# Data Anonymization and Security

M2 internship

The concept of data anonymization has recently emerged to address the privacy and security of data. Data anonymization aims at obscuring the available data in such a way that distinct records are not mapped easily to the specific individual they concern. It allows organizations to utilize the data and analyze it as they wish without legal consequences. The most direct example is the data used in medical research. Clinical results help in tracking disease symptoms and picking out the response that certain medications have on patients. The concept of $k$-anonymity was developed as a response for the problem of publishing relevant medical data without violating privacy. Many companies also analyze their customer data to figure out consumer patterns that help in targeted advertising. Such information has had a great impact on the profits of such companies. There are multiple methods of data anonymization which vary in their applications and effectiveness. K-anonymity is a commonly used method that defines a degree of anonymity "$k$." A data sample is considered anonymous if each record is indistinguishable from at least $k$ other records. In this internship we plan to work on degree-based anonymization.

A graph is said to be $k$-anonymous for an integer $k$, if for every vertex in the graph there are at least $k-1$ other vertices with the same degree. We refer to [3] for a survey of anonymization models and a discussion about the pros and cons of the $k$-anonymity concept. In [1, 2], the computational complexity of making a given undirected graph $k$-anonymous either through at most $s$ vertex deletions or through at most $s$ edge deletions was studied. The two variants are intractable from the parameterized as well as from the approximation point of view. One reason of this hardness is that the number $s$ of allowed removals and the anonymity level $k$ are independent of each other, and that a small change in one of these parameter values might lead to a large jump of the other parameter value.

The goal of the internship is to study new models for degree-based anonymization.

This subject could be a starting point for a PhD thesis.

**Supervisor**: Cristina Bazgan, LAMSADE, Université Paris-Dauphine, bazgan@lamsade.dauphine.fr

## References

[1] C. Bazgan, R. Bredereck, S. Hartung, A. Nichterlein, and G. Woeginger. Finding large degree-anonymous subgraphs is hard. *Theor. Comput. Sci.*, 622:90–110, 2016.

[2] C. Bazgan and A. Nichterlein. Parameterized inapproximability of degree anonymization. In Marek Cygan and Pinar Heggernes, editors, *Parameterized and Exact Computation - 9th International Symposium, IPEC 2014*, LNCS 8894, pages 75–84. Springer, 2014.

[3] X. Wu, X. Ying, K. Liu, and L. Chen. A survey of privacy-preservation of graphs and social networks. In *Managing and Mining Graph Data*, pages 421–453. Springer, 2010.