Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-021-00242-6

REGULAR PAPER

# The space of models in machine learning: using Markov chains to model transitions

**Vicenç Torra[1,2,3] · Mariam Taha[4] · Guillermo Navarro-Arribas[5]**

Received: 4 April 2020 / Accepted: 26 March 2021


**Abstract**
Machine and statistical learning is about constructing models from data. Data is usually understood as a set of records, a database. Nevertheless, databases are not static but change over time. We can understand this as follows: there is a space of possible databases and a database during its lifetime transits this space. Therefore, we may consider transitions between databases, and the database space. NoSQL databases also fit with this representation. In addition, when we learn models from databases, we can also consider the space of models. Naturally, there are relationships between the space of data and the space of models. Any transition in the space of data may correspond to a transition in the space of models. We argue that a better understanding of the space of data and the space of models, as well as the relationships between these two spaces is basic for machine and statistical learning. The relationship between these two spaces can be exploited in several contexts as, e.g., in model selection and data privacy. We consider that this relationship between spaces is also fundamental to understand generalization and overfitting. In this paper, we develop these ideas. Then, we consider a distance on the space of models based on a distance on the space of data. More particularly, we consider distance distribution functions and probabilistic metric spaces on the space of data and the space of models. Our modelization of changes in databases is based on Markov chains and transition matrices. This modelization is used in the definition of distances. We provide examples of our definitions.

**Keywords** Machine and statistical learning models · Space of data · Space of models · Hypothesis space · Probabilistic metric spaces

## 1 Introduction

Machine and statistical learning can be seen as a search problem. That is, we have a state space corresponding to possible models and we want to find the one that better represents our data. Here, *better* can correspond to the one with best accuracy. Different definitions of *better* as well as different search strategies to find a good solution can be considered. From this perspective, we consider operators that permit to transform one model into another one. In this case, a transformation is usually to improve accuracy (i.e., *better = better accuracy)*. Examples of these transformations include operators that expand a node in a decision tree, update weights in a deep learning model, or operators that mutate a solution in genetic algorithms.

In this paper, we consider a different perspective. We consider the space of models taking into account the space of data that have generated these models.

When we learn a model from an actual database, the database is just a database from the space of data. When databases change, we are traversing the space of data through a particular path. Different databases in this particular path can lead to different machine learning models.

✉ Vicenç Torra
vtorra@ieee.org

Mariam Taha
mariamt@cs.umu.se

Guillermo Navarro-Arribas
guillermo.navarro@uab.cat

[1] Department Computing Science, Umeå University, Umeå, Sweden

[2] Hamilton Institute, Maynooth University, Maynooth, Ireland

[3] School of Informatics, University of Skövde, Skövde, Sweden

[4] Department Computing Science, Umeå University, Umeå, Sweden

[5] Dept. Information and Communications Engineering – CYBERCAT, Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain

Published online: 12 April 2021

We claim that it is necessary to study how the space of data interacts with the space of models. More particularly, that any decision on the space of models has to take into account relationships between the space of data that generate these models.

We consider that this perspective is of great interest in the following areas.

- *In model selection for statistical and machine learning* In this area the goal is to select models that better generalize data and avoid overfitting. The study of the interaction between the space of models and the space of data can increase our understanding on the models themselves, and on the methods that generate the models (comparing their respective mappings between the two spaces). In particular, we think it is fundamental to understand the concepts of generalization and overfitting. This also relates to the effect of outliers and influential points in learning. It is important to understand generalization and overfitting in terms of the relationship between the space of models and the space of data.
- *In privacy preserving data mining and machine learning* The need to study the relationship between the two spaces was first proposed in [14] in the context of integral privacy [10,11]. In short, a model is integrally private if it can be generated by a large number of databases, and these databases are sufficiently different (e.g., they do not share records). This is to avoid some type of privacy attacks on machine learning models as, e.g., membership attacks [12].

We are interested in knowing when two models are similar, where similar does not correspond to a syntactic similarity of the models (e.g., if two decision trees have the same structure), nor on a semantic similarity of the models (e.g., if two models have the same accuracy). We are interested in knowing when models are similar because they have been generated from similar databases.

There are naturally different ways of understanding the similarity between databases. For example, one database may be similar because it is a noisy version of the other (e.g., an anonymized version of the original database). Here, we focus on changes in databases due to the natural processes a database suffers in a company. That is, we consider a database that is updated, as time passes, by means of e.g., adding and removing records. These types of changes are usual when databases are in production. In addition, these types of changes are also relevant in the framework of data privacy [13] with the right to be forgotten and the right to amend (under the GDPR).

In [14], the authors proposed the use of probabilistic metric spaces [9] for modeling the similarity between models. Informally, these spaces are defined in terms of distance distribution functions. That is, distance between pairs of objects are not a real number but a distribution on these numbers. This approach permits us to define a distance between pairs of models taking into account the distance between the set of databases that have generated these models.

In this paper, we propose the use of Markov chains and transition matrices to represent, respectively, sequences of changes in databases and the probability of changes taking place. This representation permits the definition of probabilistic metric spaces on the space of data. We use them later to define distance distribution functions for the space of models in terms of the databases that have generated them. This is a much simpler approach than the one introduced in [14].

The structure of this paper is as follows. In Sect. 2, we introduce the definitions that are needed later in the paper. In particular, we introduce Markov chains and probabilistic metric spaces. In Sect. 3, we introduce two definitions of metric spaces for databases based on Markov chains and prove some results. In Sect. 4, we introduce definitions for distance distribution functions for models based on the probabilistic metric spaces introduced in Sect. 3. We provide some examples of how these distances can be actually computed. The paper finishes with a discussion and lines for future work.

## 2 Preliminaries

In this section, we review some concepts that are needed later. We begin with Markov chains and transition matrices. We also discuss probabilistic metric spaces and distances for sets of elements.

### 2.1 Markov chains

In this paper, we will use Markov transition matrices and Markov chains to model the space of databases. Because of that, we will review in this section a few concepts that we need later. See e.g., [7] for details.

We consider a state space $S$ finite or enumerable. We will use

$$S = \{DB_1, DB_2, DB_3, \dots\}$$

to denote the space of possible databases. Thus, in our case, a finite although extremely huge set.

We will consider chains defined on the state space $S$. That is, $(Z_n)_{n\in\mathbb{N}}$ taking values in $S$, i.e., $Z_n \in S$. More particularly, we consider Markov chains. This corresponds to chains in which the probability distribution on $Z_{n+1}$ depends only on the process $Z_n$ at time $n$ and not on previous values of $Z$. In other words, there is no memory on previous transitions.

Formally,

$$P(Z_{n+1} = DB_j | Z_n = DB_i, Z_{n-1} = DB_{n-1}, \ldots, Z_0 = DB_0) = P(Z_{n+1} = DB_j | Z_n = DB_i).$$

We consider time-homogeneous Markov chains. That is, the probability of transition does not depend on time. This is expressed mathematically as

$$P(Z_{n+1} = DB_j | Z_n = DB_i) = P(Z_{m+1} = DB_j | Z_m = DB_i)$$

for any $n, m$.

As the probability does not depend on $n$, we will not use this index unless required. Then, we will use $P_{ij}$ to denote $P(Z_{n+1} = DB_j | Z_n = DB_i)$ (for any $n$). For the sake of simplicity, we will also use $P(Z_{n+1} = j | Z_n = i)$ when no confusion arises.

From the explanation above, it is clear that transition depends only on the probabilities $P_{ij}$. These probabilities for all states $i$ and $j$ define a matrix. It is known as transition matrix. Formally, a transition matrix $P$ is a $S \times S$ matrix with values in $[0, 1]$ such that (for any $n$)

$$\sum_{DB_j \in S} P_{ij} = \sum_{DB_j \in S} P(Z_{n+1} = DB_j | Z_n = DB_i) = 1.$$

We can prove that given a probability distribution $\pi$ on $S$ for time 0, say probabilities $P(Z_0 = i)$ for $i \in S$, the probability distribution for time 1, say probabilities $P(Z_1 = j)$ for $j \in S$, can be expressed in matrix notation as $\pi P$. Let us denote by $P^n$ the transition matrix defined by $P^n_{ij} = P(Z_{m+n} = DB_j | Z_m = DB_i)$. Naturally, the computation of $P^n_{ij}$ does not depend on $m$. We can prove that

$$P^{r+t}_{ij} = \sum_{k \in S} P^r_{ik} P^t_{kj},$$

or in matrix form $P^{r+t} = P^r P^t$. This is called the Chapman–Kolmogorov equation.

## 2.2 Probabilistic metric spaces

Probabilistic metric spaces [9] are a generalization of metric spaces in which a distance distribution function replaces the role of distance functions. That is, instead of considering $d(a, b)$ as a real number, it is a distribution function on the real numbers.

Recall that metric spaces are defined in terms of sets (a non-empty set) and a distance or metric for pairs of elements in this set. Formally, we denote a metric space by $(S, d)$, where $S$ is the set and $d$ for $a, b \in S$ the distance. The function $d$ is required to satisfy the following properties: (i) positiveness, (ii) symmetry, and (iii) triangle inequality (formally, $d(a, b) \leq d(a, c) + d(c, b)$ for any $a, b, c$ in $S$). Also, it is usual to require that if $a$ and $b$ are different then the distance should be strictly positive. Special names are given when some of these conditions fail. For example, when the distance does not satisfy the symmetry condition, we say that $(S, d)$ is a quasimetric space; and when the distance does not satisfy the triangle inequality, we say that $(S, d)$ is a semimetric space.

Probabilistic metric spaces are a generalization of metric spaces. As stated above, we can informally consider that we replace the function $d(a, b)$ by a distribution function $F(a, b)$ defined on $\mathbb{R}$. These functions are known as distance distribution functions. Their definition follows.

**Definition 1** [9] A nondecreasing function $F$ defined on $\mathbb{R}^+$ that satisfies (i) $F(0) = 0$; (ii) $F(\infty) = 1$, and (iii) that is left continuous on $(0, \infty)$ is a distance distribution function. $\Delta^+$ denotes the set of all distance distribution functions.

The following interpretation is usual for these functions: $F(x)$ corresponds to the probability that the distance is less than or equal to $x$. Note that this definition is a generalization of a distance.

In particular, we use $\epsilon_a$ to denote the distance distribution function that represents the classical distance $a$. This $\epsilon_a$ function is just a step function at $a$. Its definition follows.

**Definition 2** [9] For any $a$ in $\mathbb{R}$, we define $\epsilon_a$ as the function given by

$$\epsilon_a(x) = \begin{cases} 0, & -\infty \leq x \leq a \\ 1, & a < x \leq \infty. \end{cases}$$

In order to define probabilistic metric spaces we need to consider the set of distance distribution functions, and we need to define a condition on triples of functions in this set analogous to the triangle equality in metric spaces. This condition given below is based on triangle functions. Let us start defining the triangle functions.

**Definition 3** [9] Let $\Delta^+$ be defined as above, then a binary operation on $\Delta^+$ is a triangle function if it is commutative, associative, and nondecreasing in each place, and has $\epsilon_0$ as the identity.

Using triangle functions we can establish the definition of probabilistic metric spaces.

**Definition 4** [9] Let $(S, \mathcal{F}, \tau)$ be a triple where $S$ is a nonempty set, $\mathcal{F}$ is a function from $S \times S$ into $\Delta^+$, $\tau$ is a triangle function; then $(S, \mathcal{F}, \tau)$ is a probabilistic metric space if the following conditions are satisfied for all $p$, $q$, and $r$ in $S$:

- (i) $\mathcal{F}(p, p) = \epsilon_0$
- (ii) $\mathcal{F}(p, q) \neq \epsilon_0$ if $p \neq q$
- (iii) $\mathcal{F}(p, q) = \mathcal{F}(q, p)$
- (iv) $\mathcal{F}(p, r) \geq \tau(\mathcal{F}(p, q), \mathcal{F}(q, r))$.

As usual in this field, we will use $F_{pq}$ instead of $\mathcal{F}(p, q)$. This permits to express the value of the distance distribution function at $x$ by means of the expression: $F_{pq}(x)$.

### 2.3 Metrics for sets of objects

In order to define the distance between pairs of models, we will consider the set of databases that have generated these models. This permits to define the distance in terms of the distance between these sets. Let us first consider the classical setting with a standard distance.

Let $G$ be an algorithm that given a database generates a model, then, the set of generators of a model $m$ is defined by $Gen_m = \{DB | G(DB) = m\}$. In this case, given two models $m_1$ and $m_2$ we define the distance between $m_1$ and $m_2$ in terms of $Gen_{m_1}$ and $Gen_{m_2}$. In order to do so, we need to extend the distance for databases to sets of databases.

Nevertheless, given a metric space $(S, d)$, its extension to a set of elements of $S$ is not trivial. This is so because although several distances have been defined on sets, not all of them satisfy the triangle inequality. This implies that they are not valid to define a metric.

In [14], different types of functions are considered. The discussion includes the Hausdorff distance and the sum of minimum distance (which do not satisfy the triangle inequality) and the definition by Eiter and Mannila [3] that is indeed a valid definition of a distance and leads to a metric space. Nevertheless, this is a very complex function to compute.

## 3 Probabilistic metric spaces from Markov chains

We consider that transition matrices are a suitable approach to model changes on databases. In other words, we consider that for a given database there is some probability that this database is transformed by means of a modification to another database. For the sake of simplicity, we consider in this work time-homogeneous Markov chains. That is, as explained in Sect. 2 that changes on a database only depend on what is currently available in the database and that it does not depend on its previous values (history of the database). This assumption can be considered simplistic, as e.g., the probability of adding a record may depend on how many times has been already present in the database and has been removed. Nevertheless, we consider that this assumption is acceptable for this initial study.

For illustration, we consider only addition and deletion of records from a database, and that only one addition and one deletion is allowed at a time. We also assume that we have access to the whole population or that we know the size of the whole population. Then, we can define a transition matrix based on assigning a probability of having a deletion ($p_d$) and a probability of having an addition ($p_a$). Naturally, these probabilities add less than or equal to one ($p_d + p_a \leq 1$).

**Definition 5** Let $p_d$ and $p_a$ be the probability of deleting or adding a record. Then, given an arbitrary database $DB_i$, where $DB_i$ is a subset of the whole population $P$ (with $|P|$ denoting the size of this population), we define the probability of transition from $DB_i$ to any $DB_j$ as follows (here, $p_{ij}$ stands for $P(Z_{n+1} = DB_j | Z_n = DB_i)$ as above):

$$p_{ij} = \begin{cases} p_d \frac{1}{|DB_i|} & \text{if } c_1 \& c_3 \\ p_a \frac{1}{|P| - |DB_i|} & \text{if } c_2 \& c_3 \\ \frac{1}{|P|} & \text{if } (c_1 \text{ or } c_2) \& c_4 \\ 1 - (p_d + p_a) & \text{if } c_5 \& c_3 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $c_1 - c_5$ are the following conditions:

- $c_1$: $DB_j \subset DB_i$ and $|DB_i \setminus DB_j| = 1$
- $c_2$: $DB_i \subset DB_j$ and $|DB_j \setminus DB_i| = 1$
- $c_3$ : $|DB_i| \notin \{0, |P|\}$
- $c_4$: $|DB_i| \in \{0, |P|\}$
- $c_5$: $DB_i = DB_j$

Here, $c_4$ means that the database $DB_i$ is either empty or no further records can be added to it, and $c_3$ means that $DB_i$ is not one of such extreme databases.

**Lemma 1** *The above definition leads to a valid transition matrix. That is,* $\sum_j p_{ij} = 1$ *for all* $j$.

***Proof*** Observe that given $DB_i$, $p_{ij}$ is not zero for databases $DB_j \neq DB_i$ that have either one additional record more (i.e., $DB_i \subset DB_j$ and $|DB_j \setminus DB_i| = 1$) or less (i.e., $DB_j \subset DB_i$ and $|DB_i \setminus DB_j| = 1$) than $DB_i$. Then, in the general case, $DB_i$ can lead to any of the $|DB_i|$ databases that has just one record less, or $DB_i$ can lead to any of the $|P| - |DB_i|$ databases that have exactly one additional record. So, according to Equation 1, the probability of deleting a record is $p_d$ and the probability of adding a record is $p_a$. As the probability of $DB_i$ not being modified is $1 - (p_d + p_a)$, it is proved that the definition leads to a row adding to one. Then, we have conditions for the extreme cases in which $DB_i$ is empty or $DB_i$ includes all records. In this case, there are $|P|$ neighboring databases with a probability of transition equal to $\frac{1}{|P|}$. Therefore, this row also adds to one. Therefore, the matrix is a transition matrix. □

In this section, we introduce two definitions of probabilistic metric spaces for databases based on transition matrices and Markov chains.

The first definition considers the distance between two databases in terms of the probability of being transformed into the second one. This approach defines the probabilistic metric space solely based on the transition matrices. We give below both symmetric and asymmetric definitions for the distance distribution functions. See Definition 6. We call this type of space, visited database-based probabilistic metric space (VD-PMS).

The second definition considers the distance between two databases in terms of their evolution. That is, given two databases, will they be similar as time passes? In order to give a formal definition, we need to consider how databases are being modified, and what similarity means for databases. With respect to the later, the model presumes the existence of a standard distance function (a metric space, in fact) on the space of databases. We call this type of space, database distance-based probabilistic metric space (DD-PMS). See Definition 7.

We consider that both types of definitions are relevant for statistical and machine learning and, in particular, for privacy preserving data mining. We are interested in models that are valid today but that will be also valid in the future. So, the first definition states that two models are similar if we can transit from one to the other and the second definition states that two models are similar if they have a *similar future* (a similar machine learning model in the future).

### 3.1 Visited database-based probabilistic metric spaces

We consider a definition of probabilistic metric spaces for databases based on [5]. The distance between two databases depends on the probability that one database becomes the second one after a sequence of changes (there is a chain between the first to the second) within a given time frame.

The definition is based on transition matrices $P$ on the space of databases. The definition follows.

**Definition 6** Let $S$ be a state space representing the space of databases, let $P$ be the transition matrix for $S$ that defines a time-homogeneous Markov chain $(Z_n)_{n\in\mathbb{N}}$. Then, given two states $i$ and $j$ in $S$ we define

$$F_{ij}(t) = P[\text{exists a time } s < t \text{ such that } Z_s = j | Z_0 = i].$$

Formally, let $f_{ij}^s$ denote the probability that with $Z_0 = i$ (i.e., starting the chain from state $i$), the first time we visit state $j$ is exactly at time $s$. Then, $F_{ij}(t) = \sum_{s=1}^{t} f_{ij}^s$.

From the point of view of the space of databases, the definition above establishes that the distance between two databases $DB_1$ and $DB_2$ for the value $t$ is $\alpha$ (i.e., $P_{12}(t) = \alpha$) if the probability of reaching $DB_2$ from $DB_1$ in less than $t$ transitions is $\alpha$.

We can prove from this definition that $F_{ij}(t_1 + t_2) \geq F_{ik}(t_1)F_{kj}(t_2)$. From this property, we can prove the following theorem. See [5] for a proof. Observe that the formulation of the following theorem in [5] uses stationary to refer to time-homogeneous, using the notation in [2].

**Theorem 1** *Let $S$, $P$, $(Z_n)_{n\in\mathbb{N}}$ and $F_{ij}(t)$ be defined as in Definition 6. Let $\mathcal{F}$ be the mapping from $S\times S$ into the space of cumulative distribution functions defined by $\mathcal{F}(i, j) = F_{ij}$. Then, $\mathcal{F}$ satisfies properties (i) and (iii) in Definition 1, and properties (i), (ii), and (iv) in Definition 4.*

*It is a non-symmetric distance distribution function satisfying (iv) under the t-norm $T = Prod$ (i.e., $T(a, b) = ab$).*

The hitting time of a state $DB_j$ starting from state $DB_i$ is the random variable defined by

$$T_{ij} = \min\{n \geq 0 : X_n = DB_j\}$$

with the minimum of the empty set defined as $\infty$. The probability of hitting state $DB_j$ is defined by

$$h_i^j = P(T_{ij} < \infty).$$

Not all transition matrices lead to Markov chains with probabilities of hitting a state equal to 1. If this is the case, then, the Definition above will lead to a probabilistic metric space with a non-symmetric function. We establish this in the next theorem.

**Theorem 2** *Let $S$, $P$, $(Z_n)_{n\in\mathbb{N}}$ and $F_{ij}(t)$ be defined as in Definition 6. Then, the pair $(S, \mathcal{F})$ is a probabilistic metric space with a non-symmetric distance function under the t-norm $T = Prod$ when $h_i^j = 1$ for all $i, j$.*

Definition 6 gives a distance that is not necessarily symmetric. Note that accessing $j$ from $i$ at time $t$ does not mean that it is possible to access $i$ from $j$ in the same time $t$.

It is possible to define a probabilistic metric space with a symmetric distance function using $F'_{ij} = \sqrt{F_{ij}F_{ji}}$ or $F''_{ij} = 0.5(F_{ij} + F_{ji})$. The first definition is a probabilistic metric space satisfying the Menger inequality with $T = Prod$ and the second one satisfies the Menger inequality with $T = T_m$ (i.e., $T_m(a, b) = \max(a + b - 1, 0)$).

### 3.2 Computation and example

Definitions above use $f_{ij}^s$, that as explained above, denotes the probability that with $Z_0 = i$ the first time we visit state $j$ is exactly at time $s$ when we start at $Z_0 = i$. For a given transition matrix $P$, we can compute $f_{ij}^s$ as follows. If $s = 1$,

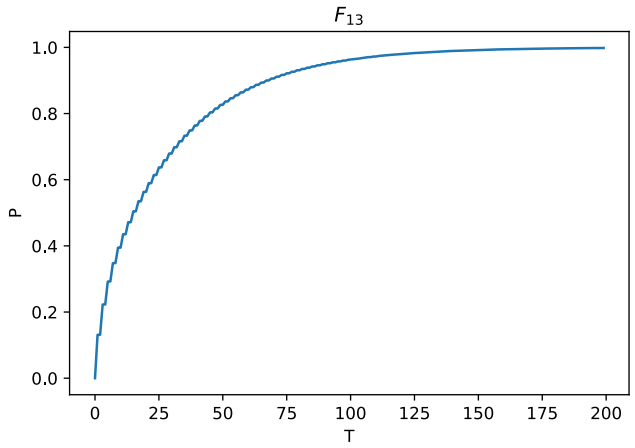


**Fig. 1** $F_{13}$ for Example 1 according to Definition 6

$f_{ij}^s = P_{ij}$. If $s > 1$ then we perform the following steps: (i) define $\bar{P}$ as $P$ and assigning $\bar{P}_{rs} = 0$ for all pairs $(j, t)$ and $(t, j)$ with $j \in S$, (ii) compute $\bar{P}^{s-1}$, (iii) compute $f_{ij}^s = \sum_{k \in S} \bar{P}_{ik}^{s-1} P_{kj}$. Finally, (iv) using the expression in Definition 6 we compute $F_{ij}(t)$ as $F_{ij}(t) = \sum_{s \leq t} f_{ij}^s$.

The rationale of this definition is that in order to reach $j$ from $i$ in exactly $s$ steps (with $s > 1$) we need to reach any other state $k \neq j$ in exactly $s-1$ steps without reaching $j$ at any moment $s' < s$, and then move from $k$ to $j$. Probabilities of reaching $k \neq j$ from $i$ in $s-1$ steps without hitting $j$ will be computed using $\bar{P}$. This computation corresponds to compute $\bar{P}^{s-1}$ as noted above.

We give an example of this computation with a very small database. The space is built from a set of 5 records. In this way, we can consider the whole database space that has a size of $2^5$ databases.

***Example 1*** Let $DB$ be the set of all databases that can be generated from 5 records. That is, $DB$ corresponds to the power set of these 5 records. Let $P$ be the transition matrix of $DB$ defined according to Definition 5 using $p_a = p_d = 0.5$. Then, we can compute $F_{13}$ according to Definition 6. $F_{13}$ is the distance between databases $DB_1 = \{a\}$ and $DB_3 = \{c\}$. Figure 1 represents this computation.

### 3.3 Results on the approximation of distance distribution functions

Computation of the distance introduced in Definition 6 is costly. Because of that we are interested in the approximation of this distance. We can prove the following results in relation to Definition 6.

**Lemma 2** *Let $DB_1$ and $DB_2$ be two databases. Let us consider the sets $DB_1 \cap DB_2$, $DB_1 \setminus DB_2$, and $DB_2 \setminus DB_1$. Let $t_a = |DB_2 \setminus DB_1|$ be the number of elements we need to add to transit from $DB_1$ to $DB_2$, and let $t_d = |DB_1 \setminus DB_2|$ be the elements we need to delete to transit from $DB_1$ to $DB_2$. Then, the shortest chain from $DB_1$ to $DB_2$ when we only consider addition and deletion of records has length*

$$t_0 = |DB_1 \setminus DB_2| + |DB_2 \setminus DB_1| = t_a + t_d$$

*and, therefore,*

$$F_{12}(t) = 0$$

*for all $t < t_0$.*

Let us consider an arbitrary order for the $t_a$ elements we add, and an arbitrary order for the $t_d$ elements we remove. Let $i$ in $\{1, \ldots, t_a\}$ represent the addition of the $i$th element according to this order and $i$ in $\{t_a + 1, \ldots, t_a + t_d\}$ the removal of the $(i - t_a)$th element according to this order. Using this interpretation, it is clear that any permutation of $\{1, \ldots, t_a + t_d\}$ represents a valid chain with only additions and deletions and with no cycles from $DB_1$ to $DB_2$. So, there are $(t_a + t_d)!$ valid chains with no cycles.

We can also prove a lemma similar to Lemma 2 when in addition to addition and deletion we allow transitions that do not change the database. Shortest chains will of course still have length $t_a + t_d$, and from this it also follows: $F_{12}(t) = 0$ for all $t < t_0$.

Let $\mathcal{C}_{12}^t$ denote all valid chains from $DB_1$ to $DB_2$ with length $t$. Then, $\mathcal{C}_{12}^{t_a+t_d}$ will represent all shortest chains. Therefore, $|\mathcal{C}_{12}^{t_a+t_d}| = (t_a + t_d)!$. It is also clear that $\mathcal{C}_{12}^t = \emptyset$ for $t < t_a + t_d$.

Let use denote a chain $c \in \mathcal{C}^t$ by $c_0, c_1, \ldots, c_t$. Here $c_i$ will correspond to a database $DB_i$. Then, the probability of transiting from $DB_0$ to $DB_t$ through the chain $c$ is naturally

$$P_c = \prod_{c_r \in c} P_{c_{r-1}, c_r}. \tag{2}$$

**Lemma 3** *Using the notation in Lemma 2 and $P_c$ as in Equation 2, we have that when only addition and deletion are allowed, or when addition, deletion and transition without change are allowed, it holds (for $t_a$ and $t_b$ as above)*

$$\begin{aligned} F_{12}(t_a + t_d) = f_{12}^{t_a+t_d} &= \sum_{c \in \mathcal{C}_{12}^{t_a+t_d}} P_c \\ &= \sum_{c \in \mathcal{C}_{12}^{t_a+t_d}} \prod_{c_r \in c} P_{c_{r-1}, c_r}. \end{aligned} \tag{3}$$

We can also prove that when only addition and deletions are allowed, $\mathcal{C}_{12}^t = \emptyset$ for any $t = t_a + t_d + 1 + 2k$ for any $k$, i.e., given a shortest chain, we can only enlarge this chain both adding and removing $k$ records. So, in this case, for all $k$ it holds $f_{12}^{t_a+t_d+2k+1} = 0$.

When addition, deletion and also non modification are allowed as transitions between databases, we have that for $t = t_a + t_d + 1 + 2k$ (for any $k$), the chains in $\mathcal{C}_{12}^t$ are the ones in $\mathcal{C}_{12}^{t-1}$ adding a transition corresponding to non-modification (say $t_{nm}$). Let $p_\emptyset$ denote the probability of a non-modification (this corresponds to the value $1 - (p_d + p_a)$ in Equation 1). Let $c \in \mathcal{C}^t$ be one of such chains with elements $c_0, c_1, \ldots, c_t$. Then, we can insert this $t_{nm}$ transition between any pair of elements of the chain (but not at the end of the chain as we are considering that is exactly at time $s$ that we reach the goal state). This means that there are $t$ options. Given $P(c)$, the probability of the chain $c \in C^t$, the probability of any of these chains is $p_\emptyset \cdot p(c)$. So, as we have $t$ new chains for a given chain $c$, the probability for this set of chains (say $\tilde{c}$) is $p(\tilde{c}) = tp_\emptyset p(c)$. Then, considering all $\tilde{c}$ generated from all $c \in \mathcal{C}^t$, we have that for $t = t_a + t_d + 1 + 2k$

$$P(\mathcal{C}_{12}^t) = \sum_{c \in \mathcal{C}_{12}^{t-1}} p(\tilde{c}) = \sum_{c \in \mathcal{C}_{12}^{t-1}} tp_\emptyset p(c)$$
$$= tp_\emptyset \sum_{c \in \mathcal{C}_{12}^{t-1}} p(c) = tp_\emptyset P(\mathcal{C}_{12}^{t-1}).$$

Using Expression 3, it is easy to prove the following lemma.

**Lemma 4** *Let $DB_i$ and $DB_j$ be two arbitrary databases, and let $\mathcal{R} = \{c\}_c$ be a set of random valid chains $c \in \mathcal{R}$ from $DB_i$ to $DB_j$ with different lengths. Then,*

$$f_{ij}^s \geq \sum_{c:|c|=s+1} P_c$$

*and*

$$F_{ij}(t) \geq \sum_{c:|c|\leq t+1} P_c.$$

This result implies that the consideration of random valid chains give lower bounds for $F_{ij}(t)$. Therefore, any decision based on a threshold $th$ on a given $t$ (i.e., $F_{ij}(t) \geq th$) valid for a set $\mathcal{R}$ will be also valid if all the set of chains is considered.

The links between triangle functions and t-norms (see e.g., [1], and Def. 7.1.3 and Section 7.1 in [9]) permit us to establish the following lemma. This lemma establishes another lower bound for distance distribution functions for any pair of databases if we can compute exact values for pairs involving a particular database (e.g., a reference one denoted by $DB_a$ below).

**Lemma 5** *Let $(S, \mathcal{F}, \tau_T)$ be a probabilistic metric space generated by a t-norm $T$ (i.e., $\tau_T(F, G)(x) = T(F(x), G(x))$ is the triangle function generated by $T$). Let $DB_a$ be a particular database for which we can calculate exactly the distance distribution function for all $DB_i \in S$. Then,*

$$\mathcal{F}(DB_i, DB_j) \geq T(F_{DB_i,DB_a}(x), F_{DB_a,DB_j}(x))$$

*is a lower bound of $\mathcal{F}_{DB_i,DB_j}$.*

***Proof*** If $T$ is a t-norm, then the triangle function $\tau_T$ generated by $T$ is

$$\tau_T(F, G)(x) = T(F(x), G(x))$$

for distance distribution functions $F$ and $G$. Then, as $(S, \mathcal{F}, \tau_T)$ is a probabilistic metric space on the space of databases, we know that for all $p$, $q$, and $r$ it holds that

$$\mathcal{F}(p, r) \geq \tau_T(\mathcal{F}(p, q), \mathcal{F}(q, r))$$

and, therefore, in particular for $p = DB_i$, $q = DB_a$ and $r = DB_j$ we have that

$$\mathcal{F}(DB_i, DB_j) \geq \tau_T(\mathcal{F}(DB_i, DB_a), \mathcal{F}(DB_a, DB_j))$$
$$= T(\mathcal{F}(DB_i, DB_a), \mathcal{F}(DB_a, DB_j)).$$

□

### 3.4 Database distance-based probabilistic metric space

We consider an alternative way to define probabilistic metric spaces in which in addition to a transition matrix we use a distance on the state space. The definition is based on [5].

**Definition 7** Let $S$ be the database space, let $P$ be the transition matrix for $S$ that defines a time-homogeneous Markov chain $(Z_n)_{n\in\mathbb{N}}$. Let $d : S \times S \to \mathbb{R}^+$ be a distance function on $S$. Then, for any given time $t \geq 0$, we define the function $F_{ij}^t(x)$ as follows:

$$F_{ij}^t(x) = Pr[d(i, j) < x \text{ at time } t]$$
$$= \sum_{k \in S} P_{ik}^t \left( \sum_{\ell : d(\ell,k) < x} P_{j\ell}^t \right).$$

Informally, for a given time $t$, this definition implies that the probability level between states $i$ and $j$ at time $x$ is computed in terms of the probability of reaching states $k$ at time $t$ from $i$ and the probability of finding states $\ell$ from $j$ at most at distance $x$.

We can prove the following result that is similar to Lemma 4.

**Lemma 6** *Let $\mathcal{DB}$ be a collection of databases sampled from the space of databases. Let*

$$\tilde{F}^t_{ij}(x;\mathcal{DB}) = \sum_{DB_k \in \mathcal{DB}} P^t_{ik} \sum_{\ell} : d(\ell,k) < x DB_\ell \in \mathcal{DB} P^t_{j\ell}.$$

*Let $\tilde{F}^t_{ij}(x;\mathcal{DB},\mathcal{R})$ correspond to $\tilde{F}^t_{ij}(x;\mathcal{DB})$ when $P^t_{ik}$ only considers a given set $\mathcal{R}$ of random valid chains as in Lemma 4. Then,*

$$\begin{aligned}\tilde{F}^t_{ij}(x;\mathcal{DB},\mathcal{R}) &\leq \tilde{F}^t_{ij}(x;\mathcal{DB}) \\ &\leq \sum_{DB_k \in \mathcal{DB}} P^t_{ik} \sum_{\ell : d(\ell,k) < x, DB_\ell \in \mathcal{DB}} P^t_{j\ell}.\end{aligned}$$

The implications of this lemma are similar to the ones of Lemma 4. That is, the consideration of random chains and sets of databases give lower bounds for $F_{ij}(t)$. Therefore, any decision based on considering two databases $DB_1$ and $DB_2$ as different based on a threshold $th$ on a given $t$ (i.e., $F_{12}(t) \geq th$) valid for a set $\mathcal{R}$ and for a set of databases $\mathcal{DB}$ will be also valid if all the sets of chains and all databases are also considered. We illustrate this distance with one example that uses the same space of databases we have considered before.

***Example 2*** Let $P$ be the transition matrix and let $DB$ be the space of databases considered in Example 1. Let $DB_1$ and $DB_3$ be the databases considered in Example 1. We can compute $F_{13}$ according to Definition 7. To do so, we use the Jaccard Index to measure the similarities between the databases. Figures 2, 3 and 4 represent these computations with different values of $t$. We display $F_{1,3}$ as in the previous example, but we also considered the computations for very different databases (i.e., $F_{0,31}$) in Figures 5, 6 and 7. Here, $F_{0,31}$ is the distance between databases $DB_0 = \{\}$ and $DB_{31} = \{a, b, c, d, e\}$.

From the three figures, Figs. 2, 3 and 4, we notice that when $t$ becomes larger, and $x$ is greater than 0.5, the probability become higher.

Figures 5, 6 and 7) show that when we have very different databases, and the Jaccard distance is less than 0.5, we need more transitions in order to increase the probability.

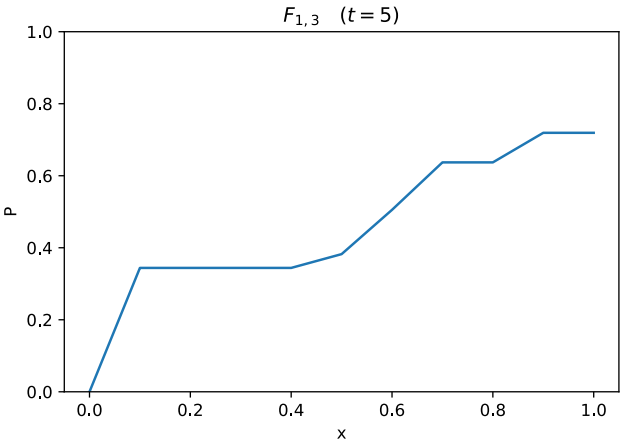


**Fig. 2** $F_{13}$ when $t = 5$ for Example 2 according to Definition 7

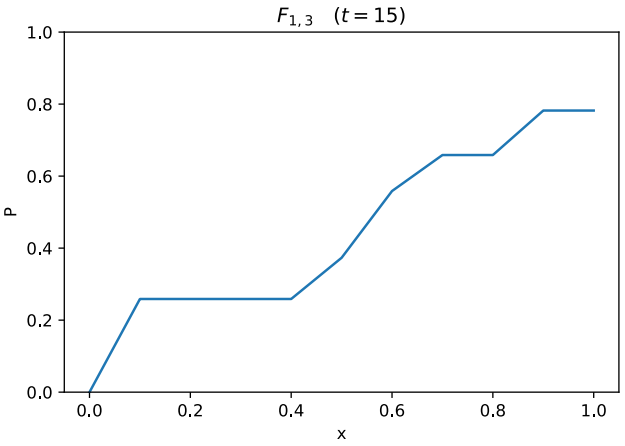


**Fig. 3** $F_{13}$ when $t = 15$ for Example 2 according to Definition 7

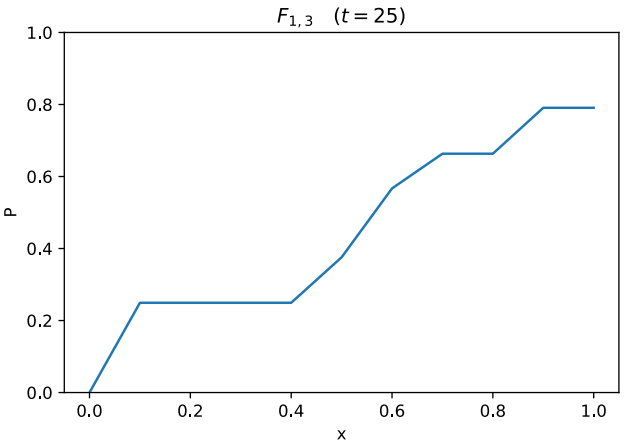


**Fig. 4** $F_{13}$ when $t = 25$ for Example 2 according to Definition 7

## 4 Construction of the distance on the space of models

The definitions above permit to extend the probabilistic metric space for databases to models. As discussed in Sect. 2.3, given two models $m_1$ and $m_2$ the goal is to define a distance based on the generators $G_{m_1}$ and $G_{m_2}$ of $m_1$ and $m_2$. In this

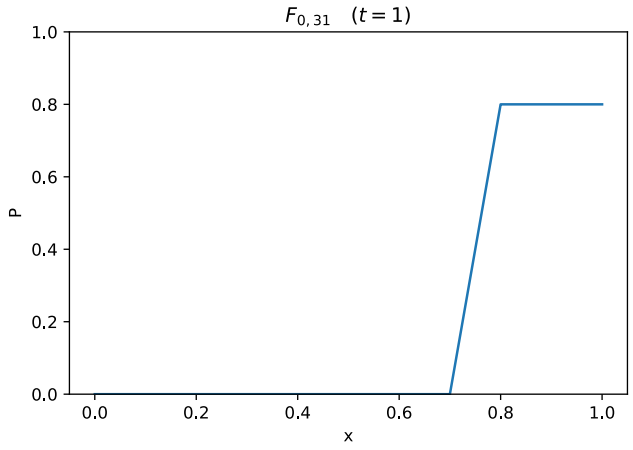


**Fig. 5** $F_{0,31}$ when $t = 1$ for Example 2 according to Definition 7

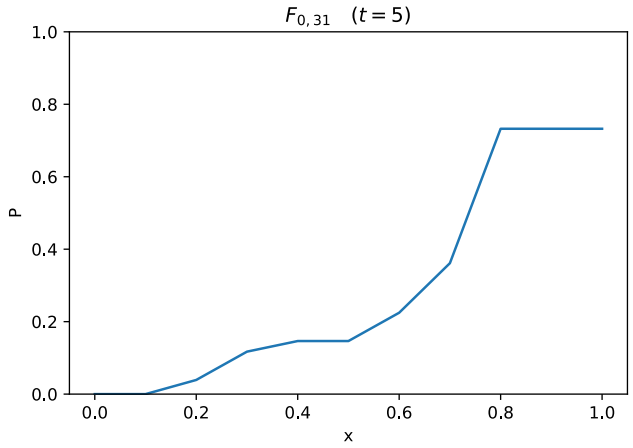


**Fig. 6** $F_{0,31}$ when $t = 5$ for Example 2 according to Definition 7

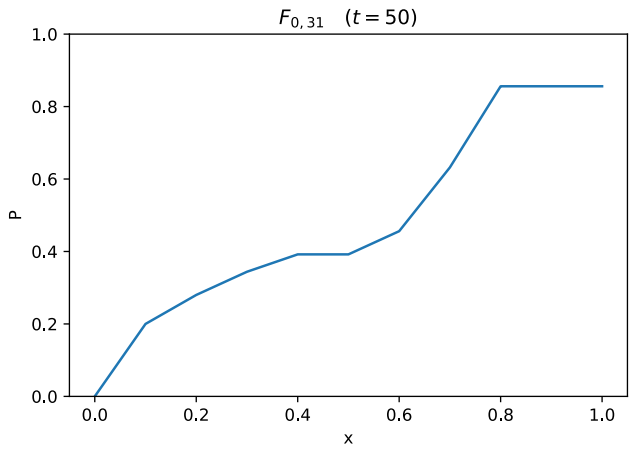


**Fig. 7** $F_{0,31}$ when $t = 50$ for Example 2 according to Definition 7

**Table 1** Set of models and their corresponding sets of Databases from Example 3

| Model $m$ | Gen($m$) |
|---|---|
| 1000 | $(a)$, $(b)$, $(a, b)$, $(a, b, c)$. |
| 1500 | $(a, c)$, $(b, c)$ |
| 2000 | $(c)$ |

case, instead of a standard distance, we consider a distance distribution function.

**Proposition 1** *Let $S$ be the space of databases, let $G$ be an algorithm to generate models from the space of databases $S$, let $\mathcal{G}$ be the space of models that can be generated by $G$. Let $m_1$ and $m_2$ be two models generated by the application of algorithm $G$ to databases in $S$. Let $Gen_{m_1}$ and $Gen_{m_2}$ be the set of databases that generate $m_1$ and $m_2$. That is, $Gen_{m_1} = \{DB \in S | G(DB) = m_1\}$ and $Gen_{m_2} = \{DB \in S | G(DB) = m_2\}$.*

*Let $(S, \mathcal{F})$ be a probabilistic metric space. Then, let $\mathcal{F}$ for pairs of models $m_1$ and $m_2$ be defined as follows:*

$$\mathcal{F}(m_1, m_2)(x) = \frac{1}{|Gen_{m_1}||Gen_{m_2}|} \sum_{DB_1 \in Gen_{m_1}} \sum_{DB_2 \in Gen_{m_2}} \mathcal{F}_{DB_1, DB_2}(x). \tag{4}$$

*Then, $\mathcal{F}$ is a distance distribution function.*

**Lemma 7** *Using Equation 4 with Definition 6, we obtain a function $\mathcal{F}(m_1, m_2)$ that is a non-symmetric distance distribution function. Using instead definitions $F'$ and $F''$ above will lead to symmetric distance functions. Using Definition 7 results into a distance distribution function.*

**Lemma 8** *When we approximate $\mathcal{F}(m_1, m_2)(x)$ using lower bounds of $\mathcal{F}_{DB_1, DB_2}$ (as considering only some chains and some databases), we will obtain lower bounds of the real $\mathcal{F}(m_1, m_2)(x)$.*

It is relevant to point out that this definition does not necessarily lead to a probabilistic metric space, as condition (iv) in Definition 4 does not always hold. We illustrate this definition above considering databases as in the previous examples based on three records/people and their salaries.

***Example 3*** Suppose we have three records $a$, $b$, and $c$, with salaries 1000, 1000 and 2000, respectively. The space of databases is the power set of theses records. If we choose $G$ to be the median function to generate the models, then the space of models is $\mathcal{G} = \{1000, 1500, 2000\}$. The models and their generators are listed in Table 1. Figure 8 displays the distance between model $m_1$=1000 and model $m_2$=2000 by

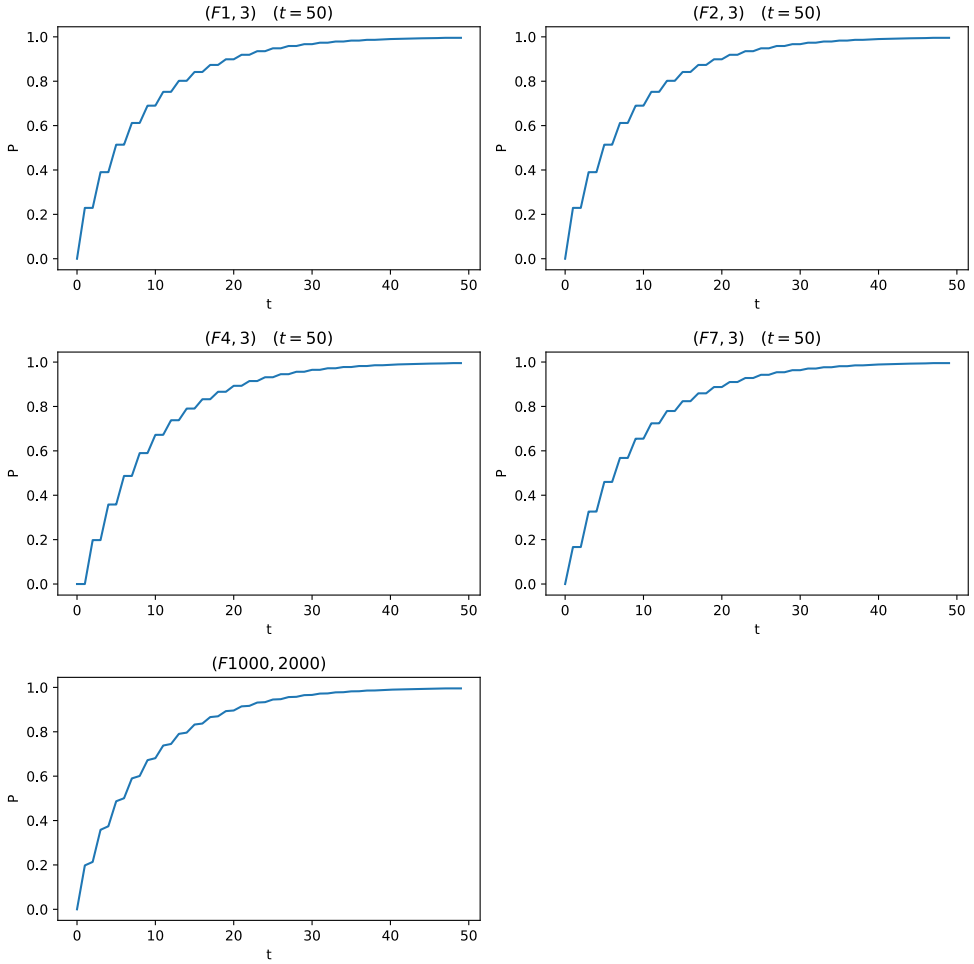


**Fig. 8** $F_{1000,2000}$ for Example 3 according to Proposition 1 and Definition 6

using Proposition 1, as well as the distance between pairs of databases according to Definition 6. Similarly, we have also computed the distance between the same pair by using the same proposition, but where the distance between databases follows Definition 7 as illustrated in Fig. 9.

From both Fig. 8, and Fig. 9, we can see that the distance distribution functions for the models and the databases are quite similar.

## 5 Summary and conclusions

In this paper, we have proposed the use of Markov chains and transition matrices to model transitions between databases, and used them to define a probabilistic metric space for models.

Our goal is to better understand the relationship between data and models. From our perspective, this requires a metric space on the space of models that reflects the relationships between the databases that can generate these models. From a machine learning perspective, a good model is one that has a good accuracy, but also that is not overfitted to data and has some level of generalization. From a data privacy perspective, a good model is one that does not lead to disclosure. This includes not leading to disclosure on the data that has been used to generate the model. In other words, we understand machine and statistical learning as a selection process. We want to select a model with good accuracy that does not have overfitting (and not vulnerable to membership attacks) and that is *near* to models with similar generators. This work is to formalize what *near* means.

As future work, we plan to develop strategies for computing these distances and for defining in practice metric spaces for real-size databases. In this paper, examples have been described for small databases because they are easier to understand but also because when considering a regular size database its power set becomes extremely large. Some initial results on boundary conditions on the distances were given in the paper. We plan to consider how to extend this approach by means of approximating the distances.

We also plan to work on the problem of model selection. Research on graphical visualization of the models and the metric spaces will be appropriate here. Sammon's map as

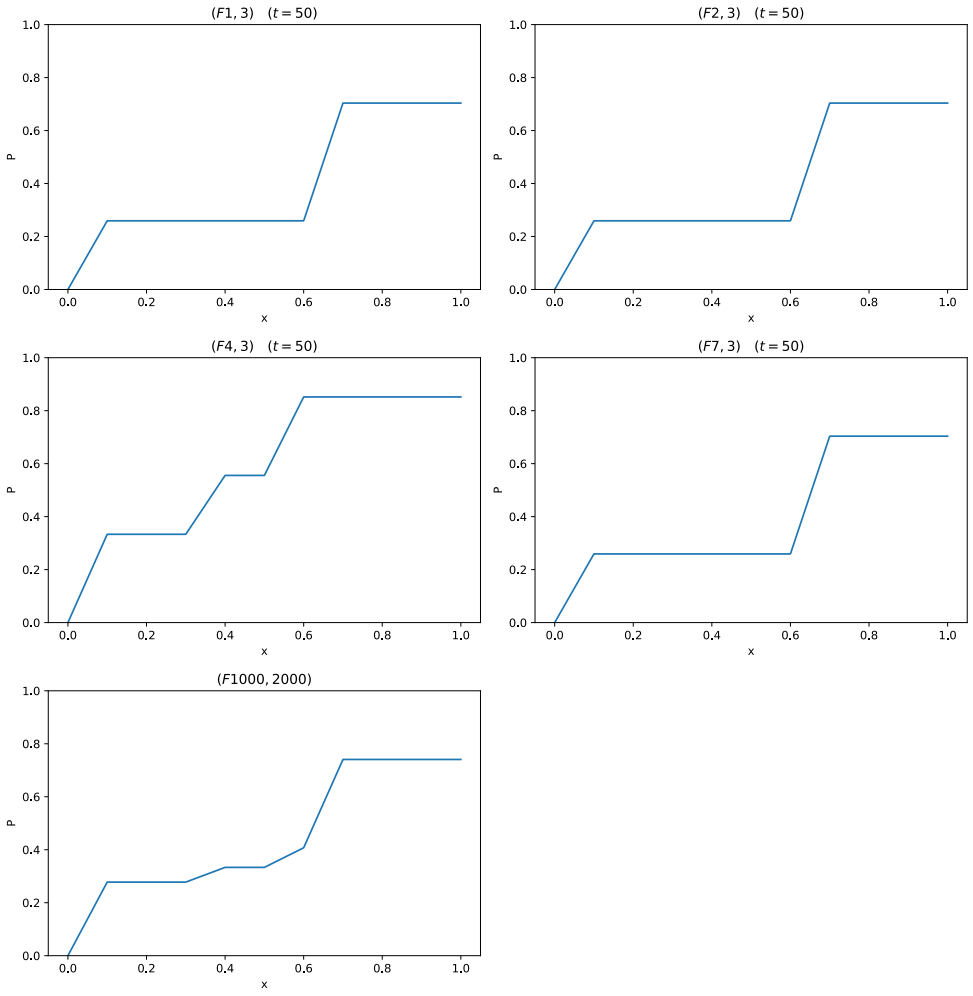


**Fig. 9** $F_{1000,2000}$ for Example 3 according to Proposition 1 and Definition 7

well as other multidimensional scaling procedures can help on this purpose.

**Acknowledgements** This study was partially funded by Vetenskapsrådet project "Disclosure risk and transparency in big data privacy" (VR 2016-03346, 2017-2020), Spanish project TIN2017-87211-R is gratefully acknowledged, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

**Funding** Open access funding provided by Umea University.



## References

1. Alsina, C., Frank, M.J., Schweizer, B.: Associative Functions: Triangular Norms and Copulas. World Scientific, Singapore (2006)
2. Doob, J.L.: Stochastic Processes. Wiley, Hoboken (1953)
3. Eiter, T., Mannila, H.: Distance measures for point sets and their computation. Acta Informatica **34**, 109–133 (1997)
4. Kent, D.C., Richardson, G.D.: Ordered probabilistic metric spaces. J. Austral. Math. Soc. **46**, 88–99 (1989)
5. Marcus, P.S.: Probabilistic metric spaces constructed from stationary Markov chains. Aequationes Mathematica **15**, 169–171 (1977)
6. Moynihan, R.: Probabilistic metric spaces induced by Markov chains. Z. Wahrscheinlichkeitstheorie. Gebiete **35**, 177–187 (1976)
7. Privault, N.: Understanding Markov Chains. Springer, Newyork (2018)
8. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
9. Schweizer, B., Sklar, A.: Probabilistic Metric Spaces. Elsevier, Amsterdam (1983)
10. Senavirathne, N., Torra, V.: Approximating robust linear regression with an integral privacy guarantee. Proc. PST **2018**, 1–10 (2018)
11. Senavirathne, N., Torra, V.: Integral privacy compliant statistics computation. Proc DPM/CBT- ESORICS **2019**, 22–38 (2019)
12. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. Proc. IEEE Symposium on Security and Privacy. (2017). arXiv:1610.05820

13. Torra, V.: Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer, Newyork (2017)
14. Torra, V., Navarro-Arribas, G.: Probabilistic metric spaces for privacy by design machine learning algorithms: modeling database changes, Proc. DPM 2018/CBT 2018. LNCS **11025**, 422–430 (2018)