

# A Conference Management System with Verified Document Confidentiality

Sudeep Kanav, Peter Lammich, and Andrei Popescu

Fakultät für Informatik, Technische Universität München, Germany

**Abstract.** We present a case study in verified security for realistic systems: the implementation of a conference management system, whose functional kernel is faithfully represented in the Isabelle theorem prover, where we specify and verify confidentiality properties. The various theoretical and practical challenges posed by this development led to a novel security model and verification method generally applicable to systems describable as input–output automata.

## 1 Introduction

Information-flow security is concerned with preventing or facilitating (un)desired flow of information in computer systems, covering aspects such as confidentiality, integrity, and availability of information. Dieter Gollmann wrote in 2005 [14]: “Currently, information flow and noninterference models are areas of research rather than the bases of a practical methodology for the design of secure systems.” The situation has improved somewhat in the past ten years, with mature software systems such as Jif [1] offering powerful and scalable information flow technology integrated with programming.

However, the state of the art in information-flow security models [23] is still far from finding its way towards applications to real-world systems. If we further restrict attention to *mechanically verified* work, the situation is even more dramatic, with examples of realistic system verification [3, 27] being brave exceptions. This is partly explained by the complexity of information-flow properties, which is much greater than that of traditional functional properties [22]. However, this situation is certainly undesirable, in a world where confidentiality and secrecy raise higher and higher challenges, often far beyond what informal reasoning can comprehend.

In this paper, we take on the task of implementing, and verifying the confidentiality of, a realistic system: a full-fledged conference management system (CMS), featuring multiple users and conferences and offering much of the functionality of widely used systems such as EasyChair [8] and HotCRP [9].

The information flow in a CMS possesses enough complexity so that errors can sneak inside implementations, sometimes with bitter–comical consequences. Recently, Popescu, as well as the authors of 267 papers submitted to a major security conference, initially received an acceptance notification, followed by a retraction [18]: “We are sorry to inform you that your paper was not one of those accepted for this year’s conference. We received 307 submissions and only accepted 40 of them ... We apologize for an earlier acceptance notification. It was due to a system error.”<sup>1</sup>

---

<sup>1</sup> After reading the initial acceptance notification, Popescu went out to celebrate; it was only hours later that he read the retraction.

```
WARNING: HotCRP version 2.47 (commit range 94ca5a0e43bd7dd0565c2c8dc7d8f710a206ab49 through 9c1b45475411ecb85d46bad1f76064881792b038) was subject to an information exposure where some authors could see PC comments. Users of affected versions should upgrade or set the following option in Code/options.inc: $Opt["disableCapabilities"] = true;
```

Fig. 1: Confidentiality bug in HotCRP

The above can be classified as an information-integrity violation: a distorted decision was initially communicated to the authors. The problem with a past version of HotCRP [9] shown in Fig. 1 is even more interesting: it describes a genuine confidentiality violation, probably stemming from the logic of the system, giving the authors capabilities to read confidential comments by the program committee (PC).

Although our methods would equally apply to integrity violations, guarding against confidentiality violations is the focus of this verification work. We verify properties such as the following (where DIS addresses the problem in Fig. 1):

PAP<sub>1</sub>: A group of users learn nothing about a paper unless one of them becomes an author of that paper or a PC member at the paper's conference

PAP<sub>2</sub>: A group of users learn nothing about a paper *beyond the last submitted version* unless one of them becomes an author of that paper

REV: A group of users learn nothing about the content of a review *beyond the last submitted version before the discussion phase and the later versions* unless one of them is that review's author

DIS: The authors learn nothing about the discussion of their paper

We will be concerned with properties restricting the information flow from the various documents maintained by the system (papers, reviews, comments, decisions) towards the users of the system. The restrictions refer to certain conditions (e.g., authorship, PC membership) as well as to upper bounds (e.g., at most the last submitted version) for information release.

We specify the kernel of a CMS using the proof assistant Isabelle [28], with which we formulate and verify confidentiality. The functional implementation of this kernel is automatically synthesized from the specification and wrapped into a web application that offers the expected behavior of a CMS as a graphical menu-based interface.

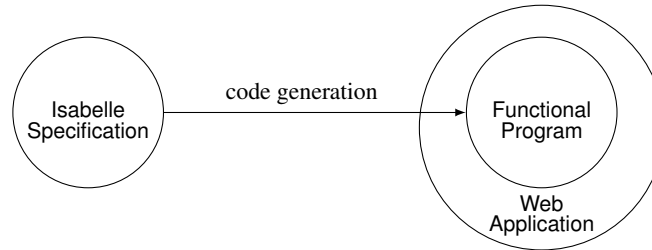
A first contribution of this paper is the engineering approach behind the system specification and implementation (§2). To keep the Isabelle specification (§3) manageable, yet faithful to the implementation and therefore reach a decent balance between trust and usability, we employ state-of-the-art theorem proving and code synthesis technology towards a security-preserving layered architecture.

A second contribution is a novel security model called *bounded-deducibility* (BD) security, born from confronting notions from the literature with the challenges posed by our system (§4). The result is a reusable framework, applicable to any IO automaton. Its main novelty is wide flexibility: it allows the precise formulation of role-based and time-based declassification triggers and of declassification upper bounds. We endow this framework with a declassification-oriented unwinding proof technique (§5).

Our third and last contribution is the verification itself: the BD security framework, its general unwinding theorem, and the unwinding proofs for the CMS confidentiality properties expressed as instances of BD security are all mechanized in Isabelle.

A usable version of the CMS deployed on a web server as well as the Isabelle proof development and the source code are available as supplementary material [19].

## 2 Overall Architecture and Security Guarantees



The architecture of our system follows the paradigm of security by design:

- We formalize and verify the kernel of the system in the Isabelle proof assistant
- The formalization is automatically translated in a functional programming language
- The translated program is wrapped in a web application

**Isabelle Specification** We specify the system as an input–output automaton (where the inputs are called “actions”). We first define, using Isabelle’s records and datatypes, the notions of state (holding all the information about users, conferences, and papers) and user action (representing user requests for manipulating documents and rights in the system: upload/download papers, edit reviews, assign reviewers, etc.). Then we define the step function that takes a state and an action and returns a new state and an output.

**Scala Functional Program** The specification was designed to fall within the executable fragment of Isabelle. This allows us to automatically synthesize, using Isabelle’s code generator [16], a program in the functional fragment of Scala [2] isomorphic to the specification. The types of data used in the specification (numbers, strings, tuples, records) are mapped to the corresponding Scala types. An exception is the Isabelle type of paper contents, which is mapped to the Scala/JVM file type.

**Web Application** Finally, the Scala program is wrapped in a web application, offering a menu-based user interface. Upon login, a user sees his conferences and his roles for each of them; the menus offer role-sensitive choices, e.g., assign reviewers (for chairs) or upload papers (for authors).

**Overall Security Guarantees** Our Isabelle verification targets information-flow properties. These properties express that for any possible trace of the system, there is no way to infer from certain observations on that trace (e.g., actions performed by designated users), certain values extracted from that trace (e.g., the paper uploads by other users). The question arises as to what guarantees we have that the properties we verified formally for the specification also hold for the overall system. E.g., if we prove in Isabelle that users never learn the content of other users’ papers, how can we be sure that this is actually the case when using the web interface? We do not have a formal answer to this, but only an informal argument in terms of the trustworthiness of two trusted steps.

First, we need to trust Isabelle’s code generator. Its general-purpose design is very flexible, supporting program and data refinement [16]. In the presence of these rich features, the code generator is only known to preserve partial correctness, hence safety properties [15, 16]. However, here we use the code generator in a very restrictive manner, to “refine” an already deterministic specification which is an implementation in its

own right—the code generator simply translates it from the functional language of Isabelle to that of Scala. In addition, all the used Isabelle functions are proved to terminate, and nontrivial data refinement is disabled. These allow us to (informally) conclude that the synthesized implementation is trace-isomorphic to the specification, hence the former leaks as little information as the latter. (This meta-argument does not cover timing channels, but these seem to be of little importance for leaking document content.)

Second, we need to trust that no further leakage occurs via the web application wrapper. To acquire this trust, we make sure that the web application acts as a stateless interface to the step function: upon a user request, all it does is invoke “step” (one or multiple times) with input from the user and then process and display the output of the step function. The only exception is that, for storage efficiency, we store paper content in files. The web application has to be trusted to create files and deliver their content back to users—this is straightforwardly implemented, hence unlikely to cause errors.

In summary, the formal guarantees we provide in Isabelle have to be combined with a few trusted steps to apply to the whole system. Our verification targets only the system’s implementation logic—lower-level attacks such as browser-level forging are out of its reach, but are orthogonal issues that could in principle be mitigated separately.

### 3 System Specification

The system behaves similarly to EasyChair [8], a popular CMS created by Andrei Voronkov. It hosts multiple users and conferences, allowing the creation of new users and conferences at any time. The system has a superuser, which we call *voronkov* as a tribute to EasyChair. The *voronkov* is the first user of the system, and his role is to approve new-conference requests. A conference goes through several successive phases.

**No-Phase** Any user can apply for a new conference, with the effect of registering it in the system with “No-Phase”. After approval from the *voronkov*, the conference moves to the setup phase, with the applicant becoming a conference chair.

**Setup** Any conference chair can add new chairs and new PC members (with a chair also being a PC member). From here on, moving the conference to successor phases can be done by the chairs.

**Submission** A user can list the conferences awaiting submissions (i.e., being in submission phase). He can submit a paper, upload new versions, or indicate other users as coauthors thereby granting them reading and editing rights.

**Bidding** Authors are no longer allowed to upload or register new papers and PC members are allowed to view the submitted papers. PC members can place bids, indicating for each paper one of the following preferences: “want to review”, “would review”, “no preference”, “would not review”, and “conflict”. If the preference is “conflict”, the PC member cannot be assigned that paper, and will not see its discussion. “Conflict” is assigned automatically to papers authored by a PC member.

**Reviewing** Chairs can assign papers to PC members for reviewing either manually or by invoking an external program to establish fair assignment based on some parameters: preferences, number of papers per PC member, and number of reviewers per paper.

**Discussion** All PC members having no conflict with a paper can see its reviews and can add comments. Also, chairs can edit the decision.

**Notification** The authors can read the reviews and the accept/reject decision, which no one can edit any longer.

### 3.1 State, Actions, and Step Function

The state stores the lists of registered conference, user, and paper IDs and, for each ID, actual conference, user, or paper information. Each paper ID is assigned a paper having title, abstract, content, and, in due time, a list of reviews, a discussion text, and a decision:  $\text{Paper} = \text{String} \times \text{String} \times \text{Paper\_Content} \times \text{List}(\text{Review}) \times \text{Dis} \times \text{Dec}$

We keep different versions of the decision and of each review, as they may transparently change during discussion:  $\text{Dec} = \text{List}(\text{String})$  and  $\text{Review} = \text{List}(\text{Review\_Content})$  where  $\text{Review\_Content}$  consists of triples (expertise, text, score).

In addition, the state stores: for each conference, the list of (IDs of) papers submitted to that conference, the list of news updated by the chairs, and the current phase; for each user and paper, the preferences resulted from biddings; for each user and conference, a list of roles: chair, PC member, paper author, or paper reviewer (the last two roles also containing paper IDs).

```
record State =
  confIDs : List(ConfID)   conf : ConfID → Conf   userIDs : List(UserID)
  pass : UserID → Pass   user : UserID → User   roles : ConfID → UserID → List(Role)
  paperIDs : ConfID → List(PaperID)   paper : PaperID → Paper
  pref : UserID → PaperID → Pref   news : ConfID → List(String)   phase : ConfID → Phase
```

Actions are parameterized by user IDs and passwords. There are 45 actions forming five categories: creation, update, undestructive update (u-update), reading and listing.

The **creation actions** register new objects (users, conferences, chairs, PC members, papers, authors), assign reviewers (by registering new review objects), and declare conflicts. E.g., `cPaper cid uid pw pid title abs` is an action by user `uid` with password `pw` attempting to register to conference `cid` a new paper `pid` with indicated title and abstract.

The **update actions** modify the various documents of the system: user information and password, paper content, reviewing preference, review content, etc. For example, `uPaperC cid uid pw pid ct` is an attempt to upload a new version of paper `pid` by modifying its content to `ct`. The **u-update actions** are similar, but also record the history of a document's versions. E.g., if a reviewer decides to change his review during the discussion phase, then the previous version is still stored in the system and visible to the other PC members (although never to the authors). Other documents subject to u-updates are the news, the discussion, and the accept-reject decision.

The **reading actions** access the content of the system's documents: papers, reviews, comments, decisions, news. The **listing actions** produce lists of IDs satisfying various filters—e.g., all conferences awaiting paper submissions, all PC members of a conference, all the papers submitted by a given user, etc.

Note that the first three categories of actions are aimed at *modifying* the state, and the last two are aimed at *observing* the state through outputs. However, the modification actions also produce a simple output, since they may succeed or fail. Moreover, the observation actions can also be seen as changing the state to itself. Therefore we can assume that both types produce a pair consisting of a new state and an action.

We define the function  $\text{step} : \text{State} \rightarrow \text{Act} \rightarrow \text{Out} \times \text{State}$  that operates by determining the type of the action and dispatching specialized handler functions. The initial state of the system,  $\text{istate} \in \text{State}$ , is the one with a single user, the voronkov, and a dummy password (which can be changed immediately). The step function and the initial state are the only items exported by our specification to the outside world.

## 4 Security Model

Here we first analyze the literature for possible inspiration concerning a suitable security model for our system. Then we introduce our own notion, which is an extension of Sutherland’s nondeducibility [36] that factors in declassification triggers and bounds.

### 4.1 Relevant Literature

There is a vast amount of literature on information-flow security, with many variants of formalisms and verification techniques. An important distinction is between notions that completely forbid information flow (between designated sources and sinks) and notions that only restrict the flow, allowing some declassification. Historically, the former were introduced first, and the latter were subsequently introduced as generalizations.

**Absence of Information Flow** The information-flow security literature starts in the late 1970s and early 1980s [6, 12, 30], motivated by the desire to express the absence of information leaks of systems more abstractly and more precisely than by means of access control [4, 20]. Very influential were Goguen and Meseguer’s notion of noninterference [12] and its associated proof by unwinding [13]. Unwinding is essentially a form of simulation that allows one to construct incrementally, from a perturbed trace of the system, an alternative “corrected” trace that “closes the leak”. Many other notions were introduced subsequently, either in specialized programming-language-based [34] or process-algebra-based [11, 33] settings or in purely semantic, event-system-based settings [24, 25, 29, 36]. (Here we are mostly interested in the last category.) These notions are aimed at extending noninterference to nondeterministic systems, closing Trojan-horse channels, or achieving compositionality. The unwinding technique has been generalized for some of these variants—McLean [26] and Mantel [22] give overviews.

Even ignoring our aimed declassification aspect, most of these notions do not adequately model our properties of interest exemplified in the introduction. One problem is that they are not flexible enough w.r.t. the observations. They state nondetectability of absence or occurrence of certain events anywhere in a system trace. By contrast, we are interested in a very controlled positioning of such undetectable events: in the property  $PAP_2$  from the introduction, the unauthorized user should not learn of preliminary (non-final) uploads of a paper. Moreover, we are not interested in whole events, but rather in certain relevant values extracted from the events: e.g., the content of the paper, and not the ID of one of the particular authors who uploads it.

A fortunate exception to the above flexibility problems is Sutherland’s early notion of *nondeducibility* [36]. One considers a set of worlds  $World$  and two functions  $F : World \rightarrow J$  and  $H : World \rightarrow K$ . For example, the worlds could be the valid traces of the system,  $F$  could select the actions of certain users (potential attackers), and  $H$  could select the actions of other users (intended as being secret). *Nondeducibility of  $H$  from  $F$*  says that the following holds for all  $w \in World$ : for all  $k$  in the image of  $H$ , there exists  $w_1 \in World$  such that  $F w_1 = F w$  and  $H w_1 = k$ . Intuitively, for all  $F$  knows about the actual world  $w$ , (the value of)  $H$  can be anything (in the image of  $H$ ). The generality of this framework allows us to fine-tune both the location of the relevant events in the trace and their values of interest. But generality is no free lunch: it is no longer clear how to provide an unwinding-like incremental proof method.

Halpern and O’Neill [17] recast nondeducibility as a property called secrecy maintenance, in a multi-agent framework of “runs-and-systems” [31] based on epistemic logic. Their formulation enables general-purpose epistemic logic primitives for deducing absence of leaks, but no unwinding or any other inductive reasoning technique.

On the practical verification side, Arapinis et al. [3] introduce ConfiChair, a CMS that improves on standard systems such as EasyChair by guaranteeing that “the cloud”, consisting of the system provider/administrator, cannot learn the content of the papers and reviews and cannot link users with their written reviews. This is achieved by a cryptographic protocol based on key translations and mixes. They encode the desired properties as strong secrecy (a property similar to Goguen-Meseguer noninterference) and verify them using the ProVerif [5] tool specialized in security protocols. Our work differs from theirs in three major aspects. First, they propose a cryptography-based enhancement, while we focus on a traditional CMS not involving cryptography. Second, they manage to encode and verify the desired properties automatically, while we use interactive theorem proving. While their automatic verification is an impressive achievement, we cannot hope for the same with our targeted properties which, while having a similar nature, are more nuanced and complex. E.g., properties like  $PAP_2$  and  $REV$ , with such flexible indications of declassification bounds, go far beyond strong secrecy and require interactive verification. Finally, we synthesize functional code isomorphic to the specification, whereas they provide a separate implementation, not linked to the specification which abstracts away from many functionality aspects.

**Restriction of Information Flow** A large body of work on declassification was pursued in a language-based setting. Sabelfeld and Sands [35] give an overview of the state of the art up to 2009. Although they target language-based declassification, they phrase some generic dimensions of declassification most of which apply to our case:

- What information is released? Here, document content, e.g., of papers, reviews, etc.
- Where in the system is information released? In our case, the relevant “where” is a “from where” (referring to the source, not to the exit point): from selected places in the system trace, e.g., the last submitted version before the deadline.
- When can information be released? After a certain trigger occurs, e.g., authorship.

Sabelfeld and Sands consider another interesting instance of the “where” dimension, namely intransitive noninterference [21, 32]. This is an extension of noninterference that allows downgrading of information, say, from High to Low, via a controlled Declassifier level. It could be possible to encode aspects of our properties of interest as intransitive noninterference—e.g., we could encode the act of a user becoming an author as a declassifying action for the target paper. However, such an encoding would be rather technical and somewhat artificial for our system; additionally, it is not clear how to factor in our aforementioned specific “where” dimension.

Recently, the “when” aspect of declassification has been included as first-class citizen in customized temporal logics [7, 10], which can express aspects of our desired properties, e.g., “unless/until he becomes an author”. Their work is focused on efficiently model-checking finite systems, whereas we are interested in verifying an infinite system. Combining model checking with infinite-to-finite abstraction is an interesting prospect, but reflecting information-flow security properties under abstraction is problematic.

## 4.2 Bounded-Deducibility Security

We introduce a novel notion of information-flow security that:

- retains the precision and versatility of nondeducibility
- factors in declassification as required by our motivating examples
- is amenable to a general unwinding technique

We shall formulate security in general, not only for our concrete CMS from §3.1, but for any IO automaton indicated by the following data. We fix sets of states,  $\text{State}$ , actions,  $\text{Act}$ , and outputs,  $\text{Out}$ , an initial state  $\text{istate} \in \text{State}$ , and a step function  $\text{step} : \text{State} \rightarrow \text{Act} \rightarrow \text{Out} \times \text{State}$ . We let  $\text{Trans}$ , the set of *transitions*, be  $\text{State} \times \text{Act} \times \text{Out} \times \text{State}$ . Thus, a transition  $\text{trn}$  is a tuple, written  $(s, a, o, s')$ ;  $s$  indicates the source,  $a$  the action,  $o$  the output, and  $s'$  the target.  $\text{trn}$  is called *valid* if it is induced by the step function, namely  $\text{step } s a = (o, s')$ .

A *trace*  $\text{tr} \in \text{Trace}$  is any list of transitions:  $\text{Trace} = \text{List}(\text{Trans})$ . For any  $s \in \text{State}$ , the set of valid traces starting in  $s$ ,  $\text{Valid}_s \subseteq \text{Trace}$ , consists of the traces of the form  $[(s_1, a_1, o_1, s_2), (s_2, a_2, o_2, s_3), \dots, (s_{n-1}, a_{n-1}, o_n, s_n)]$  for some  $n$  where  $s_1 = s$  and each transition  $(s_i, a_i, o_i, s_i)$  is valid. We will be interested in the valid traces starting in the initial state  $\text{istate}$ —we simply call these *valid traces* and write  $\text{Valid}$  for  $\text{Valid}_{\text{istate}}$ .

Besides the IO automaton, we assume that we are given the following data:

- a *value domain*  $\text{Val}$ , together with a *value filter*  $\varphi : \text{Trans} \rightarrow \text{Bool}$  and a *value producer*  $f : \text{Trans} \rightarrow \text{Val}$
- an *observation domain*  $\text{Obs}$ , together with an *observation filter*  $\gamma : \text{Trans} \rightarrow \text{Bool}$  and an *observation producer*  $g : \text{Trans} \rightarrow \text{Obs}$

We define the *value function*  $\mathbb{V} : \text{Trace} \rightarrow \text{List}(\text{Val})$  componentwise, filtering out values not satisfying  $\varphi$  and applying  $f$ :

$$\mathbb{V} [] \equiv [] \qquad \mathbb{V}([\text{trn}] \cdot \text{tr}) \equiv \text{if } \varphi \text{ trn then } (f \text{ trn}) \cdot (\mathbb{V} \text{tr}) \text{ else } \mathbb{V} \text{tr}$$

We also define the *observation function*  $\mathbb{O} : \text{Trace} \rightarrow \text{List}(\text{Obs})$  just like  $\mathbb{V}$ , but using  $\gamma$  as a filter and  $g$  as a producer.

We think of the above as an instantiation of the abstract framework for nondeducibility recalled in §4.1, where  $\text{World} = \text{Valid}$ ,  $F = \mathbb{O}$ , and  $H = \mathbb{V}$ . Thus, nondeducibility states that the observer  $\mathbb{O}$  may learn nothing about  $\mathbb{V}$ . Here we are concerned with a more fine-grained analysis, asking ourselves *what* may the observer  $\mathbb{O}$  learn about  $\mathbb{V}$ .

Using the idea underlying nondeducibility, we can answer this precisely: Given a trace  $\text{tr} \in \text{Valid}$ , the observer sees  $\mathbb{O} \text{tr}$  and therefore can infer that  $\mathbb{V} \text{tr}$  belongs to the set of all values  $\mathbb{V} \text{tr}_1$  for some  $\text{tr}_1 \in \text{Valid}$  such that  $\mathbb{O} \text{tr}_1 = \mathbb{O} \text{tr}$ . In other words, he can infer that the value is in the set  $\mathbb{V}(\mathbb{O}^{-1}(\mathbb{O} \text{tr}) \cap \text{Valid})$ , and nothing beyond this. We call this set the *declassification* associated to  $\text{tr}$ , written  $\text{Dec}_{\text{tr}}$ .

We want to establish, under certain conditions, *upper* bounds for declassification, or in set-theoretic terms, *lower* bounds for  $\text{Dec}_{\text{tr}}$ . To this end, we further fix:

- a declassification bound  $\text{B} : \text{List}(\text{Val}) \rightarrow \text{List}(\text{Val}) \rightarrow \text{Bool}$
- a declassification trigger  $\text{T} : \text{Trans} \rightarrow \text{Bool}$

The system is called *bounded-deducibility-secure* (*BD-secure*) if for all  $\text{tr} \in \text{Trace}$  such that  $\text{never } \text{T } \text{tr}$ , it holds that  $\{\text{vl}_1 \mid \text{B}(\mathbb{V} \text{tr}) \text{ vl}_1\} \subseteq \text{Dec}_{\text{tr}}$  (where “never  $\text{T } \text{tr}$ ” means “ $\text{T}$  holds for no transition in  $\text{tr}$ ”). Informally, BD security expresses the following:

*If the trigger  $\text{T}$  never holds (i.e., unless  $\text{T}$  eventually holds, i.e., until  $\text{T}$  holds),  
the observer  $\mathbb{O}$  can learn nothing about the values  $\mathbb{V}$  beyond  $\text{B}$*



We can think of  $B$  positively, as an upper bound for declassification, or negatively, as a lower bound for uncertainty. On the other hand,  $T$  is a trigger removing the bound  $B$ —as soon as  $T$  becomes true, the containment of declassification is no longer guaranteed. In the extreme case of  $B$  being everywhere true and  $T$  everywhere false, we have no declassification, i.e., total uncertainty—in other words, standard nondeducibility.

Unfolding some definitions, we can alternatively express BD security as the following being true for all  $tr \in \text{Valid}$  and  $vl, vl_1 \in \text{List}(\text{Val})$ :

$$\text{never } T \ tr \wedge \forall tr = vl \wedge B \ vl \ vl_1 \rightarrow (\exists tr_1 \in \text{Valid}. O \ tr_1 = O \ tr \wedge \forall tr_1 = vl_1) \quad (*)$$

### 4.3 Discussion

We have argued for BD security as a means to restrict the information flowing from the system to certain observers. However, as an extension of nondeducibility, BD security is in principle subject to the same criticism. The problem with nondeducibility [24,26,33] is that in some cases it is too weak, since it takes as *plausible* each possible explanation for an observation: if the observation sequence is  $ol$ , then any trace  $tr$  such that  $O \ tr = vl$  is plausible. But what if the low-level observers can synchronize their actions and observations with the actions of other entities, such as a high-level user or a Trojan horse acting on his behalf, or even a third-party entity that is neither high nor low? Even without synchronization, the low-level observer may learn from outside the system, of certain behavior patterns of the high-level users. Then the set of plausible explanations can be reduced, leading to information leak.

In our case, the low-level observers are a group of users assumed to never acquire a certain status (e.g., authorship of a paper). The other users of the system are either “high-level” (e.g., the authors of the paper) or “third-party” (e.g., the non-author users not in the group of observers). Concerning the high-level users, it does not make sense to assume that they would cooperate to leak information *through the system*, since they certainly have better means to do that outside the system, e.g., via email. Users also do not have to trust external software, since everything is filtered through the system kernel—e.g., a chair can run an external linear-programming tool for assigning reviewers, but each assignment is still done through the verified step function. As for the possible third-party cooperation towards leaks of information, this is bypassed by our consideration of arbitrary groups of observers: in the worst case, all the unauthorized users can be placed in this group.

However, the possibility to learn and exploit behavior patterns from outside the system is not explicitly addressed by BD security—it would be best dealt with by a probabilistic analysis. Here, what we could do (but have not done yet) to alleviate such refinement-attack concerns is increase the power of the observers. E.g., for  $PAP_1$  and  $PAP_2$  we could allow the observers to view not only their own actions and outputs, but also the contents of documents *different from the paper under scrutiny* manipulated by “high-level” users. Such instances of BD security would still hold, while relying less on nondeterminism (technically, in  $(*)$ ,  $tr_1$  would be constrained to be more similar to  $tr$ ).

### 4.4 Instantiation to Our Running Examples

Recall that BD security involves the following parameters:

- an IO automaton (State, Act, Out, istate, step)

- infrastructures for values ( $\text{Val}, \varphi, f$ ) and observations ( $\text{Obs}, \gamma, g$ )
- a declassification specification: trigger  $T$  and bound  $B$

In particular, this applies to our CMS IO automaton. BD security then captures our examples by suitably instantiating the observation and declassification parameters.

For all our examples, we have the same observation infrastructure. We fix UIDs, the set of IDs of the observing users. We let  $\text{Obs} = \text{Act} \times \text{Out}$ . Given a transition,  $\gamma$  holds iff the action’s subject is a user in UIDs, and  $g$  returns the pair (action,output).  $O\ tr$  thus purges  $tr$  keeping only actions of users in UIDs.

The value infrastructure depends on the considered type of document. For  $\text{PAP}_1$  and  $\text{PAP}_2$  we fix PID, the ID of the paper of interest. We let  $\text{Val} = \text{List}(\text{Paper\_Content})$ . Given a transition,  $\varphi$  holds iff the action is an upload of paper PID, and  $f$  returns the uploaded content.  $V\ tr$  thus returns the list of all uploaded paper contents for PID.

The declassification triggers and bounds are specific to each example. For  $\text{PAP}_1$ , we define  $T(s, a, o, s')$  as “in state  $s'$ , some user in UIDs is an author of PID or a PC member of some conference  $cid$  where PID is registered,” formally:

$$\exists uid \in \text{UIDs}. \text{isAut } s' \text{ uid PID} \vee (\exists cid. \text{PID} \in \text{paperIDs } s' \text{ cid} \wedge \text{isPC } s' \text{ uid cid})$$

Intuitively, the intent with  $\text{PAP}_1$  is that, provided  $T$  never holds, users in UIDs learn nothing about the various consecutive versions of PID. But is it true that they can learn *absolutely nothing*? There is a remote possibility that a user could infer that no version was submitted (by probing the current phases of the conferences in the system and noticing that none has reached the submission phase). But indeed, nothing beyond this quite harmless information should leak: any nonempty value sequence  $vl$  might as well have been any other (possibly empty!) sequence  $vl_1$ . Hence we define  $B\ vl\ vl_1$  as  $vl \neq []$ .

For  $\text{PAP}_2$ , the trigger involves only authorship, ignoring PC membership at the paper’s conference—we take  $T(s, a, o, s')$  to be  $\exists uid \in \text{UIDs}. \text{isAut } s' \text{ uid PID}$ . Here we have a genuine example of nontrivial declassification bound—since a PC member can learn the paper’s content but only as its last submitted version, we define  $B\ vl\ vl_1$  as  $vl \neq [] \neq vl_1 \wedge \text{last } vl = \text{last } vl_1$ , where the function  $\text{last}$  returns the last element of a list.

For  $\text{REV}$ , the value infrastructure refers not only to the review’s content but also to the conference phase:  $\text{Val} = \text{List}(\text{Phase} \times \text{Review\_Content})$ . The functions  $\varphi$  and  $f$  are defined similarly to those for paper contents, mutatis mutandis; in particular,  $f$  returns a pair  $(ph, rct)$  consisting of the conference’s current phase and the updated review’s content; hence  $V$  returns a list of such pairs. The trigger  $T$  is similar to that of  $\text{PAP}_2$  but refers to review authorship rather than paper authorship. The bound  $B$  is more complex. Any user can infer that the only possibilities for the phase are Reviewing and Discussion, in this order—i.e., that  $vl$  has the form  $ul \cdot wl$  such that the pairs in  $ul$  have Reviewing as first component and the pairs in  $wl$  have Discussion. Moreover, any PC member having no conflict with PID can learn  $\text{last } ul$  (the last submitted version before Discussion), and  $wl$  (the versions updated during Discussion, public to non-conflict PC members); but (until  $T$  holds) nothing beyond these. So  $B\ vl\ vl_1$  states that  $vl$  decomposes as  $ul \cdot wl$  as indicated above,  $vl_1$  decomposes similarly as  $ul_1 \cdot wl_1$ , and  $\text{last } ul = \text{last } ul_1$ .

DIS needs rephrasing to be captured as BD security. It can be decomposed into:

DIS<sub>1</sub>: An author always has conflict with his papers

DIS<sub>2</sub>: A group of users learn nothing about a paper’s discussion unless one of them becomes a PC member at the paper’s conference *having no conflict with the paper*

DIS<sub>1</sub> is a safety property. DIS<sub>2</sub> is an instance of BD security defined as expected.

	Source	Declassification Trigger	Declassification Bound
1	Paper Content	Paper Authorship	Last Uploaded Version
2		Paper Authorship or PC Membership <sup>B</sup>	Absence of Any Upload
3	Review	Review Authorship	Last Edited Version Before Discussion and All the Later Versions
4		Review Authorship or Non-Conflict PC Membership <sup>D</sup>	Last Edited Version Before Notification
5		Review Authorship or Non-Conflict PC Membership <sup>D</sup> or PC Membership <sup>N</sup> or Paper Authorship <sup>N</sup>	Absence of Any Edit
6	Discussion	Non-Conflict PC Membership	Absence of Any Edit
7	Decision	Non-Conflict PC Membership	Last Edited Version
8		Non-Conflict PC Membership or PC Membership <sup>N</sup> or Paper Authorship <sup>N</sup>	Absence of Any Edit
9	Reviewer Assignment	Non-Conflict PC Membership <sup>R</sup>	Non-Conflict PC Membership

Phase Stamps: B = Bidding, D = Discussion, N = Notification, R = Review

Fig. 2: Confidentiality properties

#### 4.5 More Instances

Fig. 2 shows an array of confidentiality properties we have formulated as BD security. They provide a classification of relevant roles, statuses and conference phases that are necessary conditions for degrees of information release. The observation infrastructure is always the same, given by the actions and outputs of a fixed group of users as in §4.4.

The table lists several information sources, each yielding a different value infrastructure. In rows 1–8, the sources are actual documents: paper content, review, discussion, decision. The properties  $PAP_1$ ,  $PAP_2$ ,  $REV$  and  $DIS_2$  form the rows 2, 1, 3, and 6. In row 9, the source is the identities of the reviewers assigned to the paper.

The declassification triggers express paper or review authorship (being or becoming an author of the indicated document) or PC membership at the paper’s conference, with or without the requirement of lack of conflict with the paper. Some triggers are also listed with “phase stamps” that strengthens the statements. E.g., row 2 contains a strengthening of the trigger discussed so far for  $PAP_1$ : “PC membership<sup>B</sup>” should be read as “PC membership and paper’s conference phase being at least bidding.” Some of the triggers require lack of conflict with the paper, which is often important for the security statement to be strong enough. This is the case of  $DIS_2$  (row 6), since without the non-conflict assumption  $DIS_2$  and  $DIS_1$  would no longer imply  $DIS$ . By contrast, lack of conflict cannot be added to PC membership in  $PAP_1$  (row 2), since such a stronger version would not hold: even if a PC member decides to indicate conflict with a paper, this happens after he had the opportunity to see the paper’s content.

Most of the declassification bounds are similar to those from §4.4. The row 9 property states that, unless one becomes a PC member having no conflict with a paper in the reviewing phase, one can’t learn anything about the paper’s assigned reviewers beyond of course what everyone knows: that reviewers are non-conflict PC members.

## 5 Verification

To cope with general declassification bounds, BD security speaks about system traces in conjunction with value sequences that must be produced by these traces. We extend the unwinding proof technique to this situation and employ the result to the verification of our confidentiality properties.

### 5.1 Unwinding Proof Method

We see from definition (\*) that to prove BD security, one starts with a valid  $tr$  (starting in  $s$  and having value sequence  $vl$ ) and an “alternative” value sequence  $vl_1$  such that  $B\ vl\ vl_1$  and one needs to produce an “alternative” trace  $tr_1$  starting in  $s$  whose value sequence is  $vl_1$  and whose observation sequence is the same as that of  $tr$ .

In the tradition of unwinding for noninterference [13, 32], we wish to construct  $tr_1$  from  $tr$  *incrementally*: as  $tr$  grows,  $tr_1$  should grow nearly synchronously. In order for  $tr_1$  to have the same observation sequence (produced by  $O$ ) as  $tr$ , we need to require that the observable transitions of  $tr_1$  (i.e., for which  $\gamma$  holds) be *identical* to those of  $tr$ .

As for the value sequences (produced by  $V$ ), we face the following problem. In contrast to the unwinding relations studied so far in the literature, we must consider an additional parameter, namely the *a priori given* value sequence  $vl_1$  that needs to be produced by  $tr_1$ . In fact, it appears that one would need to maintain, besides an unwinding relation on states  $\theta : \text{State} \rightarrow \text{State} \rightarrow \text{Bool}$ , also an “evolving” generalization of the declassification trigger  $B$ ; then  $\theta$  and  $B$  would certainly need to be synchronized. We resolve this by enlarging the domain of the unwindings to quaternary relations  $\Delta : \text{State} \rightarrow \text{List}(\text{Val}) \rightarrow \text{State} \rightarrow \text{List}(\text{Val}) \rightarrow \text{Bool}$  that generalize both  $\theta$  and  $B$ . Intuitively,  $\Delta\ s\ vl\ s_1\ vl_1$  keeps track of the current state of  $tr$ , the remaining value sequence of  $tr$ , the current state of  $tr_1$ , and the remaining value sequence of  $tr_1$ .

Let the predicate *consume*  $trn\ vl\ vl'$  mean that the transition  $trn$  either produces a value that is consumed from  $vl$  yielding  $vl'$  or produces no value and  $vl = vl'$ . Formally:

$$\text{if } \varphi\ trn \text{ then } (vl \neq [] \wedge f\ trn = \text{head } vl \wedge vl' = \text{tail } vl) \text{ else } (vl' = vl)$$

In light of the above discussion, we are tempted to define an unwinding as a relation  $\Delta$  such that  $\Delta\ s\ vl\ s_1\ vl_1$  implies *either* of the following conditions:

- REACTION: For any valid transition  $(s, a, o, s')$  and lists of values  $vl, vl'$  such that *consume*  $(s, a, o, s')\ vl\ vl'$  holds, either of the following holds:
  - IGNORE: The transition yields no observation ( $\neg \gamma\ a\ o$ ) and  $\Delta\ s'\ vl'\ s_1\ vl_1$  holds
  - MATCH: There exist a valid transition  $(s_1, a_1, o_1, s'_1)$  and a list of values  $vl'_1$  such that *consume*  $(s, a, o, s')\ vl_1\ vl'_1$  and  $\Delta\ s_1\ vl'\ s'_1\ vl'_1$  hold
- INDEPENDENT ACTION: There exist a valid transition  $(s_1, a_1, o_1, s'_1)$  that yields no observation ( $\neg \gamma\ a_1\ o_1$ ) and a list of values  $vl'_1$  such that *consume*  $a_1\ o_1\ vl_1\ vl'_1$  and  $\Delta\ s\ vl\ s'_1\ vl'_1$  hold

The intent is that BD security should hold if there exists an unwinding  $\Delta$  that “initially includes”  $B$ . A trace  $tr_1$  could then be constructed incrementally from  $tr$ ,  $vl$  and  $vl_1$ , applying REACTION or INDEPENDENT ACTION until the three lists become empty.

**Progress** However, such an argument faces difficulties. First, INDEPENDENT ACTION is not guaranteed to decrease any of the lists. To address this, we strengthen INDEPENDENT ACTION by adding the requirement that  $\varphi\ (s_1, a_1, o_1, s'_1)$  holds—this ensures

that  $vl_1$  decreases (i.e.,  $vl'_1$  is strictly shorter than  $vl_1$ ). This way, we know that each REACTION and INDEPENDENT ACTION decreases at least one list: the former  $tr$  and the latter  $vl_1$ ; and since  $vl$  is empty whenever  $tr$  is, the progress problem seems resolved.

Yet, there is a second, more subtle difficulty: after  $tr$  has become empty, how can we know that  $vl_1$  will start decreasing? With the restrictions so far, one may still choose REACTION with parameters that leave  $vl_1$  unaffected. So we need to make sure that the following implication holds: if  $tr = []$  and  $vl_1 \neq []$ , then  $vl_1$  will be consumed. Since from inside the unwinding relation we cannot (and do not want to!) see  $tr$ , but only  $vl$ , we weaken the assumption of this implication to “if  $vl = []$  and  $vl_1 \neq []$ ;” moreover, we strengthen its conclusion to requiring that only the INDEPENDENT ACTION choice (guaranteed to shorten  $vl_1$ ) be available. Equivalently, we condition the alternative choice of REACTION by the negation of the above, namely  $vl \neq [] \vee vl_1 = []$ .

**Exit Condition** The third observation is not concerned with a difficulty, but with an optimization. We note that BD security holds trivially if the original trace  $tr$  cannot saturate the value list  $vl$ , i.e., if  $\forall tr \neq vl$ —this happens if and only if, at some point, an element  $v$  of  $vl$  can no longer be saturated, i.e., for some decompositions  $tr = tr' \cdot tr''$  and  $vl = vl' \cdot [v] \cdot vl''$  of  $tr$  and  $vl$ , it holds that  $\forall tr' = vl'$  and  $\forall trn \in tr''. \varphi trn \rightarrow f trn \neq v$ . Can we detect such a situation from within  $\Delta$ ? The answer is (an over-approximated) yes: after  $\Delta s vl s_1 vl_1$  evolves by REACTION and INDEPENDENT ACTION to  $\Delta s' ([v] \cdot vl'') s'_1 vl'_1$  for some  $s', s'_1$  and  $vl'_1$  (presumably consuming  $tr'$  and saturating the  $vl'$  prefix of  $vl$ ), then one can safely exit the game if one proves that no valid trace  $tr''$  starting from  $s'$  can ever saturate  $v$ , in that it satisfies  $\forall trn \in tr''. \varphi trn \rightarrow f trn \neq v$ .

The final definition of BD unwinding is given below, where  $reach : State \rightarrow Bool$  is the state reachability predicate and  $reach_{\neg T} : State \rightarrow Bool$  is its strengthening to reachability by transitions that do not satisfy T:

$$\begin{aligned} \text{unwind } \Delta \equiv & \forall s vl s_1 vl_1. \text{reach}_{\neg T} s \wedge \text{reach } s_1 \wedge \Delta s vl s_1 vl_1 \rightarrow \\ & ((vl \neq [] \vee vl_1 = []) \wedge \text{reaction } \Delta s s vl s_1 vl_1) \vee \\ & \text{iaction } \Delta s s vl s_1 vl_1 \vee \\ & (vl \neq [] \wedge \text{exit } s (\text{head } vl)) \end{aligned}$$

The predicates  $\text{iaction}$  and  $\text{reaction}$  formalize INDEPENDENT ACTION (with its aforementioned strengthening) and REACTION, the latter being a disjunction of predicates formalizing IGNORE and MATCH. The predicate  $\text{exit } s v$  is defined as  $\forall tr trn. (tr \cdot [trn]) \in \text{Valid}_s \wedge \varphi trn \rightarrow f trn \neq v$ . It expresses a safety-like property, and therefore can be verified in a trace-free manner.

We can now prove that indeed any unwinding relation constructs an “alternative” trace  $tr_1$  from any trace  $tr$  starting in a  $P$ -reachable state:

**Lemma.**  $\text{unwind } \Delta \wedge \text{reach}_{\neg T} s \wedge \text{reach } s_1 \wedge \Delta s vl s_1 vl_1 \wedge tr \in \text{Valid}_s \wedge \text{never } T tr \wedge \forall tr = vl \rightarrow (\exists tr_1. tr_1 \in \text{Valid}_{s_1} \wedge O tr_1 = O tr \wedge \forall tr_1 = vl_1)$

*Proof idea.* By induction on length  $tr + \text{length } vl_1$  (as discussed above about progress).

**Unwinding Theorem.** If  $\text{unwind } \Delta$  and  $\forall vl vl_1. B vl vl_1 \rightarrow \Delta \text{istate } vl \text{istate } vl_1$ , then the system is BD-secure.

*Proof idea.* The formula (\*) follows from the lemma by taking  $s_1 = s$ .

According to the theorem, BD unwinding is a *sound proof method for BD security*: to check BD security it suffices to define a relation  $\Delta$  and prove that it coincides with  $B$  on the initial state and that it is a BD unwinding.

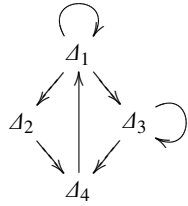


Fig. 4: A network of unwinding components

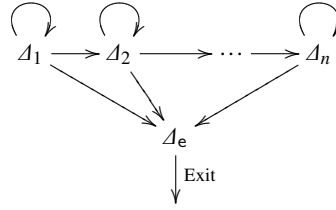


Fig. 5: A linear network with exit

## 5.2 Compositional Reasoning

To keep each reasoning step manageable, it is convenient to allow decomposing the single unwinding relation  $\Delta$  into relations  $\Delta_1, \dots, \Delta_n$ . Unlike  $\Delta$ , a component  $\Delta_i$  may unwind not only to itself but to any combination of  $\Delta_j$ 's. Technically, we define the predicate `unwind_to` just like `unwind` but taking two arguments instead of one: a first relation and a second relation to which the first one unwinds. We replace the single requirement `unwind  $\Delta$`  with a set of requirements `unwind_to  $\Delta_i$  (disj (next  $\Delta_i$ ))`, where `next  $\Delta_i$`  is a chosen subset of  $\{\Delta_1, \dots, \Delta_n\}$  and `disj` takes the disjunction of a set of predicates. This enables a form of sound compositional reasoning: if we verify a condition as above for each component  $\Delta_i$ , we obtain an overall unwinding relation `disj  $\{\Delta_1, \dots, \Delta_n\}$` .

The network of components can form any directed graph—Fig. 4 shows an example. However, our unwinding proofs will be phase-directed, and hence the following linear network will suffice (Fig. 5): each  $\Delta_i$  unwinds either to itself, or to  $\Delta_{i+1}$  (if  $i \neq n$ ), or to an exit component  $\Delta_e$  that invariably chooses the “exit” unwinding condition. For the first component,  $\Delta_1$ , we need to verify that it extends  $B$  on the initial state.

## 5.3 Verification of Concrete Instances

We have verified all the BD security instances listed in Fig. 2. For each of them we defined a suitable chain of unwinding components  $\Delta_i$  as in Fig. 5.

Recall from the definition of BD security that one needs to construct an alternative trace  $tr_1$  (which produces the value sequence  $vl_1$ ) from the original trace  $tr$  (which produces the value sequence  $vl$ ). A chain of  $\Delta_i$ 's witnesses the strategy for such a construction, although it does not record the whole traces  $tr_1$  and  $tr$  but only the states they have reached so far,  $s$  and  $s_1$ .

The separation between  $\Delta_i$ 's is guided by milestones in the journey of  $tr$  and  $tr_1$ , such as: a paper's registration to a conference, conference phases, the registration of a relevant agent like a chair, a non-conflicted PC member, or a reviewer. E.g., Fig. 6 shows the unwinding components in the proof of  $PAP_2$ , where  $B \text{ } vl \text{ } vl_1$  is the declassification bound ( $vl \neq [] \neq vl_1 \wedge \text{last } vl = \text{last } vl_1$ ) and the changes from  $\Delta_i$  to  $\Delta_{i+1}$  are emphasized.

Each property has one or more critical phases, the only phases when  $vl$  and  $vl_1$  can be produced. E.g., for  $PAP_2$ , paper uploading is only available in Submission (while for REV, there is an update action in Reviewing, and an u-update one in Discussion). Until those phases,  $tr_1$  proceeds synchronously to  $tr$  taking the same actions—consequently, the states  $s$  and  $s_1$  are equal in  $\Delta_1$ .

In the critical phases, the traces  $tr$  and  $tr_1$  will diverge, due to the need of producing different (but  $B$ -related) value sequences. As a result, the equality between  $s$  and  $s_1$  is

$\mathcal{A}_1 s vl s_1 vl_1$	$\neg (\exists cid. PID \in \text{paperIDs } s cid) \wedge s = s_1 \wedge B vl vl_1$
$\mathcal{A}_2 s vl s_1 vl_1$	$(\exists cid. PID \in \text{paperIDs } s cid \wedge \text{phase } s cid = \text{Submission}) \wedge s =_{PID} s_1 \wedge B vl vl_1$
$\mathcal{A}_3 s vl s_1 vl_1$	$(\exists cid. PID \in \text{paperIDs } s cid) \wedge s = s_1 \wedge vl = vl_1 = []$
$\mathcal{A}_e s vl s_1 vl_1$	$(\exists cid. PID \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Submission}) \wedge vl \neq []$

Fig. 6: The unwinding components for the proof of PAP<sub>2</sub>

replaced with the weaker relation of *equality everywhere except on certain components of the state*, e.g., the content of a given paper (written  $=_{PID}$  for PAP<sub>2</sub>), or of a given review, or of the previous versions of a given review, and so on.

At the end of the critical phases,  $tr_1$  will usually need to resynchronize with  $tr$  and hereafter proceed with identical actions. Consequently,  $s$  and  $s_1$  will become connected by a stronger “equality everywhere except” relation or even plain equality again. The smooth transition between consecutive components  $\mathcal{A}_i$  and  $\mathcal{A}_{i+1}$  that impose different state equalities is ensured by a suitable INDEPENDENT-ACTION/REACTION strategy. For PAP<sub>2</sub>, such a strategy for transitioning from  $\mathcal{A}_2$  to  $\mathcal{A}_3$  (with emptying  $vl$  and  $vl_1$  at the same time) is the following: by INDEPENDENT ACTION,  $tr_1$  will produce all values in  $vl_1$  save for the last one, which will be produced by REACTION in sync with  $tr$  when  $tr$  reaches the last value in  $vl$ ; this is possible since  $B$  guarantees  $\text{last } vl = \text{last } vl_1$ .

The exit component  $\mathcal{A}_e$  witnesses situations  $(s, vl)$  not producible from any system trace  $tr$  in order to exclude them via Exit. For PAP<sub>2</sub>, such a situation is the paper’s conference phase exceeding Submission with values  $vl$  still to be produced.  $\mathcal{A}_e$  is reached from  $\mathcal{A}_2$  when a change-phase action occurs.

Several safety properties are needed in the unwinding proofs. For PAP<sub>2</sub>, we use that there is at most one conference to which a paper can be registered—this ensures that no value can be produced (i.e.,  $\varphi(\text{head } vl)$  does not hold) from within  $\mathcal{A}_1$  or  $\mathcal{A}_2$ , since no paper upload is possible without prior registration.

## 6 Conclusion

Most of the information-flow security models proposed by theoreticians have not been confronted with the complexity of a realistic application, and therefore fail to address, or abstract away from, important aspects concerning the conditions for information release or restraint. We provided a case study of verified information-flow security, where we approached the problem bottom-up: we first formalized the functionality of a realistic system in such a way that we could automatically synthesize an implementation; then we identified, formulated and verified confidentiality properties. This experience, together with a screening of the theoretical literature, led to the design of a flexible verification infrastructure for restricted information flow in IO automata.

**Acknowledgement.** Tobias Nipkow made this collaboration possible. Several people made helpful comments on drafts of this paper and/or indicated related work: Jasmin Blanchette, Manuel Eberl, Lars Hupel, Fabian Immler, Steffen Lortz, Giuliano Losa, Tobias Nipkow, Benedikt Nordhoff, Martin Ochoa, Markus Rabe, and Dmitriy Traytel. The research was supported by the Deutsche Forschungsgemeinschaft (DFG) project Security Type Systems and Deduction (grant Ni 491/13-2), part of the program Reliably Secure Software Systems (RS<sup>3</sup>). The authors are listed in alphabetical order.

## References

1. Jif: Java + information flow, 2014. <http://www.cs.cornell.edu/jif>.
2. The Scala Programming Language, 2014. <http://www.scala-lang.org>.
3. M. Arapinis, S. Bursuc, and M. Ryan. Privacy supporting cloud computing: Confichair, a case study. In *POST*, pp. 89–108, 2012.
4. E. D. Bell and J. L. La Padula. Secure computer system: Unified exposition and multics interpretation. Technical Report MTR-2997, MITRE, Bedford, MA.
5. B. Blanchet, M. Abadi, and C. Fournet. Automated verification of selected equivalences for security protocols. In *LICS*, pp. 331–340, 2005.
6. E. S. Cohen. Information transmission in computational systems. In *SOSP*, pp. 133–139, 1977.
7. R. Dimitrova, B. Finkbeiner, M. Kovács, M. N. Rabe, and H. Seidl. Model checking information flow in reactive systems. In *VMCAI*, pp. 169–185, 2012.
8. The EasyChair conference system, 2014. <http://easychair.org>.
9. The HotCRP conference management system, 2014. <http://read.seas.harvard.edu/~kohler/hotcrp>.
10. B. Finkbeiner, M. N. Rabe, and C. Sánchez. A temporal logic for hyperproperties. *CoRR*, abs/1306.6657, 2013.
11. R. Focardi and R. Gorrieri. Classification of security properties (part i: Information flow). In *FOSAD*, pp. 331–396, 2000.
12. J. A. Goguen and J. Meseguer. Security policies and security models. In *IEEE Symposium on Security and Privacy*, pp. 11–20, 1982.
13. J. A. Goguen and J. Meseguer. Unwinding and inference control. In *IEEE Symposium on Security and Privacy*, pp. 75–87, 1984.
14. D. Gollmann. *Computer Security*. Wiley, 2nd ed., 2005.
15. F. Haftmann. *Code Generation from Specifications in Higher-Order Logic*. Ph.D. thesis, Technische Universität München, 2009.
16. F. Haftmann and T. Nipkow. Code generation via higher-order rewrite systems. In *FLOPS 2010*, pp. 103–117, 2010.
17. J. Y. Halpern and K. R. O’Neill. Secrecy in multiagent systems. *ACM Trans. Inf. Syst. Secur.*, 12(1), 2008.
18. IEEE Symposium on Security and Privacy. Email notification, 2012.
19. S. Kanav, P. Lammich, and A. Popescu. Supplementary material associated with this paper. [http://www21.in.tum.de/~popescua/cav2014\\_suppl.zip](http://www21.in.tum.de/~popescua/cav2014_suppl.zip), 2014.
20. B. W. Lampson. Protection. *Operating Systems Review*, 8(1):18–24, 1974.
21. H. Mantel. Information flow control and applications - bridging a gap. In *FME*, pp. 153–172, 2001.
22. H. Mantel. *A Uniform Framework for the Formal Specification and Verification of Information Flow Security*. PhD thesis, University of Saarbr 2003.
23. H. Mantel. Information flow and noninterference. In *Encyclopedia of Cryptography and Security (2nd Ed.)*, pp. 605–607. 2011.
24. D. McCullough. Specifications for multi-level security and a hook-up property. In *IEEE Symposium on Security and Privacy*, 1987.
25. J. McLean. A general theory of composition for trace sets closed under selective interleaving functions. In *In Proc. IEEE Symposium on Security and Privacy*, pp. 79–93, 1994.
26. J. McLean. Security models. In *Encyclopedia of Software Engineering*, 1994.
27. T. C. Murray, D. Maticuk, M. Brassil, P. Gammie, and G. Klein. Noninterference for operating system kernels. In *CPP*, pp. 126–142, 2012.



28. T. Nipkow, L. C. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*, vol. 2283 of *LNCS*. Springer, 2002.
29. C. O'Halloran. A calculus of information flow. In *ESORICS*, pp. 147–159, 1990.
30. G. J. Popek and D. A. Farber. A model for verification of data security in operating systems. *Commun. ACM*, 21(9):737–749, 1978.
31. Y. M. Ronald Fagin, Joseph Y. Halpern and M. Vardi. *Reasoning about knowledge*. MIT Press, 2003.
32. J. Rushby. Noninterference, transitivity, and channel-control security policies. Tech. report, dec 1992.
33. P. Y. A. Ryan. Mathematical models of computer security. In *FOSAD*, pp. 1–62, 2000.
34. A. Sabelfeld and A. C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.
35. A. Sabelfeld and D. Sands. Declassification: Dimensions and principles. *Journal of Computer Security*, 17(5):517–548, 2009.
36. D. Sutherland. A model of information. In *9th National Security Conference*, pp. 175–183, 1986.

# Appendix

This appendix gives more details on the formal specification of the system (§A), on the unwinding proof method for BD security (§B), on the formulation of confidentiality properties as instances of BD security (§C) and their unwinding proofs (§D), as well as on some safety (§E) and “forensic” (§F) properties we have proved as complements or auxiliaries to BD security.

## A More Details on the Specification

Here we present selected concrete aspects of the formalization for the reader who does not want to inspect our formal Isabelle scripts [19].

### A.1 The Roles

Each user is assigned a set of roles for each conference.

```
datatype Role = PC | Chair | Aut PaperID | Rev PaperID Nat
```

PC The user is member of the program committee

Chair The user is a chair of the conference

Aut *pid* The user is auther of the paper with id *pid*

Rev *pid n* The user is the *n*th reviewer of the paper with id *pid*

### A.2 Initial State

In the initial state of the system, there exists only a single user, which is the superuser (voronkov), and there are no conferences yet.

```
istate =
  confIDs = []
  userIDs = ["voronkov"]
  user = ( $\lambda uid.$  emptyUser)
  paperIDs = ( $\lambda cid.$  [])
  pref = ( $\lambda uid pid.$  NoPref)
  news = ( $\lambda cid.$  [])
  conf = ( $\lambda cid.$  emptyConf)
  pass = ( $\lambda uid.$  emptyPass)
  roles = ( $\lambda cid uid.$  [])
  paper = ( $\lambda pid.$  emptyPaper)
  voronkov = "voronkov"
  phase = ( $\lambda cid.$  noPh)
```

### A.3 Outputs

The output contains some generic output types, like outOK for a successful action with no other output and outErr for a failed action. Moreover, outputs for various datatypes are defined:

```
datatype Out =
  outOK | outErr | outBool Bool | outSTRP String  $\times$  String | outSTRL List(String) |
  outCONF String  $\times$  String  $\times$  List(Role)  $\times$  Phase | outPREF Pref | outCON Paper_Content |
  outNREV Nat  $\times$  Review | outREVL List(Review) | outRREVL List(UserID  $\times$  Review) |
  outDEC Decision | outDECL List(Decision) | outCIDL List(ConfID) | outUIDL List(UserID) |
  outPIDL List(PaperID) | outSTRPAL String  $\times$  String  $\times$  List(UserID)
```

## A.4 Actions

The actions are defined as an Isabelle datatype that distinguish between five categories of actions: creation, update, undestructive update, reading, and listing.

```
datatype Act = Cact cAct | Uact uAct | UUact uuAct | Ract rAct | Lact lAct
```

Each category is further refined within its own datatype, based on the object of the action: user, paper, PC member, author, conflict, review, phase, paper title and abstract, paper content, preference, news, discussion, (the changing list of) decisions, final decision, etc.

```
datatype cAct =  
  cUser UserID Pass String String | cUser ConfID UserID Pass String String  
  | cPC ConfID UserID Pass UserID | cChair ConfID UserID Pass UserID  
  | cPaper ConfID UserID Pass PaperID String String  
  | cAuthor ConfID UserID Pass PaperID UserID  
  | cConflict ConfID UserID Pass PaperID UserID  
  | cReview ConfID UserID Pass PaperID UserID
```

```
datatype uAct =  
  uUser UserID Pass Pass String String | uConfA ConfID UserID Pass  
  | uPhase ConfID UserID Pass Phase  
  | uPaperTA ConfID UserID Pass PaperID StringString  
  | uPaperC ConfID UserID Pass PaperID Paper_Content  
  | uPref ConfID UserID Pass PaperID Pref  
  | uReview ConfID UserID Pass PaperID Nat Review_Content
```

```
datatype uuAct =  
  uuNews ConfID UserID Pass String | uuDis ConfID UserID Pass PaperID String  
  | uuReview ConfID UserID Pass PaperID Nat Review_Content  
  | uuDec ConfID UserID Pass PaperID Decision
```

```
datatype rAct =  
  rAmIVoronkov UserID Pass | rUser UserID Pass UserID  
  | rConf ConfID UserID Pass | rNews ConfID UserID Pass  
  | rPaperNIA ConfID UserID Pass PaperID | rPaperC ConfID UserID Pass PaperID  
  | rPref ConfID UserID Pass PaperID | rMyReview ConfID UserID Pass PaperID  
  | rReviews ConfID UserID Pass PaperID | rDecs ConfID UserID Pass PaperID  
  | rDis ConfID UserID Pass PaperID | rFinalReviews ConfID UserID Pass PaperID  
  | rFinalDec ConfID UserID Pass PaperID  
  | rPrefOfPC ConfID UserID Pass PaperID UserID
```

```
datatype lAct =  
  lConfsUserID Pass | lAConfs UserID Pass | lSConfs UserID Pass  
  | lMyConfs UserID Pass | lAllUsers UserID Pass  
  | lAllPapers UserID Pass | lPC ConfID UserID Pass  
  | lChair ConfID UserID Pass | lPapers ConfID UserID Pass  
  | lMyPapers ConfID UserID Pass | lMyAssignedPapers ConfID UserID Pass  
  | lAssignedReviewers ConfID UserID Pass PaperID
```

## A.5 Step Function

Next we illustrate the definition of the step function by zooming into one of its subcases.

```

step  $s a \equiv$ 
case  $a$  of Cact  $ca \Rightarrow$  case  $ca$  of
  cAuthor  $cid uid pw pid uid' \Rightarrow$ 
  if e_createAuthor  $s cid uid pw pid uid'$ 
    then (outOK, createAuthor  $s cid uid pw pid uid'$ )
    else (outErr,  $s$ )
  | cUser  $cid uid pw name abs \Rightarrow \dots$ 
  ...
  | Uact  $ua \Rightarrow \dots$ 
  | UUact  $uua \Rightarrow \dots$ 
  | Ract  $ra \Rightarrow \dots$ 
  | Lact  $la \Rightarrow \dots$ 

```

Above, we only showed one subcase of the creation-action case in full. The semantics of each type of action (e.g., `cAuthor`, which is itself a subtype of creation actions) has an associated test for enabledness (e.g., `e_createAuthor`) and effect (e.g., `createAuthor`). The effect is only applied if the action is enabled; otherwise an error output is issued. The enabledness test checks if the user is allowed to perform the requested action: if his password matches the user ID, if the conference phase is appropriate, if he has an appropriate role, etc. For example, with the action `cAuthor  $s cid uid pw pid uid'$` , the user  $uid$  attempts to create (i.e., add) an author  $uid'$  to an existing paper  $pid$ , as well as a conflict in the system database between the author and the paper. Its enabledness test checks the following:

- the invoked IDs exist in the system;
- $p$  is the correct password of  $uid$ ;
- $cid$  is in the submission phase;
- $uid$  is an author of  $pid$ , and  $pid$  is associated to the conference  $cid$ ;
- $uid'$  is different from  $uid$ .

In our formalization, this reads as:

```

e_createAuthor  $s cid uid pw pid uid' \equiv$ 
  IDsOK  $s [cid] [uid, uid'] [pid] \wedge$  pass  $s uid = p \wedge$ 
  phase  $s cid =$  Submission  $\wedge$  isAut  $s uid pid \wedge uid \neq uid'$ 

```

```

createAuthor  $scid uid pw pid uid' \equiv$ 
  let  $rls =$  roles  $s cid uid'$  in
   $s$  (roles := fun_upd2 (roles  $s$ )  $cid uid'$  (insert (Aut  $pid$ )  $rls$ ),
    pref := fun_upd2(pref  $s$ )  $uid' pid$  Conflict)

```

## B More Details on the Unwinding Theorems

Here we show the formal definitions of the unwinding predicates and the statements of the unwinding theorem variations targeting compositional reasoning.

### B.1 Formal Definition of the Unwinding Predicates

Below are the formal definitions of iaction and reaction, including those of the reaction components ignore and match:

- iaction  $\Delta s \text{ vl } s_1 \text{ vl}_1 \equiv$   
 $\exists a_1 o_1 s'_1 \text{ vl}'_1. \text{ let } \text{trn}_1 = (s_1, a_1, o_1, s'_1) \text{ in}$   
 $\text{validTrans } \text{trn}_1 \wedge \text{consume } \text{trn}_1 \text{ vl}_1 \text{ vl}'_1 \wedge \varphi \text{trn}_1 \wedge \neg \gamma \text{trn}_1 \wedge \Delta s \text{ vl } s'_1 \text{ vl}'_1$
- reaction  $\Delta s \text{ vl } s_1 \text{ vl}_1 \equiv \forall a o s'. \text{ let } \text{trn} = (s, a, o, s') \text{ in}$   
 $\text{validTrans } \text{trn} \wedge \neg \top \text{trn} \wedge \text{consume } \text{trn} \text{ vl } \text{vl}' \rightarrow$   
 $\text{match } \Delta s s_1 \text{ vl}_1 a o s' \text{ vl}' \vee \text{ignore } \Delta s s_1 \text{ vl}_1 a o s' \text{ vl}'$   
 where:
  - $\text{ignore } \Delta s s_1 \text{ vl}_1 a o s' \text{ vl}' \equiv \neg \gamma (s, a, o, s') \wedge \Delta s' \text{ vl}' s_1 \text{ vl}_1$
  - $\text{match } \Delta s s_1 \text{ vl}_1 a o s' \text{ vl}' \equiv$   
 $\exists a_1 o_1 s'_1 \text{ vl}'_1. \text{ let } \text{trn} = (s, a, o, s') \text{ and } \text{trn}_1 = (s_1, a_1, o_1, s'_1) \text{ in}$   
 $\text{validTrans } \text{trn}_1 \wedge \text{consume } \text{trn}_1 \text{ vl}_1 \text{ vl}'_1 \wedge (\gamma \text{trn} \leftrightarrow \gamma \text{trn}_1) \wedge$   
 $(\gamma \text{trn} \rightarrow g \text{trn} = g \text{trn}_1) \wedge \Delta s' \text{ vl}' s'_1 \text{ vl}'_1$

### B.2 Verifiability of the exit Condition in a Trace-Free Manner

One can prove  $\text{exit } s \text{ v}$  in a trace-free manner, by exhibiting an invariant  $K : \text{State} \rightarrow \text{Bool}$  and proving that it holds for  $s$ :

**Lemma 1** Assume that the following hold for all valid transitions  $\text{trn} = (s, a, o, s')$  such that  $K s$  holds:

- $\varphi \text{trn} \rightarrow f \text{trn} \neq v$
- $K s'$

Then  $\forall s. K s \rightarrow \text{exit } s \text{ v}$

Intuitively, the invariant  $K$  ensures that the value  $v$  can never be produced. We use this lemma in all our unwinding proofs for the  $\Delta_e$  components.

### B.3 Unwinding Components

The predicate  $\text{unwind\_to}$  is a generalization of  $\text{unwind}$  allowing a relation  $\Delta$  to unwind not only to itself, but to the disjunction of any set of relations  $\Delta s = \{\Delta_1, \dots, \Delta_n\}$ :

$$\begin{aligned} \text{unwind\_to } \Delta \Delta s \equiv & \forall s \text{ vl } s_1 \text{ vl}_1. \text{reach } \neg \top s \wedge \text{reach } s_1 \wedge \Delta s \text{ vl } s_1 \text{ vl}_1 \rightarrow \\ & ((\text{vl} \neq [] \vee \text{vl}_1 = []) \wedge \text{reaction } (\text{disj } \Delta s) s s \text{ vl } s_1 \text{ vl}_1) \vee \\ & \text{iaction } (\text{disj } \Delta s) s s \text{ vl } s_1 \text{ vl}_1 \vee \\ & (\text{vl} \neq [] \wedge \text{exit } s (\text{head } \text{vl})) \end{aligned}$$

Using this we formulate a compositional version of the unwinding theorem:

**Compositional Unwinding Theorem.** For each  $\mathcal{A} \in \mathcal{A}s$ , let  $\text{next}_{\mathcal{A}} \subseteq \mathcal{A}s$  be a (possibly empty) “continuation” of  $\mathcal{A}$ , and let  $\mathcal{A}_0 \in \mathcal{A}s$  be a chosen “initial” relation. Assume the following hold:

- $\forall vl vl_1. \text{B } vl vl_1 \rightarrow \mathcal{A}_0 \text{ istate } vl \text{ istate } vl_1$
- $\forall \mathcal{A} \in \mathcal{A}s. \text{unwind\_to } \mathcal{A} \text{ next}_{\mathcal{A}}$

Then the (whole) system is BD-secure.

*Proof idea.* One can show that  $\text{unwind}(\text{disj } \mathcal{A}s)$  holds and use the original unwinding theorem.

Since in our proofs we shall only need a linear “next” network, for convenience we prove a linear variation of the above, where, assuming  $\mathcal{A}s = \{\mathcal{A}_1, \dots, \mathcal{A}_n, \mathcal{A}_e\}$ , we allow  $\mathcal{A}_i$  to only unwind to itself, to  $\mathcal{A}_{i+1}$  (if  $i < n$ ) or to an exit component  $\mathcal{A}_e$ . In addition, we employ the predicate  $\text{unwind\_cont}$  to restrict the unwinding of  $\mathcal{A}_i$  to proper continuations (i.e., no exits) and the predicate  $\text{unwind\_exit}$  to restrict the unwinding of  $\mathcal{A}_e$  to exits (as depicted in the paper’s Fig. 5):

$$\begin{aligned} \text{unwind\_cont } \mathcal{A} \mathcal{A}s &\equiv \forall s vl s_1 vl_1. \text{reach } \neg_{\top} s \wedge \text{reach } s_1 \wedge \mathcal{A} s vl s_1 vl_1 \rightarrow \\ &\quad ((vl \neq [] \vee vl_1 = []) \wedge \text{reaction } (\text{disj } \mathcal{A}s) s s vl s_1 vl_1) \vee \\ &\quad \text{iaction } (\text{disj } \mathcal{A}s) s s vl s_1 vl_1 \\ \text{unwind\_exit } \mathcal{A} &\equiv \forall s vl s_1 vl_1. \text{reach } \neg_{\top} s \wedge \text{reach } s_1 \wedge \mathcal{A} s vl s_1 vl_1 \rightarrow \\ &\quad vl \neq [] \wedge \text{exit } s (\text{head } vl) \end{aligned}$$

**Sequential Unwinding Theorem.** For each  $\mathcal{A} \in \mathcal{A}s$ , let  $\text{next}_{\mathcal{A}} \subseteq \mathcal{A}s$  be a (possibly empty) “continuation” of  $\mathcal{A}$ , and let  $\mathcal{A}_0 \in \mathcal{A}s$  be a chosen “initial” relation. Assume the following hold:

- $\forall vl vl_1. \text{B } vl vl_1 \rightarrow \mathcal{A}_0 \text{ istate } vl \text{ istate } vl_1$
- $\forall i \in \{1, \dots, n-1\}. \text{unwind\_cont } \mathcal{A}_i \{\mathcal{A}_i, \mathcal{A}_{i+1}, \mathcal{A}_e\}$
- $\text{unwind\_cont } \mathcal{A}_n \{\mathcal{A}_n, \mathcal{A}_e\}$
- $\text{unwind\_exit } \mathcal{A}_e$

Then the (whole) system is BD-secure.

*Proof idea.* Immediate from the compositional unwinding theorem.

Employing the sequential unwinding theorem in our proofs of confidentiality properties had the benefit of allowing (and encouraging!) separation of concerns: the  $\mathcal{A}_i$ ’s and the transitions between them constitute the main sequential flow of the phase-directed proof, and  $\mathcal{A}_e$  takes the “unnatural” corner cases out of our way.

Source	Observation Filter $\gamma(s, a, o, s')$	Observation Producer $g(s, a, o, s')$	Value Filter $\varphi(s, a, o, s')$	Value Producer $f(s, a, o, s')$	Declassification Trigger $T(s, a, o, s')$	Declassification Bound $B_{v/l} v/l$
Paper Content UIDs, PID	$\text{userOf } a \in \text{UIDs}$	$(a, o)$	$o = \text{outOK} \wedge$ $(\exists cid uid pw name abs pct.$ $a = \text{Uact} (\text{uPaperC } cid uid pw PID pct))$	$pct$	Paper Authorship $\exists uid \in \text{UIDs}. \exists cid. \text{isAut } s' uid PID$  Paper Authorship or PC Membership <sup>B</sup> $\exists uid \in \text{UIDs}. \exists cid. \text{isAut } s' uid PID$ $\vee$ $\text{isPC } s' uid cid \wedge \text{phase } s' cid \geq \text{Bidding}$	Last Uploaded Version $v/l \neq \square \neq v/l_1 \wedge$ $\text{last } v/l = \text{last } v/l_1$  Absence of Any Upload $v/l \neq \square$
Review UIDs, PID, N	$\text{userOf } a \in \text{UIDs}$	$(a, o)$	$o = \text{outOK} \wedge$ $(\exists cid uid pw rct.$ $a = \text{Uact} (\text{uReview } cid uid pw PID N rct)$ $\vee$ $a = \text{UuAct} (\text{uuReview } cid uid pw PID N rct))$	$(\text{phase } s cid, rct)$	Review Authorship $\exists uid \in \text{UIDs}. \exists cid. \text{isRevNth } s' uid PID N$  Review Authorship or Non-Conflict PC Membership <sup>D</sup> $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isRevNth } s' uid PID N$ $\vee$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid \neq \text{Conflict}$ Review Authorship or Non-Conflict PC Membership <sup>D</sup> or Paper Authorship <sup>N</sup> $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isRevNth } s' uid PID N$ $\vee$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid PID \neq \text{Conflict} \wedge$ $\text{phase } s' \geq \text{Discussion}$ $\vee$ $\text{isAut } s' uid PID \wedge \text{phase } s' \geq \text{Notification}$	Last Edited Version $v/l \neq \square \neq v/l_1 \wedge$ $\text{last } v/l = \text{last } v/l_1$  Absence of any Edit $v/l \neq \square$
Discussion UIDs, PID	$\text{userOf } a \in \text{UIDs}$	$(a, o)$	$o = \text{outOK} \wedge$ $(\exists cid uid pw com.$ $a = \text{UuAct} (\text{uuDis } cid uid pw PID com))$	$com$	Non-Conflict PC Membership $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid PID \neq \text{Conflict}$ Non-Conflict PC Membership $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid PID \neq \text{Conflict}$ Non-Conflict PC Membership or PC Membership <sup>N</sup> or Paper Authorship <sup>N</sup> $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{phase } s' \geq \text{Notification}$	Absence of Any Edit
Decision UIDs, PID	$\text{userOf } a \in \text{UIDs}$	$(a, o)$	$o = \text{outOK} \wedge$ $(\exists cid uid pw dec.$ $a = \text{UuAct} (\text{uuDec } cid uid pw PID dec))$	$dec$	Non-Conflict PC Membership $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid PID \neq \text{Conflict}$ Non-Conflict PC Membership or PC Membership <sup>N</sup> or Paper Authorship <sup>N</sup> $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{phase } s' \geq \text{Notification}$	Last Edited Version  Absence of Any Edit
Reviewer Assignment UIDs, PID	$\text{userOf } a \in \text{UIDs}$	$(a, o)$	$o = \text{outOK} \wedge$ $(\exists cid uid pw uid.$ $a = \text{Cact} (\text{cReview } cid uid pw PID uid))$	$f_{\text{REVA}} s' cid uid'$	isAut $s' uid PID \wedge \text{phase } s' \geq \text{Notification}$ Non-Conflict PC Membership <sup>F</sup> $\exists uid \in \text{UIDs}. \exists cid.$ $\text{isPC } s' cid uid PID \wedge \text{pref } s' uid PID \neq \text{Conflict} \wedge$ $\text{phase } s' \geq \text{Reviewing}$	Non-Conflict PC Membership $B_{\text{REVA}} v/l v/l_1$

Phase Stamps: B = Bidding, D = Discussion, N = Notification, F = Review

Fig. 7: The confidentiality properties, formally

## C More Details on the Confidentiality Properties Formalization

Here we indicate the formal definitions of the BD security parameters used for the statements in the paper’s Fig. 2.

Fig. 7 is an extension of Fig. 2 showing the formal instantiations of the observational infrastructure  $(\text{Val}, \varphi, f, \text{Obs}, \gamma, g)$  and declassification specification  $(B, T)$ . (The value and observation domains  $\text{Val}$  and  $\text{Obs}$  can be inferred from the expressions defining the functions.)

The properties are parametrized by the items shown below the sources:

- the observers/attackers: a set of users  $\text{UIDs}$
- the source: a paper ID  $\text{PID}$  and, for reviews, an additional number  $N$

$\text{PID}$  identifies a paper  $s$   $\text{PID}$  which, in turn, stores a paper content, a review list, a discussion, and a decision. Thus, in the cases of paper content, discussion, and decision,  $\text{PID}$  determines the data under scrutiny via the components of items of type  $\text{Paper}$  stored in the state.  $\text{PID}$  also determines this data for the reviewer assignment, but via the roles component of the state. On the the hand, for reviews, we need to further indicate a number  $N$  that choses one of the multiple reviews associated to paper  $s$   $\text{PID}$ .

In all cases, the observation filter simply checks that the user who performs the action,  $\text{userOf } a$ , is one of the observers, i.e., is in  $\text{UIDs}$ ; and the produced observation consists of the action and the output.

The value filter  $\varphi$  checks if the action of the given transition  $(s, a, o, s')$  has a certain form, e.g., is a paper or review update or a review creation, case in which  $f$  selects the relevant item,<sup>2</sup> e.g., the title, abstract and content of a paper. For reviews, on one occasion, besides the specific data consisting of the review content  $rct$ ,  $f$  also returns phase  $s$ , the phase of the conference—the latter is crucially used by the declassification bound for reviews, to express nuances such as “the last edited version *before discussion*.”

The declassification triggers and bounds are the ones from Fig. 2, but listed together with their formal definitions. We use the following general-purpose operators and predicates:

- $\text{fst}$  and  $\text{snd}$  return the first and second components of pairs  $(x, y)$
- $\text{Pair}$  takes  $x$ , then  $y$  (in a curries fashion), and returns the pair  $(x, y)$
- $\text{distinct } xl$  states that the list  $xl$  has no repetitions
- $\text{map } h \ xl$  returns the list obtained by applying the function  $h$  to all the elements of the list  $xl$
- $\text{conj } bl$  returns the conjunction of the list  $bl$  of booleans

The bound  $B_{\text{REV}} \ vl \ vl_1$  of the first review property is defined by decomposing  $vl$  and  $vl_1$  in the review edits during the reviewing phase and those during the discussion phase—the former should have the same last element and the latter should be identical:  $\exists ul \ ul_1 \ wl$ .

$$vl = (\text{map } (\text{Pair } \text{Reviewing}) \ ul) \cdot (\text{map } (\text{Pair } \text{Discussion}) \ wl) \wedge$$

<sup>2</sup> As a matter of variable scoping,  $f$  always refers to the items quantified by  $\varphi$ . E.g., for paper content, the definition of  $f$  refers to the item  $rct$  whose existence is asserted by  $\varphi$ —in all such cases, the referred items, if they exist, are also unique, hence  $f$  is well-defined.



$$vl_1 = (\text{map (Pair Reviewing)} ul_1) \cdot (\text{map (Pair Discussion)} wl) \wedge \\ ul \neq [] \neq ul_1 \wedge \text{last } ul = \text{last } ul_1$$

At the reviewer assignment property,  $f_{\text{REVA}} s' cid uid'$  returns a pair consisting of the ID of the assigned reviewer and a Boolean flag indicating whether or not he is a PC member having no conflict with the paper:  $(uid', \text{isPC } cid uid' \wedge \text{pref } s' uid' \text{ PID} \neq \text{Conflict})$ . The bound  $B_{\text{REVA}} vl vl_1$  states about  $vl_1$  that all its flags hold true (expressing the public knowledge that reviewers are non-conflict PC members) and that all its user IDs are distinct (expressing another, less important, public knowledge—that already assigned reviewers cannot be assigned again to the same paper PID):  $\text{conj} (\text{map snd } vl_1) \wedge \text{distinct} (\text{map fst } vl_1)$ .

## D Some Unwinding Relations Used in Proofs

We proved all the confidentiality properties from Fig. 7 using the sequential unwinding theorem. In the paper's Fig. 6 we showed the unwinding components employed in the proof of the first paper content property from Fig. 7 (PAP<sub>2</sub>). Next we show the unwinding components of a few other properties—each time B refers to the specific declassification bound from the corresponding row in Fig. 7.

### The Second Paper Content Property (PAP<sub>1</sub>)

The state-equality relation  $=_{\text{PID}}$  is the same as for the first paper content property: equality everywhere except on the content of the paper PID.

$A_1 s vl s_1 vl_1$	$\neg (\exists cid. \text{PID} \in \text{paperIDs } s cid) \wedge s = s_1 \wedge vl \neq []$
$A_2 s vl s_1 vl_1$	$(\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid = \text{Submission}) \wedge s =_{\text{PID}} s_1$
$A_3 s vl s_1 vl_1$	$(\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Submission}) \wedge s =_{\text{PID}} s_1 \wedge vl = vl_1 = []$
$A_e s vl s_1 vl_1$	$(\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Submission}) \wedge vl \neq []$

### The First Review Property (REV)

The state-equality relation  $=_{\text{PID},N}$  expresses equality everywhere except on the N'th review of the paper PID.

$A_1 s vl s_1 vl_1$	$(\forall cid. \text{PID} \in \text{paperIDs } s cid \rightarrow \text{phase } s cid < \text{Reviewing}) \wedge s = s_1 \wedge B vl vl_1$
$A_2 s vl s_1 vl_1$	$(\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid = \text{Reviewing} \wedge \\ \neg (\exists uid. \text{isRevNth } s cid uid \text{ PID } N)) \wedge \\ s = s_1 \wedge B vl vl_1$
$A_3 s vl s_1 vl_1$	$(\exists cid uid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid = \text{Reviewing} \wedge \text{isRevNth } s uid \text{ PID } N) \wedge \\ s =_{\text{PID},N} s_1 \wedge B vl vl_1$
$A_4 s vl s_1 vl_1$	$(\exists cid uid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid \geq \text{Reviewing} \wedge \text{isRevNth } s uid \text{ PID } N) \wedge \\ s = s_1 \wedge (\exists wl. vl = vl_1 = \text{map (Pair Discussion)} wl)$
$A_e s vl s_1 vl_1$	$vl \neq [] \wedge \\ ( \\ \vee \\ (\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Submission}) \\ \vee \\ (\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Reviewing} \wedge \neg (\exists uid. \text{isRevNth } s uid \text{ PID } N)) \\ \vee \\ (\exists cid. \text{PID} \in \text{paperIDs } s cid \wedge \text{phase } s cid > \text{Reviewing} \wedge \text{fst} (\text{head } vl) = \text{Reviewing}) \\ )$

### The Second Review Property

Here we use two state-equality relations different from plain equality:

- the relation  $=_{PID,N}$  from the first review property
- the relation  $=_{PID,N}^2$  which expresses equality everywhere except on the *older* versions of the N'th review of paper PID (thus,  $s =_{PID,N}^2 s_1$  implies that the *current*, i.e., last version of the review (N, PID) is the same in  $s$  as in  $s_1$ )

Below, `reviewsOf` gives the list reviews of a paper and `!N` gives the N'th element of a list; hence, `reviewsOf (paper s PID) ! N` returns the N'th review of the paper PID, review which is itself a list of its different versions.

$\Delta_1 s \text{ vl } s_1 \text{ vl}_1$	the same as for the first review property (but with the specific B)
$\Delta_2 s \text{ vl } s_1 \text{ vl}_1$	the same as for the first review property (but with the specific B)
$\Delta_3 s \text{ vl } s_1 \text{ vl}_1$	$(\exists cid \text{ uid. PID} \in \text{paperIDs } s \text{ cid} \wedge \text{phase } s \text{ cid} \in \{\text{Reviewing, Discussion}\} \wedge$ $\text{isRevNth } s \text{ uid PID } N) \wedge$ $s =_{PID,N} s_1 \wedge B \text{ vl } vl_1$
$\Delta_4 s \text{ vl } s_1 \text{ vl}_1$	$(\exists cid \text{ uid. PID} \in \text{paperIDs } s \text{ cid} \wedge \text{phase } s \text{ cid} \geq \text{Reviewing} \wedge$ $\text{isRevNth } s \text{ uid PID } N) \wedge$ $s =_{PID,N}^2 s_1 \wedge vl = vl_1 = [] \wedge$ $\text{reviewsOf (paper } s \text{ PID) ! N} \neq [] \neq \text{reviewsOf (paper } s_1 \text{ PID) ! N}$
$\Delta_e s \text{ vl } s_1 \text{ vl}_1$	$vl \neq [] \wedge$ $($ $(\exists cid. PID \in \text{paperIDs } s \text{ cid} \wedge \text{phase } s \text{ cid} > \text{Submission})$ $\vee$ $(\exists cid. PID \in \text{paperIDs } s \text{ cid} \wedge \text{phase } s \text{ cid} > \text{Reviewing} \wedge \neg (\exists uid. \text{isRevNth } s \text{ uid PID } N))$ $\vee$ $(\exists cid. PID \in \text{paperIDs } s \text{ cid} \wedge \text{phase } s \text{ cid} > \text{Discussion})$ $)$

Above, it is interesting to see how different versions of state equality are maintained throughout the sequential unwinding components. In all cases we start with plain equality. In the first case (for `PAP2`), we switch to  $=_{PID}$  (where we stay until the end). In the second case (for `REV`), we switch to  $=_{PID,N}$ , and then back to plain equality. In the third case, we first switch to  $=_{PID,N}$ , and then to  $=_{PID,N}^2$ , which is stronger than  $=_{PID,N}$  but weaker than plain equality. Roughly speaking, the state-equality strength is turned up and down throughout the proofs as follows:

- it is turned down when the power of the observer increases,<sup>3</sup> e.g., in the discussion phase when reviews become more public
- it is turned up when the declassification bound allows, e.g., when consuming the last values assumed to be equal

<sup>3</sup> This increase is kept under control by the assumed failure of the declassification trigger.

## E Complementary Safety Properties

In the unwinding proofs, we needed about 20 safety properties, among which:

- A paper is never registered at two conferences
- An author always has conflict with his papers ( $\text{DIS}_1$ )
- A paper always has at least one author
- A user never reviews a paper with which he has conflict
- A user never gets to write more than one review for a given paper

## F Complementary Forensic Properties

Our proved confidentiality properties show upper bounds on information release that are valid unless/until some trigger  $T$  occurs, e.g., authorship or PC membership. While  $T$  can in principle depend on all four components of a transition  $(s, o, a, s')$ , our formalized instances only depend on  $s'$ , e.g.,  $\text{isAut } s' \text{ uid PID}$  or  $\text{isPC } s' \text{ uid cid}$ . Two questions arise.

First, why do we consider the target state  $s'$  and not the source state  $s$ ? The answer is that typically the two choices are equivalent but our choice seems a priori more natural: never  $T$  holding for a valid trace  $[(s_1, a_1, o_1, s_2), (s_2, a_2, o_2, s_3), \dots, (s_{n-1}, a_{n-1}, o_n, s_n)]$  then means that the corresponding state condition fails for  $s_2, \dots, s_n$  (thus also including the last state); and typically the condition also fails for the initial state  $s_1 = \text{istate}$ .

Second, why do we consider for  $T$  an action-based condition and not an event-based condition? E.g., instead of defining  $T$  not as *being* an author in the target state ( $\text{isAut } s' \text{ uid PID}$ ), why not define it as successfully taking the action of *becoming* an author ( $o$  being outOK and  $a$  being one of the two types of actions that makes  $\text{uid}$  an author for PID, namely:  $\text{uid}$  creating the paper PID or another author  $\text{uid}'$  assigning  $\text{uid}$  as a coauthor for PID). A first answer to this is that the two choices are equivalent (e.g.,  $\text{isAut } s' \text{ uid PID}$  occurs on a valid trace iff one of the aforementioned author-creation actions occur successfully), and our choice is easier to state since it does not require the bookkeeping of all the actions leading to a certain situation.

The second answer touches on a more fundamental concern triggered by the question, namely: we have proved that one does not learn such and such unless one acquires a certain role, e.g., authorship; but how can we know that only “legal” (or “intended”) users acquire that role, in particular, how can we know that an arbitrary user cannot acquire that role? The answer is: regardless on how we phrase the trigger, this concern has to be addressed separately. To this end, we track back in valid traces all possible chains of events that have led to certain roles—the result is a form of integrity properties that we call *forensics*.

For instance, we prove that, if  $\text{isAut } s \text{ uid PID}$  holds at the end of a valid trace<sup>4</sup>  $[\text{trn}_1, \text{trn}_2, \dots, \text{trn}_n]$  (i.e., such that  $s$  is the target state of  $\text{trn}_n$ ), then there exists a subsequence of transitions  $[\text{trn}_{i_1}, \dots, \text{trn}_{i_k}]$  such that  $i_k = n$  and the following hold:

- in  $\text{trn}_{i_1}$  some user  $\text{uid}_1$  successfully registers (i.e., creates) the paper PID

---

<sup>4</sup> Thanks to the prefix closure of the set of valid traces, by analyzing the forensics of a property holding *at the end* of a valid trace we also cover the case of that property holding *at some point* on a valid trace.

- for all  $j \in \{1, \dots, k-1\}$ , in  $tr_{i_j}$  some user  $uid_j$  successfully adds the user  $uid_{j+1}$  as a coauthor

In other words, authorship of PID can only occur by a chain of coauthor assignment started from the paper’s creator. We prove such forensic properties for all the trigger components used in our security properties:

1.  $\text{isChair } s \text{ cid } uid$  holds at the end of a valid trace  $tr$  iff there is a subsequence of  $tr$  starting with the creation of the conference  $cid$  by a user  $uid_1$ , and continuing with a chain of successful chair creation actions having subjects  $uid_1, \dots, uid_k$  in which  $uid_k = uid$  and each  $uid_i$  assigns  $uid_{i+1}$  as a chair
2.  $\text{isPC } s \text{ cid } uid$  holds at the end of a valid trace  $tr$  iff
  - either  $\text{isChair } s \text{ cid } uid$  holds at some point in  $tr$
  - or there exists  $uid' \neq uid$  such that  $\text{isChair } s \text{ cid } uid'$  holds at some point in  $tr$  and after that point there is a successful action with  $uid'$  assigning  $uid$  as a chair or as a committee member
3.  $\text{isRevNth } s \text{ uid } \text{PID } N$  holds at the end of a valid trace  $tr$  iff  $\text{isPC } s \text{ cid } uid$  holds at some point in  $tr$ , there exists  $uid'$  such that  $\text{isChair } s \text{ cid } uid'$  holds at some point in  $tr$ , and after both these points there is a successful review-creation action by  $uid'$  that assigns  $uid$  as the  $N$ 'th reviewer of PID.
4.  $\text{pref } s \text{ uid } \text{PID} \neq \text{Conflict}$  holds at the end of a valid trace  $tr$  iff  $tr$  does not contain
  - a successful creation by  $uid$  of the paper PID
  - or a successful assignment of  $uid$  as a coauthor of PID<sup>5</sup>
  - or a successful declaration of conflict by  $uid$  with PID which is not overwritten by any successful removal of the conflict by  $uid$
  - or a successful declaration, by some  $uid'$  for which  $\text{isAut } s \text{ uid}' \text{PID}$  holds, of a conflict between  $uid$  and PID which is not overwritten by any successful removal of the conflict by  $uid$ <sup>6</sup>
5. If  $ph > \text{noPh}$ , then  $\text{phase } s \text{ cid} = ph$  holds at the end of a valid trace  $tr$  iff there is a subsequence of  $tr$  of successful  $cid$ -change-phase actions starting from  $\text{noPh}$  and advancing sequentially through phases until  $ph$  such that the first action in the sequence is performed by the voronkov and the remaining ones are performed by  $cid$  chairs.

Note that, to obtain the full forensic story of  $\text{isPC}$ , one needs to consider that of  $\text{isChair}$ ; likewise, the full forensic story of  $\text{isRevNth}$  relies on those of  $\text{isChair}$  and  $\text{isPC}$ , and similarly for the other properties.

<sup>5</sup> Either of the last two actions automatically produces an irrevocable conflict of  $uid$  with PID.

<sup>6</sup> Allowing a PC member to remove the conflict declared by an author is a feature of our system aimed at preventing deadlocks caused by authors declaring too many conflicts.