Anonymization of centralized and distributed social networks by sequential clustering

Tamir Tassa and Dror Cohen

Social networks are of interest to researchers from many disciplines, be it sociology, psychology, market research, or epidemiology. However, the data in such social networks cannot be released as is, since it might contain sensitive information. Therefore, it is needed to anonymize the data prior to its publication in order to address the need to respect the privacy of the individuals whose sensitive information is included in the data.

The methods of anonymizing networks fall into three main categories. The methods of the first category provide *k*-anonymity via a deterministic procedure of edge additions or deletions. The methods of the second category add noise to the data, in the form of random additions, deletions or switching of edges, in order to prevent adversaries from identifying their target in the network, or inferring the existence of links between nodes. The methods of the third category do not alter the graph data like the methods of the two previous categories; instead, they cluster together nodes into super-nodes of size at least *k*, where *k* is the required anonymity parameter, and then publish the graph data in that coarse resolution.

Our study is the first one that considers the problem of privacy-preservation in the distributed setting, in which the network data is split between several data holders. The goal is to arrive at an anonymized view of the unified network without revealing to any of the data holders information about links between nodes that are controlled by other data holders. To that end, we start with the centralized setting and offer two variants of an anonymization algorithm which is based on sequential clustering. We consider social network data in which the nodes are described by some quasi-identifiers (e.g. age, gender, location). The output of our algorithms provides the graph structure over a clustering of the nodes into super-nodes of size at least *k*, and a corresponding generalization of the quasi-identifiers that are present in each super-node. We then devise a secure multi-party implementation of our algorithms that enable several social network owners to securely compute an anonymized view of the unified network.