A Fuzzy Generalisation of the Hamming Distance for Temporal Sequences

Clément Moreau, Thomas Devogele, Cyril de Runz, Veronika Peralta, Evelyne Moreau, Laurent Etienne

A Fuzzy Generalisation of the Hamming Distance for Temporal Sequences

Clement Moreau, Thomas Devogele,
Cyril de Runz, Veronika Peralta, Evelyne Moreau
BDTLN, LIFAT, Université de Tours
Blois, France
firstname.lastname@univ-tours.fr

Laurent Etienne
KLaIM, LabISEN, YNCREA OUEST
Brest, France
laurent.etienne@yncrea.fr

***Abstract*—The study of temporal sequences is a main topic in different domains, especially for human mobility mining. This article defines the Fuzzy Temporal Hamming (FTH) distance between temporal sequences. This new measure generalises the Hamming distance and improves it by introducing a fuzzy time-window. This fuzzy approach tolerates temporal distortions as shifting and permutations. Moreover, the time computation of FTH is competitive with other Optimal Matching methods used for temporal sequences comparison. To validate this approach, we cluster data from a real Time-Use Survey and we compare the results obtained with other methods.**

***Index Terms*—Context awareness, Data mining, Fuzzy approach, Hamming distance, Temporal sequence, Time-Use Survey**

I. Introduction

A *temporal sequence* is a particular type of sequence where each symbol is associated to a time duration. This chronological representation of elements is used in a wide variety of domains including biology, sociology, economics, geography or signal processing. In particular, in human mobility mining temporal sequences are largely used to represent human daily mobility using time-stamped activities [12]. Therefore, these data are particularly used to identify and understand human mobility behavior, for example, with the aim to design smart cities, improve urban planning or target advertising. Experts define complex queries as how to cluster temporal sequences with close-order set of activities or how to extract patterns of human behavior. The comparison of temporal sequences plays a key role in the resolution of these complex problems.

In this paper we deal with the comparison of temporal sequences for supporting the analysis of human mobility. Many measures have been proposed to compare such sequences like Optimal Matching methods [1], Edit Distance [22] or Hamming distance [13]. Indeed, the latter is largely used for comparing equal-length sequences due to its simplicity and computation speed. Just for a reminder, the Hamming distance is commonly defined such that as the number of mismatches between the two sequences. For example $\mathcal{H}(\text{fuzzy}, \text{foggy})=3$.

In mobility mining, Hamming distance has been used to identify daily mobility behaviors [17], [23] in particular, those concerning the same activities at the same time. However other mobility properties should be taken into account: Firstly, humans have a strong tendency to return to the same locations or activities [28], [29], their movements are characterized by a certain form of redundancy [25] and particular activities start in a given fuzzy time-window [15]. Moreover, the timing of many human activities are characterized by bursts of rapidly occurring events separated by long periods of a same activity [2] (e.g., working or staying at home). Taking these elements into account, the Hamming distance seems too time sensitive to be effective for this usage. In particular, [10] claims that a good similarity measure on temporal sequences must be robust to time distortions and be able to catch the context, i.e. detect similar activities in a given imprecise time-period (e.g., afternoon, morning).

Therefore, fuzzy representations can be considered to tackle this issue, as they proved robust in several applications. While the comparison of fuzzy intervals is a difficult issue [7], [8], [26] fuzziness can also be a solution for comparing imperfect time series [3].

In this paper we propose a fuzzy generalization of the Hamming distance considering continuous time representation and contextual information. Notice that we propose a fuzzy comparison of temporal sequences and not fuzzy temporal sequences nor fuzzy time series. Our idea is to consider the cost of edit operations (transformation of symbols) instead of just counting mismatches. Such cost is context dependent and should be considered in a fuzzy temporal neighbourhood. By this consideration, our problem does not fall in the usual fuzzy temporal period comparison issue.

Several works have proposed a fuzzy approach of the Hamming distance, in various ways. For example, [9], [11], [14] propose generalisations which take into account real-valued vectors and not only binary strings. The closest work to our contribution is [5] which proposed an extension of Hamming concept to give partial credit for near misses in binary strings. However, these works are only designed for discrete series of values and not for continuous time-intervals. Therefore, to the best of our knowledge, we are the first to propose a distance which both, takes into account continuous aspects of temporal sequences and quantifies a fuzzy notion of time and semantic neighborhood of symbols in sequences.

This work is supported by Mobi'kids project which has received funding from the Agence Nationale de la Recherche and the Smart Loire project which has received funding from the Centre-Val de Loire region. .

Contributions: In this paper we propose Fuzzy Temporal Hamming (FTH), a generalisation of the Hamming distance for the comparison of temporal sequences. We use the softness of fuzzy logic to catch context dependencies of temporal sequences of activities. In particular, FTH respects the following requirements:

- It can deal with continuous temporal sequences.
- Symbols are compared using a similarity measure for catching the context.
- It is robust to temporal distortions:
  - Small time shifts produces small costs.
  - The cost of permutations of two symbols is smaller than the cost of transformation of two symbols.
  - The cost of repetition and time expansion of symbols is small according to the context (i.e., if they are temporally close and similar).

The remainder of the paper is organized as follow: Section II introduces the preliminaries on temporal sequences and provides a concise review of related work on measures for temporal sequences. The fuzzification of the Hamming distance is described in Section III. Running examples complete the section to illustrate the usefullness and the properties of our proposition. In Section IV, we validate FHT on both, a fictive and a real dataset. Section V concludes with perspectives.

II. Background

This section introduces some preliminary knowledge on sequences and temporal sequences. In a first part, we extend classical definitions of sequences to a continuous time interval framework and illustrate the problem of temporal sequences. The second part presents the classical methods used to compare sequences and temporal sequences, their properties and their limits.

A. Preliminaries

Let Σ be a finite set of symbols (e.g. daily activities), $sim : \Sigma \times \Sigma \rightarrow [0,1]$ a similarity function over Σ and $I = [0, T_{\max}), T_{\max} > 0$ the time interval for expressing temporal sequences.

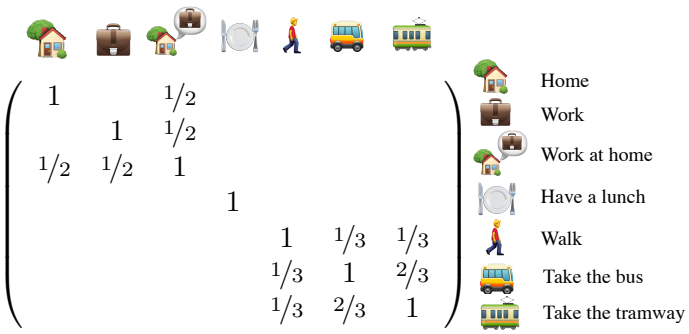

Fig. 1. Similarity between symbols of Σ, empty cells denote a similarity equal to 0.

In order to illustrate our proposal, Fig. 1 lists 7 activities (Σ) and their pair-wise similarity (sim). We consider that $T_{\max}$ represents a day (1440 minutes).

Definition 1 (Temporal sequence)**.** *A temporal sequence S_i is an ordered sequence of symbols with durations such that:*

$$S_i = \langle (x_{i1}, \delta_{i1}), ..., (x_{in}, \delta_{in}) \rangle$$

where $\forall k \in [\![1, n]\!], x_{ik} \in \Sigma$ and $\delta_{ik} > 0$ such that δ_{ik} indicates the duration of activity x_{ik} (expressed in units of times, e.g. minutes). S_i respects the two following properties:

- *There is no successive identical symbols i.e., $\forall k \in [\![1, n-1]\!], x_{ik} \neq x_{i(k+1)}$*
- *The sum of all durations is equal to $T_{\max}$ i.e., $\sum_{k=1}^{n} \delta_{ik} = T_{\max}$*

Intuitively, such a sequence indicates that an individual performed activity x_1 for δ_1 units of time, then x_2 for δ_2 units of time, ..., and finally x_n for δ_n units of time.

We note $\mathbb{S}$ the set of all temporal sequences and $\mathbb{S}^n$ the set of all temporal sequences with n symbols.

Figure 2 represents the abstraction of temporal sequences and key concepts in Definition 1.

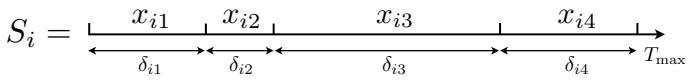

Fig. 2. Abstraction of a temporal sequence

Example 1. *We can represent Alice's daily activities thanks to the following temporal sequence:*

$$S_{alice} = \langle (🏠, 210), (🚋, 20), (🚶, 10), (💼, 250), (🚶, 15), (🍽, 60), (🚌, 15), (🏠, 290), (🏠, 570) \rangle$$

According to the emojis pictured in Fig. 1, the sequence S_{alice} formalizes the following daily mobility: Alice stayed at home (🏠) for 210min, then took the tramway (🚋) for 20min and walked (🚶) for 10min. She worked at her office (💼) for 250min. She walked for 15min and had lunch at restaurant (🍽) for 40min. She returned by bus (🚌) for 15min and worked at home (🏠) for 290min. Finally, she stayed at home the rest of the day.

Definition 2 (Time interval)**.** *Consider a temporal sequence S_i. The time interval $I(x_{ik})$ for symbol x_{ik} in S_i, is defined such that:*

$$I(x_{ik}) = [\text{begin}(x_{ik}), \text{end}(x_{ik}))$$

where:

- $\text{begin}(x_{ik}) = \sum_{j=1}^{k-1} \delta_{ij}$
- $\text{end}(x_{ik}) = \sum_{j=1}^{k} \delta_{ij} = \text{begin}(x_{ik}) + \delta_{ik}$

Therefore, for $k_1 \neq k_2$, we have: $I(x_{ik_1}) \cap I(x_{ik_2}) = \varnothing$ and $\bigcup_{k=1}^{n} I(x_{ik}) = I$

TABLE I
TIME COMPLEXITY OF MEASURES

Method	Time complexity	
	On sequences	On temporal sequences
Hamming	$O(n)$	$O\left(T_{\max}\right)$
LCSS, DTW, ED	$O(n \times p)$	$O\left(T_{\max}^2\right)$
CED	$O(n \times p \times \max(n,p))$	$O\left(T_{\max}^3\right)$

B. Related work

Many similarity measures have been proposed or adapted for comparing sequences of symbols. Most of them are part of Optimal Matching algorithms family [1] which were designed originally for DNA sequence alignment [27].

In Computer Science, the main used measures for sequence comparison are Longest Common Subsequence (LCSS) [19], Dynamic Time Warping (DTW) [4] and Edit Distance (ED) [18] (e.g., Levenshtein distance). They are computed efficiently using the dynamic programming approach [30] which guarantees a time complexity in $O(n \times p)$ where n and p are the lengths of the two compared sequences. Contrary to the Hamming distance, these measures are robust to time distortions but they do not support permutations of close symbols or cyclicity in sequences. Recently, an extension of the Edit Distance, the Contextual Edit Distance (CED) [21], proposes to take into account such requirements by modifying the cost function of the edit operations.

Nevertheless, all of these approaches are designed for discrete and non-continuous sequences which does not offer sufficient flexibility in a time framework. Therefore, to apply these algorithms, temporal sequences must be discretized by repeating a symbol as many times as its duration. For example, a temporal sequence $S_i = \langle (a,3),(b,1),(c,2) \rangle$ will be discretized such as $\langle aaabcc \rangle$.

Thus, for temporal sequences defined on a large time interval (e.g. a day) and with a small unit of time (e.g. minutes), the computation time becomes significant. Table I shows the time complexity of the previous described methods on classical sequences and temporal sequences.

III. A FUZZY APPROACH OF HAMMING DISTANCE FOR TEMPORAL SEQUENCE

This part of the section presents the fuzzification of the Hamming distance to take into account a temporal neighbourhood during the process of sequence comparison. In particular, we introduce here the concepts of edit operation, which defines the transformation of a part of a sequence, and the fuzzy context function, which quantifies a fuzzy temporal context neighborhood around the edit operation. Finally, a cost is assigned to an edit operation. The fuzzy Hamming distance on temporal sequences is defined as the maximum sum of costs to transform one sequence into another.

A. Edit operation formalisation

We define here the concept of an edit operation:

Definition 3 (EDIT OPERATION)**.** *An edit operation e is a 4-tuple defined such that:*

$$e = (x, \delta, t_{edit}, S_i) \in \Sigma \times I \times I \times \mathbb{S}$$

This means that we replace symbols in S_i, by symbol x, at time t_{edit}, for δ units of time.

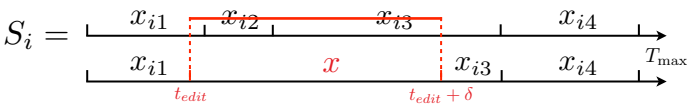

Fig. 3. Edit operation on a temporal sequence. We replace all symbols from t_{edit} to $t_{edit} + \delta$ in S_i by x.

We note $\mathbb{E}$ the set of all edit operations. Fig. 3 represents the abstraction of the edit operation given in Definition 3.

Definition 4 (FUZZY CONTEXT FUNCTION)**.** *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$, the fuzzy context function $\mu_e : I \to [0,1]$ is a fuzzy function defined for an edit operation e and such that the core $Core(\mu_e) = \{t | t \in I, \mu_e(t) = 1\} = [t_{edit}, t_{edit} + \delta)$.*

Intuitively, the fuzzy context function is used to quantify the temporal hold of an edit operation over the sequence S_i. Therefore, the hold is equal to 1 between t_{edit} and $t_{edit} + \delta$, then decreases on both sides of the interval.

Definition 5 (SIMILARITY OVER S_i)**.** *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$, the similarity function over S_i, $sim_e : I \to [0,1]$ is defined such that:*

$$sim_e(t) = \sum_{k=1}^{n} \mathbf{1}_{I(x_{ik})}(t) \times sim(x_{ik}, x) \qquad (1)$$

where $\mathbf{1}_A$ is the indicator function on I for a subset $A \subseteq I$ defined such that $\mathbf{1}_A(t) = \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{else} \end{cases}$.

The key concept in Equation 1 is that, at time t, we compute the similarity between x and the symbol x_{ik} that occurs at such time, i.e. for $t \in I(x_{ik})$. Therefore, it results a step function which is a linear combination of indicator functions weighted by the similarity of the two activities over the time interval I.

Example 2. *Consider the edit operation $e = (💼, 480, 270, S_1)$ where S_1 is the sequence described in example 1. The Fig. 4 shows the functions μ_e and sim_e of the edit operation e.*

The red function shows a fuzzy contextual function μ_e with a boundary of two hours (240min).

The blue function shows the similarity function over S_1 with the edited activity 💼. Remark that $sim_e(t)$ is equal to 1 for $t \in [240, 490)$ when Alice works at her office (💼), to $\frac{1}{2}$ for $t \in [580, 870)$ when she works at home (🏠💼) and 0 elsewhere.

B. Cost operation formalisation

Thanks to the previous definitions and in order to design a dissimilarity index between semantic temporal sequences, we assign a cost of application of an edit operation.

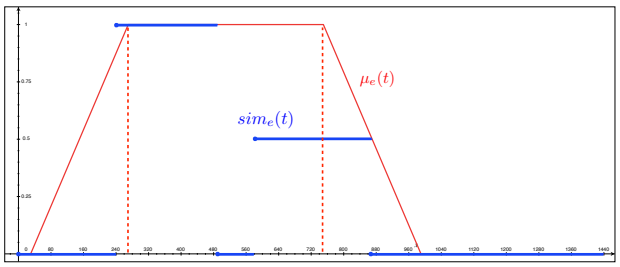

Fig. 4. Example of μ_e and sim_e functions over S_{alice}

Definition 6 (γ NORMALIZED COST)**.** *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$, the function $\gamma : \mathbb{E} \rightarrow [0, 1]$ is the normalized cost function of the application of e. It is defined such that:*

$$\gamma(e) = 1 - \sup_{\tau \in I} \left\{ \frac{1}{\delta} \int_{\tau}^{\tau+\delta} sim_e(t) \times \mu_e(t) dt \right\} \quad (2)$$

The Equation 2 is strongly inspired from the edit cost function defined in [21]. Indeed, the context is considered as similar if the $sim_e(t) \times \mu_e(t) \approx 1$ i.e., the edited activity x is temporally close to a similar one in S_i. Therefore, the key idea is, given an edit operation, to search the temporal segment $[\tau, \tau+\delta)$ over I which maximizes both similarity of the edited symbol x and the fuzzy contextual function.

At the computational level, the Equation 2 is equivalent to calculating the supremum of the convolution product between $sim_e(t) \times \mu_e(t)$ and the function $\mathbf{1}_{[0,\delta)}$ which can be computed efficiently in $O(T_{\max} \log T_{\max})$ using Fast Fourier Transform algorithms [16].

Example 3. *Consider the edit operation presented in Example 2. Fig. 5 shows the application of the γ function for $e = (\text{💼}, 480, 270, S_{alice})$. The supremum in equation 2 is achieved for $\tau = 240$ then, we have $\int_{240}^{720} sim_e(t) \times \mu_e(t) dt = 318.21$. Finally, $\gamma(e) = 1 - \frac{318.21}{480} = 0.34$*

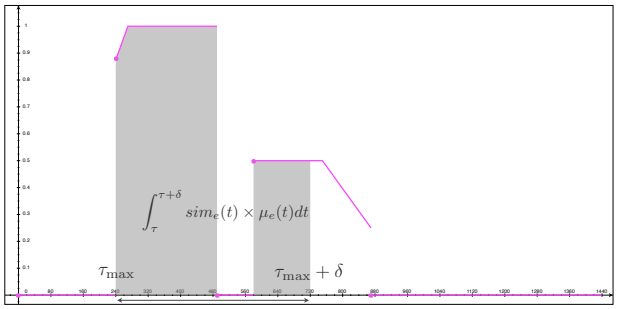

Fig. 5. Application of $\gamma(e)$

Lemma 1. *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$, we have $\gamma(e) = 0 \Leftrightarrow \exists k \in [\![1, n]\!], [t_{edit}, t_{edit} + \delta) \subseteq I(x_{ik})$ and such that $x_k = x$.*

Proof. We have:

$$\gamma(e) = 0 \Leftrightarrow \exists \tau \in I, \int_{\tau}^{\tau+\delta} sim_e(t) \times \mu_e(t) dt = \delta$$
$$\Leftrightarrow \exists \tau \in I, \forall t \in (\tau, \tau + \delta), \mu_e(t) = 1, sim_e(t) = 1$$

Yet, according to Definition 4, we have: $\mu_e(t) = 1 \Leftrightarrow t \in [t_{edit}, t_{edit} + \delta)$. Then, we know $\tau = t_{edit}$. Similarly: $sim_e(t) = 1 \Leftrightarrow \exists k \in [\![1, n]\!], t \in I(x_{ik})$ and $x_{ik} = x$ which concludes the proof. □

Lemma 2. *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$. We have: $\lim_{\delta \rightarrow 0} \gamma(e) = 1 - \sup_{t \in I} \{sim_e(t) \times \mu_e(t)\}$*

The underlined property in Lemma 2 shows that even tiny-time symbols can have an important cost in the edition process. Indeed, depending some business needs, we claim that short-time symbols are equally important than long-time symbols and the cost function must be normalized in $[0, 1]$ like the γ function. This is particularly true when the duration of symbols is highly unbalanced, for example, in the context of mobility where few activities concentrate a large part of the time during a day (e.g., home and work).

However, most of dissimilarities discussed in Section II-B give a high weight for long-duration activities and, conversely, short activities are found to have a negligible weight compared to the latter.

Aware that this usage depends on the applications, we also propose a function that weights the cost according to the duration of the edited symbol.

Definition 7 (Δ TIME-WEIGHTED COST)**.** *Given an edit operation $e = (x, \delta, t_{edit}, S_i)$, the function $\Delta : \mathbb{E} \rightarrow \mathbb{R}^+$ is the time-weighted cost function of the application of e. It is defined such that:*

$$\Delta(e) = \delta \times \gamma(e) \quad (3)$$

Thanks to this weighting, the Δ cost function respects the following properties:

Lemma 3. *For all edit operation $e \in \mathbb{E}, 0 \leqslant \Delta(e) \leqslant \delta$*

Theorem 1. *Δ is a monotone function for the time duration δ i.e., given two edit operations $e = (x, \delta, t_{edit}, S_i)$ and $e' = (x, \delta', t_{edit}, S_i)$ such that $\delta \leqslant \delta'$, then $\Delta(e) \leqslant \Delta(e')$.*

Proof. See Appendix. □

The Theorem 1 reflects the intuitive fact that, under the same editing conditions, the longer an activity is edited, the higher the editing cost.

C. Fuzzy Temporal Hamming between temporal sequences

Finally, thanks to edit operation cost functions, we can design a dissimilarity index between two temporal sequences.

Definition 8 (ONE-SIDED FUZZY TEMPORAL HAMMING)**.** *Given two temporal sequences $S_1 = \langle (x_{11}, \delta_{11}), ..., (x_{1n}, \delta_{1n}) \rangle$ and $S_2 = \langle (x_{21}, \delta_{21}), ..., (x_{2p}, \delta_{2p}) \rangle$, the one-sided Fuzzy Temporal*

TABLE II
COMPARISON OF MAIN PROPERTIES OF MEASURES ON TEMPORAL SEQUENCES

Method	Properties							
	Metric	Semi-metric	Temp. disto.	Permut.	Fuzzy context	Sim.	Continuous	Time Complexity
Hamming	×					×†		$O\left(T_{\max}\right)$
LCS	×							$O\left(T_{\max}^2\right)$
DTW			×			×†		$O\left(T_{\max}^2\right)$
ED	×‡		×	×‡		×†		$O\left(T_{\max}^2\right)$
CED		×	×	×	×	×		$O\left(T_{\max}^3\right)$
FTH		×	×	×	×	×	×	$O(\max\{n,p\}T_{\max}\log T_{\max})$

†By default discrete metric $\rho(x,y)=\begin{cases}0 & x=y\\ 1 & \text{else}\end{cases}$

‡ Only transpositions in [6] variant. The triangle inequality does not hold in this case.

Hamming from S_1 to S_2, $\mathrm{FTH}_{S_1\to S_2}:\mathbb{S}^n\times\mathbb{S}^p\to\mathbb{R}^+$ is defined such that:

$$\mathrm{FTH}_{S_1\to S_2}=\sum_{i=1}^{n}\Delta(e_i) \tag{4}$$

where $e_i=(x_{1i},\delta_{1i},\mathrm{begin}(x_{1i}),S_2)$.

Equation 4 is the total cost in order to transform the temporal sequence S_1 into S_2. It must be noted that the symmetry does not hold for $\mathrm{FTH}_{S_1\to S_2}$.

Example 4. *Let us represent Bob's daily activities as follows:*

$$S_{bob}=\langle(🏠,230),(🚶,10),(🚃,30),(💼,480),(🚶,60),(🏠,630)\rangle$$

We compute $\mathrm{FTH}_{S_{alice}\to S_{bob}}=252.31$. As the reverse, $\mathrm{FTH}_{S_{bob}\to S_{alice}}=280.22$.

Theorem 2. *$\mathrm{FTH}_{S_1\to S_2}$ is bounded by $T_{\max}$.*

Proof. By the Lemma 3, we know that for all edit operator $e\in\mathbb{E},\Delta(e)\leqslant\delta$. Consequently, and thanks to Definition 8, we know that $\mathrm{FTH}_{S_1\to S_2}\leqslant\sum_{i=1}^{n}\delta_{1i}$

Yet, thanks to Definition 1, $\sum_{i=1}^{n}\delta_{1i}=T_{\max}$ which concludes the proof. □

Note that the Theorem 2 can be interesting in order to normalize the dissimilarity in $[0,1]$.

Lemma 4. *$\mathrm{FTH}_{S_1\to S_2}$ respects the identity of indiscernibles:* $\forall S_1,S_2\in\mathbb{S},\mathrm{FTH}_{S_1\to S_2}=0\Leftrightarrow S_1=S_2$

Proof. See Appendix. □

Finally, to recover the symmetry, we applied a $\top$-conorm between the two one-sided fuzzy temporal distance.

Definition 9 (FUZZY TEMPORAL HAMMING)**.** *Given two temporal sequences $S_1\in\mathbb{S}^n$ and $S_2\in\mathbb{S}^p$, the Fuzzy Temporal Hamming measure $\mathrm{FTH}:\mathbb{S}^n\times\mathbb{S}^p\to\mathbb{R}^+$ between S_1 and S_2 is defined such that:*

$$\mathrm{FTH}(S_1,S_2)=\max\left\{\mathrm{FTH}_{S_1\to S_2},\mathrm{FTH}_{S_2\to S_1}\right\} \tag{5}$$

Theorem 3. *$(\mathrm{FTH},\mathbb{S})$ is a semi-metric space.*

Proof. By construction, from Definition 9, the symmetry of FTH is already satisfied. Also, thanks to Lemma 4 which shows $\mathrm{FTH}_{S_1\to S_2}$ respects identity of indiscernibles, then the identity of indiscernibles holds immediately for FTH. □

Theorem 4. *If for all $e\in\mathbb{E}$ the support of the fuzzy context function $Supp(\mu_e)=Core(\mu_e)$, then* FTH *is equivalent to the Hamming distance.*

Proof. See Appendix. □

Theorem 5. *Given two temporal sequences $S_1\in\mathbb{S}^n$ and $S_2\in\mathbb{S}^p$, the computation time of $\mathrm{FTH}(S_1,S_2)$ is in $O\left(\max\{n,p\}T_{\max}\log T_{\max}\right)$*

Proof. We see in Definition 6 that Equations 2 and 3 can be computed in $O(T_{\max}\log T_{\max})$. Therefore, for $S_1,S_2\in\mathbb{S}^n\times\mathbb{S}^p$, $\mathrm{FTH}_{S_1\to S_2}$ has a time complexity in $O(nT_{\max}\log T_{\max})$. Thus, Equation 5 is computed in $O(T_{\max}\log T_{\max}(n+p))=O(\max\{n,p\}T_{\max}\log T_{\max})$ □

Therefore, when n and p are well below than $T_{\max}$ (which is our case in the next Experiments section), the computation of FTH is much less greedy than other classical Optimal matching measures.

To conclude this section, the Table III-B summaries advantages and main properties of each studied methods. We note FTH checks main properties while keeping a better computation time than states of the art measures on the condition that sequences have few symbols (i.e., small n,p).

IV. EXPERIMENTS

In this section, we show the practical usage of our proposal for human mobility mining in order to identify human behaviors based on the temporal sequences of activities. The first part presents a running example which illustrates the fulfillment of the properties showed in previous section, compared to other measures. The second part exposes a comparison between FTH and Hamming measures on real temporal sequences obtained from a French household Time-Use Survey (TUS) focused on mobility, called EMD (Enquête Ménages-Déplacements). The goal of this study, as the EMD survey, is to provide a snapshot of the trips undertaken by residents of a given metropolitan area, which can aid in understanding

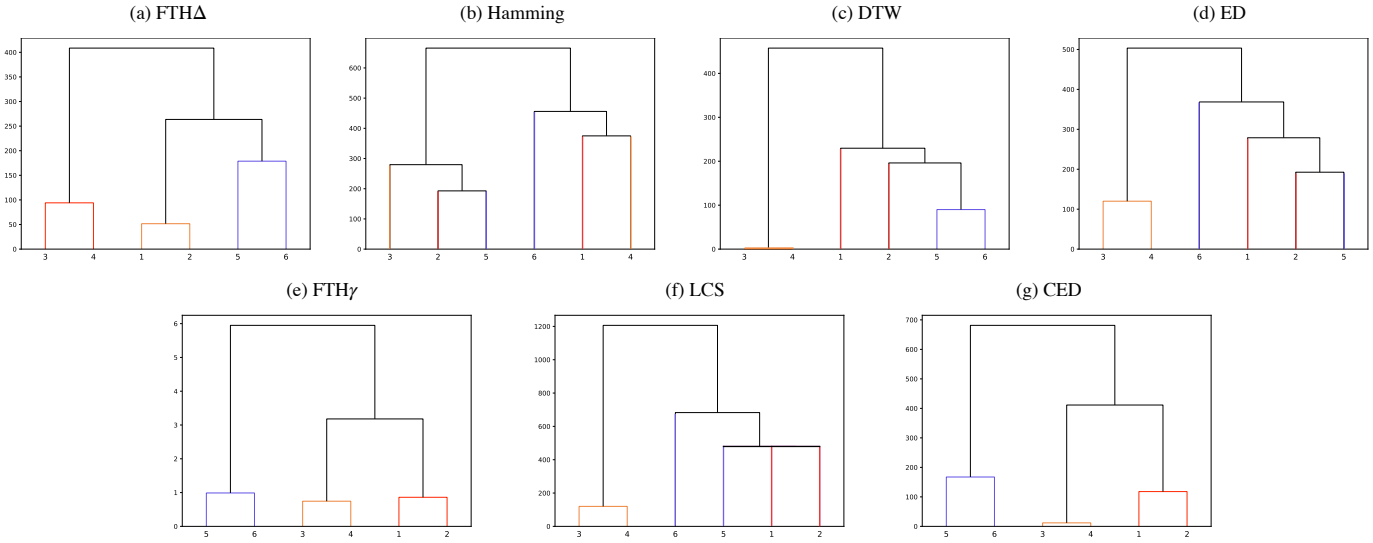

Fig. 6. Dendograms of the temporal sequences in Figure 7 for different measures. Colors indicate pairs correctly clustered.

mobility behaviors. For simplicity, we call FTHΔ and FTHγ to the instantiation of FTH with Δ and γ costs, respectively.[1]

A. Running example

Figure 7 presents a sample of 6 temporal sequences inspired in the EMD dataset. These temporal sequences are deliberately exaggerated in order to highlight the expected properties of the studied measures. Colored squares represents one hour (60 min) of activity. Sequences are made such that:

- Red couple (S_1, S_2) exhibits some permutations of activities. Indeed, both sequences contain the same activities, but shuffled: There are close permutations between bus (🚌) and work (💼), and between walk (🚶) and bus (🚌), and a more distant permutation between tramway (🚋) and work at home (🏠).
- Orange couple (S_3, S_4) shows a temporal shifting. Indeed, S_4 contains the same chaining of activities than S_3, with a shift of one hour.
- Blue couple (S_5, S_6) is similar to previous one but with a greater shifting of 90 min. Moreover, the morning work activity (💼) in S_5 is substituted by a similar one, work at home (🏠), in S_6.

Based on these temporal sequences, Figure 6 portrays the dendrograms built using Hierarchical Clustering (HC) for each measure studied in Subsection II-B. They are computed using the Linkage algorithm with the Ward agglomeration method of the Scipy library (1.4.1) in Python. Fuzzy contextual function is set with a boundary of 4 hours.

Firstly, notice that sequences with small shifts (orange couple) are well clustered with all tested methods except the Hamming distance, while sequences with larger shifts and similar activities (blue couple) are only correctly clustered

[1]Code, examples and experiments are fully available at https://colab.research.google.com/drive/1we6-mhgbQnJzdwaJX70z-ctOrTOdXJAi?usp=sharing

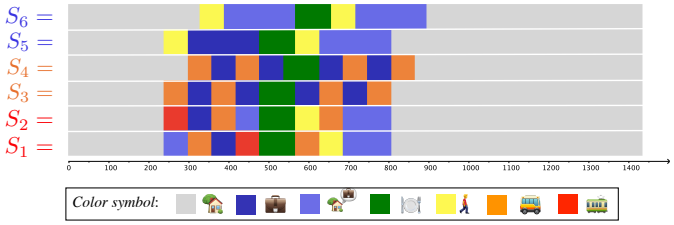

Fig. 7. Temporal sequences example

with FTHΔ, FTHγ, DTW and CED, endorsing that such methods are robust to time shifting. For permutations (red couple), sequences are well clustered with FTHΔ, FTHγ and CED. We remark that LCS groups sequences S_1 and S_2 with S_5. Therefore, we see that temporal sequences are correctly clustered with FTHΔ, FTHγ and CED.

Nevertheless, we moderate these results by recalling that the examples shown in this section are constructed to test the validity of measures on the given properties. In particular, on a real data set, it is expected that Hamming performs better. Furthermore, from a computational point of view, Hamming remains by far the best choice compared, for example, to the CED measure which is inapplicable for large $T_{\max}$.

B. Clustering mobility

1) Dataset description: To test the applicability of FTH compared to other measures, we apply them on a sample of 1200 temporal sequences extracted from the EMD dataset. Symbols used are strongly similar to the ones exhibited in Examples 1 and 4. They are organized in an semantic graph (i.e., taxonomy) and can be pairwise compared using the Wu-Palmer's similarity measure [31]. See [20] for a complete description of activities, and our code lab for more examples and details.

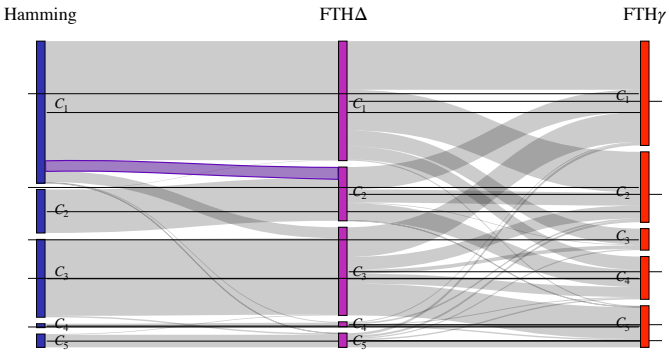

Fig. 8. Sankey diagram showing the flows between Hamming, FTHΔ and FTHγ clusterings

2) Clustering methodology: In order to compare the Hamming distance with FTHΔ and FTHγ, we applied three HC models and studied the rearrangement of clusters.

According to the Silhouette score [24] and the inertia gap, we set the number of clusters at 5 for each clustering process. The time-boundary of the fuzzy contextual function μ_e is equal to 12 hours in order to detect similar activities within morning and afternoon.

Figure 8 shows the flows between clustering results. We observe 10.5% of sequences rearrangement between Hamming and FTHΔ clusterings. In order to illustrate one of these rearrangements, Figure 9 details the flow between Hamming C_1 and FTHΔ C_2 colored in purple in Figure 8. This flow contains 46 temporal sequences detailed in (a), concerning people studying (🎓) in the morning and having leisure (📻) in the afternoon, while Hamming C_2 (b) concerns full-day students. These sequences are merged in FTH C_2 (c), evidencing that FTHΔ supports the time dilation of symbols as long as the context remains similar (in this case student typical days).

Concerning flows between FTHΔ and FTHγ, the large number of rearrangements is due to the fact that FTHγ follows a different paradigm. For FTHγ, the transformation cost can be high even if the activity duration is short. For example, the transformation of 10 min of walk can be as important as the transformation of 4 hours of work. We plan to work with experts to analyse the different clusters results and choose the best variant according to the topic.

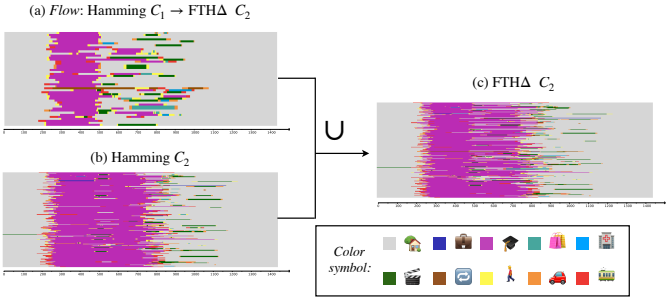

Fig. 9. Chronograms of sequences in (a) Flow from Hamming cluster C_1 to FTHΔ C_2 (b) Hamming C_2 and their merger in (c) FTHΔ C_2

V. Conclusion and Future Works

In this paper, we introduced a fuzzy extension of the Hamming distance for temporal sequences called FTH. This new measure improves the Hamming distance by introducing a fuzzy time-window in order to be robust to temporal distortions like shifts and permutations, and to catch the global context around a given period. These properties are particularly required in domains like human mobility mining in order to extract similar behaviors.

Based on an adaption of the edit operation cost function of the Hamming distance, we proved that FTH satisfy previous requirements while having a competitive time complexity regarding other states of the art measures on sequences. Moreover, FTH is generic and can be used on both continuous or discrete sequences and in various domains.

Finally, FTH was experimentally tested on two datasets. A first one of 6 tricky temporal sequences in order to check the desired properties compared to other measures. The results confirms that FTH outperformed them on the tested requirements, in particular concerning permutations of elements. The second sample on real temporal sequences was used to compare FTH with the Hamming distance in a clustering task. We showed that FTH have the ability to gather sequences whose context is close.

In future work, we plan to analyse in more detail the clusters produced by the FTH variants – with Δ and γ costs functions. Additionally, we hope to test our new measure on larger and more complex datasets.

References

[1] A. Abbott and A. Tsay. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000.

[2] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[3] Z. Ben Othmane, C. De Runz, A. A. Younes, and V. Mercelot. Quantify the variability of time series of imprecise data. In *International Conference on Flexible Query Answering Systems (FQAS)*, volume 11529, pages 203–214, Amantea, Italy, 2019.

[4] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[5] A. Bookstein, S. Tomi Klein, and T. Raita. Fuzzy hamming distance: A new dissimilarity measure. In *Combinatorial Pattern Matching*, 2001.

[6] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[7] C. De Runz, E. Desjardin, F. Piantoni, and M. Herbin. Anteriority index for managing fuzzy dates in archæological GIS. *Soft Computing*, 14(4):339–344, 2010.

[8] D. Dubois, A. HadjAli, and H. Prade. Fuzziness and uncertainty in temporal reasoning. *J. UCS*, 9(9):1168, 2003.

[9] J. Fan and W. Xie. Distance measure and induced fuzzy entropy. *Fuzzy sets and systems*, 104(2):305–314, 1999.

[10] C. Ferrero, L. Alvares, and V. Bogorny. Multiple aspect trajectory data analysis: Research challenges and opportunities. *GeoInformatica*, 17:56–67, 2016.

[11] S. D. Galbraith and L. Zobernig. Obfuscated fuzzy hamming distance and conjunctions from subset product problems. In *Theory of Cryptography*, pages 81–110, 2019.

[12] T. Hägerstraand. What about people in regional science? *Papers in regional science*, 24(1):7–24, 1970.

[13] R. W. Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[14] M. Ionescu and A. Ralescu. Fuzzy hamming distance in a content-based image retrieval system. In *Fuzz-IEEE*, volume 3, pages 1721–1726, 2004.
[15] S. Jiang, J. Ferreira, and M. C. González. Clustering daily patterns of human activities in the city. *DMKD*, 25(3):478–510, 2012.
[16] D. E. Knuth. Son of seminumerical algorithms. *ACM SIGSAM Bulletin*, 9(4):10–11, 1975.
[17] L. Lesnard. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419, 2010.
[18] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
[19] D. Maier. The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, 25(2):322–336, 1978.
[20] C. Moreau, T. Devogele, V. Peralta, L. Etienne, and C. de Runz. Methodology for mining, discovering and analyzing semantic human mobility behaviors. *arXiv preprint arXiv:2012.04767*, 2020.
[21] C. Moreau, T. Devogele, V. Peralta, and E. Laurent. Contextual edit distance for semantic trajectories. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 635–637, 2020.
[22] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.
[23] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding*, pages 14–25. Springer, 2010.
[24] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
[25] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.
[26] S. Schockaert, M. De Cock, and E. E. Kerre. Fuzzifying allen's temporal interval relations. *IEEE Transactions on Fuzzy Systems*, 16(2):517–533, 2008.
[27] T. F. Smith, M. S. Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
[28] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.
[29] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
[30] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.
[31] Z. Wu and M. Palmer. In *ACL*, pages 133–138, USA.

APPENDIX

Proof of Theorem 1

Proof. In order to simplify notations, we note $g(t) = sim_e(t) \times \mu_e(t)$. Moreover, g is Riemann-integrable and we note G such that $G' = g$.

We derive the function $\Delta(e)$ with respect to the variable δ, we have:

$$\begin{aligned}\frac{\partial \Delta}{\partial \delta}(e) &= \frac{\partial}{\partial \delta}\left(\delta - \sup_{\tau \in I}\left\{\int_{\tau}^{\tau+\delta} g(t)dt\right\}\right) \\ &= \frac{\partial}{\partial \delta}\left(\delta - \sup_{\tau \in I}\{G(\tau+\delta) - G(\tau)\}\right) \\ &= 1 - \sup_{\tau \in I}\{g(\tau+\delta) - g(\tau)\}\end{aligned}$$

By Definitions 4 and 5, we know that $\forall t \in I, 0 \leqslant g(t) \leqslant 1$, so we have: $-1 \leqslant \sup_{\tau \in I}\{g(\tau+\delta) - g(\tau)\} \leqslant 1$. Therefore, $\frac{\partial \Delta}{\partial \delta}(e) \geqslant 0$, we conclude that Δ is monotone increasing according to δ. □

Proof of Lemma 4

Proof. Let proof it by contradiction. We suppose $\exists S_1, S_2 \in \mathbb{S}^n \times \mathbb{S}^p$ such that $S_1 \neq S_2$ and $\mathrm{FTH}_{S_1 \to S_2} = 0$.

$S_1 \neq S_2$ means that there exists a time-interval $[t, t+\epsilon)$ with $\epsilon > 0$ such that S_1 and S_2 activities are different.

According to Equation 4, this result is possible if and only if $\forall i \in [\![1, n]\!], \Delta(e_i) = 0$. Because $\Delta(e_i) = \delta_{1i} \times \gamma(e_i)$ and δ_{1i} is always strictly positive, we have to show that $\forall i \in [\![1, n]\!], \gamma(e_i) = 0$.

Thanks to Lemma 1, this result is possible if and only if $\forall i \in [\![1, n]\!], \exists k \in [\![1, p]\!], I(x_{1i}) \subseteq I(x_{2k})$ and $x_{1i} = x_{2k}$.

Because $\bigcup_{i=1}^{n} I(x_{1i}) = [0, T_{\max})$ and $\forall i, j \in [\![1, n]\!], i \neq j, I(x_{1i}) \cap I(x_{1j}) = \varnothing$. Thus, we deduce that the only way to satisfy the assertion is for $I(x_{1i}) = I(x_{2k})$.

Moreover, we know that $\forall i \in [\![1, n]\!], x_{1i} = x_{2k}$. Thus, we conclude that $S_1 = S_2$ which refutes our hypothesis. □

Proof of Theorem 4

Proof. Given $(S_1, S_2) \in \mathbb{S}^n \times \mathbb{S}^p$ let compute $\mathrm{FTH}_{S_1 \to S_2}$:

$$\begin{aligned}\mathrm{FTH}_{S_1 \to S_2} &= \sum_{i=1}^{n} \Delta(e_i) \\ &= \sum_{i=1}^{n} \delta_{1i} - \sup_{\tau \in I}\left\{\int_{\tau}^{\tau+\delta_{1i}} sim_{e_i}(t) \times \mu_{e_i}(t)dt\right\}\end{aligned}$$

Yet, we know that $\forall e_i \in \mathbb{E}$ such that $e_i = (x_{1i}, \delta_{1i}, \mathrm{begin}(x_{1i}), S_2)$, we have $Supp(\mu_{e_i}) = Core(\mu_{e_i})$ so: $sim_{e_i}(t) \times \mu_{e_i}(t) = \begin{cases} sim_{e_i}(t) & \text{if } t \in I(x_{1i}) \\ 0 & \text{else} \end{cases}$. Thus, we can restrict the integral on interval $I(x_{1i})$. We have:

$$\begin{aligned}\mathrm{FTH}_{S_1 \to S_2} &= \sum_{i=1}^{n} \delta_{1i} - \int_{I(x_{1i})} sim_{e_i}(t)dt \\ &= T_{\max} - \sum_{i=1}^{n} \int_{I(x_{1i})} \sum_{j=1}^{p} \mathbf{1}_{I(x_{2j})}(t) sim(x_{1i}, x_{2j})dt\end{aligned}$$

Similarly, we can limit the indicator function $\mathbf{1}_{I(x_{2j})}$ to $I(x_{1i})$ such that:

$$\mathrm{FTH}_{S_1 \to S_2} = T_{\max} - \sum_{i=1}^{n} \int_{I(x_{1i})} \sum_{j=1}^{p} \mathbf{1}_{I(x_{1i}) \cap I(x_{2j})}(t) sim(x_{1i}, x_{2j})dt$$

Finally, and because $sim_{e_i}(t) = 0$ for $t \notin I(x_{i1})$, we can generalize the integral over I:

$$\begin{aligned}\mathrm{FTH}_{S_1 \to S_2} &= T_{\max} - \int_{I} \sum_{i=1}^{n} \sum_{j=1}^{p} \mathbf{1}_{I(x_{1i}) \cap I(x_{2j})}(t) sim(x_{1i}, x_{2j})dt \\ &= T_{\max} - \sum_{i=1}^{n} \sum_{j=1}^{p} |I(x_{1i}) \cap I(x_{2j})| sim(x_{1i}, x_{2j})\end{aligned}$$

The last expression satisfies the symmetry property and is equivalent to the Hamming distance defined for continuous temporal sequences. Therefore, it is already the same for Equation 5 which concludes the proof. □