

Tree-based Microaggregation for the Anonymization of Search Logs

Guillermo Navarro-Arribas, Vicenç Torra

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research

`{guille,vtorra}@iia.csic.es`

SAIAW-2009

Outline

- 1 Introduction
- 2 Microaggregation
- 3 Microaggregation of Search logs
- 4 Conclusion

Web search logs

A search log normally has the form:

user id, search terms, timestamp, rank, clicked url

Example (AOL)

```
24969374 orioles tickets 2006-05-31 12:31:57 2 http://www.greatseats.com
24969423 jennifer craford my space.com 2006-05-31 19:15:02
24969423 jennifer crawford my space.com 2006-05-31 19:16:05
144 boston redsoxweb page.com 2006-03-28 17:51:55
144 www.bostonredsox 2006-03-28 18:12:26 1 http://boston.redsox.mlb.com
```

- Generated by search engines.
- Very useful for profiling, marketing, and research community.

Privacy problems in search logs

- Use of sensitive information in search terms.
 - Social security number, ...
 - Normally removed (sort of manually)
- Users can be re-identified even with anonymised IDs.
 - Case of user 4417749 (AOL): Thelma Arnold



Search log anonymisation

The good

- Logs can be safely stored for future analysis.
- Log analysis can be outsourced or made public.

The bad

- Anonymisation involves losing information.

Towards Search log anonymisation

Our approach

- Use Statistical Disclosure Control techniques
 - SDC successfully used in PPDM
 - Keeps (most) statistical information
- There is always a trade off:

utility vs. privacy

Microaggregation

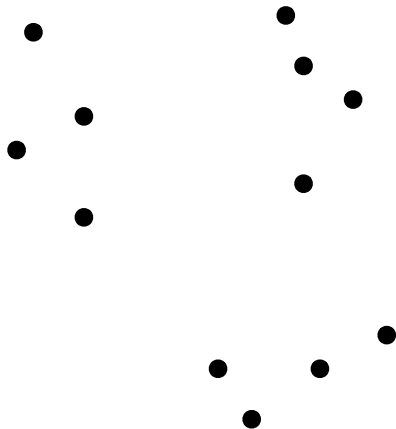
Microaggregation one of the most used SDC techniques

- Clustering with parameter k
 - k minimum number of element per cluster.
- Can provide **k-anonymity**

Each record is indistinguishable with at least $k - 1$ other records

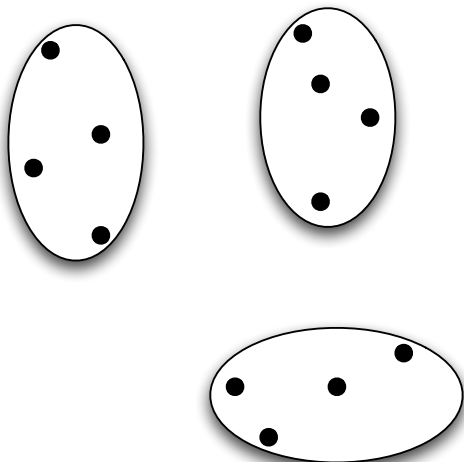
- Multivariate microaggregation is NP-hard
 - Need for heuristics \Rightarrow MDAV (Maximum Distance to Average Vector)

Microaggregation explained



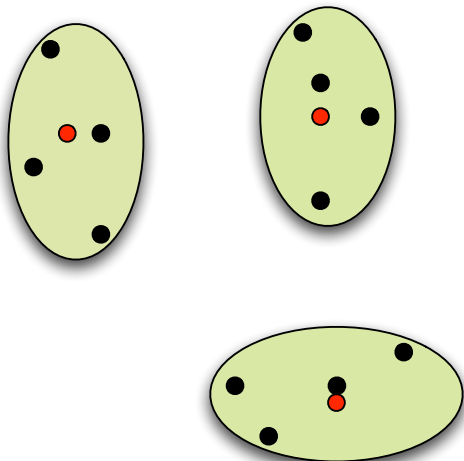
- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.
- 3 Substitute points by their centroid.

Microaggregation explained



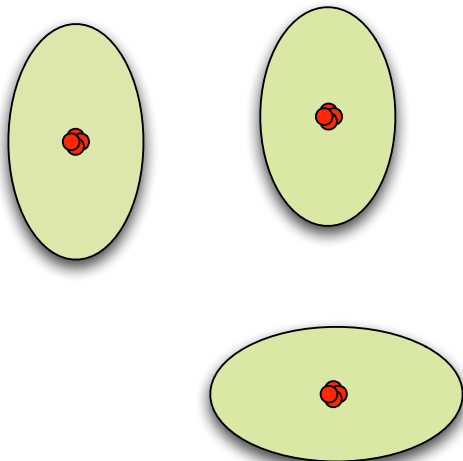
- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.
- 3 Substitute points by their centroid.

Microaggregation explained



- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.
- 3 Substitute points by their centroid.

Microaggregation explained



- 1 Make clusters \Rightarrow **distance** function.
- 2 Calculate centroid \Rightarrow **aggregator** operator.
- 3 Substitute points by their centroid.

Microaggregation of Search logs

- Provide suitable **distance** function and **aggregator** operator.
- Maintain the **order** of queries for the same user.
- Attempt to provide *k-anonymity*.

Record-based microaggregation of search logs

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	t_0	U_0
id_0	$\{\mu_0, \mu_2\}$	t_1	U_1
id_1	$\{\mu_0\}$	t_2	U_0
id_1	$\{\mu_0, \mu_2, \mu_3\}$	t_3	U_2

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2

Record-based microaggregation of search logs

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	t_0	U_0
id_0	$\{\mu_0, \mu_2\}$	t_1	U_1
id_1	$\{\mu_0\}$	t_2	U_0
id_1	$\{\mu_0, \mu_2, \mu_3\}$	t_3	U_2

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2

Record-based microaggregation of search logs

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	t_0	U_0
id_0	$\{\mu_0, \mu_2\}$	t_1	U_1
id_1	$\{\mu_0\}$	t_2	U_0
id_1	$\{\mu_0, \mu_2, \mu_3\}$	t_3	U_2

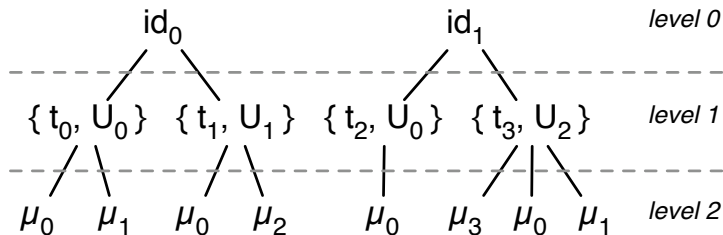
id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_0	$\{\mu_0, \mu_1\}$	$t_{0.5}$	U_0
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2
id_1	$\{\mu_0, \mu_2\}$	$t_{2.5}$	U_2

Search logs as ordered trees \Rightarrow query-tree

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	t_0	U_0
id_0	$\{\mu_0, \mu_2\}$	t_1	U_1
id_1	$\{\mu_0\}$	t_2	U_0
id_1	$\{\mu_0, \mu_2, \mu_3\}$	t_3	U_2

Search logs as ordered trees \Rightarrow query-tree

id	query terms	timestamp	clickedURL
id_0	$\{\mu_0, \mu_1\}$	t_0	U_0
id_0	$\{\mu_0, \mu_2\}$	t_1	U_1
id_1	$\{\mu_0\}$	t_2	U_0
id_1	$\{\mu_0, \mu_2, \mu_3\}$	t_3	U_2



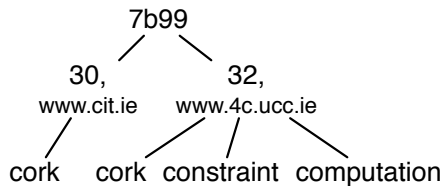
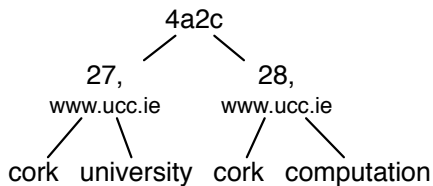
Query-tree distance

Ordered tree edit distance

- Cost function γ :
 - Penalise operations that transform nodes of different levels.
- Cluster users with similar search patterns.

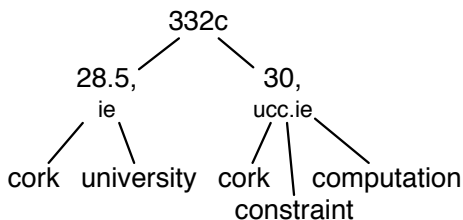
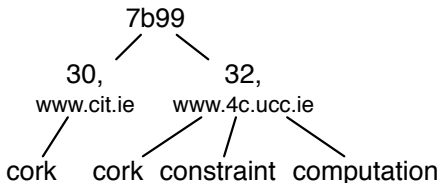
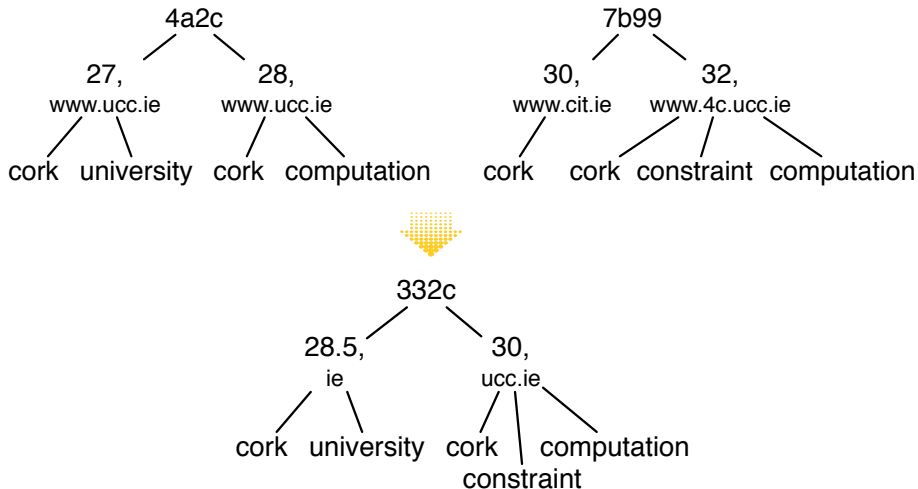
Query-tree aggregation

→ Aggregate nodes by level respecting the order.



Query-tree aggregation

→ Aggregate nodes by level respecting the order.



Tree-based microaggregation ($k = 2$)

id	query terms	tstamp	clickedURL
4a2c	{ cork, university }	27	www.ucc.ie
4a2c	{ cork, computation }	28	www.ucc.ie
7b99	{ cork }	30	www.cit.ie
7b99	{ cork, constraint, computation }	32	www.4c.ucc.ie

id	query terms	tstamp	clickedURL
332c (4adc)	{ cork, university }	28.5	ie
332c (4adc)	{ cork, constraint, computation }	30	ucc.ie
332c (7b99)	{ cork, university }	28.5	ie
332c (7b99)	{ cork, constraint, computation }	30	ucc.ie

Note (number of users): $Users = Users' * k$

Tree-based microaggregation ($k = 2$)

id	query terms	tstamp	clickedURL
4a2c	{ cork, university }	27	www.ucc.ie
4a2c	{ cork, computation }	28	www.ucc.ie
7b99	{ cork }	30	www.cit.ie
7b99	{ cork, constraint, computation }	32	www.4c.ucc.ie

id	query terms	tstamp	clickedURL
332c (4adc)	{ cork, university }	28.5	ie
332c (4adc)	{ cork, constraint, computation }	30	ucc.ie
332c (7b99)	{ cork, university }	28.5	ie
332c (7b99)	{ cork, constraint, computation }	30	ucc.ie

Note (number of users): $Users = Users' * k$

Conclusions

- Use of microaggregation to protect search logs.
- Protection **a posteriori**.
- Achieve k -anonymity at user level.
 - There are at least k indistinguishable users.
- We expect relatively low loss of information for big datasets.
 - AOL search logs are 20M web queries from 650k users.
- Still evaluating and fine-tuning the method.

Tree-based Microaggregation for the Anonymization of Search Logs

Guillermo Navarro-Arribas, Vicenç Torra

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research

`{guille,vtorra}@iia.csic.es`

SAIAW-2009