Actes de la 3ème Journée EGC & IA

« Données personnelles, vie privée et éthique »

Organisée par

Jérôme Azé et Thierry Charnois

23 Janvier 2018
Paris, France

Comités

Comité d'organisation

- Jérôme Azé, LIRMM UM-CNRS
- Thierry Charnois, Université Paris 13/LIPN

Comité de programme

Présidents : Jérôme Azé, LIRMM UM-CNRS et Thierry Charnois, Université Paris 13/LIPN

- Maxime Amblard, LORIA
- Jean-Yves Antoine, Université de Tours
- Annie Blandin, IMT Atlantique
- Gauvain Bourgne, CNRS & Sorbonnes Universités, UPMC Paris 06, LIP6
- Nora Cuppens-Boulahia, IMT Atlantique
- Gaël Dechalendar, CEA
- Karën Fort, Paris Sorbonne
- Philippe Lenca, IMT Atlantique
- Fabrice Muhlenbach, Université Jean Monnet
- Guillaume Piolle, Centrale Supelec

Avant-propos

La sélection d'articles publiés dans ce recueil constitue les actes de la 3ème Journée thématique EGC&IA qui s'est déroulé le 23 janvier 2018 à Paris. La thématique de cette 3ème journée était : "Données personnelles, vie privée et éthique".

L'objectif de cette journée EGC&IA était d'encourager les discussions et les recherches fondées sur des principes qui conduiront à l'avancement de l'analyse des données personnelles, du développement des services personnels, de la protection de la vie privée, du respect de l'éthique, de la protection des données et de l'évaluation des risques liés à la vie privée.

À l'ère de Big Data, chaque utilisateur de notre monde hyper-connecté laisse derrière lui une myriade de traces numériques tout en effectuant ses activités quotidiennes. Ces masses de données sont majoritairement exploitées par de grandes compagnies afin d'en extraire de précieuses informations permettant de modéliser le comportement humain et d'améliorer des stratégies de marketing. Chaque compagnie n'a qu'une vision limitée des données personnelles disponibles, uniquement celles collectées via les services qu'elle propose. Et surtout, chaque utilisateur a un contrôle assez limité sur les données réellement collectées, et sur l'usage qu'il pourrait en faire. Des solutions telles que les "Personal Data Store" ou les "Information Management System" apparaissent. L'objectif est de permettre aux utilisateurs de centraliser des données fréquemment utilisées par des sites marchands ou autres et ainsi de faciliter les actions de l'utilisateur lorsqu'il a besoin de transmettre ces informations en lui évitant de devoir les ressaisir (par exemple). Il y a cependant un manque réel d'algorithmes et de modèles dédiés pour le traitement des données personnelles afin d'y détecter des comportements et d'en extraire de la connaissance, tout en garantissant les aspects liés à la protection de la vie privée et les aspects liés à l'éthique dans le cas de scénario d'utilisation centré sur l'utilisateur.

Aujourd'hui, l'analyse des données personnelles, la protection individuelle de la vie privée et le respect de l'éthique sont les éléments clés pour tirer parti des services pour un nouveau type de systèmes. La disponibilité d'outils d'analyse personnelle capables d'extraire des connaissances cachées à partir de données individuelles tout en protégeant le droit à la vie privée et en respectant l'éthique peut aider la société à passer de systèmes centrés sur l'organisation à des systèmes centrés sur l'utilisateur, où l'utilisateur est le propriétaire de ses données personnelles et peut gérer, comprendre, exploiter, contrôler et partager ses propres données et les connaissances pouvant en être extraites de manière complètement sûre.

Cette problématique trouve un écho particulier dans le cadre du nouveau Règlement européen sur la Protection des Données Personnelles (RPDG) paru au journal officiel de l'Union européenne et entrant en application le 25 mai 2018. Cette réglementation renforce les droits des personnes, notamment par la création d'un droit à la portabilité des données personnelles, et responsabilise les acteurs traitant des données (responsables de traitement et sous-traitants). (https ://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels)

Jérôme Azé et Thierry Charnois
Présidents du comité de programme

Sommaire

Journée EGC&IA

A Declarative Approach to Linked Data Anonymization (Position Paper)

Angela Bonifati*, Rémy Delanaux*, Marie-Christine Rousset**,*** Romuald Thion*

*Université Lyon 1, LIRIS CNRS, 69100 Villeurbanne, France
**Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
***Institut Universitaire de France

Résumé. We present a declarative approach for privacy-preserving data publishing in Linked Open Data, in which privacy and utility policies along with basic data transformation operators are encoded as SPARQL queries. We believe that our approach can be useful in many settings in which data providers are responsible for the non-disclosure of information that could serve as quasi-identifiers whenever crossed with external data sources. Data producers need their own means to specify the privacy policies they want to enforce on their datasets to be published, but also to specify the utility policies that tell which parts of the data should be kept. They also need tools for computing the anonymization operations to apply to their datasets prior to their publication with the guarantee that the specified privacy and utility policies are verified by the resulting dataset. Our approach is designed to meet these requirements by leveraging the expressive power of SPARQL queries and the effectiveness of SPARQL query engines.

1 Introduction

Linked Open Data (LOD) provides access to continuously increasing amounts of RDF data that describe properties and links among entities referenced by means of Uniform Resource Identifiers (URIs). LOD is a worldwide effort to collect data that can be reused for free [1]. Whereas many organizations, institutions and governments participate to the LOD movement by making their data accessible and reusable to citizens, the risks of identity disclosure in this process are not completely understood. For example, in smart city applications, information about the trajectories of people in public transportation can help re-identify the individuals if they are joined with other public data sources by leveraging quasi-identifiers.

For all these reasons, data producers should have at their disposal the means to readily anonymize their data before exposure into the LOD cloud. The solutions proposed so far are mainly devoted to legacy relational database systems and thus not applicable to the context of RDF data. Such solutions rely on variants of differential privacy as surveyed in Machanavajjhala et al. (2017) or k-anonymity proposed by Sweeney (2002). Whereas differential privacy offers strong mathematical guarantees of non disclosure of any factual information by adding

1. "Linked Open Data (LOD) is Linked Data which is released under an open license, which does not impede its reuse for free". Tim Berners-Lee, Design Issues, W3C.

noise to the data, it suffers from the shortcoming that the obtained linked data as returned by the queries exhibits low utility. Several k-anonymization methods have been developed that transform the original dataset into clusters containing at least k records with similar values over quasi-identifiers. When taken into account, the utility loss is defined and minimized for a specific need depending on the application. Logical frameworks for Linked Data publishing were briefly studied, introducing the query-based policy system (Grau et Kostylev, 2016) but were not investigated further than logical foundations and theoretical problems.

In this paper, we present a novel declarative approach that (1) enables the data producers to specify as queries both the privacy and utility policies s/he wants to enforce on her/his RDF dataset, (2) checks whether they are compatible, and (3) based on a set of operators for basic data transformations, builds possibly multiple anonymization procedures satisfying the required privacy and utility policies. These anonymization procedures are then compared in terms of information loss and proposed to the data producers who may eventually choose which anonymization procedure s/he prefers to apply. In that respect, we believe that our approach paves the way to a democratization of privacy-preserving mechanisms.

In the following, we present the main ingredients of this approach and we illustrate it with some examples representative of use cases on urban data. Other application domains can be envisioned, such as healthcare data and social data among the others. We omit their illustration for space reasons.

2 Preliminaries

Let **I**, **L** and **B** be countably infinite pairwise disjoint sets representing respectively *IRIs*, *literal values* (or *'literals'*) and *blank nodes*. IRIs (Internationalized Resource Identifiers) are standard identifiers used for denoting any Web resource described in RDF within the LOD.

Definition 1 (RDF graph) *An **RDF graph** is a set of RDF triples* (s,p,o)*, where* $(s,p,o) \in (\boldsymbol{I} \cup \boldsymbol{B}) \times \boldsymbol{I} \times (\boldsymbol{I} \cup \boldsymbol{L} \cup \boldsymbol{B})$.

In our context, we refer to an RDF graph as an RDF dataset, or simply a dataset.

We denote by T the union $\mathbf{I} \cup \mathbf{L} \cup \mathbf{B}$. We also assume an infinite set V of variables disjoint from the above sets.

Definition 2 (Graph pattern) *A **triple pattern** is a tuple from* $(\boldsymbol{T} \cup \boldsymbol{V}) \times (\boldsymbol{I} \cup \boldsymbol{V}) \times (\boldsymbol{T} \cup \boldsymbol{V})$*. A **graph pattern** GP is a conjunction of triple patterns.*

We can now define the two types of queries that we consider and their answers. The first type of queries corresponds to the standard notion of *conjunctive queries*, while the second type corresponds to *counting queries* that are the basis for computing statistics on data.

Definition 3 (Conjunctive query) *A **conjunctive query** q is defined by an expression :*

SELECT $\bar{x}$ *WHERE* $GP(\bar{x}, \bar{y})$ *FILTER C*

where $\bar{x}$ *is a set of variables and* $GP(\bar{x}, \bar{y})$ *is a graph pattern possibly constrained by a FILTER condition C. Among the set* $\bar{x} \cup \bar{y}$ *of variables appearing in the query body, the variables in* $\bar{x}$ *are the result (also called distinguished) variables.*

Its evaluation over a RDF graph G consists in finding mappings θ from $\bar{x} \cup \bar{y}$ to T such that $\theta.GP(\bar{x}, \bar{y})$ is a subgraph of G and $\theta.C$ is true. The answer set of q over G is defined by :

$$\mathsf{Answer}(q, G) = \{\theta.\bar{x} | \theta.GP(\bar{x}, \bar{y}) \subseteq G \ \wedge \ \theta.C\}$$

As for the FILTER conditions, we adopt the same fragment of Pérez et al. (2009), by restricting ourselves to the built-in condition as a boolean combination of terms constructed by using $=$ and bound, and boolean connectives of the kind $\neg$, $\wedge$, $\vee$ between the built-in conditions.

Definition 4 (Counting query) *Let q be a conjunctive query. The query $\mathsf{Count}(q)$ is a **counting query**, whose answer over a graph G is defined by :*

$$\mathsf{Answer}(\mathsf{Count}(q), G) = |\mathsf{Answer}(q, G)|$$

3 Query-based specification of privacy and utility policies

Data producers can parametrize the desired properties of the anonymization process to apply to their dataset, by specifying by means of queries (i) the data for which the anonymization process must guarantee the non disclosure, and (ii) the data or results that have to remain unchanged after the application of the anonymization process. This formalism is based on the notion of privacy policy designed in (Grau et Kostylev, 2016), extends it, and situate it in a concrete application framework.

Intuitively, a couple of privacy and utility policies is satisfied by an anonymization process if it alters the dataset in a way such that all the privacy queries return no answer but all the utility queries still return exactly the same answers as before.

Definition 5 (Privacy and utility policies) *Let G be an input RDF graph, a* privacy *(resp.* utility*)* policy $\mathcal{P}$ *(resp. $\mathcal{U}$) is a set of conjunctive queries (resp. conjunctive or counting queries). Let $Anonym(G)$ be the result of an anonymization process of the graph G by a sequence of anomymisation operators.*

- *A privacy policy $\mathcal{P}$ is* satisfied *on $Anonym(G)$ if for every $p \in \mathcal{P}$:* $\mathsf{Answer}(p, Anonym(G)) = \emptyset$.
- *A utility policy $\mathcal{U}$ is* satisfied *on $Anonym(G)$ if for every $u \in \mathcal{U}$:* $\mathsf{Answer}(u, Anonym(G)) = \mathsf{Answer}(u, G)$.

Example 1 *Consider a privacy policy P_1 on data related to public transportation in a given city, and defined by the two following conjunctive queries The first privacy query expresses that users' names are sensitive and shall be protected, and the second privacy query specifies that the disclosure of users identifiers associated with location information (like latitude and longitude as given by the user ticket validation) may also pose a risk (for re-identification).*

```
SELECT ?name
WHERE {
        ?c      a       tcl:Validation;
                tcl:user        ?u.
        ?u      a       tcl:User;
                foaf:familyName ?name.
```

```
}


SELECT ?u ?lat ?long
WHERE {
        ?c      a       tcl:Validation;
                tcl:user        ?u;
                geo:latitude    ?lat;
                geo:longitude   ?long.
}
```

As a consequence, any query displaying either users' names or users' identifiers together with their geolocation information would infringe this privacy policy, violating the anonymization of the underlying dataset to be published as open data.

Using the following counting query, an utility policy U_1 would require that the number of validations of users' tickets must be preserved.

```
SELECT (COUNT(DISTINCT ?c) AS ?count)
WHERE {
        ?c      a       tcl:Validation;
                tcl:user        ?u.
}
```

This simple example shows that privacy and utility policies might impose constraints on overlapping portions of the dataset, such as the users of a public transportation system.

4 Query-based specification of anonymization operators

In our setting, we consider an anonymization process as a sequence of atomic anonymization operations applied to an input RDF graph in order to transform it in a new RDF graph containing less precise information on individuals. In this section, we just explain the atomic operators that we consider to transform graphs. In the next section, we will define how to check *a priori* or *a posteriori* whether the transformed graph satisfies the privacy and utility policies.

We consider four types of atomic anonymization operations on RDF graphs : deletion of triples, value replacement in triples involving datatype properties, IRI replacement in triples involving object properties and IRI aggregation. IRI aggregation consists in grouping IRIs of the same type with same properties in groups that are represented by blank nodes in which the specific properties values that we want to hide are replaced by their count for each group.

The point is that each of these operations can be defined with (possibly complex) queries by leveraging SPARQL 1.1 aggregate and update queries as well as calls to built-in functions.

Due to space constraints, we only report the definition of *deletion queries*. They specify as a graph pattern (denoted $GPD(\bar{x})$ in the following definition) which triples to remove from the RDF graphs G on which they are evaluated. This removal is conditioned by the existence in G of a graph pattern (denoted $GPW(\bar{x}, \bar{y})$) possibly constrained by a FILTER condition (denoted C). The evaluation of such a deletion query will remove from G all the triples in $\theta.GPD(\bar{x})$ where θ is a mapping of the variables $\bar{x} \cup \bar{y}$ such that $\theta.GPW(\bar{x}, \bar{y})$ is a subgraph of G and $\theta.C$ is true.

Definition 6 (Deletion query) *A **deletion query** q_{delete} is defined by :*
DELETE $GPD(\bar{x})$ WHERE $GPW(\bar{x}, \bar{y})$ FILTER C
where $GPD(\bar{x})$ and $GPW(\bar{x}, \bar{y})$ are graph patterns.
The result of its evaluation over a RDF graph G is defined by :

$$\mathsf{Result}(q_{delete}, G) = G \setminus \{\theta.GPD(\bar{x}) | \theta.GPW(\bar{x}, \bar{y}) \subseteq G \wedge \theta.C\}$$

Example 2 *In the spirit of Example 1 related to transportation data, the following query specifies the operation leading to the deletion of the family names of users.*

```
DELETE {?u      foaf:familyName ?name.}
WHERE { ?u      a       tcl:User;
                foaf:familyName ?name. }
```

The other three types of operations can also be specified as queries. For instance, instead of deleting the triples involving the datatype property `foaf:familyName`, we could opt for the value replacement of family names using a built-in cryptographic one-way function $crypto$ on strings, and specify it by the following DELETE-INSERT query :

```
DELETE {?u      foaf:familyName ?name.}
INSERT {?u      foaf:familyName crypto(?name).}
WHERE { ?u      a       tcl:User;
                foaf:familyName ?name. }
```

5 Towards a full-fledged anonymization system

In this section, we formalize several problems of interest of the envisioned anonymization system by leveraging declarative privacy and utility policies defined above.

Let $\mathcal{P}$ be a privacy policy and $\mathcal{U}$ be an utility policy.

The first problem of interest consists in verifying whether the privacy and utility policies specified by a data provider are incompatible. If it is the case, no anonymization process will satisfy both the privacy and utility policies. It is useful to identify such cases to explain to the data provider that he should restrict the utility or relax the privacy constraints.

Incompatible$(\mathcal{P}, \mathcal{U})$;
Input : A set $\mathcal{P}$ of privacy queries and a set $\mathcal{U}$ of utility queries
Output: `True` if and only if for every graph G either there exists $p \in \mathcal{P}$ such that $\mathsf{Answer}(p, G) \neq \emptyset$, or there exists $u \in \mathcal{U}$ such that $\mathsf{Answer}(u, G) = \emptyset$

Given compatible privacy and utility policies, and an input graph G, the second problem of interest consists in finding a sequence of anonymization queries for which the privacy and the utility policies are satisfied according to Def. 5.

CandidateAnonymization$(G, \mathcal{P}, \mathcal{U})$;
Input : A graph G, and $\mathcal{P}$, $\mathcal{U}$: compatible privacy and utility policies
Output: A sequence O of anonymization queries resulting into a modified graph $G' = O(G)$ such that $\mathcal{P}$ and $\mathcal{U}$ are satisfied on G' ; $\perp$ otherwise.

The static version of this problem is when an instance graph G is not in the input :

StaticCandidateAnonymization$(\mathcal{P},\mathcal{U})$;
Input : $\mathcal{P}$, $\mathcal{U}$: compatible privacy and utility policies
Output: A sequence O of anonymization queries resulting *for any instance graph G* into a modified graph $G' = O(G)$ such that $\mathcal{P}$ and $\mathcal{U}$ are satisfied on G' ; $\perp$ otherwise.

The third problem of interest consists in finding an optimal sequence O^{opt} of anonymization queries on a graph G.

BestAnonymization$(G, \mathcal{P},\mathcal{U}, \leq)$;
Input : A graph G, and $\mathcal{P}$, $\mathcal{U}$: user-defined compatible privacy and utility policies
Input : A partial order binary relation $\leq$ over sequences of anonymization queries
Output: An optimal sequence O^{opt} of anonymization queries resulting into a modified graph $G^{opt} = O^{opt}(G)$ such that $\mathcal{P} \cup \mathcal{U}$ are satisfied on G^{opt} and for all O' satisfying $\mathcal{P}$ and $\mathcal{U}$, we have $O^{opt}(G) \leq O'(G)$; $\perp$ otherwise.

---

6 Conclusion and Future Directions

In this paper, we have discussed the underpinnings of a declarative anonymization system that enables data producers to design and verify privacy and utility policies when publishing in Linked Open Data. We have also introduced declarative data transformation operations expressed in the same query language of the aforementioned policies and stated some problems of interest to guarantee the enforcement of these policies. Our future work is devoted to address those problems from a theoretical and practical viewpoint.

Acknowledgements Remy Delanaux would like to thank the Rhone-Alpes region and its ARC6 program for funding his Ph.D scholarship.

Références

Grau, B. C. et E. V. Kostylev (2016). Logical foundations of privacy-preserving publishing of linked data. In D. Schuurmans et M. P. Wellman (Eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pp. 943–949. AAAI Press.

Machanavajjhala, A., X. He, et M. Hay (2017). Differential privacy in the wild : A tutorial on current practices & open challenges. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pp. 1727–1730.

Pérez, J., M. Arenas, et C. Gutierrez (2009). Semantics and complexity of SPARQL. *ACM Trans. Database Syst. 34*(3), 16 :1–16 :45.

Sweeney, L. (2002). k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(5), 557–570.

Co-clustering for differentially private synthetic data generation

Tarek Benkhelif[*,**] Françoise Fessant[*]
Fabrice Clérot[*], Guillaume Raschia[**]

[*]Orange Labs
2, avenue Pierre Marzin, F-22307 Lannion Cédex, France
prénom.nom@orange.com
[**]LS2N - Polytech Nantes
Rue Christian Pauc, BP50609, 44306 Nantes Cédex 3, France
prénom.nom@univ-nantes.fr

Abstract. We propose a methodology to anonymize microdata (i.e. a table of n individuals described by d attributes). The goal is to be able to release an anonymized data table built from the original data while meeting the differential privacy requirements. The proposed solution combines co-clustering with synthetic data generation to produce anonymized data. First, a data independent partitioning on the domains is used to generate a perturbed multidimensional histogram; a multidimensional co-clustering is then performed on the noisy histogram resulting in a partitioning scheme. This differentially private co-clustering phase aims to form attribute values clusters and thus, limits the impact of the noise addition in the second phase. Finally, the obtained scheme is used to partition the original data in a differentially private fashion. Synthetic individuals can then be drawn from the partitions. We show through experiments that our solution outperforms existing approaches and we demonstrate that the produced synthetic data preserve sufficient information and can be used for several datamining tasks.

1 Introduction

There is an increasingly social and economic demand for open data in order to improve planning, scientific research or market analysis. In particular, the public sector via its national statistical institutes, healthcare or transport authorities, is pushed to release as much information as possible for the sake of transparency. Private companies are also involved in the valorization of their data through exchange or publication. Orange has recently made available to the scientific community several mobile communication datasets collected from its networks in Senegal and Ivory Coast as part of D4D challenges (Data for Development). These challenges have shown the potential added-value of analyzing such data for several application domains which address both development projects and improvement of public policies effectiveness Blondel et al. (2012). This demand for publicly available data motivated the research community to propose several privacy preserving data publishing solutions.

1.0.1 Problem statement.

The literature about privacy preserving data publishing is mainly organized around two privacy concepts i) group anonymization techniques such as k-anonymity Sweeney (2002) and ii) random perturbation methods with in particular the concept of Differential Privacy (DP) Dwork (2008). K-anonymity seeks to prevent re-identification of records by making each record indistinguishable within a group of k or more records and allows the release of data in its original form. The notion of protection defended by DP is the strong guarantee that the presence or absence of an individual in a dataset will not significantly affect the result of aggregated statistics computed from this dataset. DP works by adding some controlled noise to the computed function. There are two models for differential privacy: the interactive model and the non-interactive model. A trusted third party collects data from data owners and make it available for data users. In the interactive model, the trusted party catches the queries sent by data users and outputs a sanitized response. In the non-interactive model, the trusted party publishes a protected version of the data. In this paper, we study the problem of differentially private data generation. We consider the non-interactive model and seek to release synthetic data, providing utility to the users while protecting the individuals represented in the data.

1.0.2 Contributions.

We present an original differentially private approach that combines co-clustering, an unsupervised data mining analysis technique, and synthetic data generation. We summarize our contributions below.

— We study and implement a two-phase co-clustering based partitioning strategy for synthetic data generation,
— We experimentally evaluate the released data utility, by measuring the statistical properties preservation and the predictive performance of the synthetic data,
— We compare our approach with other existing differentially private data release algorithms;

The paper is organized as follows. Section 2 first identifies the most related efforts to our work, Sections 3 and 4 give the necessary background on differential privacy and co-clustering, in Section 5 the proposed approach is described. The utility of the produced synthetic datasets is evaluated in Section 6. The final section gathers some conclusions and future lines of research.

2 Related Work

There are many methods designed for learning specific models with differential privacy, but we briefly review here the most related approaches to our work, and we only focus on histogram and synthetic data generation.

The first propositions that started addressing the non-interactive data release while achieving differential privacy are based on histogram release. Dwork et al. (Dwork et al., 2006) proposed a method that publishes differentially private histograms by adding a Laplacian random noise to each cell count of the original histogram, it is considered as a baseline strategy. Xu et al. (Xu et al., 2013) propose two approaches for the publication of differentially private histograms: NoiseFirst and StructureFirst. NoiseFirst is based on the baseline strategy: a

Laplacian random noise is first added to each count as in (Dwork et al., 2006). It is followed by a post-optimization step in which the authors use a dynamic programming technique to build a new histogram by merging the noisy counts. StructureFirst consists in constructing an optimal histogram using the dynamic programming technique to determine the limits of the bins to be merged. The structure of this optimal histogram is then perturbed via an exponential mechanism. And finally the averages of the aggregated bins are perturbed using the Laplacian mechanism. The authors in (Acs et al., 2012) propose a method that uses a divisible hierarchical clustering scheme to compress histograms. The histogram bins belonging to the same cluster have similar counts, and hence can be approximated by their mean value. Finally, only the noisy cluster centers, which have a smaller sensitivity are released. All the mentioned contributions deal only with unidimensional and bidimensional histogram publication and are not adapted to the release of multidimensional data. The closest approach to our work is proposed in (Xiao et al., 2012), first, a cell-based partitioning based on the domains is used to generate a fine-grained equi-width cell histogram. Then a synthetic dataset D_c is released based on the cell histogram. Second, a multidimensional partitioning based on kd-tree is performed on D_c to obtain uniform or close to uniform partitions. The resulted partitioning keys are used to partition the original database and obtain a noisy count for each of the partitions. Finally, given a user-issued query, an estimation component uses either the optimal histogram or both histograms to compute an answer of the query. Other differentially private data release solutions are based on synthetic data generation (Mohammed et al., 2011)(Zhang et al., 2014). The proposed solution in (Mohammed et al., 2011) first probabilistically generalizes the raw data and then adds noise to guarantee differential privacy. Given a dataset D, the approach proposed in (Zhang et al., 2014) constructs a Bayesian network N, that approximates the distribution of D using a set P of low dimensional marginals of D. After that, noise is injected into each marginal in P to ensure differential privacy, and then the noisy marginals and the Bayesian network are used to construct an approximation of the data distribution in D. Finally, tuples are sampled from the approximate distribution to construct a synthetic dataset that is released. Our work focuses on releasing synthetic data and complements the efforts of (Xiao et al., 2012) and (Zhang et al., 2014) in the way that we also study a differentially private aggregation of multidimensional marginals. As in (Xiao et al., 2012), we use the co-clustering like a multidimensional partitioning that is data-aware. And, unlike the variance threshold used in (Xiao et al., 2012) or the θ parameter that determines the degree of the Bayesian network in (Zhang et al., 2014) our solution is parameter-free.

3 Preliminaries and Definitions

3.1 Differential Privacy

Definition 1 (ε-Differential Privacy Dwork (2006)) *A random algorithm $\mathcal{A}$ satisfies ε-differential privacy, if for any two datasets D_1 and D_2 that differ only in one tuple, and for any outcome O of $\mathcal{A}$, we have*

$$Pr[\mathcal{A}(D_1) = O] \leq e^{\varepsilon} \times Pr[\mathcal{A}(D_2) = O], \tag{1}$$

where $Pr[.]$ denotes the probability of an event.

3.1.1 Laplace Mechanism.

To achieve differential privacy, we use the Laplace mechanism that adds random noise to the response to a query. First, the true value of $f(D)$ is computed, where f is the query function and D the data set, then a $random$ noise is added to $f(D)$ and the $\mathcal{A}(D) = f(D) + noise$ response is finally returned. The amplitude of the noise is chosen as a function of the biggest change that can cause one tuple on the output of the query function. This amount defined by Dwork is called sensitivity.

Definition 2 (L_1-sensitivity) *The L_1-sensitivity of $f : D \to \mathbb{R}^d$ is defined as*

$$\Delta(f) = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2}$$

For any two datasets D_1 and D_2 that differ only in one tuple.

The density function of the Laplace distribution is defined as follows.

$$Lap(x|\mu, b) = \frac{1}{2b} \exp\left(\frac{-|x - \mu|}{b}\right) \tag{3}$$

Where μ is called the position parameter and $b > 0$ the scale parameter.

The use of a noise drawn from a Laplacian distribution, $noise = Lap(\Delta f/\varepsilon)$, with the position parameter = 0, and the scale parameter = $\Delta f/\varepsilon$ guarantees the ε-differential privacy Nissim et al. (2007).

3.1.2 Composition.

For a sequence of differentially private mechanisms, the composition of the mechanisms guarantees privacy in the following way:

Definition 3 (Sequential composition McSherry (2009)) *For a sequence of n mechanisms $\mathcal{A}_1$,...,$\mathcal{A}_n$ where each $\mathcal{A}_i$ respects the ε_i-differential privacy, the sequence of the $\mathcal{A}_i$ mechanisms ensures the ($\sum_{i=1}^{n} \varepsilon_i$)-differential privacy.*

Definition 4 (Parallel composition McSherry (2009)) *If D_i are disjoint sets of the original database and $\mathcal{A}_i$ is a mechanism that ensures the ε-differential privacy for each D_i, then the sequence of $\mathcal{A}_i$ ensures the ε-differential privacy.*

3.2 Data model

We focus on microdata. Each record or row is a vector that represents an entity and the columns represent the entity's attributes. We suppose that all the d attributes are nominal or discretized. We use d-dimensional histogram or data cube, to represent the aggregate information of the data set. The records are the points in the d-dimensional data space. Each cell of a data cube represents the count of the data points corresponding to the multidimensional coordinates of the cell.

3.3 Utility metrics

3.3.1 Hellinger distance.

In order to measure the utility of the produced data, we use the Hellinger distance between the distributions in the original data and our synthetic data. We considered the Kullback-Leibler divergence, but we found the Hellinger distance to be more robust given that multidimensional histograms are highly sparse.

Definition 5 (Hellinger distance) *The Hellinger distance between two discrete probability distributions* $P = (p_1, ..., p_k)$ *and* $Q = (q_1, ..., q_k)$ *is given by :*
$D_{Hellinger}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$.

3.3.2 Random range queries.

We use random range queries as a utility measure of the synthetic data. We generate random count queries with random query predicates over all the attributes:

Select COUNT(*) From D Where $X_1 \in I_1$ and $X_2 \in I_2$ and ... and $X_d \in I_d$.

For each attribute X_i, I_i is a random interval generated from the domain of X_i. We use the relative error to measure the accuracy of a query q, where $A_{original}(q)$ denotes the true answer of q on the original data and $A_{perturbed}(q)$ is the noisy count computed when the synthetic data generated from a differentially private mechanism are used.

Definition 6 (Relative error) $RelativeError(q) = \frac{|A_{perturbed}(q) - A_{original}(q)|}{A_{original}(q)}$

4 Co-clustering

Co-clustering is an unsupervised data mining analysis technique which aims to extract the existing underlying block structure in a data matrix Hartigan (1972). The data studied in the co-clustering problems are of the same nature as the data processed by the clustering approaches: they are composed of m observations without label, described by several variables, denoted $\{X_1, X_2, ..., X_d\}$. These variables can be continuous or nominal, then taking a finite number of different values. The values taken by the descriptive variables are partitioned in order to obtain new variables $\{X_1^M, X_2^M, ..., X_d^M\}$ that are called variables-partitions. The values of these new variables are the clusters obtained by the partitions of the values of the variables $\{X_1, X_2, ..., X_d\}$. Each of the X_i^M variables has $\{k_1, k_2, ..., k_d\}$ values which are groups of values if the variable is nominal and intervals if the variable is continuous. The MODL approach makes it possible to achieve a co-clustering on the values of d descriptive variables of the data, we will use this particular feature in our work.

4.1 MODL Co-clustering

We choose the MODL co-clustering Boullé (2010) because: First, MODL is theoretically grounded and exploits an objective Bayesian approach Robert (2007) which turns the dis-

cretization problem into a task of model selection. The Bayes formula is applied by using a hierarchical and uniform prior distribution and leads to an analytical criterion which represents the probability of a model given the data. Then, this criterion is optimized in order to find the most probable model given the data. The number of intervals and their bounds are automatically chosen. Second, MODL is a nonparametric approach according to C. Robert Robert (2007): the number of modeling parameters increases continuously with the number of training examples. Any joint distribution can be estimated, provided that enough examples are available.

4.1.1 Data grid models.

The MODL co-clustering approach allows one to automatically estimate the joint density of several (numerical or categorial) variables, by using a data grid model Boullé (2010). A data grid model consists in partitioning each numerical variable into intervals, and each categorical variable into groups. The cross-product of the univariate partitions constitutes a data grid model, which can be interpreted as a nonparametric piecewise constant estimator of the joint density. A Bayesian approach selects the most probable model given the dataset, within a family of data grid models. In order to find the best M^* model (knowing the data D), the MODL co-clustering uses a Bayesian approach called Maximum A Posteriori (MAP). It explores the space of models by minimizing a Bayesian criterion, called $cost$, which makes a compromise between the robustness of the model and its precision:

$$cost(M) = -\log(P(M|D))\alpha - log(P(M) * P(D|M)) \tag{4}$$

The MODL co-clustering also builds a hierarchy of the parts of each dimension using an ascending agglomerative strategy, starting from M^*, the optimal grid result of the optimization procedure up to $M_\emptyset$, the Null model, the unicellular grid where no dimension is partitioned. The hierarchies are constructed by merging the parts that minimize the dissimilarity index $\Delta(c_1, c_2) = cost(M_{c_1 \cup c_2}) - cost(M)$, where c_1, c_2 are two parts of a partition of a dimension of the grid M and $M_{c_1 \cup c_2}$ the grid after fusion of c_1 and c_2. In this way the fusion of the parts minimizes the degradation of the cost criterion, and thus, minimizes the loss of information.

5 DPCocGen

We present our DPCocGen algorithm, a two-phase co-clustering based partitioning strategy for synthetic data generation. First, a data independent partitioning on the domains is used to generate a multidimensional histogram, the Laplace mechanism is used as in the baseline strategy Dwork et al. (2006) to perturb the histogram. Then, a multidimensional MODL co-clustering is performed on the noisy histogram. This first phase corresponds to a differentially private co-clustering (as shown in figure 1) and aims to produce a partitioning scheme. In the second phase, DPCocGen uses the partitioning scheme to partition the original data and computes a noisy count for each of the partitions (using Laplace mechanism). Finally, the noisy counts are used to draw synthetic individuals from each partition.

The advantage of this approach lies in the fact that the partitioning scheme obtained through the co-clustering is indirectly dependent on the data structure, the intuition is that even after perturbing the multidimensional histogram, the co-clustering phase will preserve some of the

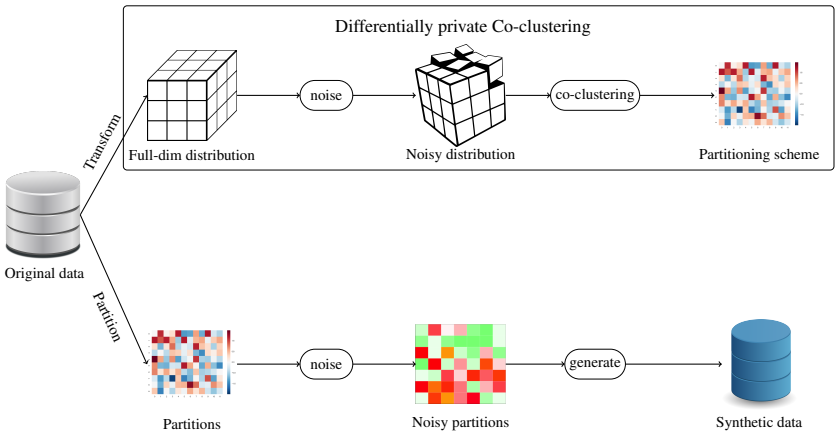


FIG. 1: DPCocGen: a two-phase co-clustering based partitioning strategy for synthetic data generation.

relations between the clusters of attribute values (partitions) of the various dimensions. The resulting cell fusions limit the impact of the noise addition in the second phase. The original data is not consulted during the co-clustering construction which saves the privacy budget that is divided between the two phases to perturb the counts. The detailed steps of DPCocGen are given in Algorithm 1.

Algorithm 1 DPCocGen algorithm

Require: Dataset D, the overall privacy budget ε

1: **Phase 1** :
2: Build a multidimensional histogram from D.
3: **Perturb** the counts of each cell using a privacy budget ε_1.
4: Perform a multidimensional co-clustering from the histogram obtained in step 3.
5: **Phase 2** :
6: Partition the data set D based on the partitioning scheme obtained from step 4.
7: **Perturb** the aggregated counts of each partition returned form step 6 using a privacy budget $\varepsilon_2 = \varepsilon - \varepsilon_1$.
8: Generate synthetic individuals from each partition using the perturbed counts returned from step 7 to build a synthetic dataset D'.

5.1 Privacy guarantee

DPCocGen follows the composability property of the differential privacy, the first and second phases require direct access to the database, Steps 3 and 7 of the Algorithm 1 are ε_1,

Algorithm 2 Perturb algorithm

Require: Count c, privacy budget ε
1: $c' = c + Lap(1/\varepsilon)$
2: **if** $c' < 0$ **then**
3: $\quad c' = 0$
4: **end if**
5: Return c'

ε_2-differentially private. No access to the original database is invoked during the sampling phase. The sequence is therefore ε-differentially private with $\varepsilon = \varepsilon_1 + \varepsilon_2$.

6 Experiments

In this section we conduct three experiments on a real-life microdata set in order to illustrate the efficiency of our proposition on a practical case. The objective is to explore the utility of synthetic data by measuring the statistical properties preservation, the relative error on a set of random range queries answers and their predictive performance.

6.1 Experimental Settings

6.1.1 Dataset.

We experiment with the Adult database available from the UCI Machine Learning Repository [1] which contains 48882 records from the 1994 US census data. We retain the attributes {age, workclass, education, relationship, sex}. We discretize continuous attributes into data-independent equi-width partitions.

6.1.2 Baseline.

We implement the baseline strategy Dwork et al. (2006) to generate a synthetic dataset, a multidimensional histogram is computed and then disturbed through a differentially private mechanism. Records are then drawn from the noisy counts to form a data set.

6.1.3 PrivBayes.

We use an implementation of PriveBayes Zhang et al. (2014) available at Zhang in order to generate a synthetic dataset, we use $\theta = 4$ as suggested by the authors.

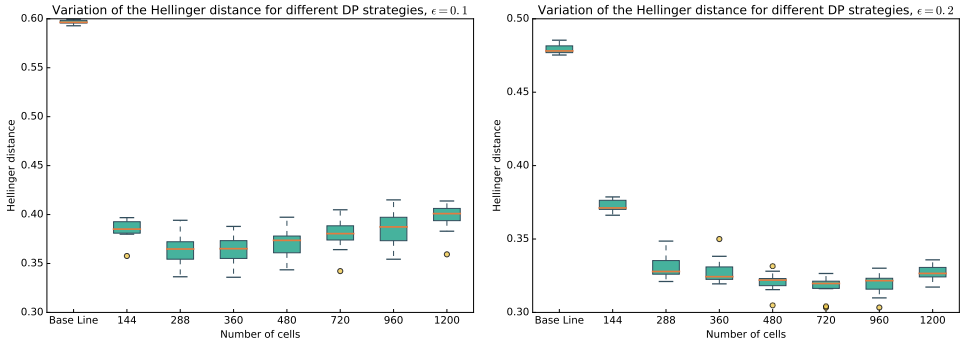
6.1.4 Privacy budget allocation.

The privacy budget is equally divided between the two phases of DPCocGen for all the experiments, $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$.

1. https://archive.ics.uci.edu/ml/

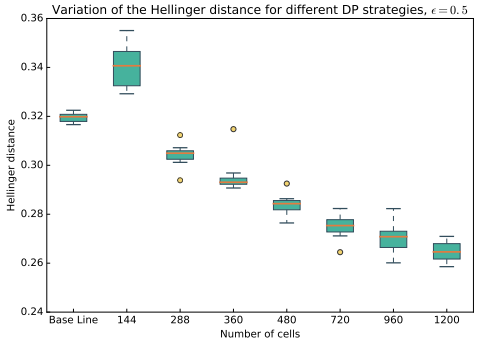
6.2 Descriptive performance

In this experiment, we are interested in preservation of the joint distribution of the original dataset in the generated synthetic data. In order to measure the difference between two probability distribution vectors we choose the Hellinger distance. First, we compute the multivariate distribution vector P of the original dataset, then, we compute the multivariate distribution vector Q of the synthetic data generated using *DPCocGen* and the multivariate distribution vector Q' of the synthetic data generated using *Base line*. Finally, the distances $D_{Hellinger}(P,Q)$ and $D_{Hellinger}(P,Q')$ are measured. For each configuration the distances are calculated through 50 synthetic data sets and represented in Figure 2. We use box-plots diagrams to represent these results where the x-axis represents the synthetic data generation method, the first box in the left represents the baseline strategy, the following boxes correspond to *DPCocGen* with different levels of granularity (number of cells). The y-axis indicates the Hellinger distance measured between the distribution calculated on the generated data and the original distribution.



(a) $\varepsilon = 0.1$

(b) $\varepsilon = 0.2$



(c) $\varepsilon = 0.5$

FIG. 2: Joint distribution distances

Regardless of the privacy budget, the joint probability distribution of the synthetic data generated with *DPCocGen* is closer to the original distribution than the distribution of the data that is obtained using *Baseline*, except when $\varepsilon = 0.5$ and for the *DPCocGen* case with a high co-clustering aggregation level (144 cells), in that particular configuration the partitioning

was too coarse and failed to correctly describe the data. The optimal aggregation level varies according to noise magnitude, but the finest aggregation level seems to offer a satisfying result for each configuration.

6.3 Random range queries

The goal of this experiment is to evaluate the utility of the produced data in terms of relative error when answering random range queries. We first generate 100 random queries. We produce synthetic datasets using *Base line*, *PrivBayes* and *DPCocGen*. We compute all the queries and report their average error over 15 runs. We use for this experiment the finer co-clustering level. Figure 3 shows that the average relative error decreases as the privacy budget ε grows for the three algorithms. One can also observe that *PrivBayes* and *DPCocGen* are close and do better than *Base line* regardless of the privacy budget.

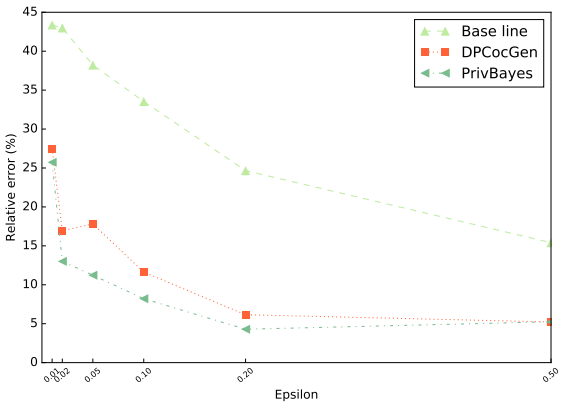


FIG. 3: Random range queries

6.4 Predictive performance

In this experiment we are interested in the classification performances obtained with a supervised classifier whose learning is based on synthetic datasets. We randomly select 80% of the observations in order to generate the synthetic data using *DPCocGen*, *Base line* and *PrivBayes*, we use the generated data to train a classifier in order to predict the value of the attributes *Sex* and *Relationship*. The remaining 20% are used for the evaluations. We use for this experiment the finer co-clustering level. The results are presented Figures 4 and 5, they represent the average on 50 runs. The privacy budget value is shown on the x-axis, the y-axis shows the area under the ROC curve (AUC) measured on the test set. The figure also indicates the performances obtained when the real data are used for learning the model (Original Data).

We retain that the classification performances obtained with *DPCocGen* are close to those obtained when the real data are used for learning the model. The performances of *DPCocGen* are always higher than those of the *Base line* and *PrivBayes*.

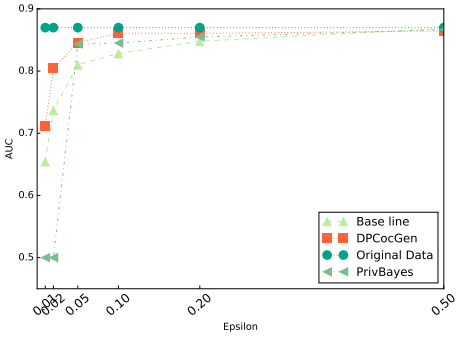


FIG. 4: Sex prediction

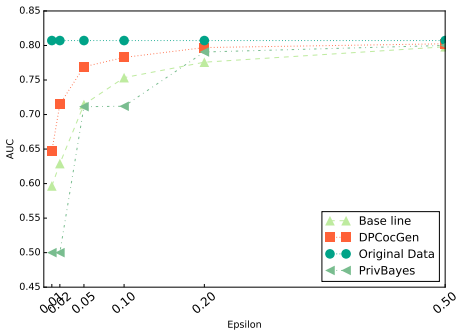


FIG. 5: Relationship prediction

7 Conclusion

The synthetic data generated using *DPCocGen* retain the statistical properties of the raw data, thus using the synthetic data for various data mining tasks can be envisaged. We have also shown that our parameter-free approach outperforms other existing differentially private data release algorithms.

References

Acs, G., C. Castelluccia, and R. Chen (2012). Differentially private histogram publishing through lossy compression. In *2012 IEEE 12th International Conference on Data Mining*, pp. 1–10. IEEE.

Blondel, V. D., M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki (2012). Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*.

Boullé, M. (2010). Data grid models for preparation and modeling in supervised learning. *Hands-On Pattern Recognition: Challenges in Machine Learning 1*, 99–130.
English

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener (Eds.), *Automata, Languages and Programming*, Volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer Berlin Heidelberg.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association 67*(337), 123–129.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Con-*

ference on Management of data, pp. 19–30. ACM.

Mohammed, N., R. Chen, B. Fung, and P. S. Yu (2011). Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–501. ACM.

Nissim, K., S. Raskhodnikova, and A. Smith (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, New York, NY, USA, pp. 75–84. ACM.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(05), 557–570.

Xiao, Y., L. Xiong, L. Fan, and S. Goryczka (2012). Dpcube: differentially private histogram release through multidimensional partitioning. *arXiv preprint arXiv:1202.5358*.

Xu, J., Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett (2013). Differentially private histogram publication. *The VLDB Journal 22*(6), 797–822.

Zhang. `https://sourceforge.net/projects/privbayes`.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2014). Privbayes: Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1423–1434. ACM.

Résumé

This work presents an approach for the anonymization of microdata sets. The goal was to be able to produce synthetic data that preserve sufficient information to be used instead of the real data. Our approach involves combining differential privacy with synthetic data generation. We use co-clustering a data joint distribution estimation technique, in order to partition the data space in a differentially private manner. Then, we use the resulting partitions to generate synthetic individuals. We now plan to compare our approach to a previous work, that is being published, which is based on a group anonymization technique and we aim to articulate the discussion around the utility/protection trade-off.

Système de gestion de l'obsolescence dans une base d'informations personnelles

Salma Chaieb[1,3,*], Véronique Delcroix [2,**]
Ali Ben Mrad[1,3,***] Emmanuelle Grislin-Le Strugeon[2,****]

1 : Faculté des Sciences de Monastir (FSM)
Université de Monastir Rue Salem BCHIR
Skanes 5000 Monastir - Tunisie
2 : LAMIH - UVHC - UMR8201
3 CES Lab, ENIS, Université de Sfax, 3038, Sfax, Tunisie
* salma.chaieb2@yahoo.com
** veronique.delcroix@univ-valenciennes.fr
*** benmradali2@gmail.com
**** emmanuelle.grislin@univ-valenciennes.fr

Résumé. L'évaluation régulière du risque de chute des personnes âgées requiert des informations fiables et nombreuses. Comme il n"est pas possible de recueillir régulièrement toutes ces informations, les observations sont faites au fil du temps et conservées, ce qui entraîne une problématique liée au vieillissement des informations. Cet article traite de la détection des informations obsolètes dans une base d'informations sur une personne âgée. Nous proposons un agent chargé de maintenir une base d'informations et capable de fournir à la demande des informations fiables et cohérentes sur la personne. L'agent est équipé d'un modèle de connaissances sur les personnes âgées, sous forme d'un réseau bayésien et d'un module de raisonnement chargé de la détection et de la gestion des contradictions et des doutes sur les informations.

1 Introduction

La chute est la première cause de décès accidentel chez les personnes âgées de plus de 65 ans. Les médecins traitants sont des acteurs majeurs de la prévention des chutes. Cependant, l'évaluation régulière du risque de chute requiert des informations fiables et nombreuses sur la personne âgée et la collecte de ces informations est coûteuse en temps. Pour pallier ce problème, nous proposons un système de gestion des informations personnelles chargé de collecter les informations au fil du temps, de les stocker et les gérer, en vue de fournir ces informations à la demande.

Dans cet article, nous traitons des cas de la vie réelle où on ne dispose que d"informations incomplètes, incertaines ou incohérentes dans des situations de prise de décision. Ces informations sont très souvent variées et évolutives et risquent d'être obsolètes à un certain moment et de contredire d'autres informations. Elles doivent ainsi être vérifiées

continuellement afin de refléter fidèlement l'état et le comportement des sujets considérés. L'objectif de ce travail est de maintenir à jour un ensemble d'informations le plus fiable et complet possible, de façon à fournir à la demande des informations avec un degré de confiance élevé. Pour cela, nous proposons un mécanisme de détection et de gestion des contradictions dans la base d'informations. Notre contribution concerne un dispositif embarqué, intelligent et autonome gérant des informations sur la personne âgée (PA) à l'aide d"agent et de réseau bayésien. Nous utilisons un réseau bayésien (RB) (Jensen, 1996; Naïm et al., 2011; Pearl, 1988) pour modéliser les connaissances sur les personnes âgées sur les facteurs de risque de chute. Ce type de modèle graphique probabiliste permet de combiner les connaissances générales sur les PA et les observations sur une PA précise pour mettre à jour les croyances sur les variables non observées.

Cet article comporte trois parties. La première explore un bref état de l'art sur la problématique de l'obsolescence des informations, les outils techniques utilisés. La deuxième partie aborde l'architecture générale de l'agent ainsi que l'algorithme de gestion de l'obsolescence et les définitions de base proposés. La dernière partie présente les résultats de la réalisation de notre application.

2 Modélisation de l'incertitude

Dans cette section, nous présentons un bref parcours de l'état de l'art sur les systèmes de gestion de l'obsolescence des informations déjà existants. Nous justifions le choix des RBs comme outils de modélisation du problème. Par la suite, nous abordons le contexte d'utilisation de l'agent responsable de la gestion des informations d'une PA.

2.1 Etat de l'art sur la gestion de l'obsolescence des informations

Une première recherche bibliographique montre qu'il n'y a pas de travaux autour de l'obsolescence des informations concernant une entité telle qu'une PA, ni d'agents logiciels utilisant les RBs à cet effet. Cependant, cet axe de recherche a également attiré certains chercheurs dans d'autres secteurs. Citons à titre d"exemple l'algorithme de bandit (Hachour et al., 2014; Louëdec et al., 2015) sur lequel se basent les systèmes de recommandation (Kembellec et al., 2014). Ce dernier permet de mettre à jour l'ensemble des informations en supprimant celles les plus anciennes (dont l'âge dépasse un certain seuil). Les systèmes de reconnaissance automatique de cibles radar (Saidi et al., 2009) ou encore les systèmes d'information hospitalier (Saidi et al., 2009). Ces systèmes sont basés sur la proposition et l'étude de différentes mesures de qualité de l'information en fonction de son évolution dans le temps (la qualité locale/en entrée et la qualité globale/en sortie) . Dans notre proposition, comme dans les cas mentionnés, les informations traitées sont accompagnées par des méta-informations permettant d'évaluer la qualité de ces dernières. Cependant, dans notre cas, il ne s'agit pas de déterminer la pertinence des observations "au moment où elles arrivent" comme cela est décrit dans ces cas ni de prendre en compte que les informations les plus récentes. Le vieillissement des informations est un problème connu dans les bases de données mais il

se présente différemment. En effet, une base de données regroupe des enregistrements concernant un grand nombre d'entités, dont chacune est décrite par un nombre limité d'informations. A l'inverse, la base d'informations que nous considérons concerne une seule personne, mais regroupe de nombreuses informations qui sont en partie dépendantes les unes des autres (dépendance causale et/ou statistique). La problématique de cet article vise à prendre en compte ces dépendances pour détecter les contradictions avec une approche probabiliste.

2.2 Réseau bayésien

Les réseaux bayésiens figurent parmi les modèles graphiques probabilistes. Un RB est défini par un couple $B = (G, P)$ où $G = (\mathbf{X}, E)$ est un graphe dirigé sans circuit caractérisé par un ensemble de nœuds $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ et un ensemble d'arcs E et P une distribution de probabilités jointes sur $\mathbf{X}$ donnée par la règle suivante :

$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$ avec $Pa(X_i)$ est l'ensemble des parents du nœud X_i dans G.

Dans ce cadre, nous précisons les concepts de connaissance, observation et croyance. Une *connaissance* est une information stable et générale qui peut être acquise par les expériences, les compétences, l'intuition et l'expertise. Dans un RB, la connaissance est modélisée par la structure du graphe et par la distribution de probabilités jointes sur l'ensemble des variables associées aux nœuds du graphe. Une *observation* est une information locale et ponctuelle liée à une situation bien déterminée. Une observation est une instanciation. Les observations sont généralement considérées comme étant synchrones, ce qui n'est pas le cas dans cet article puisque nous considérons une base d'informations sur une personne recueillies au fil du temps. Quant à la *croyance* (belief en anglais), elle représente l'état cognitif d'un observateur à propos d'une variable, dans une situation bien déterminée. Une croyance est exprimée sous la forme de distributions de probabilités *a posteriori*. La détection des observations obsolètes repose sur leur contradiction potentielle avec d'autres observations plus récentes et/ou plus jugées plus fiables portant sur des variables différentes. Détecter ce type de contradiction nécessite de combiner des connaissances sur les dépendances entre les variables et du raisonnement dans l'incertain (du fait des observations incomplètes d'une part et de l'incertitude stochastique naturelle sur les connaissances). Notre proposition est basée sur l'utilisation d'un réseau bayésien car ce type de modèle combine parfaitement modèle de connaissance et de raisonnement dans l'incertain.

2.3 Incertitude et obsolescence des observations

Lors de l'utilisation classique d'un RB, on dispose d'observations partielles mais certaines sur une situation, c'est-à-dire que toutes les variables ne sont pas observées, et le RB permet de calculer $P(V|OBS)$ où V est une variable non observée. Dans cet article, nous ajoutons un degré d'incertitude supplémentaire puisqu'en conservant des observations du passé, nous ne pouvons plus affirmer avec certitude que l'observation est encore vraie. Ce second niveau d'incertitude lié au vieillissement des observations ne peut être géré dans le RB. Il ne s'agit pas non plus d'observations incertaines au sens

des observations de *vraisemblance*[1] (Pearl, 1988) ou des observations *probabilistes* fixes ou non fixes[2] (Bloemeke, 1998; Valtorta et al., 2002; Mrad, 2015; Mrad et al., 2015) car le RB n'inclut pas de dimension temporelle. Il s'agit donc de considérer une plage de temps étendue au cours de laquelle diverses observations sont faites, avec la prise en compte du vieillissement des observations : une observation faite à un instant t reste vraie pendant un certain intervalle de temps, mais on ne sait pas précisément combien de temps elle reste vraie. On sait en revanche que les observations sur certaines variables ne vieillissent pas, et que pour d'autres variables, les observations restent valides plus ou moins longtemps.

Pour gérer l'incertitude liée au vieillissement des observations, nous proposons des fonctions spécifiques à chaque variable du RB appelées fonctions de *péremption*, qui permettent de calculer pour chaque observation un *degré de confiance* qui est fonction de la durée écoulée depuis la date de l'observation.

2.4 Contexte d'utilisation de l'agent chargé de la gestion des informations

Le système de prévention des chutes peut être utilisé dans le cadre d'un rendez-vous de la PA avec son médecin traitant. La consultation n'étant pas principalement dédiée à la prévention de la chute, ces informations sur le patient ne peuvent être toutes resaisies par le médecin. Pour pallier cela, le système de prévention des chutes communique avec l'agent responsable de la gestion des informations d'une PA : (1) le système de prévention des chutes demande des informations sur une PA pour un ensemble de variables utiles pour évaluer le risque de chute et fournir des recommandations ; (2) l'agent responsable de la gestion des informations pour la PA renvoie les informations sur les variables demandées. Deux cas sont possibles suivant les variables demandées :

1. Si l'agent dispose d'une information sur une variable V demandée par le système de prévention des chutes, il renvoie l'information sous la forme $V = v$ avec un certain degré de confiance
2. Sinon, l'agent renvoie une information sous forme de distributions de probabilités mises à jour à partir des informations sur la PA connues par l'agent $P(V|\mathbf{OBS} = \mathbf{obs})$,

3 Agent pour la gestion des informations d'une PA

Dans cette section, nous décrivons les différents modules qui composent l'agent responsable de la gestion des informations d'une PA. Nous présentons ensuite l'algorithme général de gestion d'une base d'informations ainsi que les définitions des concepts sur lesquelles repose notre système.

1. appelées virtual evidence ou likelihood evidence en anglais
2. appelées soft evidence ou fixed / not fixed probabilistic evidence en anglais

3.1 Architecture de l'agent chargé des informations d'une PA

Le rôle de l'agent est de maintenir à jour une base d'informations sur une PA, sur la base de données acquises par des dispositifs externes et/ou par interaction avec la personne. Ces activités peuvent être vues comme les différentes capacités d'un agent logiciel, à savoir un système capable de percevoir et d'interagir avec son environnement en faisant preuve de pro-activité (Wooldridge et Jennings, 1995). L'agent visé a ainsi pour but de maintenir à jour une base d'informations sur la PA. Pour cela, il doit récupérer des données issues de sources externes et les intégrer au modèle. Lorsque cela s'avère nécessaire, il peut décider d'interagir avec la PA pour confirmer les observations. Pour cela, l'agent effectue de façon cyclique les opérations suivantes :

- le recueil de données et d'informations sur la personne constituées d'observations et leurs dates à l'aide d'une interface graphique proposée ;
- la détection d'éventuelles contradictions entre ces nouvelles observations et les observations plus anciennes, et plus généralement, avec l'ensemble des croyances sur l'état des variables (y compris celles pour lesquelles on n'a pas ou plus d'observations) ;
- la mise à jour des informations sur la personne à partir des observations et du RB ;
- l'interaction avec la personne en vue de réduire les contradictions.

Comme montré dans la figure 1, notre agent est composé principalement de quatre parties que nous présentons plus en détail ci-dessous : un modèle probabiliste de connaissances sur la PA constitué d'un réseau bayésien, une base d'informations, un module d'interaction et un module de raisonnement.

Modèle probabiliste de connaissances de la personne âgée

Le modèle de connaissances de l'agent est un réseau bayésien qui encapsule les connaissances issues des experts, de la littérature et des données statistiques sur une population de personnes âgées. Ce type de modèle graphique probabiliste décrit de façon qualitative et probabiliste les liens de dépendances et d'indépendances conditionnelles parmi un ensemble de variables d'intérêt. Dans ce travail, nous proposons un petit RB à treize nœuds (voir figure 2) qui vise à implémenter un démonstrateur.

La base d'informations

La gestion classique des observations dans un RB comme de simples affectations de certaines variables du RB ne permet pas de gérer le vieillissement des observations. Nous proposons donc un module de stockage spécifique pour chaque PA qui contient les observations qui lui sont propres. Comme montré dans le tableau 1, cette base d'informations regroupe l'ensemble des variables du RB : pour les variables qui ont déjà été observées et dont l'observation est considérée comme étant toujours d'actualité, la base contient les observations associées, leurs dates de saisie, les fonctions de péremptions et donc les degrés de confiances. Pour les variables qui n'ont jamais été observées ou dont la dernière observation a été considérée comme obsolète, la base contient : la valeur la

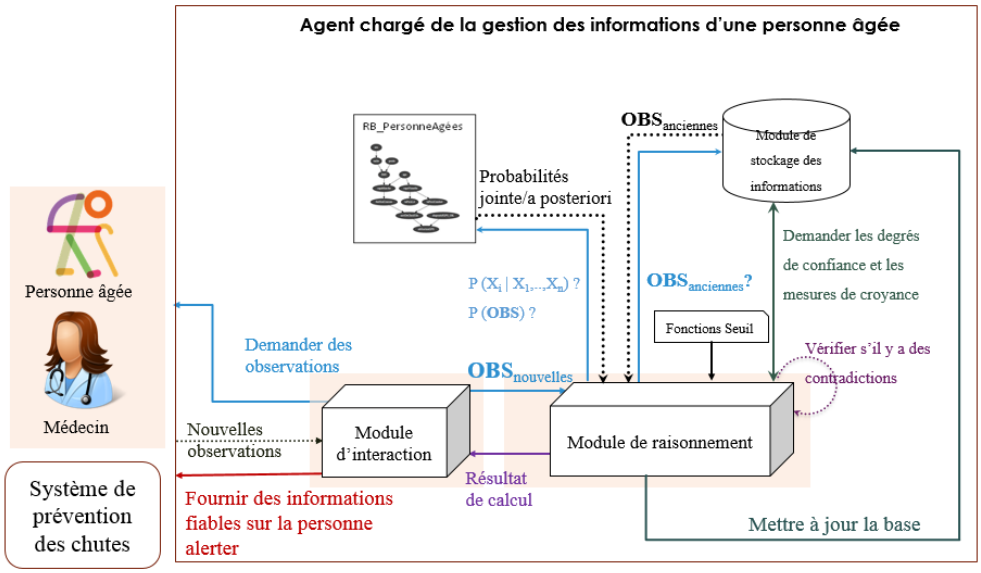


FIG. 1 – *Schéma général de l'agent chargé de la gestion des informations d'une PA*

plus probable sachant les observations récentes et la probabilité *a posteriori* associée. Ces mesures reflètent les croyances sur l'état de la PA.

Module d'interaction

Le module d'interaction permet de gérer en permanence l'arrivée de nouvelles informations que ce soit à partir d'une application mobile, ou à partir de capteurs. Ce module fournit des informations fiables sur des variables cibles à la demande. Il gère aussi les messages d'alerte concernant les contradictions afin de vérifier d'où vient le problème. Nous avons proposé des interfaces graphiques qui traduisent les interactions entre notre agent et la PA et/ou agents concernés (voir figure 2). Il s'agit d'une version initiale de notre agent ayant comme objectif de tester les méthodes proposées et de faire dérouler quelques scénarios. Cependant, nous sommes satisfaits, dans un premier temps, des résultats obtenus pour ce premier démonstrateur.

Module de raisonnement

Comme le montre la figure 1, le module de raisonnement est central et il interagit avec les 3 autres modules. Sa fonction principale est la gestion de la cohérence de la base d'information. Il s'agit de détecter la présence d'observations obsolètes dans la base d'informations. Précisons que la base d'informations contient au plus une valeur observée pour chaque variable. Lorsqu'une observation est faite sur une variable, la valeur observée remplace la valeur précédente dans la base d'informations si elle existait. La présence d'une observation obsolète dans la base d'informations peut se manifester de deux façons : d'une part, une observation obsolète entre en contradiction avec des

Variable Variable	Valeur observée	Date de l'observation	Degré de confiance	Valeur la plus probable | obs	Probabilité | obs
laPersConduit	auMoins1Fois-ParSem	Date 1 : t_1	Pm_1 (t_1)= +	-	-
faitSesCourses	oui	Date 1 : t_1	Pm_1 (t_1)= +	-	-
dispositifGPS	ok	Date 1 : t_1	Pm_1 (t_1)= +	-	-
capaVisuelle	-	-	-	correcte	0.98
sortDeChezElle	-	-	-	auMoins1Fois-ParSem	0.9
nbSortiesGPS	-	-	-	deuxOuPlus	0.73
capaMarche	-	-	-	normale	0.8
laPersLit	-	-	-	régulièrement	0.63
age	-	-	-	[60-63]	0.51
sexe	-	-	-	M	0.51
IMC	-	-	-	normale	0.38
taille	-	-	-	[160-170]	0.3
poids	-	-	-	[60-75]	0.19

TAB. 1 – *Base d'informations d'une personne âgée : les observations (avec date et niveau de confiance) et les croyances sur les variables non observées (distributions de probabilités a posteriori.*

observations sur d'autres variables. Ceci se traduit par le fait que l'observation simultanée de cet ensemble de valeur est extrêmement improbable. Nous disons ici que le la base d'informations n'est pas dans un *état possible*. D'autre part, une observation obsolète peut générer un *doute* sur une variable non observée.

Le diagramme d'activité général (figure 3) donne une vision globale du module de raisonnement. La succession des cinq phases principales se répète jusqu'à ce que la base d'informations soit dans un *état stable*. La détection des contradictions repose sur l'évaluation de la probabilité jointe de l'ensemble des observations. En dessous d'un seuil très bas, on considère que l'ensemble des observations anciennes et nouvelles est contradictoire. La valeur du seuil est spécifique à chaque sous-ensemble de variables. Une fois que les observations menant à des contradictions sont identifiées et supprimées de la base, il faut vérifier s'il y a des *doutes* sur l'une des variables non observées et mettre à jour la base d'observations. Nous proposons ci-dessous les définitions des termes « *état possible* » et « *état stable* ».

Définition 3.1.1 *On appelle* état possible *un état de la base d'informations tel que* $P(\boldsymbol{OBS} = \boldsymbol{obs}) > Seuil(\boldsymbol{OBS})$*, où* $\boldsymbol{OBS} \subset \boldsymbol{X}$ *est l'ensemble des variables observées,* $\boldsymbol{obs}$ *est l'ensemble des valeurs observées pour* $\boldsymbol{OBS}$ *et où* $Seuil(\boldsymbol{OBS})$ *est une fonction qui associe à chaque sous ensemble de variables une valeur en dessous de laquelle on considère que cet ensemble d'observations n'est "pas possible".*

Les observations concernant les variables de **OBS** ont eu lieu à des dates quelconques, mais n'ont pas (ou pas encore) été détectées comme obsolètes. Cette définition relaxe le sens strict du mot *possible* qui correspond généralement à une probabilité nulle.

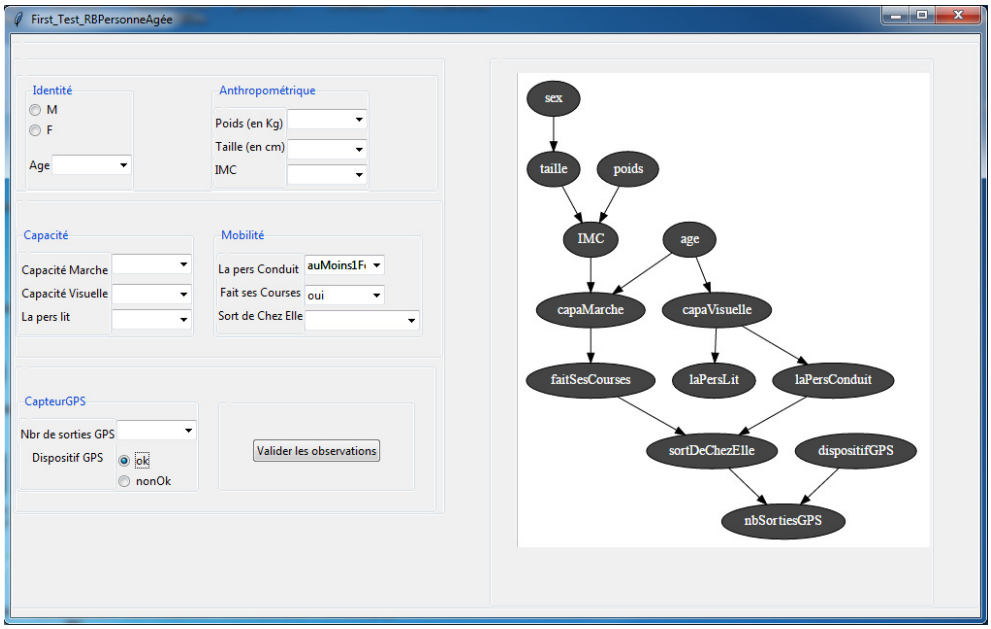


FIG. 2 – *Interfaces graphiques pour la gestion des interactions.*

Définition 3.1.2 *On dit qu'on a un* doute *sur une variable (non observée : jamais observée ou pour laquelle on a supprimé une ancienne observation) si la croyance qu'on avait sur cette variable est fortement remise en cause par les nouvelles observations.*

Définition 3.1.3 *Une base d'informations est dans un* état stable *si et seulement si elle est dans un* état possible *et on n'a aucun doute sur ses variables non observées ou sur lesquelles on a des anciennes observations.*

La valeur du seuil dépend logiquement de la taille du domaine des observations et des conventions adoptées dans la définition du RB pour caractériser les situations locales impossibles dans les tables de probabilités conditionnelles. Nous ne détaillons pas cette fonction car nous avons temporairement adopté une définition très simplifiée.

3.2 Algorithme général

Notre algorithme de gestion d'une base d'informations se compose principalement de deux phases. La première phase consiste à chercher parmi les observations contenues dans la base celles qui contredisent les nouvelles observations. La deuxième phase gère les doutes sur les variables non observées après avoir supprimé les observations obsolètes.

Algorithme de gestion de l'obsolescence des observations

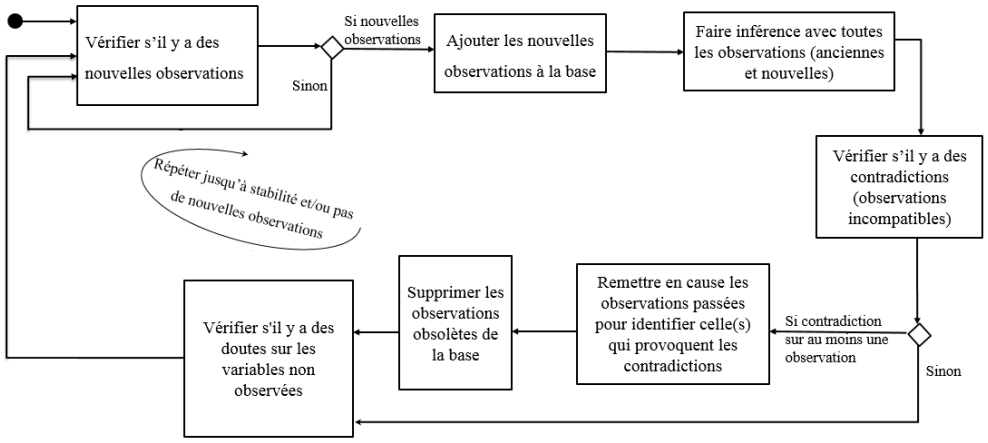


FIG. 3 – *Diagramme d'activité général.*

Procédure événementielle déclenchée par l'arrivée de nouvelles observations.

Entrée : un réseau bayésien (G, P) avec G = (**X**, **E**), une base d'informations sur **X**, l'ensemble des nouvelles observations $\mathbf{OBS}_{nouvelles}$ avec leur date d'observation.

Postcondition : cette procédure restitue la base d'informations dans un *état stable* ou bien, cette procédure aboutit à une question, dont la réponse déclenche à nouveau cette procédure.

$\mathbf{OBS} = \mathbf{OBS}_{anciennes} \cup \mathbf{OBS}_{nouvelles}$
Si *EtatPossible* (**OBS**, RB) = faux, alors
  $\mathbf{OBS}_{supprimées} = \mathit{GérerContradictions}(\mathbf{OBS}_{nouvelles}, \mathbf{OBS}_{anciennes}, RB)$
Fin Si
GérerDoutes (RB, $\mathbf{OBS}_{anciennes}$, $\mathbf{OBS}_{nouvelles}$)

3.3 Détection et gestion des contradictions

La présence de contradiction dans la base d'informations est détectée à l'aide de la fonction **EtatPossible** qui teste simplement si la probabilité jointe des observations est sous un seuil minimum. Lorsqu'il y a présence de contradiction, il s'agit de trouver parmi la ou les observations du passé, laquelle (ou lesquelles) devrait être remise(s) en cause. Nous vérifions en premier lieu parmi les observations anciennes, celles ayant le plus faible degré de confiance, susceptibles d'être supprimées. Puis nous étudions les probabilités *a posteriori* de ces dernières avant et après l'arrivée des nouvelles observations. Ainsi, on ne supprime que les observations dont la probabilité *a posteriori*

sachant les anciennes observations est très différente de leur probabilité *a posteriori* sachant les anciennes et les nouvelles observations.

4 Réalisation et résultats de l'application

Notre application a été implantée sous l'environnement Windows, avec le le logiciel pyAgrum (Gonzales et al., 2014). La visualisation graphique des résultats s'effectue à l'aide du logiciel graphviz (Gansner et North, 2000). Notre base d'observations a été implantée à l'aide du logiciel Excel.

Afin de tester la validité de notre approche, nous avons déroulé divers scénarios et réalisé un premier test comparatif d'une part avec un raisonnement plus systématique, d'autre part avec les résultats attendus en prenant en compte la signification des informations.

Le raisonnement systématique consisterait à déclarer obsolète toute information dont le degré de confiance est sous un seuil fixé. Ceci conduit à deux erreurs :

— supprimer une information qui n'est pas devenue obsolète

— ne pas supprimer des informations obsolètes.

En comparaison avec ce raisonnement, l'algorithme que nous proposons évite systématiquement le premier écueil et partiellement le second puisque c'est la contradiction avec une information plus récente qui permet de détecter l'obsolescence.

Avec la prise en compte de la signification des informations, les résultats sont ceux que pourrait donner un expert humain. Dans la majorité des cas, notre agent donne des résultats corrects et raisonne de façon très proche de la réalité. Il simule le raisonnement humain dans sa capacité de gérer l'obsolescence des informations et de traiter les contradictions. Ainsi, sur 5 scénarios testés, le résultat est satisfaisant dans 4 cas (l'agent supprime toutes les observations obsolètes de la base) et une fausse alerte dans un cas. Cette fausse alerte est une mauvaise évaluation de la fonction **EtatPossible**, que nous attribuons à la définition de la fonction seuil. Les résultats préliminaires trouvés sont encourageants mais doivent être vérifiés et améliorés afin d'améliorer la performance du système. L'évaluation de notre modèle auprès des PA et d'autres acteurs n'est pas évidente. Il est à ce stade de l'étude trop tôt pour arbitrer de tel objectif. Cette étape ne peut être franchie plus tard que par des tests significatifs en situation d'usages réels.

5 Conclusion et perspectives

Certes, le domaine de traitement des informations en termes d'obsolescence est encore naissant et il semble que cette question n'a pas fait l'objet de travaux de recherche. Cela nous amène à concevoir un agent chargé de la gestion d'obsolescence des informations qui vise en particulier les PA. Cet agent doit réaliser différentes tâches dont le recueil de données sur la personne (les observations), la détection d'éventuelles contradictions entre ces nouvelles observations et les informations précédemment recueillies sur la personne, la mise à jour de la base d'informations sur la personne à partir des observations et d'un modèle probabiliste de connaissances générales sur les PA sous forme

d'un réseau bayésien, l'interaction avec la personne en vue de réduire les contradictions, etc. Parmi les objectifs mentionnés, notre agent vise également à limiter le nombre de questions posées à la PA et exploiter au maximum des observations faites dans le passé, tout en sachant que les choses ont pu évoluer. Nos perspectives concernent en premier lieu une meilleure définition formelle du problème de la gestion d'une base d'informations non synchrones et non pérennes, en termes de définitions des concepts et de la terminologie, tels que la cohérence de la base d'informations, son niveau d'information, les concepts de contradictions, de doutes, d'état possible et d'état stable. Cette perspective vise à obtenir une terminologie solidement définie et articulée dans le contexte général de l'obsolescence des informations. En second lieu, nous prévoyons le développement d'une première version d'un système de gestion des données personnelles et d'un système de prévention des chutes des personnes âgées à domicile à destination des médecins traitants ; ce second système utilise les informations du premier système. Ces travaux impliquent une amélioration des algorithmes proposés et des tests sur un RB plus complet sur la PA, défini par des experts du domaine. Dans le cadre de ces travaux, nous prévoyons de travailler avec des médecins traitants et certains de leurs patients pour un développement centré utilisateur. Ces travaux devraient permettre d'une part de mieux cerner les questions d'éthique concernant les données personnelles (stockage, droit d'accès, sécurité des données, ...) et de proposer des réponses partielles dans un cadre restreint, et d'autre part d'avoir un premier retour sur l'acceptabilité de ces systèmes par les personnes âgées et les médecins traitants. Au cours de cette phase, le médecin traitant sera le seul utilisateur à entrer des informations dans le système de gestion des données personnelles de ses patients. Cette restriction permet d'écarter temporairement les questions liées au recueil des données via des sources multiples et incertaines. Les perspectives à plus long terme concernent l'extension du système de gestion des données personnelles à plusieurs utilisateurs gravitant autour de la personne concernée, et à plusieurs systèmes clients visant d'autres objectifs que la prévention des chutes.

Remerciements Ce travail est soutenu par le projet ELSAT2020 (Eco-mobilité Logistique Sécurité et Adaptabilité dans les Transports à l'Horizon 2020) et cofinancé par l'Union Européenne avec le Fond européen de développement régional, l'État et la Région Hauts de France. Les auteurs remercient le support de ces institutions.

Références

Bloemeke, M. (1998). *Agent encapsulated Bayesian networks.* Ph.d. thesis, Department of Computer Science, University of South Carolina.

Gansner, E. R. et S. C. North (2000). An open graph visualization system and its applications to software engineering. *Software Practice and Experience 30*(11), 1203–1233.

Gonzales, C., L. Torti, et P.-H. Wuillemin (2014). Librairie agrum : a graphical universal model. *Revue d'intelligence artificielle–no 2*(3), 1–10.

Hachour, H., S. Szoniecky, et S. Abouad (2014). Espaces sémio-cognitifs : les frontières des systèmes de recommendation.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, Volume 210. UCL press London.

Kembellec, G., G. Chartron, et I. Saleh (2014). *Les moteurs et systèmes de recommandation*. ISTE éditions.

Louëdec, J., M. Chevalier, A. Garivier, et J. Mothe (2015). Algorithmes de bandit pour les systèmes de recommandation : le cas de multiples recommandations simultanées. In *CORIA*, pp. 73–88.

Mrad, A. B. (2015). *Observations probabilistes dans les réseaux bayésiens*. Ph.d. thesis, LAMIH, Univ. de Valenciennes, France.

Mrad, A. B., V. Delcroix, S. Piechowiak, P. Leicester, et M. Abid (2015). An explication of uncertain evidence in bayesian networks : likelihood evidence and probabilistic evidence. *Appl. Intell. 43*(4), 802–824.

Naïm, P., P.-H. Wuillemin, P. Leray, O. Pourret, et A. Becker (2011). *Réseaux bayésiens*. Editions Eyrolles.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann.

Saidi, M. N., A. Toumi, B. Hoeltzener, A. Khenchaf, et D. Aboutajdine (2009). Système automatique de reconnaissance de cibles radar. *6e Atelier Fouille de Données Complexes dans un processus d'extraction de connaissances*, 15.

Valtorta, M., Y.-G. Kim, et J. Vomlel (2002). Soft evidential update for probabilistic multiagent systems. *International Journal of Approximate Reasoning 29*(1), 71–106.

Wooldridge, M. et N. R. Jennings (1995). Intelligent agents : Theory and practice. *The Knowledge Engineering Review 10*(2), 115–152.

Summary

The frequent evaluation of the risk of fall for elderly people require reliable and abundant information about the person. Since it is not possible to gather regularly all the interesting pieces of information, observations are made over time and stored, which lead to a problem of aging of information. This article deals with oudated information in a base of information about an elderly person. We propose an agent in charge of the maintenance of a base of information and able to provide on demand reliable and consistent information about the person. The agent is equiped with a model of knowledge about elderly in the form of a Bayesian network and a reasoning module in charge of the detection and the management of contradictions and doubts about information.

Index des auteurs