Linkability Estimation Between Subjects and Message Contents Using Formal Concepts

Stefan Berthold and Sebastian Clauß
TU Dresden
Fakultät Informatik
D-01062 Dresden, Germany
stefan.berthold@tu-dresden.de, sebastian.clauss@tu-dresden.de

ABSTRACT

In this paper, we examine how conclusions about linkability threats can be drawn by analyzing message contents and subject knowledge in arbitrary communication systems. At first, we define messages described by their contents as formal contexts. Then, we define subjects described by their knowledge as further formal contexts. Finally, we show that concept lattices, which are achieved by applying Formal Concept Analysis to the concatenation of these formal contexts, can be used in order to draw conclusions about correlations, and therefore linkability, between contents of messages and knowledge of subjects. The goal is to define formal specifications which can be utilized in privacy enhancing identity management systems in order to support users in the choice of data items which are to be disclosed to a communication partner.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*; E.1 [**Data Structures**]: Graphs and Networks; G.2.3 [**Discrete Mathematics**]: Applications

General Terms

Security

Keywords

Privacy-enhancing technology, Unlinkability, Anonymity, Identity management, Information management, Concept analysis, Data analysis

1. INTRODUCTION

Anonymity has various applications in several areas of daily life. Many of these interactions automatically take place in an anonymous manner. For other interactions, there is a need for laws and legal enforcement, typically in order to achieve more fairness, to enforce multilateral security, or to compensate the impact of modern information technology on social structures.

A recent example of such a law proposal on national level is a result of the anonymity board of the Swedish government ("anonymitetsutredning"). In early 2006, this board came to the conclusion that anonymous job applications would very well help to support ethnic diversity and equal rights, in general. The anonymity board proposed, therefore, a job application process driven by forms which must not contain data about gender, ethnic background, or age. According to the corresponding report [10], however, several more items of personal data are seen as problematic in order to obtain rational decisions which exclusively base on professional skills. These items include the applicant's surname, since it often yields hints about her ethnic background, and her first name, since it obviously provides information about the applicant's gender. In summary, anonymity with respect to a certain kind of information does not only depend on this particular information itself, but rather on all personal data items which may allow the adversary to draw conclusions about the information.

As a consequence, a serious drawback of an approach with anonymous job applications is the lack of ordinary return addresses, since the surname must not be revealed. Thus, such an approach would require to provide a different kind of return address which cannot be linked to a surname or other personal information which was not intended to be provided by the applicant herself. This could be a pseudonym, for instance, which at the time of usage has not been used in correlation with such personal information and should, indeed, not be used in such a correlation in future times. The anonymity board in Sweden proposed an application process where such pseudonyms are assigned by clerks at the staff departments. We, in fact, see a further improvement in a procedure where applicants do not even have to provide those data items which are going to be made anonymous by the staff later, anyway. Therefore, applicants need carefully to choose the data to be revealed or, more general, people who participate in highly interconnected systems need at least to be aware of their interactions and possible effects arising from them.

In our paper, we deal with links or correlations, respectively, between personal data and subjects (or pseudonyms, respectively) in arbitrary communication systems which use messages for data transactions. We propose an application of Formal Concept Analysis (FCA). A general introduction

DIM'07, November 2, 2007, Fairfax, Virginia, USA.
Copyright 2007 ACM 978-1-59593-889-3/07/0011 ...$5.00.

in FCA is given in [8, 14]. By means of this FCA application, we show how this approach can be used to determine which personal data has been disclosed, to whom it has been revealed, and which correlations to previously disclosed personal data arise from new messages and their contents.

Our approach is meant to support users of (privacy enhancing) identity management systems in a formal manner, i. e. particularly provide formal specifications for partial user identities, cf. [12], and their relations to communication partners. We see a plain set notion as not appropriate enough, neither in terms of expressiveness, nor in terms of efficiency, and propose a lattice based data structure, instead.

Privacy enhancing technology and identity management, particularly, has been intensively discussed in research during the last years. The main idea is to support users in preserving their anonymity. Particularly pseudonyms are suitable countermeasures against deanonymization. That is, a user can limit the part of personal information which she wants to disclose to a communication partner using pseudonyms. Supposed, personal information can be quantized, then such a part of personal information would be a subset of all available personal information about the user's identity. Thus, we speak of partial identities that can be recognized by communication partners. A pseudonym can, moreover, easily be reused or (metaphorically) thrown away. However, pseudonyms must not be linkable to users in order to preserve anonymity. A more detailed discussion of privacy enhancing identity management can be found in [3] and, for more recent developments, in [9]. An approach of a privacy enhancing identity management system is in development in the PRIME project[1]. A detailed overview can be found in [7].

There have been several approaches in literature to measure anonymity. Schneider and Sidiropoulos [15] use the modelling language CSP for a process algebraic formalisation of anonymity. Syverson and Stubblebine describe anonymity properties in formal languages based on group principals [19]. They describe the information which is to be protected and the purpose of the protection, i.e. the degree of anonymity. Both approaches are possibilistic, i.e. they do not consider a probability distribution on the basic anonymity set.

Many approaches use Shannon entropies [16]. Fischer-Hübner [6] describes a probabilistic approach to compute the risk of reidentification. The adversary is supposed to learn from a large database which contains personal data but no data about the corresponding identities. Sweeney [18] proposes k-anonymity as a measure. The adversary is supposed to learn from the connection between several database tables. He is able to do so, if the tables are linkable by a set of data items, the so called quasi-identifier. Díaz et al. [4] suggest a probabilistic metric for the degree of anonymity. This approach analyzes the quality of anonymity providing services. The adversary is supposed to learn from his observations only. Steinbrecher and Köpsell [17] work out that unlinkability is a generalization of anonymity. In their approach, they also provide a probabilistic metric based on Shannon entropies. However, they measure linkability of items where items can be messages, senders, or recipients. A user is, consequently, anonymous as long as her actions cannot be linked to her, i. e. actions do not yield a hint which points back to the user. They also point out that message contents, independently of their suggested metric, may reduce the user's anonymity to zero.

We attempt to fill the gap which arises from content data. Instead of analyzing linkability of database contents or limiting our view to traffic data, our approach addresses linkability of message contents in communication systems. For this purpose, we figure out which links exist between messages, between subjects, and between messages and subjects.

Formal Concept Analysis [8] was primarily meant to examine the lattice structure of concepts. Such concepts are described by a set of attributes and comprise a set of objects. The concept lattice $(\mathfrak{B}, \leq)$, which results from the analysis, can be computed by means of an incidence relation $I \subseteq G \times M$ between formal objects $g \in G$ and formal attributes $m \in M$. Thus, formal contexts (G, M, I) are the foundation of Formal Concept Analysis. Single concepts (A, B) consist of a set of objects $A \subseteq G$ and a set of attributes $B \subseteq M$. However, this only holds for a limited number of A and B. Given a derivation operator on object sets such that $A' \subseteq M$ is the set of all attributes which are in relation to each object in A. And given (another[2]) derivation operator on attribute sets such that $B' \subseteq G$ is the set of all objects which are in relation to each attribute in B. Then all formal concepts (A, B) require to satisfy the following closure property:

$$(A, B) = (B', B'') = (B', (B')') = (B', A')$$

The set A is the concept's extent and B is called the concept's intent.

The set of all concepts with respect to a formal context is denoted by $\mathfrak{B}$. The order, $\leq$, is defined[8, Definition 21] as

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$$

This order allows to address subordinate and superordinate concepts. An effective lattice computation algorithm has been proposed by Lindig in [11]. We present concept lattices for all examples in this paper, thus, it is not necessary to compute one by hand. In addition we try to support the understanding with so called line-diagrams, a kind of hasse diagram which can be used for representation of concept lattices. It is easy to relate to such diagrams using Toscana systems, for instance, ToscanaJ [1]. We briefly describe how to read line-diagrams in Section 2.1.

In addition to one-valued contexts which are characterized by their incidence relation, we make also use of many-valued contexts which are particularly necessary for *conceptual scaling*. A many-valued context (G, M, W, I) consists of a set of objects G, a set of attributes M, a set of attribute values W, and a ternary relation $I \subseteq G \times M \times W$.

The analogy of the word *concept* to natural languages yields, even though a deliberate, not the only possible application of FCA. Today, Formal Concept Analysis is utilized and extended in various fields, including data exploration and logics.

[1] "Privacy Enhancing Identity Management for Europe", project within 6th framework program of the European Union

[2] We address two formally different operations with the same operator in favor for intuitive notation, as proposed in [8]. This may appear mathematically inconsistent, however, it is formally feasible, as we prove in [2, Chapter 5].

In Section 2, we describe the basics of our approach. Particularly, messages are formally described in relation to their contents. In Section 3, we enhance this approach by subjects and show how subject knowledge can be deduced. In Section 4, we describe how to combine the results of the two previous sections in order to draw linkability conclusions. Finally, in Section 5 we summarize results and conclusions of our approach.

2. FORMALIZATION OF MESSAGES AND CONTENTS

In this section we describe our model of messages and propose basic ideas of formalization by means of formal objects, attributes, concepts, and concept lattices. At first, we show how concept lattices turn out to be very useful in order to sum up those messages which all contain *equal* information. And second, we apply basic scaling methods which, in addition, yield a straightforward way to address messages with *similar* contents.

2.1 Message Lattice

Suppose that messages are containers which can be basically described by their contents. In order to keep the model simple, we furthermore suppose that each message contains one data item only. This is no real restriction, since a message which should contain more than one data item can be represented by several messages which in sum contain all these contents. Then, we can grasp the relation between messages and contents as an incidence relation and use it as formal context. More precisely, we define message ids as formal objects and data items as formal attributes. Such a context is given by (G, M, I) where

$$G = \{m_1, m_2, m_3, m_4, m_5\}, M = \{d_1, d_2, d_3, d_4\},$$

and I is given by the cross-table in Figure 1(a).

Starting from this definition of formal contexts, we can compute simple structured concept lattices. From the context (G, M, I), we obtain six concepts, arranged in a concept lattice, as shown by the reduced line-diagram in Figure 1(b):

$$\begin{aligned}
\mathfrak{B} = \Big\{(G, \varnothing), \quad & \text{(supremum, no data item is commonly contained in all messages)}\\
(\{m_1\}, \{d_1\}), \quad & (m_1 \text{ contains } d_1)\\
(\{m_2, m_5\}, \{d_2\}), \quad & (d_2 \text{ is contained in } m_2 \text{ and } m_5)\\
(\{m_3\}, \{d_3\}), \quad & (m_3 \text{ contains } d_3)\\
(\{m_4\}, \{d_4\}), \quad & (m_4 \text{ contains } d_4)\\
(\varnothing, M)\Big\} \quad & \text{(infimum, no message contains all data items)}
\end{aligned}$$

A line-diagram is an easy to read graph representation of a concept lattice. All vertex labels of such a line-diagram together yield the entire concept lattice. Each concept is represented by a vertex. The extent can be derived from this vertex's object labels and all object labels of vertices which are reachable by *descending* edges. The intent can be derived from the vertex's attribute labels and all attribute labels of vertices which are reachable by *ascending* edges.

Formal concepts consist of a set of objects and a set of attributes. In case of the currently defined context, that is a set of message ids and a set of data items. We know that the set of data items is a singleton for each concept (except supremum and infimum of the concept lattice), since we suppose that each message contains one data item only. The set of message ids, however, may have a size greater than one. Such concepts simply show that all messages in the concept's extent equally contain the data items in the corresponding intent. Thus, concepts sum up equal messages.

2.2 Data Lattice

As we have seen in Section 1, data items are not necessarily independent from each other. In the job application example, for instance, the applicant's first name provides information about the gender which was, in turn, actually to be anonymous. Thus, with our model of messages we only represent this part of communication, so far, which is obvious to everybody, but not the part which is derivable by an adversary with possible background knowledge.

In fact, the relation between data items and derivable data items is an incidence relation. Any data item can be grasped as formal object which has several attributes, i. e. several derivable data items. This is, indeed, no definition of formal contexts which leads to a message lattice. We rather call lattices which have been computed from such a context *data lattices.*

Concepts in such lattices consist of a set of data items and the corresponding set of derivable data items. If the extent, i. e. the set of data items, is not a singleton, then all of these data items have the corresponding intent set of derivable data items in common.

We obtain a formal context (G, M, I) as characterized by the cross-table in Figure 2(a), for instance, if we consider four data items, surname n_s, first name n_f, male g_m, and female g_f as formal objects and a similar set as formal attributes.

$$G = \{n_s, n_f, g_m, g_f\}, M = \{n'_s, n'_f, g'_m, g'_f\}$$

Then, the concept lattice as characterized by the reduced line-diagram in Figure 2(b) can be computed from this context. The entire concept lattice which corresponds to this line-diagram is $(\mathfrak{B}, \leq)$ with

$$\begin{aligned}
\mathfrak{B} = \Big\{(G, \varnothing), \quad & \text{(supremum)}\\
(\{n_s\}, \{n'_s\}), \quad & (\text{label } n_s,\ n'_s)\\
(\{g_m, n_f\}, \{g'_m\}), \quad & (\text{label } g_m,\ g'_m)\\
(\{g_f, n_f\}, \{g'_f\}), \quad & (\text{label } g_f,\ g'_f)\\
(\{n_f\}, \{n'_f, g'_m, g'_f\}), \quad & (\text{label } n_f,\ n'_f)\\
(\varnothing, M)\Big\} \quad & \text{(infimum)}
\end{aligned}$$

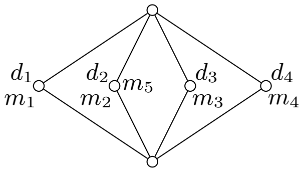
This concept lattice simply yields the fact that the gender is derivable from the first name.

2.3 Improve Message Contents by Derivable Data

Data lattices, as defined in the previous section, can be used to scale message lattices, and, therefore, enhance message lattices with derivable data items. This scaling can be done by plain scaling which is a simple method to enrich the expressiveness of concept lattices. By scaling, we add new attributes to the corresponding context and determine the incidence relation with respect to these new attributes depending on a many-valued attribute. Many-valued attributes differ from one-valued ones, those we used in cross-tables, in so far as they may provide arbitrary attribute values in relation to formal objects.

	d_1	d_2	d_3	d_4
m_1	×			
m_2		×		
m_3			×	
m_4				×
m_5		×		

(a) Formal Context

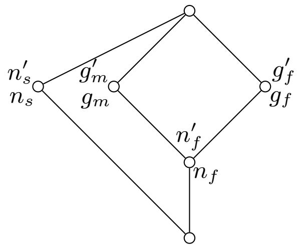


(b) Concept Lattice (shown as reduced line-diagram)

Figure 1: Representation of Messages

	n'_s	n'_f	g'_m	g'_f
n_s	×			
n_f		×	×	×
g_m			×	
g_f				×

(a) Formal Context



(b) Concept Lattice (shown as reduced line-diagram)

Figure 2: Representation of Derivable Data Items

Formal contexts of message lattices can, in fact, be transformed to contexts which contain many-valued attributes. This is basically the reverse operation of plain scaling and for sure possible, since only one single one-valued attribute stands in relation with each object, by definition. The process of plain scaling, then, replaces the many-valued attribute *data item*, with the corresponding rows of the conceptual scale.

Figure 3 shows how contexts which follow the definition from Section 2.1, cf. Figure 3(a), can be plain scaled by first constructing a many-valued context, cf. Figure 3(b), and then applying the context in Figure 2(a) as scale which yields the scaled context shown in Figure 3(c). In Figure 3(d), we see that the resulting concept lattice $(\mathfrak{B}, \leq)$ arranges messages according to data items which are derivable from their contents:

$$
\begin{aligned}
\mathfrak{B} = \big\{(G, \varnothing), & \quad \text{(no data item is contained in each message)}\\
(\{m_1, m_2, m_3\}, \{g'_m\}), & \quad (m_1,\ m_2, \text{ and } m_3 \text{ contain } g'_m)\\
(\{m_2, m_4\}, \{g'_f\}), & \quad (m_2 \text{ and } m_4 \text{ contain } g'_f)\\
(\{m_2\}, \{n'_f, g'_m, g'_f\}), & \quad (m_2 \text{ contains } n'_f,\ g'_m, \text{ and } g'_f)\\
(\{m_5\}, \{n'_s\}), & \quad (m_5 \text{ contains } n'_s)\\
(\varnothing, M)\big\} & \quad \text{(no message contains all data items)}
\end{aligned}
$$

In this particular case, the lattice structure does not differ from the scale lattice structure, because the unscaled context contains only one attribute which is right this one which is subject of the scaling process, cf. 3(b).

2.4 Intermediate Results

A simple abstraction of messages can easily be formalized using formal concepts. Concepts, then, describe all pairwise different contents and content intersections, and assign the corresponding messages to these contents.

We also propose a way to arrange and grasp *data* as superordinate and subordinate concepts in a formal manner, i. e. in concept lattices. Particularly, with respect to privacy-enhancing technology, this is an important achievement, since disclosed data may let an adversary with background knowledge link different data items with each other and draw conclusions about personal data which has not actually been sent in any relevant message.

The connection between messages, their contents, and derivable data can be established using plain scaling. The scaling method determines, in fact, the expressiveness of possible derivations. In order to keep it simple, we propose a scale which maps each single data item to a set of derivable data. This already provides a powerful method, but might be insufficient if it is required to model data items which are only derivable from two or more contents. However, such a case could easily be overcome by switching to a more complex scaling method and a more precise scale.

From the concept lattice structure, we see which messages are correlated with respect to their contents. Such messages stand in relation with respect to $\leq$, the lattice order.

3. DEDUCING SUBJECT KNOWLEDGE

Subjects or users of privacy-enhancing technology often try to act anonymously or pseudonymously. Thus, in the proposed model of communication systems, we need to consider subjects and pseudonyms. In this section, we first show how relations between subjects and their pseudonyms can be formalized using concept lattices. Then, we show that relations between subjects and messages can easily be added to context definitions of the previous section. By using such enhanced context definitions, we propose a method to deduce subject knowledge from subjects and their relation to messages.

	n_s	n_f	g_m	g_f
m_1			×	
m_2		×		
m_3			×	
m_4				×
m_5	×			

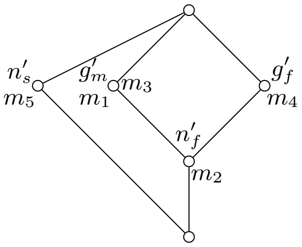
(a) One-Valued

	data item
m_1	g_m
m_2	n_f
m_3	g_m
m_4	g_f
m_5	n_s

(b) Many-Valued

	n'_s	n'_f	g'_m	g'_f
m_1			×	
m_2		×	×	×
m_3			×	
m_4				×
m_5	×			

(c) Scaled



(d) Concept Lattice

Figure 3: Plain Scaling of Message Contents by Means of Derivable Data Items

3.1 Subject–Pseudonym Lattice

So far, we considered messages and contents without any relation to users of the communication system. Users are important to look at, since in the long run we do not only need to keep track of which messages contained what personal data, but also which contents were given to which users. The word user is, however, not necessarily suitable. We speak of pseudonyms, instead, i. e. of digital representations of real persons. Additionally, we henceforth reduce real persons to subjects, since in our approach we are not interested in any fleshy attributes of them. Neither subjects nor pseudonyms are required to use the system, it is rather sufficient that they could be users. We consider both, subjects and pseudonyms, for our model of a communication system, since subjects typically cause messages to be sent, but only their pseudonyms can be recognized without applying background knowledge.

Such background knowledge, i. e. the assignment of subjects to pseudonyms, can be arranged in a lattice. By defining a corresponding formal context as conceptual scale, we can additionally use scaling methods to enrich other contexts with this supplementary knowledge. Therefore, we construct a conceptual scale (G, M, I) in which G represents the set of all known pseudonyms, M the set of subjects, and I the known part of the corresponding relation.

Using this definition, arbitrary relations between subjects and pseudonyms can be formalized. Commonly, each pseudonym will be held by one subject. The proposed conceptual scale is, however, even usable in more complex situation, for instance, if a pseudonym is shared between several subjects. This includes, indeed, the case that several pseudonyms are shared between several subjects. In Figure 4, we present a corresponding example with a set of pseudonyms $G = \{\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3, \mathbb{P}_4\}$, a set of subjects $M = \{\mathbb{S}_1, \mathbb{S}_2, \mathbb{S}_3, \mathbb{S}_4\}$, and the relation I between both given by the cross-table in Figure 4(a). From concepts of the corresponding lattice $(\mathfrak{B}, \leq)$, we see:

$$\begin{aligned}
\mathfrak{B} = \big\{(G, \varnothing), &\qquad \text{(no subject controls all pseudonyms)}\\
(\{\mathbb{P}_1, \mathbb{P}_3, \mathbb{P}_4\}, \{\mathbb{S}_1\}), &\qquad \text{($\mathbb{S}_1$ controls $\mathbb{P}_1$, $\mathbb{P}_3$, and $\mathbb{P}_4$)}\\
(\{\mathbb{P}_1, \mathbb{P}_4\}, \{\mathbb{S}_1, \mathbb{S}_4\}), &\qquad \text{($\mathbb{S}_1$ and $\mathbb{S}_4$ share $\mathbb{P}_1$ and $\mathbb{P}_4$)}\\
(\{\mathbb{P}_3, \mathbb{P}_4\}, \{\mathbb{S}_1, \mathbb{S}_3\}), &\qquad \text{($\mathbb{S}_1$ and $\mathbb{S}_3$ share $\mathbb{P}_3$ and $\mathbb{P}_4$)}\\
(\{\mathbb{P}_2\}, \{\mathbb{S}_2\}), &\qquad \text{($\mathbb{S}_2$ controls $\mathbb{P}_2$)}\\
(\{\mathbb{P}_4\}, \{\mathbb{S}_1, \mathbb{S}_3, \mathbb{S}_4\}), &\qquad \text{($\mathbb{S}_1$, $\mathbb{S}_3$, and $\mathbb{S}_4$ share $\mathbb{P}_4$)}\\
(\varnothing, M)\big\} &\qquad \text{(no pseudonym is shared between all subjects)}
\end{aligned}$$

3.2 Assigning Subjects to Messages

In Section 2.1, we consider only data items as formal attributes in context definitions for message lattices. In order to describe messages more detailed, we enhance this context definition by another attribute type, i. e. by subject ids. Alternatively, we allow pseudonym ids, either as one many-valued attribute or each as a single one-valued attribute. The first case is appropriate, in case subject ids shall be substituted for pseudonym ids by scaling using a conceptual scale such as defined in Section 3.1. The latter case is appropriate, if such a scale is not known or should not be applied and, therefore, the formal context is to be used as such for the computation of concept lattices.

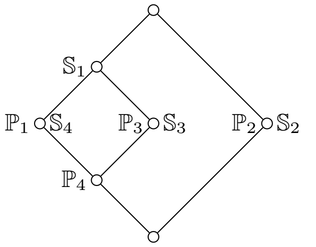
Suppose, we want to substitute subject ids for pseudonym ids. Plain scaling would require that each message is in relation to one pseudonym at most, since only this particular pseudonym id could appear as value in the corresponding many-valued attribute. Messages in a real communication system, however, are at least related to a sender and a recipient. Thus, plain scaling with one many-valued attribute for pseudonyms is only sufficient in trivial cases, for instance, if only the relation of a sender pseudonym to the subjects is of interest.

In order to overcome this limitation, we propose another scaling method, the *power set scaling*. An example application of power set scaling can be found at the end of Section 3.3. We use the context in Figure 6(b) as many-valued context, Figure 6(c) as conceptual scale, and achieve the context in Figure 7(a).

Henceforth we use $\mathcal{P}(A)$ to denote the power set of set A. Suppose, the values (in W) of many-valued attributes $m \in$

	$\mathbb{S}_1$	$\mathbb{S}_2$	$\mathbb{S}_3$	$\mathbb{S}_4$
$\mathbb{P}_1$	×			×
$\mathbb{P}_2$		×		
$\mathbb{P}_3$	×		×	
$\mathbb{P}_4$	×		×	×

(a) Context


$\mathbb{S}_1$
$\mathbb{P}_1$ $\mathbb{S}_4$
$\mathbb{P}_3$ $\mathbb{S}_3$
$\mathbb{P}_2$ $\mathbb{S}_2$
$\mathbb{P}_4$

(b) Concept Lattice (shown as reduced line-diagram)

Figure 4: Pseudonym–Subject Scale

M in a context (G, M, W, I) are power set elements of scale objects G_m with respect to conceptual scales (G_m, M_m, I_m), i. e.

$$W \subseteq \bigcup_{m \in M} \mathcal{P}(G_m)$$

Then, this scaling method substitutes all scale attributes which are related to any of the scale objects (in the corresponding attribute value) for attribute values. Formally, let $\dot{M}_m = \{m\} \times M_m$ in the *scaled* context $(\tilde{G}, \tilde{M}, \tilde{I})$ where $\tilde{G} = G$ and

$$\tilde{M} = \bigcup_{m \in M} \dot{M}_m$$

$$\tilde{I} = \left\{ (\tilde{g}, \tilde{m}) \;\middle|\; (\tilde{g}, m, w) \in I, \hat{g} \in w, (\hat{g}, \tilde{m}) \in I_m \right\}$$

Thus, for this scaling method, messages can be related to many subjects or pseudonyms, respectively.

Such a context could, for instance, be (G, M, I) with $G = \{m_1, m_2, m_3, m_4, m_5\}$, $M = \{n'_s, n'_f, g'_m, g'_f, \mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3, \mathbb{P}_4\}$, and I given by the cross-table in Figure 5(a). The corresponding concept lattice $(\mathfrak{B}, \leq)$, cf. Figure 5(b), consists of:

$$\begin{aligned}
\mathfrak{B} = \big\{ &(G, \varnothing), && \text{(no subject belong to all messages)}\\
&(\{m_1, m_2, m_3\}, \{g'_m, \mathbb{P}_1\}), && (\mathbb{P}_1 \text{ knows } m_1,\ m_2, \text{ and } m_3)\\
&(\{m_1, m_3\}, \{g'_m, \mathbb{P}_1, \mathbb{P}_3\}), && (\mathbb{P}_1 \text{ and } \mathbb{P}_3 \text{ know } m_1 \text{ and } m_3)\\
&(\{m_1, m_2\}, \{g'_m, \mathbb{P}_1, \mathbb{P}_4\}), && (\mathbb{P}_1 \text{ and } \mathbb{P}_4 \text{ know } m_1 \text{ and } m_2)\\
&(\{m_2, m_4\}, \{g'_f\}),\\
&(\{m_4, m_5\}, \{\mathbb{P}_2\}), && (\mathbb{P}_2 \text{ knows } m_4 \text{ and } m_5)\\
&(\{m_1\}, \{g'_m, \mathbb{P}_1, \mathbb{P}_3, \mathbb{P}_4\}), && (\mathbb{P}_1,\ \mathbb{P}_3, \text{ and } \mathbb{P}_4 \text{ know } m_1)\\
&(\{m_2\}, \{n'_f, g'_m, g'_f, \mathbb{P}_1, \mathbb{P}_4\}), && (\mathbb{P}_1 \text{ and } \mathbb{P}_4 \text{ know } m_2)\\
&(\{m_4\}, \{g'_f, \mathbb{P}_2\}), && (\mathbb{P}_2 \text{ knows } m_4)\\
&(\{m_5\}, \{n'_s, \mathbb{P}_2\}), && (\mathbb{P}_2 \text{ knows } m_5)\\
&(\varnothing, M) \big\} && \text{(no message is known to all subjects)}
\end{aligned}$$

3.3 Contents towards Subject Knowledge

In this section, we describe a procedure which can be understood as separate scaling method. Starting from a formal context (G, M, I), we show formally how to develop the intermediate contexts $(\tilde{G}, \tilde{M}, \tilde{I})$ and $(\hat{G}, \hat{M}, \hat{W}, \hat{I})$. And by means of a conceptual scale $(G_{\text{msg}}, M_{\text{msg}}, I_{\text{msg}})$ we develop the resulting context $(\bar{G}, \bar{M}, \bar{I})$. The main idea is to reverse formal objects and part of the attributes within a formal context.

Suppose a context where messages are formal objects and contents as well as subjects are attributes, as described in Section 3.2. We are looking for a context reflecting the same relations, but providing subjects or pseudonyms as formal objects and summed up message contents as attributes, i. e. data items as knowledge.

At first, we basically limit the view on a formal context (G, M, I) to these attributes $M_\times$ which are to be reversed. In the case as described by Figure 5(a), $M_\times$ just contains all pseudonym ids. Such a context $(G, M_\times, I)$ can easily be reversed by interchanging G and $M_\times$ and replacing I by

$$\tilde{I} = \left\{ (m, g) \;\middle|\; (g, m) \in I \right\}$$

The resulting context $(\tilde{G}, \tilde{M}, \tilde{I})$ with $\tilde{G} = M_\times$ and $\tilde{M} = G$, cf. Figure 6(a), yields pseudonyms as formal objects, but messages instead of data items as formal attributes. Thus, in the next step, we have to replace the formal attributes $\tilde{M}$, i. e. messages, by data items. This can be done by power set scaling. First, however, we need the many-valued attribute which is to be scaled. This can be achieved by constructing a many-valued context $(\hat{G}, \hat{M}, \hat{W}, \hat{I})$, cf. Figure 6(b), with $\hat{G} = \tilde{G}$, $\hat{M} = \{\text{msg}\}$ (i. e. just one attribute for message sets), $\hat{W} = \mathcal{P}(\tilde{M})$ (i. e. the power set of messages), and

$$\hat{I} = \left\{ \big(g, \text{msg}, w\big) \;\middle|\; g \in \tilde{G}, w = \left\{ m \;\middle|\; (g, m) \in \tilde{I}, m \in \tilde{M} \right\} \right\}$$

or by using the derivation operator, cf. Section 1

$$\hat{I} = \left\{ \big(g, \text{msg}, w\big) \;\middle|\; g \in \tilde{G}, w = \{g\}' \right\}$$

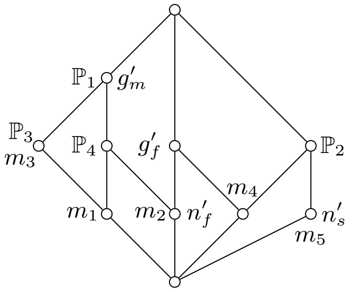
For power set scaling, we additionally need to define the conceptual scale $(G_{\text{msg}}, M_{\text{msg}}, I_{\text{msg}})$. We choose $G_{\text{msg}} = G$ as scale objects, $M_{\text{msg}} = M \setminus M_\times$ as scale attributes, which are all data items but no subjects, and

$$I_{\text{msg}} = \left\{ (g, m) \;\middle|\; (g, m) \in I, m \in M_{\text{msg}} \right\}$$

The result from applying scale $(G_{\text{msg}}, M_{\text{msg}}, I_{\text{msg}})$, cf. Figure 6(c), to the many-valued context $(\hat{G}, \hat{M}, \hat{W}, \hat{I})$ yields an one-valued context $(\bar{G}, \bar{M}, \bar{I})$, cf. Figure 7(a), with pseudonyms as formal objects and data items as formal attributes.

	n'_s	n'_f	g'_m	g'_f	$\mathbb{P}_1$	$\mathbb{P}_2$	$\mathbb{P}_3$	$\mathbb{P}_4$
m_1			×		×		×	×
m_2		×	×	×	×			×
m_3			×		×		×	
m_4				×		×		
m_5	×					×		

(a) Context



(b) Concept Lattice

Figure 5: Subject Ids as Attributes of Messages

	m_1	m_2	m_3	m_4	m_5
$\mathbb{P}_1$	×	×	×		
$\mathbb{P}_2$				×	×
$\mathbb{P}_3$	×		×		
$\mathbb{P}_4$	×	×			

(a) $(\tilde{G}, \tilde{M}, \tilde{I})$

	msg
$\mathbb{P}_1$	$\{m_1, m_2, m_3\}$
$\mathbb{P}_2$	$\{m_4, m_5\}$
$\mathbb{P}_3$	$\{m_1, m_3\}$
$\mathbb{P}_4$	$\{m_1, m_2\}$

(b) $(\hat{G}, \hat{M}, \hat{W}, \hat{I})$

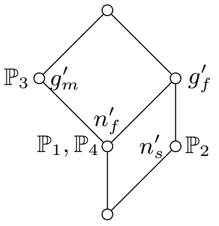
	n'_s	n'_f	g'_m	g'_f
m_1			×	
m_2		×	×	×
m_3			×	
m_4				×
m_5	×			

(c) $(G_{\text{msg}}, M_{\text{msg}}, I_{\text{msg}})$

Figure 6: Transformation Contexts, Basing on Figure 5(a)

	n'_s	n'_f	g'_m	g'_f
$\mathbb{P}_1$		×	×	×
$\mathbb{P}_2$	×			×
$\mathbb{P}_3$			×	
$\mathbb{P}_4$		×	×	×

(a) $(\bar{G}, \bar{M}, \bar{I})$



(b) Concept Lattice

Figure 7: Reversed Context

We achieve the concept lattice $(\bar{\mathfrak{B}}, \leq)$, cf. Figure 7(b):

$$\begin{aligned}
\bar{\mathfrak{B}} = \big\{(\bar{G}, \varnothing), && \text{(no data item is known to each pseudonym)}\\
(\{\mathbb{P}_1, \mathbb{P}_3, \mathbb{P}_4\}, \{g'_m\}), && (\mathbb{P}_1,\ \mathbb{P}_3, \text{ and } \mathbb{P}_4 \text{ know } g'_m)\\
(\{\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_4\}, \{g'_f\}), && (\mathbb{P}_1,\ \mathbb{P}_2, \text{ and } \mathbb{P}_4 \text{ know } g'_f)\\
(\{\mathbb{P}_1, \mathbb{P}_4\}, \{n'_f, g'_m, g'_f\}), && (\mathbb{P}_1 \text{ and } \mathbb{P}_4 \text{ know } n'_f,\ g'_m, \text{ and } g'_f)\\
(\{\mathbb{P}_2\}, \{n'_s, g'_f\}), && (\mathbb{P}_2 \text{ knows } n'_s \text{ and } g'_f)\\
(\varnothing, \bar{M})\big\} && \text{(no pseudonym knows all data items)}
\end{aligned}$$

3.4 Intermediate Results

It is easy to formalize the relation between subjects and pseudonyms in a conceptual manner. However, with respect to privacy enhancing technology, this is usually background knowledge which is not necessarily available.

The way of formalization which we propose is moreover easy to integrate in context definitions from the previous sections. Subject or pseudonym ids are basically assignable to messages as formal attributes. Then, we saw that resolving pseudonym ids to subject ids in such contexts by hindsight is not much effort, since we already formalized that knowledge and can, therefore, utilize very basic methods of Formal Concept Analysis, that is conceptual scaling.

By means of a context transformation, we saw that it is possible to deduce formal contexts with subject ids (or pseudonym ids, respectively) as formal objects rather than message ids. These contexts can be used to analyze the knowledge of subjects by means of data items. The lattice structure with respect to the order $\leq$ yields the correlations between pseudonyms (or subjects) with respect to their knowledge.

4. COMPOSING CONTENTS AND KNOWLEDGE IN ONE LATTICE

From Section 2 and 3, we achieve context definitions which have data items as formal attributes in common. The difference between both is that context definitions in Section 2 yield messages as formal objects whereas the final contexts in Section 3 yield pseudonyms, instead. In order to find correlations between message contents and subject knowledge, we need to concatenate contexts of both types. In this section, we describe how to concatenate such contexts in order to achieve a concept lattice from which we can, then, derive linkability estimations.

By concatenating two formal contexts (G_1, M_1, I_1) and (G_2, M_2, I_2), we mean mainly to unite the contained sets. However, we use disjoint object sets $\dot{G}_i = \{i\} \times G_i$ and have, therefore, also to adapt the incidence relations I_i before the union:

$$\begin{aligned}
(G_1, M_1, I_1) \cdot (G_2, M_2, I_2) &= (\dot{G}_1 \cup \dot{G}_2, M_1 \cup M_2, \dot{I}_1 \cup \dot{I}_2)\\
\dot{G}_i &= \{i\} \times G_i\\
\dot{I}_i &= \Big\{\big((i, g), m\big) \ \Big|\ (g, m) \in I_i\Big\}
\end{aligned}$$

This differs from the definition of the direct sum in [8, Definition 32] where the disjoint union is also used for attributes. In the case which we describe here, however, we want to intersect attribute sets M_i, i. e. data items, as much as possible. Informally, the concatenation extends the first context (written as cross-table) by the rows and additional columns of the second context. Furthermore, there cannot appear any conflict between the two incidence relations, since the object sets $\dot{G}_i$ are disjoint by definition.

This formal description can be used for automated computing. We will leave out the cross product in the next examples which ensures that the two object sets are disjoint and choose disjoint objects sets, instead, by hand.

The concatenation (G, M, I) of the final contexts from Section 2 and 3 is, for instance, determined by

$$\begin{aligned}
G &= \{m_1, m_2, m_3, m_4, m_5, \mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3, \mathbb{P}_4\}\\
M &= \{n'_s, n'_f, g'_m, g'_f\}
\end{aligned}$$

and the cross-table in Figure 8(a) as incidence relation I. As the corresponding concept lattice $(\mathfrak{B}, \leq)$, cf. Figure 8(b), we get

$$\begin{aligned}
\mathfrak{B} = \big\{(G, \varnothing), \quad& \tag{1}\\
(\{m_1, m_2, m_3, \mathbb{P}_1, \mathbb{P}_3, \mathbb{P}_4\}, \{g'_m\}), \quad& \tag{2}\\
(\{m_2, m_4, \mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_4\}, \{g'_f\}), \quad& \tag{3}\\
(\{m_5, \mathbb{P}_2\}, \{n'_s\}), \quad& \tag{4}\\
(\{m_2, \mathbb{P}_1, \mathbb{P}_4\}, \{n'_f, g'_m, g'_f\}), \quad& \tag{5}\\
(\{\mathbb{P}_2\}, \{g'_f, n'_s\}), \quad& \tag{6}\\
(\varnothing, M)\big\} \quad& \tag{7}
\end{aligned}$$

In this lattice, we see that no data item links all messages and pseudonyms, cf. (1), and no message or pseudonym is linkable by all data items, cf. (7). Furthermore, we see in (2) that the messages m_1, m_2, and m_3 are linkable by one data item, g'_m, to the pseudonym $\mathbb{P}_1$, $\mathbb{P}_3$, and $\mathbb{P}_4$. In (3), we also find the pseudonyms $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{P}_4$, and the messages m_2 and m_4 related by a single data item. Slightly more interesting are (4) and (5), the first concept is linking exactly one message to a pseudonym, the latter is linking messages and pseudonyms by a relatively great set of data items.

5. CONCLUSIONS

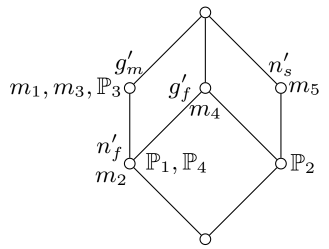
We propose specifications for several kinds of formal contexts. Formal Concept Analysis can be utilized, as we have seen, to formalize the relation between messages and contents as well as the relation between subjects and their knowledge. The resulting concept lattices are well suited for linkability estimations. The correlations are computed as side effect of the lattice structure.

In Section 4, we combine both approaches of linkability analysis by defining the concatenation of formal contexts. We see that concept lattices of concatenated contexts yield correlations between messages and pseudonyms in addition to the already discovered correlations. This linkability analysis is particularly useful when the user of an identity management system is about to create a new message and needs support in choosing appropriate data items as contents. With possible correlations to (actually unrelated) pseudonyms in mind, she can choose a reasonable set of data items, such that arising correlations cannot be used to reidentify her.

This can even be supported in an automated way by means of the lattice structure. Say, y is *greater* than x with respect to the lattice order $\leq$, if and only if it holds $x \leq y$.

	n'_s	n'_f	g'_m	g'_f
m_1			×	
m_2		×	×	×
m_3			×	
m_4				×
m_5	×			
$\mathbb{P}_1$		×	×	×
$\mathbb{P}_2$	×			×
$\mathbb{P}_3$			×	
$\mathbb{P}_4$		×	×	×

(a) Context



(b) Concept Lattice

Figure 8: Concatenation of Contexts from Figure 3(c) and 7(a)

Then, an algorithm could choose the greatest concept from the lattice which still contains an appropriate set of data items. Such a simple algorithm would yield the concept intent as a reasonable starting point for new messages.

An identity management system can moreover provide the capability to let the user browse through the concept lattice structure. That is, the user would be given the opportunity to explore her partial identities by means of concepts. The lattice structure provides, therefore, an easy to understand order of partial identities. Indeed, user acceptance first of all depends on the user interface. A promising approach could be an interface like a file system, cf. [5].

Yet, there is a reference implementation [2, Appendix A] which is still lacking the integration in an existing identity management system. However, in [2, Chapter 5] or [11], respectively, the complexity of all necessary FCA operations is determined. Furthermore, a comprehensive application example can be found in [2, Chapter 4].

Formal contexts can easily be adapted to the current requirements. We have seen that, with very basic operations, quite a lot of analysis can be done. Besides, there is a great variety of further scaling methods for Formal Concept Analysis which can be applied. In [2] we utilize relational scaling, cf. [13], in order to categorize subjects as originators or receivers of messages, for instance.

Further research has to be done on the representation of all relevant correlations in single concepts. Suppose, two messages, for instance, one linkable by the adversary to a subject and the second message linkable (by a disjunct set of data items) to the first one. The link between this subject and the second message would not be derivable from a single concept. However, this affects just transitive linkability threats which arise from relations with one or more intermediate item.

6. REFERENCES

[1] P. Becker, J. Hereth, and G. Stumme. ToscanaJ: An open source tool for qualitative data analysis. In V. Duquenne, B. Ganter, M. Liquiere, E. M. Nguifo, and G. Stumme, editors, *Advances in Formal Concept Analysis for Knowledge Discovery in Databases. Proc. Workshop FCAKDD of the 15th European Conference on Artificial Intelligence (ECAI 2002). Lyon, France.*, July 23 2002.

[2] S. Berthold. Linkability of communication contents: Keeping track of disclosed data using formal concept analysis. Master's thesis, Technische Universität Dresden, Germany, Karlstads Universitet, Sweden, 2006.

[3] S. Clauß and M. Köhntopp. Identity management and its support of multilateral security. *Computer Networks*, Special Issue on Electronic Business Systems(37):205–219, 2001. Elsevier, North-Holland.

[4] C. Díaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In R. Dingledine and P. Syverson, editors, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002).* Springer-Verlag, LNCS 2482, April 2002.

[5] S. Ferré and O. Ridoux. A file system based on concept analysis. In *Computational Logic - CL 2000: First International Conference*, page 1033, London, UK, July 2000. Springer-Verlag, LNCS 1861.

[6] S. Fischer-Hübner. *IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms*, volume 1958 of *Lecture notes in computer science.* Springer-Verlag, 2001.

[7] S. Fischer-Hübner, C. Andersson, and T. Holleboom. Framework v2. Deliverable D14.1.b, PRIME – Privacy and Identity Management for Europe, Contract No. 507591, July 2006.

[8] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer-Verlag Berlin Heidelberg, 1999.

[9] M. Hansen, P. Berlich, J. Camenisch, S. Clauß, A. Pfitzmann, and M. Waidner. Privacy-enhancing identity management. Istr, 2004.

[10] T. Linderborg. Avidentifiera jobbansökningar – en metod för mångfald. Technical Report SOU 2005:115, Sveriges regering, December 2005.

[11] C. Lindig. Fast concept analysis. In *Work with Conceptual Structures – Contributions to ICCS 2000*, pages 152–161. Shaker Verlag, August 2000.

[12] A. Pfitzmann and M. Hansen. *Anonymity, Unlinkablility, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology.* Technische Universität Dresden and ULD Kiel, v0.28 edition, May 2006.

[13] S. Prediger and R. Wille. The lattice of concept graphs of a relationally scaled context. In *ICCS '99: Proceedings of the 7th International Conference on Conceptual Structures*, pages 401–414, London, UK, 1999. Springer-Verlag.

[14] U. Priss. Formal concept analysis in information science. *Annual review of information science and technology*, 40:521–543, 2006.

[15] S. Schneider and A. Sidiropoulos. CSP and anonymity. In *ESORICS 1996*, number 1146 in Lecture Notes in Computer Science, pages 198–218, Berlin, 1996. Springer.

[16] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[17] S. Steinbrecher and S. Köpsell. Modelling unlinkability. In R. Dingledine, editor, *Proceedings of Privacy Enhancing Technologies workshop (PET 2003)*. Springer-Verlag, LNCS 2760, March 2003.

[18] L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[19] P. F. Syverson and S. G. Stubblebine. Group principals and the formalization of anonymity. In *Proceedings of the World Congress on Formal Methods (1)*, pages 814–833, 1999.