ENFrame: A Platform for Processing Probabilistic Data

Sebastiaan J. van Schaik\textsuperscript{1,2}, Dan Olteanu\textsuperscript{1}, Robert Fink\textsuperscript{1}
\{Sebastiaan.van.Schaik, Dan.Olteanu, Robert.Fink\}@cs.ox.ac.uk
\textsuperscript{1}Department of Computer Science, University of Oxford, United Kingdom
\textsuperscript{2}Oxford e-Research Centre, University of Oxford, United Kingdom

ABSTRACT
This paper introduces ENFrame, a unified data processing platform for querying and mining probabilistic data. Using ENFrame, users can write programs in a fragment of Python with constructs such as bounded-range loops, list comprehension, aggregate operations on lists, and calls to external database engines. The program is then interpreted probabilistically by ENFrame.

The realisation of ENFrame required novel contributions along several directions. We propose an event language that is expressive enough to succinctly encode arbitrary correlations, trace the computation of user programs, and allow for computation of discrete probability distributions of program variables. We exemplify ENFrame on three clustering algorithms: k-means, k-medoids, and Markov clustering. We introduce sequential and distributed algorithms for computing the probability of interconnected events exactly or approximately with error guarantees.

Experiments with k-medoids clustering of sensor readings from energy networks show orders-of-magnitude improvements over naive clustering in each possible world, of approximate over exact, and of distributed over sequential algorithms.

1. INTRODUCTION
Recent years have witnessed a solid body of work in probabilistic databases with sustained systems building effort and extensive analysis of computational problems for rich classes of queries and probabilistic data models of varying expressivity \cite{29}. In contrast, most state-of-the-art probabilistic data mining approaches so far consider the restricted model of probabilistically independent input and produce hard, deterministic output \cite{1}. This technology gap hinders the development of data processing systems that integrate techniques for both probabilistic databases and data mining.

The ENFrame data processing platform aims to close this gap by allowing users to specify iterative programs to query and mine probabilistic data. The semantics of ENFrame programs is based on a unified probabilistic interpretation of the entire processing pipeline from the input data to the program result. It features an expressive set of programming constructs, such as assignments, bounded-range loops, list comprehension, aggregate operations on lists, and calls to database engines, coupled with aspects of probabilistic databases, such as possible worlds semantics, arbitrary data correlations, and exact and approximate probability computation with error guarantees. Existing probabilistic data mining algorithms do not share these latter aspects.

Under the possible worlds semantics, the input is a probability distribution over a finite set of possible worlds, whereby each world defines a standard database or a set of input data points. The result of a user program is equivalent to executing it within each world and is thus a probability distribution over possible outcomes (e.g., partitionings). ENFrame exploits the fact that many of the possible worlds are alike, and avoids iterating over the exponentially many worlds.

Correlations occur naturally in query results \cite{29}, after conditioning probabilistic databases using constraints \cite{19}, and are supported by virtually all mainstream probabilistic models. If correlations are ignored, the output can be arbitrarily off from the expected result \cite{22, 2}. For instance, consider two similar, but contradicting sensor readings (mutually exclusive data points) in a clustering setting. There is no possible world and thus no cluster containing both points, yet by ignoring their negative correlation, we would assign them to the same cluster.

The user is oblivious to the probabilistic nature of the input data, and can write programs as if the input data were deterministic. It is the task of ENFrame to interpret the program probabilistically. The approach taken here is to trace the user computation using fine-grained provenance, which we call events. The event language is a non-trivial extension of provenance semirings \cite{11} and semimodules \cite{4} that are used to trace the evaluation of positive relational algebra queries with aggregates and to compute probabilities of query results \cite{10}. It features events with negation, aggregates, loops, and definitions. The language is expressive enough to succinctly encode arbitrary correlations occurring in the input data (e.g., modelled on Bayesian networks and pc-tables) and in the result of the user program (e.g., co-occurrence of data points in the same cluster), and trace the program state at any time. By annotating each computation in the program with events, we effectively translate it into an event program: variables become random variables whose possible outcomes are conditioned on events. Selected events represent the probabilistic program output, e.g. in case of clustering: the probability that a data point is a medoid, or the probability that two data points are assigned to the same cluster. Events are also essential for sensitivity analysis and explanation of the program result.

The most expensive task supported by ENFrame is probability computation for event programs, which is \#P-hard
in general. We developed sequential and distributed algorithms for both exact and approximate probability computation with error guarantees. The algorithms operate on graph representations of the event programs called event networks. Expressions common to several events are only represented once in such graphs. Event networks for data mining tasks are very repetitive and highly interconnected due to the combinatorial nature of the algorithms: the events at each iteration are expressions over the events at the previous iteration and have the same structure at each iteration. Moreover, the event networks can be cyclic, so as to account for program loops. While it is possible to unfold bounded-range loops, this can lead to prohibitively large event networks.

The key challenge faced by ENFrame is to compute the probabilities of a large number of interconnected events that are defined over several iterations, exhibit deep nesting structures, and use novel expressive constructs. This contrasts with earlier work on probability computation for propositional events, e.g. [14, 15], which only considers one event in disjunctive normal form at a time. Clustering events are many, highly interconnected, and have deep nesting structures. Rather than separately computing the probability of each event, ENFrame’s algorithms employ a novel bulk-compilation technique, using Shannon expansion to depth-first explore the decision tree induced by the input random variables and the events in the program. The approximation algorithms use an error budget to prune large fragments of this decision tree that only account for a small probability mass. We introduce three approximation approaches (eager, lazy, and hybrid), each with a different strategy for spending the error budget. The distributed versions of these algorithms divide the exploration space into fragments for concurrent exploration.

While the computation time can grow exponentially in the number of input random variables in worst case, the structure of correlations can reduce it dramatically. As shown experimentally, ENFrame’s algorithm for exact probability computation is orders of magnitude faster than executing the user program in each possible world.

To sum up, the contributions of this paper are as follows:

- We propose the ENFrame platform for processing probabilistic data. ENFrame can evaluate user programs on probabilistic data with arbitrary correlations following the possible worlds semantics.
- User programs are written in a fragment of Python that supports bounded-range loops, list comprehension, aggregates, and calls to external database engines. We illustrate ENFrame’s features by giving programs for three clustering algorithms (k-means, k-medoids, and Markov clustering) and provide a formal specification of ENFrame’s user language which can be used to write arbitrary programs for the platform.
- User programs are annotated by ENFrame with events that are expressive enough to capture correlations in the input data, trace the program computation, and allow for probability computation.
- ENFrame uses novel sequential and distributed algorithms for exact and approximate probability computation of event programs.
- We implemented ENFrame’s probability computation algorithms in C++.
- We report on experiments with k-medoids clustering of readings from partial discharge sensors in energy networks [22]. We show orders-of-magnitude performance improvements of ENFrame’s exact algorithm over the naïve approach of clustering in each possible world, of approximate over exact clustering, and of distributed over sequential algorithms. Experiments also confirm that ENFrame’s clustering accuracy is identical to that of the naïve approach and that this is not the case for clustering in the top-k most probable worlds.

The paper is organised as follows. Section 2 introduces the Python fragment supported by ENFrame along with encodings of clustering algorithms. Section 2 introduces our event language and shows how user programs are annotated with events. Our probability computation algorithms are introduced in Section 4 and experimentally evaluated in Section 5.
1. It should naturally express common mining algorithms, count the number of True values in a one-dimensional array.

2. ENFRAME’S USER LANGUAGE

This section introduces the user language supported by ENFrame. Its design is grounded in three main desiderata:

1. It should naturally express common mining algorithms, allow to issue queries and manipulate their results.
2. User programs must be oblivious to the deterministic or probabilistic nature of the input data and to the probabilistic formalism considered.
3. It should be simple enough to allow for an intuitive and straightforward probabilistic interpretation.

We settled on a subset of Python that can express, among others, k-means, k-medoids, and Markov clustering. In line with query languages for probabilistic databases, where a Boolean query $Q$ is a map $Q : D \rightarrow \{true, false\}$ for deterministic databases and a Boolean random variable for probabilistic databases, every user program has a sound semantics for both deterministic and probabilistic input data: in the former case, the result of a clustering algorithm is a deterministic clustering, in the latter case it is a probability distribution over possible clusterings.

The user language comprises the following constructs:

**Variables and arrays.** Variables can be of scalar types (real, integer, or Boolean) or arrays. Examples of variable assignments: $V = 2$, $W = V$, $M[2] = True$, or $M[i] = W$. Arrays must be initialised, e.g., for array $M$ of cardinality $k$: $M = [None] * k$. Additionally, the expression $range(0,n)$ specifies the array $[0,\ldots,n-1]$.

**Functions.** Scalar variables can be multiplied, exponentiated (pow$(B,r)$ for $B^r$), and inverted (invert$(B)$ for $1/B$). The function dist$(A,B)$ is a distance measure on the feature space between the arrays $A,B$ of reals; scalar_mult computes the component-wise multiplication of an array with a scalar.

**Reduce.** Given a one-dimensional array $M$ of some scalar type, it can be reduced to a scalar value using one of the functions reduce_or, reduce_and, reduce_sum, reduce_mult, reduce_count. For instance, for an array $B$ of Booleans, the expression reduce_and$(B)$ computes the conjunction of the truth values in $B$, and the expression reduce_count$(B)$ computes the number of elements in $B$. For a two-dimensional array of reals or integers (an array of vectors), reduce_sum computes the component-wise sum of the vectors.

**List comprehension.** Inside a reduce-function, anonymous arrays may be defined using list comprehension. For example, given an array $B$ of Booleans of size $n$, the expression reduce_sum([1 for i in range(0,n) if B[i]]) counts the number of True values in $B$.

**Loops.** We only allow bounded-range loops; for any fixed integer $n$ and counter variable $i$, for-loops can be defined by:

\[
\forall i \in 0..n-1 : O^i \equiv \Phi(\alpha_0) \otimes \alpha_1 M_{i+1}^0 \equiv \Phi(\alpha_{(k-1)}) \otimes \alpha_{(k-1)}
\]

\[
\forall i \in 0..iter-1 : M_i \equiv \bigwedge_{l=0}^{n-1} \left[ \text{dist}(O^i, M_{i+l-1}^o) \leq \text{dist}(O^i, M_{i+l-1}^o) \right]
\]

\[
\forall i \in 0..k-1 : M_i^o \equiv \left( \sum_{l=0}^{n-1} \text{InCl}_{i+l}^1 \right) \ominus \left( \sum_{l=0}^{n-1} \text{InCl}_{i+l}^1 \right)
\]

Figure 2: K-means clustering specified as user program (left) and simplified event program (right).
1: \( (0, n, M) = \text{loadData()} \) \# M is a stochastic \( n \times n \) matrix of weights
2: \# edge weights between the \( n \) nodes, \( O \) is list of nodes
3: \( (r, \text{iter}) = \text{loadParams()} \) \# Hadamard power, number of iterations
4: for \( i \) in range(0, iter):
5:   \( N = [\text{None}] \times n \) \# expansion phase
6:   for \( l \) in range(0, n):
7:     \( M[i] = [\text{None}] \times n \)
8:   for \( j \) in range(0, n):
9:     \( M[i][j] = \text{reduce_sum}([M[i][k] \ast M[k][j] \text{ for } k \text{ in range(0,n)}]) \)
10: \( M = [\text{None}] \times n \) \# inflation phase
11: for \( i \) in range(0, n):
12:   \( M[i] = [\text{None}] \times n \)
13: for \( j \) in range(0, n):
14:   \( \text{for } j \text{ in range(0,n)}: \)
15:     \( \text{N}[i][j] = \text{reduce_sum}([\text{M}[i][k][r] \ast \text{invert}\langle \text{reduce_sum}([\text{pow}\langle \text{[N][k],[r]\rangle} \ast \text{M}[k][j] \text{ for } k \text{ in range(0,n)}]) \rangle \text{ for } k \text{ in range(0,n)}]) \)

**Figure 3:** Markov clustering specified as user program (left) and simplified event program (right).

ters. Next, the new cluster centroids \( M[i] \) are computed as the centroids of each cluster (lines 12–16). The assignment and update phases are repeated \( \text{iter} \) times (line 4).

**k-medoids clustering.** The \( k \)-medoids algorithm is almost identical to \( k \)-means, but elects \( k \) cluster medoids rather than centroids: these are cluster members that minimise the sum of distances to all other objects in the cluster. The assignment phase is the same as for \( k \)-means (lines 5–11), while the update phase is more involved: we first compute an array \( \text{DistSum} \) of sums of distances between each cluster medoid and all other objects in its cluster (lines 12–17), then find one object in each cluster that minimises this sum (lines 18–24), and finally elect these objects as the new cluster medoids \( M \) (lines 25–27). The last step uses \( \text{reduce}_\text{sum} \) to select exactly one of the objects in a cluster as the new medoid, since for each fixed \( i \) only one value in \( \text{Centre}[i][l] \) is \text{True} due to the tie-breaker in line 24.

**Markov clustering (MCL).** MCL is a fast and scalable unsupervised clustering algorithm for graphs based on simulation of stochastic flow [31]. Natural clusters in a graph are characterised by the presence of many edges within a cluster and few edges across clusters. MCL simulates random walks within a graph by alternating two operations: **expansion** and **inflation**. Expansion corresponds to computing random walks of higher length. It associates new probabilities with all pairs of nodes, where one node is the point of departure and the other is the destination. Since higher length paths are more common within clusters than between different clusters, the probabilities associated with node pairs lying in the same cluster will, in general, be relatively large as there are many ways of going from one to the other. Inflation has the effect of boosting the probabilities of intra-cluster walks and demoting inter-cluster walks. This is achieved without a priori knowledge of cluster structure; it is the result of cluster structure being present.

Figure 8 gives the MCL user program. Expansion coincides with taking the power of a stochastic matrix \( M \) using the normal matrix product (i.e. matrix squaring). Inflation corresponds to taking the Hadamard power of a matrix (taking powers entry-wise). It is followed by a scaling step to maintain the stochastic property, i.e., the matrix elements correspond to probabilities that sum up to 1 in each column.

Section 2.2 discusses the probabilistic interpretation of the computation of the above three clustering algorithms.

### 2.2 Syntax of the User Language

Figure 4 specifies the formal grammar for the language of user programs. A program is a sequence of declarations (DECL) and loop blocks (LOOP), each of which may again contain declarations and nested loops. The language allows to assign expressions (EXPR) to variable identifiers (ID).

**Figure 4:** The grammar of the user language.

An expression may be a Boolean, integer, or float constant (LIT), an identifier, an array declaration, the result of a Boolean comparison between expressions, or the result of such operations as sum, product, inversion, or exponentiation. The result of a reduce operation on an anonymous array created through list comprehension (LCOMPR), and the result of breaking ties in a Boolean array give rise to expressions; we elaborate on these two constructions below.

In addition to the syntactic structure as defined by the grammar, programs have to satisfy the following constraints:

**Bounded-range loops.** The parameters to the range function \( \text{iter} \) must be integer constants (or immutable integer-valued variables). This restriction ensures that for-loops (LOOP) and list comprehensions (LCOMPR) are of bounded size that is known at compile time.

**Anonymous arrays via list comprehension.** List comprehension may only be used to construct one-dimensional arrays of base types (i.e., integers, floats, or Booleans).

**Breaking ties.** Clustering algorithms require explicit handling of ties: For instance, if two objects are equidistant to two distinct cluster centroids in \( k \)-means, the algorithm has to decide which cluster the object will be assigned to. In ENFrame programs, the membership of objects to clusters can be encoded by a Boolean array like \( \text{InCl}[i][l] \) which is \text{True} if and only if object \( i \) is in cluster \( l \). In this context, a tie is a configuration of \( \text{InCl} \) in which for a fixed object \( i \), \( \text{InCl}[i][l] \) is \text{True} for more than one cluster \( l \). We explicitly break such ties using the function \text{breakTies}(). For each fixed value \( i \) of the second dimension (hence the 2 in the function name) of the 2-dimensional array \( M \), it iterates over the first dimension of \( M \) and sets all but the first \text{True} value of \( M[i][l] \) to \text{False}. Similarly, the
function \texttt{breakTiles}(\mathcal{M})\) fixes the first dimension and breaks ties in the second dimension of \(\mathcal{M}\), and \texttt{breakTiles}(\mathcal{M})\) breaks ties in a one-dimensional array.

### 3. TRACING COMPUTATION BY EVENTS

The central concept for representing user programs in ENFrame is that of \textit{events}. Each event is a concise syntactic encoding of a random variable and its probability distribution. This section describes the syntax and semantics of events and event programs, and finally explains how ENFrame programs written in the user language from Section 2 can be translated to event programs.

The key features of events and event programs are:

- Events can encode arbitrarily correlated, discrete probability distributions over input objects. In particular, they can succinctly encode instances of such formalisms as Bayesian networks and pc-tables. The input objects and their correlations can be explicitly provided, or imported via a positive relational algebra query with aggregates over pc-tables [10].

- By allowing non-Boolean events, our encoding is exponentially more succinct than an equivalent purely Boolean description.

- Each event has a well-defined probabilistic semantics that allows to interpret it as a random variable.

- The iterative nature of many clustering algorithms carries over to event programs, in which events can be defined by means of nested loops. This construction together with the ability to reuse existing, named events in the definition of new, more complex events leads to a concise encoding of a large number of distinct events.

The events generated by user programs are compositions of events associated with objects (data points) in the input data set: each object \(o_i\) is annotated with an event over a set \(X\) of independent Boolean random variables.

#### Example 1. Clustering in possible worlds.

We start by presenting an instructive example of \(k\)-medoids clustering under possible worlds semantics. Let \(o_0, \ldots, o_3\) be objects in the feature space as shown below. They can be clustered into two clusters using \(k\)-medoids with medoids \(o_1\) and \(o_2\).

![Possible worlds](image)

The possible valuations \(\nu : X \to \{\text{true}, \text{false}\}\) define the possible worlds of the input objects: for each valuation \(\nu\), there exists one world that contains exactly those objects \(o_i\) for which \(\Phi(o_i)\) is \textit{true} under \(\nu\). The probability of a world is the product of the probabilities of the variables \(x \in X\) taking a truth value \(\nu(x)\).

Assume that the input data specifies the following events:

\[
\Phi(o_0) = x_1 \lor x_3, \quad \Phi(o_1) = x_2, \quad \Phi(o_2) = x_3, \quad \Phi(o_3) = \neg x_2 \land x_4.
\]

Distinct worlds can have different clustering results, as exemplified next. The world defined by \(\{x_1 \mapsto \top, x_2 \mapsto \bot, x_3 \mapsto \top, x_4 \mapsto \top\}\) consists of objects \(o_0, o_2,\) and \(o_3\), for which \(k\)-medoids clustering yields:

![Clustering result](image)

Similarly, the worlds defined by \(\{x_1 \mapsto \bot, x_2 \mapsto \top, x_3 \mapsto \top\}\) and any assignment for \(x_4\), yields:

![Clustering result](image)

The probability of a query “\(\text{Are } o_1\) and \(o_2\) in the same cluster?\)” is the sum of the worlds in which \(o_1\) and \(o_2\) are in the same cluster.

Events do not only encode the correlations and probabilities of input objects, but can symbolically encode the entire clustering process. We illustrate this in the next example.

#### Example 2. Symbolic encoding of \(k\)-means by events.

We again assume four input objects \(o_0, \ldots, o_3\) with their respective events \(\Phi(o_i)\). This example introduces \textit{conditional values} (c-values) which are expressions of the form \(\Phi \odot v\), where \(\Phi\) is a Boolean formula and \(v\) is a vector from the feature space. Intuitively, this c-value takes the value \(v\) whenever \(\Phi\) evaluates to true, and a special \textit{undefined} value when \(\Phi\) is false. C-values can be added and multiplied; for example, the expression \(\Phi \odot v + \Psi \odot w\) evaluates to \(v + w\) if \(\Phi\) and \(\Psi\) are \textit{true}, or to \(v\) if \(\Phi\) is \textit{true} and \(\Psi\) is \textit{false}, etc.

Equipped with c-values, an initialisation of \(k\)-means with \(k = 2\) can for instance be written in terms of two expressions:

\[
M^0 = \Phi(o_0) \odot o_0 + \Phi(o_1) \odot o_2 \quad \text{and} \quad M^1 = \top \odot 0.5 \cdot (o_1 + o_3).
\]

The iterative nature of many clustering algorithms carries over to event programs, in which events can be defined by means of nested loops. This construction together with the ability to reuse existing, named events in the definition of new, more complex events leads to a concise encoding of a large number of distinct events.

In the assignment phase (lines 5–11 of Figure 2), each object is assigned to its nearest centroid. The condition \(\text{InCl}^{i,t}(o_i\) being closest to \(M^t)\) can be written as the Boolean event

\[
\text{InCl}^{i,t} \equiv \bigwedge_{l=0}^{2} \left[\text{dist}(\Phi(o_i) \odot o_l, M^t) \leq \text{dist}(\Phi(o_i) \odot o_l, M^t)\right],
\]

which encodes that the distance from \(o_i\) to centroid \(M^0\) is smaller than the distance to centroid \(M^1\).

Given the Boolean events \(\text{InCl}^{i,t}\), we can represent the centroid of cluster \(i\) for the next iteration by the expression

\[
\left(\sum_{o=0}^{3} \text{InCl}^{i,t} \odot o_i\right)^{-1} \cdot \left(\sum_{o=0}^{3} \text{InCl}^{i,t} \odot o_i\right),
\]

which specifies a random variable over possible cluster centroids conditioned on the assignments of objects to clusters as encoded by \(\text{InCl}^{i,t}\) (lines 12–16). This expression is exponentially more succinct than an equivalent purely Boolean encoding of centroids, since the latter would require one Boolean expression for each subset of the input objects.

The event programs corresponding to the three user programs for \(k\)-means, \(k\)-medoids, and MCL are given on the right side of Figures 1–3. In addition to the constructs introduced in Example 2 they use event declarations that assign identifiers to event expressions, and \(\forall\)-loops that specify sets of events parametrised by \(i\). The remainder of this section specifies the formal syntax and semantics of event programs, and gives a translation from user to event programs.

#### 3.1 Syntax of Event Expressions

The grammar for event expressions is as follows:

\[
\begin{align*}
\text{VAL} &::= \text{A scalar or feature vector} \\
\text{INT} &::= \text{Any integer} \\
\text{CVAL} &::= \text{EVENT} \odot \text{VAL} \mid \text{CVAL}^{-1} \mid \text{CVAL}^\land \text{CVAL}^\lor \mid \text{CVAL}^{-1} \\cdot \text{CVAL}^{-1} \mid \text{CVAL}^{-1} \cdot \text{CVAL}^{-1} \mid \text{CVAL}^{-1} \cdot \text{CVAL}^{-1} \\
\text{COMP} &::= \leq \mid \geq \mid = \mid \neq \mid > \\
\text{ATOM} &::= \text{CVAL} \mid \text{COMP} \mid \text{CVAL} \\
\text{EID} &::= \text{Elements of a set of event identifiers} \\
\text{EVENT} &::= \text{Propositional formula over } X, \text{ EID, ATOM}
\end{align*}
\]

Conditional values. Reals and feature vectors are denoted by VAL. Together with a propositional formula, they give rise to a \textit{conditional value} (CVAL), c-value for short.
3.2 Semantics of Event Expressions

The semantics of event expressions is defined by extending a Boolean valuation \( \nu : X \rightarrow \{ \text{true}, \text{false} \} \) to a valuation of c-values and event expressions. We define in the sequel how \( \nu \) acts on each of the expression types in the grammar.

The base cases of this mapping are the standard algebraic operations on scalars and the feature space, extended by operations on scalars and the feature space, extended by the feature vector, \( \mathbf{u} \).

\[
\nu(CVAL) = \begin{cases} 
X, & \text{if } \nu(\text{EVENT}) = \text{true} \\
\mathbf{u}, & \text{if } \nu(\text{EVENT}) = \text{false} \\
\mathbf{u} \cdot \mathbf{x} = \mathbf{u}, & \text{if } \nu(\text{EVENT}) \text{ is undefined} \end{cases}
\]

\[
\nu(\text{EVENT} \odot \text{VAL}) = \begin{cases} 
\text{VAL}, & \text{if } \nu(\text{EVENT}) = \text{true} \\
\mathbf{u}, & \text{if } \nu(\text{EVENT}) = \text{false} \\
\mathbf{u} \cdot \mathbf{x} = \mathbf{u} \cdot \mathbf{x} = \mathbf{x}, & \text{if } \nu(\text{EVENT}) \text{ is undefined} \end{cases}
\]

\[
\nu(\text{dist}(CVAL_1, CVAL_2)) = \begin{cases} 
\mathbf{u}, & \text{if } \nu(CVAL_1) = \nu(CVAL_2) \\
\mathbf{u} \cdot \mathbf{x} = \mathbf{u} \cdot \mathbf{x} = \mathbf{x}, & \text{if } \nu(CVAL_1) \neq \nu(CVAL_2) \text{ or } \nu(CVAL_2) \neq \nu(CVAL_1) \\
\mathbf{u} \cdot \mathbf{x} = \mathbf{u} \cdot \mathbf{x} = \mathbf{x}, & \text{if } \nu(CVAL_1) \neq \nu(CVAL_2) \text{ or } \nu(CVAL_2) \neq \nu(CVAL_1) \text{ and the comparison does not hold} \end{cases}
\]

\[
\nu(\text{ATOM}) = \begin{cases} 
\nu(CVAL), & \text{if } \nu(\text{EVENT}) = \text{true} \\
\mathbf{u}, & \text{if } \nu(\text{EVENT}) = \text{false} \end{cases}
\]

ATOM, EVENT.

3.3 Probabilistic Semantics of Events

We next give a probabilistic interpretation of event expressions that explains how they can be understood as random variables: Boolean event expressions (EVENT) give rise to Boolean random variables, and conditional values (CVAL) give rise to random variables over their respective domain.

For every random variable \( x \in X \), we denote by \( P_x[\text{true}] \) and \( P_x[\text{false}] \) the probability that \( x = \text{true} \) or \( x = \text{false} \), respectively; we also simply write \( P_x \) for \( P_x[\text{true}] \).

\[
\Omega = \{ \nu : X \rightarrow \{ \text{true}, \text{false} \} \}
\]

The set of mappings from the random variables \( X \) to \( \{ \text{true}, \text{false} \} \).

Definition 1 (Induced Probability Space).

The probability mass function \( \Pr(\nu) = \prod_{x \in \mathbf{x}} P_x[\nu(x)] \) for every sample \( \nu \in \Omega \), and the probability measure \( \Pr(E) = \sum_{\nu \in \Omega} \Pr(\nu) \) for \( E \subseteq \Omega \) define a probability space \((\Omega, \mathcal{P}^2(\Omega), \Pr)\) that we call the probability space induced by \( X \).

An event expression \( E \) is a random variable over the probability space induced by \( X \) with probability distribution

\[
P_E[s] = \Pr(\{ \nu \in \Omega \mid \nu(E) = s \})
\]

By virtue of this definition, every Boolean event expression becomes a Boolean random variable, and real-valued (vector-valued) c-values become random variables over the feature space.

3.4 Event Programs

Event programs consist of a sequence of event declarations (DECL) and nested loops (LOOP) of event declarations.

A central concept is that of event identifiers (EID); it is required that event declarations be immutable, i.e., each distinct EID may only be assigned once to an event expression. Inside a Vi-loop, identifiers can be parametrised by \( i \) to create a distinct identifier in each iteration of the loop.

The meaning of an event program is simply the set of all named and grounded c-values and event expressions defined by the program; here, grounded means that all identifiers in expressions are recursively resolved and replaced by the referenced expressions. For declarations outside of loops, this is clear; each declaration inside of (nested loops) is instantiated for each value of the loop counter variables.

3.5 From User Programs to Event Programs

The translation of user to event programs has two main challenges: (1) Translating mutable variables and arrays to immutable events, and (2) translating function calls such as reduce_. We cover these two issues separately.

From mutable variables to immutable events. It is natural to reassign variables in user language programs, for example when updating k-means centroids in each iteration based on the cluster configuration of the previous iteration. In contrast, events in event programs are immutable, i.e., can be assigned only once. The translation from the user language to the event language utilises a function getLabel that generates for each user language variable \( M \) a sequence of unique event identifiers whose lexicographic order reflects the sequence of assignments of \( M \).

The idea of getLabel is to first identify the nested loop blocks of the user language program, and then to establish a counter for each distinct variable symbol \( M \) and each block.
An assignment of a variable within \( k \) nested blocks corresponds to an event identifier of the form \( M_{i_1} \ldots i_k \) where \( c_1, \ldots, c_k \) are the \( k \) counters for the \( k \) blocks. Within each block, its corresponding counter is incremented for every assignment of its variable symbol. When going from one block into a nested inner block, the counters for the outer blocks are kept constant while the counter for the inner block is incremented as \( M \) is reassigned in the inner block.

Special attention must be paid to the encoding of entering and leaving a block: in order to carry over the reference to a variable \( M_{i_1} \ldots i_k \) to the next block at level \( k+1 \), we establish a copy \( M_{i_1} \ldots i_k(-1) \equiv M_{i_1} \ldots i_k \), such that the first access to \( M \) in the block may access its last assignment of \( M \) via \( M_{i_1} \ldots i_k(-1) \). Similarly, the last assignment of a variable in the inner block is passed back to the outer block by copying the last identifier of an inner block to the next identifier of the outer block.

**Example 3.** Consider the following user language program (left) and its translation to an event program (right).

```
1: \( M \equiv 7 \)
2: \( \forall i \in \text{range}(0,2) \) do \( C \)
3: for \( i \) in \( \text{range}(0,2) \) do \( C \)
4: \( \forall i \in \text{range}(0,1) \) do \( B \)
5: if \( j \) in \( \text{range}(0,3) \) then \( B \)
6: \( k \equiv \{ M \} \)
7: \( k \equiv \{ M \} \)+1
```

The program has three nested blocks. Within each block, the respective counter is incremented for each assignment of \( M \): \( M \equiv M_0 + 2 \). The encoding for entering and leaving a block are in lines \( C \) and \( F \), and lines \( I \) and \( J \), respectively.

**Translation of arrays.** Since arrays in user language programs have a known fixed size, their translation is straightforward: a \( k \)-dimensional array \( M[i_1], \ldots, [i_k] \) translates to \( \prod_{i_1 \in \text{range}(0,k)} M \). According to the grammar in Figure 4, reduce operations can only be applied to anonymous arrays created by list comprehension. The expression `reduce_and(EXPR for ID in \( \text{range}(\text{FROM}, \text{TO}) \) if COND)` is translated to the Boolean event \( \bigwedge_{\text{ID} \in \text{FROM} \ldots \text{TO}} \text{COND} \lor \text{EXPR} \). Symmetrically, `reduce_or(EXPR for ID in \( \text{range}(\text{FROM}, \text{TO}) \) if COND)` translates to \( \bigvee_{\text{ID} \in \text{FROM} \ldots \text{TO}} \text{COND} \lor \text{EXPR} \), and `reduce_sum(EXPR for ID in \( \text{range}(\text{FROM}, \text{TO}) \) if COND)` translates to the event \( \sum_{\text{ID} \in \text{FROM} \ldots \text{TO}} \text{COND} \lor \text{EXPR} \).

4. **PROBABILITY COMPUTATION**

The probability computation problem is \#P-hard already for simple events representing propositional formulas such as positive bipartite formulas in disjunctive normal form [20]. In ENFrame, we need to compute probabilities of a large number of interconnected complex events. Although the worst-case complexity remains hard, we attack the problem with three complementary techniques: (1) bulk-compile all events into one decision tree while exploiting the structure of the events to obtain smaller trees, (2) employ approximation techniques to prune significant parts of the decision tree, and ultimately (3) distribute the compilation by assigning distinct distributed workers to explore disjoint parts of the tree.

---

**Algorithm 1:** Exact and approx. compilation of network

```
\[\text{DFS}(D, M, \{\}, B) \quad \text{\textbf{\textcolor{red}{error budget}} B) \quad \text{\textbf{\textcolor{blue}{do}}}}\]
```

We next introduce the bulk-compilation technique, look at three approximation approaches, and discuss how to distribute the probability computation.

4.1 **Compilation of event programs**

The event programs consist of interconnected events which are represented in an event network: a graph representation of the event programs, in which nodes are, e.g., Boolean connectives, comparisions, aggregates, and c-values.

The goal is to compute probabilities for the top nodes in the network, which are referred to as *compilation targets*. These nodes represent events such as "object \( o_i \) is assigned to cluster \( C_j \) in iteration \( I \)". We keep lower and upper bounds for the probability of each target. Initially, these bounds are \([0, 1]\) and they eventually converge during computation.

The bulk-compilation procedure is based on Shannon expansion: select an input random variable \( x \) and partially evaluate each compilation target \( \Phi \) to \( \Phi[x] \) for \( x \) being set to true \((\top)\) and \( \Phi[x] \) for \( x \) being set to false \((\bot)\). Then, the probability of \( \Phi \) is defined by \( Pr[\Phi] = Pr[\Phi[x] \cdot Pr[\Phi[x]] + Pr[\neg x] \cdot Pr[\Phi[\neg x]] \). We are now left with two simpler events \( \Phi[x] \) and \( \Phi[\neg x] \). By repeating this procedure, we eventually resolve all variables in the events to the constants true or false. The trace of this repeated expansion is the decision tree. We need not materialise the tree. Instead, we just explore it depth-first and collect the probabilities of all visited branches as well as record for each event \( \Phi \) the sums \( L(\Phi) \) and \( N(\Phi) \) of probabilities of those branches that satisfied and respectively did not satisfy the event. At any time, \( L(\Phi) \) and \( 1 - N(\Phi) \) represent, respectively, a lower and upper bound on the probability of \( \Phi \). This compilation procedure needs time polynomial in the network size (and in the size of the input data set), yet in worst case (unavoidably) exponential in the number of variables used by the events.

For practical reasons, we do not construct \( \Phi[x] \) and \( \Phi[\neg x] \).
Figure 5: Simplified example of an event network.

Example 4. Figure 5 shows a simplified event network under the assignment $f_0 \mapsto \top, x_1 \mapsto \bot$. The masks of $f_0$ and $x_1$ have propagated to event nodes $\Phi(o_0), \Phi(o_1), \Phi(o_2)$, which are also masked. The red nodes are masked for $\bot$, whereas the green nodes are masked $\top$.

Algorithm 4 gives the pseudocode for the DFS-traversal of the decision tree. The blue lines are necessary for approximating the probability computation and will be explained later. Compilation starts with an empty branch (variable assignment) $\nu$; the probability $Pr(\nu)$ is added to its lower bound if $\nu(x) = \top$, or subtracted from its upper bound if $\nu(x) = \bot$. If one or more targets are left unmasked, a next variable $x'$ is chosen and the process is repeated with $\nu' = \nu \cup \{x' \mapsto c\}$, where $c$ is either $\top$ or $\bot$. The algorithm chooses a next variable $x'$ such that it influences as many events as possible.

Once all compilation targets are masked by an assignment $\nu$, the compilation backtracks and selects a different assignment for the most recently chosen variable whose assignments are not exhausted. If all branches of the decision tree have been investigated, the probability bounds of the targets have necessarily converged and the algorithm terminates.

Algorithm 4 performs mask propagation: a variable assignment (mask) $M$ is inserted into the network, and the variable node propagates the mask to its parents. Depending on the node, its mask is either updated and propagated, or propagation is stopped in case the node cannot be masked. Convergence of the algorithm (e.g., clustering) can be detected by comparing the mask values at network nodes corresponding to iteration $t$ with the masks of nodes for iteration $t + 1$. If none of the mask assignments has changed between iterations, then the algorithm has converged.

4.3 Approximation with error guarantees

The compilation procedure can be extended to achieve an anytime absolute $\varepsilon$-approximation with error guarantees. The idea is to stop the probability computation as soon as the bounds of all compilation targets are sufficiently tight.

Definition 2. Given a fixed error $0 \leq \varepsilon \leq 1$ and events $(t_0, \ldots, t_{n-1})$ with probabilities $(p_0, \ldots, p_{n-1})$, an absolute $\varepsilon$-approximation for these events is defined as a tuple $(p_0, \ldots, p_{n-1})$ such that $0 \leq i < n : p_i - \varepsilon \leq p_i \leq p_i + \varepsilon$.

The compilation of the network yields probability bounds $[L_0, U_0], \ldots, [L_{n-1}, U_{n-1}]$ for the targets $(t_0, \ldots, t_{n-1})$. It can be easily seen that an absolute $\varepsilon$-approximation can be defined by any tuple $(p_0, \ldots, p_{n-1})$ such that $0 \leq i < n : U_i - \varepsilon \leq p_i \leq L_i + \varepsilon$. We thus need to run the algorithm until $U_i - L_i \leq 2\varepsilon$ for each target $t_i$.

There exist multiple strategies for investing this $2\varepsilon$ error budget for every target. We next discuss three such strategies. The lazy scheme follows the exact probability computation approach and stops as soon as the bounds become tight enough. Effectively, this results in investing the entire error budget into the rightmost branches of the decision tree. The eager scheme spends the entire error budget as soon as possible, i.e., on the leftmost branches of the decision tree, and then continues as in the case of exact computation. At each node in the decision tree, the hybrid scheme divides the current error budget equally over the two branches. Any unused error budget is transferred to the next branch.

The blue lines in Algorithm 4 show how the DFS procedure can be extended to support anytime absolute $\varepsilon$-approximation with error guarantees using the hybrid scheme. The DFS procedure is called using a non-zero error budget $2\varepsilon$, and it assigns half of the budget to the newly created left branch of the decision tree. The recursive DFS call returns the residual error budget of each target, which is then added to the budget for the right branch.

Other approximation techniques can be added to ENFrame, e.g., the randomised $\varepsilon$-approximation Oracle average, and probability computation using top-$k$ most probable worlds.

4.4 Distributed probability computation

By splitting the task of exploring the decision tree in a number of jobs, the compilation can be performed concurrently by multiple threads or machines. A worker explores a tree fragment of a given maximum size. For simplicity, we define the size of a job to be the depth $d$ of the sub-tree
to explore. The computation then proceeds as follows. One worker explores the tree from the root and every time it reaches depth $d$, it forks a new job that continues from that node as its root. Given that the maximum depth of the tree is the number of variables $m$, the number of jobs created is at most $\sum_{i=0}^{\lfloor \log_2 d \rfloor} 2^m$, where the cost of each job would propagate at most $2^m$ variable valuations $\nu$ into the event network. Each job incurs the cost of communicating the mask at job creation and the probability bounds for each target at the end of the job. In case of approximation, the error budgets need to be synchronised both at the start and end of a job.

5. EXPERIMENTAL EVALUATION

This section describes an experimental evaluation of clustering probabilistic data using ENFrame. The focus of this evaluation is a preliminary benchmark of the performance of the probability computation algorithms introduced in Section 4. At the end of this section, we comment on further experimental considerations that could not be included in full due to space limitations.

Data. We use a data set describing network load and occurrences of partial discharge in energy distribution networks. This data is gathered from two different types of sensors: partial discharge sensors installed on switchgear and cables in substations of the distribution network, and network load sensors in substations. We aggregate the number of partial discharge occurrences over the duration of an hour and subsequently pair this value with the average network load during that hour. Clustering can assist in detecting anomalies and predicting failures in the energy networks.

Uncertainty. Our goal is to show that ENFrame can deal with common correlations patterns that occur in probabilistic data. Each data point is associated with an event described by Boolean random variables, whose probabilities for true are chosen at random from the range $[0.5, 0.8]$. Different values would make the probabilities of clustering events too close to 0 or 1 which are then easily approximable. The experiments were carried out using three types of correlations to illustrate ENFrame’s capability to process arbitrarily correlated data.

The positive correlations scheme yields events such that two data points are either positively correlated or independent. Each event is a disjunction of $l$ distinct positive literals. In the mutex correlations scheme, the data points are partitioned in mutex sets of cardinality (at most) $m$: any two points are mutually exclusive within a mutex set and independent across the sets. The conditional correlations scheme expresses uncertainty as a Markov chain, using one node per data point. Let $\Phi$, be the event that the data point $a_i$ exists. The event $\Phi_{i+1}$ becomes $\Phi_{i+1} = \Phi_i \land x_{i+1}$: it is a conjunction of two events, for the cases that $a_i$ exists or not. We thus introduce two new Boolean random variables $x_{i+1}$ and $x'_{i+1}$ per data point $a_{i+1}$. For every correlation scheme, a group size of 4 has been used, i.e. data points were divided in groups with identical lineage. This is realistic for uncertain time-series sensor data: readings from a small time window have identical correlations and uncertainty. Additionally, we show experiments with a varying fraction of certain data points.

ENFrame algorithms. We report on performance benchmarks for $k$-medoids clustering on the energy network data set, comparing ENFrame to naive clustering in the top-$k$ worlds. The naive approach computes a clustering by explicitly iterating over all possible worlds. We show the performance of multiple probability computation algorithms of ENFrame: the sequential exact approach, three sequential approximation schemes (eager, lazy, hybrid), and distributed hybrid approximation (hybrid-d). All approximation algorithms are set to compute probabilities with an (absolute) error of at most $\varepsilon = 0.1$, the compilation targets are the events that represent medoid selection.

The literature describes both existing platforms for processing of uncertain data and various algorithms for clustering of uncertain data. Only the newest version of SPROUT supports a form of conditional values, yet only for query evaluation; no other probabilistic DBMS (including MCDB, MayBMS) provides this functionality, which makes it impossible to compare ENFrame to any of the platforms described in Section 6. In addition, work on probabilistic clustering only considers basic uncertainty models based on independence. Those algorithms might outperform our sequential algorithms, at the cost of producing an output that can be arbitrarily off from the golden standard of clustering in every possible world. None of the prototypes was available for testing at the time of writing.

Setup. The experiments were carried out on an Intel Xeon X5660/2.80GHz machine with 4GB of RAM, running Ubuntu with Linux kernel 3.5. The timings reported for hybrid-d were obtained by simulating distributed computation on a single machine. The algorithms are implemented in C++ and compiled using GCC 4.7.2. Each plot in Figures 6 and 7 depicts average performance with min/max ranges of five runs with randomly generated event expressions, different probabilities, and three clustering iterations (using Euclidean distance).

Sequential algorithms. Figures 6 and 7 show that all of ENFrame’s probability computation algorithms outperform the naive algorithm by up to six orders of magnitude for each data set with more than 10 variables. Furthermore, the
**Figure 6:** Positively correlated data. On the left: scalability in terms of variables, on the right: scalability of approximations in terms of size of the data set (hybrid-d not shown for visibility).

**Figure 7:** Mutex and conditionally correlated data (legend: see Fig. 6). Algorithms with exact and lazy overlap with exact, and are not shown. Grey dashed line indicates number of variables.

The hybrid approximation can be up to four orders of magnitude faster than exact computation. Indeed, for a very small number of possible worlds (i.e., a small number of variables), it pays off to cluster individually in each world and avoid the overhead of the event networks. For a larger number of worlds, our exact and approximate approaches are up to six orders of magnitude faster. The naive method times out for over 25 variables in every correlation scheme.

The reason why our approximation schemes outperform exact is as follows. For a given depth \( d \), there are up to \( 2^d \) nodes in the decision tree that contribute to the probability mass of a node in the event network. The contributed mass decreases exponentially with an increase in depth, suggesting that most nodes in the decision tree only contribute a small fraction of the total mass. Depending on the desired error bound, a shallow exploration of the decision tree could be enough to obtain a sufficiently large probability mass.

Among the approximation algorithms, hybrid performs best; it outperforms exact by up to four orders of magnitude since it does only need to traverse a shallow prefix of the decision tree. The algorithm invests the error budget over the entire width of the decision tree, cutting branches of the tree after a certain depth. The other two methods (eager and lazy) use the budget to respectively cut the first and last branches, while exploring other branches in full depth.

For positive correlations, lazy performs very well, because the decision tree is very unbalanced under this scheme. The left branches of the tree correspond to variables being set to true, which quickly satisfy the (disjunctive) input events and allow for compilation targets to be reached. Further to the right, branches correspond to variables being set to false. More variables need to be set to (un)satisfy the disjunctive input event, thus leading to longer branches. The lazy algorithm saves the error budget until the very last moment and can therefore prune the deep branches whilst maintaining the \( \varepsilon \)-approximation. The decision trees for the mutex and conditional correlation schemes are more balanced, resulting in both lazy and eager to perform almost identically to exact. Hence, they are not shown in Figure 7.

**Distributed algorithms.** By distributing the probability computation task, we can significantly increase ENFrame’s performance. Figures 6 and 7 show the timings for hybrid-d using \( w = 16 \) workers and job size \( d = 3 \). For all correlation schemes, hybrid-d gets increasingly faster than hybrid as we increase the number of variables. For small numbers of variables, the overhead of distribution does not pay off. The benefits are best seen for mutex correlations and over 100 objects (over 60 variables), where hybrid-d becomes more than one order of magnitude faster than hybrid. For readability reasons, the performance of hybrid-d is not depicted in Figure 6; its performance is up to one order of magnitude better than hybrid, as can be seen in Figure 8. For ten variables, there is only a small performance gain when compared to the single-threaded hybrid approximation: the decision tree remains small, as is the number of jobs that can be generated. However, for 30 and 50 variables, hybrid-d yields a performance improvement of more than one order of magnitude over hybrid.

Figure 10 shows the influence of the number of workers on hybrid-d’s performance for varying job sizes. A job is the work unit allocated to a worker at any one time; a size of \( d \) means that the worker has to explore a fragment of the decision tree of depth at most \( d \) and would need to traverse the event network at most \( 2^d \) times. For large job sizes, the overall number of jobs decreases; in the case of positive correlations, the number of jobs of size 9 is small since the decision tree is very unbalanced and only a few branches on the right-hand side of the tree grow deeper than nine variables. Therefore, increasing the number of workers would not help; indeed, there is no improvement for more than four workers for job sizes larger than 5. However, for a job size of 3, up to 16 workers can still be beneficial. In our experiment, smaller job sizes led to a performance gain of up to one order of magnitude, since they allow for a more...
equal distribution of the work over the available workers. Synchronisation did not play a significant role in our setup.

**Certain data points.** Figure 8 shows that performance improves as the number of certain data points (objects that occur in all possible worlds) increases. The speed-up in such cases is explained by the fact that the distance sums of medoids to data points in a cluster become less complex and can be initialised using the distances to objects that certainly exist. Consequently, fewer variables assignments are needed to decide on a cluster medoid, resulting in a shallower decision tree and a speedup in the compilation time.

**Clustering accuracy.** We investigated the accuracy of ENFrame versus clustering in the top-k most probable worlds and naive clustering. We consider the latter as golden standard, since it respects correlations and does not introduce errors due to approximations of probabilities. To this end, we developed an external evaluator for clustering of uncertain data: the pairwise assignment probability metric (PAPM). Cluster evaluation using pairwise object assignments is a well-described technique for deterministic clustering. We compare two clustering results \( C_1, C_2 \) using the probability that pairs of distinct objects \( a_n, a_o \) are assigned to the same cluster (pairwise assignment probability: \( \text{Pr}(a_n \sim a_o) \)). PAPM reports the maximum difference in this probability between \( C_1 \) and \( C_2 \) over all pairs of objects:

\[
\text{PAPM}(C_1, C_2) = \max_{a_n, a_o} \left| \text{Pr}_{C_1}(a_n \sim a_o) - \text{Pr}_{C_2}(a_n \sim a_o) \right|
\]

Figure 9 experimentally confirms that the accuracy of ENFrame's exact clustering is equal to the golden standard of clustering in all possible worlds (naive). The accuracy of clustering in the top-k most probable worlds if far off from the golden standard, even for large \( k \). Already for \( k > 1000 \), ENFrame outperforms top-k clustering (not shown in this figure). As expected, the difference between hybrid and naive never exceeds 0.1. The data set for this experiment was restricted due to the limited scalability of naive.

A more extensive accuracy comparison is out of scope of this paper, but is part of future work. A common approach to assess the accuracy measure of a probabilistic method like ours, is to assume a notion of ground truth. It is unclear how to deal with probabilistic data which represents inherently contradictory information for which no ground truth exists or is known. In turn, we proposed PAPM, which takes correlations and probabilities into account.

**Further findings.** We have investigated the influence of the number of dimensions, data point coordinates, the error budget, the numbers of iterations, and alternative clustering compilation targets on the performance of ENFrame, as well as its total memory usage. As is the case with k-medoids on certain data, the number of dimensions has no influence on the computation time. The reported performance gap between exact and hybrid shows that performance is highly sensitive to the error budget. The number of iterations has a linear effect on the running time of the algorithm. The number of targets (including those representing co-occurrence queries) has a minor influence on performance; due to the combinatorial nature of k-medoids, events are mostly satisfied in bulk and it is thus very rare that one event alone is satisfied at any one time. This also explains why experiments with other types of compilation targets (e.g., object-cluster assignment, pairwise object-cluster assignment) show very similar performance. In our experiments, the size of the event networks grows linearly in the number of objects and clusters and the memory usage of ENFrame is under 1GB.

## 6. RELATED WORK

Our work is at the confluence of several research areas: probabilistic data management, data analytics platforms, and provenance data management. A key aspect that differentiates ENFrame from the algorithms and platforms described in this section, is the probabilistic and correlated nature of input data and the entire processing pipeline. This calls for tailored algorithms. As a platform for expressing algorithms to process uncertain data, ENFrame goes beyond any of the single data mining algorithms described here.

**Probabilistic data mining and querying.** Our work adds to a wealth of literature on this topic [1, 29] along two directions: distributed probabilistic computation techniques and a unified formalisation of several clustering algorithms in line with work on probabilistic databases. Distributed probabilistic computation has been approached only in the context of the SimSQL/MCDB system, where approximate query results are computed by Monte Carlo simulations [17, 18]. This contrasts with our approach in that MCDB was not designed for exact and approximate computation with error guarantees and does not exploit correlations allowed by pc-tables and ENFrame.

Early approaches to mining uncertain data are based on imprecise (fuzzy) data, for example using intervals, and produce fuzzy (soft) and hard output. Follow-up work shifted to representation of uncertainty by (independent) probability density functions per data point. In contrast, we allow for arbitrarily correlated discrete probability distributions. The importance of correlations has been previously acknowl-
edged for clustering [32] and frequent pattern mining [30]. A further key aspect of our approach that is not shared by existing uncertain data mining approaches is that we follow the possible worlds semantics throughout the whole mining process. This allows for exact and approximate computation with error guarantees and sound semantics of the mining process that is compatible with probabilistic databases. This cannot be achieved by existing work; for instance, most existing k-means clustering approaches for uncertain data define cluster centroids using expected distances between data points [8, 25, 12, 20, 14, 18] or the expected variance of all data points in the same cluster [13]; they also compute hard clustering where the centroids are deterministic. The recently introduced UCPC approach to k-means clustering [15] is the first work to acknowledge the importance of probabilistic cluster centroids. However, it assumes independence in the input and does not support correlations.

Data analytics platforms. Support for iterative programs is essential in many applications including data mining, web ranking, graph analysis, and model fitting. This has recently led to a surge in data-intensive computing platforms with built-in iteration capability. REX supports iterative distributed computation along database operations in which changes are propagated between iterations [29]. MADlib is an open-source library for in-database analytics [16]. Similarly, Bismark is an architecture for in-database analytics [9]. GraphLab [21] uses graph representations for scalable parallel programming. The Iterative Map-Reduce-Update programming abstraction for machine learning compiles programs into declarative Datalog code [9]. Infer.NET [23] originates from the programming languages community, but has a closed nature with restricted availability.

Provenance in database and workflow systems. To enable probability computation, we trace fine-grained provenance of the user computation. This is in line with a wealth of work in probabilistic databases [29]. Our event language is influenced by work on provenance semirings [11] and semi-modules [10] that capture provenance for positive queries with aggregates in relational databases. The construct $\Phi \otimes v$ resembles the algebraic structure of a semimodule that is a tensor product of the Boolean semiring $B[X]$ freely generated by the variable set $X$ and of the SUM monoid over the real numbers $\mathbb{R}$. There are two differences between our construct and these structures. Firstly, we allow negation in events, which is not captured by the Boolean semiring. Secondly, even for positive events, $B[X] \otimes R$ is not a semimodule since it violates the following law: $(\Phi_1 \lor \Phi_2) \otimes v = \Phi_1 \otimes v + \Phi_2 \otimes v$. Indeed, under an assignment that maps both $\Phi_1$ and $\Phi_2$ to $\top$, the left side of the equality evaluates to $v$, whereas the right side becomes $v + v$. Furthermore, our event language allows to define events via iterations, as needed to succinctly trace data mining computation.

7. REFERENCES