**Privacy in Multiple On-line Social Networks – Re-identification and Predictability**

This software demonstrates how overlapping Online Social Networks can affect user privacy. The full paper will be published soon and will be made available:

> *David F. Nettleton, Vladimir Estivill-Castro, Julian Salas, "Privacy in Multiple On-line Social Networks – Re-identification and Predictability", Transactions on Data Privacy (in press).*

The main github repository, "**OverlappingOSNsUserPrivacy**", contains three Java projects:

- **IdentifyCrossSAN**
- **SplitSAN**
- **InformationGainReal3**

which correspond to the empirical Sections 5.2 (Re-identification using a Set-theoretical Approach) and 5.3 (Predictive Modeling Approach using SAN Metrics) of the paper, respectively.

The project "InformationGainReal3" is copyright David F. Nettleton (GNU General Public License V3.0) and the projects "IdentifyCrossSAN", "SplitSAN" are copyright Vladimir Estivill-Castro (GNU General Public License V3.0).

InformationGainReal3
This program essentially preprocesses the raw graph data (structure in one file and data one record per user) for its use in Weka to predict based on different dependent/independent attribute combinations. First it inputs the graph into an appropriate data structure, then it calculates the SAN metrics, and then creates the flags which indicate if a user has or not a given attribute. This file is then output to disc and is used as input to Weka in which we used the J48 rule induction algorithm and SVM-SMO algorithm to build supervised models. Models were built for different combinations of metrics/predictors.

The main input files are "SEP4.csv" and "ATTRIBDATA.csv".