Big Data Analytics - Twitter Social Graph

Dr. Giulio Pasqualetti

August 23, 2023

Introduction

Problem: Analysis of the **Twitter** social graph.

Goals:

- 3 **most followed IDs**;
- **Highest follower** (user f);
- **Shortest path** from user f to other users;
- **Longest cycle** from user f;
- **Density**/**Sparsity** of the graph;
- Analogies to **money-laundering**.

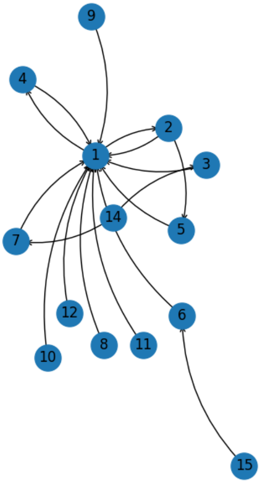


Figure: A social graph

Understanding the Data

Data Source: https://github.com/ANLAB-KAIST/traces/releases

- twitter_rv.net.00, twitter_rv.net.01, . . . , twitter_rv.net.03

Data Format: USERID \t FOLLOWERID \n

Volume: twitter_rv.net.00 ∼ **400 million** lines.

Restriction: First **10 million** lines.

12	13
12	14
12	15
⋮	
16	12

Figure: File format

Most Influential Users

Findings: Users **20**, **13** and **10350** are the most followed users, with respectively 1213787, 1031830 and 1003728 followers.

Method: Construction, during data acquisition, of a "followers" **dictionary** and subsequent analysis of it.

Dictionary format: {user: m}, where m is the number of followers.

$$\begin{array}{|cc|}\hline 12 & 13 \\ 12 & 14 \\ 16 & 12 \\ \vdots & \\ \hline \end{array} \Longrightarrow \begin{array}{|c|}\hline 12 \mapsto 2 \\ 16 \mapsto 1 \\ \vdots \\ \hline \end{array}$$

Figure: Dict. construction

Highest Following User: f

Let **f** be the user who follows the **highest number** of IDs.

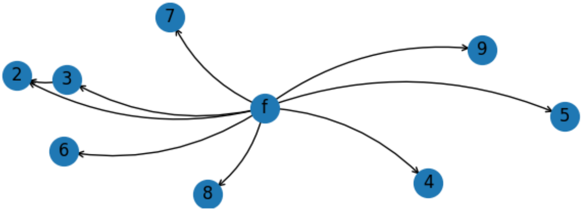


Figure: Example

Findings: **10316422**, with 2583 IDs followed, is the user **f**.

Remark: User f is **unique** (the second highest user follows 940 accounts).

Highest Following User: f

Method: Construction, during data acquisition, of a "following" **dictionary**, and subsequent analysis of it.

Dictionary format: {follower: { u_1, . . . , u_p}}, where u_1, . . . , u_p are the **followed users**.

12	13
16	13
12	14
13	14
15	14
⋮	

$$\Longrightarrow$$

$$\begin{array}{c} 13 \mapsto \{12, 16\} \\ 14 \mapsto \{12, 13, 15\} \\ \vdots \end{array}$$

Figure: Dictionary construction

Shortest Paths

Given a user u, the **shortest path** from user f to u is the **lowest number of edges** directed from f to u, when it is reachable.

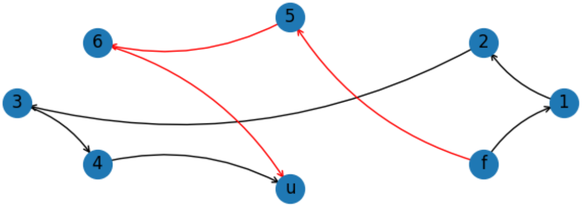


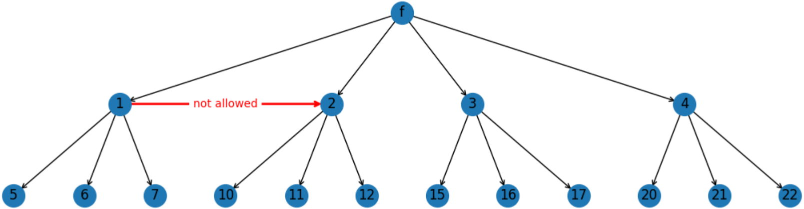
Figure: Example

Findings: 3975 users are **reachable** from user f.

The **four farthest** nodes are 342, 205, 435 and 928, at distance 6, 5, 5 and 5 respectively.

Shortest Paths

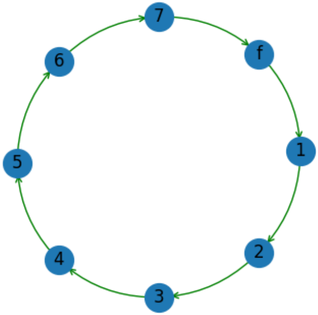
Method: Construction of a "following" **tree**, in which each floor "follows" the next one, with **no node repetitions**.



Remark: Depending on the structure of the graph, more **efficient algorithms** could be implemented, e.g., the graph already is a tree.

Optional: Cycles

Goal: The longest **cycle**, e.g. path **starting and ending at user f**, but visiting other users at most once.



Findings: User f has **no followers** $\Rightarrow$ No path to him.

General approach: Strongly dependent on the graph **structure**

Sparse vs. Dense: Impact on Analysis

In mathematics, a **dense** graph is a graph in which the number of edges is close to its **maximal possible value**. If it is significantly lower than that, the graph is **sparse**.

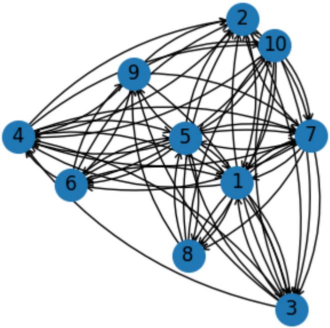


Figure: A dense graph

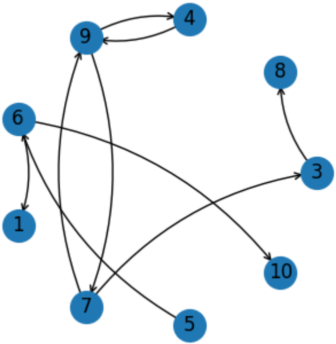


Figure: a sparse graph

Sparse vs. Dense: Impact on Analysis

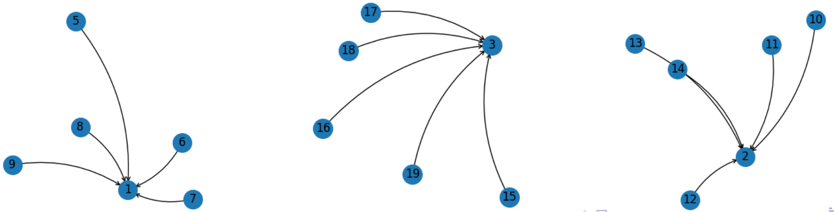
Recall: If a graph has N nodes, the **maximal number** of edges is

$$\binom{N}{2} = \frac{N(N-1)}{2}.$$

Remark: Our edges are **oriented** $\Rightarrow N(N-1)$.

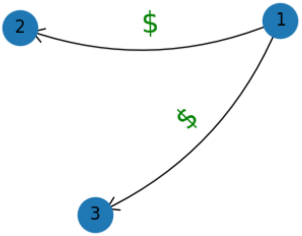
Findings:

- $N \approx 5.15 * 10^6$, $\#\text{edges} = 10^7 \approx 2N \Rightarrow$ **sparse**.
- The first **8 most followed IDs** account for 73.8% of the edges.



Money Laundering Detection

Setting: Edges represent **money transfers**.

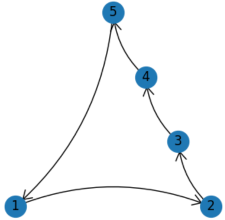


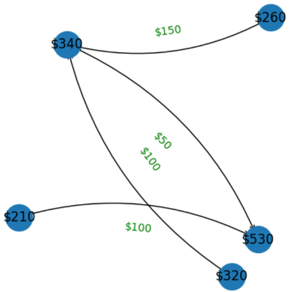
Questions:

- What patterns suggest **money laundering**?
- **Additional information** to enhance identification?

Money Laundering Detection: Possible Solutions

Money laundering is typically performed through **circular payments**, i.e. cycles.

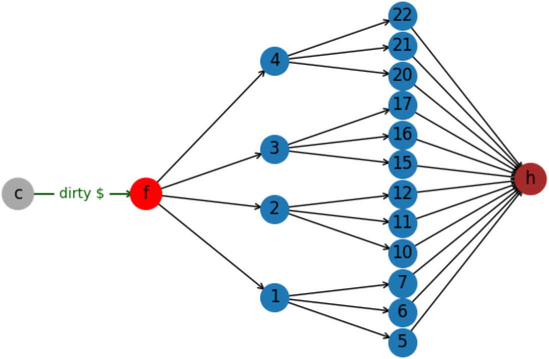




The graph should include transfer **amounts**, node **balances**, account holder **names** and **countries**.

Money Laundering Detection: Possible Solutions

Another common technique is **layering**.



The aim is to **avoid detection** through multiple **small money transfers**, converging to the same **final account**.

Remark: It is hard for banks to detect such complex patterns, as they only have access to **their own transactions**.

Thank you for your attention!