**TITLE:**

**An analysis of the *friendship paradox* on different type of social networks**

**Motivations:**

In recent decades, social media has become increasingly popular, transforming the way people connect and interact, and concerns us all closely. Therefore, examining these platforms closely from a graph network perspective can provide valuable insights into the social dynamics they enclose.  
Among the vast landscape of social networks, the platforms Facebook, LinkedIn, and YouTube were selected for this analysis, due to their accessible datasets and their widespread popularity.  
One particularly interesting phenomenon that was decided to explore is the *friendship paradox*, a concept formulated by sociologist Scott L. Feld, which suggests that, "on average, an individual's friends have more friends than the individual does.".

This analysis will investigate how each platform reflects this paradox, aiming to understand the extent to which the phenomenon appears across these networks.

**Datasets**:

We’ve chosen three graphs, each one representing the friendship in a different social network:

For Youtube we chose the dataset present at <https://snap.stanford.edu/data/com-Youtube.html> which contains information about the networks of followers on Youtube communities, for Linkedin we chose the dataset present at <https://networkrepository.com/soc-linkedin.php> which contains the user-to-user connections while for facebook we chose the dataset at <https://networkrepository.com/socfb-wosn-friends.php> which contains a network of users friendships

We think that Youtube, Linkedin and Facebook are different enough in nature to have different features between each other; for example we expect that a “popular user” on youtube will have many more “friends” than the average one, while “popular users” on facebook will have less friends in proportion due to the fact that often unpopular users are still friends with a fair amount of people (like real life friends and family). We define a **popular node** (and consequently a **popular user**) a node that has more neighbors than the average number of neighbors of its neighbors.

**Methods**

The program responsible for the analysis will be written in *Python* using the *NetworkX* library (and other ones, if necessary), which contains different utilities to deal with graphs. The files containing edge information, downloaded from the datasets, will serve as inputs to our programs. We aim to parallelize the process to improve performances but if that poses challenges and the time spent trying to enhance speed becomes excessive, we will revert to a serialized approach.

Initially, we will try to execute the algorithm exactly; however, if we discover that the graph is too large and processing requires excessive time and memory consumption, we will try to use approximated algorithms. Since there seem to be a huge number of edges (greater than two million) the exact computation could in fact be very well challenging. However, since each edge is represented by two integers, the system should be able to read the data quickly, so we’ll test before deciding definitely

**Intended experiments:**

We will conduct several experiments focused on analyzing graph structures representing friendships/follows: we aim at identifying key patterns among different types of social networks and also their differences. To do this we will need to compute some features of the nodes of each graph as well as features of the whole graph.

We will use the **degree** of the nodes and the degree of their neighbors to verify the friendship paradox and also check if the paradox is more accentuated in a social network than in another (i.e. the difference in “friends” between popular and unpopular users is greater or lower in proportion). We expect that the paradox will be “more accentuated” in Youtube and Linkedin than in Facebook for example (as explained before). We will also use the degree to understand which nodes correspond to popular users: we define a **popular node** (and consequently a **popular user**) a node that has more neighbors than the average number of neighbors of its neighbors.

We will compute other node/graph features such as **clustering coefficient** and different types of centralities such as **closeness centrality**, **PageRank centrality** and **betweenness centrality**. This will allow us to understand if some of those measures correlate with the “popularity” of a node or, in the case of graph measures, correlate with having “a more pronounced” friendship paradox. If the computation of those measures will prove too computationally intensive we will result in methods such as sampling to reduce the time needed providing an approximated result

We will compare our results with **random graphs** to check if the paradox applies even in networks that don’t represent friendships

**Machine specifications:**

- 8 cores CPU, 4.8 Ghz sustained clock while boosting

- 32 GB of 3600 Mhz

- SSD ≈ 3000 MB/s in read and write

**Contributions:**

In this first project proposal part we more or less worked together, most of the time was spent doing organization, sharing ideas and compiling this report so the contribution of each member was approximately ⅓.

**Lavorati Ippolito**: Helped writing the report, helped finding the datasets and assess their quality, research and presented ideas, some of which were discarded

**Orsolon Ludovico**: Helped writing the report, gave the main idea about the *friendship paradox*, formulated of the hypothesis and researched about experiments to test it

**Stefani Patrizia**: Helped writing the report, helped finding the datasets, research on the methods to implement the experiments and relative python functions