Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Supervisors: Andreas Plesner, Prof. R. Wattenhofer
Student: Stuart Heeb

February 26, 2024

Master Thesis Proposal

Finding Sybils

1 Background

1.1 The Problem

A social graph is a graph where each node represents a person and each edge is a connection between two people. A "sybil" node refers to the representation of a fake[1] person (a fake account generated by a person or program). The problem of finding sybils describes the identification of such nodes in a social graph.

1.2 Existing Work

Research in this field has been going on for many years. Most of it has revolved around developing non-data-driven graph-based algorithms, that may or may not involve using additional data about the social graph, rather than just utilizing the connectivity (e.g., utilizing information about sent friend requests and the acceptance/denial thereof). Recently, there has been some research that uses machine learning techniques to find sybils. These usually involve a great set of features (e.g., information about likes, comments, profile description and many more).

1.3 Data

Previous works in this area have almost exclusively used synthetic data based on the social network platforms Facebook, Instagram and Twitter/X. The reason for this is that real data is very hard to come by. Further, synthetized datasets tend to be quite small, making data-driven approaches more difficult. Sometimes, the specific feature selection employed by research using machine learning described in Section 1.2 limits the applicability to a subset of these platforms (because some features might not exist in a platform or a platform might have a different underlying use-case regarding certain features).

[1] Fake may refer to things such as malicious, spam, misinformation.

2 Project Description

2.1 Finding Sybils

The objective is to identify sybils in social graphs, as described in Section 1.1. The knowledge obtained by studying existing graph-based and machine learning algorithms will hopefully lead to novel or optimized ways of solving this problem.

Focus on machine learning The main goal of this thesis is to evaluate whether machine learning methods, specifically using graph neural networks (GNNs), can be used to detect sybils reasonably well while still being computationally efficient.

2.2 Data Synthetization

As described in Section 1.3, real data is usually not obtainable. For this, researchers have fallen back on synthetizing data in order to develop algorithms to find sybils. These datasets are usually rather small. For this projects, there are three possible sources of data:

1. real data (if available)
2. pre-existing synthetic data
3. data synthetization (i.e., graph generation)

It's not clear whether a social graph by itself (solely the connectivity) is sufficient to achieve good results. It may be necessary to utilize additional features.

The study of data synthetization might play a notably large role in solving the problem of finding sybils, as this amounts to figuring out what structure these social graphs typically have when they contain both real nodes and sybil nodes.

3 Project Outline

The official start of this thesis is March 1st, 2024 and it will last for at most six months. Half-way into the thesis there will be a midterm presentation with Prof. Wattenhofer, Andreas Plesner and other PhDs from the group. The following describes a tentative and rough project outline.

In the first 2-3 months, the focus will be on studying existing papers and their algorithms, obtaining or creating datasets, and using this data to reproduce the results in said papers (depending on the availability of code, some algorithms might need to be implemented from scratch). During this process a code pipeline will be created which allows running different algorithms and comparing results.

In the remaining 3-4 months, the focus will be on finding machine learning apporaches to solve the problem, which is the main goal of this thesis. This objective will be aided by the previous study and code implementation of existing research. Emphasis will be on methods that involve only the graph structure, such as GNNs, but utilizing other features might be explored. These findings will then be compared to the existing algorithms.