# ABSTRACT

Big data is predominantly associated with data retrieval, storage, and analytics. The world is creating a massive data size, which increases exponentially. Since the dawn of time until 2015, human had created 7.9 Zettabyte. This number will be exponentially raised up to 40.9 Zettabyte by 2020. Analytics in big data is maturing and moving towards mass adoption. The emergence of analytics increases the need for innovative tools and methodologies to protect data against privacy violation. Data analytics is prone to privacy violations and data disclosures, which can be partly attributed to the multi-user characteristics of big data environments. Adversaries may link data to external resources, try to access confidential data, or deduce private information from the large number of data pieces that they can obtain. Many data anonymisation methods were proposed to provide some degree of privacy protection by applying data suppression and other distortion techniques. However, currently available methods suffer from poor scalability and performance, low granularity, and lack of framework standardization etc. Current anonymisation methods are unable to cope with the processing of massive size of data. Some of these methods were especially proposed for MapReduce framework to operate in big data. However, they still operate in conventional data management approaches. Therefore, there were no remarkable gains in the performance.

To fill this gap, this thesis introduces a sensitivity-based anonymity framework that can operate in MapReduce environment to benefit from its advantages, as well as from those in Hadoop ecosystems. The framework provides a granular user’s access that can be tuned to different authorization levels. The proposed solution provides a fine-grained alteration based on the user’s authorization level to access a domain for analytics. The framework’s core concept was derived from k-anonymisation techniques, which was proposed by Sweeney on 1998 for data protection. Using well-developed role-based access control approaches, this framework is capable of assigning roles to users and mapping them to relevant data attributes. Moreover, the thesis introduces a simple classification technique that can properly measure the anonymisation extent in any anonymized data. Various expirments show promising results on applying the framework porposed in this thesis. The framework anonymisation expirements demonstrate fine graduality, good perfomance of parallel processing with a high scalability and a low distortion.

To examine the effectiveness of the proposed framework in protecting privacy and reducing data loss, a diverse range of experimental studies are carried out. The experimental studies aimed to demonstrate the capability of the framework’s fine granularity by applying gradual levels of anonymisation for data analysers. The experiments also meant to compare between the proposed anonymisation framework and the current available frameworks. In addition, the performance and data loss in big data are related to operational tools, such as MapReduce and Spark. The comparison has been done in both systems. The experiments’ results showed higher performance output, in general, when anonymisation was conducted in Spark. However, in some limited cases, MapReduce is preferable when the cluster resources are limited and the network is non-congested.

The experiments unveil several facts regarding big data behaviour. For instance, big data tends to be more equivalent as the data size increases. Moreover, the major concern on big data is security, hence, focusing on security side should be the main target. The few obfuscated records do not have a major impact on the overall statistical results. Therefore, the trade-off between security and information gain tends to give security a higher priority. It is expected that big data access is requested by a great number of users. This massive demand has recently increased with the social media blossom over the Internet. Personal and contextual information are available online publicly. Thus, personal re-identification has never been easier now. For this reason, we believe that security should be the major focus of anonymisation algorithms.

The experiments have also shown a high performance of processing and an average information loss for the proposed anonymisation framework. The anonymised data has gained a low classification error by Bayesian classifier. In comparison to the current anonymisation methods, the proposed framework has a little lower classification error by 0.12%. From the performance perspective, the proposed framework has reached up to 40% faster than the current anonymisation frameworks. For the security side, it was strengthened by increasing the *k-anonymity* value, and assigning granularity for user’s access.