Using *k*-anonymisation for analysis of registry data: pitfalls and alternatives

Sten Anspal

The Estonian Centre for Applied Research (CENTAR), Estonia,
sten.anspal@centar.ee

Keywords: k-anonymisation, privacy-preserving computation, linked registry data

We describe an applied study of the ICT students employment in Estonia based on linked official registry data. The study offered an opportunity to compare results from both k-anonymised data as well as those from privacy-preserving computations on the Sharemind platform ([1], [2]), which offers a way to use confidential data for research without loss of observations.

The research question was simple: What is the employment rate in general, and specifically in ICT companies, among ICT students during their studies? The question was motivated by the low rate of timely graduation (substantially lower than for non-ICT students), examining the hypothesis that the problems of dropping out and delays in graduation is a worse problem among ICT students than others due to high labour market demand for their skills and, therefore, their higher employment rates.

The study was carried out based on linked data from two registries: the Estonian Education Registry data for all students in higher education from 2006-2012 was used as the source of information on persons studies on various curricula and Tax Board data on social tax declarations in 2006-2013 was used for information on employment.

However, as a condition of using these datasets, the problem of preserving subjects privacy had to be followed. Two approaches were used and compared. The first was k-anonymisation: cases for which there were less than 3 persons with the same unique combination of characteristics were removed from query results. The second was the Sharemind platform for privacy-preserving secure computing, which made it possible to analyse the same dataset without the loss of observations due to k-anonymization, but without access to individual observations. The results of the k-anonymized and lossless analyses indicate substantial differences in employment rates of ICT and non-ICT students. Depending on the level of study (Bachelor's or Master's studies) and institution of higher education, differences in employment rates using the two approaches range from 1 to more than 10 percentage points.

The results illustrate, on the basis of a real-world study, how the effects of k-anonymization can be drastic and unpredictable in terms of inference. While the use of a share computing based privacy-preserving does entail time costs related to unobservability of individual observations and therefore additional efforts to verify the computations, these are offset by greater confidence in the results.

References

[1] Bogdanov, D., Laur, S., Willemson, J. (2008). Sharemind: A framework for fast privacy-preserving computations. *Computer Security-ESORICS 2008*, 192–206.

[2] Bogdanov, D. (2013). *Sharemind: A framework for fast privacy-preserving computations*. University of Tartu, Tartu.