ABSTRACT

Nowadays, the users depend on the credibility of the Online Social Networks (OSN), on which they share their personal data with and their guarantees of protecting them. However, the users do not know about the lack of protection that OSNs have against fake accounts. People create fake for various reasons. For instance, a company can create an account for advertising reasons or for retrieving any information, they can from the users, which in turn will use for their advantage. These accounts can belong to anyone and anything (company, animal, etc.). The problem lies not to the reason of the creation of fake accounts but to the creation itself.

Nonetheless, OSNs are trying to find and deactivate these accounts with the help of human personnel. The detection of these accounts is a great task because thousands of accounts are activated in a day but only a small part of them are deactivated, according to the designers of SybilRank (Cao, Sirivianos, Yang, & Pregueiro, 2010, p. 3). SybilRank is a new tool, which was developed by researchers from CUT, Telefonica and Duke University. With this tool, OSN providers can take advantage of the social graph and rank accounts according to their probability of being fake (Sybils) therefore making easier their detection.

Nevertheless, the processing of social graphs is not such an easy task. They are large amounts of data, which need large-scale parallel computing frameworks for their processing. The current frameworks that have been chosen for this purpose, such as MapReduce (Dean & Ghemawat, 2004), they do not efficiently utilize the shared memory of computers. This drawback makes them useless for computations that use intermediate results, such as PageRank (Wills, 2006) and SybilRank. The solution is a new abstraction called Resilient Distributed Datasets (RDDs), which was developed in UC Berkeley (Zaharia, et al., 2011). RDDs use mainly the RAM of the computers instead of the hard disk during the large-scale parallel computations.

In this thesis, we discuss we created an enhanced SybilRank in order to take advantage of this abstraction. Moreover, we demonstrate the implementation of SybilRank on Spark, a framework, which was developed in UC Berkeley by the designers of RDDs. Finally, we discuss about future work and ways to improve the work done in this Thesis, by taking advantage of Amazon EC2 clusters and taking full advantage of the capabilities of Spark and resilient distributed datasets.