LIVES
Swiss National Centre of Competence in Research

FNSNF
Swiss National Science Foundation

Elzinga, C. H. & M. Studer (2016)

Normalization of Distance and Similarity in Sequence Analysis

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 445-468.

UNIVERSITÉ DE GENÈVE

FONDATION POUR L'UNIVERSITÉ DE LAUSANNE

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

UNIL | Université de Lausanne
Institut des sciences sociales

Normalization of Distance and Similarity in Sequence Analysis

C. (Cees) H. Elzinga, M. (Matthias) Studer

Abstract We explore the relations between the notion of distance and a feature set based concept of similarity and show that this concept of similarity has a spatial interpretation that is complementary to distance: it is interpreted as "direction". Furthermore, we show how proper normalization leads to distances that can be directly interpreted as dissimilarity: closeness in normalized space implies and is implied by similarity of the same objects while remoteness implies and is implied by dissimilarity. Finally, we show how, in research into de-standardization of the life course, properly normalizing may drastically and unequivocally change our interpretation of inter-cohortal distances.[1]

Key Words: distance, similarity, normalization, sequence analysis.

C. (Cees) H. Elzinga
Netherlands Interdisciplinary Demographic Institute, The Hague,
VU University Amsterdam, the Netherlands e-mail: c.h.elzinga@vu.nl

M. (Matthias) Studer
Faculty of Social Sciences, University of Geneva
NCCR Program "LIVES", Switzerland e-mail: matthias.studer@unige.ch

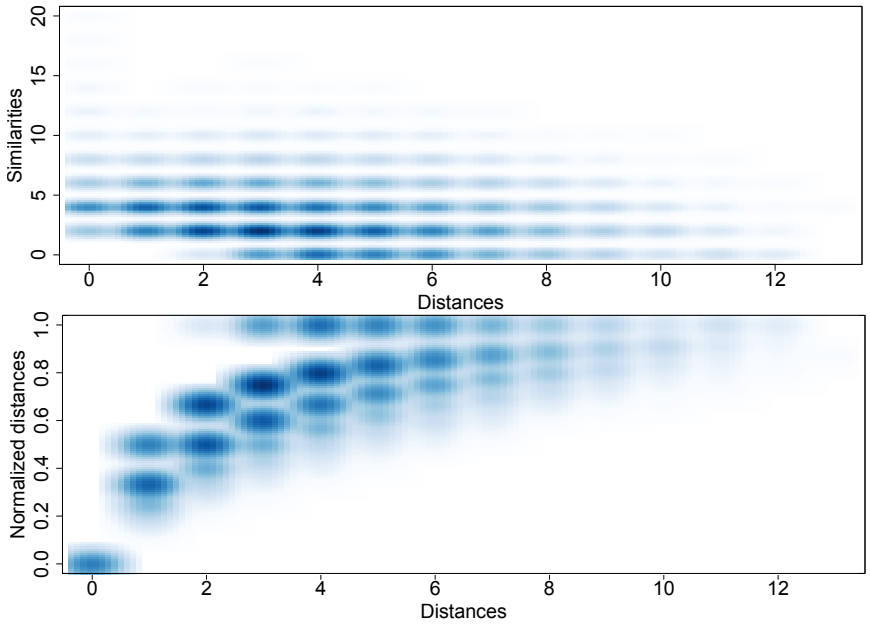
[1] This paper has been accepted for publication in *Sociological Methods & Research*.

1 Introduction

In science, the concept of distance is used in two, quite different ways. First and oldest, in physics, distance or length is one of the fundamental dimensions that are used to express physical quantities (see e.g. Pfanzagl, 1968; Krantz et al., 1971). These quantities in turn are used to formulate models of natural phenomena. For example, the gravitational force that two masses exert on each other is inversely proportional to the squared distance between the masses.
In the social and behavioral sciences, spatial concepts are abundantly applied in different methodologies like multidimensional scaling (Borg and Groenen, 2005), multivariate linear statistics (Rencher and Cristensen, 2012), classification (Shawe-Taylor and Cristianini, 2004), clustering (Hennig et al., 2015) and sequence analysis (Blanchard et al., 2014). Spatial concepts are used in fields as diverse as clinical psychology, social demography and political science. However, in these sciences, distance is not one of the fundamental dimensions of theory. Instead, social scientists and psychologist calculate distances between personalities, political programs or labour market careers with the objective to judge or gauge the similarity between their objects of interest, eventually sorting or grouping these objects into more or less homogeneous subsets or projecting the objects onto a limited number of dimensions. In doing so, these colleagues assume that similarity is the opposite of distance: the more distant, the less similar and vice versa. This way of using distances is not unique for the social sciences; for example, chemists calculate distances between graph-representations of molecules (see e.g. Borgwardt, 2011) with the objective of finding similar but cheaper or more effective variants.
However, the relation between distance and similarity is not obvious as distance relations derive from spatial considerations and similarity relations derive from considering common and non-common feature sets (see e.g. Tversky, 1977). For a qualitative review of the concept of similarity, the reader is referred to Decock and Douven (2011).
So, grouping objects on the basis of distance may not result in groups of objects that are similar. Peeking around the corner at the issues dealt with in the next sections, we plotted distances versus similarities for the same pairs of objects in Figure 1. Indeed, distance and similarity are not simply anti-monotone and equal distance does not imply nor is implied by equal similarity. On the other hand, given a distance d, we can always calculate a similarity $s = h(d)$ and the reverse is also possible: given a similarity s, we can construct[2] a distance d from it through $s = h'(d)$. However, these calculations are not very intuitive. Fortunately, when we properly convert a distance d into a normalized distance D, we will obtain a normalized similarity S through the intuitive $S = 1 - D$. This latter transformation is the most simple and straightforward expression that formalizes the widely held beliefs that close objects are similar, that dissimilar objects are remote, etc. But apparently, we first have to normalize either the distance or the similarity before we can apply such an intuitive transform. In Figure 2 we illustrate the conceptual relations between (normalized) distance and similarity.
In this paper, we build on what was discussed in e.g. Chen et al. (2009); Elzinga et al. (2011) and in Elzinga (2014a). The discussion is quite general in the sense that it pertains to all measures of distance, whichever the context or application. However, we will focus on sequence analysis and to illustrate, we will use a publicly available data set and a measure of distance that has become popular in that context. For an introduction to sequence analysis, the reader is referred to Martin and Wiggins (2009) or to Cornwell (2015); an overview of distance measures used in this context is provided in

[2] It should be noted that $h' \neq h^{-1}$. It will appear that neither h nor h' is a 1-1 function: different distances may be mapped onto the same similarity and different similarities may be mapped onto the same distance.

Fig. 1 Upper panel: scatter plot of distances (horizontal axis, metric: OMspell) versus similarities (vertical axis). Pearson's $r = -.29$. Lower panel: scatter plot of distances (horizontal axis) versus normalized distance. Pearson's $r = .75$. Darker areas contain more points. The data used in both panels were published in McVicar and Anyadike-Danes (2002).



Studer and Ritschard (2015). For now, it suffices to define a sequence as an ordered set of labeled states or events that, depending on the specific application, may have an associated"“time-stamp". In the case of states like "living single" or "being unemployed", the time-stamp is interpreted as "duration" and in case of events like "becoming a parent" or "starting parental leave", the time-stamp is interpreted as a date or amount of lapsed time.

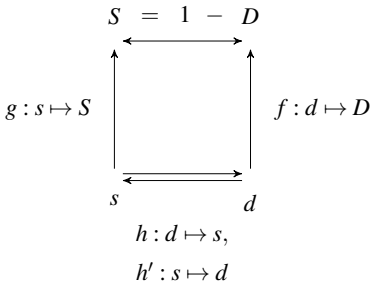
The purpose of this paper is threefold. First, we will explore the relations between distance as a spatial concept and similarity as a feature set based concept. We will conclude that the spatial interpretation of similarity is "direction" or angle, similar objects being located in (almost) the same direction, relative to some arbitrary but fixed reference object. This interpretation allows us to consider distance and similarity as complementary in the spatial structuring of a set of objects.

Second, we will discuss normalization as a means to transform the object space in such a way that distance and similarity become opposites in the sense that similar objects are close and remote objects are dissimilar, etc..

Third, we will show that normalization may affect the way that we interpret differences between life courses of different age-cohorts.

The paper is structured as follows: In the next Section 2, we discuss the concepts of distance and

Fig. 2 Formal relations between distance d, similarity s and their normalized versions D and S. Neither of the mappings f, g, h or h' is 1-1. Therefore, we do not write $h^{-1} = h'$.



similarity and illustrate them by discussing classes of distance measures that have become popular in the context of sequence analysis. In its final subsection 2.4, we deal with the spatial interpretation of similarity through discussing transforms from distance to similarity and vice versa. In Section 3, we discuss normalization and in Section 4, we will apply normalization and discuss some of its effects. In Section 5, we make some concluding remarks.

2 Distance and Similarity

2.1 Distance

We briefly discuss the formal concept of distance. A function $d : X \times X \mapsto \mathbb{R}$ that maps pairs of objects from an object set X onto nonnegative (real) numbers is called a "distance" or, equivalently, a "(distance) metric" if it satisfies the following conditions or axioms for all triples of elements from the object set:

D1: $d(x,x) = 0$	"one location per object"
D2: $d(x,y) > 0$	"one object per location"
D3: $d(x,y) = d(y,x)$	symmetry
D4: $d(x,y) \leq d(x,z) + d(z,y)$	triangular inequality (TI)

The first three axioms are quite intuitive to most readers so we only elaborate on D4, the triangular inequality. Axiom D4 says that if two objects x and y are close to a third object, say z, the objects cannot be far apart. At the same time, this is a way of saying that the space, i.e. the set of objects structured by d, cannot be irregular in the sense that all distances are confined to boundaries imposed by other distances. From the TI, it follows that, for all triples of objects (x,y,z), we have

$$|d(x,z) - d(z,y)| \leq d(x,y) \leq d(x,z) + d(z,y). \quad (1)$$

Hence, the distance between a particular pair (x,y) is bounded by the distances $d(x,z)$ and $d(z,y)$ for *all* possible objects z in $X - \{x,y\}$ (see e.g. Elzinga and Studer, 2015). Without such boundaries, actually gauging distances between objects would not be very informative since, when presented with a new object, we would have no idea about its possible distance to the objects already known, i.e. d would not impose much structure on the object set.
The reader notes that the axioms D1-D4 only restrict the method of actually measuring distances, they do not specify or favour any method in particular; they just formalize our intuitions about a particular spatial relation.

2.2 Distances for Sequences

Here, we discuss two broad classes of distance measures that are often used in the context of sequence analysis. For a detailed overview of such measures, the reader is referred to Deza and Deza (2014); Studer (2012); Studer and Ritschard (2015).
The first class, and by far the most popular one in applications of sequence analysis, is the class of so-called edit-distances: an edit distance is a function that counts the minimum number of (weighted) edit-operations that is necessary to turn one sequence into a perfect copy of the other sequence (see e.g. Navarro, 2001). The smaller this number, the smaller the distance between the sequences. In the social sciences, edit-distances were introduced by Abbott and Forrest (1986) and are known as "OM-distances"; "OM" being the acronym for "Optimal Matching". Quite a variety of edit distances has been proposed, each variant defining a distinct way of weighing the edit-operations and the pairs of characters involved. Here, we will denote an edit-distance by writing d_E.
An alternative class of distance functions for sequences has been proposed by Elzinga and Studer (2015) and derives from representing sequences as non-negative, infinite-dimensional vectors and defines the distance between the sequences as the Euclidean distance between the vector-representations. Formally, let $X\{x,y,\ldots\}$ denote a set of sequences and let $\mathbf{X} = \{\mathbf{x},\mathbf{y},\ldots\}$ denote the set of vectors, representing the sequences in X. Furthermore, let $\mathbf{x}'\mathbf{y}$ denote an inner product of the vectors in $\mathbf{X}$. Then the distance d_K is defined (see e.g. Golub and Van Loan, 2013) as the Euclidean vector-distance

$$d_K(x,y) = \sqrt{\mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{y}}. \quad (2)$$

Within this class, distance measures differ because of the different methods used to define the vectors and/or the inner vector-product. The algorithms required to evaluate the inner products are called "kernels" (see e.g. Shawe-Taylor and Cristianini, 2004) and hence, the distances obtained from such algorithms are subscripted as d_K.

2.3 Similarity

Similarity too is a function that maps pairs of objects to the real numbers but its properties do not derive from spatial relations. Instead, they derive from the general notion that similarity between objects is determined by the number of features that the objects share. A function $s : X \times X \mapsto \mathbb{R}$ that maps pairs[3] of objects to non-negative real numbers is a similarity if it satisfies, for *all* triples of objects $x, y, z \in X$:

S1 $s(x,y) \geq 0$	nonnegativity
S2 $s(x,y) \leq \min\{s(x,x), s(y,y)\}$	"bounded by self-similarity"
S3 $s(x,y) = s(y,x)$	symmetry
S4 $s(x,y) + s(z,z) \geq s(x,z) + s(z,y)$	covering inequality (CI)

Here, we elaborate on S2 and S4. If similarity somehow depends on sets of common and non-common features, S2 states that the set of features that x and y have in common cannot be bigger than the smallest of the feature sets of each of x and y. S4 is an inequality that, like TI, regularizes or smoothes the similarity space since similarities are bounded by other sums of similarities. Together, axioms S2 and S4 imply, for all triples $x, y, z \in X$, that $s(x,y)$ satisfies

$$\min\{s(x,x), s(y,y)\} \geq s(x,y) \geq s(x,z) + s(z,y) - s(z,z). \tag{3}$$

Two issues deserve some comment. The first pertains to "self-similarity" $s(x,x)$. Unlike D1: $d(x,x) = 0$, the system S1-S4 only limits $s(x,x)$ in S2 and allows for $0 < s(x,x) \neq s(y,y)$. At first sight, this may seem counter-intuitive: "x is more (or less) similar to itself than y". In spatial models of similarity (see e.g. Torgerson, 1965), this representational issue does not arise since $d(x,x) = 0$ for *all* objects. The system S1-S4 is based on the counting of feature sets and the result is the possibility of non-identical self-similarities: similarity depends on the set of common features. If however we interpret such counts as "description lengths" or "complexities" (Elzinga, 2010; Gabadinho et al., 2010), unequal self-similarities become quite natural and if x has more features than y, we have that $s(x,x) > s(y,y)$. We will come back on this in our discussion of the spatial interpretation of similarity (subsection 2.5) and the issue will be resolved when we discuss normalized similarity in subsection 3.2.
The second issue pertains to S3, the symmetry of similarity. We know that this axiom often fails when s is interpreted as *perceived* similarity, perhaps due to "context-switching" (see e.g Gärdenfors, 2000). Here we consider models of cognitive similarity as irrelevant although we are aware of the fact that the concept of similarity should not be fully decoupled from its role in our everyday experience.
For a more detailed account of the axiomatisation of the concept of similarity, the reader is referred to Chen et al. (2009) or Elzinga (2014a). We confine to discussing two simple similarities, one edit-based and one kernel-based.

[3] Similarity between more than two objects is dealt with in e.g. Elzinga et al. (2011)

2.3.1 An edit-based similarity

The concept of similarity has been widely ignored among those applying sequence analysis, probably because of two reasons. First, most sequence analysis software generates distances, not similarities. Second, there is a generally held belief that remoteness implies dissimilarity and that closeness implies similarity. But despite this belief, it is not clear if and how similarity and distance are related. However, when we calculate the length ℓ of the longest common subsequence (lcs), then (Wagner and Fisher, 1974, Section 5)

$$\ell(x,y) = \frac{1}{2}(|x| + |y| - d_E(x,y)), \tag{4}$$

and $\ell(x,y)$ satisfies the similarity axioms S1-S4. Here, $|x|$ denotes the length of sequence x and $d_E(x,y)$ denotes the edit-distance based on unit insertion- and deletion-cost. It is almost immediate that ℓ satisfies axioms S1-S3 but it is not trivial to see that ℓ satisfies S4 too. A proof of this claim is presented in the Appendix 1. Hence, although there are many faster algorithms to calculate ℓ (see e.g. Bergroth et al., 2000, 2003), it can be considered as an edit-based similarity. Later we will generalize to the case of a metric edit-cost structure.

2.3.2 Kernel-based similarity

Kernel algorithms evaluate inner products of vectors, i.e. they evaluate $\mathbf{x}'\mathbf{y} = \sum_i \mathrm{x}_i \mathrm{y}_i$ for vectors $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots)$ and $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots)$ without directly using the vectors themselves. Thus, if we want to use kernels to define and/or evaluate similarities, we first have to represent objects by vectors, so called "feature-vectors". Let $\mathcal{F} = \{u, v, \ldots\}$ denote a set of features and when an object x possesses a feature u, we denote this fact by writing $u \sqsubseteq x$. To construct feature vectors, we first index the features by assigning to each and every feature u a unique positive integer $r(u)$ that is as small as possible. So, with n distinct features, the features will be indexed by the integers $1, \ldots, n$. Now we may construct a vector representation $x \mapsto \mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots)$ for an object x by assigning to each coordinate a non-negative value through

$$\mathrm{x}_{r(u)} = \begin{cases} c(u) \text{ if } u \sqsubseteq x \\ \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

This in effect means that we assign a value $c(u)$ to the i^{th} coordinate if the object x possesses the i^{th} feature u and if x does not possess that feature, the coordinate is set to zero. If, for example $c(u) = 1$ for all features, the resulting binary vector shows which features are or are not possessed by the object represented. The inner product of such binary vectors is then a count of the features that the two objects have in common. If $c(u)$ varies with the features, the inner product evaluates a weighted count of common features. In all cases where the coordinate values *only* depend on the features, the resulting inner product $\mathbf{x}'\mathbf{y}$ is a similarity, i.e. $s(x,y) = \mathbf{x}'\mathbf{y}$ satisfies the similarity axioms S1-S4. In the context of sequence analysis, the set of features can be taken to be the set of all possible subsequences (see e.g. Elzinga and Studer, 2015), in which case an inner product counts the number of distinct common subsequences.

To actually calculate the inner product of feature vectors that use the possible subsequences as coordinates, kernel algorithms have been designed that allow for calculation times that are roughly proportional to the third power of the length of the sequences involve. These kernels have been described and analysed in e.g. Elzinga and Wang (2013); Elzinga and Studer (2015) and in Shawe-Taylor and Cristianini (2004). If however the coordinate values not only depend on the features as such, but also depend on the way these features are "embedded" in the pertaining sequences, i.e. when

$$\mathrm{x}_{r(u)} = \begin{cases} c(u,x) \text{ if } u \sqsubseteq x \\ \\ 0 \text{ otherwise} \end{cases}, \tag{6}$$

the inner products do not necessarily satisfy the axioms of similarity. For example, if $c(u,x)$ is a value that depends on the embedding frequency or the duration of u in x, $\mathbf{x}'\mathbf{y}$ will *not* satisfy the axioms of similarity[4]. Therefor, in the next subsection, we discus how to construct a similarity given a distance and vice versa.

2.4 From Distance to Similarity and Back

We already discussed a particular construction of a similarity from a distance: in Equation 4 , we transformed an edit-based distance d_E into an edit-based similarity $\ell(x,y)$. Unfortunately, there is no general and explicit solution to this problem, i.e. there is no (known) general specification of the map $h : d \mapsto s$, nor is there a general solution for the reverse problem - constructing a similarity from a distance. Some quite broad classes of solutions were derived by Chen et al. (2009) and some results of Yianilos (2002) are useful too. Here we discuss simple instances of these classes: one instance of $h(d) = s$ and one instance of $h'(s) = d$. For most applications, these examples will suffice.
First we discuss a solution to $s = h(d)$: let $X = \{x, y, z, r, \ldots\}$ denote an object set and d a metric on $X \times X$. Then, for an arbitrary but fixed reference object r,

$$s(x,y) = d(x,r) + d(y,r) - d(x,y) \tag{8}$$

is a similarity[5]. To see that this is true indeed, we check if the axioms S1-S4 hold.
Because of $d(x,y) \leq d(x,r) + d(y,r)$ (TI), $s(x,y) \geq 0$ so S1 holds. Using Inequality 1 and the above Equation 8, we derive

[4] As an example, we define $c(u,x)$ as the embedding frequency of the subsequence u of sequence x. Consider the sequences $x = aabab$ and $y = aabb$. Then, for example, we have that $c(ab,x) = 5$ and $c(abb,y) = 4 = c(abb,x)$. Now the inner product $\mathbf{x}'\mathbf{y}$ can be computed through

$$\mathbf{x}'\mathbf{y} = \sum_u c(u,x)c(u,y). \tag{7}$$

By constructing a table that lists all common subsequences of x and y and their embedding frequencies, the reader will discover that $\mathbf{x}'\mathbf{y} = 47$, that $\mathbf{x}'\mathbf{x} = 83$ and that $\mathbf{y}'\mathbf{y} = 35$. Taking the inner product for a similarity would violate axiom S2 since $\mathbf{x}'\mathbf{y} > \mathbf{y}'\mathbf{y}$

[5] We should take the reference object r into account in our notation by writing $s_r(\cdot,\cdot)$, but we do not since in this paper, the simpler notation does not lead to ambiguities.

$$s(x,y) \leq d(x,r) + d(y,r) - |d(x,r) - d(y,r)| \tag{9}$$
$$= 2\min\{d(x,r), d(y,r)\} \tag{10}$$
$$= \min\{s(x,x), s(y,y)\}, \tag{11}$$

so $s(x,y)$ is bounded by self-similarity (S2). Symmetry is trivial (S3) and the CI (S4) follows from the TI. Apparently, s as constructed through Equation 8 is a similarity.
A way to construct a distance d from a similarity s is through (Chen et al., 2009)

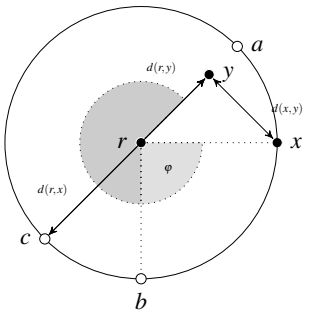
$$d(x,y) = s(x,x) + s(y,y) - 2s(x,y). \tag{12}$$

d is a distance that depends on the size of the symmetric set difference $|\triangle(\mathcal{X},\mathcal{Y})|$. That it is a proper distance is checked by verifying the distance axioms D1-D4. Clearly, $d(x,x) = 0$ and if $d(x,y) = 0$, we must have that $x = y$ because of S2. S2 also ensures that $d(x,y) > 0$ (D2) and symmetry (D3) is trivial. That d satisfies the TI immediately follows from the CI.
However, this formal reasoning does not clarify the conceptual relations between the spatial concept of distance and the set-based concept of similarity. Therefore, in the next subsection, we will provide for a spatial interpretation of similarity and show how the spatial and set-based concepts touch through binary feature vectors.

2.5 Similarity as angle or direction

To spatially interpret the concept of similarity, Equation 8 is a good starting point and to explore its behavior, we constructed Figure 3. Figure 3 shows a plane[6] in which we fixed two objects r and y at

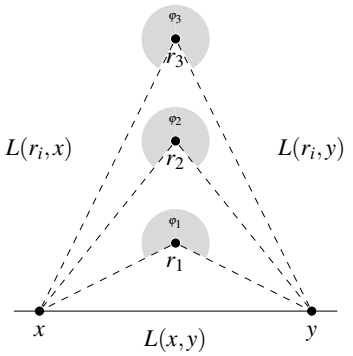
Fig. 3 The similarity $s(x,y)$ is determined by the (fixed) distances $d(x,r) + d(y,r)$ and the *obtuse* angle φ between the lines $L(r,y)$ and $L(r,x)$. See text for a detailed explanation.



[6] not necessarily a coordinated plane

distance $d(r,y)$. Furthermore, we presumed a third object x at distance $d(r,x)$. This latter assumption determines a circle around r that includes x. Having drawn that circle, we fixed an x on it, therewith fixing $d(y,x)$ too. The picture then contains enough information to determine $s(x,y)$ according to Equation 8. Furthermore, we determined the *obtuse* angle φ between the lines $L(r,x)$ and $L(r,y)$.
When we now move the object x along the circle in the direction of location b, φ will decrease and $s(x,y)$ decreases since $d(x,r)$ and $d(y,r)$ are fixed and $d(x,y)$ increases. Once x has reached location c, φ will attain its minimum-value: passing c will cause the obtuse and the acute angle to flip. Also, $d(x,y)$ will reach its maximum value and $s(x,y) = 0$ since, at c, $d(x,y) = d(x,r) + d(r,y)$. Moving x clockwise from c will increase the *obtuse* φ and, since $d(x,y)$ will get smaller, $s(x,y)$ will increase too. At a, φ will reach its maximum value and $s(x,y)$ will be maximal too.
So it seems that φ is monotone with $s(x,y)$: when $s(x,y)$ is maximal, φ is maximal and when $s(x,y)$ is minimal, then x and y are opposite relative to r and φ is minimal. However, in Figure 3, φ is anti-monotone with $d(x,y)$ but this is not necessary. In Figure 4, we fixed the objects x and y and move the reference relative to the line $L(x,y)$. We see that when the reference moves away from the line $L(x,y)$, the obtuse angle φ_i between the lines $L(r_i,x)$ and $L(r_i,y)$ gets bigger and, simultaneously, $s(x,y)$ gets bigger since $d(r_i,x)$ and $d(r_i,y)$ get bigger while $d(x,y)$ remains fixed.
From Figures 3 and 4, we conclude that $s(x,y)$ is monotone with the angle φ: the bigger it gets, the more similar the objects. Equivalently, we could say that x and y are similar to the degree that the *directions* of the lines $L(r,x)$ and $L(r,y)$ coincide. The difference $d(x,r_i) + d(y,r_i) - d(x,y) = s_{r_i}$

Fig. 4 The similarity $s_{r_i}(x,y)$ is measured as the distance from the reference r_i to the line L between x and y, i.e. as monotone with the obtuse angle between the lines $L(r_i,x)$ and $L(r_i,y)$. See text for a detailed explanation.

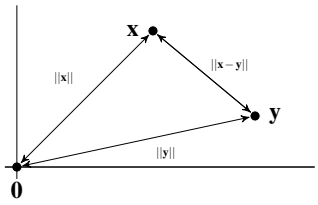


measures the distance of r_i to the line L between x and y. If this distance equals zero, x and y are opposite relative to r_i and hence $s_{r_i} = 0$. The greater the distance of r_i to L, the greater the similarity s_{r_i} the more the directions of x and y, relative to r_i coincide.
In the above interpretation, we did not assume anything about the properties of the object-space (X,d). However, when we assume that the space is a Euclidean coordinate space and we take $\mathbf{0} = (0,0,\ldots)$ as our reference, Equation 8 implies that

$$s(x,y) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|, \tag{13}$$

where $\|x\|$ denotes the "length" of the vector $\mathbf{x}$, or, equivalently, the distance of $\mathbf{x}$ to the reference $\mathbf{0}$, i.e. for $\mathbf{x} = (x_1, x_2, \ldots)$, $\|\mathbf{x}\| = \sqrt{\sum_i x_i^2}$. Hence in a Euclidean vector space, $s(x,y)$ equals the sum of the lengths of $\mathbf{x}$ and $\mathbf{y}$ in as far as this sum is *not* due to the distance between $\mathbf{x}$ and $\mathbf{y}$ (see Figure 5). When

Fig. 5 Illustration of $s(x,y) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|$, which holds when sequences x and y are represented in Euclidean vector space and $\mathbf{0}$ is used as the reference object. In such cases, the similarity is interpreted as the sum of the lengths of the representations $\mathbf{x}$ and $\mathbf{y}$ in as far as this sum is not due to the distance between the two.



the vectors are binary feature vectors, the spatial and set-based interpretation of similarity coincide. For let $\{\mathcal{X}, \mathcal{Y}, \ldots\}$ denote the feature sets of the objects $\{x, y, \ldots\}$ and $\{\mathbf{x}, \mathbf{y}, \ldots\}$ their representing vectors, then we have that $\|\mathbf{x}\| = \sqrt{|\mathcal{X}|}$ and $\|\mathbf{x}-\mathbf{y}\| = \sqrt{|\triangle(\mathcal{X},\mathcal{Y})|}$ with $\triangle(X,Y) = (X-Y) \cup (Y-X)$ and hence that

$$s(x,y) = \sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{Y}|} - \sqrt{|\triangle(\mathcal{X},\mathcal{Y})|}. \tag{14}$$

So, for binary feature vectors, the similarity defined by Equation 8 depends in a nonlinear way on the size of the set of features common to both x and y.

3 Normalizing Distance and Similarity

As will appear later, once we have constructed a normalized distance, denoted as D, the evaluation of a normalized similarity S simply amounts to evaluating the quantity $S = 1 - D$. So, once we have distances available and know how to normalize them, the calculation of similarity is easy.

The simple relation between normalized distances and similarities is the main, but not the only reason to discuss the subject of normalizing distances. The other reason is that normalizing greatly facilitates the interpretation of the numbers generated by the algorithm or procedure that generates the distance measurements.

3.1 The Normalization of Distance

Normalizing a distance(-scale) not only requires that we divide out units and map to a closed interval but also that the normalized scale still adheres to the metric axioms D1-D4. Simultaneously imposing these three requirements is not trivial.
A simple normalization[7] of distance is provided by the transform

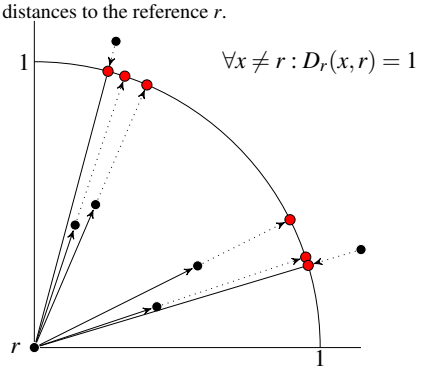
$$D_r(x,y) = \frac{d(x,y)}{(d(x,y)+d(x,r)+d(y,r))/2}. \tag{15}$$

The choice of r determines the scale D_r. The effect of normalization can be understood through looking at the quantities $D_r(x,r)$: according to Equation 15, we have that $D_r(x,x) = 1$ since $d(r,r) = 0$. Because this holds for all $x \neq r$, normalization according to Equation 15 has the effect of projecting all objects other than r on a unit sphere with r at the center and gauging distances between the objects through evaluating distances between the projections. This geometrical interpretation is illustrated in Figure 6.
The effect of remoteness from the reference object can also be understood as follows: assume that both $d(x,r)$ and $d(y,r)$ are equal to, say, $\frac{1}{2}a$. Then $D_r(x,y)$ reduces to

$$D_r(x,y) = \frac{d(x,y)}{(d(x,y)+a)/2}. \tag{16}$$

In Figure 7, we plot this function for an arbitrary but fixed distance $d(x,y)$ and $a \geq d(x,y)$. The plot shows that the normalized distance D_r gets smaller, the more remote x and y are from r, i.e. the bigger a gets.

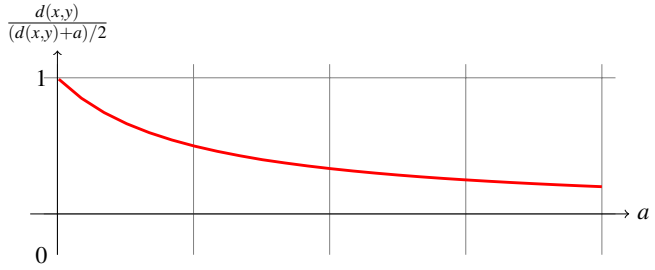
Fig. 6 Normalization amounts to projecting all objects on an r-centered unit-sphere. The reader notes that D_r may produce a different ordering of pairs of objects than d. Normalization is a way to control for variation among the distances to the reference r.



[7] In the sequel, we will write D instead of d to denote a normalized distance

A proof that D_r satisfies the axioms D1-D4 is presented in the Appendix. Summarizing, we have that D_r satisfies the axioms of distance D1-D4 and $0 < D_r(x,y) \leq 1$ if and only if $x \neq y$.
Finally, we note that D_r is compressive with respect to d in the sense that differences between big d-distances tend to be smaller on the D_r-scale (see Figure 8).

Fig. 7 The effect of normalizing a fixed distance $d(x,y)$ with respect to a reference at a distance $\frac{1}{2}a$ from x and y: the bigger a gets, the smaller the normalized distance $D(x,y) = \frac{d(x,y)}{(d(x,y)+a)/2}$ (vertical axis) gets.



We now have discussed normalisation of distances, thereby confining ourselves to just one technique: the one embodied in Equation 15. The reader might be interested in different ways of normalizing distance and indeed, alternatives to Equation 15 can be found in Chen et al. (2009) or in Elzinga (2014a). The behavior of the resulting scales is roughly the same as the properties that we discussed here, so we leave it to the interested reader to explore these alternatives. We now turn our attention to the normalization of similarity.

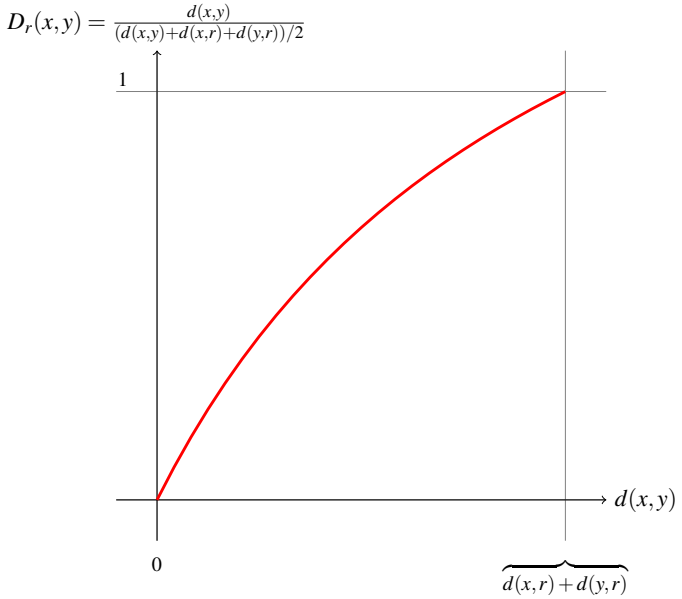
3.2 The Normalization of Similarity

There are two ways to obtain a normalized similarity, the direct and the indirect way. The indirect route is via a normalized distance according to the simple formula

$$S_r(x,y) = 1 - D_r(x,y), \tag{17}$$

and the direct way uses "raw" or non-normalized similarity through one of the many transforms that yield a normalization. Here we start discussing the indirect way, i.e. via Equation 17, since it is simple and because it directly connects similarity to *any* normalized distance. Thereafter, we will just discuss just one of the many possibilities to normalize a similarity, a generalization of the Tanimoto-coefficient. Again, we confine ourselves to just one technique because we believe that discussing more different techniques will not be very productive for social science sequence analysts and again, the interested reader is referred to Elzinga (2014a) for different techniques and more

Fig. 8 Plot of D_r vs d, showing that normalizing is compressive. Because of the triangular inequality, the domain (horizontal axis) is limited to $d(x,r)+d(y,r)$.



literature.
Using Equation 17, the axioms D'1-D'4 can be written in terms of S_r:

S'1 $S_r(x,x) = 1$ for all x,
S'2 $1 > S_r(x,y) \geq 0$ if and only if $x \neq y$,
S'3 $S_r(x,y) = S_r(y,x)$,
S'4 $S_r(x,y) + 1 \geq S_r(x,z) + S_r(z,y)$.

Comparing the system S'1-S'4 with the system S1-S4 reveals that S'1-S'4 is a special case of S1-S4: S2 states that a similarity can never exceed the self-similarity of the pertaining sequences while the same is implied in S'1-S'2. Only S'1-S'2 is more explicit through specifying a uniform upper boundary for all self-similarities and this is reflected in the difference between S'4 and S4. So we conclude that $S_r(x,y)$ as defined in Equation 17 is a normalized similarity: it adheres to the axioms of similarity and it is dimensionless and bounded in $[0,1]$ because D_r is dimensionless and bounded in the same interval.
The reader notes that, since $D_r(x,r) = 1$ for all $x \neq r$, $S_r(x,r) = 0$ for all $x \neq r$, again expressing that normalizing with respect to a reference implies "controlling for differences in distance/similarity to the reference object r".
Next we turn our attention to normalizing similarity without explicitly using a reference object. This

is possible, given a non-normalized similarity s, through evaluating

$$S(x,y) = \frac{s(x,y)}{s(x,x)+s(y,y)-s(x,y)}. \tag{18}$$

This formulation generalizes Tanimoto's coefficient (see e.g. Duda et al., 2001) for similarity in biological taxonomy: if $s(x,y)$ is interpreted as a count of common features, $S(x,y)$ expresses this count as a fraction of the total number of features of x and y.
In Appendix A3, we show that S as defined in Equation 18 satisfies the Covering Inequality S'4.
Normalizing similarity does not involve a reference object, at least not according to Equation 18. But Equation 18 does not say how the similarity s was constructed; if s does not directly derive from a (weighted) set count, s will be derived from a scale that derives from some spatial concept. But then the concept of "direction" will require a reference object: direction is always relative to some fixed point in space.
We discuss one example: we set $s(x,y) = llcs(x,y)$ and hence obtain

$$S(x,y) = \frac{llcs(x,y)}{|x|+|y|-llcs(x,y)} \tag{19}$$

since $llcs(x,x) = |x|$, the length of the sequence x. This formulation shows that normalized similarity can be computed in an edit-oriented sequence analysis. Using Equation 4, we elaborate the above expression to

$$S(x,y) = \frac{|x|+|y|-d_{OM}(x,y)}{|x|+|y|+d_{OM}(x,y)}, \tag{20}$$

expressing the normalized coefficient directly in terms of OM-distances. Distances, plural, since $|x| = d_{OM}(x,\lambda)$, the sum of unit deletion costs when transforming x into λ, the empty sequence. However, when a cost-metric other than the standard unit-cost is used, we have that $d_{OM}(x,\lambda) = \sum_i c(x_i,-)$ where $c(x_i,-)$ denotes the deletion cost of the i-th character of x. This leads to a generalization of $llcs$ to the general OM-context with any metric cost-structure. Thereto, we define $\delta(x) = d_{OM}(x,\lambda)$ and introduce the concept of the "cost of a most expensive common subsequence", abbreviated as *cmcs*. Formally,

$$cmcs(x,y) = \max\{\delta(u) : u \sqsubseteq (x,y)\}. \tag{21}$$

Clearly, $cmcs(x,\lambda) = \delta(x)$ and $\delta(x) = |x|$ in case of unit-indel cost. Just like $llcs$, *cmcs* is a similarity and it can be computed through

$$cmcs(x,y) = \frac{1}{2}(\delta(x)+\delta(y)-d_{OM}(x,y)) \tag{22}$$

or through a variant of an algorithm to evaluate $llcs(x,y)$. To normalize *cmcs*, we apply

$$S(x,y) = \frac{cmcs(x,y)}{\delta(x)+\delta(y)-cmcs(x,y)}. \tag{23}$$

3.3 The use of normalization in sequence analysis

In the sequence analysis literature, little attention has been paid to normalization although it has been used by several authors. We mention a few examples.
In a paper on gendered trajectories on the labor market and household status, Levy et al. (2006) applied a form of normalization to compensate for the very different lengths of their sequences and thus to avoid clustering on the basis of the lengths of the trajectories: for each pair of trajectories, they they divided the distance by the length of the longest sequence:

$$d_E^*(x,y) = \frac{d_E(x,y)}{\max\{|x|,|y|\}}. \tag{24}$$

This is not a proper normalization since it does not map to $[0,1]$ and it does not satisfy TI. Furthermore, normalization should not be used to mask or compensate for unequal censoring of the trajectories: if sequences differ in length due to differences in censoring, there is a problem caused by missing data and that should be handled by appropriate imputation methods. Unfortunately, the frequently occurring problem of missing data has hardly received attention (but see Halpin, 2012).
Elzinga and Liefbroer (2007) and Bras et al. (2010) proposed to use

$$s_K^*(x,y) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}\cdot\mathbf{y}'\mathbf{y}}} \tag{25}$$

to compare average similarities of family formation trajectories from different countries or different epochs. Indeed, this index is bounded by $[0,1]$ but it is not a proper similarity since it fails to satisfy the Covering Inequality (see Elzinga, 2014a).
Gabadinho et al. (2011) proposed to use

$$d^*(x,y) = \frac{d(x,y)}{d(x,r)+d(y,r)} \tag{26}$$

with $r = \lambda$. This maps to $[0,1]$ but it fails to satisfy TI unless d is a Euclidean distance (see Yianilos, 2002).
Finally, in a comparative study of distance metrics for sequence analysis, Ritschard et al. (2014) proposed to use Equation 15.

How can we use proper normalization in sequence analysis, other than for creating a scale that can be interpreted both as a similarity and as a distance? We see three possibilities.
First, normalization can be used when analyzing sequences while ignoring durations, i.e. analyzing the sequence of distinct subsequent subsequences (DSS). As Studer and Ritschard (2015) demonstrate, such analysis can be very useful when one wants to focus on order of states or events. However, such sequences will be of very different lengths. For example, a sequence on the labor market that just consists of uninterrupted unemployment then has only one element while a sequence that consists of an alteration of employment and unemployment may consist of many states and thus constitutes a (much) longer sequence. Normalization may be used to weigh the differences by the length of such DSS's by choosing the empty sequence as the reference object.

Second, normalization can be used to compare scales. For example, when we want to compare average distances or similarities between object from different regions or different epochs as was done, for example, in the study of Bras et al. (2010).
Finally, normalization can be used to focus on the deviations from a particular template sequence, e.g. uninterrupted employment or a career that uses the full length of legally facilitated parental leave (see e.g. Zhelyazkova, 2015). Using such a template will enlarge the various kinds of ways in which the objects differ from the reference. An alternative is to pick a medoid sequence[8] as the reference object, therewith simultaneously enlarging the various ways in which the sequences differ from this medoid.
In the next subsection, we will demonstrate some of the effects mentioned. However, the reader should be aware that such effects, with different data, may not show up when the topological structure of the distance matrix is equal to or very close to the topological structure of the similarity matrix.

4 Balancing Distance and Similarity through Normalizing

Table 1 Average OM-distances, similarities (Equation 8) and normalized distances between the household trajectories of three different cohorts. The columns labeled with "se" show the corresponding bootstrapped standard errors of these quantities.

Cohortal Birthyear	Distance	se	Similarity	se	Normalized Distance	se
<1955	2.24	.052	4.90	.053	.43	.008
1955-1964	2.93	.045	4.95	.056	.51	.006
1965-1974	3.44	.047	5.21	.062	.54	.005

In this section we demonstrate how normalizing can be used to compare data from different cohorts. Thereto, we use data on family formation in The Netherlands as collected in the 2008-wave of the Family Formation Research Program (CBS, 2008) by Statistics Netherlands. The data pertain to the retrospective household histories of 5287 Dutch, born between 1945 and 1989. We encoded these histories using 8 different states: single and childless living in the parental home, living in the parental home in all other circumstances, living single, single with child, cohabiting with child(ren), cohabiting without child(ren), married with child(ren), married without child(ren). Ignoring durations, we used the OM-metric with unit-cost to calculate the distance between trajectories.
We distinguished three birth-cohorts as shown in the leftmost column of Table 1. If indeed de-standardization of the life course (see e.g. Brückner and Mayer, 2005) took place, one would expect that, on the average, distances between the household histories of younger Dutch are greater than the distances between the trajectories of the older ones. Table 1 shows the averages and bootstrapped

[8] A medoid is an observed object of which the sum of distances to all other sequences is minimal; thus a medoid may not be unique - the centroid may not have been observed or even may not be an observable object. On the other hand, given a full distance matrix, the distances to the centroid are always computable (see Elzinga et al., 2011, section 5.3).

standard errors[9] of these estimates. We also calculated similarities (according to Equation 8) between the household trajectories, the averages of which are also shown in the middle column of Table 1. According to the same hypothesis of de-standardization, one would expect that average similarity would increase with cohortal age.

Surprisingly, Table 1 shows that both the average distance and the average similarity decrease with cohortal age. One would expect an *in*crease of similarity but a *de*crease of distance. How can we explain these results? How can we explain their apparent contradiction? Stated in terms of common and non-common features, the results show that both, the number of common and the number of non-common features increase with younger cohorts, implying that the *total* number of feature also increased. In other words, we observe an increase in the number of common features (revealed by the increase of similarity) in a general context of an increasing number of features. This result is interesting. The de-standardization could have been the result of the emergence of new phenomena like unmarried cohabitation with or without children and/or because divorce has become much more common than it was in the past (Shanahan, 2000). These phenomena apparently have resulted in longer, more complex sequences, with, on the average, more features. Indeed, the average complexity index (see e.g. Gabadinho et al., 2010) of the sequences significantly increases (figures not shown) for younger cohorts.

By normalizing the distances according to Equation 15 and using the empty sequence as the reference, we take the total number of features and therefore the complexity of the sequences into account. Average normalized distance measures the evolution of the number of non-common features, while taking into account that the maximum (potential) number of features, which is not necessarily the same in every context. The results are shown in the rightmost column of Table 1. Now it appears that the average normalized distances, and thus the normalized similarities, again are significantly different in the direction predicted by the assumption of de-standardization. This implies that the increase of the number of common features was less important than the increase of the number of non-common features.

5 Conclusions

In the previous sections, we have explored the relation between the concepts of distance and similarity and shown that adopting the axiomatic definition of similarity as presented here, leads to a spatial interpretation of similarity as "direction", complementary to distance.

We also discussed normalisation as a means to create distances that may directly be interpreted as similarities in the sense that similarity is linearly anti-monotone with distance: the more remote, the less similar and vice versa.

[9] The standard errors have been computed as follows: We took 500 samples (with replacement) of 1000 individuals each and for each of these samples, we computed the average distance. The standard deviation of these estimates is the standard error reported here. By sampling individuals, the estimated se's are bigger than when sampling distances from the original distance matrix since therein, the distances are correlated; sampling individuals instead of distances avoids this problem. The time required for this procedure is limited to the sampling time as the required distances, given the sampled individuals, can be taken from the original distance matrix. The computation time for this procedure is negligible. For an introduction to bootstrapping techniques, the reader is referred to Davison and Hinkley (1997).

Now there are a number of questions that can be put up. First of all the question of why we need this axiomatic concept of similarity at all. And if we really need this concept of similarity, does it really matter, wether or not we apply it, will it lead to new interpretations and new insights about our data? Below, we will discuss these matters one by one.

In research into choice behavior and decision making, similarity of (sets of) stimuli has long been dealt with through spatial representation of the objects and analyzing distances as if these distances were actually gauging similarity. This long tradition originates from the seminal work of Torgerson (1965), Guttman (1968) and Lingoes (1968) on multidimensional scaling. It rests on the idea that, given the multidimensional representation is precise enough, the number of significant dimensions is limited and the dimensions can be assigned a rigorous substantive interpretation in terms of meaningful properties of the stimulus- or item-set, similarity and (lack of) distance coincide. Some even considered the Minkowski-metrics, in particular for the exponent values 1, 2 and ∞, as formalisations of psychological composition rules for the generation of preferences, similarity judgments and item-correlations, to mention only a few of the variety of object sets and applications studied in this tradition.
However, in the context of sequence analysis, the dimensions of the spatial representations created through applying an edit-based or kernel-based algorithm are hard to isolate. Edit-distances are not defined through an equation that relates to a notion of dimension. Kernel-based algorithms evaluate inner-products in high-dimensional vector-spaces where each and every feature, e.g. a particular subsequence, defines its own dimension. In general, it does not make much sense to try to substantively and separately interpret these dimensions. Therefore, directly interpreting distance as similarity is hard to justify in the context of sequence analysis applications.
The alternative is a feature-based approach and it would have been appealing to adopt an axiomatisation like Tversky's. However, Tversky's axioms allow for a asymmetric similarity, therewith cutting the possibility of a structural bond between distance and similarity. Therefore we adopted an approach embodied in the axioms S1-S4: it retains Tversky's matching condition, stating that similarity is a function of both common and non-common feature sets and it has a direct connection to spatial representations. However, we do not claim that the axioms S1-S4 constitute a unique solution to the problem. Perhaps a relaxation of the system, for example by replacing S2 with the weaker $s(x,y) \leq \max\{s(x,x), s(y,y)\}$, would be interesting in some applications. Here, we will not pursue such an alternative.
Several researchers have published interesting results despite the fact that they used distance measures or normalizations that (potentially) violate the triangular inequality. Understandably, the question has been put up why we need such formal restrictions at all when, without them, we seem to have interesting results too. Some (e.g. Lesnard, 2006; Gauthier, 2015) even doubt that a formal, axiomatic approach is necessary at all. We believe that an axiomatic approach helps us to better understand the methods that we apply, either to explore data structures for patterns or to test hypothesis about these patterns. Axioms like the triangular inequality (D4) or the covering inequality (S4) allow us to map relations onto real or rational numbers and assure that the spatial structure is "smooth" in the sense that new data points are restricted by the ones that we already observed (Elzinga and Studer, 2015). Applying methods that violate such axioms not necessarily are invalid but we cannot trust them to be generalizable beyond the data collected and the representation constructed from them.

Appendix 1

Here, we are concerned with the length of the longest common subsequence as a similarity in the sense of the axioms S1-S4. A set of sequences may have many distinct lcs's, see e.g. Elzinga (2014b), but their length is a unique integer. Let us write $\mathcal{L}(x,y,\ldots)$ to denote any lcs of the sequences $\{x,y,\ldots\}$ and let $\ell(x,y,\ldots)$ denote the length of such an lcs. Here we prove that ℓ satisfies the covering inequality S4.
To prove this, we write S4, using ℓ as

$$\ell(x,y)+\ell(y,z)\leq\ell(x,z)+\ell(y,y) \tag{27}$$

and derive its correctness. First, we consider the right hand side of Inequality 27. Clearly, we have that $\ell(y,y)=|y|$. Next, we write $\ell_y(x,z)$ for the length of the longest common subsequence $\mathcal{L}_y(x,z)$ that only consists of elements that are common to y. Similarly, we write $\mathcal{L}_{\bar{y}}(x,z)$ for a longest common subsequence that does not contain any character that is common to y and we let $\ell_{\bar{y}}(x,z)$ denote its length. Since and $\mathcal{L}_y(x,z)$ and an $\mathcal{L}_{\bar{y}}(x,z)$ have no common subsequences, we must have that $\ell(x,z)=\ell_y(x,z)+\ell_{\bar{y}}(x,z)=\ell(x,y,z)+\ell_{\bar{y}}(x,z)$. Hence the right side of Inequality 27 is equivalent to

$$\ell(x,y,z)+\ell_{\bar{y}}(x,z)+|y|. \tag{28}$$

Next, we consider the left had side of Inequality 27: we have that $\ell(x,y)=\ell_z(x,y)+\ell_{\bar{z}}(x,y)$ and $\ell(y,z)=\ell_x(y,z)+\ell_{\bar{x}}(y,z)$ and we know that $\ell_z(x,y)+\ell_x(z,y)\leq\ell(x,y,z)$ and that $\ell_{\bar{z}}(x,y)+\ell_{\bar{x}}(z,y)\leq\ell(y,y)$. Therefore, the right hand side of Inequality 27 cannot exceed its right hand side and this proves the claim.

Appendix 2

To see that D_r as defined in Equation 15 is a distance, we have to investigate whether or not it satisfies the axioms D1-D4. Here we confine ourselves to showing that D_r satisfies the triangular inequality D4. We first present the proof and then comment:

$$\begin{aligned}
D_r(x,y) &= \frac{d(x,y)}{(d(x,y)+d(r,x)+d(r,y))/2} \\
&\leq \frac{d(x,z)+d(z,y)}{(d(x,z)+d(z,y)+d(r,x)+d(r,y))/2} &(29)\\
&\leq \frac{d(x,z)}{(d(x,z)+d(r,x)+d(r,y))/2}+\frac{d(z,y)}{(d(z,y)+d(r,x)+d(r,y))/2} &(30)\\
&= D_r(x,z)+D_r(z,y)
\end{aligned}$$

In inequality 29, we replaced $d(x,y)=a$ by one of its upper bounds: $d(x,z)+d(z,y)=b$. Writing $c=d(r,x)+d(r,y)$, inequality 29 states $\frac{a}{a+c}\leq\frac{b}{b+c}$ and we know this to be true for all positive a, b

and c with $a \leq b$.
In the next step, we split the sum of ratio's and then remove a positive term from each of the denominators: this must yield bigger ratio's and hence inequality 30 is correct.
The reader notes that we used the premiss that d is a distance, i.e. that d satisfies the triangular inequality in generating inequality 29.

Appendix 3

To prove that S as defined in Equation 18 satisfies S'4, we use an equivalent of S4: $s(x,y) \geq s(x,z)+s(z,y)-s(z,z)$.
Using Equation 18 in the left hand side of S'4, we write

$$\begin{aligned}
&\frac{s(x,y)}{s(x,x)+s(y,y)-s(x,y)}+1 \\
&\qquad \geq \frac{s(x,z)+s(z,y)-s(z,z)}{s(x,x)+s(y,y)-(s(x,z)+s(z,y)-s(z,z))}+1 && (31)\\
&\qquad = \frac{s(x,x)+s(y,y)}{s(x,x)+s(y,y)-(s(x,z)+s(z,y)-s(z,z))} && (32)\\
&\qquad = \frac{s(x,z)+s(y,y)-s(x,z)+s(z,y)+s(x,x)-s(z,y)}{s(x,x)+s(y,y)-s(x,z)-s(z,y)+s(z,z)} && (33)\\
&\qquad = \frac{s(x,z)+s(y,y)-s(z,y)}{s(x,x)+s(y,y)-s(x,z)-s(z,y)+s(z,z)} \\
&\qquad\quad + \frac{s(z,y)+s(x,x)-s(x,z)}{s(x,x)+s(y,y)-s(x,z)-s(z,y)+s(z,z)} && (34)\\
&\qquad \geq \frac{s(x,z)}{s(x,x)+s(z,z)-s(x,z)}+\frac{s(z,y)}{s(z,z)+s(y,y)-s(z,y)} && (35)\\
&\qquad = S(x,z)+S(z,y),
\end{aligned}$$

as required.
We concisely comment on the above steps as follows: In going from Equation 31 to 31, we used S4 to replace $s(x,y)$ by the smaller quantity $s(x,z)+s(z,y)-s(z,z)$. In Equation 33 we added $s(x,z)-s(x,z)$ and $s(z,y)-s(z,y)$ to the numerator of Equation 32 and in Equation 34, we split the ratio of Equation 33 into two appropriately chosen ratio's. Then, in passing from Equation 34 to 35, we used the general principle that, for nonnegative numbers p, q and a with $q > p$, we have that $\frac{p+a}{q+a} \geq \frac{p}{q}$.

References

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, XVI(3):471–494.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings of the Seventh International Symposium on String Processing and Information Retrieval (SPIRE'00)*, pages 39–48. IEEE Computer Society.

Bergroth, L., Hakonen, H., and Väisänen, J. (2003). New refinement techniques for longest common subsequence algorithms. In Nascimento, M. A., Silva de Moura, E., and Oliveira, A. L., editors, *String Processing and Information Retrieval, 10th International Symposium,SPIRE 2003*, pages 287–303. Springer, New York.

Blanchard, P., Bühlmann, F., and Gauthier, J.-A. (2014). *Advances in Sequence Analysis: Theory, Method, Applications*. Life Course Research and Social Policies. Springer, New York.

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer, 2 edition.

Borgwardt (2011). Kernel methods in bioinformatics. In Horng-Shing Lu, H., Schölkopf, B., and Zhao, H., editors, *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, chapter 15, pages 317–334. Springer, New York.

Bras, H., Liefbroer, A. C., and Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography*, 47(4):1013–1034.

Brückner, H. and Mayer, K.-U. (2005). De-standardization of the life course: What might it mean? and if it means anything, whether it actually took place. In *The Structure of the Life Course: Standardized? Individualized? Differentiated?(Advances in Life Course Research, Volume 9)*, pages 27–53. Elsevier, Amsterdam.

CBS (2008). *Onderzoek Gezinsvorming 2008*. Data Archiving and Network Services (DANS, https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:39585), The Hague.

Chen, S., Ma, B., and Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24-25):2365–2376.

Cornwell, B. (2015). *Social Sequence Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Decock, L. and Douven, I. (2011). Similarity after goodman. *Review of Philosophy and Psychology*, 2(1):61–75.

Deza, M. M. and Deza, E. (2014). *Encyclopedia of Distances*. Springer, Berlin, 3rd edition.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, Chichester (UK).

Elzinga, C. H. (2010). Complexity in categorical time series. *Sociological Methods & Research*, 38(3):463–481.

Elzinga, C. H. (2014a). Distance, similarity and sequence comparison. In Blanchard, P., Bühlmann, F., and Gauthier, J.-A., editors, *Advances in Sequence Analysis: Theory, Method, Applications*, number 2 in Life Course Research and Social Policies, chapter 4, pages 51–73. Springer, New York.

Elzinga, C. H. (2014b). Sequence A152072. The On-Line Encyclopedia of Integer Sequences (2014), published electronically at http://oeis.org.

Elzinga, C. H. and Liefbroer, A. C. (2007). De-standardization and differentiation of family life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population*, 23(3-4):225–250.

Elzinga, C. H. and Studer, M. (2015). Spell sequences, state proximities and distance metrics. *Sociological Methods & Research*, 44(1):3–47.

Elzinga, C. H. and Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science*, 495:50–65.

Elzinga, C. H., Wang, H., Lin, Z., and Kumar, Y. (2011). Concordance and consensus. *Information Sciences*, 181:2529–2549.

Gabadinho, A., Ritschard, G., Matthias, S., and Müller, N. M. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des Nouvelles Technologies de l'Information*, E-19:61–66.

Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

Gärdenfors (2000). *Conceptual Spaces: The geometry of thought.* MIT Press.

Gauthier, J.-A. (2015). Comment: How to make a long story short. *Sociological Methodology*, 45(1):1–3.

Golub, G. H. and Van Loan, C. H. (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, 4th edition.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33:341–353.

Halpin, B. (2012). Multiple imputation for life-course sequence data. Technical report, University of Limerick, Dept. of Sociology, http://hdl.handle.net/10344/3639.

Hennig, C. M., Meila, M., Murtagh, F., and Rocci (Eds.), R. (2015). Handbook of cluster analysis. Chapman & Hall, London.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Academic Press, San Diego.

Lesnard, L. (2006). Optimal matching and social sciences. Technical Report 01, Institut National de la Statistique et des Etudes Economiques.

Levy, R., Gauthier, J.-A., and Widmer, E. (2006). Entre contraintes institutionelle et domestique: les parcours de vie masculins et féminins en Suisse. *Canadian Journal of Sociology*, 31(4):71–92.

Lingoes, J. C. (1968). The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3(1):61–94.

Martin, P. and Wiggins, R. D. (2009). Optimal matching analysis. In Wiliams, M. and Vogt, W. P., editors, *Sage Handbook of Methodological Innovations*. Sage.

McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society. Series A*, 165(2):317–334.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Pfanzagl, J. (1968). *Theory of measurement*. Wiley, Oxford, UK.

Rencher, A. C. and Cristensen, W. F. (2012). *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

Ritschard, G., Bürgin, R., and Studer, M. (2014). Exploratory mining of life event histories. In McArdle, J. and Ritschard, G., editors, *Contemporary Issues in Exploratory Data Mining in the*

Social Sciences, Quantitative Methodology Series, chapter 9, pages 221–254. Routledge, New York.

Shanahan, M. J. (2000). Pathways to adulthood: Variability and mechanisms in life course perspective. *Annual Review of Sociology*, 26:667–692.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Recognition*. Cambridge University Press, Cambridge.

Studer, M. (2012). *Analyse de données séquentielles et application à l'étude des inégalités sociales en début de carrière académique*. PhD thesis, Faculté des sciences économiques et sociales de l'Université de Genève http://archive-ouverte.unige.ch/vital/access/manager/Repository/unige:22054.

Studer, M. and Ritschard, G. (2015). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A*. published ahead of print.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

Wagner, R. A. and Fisher, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.

Yianilos, P. N. (2002). Normalized forms for two common metrics. Technical Report 91-082-3-9027-1 Rev 7/7/2002, NEC Research Institute, Princeton, NJ.

Zhelyazkova, N. (2015). *Work-Family Reconciliation and Use of Parental Leave in Luxembourg. Empirical Analysis of Administrative Records*. PhD thesis, Maastricht University.

Acknowledgements

Matthias Studer benefitted from the support of a postdoctoral fellowship granted by the Swiss National Science Foundation at the VU University of Amsterdam. This publication is part of the research Matthias Studer conducted at the Swiss National Center of Competence in Research "LIVES - Overcoming vulnerability: life course perspectives," which is financed by the Swiss National Science Foundation.

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n.324178(Project: Context of Opportunity).