Promoting Data Provenance Tracking in the Archaeological Interpretation Process

Sara Migliorini
Dept. of Computer Science,
University of Verona
sara.migliorini@univr.it

Alberto Belussi
Dept. of Computer Science,
University of Verona
alberto.belussi@univr.it

Elisa Quintarelli
Dept. of Computer Science,
University of Verona
elisa.quintarelli@univr.it

CEUR-WS.org/Vol-2578/PIE5.pdf

ABSTRACT

In this paper we propose a model and a set of derivation rules for tracking data provenance during the archaeological interpretation process. The interpretation process is the main task performed by an archaeologist that, starting from ground data about evidences and findings, tries to derive knowledge about an ancient object or event. In particular, in this work we concentrate on the dating process used by archaeologists to assign one or more time intervals to a finding in order to define its lifespan on the temporal axis and we propose a framework to represent such information and infer new knowledge including provenance of data. Archaeological data, and in particular their temporal dimension, are typically vague, since many different interpretations can coexist, thus we will use Fuzzy Logic to assign a degree of confidence to values and Fuzzy Temporal Constraint Networks to model relationships between dating of different findings.

KEYWORDS

Provenance, Temporal Constraints, Information discovery

1 INTRODUCTION

Interpretation and knowledge discovery represent a significant amount of the archaeological activity. Such interpretation process is usually based on direct and indirect observations of domain experts (archeologists) which also consider previous interpretations performed by themselves or other colleagues. Spatial and temporal dimensions are usually of considerable interest for archaeological research, because they allow to derive new important relationships between findings, in particular as concern to stratigraphic analysis. A typical example involving such interpretation process is represented by the dating activity. Considering the process through which objects are usually manually dated by archaeologists, some proposals in literature (e.g. [5, 7]) apply existing automatic techniques for time reasoning, in order to automatically derive new temporal knowledge or validate existing interpretations based on the available spatial and temporal information.

We can observe that archaeological interpretations depend not only from direct observations, but also from past interpretations performed by the same archaeologist or other colleagues. In general archaeological data, and more specifically the temporal dimension, are typically vague since many different interpretations can coexist; each one has its own degree of confidence and consequently several different global interpretations can be derived from them. Each interpretation is typically identified by its author; moreover, the confidence greatly depends on the archaeologist's reputation in the field. For these reasons, during the interpretation process, it is necessary not only to infer new knowledge but also to track the provenance of the information that has affected the inference. More specifically, it is necessary to keep track from which pieces of information (past interpretations) the current new knowledge has been originated, together with their authorship.

In computer science, *provenance* is the ability to record the history of data and its place of origin, and is useful to determine the chronology of the ownership, custody or location of any object and to provide a critical foundation for assessing authenticity and enabling trust. As highlighted in [9], *data provenance* is separable from other forms of provenance. In our specific archaeological scenario, the term provenance comes originally from the art world and it has been applied in archaeology and paleontology as well, where it refers to having trace of all the steps involved in producing a scientific result, such as a finding, from experiment design through acquisition of raw data, and all the subsequent steps of data selection, analysis and visualization. Such information is necessary for the reproduction of a given result, it can be useful to establish precedence (in case of patents, Nobel prizes, etc.) [11] and is different from that of provenience.

In the recent years there have been different proposals of formal models for provenance storage, maintenance, and querying; PROV is the W3C recommendation for provenance data model and language [1]. Data provenance [8] differs from other forms of meta-data because it is based on relationships among objects. Indeed, the ancestry relationships, used in provenance for correlated objects, forms a directed graph that can be represented though semistructured data models. In [12] the authors have encoded provenance graphs into Datalog and expressed inference rules and constraints with the same declarative language, in order to determine inconsistencies with respect to temporal constraints or provenance information (e.g. inconsistent cycles).

The aim of this paper is to propose a model and a set of derivation rules that are able to track the data provenance during the archaeological interpretation process. More specifically, we concentrate on the dating process used by archaeologists to assign one or more lifespans to a finding. Such process was initially modelled in [5, 7] for checking the temporal data consistency and vagueness reduction based on the use of Fuzzy Temporal Constraint Networks (FTCN) [4, 13], here we extend it in order to manage and infer new knowledge including provenance of data and complex inferences.

The remainder of the paper is organized as follows: Sect. 2 provides a formal description of the problem, while Sect. 3 describes the proposed solution; Sect. 4 exemplifies the application of this solution to a real-world case scenario.

2 PROBLEM FORMULATION

This paper refers to the Spatio-Temporal ARchaeological model ($\mathcal{S}tar$) presented in [5, 7]. In the $\mathcal{S}tar$ model three main objects

© 2020 Copyright held by the owner/author(s). Published in the Workshop Proceedings of the EDBT/ICDT 2020 Joint Conference, March 30-April 2, 2020 on CEUR-WS.org Distribution of this paper is permitted under the terms of the Creative Commons license CC BY 4.0.

of interest can be recognized: `ST_InformationSource`, `ST_ArchaeoPart` and `ST_ArchaeoUnit`. An `ST_ArchaeoUnit` is a complex archaeological entity obtained from an interpretation process performed by the responsible officer. Such an interpretation is done based on some findings (represented by `ST_ArchaeoPart` instances) retrieved during an excavation process or a bibliographical analysis (represented by `ST_InformationSource` instances). Therefore, each `ST_ArchaeoUnit` is connected to one or more constituent `ST_ArchaeoParts`, each one representing a single result of an excavation or other investigation processes.

As regards to the dating process, we can observe that the dating of an `ST_ArchaeoPart` instance (when not available from other objective measures) can also be determined from the dating of other correlated instances, or the dating of an overall `ST_ArchaeoUnit` can be obtained starting from the dating of its constituent partitions. In this paper we extend the model proposed in [5, 7] in order to keep track of the provenance of such information and to provide a measure of the contribution provided by each author of the considered past interpretations.

In the $\mathcal{S}tar$ model, temporal information regarding an archaeological finding can be quantitative or qualitative: a quantitative temporal information is represented by time instants, while a qualitative information is a temporal information defined using the well-known Allen's interval algebra [2]. Through the use of quantitative and qualitative temporal information it is possible to derive a topological structure composed of a set of related objects. Notice that inside a topological structure, some instants can be *realized*, namely they have an associated quantitative characterization (i.e., an associated time instant value), while others can be defined only qualitatively by means of relations with other nodes (i.e., represented as dummy nodes connected to other nodes).

Example 2.1. Let us consider four archaeological findings labeled as f_1, f_2, f_3 and f_4 which are coarsely dated as follows: f_1, f_2 have been located in the 19th century by archaeologist a_1, while f_3 has been dated 1850 by a_2 and f_4 has been dated 1820 by a_3. Besides these geometrical values, the following temporal relations have been detected: f_1 before f_2 and f_3 by a_4, while f_2 before f_3 and after f_4 by a_5. This knowledge can be represented by the topological complex in Fig. 1. Dates associated to nodes f_3 and f_4 are realized as the years 1850 and 1820, respectively. Conversely, dates related to nodes f_1 and f_2 are not realized, but they are located between two dummy nodes representing the years 1800 and 1899. Notice that both nodes and arcs can have an additional label representing the archaeologists that define such quantitative or qualitative temporal information. Given such topological relations some automatic reasoning techniques can be applied in order to specialize some coarse-grained dates and realize the dummy nodes. For instance, as regards to this example, the geometric temporal value associated to f_2 can be restricted from 1800-1899 to 1820-1850, and consequently the dating of f_1 can be restricted from 1800-1899 to 1800-1820. When considering the provenance propagation, we can observe that the new dating of f_2 is determined by archaeologist a_2 who generally locates it in the 19th century, but also more specifically by a_5 who defines the relations with f_3, f_4 and by a_2 and a_3 who give a precise date to f_3 and f_4. Similar considerations can be done also for the new dating of f_1. □

3 PROPOSED SOLUTION

Temporal Constraint Network (TCN) [10] is a formalism for representing temporal knowledge based on *metric* constraints among pairs of time-points. A TCN can be represented by a directed graph, where each node is associated with a variable and each arc corresponds to the constraint between the connected variables. However, in the archaeological domain, temporal knowledge is generally characterized by a level of vagueness and dates are usually expressed as periods of great confidence together with an additional interval, i.e. the safety interval. For instance, the construction date of a building can be expressed as: between 1830-1850 with more confidence plus or minus 10 years of safety.

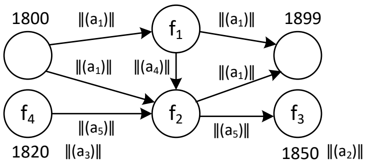


Figure 1: Example of topological complex representing temporal relations between archaeological partition.

Fuzzy set theory has been used to model the uncertainty of natural language and is able to handle the concept of partial truth (or degree of truth). In particular, given a fuzzy set F, the term *support* denotes the set of elements with a possibility greater than zero, while the term *core* denotes the set of elements with a possibility equal to 1. Therefore, a fuzzy representation of time seams to be the most appropriate solution for representing time dimensions in the archaeological context.

A fuzzy temporal constraint network (FTCN) is a generalization of TCN where a *degree of possibility* is associated with each possible value of a temporal constraint. In particular, a constraint between a pair of time-points represents a possibility distribution over temporal distances [13].

Definition 3.1 (fuzzy temporal constraint). Given two temporal variables x_i and x_j, a *fuzzy temporal constraint* C_{ij} between them is represented as a *possibility distribution function* $\pi_{ij} : \mathbb{R} \longrightarrow [0, 1]$ that constraints the possible values for the temporal distance $x_j - x_i$. □

In other words, $\pi_{ij}(d)$ is the possibility degree for the distance $x_j - x_i$ to take the value d under the constraint C_{ij}. As done in our previous work [5, 7], this paper considers only trapezoidal distributions which are sufficiently expressive in practical contexts, while computationally less expensive during the reasoning. They can be represented as a 4-tuple $\langle a, b, c, d \rangle$, where the intervals $[b, c]$ and $[a, d]$ represent the core and the support of the fuzzy set, respectively. Such tuple representation is enriched with a value α_k, called *degree of consistency*, which denotes the height of the trapeze and allows the representation of non-normalized distributions. This is necessary in the general case, because even if the initial knowledge is always represented by a trapeze with unitary height, during the reasoning the conjunction of some constraints can produce trapezes with an height less than one.

Starting from this representation, in this paper we introduce the possibility to specify for each temporal constraint also its provenance (authorship). Moreover, we introduce a modified set of operations on these constraints which allow to track and update provenance information during the interpretation process. Given such considerations, the notion of *provenance-aware fuzzy temporal constraint* (PA-FTCN) can be defined as follows.

Definition 3.2 (provenance-aware fuzzy trapezoidal constraint). Given two variables x_i and x_j, a provenance-aware fuzzy trapezoidal temporal constraint $C_{ij} = \{T_1, \ldots, T_m\}$ is a disjunction of trapezoidal distributions π_{T_k}, each one denoted by a trapeze $T_k = \langle a_k, b_k, c_k, d_k \rangle [\alpha_k] [\![\Omega]\!]$, where the characteristics 4-tuple is enriched with a degree of consistency α_k representing its height [3] and a set of provenance statements $\Omega = \{(o_1, d_1), \ldots, (o_n, d_n)\}$. Each provenance statement $\omega_i = (o_i, d_i)$ contains a label o_1 identifying the data owner and a number $d_i \in [0, 1]$ representing the degree of ownership. □

The components of a trapeze T_k take values as follows: $a_k, b_k \in \mathbb{R} \cup \{-\infty\}$, $c_k, d_k \in \mathbb{R} \cup \{+\infty\}$, $\alpha_k \in [0, 1]$, $\Omega \subseteq \mathcal{A} \times [0, 1]$ where $\mathcal{A}$ is a set of labels representing known data owners. As mentioned before, the support of π is defined as $supp(\pi_{T_k}) = \{x : \pi_{T_k}(x) > 0\} = [a_k, d_k]$, while the core as $core(\pi_{T_k}) = \{x : \pi_{T_k}(x) = \alpha_k\} = [b_k, c_k]$. Moreover, this paper considers only well-formed trapezes: a trapeze $T = \langle a, b, c, d \rangle$ is *well-formed*, if $a \leq b \leq c \leq d$. In the following the set of well-formed trapezes is denoted as $\mathcal{T}$. From this definition several shapes are allowed, as illustrated in Fig. 2.

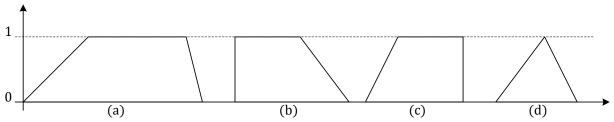


Figure 2: Possible shapes of a trapezoidal possibility distribution function: (a) $a < b < c < d$, **(b)** $a = b < c < d$, **(c)** $a < b < c = d$, **and (d)** $a < b = c < d$.

The semantics of a constraint $C_{ij} = \{T_1, \ldots, T_m\}$ is the possibility distribution function $\pi_{C_{ij}}$ corresponding to the disjunction of the trapezoidal distribution $\pi_{T_k} : \mathbb{R} \longrightarrow [0, 1]$ for $k = 1, \ldots, m$.

Definition 3.3 (trapezoid possibility distribution function). The *possibility distribution function* of a generic trapeze $T_k \in \mathcal{T}$ can be written as:

$$\pi_{T_k}(x) = \begin{cases} 0 & \text{if } x < a_k \vee x > d_k \\ \alpha_k \cdot ((x - a_k)/(b_k - a_k)) & \text{if } a_k \leq x < b_k \\ \alpha_k \cdot ((d_k - x)/(d_k - c_k)) & \text{if } c_k < x \leq d_k \\ \alpha_k & \text{otherwise} \end{cases}$$

□

Definition 3.4 (solution). Let $\mathcal{P} = \langle \mathcal{X}, C \rangle$ be a provenance-aware fuzzy temporal constraint network. An n-tuple $S = \{s_1, \ldots s_n\}$, where $s_i \in \mathbb{R}$, is a *possible solution* of $\mathcal{P}$ at degree α if and only if: $\deg(S) = \min_{i,j}\{\pi_{C_{ij}}(s_j - s_i)\} = \alpha$, where π_{ij} stands for the possibility distribution associated to the constraint C_{ij} and the degree corresponds to the least satisfied constraint [13]. □

In the case of a PA-FTCN, each solution is characterized by a *degree of satisfaction* reflecting a trade-off among potentially conflicting constraints, and a set of *provenance statements* characterizing the ownership of each constraint.

The most widely used algorithm for constraint propagation is the *path-consistency algorithm.*

Definition 3.5 (path-consistency algorithm). Given three variables x_i, x_k and x_j of a PA-FTCN $\mathcal{P}$ and a local instantiation $x_i = d_i$, $x_j = d_j$, a new constraint between x_i and x_j can be induced from pre-existing constraints by the path consistency algorithm as follows: $\pi_{ij} \otimes (\pi_{ik} \circ \pi_{kj})(x)$, where $(\pi_{ik} \circ \pi_{kj})$ is the composition (addition between fuzzy sets) of the constraints between $x_i - x_k$ and $x_k - x_j$, while π_{ij} is the existing constraints between $x_i - x_j$. □

In order to determine the result of the previous definition, it is necessary to define the required operations. More specifically, it is necessary to specialize some operations on fuzzy sets to operations on trapezoids with provenance statement. In particular, the specialization of the inversion (T_k^{-1}), composition ($T_1 \circ T_2$), conjunction ($T_1 \otimes_a T_2$) and disjunction ($T_1 \oplus_a T_2$) operations on trapezoidal distributions can be found in [6]. Here we specialize them in order to take care also of the provenance information. In particular, our aim is from one side to propagate provenance labels, but also to provide a degree of ownership to each author, thus, we need to define the concept of similarity between two trapezes.

Definition 3.6 (trapeze similarity). Given two trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle [\alpha_1] [\![\Omega_1]\!]$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle [\alpha_2] [\![\Omega_2]\!]$, the degree of similarity $sim(T_1, T_2) \in [0, 1]$ between them is defined as:

$$sim(T_1, T_2) = \frac{area(T_1 \cap T_2)}{area(T_1 \cup T_2)} \tag{1}$$

In other words the similarity is maximum (equal to 1) when the two trapezes coincide, while it is minimium (equal to 0) when the two trapezes are completely disjoint, otherwise it is proportional to the degree of overlap between them. Notice that there can be two cases where the degree of similarity is equal to 0: i) when the intersection is empty, and ii) when the union of the two trapezes generates an infinite trapeze. This second case is possibile, for instance, when one of the trapezes represents a qualitative precedence constraint. In order to distinguish these two situations, we use the symbol 0 when the intersection is empty (no similarity at all), and the symbol $\perp$ when the union is infinite (very low similarity).

During the various operations the degree of ownership assigned to each author is computed on the basis of the starting degree of ownership and the similarity between the original constraint and the new obtained one.

Definition 3.7 (inversion). Given a constraint $C_{ij} = \{T_1, \ldots, T_m\}$ between variables x_i and x_j, the constraint C_{ij}^{-1} represents the equivalent constraint holding between x_j and x_i. Such constraint can be obtained by making the inversion of each constituent trapezoids $T_k = \langle a_k, b_k, c_k, d_k \rangle [\alpha_k] [\![\Omega]\!]$ contained in C_{ij}, as follows: $T_k^{-1} = \langle -d_k, -c_k, -b_k, -a_k \rangle\ [\alpha_k] [\![\Omega]\!]$. □

Notice that in this case the provenance information is not affected by the operation.

Definition 3.8 (composition $\circ$). Given two constraints C_1 and C_2, the composition of two generic trapezoids $T_1 = \langle a_1, b_1, c_1, d_1 \rangle [\alpha_1] [\![\Omega_1]\!] \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle [\alpha_2] [\![\Omega_2]\!] \in C_2$, assuming that $\alpha_1 \geq \alpha_2$, is defined as: $T_1 \circ T_2 = \langle a_1 + a_2, b_1' + b_2, c_1 + c_2', d_1 + d_2 \rangle [\min\{\alpha_1, \alpha_2\}] [\![\Omega_1 \cup \Omega_2]\!]$, where $b_1' = a_1 + (\alpha_2/\alpha_1)(b_1 - a_1)$ and $c_2' = d_2 - (\alpha_2/\alpha_1)(d_2 - c_2)$ and

$$[\![\Omega_1 \cup \Omega_2]\!] = \{(o_i, d_i) \mid (o_i, d_i) \in \Omega_1 \vee (o_i, d_i) \in \Omega_2\} \tag{2}$$

□

The composition of two constraints produces a bigger trapezoid w.r.t. the source trapezoids, thus, the provenance information is the union of the input ones, with the same degree of ownership.

The conjunction of two generic fuzzy possibility distribution functions π_1 and π_2 is defined as: $\forall d \in \mathbb{R}\ (\pi_1 \otimes \pi_2(d) = \min\{\pi_1, \pi_2\})$. Unfortunately, this operation cannot be directly applied to trapezoids and is more complex to specialize than composition, because given two generic trapezoids T_1 and T_2, the function $T_1 \otimes T_2 = \min\{T_1, T_2\}$ is not always a trapeze: Fig. 3.a

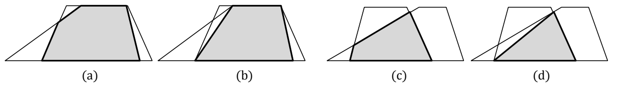
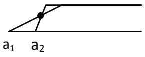
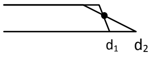
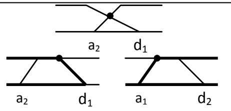
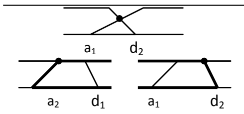


Figure 3: Two examples of approximated conjunction operation $\otimes_a$ between trapezoids: in (a) and (c) the result of the classical conjunction operation between fuzzy possibility distribution functions, and in (b) and (d) the corresponding approximation which produces a trapeze.

and Fig. 3.c contain two examples of such situation. Therefore, some sort of approximation of $T_1 \otimes T_2$ has to be defined to obtain a trapeze. For the application context considered by this paper, the following approximation criteria formulated in [3] are appropriate, where T is the result of the approximated conjunction: $core(\pi_T) = core(\pi_{T_1} \otimes \pi_{T_2})$, $h(\pi_T) = h(\pi_{T_1} \otimes \pi_{T_2})$, and $supp(\pi_T) \subseteq supp(\pi_{T_1} \otimes \pi_{T_2})$. In other words, the approximation shall ensure that the core of the obtained distribution is maintained while the possibility of the support elements outside the core can be sightly modified. This operation is formalized as follows.

Table 1: Possible intersection between two trapezes and corresponding element of the conjunction result.

Situation	Result
$a_2 \in (a_1, b_1)$  	$b' = \begin{cases} b_1 & \text{if } \alpha_1 = \alpha_2 \wedge b_1 > b_2 \\ b_1 & \text{if } \alpha_1 < \alpha_2 \\ b_2 & \text{otherwise} \end{cases}$
$d_1 \in (c_2, d_2)$  	$c' = \begin{cases} c_1 & \text{if } \alpha_1 = \alpha_2 \wedge c_1 > c_2 \\ c_1 & \text{if } \alpha_1 < \alpha_2 \\ c_2 & \text{otherwise} \end{cases}$
 	b' is the highlighted intersection point.
 	c' is the highlighted intersection point.

Definition 3.9 (conjunction $\otimes_a$). Given two constraints C_1 and C_2, the conjunction between two trapezoids $T_1 = \langle a_1, b_1, c_1, d_1 \rangle[\alpha_1]$ $[\![\Omega_1]\!] \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle[\alpha_2]$ $[\![\Omega_2]\!] \in C_2$ is defined as follows: $T_1 \otimes_a T_2 = T \in \mathcal{T}^{inf}(T_1, T_2) : \forall T_1 \in \mathcal{T}^{inf}(T_1, T_2), \pi_{T_i} \leq \pi_T$, where $\mathcal{T}^{inf}(T_1, T_2) = \{T \mid \pi_T \leq \pi_{T_1} \otimes \pi_{T_2} \;\wedge\; h(\pi_T) = h(\pi_{T_1} \otimes \pi_{T_2})\}$ [3]. The trapezoid T can be computed as follows: $T = (\max\{a_1, a_2\}, b', c', \min\{d_1, d_2\})[\min\{\alpha_1, \alpha_2\}][\![\Omega_1 \cup \Omega_2]\!]$ where b' and c' depends on the 8 possible intersections between T_1 and T_2 illustrated in Table 1 and

$$[\![\Omega_1 \cup \Omega_2]\!] = \{(o_i, d_i) \mid o_i \in \mathcal{A}_1 \cup \mathcal{A}_2 \wedge d_i = sim(T_i, T)\} \quad \Box$$

The set $\mathcal{T}^{inf}$ is the set of trapeze that approximate the conjunction from “below”, the result of the conjunction is the greatest trapeze in this set. Some examples of $T_1 \otimes_a T_2$ are illustrated in Fig. 3. In case (d) it is evident that the height of the resulting trapeze can become less than one, hence the degree of consistency α becomes necessary.

As regards to the provenance, in this case, we keep track of all authors who contribute to the trapeze conjunction, but we update the degree of ownership on the basis of the similarity between the original information and the obtained one. Notice that, when the same author o_i is present in both the two trapezoids T_1 and T_2, we will compute its degree of ownership as $d_i = max(sim(T_1, T), sim(T_2, T))$. Moreover, $max(\perp, sim(T_i, T)) = sim(T_i, T)$.

Finally, the disjunction operation is not required by the path consistency algorithm, but it can be useful for eliminating redundant trapezes that are accidentally introduced by users or are due to constraint propagation. Thus, it is an operation useful for compressing available information.

The disjunction of two general fuzzy distribution functions π_1 and π_2 is defined as $\forall d \in \mathbb{R} : \pi_1 \oplus \pi_2(d) = \max\{\pi_1(d), \pi_2(d)\}$. However, like conjunction, disjunction is not closed in the algebra of trapezoids. Therefore, the idea is to compute a tentative trapeze, and then check whether it corresponds to the disjunction of the involved constraints (i.e., correspond of one of the two involved trapezes), otherwise the constraints will be maintained separated.

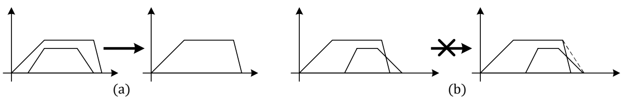


Figure 4: Two examples of approximated disjunction operation $\oplus_a$ between trapezoids: in (a) the operation can be performed, while in (b) the operation cannot be performed.

Definition 3.10 (disjunction $\oplus_a$). Given two constraints C_1 and C_2, the disjunction between two trapezes $T_1 = \langle a_1, b_1, c_1, d_1 \rangle[\alpha_1]\Omega_1 \in C_1$ and $T_2 = \langle a_2, b_2, c_2, d_2 \rangle[\alpha_2] \in C_2'$ is defined as follows [3]: $T_1 \oplus_a T_2 = \langle a, b, c, d \rangle[\max\{\alpha_1, \alpha_2\}][\![\Omega_1 \cup \Omega_2]\!]$ where $a = \min\{a_1, a_2\}$, $b = b_1$ if $\alpha_1 > \alpha_2$ or $b = b_2$ if $\alpha_2 > \alpha_1$ or $b = \min\{b_1, b_2\}$ otherwise, $c = c_1$ if $\alpha_1 > \alpha_2$ or $c = c_2$ if $\alpha_2 > \alpha_1$ or $c = \min\{c_1, c_2\}$ otherwise, $d = \max\{d_1, d_2\}$ and

$$[\![\Omega_1 \cup \Omega_2]\!] = \{(o_i, d_i) \mid o_i \in \mathcal{A}_1 \cup \mathcal{A}_2 \wedge d_i = sim(T_i, T)\} \quad \Box$$

Fig. 4.a illustrates a case where the disjunction is executed, while Fig. 4.b illustrates a case where it cannot be executed. The disjunction has the same behaviour on data provenance of the conjunction.

4 CASE STUDY

This section illustrates an example of reasoning performed on archaeological data that allows the identification of some new temporal and data provenance knowledge. It regards an archaeological object called *Porta Borsari* which is an ancient Roman gate in Verona. This object has been modeled as an `ST_ArchaeoUnit` by author a_1, who also identifies and dates three distinct phases into its life:

- Phase A – first foundation as *Porta Iovia* during the Late Republican Time, which spans from 200 B.C. to 27 B.C.;
- Phase B – reconstruction during the Claudian Time, which spans from 41 A.C. to 54 A.C.;
- Phase C – Teodorician changes during the Middle-Age, which spans from 312 A.C. to 553 A.C.

This information is represented in Fig. 5-7 by using two nodes for each phase X, a node X_s denoting the phase start and a node X_e denoting the phase end. An arrow connects X_s with the network start node s, while another arrow connects X_s with X_e. The labels on these arrows is derived from the phase duration and its relation with date associated to the start node s (in our example 200 B.C.).

Subsequently, other archeologists have identified some findings as archaeological partitions belonging to this archaeological unit. Table 2 reports some information about them together with the associated dating. As regards to the dating, we assume that the first archaeologist who found an archaeological partition simply assigns it to one of the identified phases, while later the same or other authors will restrict such dating as soon as new information becomes available. The author responsible for the identification of the phase membership is reported in column **Ph** inside round brackets together with the phase name, while the author(s) responsible for the fine-grained dating is (are) reported in column **Dating**. Notice that in order not to cluttering the notation, we have omitted to report the original unitary height of the trapeze (namely [1]). Moreover, since the table reports initial information, we assume that when more than one author is present in Tab. 2, the contribution provided by each author is equal, i.e. the reporting date is the result of a joint work.

Table 2: Dating of each partition and associated phase.

Archaeo. Partition		Ph	Dating
P208	Foundation and North Tower	A (a_1)	$\langle -110, -100, -1, +9\rangle [\![(a_2, 1)]\!]$ I B.C. ± 10 years
P263	Structures of eastern facade	A (a_1)	$\langle -60, -50, -45, -35\rangle [\![(a_3, 1)]\!]$ Middle of I B.C. ± 10 years
P214	Front of the external facade	B (a_1)	$\langle 35, 45, 50, 60\rangle [\![(a_4, 1)]\!]$ Middle of I A.C. ± 10 years
P248	External Foundations	B (a_1)	$\langle -9, 1, 100, 110\rangle [\![(a_1, 0.5), (a_4, 0.5)]\!]$ I A.C. ± 10 years
P275	Internal Foundations	B (a_1)	$\langle -10, 1, 50, 100\rangle [\![(a_2, 0.5), (a_3, 0.5)]\!]$ Middle of I A.C. ± 5 years
P250	Defensive structures	C (a_1)	$\langle 401, 450, 500, 500\rangle [\![(a_2, 1)]\!]$ 2nd middle of V A.C.

Finally, author a_5 identifies the following temporal relations between partitions: P208 terminates before P263 starts, and P248 terminates before P214 starts. These precedence relations have to be modeled with an arc $\langle 0, 0, \infty, \infty\rangle [\![(a_5, 1)]\!]$; however, for not cluttering the diagram, the figure reports only the author label.

Accordingly with the transformation rules of the previous section, the first operation to perform is the definition of a common coordinate reference system. The origin of such system is placed to 200 B.C., since it is the earliest date in the model, while the granularity is the year, since for all dates the minimum granularity is at least a year. In order to simplify the presentation, the resulting network is presented through three portions, each one corresponding to a different phase. The overall network can be obtained by combining the three sub-networks and by adding an edge from phase A to phase B and an edge from phase B to phase C, both labeled with $\langle 0, 0, +\infty, +\infty\rangle [\![(a_1, 1)]\!]$. These edges represent the precedence relations between phases.

Fig. 5 illustrates the subnetwork related to phase A: node s represents the starting point, nodes A_s and A_e represent the start and end points of the phase respectively, while nodes P263 and P208 represent the dating of the corresponding archaeological partitions. This portion of FTCN allows to compute some derived constraints for the nodes based on the declared one, using the formula in Def. 3.5: $\pi'_{ij}(x) = \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}(x))$.

In particular, a more precise relation can be derived between partition P208 and partition P263, which is initially represented simply as an edge labeled with the constraint $\langle 0, 0, +\infty, +\infty\rangle$. In particular, by assuming i = P208, $k = s$ and j = P263, the following new constraint π'_{ij} can be derived between P208 and P263:

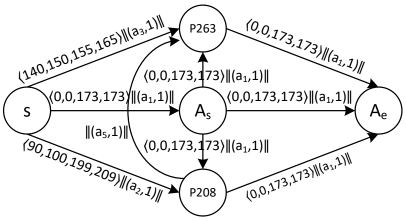


Figure 5: Portion of FTCN related to phase A.

$$
\begin{aligned}
\pi'_{ij} &= \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}) \\
&= \pi_{ij} \otimes_a (\pi_{ki}^{-1} \circ \pi_{kj}) \\
&= \langle 0, 0, \infty, \infty\rangle [\![(a_1, 1)]\!] \otimes_a \\
&\qquad (\langle -209, -199, -100, -90\rangle [\![(a_2, 1)]\!] \circ \langle 140, 150, 155, 165\rangle [\![(a_3, 1)]\!]) \\
&= \langle 0, 0, \infty, \infty\rangle [\![(a_1, 1)]\!] \otimes_a \langle -69, -49, 55, 75\rangle [\![(a_2, 1), (a_3, 1)]\!] \\
&= \langle 0, 0, 55, 75\rangle [\![(a_1, \bot), (a_2, 0.52), (a_3, 0.52)]\!]
\end{aligned}
$$

From this derivation follows that the distance between P208 and P263 can be from 0 to 75 years, with great possibility until 55. This is consistent with the observation that P208 is located in I B.C., but it shall precede P263 which is located in the middle of I B.C. As regards to the authors' ownership, we can observe that all three authors partecipate to the final result, but with different degrees of ownership. In particular, the final degree of ownership for a_2 and a_3 is 0.52, computed using Def. 3.6, while a_1 is reported with a degree of similarity equal to $\bot$, since the union operator produces a figure with an infinite area.

A similar operation can be performed on the FTCN portion in Fig. 6, where B_s and B_e represent the start and end points of phase B, respectively. The constraint between partition PA-248 and PA-214 can be restricted as follows where i = P248, $k = s$ and j = P214:

$$
\begin{aligned}
\pi'_{ij} &= \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}) \\
&= \pi_{ij} \otimes_a (\pi_{ki}^{-1} \circ \pi_{kj}) \\
&= \langle 0, 0, \infty, \infty\rangle [\![(a_1, 1)]\!] \otimes_a \\
&\qquad (\langle -209, -199, -100, -90\rangle [\![(a_1, 0.5), (a_4, 0.5)]\!] \circ \\
&\qquad \langle 140, 150, 155, 165\rangle [\![(a_4, 1)]\!]) \\
&= \langle 0, 0, \infty, \infty\rangle [\![(a_1, 1)]\!] \otimes_a \langle -69, -49, 55, 75\rangle [\![(a_1, 0.5), (a_4, 1)]\!] \\
&= \langle 0, 0, 55, 75\rangle [\![(a_1, 0.52), (a_4, 0.52)]\!]\rangle
\end{aligned}
$$

The consideration is similar to the previous one, since P214 happens in the middle of the I A.C. and P248 is generally dated I A.C. but has to finish before P214 starts living. Notice that in this case we have two ownership information for a_1, thus we choose the maximum one.

Finally, as regards to phase C, whose corresponding sub-network is reported in Fig. 7, the dating of its partition can determine a restriction of the phase start as follows, by considering $i = s$, k =

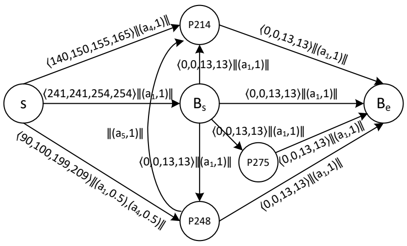


Figure 6: Portion of FTCN related to phase B.

P250 and $j = C_s$:

$$\begin{aligned}
\pi'_{ij} &= \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}) \\
&= \pi_{ij} \otimes_a (\pi_{ik} \circ \pi_{kj}^{-1}) \\
&= \langle 512, 512, 753, 753 \rangle \llbracket (a_1, 1) \rrbracket \otimes_a \\
&\quad (\langle 601, 650, 700, 700 \rangle \llbracket (a_2, 1) \rrbracket \circ \\
&\quad \langle -241, -241, 0, 0 \rangle \llbracket (a_1, 1) \rrbracket) \\
&= \langle 512, 512, 753, 753 \rangle \llbracket (a_1, 1) \rrbracket \otimes_a \\
&\quad \langle 360, 409, 700, 700 \rangle \llbracket (a_1, 0.5), (a_2, 0.5) \rrbracket \\
&= \langle 512, 512, 700, 700 \rangle \llbracket (a_1, 0.78), (a_2, 0.30) \rrbracket
\end{aligned}$$

Clearly, these are only examples of the derivations that can be obtained by executing the path-consistency algorithm on the overall network and considering all the triangles. However, these examples make clear the utility of applying existing temporal reasoning techniques on archaeological data.

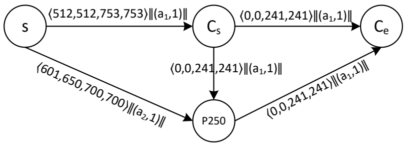


Figure 7: Portion of FTCN related to phase C.

5 CONCLUSION

In this paper we have proposed an extension of a model, able to store temporal information about archeological findings, for managing also the data provenance during the archaeological interpretation process. In particular, we have extended a set of fuzzy operators in order to represent and infer new knowledge including provenance of data and its degree of truth.

ACKNOWLEDGMENTS

This work was partially supported by the Italian National Group for Scientific Computation (GNCS-INDAM) and by "Progetto di Eccellenza" of the Computer Science Dept., Univ. of Verona, Italy.

REFERENCES

[1] 2013. World Wide Web Consortium - PROV-DM: The PROV Data Model. https://www.w3.org/TR/prov-dm/.

[2] J. F. Allen. 1983. Maintaining Knowledge About Temporal Intervals. *Communications of the ACM* 26, 11 (1983), 832–843.

[3] S. Badaloni, M. Falda, and M. Giacomin. 2004. Integrating Quantitative and Qualitative Fuzzy Temporal Constraints. *AI Communications* 17, 4 (2004), 187–200.

[4] S. Badaloni and M. Giacomin. 2006. The Algebra IA^{fuz}: A Framework for Qualitative Fuzzy Temporal Reasoning. *Artificial Intelligence* 170, 10 (2006), 872–908.

[5] A. Belussi and S. Migliorini. 2014. A Framework for Managing Temporal Dimensions in Archaeological Data. In *Proceedings of 21st International Symposium on Temporal Representation and Reasoning (TIME).* 81–90. https://doi.org/10.1109/TIME.2014.15

[6] A. Belussi and S. Migliorini. 2014. *Modeling Time in Archaeological Data: the Verona Case Study.* Technical Report RR 93/2014. Department of Computer Science, University of Verona. http://www.di.univr.it/report

[7] A. Belussi and S. Migliorini. 2017. A spatio-temporal framework for managing archeological data. *Annals of Mathematics and Artificial Intelligence* 80, 3 (Aug 2017), 175–218. https://doi.org/10.1007/s10472-017-9535-0

[8] P. Buneman, S. Khanna, and W. C. Tan. 2001. Why and Where: A Characterization of Data Provenance. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings.* 316–330. https://doi.org/10.1007/3-540-44503-X_20

[9] P. Buneman and W. C. Tan. 2018. Data Provenance: What next? *SIGMOD Record* 47, 3 (2018), 5–16. https://doi.org/10.1145/3316416.3316418

[10] R. Dechter, I. Meiri, and J. Pearl. 1991. Temporal Constraint Networks. *Artificial Intelligence* 49, 1-3 (1991), 61–95.

[11] C. C. Kolb. 2014. *Provenance Studies in Archaeology.* Springer New York, New York, NY, 6172–6181. https://doi.org/10.1007/978-1-4419-0465-2_327

[12] P. Missier and K. Belhajjame. 2012. A PROV Encoding for Provenance Analysis Using Deductive Rules. In *Provenance and Annotation of Data and Processes - 4th International Provenance and Annotation Workshop, IPAW.* 67–81. https://doi.org/10.1007/978-3-642-34222-6_6

[13] Lluis V. and Lluis G. 1994. On Fuzzy Temporal Constraint Networks. *Mathware and Soft Compunting* 3 (1994), 315–334.