State Statistical Institute Berlin-Brandenburg

Morpheus – Remote Data Access with a Quality Measure

Joint UNECE /Eurostat Work Session on Statistical Data Confidentiality

26 – 28 October 2011

Jörg Höhne / Julia Höninger Research Data Centre

Need for automatic output checking

Current situation: remote processing

- Researchers want results fast
- Manual control time-consuming

→ In need for a system that can check results automatically

- On the way to real remote access
- Work part of the project „infinitE - an informational infrastructure for the E-Science Age“
- funded by German Federal Ministry of Education and Research

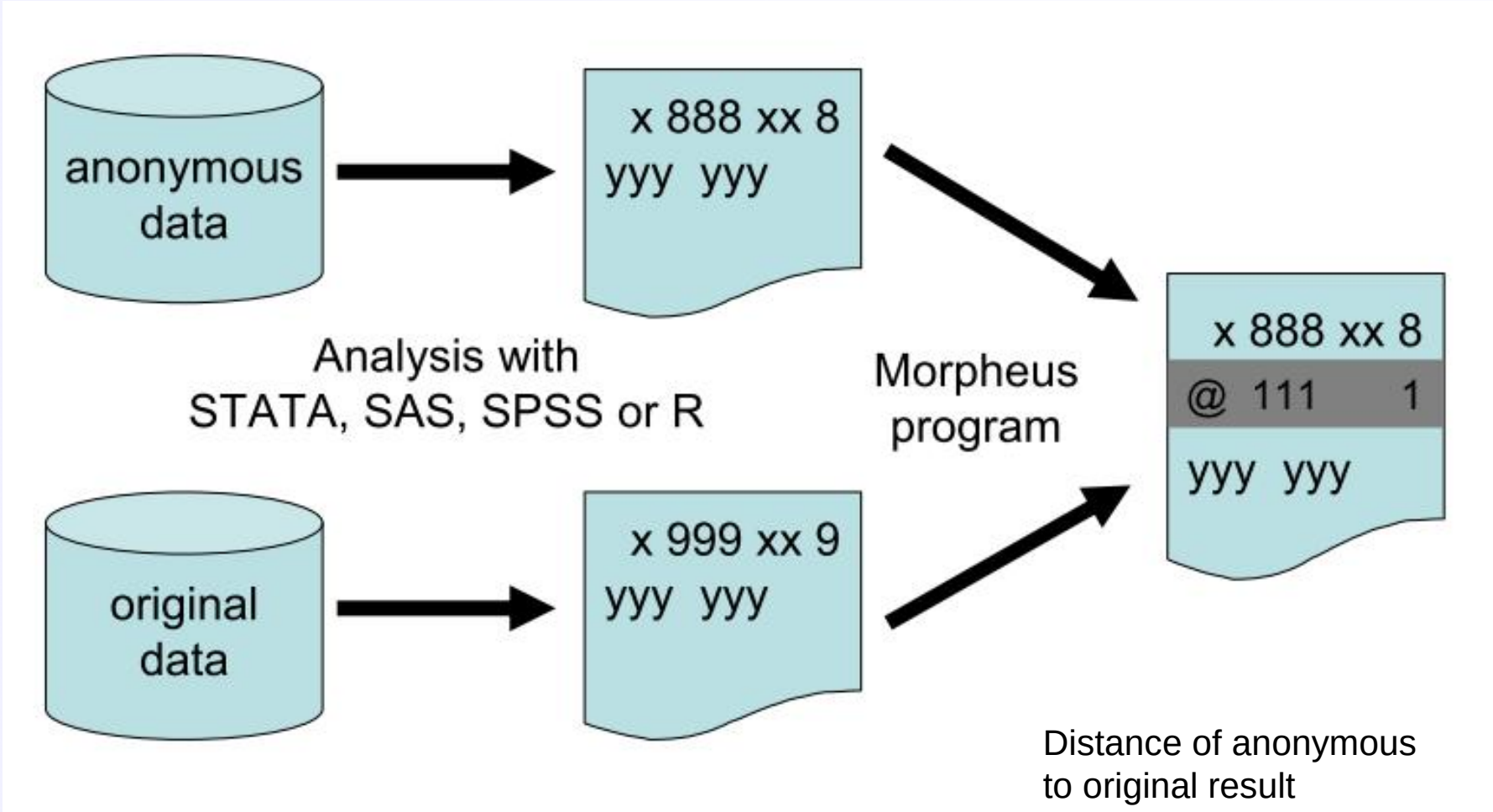
Morpheus – general idea

- Users run calculations on anonymous data
- Additional to anonymous results a measure of quality
- Results (almost) in real time

For final publication:

- Results on original data, checked manually

Morpheus – functionality



Advantages for users

Compared to current remote processing:

- Fast return of results
- Browsing in data set allowed
- (almost) all analysis permitted

Compared to scientific use files:

- Measure of quality for all results

Challenge:

- Reading output

Challenges for RDC staff

- Generate anonymous files for all data sets
  - Census data
  - Business data
  - Panel data
- Perturbative anonymisation methods or synthetic data

→ Reduction of manual work load for RDC staff if the anonymous data set is good enough

Morpheus – a first prototype

- First case study with Morpheus
- Statistical software Stata
- Anonymous file for business statistics
- Program from an actual user

```
. tabstat expshare2000, stats(N mean sd p25 p50 p75)

    variable |         N      mean        sd       p25       p50       p75

-------------+--------------------------------------------------------------

expshare2000 |     48305  14.71019  22.10718         0  1.776615  22.77032

@            |         0   0.06056   0.22858         0  0.011463   0.11264

-------------+--------------------------------------------------------------
```

- Works for output from SPSS and SAS

Morpheus - regression output

```
. xtreg lnapro export pers perssq hc, fe r
```

lnapro	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
@					0	
export	.0882477	.0076721	11.50	0.000	.0732102	.1032851
@	0.0002211	0.0000887	0.15	0.000	0.0004296	0.0000251
pers	-.0000162	2.99e-06	-5.42	0.000	-.000022	-.0000103
@	0.0000066	9.11E-07	3.32	0.010	0.000004	0.0000089
perssq	1.37e-11	3.28e-12	4.17	0.000	7.24e-12	2.01e-11
@	1.85E-12	2.4E-12	2.1	0.032	7.1E-12	3.86E-12
hc	.0001662	.0000164	10.15	0.000	.0001341	.0001983
@	0.0000163	0.0000040	2.49	0.000	0.0000235	0.0000068
_cons	8.696176	.0413789	210.16	0.000	8.615073	8.777278
@	0.025095	0.0081233	35.25	0.000	0.011238	0.036079

Disclosure risk of absolute deviation

Risks can arise if direction of deviation is clear because of logical reasons:

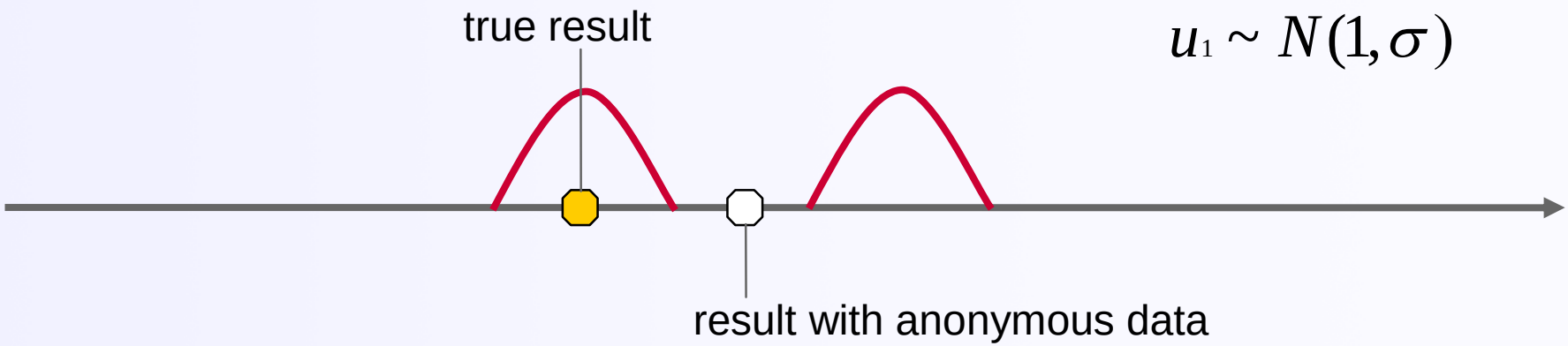
- Variable must be non-negative,
  but distance is bigger than absolute value
  - Example:
    average number of employees = 1000

@	1500

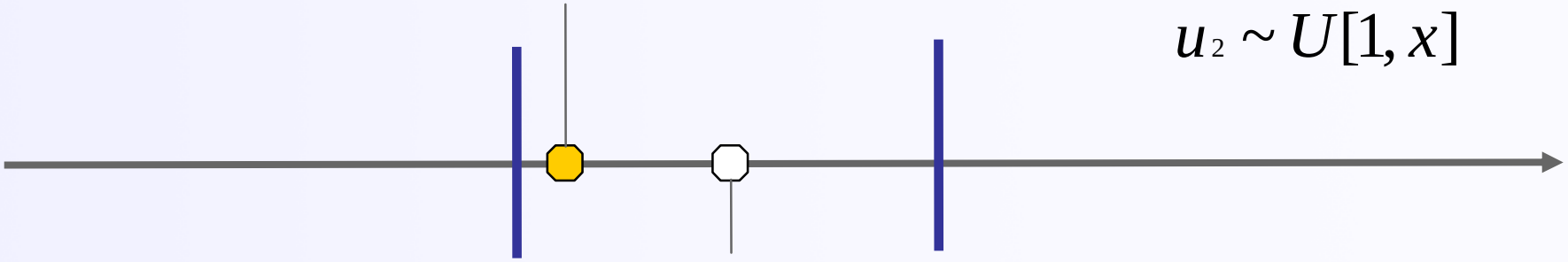
→ Adding stochastic noise to the quality measure

Variants of modifying the distance stochastically

a) Unbiased point estimate



b) Maximum distance



Remaining questions

- Remaining disclosure risks
- How to deal with distance 0?
- Testing acceptance among users
  - „use Morpheus or wait“
- Independent of anonymisation technique
- IT server environment

Thank you for your attention!

Contact

Dr. Jörg Höhne
Joerg.Hoehne@statistik-bbb.de

Julia Höninger
Julia.Hoeninger@statistik-bbb.de