$

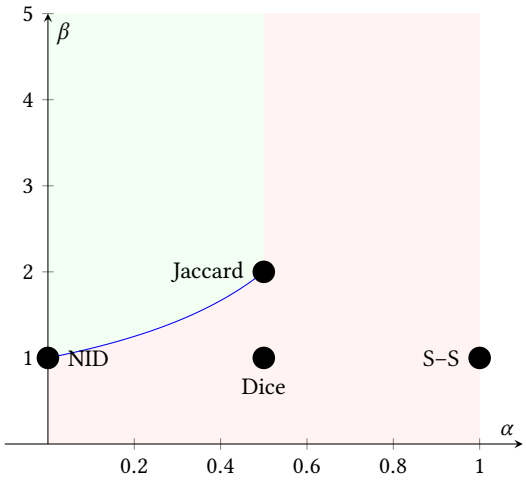

Figure 1: A Tversky semimetric $D'_{\alpha,\beta}$ is a metric if and only if (α, β) belongs to the green region. The parameter values corresponding to the Jaccard distance, the analogue of Normalized Information Distance (NID), the Sørensen–Dice semimetric, and the Szymkiewicz–Simpson semimetric are indicated.

$$= \gamma \frac{|X \setminus Y| + |Y \setminus X|}{|X \cap Y| + \gamma\,[|X \setminus Y| + |Y \setminus X|]} = 1 - \frac{|X \cap Y|}{|X \cap Y| + \gamma |X \setminus Y| + \gamma |Y \setminus X|}$$

$= d^T_{\gamma,\gamma}$. By Theorem 10, $d^T_{\gamma,\gamma}$ is a metric if and only if $\beta_0 \geq 1/(1-\alpha_0)$. This is equivalent to $2\gamma \geq 2$, i.e., $\gamma \geq 1$. □

The truth or falsity of Corollary 11 does not arise in Gragera and Suppakitpaisarn's work, as they require $\alpha, \beta \leq 1$ in their definition of Tversky index. We note that Tversky [15] only required $\alpha, \beta \geq 0$.

2 RESULTS

Lemma 12. *Let $u, v, w, \epsilon > 0$. Then*

$$\frac{1}{u} \leq \frac{1}{v} + \frac{1}{w} \implies \frac{1}{u+\epsilon} \leq \frac{1}{v+\epsilon} + \frac{1}{w+\epsilon}.$$

Proof. It is of course equivalent to show

$$vw \leq uw + uv \implies (v+\epsilon)(w+\epsilon) \leq (u+\epsilon)(w+\epsilon) + (u+\epsilon)(v+\epsilon),$$

which reduces to

$$(v+w)\epsilon \leq (u+w)\epsilon + (u+v)\epsilon + \epsilon^2,$$

which is clearly the case. □

Lemma 13. *Suppose $a(x,y) = a_{xy}$ and $b(x,y) = b_{xy}$ are functions. Suppose the function d given by $d(x,y) = a_{xy}/b_{xy}$ is a metric, and $\epsilon > 0$ is a real number. Let $\hat{d}(x,y) = \frac{a_{xy}}{b_{xy}+\epsilon a_{xy}}$. Then $\hat{d}$ is also a metric.*

Proof. The only nontrivial task is to verify the triangle inequality. Define further functions u, v, w by

$$u = b_{xy}/a_{xy}, \quad v = b_{xz}/a_{xz}, \quad w = b_{zy}/a_{zy}.$$

Since d is a metric we have

$$\frac{a_{xy}}{b_{xy}} \leq \frac{a_{xz}}{b_{xz}} + \frac{a_{zy}}{b_{zy}}$$

and hence $\frac{1}{u} \leq \frac{1}{v} + \frac{1}{w}$. We proceed by forward reasoning: we need the truth of the following equivalent conditions:

$$\frac{a_{xy}}{b_{xy}+\epsilon a_{xy}} \leq \frac{a_{xz}}{b_{xz}+\epsilon a_{xz}} + \frac{a_{zy}}{b_{zy}+\epsilon a_{zy}},$$
$$\frac{1}{u+\epsilon} \leq \frac{1}{v+\epsilon} + \frac{1}{w+\epsilon}.$$

By Lemma 12, we are done. □

Theorem 14. *For each α, the set of β for which $D'_{\alpha,\beta}$ is a metric is upward closed.*

Proof. Suppose D'_{α,β_0} is a metric and $\epsilon = \beta - \beta_0 \geq 0$. Let $a_{XY} := \alpha m(X,Y) + (1-\alpha)M(X,Y)$. Since

$$D'_{\alpha,\beta}(X,Y) = \beta \frac{a_{XY}}{|X \cap Y| + \beta a_{XY}} = \beta \frac{a_{XY}}{|X \cap Y| + \beta_0 a_{XY} + \epsilon a_{XY}},$$

and since the upfront factor of β may be removed without loss of generality, the question reduces to Lemma 13. □

Some convenient notation to be used below includes $\overline{\alpha} = 1 - \alpha$; $\gamma := \beta\alpha \leq 1$ with $\beta = 1/\overline{\alpha}$; $xy = |X \cap Y|$, $x = |X|$ etc.;

- $x_y = |X \setminus Y|$, $x_{zy} = |X \setminus (Z \cup Y)| = |(X \setminus Z) \setminus Y|$,
- $x_{000} = |\overline{X} \cap \overline{Y} \cap \overline{Z}|$, $x_{001} = |\overline{X} \cap \overline{Y} \cap Z|$, $x_{010} = |\overline{X} \cap Y \cap \overline{Z}|$, $x_{011} = |\overline{X} \cap Y \cap Z|$, $x_{100} = |X \cap \overline{Y} \cap \overline{Z}|$,$x_{101} = |X \cap \overline{Y} \cap Z|$, $x_{110} = |X \cap Y \cap \overline{Z}|$, $x_{111} = |X \cap Y \cap Z|$.

Theorem 15. *$\delta := \alpha m + \overline{\alpha} M$ satisfies the triangle inequality if and only if $\alpha \leq 1/2$.*

Proof. We first show the *only if* direction. Let $X = \{0\}$, $Y = \{1\}$, $Z = \{0,1\}$. Then

$$\alpha m(X,Y) + \overline{\alpha} M(X,Y) = 1,$$
$$\alpha m(X,Z) + \overline{\alpha} M(X,Z) = \alpha m(Y,Z) + \overline{\alpha} M(Y,Z) = 0 + \overline{\alpha}.$$

The triangle inequality then is equivalent to $1 \leq 2\overline{\alpha}$, i.e., $\alpha \leq 1/2$.

Now let us show the *if* direction. The triangle inequality says

$$\alpha \min\{x_y, y_x\} + \overline{\alpha} \max\{y_x, x_y\} \leq \alpha \min\{x_z, z_x\} + \overline{\alpha} \max\{z_x, x_z\} + \alpha \min\{z_y, y_z\} + \overline{\alpha} \max\{y_z, z_y\}$$

By symmetry between x and y, we may assume that $y \leq x$. Hence either $y \leq z \leq x$, $y \leq x \leq z$, or $z \leq y \leq x$. Thus our proof splits into three Cases, I, II, and III.

Case I: $y \leq z \leq x$: we must show that $\alpha y_x + \overline{\alpha} x_y \leq \alpha z_x + \overline{\alpha} x_z + \alpha y_z + \overline{\alpha} z_y$. Since $y_x \leq y_z + z_x$ and $x_y \leq x_z + z_y$, this holds for all α.

Case II: $y \leq x \leq z$: We want to show that $\alpha y_x + \overline{\alpha} x_y \leq \alpha x_z + \overline{\alpha} z_x + \alpha y_z + \overline{\alpha} z_y$. Let us add αy_x to both sides:

$$2\alpha y_x + \overline{\alpha} x_y \leq \alpha x_z + \overline{\alpha} z_x + \alpha y_z + \overline{\alpha} z_y + \alpha y_x.$$

The identity $x_y + y_z + z_x = x_z + z_y + y_x$ holds generally since both sides counts the elements that belong to exactly one of X, Y, Z once each, and counts the elements that belong to exactly two of X, Y, Z once each. Since $\alpha \leq \overline{\alpha}$, let us write $\overline{\alpha} = \alpha + \epsilon$. Thus we must show

$$2\alpha y_x + (\alpha+\epsilon) x_y \leq \alpha x_z + (\alpha+\epsilon) z_x + \alpha y_z + (\alpha+\epsilon) z_y + \alpha y_x.$$

Replacing $\alpha(y_x + x_z + z_y)$ by $\alpha(y_z + z_x + x_y)$,

$$2\alpha y_x + (\alpha+\epsilon) x_y \leq (\alpha+\epsilon) z_x + \alpha y_z + \epsilon z_y + \alpha(y_z + z_x + x_y).$$

$$2\alpha y_x + (\alpha+\epsilon)x_y \le (\epsilon)z_x + \epsilon z_y + \alpha(x_y) + 2\alpha(z_x + y_z).$$

But now we seem to need $x_y \le z_x + z_y$, which follows from $x_y \le x_z + z_y$ and $x_z \le z_x$.

Case III: $z \le y \le x$: We now need

$$\alpha y_x + \overline{\alpha} x_y \le \alpha z_x + \overline{\alpha} x_z + \alpha z_y + \overline{\alpha} y_z$$

Let us do it the same way:

$$2\alpha y_x + \overline{\alpha} x_y \le \alpha z_x + \overline{\alpha} x_z + \alpha z_y + \overline{\alpha} y_z + \alpha y_x$$
$$2\alpha y_x + \overline{\alpha} x_y \le \alpha z_x + \epsilon x_z + \overline{\alpha} y_z + \alpha(y_x + x_z + z_y)$$
$$2\alpha y_x + \overline{\alpha} x_y \le \alpha z_x + \epsilon x_z + \overline{\alpha} y_z + \alpha(y_z + z_x + x_y)$$
$$2\alpha y_x + \overline{\alpha} x_y \le \epsilon x_z + \epsilon y_z + \alpha(x_y) + 2\alpha(y_z + z_x)$$
$$2\alpha y_x + \epsilon x_y \le \epsilon x_z + \epsilon y_z + 2\alpha(y_z + z_x)$$

Now we need $x_y \le x_z + y_z$ which follows from $x_y \le x_z + z_y$ and $z_y \le y_z$. Written without "forward reasoning":

$$\begin{aligned} 2\alpha y_x + \overline{\alpha} x_y &\le 2\alpha(y_z + z_x) + \alpha x_y + \epsilon(x_z + z_y) \\ &\le 2\alpha(y_z + z_x) + \alpha x_y + \epsilon(x_z + y_z) \\ &= \alpha(y_z + z_x) + \epsilon(x_z + y_z) + \alpha(x_y + y_z + z_x) \\ &= \alpha(y_z + z_x) + \epsilon(x_z + y_z) + \alpha(x_z + z_y + y_x) \\ &= \alpha z_x + \overline{\alpha} x_z + \alpha z_y + \overline{\alpha} y_z + \alpha y_x \end{aligned}$$

□

Theorem 16. *The function $D'_{\alpha,\beta}$ is a metric only if $\beta \ge 1/(1-\alpha)$.*

Proof. Consider the same example as in Theorem 15. Ignoring the upfront factor of β, we have

$$D' = \frac{\delta}{|X \cap Y| + \beta\delta}.$$

In our example,

$$\begin{aligned} D'(X,Y) &= \frac{1}{0 + \beta \cdot 1} = \frac{1}{\beta}, \\ D'(X,Z) = D'(Y,Z) &= \frac{\overline{\alpha}}{1 + \beta \cdot \overline{\alpha}} = \frac{\overline{\alpha}}{1 + \beta\overline{\alpha}}. \end{aligned}$$

The triangle inequality is then equivalent to:

$$\frac{1}{\beta} \le 2\frac{\overline{\alpha}}{1+\beta\overline{\alpha}} \iff \beta \ge \frac{1+\beta\overline{\alpha}}{2\overline{\alpha}} \iff \beta \ge 1/(1-\alpha). \qquad \square$$

Theorem 17. *The function $D'_{\alpha,\beta}$ is a metric on all finite power sets only if $\alpha \le 1/2$.*

Proof. Suppose $\alpha > 1/2$. Let $Z_n = \{-(n-1), -(n-2), \ldots, 0\}$, a set of cardinality n disjoint from $\{1,2\}$, and let $Y_n = Z_n \cup \{1\}$, $X_n = Z_n \cup \{2\}$. The triangle inequality says

$$\begin{aligned} \beta\frac{1}{n+\beta\cdot 1} = D'(X_n, Y_n) &\le D'(X_n, Z_n) + D'(Z_n, Y_n) = 2\beta\frac{\overline{\alpha}}{n+\beta\overline{\alpha}} \\ n + \beta\overline{\alpha} &\le 2\overline{\alpha}(n+\beta) \\ n(1-2\overline{\alpha}) &\le \beta\overline{\alpha} \end{aligned}$$

Since $\alpha > 1/2$, we have $2\overline{\alpha} < 1$. Let $n > \frac{\beta\overline{\alpha}}{1-2\overline{\alpha}}$. Then the triangle inequality does not hold, so $D'_{\alpha,\beta}$ is not a metric on the power set of $\{-(n-1), -(n-2), \ldots, 0, 1, 2\}$. □

Proof of Theorem 10. We saw in Theorem 15 that δ is a metric for $0 \le \gamma \le 1$. (Recall that $\beta = 1/(1-\alpha)$, so that $\gamma = \alpha/\overline{\alpha}$.) In general if d is a metric and a is a function, we may hope that $d/(a+d)$ is a metric. We shall use the observation, mentioned by [14], that in order to show

$$\frac{d_{xy}}{a_{xy}+d_{xy}} \le \frac{d_{xz}}{a_{xz}+d_{xz}} + \frac{d_{yz}}{a_{yz}+d_{yz}},$$

it suffices to show the following pair of inequalities:

$$\frac{d_{xy}}{a_{xy}+d_{xy}} \le \frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \tag{2.0.1}$$

$$\frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \le \frac{d_{xz}}{a_{xz}+d_{xz}} + \frac{d_{yz}}{a_{yz}+d_{yz}} \tag{2.0.2}$$

Here (2.0.1) follows from d being a metric, i.e., $d_{xy} \le d_{xz} + d_{yz}$, since $c \ge 0 < a \le b \implies \frac{a}{a+c} \le \frac{b}{b+c}$.

Next, (2.0.2) would follow from $a_{xy} + d_{yz} \ge a_{xz}$ and $a_{xy} + d_{xz} \ge a_{yz}$. By symmetry between x and y and since $a_{xy} = a_{yx}$ in our case, it suffices to prove the first of these, $a_{xy} + d_{yz} \ge a_{xz}$. This is equivalent to

$$x\cap y + \beta(\alpha\min\{z_y, y_z\} + \overline{\alpha}\max\{z_y, y_z\}) \ge x\cap z,$$

i.e.,

$$x\cap y + \gamma\min\{z_y, y_z\} + \max\{z_y, y_z\} \ge x\cap z,$$

which holds for all $0 \le \gamma \le 1$ if and only if $x\cap y + \max\{z_y, y_z\} \ge x\cap z$. There are now two cases.

Case $z \ge y$: We have

$$x\cap y + z_y \ge x\cap z$$

since any element of $X \cap Z$ is either in Y or not:

$$x\cap z = x\cap y\cap z + x\cap z_y, \quad x\cap y\cap z \le x\cap y.$$
$$x\cap z \le x\cap y + z_y$$
$$x\cap y\cap z + x\cap z_y \le (x\cap y_z + x\cap y\cap z) + (z\cap x_y + z_{yx})$$
$$x\cap z_y \le (x\cap y_z) + (z\cap x_y + z_{yx})$$
$$0 \le x\cap y_z + z_{yx}$$

which is true. Case $y \ge z$:

$$\begin{aligned} x\cap y + y_z &\ge x\cap z \\ x_{110} + x_{111} + x_{110} + x_{010} &\ge x_{101} + x_{111} \\ x_{110} + x_{110} + x_{010} &\ge x_{101} \end{aligned}$$

This is true since $x_{101} = x\cap z_y \le z_y \le y_z = x\cap y_z + y_{xz} = x_{110} + x_{010}$. □

3 APPLICATION TO PHYLOGENY

The mutations of spike glycoproteins of coronaviruses are of great concern with the new SARS-CoV-2 virus causing the disease CoViD-19. We calculate several distance measures between peptide sequences for such proteins. The distance

$$Z_{2,\alpha}(x_0, x_1) = \alpha\min(|A_1|, |A_2|) + \overline{\alpha}\max(|A_1|, |A_2|)$$

where A_i is the set of subwords of length 2 in x_i but not in x_{1-i}, counts how many subwords of length 2 appear in one sequence and not the other. Our calculations are available in a URL format as follows:

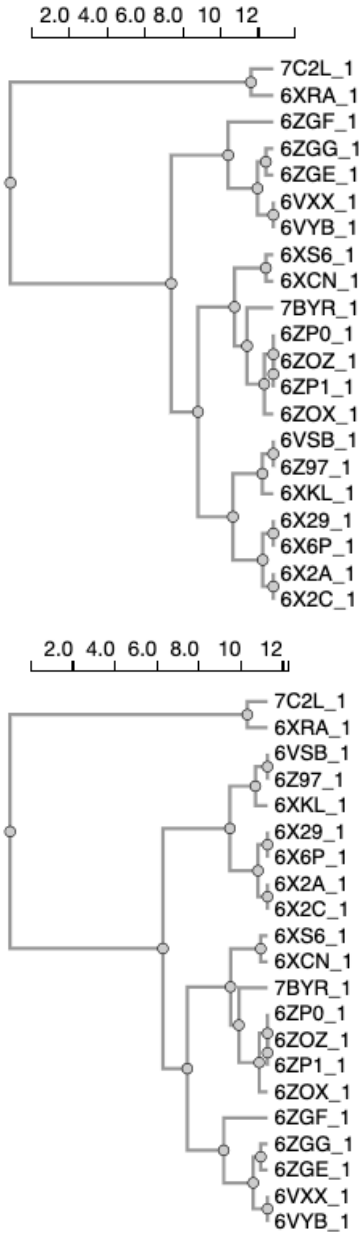

Figure 2: $\alpha = 0.21$ **and** 0.36**.**

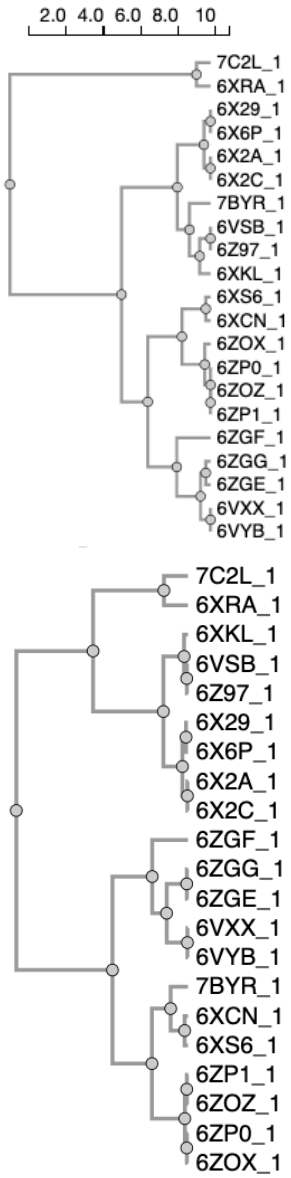

Figure 3: $\alpha = 0.5$ **and edit distance.**

counter-automata.appspot.com/spike?metric=z2&alpha=0.36

We used the Ward linkage criterion for producing Newick trees using the `hclust` package for the Go programming language. The calculated phylogenetic trees were based on the metric $Z_{2,\alpha}$.

We found one tree isomorphism class each for $0 \leq \alpha \leq 0.21$, $0.22 \leq \alpha \leq 0.36$, and $0.37 \leq \alpha \leq 0.5$, respectively (Figure 2, Figure 3). In Figure 3 we are also including the tree produced using the Levenshtein edit distance in place of $Z_{2,\alpha}$. We see that the various intervals for α can correspond to "better" or "worse" agreement with other distance measures. Thus, we propose that rather than focusing on $\alpha = 0$ and $\alpha = 1/2$ exclusively, future work may consider the whole interval $[0, 1/2]$.

4 FORMALIZATION IN A PROOF ASSISTANT

We formally prove Theorem 10 in the Lean theorem prover. The Github repository can be found at [5].

Our formal version asserts that `d(X,Y)= m min(|XY|, |YX|) + M max(|XY|, |YX|)` and also `d(X,Y)= m min(|X\Y|, |Y\X|) + M max(|X\Y|, |Y\X|)` is a metric if and only if `m ≤ M`.

Theorem seventeen is the proof of Theorem 10 in the present paper.

4.1 Formal development of the function δ

We include definitions and theorem statements. As usual in Lean development, missing proofs are indicated by the keyword sorry. Full formal proofs are available on Github [5].

For our implementation we change variables to m and M with $\beta\alpha = m$ and $\beta(1-\alpha) = M$.

```
import data.finset  -- finite set
import data.set -- to make backslash work as set
    difference
import data.finset.basic
import topology.metric_space.basic

import data.real.basic

open finset
local notation |X| := X.card

variables {α : Type*} [decidable_eq α]
variables {β : Type*}

lemma disj₁ (X Y Z : finset α) : disjoint ((X \ Y) ∪
    (Y \ Z)) (Z \ X) := sorry
lemma disj₂ (X Y Z : finset α) : disjoint (X \ Y) (Y \
    Z) := sorry

lemma union_rot_sdiff (X Y Z : finset α) :
    (X \ Y) ∪ (Y \ Z) ∪ (Z \ X) = (X \ Z) ∪ (Z \ Y) ∪
    (Y \ X) := sorry

lemma card_rot (X Y Z : finset α) : |X \ Y| + |Y \ Z| +
    |Z \ X| = |X \ Z| + |Z \ Y| + |Y \ X| := sorry

lemma card_rot_cast (X Y Z : finset α) : ((|X\Y| + |
    Y\Z| + |Z\X|):ℝ) = ((|X\Z| + |Z\Y| + |Y\X|):ℝ) :=
    sorry

variables {m M : ℝ}

/-- The function δ is a stepping stone towards Jaccard
    inequality. It is not normalized to be between 0
    and 1. -/
def δ : ℝ → ℝ → finset α → (finset α → ℝ) :=
    λ m M A B, M * ↑ (max (|A\B|) (|B\A|)) + m * ↑ (min
    (|A\B|) (|B\A|))

theorem delta_cast {m M:ℝ} {A B : finset α} :
    δ m M A B = M * (max ↑(|A\B|) ↑(|B\A|)) + m * (min ↑
    (|A\B|) ↑(|B\A|)) := by norm_cast

theorem delta_comm {m M : ℝ} {A B : finset α}: δ m M A
    B = δ m M B A := sorry

lemma card_sdiff_self (X : finset α) : |X \ X| = 0 :=
    sorry

theorem delta_self (X : finset α): δ m M X X = 0 :=
    sorry

lemma subseteq_of_card_zero (x y : finset α) : |x \ y| =
     0 → x ⊆ y :=  sorry

lemma card_zero_of_not_pos (X : finset β) : ¬ 0 < |X|
    → |X| = 0 := sorry
lemma eq_zero_of_nonneg_of_nonneg_of_add_zero {a b : ℝ}
    : 0 ≤ a → 0 ≤ b → 0 = a + b → 0 = a := sorry

theorem subset_of_delta_eq_zero (hm: 0 < m) (hM: m ≤
    M) (X Y : finset α) (h: δ m M X Y = 0) : X ⊆ Y :=
    sorry

 theorem eq_of_delta_eq_zero (hm: 0 < m) (hM: m ≤ M)
    (X Y : finset α) (h: δ m M X Y = 0) :
 X = Y := sorry

theorem sdiff_covering {A B C : finset α}: A\C ⊆ (A\B)
    ∪ (B\C) := sorry

theorem sdiff_triangle (A B C: finset α): |A\C| ≤ |
    A\B| + |B\C| := sorry
lemma venn (X Y : finset α): X = X\Y ∪ (X ∩ Y) := sorry

lemma venn_card (X Y : finset α): |X| = |X\Y| + |X ∩
    Y| := sorry

lemma sdiff_card (X Y : finset α): |Y| ≤ |X| → |Y\X| ≤
     |X\Y| := sorry

lemma maxmin_1 {X Y : finset α}: |X| ≤ |Y| → δ m M X
    Y = M * |Y\X| + m * |X\Y| := sorry

lemma maxmin_2 {X Y : finset α}: |Y| ≤ |X| → δ m M X
    Y = M * |X\Y| + m * |Y\X| := sorry

theorem casting {a b : ℕ}: a ≤ b → (a:ℝ) ≤ (b:ℝ) :=
    sorry

theorem mul_sdiff_tri (m : ℝ) (hm: 0 ≤ m) (X Y Z :
    finset α):
    m * ↑ |X\Z| ≤ m * (↑ |X\Y| + ↑ |Y\Z|) := sorry

/-- The triangle inequality for δ -/
def triangle_inequality (m M :ℝ) (X Y Z: finset α) :
    Prop :=
    δ m M X Y ≤ δ m M X Z + δ m M Z Y

lemma seventeen_right_yzx {m M :ℝ} {X Y Z: finset α} :
    0 ≤ m → m ≤ M → |Y| ≤ |Z| ∧ |Z| ≤ |X| →
    triangle_inequality m M X Y Z
    := sorry

lemma co_sdiff (X Y U : finset α):
X ⊆ U → Y ⊆ U → (U\X)\(U\Y) = Y\X := sorry

lemma co_sdiff_card (X Y U : finset α):
X ⊆ U → Y ⊆ U → ((U\X)\(U\Y)).card = (Y\X).card :=
    sorry

lemma co_sdiff_card_max (X Y U : finset α):
X ⊆ U → Y ⊆ U → max (|(U\Y)\(U\X)|) (|(U\X)\(U\Y)|) =
     max (|X\Y|) (|Y\X|) := sorry

lemma co_sdiff_card_min (X Y U : finset α):
```

```
X ⊆ U → Y ⊆ U → min (|(U\Y)\(U\X)|) (|(U\X)\(U\Y)|) =
     min (|X\Y|) (|Y\X|) := sorry

theorem delta_complement (X Y U : finset α):
    X ⊆ U → Y ⊆ U → δ m M X Y = δ m M (U\Y) (U\X) :=
    sorry

theorem seventeen_right_yxz {X Y Z : finset α}:
    0 ≤ m → m ≤ M → |Y| ≤ |X| ∧ |X| ≤ |Z| →
    triangle_inequality m M X Y Z := sorry

lemma sdiff_card_le (X Y U : finset α) (hx: X ⊆ U)
    (hy: Y ⊆ U) (h:|X| ≤ |Y|):
    |U \ Y| ≤ |U \ X| := sorry

theorem seventeen_right_zyx{m M : ℝ} {X Y Z : finset α
    }:
    0 ≤ m → m ≤ M → |Z| ≤ |Y| ∧ |Y| ≤ |X| →
    triangle_inequality m M X Y Z := sorry

theorem seventeen_right_zxy {X Y Z : finset α}:
    0 ≤ m → m ≤ M → |Z| ≤ |X| ∧ |X| ≤ |Y| →
    triangle_inequality m M X Y Z := sorry

theorem three_places {x y z : ℕ} :
    y ≤ x → (z ≤ y ∧ y ≤ x) ∨ (y ≤ z ∧ z ≤ x) ∨ (y
    ≤ x ∧ x ≤ z) := sorry

theorem seventeen_right_y_le_x {m M : ℝ} {X Y Z :
    finset α} :
    |Y| ≤ |X| → 0 ≤ m → m ≤ M →
    triangle_inequality m M X Y Z := sorry

theorem seventeen_right_x_le_y {m M : ℝ} {X Y Z :
    finset α} :
    |X| ≤ |Y| → 0 ≤ m → m ≤ M →
    triangle_inequality m M X Y Z := sorry

theorem mem_pair {x y z : α} : x ∈ ({y, z} : finset α)
    ↔ x = y ∨ x = z := sorry

lemma sdiff_singleton_pair (x y : α) (hne : x ≠ y):
    ({x}: finset α) \ ({x,y}: finset α)  = (∅: finset
    α) := sorry

lemma sdiff_singleton_pair' (x y : α) (hne : x ≠ y):
    ({y}: finset α) \ ({x,y}: finset α)  = (∅: finset
    α) := sorry

lemma sdiff_pair_singleton (x y : α) (hne : x ≠ y):
    ({x,y}: finset α) \ ({x}: finset α)  = ({y}:
    finset α) := sorry

lemma sdiff_pair_singleton' (x y : α) (hne : x ≠ y):
    ({x,y}: finset α) \ ({y}: finset α)  = ({x}:
    finset α) := sorry

lemma sdiff_singleton (x y : α) (hne : x ≠ y):  ({x}:
    finset α) \ ({y}: finset α)  = ({x}: finset α) :=
    sorry

theorem seventeen_right {m M : ℝ} { X Y Z : finset α}:
    0 ≤ m → m ≤ M → triangle_inequality m M X Y Z :=
    λ hm: 0 ≤ m, λ h: m ≤ M,
    (le_total (|X|) (|Y|)).elim (
        λ h1: |X| ≤ |Y|, seventeen_right_x_le_y h1 hm h
    )(
        λ h1: |Y| ≤ |X|, seventeen_right_y_le_x h1 hm h
    )

def s_0 : finset ℕ := ({0}: finset ℕ)
def s_1 : finset ℕ := ({1}: finset ℕ)
def s01 : finset ℕ := ({0,1} : finset ℕ)
theorem seventeen: (∃ x y : α, x ≠ y) →
    0 ≤ m → (m ≤ M ↔ ∀ X Y Z : finset α,
     triangle_inequality m M X Y Z) := sorry

    def delta_triangle (X Y Z: finset α) (hm: 0 < m)
     (hM: m ≤ M):
    triangle_inequality m M X Z Y
    --δ m M X Y ≤ δ m M X Z + δ m M Z Y
    := seventeen_right (le_of_lt hm) hM

    section jaccard_numerator
    /-- Instantiate finset ℕ as a metric space. -/

    def protein {m M : ℝ} (hm : 0 < m) (hM : m ≤ M) :=
    finset α

    noncomputable instance
     jaccard_numerator.metric_space (typ: (∃ x y : α, x
     ≠ y)) (hm : 0 < m) (hM : m ≤ M): metric_space
     (protein hm hM) := {
        dist               := λx y, δ m M x y,
        dist_self          := delta_self,
        eq_of_dist_eq_zero := eq_of_delta_eq_zero hm hM,
        dist_comm          := λ x y, delta_comm,
        dist_triangle      := λ x y z, ((iff.elim_left
     (seventeen typ (le_of_lt hm))) hM) x z y
    }
end jaccard_numerator
```

The norm_cast tactic [9], developed to simplify expressions containing type coercions, proved useful to handle the many embeddings between $\mathbb{N}$ and $\mathbb{Z}$.

4.2 Implementation of the generalizations of Jaccard distance

We give a formal proof that in terms of

$$d(X,Y) = m\min(|X \setminus Y|, |Y \setminus X|) + M\max(|X \setminus Y|, |Y \setminus X|)$$

for $m \geq 0$ and $M \geq 0$, the function

$$D(X,Y) = d(X,Y)/(|X \cap Y| + d(X,Y))$$

is a metric if and only if $m \leq M$ and $1 \leq M$. In particular, taking $m = M = 1$, the Jaccard distance is a metric on the set of finite subsets of $\mathbb{N}$.

```
import data.finset  -- finite set
import data.set -- to make backslash work as set
    difference
import data.finset.basic
import topology.metric_space.basic
```

```
import data.real.basic
import delta
import data.set.basic

open finset

local notation |X| := X.card

variables {m M : ℝ} -- in delta.lean but can't import
    variables

section jaccard_nid

variables {α : Type*} [decidable_eq α]
--#check δ

/-- using Lean's "group with zero" to hand the case
    0/0=0 -/
noncomputable def D : ℝ → ℝ → finset α → (finset α
    → ℝ) :=
  λ m M x y, (δ m M x y) / (|x ∩ y| + δ m M x y)

theorem twelve_end (X Y Z : finset α) : |X ∩ Z| ≤ |X ∩
     Y| + max (|Z \ Y|) (|Y \ Z|) := sorry

theorem twelve_middle (hm: 0 ≤ m) (hM: 0 < M) (X Y Z :
    finset α) :
let y_z := |Y\Z|, z_y := |Z\Y|, xy := |X ∩ Y|, xz := |X
    ∩ Z| in
(|X ∩ Z|:ℝ) ≤ (xy:ℝ) + (max z_y y_z:ℝ) + (m/M) * (min
    z_y y_z:ℝ) := sorry

theorem jn_self  (X : finset α): D m M X X = 0 := sorry

theorem delta_nonneg {x y : finset α} (hm: 0 ≤ m) (hM:
    m ≤ M): 0 ≤ δ m M x y := sorry

theorem jn_comm (X Y : finset α): D m M X Y = D m M Y X
    := sorry

lemma card_inter_nonneg (X Y : finset α):
  0 ≤ (|X ∩ Y|:ℝ) := sorry

lemma D_denom_nonneg (X Y : finset α) (hm: 0 ≤ m) (hM:
    m ≤ M):
  0 ≤ (|X ∩ Y|:ℝ) + δ m M X Y := sorry

theorem eq_of_jn_eq_zero (hm: 0 < m) (hM: m ≤ M) (X Y :
     finset α) (h: D m M X Y = 0) : X = Y := sorry

theorem D_nonneg (X Y : finset α) (hm: 0 ≤ m) (hM: m ≤
    M): 0 ≤ D m M X Y := sorry

theorem D_empty_1 (m M : ℝ) {X Y : finset α} (hm: 0 <
    m) (hM: m ≤ M):
  X = ∅ → Y ≠ ∅ → D m M X Y = 1 := sorry

theorem D_empty_2  (m M : ℝ) {X Y : finset α} (hm: 0 <
    m) (hM: m ≤ M):
  X ≠ ∅ → Y = ∅ → D m M X Y = 1
:= sorry

theorem D_bounded (m M : ℝ) (X Y : finset α) (hm: 0 ≤
    m) (hM: m ≤ M): D m M X Y ≤ 1 := sorry

theorem intersect_cases (m M : ℝ) (Y Z : finset α)
    (hm: 0<m) (hM: m≤ M) (hy: Y ≠ ∅) (hz: Z ≠ ∅): let
    ayz := (|Z ∩ Y|:ℝ), dyz :=  (δ m M Z Y) in 0 <
    (ayz + dyz) := sorry

lemma four_immediate_from (m M : ℝ) (X Y Z : finset α)
  (hm: 0 < m) (hM: m ≤ M) (h1M: 1 ≤ M)
  (hx: X ≠ ∅) (hz: Z ≠ ∅):
  let axy := (|X ∩ Y|:ℝ), dxz := δ m M X Z, dyz := δ m
    M Z Y,
     axz := (|X ∩ Z|:ℝ), denom := axy+dxz+dyz in
  dxz/denom ≤ dxz/(axz + dxz) := sorry

 lemma four_immediate_from_and  (m M : ℝ) (X Y Z :
    finset α)
  (hm: 0 < m) (hM: m ≤ M) (h1M: 1 ≤ M)
  (hy: Y ≠ ∅) (hz: Z ≠ ∅):
  (δ m M Z Y)/((|X ∩ Y|:ℝ) + δ m M X Z + δ m M Z Y) ≤
    (δ m M Z Y)/((|Z ∩ Y|:ℝ) + δ m M Z Y) := sorry

lemma mul_le_mul_rt {a b c : ℝ} (h : 0 ≤ c) : a ≤ b →
    a * c ≤ b * c := sorry

lemma abc_lemma {a b c : ℝ} (h : 0 ≤ a) (hb : a ≤ b)
    (hc : 0 ≤ c) : (a/(a+c)) ≤ (b/(b+c)) := sorry

theorem three (X Y Z : finset α) (hm: 0 < m) (hM: m ≤
    M):
let axy := (|X ∩ Y| : ℝ), dxy := δ m M X Y, dxz := δ m
    M X Z,
   dyz := δ m M Z Y, denom := (axy+dxz+dyz) in

dxy/(axy + dxy) ≤ (dxz+dyz)/denom := sorry

 theorem jn_triangle_nonempty
  (m M : ℝ) (X Y Z : finset α) (hm: 0 < m) (hM: m ≤ M)
    (h1M: 1 ≤ M)
  (hx: X ≠ ∅) (hy: Y ≠ ∅) (hz: Z ≠ ∅): D m M X Y ≤ D
    m M X Z + D m M Z Y := sorry

theorem jn_triangle (m M : ℝ) (X Y Z : finset α)
(hm: 0 < m) (hM: m ≤ M) (h1M: 1 ≤ M): D m M X Y ≤ D m
    M X Z + D m M Z Y := sorry

noncomputable instance jaccard_nid.metric_space
(hm : 0 < m) (hM : m ≤ M) (h1M: 1 ≤ M): metric_space
    (finset α) := {
       dist               := λx y, D m M x y,
       dist_self          := jn_self,
       eq_of_dist_eq_zero := eq_of_jn_eq_zero hm hM,
       dist_comm          := λ x y, jn_comm x y,
       dist_triangle      := λ x z y, jn_triangle m M
    x y z hm hM h1M
    }

noncomputable def J : finset α → (finset α → ℝ) :=
  λ x y, (δ 1 1 x y) / ((|x ∩ y|:ℝ) + δ 1 1 x y)

noncomputable instance jaccard.metric_space
```

```
(hm : (0:ℝ) < (1:ℝ)) (hM : (1:ℝ) ≤ (1:ℝ)) (h1M: (1:ℝ)
     ≤ (1:ℝ)): metric_space (finset ℕ) := {
        dist               := λx y, D 1 1 x y,
        dist_self          := jn_self,
        eq_of_dist_eq_zero := eq_of_jn_eq_zero hm hM,
        dist_comm          := λ x y, jn_comm x y,
        dist_triangle      := λ x z y, jn_triangle 1 1
     x y z hm hM h1M
     }
end jaccard_nid
```

ACKNOWLEDGMENTS

This work was partially supported by grants from the Simons Foundation (#704836 to Bjørn Kjos-Hanssen) and Decision Research Corporation (University of Hawai'i Foundation Account #129-4770-4).

For the formalization in Lean we received Zulip chat help from: Johan Commelin, Kyle Miller, Pedro Minicz, Reid Barton, Scott Morrison, Heather Macbeth. Jason Greuling also improved our formal definitions.

REFERENCES

[1] Michel Marie Deza and Elena Deza. 2016. *Encyclopedia of distances* (fourth ed.). Springer, Berlin. xxii+756 pages. https://doi.org/10.1007/978-3-662-52844-0

[2] Alonso Gragera and Vorapong Suppakitpaisarn. 2016. Semimetric properties of Sørensen-Dice and Tversky indexes. In *WALCOM: algorithms and computation*. Lecture Notes in Comput. Sci., Vol. 9627. Springer, [Cham], 339–350. https://doi.org/10.1007/978-3-319-30139-6_27

[3] Alonso Gragera and Vorapong Suppakitpaisarn. 2018. Relaxed triangle inequality ratio of the Sørensen-Dice and Tversky indexes. *Theoret. Comput. Sci.* 718 (2018), 37–45. https://doi.org/10.1016/j.tcs.2017.01.004

[4] Sergio Jiménez, Claudia Jeanneth Becerra, and Alexander F. Gelbukh. 2013. SOFTCARDINALITY-CORE: Improving Text Overlap with Distributional Measures for Semantic Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*, Mona T. Diab, Timothy Baldwin, and Marco Baroni (Eds.). Association for Computational Linguistics, 194–201. https://www.aclweb.org/anthology/S13-1028/

[5] Bjørn Kjos-Hanssen. 2021. jaccard. https://github.com/bjoernkjoshanssen/jaccard.

[6] Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger. 2003. Hierarchical Clustering Based on Mutual Information. *ArXiv* q-bio.QM/0311039 (2003).

[7] A Kraskov, H Stögbauer, R. G Andrzejak, and P Grassberger. 2005. Hierarchical clustering using mutual information. *Europhysics Letters (EPL)* 70, 2 (apr 2005), 278–284. https://doi.org/10.1209/epl/i2004-10483-y

[8] Abraham Lempel and Jacob Ziv. 1976. On the complexity of finite sequences. *IEEE Trans. Inform. Theory* IT-22, 1 (1976), 75–81. https://doi.org/10.1109/tit.1976.1055501

[9] Robert Y. Lewis and Paul-Nicolas Madelaine. 2020. Simplifying Casts and Coercions. arXiv:2001.10594 [cs.PL]

[10] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul E. Kearney, and Haoyong Zhang. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17 2 (2001), 149–54.

[11] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. 2004. The similarity metric. *IEEE Trans. Inform. Theory* 50, 12 (2004), 3250–3264. https://doi.org/10.1109/TIT.2004.838101

[12] Edward Raff and Charles K. Nicholas. 2017. An Alternative to NCD for Large Sequences, Lempel–Ziv Jaccard Distance. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017).

[13] C. Rajski. 1961. Entropy and metric spaces. In *Information theory (Symposium, London, 1960)*. Butterworths, Washington, D.C., 41–45.

[14] Suvrit Sra. [n.d.]. Is the Jaccard distance a distance? MathOverflow. arXiv:https://mathoverflow.net/q/210750 https://mathoverflow.net/q/210750 URL:https://mathoverflow.net/q/210750 (version: 2015-07-03).

[15] A. Tversky. 1977. Features of similarity. *Psychological Review* 84, 4 (1977), 327––352. https://doi.org/10.1037/0033-295X.84.4.327

[16] Jacob Ziv and Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory* IT-23, 3 (1977), 337–343. https://doi.org/10.1109/tit.1977.1055714

[17] Jacob Ziv and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory* 24, 5 (1978), 530–536. https://doi.org/10.1109/TIT.1978.1055934