Semimetric Properties of Sørensen-Dice and Tversky Indexes

Alonso Gragera[1] and Vorapong Suppakitpaisarn[1,2]

[1] Department of Computer Science, The University of Tokyo, Japan
alonso@is.s.u-tokyo.ac.jp, vorapong@is.s.u-tokyo.ac.jp
[2] JST ERATO Kawarabayashi Large Graph Project

Abstract. In this work we prove a semimetric property for distances used for finding dissimilarities between two finite sets such as the Sørensen-Dice and the Tversky indexes. The Jaccard-Tanimoto index is known to be one of the most common distances for the task. Because the distance is a metric, when used, several algorithms can be applied to retrieve information from the data. Although the Sørensen-Dice index is known to be more robust than the Jaccard-Tanimoto when some information is missing from datasets, the distance is not a metric as it does not satisfy the triangle inequality. Recently, there are several machine learning algorithms proposed which use non-metric distances. Hence, instead of the triangle inequality, it is required that the distance satisfies the approximate triangle inequality with some small value of ρ. This motivates us to find the value of ρ for the Sørensen-Dice index. In this paper, we prove that this value is 1.5. Besides, we can find the value for some of the Tversky index.

Keywords: distance, metric, approximate triangle inequality, Sørensen-Dice index, Tversky index

1 Introduction

In this work, we consider distances used for finding a dissimilarity between two finite sets. The most common dissimilarity between sets is called Jaccard-Tanimoto index (JT index) [1, 2]. The Jaccard-Tanimoto distance between a finite set A and a finite set B, $d_{JT}(A, B)$, can be defined as follows:

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

Because the JT distance is known to be a metric [3], we can use algorithms proposed for any metric space to obtain information from the data. Examples of those algorithms include the approximation algorithms for facility location problem [4], nearest neighbor problem [5], and the Steiner tree problem [6].

Besides the JT index, there are other dissimilarity indexes considered in literature. Among them, the most well-known indexes include the Sørensen-Dice

index [7, 8] (SD index), in which the distance between a finite set A and a finite set B, $d_{SD}(A, B)$, can be defined as follows:

$$d_{SD} := 1 - \frac{2|A \cap B|}{|A| + |B|} = 1 - \frac{2|A \cap B|}{|A \cup B| + |A \cap B|}.$$

The SD dissimilarity is known to be robust in datasets of which some data points are missing [9]. Because of that, the distance is widely used in the ecological community data [10]. In our previous work [11], for some Go board position G, we know the set of positions P_G which is the set of positions that professional Go players choose to play when the board position is G. Because there is a lot of information missing in the database, we need to use a robust dissimilarity such as the SD distance.

Consider the case when $A = B = \{1, 2, 3\}$, suppose not all elements can be observed so A is observed as $A' = \{1, 2\}$ and B as $B' = \{2, 3\}$. For such case, the JT dissimilarity d_{JT} is $1 - 1/3 \approx 0.67$, while the SD dissimilarity is $1-(2\cdot1)/(2+2) \approx 0.5$. We can see from this example that with only one element missing the JT dissimilarity can increase to $2/3$, while the ST dissimilarity is still as small as $1/2$.

Although the ST index is robust, the dissimilarity is not a metric. That is because the dissimilarity does not satisfy the triangle inequality [12]. For example, when $A = \{1\}$, $B = \{1, 2\}$ and $C = \{3\}$, we have $d_{ST}(A, C) = 1$, $d_{ST}(A, B) = 1/3$, and $d_{ST}(B, C) = 1/3$. By that, we have $d_{ST}(A, C) > d_{ST}(A, B) + d_{ST}(B, C)$. Because the distance is not a metric, we cannot use the algorithms devised for metric spaces.

Recently, there are several algorithms proposed for semi-metric distances that satisfy the ρ-approximate triangle inequality [13–15] for some $\rho > 1$. A distance D satisfies the inequality, if for any finite sets A, B, C,

$$d(A, C) \leq \rho \left(d(A, B) + d(B, C) \right).$$

Those algorithms are more efficient, when the value of ρ is smaller. Knowing the value of ρ for a specific dissimilarity can help in analyzing the efficiency of the algorithms, when it is applied to the distance.

1.1 Our contribution

By the previous example, we know that the value of ρ for the SD index must be at least $3/2$. In this paper, we will show that the lower bound is tight, meaning that the SD index satisfies the $3/2$-approximate triangle inequality.

As discussed previously, the SD index is more robust than the JT index. We extend that idea to propose a dissimilarity called **robust Jaccard index (RJ index)**. We define the RJ dissimilarity between a finite set A and a finite set B, $d_{RJ,\alpha}(A, B)$, as follows:

$$d_{RJ,\alpha}(A, B) := 1 - \frac{\alpha|A \cap B|}{|A \cup B| + (\alpha - 1)|A \cap B|}.$$

Clearly, when $\alpha = 1$, the dissimilarity $d_{RJ,1}$ is equal to the JT index. When $\alpha = 2$, the dissimilarity $d_{RJ,2}$ is equal to the SD index. When $\alpha \to \infty$, $d_{RJ,\infty}(A, B)$ is always 0 when $A \cap B \neq \emptyset$ and the distance is 1 when $A \cap B = \emptyset$. The dissimilarity is very robust because we still get the distance equals 0 even when a lot of information are missing and $|A \cap B| = 1$. When $A' = \{1, 2\}$, $B' = \{2, 3\}$, $d_{RJ,\alpha} = 2/(2 + \alpha)$. The distance will get smaller and more robust when α is larger.

The robust Jaccard index is known to be a subclass of the Tversky index [16]. The Tversky index is proposed for differentiating the importance between two sets obtained as inputs [17]. They do not consider the robustness as one of the applications of the distance, but we believe that their index can also be used for this purpose.

The RJ index is a semi-metric. We can find an example to show that the value of ρ of $d_{RJ,\alpha}$ is at least $(\alpha + 1)/2$. We also show that the lower bound is tight. The distance $d_{RJ,\alpha}$ satisfies the $(\alpha+1)/2$-approximate triangle inequality.

The remainder of this paper is divided into five sections. Section 2 describes the robust Jaccard index and its interpretation and discusses the relation of our proposed coefficient with other similarity indexes. Section 3 consist of the proof of its semimetric properties. Then, the final section summarizes this paper and present the future direction of our research.

2 Robust Jaccard index

In order to be able to properly study how the concept of robustness of a set similarity index, and how it affects its metric properties, we propose a new coefficient that captures its fundamental idea.

Definition 1 (Robust Jaccard index). *For sets X and Y the robust Jaccard index is a number between 0 and 1 given by:*

$$S_{RJ,\alpha}(X, Y) := \frac{\alpha|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}.$$

By using this index, we not only expect to make the following proofs clearer to the reader; but also provide a powerful yet easy to use tool that can be directly applied when a more customized approach for dealing with uncertainty in experimental data is required.

2.1 Relation with other indexes

In this section, we proceed to investigate the relation between our similarity index and the three best-known indexes on sets; the Jaccard-Tanimoto, Sørensen-Dice and Tversky.

Definition (Jaccard-Tanimoto index) *For sets X and Y the Jaccard-Tanimoto index is a number between 0 and 1 given by:*

$$S_{JT}(X, Y) := \frac{|X \cap Y|}{|X \cup Y|}.$$

The robust Jaccard index is equivalent to the Jaccard-Tanimoto index when $\alpha = 1$.

Definition (Sørensen-Dice index) *For sets X and Y the Sørensen-Dice index is a number between 0 and 1 given by:*

$$S_{SD}(X, Y) := \frac{2|X \cap Y|}{|X| + |Y|}.$$

The robust Jaccard index is equivalent to the Sørensen-Dice index when $\alpha = 2$.

Definition (Tversky index) *For sets X and Y the Tversky index is a number between 0 and 1 given by:*

$$S_{T,\beta,\gamma}(X, Y) := \frac{|X \cap Y|}{|X \cap Y| + \beta|X - Y| + \gamma|Y - X|}.$$

Proposition 1. *The Tversky index is equivalent to the robust Jaccard index when $\beta = \gamma = \frac{1}{\alpha}$, i.e. $S_{T,1/\alpha,1/\alpha} = S_{RJ,\alpha}$.*

Proof. By the definitions of Tversky and robust Jaccard index, we have

$$\begin{aligned} S_{T,1/\alpha,1/\alpha}(X, Y) &= \frac{|X \cap Y|}{|X \cap Y| + \frac{1}{\alpha}|X - Y| + \frac{1}{\alpha}|Y - X|} \\ &= \frac{|X \cap Y|}{|X \cap Y| + \frac{1}{\alpha}(|X \cup X| - |X \cap Y|)} \\ &= \frac{\alpha|X \cap Y|}{|X \cup X| - (\alpha - 1)|X \cap Y|} = S_{RJ,\alpha} \end{aligned}$$

□

3 Metric properties

Once that we have defined the robust Jaccard similarity index, it is only natural to define a distance (or dissimilarity) coefficient as well.

Definition 2 (Robust Jaccard distance). *For sets X and Y and the robust Jaccard similarity index, the expression as a distance is given by:*

$$d_{RJ}(X, Y) = 1 - S_{RJ}(X, Y)$$

Then one could start questioning about the metric properties of this distance. In our case, since a counter example that show that it is not a metric can be easily found, we are more interested on the possibility of it being a semimetric.

Definition (Semimetric) *A semimetric on X is a function $d : X \times X \to \mathbb{R}$ that satisfies the first three axioms of a metric, but not necessarily the triangle inequality:*

(1) $d(x, y) \geq 0$
(2) $d(x, y) = 0 \iff x = y$
(3) $d(x, y) = d(y, x)$

Among all possible semimetrics, some of the most interesting ones are the so called "near-metrics", that are useful to guarantee a good performance in several approximation algorithms [13–15].

Definition (ρ-relaxed semimetric) *ρ-relaxed semimetric is a semimetric that also satisifies a ρ-relaxed triangle inequality:*

(4) $d(x, z) \leq \rho(d(x, y) + d(y, z))$.

In order to make the following proofs simpler, let us define $X_{A,B}, Y_{B,C}$, and $Z_{A,C}$ as follows:

$$X_{A,B} = \frac{S_{RJ,\alpha}(A, B)}{\alpha} = \frac{|A \cap B|}{|A \cup B| + (\alpha - 1)|A \cap B|},$$

$$Y_{B,C} = \frac{S_{RJ,\alpha}(B, C)}{\alpha} = \frac{|B \cap C|}{|B \cup C| + (\alpha - 1)|B \cap C|},$$

$$Z_{A,C} = \frac{S_{RJ,\alpha}(A, C)}{\alpha} = \frac{|A \cap C|}{|A \cup C| + (\alpha - 1)|A \cap C|}.$$

Also, $f(A, B, C) := \frac{\frac{1}{\alpha} - Z_{A,C}}{\frac{2}{\alpha} - X_{A,B} - Y_{B,C}}$, and let

$$A^*, B^*, C^* := \arg \max_{A,B,C} f(A, B, C).$$

In Lemma 3, we show that

$$f(A^*, B^*, C^*) = \frac{\alpha + 1}{2}.$$

Using that result, we can obtain our main result stated in the following theorem.

Theorem 1. *For $\alpha \in \mathbb{Z}_+$, $d_{RJ,\alpha}$ is a ρ-relaxed semimetric, with $\rho = (\alpha + 1)/2$, i.e., for any finite sets A, B, C,*

$$d_{RJ,\alpha}(A, C) \leq \frac{\alpha + 1}{2} \left(d_{RJ,\alpha}(A, B) + d_{RJ,\alpha}(B, C) \right).$$

Furthermore, there are finite sets A, B, C such that

$$d_{RJ,\alpha}(A, C) = \frac{\alpha + 1}{2} \left(d_{RJ,\alpha}(A, B) + d_{RJ,\alpha}(B, C) \right).$$

Proof. Equation (4) can be rewritten as

$$\begin{aligned} 1 - \alpha Z_{A,C} &\leq \rho(1 - \alpha X_{A,B} + 1 - \alpha Y_{B,C}) \\ \rho &\geq \frac{1 - \alpha Z_{A,C}}{2 - \alpha X_{A,B} - \alpha Y_{B,C}} \\ &= \frac{\frac{1}{\alpha} - Z_{A,C}}{\frac{2}{\alpha} - X_{A,B} - Y_{B,C}}. \end{aligned}$$

By the above inequality and Lemma 3, we can prove this theorem. □

By substituting the value of α in the previous theorem by 2, we have the following corollary.

Corollary 1. *d_{SD} is a 3/2-relaxed semimetric, i.e., for any finite sets A, B, C,*

$$d_{SD}(A, C) \leq \frac{3}{2} \left(d_{SD}(A, B) + d_{SD}(B, C) \right).$$

Furthermore, there are finite sets A, B, C such that

$$d_{SD}(A, C) = \frac{3}{2} \left(d_{SD}(A, B) + d_{SD}(B, C) \right).$$

In order to prove Lemma 3, we show several properties of A^*, B^*, C^* though Propositions 2-5, Lemmas 2,3. As shown in Figure 1(a), we show that $B^* \backslash (A^* \cup B^*)$, $(A^* \cup C^*) \backslash B^*$, $A^* \backslash (B^* \cup C^*)$, and $C^* \backslash (A^* \cup B^*)$ must be empty sets in Proposition 2,3,4,5 respectively. By those propositions, we can consider A^*, B^*, C^* in the form shown in Figure 1(b). Then, we will prove that all sets A^*, B^*, C^* in that form must have ρ less than or equal to $(\alpha + 1)/2$ in Lemmas 1,2.

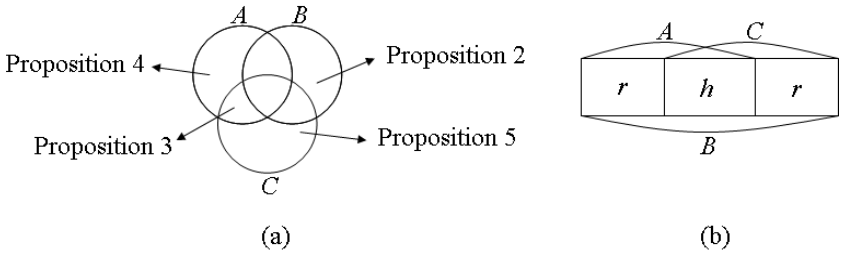

Fig. 1. An outline of our proof of Theorem 1

Proposition 2. *If $B' \setminus (A' \cup C') \neq \emptyset$ then $f(A', B', C') < \max_{A,B,C} f(A, B, C)$.*

Proof. Assume $e \in B' \setminus (A' \cup C')$. Let $A'' = A'$, $B'' = B' \setminus \{e\}$ and $C'' = C'$.

Since $|A' \cap B'| = |A'' \cap B''|$ and $|A' \cup B'| - 1 = |A'' \cup B''|$, we have

$$\begin{aligned} X_{A'',B''} &= \frac{|A'' \cap B''|}{|A'' \cup B''| + (\alpha - 1)|A'' \cap B''|} \\ &= \frac{|A' \cap B'|}{|A' \cup B'| + (\alpha - 1)|A' \cap B'| - 1} \\ &> \frac{|A' \cap B'|}{|A' \cup B'| + (\alpha - 1)|A' \cap B'|} = X_{A',B'}. \end{aligned}$$

Similarly, $|B' \cap C'| = |B'' \cap C''|$ and $|B' \cup C'| - 1 = |B'' \cup C''|$, we have

$$\begin{aligned} Y_{B'',C''} &= \frac{|B'' \cap C''|}{|B'' \cup C''| + (\alpha - 1)|B'' \cap C''|} \\ &= \frac{|B' \cap C'|}{|B' \cup C'| + (\alpha - 1)|B' \cap C'| - 1} \\ &> \frac{|B' \cap C'|}{|B' \cup C'| + (\alpha - 1)|B' \cap C'|} = X_{B',C'}. \end{aligned}$$

Since $A'' = A'$ and $C'' = C'$, we have $Z_{A'',C''} = Z_{A',C'}$. Then,

$$\begin{aligned} f(A'', B'', C'') &= \frac{1/\alpha - Z_{A'',C''}}{2/\alpha - X_{A'',C''} - Y_{A'',C''}} \\ &> \frac{1/\alpha - Z_{A',C'}}{2/\alpha - X_{A',C'} - Y_{A',C'}} = f(A', B', C') \end{aligned}$$

□

To show the next proposition, we need the following lemma.

Lemma 1. *Let N, M be positive integers such that $M \geq hN$ and c be a natural number smaller or equal to h. We have*

$$\frac{N+1}{M+c} \geq \frac{N}{M}.$$

Furthermore, when c is strictly smaller than h. We have

$$\frac{N+1}{M+c} > \frac{N}{M}.$$

Proof. From the first part of this lemma statement, we have

$$\begin{aligned} (N+1)M &\geq N(M+c) \\ NM + M &\geq NM + Nc \\ M &\geq Nc. \end{aligned}$$

By the same argument, we can also prove the second part of this lemma. □

Proposition 3. *If* $(A' \cap C') \setminus B' \neq \emptyset$ *then* $f(A', B', C') < \max_{A,B,C} f(A, B, C)$.

Proof. Assume $e \in (A' \cap C') \setminus B'$. Let $A'' = A'$, $B'' = B' \cup \{e\}$ and $C'' = C'$. Then, $|A' \cap B'| + 1 = |A'' \cap B''|$ and $|A' \cup B'| = |A'' \cup B''|$. By Lemma 1, we have

$$\begin{aligned} X_{A'',B''} &= \frac{|A'' \cap B''|}{|A'' \cup B''| + (\alpha - 1)|A'' \cap B''|} \\ &= \frac{|A' \cap B'| + 1}{|A' \cup B'| + (\alpha - 1)|A' \cap B'| + \alpha - 1} \\ &> \frac{|A' \cap B'|}{|A' \cup B'| + (\alpha - 1)|A' \cap B'|} = X_{A',B'}. \end{aligned}$$

Similarly, $|B' \cap C'| + 1 = |B'' \cap C''|$ and $|B' \cup C'| = |B'' \cup C''|$. By Lemma 1,

$$\begin{aligned} Y_{B'',C''} &= \frac{|B'' \cap C''|}{|B'' \cup C''| + (\alpha - 1)|B'' \cap C''|} \\ &= \frac{|B' \cap C'| + 1}{|B' \cup C'| + (\alpha - 1)|B' \cap C'| + \alpha - 1} \\ &> \frac{|B' \cap C'|}{|B' \cup C'| + (\alpha - 1)|B' \cap C'|} = Y_{B',C'}. \end{aligned}$$

Since $A'' = A'$ and $C'' = C'$, we have $Z_{A'',C''} = Z_{A',C'}$. Then,

$$\begin{aligned} f(A'', B'', C'') &= \frac{1/\alpha - Z_{A'',C''}}{2/\alpha - X_{A'',C''} - Y_{A'',C''}} \\ &> \frac{1/\alpha - Z_{A',C'}}{2/\alpha - X_{A',C'} - Y_{A',C'}} = f(A', B', C') \end{aligned}$$

□

Up until now, we know that $B^* - (A^* \cup C^*)$ and $(A^* \cup C^*) - B^*$ must be an empty set. In the next proposition, we will show that there must be sets A^*, B^*, C^* that maximize function f and $A^* - (B^* \cup C^*)$ is also an empty set.

Proposition 4. *If* $A' \setminus (B' \cap C') \neq \emptyset$ *then there exists* A, B, C *such that* $B - (A \cup C) = \emptyset, (A \cup C) - B = \emptyset, A - (B \cup C) = \emptyset$ *and* $f(A, B, C) \geq f(A', B', C')$.

Proof. Assume $e \in A' \setminus (B' \cap C')$. Let $A'' = A' \setminus \{e\}$, $B'' = B' \cup \{e\}$ and $C'' = C' \cup \{e\}$. Since $|A'' \cap B''| = |A' \cap B'|$ and $|A'' \cup B''| = |A' \cup B'|$, we have $X_{A'',B''} = X_{A',B'}$.

By $|B'' \cap C'''| = |B' \cap C'| + 1$, $|B'' \cup C'''| = |B'' \cup C'''| + 1$, and Lemma 1, we have

$$\begin{aligned} Y_{B'',C''} &= \frac{|B'' \cap C''|}{|B'' \cup C''| + (\alpha - 1)|B'' \cap C''|} \\ &= \frac{|B' \cap C'| + 1}{|B' \cup C'| + (\alpha - 1)|B' \cap C'| + \alpha} \\ &\geq \frac{|B' \cap C'|}{|B' \cup C'| + (\alpha - 1)|B' \cap C'|} = Y_{B',C'}. \end{aligned}$$

Since $|A'' \cap C''| = |A' \cap C'|$ and $|A'' \cup C''| = |A' \cup C'|$, we have $Z_{A'',C''} = Z_{A',C'}$. Then,

$$f(A'', B'', C'') = \frac{1/\alpha - Z_{A'',C''}}{2/\alpha - X_{A'',C''} - Y_{A'',C''}} \geq \frac{1/\alpha - Z_{A',C'}}{2/\alpha - X_{A',C'} - Y_{A',C'}} = f(A', B', C').$$

If $A'' - (B'' \cup C'')$ is an empty set, then we can prove this proposition. If not, we can take another element out of $A'' - (B'' \cup C'')$ and add that element to $(B'' \cap C'') - A''$. By the same argument, we know that the value of function f is not decreased. We can do the same action until the set $A'' - (B'' \cup C'')$ becomes an empty set. After the loop, we will have A, B, C that satisfy the lemma statement.

By the previous proposition, we know there exists at least one set of A^*, B^*, C^* such that $B^* - (A^* \cup C^*) = \emptyset, (A^* \cup C^*) - B^* = \emptyset, A^* - (B^* \cup C^*) = \emptyset$ and $A^*, B^*, C^* = \arg\max A, B, C f(A, B, C)$. The next theorem will show that $C^* - (A^* \cup B^*)$ can also be an empty set.

Proposition 5. *Suppose that $B' - (A' \cup C') = \emptyset, (A' \cup C') - B' = \emptyset, A' - (B' \cup C') = \emptyset$. Then, if $C' - (A' \cup B') \neq \emptyset$, there are finite sets A, B, C such that $B - (A \cup C) = \emptyset, (A \cup C) - B = \emptyset, A - (B \cup C) = \emptyset$, and $f(A, B, C) \geq f(A', B', C')$.*

Proof. Assume $e \in C' \setminus (A' \cup B')$. Let $A'' = A' \cup \{e\}$, $B'' = B' \cup \{e\}$ and $C'' = C' \setminus \{e\}$. By $|A'' \cap B''| = |A' \cap B'| + 1$, $|A'' \cup B''| = |A'' \cup B''| + 1$, and Lemma 1, we have

$$\begin{aligned} X_{A'',B''} &= \frac{|A'' \cap B''|}{|A'' \cup B''| + (\alpha - 1)|A'' \cap B''|} \\ &= \frac{|A' \cap B'| + 1}{|A' \cup B'| + (\alpha - 1)|A' \cap B'| + \alpha} \\ &\geq \frac{|A' \cap B'|}{|A' \cup B'| + (\alpha - 1)|A' \cap B'|} = X_{A',B'}. \end{aligned}$$

Since $|B'' \cap C''| = |B' \cap C'|$ and $|B'' \cup C''| = |B' \cup C'|$, we have $Y_{B'',C''} = Y_{B',C'}$.

Since $|A'' \cap C''| = |A' \cap C'|$ and $|A'' \cup C''| = |A' \cup C'|$, we have $Z_{A'',C''} = Z_{A',C'}$. Then,

$$f(A'', B'', C'') = \frac{1/\alpha - Z_{A'',C''}}{2/\alpha - X_{A'',C''} - Y_{A'',C''}} \geq \frac{1/\alpha - Z_{A',C'}}{2/\alpha - X_{A',C'} - Y_{A',C'}} = f(A', B', C').$$

If $A'' - (B'' \cup C'')$ is an empty set, then we can prove this proposition. If not, we can take another element out of $A'' - (B'' \cup C'')$ and add that element to $(B'' \cap C'') - A''$. By the same argument, we know that the value of function

f is not decreased. We can do the same action until the set $A'' - (B'' \cup C'')$ becomes an empty set. After the loop, we will have A, B, C that satisfy the lemma statement.

By Propositions 2-5, we know that that if $e \in (A^* \cup B^* \cup C^*)$, either $e \in (A^* \cap B^*) \setminus C^*$, $e \in (A^* \cap B^* \cup C^*$, or $e \in (B^* \cap C^*) \setminus A^*$. Meaning that

$$B^* = A^* \cup C^*.$$

With this result, we can state the following lemmas:

Lemma 2. *When $\alpha > 1$, $|A^*| = |C^*|$*

Proof. Let assume that $|A^* \cap C^*| = h$ and $|A^*| + |C^*| = m$. With this, Z_{A^*,C^*} is always equal to $h/(m + (\alpha - 2)h)$, and $A^*, B^*, C^* = \arg\max_{A,B,C}(X_{A,B} + Y_{B,C})$.

Let $0 \leq k \leq 1$. We can denote now $|A^*| = km$, and $|C^*| = (1-k)m$. Thus,

$$X_{A,B} + Y_{B,C} = \frac{km}{(\alpha-1)km + m - h} + \frac{(1-k)m}{(\alpha-1)(1-k)m + m - h}.$$

Because

$$\frac{d(X_{A,B} + Y_{B,C})}{dk} = \frac{(\alpha-1)(2k-1)m^2(h-m)((\alpha+1)m - 2h)}{(m(\alpha(k-1)-k)+h)^2(m(-\alpha k + h - 1) + h)^2},$$

and for $\alpha > 1$ it can only be equal to 0 when $k = \frac{1}{2}$.

Therefore $|A^*| = \frac{1}{2}m = |C^*|$. □

Lemma 3. *When $\alpha > 1$, $f(A^*, B^*, C^*) = (\alpha + 1)/2$.*

Proof. Recall that $|A_* \cap C_*| = h$ and $|A^*| = |C^*|$.

In that case $|A^* \setminus C^*| = |A^*| - |A^* \cap C^*| = |C^*| - |A^* \cup C^*| = |C^* \setminus A^*|$. Thus, we can denote $|A^* \setminus C^*| = |C^* \setminus A^*| = r$.

Also recall that $\rho = \frac{\frac{1}{\alpha} - Z_{A^*,C^*}}{\frac{2}{\alpha} - X_{A^*,B^*} - Y_{B^*,C^*}}$, so

$$X_{A^*,B^*} = \frac{r+h}{(\alpha+1)r + \alpha h},$$

$$Y_{B^*,C^*} = \frac{r+h}{(\alpha+1)r + \alpha h},$$

$$Z_{A^*,C^*} = \frac{h}{2r + \alpha h}.$$

Then

$$\rho = \frac{\frac{1}{\alpha} - \frac{h}{2r+\alpha h}}{\frac{2}{\alpha} - \frac{2r+2h}{(\alpha+1)r+\alpha h}} = \frac{(\alpha+1)r + \alpha h}{2r + \alpha h}.$$

Therefore ρ is maximized when $|A^* \cap C^*| = h = 0$. □

The proof of Lemma 2 works only for the case when $\alpha > 1$, but we know from [3] that the RJ distance is a metric ($\rho = 1$) when $\alpha = 1$. That makes our theorem hold for any positive integer α.

We can easily construct an example to show that the bound obtained in Theorem 1 is tight. That is when $A = \{1\}$, $B = \{1, 2\}$, $C = \{2\}$. We have

$$d_{RJ,\alpha}(A, B) = d_{RJ,\alpha}(B, C) = d_{RJ,\alpha}(B, C) = 1 - \frac{\alpha}{1+\alpha} = \frac{1}{\alpha},$$

and $d_{RJ,\alpha}(B, C) = 1$. Then,

$$1 \leq \rho\left(\frac{1}{\alpha} + \frac{1}{\alpha}\right)$$
$$\rho \geq \frac{\alpha}{2}.$$

4 Conclusions and future work

In this paper, we have proposed a family of similarity indexes, named as the robust Jaccard index, for datasets with missing information. The only parameter of the index is α. When α gets larger, the dissimilarity becomes more robust. However, we have shown in this paper that the value of ρ in the approximate triangle inequality also becomes larger, when the value of α increases. Because, when the value of ρ gets larger, the algorithms proposed for this semi-metric spaces will be less efficient, we have to trade between the robustness and the efficiency of algorithms.

Because of that, we plan to perform experiments to see what is the optimal value of α in each dataset. Besides, We are aiming to find the value of ρ for the general Trevsky distance in our future work. We want to find the relationship between the symmetricity, robustness, and efficiency with those results.

Acknowledgement: The authors would like to thank Mr. Naoto Osaka and Prof. Hiroshi Imai for several useful comments during the course of this research.

References

1. Jaccard, P.: Lois de distribution florale dans la zone alpine. Corbaz (1902)
2. Tanimoto, T.: An elementary mathematical theory of classification and prediction. Technical report, IBM Report (1958)
3. Lipkus, A.H.: A proof of the triangle inequality for the Tanimoto distance. Journal of Mathematical Chemistry **26**(1-3) (1999) 263–265
4. Jain, K., Vazirani, V.V.: Primal-dual approximation algorithms for metric facility location and k-median problems. In: FOCS'99. (1999) 2–13
5. Ruiz, E.V.: An algorithm for finding nearest neighbours in (approximately) constant average time. Pattern Recognition Letters **4**(3) (1986) 145–157
6. Sankoff, D., Rousseau, P.: Locating the vertices of a Steiner tree in an arbitrary metric space. Mathematical Programming **9**(1) (1975) 240–246

7. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. Biologiske Skrifte **5** (1948) 1–34
8. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3) (1945) 297–302
9. McCune, B., Grace, J.B., Urban, D.L.: Analysis of ecological communities. Volume 28. MjM software design Gleneden Beach, OR (2002)
10. Looman, J., Campbell, J.: Adaptation of Sorensen's K (1948) for estimating unit affinities in prairie vegetation. Ecology (1960) 409–416
11. Gragera, A.: Approximate matching for Go board positions. In: GPW'15. (2015)
12. Schubert, A., Telcs, A.: A note on the Jaccardized Czekanowski similarity index. Scientometrics **98**(2) (2014) 1397–1399
13. Braverman, V., Meyerson, A., Ostrovsky, R., Roytman, A., Shindler, M., Tagiku, B.: Streaming k-means on well-clusterable data. In: SODA'11. (2011) 26–40
14. Mettu, R.R., Plaxton, C.G.: The online median problem. SIAM Journal on Computing **32**(3) (2003) 816–832
15. Jaiswal, R., Kumar, M., Yadav, P.: Improved analysis of D2-sampling based PTAS for k-means and other clustering problems. Information Processing Letters **115**(2) (2015) 100–103
16. Tversky, A., Gati, I.: Similarity, separability, and the triangle inequality. Psychological review **89**(2) (1982) 123
17. Jimenez, S., Becerra, C., Gelbukh, A., Bátiz, A.J.D., Mendizábal, A.: Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In: *SEM'13. (2013) 194–201