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

Ippolito Lavorati, Ludovico Orsolon, Patrizia Stefani

November 2024 - January 2025

1 Motivations

In recent decades, social media have become increasingly popular, transforming the way people connect and interact; examining these platforms from a graph network perspective can thus provide valuable insights into the social dynamics they enclose.
One particularly interesting phenomenon that we 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 considered platform reflects this paradox, aiming to understand the extent to which the phenomenon appears across these networks.

1.1 Datasets

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

- Youtube: the dataset at `https://snap.stanford.edu/data/com-Youtube.html` contains information about the networks of followers on Youtube communities
- Linkedin: the dataset at `https://networkrepository.com/soc-linkedin.php` contains the user-to-user connections
- Facebook: the dataset at `https://networkrepository.com/socfb-wosn-friends.php` contains a network of users' friendships

We think that these three social networks are different enough in nature to have different characteristics 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 fewer friends in proportion due to the fact that often unpopular users are still friends with a fair number of people (such as friends and family in real life).
Furthermore, since the datasets utilized are huge, we designed a smaller graph to debug our program and verify our measures (Figure **??**). We called this graph *TestGraph*.

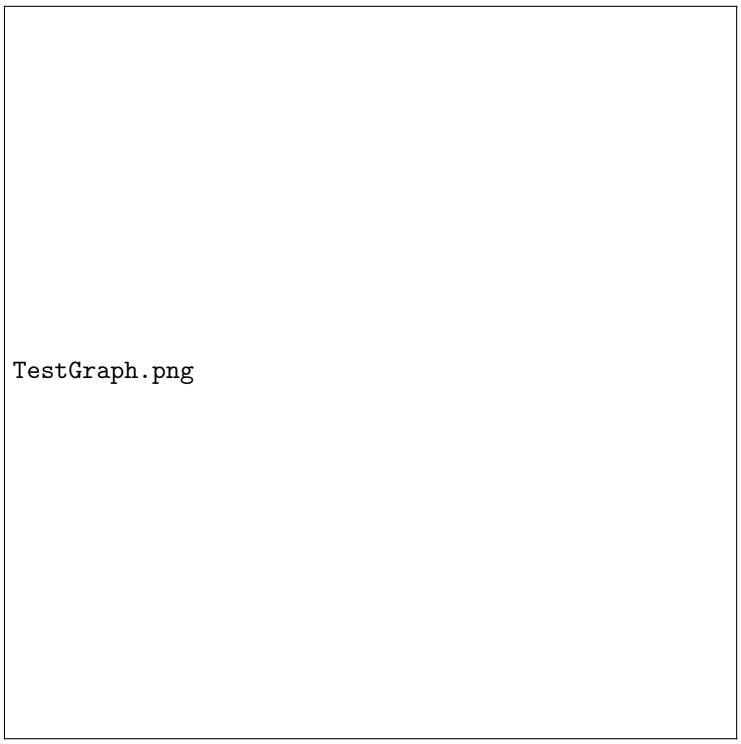

Figure 1: *TestGraph*

We created a random graph for each dataset in order to better verify the statistical analysis. The random graphs are constructed with the same number of nodes and edges of the real graph utilized using the *gnm_random_graph* function from the *NetworkX* library.
Generating a single random graph for each social network is certainly not optimal; we would need many random graphs to better estimate random features. We chose not to do so for the sake of computational time: computing even a single measure on our machines was a long process, so we decided to employ our time trying to understand results rather than computing tons of measures. The graphs employed, as well as the respective random graphs are characterized by the following properties:

	TestGraph	Facebook	Linkedin	Youtube
Nodes	22	63 731	6 726 011	1 134 890
Edges	34	817 090	19 360 690	2 987 624
Avg Degree	3.09	25.64	5.76	5.26

Table 1: Some graphs' properties

2 Methods

The program responsible for the analysis is written in *Python* using the *NetworkX* library for graphs, the *tqdm* library to keep track of long processes through progress bars, *pyplot* from *matplotlib* to visualize information, *pandas* to deal with csv files. The files containing edge information, downloaded from the datasets, will serve as input to our programs.Some statistical analysis were also performed in R The program was developed collaboratively on GitHub[**?**] and all computations provided in this paper can be found in one of the functions provided on it.

2.1 The fship_score

We decided to encapsulate the concept of the friendship score in a measure that we called *fship_score*, computed for each node. The *fship_score* is defined as:

$$\textbf{\textit{fship\_score(u)}} = \deg(u) \cdot \left(\frac{\sum_{v \in N(u)} \deg(v)}{\deg(u)} \right)^{-1} = \frac{\deg(u)^2}{\sum_{v \in N(u)} \deg(v)}$$

That is, the degree of the considered node divided by the average degree of its neighbors. Intuitively if the score is less or equal to 1 the friendship paradox is true for the node, otherwise we consider it an outlier.

3 Experiments

3.1 Verifying the *friendship paradox*

First of all we need to check if the *friendship paradox* is present in our datasets: below a table with the proportion of outliers is reported (*number of nodes with* $fship\_score > 1$ / *total_number_of_nodes*), this table contains results from one of our R scripts.

	TestGraph	Facebook	Linkedin	Youtube
Real	0.2777	0.0865	0.2927	0.0325
Random	0.3636	0.4086	0.3556	0.3544

Table 2: Proportion of outliers

We can immediately see that the *frienship paradox* is true for both real and random graphs but we can also see that it is more accentuated on the real graphs than on the random ones: this suggests that the *friendship paradox* is a characteristic of real "social network" graphs even though it can be also present in random graphs to some degree. Furthermore, we can see that the paradox is more accentuated on Youtube and less on Linkedin as expected. From the data, the random graphs seem to have a different proportion of outliers than the real

graphs (H1), to verify this hypothesis we performed the z-test on the values in the above table using [real graph - random corresponding graph] pairs.

$$z = \frac{p_1 - p_2}{\sqrt{p_{comb}(1 - p_{comb})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

$$n_i = number\ of\ nodes\ of\ i$$

$$k_i = number\ of\ nodes\ with\ fship\_score > 1$$

$$p_i = \frac{k_i}{n_i}$$

$$p_{comb} = \frac{k_1 + k_2}{n_1 + n_2}$$

These are the resulting p-values:

	TestGraph	Facebook	Linkedin	Youtube
Real vs Random	0.5	0	0	0

Table 3: p-values Real vs Random z-test (computed by our R function)

The p-values are all zero (reject H_0) except for the artificial graph suggesting that the proportion of outliers are different between the random and the real graphs, this means that the *fship_score* is probably not due to randomness. We also compared datasets against each other to see if there are common proportions between them, but the p-values are zero for all combinations (facebook-Linkedin, Linkedin-Youtube and Youtube-Facebook) so we conclude that such correlation doesn't exists.

3.2 *Friendship paradox* correlations

The *friendship paradox* reveals intriguing insights into the structural dynamics of social networks. To explore this phenomenon, we analyzed three key metrics: *degree*, *PageRank*, and *clustering coefficient*, each of which offers a unique perspective on how the paradox manifests. The results varied across different social networks, as shown in *Table 4*, reflecting their distinct structures.

In Facebook, PageRank showed the strongest correlation with the friendship paradox, underscoring the influence of highly connected hubs. Degree also showed a significant positive correlation, reinforcing its central role in the paradox, while clustering coefficient had minimal negative correlation, suggesting that local cohesion has little impact.
For Linkedin, PageRank again emerged as the most significant factor, with stronger correlations than those observed in Facebook. Degree correlations were

	Facebook	Linkedin	Youtube
Degree Correlation			
Pearson	0.784	0.728	0.575
Spearman	0.747	0.865	0.185
PageRank Correlation			
Pearson	0.898	0.967	0.836
Spearman	0.906	0.980	0.743
Clustering Correlation			
Pearson	-0.106	-0.355	-0.003
Spearman	0.111	-0.230	0.064

Table 4: Metrics correlation

also substantial, reflecting Linkedin's emphasis on direct professional connections. The clustering coefficient showed a slightly more negative correlation, but remained a secondary factor.
In youtube, PageRank remained a strong driver of the friendship paradox, although its correlation was weaker than Facebook and Linkedin. Degree correlations were less pronounced, indicating different dynamics shaped by content creation and consumption. The clustering coefficient showed an almost negligible correlation, underscoring its limited relevance in this network.

To quantify these relationships, we used several statistical methods. Linear regression provided a clear model of how the friendship paradox score varies with each metric, offering insights into the strength and direction of these associations. Pearson correlation measured the degree of linear relationships, while Spearman correlation assessed monotonic associations, uncovering potential nonlinear trends. In particular, we excluded *closeness centrality* and *betweenness centrality* from our analysis due to computational constraints, as computing these metrics for large-scale networks such as Facebook, Linkedin, and Youtube was not feasible within this project.
Our results highlight the multifaceted nature of the friendship paradox, driven primarily by global metrics such as PageRank and degree, while local metrics such as clustering have a more subtle impact.

4 Possible improvements

With more powerful computational machines and more time there are a few ideas that could be developed to expand on this paper results:

- Generating many random graphs for each social network would provide a better approximation of the randomly-generated features and also, generating other types of random graphs (Chung-Lu model for example), could highlight new correlations we may have missed.

- An obvious improvement could be simply using more data, for example finding new graphs from other social networks, we could also check if in different graphs obtained from the same social network we find significative differences on *fship_score*.

5 Machine specifications

- 8 cores CPU, 4.8 Ghz sustained clock while boosting
- 32 GB of 3600 Mhz
- SSD $\approx$ 3000 MB/s in read and write

Contributions

We worked together, we belive that the the difference in contribution amount between all of us was negligable, hence we report a contibution of 1/3 for each member.

Lavorati Ippolito: Helped writing the report, helped finding the datasets, researched and presented ideas, Statistical tests, pyhton programming

Orsolon Ludovico: Helped writing the report, gave the main idea about the friendship paradox, formulated of the hypothesis, researched about experiments, python programming, statistical test idea and interpretation

Stefani Patrizia: Helped writing the report, helped finding the datasets, researched on the methods to implement the experiments and relative python functions, implemented some methods for the computation of graph statistics

References

[1] Scott L. Feld (1991) *Why Your Friends Have More Friends than You Do*, The University of Chicago.

[2] The program can be found on GitHub at `https://github.com/OrsolonLudovico/LFN_project.git`