Trusted Compliance Enforcement Framework for Sharing Health Big Data

Dae-young Kim, Lavanya Elluri, Karuna P. Joshi
University of Maryland, Baltimore County, Baltimore, MD 21250 USA
{leroy.kim, lelluri1, karuna.joshi}@umbc.edu

***Abstract*—COVID pandemic management via contact tracing and vaccine distribution has resulted in a large volume and high velocity of Health-related data being collected and exchanged among various healthcare providers, regulatory and government agencies, and people. This unprecedented sharing of sensitive health-related Big Data has raised technical challenges of ensuring robust data exchange while adhering to security and privacy regulations. We have developed a semantically rich and trusted Compliance Enforcement Framework for sharing large velocity Health datasets. This framework, built using Semantic Web technologies, defines a Trust Score for each participant in the data exchange process and includes ontologies combined with policy reasoners that ensure data access complies with health regulations, like Health Insurance Portability and Accountability Act (HIPAA). We have validated our framework by applying it to the Centers for Disease Control and Prevention (CDC) Contact Tracing Use case by exchanging over 1 million synthetic contact tracing records. This paper presents our framework in detail, along with the validation results against Contact Tracing data exchange. This framework can be used by all entities who need to exchange high velocity-sensitive data while ensuring real-time compliance with data regulations.**

***Index Terms*—Trust Management, Secure Data Sharing, Contact Tracing, Semantic Web, Access Control**

I. Introduction

Recently, the rampancy of the COVID-19 pandemic magnified the importance of distributed trust management of sharing sensitive and high volume and velocity health data. To conduct collaborative prevention and cope with localized mutation of the virus, data exchange between the countries, organizations, and people became indispensable. In this circumstance, trust between various stakeholders is crucial because the result depends on the abundance and quality of the data. Also, it is mandatory to establish a bond of trust that Personally Identifiable Information (PII) is treated securely during the data exchange to expedite the health data exchange.

Regulation can contribute to establishing the trusted ecosystem by protecting the privacy of patients. For example, the Health Insurance Portability and Accountability Act (HIPAA) regulates the healthcare data exchange to protect patient rights in the US. To achieve the goal, it identifies Protected Health Information (PHI) and conditions when stakeholders related to PHI can share the information. However, regulations are not perfect, and there is always the possibility of grey areas for evasion of the law.

There are some obvious cases. First, HIPAA does not cover metadata or non-health data that can indicate a patient's health status. For example, a third-party service can access the transaction history of buying an inhaler from an online shop. Second, the definitions of PHI in 45 CFR § 160.103 stated as "Is created or received by a health care provider, health plan, employer, or health care clearinghouse." Therefore, companies that produce health wearables, mobile apps, fitness trackers, smartwatches, and other analogous products are outside HIPAA's scope unless they partner with covered entities. Consequently, they avoid partnerships with covered entities for less legal liabilities yet collect and process health-related data [1].

We have developed a novel decentralized data exchange framework to enforce healthcare data regulations and evaluate participants' trust based on the compliance history. Our framework facilitates sharing large velocity Health datasets, like vaccination drives or contact tracing, with multiple distributed stakeholders. Our objective is to reason over data access policies on behalf of users and maintain distributed trust of users based on their history in the network. This framework, built using Semantic Web technologies, defines a Trust Score for each participant in the data exchange process and then reasons over policies and regulations combined with the trust scores to control data access. We have developed and integrated the detailed ontology of HIPAA health regulation into our system. We have validated our framework by applying it to the Centers for Disease Control and Prevention (CDC) Contact Tracing Use case by exchanging over 1 million synthetic contact tracing records. This paper will illustrate our results with COVID-19 contact tracing and organizational data exchange as use cases. We believe that our research will contribute to the paradigm change in health data from "needs to know" to "need to share" and solution to the data blocking issues tossed by the 21st Century Cures Act [2].

The rest of the paper is organized as follows: Section II discusses the background and related work. Section III describes our framework in detail, including the knowledge graphs, Trust Score measures, and Trust compliance enforcement. Section IV describes the results of our validation on applying the framework to the CDC contact tracing use case. Finally, we conclude the paper in Section V and describe some of our ongoing work.

II. Related Work

One of the main requirements for our framework is that it must reason over healthcare regulations, such as HIPAA and the 21st Century Cures Act (Cures Act), to achieve trusted compliance enforcement of healthcare organizations. Also, it should provide appropriate security measure which considers the attributes of users who access and share the data. This section listed the current work done in trust management, representing knowledge in HIPAA and Semantic Web technologies.

A. HIPAA Knowledge Representation

A proper health regulation knowledge graph (or ontology) is a critical element of reasoning to achieve automatic compliance enforcement. It is because a knowledge graph explicitly specifies concepts of the real world [3]. Reasoners such as Pellet [4], HermiT [5], and Apache Jena [6] can reason over the ontology to validate the relationships or determine the implicit relationships between instances based on the knowledge graph.

However, there is a lack of research on the HIPAA knowledge graph. In our previous research, we presented a systematic literature review of the HIPAA knowledge graph for the compliance automation [7]. It turned out that there is much research on the HIPAA, but little adopted ontological approaches to analyze it.

Joshi et al. developed a HIPAA knowledge graph for the regulation-compliant cloud services [8]. The ontology demonstrated the three main categories' hierarchical orders: Privacy rule, security rule, and stakeholders. Also, it illustrated the concepts specified in the act in detail to help define security and privacy measures of the healthcare domain.

In the subsequent research, we expanded the ontology and illustrated practicable usage example [7]. The expanded ontology included the Health Information class and in-depth sub-classes of the existing three fundamental classes. Also, it introduced object property between the classes to delineate the relationship between the classes. Therefore, it is possible to describe more context in health data exchange concerning HIPAA.

In paper [9], authors developed a knowledge graph to describe COVID-related security and privacy rules, such as those specified in HIPAA. This ontology expands in designing a HIPAA ontology for automatically populating HIPAA guidelines to access patient records [8]. It facilitates to differentiation of healthcare domain-specific security and privacy measures. HIPAA knowledge graph relates notions required in the regulation not related to COVID. They state that by expanding the knowledge graph merged with COVID guidelines and incorporating the COVID and HIPAA compliance regulations, healthcare organizations can promptly verify and implement HIPAA and COVID privacy requirements. They illustrate the improved and revised ontology by describing all the classes. Health center organizations employing COVID-19 patient information can use their knowledge graph to ensure that privacy policy documents have all the regulations stated by the HIPAA-COVID agreement. Their semantically rich, machine-processable knowledge graph describes all the regulations stated in HIPAA-COVID. They mentioned that their ontology could also assist in recognizing missing guidelines in the health organization's privacy policy, which can then be included as required. However, their work does not restrict access to ontology.

B. Trust Management

There have been many studies to address trust management, especially the quantification of trust. Most of the research followed the distributed trust maintained by the recommendation chain based on Pretty Good Privacy (PGP) [10], [11]. The research proposed trust management and evaluation mechanism consist of direct trust, recommender trust concepts. The critical contribution of the research was to cover insufficient direct trust with a chain of trust, which is discrete trust evaluation scores from other participants in the network. On the other hand, Blaze et al. proposed decentralized trust management based on security policy [12]. The policies consist of simple language specifying trusted actions and trusted relationships.

Until today, most trust management research in the information systems domain follows the general structure of those distributed and decentralized trust models. Their originality came from so-called "more precise" trust evaluation methods based on the recommendation score of trusters on a trustee. However, in the real world, factors that consist of trust are too complicated yet subjective. As a result, it is nearly impossible to unravel all the elements and approximate interplay between them.

Ray and Chakraborty proposed the trust evaluation method that considers time factor [13]. The implicit rationale was Nietzsche's famous quote, "The human being is a forgetful animal." They claimed that trust also decays over time. Their approach well reflected the nature of human trust, but they evaluated trustworthiness based on the experience scores of each event only. Therefore, the evaluation method could not reflect the elements of trust.

Zhou et al. claimed that it is possible to solve this problem with a deep learning method, but they have the same shortcomings regarding the evasiveness of trust factors [14]. Most importantly, they are black-box, and it is a significant impediment to credibility in sensitive issues such as trust.

C. Trust Knowledge Graph

Discrete recommendation scores and policies specifying trusted actions and relationships were not enough to capture complex dimensions of trust in the real world. Therefore, there has been much research to explain trust in a knowledge graph to supplement this shortcoming.

A knowledge graph is an explicit specification of real-world conceptualization [3]. Generally, it consists of triples, which consist of subject, predicate, and object in turn. In other words, a predicate specifies a relationship between subject and object. The final goal of the knowledge graph is to create a map that illustrates relationships between real-world concepts. Some studies on building knowledge graphs for trust include -

- Towards an ontology of trust [15]

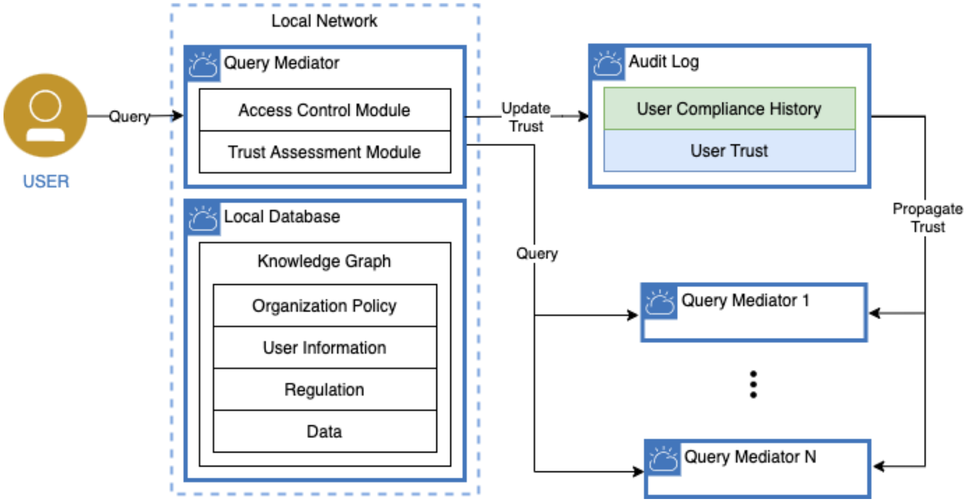


Fig. 1. System Architecture includes Query mediator, Knowledge graphs and Audit Log

- An Ontology of Trust – Formal Semantics and Transitivity [16]
- Trust Networks on the Semantic Web [17]
- Ontology based Approach in Knowledge Sharing Measurement [18]
- A Trust Ontology for Semantic Services [19]

D. Semantic Web

We utilized Semantic Web technologies to develop our system's knowledge graph and the reasoning component. These enable us to build the schema using W3C standardized languages that support our design requirements, including interoperability, sound semantics, Web integration, and the availability of tools and system components. It is possible to model the classes of data and relationships between them using semantic web technologies. As a result, the information is stored in a machine-understandable format, allowing machines to identify the correct context of data usage or retrieval. Resource Description Framework (RDF) [20] and Web Ontology Language (OWL) [21] are the most popular Semantic Web data modeling languages. Among them, we adopted OWL for knowledge graph language.

OWL has well-defined semantics grounded in first-order logic and model theory, allowing programs to draw inferences with the assurance that the subsequent interpretation is sound. Therefore, it was possible to satisfy our most fundamental requirement - a representation that supports interoperability at both the syntactic and semantic levels to facilitate easy data exchange.

Also, OWL is built on basic Web standards and protocols and is evolving to remain compatible with them. It is possible to embed RDF and OWL knowledge in HTML pages, and several search engines (including Google) will find and process some embedded RDF. Furthermore, it integrates well with the Web and Cloud, becoming the dominant technology for today's digital health systems. In other words, it can provide standard semantics of service information and policies enabling all agents who understand essential Semantic Web technologies to communicate and use each other's data and services effectively.

III. Trusted Compliance Enforcement Framework

Our framework consists of two main components: Query Mediators (QMs) and Audit Log (AL). QM gets user input queries and sends out a query to other QM in the network to retrieve healthcare data. During the process, QM reasons over related regulations, organization policy, target data attributes, and user attributes to check that the user is compliant with regulation. Related regulations include, but are not limited to HIPAA, HITECH, Cares Act, and 21st Century Cures Act. When it turns out that the user violated one or more regulations, it halts the data retrieval process and records violation details in the AL. At the same time, QM evaluates the trust value of the user based on the updated compliance history and updates it to AL in real-time.

QM has two components: Access Control Module (ACM) and Trust Assessment Module (TAM), and it has access to knowledge graph databases in the same local network. First, Knowledge graphs are a fundamental part of the system, and the other two modules refer to the knowledge graphs to accomplish their primary purpose. The knowledge graph includes organization policy, user trust, user information, regulation, and data ontology. Second, ACM decides whether or not a user can access specific data based on the regulations, organization policy, and attributes of the user and target data. To achieve this, ACM reasons over the knowledge graph of them with a reasoner. If the user violates the regulation, ACM records the compliance history and updates it to AL. Third, TAM assesses the trustworthiness of the user based on the latest compliance history in real-time and updates it to AL. We assume that QM always functions correctly and is trusted by every other QMs.

AL stores user compliance history and user trust. The green box in figure 1 indicates that user compliance history is synced between a QM and AL and not propagated to other QMs. On

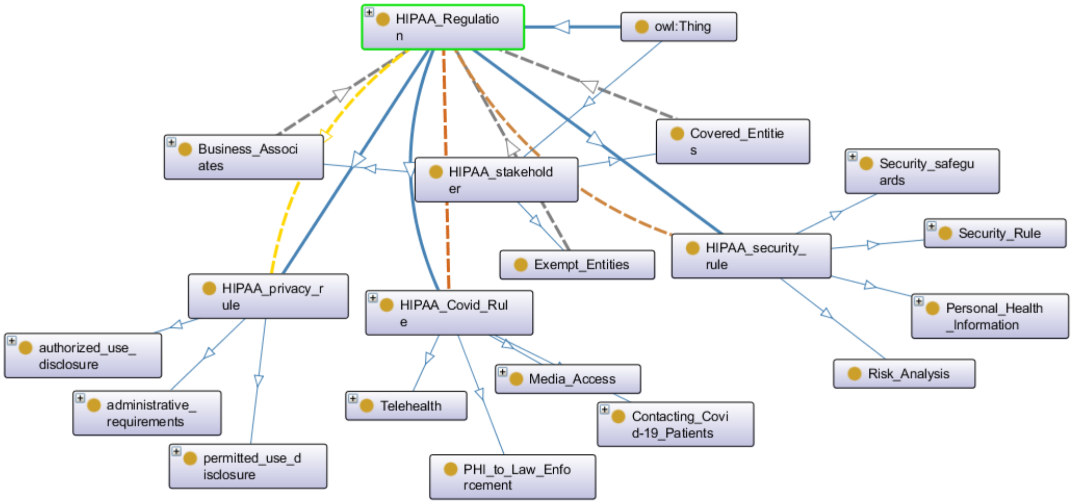


Fig. 2. HIPAA Ontology describes relationships between HIPAA stakeholders COVID-19 related provisions

the other hand, the blue box indicates that user trust is also propagated to other QMs.

A. Knowledge Graph

The knowledge graph is an essential part of the framework. It provides the underlying knowledge base when the other three components of the TM process their duties. We included HIPAA, CDC contact tracing data elements, and trust ontology in the knowledge graph in this research. Trust ontology will be illustrated in the subsection III-B.

1) HIPAA Knowledge Graph: As part of our previous work [9], we identified the most important words in the HIPAA COVID-19 regulation. Most of the semantically similar key terms associated with regulation must be referred in an organization's privacy policy. Therefore, the occurrence of these terms or words associated with them is a significant indication of an organization policy's agreement with the HIPAA regulations [22]. We assessed the occurrence of HIPAA key terms and associated terms in a privacy policy document. We use the vector representation of key terms to classify semantically similar terms in a privacy policy. We demonstrated this procedure by assessing the occurrence of HIPAA key terms in the privacy policy documents of ten health centers and organizations that deal with COVID-19 patient data. Also, we showed the occurrence of semantically similar HIPAA Key Terms in all the organizational privacy policies. The higher occurrence of HIPAA key terms or semantically similar phrases in an organization's privacy policies implies that the privacy policy is more in compliance with the HIPAA guidelines.

To develop the HIPAA ontology, we used the extracted key terms from the HIPAA regulation document. We then extracted the rules from the HIPAA regulation that addresses COVID-19 rules. The rules were then examined in a bag-of-words model. First, we removed the stop words from the list of key terms. We also got rid of certain words modal verbs like 'should,' 'can,' 'could,' 'must' 'will,' 'shall.' These modal words were used to extract guidelines represented in deontic logic from the organizational policies [23]. Finally, we identified the most commonly occurring terms from the remaining list of words in the HIPAA COVID-19 regulation. This list of words is the key terms in the HIPAA repository related to COVID-19. These key terms facilitated us in creating the HIPAA knowledge graph schema as shown in figure 2. Also, these key terms helped us in checking compliance with organizational privacy policies.

2) CDC Contact Tracing Knowledge Graph: We designed a contact tracing knowledge graph based on the Interim Guidance on Developing a COVID-19 Case Investigation & Contact Tracing Plan provided by CDC [24]. The primary purpose of the ontology is to explain patients involved in COVID-19 cases.

Therefore, as illustrated in the figure 3, seven classes that represent sub-categories of the information describe a patient. The seven classes are contact tracing, interview, locating information, pre-existing conditions, risk factors, SARS-CoV-2 test, and symptoms and clinical course.

Bulleted lists inside each class are data properties related to the class. Objects of the data properties follow types and codes specified in Appendix C - Data Elements for Case Investigation and Contact Tracing Forms of the CDC document - open text, date, numeric and categorical value. For example, the data property "Loss of sense of smell" has categorical values Y, N, U, and, R which stand for yes, no, unknown, and refused. In the case of SARS-CoV-2 tests, they can have Pos, Neg, Equi, or Unk for positive, negative, equivocal, or unknown.

B. Trust Management

1) Definition of Trust: The majority of trust management theory follows a philosophy that we seldom trust people completely. The philosophy claim that "A trusts B to do X" [25] or "A trusts B with valued item C" [26]:

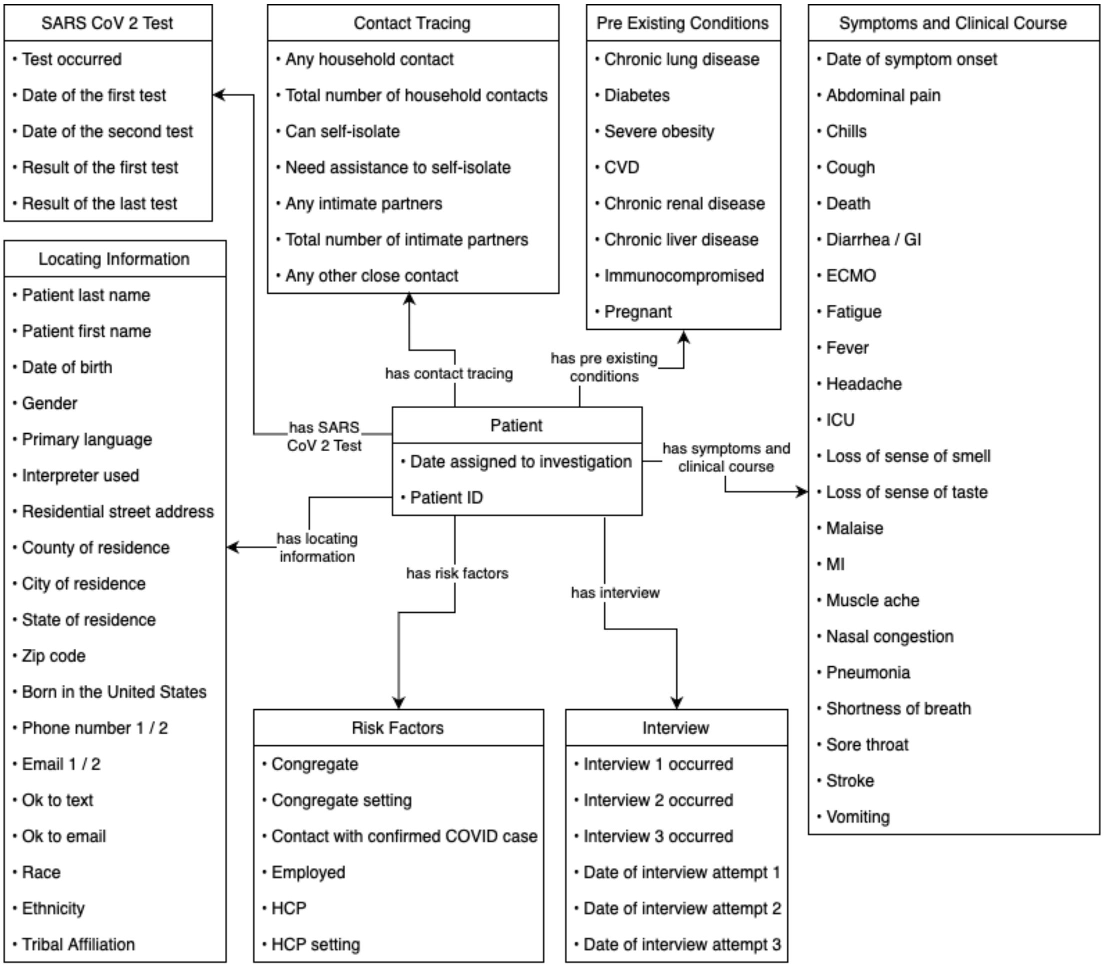


Fig. 3. CDC Contact Tracing Ontology illustrates data elements for case investigation and contact tracing

Definition 1. *A trust relationship is not absolute. (Hardin, Baier)*

Grandison & Sloman adopted similar idea to information system and claimed that [27] trust depends on contexts:

Definition 2. *Trust is the firm belief in the competence of an entity to act dependably, securely, and reliably within a specified context. (Grandison & Sloman)*

For the definition of distrust, Ray & Chakraborty put more resolute nuance to exclude ambivalence in making a decision based on the Grandison & Sloman's definition [13]:

Definition 3. *Distrust is defined as the firm belief in the incompetence of an entity to act dependably, securely, and reliably within a specific context. (Ray & Chakraborty)*

In the previous research of them, the context was the solution to describe the relativity of trust. They defined context as a situation, specific action, or service. It was a static environmental factor that we could quickly identify and classify. For example, we can say trust relationships in the context of data exchange and modification should be different.

Our research follows the abstract idea of past research but unfolds it from a different point of view. We believe that genuine trust or distrust is the mixture of themselves. We agree that the context is the crucial factor to represent relativity of trust, but in different connotations defined below:

Definition 4. *Trustworthiness results from a decision-making process that involves trust and distrust in an entity considering various elements that constitute a trust.*

Definition 5. *For each context of trust, a different combination of trust elements affects the trust assessment.*

2) Trust Representation: This research follows the trust representation of Ray & Chakraborty [13]. They adopted vector form to represent trust between truster A and trustee B in a particular context c:

$$A \xrightarrow{c} B \tag{1}$$

3) Trust Assessment: Let C denote the set of all context that will be involved in trust assessment. For $i \in \mathbb{N}$, c_i is distinct context of trusts in C. Also, let E denote the set of elements which constitutes trust. For $j \in \mathbb{N}$, e_i is distinct element of trust. For each context of trust, different subset of E is taken into account during the test assessment process. Lastly, let W denote the set of weights of elements. For $i \in \mathbb{N}$, w_i will be the weight of element e_i.

$$C = \{c_i \in \mathbb{R}, i \in \mathbb{N} | -1 \leq c_i \leq 1\} \tag{2}$$

$$\begin{gathered} E = \{j \in \mathbb{N} | e_j = \textit{a distinct element of trust}\} \\ E \supset \{i \in \mathbb{N} | E_i = \textit{subset of } E\} \end{gathered} \tag{3}$$

$$W = \{w_i \in \mathbb{R}, i \in \mathbb{N} | 0 \leq w_i \leq 1 \sum_{i=1}^{n} w_i = 1\} \tag{4}$$

Therefore, it is possible to specify Trust Assessment Conditions (TAC) that illustrates which elements of trust is considered in what context:

$$TAC = (c_i, E_i) \tag{5}$$

Finally, based on the TAC, there should be a Trust Assessment Function (TAF) that assesses A's trustworthiness based on the TAC. The f denotes arbitrary formula which a policymaker will define. It also illustrates the weights of each trust element.

$$TAF = f(TAC) = f(E_k) \tag{6}$$

TAF elucidates interactions of elements during the trust assessment process. For example, let there be a trust element "identity trust" regarding the identity of a user in the system. In a normal situation, people usually refer to the user's role and organization to check the user's identity. In this case, if one regards the user's role in what organization is important, it is possible to represent interactions of identity trust of role (I_R) and organization (I_O) as $I_O \times I_R$. In contrast, if one treats each role equally regardless of the organization, the interactions of I_R and I_O can be $I_R + I_O$.

Since this research aims to provide framework and proof of concepts, we defined the simplest form of TAF in the equation 7, the weighted sum of each trust element score. Therefore, the maximum trust score is 1.

$$T(CA, B) = (CA \xrightarrow[E]{C} B) = \sum_{i=1}^{n} w_i e_i \tag{7}$$

T is the trustworthiness of B in the perspective of CA, and e_i and w_i are trust elements and their weight. We assumed that the trust score evaluations of each element are available in advance. By this equation, CA can set the importance of each context and assess overall trustworthiness.

4) Trust Ontology: Based on the definition 4, we specified three contexts of trust, which are identity trust, behavioral trust, veracity for the proof of concept in figure 4. The rationale behind the three contexts is that they are the most basic information that humans seek when identifying the trustworthiness of others.

Also, we classified which components of the system can be evaluated by what element of trust. For example, the user class has behavioral trust because the user is the subject of action in the system. On the other hand, the organization and role class have identity trust, and it will affect the overall trust of the user because every user has a role and organization.

C. Trusted Compliance Enforcement

Suppose we have reviews of organizations' policy and scores of the three trust elements, we can enforce compliance with the user's trust level. Figure 5 illustrates the access control module that enforces compliance and limits the access based on the trustworthiness of the user. First, it takes input query from the user and injects SPARQL triple that checks if the user's organization is compliant to HIPAA based on the purpose of the query. Then, it injects triples that cover trust assessment based on the purpose of the query.

For example, a junior physician A in a hospital wants to query a list of patients whose last result of SARS-CoV-2 is positive to process PHI:

```
SELECT DISTINCT ?patient
WHERE {
  ?patient ctt:hasSARSCoV2Test ?test.
  ?test ctt:ResultOfLastSARSCoV2Test 'Pos'.}
```

The compliance checker and trust checker inject triples that can enforce PHI compliance and trust level policy. Triples for each purpose are commented on in the SPARQL query below.

```
SELECT DISTINCT ?patient
WHERE {
  # compliance
  ctt:A ctt:hasOrganization ?hospital.
  ?hospital ctt:PHI ?phi_compliance.
  # trust
  ctt:A ctt:BehavioralTrust ?behavior.
  ?org ctt:IdentityTrust ?org_identity .
  ctt:A ctt:hasRole ?role.
  ?role ctt:IdentityTrust ?role_identity.

  BIND(0.1 * ?behavior + 0.6 * ?org_identity +
    0.3 *?role_identity AS ?tscore)
  FILTER(?tscore > 0.75)
  # input query
  ?patient ctt:hasSARSCoV2Test ?test.
  ?test ctt:ResultOfLastSARSCoV2Test 'Pos'.}
```

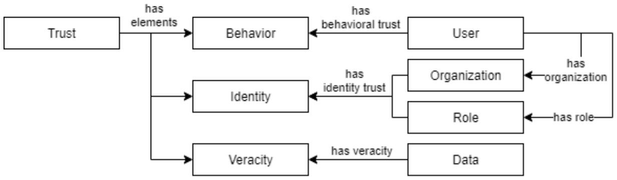


Fig. 4. Trust Ontology describes relationships between trust elements and system components

IV. Validation using Contact Tracing Usecase

To demonstrate the proof of concept of our research, we created datasets for use cases and applied them against our framework. We created a large synthetic dataset of 1 million records based on the CDC Contact tracing, HIPAA, and Trust ontologies.

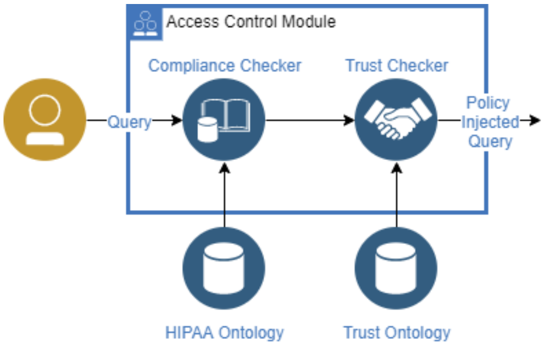


Fig. 5. Detailed view of the Access Control Module, which injects policy queries to an input query

A. Ontology Integration

To reason over user compliance status and trust level during the query process, we integrated the three ontologies illustrated earlier in this paper - CDC contact tracing data element, HIPAA, and trust ontologies as illustrated in figure 6. It was essential to find the intersection of ontologies, especially between the HIPAA and trust ontology. It turned out that stakeholder classes in the HIPAA ontology have characteristics of organization class in the trust ontology. In most cases, covered entities and business associates - stakeholders of HIPAA - are organizations dealing with protected health information. Therefore, in the integrated ontology, stakeholder has identity trust.

B. Synthetic Data Generation

We used the Python 3 script with *lxml* library for synthetic data generation to create an OWL file in XML format. The data property values of symptoms, risk factors, and pre-existing conditions are from the CDC official contact tracing guideline [24]. In addition, we followed HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language, and Disability Status [28] for the values of race and ethnicity. Lastly, we used the *faker* library to generate other data such as first name, last name, email, phone number, and address. The script randomly generated each patient's data, and it also covered dependency between the data, for example, gender and pregnancy status. For the efficiency of data processing, we divided the data into 20 files of 442 MB, and the overall size of the data was 8.63 GB.

C. Data Ingestion

We uploaded our data to Apache Jena Fuseki server running in the Docker container. The default 1200 MB of Java heap size was too small to process the data, so we set the maximum memory allocation pool as 9 GB. Another reason we set a massive amount of heap size is the data modification algorithm of TDB, which is the database that supports the Jena Fuseki server. TDB uses write-ahead logging, so new data is written to an on-disk journal and kept in memory. Therefore, in the worst case, all the data files remain in memory during the data posting process and consume 8.63 GB of memory space. The average time taken to upload each file was 117.6309 s.

D. Use Case Scenario

1) Department of Education querying K-12 school student patients information: The first use case scenario got its motif from the official CDC document "Considerations for Case Investigation and Contact Tracing in K-12 Schools and Institutions of Higher Education (IHEs) [29]." Senior staff John in the Department of Education wants to query a record of patients under age 18 whose first SARS-CoV-2 test result is positive. The purpose of the query is to retrieve the Date of Birth (DOB) and state of residence and create a report to decide whether the cancellation of classes or closure of buildings and facilities is necessary for each state. In this case, he needs PHI compliance and a trust score threshold for PHI compliance of 0.8. Weights for behavior, role identity, and organization trust are 0.1, 0.2, and 0.7.

```
SELECT DISTINCT  ?patient ?dob ?state
WHERE {
  # HIPAA PHI Compliance
  ctt:John ctt:hasOrganization ?org.
  ?org ctt:PHI ?phi_compliance.

  # Trust policy
  ctt:John ctt:BehavioralTrust ?behavior.
  ctt:John ctt:hasRole ?role.
  ?org ctt:IdentityTrust ?org_identity .
  ?role ctt:IdentityTrust ?role_identity.

  BIND(0.1 * ?behavior + 0.2 * ?role_identity
    + 0.7 *?org_identity AS ?tscore)
  FILTER(?tscore > 0.8)

  # Input query
  ?patient ctt:hasSARSCoV2Test ?test.
  ?patient ctt:hasLocatingInformation ?loc.
  ?test ctt:ResultOfFirstSARSCoV2Test 'Pos'.
  ?loc ctt:StateOfResidence ?state.
  ?loc ctt:DOB ?dob.
  FILTER (?dob >
    "2003-09-04T00:00:00"^^xsd:dateTime)}
```

2) Insurance company querying potential COVID-19 patients with pre-existing conditions: In the second scenario, junior staff Jane who works for B&B Insurance Company wants pre-existing conditions of patients whose first SARS-CoV-2 test result is negative. The exact pre-conditions she wants are patients' diabetes and chronic lung disease status. Her primary purpose is to summarize the information and plan strategic marketing to avoid states with high COVID-19 risks. Her organization must have media access compliance, and the trust score threshold is 0.6. Weights for behavior, role identity, and organization trust are 0.5, 0.2, and 0.3.

```
SELECT
DISTINCT
  ?patient ?diabetes ?lung_disease ?state
WHERE {
  # HIPAA Media Access Compliance
  ctt:Jane ctt:hasOrganization ?org.
```

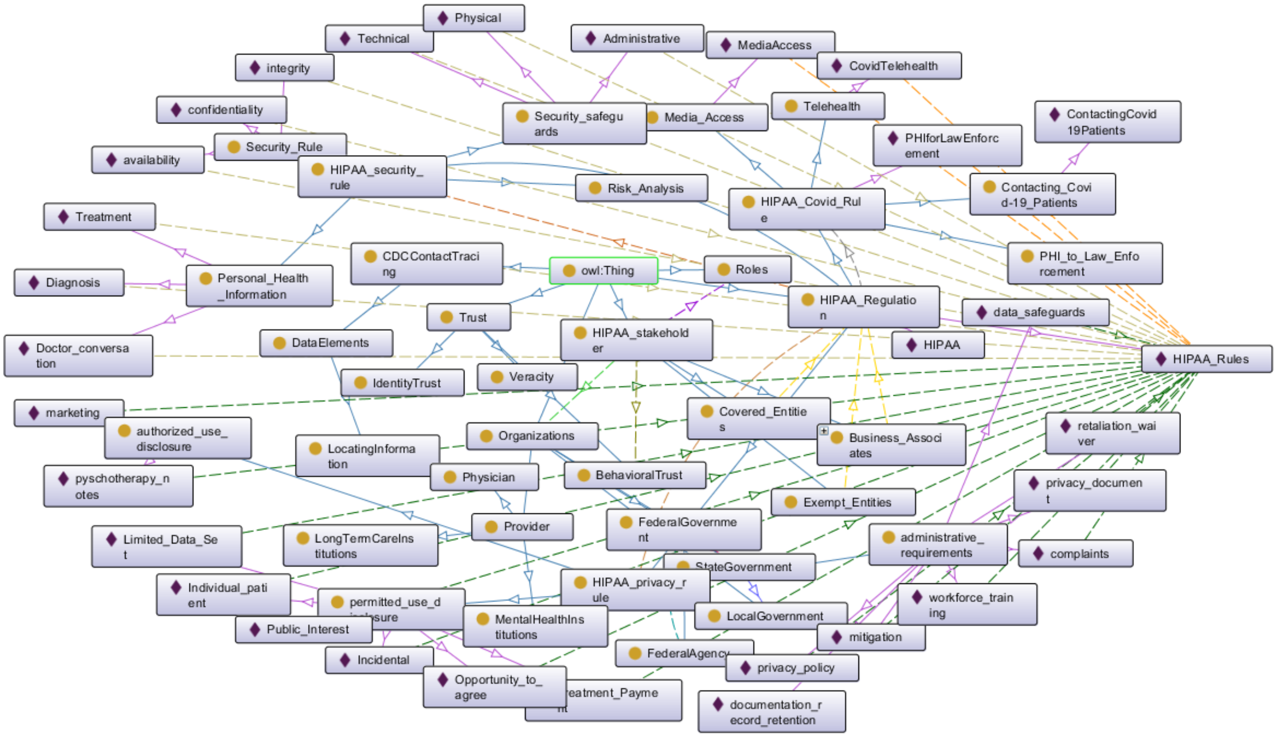


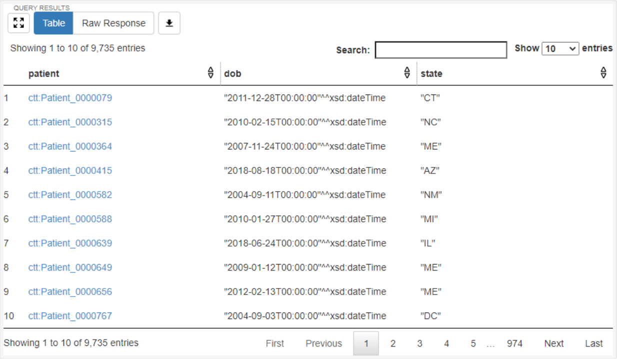
Fig. 6. Merged Ontology incorporates HIPAA Ontology, Trust Ontology, and CDC Contact Tracing Ontology

```
?org ctt:DP_MediaAccess ?media_compliance.

# Trust policy
ctt:Jane ctt:BehavioralTrust ?behavior.
ctt:Jane ctt:hasRole ?role.
?org ctt:IdentityTrust ?org_identity .
?role ctt:IdentityTrust ?role_identity.

BIND(0.5 * ?behavior + 0.2 * ?role_identity
  + 0.3 *?org_identity AS ?tscore)
FILTER(?tscore > 0.6)

# Input query
?patient ctt:hasSARSCoV2Test ?test.
?patient ctt:hasPreExistingConditions ?con.
?patient ctt:hasLocatingInformation ?loc.
?test ctt:ResultOfFirstSARSCoV2Test 'Neg'.
?con ctt:Diabetes ?diabetes.
?con ctt:ChronicLungDisease ?lung_disease.
?loc ctt:StateOfResidence ?state.
FILTER(?diabetes='Y' || ?lung_disease = 'Y')}
```

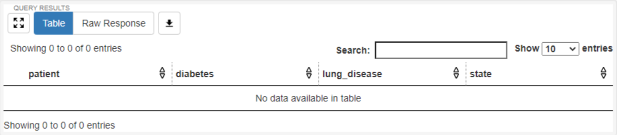


Showing 1 to 10 of 9,735 entries

	patient	dob	state
1	ctt:Patient_0000079	"2011-12-28T00:00:00"^^xsd:dateTime	"CT"
2	ctt:Patient_0000315	"2010-02-15T00:00:00"^^xsd:dateTime	"NC"
3	ctt:Patient_0000364	"2007-11-24T00:00:00"^^xsd:dateTime	"ME"
4	ctt:Patient_0000415	"2018-08-18T00:00:00"^^xsd:dateTime	"AZ"
5	ctt:Patient_0000582	"2004-09-11T00:00:00"^^xsd:dateTime	"NM"
6	ctt:Patient_0000588	"2010-01-27T00:00:00"^^xsd:dateTime	"MI"
7	ctt:Patient_0000639	"2018-06-24T00:00:00"^^xsd:dateTime	"IL"
8	ctt:Patient_0000649	"2009-01-12T00:00:00"^^xsd:dateTime	"ME"
9	ctt:Patient_0000656	"2012-02-13T00:00:00"^^xsd:dateTime	"ME"
10	ctt:Patient_0000767	"2004-09-03T00:00:00"^^xsd:dateTime	"DC"

Showing 1 to 10 of 9,735 entries

Fig. 7. John's query result



Showing 0 to 0 of 0 entries

patient	diabetes	lung_disease	state
No data available in table			

Showing 0 to 0 of 0 entries

Fig. 8. Jane's query result

V. Results

A. Department of Education querying K-12 school student patients information

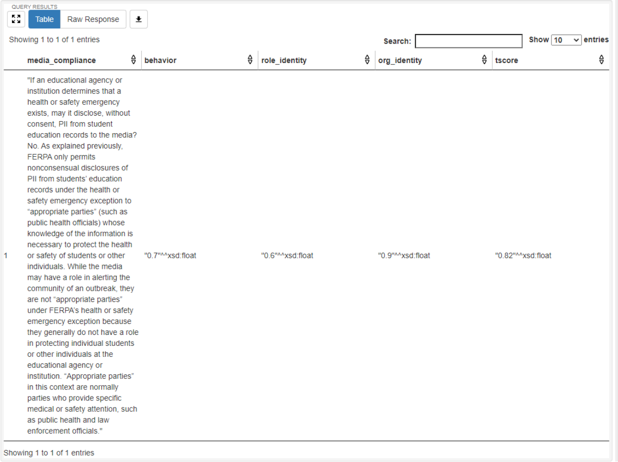
Figure 9 illustrates compliance policy of Department of Education and trust scores of John. Given weights of each trust element, John's trustworthiness is $\sum_{i=1}^{n} w_i e_i$ $= 0.1 \times$ (behavioral trust) $+ 0.2 \times$ (role identity trust) $+ 0.7 \times$ (organization identity trust) $= 0.82$. Since trust score threshold for PHI is 0.8, he satisfies trust policy. Figure 7 demonstrates John's query result. Date of birth and state of residence data of 9,735 adolescents are retrieved and it took 2.199 s.

The logic for the weight distribution is that organizations' capability to treat PHI is essential when processing it, such as workforce training and management, risk mitigation, and data safeguards. Therefore the weight of the organization identity trust prevails other elements. John could retrieve the data

TABLE I
ELAPSED TIME COMPARISON WITH/WITHOUT TRUST POLICY

Use Case	with Trust Policy		without Trust Policy	
	Elapsed Time	No. Results	Elapsed Time	No. Results
Department of Education	2.199 s	9,735	1.022 s	9,735
Insurance Company	16 ms	0	2.305 s	27,571

because the Department of Education has a high identity trust score of 0.8.



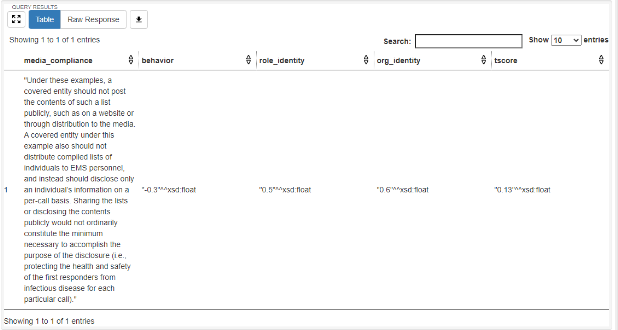
Showing 1 to 1 of 1 entries

Search: Show 10 entries

	media_compliance	behavior	role_identity	org_identity	tscore
1	"If an educational agency or institution determines that a health or safety emergency exists, may it disclose, without consent, PII from student education records to the media? No. As explained previously, FERPA only permits nonconsensual disclosures of PII from students' education records under the health or safety emergency exception to "appropriate parties" (such as public health officials) whose knowledge of the information is necessary to protect the health or safety of students or other individuals. While the media may have a role in alerting the community of an outbreak, they are not "appropriate parties" under FERPA's health or safety emergency exception because they generally do not have a role in protecting individual students or other individuals at the educational agency or institution. "Appropriate parties" in this context are normally parties who provide specific medical or safety attention, such as public health and law enforcement officials."	"0.7"^^xsd:float	"0.6"^^xsd:float	"0.9"^^xsd:float	"0.82"^^xsd:float

Showing 1 to 1 of 1 entries

Fig. 9. John's trust score



Showing 1 to 1 of 1 entries

Search: Show 10 entries

	media_compliance	behavior	role_identity	org_identity	tscore
1	"Under these examples, a covered entity should not post the contents of such a list publicly, such as on a website or through distribution to the media. A covered entity under this example also should not distribute compiled lists of individuals to EMS personnel, and instead should disclose only an individual's information on a per-call basis. Sharing the lists or disclosing the contents publicly would not ordinarily constitute the minimum necessary to accomplish the purpose of the disclosure (i.e., protecting the health and safety of the first responders from infectious disease for each particular call)."	"-0.3"^^xsd:float	"0.5"^^xsd:float	"0.6"^^xsd:float	"0.13"^^xsd:float

Showing 1 to 1 of 1 entries

Fig. 10. Jane's trust score

We set states as a geographic unit of this use case because previous research reviewed the federal Department of Education policy compliance [9]. Hence, if we have state-level data, it would be possible to query data with smaller geographic units and help policy decision-making at the states or county level.

B. Insurance company querying potential COVID-19 patients with pre-existing conditions

Figure 10 illustrates trust scores of Jane and her company, B&B Insurance Company. Given weights of each trust element, Jane's trustworthiness is $0.5 \times$ (behavioral trust) $+ 0.2\times$(role identity trust)$+0.3\times$(organization identity trust) ≈ 0.12. She could not pass the trust policy because the trust threshold is 0.5. Therefore, as demonstrated in the figure 8, she could not get the result and it took 16 ms.

Behavioral trust has the highest weight in this use case because how a person treats PHI is vital to prevent disclosure via written, electronic, oral, or other visual or audio forms regarding media access compliance. Unfortunately, Jane could not retrieve the data because her behavioral trust score was -0.3 and surpassed other trust elements.

This use case aims to show how our framework can prevent law evasion in the grey area. In ordinary circumstances, when the organization complies with regulations and an employee inquires about the information, it is not against the law until data breaches happen. We can prevent this by considering the user's behavioral trust score. For example, the weight distribution based on this idea prevented Jane from accessing PHI though the insurance company has a good identity score of 0.6.

C. Performance Evaluation

Table I describes elapsed time in each Use Case comparing SPARQL query with/without trust policy. For example, in the Department of Education use case, a query without trust policy took one second less. This implies that the processing trust policy creates some additional loads to the Jena Fuseki server. However, a query without the trust policy took more time in the insurance company's case. This is because the trust policy did not limit the access, and 27,571 entries are retrieved for the result.

VI. CONCLUSION AND FUTURE WORK

This paper developed a novel decentralized data exchange framework to enforce healthcare data regulations and assess participants' trustworthiness based on compliance history. We introduced the trust element concept to add one more layer of explainability of trust decision-making and fine-tune the data access policy based on the interaction between the elements that constitute trustworthiness in a specific context. Also, we integrated the HIPAA knowledge graph particularly tuned for COVID-19 data exchange for compliance verification. We have incorporated the detailed ontology of HIPAA health regulation in our system. This knowledge graph is developed using Semantic Web technologies and identifies each individual's trust score in the data exchange procedure. Reasoners can reason over the knowledge graphs with graph query languages that define related rules and trust scores threshold to control data access. As a proof of concept of our framework, we generated synthetic data with the size of 1M based on the CDC contact tracing data element ontology and demonstrated use case based on the CDC COVID-19 prevention guidelines for K-12 schools and the user with maleficent purpose. We demonstrated how our framework could help share large velocity Health datasets, like vaccination drives or contact tracing, with multiple distributed stakeholders.

To the best of our knowledge, our research is the first to

1) formally propose the concept of trust elements for real-time data sharing,
2) explicitly illustrate interactions between trust elements in a particular context,
3) introduce how to fine-tune the interaction for different compliance situations,
4) integrate the trust assessment with the compliance verification.

As part of our ongoing work, we will keep extending this framework to reflect the nature of trust more precisely and make it work in a real-world environment. Since trust has been an interest of philosophy for a long time, we will explore more on trust from a philosophy perspective to determine more trust elements. We strive to answer the questions include these, but not limited to: what are other trust elements to consider during the data exchange? How can we decide the subset of trust elements for different contexts? How can we evaluate trust elements? What should be the appropriate guideline for the weight assignment for each trust element? Also, we plan to include other concepts in trust management research such as delegation, recommendation, trust decay over time, and others to improve our framework. Finally, from the "need to know" to the "need to share" paradigm shift, we also plan to consider other healthcare standards for data interoperability, such as HL7 FHIR, LOINC, SNOMED CT, and CDS hooks. Our final goal is to achieve a trusted compliance enforcement system for health big data exchange that is practical yet supported by a solid theoretical foundation.

Acknowledgment

This research was partially supported by a DoD supplement to the NSF award 1747724, Phase I IUCRC UMBC: Center for Accelerated Real time Analytics (CARTA).

References

[1] K. Montgomery, J. Chester, and K. Kopp, "Health wearables: ensuring fairness, preventing discrimination, and promoting equity in an emerging internet-of-things environment," *Journal of Information Policy*, vol. 8, pp. 34–77, 2018.

[2] Office of the National Coordinator for Health Information Technology, "The ONC Cures Act final rule," https://www.healthit.gov/sites/default/files/cures/2020-03/TheONCCuresActFinalRule.pdf, 2020, accessed: 2021-09-04.

[3] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[4] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical owl-dl reasoner," *Journal of Web Semantics*, vol. 5, no. 2, pp. 51–53, 2007.

[5] R. Shearer, B. Motik, and I. Horrocks, "Hermit: A highly-efficient owl reasoner." in *Owled*, vol. 432, 2008, p. 91.

[6] The Apache Software Foundation, "Aapche Jena," http://jena.apache.org/, accessed: 2021-09-04.

[7] D. Kim and K. P. Joshi, "A semantically rich knowledge graph to automate hipaa regulations for cloud health it services," in *2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2021, pp. 7–12.

[8] K. P. Joshi, Y. Yesha, T. Finin *et al.*, "An ontology for a hipaa compliant cloud service," in *4th International IBM Cloud Academy Conference ICACON 2016*, 2016.

[9] L. Elluri, A. Piplai, A. Kotal, A. Joshi, and K. P. Joshi, "A policy-driven approach to secure extraction of covid-19 data from research papers," *Frontiers in Big Data*, vol. 4, 2021.

[10] P. R. Zimmermann, *The official PGP user's guide*. MIT press, 1995.

[11] A. Abdul-Rahman and S. Hailes, "A distributed trust model," in *Proceedings of the 1997 workshop on New security paradigms*, 1998, pp. 48–60.

[12] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized trust management," in *Proceedings 1996 IEEE Symposium on Security and Privacy*. IEEE, 1996, pp. 164–173.

[13] I. Ray and S. Chakraborty, "A vector model of trust for developing trustworthy systems," in *European Symposium on Research in Computer Security*. Springer, 2004, pp. 260–275.

[14] P. Zhou, X. Gu, J. Zhang, and M. Fei, "A priori trust inference with context-aware stereotypical deep learning," *Knowledge-Based Systems*, vol. 88, pp. 97–106, 2015.

[15] L. Viljanen, "Towards an ontology of trust," in *International conference on trust, privacy and security in digital business*. Springer, 2005, pp. 175–184.

[16] J. Huang and M. S. Fox, "An ontology of trust: formal semantics and transitivity," in *Proceedings of the 8th international conference on Electronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet*, 2006, pp. 259–270.

[17] J. Golbeck, B. Parsia, and J. Hendler, "Trust networks on the semantic web," in *International workshop on cooperative information agents*. Springer, 2003, pp. 238–249.

[18] B. ZadJabbari, P. Wongthongtham, and F. K. Hussain, "Ontology based approach in knowledge sharing measurement," *Journal of Universal Computer Science*, vol. 16, no. 6, pp. 956–982, 2010.

[19] W. Sherchan, S. Nepal, J. Hunklinger, and A. Bouguettaya, "A trust ontology for semantic services," in *2010 IEEE International Conference on Services Computing*. IEEE, 2010, pp. 313–320.

[20] O. Lassila and R. R. Swick, "Resource description framework (rdf) model and syntax specification, w3c recommendation 22 february 1999," 1999.

[21] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[22] L. Elluri, K. P. Joshi, and A. Kotal, "Measuring semantic similarity across eu gdpr regulation and cloud privacy policies," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3963–3978.

[23] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate gdpr and pci dss compliance," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1266–1271.

[24] Centers for Disease Control and Prevention (CDC), "Interim guidance on developing a covid-19 case investigation & contact tracing plan," https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/overview.html, Feb. 2021, accessed: 2021-09-04.

[25] R. Hardin, *Trust and trustworthiness*. Russell Sage Foundation, 2002.

[26] A. Baier, "Trust and antitrust," *ethics*, vol. 96, no. 2, pp. 231–260, 1986.

[27] T. Grandison and M. Sloman, "A survey of trust in internet applications," *IEEE Communications Surveys & Tutorials*, vol. 3, no. 4, pp. 2–16, 2000.

[28] Assistant Secretary for Planning and Evaluation (ASPE), "HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language, and Disability Status," https://aspe.hhs.gov/reports/hhs-implementation-guidance-data-collection-standards-race-ethnicity-sex-primary-language-disability-0, Oct. 2011, accessed: 2021-09-04.

[29] Centers for Disease Control and Prevention (CDC), "Considerations for Case Investigation and Contact Tracing in K-12 Schools and Institutions of Higher Education (IHEs)," https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/contact-tracing.html, Aug. 2021, accessed: 2021-09-04.