Statistical Disclosure Control meets Recommender Systems: A practical approach

Fran Casino and Agusti Solanas
{franciscojose.casino, agusti.solanas}@urv.cat
Smart Health Research Group
Universitat Rovira i Virgili

Outline

- **Background**
  - Recommender Systems and Collaborative Filtering
  - Limitations and Countermeasures
  - Statistical Disclosure Control and Privacy-Preserving Collaborative Filtering
  - Evaluation Tools
- **Contributions to Privacy-Preserving Collaborative Filtering**
  - Evaluated Methods
  - Experiments and Comparisons
- **Conclusions**

Recommender Systems

- **Recommender Systems evolve** from the **Knowledge Discovery in Databases** field.
- In a typical recommender system, **people provide opinions/evaluations as inputs**, which the **system** then **aggregates and directs** to appropriate recipients [Resnick et. al.].
- The **main advantage** of Recommender Systems (RS) is that they help us to **deal with/overcome** information **overload**.

P. Resnick, H. Varian, **"Recommender Systems"** *Communications of the ACM 40(3), 56 (1997)*

Collaborative Filtering

Collaborative Filtering (CF) is a **crowdsourcing-based recommender system** which aims to make **suggestions on items** (books, music, movies or routes) **based on preferences of users** that have **already** acquired and/or rated these items.

CF Philosophy

- The **recommendations** provided by CF methods are **based** on the assumption that **similar users** will be interested in the **same items.**
- Users **collaborate** in order to obtain more **quality recommendations**.

CF Families

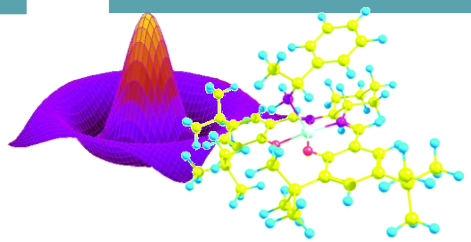
Collaborative Filtering

Memory

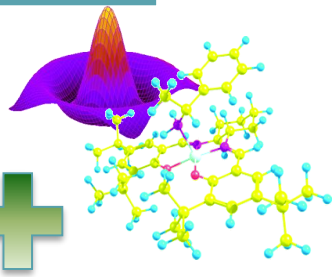
Model

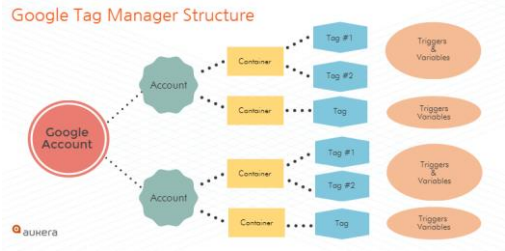
Hybrid

$U \backslash I$	i_1	i_2	$\dots$	i_j	$\dots$	i_{m-1}	i_m
u_1	2	4	$\dots$	1	$\dots$	1	1
u_2	2	3	$\dots$	5	$\dots$	5	2
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_i	1	5	$\dots$	2	$\dots$	4	2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
u_{n-1}	1	2	$\dots$	4	$\dots$	1	1
u_n	3	5	$\dots$	5	$\dots$	1	1

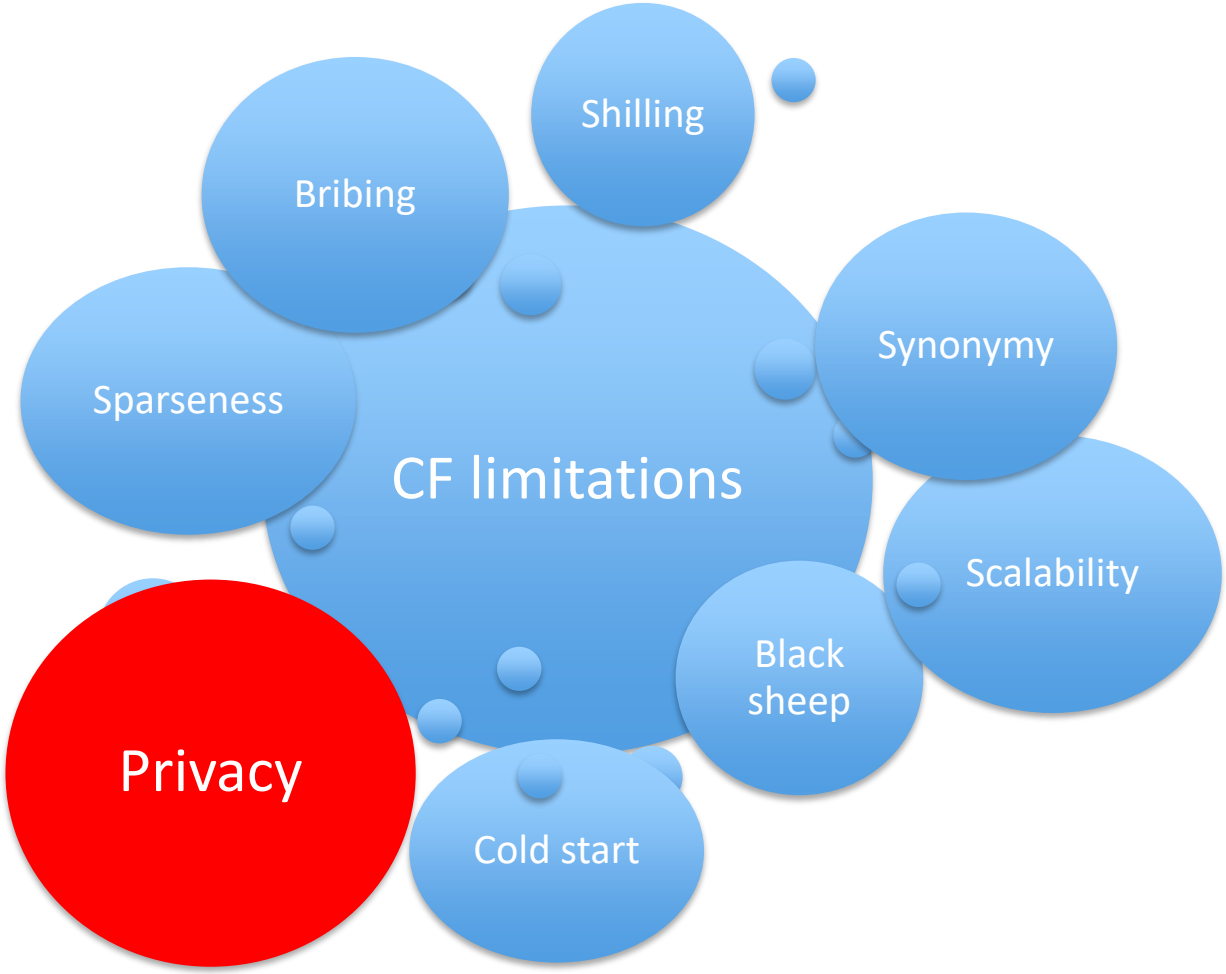


$U \backslash I$	i_1	i_2	$\dots$	i_j	$\dots$	i_{m-1}	i_m
u_1	2	4	$\dots$	1	$\dots$	1	1
u_2	2	3	$\dots$	5	$\dots$	5	2
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_i	1	5	$\dots$	2	$\dots$	4	2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
u_{n-1}	1	2	$\dots$	4	$\dots$	1	1
u_n	3	5	$\dots$	5	$\dots$	1	1

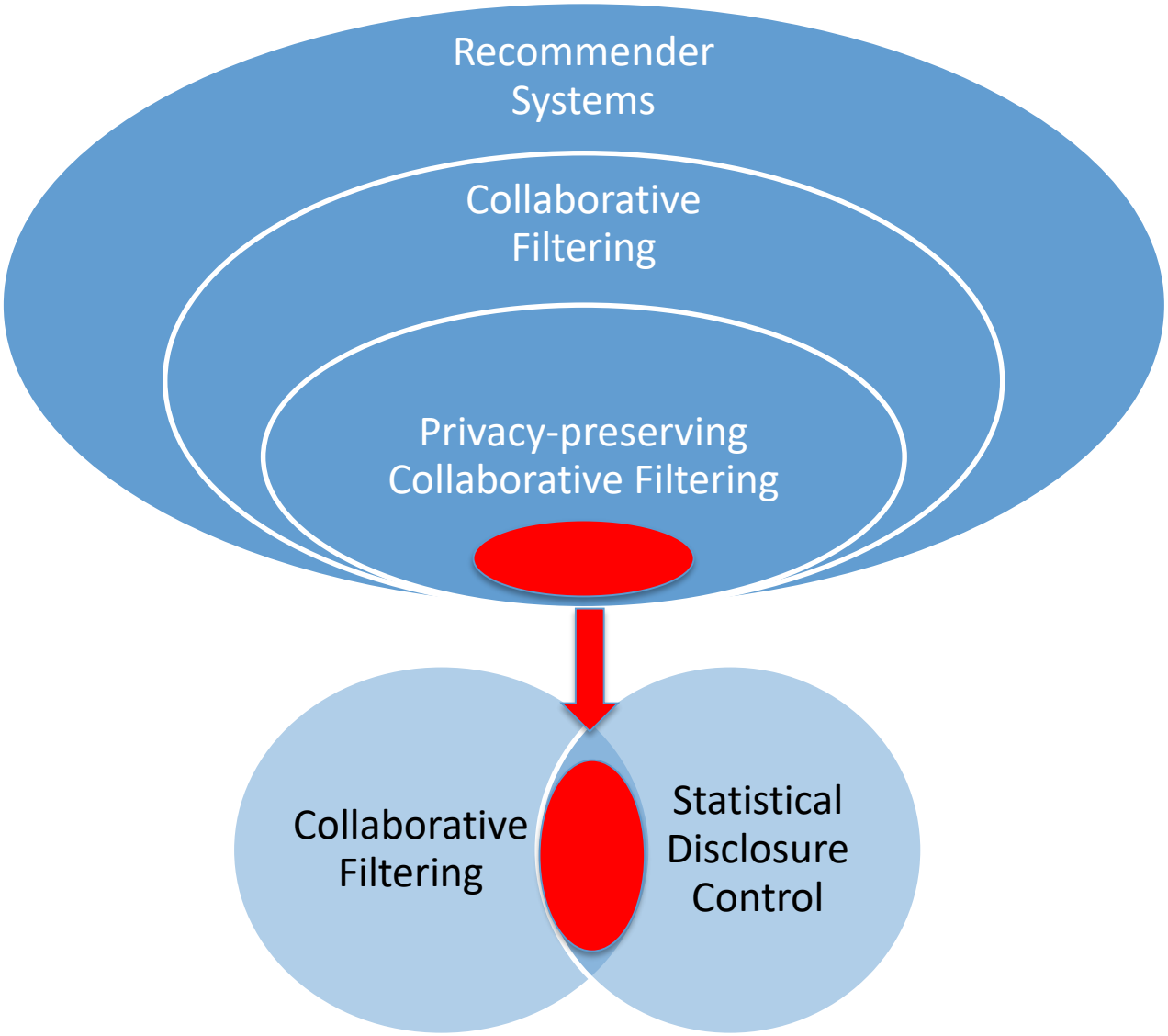




Limitations & Privacy



Collaborative Filtering & Privacy



Statistical Disclosure Control

- Statistical Disclosure Control (SDC, [Hunderpool et. al.]), seeks to **anonymise microdata sets** (i.e. datasets consisting of multiple records corresponding to individual respondents) in order to **prevent their disclosure**.

<u>Types of disclosure</u>

- ***Identity Disclosure*** – Identification of an entity (person, institution).
- ***Attribute Disclosure*** – The intruder finds something new about the target entity.

A. Hundepool, et al. "**Statistical Disclosure Control**". *Wiley, 2012.*

Data Anonymisation Techniques Overview

- Top/bottom coding
- Rounding
- Sampling
- Suppression
- Generalisation
- Limitation of detail
- Anatomisation
- Data swapping
- Noise addition
- **Microaggregation**

Microaggregation

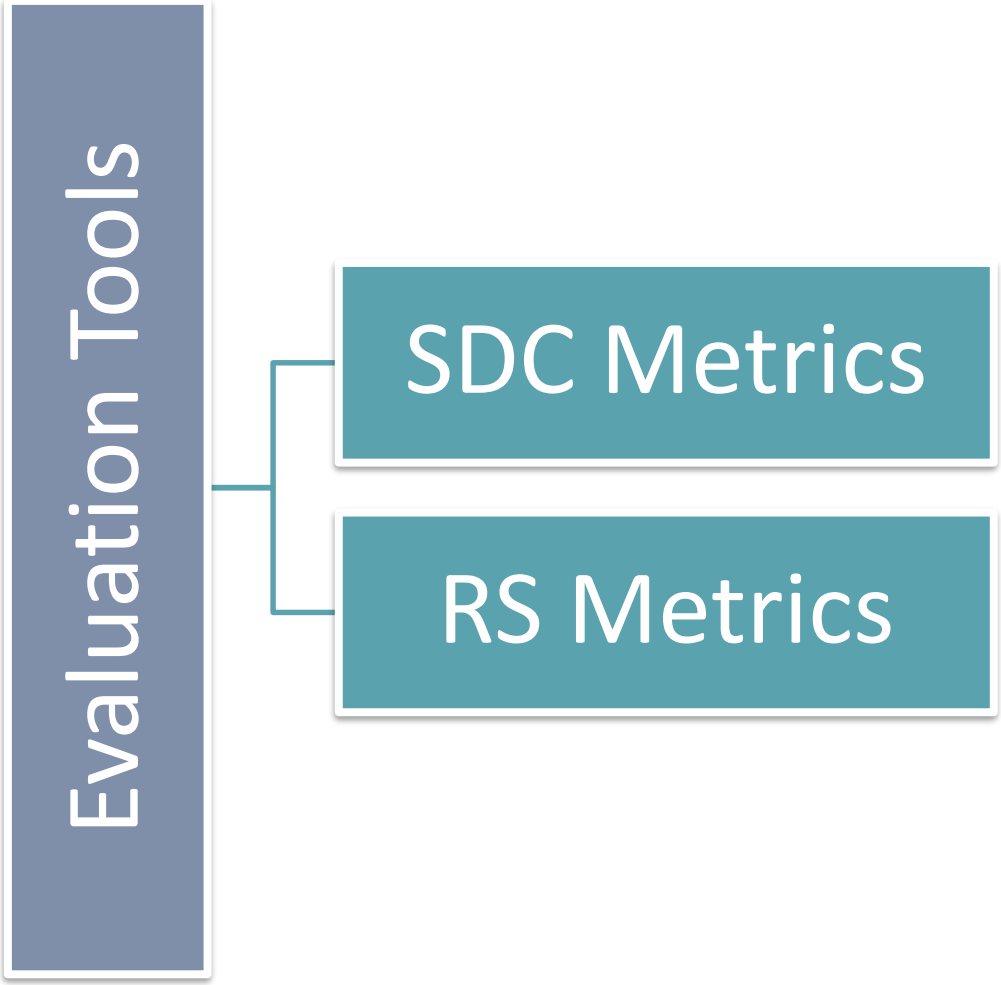
- **Microaggregation is a family of SDC algorithms** for datasets used to prevent against re-identification, which **works in two stages**:

In the case of RS…

- We consider all ratings as quasi-identifiers.
- Therefore, we anonymise all ratings in order to achieve **k-anonymity**.

2. **Records** within each cluster **are replaced by** a representative of the cluster, typically **the centroid** record (i.e. the average of the cluster).

Evaluation Tools



SDC – Information Loss

The quantity of **information which exist in the initial** microdata and because of disclosure control methods **does not occur in masked microdata** [Willemborg et. al.].

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mathbf{x}}_i)'(x_{ij} - \bar{\mathbf{x}}_i)$$

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mathbf{x}})'(x_{ij} - \bar{\mathbf{x}})$$

$$IL = \frac{SSE}{SST}$$

Willemborg L., Waal T. "**Elements of Statistical Disclosure Control**". *Springer Verlag.*

SDC – Disclosure Risk

- The **risk** that a given form of **disclosure** will **arise** if a masked microdata is **released** [Chen et. al.].
  - Value/attribute disclosure
  - Identity disclosure

- ***Individual measures*** - The risk per record or the probability of **correctly re-identifying a unit**. [Willemborg et. al.]

- ***Global measures*** - The risk for the **entire dataset**. Number of **correct re-identifications** according to a linking measure. [Domingo-Ferrer et. al.]

Chen G., Keller-McNulty S. "**Estimation of Deidentification Disclosure Risk in Microdata"**. *Journal of Official Statistics, Vol 14. No. 1, 79-95.*

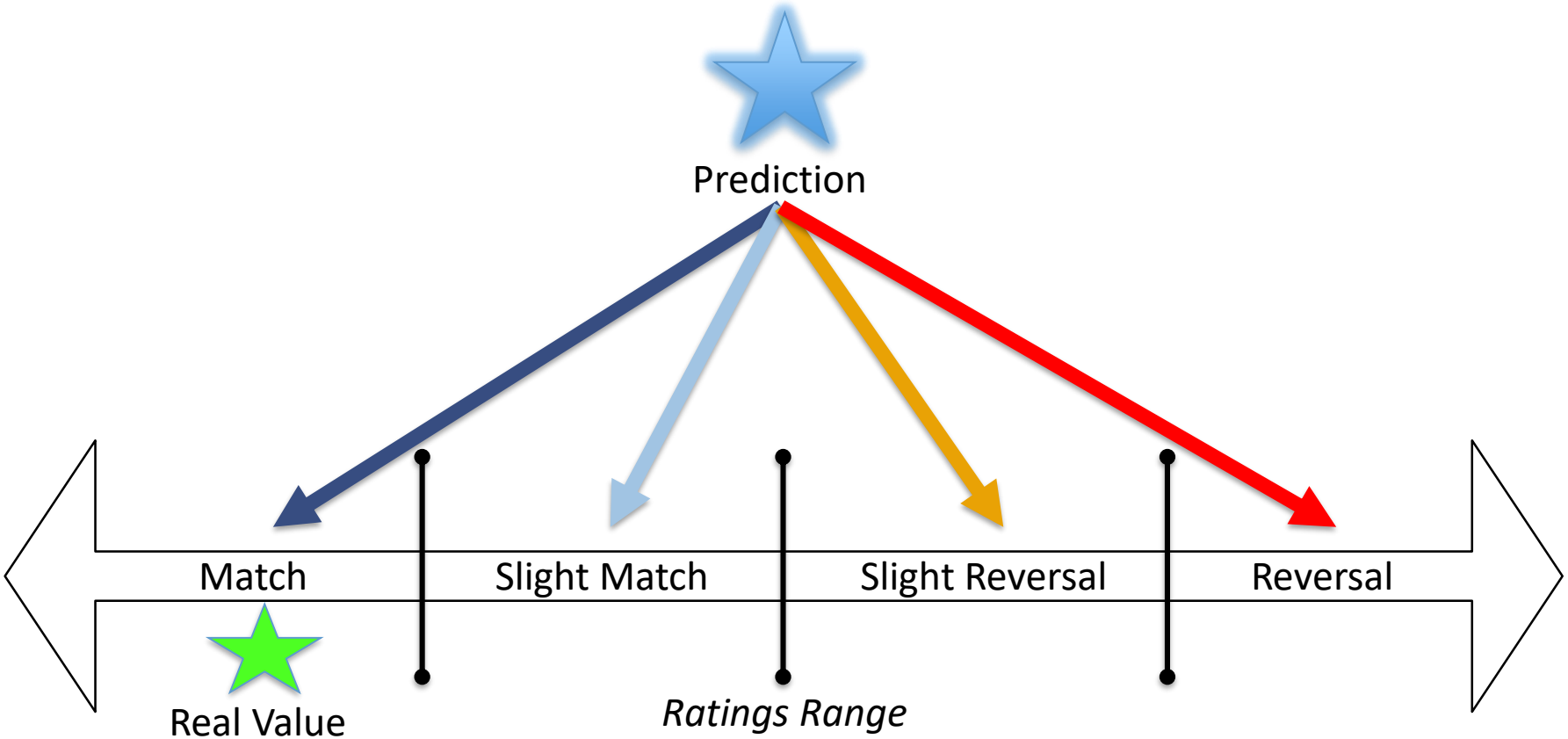
Willemborg L. Waal T. "**Elements of Statistical Disclosure Control",** *Springer Verlag.*

Domingo-Ferrer J. Torra V. **"Disclosure Risk Assessment in Statistical Microdata Protection Via Advanced Record Linkage"** *Statistics and Computing, vol 13, no 4, pp- 343-354*

RS Metrics

$$MAE = \frac{\sum_{i=1}^{n} |p_i - r_i|}{n}$$

$$\text{T-score} = \frac{MAE + DR}{2}$$



Outline

- **Background**
  - Recommender Systems and Information Overload
  - Limitations of Collaborative Filtering and Countermeasures
  - Statistical Disclosure Control and Privacy-Preserving Collaborative Filtering
  - Evaluation Tools
- **Contributions to Privacy-Preserving Collaborative Filtering**
  - Evaluated Methods
  - Experiments and Comparisons
- **Conclusions**

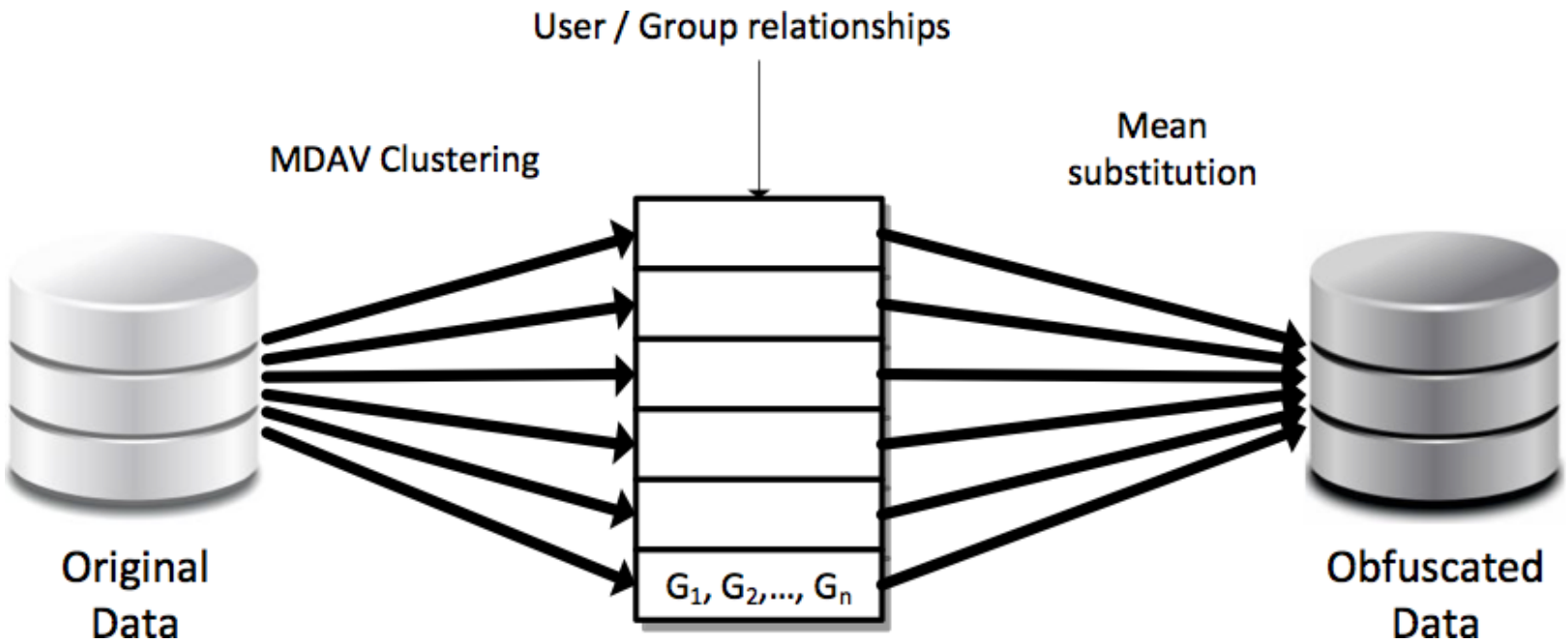
PPCF Methods

- Gaussian Noise Addition with zero mean. $\mathcal{N}(0, \sigma)$
- Maximum Distance to Average Vector (MDAV) [Domingo-Ferrer et. al.]
- Variable MDAV (V-MDAV) [Solanas et. al.]

J. Domingo-Ferrer and J. M. Mateo-Sanz. "**Practical data-oriented microaggregation for statistical disclosure control"**, *IEEE Transactions on Knowledge and data Engineering*, 2002.

A. Solanas and A. Martínez-Ballesté. **V-MDAV : A Multivariate Microaggregation With Variable Group Size.** *Seventh COMPSTAT Symposium of the IASC, 2006.*

MDAV



Fixed-size groups & k-anonymity

V-MDAV

- After each iteration, a **heuristic evaluates** whether to **include** a new record *r* to a group:
  - If *r* is **closer** to the **actual group** than to the rest of records, according to its **distance** and a **gain factor**.
  - If the **actual group size is < 2k-1**, because the **optimal k-partition** is achieved when groups consists of **k to 2k-1** records [Domingo-Ferrer et. al.].
  - The **gain factor** can be **tuned** in order to fit the **data distribution**.

Variable-sized Groups & k-anonymity

J. Domingo-Ferrer and V. Torra. **Ordinal, continuous and heterogenerous k-anonymity through microaggregation**. *Data Mining and Knowledge Discovery, 11(2):195–212, 2005.*

Data Preprocessing

- Matrices are **filled** and **stantardised** (z-scores).

$$z - score = \frac{x_i - \mu}{\sigma}$$

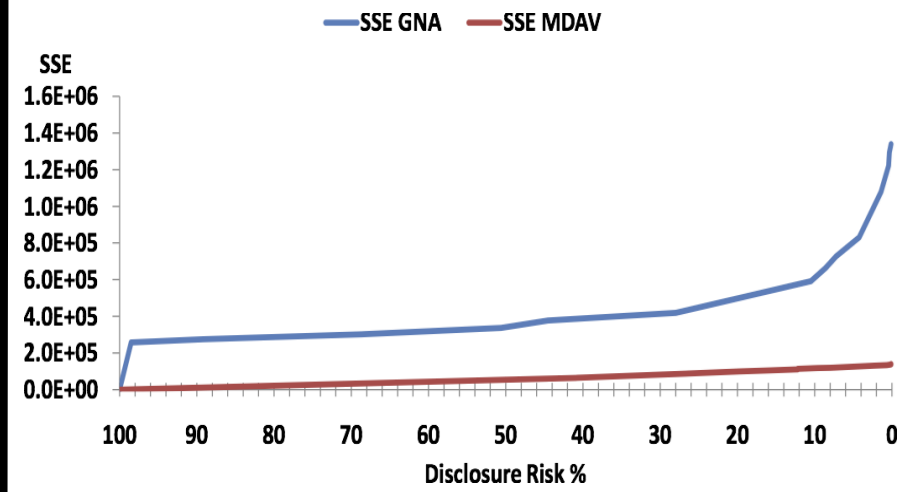
where x_i is the *i*-th value of item *x* and μ and σ are the mean and the standard deviation of item *x*, respectively.

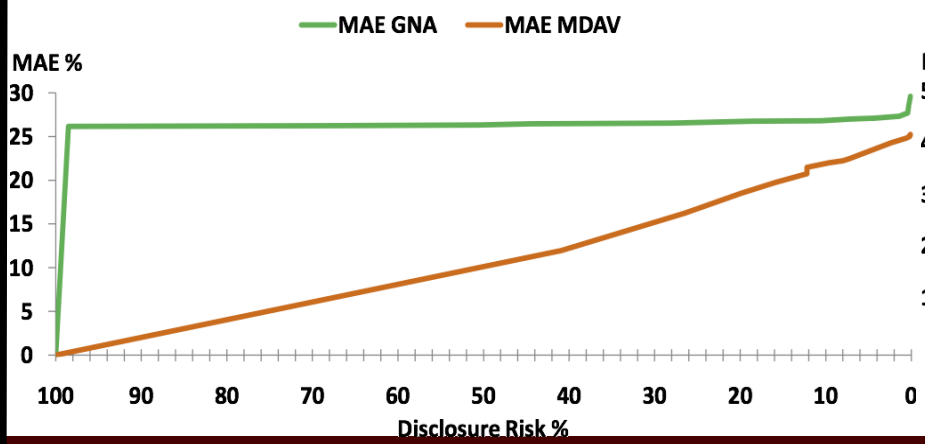
- Next, the corresponding **method is applied**.

- **Comparison** between methods in terms of data utility and privacy **using well-known metrics**.

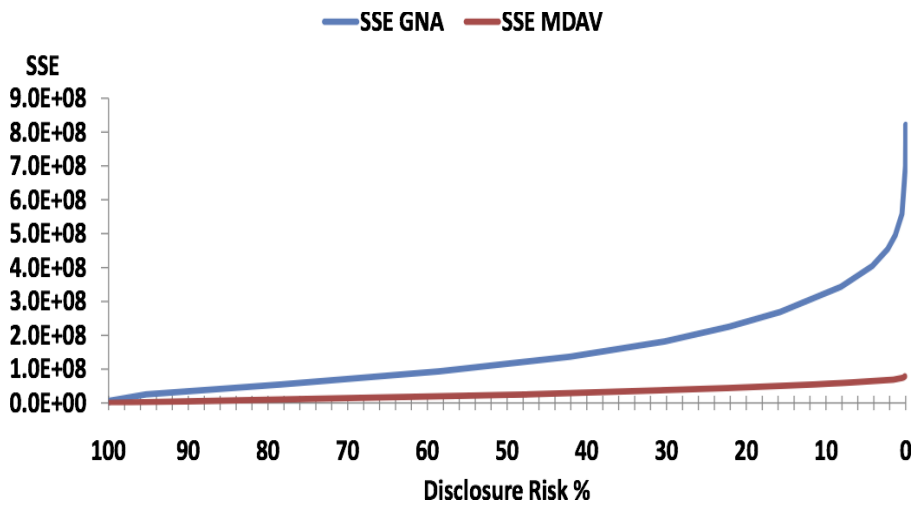
GNA & MDAV

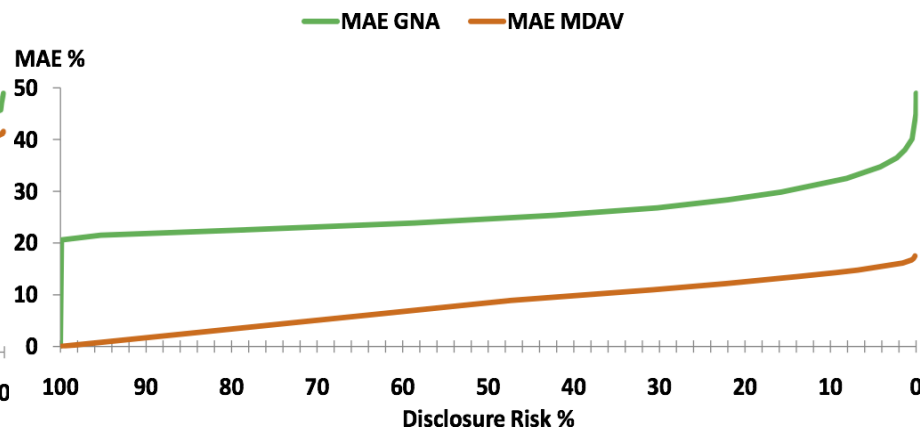
Movielens 100k



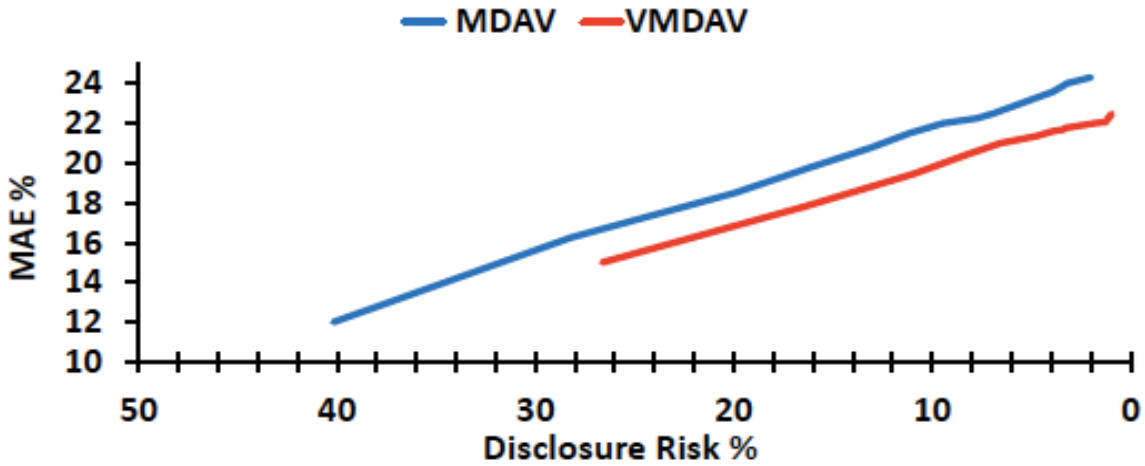


Jester

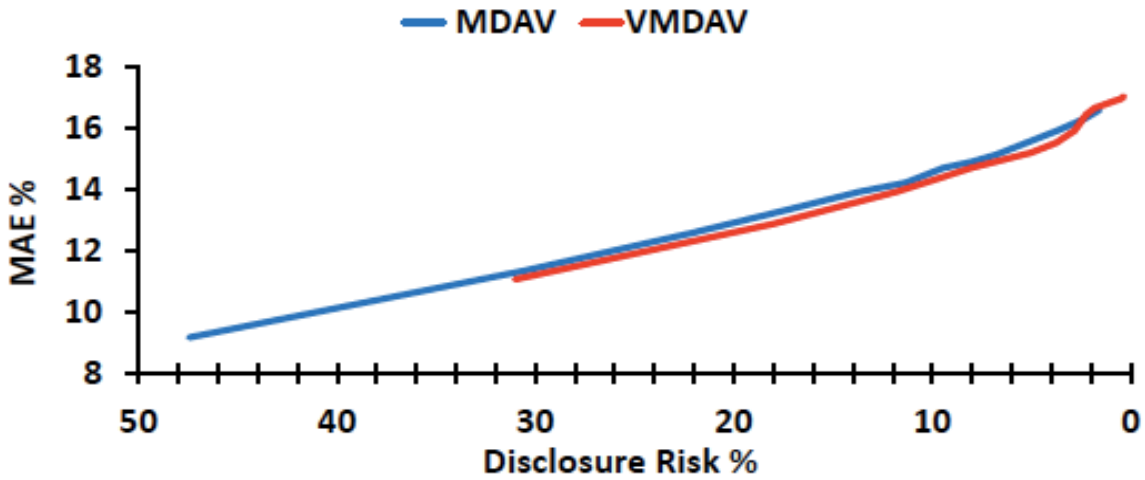




MDAV & V-MDAV (I)

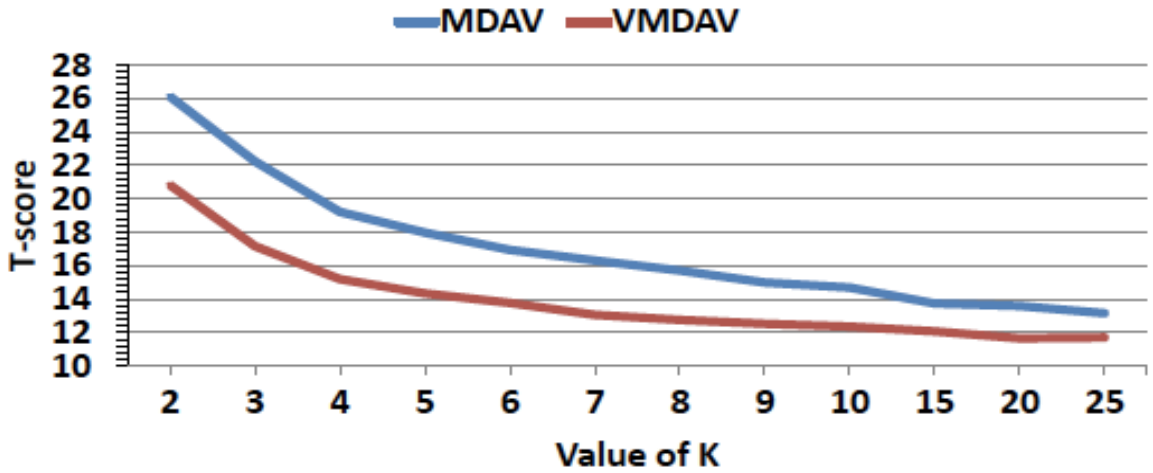


(a) MAE and DR comparison - Movielens 100k

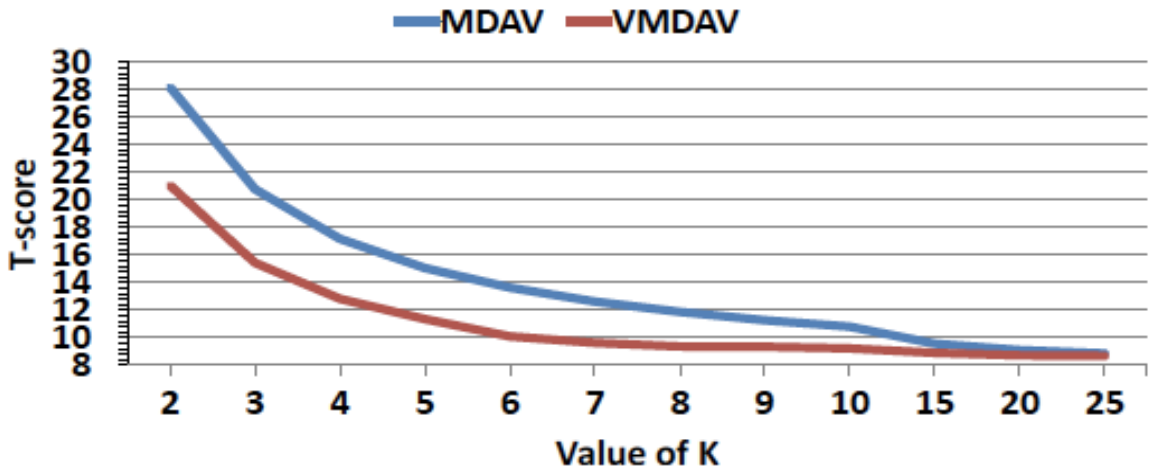


(b) MAE and DR comparison - Jester

MDAV & V-MDAV (II)

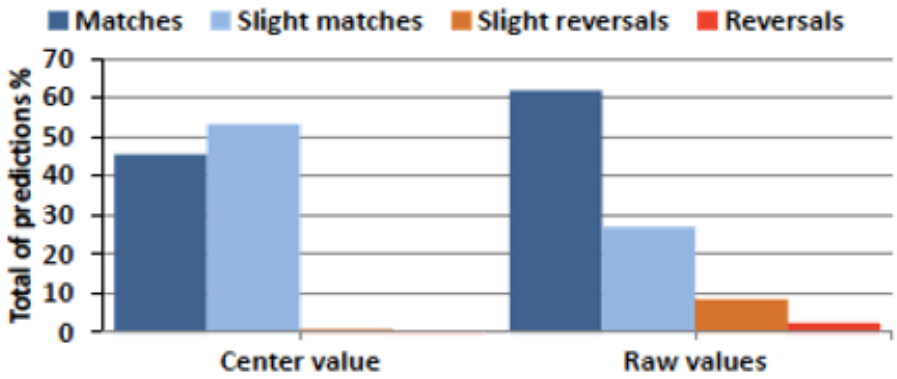


(a) Efficiency of the applied noise - Movielens 100k

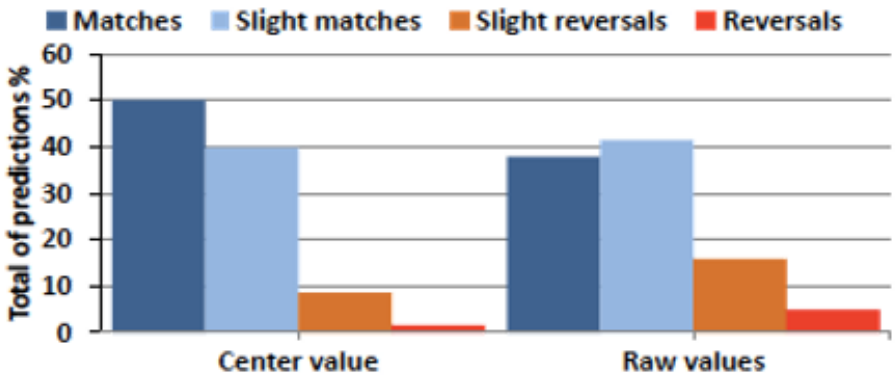


(b) Efficiency of the applied noise - Jester

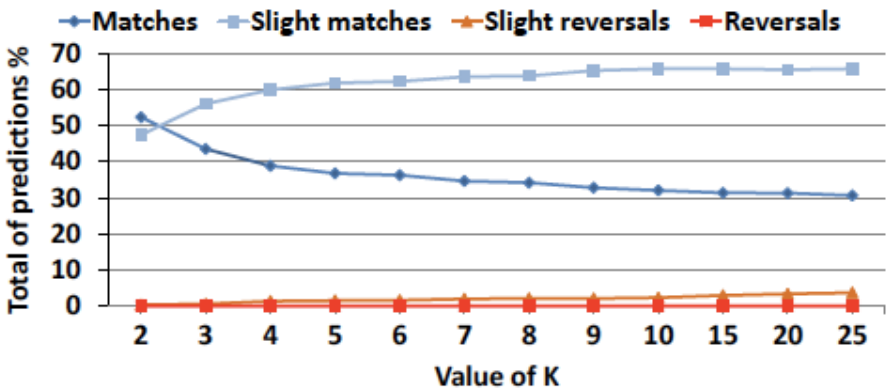
Behavioural Precision B/A



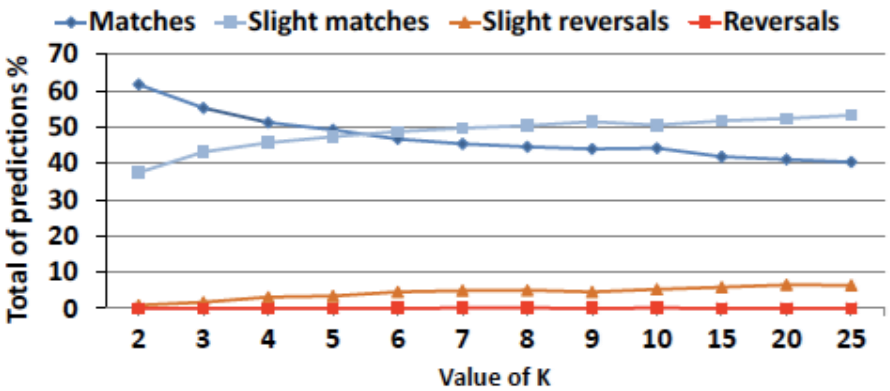
(c) 4L behavioural prediction accuracy - Movielens 100k



(d) 4L behavioural prediction accuracy - Jester



(c) 4L behavioural prediction accuracy - Movielens 100k



(d) 4L behavioural prediction accuracy - Jester

Conclusions - Highlights

- Despite the great **advantages** of using CF, we have **highlighted** its downside regarding users' **privacy**.
- We have analysed/discussed how **V-MDAV** obtains **better** results and provides both **more privacy and data usability** than well-known methods such as **MDAV** and **Gaussian noise addition**.
- Both **microaggregation-based proposals achieve k-anonymity**, which guarantees privacy by design, a feature **not offered by GNA**.
- Moreover, for **low cardinality values**, **recommendations** were **more accurate** than these obtained when using **data without obfuscation**, showing the efficacy of our proposal.
- The use of **behavioural measures** allowed us to **better analyse** data and increase its usability.

Statistical Disclosure Control meets Recommender Systems: A practical approach

Fran Casino and Agusti Solanas
{franciscojose.casino, agusti.solanas}@urv.cat
Smart Health Research Group
Universitat Rovira i Virgili