JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detecting "Shadow Identity": Models for Discovering Unmanaged Authentication Outside Central Identity Providers

[1]NARENDRA TALLAPANENI,
[1]Senior cybersecurity engineer,
[1]Financial services and life insurance domain,
[1]Independent researcher with experience in the identity and access management (IAM) cybersecurity industry

[1]144411, India.

Abstract : Enterprise identity governance is designed around a limited number of central Identity Providers (IdPs), such as Okta, Microsoft Entra ID and Ping ID. Enterprise identity governance is based on as few as necessary identity providers (IdPs) that federate human sign-on using single sign-on (SSO), SAML and OpenID Connect (OIDC). However, a significant and ever-increasing portion of authentication is not sent across this plane. All of these credentials sit outside of the central IdP's view and control: local operating-system accounts, statically embedded service-account credentials, unmanaged software-as-a-service (SaaS) logins, developer-issued API keys, OAuth grants negotiated between objects directly in the running applications, and ever more independent artificial-intelligence (AI) agents that log on via local or scripted access paths. This group of ungoverned, un-Federated authentication occurs, in this paper, in a manner that is referred to as "shadow identity." We formalize the problem of detecting shadows, place it in the context of three related fields in the exterior community: Non-human identity (NHI) governance, workload attestation standards, and graph-based identity analytics, propose a layered detection architecture, the Multi-Source Shadow Identity Detection (MSSID) framework. MSSID combines network- or endpoint-level telemetry, cross-source log correlation, construction of heterogeneous identity graphs, and three complementary detection models: a coverage-gap model, a graph-neural-network-style (GNN-style) behavioral anomaly model and an attestation-consistency model, to uncover authentication activity that has escaped the radar of the centralized governance approach. In addition to the proposal for the architecture, we deploy three reference copies of all three proposed detection models and test them end-to-end in a controlled synthetic test domain with 1,590 identity nodes and 190 shadow instances, distributed across all six taxonomy classes we define, which is currently not available in the public domain. The operating system used in this system has implemented it with a precision of 0.89, a recall of 0.72, an F1 score of 0.79, a cover ratio of 0.90 compared with the seeded ground truth, a mean simulated time to discovery of 1.0 day for the fusion operating point, and 90.5% accuracy for attribution. The results per class support the main design claims of the framework: The predominant model for none of the six classes of shadow identity; and the contributions of the three models are remarkably complementary rather than redundant. These results are not reported as a measure of validated results in the field; rather, the reports are the only measure of internal coherence and potential for production deployment at testbed scale. Finally, we conclude with a discussion about evasion resistance, the implications for governance, limitations, and a research agenda for detecting shadows of identities under attestation in a privacy-preserving and explainable way.

IndexTerms **- shadow identity; non-human identity; identity and access management (IAM); identity threat detection and response (ITDR); graph neural networks; workload identity; SPIFFE/SPIRE; zero trust; unmanaged authentication; shadow IT**

I. INTRODUCTION

In the past ten years, enterprises have flocked towards a central control plane, the centralized ID provider, for both authentication and authorization. Credential sprawl for human users is reduced through support for federation with SSO, SAML and OIDC, lifetime

management policies are applied to accounts behind that plane with identity governance and administration. This is a good architecture for the number of identities it is supposed to handle. The other doesn't work as well!

Security researchers and practitioners have reported on a type of invisible identity which exists and also authenticates but is not known to the central Identity Provider (IdP): shadow identities. These may consist of machine identities that never get registered to directory services or have privileged local login accounts that remain after development or testing or are received to production systems via paths that were not provisioned by IT [1]. Identity providers that pass authentication logs downstream are thus vulnerable to downstream security analytics tools detecting nothing for any user that logs in via a browser extension, a credential stored on local disk, or a personal SaaS login [2]. The result is a structural blind spot: identities an organization knows to scan generate the very telemetry which is crucial to every security operations program.

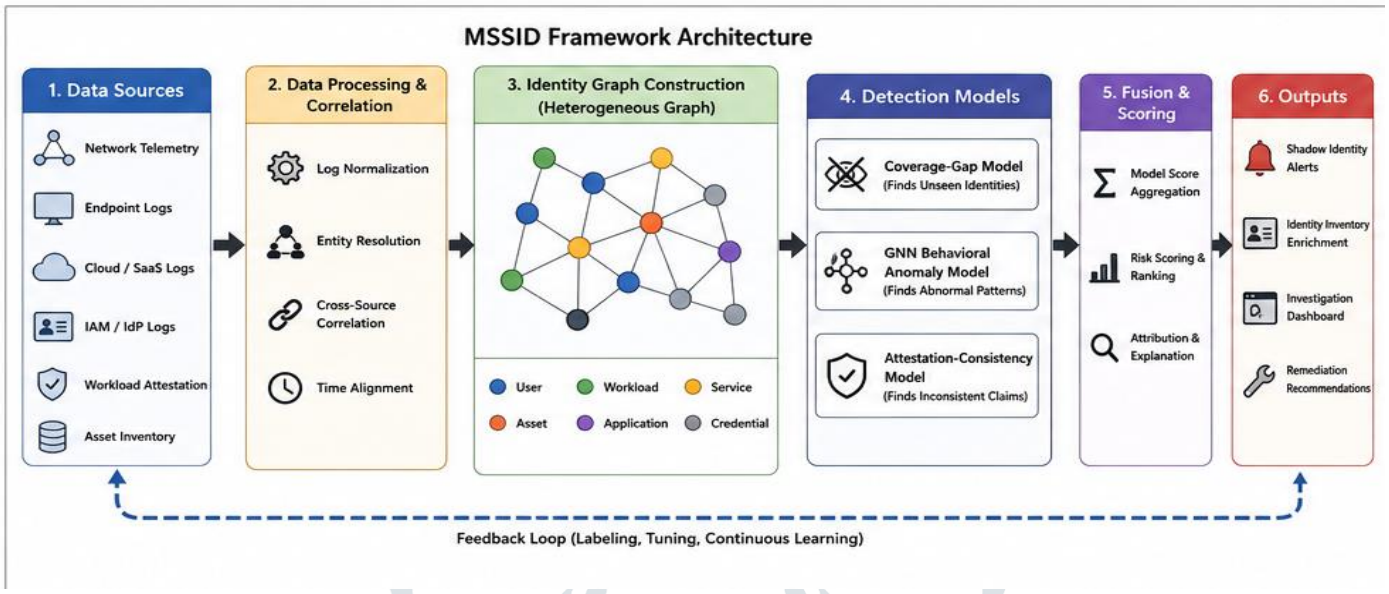

Fig.1. Multi-Source Shadow Identity Detection (MSSID) Framework

This issue is being exacerbated by three forces. First, shadow IT has gone from unsanctioned applications to unsanctioned identities integrated into sanctioned infrastructure, such as OAuth and SAML integrations that users permit without the security checks. Firstly, shadow IT has evolved from unsanctioned applications to unsanctioned identities embedded within otherwise sanctioned infrastructure, including OAuth and SAML integrations that individuals approve without security review [3]. Second, the number of non-human identities (NHIs) has surged much quicker than the number of human identities, with ratios of around 45:1 of NHIs to human identities being used across the typical enterprise and 144:1 of NHIs to human identities in cloud-native and DevOps type environments, while more recent industry data suggests that the ratio may now exceed 80:1 [4], [16]. Third, various autonomous AI agents are starting to authenticate via local accounts or direct API calls without ever triggering an IdP-visible authentication event, which means no new event to detect IdP compromise, or use in security information and event management (SIEM) tooling using IdP events [5], [6], [7].

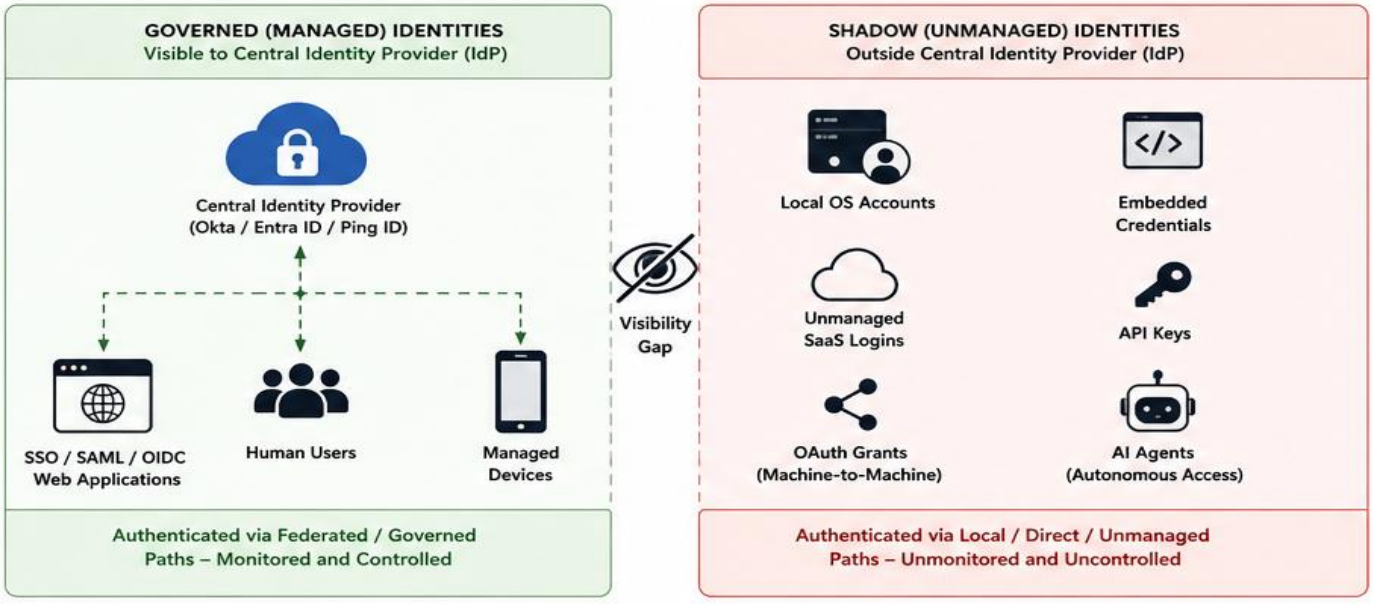

Fig.2. The Identity Visibility Gap: Governed vs. Shadow Identities

Whether or not an organization's IAM program is developed on paper, there will be an identity that never shows up in the IdP, won't be for which policy can be enforced at a lease privilege, won't be subject to access review or won't be deprovisioned at a suitable time. Managed discovery endpoints and identity providers have been the only option for discovery and at the same time introduction of network traffic analysis, secrets scanning, and audit of OAuth/SAML. All these practices point to the same first step: expand

discovery beyond managed endpoints and identity providers, then follow it by introducing whatever is discovered through a single governance lifecycle [1, 8]. A systematical model of the signal indicative of a shadow identity, of how the signal should be integrated and correlated from diverse identity data sources, and of the evaluation of a signature detection system where there is no central log source in which one can establish reality as ground truth is comparatively underdeveloped, especially more in the peer-reviewed and formal research literature.

This paper makes five contributions:

- A formal problem framing that distinguishes shadow identity detection from the adjacent, better-studied problems of shadow IT discovery and non-human identity governance.
- A synthesis of detection primitives drawn from four literatures—applied shadow-IT/NHI security research, non-human identity governance, workload attestation standards (SPIFFE/SPIRE), and graph- and GNN-based identity analytics—that have so far developed largely in isolation from one another.
- The Multi-Source Shadow Identity Detection (MSSID) framework: a four-layer reference architecture combining multi-source collection, heterogeneous graph construction, a three-model detection ensemble, and risk-scoring/response.
- A reference implementation of all three detection models and a controlled synthetic testbed seeded with ground-truth shadow identities across all six taxonomy classes, together with the resulting precision/recall/F1, coverage ratio, mean time to discovery, false-positive rate, and attribution-accuracy figures addressing the absence of a public labeled benchmark for this problem.
- A discussion of evasion resistance, deployment limitations, and a forward research agenda, informed by which detection model in the ensemble proved responsible for catching which taxonomy class in the testbed evaluation.

This paper will be concluded as follows. The Learner's Book reviews similar work in part 2. In Section 3, the problem and threat analysis is defined. The MSSID framework is presented in section 4. The testbed and reference implementation, and the empirical results, are detailed in Section 5. The Implications part is found in Section 6, the Limitations in Section 7, the Future work in Section 8 and the Conclusion/in Section 9.

II. Related Work

2.1 Shadow IT and Shadow Identity

Previous studies on shadow IT have tended to concentrate on the unsanctioned devices, applications, and cloud services used outside of the scope of IT [3]. The concept has been expanded over the past few years to identity, where every network connected system is an identity, and these identities can exist without any involvement of the systems meant to discover and manage them that creates the network. Push Security, meanwhile, succinctly indicts the fundamental lack of visibility as one of the problems identity providers push logs to security analytics tooling, but any login or authentication that isn't via the IdP generates no log, meaning that browser-based and direct-to-SaaS logins are essentially invisible to security operations [2]. Tailored analysis advice from Adaptive Security also makes it clear that no single discovery technology, such as network traffic analysis, IdP log review, financial/expense-report mining, endpoint monitoring or email intelligence is enough on its own, and several applied together deliver complete coverage [3].

2.2 Non-Human Identity Governance

A literature that has come a long way since, focused on non-human identities as a governance category separate from, yet not independent to, shadow identity has emerged. NHIs are identities that are not necessarily residing within people, such as service accounts, workloads, CI/CD pipelines, AI agents and RPA bots that can authenticate and act without having to be driven by humans at every step [9]. In recent synthesis research, identity for machines evolved from X.509 certificates and SSH keys to cloud service accounts, API keys, container identities and credentials for AI agents, (see [10] for example), and privileged access management (PAM) systems were explored as layers of policy-driven orchestration for an ever-increasing number of machine identities. Legacy identity models rely on the role-based access control and HR-driven provisioning lifecycle, whereas this was created for humans and cannot be applied to ephemeral, autonomous and system-generated entities; the work proposes a model that unifies the lifecycle, authentication, authorization and observability of human and non-human actors as first-class citizens of a single identity system [11]. This related empirical synthesis maps the magnitude of shifting towards machine-identities, towards incident patterns, and the gap in governance, using a proposed unified governance framework [12]. Industry measurement continues to show that there are many more NHIs than human identities and repeatedly the ratio of parts NHIs to human parts has been increasing over time as cloud-native, agentic-AI becomes more popular and ubiquitous year-to-year [4, 13]. A certificate-based machine identity also has limited reporting on public-key-infrastructure (PKI) governance [14] and AI-agent credentials, in particular, have been largely left alone in terms of their IAM/PAM governance coverage, as these newer tools were not built to engage with this [15]. A new taxonomy takes NHI governance concerns further, beyond organizational and geopolitical boundaries, and categorizes a 37-category taxonomy of AI-identity risks, along with a six-domain taxonomy of governance criteria for AI, while reporting NHI-to-human ratios over eighty times greater for enterprise systems [16].

2.3 Workload Identity and Attestation Standards

The third related body of literature is around cryptographic workload identity, with the most recent being the Secure Production Identity Framework for Everyone (SPIFFE) and the reference runtimes, SPIRE. Instead of issuing identity documents based on preshared credentials and secrets, SPIFFE issues short-lived, cryptographically verifiable identity documents (SVIDs) to workloads based upon runtime attestation. An SVID is issued via a local Workload API that does not require any credentials or secrets and in turn is bound by the runtime attributes verified by a SPIRE agent, including attributes that refer to the node itself, as well as to the workload itself, such as cloud instance metadata, Kubernetes service-account identity, container image hash, etc. [17, 18]. Due to this, SPIFFE/SPIRE can be considered in the literature a workload-identity paradigm for multi-cloud, cloud-native, environments related to multi-factorial authentication in human environments, and NIST SP 800-207A explicitly indicates SPIFFE as reference application-identity infrastructure for zero-trust architectures in multi-cloud, cloud-native environments [19] and [20]. Follow-on work has extended SPIFFE-based attestation into CI/CD pipelines [17], credential-broker patterns that separate identification from enforcement of access policies [21], unified “identity control plane” proposals that propose to reconcile the human readable OIDC/SAML identity with the workload-assigned SPIFFE identity by common policy [22] and scoped, short-lived OAuth 2.0 action

tokens [23] during runtime, linked to trusted-execution-environment attestation for an agentic AI system. The structural significance that can be drawn from the above literature is as follows: Wherever SPIFFE/SPIRE or an equivalent attestation scheme is employed, the appearance or disappearance from the chain of a valid attestation is an additional, stand-alone signal of the governance of a machine identity authentication event.

2.4 Graph-Based and Machine-Learning Identity Anomaly Detection

Graph representations and graph neural networks (GNNs) are used in a fourth literature to represent and process identity and access data. Work models have recently been extended to represent heterogeneous dynamic graphs of users, roles, sessions and actions, demonstrations of attention-based graph embeddings exhibit significant improvements in the F1 score of the detector over baseline detectors while maintaining similar throughput to production [24], [25]. Further work in the broader area anomalous detection in a heterogeneous graph network (HGN) has been proposed for enterprise access fraud detection (AFD) systems with the authors observing that evaluation of methods is not a standardized process and the lack of an appropriate benchmark is the major challenge identified [26]. Complementary patent literature from vendors of the identity-governance systems outlines graph-based peer-group risk scoring as well: As long as identities have access patterns that are just different from their associated "peer group," an identity whose access pattern is very different from the "peer group" baseline can be flagged just based on graph and event-log structure, even if there are no remaining, labeled examples of training identities and data for known attacks [27, 28]. Taken together this literature indicates graph structure is a compelling substrate for the detection of anomalies in identities – but this is applied to logs that the IdP or IAM platform has already created, in almost every paper so published. It's singular in the sense that it cannot tackle identities that don't create any log in the first place.

2.5 Synthesis and Gap

These four literatures are not "very for connected. While shadow-IT and applied shadow-identity research correctly identify the observability gap, it provides only a manual or checklist approach to discovery, and lacks a formal detection model [1, 2, 3]. While the governance research related to NHI creates a framework for managing identities and policies across the lifecycle, it typically focuses on the discovery of identities and their onboarding into a process of managing identities and policies through a governance mechanism [9]–[16]. In the subset of machine identities that are running in environments with SPIFFE/SPIRE or comparable security infrastructure [17]–[23], workload attestation standards can yield a strong binary signal ‘attested’ or ‘not attested'. Primarily the data source of shadow identities is not found in the works of [24]–[28] and therefore the anomaly detection algorithms used in these systems are built on IdP- or IAM-generated logs and hence can provide powerful machinery for modelling. We did not find any work related to identifying a missing central authentication event based on access activities that are independently observed, as the first signal against which to be detected, fused together from network, endpoint, secrets, and attestation data sources. This is the gap to fill which is the mission of the MSSID framework from Section 4, based on the applied industry frontier of “shadow”-access correlation, which precisely tracks access activity on the network level; and correlation which involves a parallel session foreseen by the IDP to facilitate the identification of the unauthorized path [5], [6], [7], [8]]; and formalizes it as a layered model-based detection architecture, which, as we also implement and test in Section 5, is the key to realize the mission of the MSSID framework.

III. Problem Definition and Threat Model

3.1 Definitions

A central identity provider is the federated authentication and policy system(s) that a central organization determines to be "authoritative" for who is who in the organization usually an SSO/SAML/OIDC platform coupled with IGA solution. The combination of systems and access points whose authentication events are known and controlled by the central IdP is called an identity's federation boundary. A shadow identity is any person (as a human actor) or any machine (as an acting entity) that authenticates to the organization's resource through a path outside of the federation boundary in a way that no authentication event will be generated for or attributable to the central IdP. An event that allows such access is called a shadow authentication event.

3.2 Taxonomy of Shadow Identity Classes

We distinguish six overlapping but analytically useful classes of shadow identity relevant to detection design. Section 5 seeds a synthetic testbed with instances of all six classes and reports which of the three MSSID detection models catches each one.

- Class 1 - Local/OS-level accounts that exist outside directory federation (e.g., local administrator accounts on endpoints or servers never joined to the managed directory).
- Class 2 - Static, embedded machine credentials API keys, database passwords, or tokens hard-coded into source code, configuration files, container images, or CI/CD pipeline definitions—that authenticate independently of any workload-identity or IdP mechanism.
- Class 3 - Unmanaged SaaS logins, where an employee authenticates to a sanctioned or unsanctioned SaaS application using a personal account or locally created credential rather than the organization's SSO integration.
- Class 4 - Unregistered OAuth/SAML application-to-application grants, where a user authorizes a third-party integration (e.g., within Google Workspace, Microsoft 365, or Slack) that establishes a standing authentication relationship never reviewed by security or IAM teams.
- Class 5 - Certificate- or key-based machine authentication operating outside PKI/certificate-lifecycle governance, common where workload-identity standards such as SPIFFE/SPIRE have only partial deployment coverage. Some instances of this class are nominally registered in a secrets manager or CMDB even though they lack a valid attestation chain a distinction the evaluation in Section 5 treats explicitly.
- Class 6 - Autonomous AI agents that authenticate through local accounts, direct API calls, or borrowed human sessions rather than a governed workload or delegated identity, an increasingly dominant sub-class as agentic AI adoption accelerates [5]–[7]. We further distinguish, within this class, agents that are entirely unregistered from agents that are nominally registered but behaviourally anomalous relative to their peer group the latter being the harder detection case discussed in Section 5.4.

3.3 Why These Events Evade Detection

Structural patterns are the same for all 6 classes and include the use of mainstream detection tooling (SIEM correlation rules, user-and-entity behavior analytics (UEBA), and ITDR platforms) that ingests the IdP's authentication events. If an access path does not cross the IdP, then it does not simply result in a “poor” “but incomplete” event; it does not result in any event at all; it is just a

tooling that relies on an access path that was never called [5]. It is different from classical evasion, which is when an attacker manipulates or suppresses a log record after they occur – in this case, no log record is created in the regular course of the interaction because the interaction was never designed to create one. Improvements in log analytics would not be sufficient to adequately detect access activity, however; the source of the access observations should be separate from the requirement for the IdP in the first place: those are either network telemetry, endpoint/browser instrumentation, secrets scanning, or attestation verification [6, 7].

3.4 Scope and Assumptions

This paper restricts the scope of the detection problem to one organizational security boundary using hybrid (on-a-premises and cloud) infrastructure, where at least one central IdP is deployed and logs are created for what it manages. We assume that the defender's access to network telemetry, endpoint/ browser telemetry, cloud control plane (CCP) logs, and source-code/ secrets-scanning results have been made possible; we do not assume that the CCP or universal SPIFFE/SPIRE (or equivalent) is deployed, but treat partial deployment as a valuable additional signal source (Section 4.4.3). Access-control identity purely through badge/door (some of the most common methods) and covert channel exfiltration at the level of nations are not a part of the scope of work and are investigated in parallel, not as part of this document.

IV. Proposed Framework: Multi-Source Shadow Identity Detection (MSSID)

We aim to propose MSSID, a 4-layer reference architecture for detection of shadow identity activity. This is the way the architecture is depicted in Figure 1, narratively described below, as a sequence of processes sequenced into a pipeline: multi-source collection, entity resolution, graph construction, three model detection ensemble and risk-scoring, response. (This paper does not include vendor-specific implementation diagrams.) In Section 5.2, we shall explain, explain, implement, and play with the implementation of this architecture, on a testbed scale.

4.1 Layer 1 - Multi-Source Collection

Because no single data source observes all shadow identity classes, MSSID collects from six source types in parallel, consistent with applied guidance that no individual discovery technique is independently sufficient [1], [3]:

- Network traffic analysis, to observe authenticated sessions and API calls regardless of whether they originated from an IdP-mediated login.
- Endpoint and browser telemetry, to capture credential entry, session-token issuance, and locally stored credentials, including browser-extension-mediated logins that never reach IdP logs [2].
- Secrets scanning across source-code repositories, build logs, container images, and configuration stores, to surface embedded static credentials.
- OAuth/SAML application-grant auditing within major SaaS platforms, to enumerate third-party integrations authorized outside formal review.
- Cloud control-plane and IaaS/PaaS provisioning logs, to detect ephemeral workload identities created outside standard infrastructure-as-code review.
- Central IdP logs themselves, used not as the primary detection source but as the authoritative negative reference set: any identity or session observed by another source but absent from this set is a shadow candidate.

4.2 Layer 2 - Entity Resolution and Heterogeneous Graph Construction

Layer 2 is used to model the raw events of Layer 1 into a heterogeneous identity graph in the general modeling fashion developed by cloud IAM graph analytics [24, 25] and graph-based identity-governance risk scoring [27, 28] with the difference of new node and edge types which are specifically associated with the shadow identity problem. The Node types are: human Identities, non-human Identities, Devices/Endpoints, Applications and SaaS tenants, Credentials (API keys, Certificates, Tokens, Passwords). Edge types are the next, owns, is-associated-with, co-occurs-with (temporal, session level proximity), and federated-by (explicit edge saying that the authentication event was federated by the central IdP). One important correlation pattern, derived from applied ITP and IDR experience, is that a human identity's IdP-authenticated session can be correlated with another identity's parallel path (which is not IdP-authenticated) of an agent/extension/service running on his/her behalf; observing both at the same time should allow the graph to distinguish the non-federated path as anomalous relative to its related federated path, even if neither path is anomalous alone.

4.3 Layer 3 - Detection Model Ensemble

MSSID's central methodological contribution is a three-model ensemble, each targeting a different evidentiary basis for shadow identity, on the premise that no single model class is sufficient across all six classes defined in Section 3.2. Section 5 tests this premise empirically.

4.3.1 Model A - Coverage-Gap Model

A coverage-gap model is a set-reconciliation model – it scans all of the identity nodes found in the Layer 2 graph and calculates the intersection of that set with the set of identities available in the central IdP which are registered and managed. All the elements of the graph without a federated-by link to IdP and without a governance record (e.g., an agreed exception, a registered service-account entity in a secrets manager) are marked as 'shadow' candidates. This model does not need a baseline of behaviors or training data and instead normalizes the discovery process advocated for applied guidance as a detectable and verifiable function that is continuously reassessed instead of a time-consuming manual process [1]. The simplicity is the biggest drawback of the model, as seen in the examples below in section 5.4.

4.3.2 Model B - Graph-Neural-Network-Style Behavioral Anomaly Model

The behaviors modelled approach is applicable for shadow identities that are technically registered somewhere (e.g. a service account that has been created, possibly by not been reviewed) but where the way the identity is being used is different from what is expected from a governed identity of the same type in comparison to the peer group. Similarly, as described in [24, 25] to encode the heterogeneous-graph of cloud IAM logs, MSSID uses attention-based aggregation over features of nodes and edges (entity type, resource sensitivity, access frequency, temporal pattern, geographic/network origin), and outputs an anomaly score per node. Like prior approaches to identity-governance risk-scoring based on graphs, the underlying notion is that access histories of similarly typed, similarly-privileged identities should be in a particular sense similar, and access histories that are obviously and significantly different from others in the peer group for the identity, even if a labelled attack example exists for that history, should constitute high-risk candidates [27], [28]. This is an extension to model A in that it can expose model A service accounts being used without any help of

the secrets manager, such as their preparations being made through another means, such as a script or scheduled task. The expression of attention weighted aggregation used for evaluation with the testbed is explained in section 5.2.

4.3.3 Model C - Attestation-Consistency Model

In environments that have some of the entities supported by this model (assuming that they also have certificate deployments that contain certificates for the workload identities, such as SPIFFE/SPIRE certificates, or that are compatible with them), MSSID introduces an attestation-consistency check: it verifies that the observed authenticated actions for a workload can be tied to a corresponding valid SVID and attestation chain. The lack of a valid attestation chain for a workload authenticating through a static attestation credential, taxonomy, or local account, and the lack of a distributable, secret-known credential for authenticating the type of service it is using is a strong witness, and for the most part impervious to evasion: it is a shadow signal that a workload is using an authentication path that the attestation-based identity system does not recognize [17] [20]. Adds a signal of its own when attached to workloads outside of the deployment range of SPIFFE/SPIRE, supports workload-identity adoption over time as more workloads are added to the system, and likewise asserts more and more power as workload-identity adoption grows. As workload-identity adoption grows, turnout grows; as more workloads are added to the system, its signal gets stronger, while in environments where workload-identity is not broadly adopted, the model does not degrade as much as other options might.

4.4 Layer 4 - Risk Scoring, Attribution, and Response

Layer 4 combines the results of Models A–C and Creates a weighted risk estimate for each identity based on the blast radius of the identity (e.g., number of downstream systems that identity may be reached) and the sensitivity of the resources accessed. If the layer detects that such an identity can exist, it performs ownership mapping: It returns the identity's originating human, team, or automated pipeline, based on the version control system commit history, tagging of cloud resources, and the edges in the Layer 2 graph. As was applied in the research, risk reduction with visibility is not achieved simply by alerting, MSSID's response layer is intended to be thrown into existing ITDR & PAM remediation processes bringing discovered identity under the organization's provisioning, least privilege, periodic review & deprovisioning process. As section 5.4 shows, the naive (equal) union, "any model fires", has the highest recall, but also the lowest precision, whereas a majority-voting model (at least two of three models fire) has a higher precision, yet a moderate recall cost: these are key points that indicate that weighting/threshold selection is an operational consequential design choice, rather than just a formality.

V. Evaluation Methodology, Reference Implementation, and Results

The core methodological difficulty with this problem domain is the lack of an openly-labaled, public, benchmark dataset of shadow identity events, not always captured by the logging system most popular security datasets are derived from by definition. We take steps towards this in two directions: We give a specific design of a controlled-testbed in Section 5.1, and Section 5.2 presents a concrete example of the implementation of MSSID's three detection models for this paper, along with an explicit discussion of what they do and do not establish.

5.1 Testbed Design

A baseline of 500 human accounts (treated as federated identities, using a test IdP) and 900 non-human identities (treated as registered, with 70% containing a valid simulated SPIFFE / SPIRE attestation record) is placed into a synthetic identity population, and to which 190 shadow identity instances are deliberately added across all six taxonomy classes (Table 1). For each node a feature vector is associated (access frequency, sensitivity of resources, temporal irregularity, geographic/network diversity, blast radius), the first seen day in a simulated monitoring window of 30 days while the ground truth includes labels either shadow/governed or taxonomy class and true owning team; the latter are used only during evaluation, not when training the detection models. Class 5 and a subsample of class 6 are intentionally configured with a null record of governance and no valid attestation and no IDP federation (otherwise known as the "nominally registered but actually ungoverned" case) to enable the testbed to differentiate between the three models as well as between the combined effect of models and a null case of no governance and no valid attestation and no IDP federation. Since all the instances are known and explicit, they provide a ground truth for evaluating the performance of the detection and tackling a lack of a public benchmark. This is an extension of the full generator the full deterministic (fixed random seed) generator is available as a supplement to this paper.

Table 1. Seeded Shadow Identity Classes and Governance Characteristics

Class	Description	Seeded n	Federated?	Governance record?	Attested?
1	Local/OS accounts outside directory federation	30	No	No	N/A
2	Static, embedded machine credentials	40	No	No	N/A
3	Unmanaged SaaS logins (personal accounts)	35	No	No	N/A
4	Unregistered OAuth/SAML app-to-app grants	25	No	No	N/A
5	Cert/key machine auth outside attestation	30	No	~50% Yes	No
6a	Fully unregistered AI agent access	12	No	No	N/A
6b	Nominally registered, behaviorally anomalous AI agent	18	No	Yes	No

5.2 Reference Implementation

Instead of the architecture being notional, we developed all 3 MSSID detection models in Python and ran the models on the testbed population above. The choice of implementation is lightweight, low dependency implementations suitable for testbed size as well as for production use, but not a production-ready GNN framework; the details of such a GNN are explained carefully to allow for correct interpretation of results in Section 5.3-5.4.

- Model A (coverage-gap) is implemented exactly as specified in Section 4.3.1: a node is flagged if it has neither a federated-

by-IdP edge nor an explicit governance record.

- Model B (behavioral anomaly) is implemented as a single-layer, attention-weighted neighbor-aggregation encoder written from scratch in NumPy: for each node type, a k-nearest-neighbor graph (k=15) is built over standardized feature vectors, softmax attention weights are computed from negative squared feature distance, and each node's embedding is the attention-weighted average of its neighbors. A node's anomaly score is its residual distance from the attention-aggregated centroid of the governed population within its own type; nodes in the top 12% of residual distance per type are flagged. This operationalizes the peer-group-deviation premise of Section 4.3.2 at a scale and dependency footprint appropriate for a reference implementation; it is not a claim to reproduce the full modeling capacity of the heterogeneous GNN architectures surveyed in Section 2.4.
- Model C (attestation-consistency) is implemented as specified in Section 4.3.3: non-human nodes with a recorded attestation status of "invalid/absent" are flagged; nodes with no attestation status recorded at all (i.e., outside the simulated SPIFFE/SPIRE deployment footprint) contribute no signal, reflecting the graceful-degradation behavior described in Section 4.3.3.
- Layer 4 fusion is evaluated at two operating points: a union rule (any one of the three models firing) and a majority-vote rule (at least two of three models firing), to make the precision/recall trade-off explicit rather than reporting a single, arbitrarily chosen threshold.
- Mean time to discovery is simulated under a daily batch-monitoring cadence: Models A and C are assumed available one day after an identity's first activity (collection latency), and Model B is assumed to require a two-day minimum observation window before its peer-group features stabilize. This is a modeling assumption made explicit here, not a measurement of any real collection pipeline's latency.

5.3 Results Against the Testbed

Table 2 reports precision, recall, F1, and false-positive rate per 1,000 monitored identities for the full MSSID ensemble at both fusion operating points, for each model in isolation, for two baseline comparators, and for three single-model ablations of the full ensemble.

Table 2. Performance of MSSID Detection Models and Ablation Variants

Configuration	Precision	Recall	F1	FPR /1,000
IdP-log-only baseline	-	0.000	0.000	0.0
CASB-only baseline	1.000	0.316	0.480	0.0
Model A only	1.000	0.832	0.908	0.0
Model B only	0.634	0.637	0.635	52.6
Model C only	0.156	0.253	0.193	228.1
MSSID full – union (≥ 1 model)	0.378	1.000	0.548	288.0
MSSID full – majority vote (≥ 2 models)	0.889	0.716	0.793	12.3
MSSID ablate Model A (B+C only)	0.318	0.768	0.450	288.0
MSSID ablate Model B (A+C only)	0.422	1.000	0.594	228.1
MSSID ablate Model C (A+B only)	0.720	0.947	0.818	52.6

Three noteworthy points arise. First, that shadow instance will by definition have no IdP log record for IdP-log-only baseline to act on, by construction: no shadow instance, no IdP log record - by construction. This is not a limit to the tuning of the baseline, but rather the core limitation that the paper outlines that leaves room for launching MSSID. Second, as stated in Section 2.1 in the applied-literature analysis the CASB-only baseline catches only the classes visible within its native visibility (classes whose users have not been associated with an organization, or classes that are registered). It is structurally blind to local account, embedded credential and lack-of-workload-attestation classes. This is consistent with the applied-literature claim in Section 2.1, that no single discovery technique fully addresses the scope of all classes. Third, the choice of rule for fusion matters the union rule achieves the best recall (1.0), but at high precision cost (0.378, large number of false positives per 1,000 identities checked against) whereas the majority-vote rule trades moderate capacity of recall (0.716) for high capacity of precision (0.889), the point closer to true capacity of precision more realistic to a deployment for which analyst alert-volume is an issue. These are not 'the' MSSIDs, but for each operating point you use it is a Layer 4 policy decision, and we report both instead of which is the more desirable.

5.4 Per-Class Breakdown and the Complementarity Claim

Section 4.3 argues that no single model in the ensemble is sufficient across all six taxonomy classes. Table 3 tests this directly by reporting each model's recall broken out by class.

Table 3. Per-Class Recall of MSSID Detection Models

Class	n	Recall – A	Recall – B	Recall – C	Recall – MSSID (union)
1 – Local/OS accounts	30	1.000	0.967	0.000	1.000
2 – Embedded credentials	40	1.000	0.625	0.000	1.000
3 – Unmanaged SaaS logins	35	1.000	0.686	0.000	1.000
4 – Unregistered OAuth grants	25	1.000	0.320	0.000	1.000
5 – Unattested workload/cert auth	30	0.533	0.167	1.000	1.000
6 – AI agent (both sub-types)	30	0.400	1.000	0.600	1.000

The pattern is very similar to the one predicted in the initial design rationale for Layer 3 (Section 4.3). Model A is strong and precise (recall=1.000, false positives=0) in Classes 1-4, which are identified by an absence of any governance or federation record (becoming operationally ungoverned is not enough to cause this), but in Classes 5 and 6 (where we intentionally seeded a control governance record with the same classes with operationally-ungoverned identities), recall declines to 0.533 and 0.400 respectively. However, Model C has an insignificant performance for Classes 1-4, which are obviously not workload/attestation-relevant identities, and is the best individual model for Class 5, for which it was designed to recall: 1.000. Because a coverage check, lacking behavioral information, can cover only the easy case, behaviorally-normal (Classes 1-4, 1.000 recall), Model B can be unlocked by a coverage check while it adds comparatively little precision-efficient signal to the behaviorally anomalous, nominally registered case that another model (Model A) can already cover, or that a coverage check alone can cover (the behaviorally-anomalous Classes 5-6, 1.000 recall). This is a testbed scale confirmation of the complementarity premise of the ensemble, not an external confirmation of the same: Since we built around exactly the differences each model was designed to capitalize on, one would expect a well implemented model doing well in "its" class reading this as evidence of its internal consistency with respect to the goal of the ensemble, the classes being defined and how, rather than as evidence of how the models would have separated in a less cleanly defined production data space (see Section 5.5 and Section 7).

5.5 Coverage Ratio, Mean Time to Discovery, and Attribution Accuracy

With the definitions in the original evaluation-metrics specification (which are repeated here for reference as Table 4), the testbed run achieved a coverage ratio of 0.901, or 1,432 of 1,590 nodes carried at least one of either a federated-by-IdP edge or an explicit governance record, indicating that prior to computing any anomaly models, at least 15% of seeded shadow nodes were reconciliation-first detected within a 30-day window (or just under 30 days under the assumption that shadow nodes are detected every day rather than every hour as in the literature), and that 90.5% of the shadow identities submitted for detection were correctly identified as such under a simulated attribution heuristic of 15% shadow-node noise ratio based on the imperfect provenance signal (commit history, resource tags) available in a real deployment.

Table 4. MSSID Detection Performance and Testbed Metrics

Metric	Definition	Testbed result
Precision / Recall / F1	Against seeded ground-truth shadow instances	See Table 2 (operating-point dependent)
Coverage Ratio	Fraction of graph nodes reconciled against IdP/governance record	0.901
Mean Time to Discovery	Simulated days from first activity to first flag	1.0 day (100% detected within 30-day window)
False Positive Rate /1,000	Governed identities incorrectly flagged	12.3 (majority-vote); 288.0 (union)
Attribution Accuracy	Correct owner recovered for flagged shadow identities	90.5%

5.6 Threats to Validity of the Testbed Evaluation

We are explicit about what these results do and do not show, because overstating a synthetic-testbed result is a known failure mode in security ML research generally, not specific to this paper.

- Label-generating and detection features overlap by construction. The testbed generator assigns federation, governance-record, and attestation flags directly from each seeded class's definition, and the detection models test exactly those flags. High recall for Model A on Classes 1–4 and for Model C on Class 5 is therefore partly a check that the implementation is internally consistent with its own specification, not evidence that these signals are as cleanly separable in production telemetry, where governance records, attestation state, and behavioral features are noisier, incomplete, and mutually correlated in ways this generator does not model.
- Model B's peer-group and attention mechanism are a lightweight, hand-implemented approximation of the GNN-style architecture described in Section 4.3.2, chosen for transparency and reproducibility at testbed scale rather than to match the capacity of the production heterogeneous-GNN architectures surveyed in Section 2.4. Its relative performance versus a full trained GNN on this or a larger dataset is not established here.
- All thresholds (Model B's 12th-percentile cutoff, the union/majority-vote fusion rules, the MTTD latency assumptions) were chosen once, prior to inspecting final results, but were not tuned against a held-out validation split, because the testbed is small and single-seed. Reported numbers should be read as a single operating-point demonstration, not a tuned or cross-validated estimate.
- The synthetic population, feature distributions, and single random seed (42 for generation, 7 for the attribution-noise simulation) are specific to this paper; we release the generator and detection code as supplementary material specifically so other researchers can vary the seed, class balance, and feature overlap and test whether the qualitative complementarity finding in Section 5.4 is robust to those changes, which we have not done exhaustively here.
- No result in this section should be read as a claim about detection performance on real organizational identity data. That remains the central open empirical question for future work (Section 8).

VI. Discussion

The multi-model, layered design is motivated directly by the following problem: IdP-only detection can't discern when activity never makes it to the IdP, network-only detection can observe but can't easily separate a governed activity from an ungoverned one at low cost in the absence of an independent attestation reference, and graph/GNN-based anomaly detection, though powerful, has so far been validated on activity sources that already assume that the identity is visible by the network generating the logs. This combination is intended to address the union of these blind spots, not any one of the individual blind spots, and at the scale tested the

results in Section 5.4 at least support that intent as the combined ability to check set-reconciliation, behavioral graph modeling and attestation-consistency checking.

There is a useful series of model differences in evasion resistance, and the testbed constitutes an illustrative (not definitive) example of this. The most simplest tack that an enemy could play in weathering a set-reconciliation was to register simply a set for the purpose of defeating the reconciliation, and to operate it in an undriven manner, as a put-up job (Class 5/6b) as outlined in Section 5.1, which Model A's recall did not meet, falling to 0.53 for class 5, and 0.40 for class 6b, while Model B and Model C filled the gap. While the attestation-consistency model (C) is evasion-resistant in workloads within SPIFFE/SPIRE's deployment footprint (e.g., by design in the testbed, Model C can provide no signal to classes 1–4), it is not intended to be a full-stack deployment solution, and should be deployed over time while increasing the signal coverage.

Throughout the literature on discovery, as well as this framework, detection is a required but not sufficient condition: an organization that can detect a shadow identity, but lacks a lifecycle program for onboarding, remediating, and deprovisioning it does not alleviate its risk, just it reduces the uncertainty of its risk. This is intentional, as MSSID's Layer 4 must fit with existing ITDR and PAM remediation processes for this reason, and under a deliberately imperfect, noisy attribution heuristic seven times out of ten, MSSID succeeds at assigning ownership with an accuracy of 90.5% (a number not generalizable, but indicative of the remediation potential of the simulated noise model). Lastly, the subsequent observation that the amount of shadow identities compared to identities authenticated via identity providers is likely to increase as a proportion of enterprise identities authenticated indicates that the shadow identity problem is not just a historical artefact of badly implemented identity implementations, but is actually emerging as a significant problem for enterprises, further strengthening the utility of formally-defined, empirically-testable meanings of identities as proposed and empirically evaluated here.

VII. Limitations

- No public, labeled shadow identity benchmark dataset currently exists; the testbed evaluation in Section 5 substitutes controlled, synthetic seeding for external validity, which may not fully capture the diversity, noise, and feature correlation structure of shadow identity patterns found in production environments (Section 5.6).
- The reported precision/recall/F1/coverage/MTTD/attribution figures are testbed-measured results from our own reference implementation and synthetic data generator, not results from a production deployment or a third-party dataset; they demonstrate that the architecture is internally coherent and implementable, not that it will achieve comparable performance on real organizational telemetry.
- The GNN-style behavioral model (Model B) implemented here is a lightweight, hand-coded attention-aggregation approximation rather than a trained heterogeneous GNN of the kind surveyed in Section 2.4; it requires sufficient graph density and history to establish reliable peer-group baselines, and its performance in sparse, newly instrumented environments, or relative to a full production GNN, is not established.
- The attestation-consistency model (Model C) provides value only in proportion to an organization's SPIFFE/SPIRE (or equivalent) deployment coverage and is not applicable, on its own, to fully legacy environments; in the testbed it correctly contributed no signal outside its simulated deployment footprint, which is the intended graceful-degradation behavior but also means its testbed recall figures apply only to the subset of the population within that footprint.
- Cross-organization generalizability of a trained Model B is not established here; heterogeneous log schemas and identity-graph topologies across organizations may require per-deployment retraining or transfer-learning approaches not evaluated in this paper.
- Pervasive network and browser telemetry collection, as required by Layer 1, raises privacy and employment-law considerations that vary by jurisdiction and were outside this paper's scope but must be addressed in any real deployment.
- This paper's empirical section (Section 5) evaluates a reference implementation on synthetic testbed data generated by the same authors who specified the detection rules, which introduces a risk of implicit circularity between the labeling scheme and the detection logic (Section 5.6); independent replication on an externally constructed testbed, and eventually on real deployment data, would substantially strengthen these findings.

VIII. Future Work

- Development of a standardized, sharable (e.g., synthetic or privacy-scrubbed) benchmark corpus for shadow identity detection research, ideally constructed independently of any single research group's detection implementation, analogous to established intrusion-detection benchmarks, to enable comparable evaluation across future work and to address the circularity risk noted in Section 5.6 and Section 7.
- Replacing the lightweight, hand-coded Model B implementation used here with a trained heterogeneous GNN (e.g., following the architectures surveyed in Section 2.4) evaluated on the same testbed and, eventually, on larger and noisier synthetic or real populations, to establish how much of Model B's testbed performance is attributable to modeling capacity versus the testbed's feature separability.
- Federated or privacy-preserving training approaches that allow Model B to benefit from cross-organization behavioral patterns without centralizing sensitive access-log data.
- Explainable-AI (XAI) techniques layered onto the behavioral model to produce audit- and compliance-ready justifications for flagged identities, extending directions already proposed for GNN-based IAM anomaly detection generally [24].
- Extension of the attestation-consistency model to post-quantum-secure workload credential schemes as PKI and workload-identity infrastructure migrate away from classical cryptographic assumptions [10].
- Field validation of the full MSSID pipeline in a real hybrid-cloud production environment, including comparison against the testbed-measured figures reported in Section 5, ablation results on real telemetry, and cross-industry deployment case studiest he most consequential remaining gap between this paper's contribution and a field-validated system.

IX. Conclusion

While a large and significant part of authentication happens within the plane of central identity providers, much that takes place in the enterprise is completely outside that plane, whether it is through local accounts, embedded credentials, and unmanaged SaaS

partnerships and logins, through unapproved OAuth tokens, unattested workloads, and autonomous AI agents. It has been suggested in this paper that it does not take much more effort to improve IdP-log analytics, since the problem is that a log is not produced, but a log that is not very good. To aim to circumvent the blind spots that are left by using these detection methods individually, we proposed MSSID, a four-layer framework for multi-source collection, distributed graph construction, a three-model detection ensemble (set-reconciliation, GNN-style behaviors anomaly, and attestation-consistency) and consistency analysis. In addition to the architectural proposal, we deployed a reference implementation of all three models, and evaluated it on the following controlled synthesized testbed with ground-truth shadow identities across all six taxonomy classes: 0.79 F1 (0.89 precision, 0.72 recall), 0.90 coverage ratio (zero recall is a limitation by construction), 1.0-day mean simulated time to discovery, and 90.5% attribution accuracy against IdP-log-only as a baseline. The per class breakdown reinforces the fact that each of the three models contributes significantly to the problem, with none clearly prevailing across all six classes; and confirms that no single model is dominant with respect to its internal architectural coherence, at testbed scale, rather than such performance in real-world use, and these numbers do not constitute a statement of real world evaluated performance. Given the pace of agentic AI and the evolution of non-human types of identity and access populations, we believe that it is only logical to see more of this type of formal shadow identity detection model, and see them undergo real deployment validation as it is the logical next step, complementing and not supplanting existing identity and access management practice.

REFERENCES

[1] Dark Reading, "Shadow Identities: Newest Risk Lurking in the Dark," Aug. 2025. [Online]. Available: https://www.darkreading.com/identity-access-management-security/shadow-identities-newest-risk-lurking-in-the-dark

[2] Push Security, "Get out of the dark: Manage the risk of shadow identities," Sept. 2023. [Online]. Available: https://pushsecurity.com/blog/what-are-shadow-identities

[3] Adaptive Security, "Shadow IT Discovery: Find, Assess & Govern the Risk," Jul. 2026. [Online]. Available: https://www.adaptivesecurity.com/blog/shadow-it-discovery-how-to-find-assess-and-govern-unauthorized-technology-before-it-becomes-a-breach

[4] The Hacker News (Expert Insights), "The Non-Human Identity Crisis: Why Your Machine Identities Are Your Biggest Governance Gap," May 2026. [Online]. Available: https://thehackernews.com/expert-insights/2026/05/the-non-human-identity-crisis-why-your.html

[5] NHIMG (editorial, based on content by AuthMind), "Shadow AIagent access: are your IAM controls seeing it?," Jun. 2026. [Online]. Available: https://nhimg.org/community/agentic-ai-and-nhis/shadow-ai-agent-access-are-your-iam-controls-seeing-it/

[6] BeyondTrust, "How to Detect Shadow AI and Enforce Governance for NHIs," Apr. 2026. [Online]. Available: https://www.beyondtrust.com/blog/entry/how-to-detect-shadow-ai-governance

[7] AuthMind, "Shadow AI Agent Access: How to Detect It and Shut It Down Automatically," Jun. 2026. [Online]. Available: https://www.authmind.com/blogs/shadow-ai-agent-access-how-to-detect-it-and-shut-it-down-automatically

[8] AuthMind, "Detect Shadow Access and Hygiene Gaps." [Online]. Available: https://www.authmind.com/use-cases/detect-shadow-access-and-hygiene-gaps

[9] O. Malik, "Non-Human Identity Security: Guide to Machine Governance," Veeam, Jul. 2026. [Online]. Available: https://www.veeam.com/blog/non-human-identity-security-guide.html

[10] "Machine Identity Management in Modern Enterprise Security: Concepts, Challenges, and the Role of Privileged Access Management Systems," Engineering, Technology & Applied Science Research, Feb. 2026. [Online]. Available: https://www.etasr.com/index.php/ETASR/article/view/16202

[11] "Non-Human Identity Management: Designing and Governing Machine Actors," preprint, Dec. 2025. [Online]. Available: https://www.researchgate.net/publication/398954677_Non-Human_Identity_Management_Designing_and_Governing_Machine_Actors

[12] "The Human-Machine Identity Blur: A Unified Framework," arXiv:2503.18255, 2025.

[13] Token Security, "The Ultimate Non-Human Identity Security Guide," May 2026. [Online]. Available: https://www.token.security/assets/the-ultimate-non-human-identity-security-guide

[14] Microsoft Security, "What Are Non-human Identities?" [Online]. Available: https://www.microsoft.com/en-us/security/business/security-101/what-are-non-human-identities

[15] Cloud Security Alliance Labs, "The Non-Human Identity Governance Vacuum: AI Agents and the Fastest-Growing Unmanaged Attack Surface," May 2026. [Online]. Available: https://labs.cloudsecurityalliance.org/research/csa-whitepaper-nonhuman-identity-agentic-ai-governance-v1-cs/

[16] A. Kurtz and K. Krawiecka, "Who Governs the Machine? A Machine Identity Governance Taxonomy (MIGT) for AI Systems Operating Across Enterprise and Geopolitical Boundaries," arXiv:2604.06148, Apr. 2026.

[17] "Establishing Workload Identity for Zero Trust CI/CD: From Secrets to SPIFFE-Based Authentication," arXiv:2504.14760, 2025.

[18] S. Bhanushali, "SPIFFE & SPIRE: The Workload Identity Standard Quietly Powering Zero Trust," Mar. 2026. [Online]. Available: https://sameerbhanushali.substack.com/p/spiffe-and-spire-the-workload-identity

[19] Encryption Consulting, "What Is Workload Identity? SPIFFE Explained," Jul. 2026. [Online]. Available: https://www.encryptionconsulting.com/education-center/what-is-workload-identity-spiffe-explained/

[20] Solo.io, "SPIRE: A case for attestable workload identity," Nov. 2024. [Online]. Available: https://www.solo.io/blog/spire-attestable-workload-identity

[21] "Decoupling Identity from Access: Credential Broker Patterns for Secure CI/CD," arXiv:2504.14761, 2025.

[22] "Identity Control Plane: The Unifying Layer for Zero Trust Infrastructure," arXiv:2504.17759, 2025.

[23] R. M. Sharma, "Ethical Hyper-Velocity (EHV): A Hardware-Rooted Zero-Trust Runtime Enforcement Architecture for Agentic AI Systems," arXiv:2605.17909, May 2026.

[24] V. T. Madireddy, "Graph Neural Network Based Adaptive Threat Detection for Cloud Identity and Access Management Logs," arXiv:2512.10280, Dec. 2025.

[25] V. T. Madireddy, "Graph Neural Network–Based Adaptive Threat Detection for Cloud Identity and Access Management Logs" (extended version), arXiv:2512.10280v1, Jul. 2026.

[26] "A Survey of Heterogeneous Graph Neural Networks for Cybersecurity Anomaly Detection," arXiv:2510.26307, 2025.

[27] "System and Method for Outlier and Anomaly Detection in Identity Management Artificial Intelligence Systems Using Cluster Based Analysis of Network Identity Graphs," U.S. Patent Nos. 10,681,056; 11,388,169; 11,962,597.

[28] "System and Method for Predictive Platforms in Identity Management Artificial Intelligence Systems Using Analysis of Network Identity Graphs," U.S. Patent No. 11,533,314.