Information Flow in Credential Systems

Moritz Y. Becker
Microsoft Research
Cambridge, United Kingdom
moritzb@microsoft.com

***Abstract*—This paper proposes a systematic study of information flow in credential-based declarative authorization policies. It argues that a treatment in terms of information flow is needed to adequately describe, analyze and mitigate a class of probing attacks which allow an adversary to infer any confidential fact within a policy. Two information flow properties that have been studied in the context of state transition systems, non-interference and opacity, are reformulated in the current context of policy languages. A comparison between these properties reveals that opacity is the more useful, and more general of the two; indeed, it is shown that non-interference can be stated in terms of opacity. The paper then presents an inference system for non-opacity, or detectability, in Datalog-based policies. Finally, a pragmatic method is presented, based on a mild modification of the mechanics of delegation, for preventing a particularly dangerous kind of probing attack that abuses delegation of authority.**

I. Introduction

This paper is motivated by a class of attacks on access control systems. The systems that are susceptible to such attacks are the ones that subscribe to the trust management principle [8]: authorization decisions are based on the service's local policy (consisting of declarative assertions) in union with a (potentially empty) set of supporting credentials (containing assertions) submitted by the requester. Access requests are mapped to queries, so that access is granted if the corresponding query succeeds when evaluated in the context of the policy and the submitted credentials.

It turns out that if the language supports credentials of sufficient expressiveness, then by creating and submitting certain credentials and by observing the corresponding results of a series of access requests, an adversary can potentially gain knowledge about any confidential fact in the service's local policy. Such probing attacks are particularly dangerous when mounted against policies written in languages supporting decentralized delegation of authority (e.g. Delegation Logic [27], SPKI/SDSI [12], RT [29], SD3 [25], Binder [15], Cassandra [7], SecPAL [4]). Gurevich and Neeman [21] recently described an example of such an attack on SecPAL. We briefly recall their example (in a paraphrased, but equivalent form). Imagine a service with a policy that gives parking permissions to principals that consent to the relevant terms and conditions, but additionally contains confidential facts (here hyperbolically represented by secret agent memberships). The adversary Alice submits two self-issued credentials containing the following assertions to the service, together with her request for a parking permission:

(1) Alice says Alice consents to parking rules if
Bob is a secret agent

(2) Alice says Service can say
Bob is a secret agent

Suppose the access request succeeds: in other words, the corresponding query $\langle$Service says Alice can park$\rangle$ evaluates to true in the service's policy augmented by (1) and (2), and this positive result is observed by Alice. Gurevich and Neeman [21] argue that Alice can now *infer* a secret from this observation, namely that (3) Service says that Bob is a secret agent: (3) and (2) together would imply that Alice also says that Bob is a secret agent, hence the condition of (1) would be satisfied and Alice's consent would be valid, which in turn would explain why the query succeeds. However, a closer look reveals that Alice cannot be absolutely certain that (3) holds in the service's policy. For instance, it may be the case that (3) is not true, but the fact $\langle$Alice says Bob is a secret agent$\rangle$ is true in the policy, which would result in the same observation from Alice's restricted point of view.

But the attack can be made more precise by Alice conducting a second *probe*, again with the same request (and thus the same query), but submitting only the first of the two assertions above. If this second probe yields a negative response, Alice can be certain that Bob is indeed a secret agent (according to Service), for then she knows that (2) is essential in making the query succeed. A similar attack can be mounted in order to check for the *absence* of any secret agent fact. Such attacks, where successive probing effectively provides a covert channel for confidential facts, work with a wide range of modern policy systems, including the ones mentioned above, as well as any of their derivatives.

Gurevich and Neeman propose a new policy language,

DKAL [21], [22], with a "says to" construct, to mitigate such unintended information leaks. However, as we will argue towards the end of this paper, their solution does not offer a good tradeoff between protection, expressiveness and usability. But more fundamentally, we currently lack a formal framework that could state which security properties an effective mitigation mechanism should guarantee. Indeed, there has so far been no systematic description of what is meant by "probing attack", and no analysis on what information is leaked. As the example above shows, informal reasoning easily leads to false inferences, and later examples in this paper will illustrate that the general inference problem is far from trivial.

The problem is clearly closely related to the well-studied research area of information flow. However, information flow in credential-based policy systems is distinctly different from that in state transition systems and executable programs, which have traditionally been the focus of such research. In particular, in the current setting there is no notion of state, state transition, or trace, and, most importantly, adversaries have the power to inject credentials into the policy. As such, the established techniques cannot be directly transferred to credential-based policy systems.

We wish to initiate a systematic study of information flow in the context of credential policies. The current paper is a first step in this direction. Our main contributions can be summarized as follows.

- A reformulation of the well-known information flow property, *non-interference* [20], in the context of credential policies, based on observational equivalence. We then consider another property, *opacity* [10], which so far has received somewhat less attention. We redefine opacity as the inability to infer a specific predicate about a policy and prove that non-interference can be stated in terms of opacity. Opacity turns out to be far more useful (but harder to check) for our purposes. (Section II)
- An inference system for checking the negation of opacity, which we call *detectability*, in policies based on Datalog. The system can be used to analyze probing attacks and to prove that certain properties about a policy are leaked. (Section III)
- A pragmatic method for mitigating information leaks caused by delegation-based probing. The method consists of a mild modification of the delegation mechanism in policy languages. The concepts developed in this paper allow us to precisely describe the security guarantees from this modification. (Section IV)

Connections to traditional information flow, the database inference problem, automated trust negotiation, and hypothetical logic programming are pointed out in Section V. We discuss further issues and conclude with a critical examination of the mitigation mechanism in Cassandra and DKAL/DKAL2 in Section VI.

II. Non-interference and Opacity

The goal of this section is to establish basic concepts for reasoning about information flow in declarative credential policies. We define a *policy language* $\mathcal{L}$ as a triple $(\mathcal{A}, \mathcal{Q}, \vdash)$ where $\mathcal{A}$ and $\mathcal{Q}$ are countable sets called *assertions* and *queries*, respectively, and the binary relation $\vdash \subseteq \wp(\mathcal{A}) \times \mathcal{Q}$ is the *decision relation*. Note that the decision relation need not be monotonic, although the concrete ones considered in later sections of this paper are all monotonic. A *policy* A is a subset of $\mathcal{A}$.

At this stage, our definition of policy language is kept very abstract in order to cover a wide range of existing declarative policy languages. In later sections, we will concretise this concept and consider concrete languages or specific classes of languages.

For instance, in the language of Datalog (cf. Section III), assertions are Horn clauses, queries are boolean formulas over ground atoms, and $A \vdash q$ holds iff q holds in the unique minimal model of A. To model XACML [30], assertions would correspond to Rules, Policy Target declarations and references to Combining Algorithms; a query would model a Request together with a Response (Permit, Deny, Indeterminate, NotApplicable); and the decision relation would model the Policy Determination Point (PDP).

We fix a complete lattice of *security labels* $(\Lambda, \leq)$, and a *security environment* $\Gamma = (C, I)$, where $C : \mathcal{A} \cup \mathcal{Q} \to \Lambda$ is called the *confidentiality mapping*, and $I : \mathcal{A} \to \Lambda$ is the *integrity mapping*. We generalize C and I to sets by computing the least upper bound on the labels of the elements, e.g. $C(A) = \mathbf{lub}\{C(x) \mid x \in A\}$.

All definitions in this section are parameterized by $\mathcal{L}$, Λ and Γ; they are left implicit for the sake of readability. For the remainder of this paper, let ℓ denote a label in Λ. Given a policy A, we define $\mathbf{visible}_\ell(A) = \{a \in A \mid C(a) \leq \ell\}$.

The security environment can be seen as defining a multi-level security meta-policy on the assertions within a service's local policy. The labels represent adversaries with varying clearance levels. The confidentiality mapping $C(a)$ specifies the minimal label, or the weakest adversary, capable of viewing (i.e., knowing) a specific assertion a inside a service's policy. The set $\mathbf{visible}_\ell(A)$ therefore denotes the part of A that can be passively observed (and reasoned about) by an adversary of level ℓ. In many systems, this set will be empty, or only contain the public access rules but not the extensional facts.

Additionally, an adversary may actively infer information about a policy by running a *probe*, that is,

submitting a set of credentials (which are assertions) and evaluating a query against the *union* of the policy *and* the credentials. An adversary can gain more information by running a sequence of probes, but the set of probes available to the adversary is typically restricted: $C(q)$ is the minimal label needed to evaluate (and read the result of) the query q. For instance, in SecPAL [4], this part of C would model the Authorization Query Table, which exposes a list of queries to the public and prohibits evaluation of any other query.

The integrity mapping $I(a)$ denotes the minimal label required for being able to submit an assertion a as a credential. To model a typical setting for decentralized systems where assertions are statements "said" (or issued) by a principal, Λ could be defined as the powerset of the set of principals, ordered by the superset ordering (so the set of all principals is bottom and the empty set is top). The integrity label $I(a)$ of a credential a issued and signed by a principal p would typically always include p (since p can freely create such an assertion), and additionally any other principal p' that is in possession of a (for instance, because p has issued and given credential a to p').

Definition II.1 (Alikeness, observational equivalence)**.** Two policies A and A' are *alike up to* ℓ ($A \simeq_\ell A'$) iff $\mathbf{visible}_\ell(A) = \mathbf{visible}_\ell(A')$.

Two policies A and A' are *observationally equivalent up to* ℓ ($A \equiv_\ell A'$) iff

1) $A \simeq_\ell A'$, and
2) for all assertion sets $A'' \subseteq \mathcal{A}$ and queries $q \in \mathcal{Q}$ such that $I(A'') \leq \ell$ and $C(q) \leq \ell$: $A \cup A'' \vdash q \iff A' \cup A'' \vdash q$. □

A passive adversary of level ℓ can only see the ℓ-visible assertions, and hence cannot distinguish policies that are alike up to ℓ. An active adversary can see the visible assertions *and* run probes against the policy. These two capabilities are represented by conditions 1 and 2, respectively, in the definition of observational equivalence. Hence an active adversary of level ℓ cannot distinguish policies within any ($\equiv_\ell$)-equivalence class.

Example. Consider the assertions and the query from Fig. 1. Under the given security environment, $\{a_1\} \simeq_{\text{Lo}} \{a_1, a_2\} \simeq_{\text{Lo}} \{a_1, a_2, a_3\}$. As for observational equivalence, we have $\{a_1\} \equiv_{\text{Lo}} \{a_1, a_2\}$ because q_1 has the same outcomes in both policies, no matter whether the Lo-integrity a_4 is injected or not. However, $\{a_1, a_2\} \not\equiv_{\text{Lo}} \{a_1, a_2, a_3\}$ because the latter satisfies q_1 (without injection of any assertion), whereas the former does not. □

Alikeness and observational equivalence are essential in the following definitions of two information flow properties, non-interference and opacity.

Definition II.2 (Non-interference)**.** A policy A has the *non-interference property for* ℓ iff for all policies A':

$$A \simeq_\ell A' \Rightarrow A \equiv_\ell A'. \qquad \square$$

Informally, a policy has the non-interference property if any policy that looks the same to a passive adversary also behaves the same way when probed by an active adversary. This formulation based on observational equivalence is inspired by the definition of non-interference in transition systems by Zdancewic and Myers [39]. But the definition is also equivalent to the more informal way of describing non-interference: low output (i.e., the results of evaluating low queries) only depends on low input (the immutable visible part of the policy and the submitted assertions); in particular, it is independent of high input (i.e., the confidential assertions in the policy).

Example. Continuing the example above, neither $\{a_1\}$ nor $\{a_1, a_2\}$ nor $\{a_1, a_2, a_3\}$ has the non-interference property for Lo, but adding the Lo-visible a_4 to either of these policies results in a policy that has the non-interference property. This is because whenever a policy contains both a_1 and a_4, the only Lo-query q_1 is always true (since SecPAL is monotonic). □

Non-interference is a very restrictive property that, in the case of programming languages, rules out many innocuous programs that intentionally declassify confidential information. In our context, there are many useful policies in which the result of a low query intentionally depends on the presence or absence of some confidential information. For example, a patient may read an item from his medical record if it does not contain any confidential information about a third party. The result of checking if the patient's read request is permitted thus legitimately depends on a confidential fact, even if this fact should not be directly disclosed.

Another problem with non-interference is that it is too coarse. The policy mentioned above breaks non-interference, but all this tells us is that there is some dependency between the confidential information and the query result. It does not tell us what exactly the patient learns from the result. Indeed, he might not learn anything substantial from the fact that his read request was denied, because there may be several potential reasons for access denial, some of which may be unknown to him. Conversely, we could try to restore non-interference by classifying the confidential third party fact down to the patient's security label. This measure, however, would be too drastic, as it would fully disclose the third party information to the patient even without him probing.

What is needed is a property that provides a more fine-grained handle on what an active adversary may infer

(a_1) Service says x can park if x consents to parking rules

(a_2) Service says x can park if x is a secret agent

(a_3) Service says Bob is a secret agent

(a_4) Service says Bob consents to parking rules

(q_1) Service says Bob can park

$$C(x) = \text{Lo if } x \in \{a_1, a_4, q_1\}, \text{ and Hi otherwise.}$$
$$I(x) = \text{Lo if } x = a_4, \text{ and Hi otherwise.}$$

Figure 1. Running example: the policy language is SecPAL [4], and Λ is the two-point lattice Lo $\leq$ Hi.

from a series of probes. This leads us to the information flow property of *opacity*, and its complement, which we call *detectability*. Opacity has been studied in the context of Petri nets and state transition systems [33], [10], where it is described as "the inability of an observer to establish the truth of a predicate over system traces". We reformulate this concept by replacing "system traces" by "policies".

Definition II.3 (Detectability, opacity)**.** A *policy property* Φ is a set of policies. We often identify Φ with its indicator function, i.e., $\Phi(A)$ is equivalent to $A \in \Phi$.

A policy property Φ is ℓ*-detectable* in a policy A iff for all policies A':

$$A \equiv_\ell A' \Rightarrow \Phi(A').$$

A policy property Φ is ℓ*-opaque* in a policy A iff it is not ℓ-detectable in A, or equivalently, iff there exists a policy A' such that $A \equiv_\ell A'$ and $\neg\Phi(A')$. □

Intuitively, a policy property is ℓ-detectable in A if the adversary ℓ can infer (from the visible part of A and from running his probes against A) that A must have that property. Conversely, a policy property that holds in A is ℓ-opaque in A if the adversary cannot be sure that A has that property, because there exists some policy A' in which the property does not hold and that masquerades as A with respect to all probes available to ℓ.

Note that detectability requires knowledge with absolute certainty, and not simply high probability. It may for instance be the case that "Alice is a secret agent or Bob is a nurse" is the smallest (i.e., most accurate) property that is detectable in the policy. From this it follows that the stricter property "Alice is a secret agent" is deemed opaque, simply because there is a possibility that Alice is not a secret agent but Bob is a nurse. In other words, opacity is based on a notion of uncertainty that is *possibilistic* as opposed to probabilistic.

Example. Under the given security environment from Fig. 1, the property that, according to the policy, Service says that Bob is a secret agent is Lo-opaque in $\{a_1, a_2, a_3\}$, and the property that Service does *not* say Bob is a secret agent is Lo-opaque in $\{a_1, a_2\}$. The property that Service does *not* say that Bob consents to parking rules is Lo-detectable in a_1 (by first injecting nothing, and then a_4), but Lo-opaque in $\{a_1, a_2, a_3\}$. □

The following is a list of basic sanity properties that follow directly from the definitions.

Proposition II.4. Let $\ell, \ell' \in \Lambda$ with $\ell \leq \ell'$. The following propositions hold:

1) $A \simeq_{\ell'} A' \Rightarrow A \simeq_\ell A'$.
2) $A \equiv_{\ell'} A' \Rightarrow A \equiv_\ell A'$.
3) If A has the non-interference property for ℓ', then A also has it for ℓ.
4) If Φ is ℓ'-opaque in A, then Φ is also ℓ-opaque in A.
5) If Φ is ℓ-detectable in A, then Φ is also ℓ'-detectable in A.
6) If $\neg\Phi(A)$, then Φ is ℓ-opaque in A.
7) If Φ is ℓ-detectable in A, then $\Phi(A)$ holds.
8) If $\Phi' \subseteq \Phi$ and Φ is ℓ-opaque in A, then Φ' is ℓ-opaque in A.
9) If $\Phi' \supseteq \Phi$ and Φ is ℓ-detectable in A, then Φ' is ℓ-detectable in A.

What is the relationship between non-interference and opacity? Ryan and Peacock [33] showed, in the context of transition systems, that *non-inference* [31] can be cast in terms of opacity, but conjectured that the same cannot be easily done for non-interference. Theorem II.6 below shows that non-interference can be precisely characterized in terms of opacity.

Definition II.5 (Discriminating property)**.** Let A be a policy. A policy property Φ is (A, ℓ)*-discriminating* iff there exists a policy $A' \simeq_\ell A$ such that $\Phi(A) \iff \neg\Phi(A')$.

Theorem II.6. A policy A has the non-interference property for ℓ iff all (A, ℓ)-discriminating policy properties are ℓ-opaque in A.

Proof: First, assume A has the non-interference property, and consider any (A, ℓ)-discriminating policy property Φ. If $\neg\Phi(A)$, then Φ is opaque in A anyway. In the other case, $\Phi(A)$, and by Def. II.5, there

exist $A' \simeq_\ell A$ such that $\neg\Phi(A')$. By non-interference, $A \equiv_\ell A'$. Hence Φ is ℓ-opaque in A.

For the other direction, assume the right hand side of the proposition, and consider any $A' \simeq_\ell A$. For the sake of contradiction, suppose $A \equiv_\ell A'$ does not hold. Then $\Phi = \wp(\mathcal{A}) \setminus \{A'\}$ is a (A, ℓ)-discriminating policy property. By opacity, there exists $A'' \equiv_\ell A$ such that $\neg\Phi(A'')$. By construction of Φ, $A'' = A'$, hence $A \equiv_\ell A'$, which contradicts the assumption. ■

We have now defined two information flow properties for credential policies, non-interference and opacity. There is a crude and simple way of deriving a non-interference-enforcing decision relation $\vdash_\ell$ from the original decision relation $\vdash$: when evaluating a probe for ℓ, ignore all assertions in the policy that are higher than ℓ. More formally, for all policies A: $A \vdash_\ell q$ iff $\mathbf{visible}_\ell(A) \vdash q$. As a result, all A have the non-interference property for ℓ under $\vdash_\ell$. This method is safe if the language is monotonic in the sense that fewer assertions lead to fewer permissions.

Opacity (or likewise, detectability), on the other hand, is much harder to check, but also more flexible and suitable for our purpose. In fact, it is easy to see that opacity is generally undecidable for any policy language, as opacity is defined with respect to any arbitrary policy property, which may itself be undecidable.

III. Inferring Detectability

The scenario in Section I is an example of a probing attack through which an adversary can gain knowledge of confidential parts of a service's policy. The tools developed in the last section allow us to formally define and reason about this class of attacks. To do so effectively, the concepts of policy language and policy properties have to be concretized somewhat, while keeping them still abstract enough to be applicable to many concrete existing policy languages.

In this section, we will instantiate the policy language to Datalog [36], [11]. Datalog has a small syntax and semantics and can thus be reasoned about without much overhead, yet it is expressive enough to express a wide range of policies. It is the underlying semantics (though sometimes extended with constraints) for a number of policy languages [27], [25], [29], [15], [28], [7], and a number of others such as SecPAL [7] or fragments of DKAL [21] can be translated into it.

The concepts of opacity and detectability allow us to precisely quantify the information obtained through a probing attack. The last section ended on the remark that the generality of policy properties made opacity and detectability hard to check. In this section, we will therefore restrict policy properties to those linked directly to Datalog queries about the Datalog policy. Such properties include, for instance, "Bob is a secret agent *or* Mary does not have read access", but not "the policy has an odd number of assertions".

The main technical result of this section is an inference system for detectability of policy properties in Datalog policies. As was hinted at in Section I, and which will become more evident in the following, analysing detectability is highly non-trivial, and it is easy to make mistakes. Given a set of probes, our inference system allows us to deduce what information about the policy is leaked. To be more precise, Theorem III.8 is a soundness result in that it allows us to prove that an adversary can detect some policy property.

For safety analysis, it would be useful also to have the completeness result in order to prove opacity, i.e., that no adversary can detect some secret property. Whether the inference system for Datalog is complete or not remains an open problem (see discussion in Section VI), but Section IV presents a more restrictive language that enforces certain opacity guarantees.

A. Datalog

We fix a function-less first order signature with predicate names and constants. An *expression* e is a variable or a constant. An *atom* (or *fact*) $f = p(\vec{e})$ is a predicate name p applied to an expression tuple $\vec{e}$ of the correct arity. A *clause* is of the form

$$f_0 \leftarrow f_1, ..., f_n,$$

where $n \geq 0$, and f_0 is called the *head* and $\vec{f_i}$ the *body*. The arrow is usually omitted if $n = 0$. If a is a clause, then we write $\mathbf{hd}(a)$ to denote its head. A *program* is a set of clauses. A *query* q is either **true**, **false** or a ground (i.e., variable-free) boolean formula (i.e., involving connectives $\neg$, $\wedge$ and $\vee$) over atoms.

For the remainder of this section, we instantiate the set of assertions $\mathcal{A}$ to the set of Datalog clauses, and the set of queries $\mathcal{Q}$ to the set of Datalog queries. A policy is therefore a Datalog program. There are several equivalent definitions of the decision relation $\vdash$. The model-theoretic formulation is quite intuitive: given a policy $A \subseteq \mathcal{A}$, close each clause by universal quantifers and form the conjunction of all such quantified clauses. This formula has a unique minimal Herbrand model M. Then we define $A \vdash q$ iff $M \models q$.

For our proofs, a more operational definition is useful that is based on the intuition that Datalog programs are inductive definitions. Given a policy A, we define the *consequence operator* $\mathbf{T}_A$ as a monotonic mapping between sets S of ground atoms. In the following definition[1], let γ be a ground substitution (a total mapping

[1]Throughout the paper, we will occasionally use pointed brackets $\langle \_ \rangle$ as meta-level parentheses to delimit phrases of concrete syntax.

from variables to constants).

$$\mathbf{T}_A(S) = \{\ f_0' \ \mid \exists\gamma\ \exists\langle f_0 \leftarrow f_1, ..., f_n\rangle \in A, \\ \gamma(\{f_1, ..., f_n\}) \subseteq S, \\ f_0' = \gamma(f_0)\ \}$$

The least fixed point of the consequence operator, $\mathbf{T}_A^\omega(\emptyset)$, is equal to the model M above, so we can equivalently define $A \vdash q$ iff $\mathbf{T}_A^\omega(\emptyset) \models q$.

Datalog lends itself well to specifying policies because many policy rules can be naturally expressed as if-statements. For example, the SecPAL delegation assertion

A says B can say x has age n if x is a User

could be translated[2] into two Datalog assertions:

$$\mathsf{canSayHasAge}(\mathrm{A}, \mathrm{B}, x, n) \leftarrow \mathsf{isA}(\mathrm{A}, x, \mathrm{User})$$
$$\mathsf{hasAge}(\mathrm{A}, x, n) \leftarrow \mathsf{hasAge}(\mathrm{B}, x, n), \\ \mathsf{canSayHasAge}(\mathrm{A}, \mathrm{B}, x, n)$$

B. Probing environments

In Section II, the adversary was characterized by a security label ℓ and a security environment $\Gamma = (C, I)$. The confidentiality mapping C defined which assertions were visible to her and which queries she could run, and the integrity mapping I defined which assertions she could submit and inject into the policy. But even though the adversary may have a large or infinite number of probes at her disposal, she will typically only use a subset of them in a particular probing attack. In this section, we will specify the adversary in a more fine-grained fashion as a specific set of probes, corresponding to a concrete probing attack.

Definition III.1 (Probe, probing environment)**.** A *probe* $\pi = (A, q)$ is a pair consisting of an assertion set $A \subseteq \mathcal{A}$ and a query $q \in \mathcal{Q}$.

A *probing environment* $\Pi = (C, \ell, A_0, P)$ is a 4-tuple consisting of a confidentiality mapping $C : \mathcal{A} \to \Lambda$, a security label $\ell \in \Lambda$, a policy $A_0 \subseteq \mathcal{A}$, and a set P of probes. □

A probing environment (C, ℓ, A_0, P) is a complete specification of a probing attack. A_0 is the service's policy that is under attack; as before, C determines which assertions in A_0 are visible to the adversary, who is assumed to have label ℓ; and P is the set of probes that are run against A_0. The probe set P could be seen as a fine-grained instantiation of both the part of C that defined the queries available to the adversary and the integrity mapping I from Section II.

We say that a probe (A, q) is *positive* if $A_0 \cup A \vdash q$, and *negative* otherwise. For each probe π in P, the adversary can observe if π is positive or negative. Based on the notion of probing environment, we generalize the definition of detectability.

Definition III.2 (Observational equivalence)**.** Two policies A, A' are *observationally equivalent* under a probing environment Π (we write $A \equiv_\Pi A'$) iff

1) $A \simeq_\ell A'$, and
2) for all probes (A'', q) in P:
   $A \cup A'' \vdash q \iff A' \cup A'' \vdash q$. □

Definition III.3 (Detectability)**.** A query q is Π-*detectable* iff for all policies A_0':

$$A_0' \equiv_\Pi A_0 \Rightarrow A_0' \vdash q.$$

A query q is Π-*opaque* iff it is not Π-detectable. □

This refined definition of detectability allows us to state succinctly what information about a policy is leaked to the adversary in a given probing attack.

C. An informal example

We now develop an inference system that proves if a query is detectable in a probing environment. To give an intuition for the problem, we first consider an example that is small and yet illustrates the surprising complexity of this problem. Suppose that the adversary cannot read anything from the policy A_0, so $\mathbf{visible}_\ell(A_0) = \emptyset$. Suppose further that the probe set consists of three probes that all have the same query, (A_1, ok), (A_2, ok), and (A_3, ok), where

$$A_1 = \{\langle \mathsf{a}\rangle\},\ \ A_2 = \{\langle \mathsf{b}\rangle\},\ \text{and} \\ A_3 = \{\langle \mathsf{a} \leftarrow \mathsf{c}\rangle, \langle \mathsf{b} \leftarrow \mathsf{a}\rangle\}.$$

(The nullary predicates a, b, c and ok in this example could represent any security-relevant facts in the policy.)

Now suppose that the first two probes are negative and the third is positive, i.e.,

$$A_0 \cup A_1 \nvdash \mathsf{ok},\ \text{and}\ A_0 \cup A_2 \nvdash \mathsf{ok}, \tag{1}$$

and

$$A_0 \cup A_3 \vdash \mathsf{ok}.$$

From these three probes alone, what can the adversary infer with certainty about the policy A_0?

First of all, what does the fact $A_0 \cup A_3 \vdash \mathsf{ok}$ convey? The probe would be trivially positive if ok were already true in A_0. But from (1) we can infer that actually $\neg\mathsf{ok}$ holds in $A_0 \cup A_1$, and thus also in A_0, by anti-monotonicity. So it must be the case that at least one of

[2]The actual translation according to [4] is actually a bit more complicated due to the fact that SecPAL supports a second kind of delegation construct, can say_0, that forbids redelegation.

the two assertions in A_3 play an essential role in causing the third probe to be positive. Either the first assertion "fired", in which case a could not have already been true in A_0, and furthermore, c must hold. Or else the second assertion fired, in which case b is false in A_0, and either a or c are true, for these are the only two possibilities for firing the second assertion (we say: a and c form the *support of* b *in* A_3). In summary, $A_0 \cup A_3 \vdash \mathsf{ok}$ (together with (1)) implies

$$A_0 \vdash (\neg\mathsf{a} \wedge \mathsf{c}) \vee (\neg\mathsf{b} \wedge (\mathsf{a} \vee \mathsf{c})). \tag{2}$$

But there is more to be inferred. We can widen the two assertions in A_3 to $A_3' = \{\langle\mathsf{a}\rangle, \langle\mathsf{b}\rangle\}$ (we say: A_3 *is contained in* A_3') and, by monotonicity, be certain that

$$A_0 \cup A_3' \vdash \mathsf{ok}. \tag{3}$$

Note that A_1 and A_2 are proper subsets of A_3'. This fact can be exploited: from (3), we can infer (again by arguing that either ok already holds or that the other assertion has fired):

$$\begin{array}{l} A_0 \cup \{\langle\mathsf{a}\rangle\} \vdash \mathsf{ok} \vee \neg\mathsf{b}, \text{ and symmetrically,} \\ A_0 \cup \{\langle\mathsf{b}\rangle\} \vdash \mathsf{ok} \vee \neg\mathsf{a}. \end{array} \tag{4}$$

Combining these with the two results from (1), we get $A_0 \cup \{\langle\mathsf{a}\rangle\} \vdash \neg\mathsf{b}$, and $A_0 \cup \{\langle\mathsf{b}\rangle\} \vdash \neg\mathsf{a}$. Then, by anti-monotonicity of the last two queries, we can infer $A_0 \vdash \neg\mathsf{a} \wedge \neg\mathsf{b}$. Finally, this can be combined with (2) and our knowledge of $\neg\mathsf{ok}$ to yield (after some algebraic simplification)

$$A_0 \vdash \neg\mathsf{ok} \wedge \neg\mathsf{a} \wedge \neg\mathsf{b} \wedge \mathsf{c}.$$

The adversary can thus infer that ok, a and b are false, and c is true in A_0.

D. The inference system

We now formalize the intuition given in the example above. In particular, we first need to make the notions of *monotonicity*, *containment*, *support*, and *firing* precise.

Definition III.4 (Monotonicity, containment)**.** A query is *monotonic* iff it is equivalent to one without negation. A policy A is *contained in* a policy A' (we write: $A \preceq A'$) iff for all sets S of ground atoms, $\mathbf{T}^{\omega}_{A \cup S}(\emptyset) \subseteq \mathbf{T}^{\omega}_{A' \cup S}(\emptyset)$. □

A useful proposition that follows from this definition is that if q is monotonic and $A \preceq A'$ then for all policies A_0, $A \cup A_0 \vdash q$ implies $A' \cup A_0 \vdash q$. Conversely, if $A' \cup A_0 \vdash \neg q$ holds, then so does $A \cup A_0 \vdash \neg q$.

Containment is an undecidable problem in unrestricted Datalog. However, we will only check containment on ground policies, for which the problem is decidable (this follows from decidability results on containment in monadic Datalog [13]). Moreover, any conservative approximation of containment can be used without affecting soundness of the inference system (such as the syntactic widening hinted at in the example, or simply the subset relation).

Definition III.5 (Support)**.** Let **support** be a function from any assertion set A and any ground atom f to a set of sets Δ of ground atoms such that the following holds:

1) If $\Delta \in \mathbf{support}(A, f)$ then $A \cup \Delta \vdash f$.
2) If $A \cup \Delta \vdash f$, then there exists $\Delta' \subseteq \Delta$ such that $\Delta' \in \mathbf{support}(A, f)$.
3) If $\Delta, \Delta' \in \mathbf{support}(A, f)$ and $\Delta' \subseteq \Delta$ then $\Delta = \Delta'$. □

The first two requirements represent soundness and completeness of **support**, respectively, and the third is minimality. This function can be computed directly using standard abduction [26]. Abduction has been used in AI applications such as planning and fault diagnosis, and has also been applied to security policy analysis for explaining access denial, for policy debugging [5] and for distributed credential gathering [2]. For unrestricted Datalog, $\mathbf{support}(A, f)$ may be infinite, which would result in infinite queries in our inference system. However, if A is finite and ground, then for all ground atoms f, $\mathbf{support}(A, f)$ is a finite set and all members of the set are finite sets of ground atoms. For this reason, we will restrict submitted assertions in probes to be ground, to prevent the inference system generating judgements of infinite size.

Definition III.6 (Firing)**.** Let A be a ground assertion set, f a ground atom, and $S = \mathbf{support}(A, f)$. Then $\mathbf{fired}(A, f)$ denotes the query

$$\neg f \wedge \bigvee_{\Delta \in S} \bigwedge \Delta. \qquad \square$$

Suppose A_1 is a partially unknown policy and A_2 a known ground policy. If we know that $A_1 \cup A_2 \vdash q$, what does this tell us about A_1? Either q already holds in A_1, or at least one of the assertions in A_2 is essential in proving q: it is "fired" in the context of A_1. More precisely, if f is the head of the fired assertion, then $\mathbf{fired}(A_2, f)$ holds in A_1. Lemma III.7 formalizes this argument, and is the cornerstone of the correctness proof for the inference system.

Lemma III.7. Let q be a query, and A_1, A_2 be policies where A_2 is ground. If $A_1 \cup A_2 \vdash q$ then $A_1 \vdash q \vee \bigvee_{a \in A_2} \mathbf{fired}(A_2, \mathbf{hd}(a))$.

Proof: Assume the left hand side of the proposition, and $A_1 \nvdash q$. It remains to show that the big disjunction is derivable in A_1. Let n be the smallest integer such that $\mathbf{T}^{n}_{A_1 \cup A_2}(\emptyset) \neq \mathbf{T}^{n}_{A_1}(\emptyset)$. Then by definition of the

$$\text{(PEEK)}\ \frac{\mathbf{visible}_\ell(A_0) \cup A \vdash q \quad q \text{ is monotonic} \quad A \text{ is ground}}{(C, \ell, A_0, P), A \Vdash q} \qquad \text{(WEAK)}\ \frac{\Pi, A \Vdash q \quad \models q \Rightarrow q'}{\Pi, A \Vdash q'}$$

$$\text{(POKE1)}\ \frac{A_0 \cup A \vdash q \quad (A, q) \in P}{(C, \ell, A_0, P), A \Vdash q} \qquad \text{(POKE2)}\ \frac{A_0 \cup A \nvdash q \quad (A, q) \in P}{(C, \ell, A_0, P), A \Vdash \neg q}$$

$$\text{(MONO1)}\ \frac{\Pi, A \Vdash q \quad A' \succeq A \quad q \text{ is monotonic} \quad A' \text{ is ground}}{\Pi, A' \Vdash q} \qquad \text{(MONO2)}\ \frac{\Pi, A \Vdash q \quad A' \preceq A \quad \neg q \text{ is monotonic} \quad A' \text{ is ground}}{\Pi, A' \Vdash q}$$

$$\text{(CONJ)}\ \frac{\Pi, A \Vdash q_1 \quad \Pi, A \Vdash q_2}{\Pi, A \Vdash q_1 \wedge q_2} \qquad \text{(DIFF)}\ \frac{\Pi, A_1 \cup A_2 \Vdash q}{\Pi, A_1 \Vdash q \vee \bigvee_{a \in A_2} \mathbf{fired}(A_2, \mathbf{hd}(a))}$$

Figure 2. Inference system for detectability.

consequence operator, there must exist a ground assertion $\langle f \leftarrow f_1, ..., f_m \rangle$ in A_2 such that $A_1 \vdash \neg f$ and all f_i are in $\mathbf{T}^{n-1}_{A_1 \cup A_2}(\emptyset) = \mathbf{T}^{n-1}_{A_1}(\emptyset)$ (the equality holds by construction of n). Moreover, by definition of **support**, there exists $\Delta \in \mathbf{support}(A_2, f)$ such that all f_i are in Δ. Hence the big disjunction holds. ■

We now have all the tools needed to assemble the inference system for checking detectability (Fig. 2). The inference system generates judgements of the form $\Pi, A \Vdash q$ where $\Pi = (C, \ell, A_0, P)$ is a probing environment, A a set of ground assertions, and q a query. This judgement states that the adversary can learn that q holds in $A_0 \cup A$, just from looking at the visible part of the policy and from running the probes P. More precisely, the essential property of the inference system is as follows:

$$\begin{aligned} &\text{If } \Pi, A \Vdash q \text{ then for all policies } A'_0\text{:} \\ &\quad A'_0 \equiv_\Pi A_0 \ \Rightarrow\ A'_0 \cup A \vdash q. \end{aligned} \tag{5}$$

The axiom (PEEK) models the knowledge gained from just passively reading the visible part of the policy. (WEAK) allows the conclusion to be weakened; this rule is useful when the premise of another rule requires the query to be monotonic. The axioms (POKE1) and (POKE2) model the positive or negative result from a single probe from P. (MONO1) and (MONO2) exploit the properties of monotonicity and containment. (CONJ) allow conclusions to be conjoined if they hold in the same credential context A. The last rule, (DIFF), is the one that does most of the work: it encapsulates the implication of Lemma III.7.

Finally, Theorem III.8 formalizes the correctness of the inference system.

Theorem III.8. Let q be a query and $\Pi = (C, \ell, A_0, P)$ be a probing environment where P is ground. If $\Pi, \emptyset \Vdash q$ then q is Π-detectable.

Proof: (Sketch.) We prove the more general statement (5) above. The proof proceeds by rule induction on $\Vdash$. The interesting case is (DIFF). Suppose $\Pi, A_1 \cup A_2 \Vdash q$. Consider any $A'_0 \equiv_\Pi A_0$. By the induction hypothesis, $A'_0 \cup A_1 \cup A_2 \vdash q$. Hence Lemma III.7 can be applied to yield $A'_0 \cup A_1 \vdash q \vee \bigvee_{a \in A_2} \mathbf{fired}(A_2, \mathbf{hd}(a))$. ■

Example. Returning to the example from Section III-C, we sketch how the proof of detectability proceeds according to the inference system. First of all, the basic facts $A_0 \cup A_1 \nvdash \mathsf{ok}$, $A_0 \cup A_2 \nvdash \mathsf{ok}$, and $A_0 \cup A_3 \vdash \mathsf{ok}$ are obtained by applying (POKE2) to (A_1, ok) and (A_2, ok), and (POKE1) to (A_3, ok).

Then (MONO2) is applied to the basic fact from (A_1, ok) (with $\emptyset \preceq A_1$) to get the subresult $\Pi, \emptyset \Vdash \neg\mathsf{ok}$. The result from applying (DIFF) to the basic fact from (A_3, ok) (with $\emptyset$ in the conclusion) is conjoined with the latter result by (CONJ), yielding the subresult (2). Now we apply (MONO1) to the basic fact of (A_3, ok) (with $A'_3 \succeq A_3$) and obtain (3). Applying (DIFF) in two different ways (with A_1 or with A_2 in the conclusion) yields (4). The $\neg\mathsf{ok}$ in the conclusions of these two judgements is eliminated by (CONJ) by conjoining with the basic facts from (A_1, ok) and (A_2, ok). Applying (MONO2) on the two results and then combining them with (CONJ) yields the subresult $\Pi, \emptyset \Vdash \neg\mathsf{a} \wedge \neg\mathsf{b}$. Finally, repeated application of (CONJ) on all the subresults yields $\Pi, \emptyset \Vdash \neg\mathsf{ok} \wedge \neg\mathsf{a} \wedge \neg\mathsf{b} \wedge \mathsf{c}$. By Theorem III.8, the formula in the conclusion is Π-detectable.

IV. Information flow control

We now turn to the question as to how information flow can be controlled in a "real" policy language, namely SecPAL. SecPAL lends itself well to this purpose, as it has been designed with the goal of maximising generality and expressiveness while keeping the number of primitive constructs minimal. But more importantly, it is paradigmatic in its susceptibility to probing attacks

based on delegation. This vulnerability is shared by almost the entire family of related languages that support the decentralized delegation of authority via the says-operator, first introduced in the ABLP logic et al. [1], or any equivalent construct.

We briefly recall SecPAL, before taking a closer look at what causes the vulnerability. We then show how this vulnerability can be mitigated by a mild modification of the way delegation works in SecPAL. This method can be easily transferred to other policy languages that support delegation.

A. SecPAL

We give a very brief overview of SecPAL; for a more careful treatment, see [3] and [4].

Syntax. We fix an arbitrary first-order function-less signature with countably infinite sets of predicate names and constants (including principals). An *expression* e is either a variable or a constant. A *fact* f is either *flat*, i.e., it is a predicate atom (which we often write in infix notation, e.g. e can read e'), or *nested*, i.e., it is of the form e can say f' or e can say$_0$ f', where f' may be nested itself.

We further fix an arbitrary constraint language. A constraint c is a first-order formula over atomic constraints (such as equality, inequalities, arithmetic constraints or regular expression constraints). The only requirement on the constraint language is that it be equipped with a computable unary relation $\models$, such that $\models c$ holds whenever the ground constraint c is true. SecPAL can thus be flexibly adapted to various specific domains by choosing a domain-specific predicate name set and constraint language.

An *assertion* a is of the form

$$e \text{ says } f \text{ if } f_1, ..., f_n \text{ where } c,$$

where $n \geq 0$, e is ground and the f_i are flat. We say that e is the *issuer* of this assertion. We omit the "if" if $n = 0$ and the "where" if $c =$ **true**. An assertion of the form $\langle e \text{ says } f \rangle$ is an *atomic* assertion.

A *query* is a boolean formula over ground atomic assertions involving only flat facts. (SecPAL actually allows non-ground, constrained and quantified queries, but we omit these for simplicity.)

For the remainder of this section, we instantiate $\mathcal{A}$ to the set of SecPAL assertions and $\mathcal{Q}$ to the set of SecPAL queries.

Decision relation. Fig. 3 shows the decision relation $\vdash$ for atomic queries (i.e., ground atomic assertions). In the modus ponens rule (Cond), γ is a ground substitution. Note that (Cond) requires the conditional facts f_i to be issued by the same principal e that issues the original assertion and the conclusion fact. The only way that foreign facts may enter the derivation is via (CanSay) and (CanSay0). The former implements standard delegation of authority: e_1 delegates authority over f to e_2, so e_1 is willing to vouch for f whenever e_2 says it. The latter is similar, but prohibits redelegation: the delegator e_2 must say f directly, without any dependencies on what other principals say.

If A is a policy, then let $[\![A]\!]$ denote the set of atomic queries a such that $A \vdash a$. Then for a general query q, we define $A \vdash q$ iff $[\![A]\!] \models q$. We refer the interested reader to [4] for a collection of example policies and policy idioms expressed in SecPAL.

B. Probing attacks revisited

From Section III, it is clear that in all but the most trivial circumstances, any confidential fact in the policy can be detected with the right combination of probes, even if the query does not mention the fact at all. Consider for instance the following two SecPAL probes:

$$\pi_1 = (\{\text{A says ok if secret}\}, \text{A says ok})$$
$$\pi_2 = (\emptyset, \text{A says ok})$$

If π_1 is positive and π_2 is negative, then $\langle$A says secret$\rangle$ can be detected by any B running these probes. But since the essential credential in π_1 is issued by A herself, one could argue that the only one who could run this probe would be A herself, or perhaps A intended this information flow since she allowed this obviously dangerous credential to get into somebody else's possession (or else A is grossly careless).

But the real danger of probing lies in the fact that any confidential fact issued by A can be detected using probes that do not contain any assertions issued by A. Indeed, even the queries in the probes need not mention A at all, as the following example shows:

$$\pi'_1 = (\{\text{B says ok if secret},\ \text{B says A can say secret}\}, \text{C says foo})$$
$$\pi'_2 = (\{\text{B says ok if secret}\}, \text{C says foo})$$

If these assertions are translated into Datalog [4], the inference system from Section III can be used to infer that $\langle$A says secret$\rangle$ is detectable if π'_1 is positive and π'_2 is negative.

Clearly, the underlying mechanism for this attack (and similar attacks) works by exploiting the ability to delegate authority over facts to others. In other words, these attacks depend on the rules (CanSay) and (CanSay0). Taking a closer look at the (CanSay) rule, note its original intention is for e_1 to specify who (e_2) can be trusted on saying f. The principal e_1 is willing to import f from e_2 if e_1 has issued the corresponding can say assertion. But the rule also reveals an apparent asymmetry between

$$\text{(Cond)}\ \frac{\langle e \text{ says } f \text{ if } f_1, ..., f_n \text{ where } c\rangle \in A \quad \models \gamma(c) \quad A \vdash \gamma(e \text{ says } f_i) \text{ for all } i \in \{1...n\}}{A \vdash \gamma(e \text{ says } f)}$$

$$\text{(CanSay)}\ \frac{A \vdash e_1 \text{ says } e_2 \text{ can say } f \quad A \vdash e_2 \text{ says } f}{A \vdash e_1 \text{ says } f} \qquad \text{(CanSay0)}\ \frac{A \vdash e_1 \text{ says } e_2 \text{ can say}_0 f \quad \{a \in A \mid e_2 \text{ is the issuer of } a\} \vdash e_2 \text{ says } f}{A \vdash e_1 \text{ says } f}$$

Figure 3. SecPAL proof system.

e_1 and e_2: e_1 has to explicitly agree to importing e_2's fact f, but there is no premise that requires e_2 to agree to exporting the fact to e_1. Hence e_1 can import from e_2 without e_2's consent, under the guise of delegating authority over f to e_2.

C. SecPAL$^+$

Based on the intuition developed above, we now propose a mild modification to the two delegation rules which provides a more symmetric protection of both the delegator – the one who imports a fact – and the delegatee – the one who exports the fact. We first introduce a new type of nested fact of the form $\langle e \text{ can listen to } f\rangle$, where f is a (possibly nested) fact. The rules (CanSay) and (CanSay0) are then replaced by (CanSay$^+$) and (CanSay0$^+$) (below, A_{e_2} denotes $\{a \in A \mid e_2 \text{ is the issuer of } a\}$):

$$\text{(CanSay}^+\text{)}\ \frac{A \vdash e_1 \text{ says } e_2 \text{ can say } f \quad A \vdash e_2 \text{ says } e_1 \text{ can listen to } f \quad A \vdash e_2 \text{ says } f}{A \vdash e_1 \text{ says } f}$$

$$\text{(CanSay0}^+\text{)}\ \frac{A \vdash e_1 \text{ says } e_2 \text{ can say}_0 f \quad A_{e_2} \vdash e_2 \text{ says } e_1 \text{ can listen to } f \quad A_{e_2} \vdash e_2 \text{ says } f}{A \vdash e_1 \text{ says } f}$$

We call the language with the modified delegation rules SecPAL$^+$. Consider the following (re-)delegation chain:

A says B can say f
B says C can say f
C says D can say f
D says f

In standard SecPAL, these assertions would suffice to derive A says f. In SecPAL$^+$, we need additional assertions to make the delegation succeed:

B says A can listen to f
C says B can listen to f
D says C can listen to f

This conforms nicely with the idea of a delegation chain, in which each principal only knows the peer above and below the chain. The final delegatee D, for instance, need not specify that B and A can listen to f.

Opacity in SecPAL$^+$. The goal of modifying SecPAL's delegation mechanism was to rule out probing attacks based on delegation. But what exactly do our modifications achieve? We can answer this question in terms of opacity and detectability.

In the following, let Λ be the two-point lattice Lo $\leq$ Hi. For simplicity, we let C be a confidentiality mapping such that $C(a) =$ Hi, for all assertions a. Hence for any policy A_0, $\mathbf{visible}_{\text{Lo}}(A_0) = \emptyset$; we thus assume that the entire policy is invisible to a low passive adversary.

Theorem IV.1 below provides a simple opacity guarantee for a confidential fact q_0 that is issued by some e_1. Assuming that (1) the adversary is not permitted to directly query q_0, (2) the adversary is not in possession of any obviously compromising assertion issued by e_1 (cf. Section IV-B), and (3) the adversary is not in possession of a $\langle$can listen to$\rangle$-assertion for q_0 issued by e_1, the theorem guarantees opacity of q_0.

Theorem IV.1. Let q_0 be a possibly negated atomic query of the form $(\neg)(e_0 \text{ says } p(\vec{e}))$, and $\Pi = (C, \text{Lo}, A_0, P)$ a (SecPAL$^+$) probing environment. If for all $(A, q) \in P$,

1) q_0 is not a subquery of q, and
2) p does not occur in any e_0-issued assertion $a \in A$, and
3) for all $a \in A$:
   $A_0 \cup A \nvdash e_0 \text{ says } e_a \text{ can listen to } f$,
   where e_a is the issuer of a,

then q_0 is Π-opaque.

Proof: (Sketch.) The main idea of the proof is to construct some A_0' such that (a) $A_0' \nvdash q_0$ and (b) $A_0' \equiv_\Pi A_0$. In the case where q_0 is positive, A_0' is obtained from A_0 by renaming all occurrences of p to some p' that does not occur in Π. Clearly, (a) holds. Next, suppose for sake of contradiction that (b) does not hold. Then we can show that q_0 is crucial to one of the probes: there would exist some $(A, q) \in P$

such that $A'_0 \cup A \vdash q$ iff $A'_0 \cup A \cup \{q_0\} \nvdash q$. With assumption (1), q_0 would have to appear as a premise in the inference of either q or $\neg q$ on the right hand side of this equivalence. A simple rule induction over $\vdash$ shows that, with assumption (2), this particular inference step cannot involve (COND). The only possible inference step would thus be an instance of (CANSAY$^+$) or (CANSAY0$^+$) that would contradict assumption (3). Therefore, (b) holds.

In the other case where q_0 is negated, A'_0 is constructed from A_0 by adding e_0 says $p(\vec{e})$ (and thus making (a) true) and by applying the additional following transformations (for making (b) true). For each e_0-issued assertion in which $p(\vec{e}')$ occurs in the body, replace the original constraint c by $c \wedge \vec{e} \neq \vec{e}'$. Furthermore, for each assertion in which $\langle e' \text{ can say } p(\vec{e}'')\rangle$ occurs in the head, replace the original constraint c by $c \wedge e' \neq e_0 \wedge \vec{e}'' \neq \vec{e}$. The rest of the argument is similar to the positive case. ■

V. RELATED WORK

To our knowledge, this is the first systematic study of information flow in credential-based policy systems, in particular with respect to probing attacks. This section gives a brief overview of related areas of research.

Automated trust negotiation (ATN), first introduced by Winsborough et al. [38], is concerned with negotiation strategies for exchanging confidential credentials between mutual strangers. The notion of negotiation safety is formalized in Winsborough and Li [37], which also provides a comprehensive overview of research efforts in this area. Informally, a safe negotiation strategy does not reveal anything about the presence or absence of a credential (or an attribute) to the negotiation partner before the latter has proved that he satisfies the necessary disclosure conditions, specified by some policy. There are two main differences between the ATN setting and the current one which make the two hard to directly compare. Firstly, ATN considers the confidentiality of credentials of all involved parties, which necessitates a sequential and gradual negotiation process. In contrast, we are only concerned with the confidentiality of the service's policy. Secondly, work in ATN has so far not considered the possibility of agents *injecting* objects into the policy. Our adversaries, in contrast, can submit credentials that, in effect, logically extend the service's policy for the duration of one probe. This ability is natural in the context of trust management [8], but it also complicates the analysis.

Information flow has been studied since the mid 1970s [14], though research has mainly been concerned with stateful, temporal computations. A good survey is found in [34]. The current setting is quite different as there is no notion of state, state transition, run or trace, and adversaries are traditionally not permitted to inject code into the program. Nevertheless, on an abstract level, many of the traditional concepts can be adapted, such as non-interference (introduced by Goguen and Meseguer [20]). In fact, our definition is inspired by Zdancewic and Myers' definition [39], which is also based on observational equivalence (albeit on traces). Opacity [33], [10] is another, less known, information flow property that we reformulated and that proved very useful for our purposes; it is also closely related to non-inference [31] and non-deducibility [35].

The knowledge operator K_i in epistemic modal logic [17] bears some resemblance to our definition of detectability, and has been used for reasoning about secrecy in (stateful and temporally evolving) multi-agent systems [23]. We leave it to future work to investigate if reformulating our framework in such a logic may be fruitful.

The database inference problem is concerned with indirect inference channels through which confidential information from a database can leak to a database user. A wide variety of such channels have been identified and studied [24], [18]. In particular, Bonatti et al. have studied the inference problem in deductive databases [9], which are similar to declarative policies. However, they do not consider users who can temporarily inject new rules and relations into the database, as this is not natural in the database context.

The notions of probing attacks and detectability in Datalog policies (Section III) are related to hypothetical logic programming [16], where goals (corresponding to our queries) may be supplemented by hypothetical clauses which are temporarily added to the logic program (corresponding to the submitted credentials). However, we are not just interested in evaluating such hypothetical goals (i.e., probes), but in finding an *explanation* for the result of these goals. Abduction, in its most general form, is about finding explanations to a set of observations [32], and has been extensively studied in the context of standard logic programming [26], but so far not of hypothetical reasoning. Indeed, the inference system in Section III could be viewed as an abduction method for hypothetical goals in Horn logic programs.

VI. DISCUSSION

In Section II, we developed an information flow framework for credential-based policy systems, on the basis of alikeness (with respect to the visible part of a policy) and observational equivalence (with respect to probes). This abstract setting gave rise to very general and elegant formulations of the relevant information flow properties: a policy has the non-interference property if

any policy that looks alike is observationally equivalent; and a policy property is detectable in a policy if any observationally equivalent policy has that property. We leave it to future work to examine if some of the many other information flow properties [19] can also be usefully adopted.

Completeness. This framework was then put to good use in Section III where we formalized probing attacks and developed an inference system for checking detectability in Datalog policies. Theorem III.8 proves the *soundness* of the system: whenever a query is derivable in the system, then it is also detectable. However, the corresponding *completeness* statement remains a conjecture:

Conjecture VI.1. Let q be a query and $\Pi = (C, \ell, A_0, P)$ be a probing environment where A_0 is finite and P is ground and finite. If q is Π-detectable then $\Pi, \emptyset \Vdash q$ holds.

Completeness would be a useful property because it would imply at least semidecidability of detectability (for ground Datalog probes), and thus detectability could be positively checked in finite time. Furthermore, if $\Pi, \emptyset \nVdash q$ holds, then q would also be known to be opaque in the policy. Of course, to check $\Pi, \emptyset \nVdash q$ in general would require full decidability, which is not automatically implied by completeness.

The reason why completeness is not easy to prove is that the premise of detectability, being a universally quantified implication, does not provide any obvious information that would guide the inference proof. In any case, the inference proofs do not seem to be very goal-directed, as the small example in Section III-C illustrated. In particular the combination of (WEAK), (MONO1) and (DIFF) contribute to the apparent unconstructiveness of the inference system.

A more promising strategy may be to prove the *modus tollens* direction: assume $\Pi, \emptyset \nVdash q$, and then construct a policy that masquerades as A_0 with respect to the probes, but in which q does not hold. While it is easy to modify A_0 in such a way that $\neg q$ holds, the difficulty here lies in the fact that it is not clear how to "repair" the modified policy in order to successfully masquerade as A_0.

Of course, it may turn out that completeness does not hold. However, a simple informal argument suggests that ground finite detectability is actually fully decidable: it seems almost certain that all maximally strong queries that are detectable only mention predicate names and constants that occur in the probes or the visible part of the policy. If this is true, then there are only finitely many maximally strong detectable queries, and all other detectable queries are implications of these. Hence the set of detectable queries is decidable. So it seems that at least some complete finite axiomatization of detectability exists, although it may be an extension of the given inference system.

There are other interesting open problems apart from completeness. In particular, we have not analyzed the complexity of the inference system, nor have we explored detectability in the context of a non-monotonic language such as Datalog with negation.

Opacity in DKAL *et al.* In Section IV we then proposed a modification to SecPAL's delegation mechanism in order to control information leak from probing.

Cassandra [7], [6] is an earlier authorization language in which probing attacks are mitigated by design, albeit by much cruder means. In Cassandra, submitted credentials must not have any conditional facts. Furthermore, as in most other languages, delegation can only be expressed using conditional facts. Therefore, the only information that may be leaked by probing is the result of the query, and possibly that the submitted credential does not exists in the local policy as an assertion. The latter leakage is also possible in SecPAL$^+$: suppose the adversary submits no credentials and gets a negative result for a query, and subsequently submits a single credential a that does not contain any conditional facts, and receives a positive result for the same query. Then $\neg a$ is detectable in the policy. However, this scenario is usually not problematic, firstly, as the local policy must have been written explicitly to be dependent on a, and secondly, as the adversary's possession of a usually implies that a is not confidential to the adversary. But while Cassandra provides opacity guarantees similar to SecPAL$^+$, it does so at the cost of severely restricting the expressiveness of submitted credentials.

DKAL is a language specifically designed to avoid probing attacks, but the paper [21] does not specify exactly what it is trying to protect against and how, as it lacks a formal framework for quantifying information flow. DKAL does seem to provide similar opacity guarantees as SecPAL$^+$, and at first sight, the mechanism that enforces these is the introduction of a "saysto" operator, which specifies and restricts the *audience* of what is said. But saysto actually does more than that; a statement of the form[3]

$$\text{A saysto B}: \ f_0 \ \text{if} \ f_1$$

has a non-declarative, operational meaning in DKAL. Essentially, it means that as soon as A manages to derive the condition f_1 from local knowledge, it will send a credential stating f_0 via a secured communication channel to B, who will then know that A said f_0. It

[3]The syntax for DKAL/DKAL2 statements is slightly modified here for simplicity. For example, "if" is actually written "←", and says is written said.

follows that the only credentials that can be sent over the network are ones that do not contain any conditions that may depend on other principals' utterances. Therefore, the opacity guarantees in DKAL are achieved by similar means as in Cassandra, and at the same high cost.

The main motivation of DKAL2, then, is to remove DKAL's implicit restriction on credentials [22]. Assertions can now be of the form

$$\mathsf{A\ saysto\ B}:\ [f_0\ \leftarrow\ f_1]\ \mathsf{if}\ f_2$$

So when condition f_2 is locally met, A sends the whole *conditional* credential $f_0 \leftarrow f_1$ to B. But of course, this opens the door to probing attacks. In an attempt to shut the door again, DKAL2 lets B define whitelist filters of the form

$$\mathsf{B\ from}\ e:\ [f_0'\ \leftarrow\ f_1'].$$

So A's conditional credential above is only imported into B's local knowledge if $\langle e : f_0' \leftarrow f_1' \rangle$ is unifiable with $\langle A : f_0 \leftarrow f_1 \rangle$. This purely syntactic ad hoc matching mechanism is problematic as it breaks declarativeness (e.g., $\langle \mathsf{ok} \leftarrow \mathbf{true} \rangle$ may cause a behaviour that is significantly different from the logically equivalent $\langle \mathsf{ok} \rangle$). Also, it is unreasonable to assume that B can foresee all possible syntactic forms of incoming credentials that may be acceptable. But, more importantly for the present discussion, it does not provide adequate protection against probing attacks. Suppose B wishes to delegate authority over a fact f to A, permitting A to redelegate (e.g., A could redelegate f to C, and C actually says f, and A submits the whole redelegation chain to B). Then B needs to provide filters of the form

$$\mathsf{B\ from}\ x:\ [f \leftarrow y\ \mathsf{says}\ f].$$

But this allows anyone in A's redelegation chain to detect $\langle \mathsf{Y\ says}\ f \rangle$ with a self-signed probe, for any Y, which clearly is a breach of confidentiality.

Despite the evident shortcomings of the saysto approach, we initially explored if a similar mechanism could be used in SecPAL to prevent malicious probing attacks. Indeed, we found that this is possible without needing to commit to DKAL's operational, non-declarative semantics for "saysto". One possible solution in this direction would be to replace (CANSAY) by the following rule (and similarly for (CANSAY0)):

$$(\textsc{CanSay}')\ \frac{A \vdash e_1\ \mathsf{saysto}\ e_3:\ e_2\ \mathsf{can\ say}\ f \qquad A \vdash e_2\ \mathsf{saysto}\ e_1:\ f}{A \vdash e_1\ \mathsf{saysto}\ e_3:\ f}$$

This approach was eventually abandoned because the SecPAL$^+$ approach scores higher in several dimensions of usability. A detailed exploration of the design space and a comparative usability analysis will be presented in a future paper.

Conclusion. The short case study of previous attempts to mitigate the probing attack illustrates what can go wrong if the fundamental security questions are not systematically answered [9]: what do we protect, against whom do we protect, what does "protect" mean, and how do we protect?

In this paper, we provided a formal framework as a first step towards answering these questions. The design of the framework was driven by concepts from traditional information flow research. *What we protect* is expressed in terms of multi-level security labels on policy assertions and queries. *Against whom we protect* is specified by a formal definition of probe as a query together with a set of supporting credentials that are temporarily injected into the policy, and adversaries that differ in which probes are available to them, again induced by the security labels. We quantify the effective power of a specific adversary by means of an inference system that tells us what the adversary can detect about the policy. *What protection means* is stated in terms of non-interference, which we found to be too restrictive and coarse, and opacity, both based on a notion of observational equivalence. Finally, we provided a simple and elegant *protection method* that works by restoring symmetry in delegation of authority and comes with strong opacity guarantees.

Acknowledgements. I thank Cédric Fournet, Arne Heizmann, Andy Gordon, and the anonymous referees for helpful comments.

REFERENCES

[1] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin. A calculus for access control in distributed systems. *ACM Transactions on Programming Languages and Systems*, 15(4):706–734, 1993.

[2] M. Becker, J. Mackay, and B. Dillaway. Abductive authorization credential gathering. In *IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY 2009)*, 2009.

[3] M. Y. Becker. SecPAL formalisation and extensions. Technical Report MSR-TR-2009-127, Microsoft Research, 2009.

[4] M. Y. Becker, C. Fournet, and A. D. Gordon. Design and semantics of a decentralized authorization language. In *IEEE Computer Security Foundations Symposium*, pages 3–15, 2007.

[5] M. Y. Becker and S. Nanz. The role of abduction in declarative authorization policies. In *10th International Symposium on Practical Aspects of Declarative Languages (PADL)*, 2008.

[6] M. Y. Becker and P. Sewell. Cassandra: distributed access control policies with tunable expressiveness. In *IEEE International Workshop on Policies for Distributed Systems and Networks*, pages 159–168, 2004.

[7] M. Y. Becker and P. Sewell. Cassandra: Flexible trust management, applied to electronic health records. In *IEEE Computer Security Foundations*, pages 139–154, 2004.

[8] M. Blaze, J. Feigenbaum, and A. D. Keromytis. The role of trust management in distributed systems security. In *Secure Internet Programming*, pages 185–210, 1999.

[9] P. Bonatti, S. Kraus, and V. Subrahmanian. Foundations of secure deductive databases. *IEEE Transactions on Knowledge and Data Engineering*, 7(3):406–422, 1995.

[10] J. Bryans, M. Koutny, L. Mazaré, and P. Ryan. Opacity generalised to transition systems. *International Journal of Information Security*, 7(6):421–435, 2008.

[11] S. Ceri, G. Gottlob, and L. Tanca. What you always wanted to know about Datalog (and never dared to ask). *IEEE Transactions on Knowledge and Data Engineering*, 1(1):146–166, 1989.

[12] D. Clarke, J. E. Elien, C. Ellison, M. Fredette, A. Morcos, and R. L. Rivest. Certificate chain discovery in SPKI/SDSI. *Journal of Computer Security*, 9(4):285–322, 2001.

[13] S. Cosmadakis, H. Gaifman, P. Kanellakis, and M. Vardi. Decidable optimization problems for database logic programs. In *ACM Symposium on Theory of Computing*, pages 477–490, 1988.

[14] D. E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19(5), 1976.

[15] J. Detreville. Binder, a logic-based security language. In *IEEE Symposium on Security and Privacy*, pages 105–113, 2002.

[16] P. Dung. Declarative semantics of hypothetical logic programming with negation as failure. *Lecture Notes in Computer Science*, pages 45–58, 1993.

[17] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. Reasoning About Knowledge. 2003.

[18] C. Farkas and S. Jajodia. The inference problem: a survey. *ACM SIGKDD Explorations Newsletter*, 4(2):6–11, 2002.

[19] R. Focardi and R. Gorrieri. A classification of security properties. *Journal of Computer Security*, 3(1):5–33, 1995.

[20] J. Goguen and J. Meseguer. Security policies and security models. In *IEEE Symposium on Security and privacy*, volume 12, 1982.

[21] Y. Gurevich and I. Neeman. DKAL: Distributed-knowledge authorization language. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 149–162, 2008.

[22] Y. Gurevich and I. Neeman. DKAL 2 – a simplified and improved authorization language. Technical Report MSR-TR-2009-11, Microsoft Research, 2009.

[23] J. Halpern and K. O'Neill. Secrecy in multiagent systems. *ACM Transactions on Information and System Security (TISSEC)*, 12(1), 2008.

[24] S. Jajodia and C. Meadows. Inference problems in multilevel secure database management systems. *Information Security: An Integrated Collection of Essays*, pages 570–584, 1995.

[25] T. Jim. SD3: A trust management system with certified evaluation. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pages 106–115, 2001.

[26] A. C. Kakas, R. A. Kowalski, and F. Toni. The role of abduction in logic programming. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 5, pages 235–324, 1998.

[27] N. Li, B. Grosof, and J. Feigenbaum. A practically implementable and tractable delegation logic. In *IEEE Symposium on Security and Privacy*, pages 27–42, 2000.

[28] N. Li and J. C. Mitchell. Datalog with constraints: A foundation for trust management languages. In *Practical Aspects of Declarative Languages*, pages 58–73, 2003.

[29] N. Li, J. C. Mitchell, and W. H. Winsborough. Design of a role-based trust management framework. In *Symposium on Security and Privacy*, pages 114–130, 2002.

[30] OASIS. *eXtensible Access Control Markup Language (XACML) Version 2.0 core specification*, 2005.

[31] C. O'Halloran. A calculus of information flow. In *Proceedings of the European Symposium on Research in Computer Security, Toulouse, France*, 1990.

[32] C. S. Peirce. Abduction and induction. In J. Buchler, editor, *Philosophical Writings of Peirce*. Dover Publications, Oxford, 1955.

[33] P. Ryan and T. Peacock. Opacity – Further Insights on an Information Flow Property. *Technical Report Series – University of Newcastle Upon Tyne Computing Science*, 958, 2006.

[34] A. Sabelfeld and A. Myers. Language-based information-flow security. *IEEE Journal on selected areas in communications*, 21(1):5–19, 2003.

[35] D. Sutherland. A model of information. In *Proceedings of the 9th National Computer Security Conference*, volume 247, 1986.

[36] J. Ullman. Implementation of logical query languages for databases. *ACM Transactions on Database Systems (TODS)*, 10(3):289–321, 1985.

[37] W. Winsborough and N. Li. Safety in automated trust negotiation. *ACM Transactions on Information and System Security (TISSEC)*, 9(3), 2006.

[38] W. H. Winsborough, K. E. Seamons, and V. E. Jones. Automated trust negotiation. In *DARPA Information Survivability Conference and Exposition*, volume 1, 2000.

[39] S. Zdancewic and A. Myers. Robust declassification. In *IEEE Computer Security Foundations Workshop*, pages 15–23, 2001.