# Part V

# Conclusion

# Chapter 11

# Conclusion

In this thesis, research is reported that is of both fundamental and practical value. Fundamentally, we show that a number of theoretical problems can be solved. Our solutions to these fundamental problems can be applied to resolve practical real-world problems.

In the following sections, we will summarize our results and their relevance. Directions for future research are also given.

## 11.1 Information Designators

Linking information which stems from various sources, information integration, is by itself a difficult problem. Even within 'relaxed' conditions where no data has to be kept secret, linking existing information sources consistently and reliably is hard. In practice, information integration does not enjoy relaxed conditions: the information is differently encoded, inconsistent and asynchronously updated. To cope with these conditions, current techniques for information integration take essentially the approach to expose as much information as deemed possibly helpful.

If information has to be kept secret, however, it seems that one faces the choice between either not integrating the information, or sacrificing confidentiality. Of course, if one has a secret, the best way to keep it secret is to tell it to nobody. But if one *has* to tell it to some people, one would certainly like those people to protect the secret. Current information integration technology cannot offer such guarantees.

In Chapter 7 we analyze this problem and identify two causes for this problem:

1. When information is integrated across databases, this is done by literally

copying information from one database to the other (or a process which is to some extent equivalent to this). We call this the *raw data* problem.

2. The ontologies of the databases that are to be integrated overlap. With the integration of the ontologies, insufficient care is taken to identify which contributor 'owns' (is responsible for) the information that overlaps.

We propose a solution to these problems which may seem very simple, but has in effect intricate, if not dramatic effects. The solution is to *never* replicate raw data, and to *always* refer to the original author ('owner') of the information. When phrased in a slogan, it becomes:

> Don't propagate, but link!

Instead of using raw data, *information designators* are used. An information designator is a pseudonym for a piece of information. Owners of information may use any number of pseudonyms for any piece of information. They can precisely control the extent to which others can use the information designators to reason about and recombine the information.

The information designator is a new concept, and is not an instantly applicable technique. But as a proof of concept, it demonstrates that linking information and protecting it against dissemination can go hand in hand.

The information designator approach needs extension in future research in the following ways:

- The information designator has to be fleshed out. Prototype production systems have to be built in order to reveal yet unknown intricacies of the approach. Basic understanding of precisely what bottlenecks will appear when the information designator systems are scaled up have to be identified and addressed.
- For practical 'real-world' deployment, an elegant way has to be found for incorporating information that is not stored using the information designator approach.

## 11.2 Knowledge Authentication

When one wants to compare two pieces of information, it may seem that it is necessary to have the two pieces of information available at hand. Consider the following problem "Comparing Information Without Leaking It" (CIWLI) [FNW96]:

> Two players want to test whether their respective secrets are the same, but they do not want the other player to learn the secret in case the secrets do not match.

In Chapter 8, we identify a number of variations of the problem, depending on the following properties:

- How untrustful and untrustworthy are the players? (i.e., what *adversary model* is appropriate?)
- How many *possible secrets* exist? (i.e., what is the domain size $|\Omega|$?)
- How many secrets need to be compared? Just one secret against one other secret (*1-to-1*), one secret against many other secrets (*1-to-many*), or many secrets against many secrets (*many-to-many*)?
- Does 'secret' mean that it is difficult to guess the string that represents secret (as with the string 'arkjjhhg bwr ufkng'), or does it mean that the player has attributed some stance to a commonly known string? (as with 'I voted for Pim Fortuyn'). We call the former *CIWLI without reference*, and the latter *CIWLI with reference*.

We argue that for CIWLI, the only appropriate adversary model is the *malicious adversary model*. In other adversary models, the adversary may infer the secret by using a feasible amount of computation power. Of the many protocols for CIWLI that exist in literature, only a few use the malicious adversary model.

We observe that for all protocols for CIWLI that exist in literature, the communication complexity (measured in bits) contains a factor $\ln|\Omega|$ or worse, which renders these protocols infeasible for comparing secrets from large domains (for example domains containing all possible files which are 16 megabyte large).

We present two new protocols, T-1 and T-2. Both protocols assume a malicious adversary, and solve CIWLI without reference. The term $\Omega$ does not occur in their communication complexity, which means that the protocols remain feasible when the domain of possible secrets $\Omega$ is huge.

The T-1 protocol, presented in Chapter 9, solves the 1-to-many case, with a communication complexity of only $O(1)$. We prove the T-1 protocol correct using an extended version of GNY logic.

The T-2 protocol, presented in Chapter 9, solves the many-to-many case, and can be seen as a parallel composition of the T-1 protocol. It has an average case communication complexity of

$$c_1 \cdot |KB_A \cup KB_B| + c_2 \cdot |KB_A \cap KB_B|$$

where $KB_A$ ($KB_B$) is the set of secrets possessed by player $A$ ($B$), the constant $c_1$ has an upper bound $c_1 < 3$ and the constant $c_2$ depends on chosen security parameters. This is particularly efficient, as every extra secret of $A$ or $B$ (which is not mutually shared) results in a communication increase of on average less than three bits. This complexity has not yet been formally derived, but experiments point strongly to the above relation.

In future research on knowledge authentication, the following issues need to be addressed:

- The theoretical framework of knowledge authentication may benefit from further development and consolidation.
- The T-1 protocol is currently only formally analyzed using our extended version of GNY logic. Appropriate would be additional analysis using other methodologies, such as strand spaces [THG98, THG99], spi calculus [AG99] or Datta-Derek-Mitchell-Pavlovic logic [DDMP03].
- The communication complexity of the T-2 protocol has been established experimentally. Though the results point in a very positive direction, they do not provide strong formal guarantees. For formal guarantees on the communication complexity, the communication complexity has to be formally derived.

## 11.3 Hash Functions and Authentication Logics

Both knowledge authentication and information designators use cryptographic hash functions in new, unprecedented ways. In common applications of cryptographic hash functions, the pre-image of a particular hash value is not considered to be secret. In our applications, the pre-image is often secret, while the corresponding hash value is not secret.

For our applications, we need cryptographic hash functions which satisfy an uncommon property, namely that they are *non-incremental*. A cryptographic hash function is non-incremental, if it is always necessary to have the full pre-image at hand to compute the hash value of this pre-image. None of the current standard cryptographic hash functions is non-incremental, but one can construct a non-incremental cryptographic hash function quite easily from any Merkle-Damgård cryptographic hash function, such as SHA-512.

At various places in the literature, it is assumed that the possession of a hash value counts as a proof of the corresponding pre-image, but this not the case. We show that BAN logic, a highly influential method for analyzing security protocols, relies on this false assumption. As a result, BAN logic is not 'sound': it is possible to derive false beliefs from true ones. As such, we demonstrate that properly modeling cryptographic primitives can be very difficult.

We extend GNY logic, a particular authentication logic, to properly model cryptographic hash functions. We prove correctness of the T-1 protocol using GNY logic.

The following issues with cryptographic hash functions require future research:

- The concept of *non-incrementality* for cryptographic hash functions is in need of a formal definition. Given that the formal definition of a cryptographic hash functions itself is already rather cumbersome (see Sections 3.2 and 3.3), the exercise of defining non-incrementality will probably be very difficult.

- The false assumption that possession of a hash value counts as a proof of the corresponding pre-image has trickled through some parts of the literature on computer security. Results that rely on this assumption may turn out to be incorrect. Literature in which the false assumption is made needs to be identified and the results in these publications need verification.

  In particular, the SET protocol [MV97a, MV97b, MV97c] and its analysis in BAN logic [AvdHdV01] need close re-examination.

## 11.4 Relevance to the Privacy Debate

We have demonstrated that for a number of problems, confidentiality (read: privacy protection) and availability (read: fighting terrorism) can go hand in hand. A number of techniques has been developed:

**Methods** The *information designator* is a solution which demonstrates that linking databases does not imply the abundant dissemination of sensitive information. On the contrary, if information designators are used, linking databases can enhance confidentiality.

**Protocols** *Knowledge authentication*, as exemplified in the T-1 and T-2 protocols, provides solutions demonstrating that comparing information for equality (a simple and elementary action) can be done without disclosing the information.

The methods and protocols demonstrate that for linking and comparing information, the information does not need to be disclosed. Thus, for linking or comparing secrets without disclosing them, there is no longer a need for a trusted third party, which is a gain. For application domains where it is not possible to find a trusted third party, our contributions offer solutions which were impossible before.

Our results warrant an existential statement: in relevant cases, it *is* possible to reconcile information exchange and confidentiality. Thus, the idea that there is an intrinsic *trade-off* between information exchange and confidentiality is wrong and misleading. This is relevant to the privacy debate, since the goal of the privacy debate is to find a balance in this supposed trade-off.

It may take a long time before the techniques presented in this thesis are applied to the issues of the privacy debate. For one thing, policy makers must understand the basic properties of our presented solutions and the possible future solutions. We do not cherish any illusions about this. The personal experience of the author is that policy makers often have an abominable knowledge of IT, information systems and epistemic logic[1], and that the knowledge of 'IT

[1] Epistemic logic is roughly the logic of knowledge about other people's knowledge. It analyzes constructs like 'I know that you know it, but you do not know that', which are essential if one wants to protect information against inappropriate dissemination. [FHMV95, MvdH95]

consultants' of privacy and related security issues is similarly depressing. In our opinion, when a policy maker or IT consultant states that it is necessary that privacy is sacrificed for some righteous task, this likely expresses either ignorance, unwillingness or insufficient priority.[2]

Not *all* privacy problems which are caused by anti-terrorism activities can be solved with the solutions offered in this thesis, only *some* of them. There is no reason to suppose that this thesis has exhausted all solutions for reconciliation. Future research by us and others may provide many more results which help to reconcile information exchange and confidentiality.

In general, security and cryptography research has mainly focused on facilitating a situation in which there are only *good guys* and *bad guys*. In this situation, the bad guys need to be avoided, and need to be kept ignorant while the good guys can be almost fully trusted. In practice, one considers only very few of the organizations one needs to interact with as unequivocally *good guys*. Thus, we need security solutions and cryptographic methods for interacting with *so-so guys*: those not intrinsically bad, but not to be trusted more than strictly necessary. Such solutions and methods are essential for addressing privacy issues.

[2] Of course, there is nothing wrong with a policy maker who assigns only a humble priority to the issue of privacy protection, when he clearly acknowledges this.