Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil

Rommel N. Carvalho[1], Shou Matsumoto[1], Kathryn B. Laskey[1], Paulo C.G. Costa[1], Marcelo Ladeira[2], and Laécio L. Santos[2]

[1] Department of SEOR / Center of Excellence in C4I
George Mason University
4400 University Drive
Fairfax, VA 22030-4400 USA
{rommel.carvalho,cardialfly}@gmail.com, {klaskey,pcosta}@gmu.edu
http://www.gmu.edu

[2] Department of Computer Science
University of Brasilia
Campus Universitario Darcy Ribeiro
Brasilia DF 70910-900 Brazil
mladeira@unb.br, laecio@gmail.com
http://www.unb.br

Abstract. To cope with citizens' demand for transparency and corruption prevention, the Brazilian Office of the Comptroller General (CGU) has carried out a number of actions, including: awareness campaigns aimed at the private sector; campaigns to educate the public; research initiatives; and regular inspections and audits of municipalities and states. Although CGU has collected information from hundreds of different sources - Revenue Agency, Federal Police, and others - the process of fusing all this data has not been efficient enough to meet the needs of CGU's decision makers. Therefore, it is natural to change the focus from data fusion to knowledge fusion. As a consequence, traditional syntactic methods should be augmented with techniques that represent and reason with the semantics of databases. However, commonly used approaches, such as Semantic Web technologies, fail to deal with uncertainty, a dominant characteristic in corruption prevention. This paper presents the use of probabilistic ontologies built with Probabilistic OWL (PR-OWL) to design and test a model that performs information fusion to detect possible frauds in procurements involving Federal money in Brazil. To design this model, a recently developed tool for creating PR-OWL ontologies was used with support from PR-OWL specialists and careful guidance from a fraud detection specialist from CGU. At present, the task of procurement fraud detection is done manually by an auditor. The number of suspicious cases that can be analyzed by a single person is small. The experimental results obtained with the presented approach are preliminary, but show the viability of developing a tool based on PR-OWL ontologies to automatize this task. This paper also examplifies how to use PR-OWL 2.0 to provide a link between the deterministic and probabilistic parts of the ontology.

F. Bobillo et al. (Eds.): URSW 2008-2010/UniDL 2010, LNAI 7123, pp. 19–40, 2013.

Keywords: Probabilistic Ontology, PR-OWL, Ontology, Procurement, Fraud Detection, Fraud Prevention, Knowledge Fusion, MEBN, UnBBayes.

1 Introduction

A primary responsibility of the Brazilian Office of the Comptroller General (CGU) is to prevent and detect government corruption. To carry out this mission, CGU must gather information from a variety of sources and combine it to evaluate whether further action, such as an investigation, is required. One of the most difficult challenges is the information explosion. Auditors must fuse vast quantities of information from a variety of sources in a way that highlights its relevance to decision makers and helps them focus their efforts on the most critical cases. This is no trivial task. Brazil's Growing Acceleration Program (PAC) alone has a budget greater than 250 billion dollars with more than one thousand projects in the state of Sao Paulo alone[1]. Each of these projects must be audited and inspected by CGU – yet CGU has only three thousand employees. Therefore, CGU must optimize its processes in order to carry out its mission.

The Semantic Web (SW), like the document-based web that preceded it, is based on radical notions of information sharing. These ideas [1] include: (i) the Anyone can say Anything about Any topic (AAA) slogan; (ii) the open world assumption, in which we assume there is always more information that could be known, and (iii) non-unique naming, which acknowledges that different authors on the Web might use different names to define the same entity. In a fundamental departure from assumptions of traditional information systems architectures, the Semantic Web is intended to provide an environment in which information sharing can thrive and a network effect of knowledge synergy is possible. Although a powerful concept, this style of information gathering can generate a chaotic landscape rife with confusion, disagreement, and conflict.

We call an environment characterized by the above assumptions a Radical Information Sharing (RIS) environment. The challenge facing SW architects is therefore to avoid the natural chaos to which RIS environments are prone, and move to a state characterized by information sharing, cooperation, and collaboration. According to [1], one solution to this challenge lies in modeling, and this is where ontology languages such as Web Ontology Language (OWL) come in.

As noted in Section 4 below, procurement fraud detection is carried out within a RIS environment. The ability to deal with uncertainty is especially important in applications such as fraud detection, in which perpetrators seek to conceal illicit intentions and activities, making crisp assertions problematic. In such environments, partial or approximate information is more the rule than the exception.

Bayesian networks (BNs) have been widely applied to information and knowledge fusion in the presence of uncertainty. However, BNs are not expressive enough for many important applications [11]. Specifically, BNs assume a simple attribute-value representation – that is, each problem instance involves reasoning

[1] `http://www.brasil.gov.br/pac/`

about the same fixed number of attributes, with only the evidence values changing from problem instance to problem instance. Complex problems on the scale of the Semantic Web often involve intricate relationships among many variables. The limited representational power of BNs is insufficient for models in which the variables and relationships are not fixed in advance.

To address this weakness of BNs it is common to extend this formalism with approaches based on first-order logic (FOL). FOL is highly expressive but has no built-in capability to reason with uncertainty. To combine the strengths of both approaches, researchers have used FOL expressions to specify relationships among fragments of BNs. The resulting model specifies a probability distribution over many different "ground models" obtained by instantiating the fragments as many times as needed for the given situation and combining into a Bayesian network. Multi-Entity Bayesian Network (MEBN) is an example of this style of language. The ground BN generated after instantiating the variables with domain objects has been called a Situation-Specific Bayesian Network (SSBN). Inference in the SSBN can be performed with a standard belief updating algorithm.

Multi-Entity Bayesian Network (MEBN) logic can represent and reason with uncertainty about any propositions that can be expressed in first-order logic [19]. Probabilistic OWL (PR-OWL), an OWL upper ontology for expressing MEBN theories, is a language for expressing probabilistic ontologies (PO) [21]. The ability to represent and compute with probabilistic ontologies represents a major step towards semantically aware, probabilistic knowledge fusion systems. Although compatibility with OWL was a major design goal for PR-OWL [8], there are several ways in which the initial release of PR-OWL fell short of complete compatibility [4,5]. These shortcomings were addressed in PR-OWL 2.0, which extends PR-OWL by formalizing the relationship between the probabilistic and deterministic parts of a probabilistic ontology [2]. Therefore, PR-OWL 2.0 provides better integration between the probabilistic and deterministic parts of an ontology.

As a result of its focus on the probabilistic aspecs of the ontology, previous literature on PR-OWL did not discuss its relationship to the deterministic part of the ontology as defined by OWL semantics. (e.g., [17,18,22,21,10,9]). This paper uses PR-OWL 2.0 to design and test a model for fusing knowledge to detect possible frauds in procurements involving Federal funds. Unlike previous literature, this paper explicitly addresses the use of PR-OWL 2.0 to provide the link between the deterministic and probabilistic parts of the ontology. We discuss how the new features of PR-OWL 2.0 enable a more natural fusion of information available from multiple sources (see Section 4 for details).

The major contribution of this paper is to clarify how to map properties of entities of a deterministic ontology into random variables of a probabilistic ontology and how to perform hybrid ontological and probabilistic reasoning using PR-OWL 2.0. This approach can be used for the task of information fusion based on reuse of available deterministic ontologies.

This paper is organized as follows. Section 2 introduces MEBN, an expressive Bayesian logic, and PR-OWL, an extension of the OWL language that can represent probabilistic ontologies having MEBN as its underlying logic. Section 3 presents a case study from CGU to demonstrate the power of PR-OWL ontologies for knowledge representation and inferring rare events like fraud. Then, Section 4 describes how to extend this PO to gather information from other sources and perform knowledge fusion in order to improve the likelihood of finding frauds. Finally, Section 5 presents some concluding remarks.

2 MEBN and PR-OWL

Multi-Entity Bayesian Networks (MEBN) [17,20] extend BNs to achieve first-order expressive power. MEBN represents knowledge as a collection of MEBN Fragments (MFrags), which are organized into MEBN Theories (MTheories).

An MFrag contains random variables (RVs) and a fragment graph representing dependencies among these RVs. An MFrag is a template for a fragment of a Bayesian network. It is instantiated by binding its arguments to domain entity identifiers to create instances of its RVs. There are three kinds of RV: context, resident and input. Context RVs represent conditions that must be satisfied for the distributions represented in the MFrag to apply. Input nodes represent RVs that may influence the distributions defined in the MFrag, but whose distributions are defined in other MFrags. Distributions for resident RV instances are defined in the MFrag. Distributions for resident RVs are defined by specifying local distributions conditioned on the values of the instances of their parents in the fragment graph.

A set of MFrags represents a joint distribution over instances of its random variables. MEBN provides a compact way to represent repeated structures, which can then be instantiated as many times as needed to build an actual BN tailored for the specific situation at hand. An important advantage of MEBN is that there is no fixed limit on the number of RV instances, and the random variable instances are dynamically instantiated as needed.

An MTheory is a set of MFrags that satisfies conditions of consistency ensuring the existence of a unique joint probability distribution over its random variable instances.

To apply an MTheory to reason about particular scenarios, one needs to provide the system with specific information about the individual entity instances involved in the scenario. On receipt of this information, Bayesian inference can be used both to answer specific questions of interest (*e.g.*, how likely is it that a particular procurement is being directed to a specific enterprise?) and to refine the MTheory (*e.g.*, each new situation includes additional data about the likelihood of fraud for that set of circumstances). Bayesian inference is used to perform both problem specific inference and learning in a sound, logically coherent manner (for more details see [20,24]).

State-of-the-art systems are increasingly adopting ontologies as a means to ensure formal semantic support for knowledge sharing [6,7,12,3,13,15,28]. Representing and reasoning with uncertainty is becoming recognized as an essential

capability in many domains. In fact, the W3C created the Uncertainty Reasoning for the World Wide Web Incubator Group (URW3-XG) to research the use of uncertainty in semantic technologies. The group was created in 2007 and, one year later, presented its conclusion that standardized representations were needed to express uncertainty in Web-based information [23].

A candidate representation for uncertainty reasoning in the Semantic Web is Probabilistic OWL (PR-OWL) [8], an OWL upper ontology for representing probabilistic ontologies based on Multi-Entity Bayesian Networks (MEBN) [20]. More specifically, PR-OWL is an upper ontology (*i.e.* an ontology that represents fundamental concepts common to various disciplines and applications) for probabilistic systems. It consists of a set of classes, subclasses and properties that collectively form a framework for building probabilistic ontologies.

There are several ways in which the initial release of PR-OWL fell short of fully integrating the deterministic and probabilistic parts of an ontology [4,5]. In fact, Poole *et al.* [27] emphasizes that it is not clear how to match the formalization of random variables from probabilistic theories with the concepts of individuals, classes and properties from current ontological languages like OWL. However, Poole *et al.* [27] says "We can reconcile these views by having properties of individuals correspond to random variables." This is the approach used in PR-OWL 2.0 [2] to integrate MEBN and OWL.

Matsumoto [25] describes a Java implementation of PR-OWL 2.0, including a GUI, API and inference engine in the UnBBayes framework [26]. With this tool it is possible to drag and drop OWL properties into MFrags. The MFrag designer can define and edit the probabilistic definition for that property as shown in Figure 1. This action creates a RV that represents a probability distribution for the OWL property being mapped into the MFrag. Of course, an OWL property with no uncertainty can be mapped with a probability distribution with probability 0 or 1.

With this tool it is possible to drag-and-drop OWL properties into MFrags, which will automatically create a RV, allowing the definition of the probabilistic definition for that property within the context of the MFrag it is defined in as shown in Figure 1. The tool has proven to be a simple, yet powerful, asset for designing probabilistic ontologies and for uncertain reasoning in complex situations such as procurement fraud detection.

3 Procurement Fraud Detection

A major source of corruption is the procurement process. Although laws were enacted to ensure a competitive and fair process, perpetrators find ways to turn the process to their advantage while appearing to be legitimate. To better understand these many, usually creative ways of circumventing the laws, a specialist from CGU has didactically structured the different kinds of procurement frauds that CGU has dealt with in past years. Those different kinds of procurement frauds have resulted in several MFrags built by the authors with the use of the tool UnBBayes.

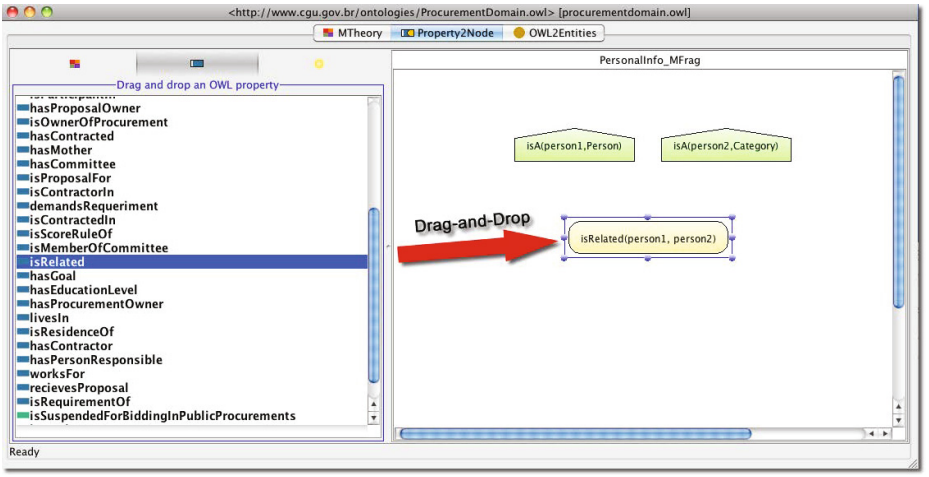

Fig. 1. Drag-and-drop of an OWL property for defining its probabilistic semantics

Different fraud types are characterized by criteria, such as business owners who work as a front for the company, use of accounting indices that are not common practice, among others. Indicators have been established by the CGU specialist to help identify cases of each of these fraud types. For instance, one principle that must be followed in public procurement is that of competition. Every public procurement should establish minimum requisites necessary to guarantee the execution of the contract in order to maximize the number of participating bidders. Nevertheless, it is common to have a fake competition when different bidders are, in fact, owned by the same person. This is usually done by having someone as a front for the enterprise, which is often someone with little or no education. Instead of calling this person a front, a common word used in Brazil is "laranja" (Portuguese for orange)[2].

Computerized support for procurement fraud detection must represent and reason about this kind of domain knowledge. The goal of this case study is to show how to structure the specialist's knowledge in a way that an automated system can reason with the evidence in a manner similar to the specialist. Such an automated fraud detection system is intended to be a decision support system to support specialists in carrying out their tasks. The system could also be used to help train new specialists. The case study focuses on a few selected criteria as a proof of concept. It is shown that the model can be incrementally updated to incorporate new criteria. In this process, it becomes clear that a number of different sources should be consulted to come up with the necessary indicators to create new and useful knowledge for decision makers about procurements.

[2] After a large chain letters hoax that happened in the late seventies in Brazil. People at the losing end were called the "laranjas," while the perpretators were called the "limões" (Portuguese for limes).

Fig. 2. Procurement fraud detection and prevention overview

Figure 2 presents an overview of the procurement fraud detection process. The data for our case study represent several requests for proposal and electronic auctions that are issued by the Federal, State, and Municipal Offices (Public Notices – Data). Our focus is on representing the specialist's knowledge and reasoning through probabilistic ontologies. We assume that analysts collect information (Information Gathering) through questionnaires specifically designed to capture indicators of the selected criteria. These questionnaires can be created using a system that is already in production at CGU. The questionnaire results provide the necessary information (DB – Information). UnBBayes, using the probabilistic ontology designed by experts (Design – UnBBayes), will collect these millions of items of information and transform them into dozens or hundreds of items of knowledge. This will be achieved through logic and probabilistic inference. For instance, procurement announcements, contracts, reports, etc. - an enormous amount of data - are analyzed to obtain relevant relations and properties - a large amount of information. Then, these relevant relations and properties are used to draw conclusions about possible irregularities - a smaller number of items of knowledge (Inference – Knowledge). This knowledge can be filtered so that only the procurements that show a probability higher than a threshold, *e.g.* 50%, are automatically forwarded to the responsible department along with the inferences about potential fraud and the supporting evidence (Report for Decision Makers).

For this proof of concept, the criteria selected by the specialist were the use of accounting indices and the demand for experience in just one contract. There are four common types of indices (acronyms in Portuguese) that are usually used as requirements in procurements (ILC for current ratio, ILG for general

liquidity index, ISG for general solvency index, and IE for indebtedness index). Any other type could indicate a made-up index specifically designed to direct the procurement to some specific company. As the number of uncommon accounting indices used in a procurement increases, the chance of fraud increases. In addition, a procurement specifies a minimum value for these accounting indices. The minimum value that is usually required is 1.0. The higher this minimum value, the more the competition is narrowed, and therefore the higher the chance the procurement is being directed to some enterprise.

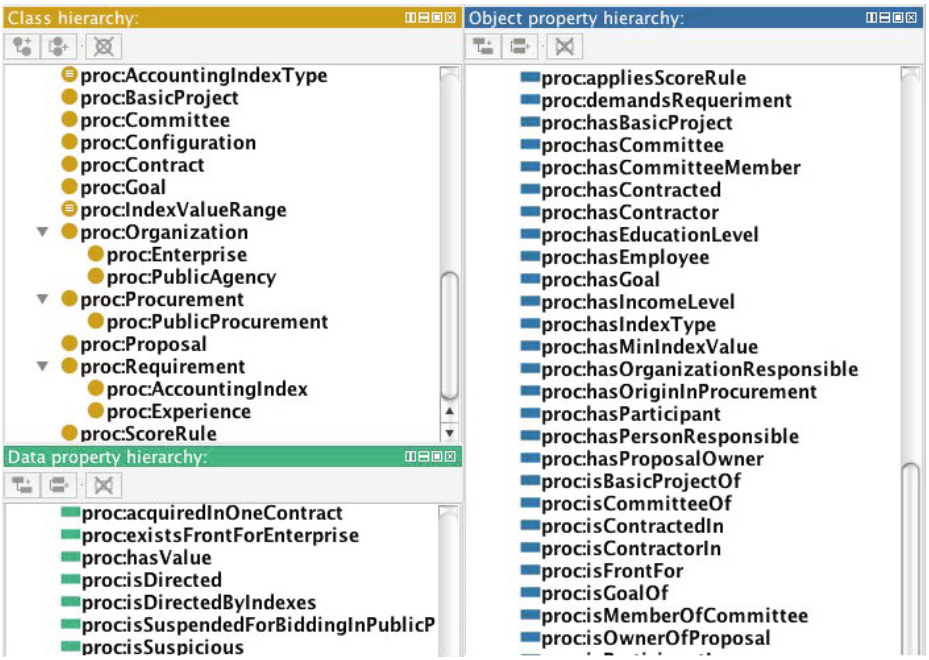

Fig. 3. A few classes, object and data properties of the OWL ontology for the procurement domain

The other criterion, demanding proof of experience in only one contract, is suspect because in almost every case, competence is attained not from a specific contract, but by repeatedly performing a given kind of work. It does not matter whether one has built 1,000 ft^2 of wall in just one contract or 100 ft^2 in 10 different contracts. The experience gained is basically the same.

Before implementing the probabilistic rules described above, we start by looking for an existing ontology that describes the procurement domain. The focus of this chapter is on how to model probabilistic ontologies and not OWL ontologies. Therefore we assume that the ontology depicted in Figure 3 is an existing ontology created by CGU and available at `http://www.cgu.gov.br/ontologies/ProcurementDomain.owl`.

Using UnBBayes PR-OWL 2.0 plugin [25] we are able to drag and drop OWL properties from a Protégé OWL ontology [16] to create corresponding RVs in our MEBN model. This provides a means to define the probabilistic rules described above. These rules were implemented in three different MFrags, built under the supervision of the CGU specialist.

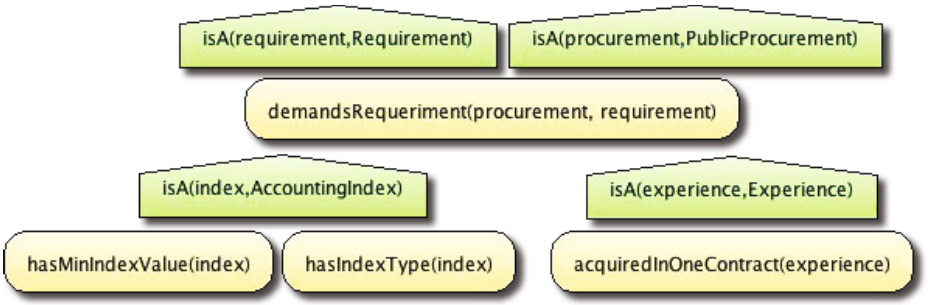

Fig. 4. Procurement Requirement MFrag

The first, Figure 4, represents the criteria required from an enterprise to participate in the procurement. The MFrag represents the type of accounting index (defined by the RV `hasIndexType(index)`, which has individuals of class `AccountingIndexType` as its possible values), as well as the minimum required value (defined by the RV `hasMinIndexValue(index)`, which has individuals of class `IndexValueRange` as its possible values). This MFrag also represents the type of requirement demanded by the procurement (defined by the RV `demandsRequirement(procurement, requirement)`, which has the datatype Boolean as its possible value), as well as whether the procurement demands experience in only one contract (defined by the RV `acquiredInOneContract(experience)`, which has the datatype Boolean as its possible value). As presented in Section 2, there are three kinds of RV: context, resident and input. Context nodes are depicted as green pentagons and represent conditions that should be satisfied for the distributions represented in the MFrag to apply. Resident nodes are depicted as yellow rounded rectangles. Probability distributions for the resident RVs are defined in the MFrag conditioned on the values of the instances of their parents in the fragment graph. Input nodes are depicted as gray trapezoids. They point to RVs that are resident in another MFrag but influence the distribution of RVs resident in this MFrag.

Both `AccountingIndexType` and `IndexValueRange` are nominal classes defined in OWL. The first has `ILC`, `ILG`, `ISG`, `IE`, and `other` as its possible individuals and the second has `between0And1`, `between1And2`, `between2And3`, and `greaterThan3` as its possible individuals.

The second MFrag, shown in Figure 5, represents whether the procurement is being directed to a specific enterprise by the use of unusual accounting indices (defined by the RV `isDirectedByIndexes(procurement)`, which has the

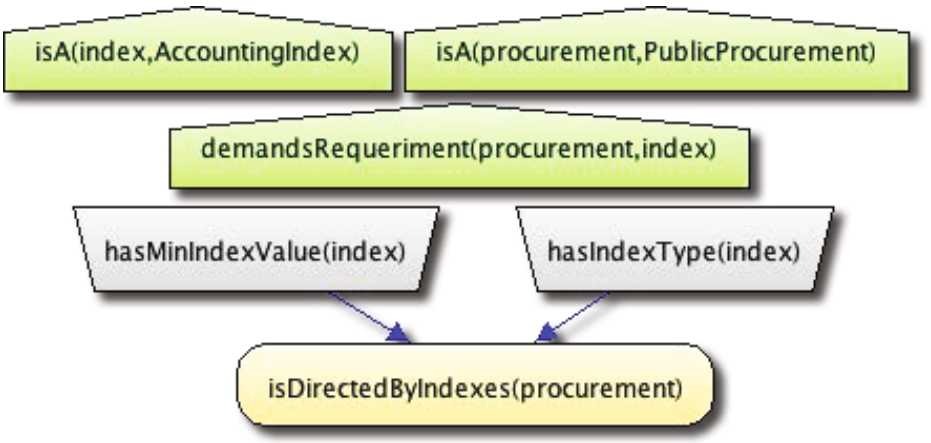

Fig. 5. Directing Procurement by Indexes MFrag

datatype Boolean as its possible value). As explained before, this analysis is based on the type of the index and the minimum value it requires (represented by the `hasIndexType(index)` and `hasMinIndexValue(index)` input nodes, respectively). This evaluation takes into consideration only the indices demanded as requirements for this specific procurement. This is represented by the context node `demandsRequirement(procurement, index)`. Notice that this RV is defined in Figure 4 as `demandsRequirement(procurement, requirement)`, where the second argument is a `Requirement`. However, in Figure 5 the second argument is an `AccountingIndex`. This is a new feature of UnBBayes PR-OWL 2.0 plugin, which allows the use of subtypes in our probabilistic ontology (in our OWL ontology in Figure 3 `AccountingIndex` is defined as a subtype of `Requirement`, and this semantics is inherited in PR-OWL 2.0).

The last MFrag, Figure 6, represents the overall possibility that the procurement is being directed to a specific enterprise (defined by the RV `isDirected(procurement)`, which has the datatype Boolean as its possible value) based on the result of it being directed by the use of unusual indices (represented by the input node `isDirectedByIndexes(procurement)`) and by the requirement of experience in only one contract (represented by the input node `acquiredInOneContract(experience)`), as explained before. Notice that we also make use of subtyping in this MFrag by considering only the experiences demanded as requirements for this specific procurement (in our OWL ontology in Figure 3 `Experience` is defined as a subtype of `Requirement`).

These three MFrags represent knowledge fragments for the domain of procurement fraud detection. The goal is to quantify the probability distribution of the resident RV `isDirected(procurement)` in order to use it to make a decision about whether a procurement is or not suspicious. The next step is to join those MFrags, respecting the logical conditions defined by the context nodes, to generate a Situation-Specific Bayesian Network (SSBN). The algorithm to generate

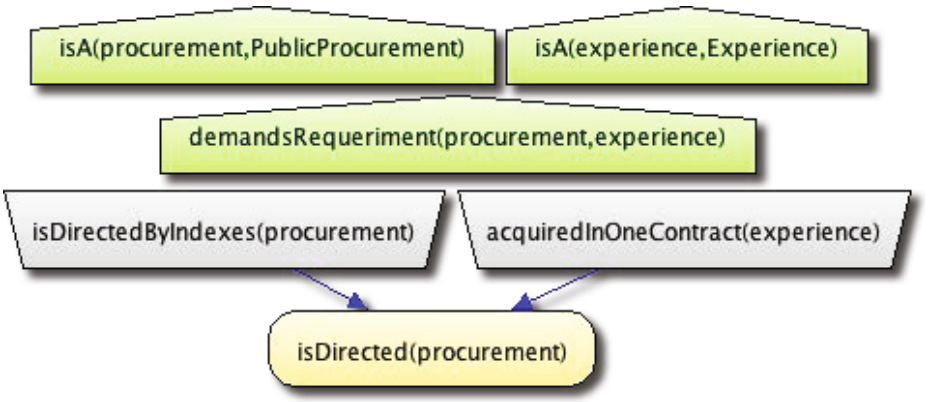

Fig. 6. Directing Procurement MFrag

SSBN proposed by Laskey [20] was implemented in UnBBayes. An SSBN is also a BN, so one can use a belief updating algorithm to do probabilistic inference after entering all available findings. The UnBBayes belief updating is exact and is performed through the strong junction tree algorithm [14].

The probability distributions for resident RVs were estimated with the CGU specialist support based on his knowledge of real cases registered at CGU.

To test the model, two scenarios, that represent the two groups of suspect and non suspect procurements, were chosen from a set of real cases, as shown:

- Suspect procurement (`procurement1`):
  - `index1` = `ILC` >= 2.0;
  - `index2` = `ILG` >= 1.5;
  - `index3` = `other` >= 3.0.
  - It demands experience in only one contract.
- Non suspect procurement (`procurement2`):
  - `index4` = `IE` >= 1.0;
  - `index5` = `ILG` >= 1.0;
  - `index6` = `ILC` >= 1.0;
  - It does not demand experience in only one contract.

The information above was introduced in our model as known individuals and evidence (as simple RDF triples defined in our OWL ontology). After that we queried the system to give us information about the node `isDirected(procurement)` for both `procurement1` and `procurement2`. UnBBayes PR-OWL 2.0 plugin then executed the SSBN algorithm and generated the same node structure as shown in Figure 7, because both procurements have three accounting indices and information about whether the demanding experience is in only one contract or not. However, as expected, the parameters and findings are different giving different results for the query, as shown below:

- Non suspect procurement:
  - 0.01% that the procurement was directed to a specific enterprise by using accounting indices;
  - 0.10% that the procurement was directed to a specific enterprise.
- Suspect procurement:
  - 55.00% that the procurement was directed to a specific enterprise by using accounting indices;
  - 29.77%, when the information about demanding experience in only one contract was omitted, and 72.00%, when it was given, that the procurement was directed to a specific enterprise.

The specialist from CGU analyzed and agreed with the knowledge generated by the probabilistic ontology developed using PR-OWL/MEBN in UnBBayes. By interpreting the resulting probabilities as high, medium, and low chances of something being true, he was able to state that the probabilities represented what he would think when analyzing the same individuals and evidence.

The SSBNs generated for this proof of concept model have the same structure. In practice, the context commonly varies from procurement to procurement in a way that would require SSBNs with different structures. For instance, we have come across several procurements that, in addition to the four common indices, include other indices as well. In this case, if there are two additional indices (`index5` and `index6`), then the resulting SSBN would have two more copies for nodes `hasIndexType(index)` and `hasMinIndexValue(index)`. Standard BNs cannot be used for such problems with varying structures. The ability to make multiple copies of nodes based on a context is only available in a more expressive formalism, such as MEBN.

4 Probabilistic Ontology Knowledge Fusion

From the criteria presented and modeled in Section 3, we can clearly see the need for a principled way of dealing with uncertainty. But what is the role of Semantic Web in this domain? Well, it is easy to see that our domain of fraud detection is a RIS environment. The data CGU has available does not come only from its audits and inspections. In fact, much complementary information can be retrieved from other Federal Agencies, including Federal Revenue Agency, Federal Police, and others. Imagine we have information about the enterprise that won the procurement, and we want to know information about its owners, such as their personal data and annual income. This type of information is not available at CGU's Data Base (DB), but should be retrieved from the Federal Revenue Agencys DB. Once the information about the owners is available, it might be useful to check their criminal history. For that (see Figure 8), information from the Federal Police (Polícia Federal) must be used. In this example, we have different sources saying different things about the same person: thus, the AAA slogan applies. Moreover, there might be other Agencies with crucial information related to our person of interest; in other words, we are operating in an

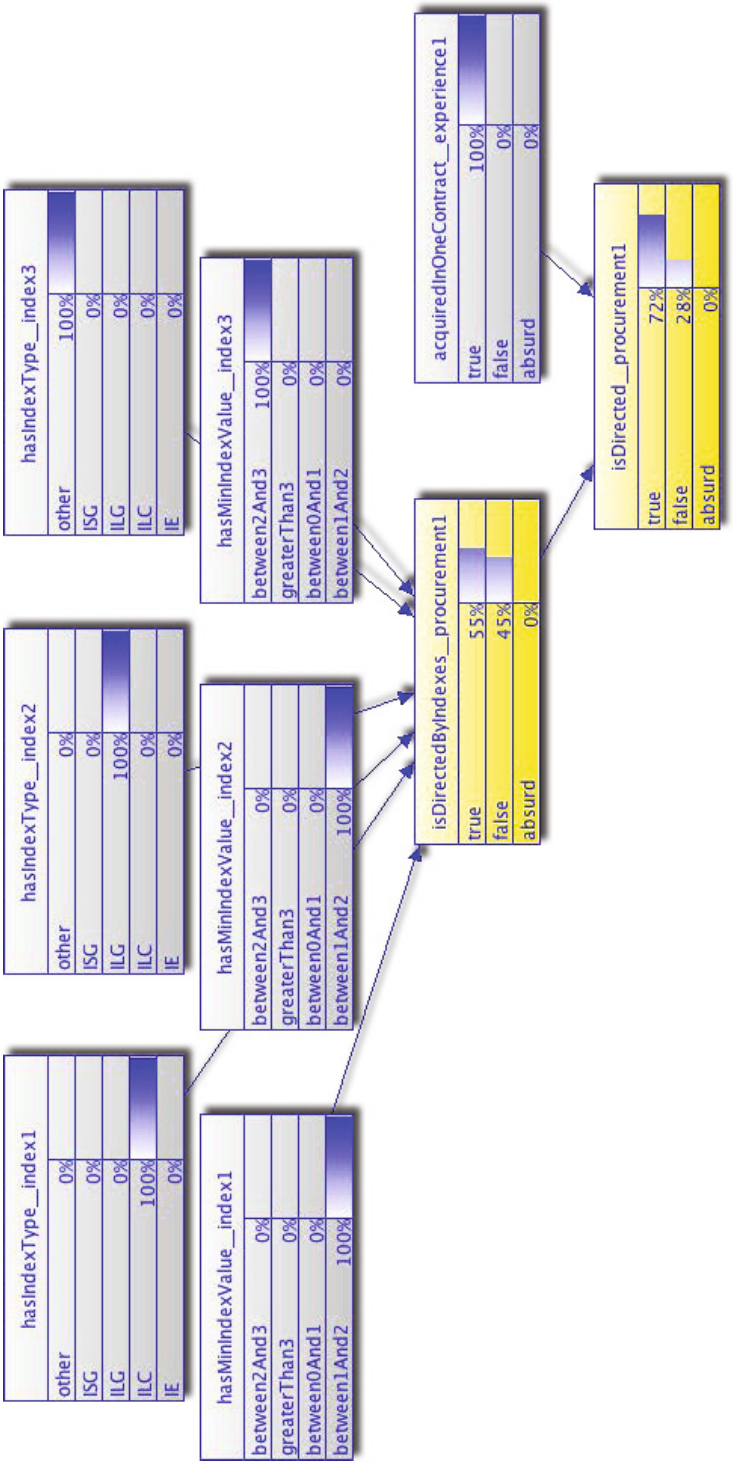

Fig. 7. SSBN generated for query `isDirected(procurement1)`

open-world enviroment which introduces uncertainty. Finally, to make this sharing and integration process possible, we have to make sure we are talking about the same person, who may (especially in case of fraud) be known by different names in different contexts.

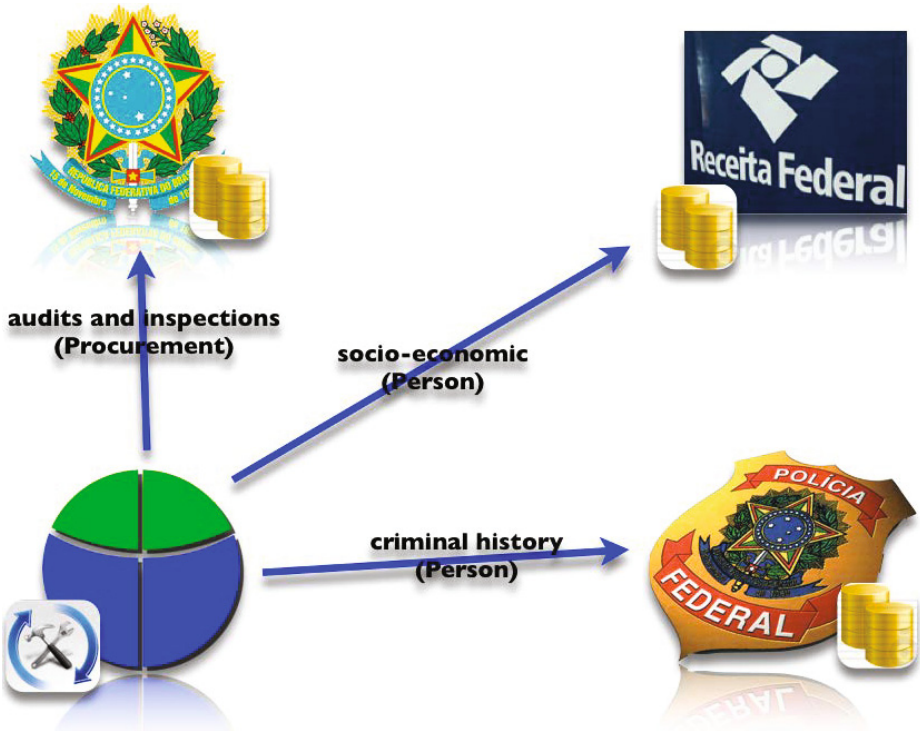

Fig. 8. Knowledge fusion from different Government Agencies DBs

We illustrate the need to fuse knowledge from different sources through the introduction of a new probabilistic reasoning rule. This rule, mentioned in Section 3, addresses the question of whether a person is front for some enterprise. Typically, a person acting as a front has a low annual income, and therefore is unlikely to not own properties such cars and houses. In fact, fronts are often gardeners or maids who work for the person who really makes the decisions for the enterprise.

So, by looking at a person's education level, annual income, and lack of properties (*e.g.*, whether the person has a car) we can determine whether this person is more likely to be a front for the enterprise for which he/she is listed as responsible. However, CGU does not have information about a person's education level, annual income, and property ownership. This information is available, but it is collected by other Federal Agencies. Information about education level can be retrieved from the Education Ministry (MEC). Information about annual

income can be retrieved from the Federal Revenue Agency (Receita Federal). Finally, information about property ownership can be retrieved from the relevant agencies, such as the Department of Motor Vehicles (DENATRAN) for the case of motor vehicles.

CGU has been engaging in collaborations with different Agencies for some years now in order to gather more information that might help identify and prevent frauds in public procurements. In this Section we show how CGU can exploit SW technologies in order to add a new probabilistic rule to our probabilistic ontology and reason with the information provided from other Agencies.

In our proof of concept architecture we assume each Agency has its own ontology with focus on its domain of application. Furthermore, we assume that all Government Agencies use a common ontology with basic concepts for people (name, address, relationship, etc), which is the ontology created by the Federal Government and available at `http://www.brasil.gov.br/ontologies/People.owl`[3]. The Education Ministry (MEC) provides an ontology for education available at `http://www.mec.gov.br/ontologies/Education.owl`. The Department of Motor Vehicles (DENATRAN) provides an ontology for motor vehicle information (*e.g.*, ownership and license) available at `http://www.denatran.gov.br/ontologies/MotorVehicle.owl`. The Federal Revenue Agency (Receita Federal) provides an ontology for internal revenue services available at `http://www.receita.fazenda.gov.br/ontologies/InternalRevenue.owl`.

Because we need to use concepts from all these ontologies to define our new probabilistic rule for identifying a front, we need to import them into our probabilistic ontology. Once they have been imported, we can start creating our MFrags. Below we describe a set of MFrags representing information imported from other ontologies, rules for using this information to determine the likelihood that a person is a front, and using this information to reason about whether a procurement is fraudulent.

Figure 9 depicts an MFrag representing information associated with an enterprise. For our proof of concept, this MFrag contains a single RV for identifying the responsible person for an organization, which is the RV `isResponsibleForOrganization(person, enterprise)`. Although the range of the OWL ontology for the property `isResponsibleForOrganization` is an `Organization`, we define it here as an `Enterprise`, which is a subtype of `Organization`, because our procurement rule concerns enterprises. The ability to reason with subtypes in probabilistic ontologies is a new feature in UnBBayes PR-OWL 2.0 plugin.

Figure 10 shows an MFrag representing information associated with a procurement. The MFrag defines a RV `hasParticipant(procurement, enterprise)` for identifying whether an enterprise is participating in a procurement. Again,

[3] The ontologies presented in this chapter were created by the authors. In order to illustrate the idea of information fusion we will present them as being ontologies created and distributed by different agencies of the Brazilian Government. Thus, the URI provided here is for illustration only.

Fig. 9. Enterprise Information MFrag

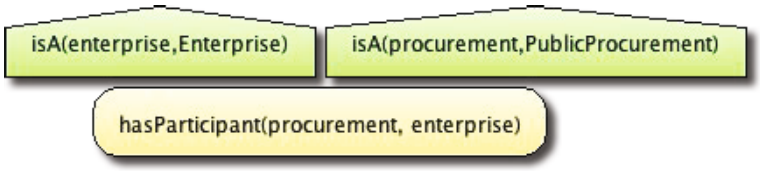

Fig. 10. Procurement Information MFrag

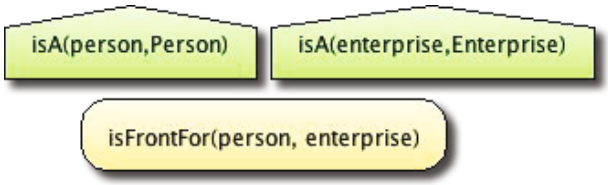

Fig. 11. Front for Enterprise MFrag

we make use of subtypes instead of the more general type defined in our OWL ontology.

The MFrag of Figure 11 defines the RV for verifying whether a person is a front for a specific enterprise, which is the RV `isFrontFor(person, enterprise)`. Here we also make use of subtyping by using `Enterprise` instead of the more general class `Organization`.

Figure 12 presents the main rule for defining whether a person is front for an enterprise. As discussed above, the idea is that if a person is front for an enterprise, then this person is more likely to have a low annual income (defined by the RV `hasIncomeLevel(person)`), no motor vehicle (defined by the RV `hasMotorVehicle(person)`), and little or no education level (defined by the RV `hasEducationLevel(person)`). These RVs are mapped to OWL properties from the Federal Revenue Agency (Receita Federal), Department of Motor Vehicle (DENATRAN), and Education Ministry (MEC) ontologies, respectively. Notice that the only persons analyzed are the ones responsible for that enterprise (constrained by the context node `isResponsibleForOrganization(person, enterprise)`).

Figure 13 defines a RV collecting all information about potential fronts to assess whether there is a front for a given enterprise. This *existential* assertion is a built-in RV in MEBN, but had to be defined manually in UnBBayes because

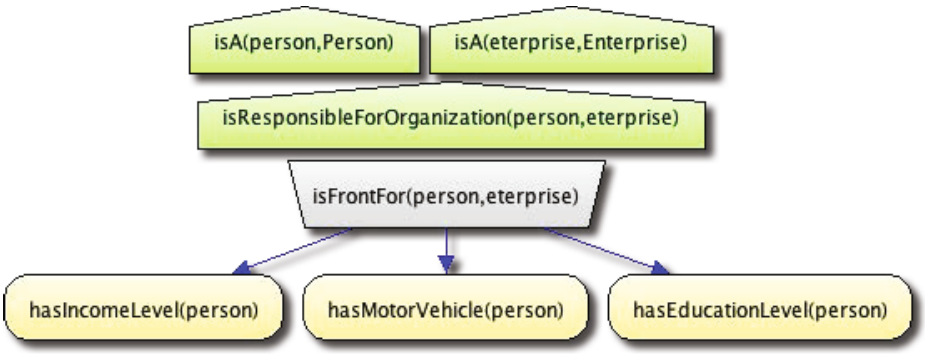

Fig. 12. Personal Information MFrag

this feature of MEBN has not yet been implemented there. The logic of this RV has the same logic as the built-in RV as defined in PR-OWL and MEBN [2]. That is, at least one of the potential fronts for an enterprise actually is a front, then there exists a front for the enterprise. Notice that the only persons included in the existential assertion are those responsible for that enterprise (i.e., the slot fillers are constrained by the context node `isResponsibleForOrganization(person, enterprise)`).

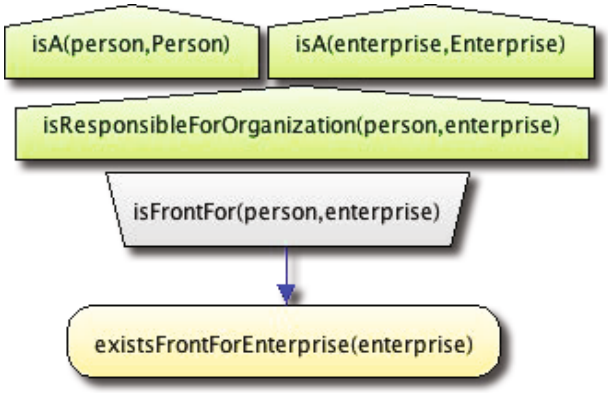

Fig. 13. Exists Front for Enterprise MFrag

Finally, Figure 14 integrates the two major probabilistic rules we have in our probabilistic ontology, namely identifying whether a procurement is being directed for a specific enterprise (represented by the input node `isDirected(procurement)`) and whether an enterprise has a front (represented by the input node `existsFrontForEnterprise(enterprise)`). The resident RV `isSuspicious(procurement)` of this MFrag represents whether the procurement

is suspicious. Notice that the only enterprises analyzed are the ones participating in this procurement (i.e., the slot fillers are constrained by the context node `hasParticipant(procurement, enterprise)`).

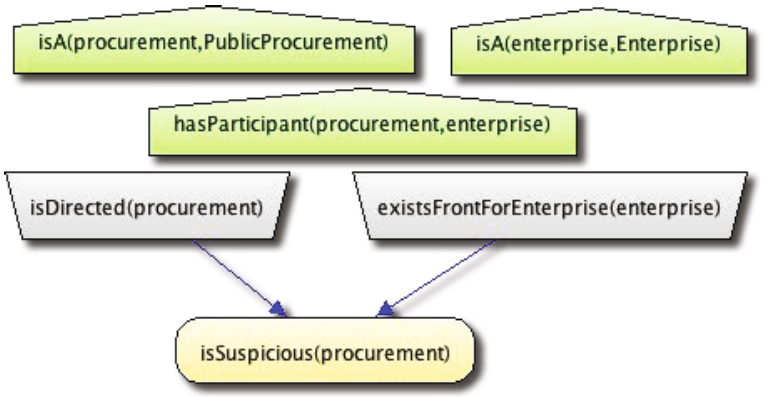

Fig. 14. Suspicious Procurement MFrag

The above MFrags represent modular pieces of knowledge that can be reused when necessary. This fact allows incremental enhancement of the model being designed by the fraud detection analyst. So, Figures 4 to 6 represent the first model of fraud detection based on the information available at CGU records only. The MFrags depicted in Figures 9 to 13 present general probabilistic rules representing concepts that can be aggregated to form the more complex fraud detection model represented in Figure 14 that considers whether the procurement is being directed or whether one of the participants is a front for an enterprise. In either of these cases, the procurement is considered suspicious.

This more complex model was built using both probabilistic models of the procurement being directed (Figure 6) and the existence of a front for an enterprise (Figure 11) which is a participant of the procurement (Figure 10).

To validate our knowledge fusion architecture we published each ontology on a different computer, but accessible via the network. We then have a user enter a query to our fraud detection and prevention ontology, which gathers information from the external ontologies for use in probabilistic inference. Unlike the example described in Section 3, the evidence collected in this Section is fictitious.

Figure 15 presents the SSBN generated with information about `John_Doe` who is responsible for `ITBusiness` and `Jane_Doe` who is responsible for `TechBusiness`. Both enterprises are participating in `procurement1` from Section 3. It can be seen that the information about `Jane_Doe` favors the hypothesis of `procurement1` being suspicious, since she seems to be a front for `TechBusiness`. Information about `John_Doe`, on the other hand, does not favor this hypothesis, since he does not seem to be a front for `ITBusiness`.

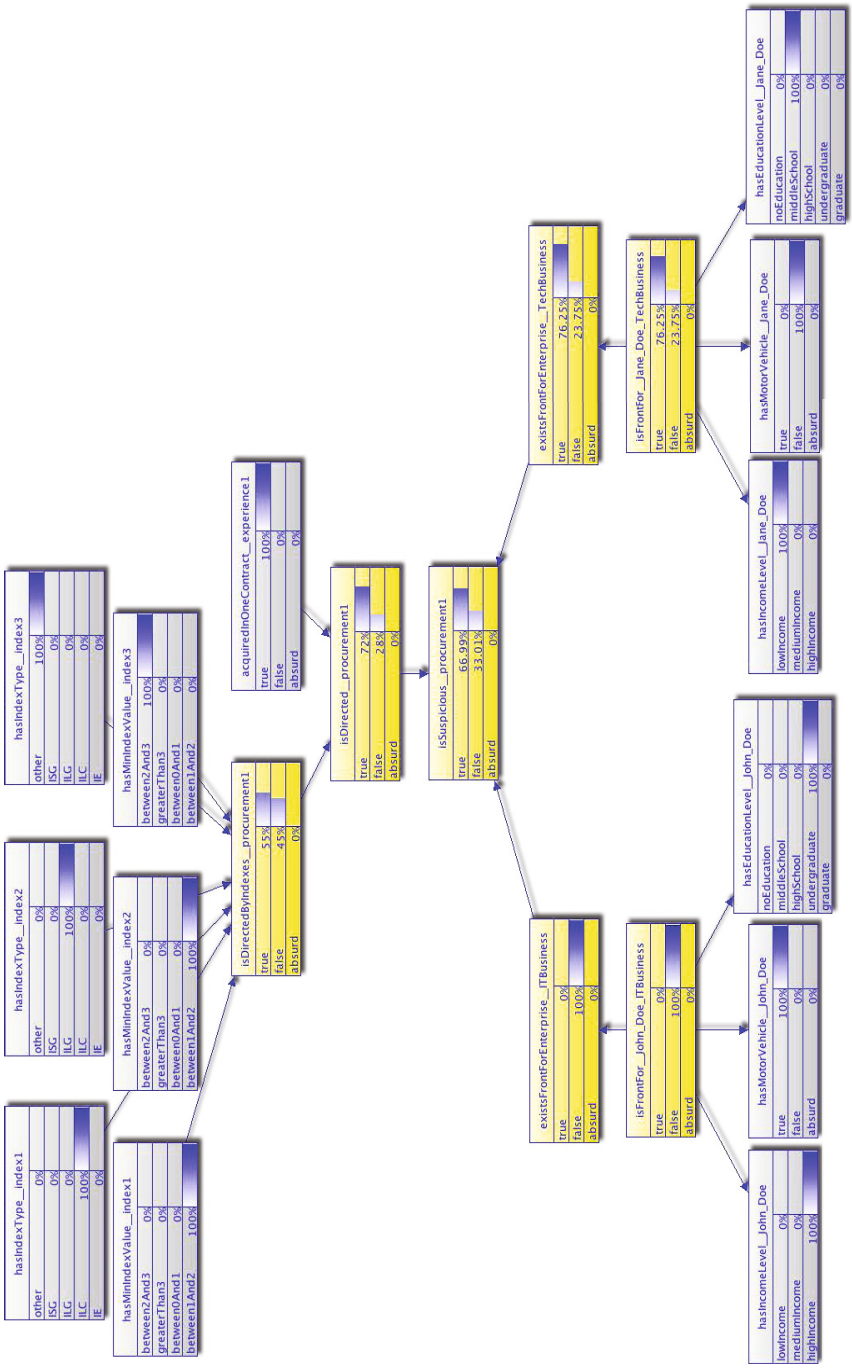

Fig. 15. SSBN generated for query `isSuspicious(procurement1)`

The importance of knowledge fusion is noticed when we compare the results of fusion with what we would be able to infer by considering each source separately. Having only the annual income of Jane Doe gives us a posterior probability of 8.26% that Jane Doe is a front. Considering only the information that Jane Doe does not have a motor vehicle gives us a posterior probability of 0.05% that she is a front. Finally, considering only the information about her education level gives us a posterior probability of 0.07% that she is a front. It is easy to see that separately these items of evidence do not provide strong evidence that Jane Doe is a front. However, if we fuse all the information, we now have strong evidence that Jane Doe is a front. This is shown in Figure 15 by the posterior probability of 76.25% that Jane Doe is a front for `TechBusiness`.

5 Conclusion

The problem that CGU and many other Agencies have faced of processing all the available data into useful knowledge is starting to be solved with the use of probabilistic ontologies, as the procurement fraud detection model showed. In addition to enabling fusion of available information from multiple external sources, the designed model was able to represent the specialist's knowledge for the two real cases we evaluated. UnBBayes reasoning given the evidence and using the designed model were accurate both in suspicious and non suspicious scenarios. These results are encouraging, suggesting that a fuller development of our proof of concept system is promising.

In addition, it is straightforward to introduce new criteria and indicators in the model in an incremental manner. That is, new rules for identifying fraud can be added without rework. After a new rule is incorporated into the model, a set of new tests can be added to the previous one with the objective of always validating the new model proposed, without doing everything from scratch. This was shown in Section 4 where we added a new rule that uses knowledge fusion to identify whether a person is a front for an enterprise. The new rule could be added, without making any changes to the existing MFrags.

Furthermore, the use of this formalism through UnBBayes allows advantages such as impartiality in the judgment of irregularities in procurements (given the same conditions the system will always deliver the same result), scalability (automatization implies expanding the capacity of the specialist to analyze more procurements in a short period of time) and a joint analysis of large volumes of indicators (the higher the number of indicators to examine jointly the more difficult it is for the specialist's analysis to be objective and consistent). The results described here are preliminary, but show that the development of a tool based on PR-OWL ontologies to automatize this task on CGU is viable. This paper also illustrates how to use PR-OWL 2.0 to provide a link between the deterministic and probabilistic parts of the ontology.

As a next step, CGU is choosing new criteria to be incorporated into the designed probabilistic ontology. This next set of criteria will require information from different Brazilian Agencies databases, as shown in Section 4. Therefore,

combining the semantic power of ontologies with the uncertainty handling capability of PR-OWL will be extremely useful for fusing information from different sources.

Acknowledgments. Rommel Carvalho gratefully acknowledges full support from the Brazilian Office of the Comptroller General (CGU) for the research reported in this paper, and its employees involved in this research, especially Mário Vinícius Claussen Spinelli, the domain expert.

References

1. Allemang, D., Hendler, J.A.: Semantic Web for the Working Ontologist. Morgan Kaufmann (2008)
2. Carvalho, R.N.: Probabilistic Ontology: Representation and Modeling Methodology. PhD, George Mason University, Fairfax, VA, USA (2011)
3. Carvalho, R.N., Haberlin, R., Costa, P.C.G., Laskey, K.B., Chang, K.-C.: Modeling a probabilistic ontology for maritime domain awareness. In: Proceedings of the 14th International Conference on Information Fusion, Chicago, USA (July 2011)
4. Carvalho, R.N., Laskey, K.B., Costa, P.C.G.: Compatibility formalization between PR-OWL and OWL. In: Proceedings of the First International Workshop on Uncertainty in Description Logics (UniDL) on Federated Logic Conference (FLoC) 2010, Edinburgh, UK (July 2010)
5. Carvalho, R.N., Laskey, K.B., Costa, P.C.G.: PR-OWL 2.0 - Bridging the Gap to OWL Semantics. In: Bobillo, F., Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2008-2010/UniDL 2010. LNCS (LNAI), vol. 7123, pp. 1–18. Springer, Heidelberg (2013)
6. Chen, H., Wu, Z.: On Case-Based knowledge sharing in semantic web. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2003, pp. 200–207. IEEE Computer Society, Los Alamitos (2003)
7. Chen, H., Wu, Z., Xu, J.: KB-Grid: enabling knowledge sharing on the semantic web. In: International Workshop on Challenges of Large Applications in Distributed Environments, p. 70. IEEE Computer Society, Los Alamitos (2003)
8. Costa, P.C.G.: Bayesian Semantics for the Semantic Web. PhD, George Mason University, Fairfax, VA, USA (July 2005)
9. Costa, P.C.G., Chang, K.-C., Laskey, K.B., Carvalho, R.N.: High level fusion and predictive situational awareness with probabilistic ontologies. In: Proceedings of the AFCEA-GMU C4I Center Symposium, George Mason University, Fairfax, VA, USA (May 2010)
10. Costa, P.C.G., Laskey, K.B., Laskey, K.J.: Probabilistic ontologies for efficient resource sharing in semantic web services. In: Proceedings of the Second Workshop on Uncertainty Reasoning for the Semantic Web, URSW 2006, Athens, GA, USA (November 2006)
11. Costa, P.C.G., Laskey, K.B., Takikawa, M., Pool, M., Fung, F., Wright, E.J.: MEBN logic: A key enabler for network centric warfare. In: Proceedings of the 10th International Command and Control Research and Technology Symposium, 10th ICCRTS. CCRP Publications, McLean (2005)
12. Costa, P.C.G., Chang, K.-C., Laskey, K.B., Carvalho, R.N.: A Multi-Disciplinary approach to high level fusion in predictive situational awareness. In: Proceedings of the 12th International Conference on Information Fusion, Seattle, Washington, USA, pp. 248–255 (July 2009)

13. Dadzie, A.-S., Bhagdev, R., Chakravarthy, A., Chapman, S., Iria, J., Lanfranchi, V., Magalhães, J., Petrelli, D., Ciravegna, F.: Applying semantic web technologies to knowledge sharing in aerospace engineering. Journal of Intelligent Manufacturing 20(5), 611–623 (2008)
14. Jensen, F., Jensen, F.V., Dittmer, S.L.: From influence diagrams to junction trees. In: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (1994)
15. Kings, N.J., Davies, J.: Semantic web for knowledge sharing. In: Semantic Knowledge Management, pp. 103–111. Springer, Heidelberg (2009), doi:10.1007/978-3-540-88845-1_8
16. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 229–243. Springer, Heidelberg (2004)
17. Laskey, K.B., Costa, P.C.G.: Of starships and klingons: Bayesian logic for the 23rd century. In: Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence, UAI 2005. AUAI Press, Arlington (2005)
18. Laskey, K.B., Costa, P.C.G., Wright, E.J., Laskey, K.J.: Probabilistic ontology for Net-Centric fusion. In: Proceedings of the 10th International Conference on Information Fusion, pp. 1–8 (2007)
19. Laskey, K.B., Mahoney, S.M., Wright, E.: Hypothesis management in Situation-Specific network construction. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI 2001, pp. 301–309. Morgan Kaufmann Publishers Inc., San Francisco (2001), ACM ID: 720228
20. Laskey, K.B.: MEBN: a language for First-Order bayesian knowledge bases. Artificial Intelligence 172(2-3), 140–178 (2008)
21. Laskey, K.B., da Costa, P.C.G., Janssen, T.: Probabilistic ontologies for knowledge fusion. In: Proceedings of the 11th International Conference on Information Fusion, pp. 1–8 (2008)
22. Laskey, K.B., da Costa, P.C.G., Janssen, T.: Probabilistic ontologies for Multi-INT fusion. Technical report, George Mason University C4I Center (May 2008)
23. Laskey, K., Laskey, K.B.: Uncertainty reasoning for the world wide web: Report on the URW3-XG incubator group. URW3-XG, W3C (2008)
24. Mahoney, S., Laskey, K.B.: Constructing situation specific belief networks. In: Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, UAI 1998. Morgan Kaufmann, San Francisco (1998)
25. Matsumoto, S.: Framework Based in Plug-ins for Reasoning with Probabilistic Ontologies. M.Sc., University of Brasília, Brasilia, Brazil (forthcoming)
26. Matsumoto, S., Carvalho, R.N., Ladeira, M., da Costa, P.C.G., Santos, L.L., Silva, D., Onishi, M., Machado, E., Cai, K.: UnBBayes: a java framework for probabilistic models in AI. In: Java in Academia and Research. iConcept Press (2011)
27. Poole, D., Smyth, C., Sharma, R.: Semantic Science: Ontologies, Data and Probabilistic Theories. In: da Costa, P.C.G., d'Amato, C., Fanizzi, N., Laskey, K.B., Laskey, K.J., Lukasiewicz, T., Nickles, M., Pool, M. (eds.) URSW 2005-2007. LNCS (LNAI), vol. 5327, pp. 26–40. Springer, Heidelberg (2008)
28. Veres, G.V., Huynh, T.D., Nixon, M.S., Smart, P.R., Shadbolt, N.R.: The military knowledge information fusion via semantic web technologies. Technical report, School of Electronics and Computer Science, University of Southampton (2006)