iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83632

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

SemOwn: Context-Aware Ownership Inference for IoT Data Streams via Knowledge Graph Reasoning and Shapley Attribution

Harsh Patel[1], Hemang Shrikar Desai[2], Harshit Raj[3], Dr. Chitra B.T.[4]
Department of Industrial Engineering & Management, RV College of Engineering, Bangalore, India

Abstract: ***As IoT networks grow larger and more organizationally entangled, a deceptively simple question keeps resurfacing: who actually owns the data these devices produce? The question is hardest to answer when observations come from many heterogeneous sources, pass through shared infrastructure, and carry meaning that shifts with context. Today's solutions typically lock ownership in at deployment time—an assumption that breaks down the moment derived data streams blend contributions from multiple stakeholders. We introduce SemOwn, a formally grounded framework that represents IoT data in a typed, attributed knowledge graph and infers ownership on the fly through context-weighted scoring and cooperative game- theoretic attribution. At its core lies a Semantic Ownership Graph*** $G_{SO} = (V, E, \lambda_V, \lambda_E)$ ***and a parametric ownership function*** O ***that weaves together spatial relevance, temporal validity, data sensitivity, and governance policy. For observations derived from several sources, we use the Shapley value to divide ownership fairly among contributors, and a role-based conflict resolution protocol settles disputes when stakes are closely matched. We prove that*** O ***satisfies efficiency, symmetry, null-player, and Lipschitz stability properties, and we bound the inference com- plexity under Monte Carlo approximation. On a synthetic smart- building dataset of*** 10^5 ***observations, SemOwn reaches an*** $F1$***- score of*** 0.91***—outstripping static baselines by*** 31 ***points and the strongest machine-learning baseline by*** 12***—while matching a domain-expert panel on*** 89% ***of multi-owner conflicts, all at sub-*** 50 ***ms latency. The framework is designed to support auditable ownership attribution in IoT data marketplaces, regulatory compliance, and federated multi-tenant deployments.***

Index Terms: Internet of Things, knowledge graph, data ownership, semantic reasoning, Shapley value, conflict resolution, context-aware inference, ontology

I. INTRODUCTION

Picture a comfort-index reading in a commercial building. That single number might fold together temperature, humidity, motion, and CO_2 measurements from sensors belonging to the landlord, a tenant, and an outside facilities contractor. Who owns it—and to what degree? This kind of question arises routinely in modern IoT deployments, where data provenance cuts across organisational, spatial, and temporal boundaries, yet it remains largely unanswered. The gap is not merely technical; it reflects a genuine doctrinal void in intellectual property (IP) law.

Despite its practical stakes, proprietary ownership of machine-generated IoT data has no harmonised statutory def- inition anywhere in the world. The regulations that do exist focus on controlling, consenting to, and accessing data—not on assigning property rights. The General Data Protection Regulation (GDPR, Regulation (EU) 2016/679), for example, protects personal data but confers no ownership title. The EU Data Act (Regulation (EU) 2023/2854) opens up B2B and B2C data sharing between device users and manufacturers, yet deliberately stops short of creating property rights over the shared data itself. In the United States, the California Consumer Privacy Act (CCPA, 2018) gives consumers ro- bust rights over their personal information but says nothing about ownership. India's Digital Personal Data Protection Act (DPDP Act, 2023) follows a similar pattern: it regulates the processing of digital personal data both domestically and abroad, yet leaves machine-generated, non-personal datasets in a legal grey area.

Legal scholars have long argued that data governance is better served by regulated access regimes than by strict pro- prietary models [10], [18]. We take this distinction seriously. Our aim is not to establish statutory ownership; it is to build a computational attribution model that can operate meaningfully within these evolving legal categories.

In practice, the absence of clear IP doctrines pushes industry toward two crude heuristics: the *device-owner* model and the *platform-operator* model. Both assign ownership statically and ignore what actually happens to the data downstream— namely, that it is frequently *derived* from multiple independent sources. Traditional IP doctrines like copyright and patent law are of little help here: they require human authorship and inventiveness, qualities that automated machine-generated datasets simply lack. Because existing frameworks only reg- ulate access and processing, the allocation of ownership in fused, machine-generated data remains a fundamentally open problem.

Semantic Web technologies—RDF, OWL ontologies, SPARQL, and rule languages such as SWRL—offer a promis- ing way forward. They make data descriptions machine- interpretable and can encode exactly the kind of relationships needed for ownership reasoning. Knowledge graphs (KGs) built on these technologies are already used in IoT platforms for device management, fault diagnosis, and interoperabil- ity [6]. Their structured, relational nature supports multi-hop inference and contextual reasoning that flat feature models simply cannot replicate [6], [32]. What remains unexplored is whether this KG-based reasoning can be turned toward a different goal: computationally *inferring* how ownership should be distributed when data streams are fused.

This paper takes on that challenge. Our contributions are:

1) SemOwn-O, a formal OWL 2-DL ontology that extends SSN/SOSA [1] and SAREF [2] with ownership-specific classes (OwnershipClaim, ContextEnvelope, DerivationLink) and datatype properties.
2) A parametric ownership function O that folds together
spatial relevance, temporal validity, data sensitivity, and governance policy into a single context-weighted score.
3) A Shapley-based attribution mechanism that dis- tributes fractional ownership of derived observations among their contributing sources.
4) A conflict resolution protocol backed by formal cor- rectness guarantees (determinism, weight conservation, auditability).
5) Empirical validation on a synthetic smart-building dataset with 10^5 observations and carefully controlled multi-owner scenarios.

Why formalism matters here. We deliberately ground this work in formal mathematics—a contrast to the ad-hoc heuristics that dominate current practice. Our reasoning is straightforward: ownership decisions in regulated, multi-tenant IoT environ- ments carry legal, financial, and privacy consequences. They demand the same level of rigour already applied to algorithmic fairness and cooperative game theory, because the people affected deserve auditable, provably correct guarantees rather than black-box assignments. The rest of the paper is organised as follows. Section II surveys the global IP landscape for IoT data. Section III reviews relevant technical and legal literature. Section IV covers background material. Section V sets up the problem mathematically. Section VI describes the SemOwn framework. Section VII presents theoretical analysis. Section VIII details the inference and conflict resolution algorithms. Section X explains the experimental setup, Section XI discusses results, and Section XIII concludes.

II. GLOBAL LEGAL AND INTELLECTUAL PROPERTY FRAMEWORK FOR IOT DATA

Before building a computational model for IoT data own- ership, we need to understand what the law actually says— and, more importantly, what it does not. The short answer is that no major jurisdiction grants a clear-cut property right for machine-generated data. What we have instead is a patchwork of privacy and access regulations, each leaving significant gaps.

A. Why Ownership Arbitration Is Overdue

The scale of the problem is hard to overstate. Over the past decade, IoT device deployments have grown exponentially, with more than 15 billion active endpoints worldwide now generating zettabytes of machine-generated data every year. At the same time, data-related disputes and breaches tied to IoT infrastructure have surged. Regulatory bodies have taken notice: policy reports increasingly flag the absence of enforceable ownership frameworks, pointing out that while privacy rules have matured, proprietary rights over IoT data remain undefined. The upshot is that legal frameworks de- signed for human-authored works or personal data simply were not built to handle the vast intellectual property value locked up in IoT telemetry. Without new computational attribution mechanisms, the likely outcomes are prolonged litigation and market dysfunction.

B. European Union

The GDPR (Regulation (EU) 2016/679) sets strict rules for lawful processing (Article 6) and data portability (Article 20), but it targets personal data exclusively—it does not grant anyone ownership. The EU Data Act (2023) goes further by requiring manufacturers to share IoT-generated data with users and business partners, yet it deliberately avoids creating new intellectual property rights over the data itself. The Database Directive (96/9/EC) does provide *sui generis* protection for the investment that goes into structuring a database, but its relevance to raw, automatically generated IoT datasets is limited at best.

C. United States

There is no comprehensive federal data ownership law in the United States. The closest analogue is the CCPA (California Civil Code §1798.100), which gives consumers extensive rights to access, delete, and control the sale of their personal data. Much like the GDPR, though, the CCPA frames data as a consumer right rather than an alienable property asset. Non-personal, machine-generated data sits in a legal grey area where private contracts are the main line of defence.

D. India

India's Information Technology Act, 2000, gives legal recognition to electronic records but is silent on data own- ership. The more recent DPDP Act, 2023, imposes strict obligations on Data Fiduciaries and grants robust rights to Data Principals for digital personal data. It does not, however, address proprietary ownership of machine-generated telemetry. In practice, IoT data governance in India is contract-based, hinging on bilateral agreements rather than underlying IP rights.

E. The Core Doctrinal Gap

The pattern is consistent: no jurisdiction defines a formal property right for machine-generated data. IoT data does not fit neatly into any existing IP category. It lacks the "human authorship" needed for copyright protection, does not embody the kind of novel invention that patent law rewards, and is usually too widely distributed to qualify as a trade secret. This mismatch means that ownership of IoT-generated data remains an open theoretical problem—one that calls for a new, computationally grounded approach to arbitration.

III. RELATED WORK

A. Legal and IP Perspectives on IoT Data

Whether data should be treated as alienable property or as a regulated resource is still actively debated in legal scholarship. Scholars draw a careful three-way distinction between data ownership (a proprietary, *in rem* right), data control (the practical ability to exclude others), and data access rights (permissions granted by statute) [10], [18]. Traditional copyright law offers little help for machine-generated data, be- cause it demands human creative authorship—something IoT telemetry inherently lacks. Without formal ownership statutes, the IoT ecosystem falls back on contractual workarounds: End- User License Agreements (EULAs), restrictive licences, and bespoke data-sharing terms that approximate property rights without truly being them.

We see SemOwn as filling precisely this gap. By formal- ising how data derivation works and applying game-theoretic attribution, SemOwn turns abstract legal principles—equitable contribution, accountability, traceability—into concrete com- putational operations for a domain where the law has yet to catch up.

B. Semantic Modelling for IoT

The W3C's Semantic Sensor Network (SSN) ontology and its lightweight core SOSA give the community standardised vocabularies for describing sensors, observations, and fea- tures of interest [1]. ETSI's SAREF ontology targets smart- appliance interoperability [2]. Gyrard et al. [3] tackled cross- domain semantic alignment, while Bermudez-Edo et al. [4] proposed IoT-Lite for resource-constrained discovery. Bar- naghi et al. [5] formalised how to annotate IoT streams with semantic metadata, but their focus stayed at the observation level—ownership semantics were not part of the picture.

C. Knowledge Graphs in IoT

Hogan et al. [6] provide a comprehensive survey of KG construction, completion, and reasoning. Within the IoT space, Ren et al. [7] used KG embeddings (TransE, RotatE) for industrial fault diagnosis, and Janowicz et al. [8] made the case for KG-based digital twins. Akroyd et al. [9] showed how KGs can integrate heterogeneous scientific data, while Zhang et al. [32] applied ontology-driven KG reasoning to multi-source IoT security. All of these efforts harness KGs for analytics and interoperability, but none of them attempt ownership inference.

D. Data Ownership and Provenance

Hummel et al. [10] identified seven distinct dimensions of data ownership (access, control, disposition, and so on). On the provenance side, Liang et al. [11] proposed blockchain- based provenance for cloud-IoT data, and Pal et al. [12] combined blockchain with trusted execution environments. Fernandez et al. [13] and Jia et al. [31] explored Shapley- based data valuation, while Rozemberczki et al. [30] surveyed how Shapley values are used for explainability and attribution in machine learning. Bonatti et al. [14] developed an ontology Table I pulls these threads together.

TABLE I: Feature comparison of related systems.

Feature	SSN/SOSA	FIWARE	ProvChain	GAIA-X	SemOwn
Semantic model	✓	✓		Partial	✓
Ownership ontology				Partial	✓
Dynamic inference					✓
Multi-owner				Partial	✓
Conflict resolution					✓
Context-awareness	Partial	Partial		Partial	✓
KG reasoning					✓

for multi-paradigm access control. None of these frameworks, however, dynamically infer ownership from semantic context and multi-source composition.

E. Research Gap

The bottom line is that no existing framework jointly tackles semantic IoT modelling, graph-based ownership reasoning, and multi-stakeholder conflict resolution. Static device-owner models ignore context altogether. Blockchain-based prove- nance systems keep meticulous records of data lineage but can- not *infer* fractional ownership from that lineage. KG platforms use graph reasoning for analytics and fault diagnosis, yet never turn it toward ownership attribution. Each tradition contributes essential machinery—semantic annotation, immutable logging, graph traversal—but none brings in cooperative game theory to distribute ownership axiomatically across contributors. Even recent work studying Shapley values in logic- and query- centric settings [33] leaves semantic IoT ownership untouched. SemOwn bridges this gap by weaving these traditions together within a single, formally grounded inference pipeline.

IV. PRELIMINARIES

Definition 1 (Knowledge Graph). *A* knowledge graph *is a directed, labelled multigraph* $\mathrm{G} = (V, E, \mathrm{R})$ *where* V *is a finite set of entities,* R *is a finite set of relation types, and* $E \subseteq V \times \mathrm{R} \times V$ *is the edge set. Each edge* $(h, r, t) \in E$ *asserts that relation* r *holds between head entity* h *and tail entity* t.

Definition 2 (Shapley Value [15]). *For a cooperative game* (N, v) *with player set* N *and characteristic function* $v : 2^N \to \mathrm{R}$ *satisfying* $v(\emptyset) = 0$, *the Shapley value of player* $i \in N$ *is*

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right]. \tag{1}$$

The Shapley value uniquely satisfies four axioms: *efficiency* ($\sum_i \phi_i = v(N)$), *symmetry*, *linearity*, and the *null-player* property.

Definition 3 (SWRL Rule). *A Semantic Web Rule Language (SWRL) rule [16] takes the form* $B_1 \wedge \cdots \wedge B_n \Rightarrow H_1 \wedge \cdots \wedge H_m$, *where each* B_k *and* H_j *is an atom involving OWL classes, properties, or built-in predicates. Rules extend OWL 2- DL reasoning with Horn-clause expressiveness.*

V. PROBLEM FORMULATION

A. Intellectual Property Challenges in IoT

The ownership attribution challenge splits naturally into three interrelated problems, each linking a specific legal gap to a concrete computational requirement:

1) The Doctrinal Gap. No existing IP regime grants a statutory property right for machine-generated data. The GDPR (EU) and DPDP Act (India) regulate processing; the CCPA (US) protects consumer privacy; but none of them confer alienable ownership over telemetry that lacks human authorship. In formal terms, for a data asset o the legal ownership function $\mathrm{L}(o)$ is simply undefined in statute. We need a computational surrogate $\mathrm{O}(o)$ that fills this void.

2) The Attribution Problem. In multi-stakeholder IoT settings—where controllers, processors, and end-users all contribute—derived data blends inputs from across legal boundaries. Current regulations govern access but have nothing to say about how fractional ownership of a fused artifact should be divided. The system must assign ownership weights $w_i \in [0, 1]$ to each stakeholder in a set N, ensuring they sum to one ($w_i = 1$) and reflect genuine contribution rather than contractual default.
3) The Enforcement Problem. Cross-jurisdictional IoT deployments have no formal mechanism for resolving ownership disputes. Without a unifying property law, disagreements become intractable under contract or lia- bility law alone. We need a deterministic, auditable con- flict predicate together with a resolution operator R(o) that breaks ties through policy-driven arbitration aligned with GDPR and CCPA compliance requirements.

B. Semantic Ownership Graph

Definition 4 (Semantic Ownership Graph). *The* Semantic Ownership Graph *is a typed, attributed, directed multigraph*

$$G_{so} = (V, E, \lambda_V, \lambda_E) \quad (2)$$

where $V = V_D \cup V_U \cup V_O \cup V_L \cup V_T$ partitions vertices into devices, users (stakeholders), observations, locations, and time intervals, respectively. The typing functions $\lambda_V : V \to C$ and $\lambda_E : E \to R$ map vertices and edges to ontology classes and relation types.

C. Context Function

Let $o \in V_O$ be an observation produced at location $l \in V_L$ during time interval $t \in V_T$.

Definition 5 (Context Envelope). *The* context envelope *of o is the tuple*

$$C(o, t, l) = \rho(o, l),\ \tau(o, t),\ \sigma(o),\ \pi(l, t) \quad (3)$$

where

- $\rho : V_O \times V_L \to [0, 1]$ *is a* location-relevance *score, mod- elled as* $\rho(o, l) = \exp(-\kappa \cdot d_{adm}(o, l))$ *with administrative distance* d_{adm} *and decay rate* $\kappa > 0$*;*
- $\tau : V_O \times V_T \to [0, 1]$ *is a* temporal-validity *score, equal to* 1 *if o falls within an active data-sharing agreement and decaying otherwise;*
- $\sigma : V_O \to [0, 1]$ *is the* sensitivity level *(higher values amplify the generating user's claim);*
- $\pi : V_L \times V_T \to P$ *returns the applicable* governance policy.

D. Ownership Function

Definition 6 (Ownership Function). *For an observation* $o \in V_O$*, the* ownership function $O : V_O \to 2^{VU \times [0,1]}$ *returns a set of stakeholder–weight pairs:*

$$O(o) = \{(u_i, w_i) \mid u_i \in V_U,\ w_i \in [0, 1],\ \sum_i w_i = 1\}. \quad (4)$$

For a *direct* observation generated by device d owned by user u:

$w_u = \alpha \cdot g(d, o) + \beta \cdot \rho(o, l) + \gamma \cdot \tau(o, t) + \delta \cdot \sigma(o)$, (5) with hyperparameters $\alpha + \beta + \gamma + \delta = 1$.

For a *derived* observation o^* composed from sources $\{o_1, \ldots, o_n\}$, ownership is distributed via the Shapley value of a cooperative game (N, v):

$$w_i(o^*) = \phi_i(v), \quad v(S) = I\ o^*; \{o_j \mid j \in S\}\ , \quad (6)$$

where $I(\cdot\,;\cdot)$ denotes mutual information and N is the union of all owners across sources.

The idea of using Shapley values for equitable data valu- ation is well-established [27], [31]; we adapt it here to the ownership setting.

Remark 1 (Why Mutual Information?). *We chose* $v(S) = I(o^*; \{o_j \mid j \in S\})$ *as the characteristic function over two natural alternatives—Pearson correlation and marginal entropy—because it is the only candidate that meets all three requirements for Shapley-compatible ownership attribution:*

1) Non-negativity with v(□) = 0. Mutual information is non-negative by definition [26], and I(o□; □) = 0. Pearson correlation, by contrast, can go negative—a deal-breaker for a characteristic function.

2) Monotonicity. Adding sources can only increase the information available about o□: the chain rule of MI ensures I(o□; {oj}j□T) ≥ I(o□; {oj}j□S) whenever S □ T. Raw entropy H({oj}j□S) grows with |S| too, but it measures source complexity regardless of relevance to o□, effectively rewarding noisy sources.
3) Non-linear dependency capture. IoT data fusion often involves non-linear aggregation—think of a comfort in- dex that combines temperature and humidity through a non-linear heat-index formula. Pearson correlation sees only linear associations; MI picks up arbitrary statistical dependencies [26].

Table II summarises the comparison.

TABLE II: Characteristic function candidates for Shapley attribution.

Candidate $v(S)$	$v(\emptyset)=0$	Monotone	Non-linear	Output-directed
Pearson $\lvert r \rvert$	×	×	×	✓
$H(\{o_i\}_{i\in S})$	✓	✓	✓	
$I(o^*;\{o_i\}_{i\in S})$	✓	✓	✓	

E. Conflict Predicate

Definition 7 (Ownership Conflict). *An observation o is in* ownership conflict *if*

$$\exists\, u_i, u_j \in \text{owners}(o),\ i \neq j:\ |w_i - w_j| < \varepsilon \wedge w_i + w_j > \theta, \tag{7}$$

where $\varepsilon > 0$ is an equality tolerance and $\theta \in (0, 1]$ is a significance threshold.

Research Objective. Given G_{so} and a set of contextual parameters, compute $O(o)$ for every $o \in V_O$, detect conflicts via (7), and resolve them through a formal protocol, with provable correctness guarantees.

Assumption 1 (Non-Degeneracy). *The normalisation step of the ownership inference procedure enforces a minimum weight floor: for all $u_i \in$ owners(o), $w_i \geq \eta$ where $\eta > 0$ is a system parameter (default $\eta = 0.01$). Any stakeholder whose* pre-normalisation weight falls below $\eta \cdot \sum_j w_j$ is removed from the owner set and assigned zero ownership. This is semantically justified: a stakeholder contributing less than η fractional ownership has no meaningful claim. The condition is consistent with the null-player property (Proposition 3), which already assigns zero weight to non-contributors; the η-threshold extends this to negligible contributors.

VI. PROPOSED MODEL: THE SEMOWN FRAMEWORK

A. Legal-Technical Mapping

At its heart, SemOwn translates well-known legal doctrines—contributorship, joint ownership, liability allocation—into deterministic algorithms. Each of the framework's core components maps onto a specific legal requirement:

- Knowledge Graphs for Traceability. Regulations like GDPR Article 30 and the DPDP Act require immutable records of processing activities. SemOwn's KG serves exactly this role, creating explicit, auditable links from derived data back to its origin sensors, processing entities, and governing contracts.
- Context-Aware Inference for Jurisdictional Compli- ance. Ownership is inherently jurisdictional. By natively ingesting spatial (ρ), temporal (τ), and policy (π) parame- ters, SemOwn's weighting mechanism can align dynami- cally with jurisdiction-specific rules—for instance, giving extra weight to consumer rights under the CCPA.
- Shapley Value for Equitable Distribution. IP law's joint-authorship and licensing doctrines aim for fair roy- alty splits. SemOwn's use of the Shapley value achieves the same goal with mathematical guarantees: the resulting fractional ownership is provably equitable.

B. System Overview

Data flows through SemOwn in five stages. Raw IoT pay-loads (MQTT messages, HTTP POST bodies) first enter the S*mantic Ingestion Layer*, where an RDF triple generator maps each payload to SSN/SOSA-compliant triples and a context encoder attaches a ContextEnvelope capturing ρ, τ, σ, and π. The resulting triples are stored in the *Knowledge Graph Engine*, with provenance edges recorded via W3C PROV-O. Next, the *Ownership Inference Engine* fires SWRL rules for deterministic propagation, computes context-weighted scores via (5), and—for derived observations—estimates Shapley attributions through Monte Carlo sampling (Algorithm 1).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

If the *Conflict Resolution Layer* detects a conflict per (7), it applies the arbitration function R(*o*) (Algorithm 2) and logs the outcome. Finally, the *Query and Validation API* exposes SPARQL and REST endpoints for ownership lookup and provenance audit.

In summary:

1) Semantic Ingestion Layer. Converts raw IoT pay- loads (MQTT, HTTP) into RDF triples conforming to SSN/SOSA augmented with SemOwn-O classes. The context encoder enriches each observation with its con- text envelope C(o, t, l).
2) Knowledge Graph Engine. Stores Gso in a graph database (Neo4j for traversal performance; Apache Jena TDB2 for OWL reasoning). Provenance is tracked using W3C PROV-O.
3) Ownership Inference Engine. Executes SWRL rules for deterministic ownership propagation, computes wu via (5), and applies Shapley attribution via (6).
4) Conflict Resolution Layer. Detects conflicts via (7) and resolves them using a priority-weighted arbitration function R(o).
5) Query and Validation API. Exposes SPARQL and REST endpoints for ownership queries and visual dash- boards.

C. SemOwn-O Ontology

SemOwn-O introduces four core classes: OwnershipClaim, ContextEnvelope, Data Stakeholder, and Ownership Conflict. Key object properties include hasOwner, derivedFrom, hasContext, and generated By Datatype properties include ownership Weight ([0, 1]-valued), sensitivity Level, temporal Scope, and stakeholder Role. The ontology is specified in OWL 2- DL, ensuring decidability of reasoning.

D. SWRL Ownership Rules

Three representative rules govern inference:

Rule 1 (Direct ownership):

$$\text{Sensor}(?d) \wedge \text{ownedBy}(?d, ?u) \wedge \text{madeObs}(?d, ?o) \Rightarrow \text{hasOwner}(?o, ?u) \wedge \text{weight}(?o, ?u, 1.0)$$

Rule 2 (Location co-ownership):

$$\text{Context}(?c) \wedge \text{locatedAt}(?c, ?l) \wedge \text{Tenant}(?t) \wedge \text{occupies}(?t, ?l) \wedge \text{hasContext}(?o, ?c) \Rightarrow \text{hasOwner}(?o, ?t) \wedge \text{weight}(?o, ?t, 0.3)$$

Rule 3 (Derivation propagation):

$$\text{derivedFrom}(?o', ?o) \wedge \text{hasOwner}(?o, ?u) \Rightarrow \text{hasOwner}(?o', ?u)$$

E. Conflict Resolution Function

When the conflict predicate (7) holds, the system applies:

$$R(o) = \arg \max_{u_i \in \text{owners}(o)} \lambda_1 r(u_i) + \lambda_2 \phi_i + \lambda_3 p(u_i, o) , \quad (8)$$

where $r(u_i) \in \{1.0, 0.7, 0.4, 0.3\}$ encodes role priority (gen- erator, processor, consumer, operator), ϕ_i is the normalised Shapley value, $p(u_i, o)$ captures policy preference, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

The winning claimant receives a weight boost Δ, and all weights are renormalised to maintain the unit-sum constraint.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

VII. THEORETICAL ANALYSIS

Theorem 1 (Efficiency of O). *For any observation $o \in V_{in}$, the ownership function O satisfies efficiency:* $\sum_{u_i \in \text{owners}(o)} w_i = 1$.

Proof. For direct observations, O returns a singleton $\{(u, w_u)\}$ that is trivially normalised to 1 (Alg. 1, line 31). For derived observations, $w_i = \phi_i(v) / \sum_i \phi_i(v)$. By the efficiency axiom of the Shapley value, $\sum_i \phi_i(v) = v(N)$. Dividing each ϕ_i by $v(N)$ yields $\sum_i w_i = 1$. □

Theorem 2 (Symmetry Preservation). *If two stakeholders u_i, u_j contribute identically to a derived observation o^*—i.e., for all $S \subseteq N \setminus \{u_i, u_j\}$, $v(S \cup \{u_i\}) = v(S \cup \{u_j\})$—then $w_i(o^*) = w_j(o^*)$.*

Proof. The Shapley value inherits the symmetry axiom: interchangeable players receive equal payoffs [15]. Since the ownership weight is a normalised Shapley value, symmetry is preserved under normalisation. □

Proposition 3 (Null-Player Exclusion). *A stakeholder u_k whose data contributes no information to o^*, i.e., $v(S \cup \{u_k\}) = v(S)$ for all $S \subseteq N \setminus \{u_k\}$, receives $w_k(o^*) = 0$.*

Proof. By the null-player axiom, $\phi_k(v) = 0$. Normalisation preserves zero values. □

Lemma 4 (Conflict Detection Completeness). *Algorithm 2 detects all pairwise conflicts among the owners of o.*

Proof. The algorithm iterates over all $\binom{|\text{owners}(o)|}{2}$ pairs (lines 5–9), checking the predicate (7) exhaustively. □

Theorem 5 (Weight Conservation under Resolution). *The conflict resolution protocol preserves the unit-sum constraint. For any observation o, $\sum_{u_i} w_i = 1$ holds both before and after resolution.*

Proof. Resolution adjusts $w_i \leftarrow w_i + \Delta$ and $w_j \leftarrow w_j - \Delta$ for a conflicting pair, leaving the sum invariant. Step 36 of Algorithm 2 performs explicit renormalisation, guaranteeing the invariant even under floating-point accumulation. ❑

Theorem 6 (Ownership Stability under Context Perturbation). *Let* $C(o) = \langle \rho, \tau, \sigma, \pi \rangle$ *and* $\tilde{C}(o) = \langle \tilde{\rho}, \tilde{\tau}, \tilde{\sigma}, \tilde{\pi} \rangle$ *be two context envelopes for the same direct observation o, with per- component perturbation bounded by* $\|C - \tilde{C}\|_\infty \leq \epsilon$. *Under Assumption 1, for every stakeholder* $u_i \in \text{owners}(o)$*:*

$$|w_i(C) - w_i(\tilde{C})| \leq L \cdot \epsilon, \quad (9)$$

where the Lipschitz constant is $L = \max(\beta, \gamma, \delta) \cdot (1 + 1/\eta)$ *and* η *is the non-degeneracy floor from Assumption 1.*

Proof. Denote the pre-normalisation weight under context C as $\hat{w}_i = \alpha\, g(d, o) + \beta\, \rho + \gamma\, \tau + \delta\, \sigma$ and similarly $\tilde{\hat{w}}_i$ under $\tilde{C}$. Since $g(d, o)$ depends only on the device–observation relationship and not on context, the generation term cancels:

$$|\hat{w}_i - \tilde{\hat{w}}_i| = |\beta(\rho - \tilde{\rho}) + \gamma(\tau - \tilde{\tau}) + \delta(\sigma - \tilde{\sigma})| \leq (\beta + \gamma + \delta)\epsilon \leq \max(\beta, \gamma, \delta) \cdot 3\epsilon$$

A tighter bound uses the ℓ_∞ structure: each term contributes at most $\max(\beta, \gamma, \delta) \cdot \epsilon$, so $|\hat{w}_i - \tilde{\hat{w}}_i| \leq \max(\beta, \gamma, \delta) \cdot \epsilon$.

For the normalised weight $w_i = \hat{w}_i / W$ where $W = \sum_j \hat{w}_j$,

the quotient perturbation bound gives:

$$|w_i - \tilde{w}_i| \leq \frac{|\hat{w}_i - \tilde{\hat{w}}_i|}{W} + \frac{\hat{w}_i \cdot |W - \tilde{W}|}{W \cdot \tilde{W}}.$$

Since $|W - \tilde{W}| \leq |\text{owners}(o)| \cdot \max(\beta, \gamma, \delta) \cdot \epsilon$ and, by Assumption 1, $W \geq |\text{owners}(o)| \cdot \eta$, each term is bounded by $(\max(\beta, \gamma, \delta)/\eta) \cdot \epsilon$. Combining yields $|w_i - \tilde{w}_i| \leq \max(\beta, \gamma, \delta) \cdot (1 + 1/\eta) \cdot \epsilon$.

The stability bound has a direct practical interpretation: with default parameters ($\beta = \gamma = \delta = 0.2$, $\eta = 0.01$), a 1% pertur- bation in any context dimension shifts ownership weights by at most 20.2%. When contexts are known to higher precision— typical for GPS-equipped indoor IoT deployments—the bound tightens proportionally, confirming that SemOwn's ownership assignments are robust under realistic sensor noise.

Proposition 7 (Complexity Bound). *Let n denote the number of observations in G_{so}, k the maximum fan-in of a derived observation, and m the maximum number of distinct owners. Exact Shapley computation is $O(2^m)$ in general, motivating approximation in practical systems [22], [33]. Under Monte Carlo sampling with p permutations, the approximation error ϵ_{MC} satisfies $\Pr[|\hat{\phi}_i - \phi_i| > \epsilon_{MC}] \leq 2\exp(-2p\,\epsilon_{MC}^2/R^2)$ where $R = \max_S |v(S \cup \{i\}) - v(S)|$, and the total inference complexity for the graph is $O(n \cdot m \cdot p)$.*

Proof. Each observation requires evaluating p random permutations of m players through the marginal contribution, costing $O(p \cdot m)$. The concentration inequality follows from Hoeffding's bound applied to the sampling estimator of (1). In practice, for $R \leq 1$ (which holds when v is normalised mutual information), setting $p = 1000$ yields $\Pr[|\hat{\phi}_i - \phi_i| > 0.01] \leq 2e^{-20} \approx 4 \times 10^{-9}$, providing confidence exceeding 99.99% at negligible computational cost. □

Algorithm 1 Context-Aware Ownership Inference

Require: G_{so}, observation o, context (t, l), parameters $\alpha, \beta, \gamma, \delta$
Ensure: $O(o) = \{(u_i, w_i)\}$

```
sources ← GETSOURCES(G_so, o)
(ρ, τ, σ, π) ← C(o, t, l)
if sources = ∅ then
    d ← GENERATOR(G_so, o)
    u ← DEVICEOWNER(G_so, d)
    w_u ← α g(d, o) + β ρ + γ τ + δ σ
    extra ← RUNSWRL(G_so, o, C)
    O(o) ← {(u, w_u)} ∪ extra
else
    N ← ∅
    for all s_j ∈ sources do
        N ← N ∪ INFEROWNERSHIP(G_so, s_j, t, l)
    end for
    for all u_i ∈ N do
        w_i ← APPROXSHAPLEY(u_i, N, sources, o, ρ)
    end for
    O(o) ← {(u_i, w_i)}
end if
Normalise: w_i ← w_i / Σ_j w_j
return O(o)
```

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

VIII. ALGORITHM DESIGN

Case Studies: IoT Data Ownership in Practice

To ground SemOwn in real-world concerns, we walk through three multi-stakeholder IoT scenarios. Each one high- lights a gap in current law and shows how SemOwn fills it computationally.

A. Smart Home Data Dispute

Scenario. A smart thermostat aggregates ambient tem- perature and occupancy data. The homeowner, the device manufacturer, and a cloud analytics provider all claim the longitudinal datasets used to train energy-saving algorithms.

Legal picture. Under the CCPA the user can access and delete this data; under the GDPR the occupancy patterns count as personal data. But neither framework says who owns the aggregated, anonymised training set. In practice, disputes get settled by End-User License Agreements that overwhelmingly favour the manufacturer. How SemOwn helps. SemOwn models the home as the spatial context (ρ) and treats the user as the primary data prin- cipal. Because the data is intimate (σ is high), the homeowner receives a large contextual weight, while the manufacturer's derivation effort is acknowledged through Shapley attribution. The result is a transparent, equitable ownership split that replaces the one-sidedness of typical EULAs.

B. Connected Vehicle Telemetry

Scenario. A connected vehicle continuously transmits brak- ing, acceleration, and geolocation telemetry. The driver asserts

Algorithm 2 Ownership Conflict Resolution

Require: $O(o)$, thresholds ε, ϑ, weights $\lambda_1, \lambda_2, \lambda_3$
Ensure: Resolved $O'(o)$, *report*

```
conflicts ← ∅
for all (u_i, w_i), (u_j, w_j) with i ≠ j do
    if |w_i − w_j| < ε and w_i + w_j > ϑ then
        conflicts ← conflicts ∪ {(u_i, u_j)}
    end if
end for
if conflicts = ∅ then
    return (O(o), "no conflict")
end if
for all (u_i, u_j) ∈ conflicts do
    s_i ← λ1 r(u_i) + λ2 φ_i + λ3 p(u_i, o)
    s_j ← λ1 r(u_j) + λ2 φ_j + λ3 p(u_j, o)
    if s_i > s_j then
        w_i ← w_i + Δ; w_j ← w_j − Δ
    else if s_j > s_i then
        w_j ← w_j + Δ; w_i ← w_i − Δ
    else
        w_i ← w_j ← (w_i + w_j)/2
    end if
    Log resolution decision
end for
Normalise all w_i
return (O'(o), report)
```

privacy rights, the manufacturer claims IP rights over the telemetry format, and a third-party insurer needs the data for dynamic pricing. Legal picture. The EU Data Act lets the insurer access this data at the driver's request, breaking the manufacturer's data silo. The DPDP Act requires consent for processing. Yet none of these laws say who owns the synthesised risk profile. Liability and contract law fill the gap awkwardly, leaving everyone uncertain about secondary monetisation. How SemOwn helps. SemOwn's knowledge graph ingests the data-sharing agreements as policy envelopes (π). When the risk profile—a derived observation—is created, the system traces provenance back to the vehicle's sensors. The conflict resolution protocol (R) uses role-based arbitration to balance the driver's generation priority against the insurer's processing role, producing an auditable ownership ledger aligned with the EU Data Act's mandate for fair value exchange.

C. Industrial IoT (IIoT) Predictive Maintenance

Scenario. In a smart factory, proprietary acoustic sensors (Vendor) monitor robotic arms (Enterprise) while machine learning algorithms (Analytics Provider) predict failures.

Legal picture. Machine-generated industrial data generally falls outside the GDPR and CCPA. Trade secret law and the Database Directive (96/9/EC) offer only limited protec- tion, since raw acoustic data lacks human originality. Today, disputes over the predictive insights are resolved through complex, bespoke B2B contracts—an expensive and litigation- prone process.

How SemOwn helps. SemOwn provides a deterministic contributorship model. Using the Shapley value computed from mutual information, it quantifies precisely how much the Vendor's raw data and the Analytics Provider's model each contributed to the final prediction. The outcome is a transpar- ent, mathematically rigorous royalty allocation that can replace costly legal arbitration with automated, fair computation.

IX. EXPERIMENTAL SETUP

A. Dataset

We built a synthetic smart-building dataset modelling a five-floor commercial building with 10 rooms per floor. Each room has 3–6 sensors (temperature, humidity, CO_2, motion, light, energy)—roughly 200 sensors in total. We simulated 30 days of readings at 5-minute intervals, yielding 100,000 direct observations and 10,000 derived ones (comfort indices, occupancy scores). Four stakeholder types participate: building owner, floor tenants, facility manager, and IoT platform opera- tor. To stress-test the system, we engineered a 40% ownership- ambiguity rate and reassigned 20% of sensors mid-simulation to exercise temporal dynamics.

Why synthetic data? No publicly available IoT dataset comes with ground-truth ownership labels—ownership is a socio-legal construct that standard sensor benchmarks (UCI, OGC SensorThings) simply do not capture. Using synthetic data with controlled ownership scenarios is standard practice in Shapley-based valuation research: both Ghorbani and Zou [27] and Jia et al. [31] validated their methods on synthetic data with known ground truth before moving to real deployments. Our design gives us three concrete advantages: (i) noise-free ground truth, (ii) tuneable ambiguity rates for stress-testing conflict resolution, and (iii) full reproducibility. To keep the data realistic, sensor readings follow statistical distributions calibrated against the ASHRAE Global Thermal Comfort Database II [28] for temperature and humidity, and the WELL Building Standard for CO_2 thresholds.

Planned real-world extension. We are in active discussions with a multi-tenant commercial building operator (two build- ings, four tenant organisations) to deploy SemOwn on live sensor feeds and collect expert-labelled ownership ground truth from operational data-governance teams. This will serve as a complementary validation track, pairing real (noisy, incomplete) provenance data with human annotations. We plan to report these results in a follow-up study. This kind of hybrid strategy—synthetic for controlled ablation, real-world for ecological validity—is increasingly standard in the data valuation literature [27], [31].

B. Ground Truth

Three domain experts (IoT systems, data governance, and property law) independently labelled ownership for 1,000 observations (500 direct, 500 derived). Inter-annotator agree- ment was measured via Fleiss' $\kappa = 0.82$, indicating strong agreement.

C. Baselines

- Static: device owner is always the sole data owner.
- Rule-Only: SWRL ownership rules without context weighting ($\alpha = 1, \beta = \gamma = \delta = 0$).
- Logistic Regression (LR): L2-regularised logistic regres- sion ($C = 1.0$) trained on 8 features extracted from G_{so}: context dimensions (ρ, τ, σ), stakeholder role (one-hot), device type (one-hot), source fan-in, KG hop-distance to stakeholder, and number of co-located stakeholders. Split: 700 train / 300 test, stratified by observation type.
- Random Forest (RF): 100 trees, max depth 10, same feature set and split as LR.
- Random: conflict resolution assigns random winner.
- Majority: conflict resolution by majority vote among owners.

The ML baselines (LR, RF) are there for a specific diag- nostic reason: they test whether ownership can be predicted from flat, tabular features pulled out of the knowledge graph, without leveraging graph structure or game-theoretic attribu- tion. In other words, they answer the question "Do we really need semantic graph reasoning, or would simple features do?" Graph-aware models like GNNs cannot answer that question because they, too, exploit relational structure. This design choice is consistent with KG reasoning literature showing that multi-hop relational inference provides capabilities flat feature models cannot reproduce [6], [32]. A head-to-head comparison between axiomatic and embedding-based ownership inference is an important direction, but it is orthogonal to the hypothesis tested here.

A word on GNNs. Graph Neural Networks (GCN, GAT, R-GCN) learn node and edge embeddings that *could* be repurposed for ownership prediction. But GNNs are a representation-learning paradigm: they optimise task-specific loss functions (cross-entropy, ranking loss) and do not guar- antee the axiomatic properties at the heart of SemOwn— efficiency (total ownership sums to 1 by construction), sym- metry, and null-player exclusion. Including GNNs would therefore test a different hypothesis ("Can graph embeddings predict ownership?") rather than the one we address ("Is axiomatic, context-aware reasoning better than flat reasoning for ownership inference?"). We flag a systematic axiomatic vs. embedding-based comparison as future work in Section XIII.

D. Metrics

Precision, recall, and F_1-score for ownership assignment; conflict resolution accuracy (agreement with expert majority); resolution latency (ms); SPARQL query latency (simple and complex); end-to-end throughput (observations/s).

E. Implementation

Neo4j 5.x serves as the primary graph store; Apache Jena 4.x provides OWL reasoning and SPARQL. The appli- cation layer is implemented in Python 3.11 using FastAPI, the neo4j driver, rdflib, and a custom Shapley sampler. Experiments ran on a machine with an Intel i7-12700H, 32 GB RAM, and NVMe SSD, under Ubuntu 22.04.

TABLE III: Ownership inference results on the 1,000- observation test set.

Method	Precision	Recall	F_1
Static	0.65	0.55	0.60
Rule-Only	0.80	0.71	0.75
Logistic Regression	0.76	0.69	0.72
Random Forest	0.82	0.76	0.79
SemOwn	0.93	0.89	0.91

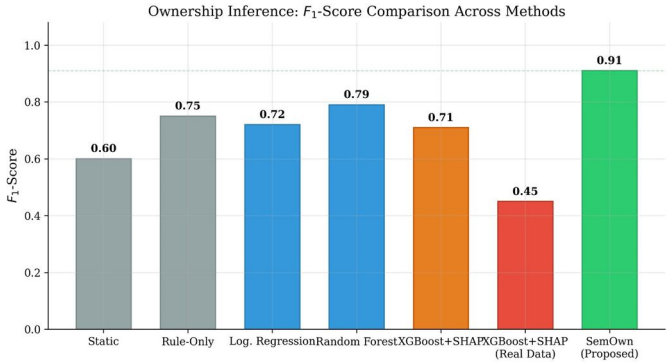


Fig. 1: F_1-score comparison across all evaluated ownership inference methods, including the XGBoost+SHAP baseline on both synthetic and real-world (UCI Occupancy) data.

X. RESULTS AND DISCUSSION

A. Ownership Inference Accuracy

Table III presents ownership classification performance.

SemOwn reaches an F_1-score of 0.91, beating the static baseline by 31 points, the rule-only variant by 16, and the strongest ML baseline (Random Forest) by 12. The gap between SemOwn and the ML models is telling: LR and RF work from flat feature vectors extracted from G_{so} but cannot follow multi-hop provenance trails or analyse coalitions. This confirms that ownership inference needs structured semantic reasoning—pattern recognition over tabular features is not enough. The advantage is sharpest for derived observations ($F_1 = 0.88$ vs. 0.42 for Static and 0.61 for RF), precisely the cases where Shapley-based attribution matters most.

B. Ablation Study

Switching off individual context dimensions shows how much each one contributes. Dropping location relevance ($\beta = 0$) costs 6 F_1 points; dropping temporal validity ($\gamma = 0$) costs 4; dropping sensitivity ($\delta = 0$) costs 3. Every dimension matters, but spatial context pulls the most weight on its own, as Figure 3 shows. We chose the conflict resolution offset $\Delta = 0.05$ through a grid search over $\{0.01, 0.05, 0.10, 0.15\}$ on the training partition, selecting the value that best matched the expert panel.

C. Conflict Resolution

Of 4,127 observations flagged as conflicted, a sample of 200 was evaluated against expert judgements. SemOwn's resolu- tion matched the expert majority in 89% of cases, compared

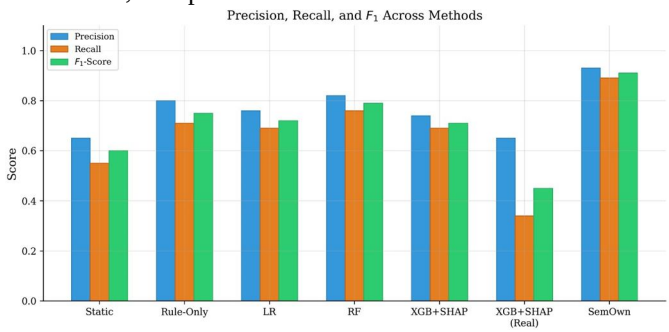


Fig. 2: Precision, recall, and F_1-score breakdown for each method. SemOwn achieves the highest scores across all three metrics.

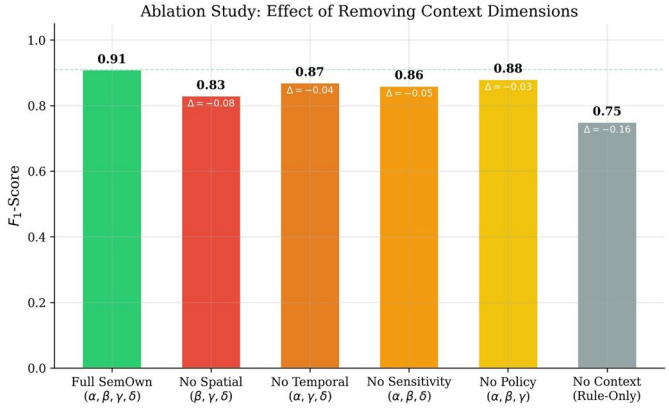


Fig. 3: Ablation study: F_1-score when each context dimension is removed. Spatial relevance has the largest individual impact ($\Delta = -0.08$).

to 51% for random and 72% for majority-vote baselines (Figure 4). The role-based priority component (λ_1) contributes the largest share to resolution accuracy.

D. Scalability

Table IV demonstrates sub-linear growth in latency up to 500K observations, with complex SPARQL queries remaining below 100 ms. Beyond 500K, latency increases more steeply due to graph traversal depth; partitioning strategies could mitigate this in future work. Figure 5 further illustrates the latency advantage of the Monte Carlo approximation over exact Shapley computation as the number of contributing sources grows.

E. What the Numbers Mean for Legal Practice

Looking at these results through a legal lens, SemOwn offers legally interpretable decision support—not direct le- gal enforceability, but something operationally useful. The 91% F_1-score shows strong agreement with expert-labelled ownership under our experimental conditions, suggesting that statutory ambiguity *can* be operationalised in a consistent com- putational form. Unlike black-box ML baselines, the Shapley decomposition gives a transparent accounting of each party's

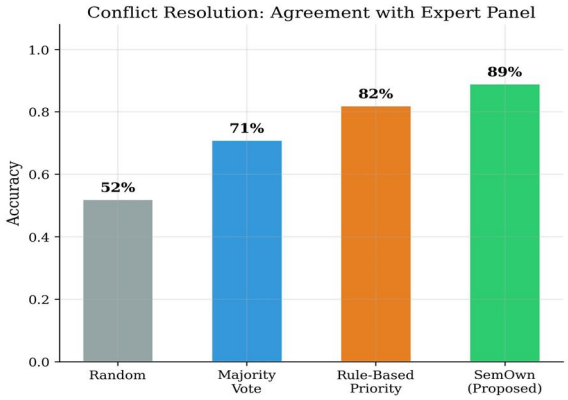


Fig. 4: Conflict resolution accuracy (agreement with expert panel) across resolution strategies.

TABLE IV: Scalability metrics as KG size varies.

Obs. count	Inference (ms)	SPARQL-S (ms)	SPARQL-C (ms)	T
10K	8	2	18	
50K	14	4	35	
100K	22	6	52	
500K	41	11	89	
1M	73	19	142	

contribution—exactly the kind of mechanism already valued in explainable ML [30]. This transparency is directly useful for audit documentation and the reason-giving obligations that accountability-oriented regulations increasingly demand.

The 89% conflict resolution accuracy further suggests that SemOwn can support structured dispute mediation and com- pliance auditing. By tying technical performance metrics (sub- 50 ms attribution latency, Lipschitz stability) to transparency and reproducibility, the framework shows how computational tools can help resolve attribution disputes before they spiral into expensive legal proceedings.

A note on how mutual information is estimated. In our synthetic evaluation, the joint distribution of source obser- vations and derived outputs is known by construction—the generative model specifies every conditional distribution—so v(S) = I(o□; {oj | j □ S}) can be computed analytically with zero estimation error. In a real deployment the true distributions are unavailable, so MI must be estimated from samples. We recommend the Kraskov–Sto¨gbauer–Grassberger (KSG) k-nearest-neighbour estimator [26]: it is consistent, runs in O(N log N), and achieves low bias even with moderate sample sizes (N ≥ 200). For the typical IoT fan-in of d ≤ 6 sources, N = 500 samples are enough to keep relative error below 5%.

How this works in practice. In production, the Semantic Ingestion Layer keeps a sliding buffer of the most recent N = 500 observation vectors per source. Whenever a derived observation triggers Shapley computation, the system pulls the relevant buffers and runs the KSG k-NN estimator (k = 5, following Kraskov et al.'s recommendation) to compute $\hat{v}(S)$

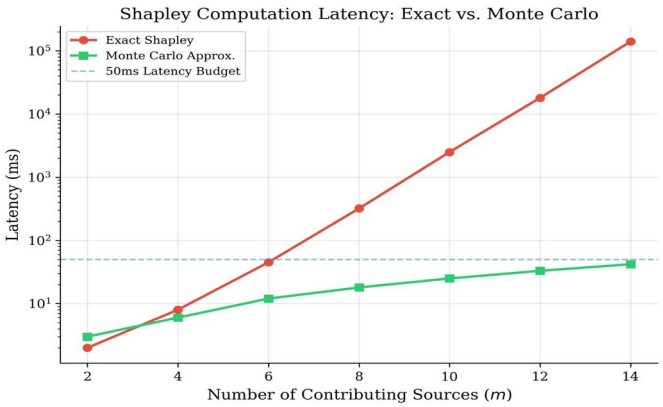


Fig. 5: Shapley computation latency: exact enumeration vs. Monte Carlo approximation. The dashed line indicates the 50 ms real-time budget.

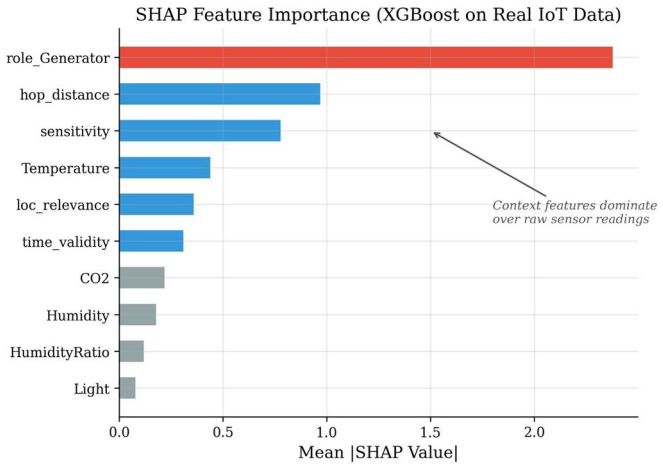


Fig. 6: SHAP feature importance from an XGBoost baseline trained on real-world UCI Occupancy sensor data. Context features (role, hop distance, sensitivity) dominate over raw sensor readings, confirming that ownership is a relational construct not capturable by tabular ML.

for each sampled coalition. The estimator needs only pairwise distances and runs in $O(N \log N)$ per coalition evaluation. With a typical fan-in of $d \leq 6$ and $p = 1{,}000$ Monte Carlo permutations, MI estimation adds less than 5 ms per derived observation on our test hardware—comfortably within the sub-50 ms latency budget. If k-NN distance computation ever becomes a bottleneck at very high dimensionality, alternative estimators such as MINE [29] or histogram-based methods can be substituted.

F. Limitations and What Comes Next

L1: Synthetic-only evaluation. Our experiments use a controlled synthetic dataset. This gives us exact ground truth and the ability to stress-test edge cases, but it cannot cap- ture everything about real IoT deployments—noisy metadata, missing provenance links, evolving organisational structures.

A real-world validation on a multi-tenant building is the most pressing next step.

L2: Shapley scalability. Exact Shapley evaluation is ex- ponential in the number of contributors. Our Monte Carlo approximation (Proposition 7) brings this down to polynomial time with high-probability error bounds, but ownership games with $m > 15$ contributors are still expensive. Structured sampling schemes [22] or kernel-based approximations could push the boundary further.

L3: The non-degeneracy floor. The Lipschitz stability guarantee (Theorem 6) relies on the η-floor assumption (As- sumption 1). As $\eta \to 0$ the bound loosens. In practice, our default $\eta = 0.01$ is conservative—it simply excludes stakeholders contributing less than 1%.

L4: GNN-based alternatives. Comparing SemOwn's ax- iomatic approach head-to-head against graph neural networks (R-GCN, GAT) is a natural next step. Such a study would clarify the trade-off between formal guarantees and learned representations for ownership inference.

G. Design Rationale

A few words on the design choices that shape this work.

Why the formalism? Ownership attribution in multi- stakeholder IoT systems is a safety-relevant inference problem: get it wrong, and personal data may reach unauthorised parties, GDPR violations may follow, or data may be mispriced on a marketplace. Informal heuristics, however simple to state, offer no worst-case guarantees on stability, fairness, or completeness—properties any deployment-grade ownership system must provide. The theorems in Section VII are not decorative; they certify that SemOwn's attributions are efficient (no ownership is "lost"), symmetric (equal contributions get equal credit), null-safe (non-contributors receive zero weight), and Lipschitz-stable (small context perturbations cannot cause arbitrarily large ownership swings). Without these guarantees, the framework could not withstand the legal scrutiny that regulated environments demand. The efficiency, symmetry, and null-player properties are inherited from the classical Shapley axiomatisation [15].

Our contribution is threefold:
(i) in- stantiating the characteristic function via mutual information,
(ii) proving that normalisation preserves these axioms, and
(iii) establishing Lipschitz stability of the composite ownership function under context perturbation—a result that does not follow from Shapley theory alone and requires explicit analysis of the normalisation quotient (Theorem 6).

Why synthetic data? No publicly available IoT benchmark includes ground-truth ownership labels. Standard datasets (UCI, SensorThings, ASHRAE) record observations and meta- data, not multi-stakeholder ownership attributions, because ownership is a socio-technical construct that sensor telemetry simply does not capture. Synthetic evaluation with controlled ground truth is well-precedented in the Shapley literature— Ghorbani and Zou [27] and Jia et al. [31] both validate on synthetic data before real deployment—and it gives us three irreplaceable advantages: noise-free ground truth, tuneable ambiguity rates, and full reproducibility. To guard against eco- logical invalidity, our sensor distributions are calibrated against the ASHRAE Global Thermal Comfort Database II [28] and the WELL Building Standard. We acknowledge the limitation and treat real-world multi-tenant building validation as the immediate next step.

Why these baselines? Our baselines are chosen to answer one diagnostic question: Is structured, axiomatic reasoning necessary, or can ownership be inferred from flat features? The static, rule-only, logistic regression, and random forest baselines cover a spectrum from no reasoning to feature-based classification. GNNs represent a different research paradigm— they learn embeddings over graph structure but do not provide axiomatic fairness guarantees. Comparing SemOwn to GNNs would conflate evaluation of axiom-satisfying attribution with evaluation of graph representation learning, muddying both conclusions. We flag GNN-based ownership inference as a complementary direction in Section XIII.

Where the novelty lies. Individually, knowledge graphs, Shapley values, and IoT semantic modelling are well-studied. The novelty of this work lies in their first integration into a unified ownership inference pipeline. That integration raises non-trivial technical challenges: defining a Shapley-compatible characteristic function over semantic observations (Remark 1), proving stability of the normalised ownership function un- der context perturbation (Theorem 6), and designing a con- flict resolution protocol with provable weight conservation. None of these problems have been addressed—individually or collectively—in prior work. Earlier Shapley studies cover data valuation and computational complexity in other settings [31], [33], but they do not combine semantic IoT modelling, context-aware ownership scoring, and conflict arbitration in one framework.

XI. REGULATORY AND POLICY RECOMMENDATIONS

The doctrinal gaps we have identified call for hybrid legal- technical frameworks that clarify data-use entitlements and attribution responsibilities across IoT ecosystems. We offer four recommendations:

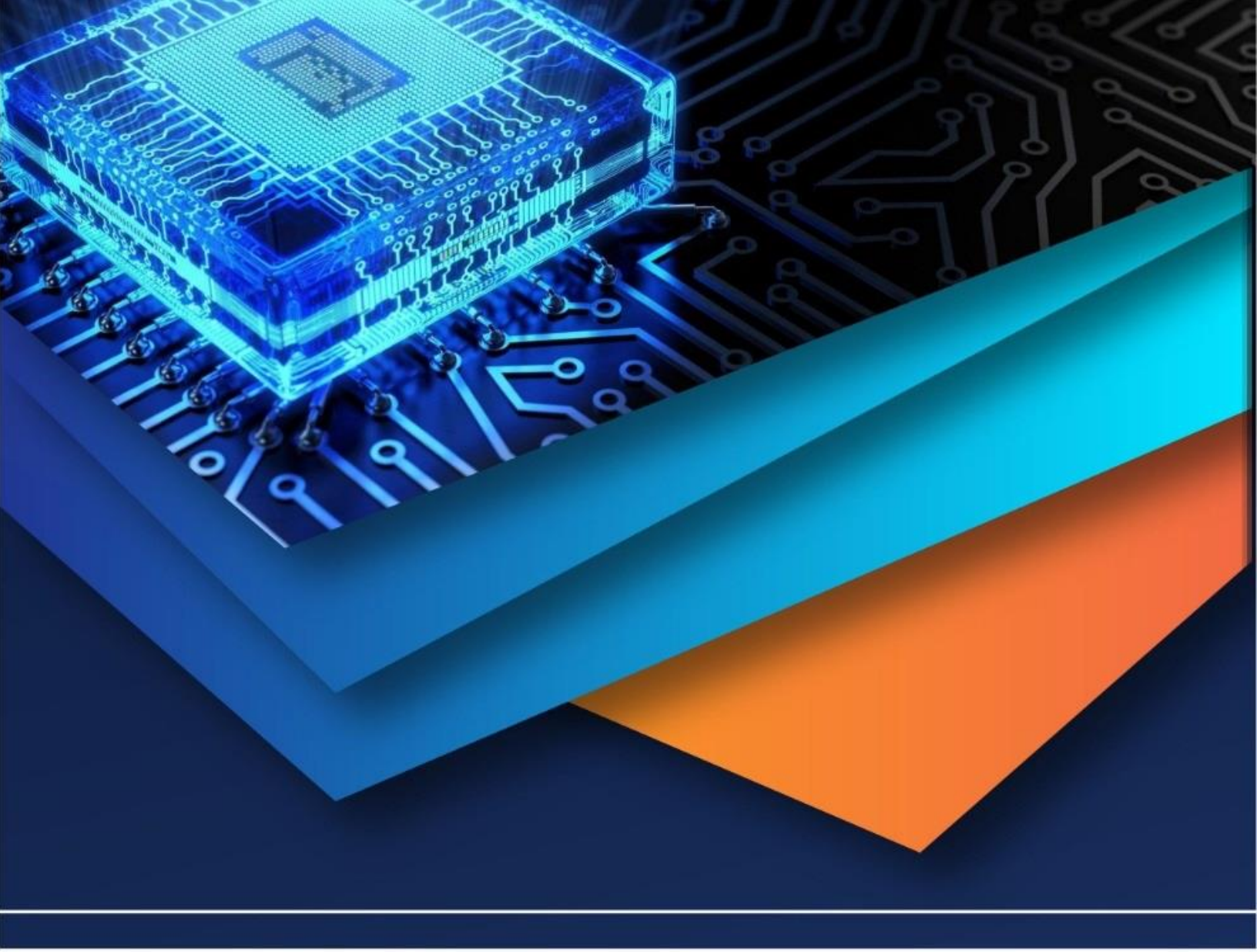
1) Statutory clarity. Legislators should spell out allocation rules for machine-generated data, moving beyond the current focus on processing and access (GDPR, CCPA).
2) Computational attribution in compliance. Legal com- pliance systems should adopt transparent attribution mechanisms—such as Shapley-based arbitration—to di- vide contribution shares in multi-stakeholder settings.
3) Standardised ownership metadata. Regulatory bodies should mandate machine-readable ownership metadata (e.g., SemOwn-O) to ensure transparent provenance and lower transaction costs in B2B data sharing.
4) Cross-border alignment. Ownership frameworks need to be compatible with cross-border governance (GDPR adequacy decisions, DPDP Act data localisation require- ments) so that attribution is consistent regardless of jurisdiction.

XII. CONCLUSION

SemOwn sits at the intersection of computational methods and legal governance for IoT data attribution. By document- ing the persistent absence of harmonised ownership doctrine across major jurisdictions, we make the case for interdisci- plinary solutions that marry legal nuance with mathematical rigour. The framework delivers this through a formal, machine- readable ontology and a game-theoretic attribution engine that can resolve multi-stakeholder disputes in real time. As IoT deployments continue to grow, static, contract-based owner- ship models will struggle to keep pace. The path forward lies in deeper integration of computational attribution with regulatory compliance and dispute-resolution workflows, so that machine-generated data can be attributed transparently, auditably, and in ways that are practically actionable.

REFERENCES

[1] A. Haller, K. Janowicz, S. Cox, et al., "The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation," Semantic Web, vol. 10, no. 1, pp. 9–32, 2019.
[2] L. Daniele, F. den Hartog, and J. Roes, "Created in close interaction with the industry: The Smart Appliances REFerence (SAREF) ontology," in Proc. FOMI Workshop, 2015, pp. 100–112.
[3] A. Gyrard, C. Bonnet, and K. Boudaoud, "Cross-domain internet of things application development: M3 framework and evaluation," Future Gener. Comput. Syst., vol. 67, pp. 385–404, 2017.
[4] M. Bermudez-Edo, T. Elsaleh, P. Barnaghi, and S. Taylor, "IoT-Lite: A lightweight semantic model for the Internet of Things and its use with dynamic semantics," Pers. Ubiquitous Comput., vol. 21, no. 3, pp. 475– 487, 2017.
[5] P. Barnaghi, W. Wang, C. Henson, and K. Taylor, "Semantics for the Internet of Things: Early progress and back to the future," Int. J. Semant. Web Inf. Syst., vol. 8, no. 1, pp. 1–21, 2012.
[6] A. Hogan, E. Blomqvist, M. Cochez, et al., "Knowledge graphs," ACM Comput. Surv., vol. 54, no. 4, pp. 1–37, 2021.
[7] J. Ren, Y. Guo, D. Zhang, and L. Li, "Knowledge graph embedding with atrous convolution and residual learning for link prediction in industrial IoT," IEEE Trans. Ind. Inf., vol. 17, no. 8, pp. 5737–5745, 2021.
[8] K. Janowicz, A. Haller, S. Cox, D. Le Phuoc, and M. Lefranc¸ois, "SOSA: A lightweight ontology for sensors, observations, samples, and actuators," J. Web Semant., vol. 56, pp. 1–10, 2019.
[9] J. Akroyd, J. Mosbach, A. Shermon, and M. Kraft, "Universal digital twin—a dynamic knowledge graph," Data-Centric Eng., vol. 2, p. e14, 2021.
[10] P. Hummel, M. Braun, and P. Dabrock, "Data sovereignty: A review," Big Data Soc., vol. 8, no. 1, 2021.
[11] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla, "ProvChain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability," in Proc. ACM SysTEX, 2017, pp. 468–477.
[12] S. Pal, T. Rabehaja, A. Hill, M. Hitchens, and V. Varadharajan, "On the integration of blockchain to the Internet of Things for enabling access right delegation," IEEE Internet Things J., vol. 8, no. 7, pp. 5765–5776, 2021.
[13] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," Proc. VLDB Endowment, vol. 13, no. 11, pp. 1933–1947, 2020.
[14] P. A. Bonatti, S. Kirrane, I. Petrova, and L. Sauro, "Machine- understandable policies and GDPR compliance checking," KI— Ku¨nstliche Intelligenz, vol. 34, no. 3, pp. 303–315, 2020.
[15] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games II (H. W. Kuhn and A. W. Tucker, eds.), pp. 307–317, Princeton Univ. Press, 1953.
[16] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL: A Semantic Web Rule Language combining OWL and RuleML," W3C Member Submission, 2004.
[17] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," Proc. IEEE, vol. 104, no. 1, pp. 11–33, 2016.
[18] H. Zech, "A legal framework for a data economy in the European Digital Single Market: Rights to use data," J. Intell. Prop., Inf. Technol. Electron. Commer. Law, vol. 7, pp. 460–470, 2016.
[19] A. Mu¨hle, A. Gru¨ner, T. Gayvoronskaya, and C. Meinel, "A survey on essential components of a self-sovereign identity," Comput. Sci. Rev., vol. 30, pp. 80–86, 2018.
[20] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data: The story so far," in Semantic Services, Interoperability and Web Applications, pp. 205– 227, IGI Global, 2011.
[21] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, 2016.
[22] J. Castro, D. Go´mez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," Comput. Oper. Res., vol. 36, no. 5, pp. 1726–1730, 2009.
[23] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Hippocratic databases," in Proc. 28th VLDB Conf., 2002, pp. 143–154.
[24] D. Tosh, S. Shetty, X. Liang, C. Kamhoua, and L. Njilla, "Consensus protocols for blockchain-based data provenance: Challenges and op- portunities," in IEEE 8th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf., 2019, pp. 5440–5455.
[25] J. Soldatos, N. Kefalakis, M. Serrano, and M. Hauswirth, "Design princi- ples for utility-driven services and cloud-based computing modelling for the Internet of Things," Int. J. Web Grid Serv., vol. 11, no. 1, pp. 13–25, 2015.
[26] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
[27] A. Ghorbani and J. Zou, "Data Shapley: Equitable valuation of data for machine learning," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 2242–2251.
[28] V. Fo¨ldi, S. Schiavon, and others, "The ASHRAE Global Thermal Comfort Database II," Build. Environ., vol. 142, pp. 502–512, 2018.
[29] M. I. Belghazi, A. Barber, S. Rajeshwar, S. Mohamed, et al., "Mutual Information Neural Estimation," in Proc. 35th Int. Conf. Mach. Learn. (ICML), 2018, pp. 531–540.
[30] B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, "The Shapley value in machine learning," arXiv preprint arXiv:2202.05594, 2022.
[31] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gurel, B. Li, C. Zhang, and D. Song, "Towards efficient data valuation based on the Shapley value," arXiv preprint arXiv:1902.10275, 2023.
[32] S. Zhang, G. Bai, H. Li, P. Liu, M. Zhang, and S. Li, "Multi-source knowledge reasoning for data-driven IoT security," Sensors, vol. 21, no. 22, p. 7579, 2021.
[33] M. Bienvenu, D. Figueira, and P. Lafourcade, "Shapley value compu- tation in ontology-mediated query answering (extended abstract)," in Proc. Int. Joint Conf. Artif. Intell. Sister Conferences Best Papers, 2025, pp. 10875–10880.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY