

2. Privacy Enhancing Security Criteria

Privacy Enhancing Security Aspects for:

- Protecting the User-Identities providing *Anonymity*, *Pseudonymity*, *Unlinkability*, *Unobservability* of users
- Protecting Usee-Identities providing *Anonymity*, *Pseudonymity* of data subjects
- Protecting *Confidentiality* and *Integrity* of personal data

2.1 Privacy-Enhancing Security Criteria for Protecting User Identities:

Anonymity:

"ensures that a user may use a resource or service without disclosing the user's identity" [CC 1998]

Definition:

R_U : "user U performs a role R during an event E"

A: attacker

NC_A : set of users, who are not cooperating with A

U ($U \in NC_A$) is anonymous in role R for an event E to an attacker A if for each observation B:

$$\forall U' \in NC_A: 0 \ll P(R_{U'} | B) \ll 1$$

U is perfectly anonymous if

$$\forall U' \in NC_A: P(R_U) = P(R_{U'} | B)$$

Sender anonymity: the user is anonymous in the role of a sender of a message

Receiver anonymity: the user is anonymous in the role of a receiver of a message

Unobservability:

" ensures that a user may use a resource or service without others being able to observe that the resource or service is being used" [CC 1998]

An event E is **unobservable** for an attacker A if for each observation B that A can make:

$$0 \ll P(E | B) \ll 1$$

An event E is **perfectly unobservable**

$$P(E) = P(E | B)$$

Unlinkability of sender and recipient:

sender and recipient cannot be identified as communicating with each other.

Unlinkability:

"ensures that a user may make use of resources and services without others being able to link these uses together"

[CC 1998]

$X_{E,F}$: "events E and F have a corresponding characteristic"

Two events, E and F, are **unlinkable** in regard of a characteristic for an attacker A if for each observation B that A can make:

$$0 \ll P(X_{E,F} | B) \ll 1$$

E and F are **perfectly unlinkable** if:

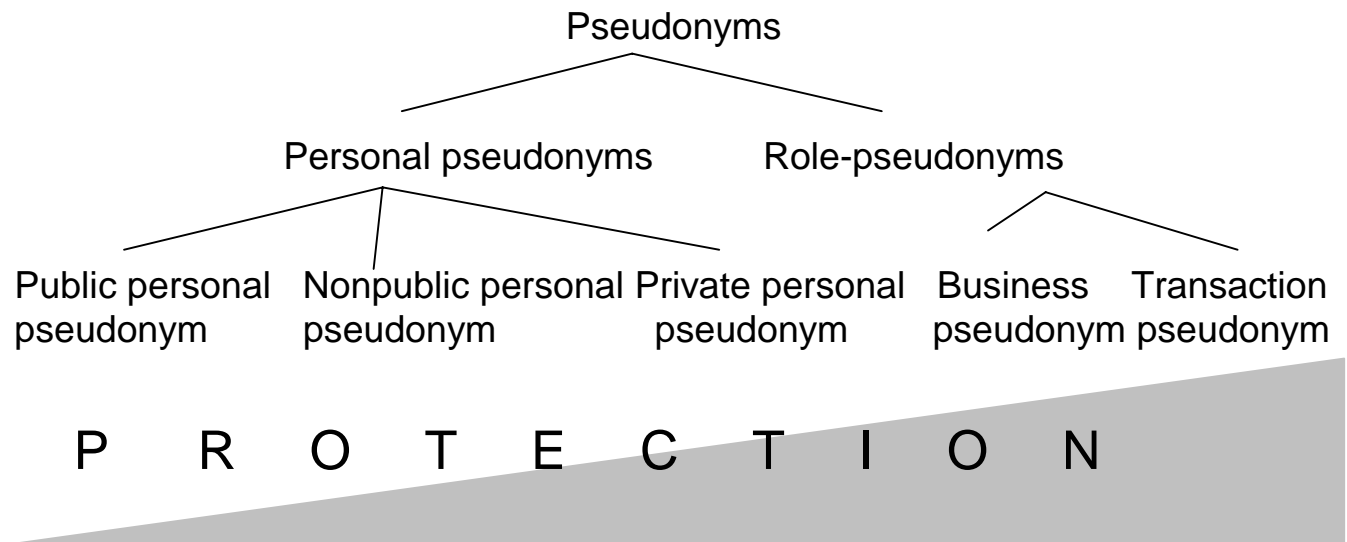
$$P(X_{E,F} | B) = P(X_{E,F})$$

Pseudonymity:

"ensures that a user acting under a pseudonym may use a resource or service without disclosing his identity"

[CC 1998]

Classification of pseudonyms according to their degree of protection:



Classification of pseudonyms according to their generation:

- Self-generated pseudonyms
- Reference pseudonyms
- Cryptographic pseudonyms
- One-way pseudonyms

2.2 Privacy-Enhancing Security Criteria for Protecting Use Identities:

2.2.1 Depersonalisation (anonymisation)

Perfect depersonalisation:

data are rendered anonymous in such a way that the data subject is no longer identifiable

Practical depersonalisation:

the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual

2.2.2 The Risk of Reidentification

Data records collected for statistical purposes contain:

- *Identity data* (e.g., name, address, personal number)
- *Demographic data* (e.g., sex, age, nationality)
- *Analysis data* (e.g., diseases, habits)

The degree of anonymity of statistical data depends on

- the size of the database
- the entropy of the demographic data attributes that can serve as supplementary knowledge of an attacker.

The entropy of the demographic data attributes depends on

- the number of attributes
- the number of possible values of each attribute
- frequency distribution of the values
- dependencies between attributes

Definition 2.1.a:

Given m attributes X_1, \dots, X_m , with values x_{i1}, \dots, x_{in_i} for X_i

n_i

and $\sum_{j=1}^{n_i} p(x_{ij}) = 1$.

$j=1$

The entropy $H(X_i)$ of attribute X_i is defined by [Shannon 1948]:

$$0 \leq H(X_i) = \sum_{j=1}^{n_i} p(x_{ij}) * \text{ld}(1/p(x_{ij})) \leq \text{ld}(n_i)$$

The entropy of an attribute X_i :

1. increases as the number n_i of possible values increases
2. decreases as the distribution of attribute values becomes more and more skewed (for a given n_i):
 $H(X) = 0$ when $p(x_{ij}) = 1$ for some value x_{ij}
3. is maximal (for a given n_i) if all values are equally likely:
 $H(X_i) = n_i * ((1/n_i) * \text{ld}(n_i)) = \text{ld}(n_i)$

Example:

Given the attribute sex with

$$p(\text{male}) = 0.469,$$

$$p(\text{female}) = 0.531.$$

Then:

$$H(\text{sex}) = p(\text{male}) * \text{ld} (1/p(\text{male})) + p(\text{female}) * \text{ld} (1/p(\text{female}))$$

=

$$0.469 * \text{ld} (1/0.469) + 0.531 * \text{ld} (1 / 0.531) = 0.997.$$

Given the attribute nationality with the two possible values S (Swedish) and F (Foreigner), $p(S) = 0.9$, $p(F) = 0.1$.

Then:

$$H(\text{nationality}) = 0.9 * \text{ld} (1/0.9) + 0.1 * \text{ld} (1/0.1) = 0.465.$$

Definition 2.1.b:

The entropy of a combination of attributes X_1, \dots, X_m is defined by:

$$\begin{aligned} 0 &\leq H(X_1 \dots X_m) \\ &= \sum_{j=1}^{n_1} \sum_{j_m=1}^{m_m} p(x_{1j_1} \dots x_{mj_m}) \cdot \text{ld}(1/p(x_{1j_1} \dots x_{mj_m})) \\ &\leq H(X_1) + \dots + H(X_m) \end{aligned}$$

Definition 3:

The function ANVC (“**a**verage number of **v**alue **c**ombinations”) is defined by:

$$H(X_1, \dots, X_m)$$

$$\text{ANVC}(X_1, \dots, X_m) = 2$$

= measure for the average number of value combinations for attributes X_1, \dots, X_m , which can actually be used for re-identification.

Example: $\text{ANVC}(\text{sex}) = 2^{0.957} \approx 1.996,$
 $\text{ANVC}(\text{nationality}) = 2^{0.465} \approx 1.38$

Since $0 \leq H(X_1, \dots, X_m) \leq H(X_1) + \dots + H(X_m) \leq \text{ld}(n_1) + \dots + \text{ld}(n_m) = \text{ld}(\prod n_i)$:

$$1 \leq \text{ANVC}(X_1, \dots, X_m) \leq \prod n_i$$

If there are no dependencies between the attributes X_1, \dots, X_m :

$$\text{ANVC}(X_1, \dots, X_m) = \prod 2^{H(X_i)}$$

If in addition for each attribute X_i all values x_{i1}, \dots, x_{in_i} are equally likely:

$$\text{ANVC}(X_1, \dots, X_m) = \prod n_i$$

Definition 4: The function RR to estimate the average re-identification risk is defined by:

$$RR(X_1, \dots, X_m) = \begin{cases} ANVC(X_1, \dots, X_m)/N & \text{if } ANVC(X_1, \dots, X_m) \leq N \\ 1 & \text{else.} \end{cases}$$

$$1 / N \leq RR (X_1, \dots, X_m) \leq 1$$

Example:

Given

- *database with N= 100 records for N different individuals*
 - *demographic attributes marital status and sex*
- H(sex, marital status) = 2.61*

Then:

$$ANVC (sex, marital status) = 2^{H(sex, marital status)} = 2^{2.6} = 6.105$$

→ *6.105 of eight possible value combinations can be used on the average*

The re-identification risk can be estimated by:

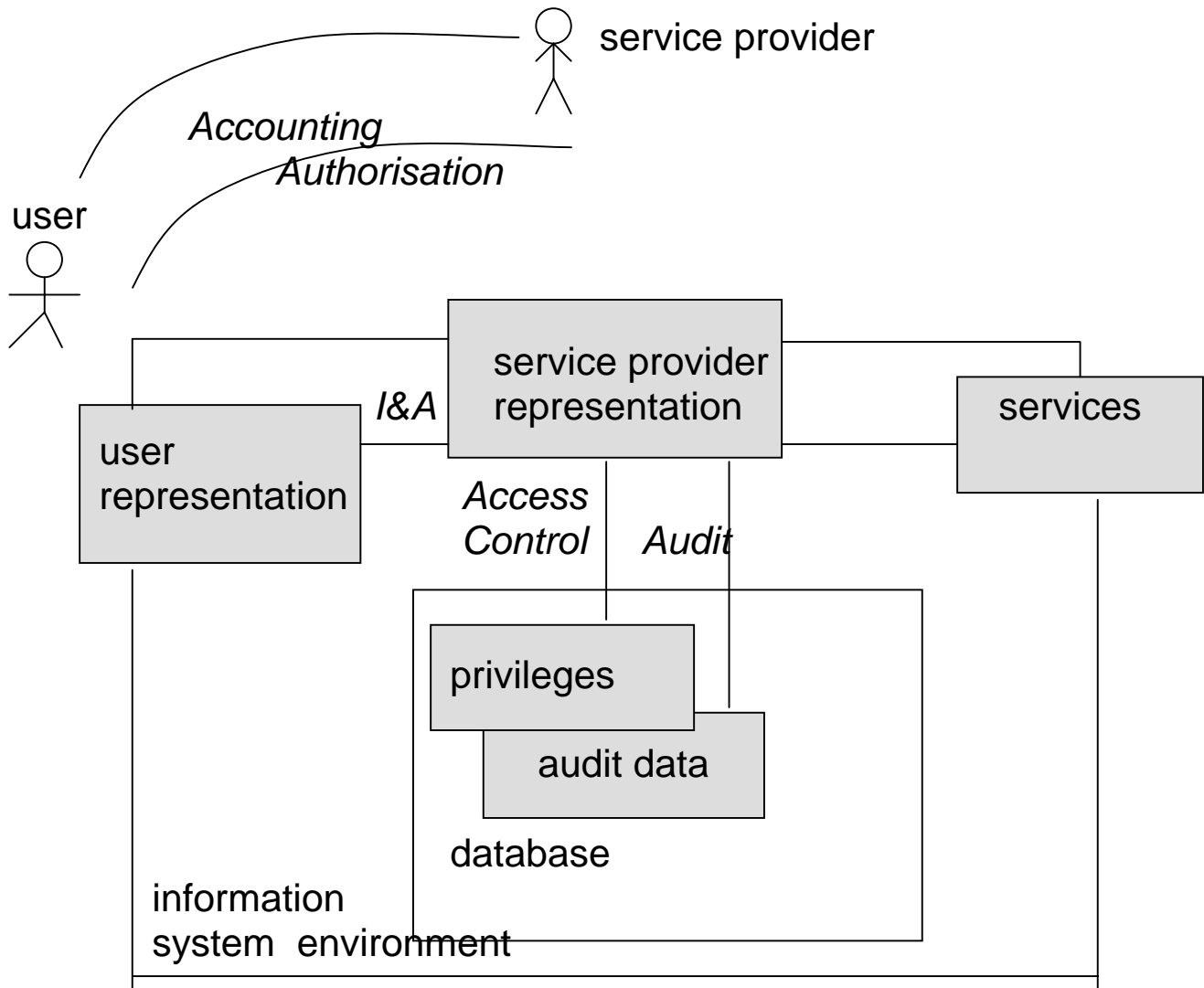
$$RR(sex, marital status) = 6.105 / 100 = 0.0615.$$

Estimation of percentage of re-identifiable individuals:

$$RR(X_1, \dots, X_m) * 100 \%$$

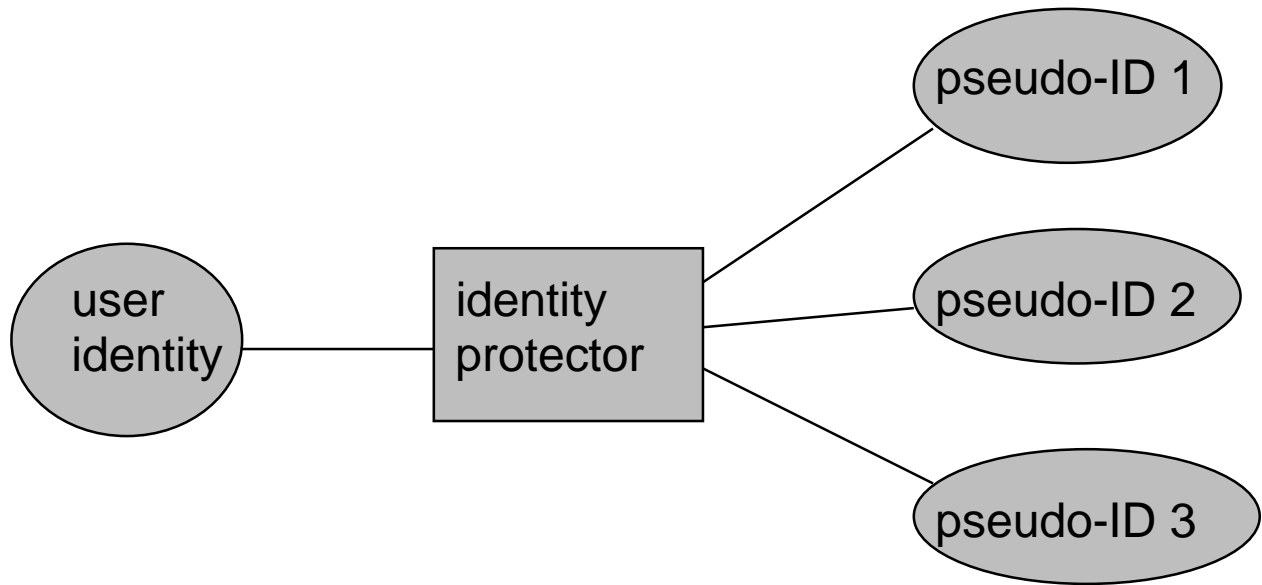
3. PET- Study of the Registratiekamer (1995)

3.1 Need for Identification within the Information System



processes	The use of identity in a conventional system	The use of identity in a privacy system
authorisation	yes	sometimes
identification & authentication	yes	no
access control	yes	no
audit	yes	no
accounting	yes	sometimes

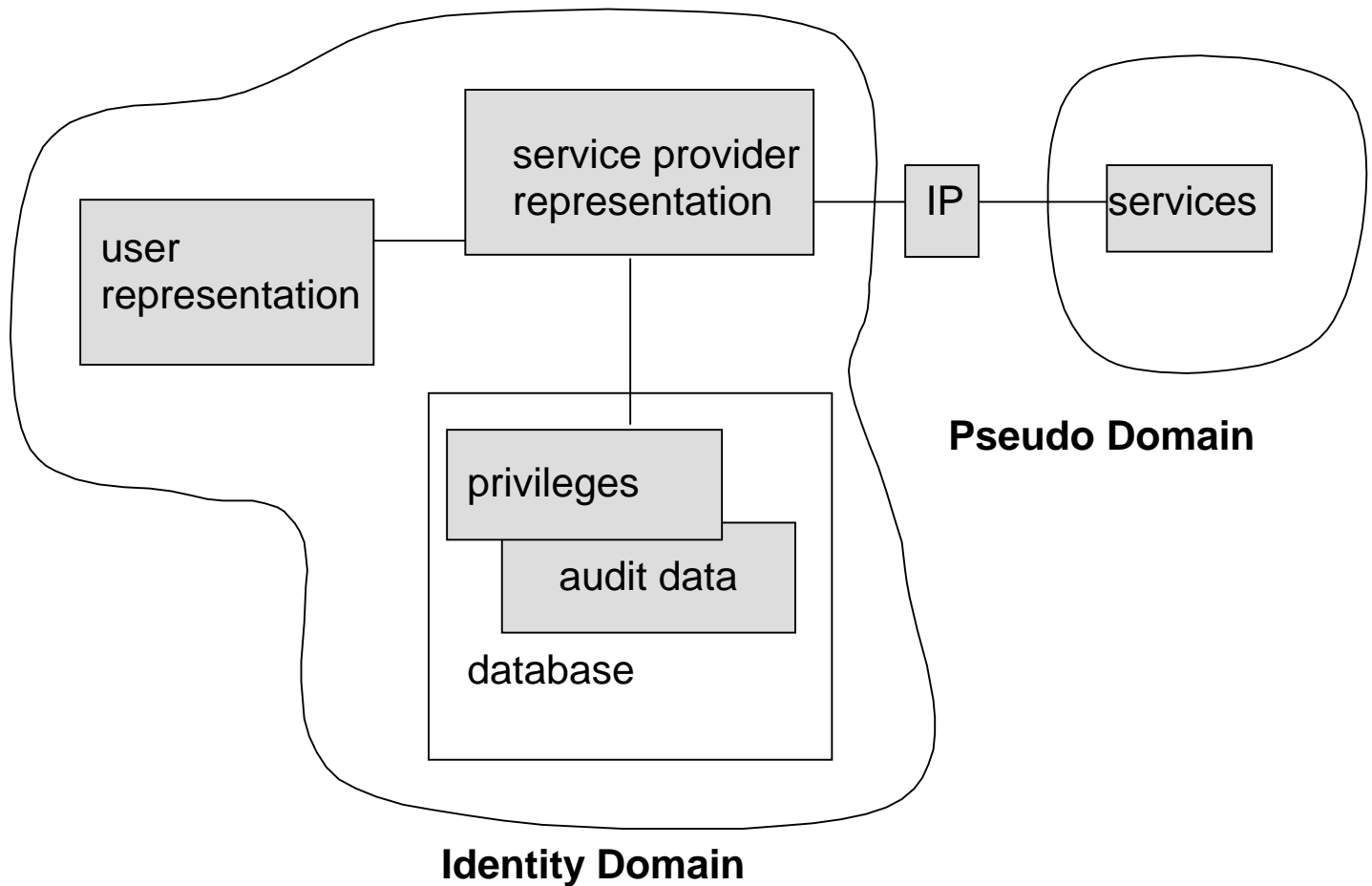
3.2 Concept of the "Identity Protector" [Registriatiekamer 1995]



The identity protector offers the following functions:

- generates pseudo-identities
- translates pseudo-identities into identities and vice versa
- converts pseudo-identities into other pseudo-identities
- reports and controls instances when identity is revealed
- combats misuse

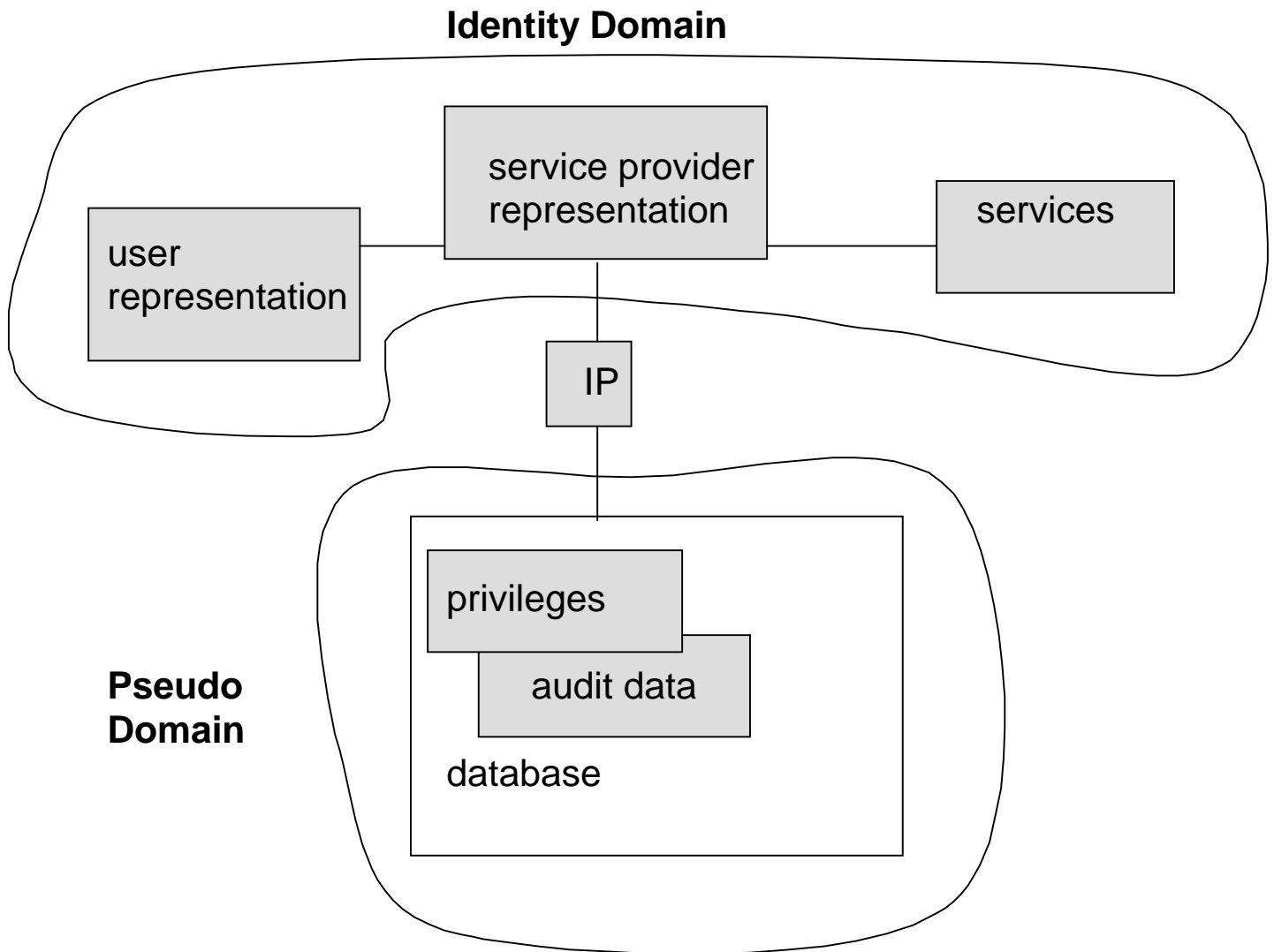
3.3 Cordoning off areas of services and other users



Examples:

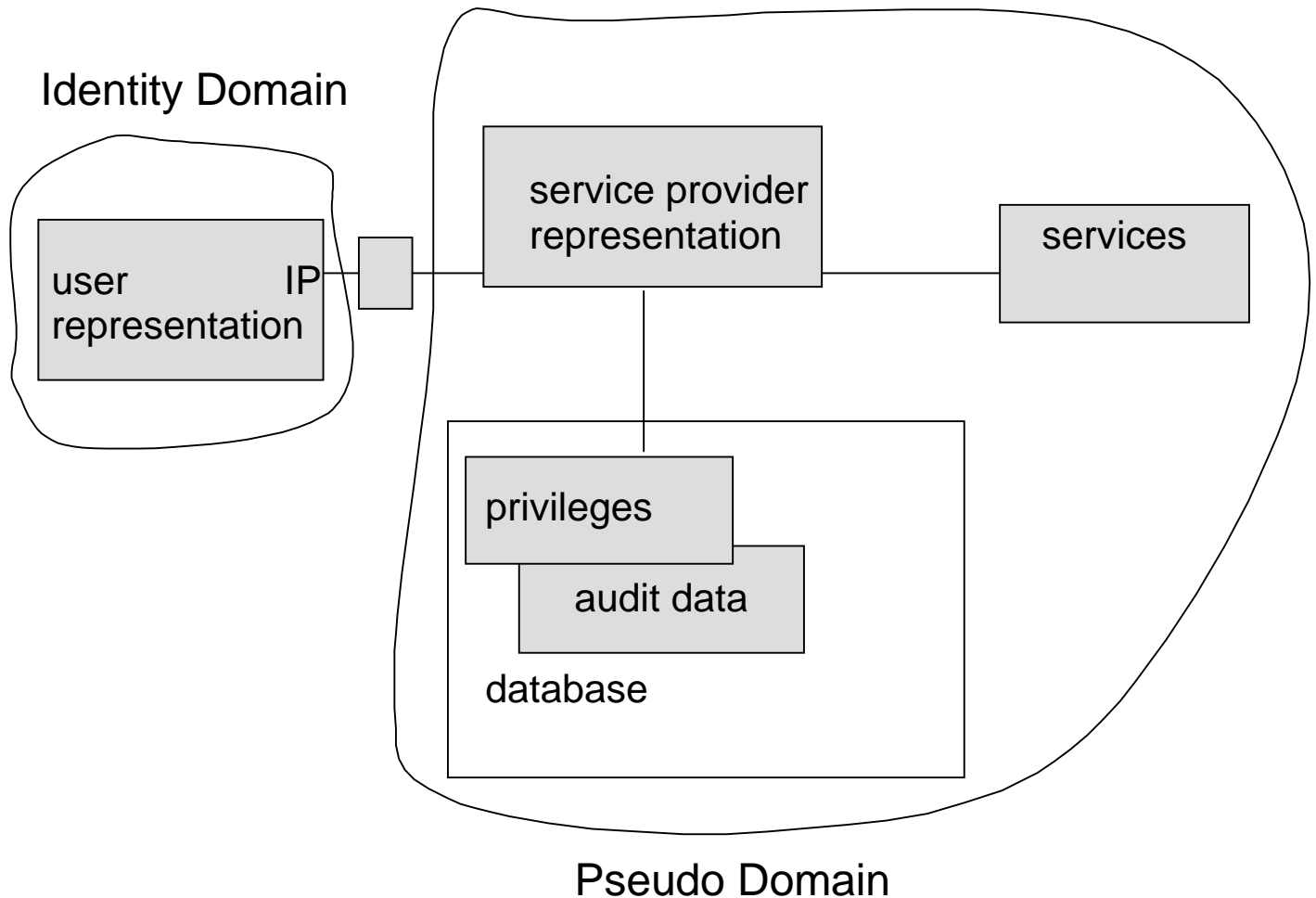
- ISDN calling line identification with possibilities of blocking phone numbers (all or selectively)
- Generation of pseudo-identities by Internet service provider

3.4 Protection of registration in the database



Example:
Pseudonymous Auditing

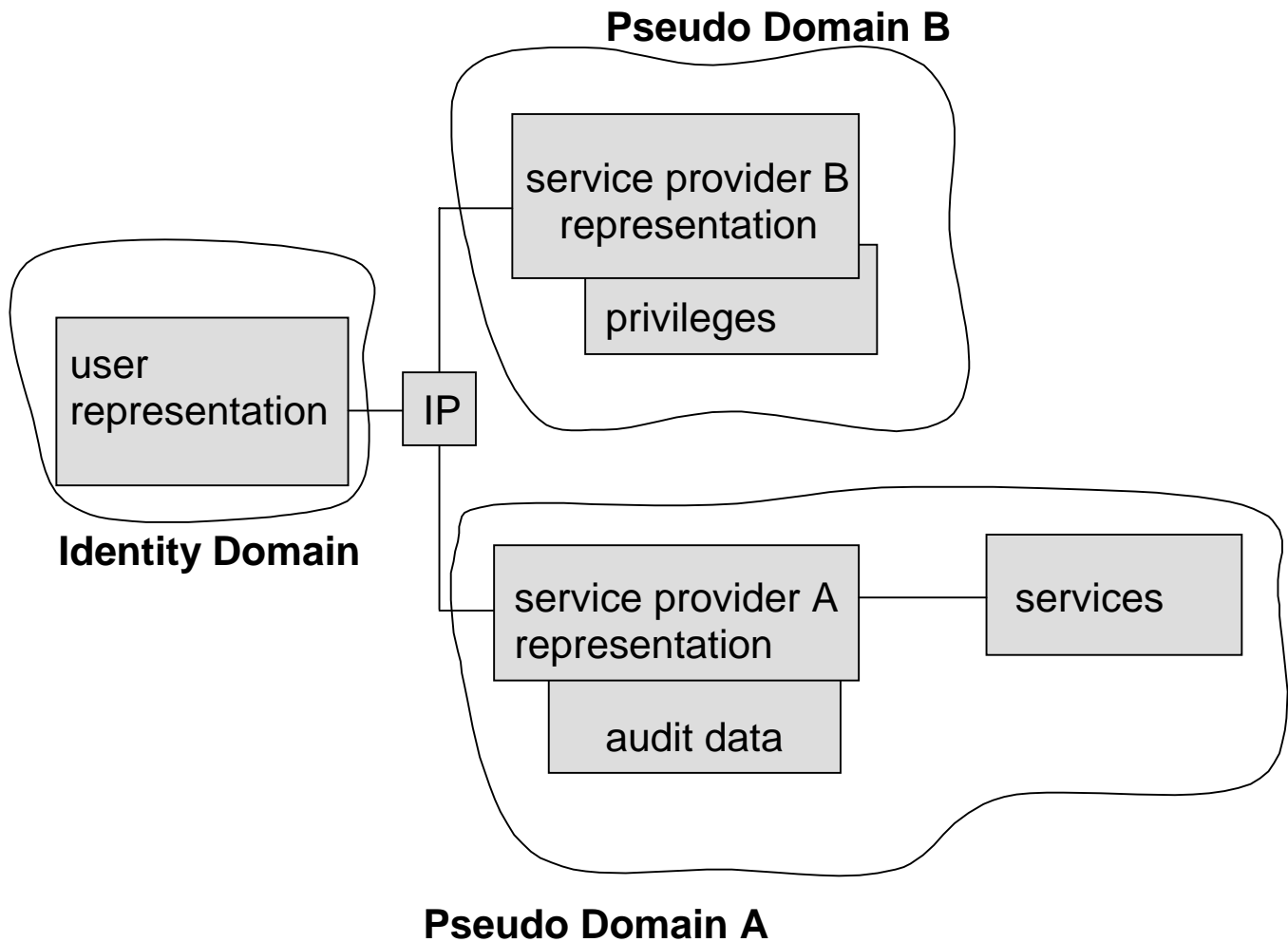
3.5 Cordoning off the entire information system



Examples:

- Pseudonymous system accounts:
department head draws up pseudonymous access profile for new employee, system manager implements authorisations
- DC-nets
- Mix-nets, Anonymous remailers/ browsers

3.6 Situation with several service providers



Example:

David Chaum's ECash

(service provider A: shopkeeper, service provider B: bank)