Assignment 1 – Reddit's Hyperlinks network

Miggiano Davide 4840761
Morando Andrea 4604844

Introduction

This report contains a social network analysis related to the Reddit's Hyperlinks network, which is a directed graph where nodes are subreddits and edges are hyperlinks between them; the main library used to compute the measurements on the graph is NetworkX.

The dataset is available at this link: https://snap.stanford.edu/data/soc-RedditHyperlinks.html and it's composed of two files: '*soc-redditHyperlinks-title.tsv*' and '*soc-redditHyperlinks-body.tsv*'.

The first file contains the hyperlinks between the subreddits, while the second file contains the hyperlinks between the posts.

The two files have the same structure, with the following columns:

- 'SOURCE_SUBREDDIT': the source subreddit of the hyperlink
- 'TARGET_SUBREDDIT': the target subreddit of the hyperlink
- 'POST_ID': the ID of the post
- 'TIMESTAMP': the timestamp of the post
- 'LINK_SENTIMENT': the sentiment of the hyperlink, which can be '1' (positive), '-1' (negative), or '0' (neutral)

We have chosen to consider the Hyperlinks-body file because it was smaller than the other one and, to create the actual graph, we have taken in consideration only the SOURCE and TARGET subreddit columns.

The general goal of this analysis is to understand the structure of the network and to identify the most important subreddits in terms of different centrality measurements and also the analysis of the most important communities.

Algorithm and measures

In the next sections, the structural properties and dynamics of our network is analyzed using various algorithms and measurements. These analytical tools offer insights into the connectivity patterns, node influence, and overall cohesion of the network.

A brief explanation of each measure:

- The **degree** and **degree centrality** metrics reveal the immediate influence of nodes based on their direct connections;
- Network **density** provides a measure of the overall interconnectedness within the network;
- The **degree distribution** helps identify the presence of hubs and understand the topology;
- The **betweenness** centrality highlights nodes acting as critical bridges within the network;
- The **closeness** centrality assesses the efficiency of nodes in disseminating information;
- **Connected components** are identified to understand the network's fragmentation and isolate the giant component to examine the largest cohesive subgraph;
- **PageRank** is utilized to rank nodes based on their importance, reflecting influence within the network;
- **Hub** identification and **authority** scores further highlight key influential nodes and trusted sources of information.
- **Community** detection shows how the network is made up of different groups of nodes with dense interconnections.

These measurements enhance our understanding of the network's architecture and dynamics.

Graph preparation

The first step was to visualize the graph, to check if there was some interesting structure to be analyzed. As we can see we a huge central component surrounded by tightly knit groups characterized by a relatively high density of ties; these components seems to be connected with the central component.

Figure 1 shows the graph visualization:

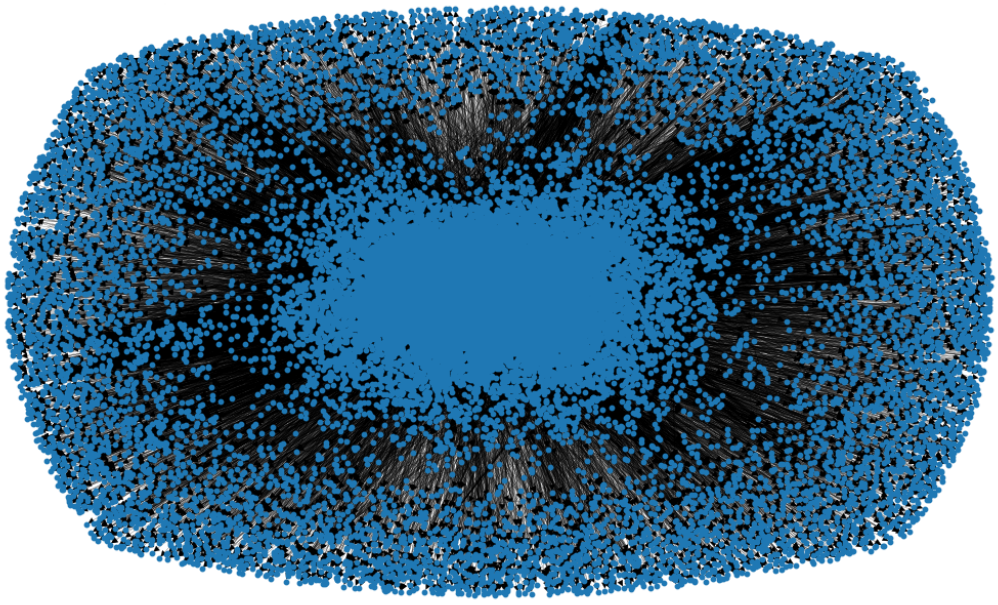

Figure 1: Entire graph structure

The main information calculated (Table 1) is:

Metric	**Value**
Number of Nodes	35776
Number of Edges	137821
Number of Strongly Connected Components	24071
Number of Weakly Connected Components	497
Average Node Degree	7.70466
Average In-Degree	3.8523311
Average Out-Degree	3.852331
Density of the Graph	0.000107

Table 1: Graph Metrics

Even if the number of nodes and links is high, it can be deduced from the density that the graph is sparse, as this value tends towards 0. This make sense because to have an higher density each connection between nodes should be bi-directional.

Looking at the number of SCCs and WCCs, we can say that the graph tends to have many islands with a high internal density. Our assumptions have a plausible match given the similarity between subreddit network structure with the well known Web graph.

The similarity is also confirmed by visualizing the degree distribution of the graph, that it's a power law distribution (Figure 2), which confirms the presence of a few central nodes (hubs) with a large number of connections and several isolated nodes with few connections.

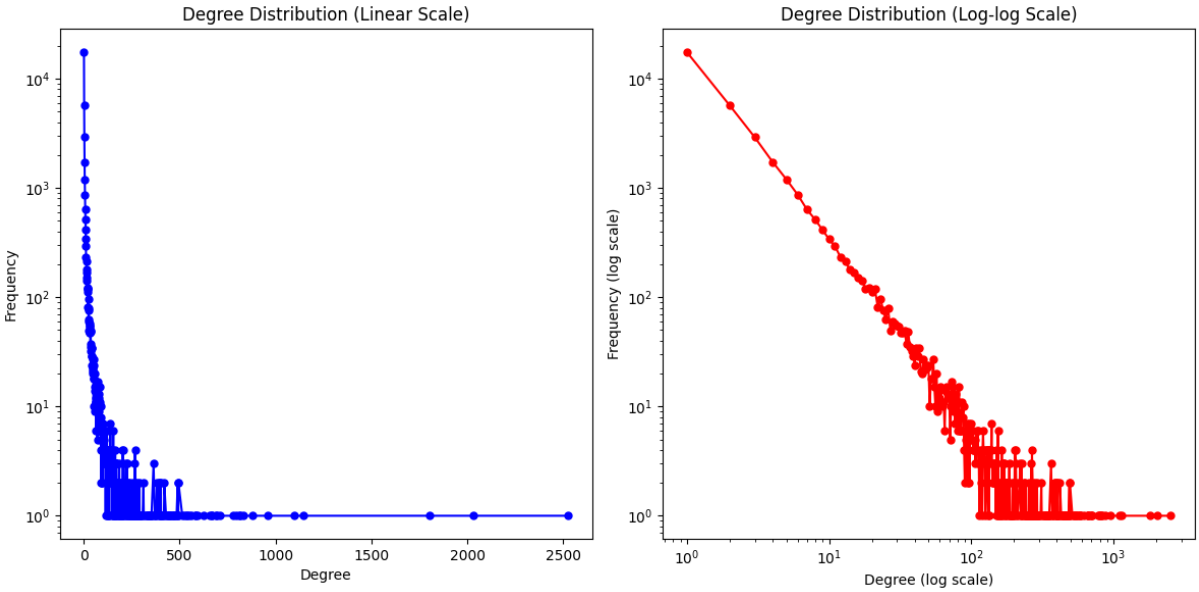

Figure 2: Entire graph degree distribution

Given the structure of the graph, it is very complex, if not impossible to calculate various measurements such as diameter, betweenness and closeness in a reasonable amount of time. This is why it is useful to analyses the structure of the SCCs and the connections within them to better understand the nature and dynamics of the network.

Since it is a power law distribution and we know that it is very robust to random 'attacks', we can reduce its size while maintaining the main characteristics of the network by performing a targeted removal of non-fundamental elements.

For this purpose, it was initially decided to reduce the graph to its SCC with the largest number of nodes in order to remove islands and SCCs of lesser significance.

Reduce to SCC

After performing the reduction, the structure of the graph and the information associated with it are as in Figure 3:

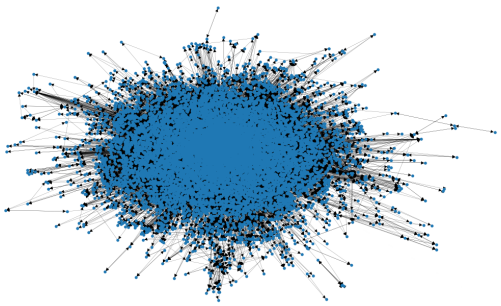

Figure 3: SCC structure

Metric	Value
Number of Nodes	11564
Number of Edges	98166
Number of Strongly Connected Components	1
Number of Weakly Connected Components	1
Average Node Degree	16.977
Average In-Degree	8.488
Average Out-Degree	8.488
Density of the Graph	0.0007341

Table 2: SCC Graph Metrics

The number of nodes and arcs is greatly reduced compared to its original version, but is still high; moreover, even though the density of the graph has increased, it still tends towards zero.

From the distribution, we can see that we have lost some big hubs but in general the situation has remained almost unchanged.

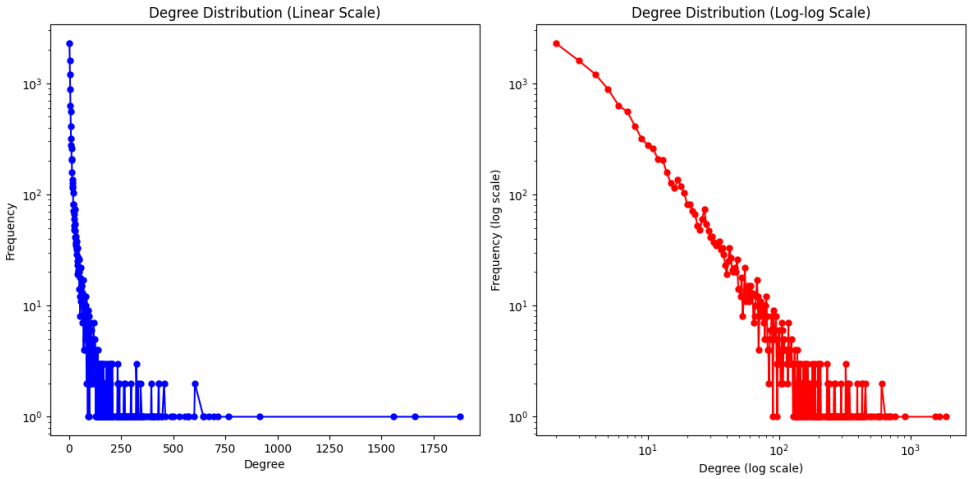

Figure 4: SCC degree distribution

Reduce based on Pageranking

Against this, it was decided to further reduce the size of the graph by keeping only 40% of it through a reduction based on the page ranking of the graph nodes. We have chosen reduction based on Pageranking because it provides clearer insights by concentrating on the most significant parts of the network.

After performing this operation, the Figure 5 shows the graph obtained:

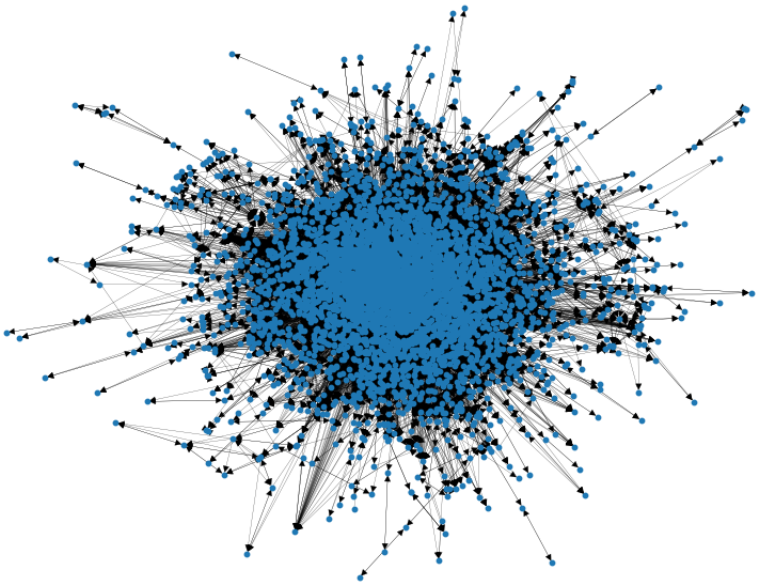

Figure 5: Pageranking reduced graph structure

Metric	**Value**
Number of Nodes	4625
Number of Edges	66136
Number of Strongly Connected Components	20
Number of Weakly Connected Components	3
Average Node Degree	28.599
Average In-Degree	14.299
Average Out-Degree	14.299
Density of the Graph	0.003092

Table 3: Reduced Graph Metrics

As we have extracted only the most important nodes, the average degree is increased by 75% and the density is slightly higher but it's still near zero.

The degree distribution (Figure 6) is:

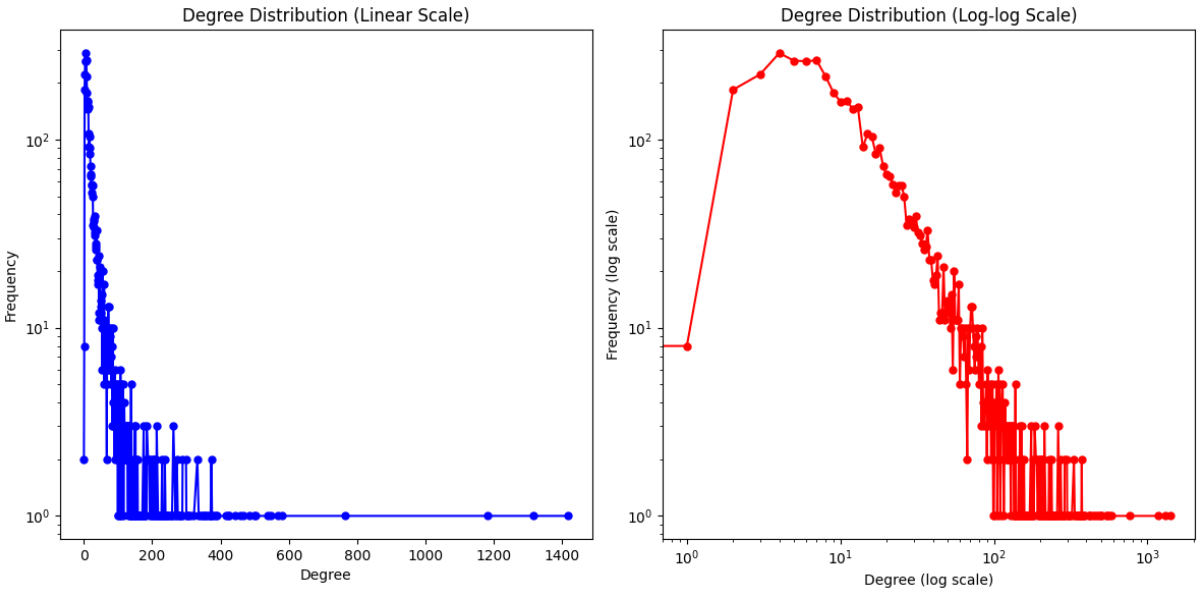

Figure 6: Pageranking reduced graph degree distribution

Although there has been a sharp reduction in nodes and arcs (for which the average degree has risen considerably), it can be seen that the distribution has still retained the typical shape of a power law distribution. Obviously, unlike before, most of the nodes will not have a very low degree and therefore the peak is slightly shifted compared to the previous graphs.

In all these reduction operations, the graph has remained directed in such a way as to preserve the links between the various SubReddits so that the importance of the nodes can be better determined, only for some metrics a version of the undirected graph will be used.

Centrality metrics and node importance

After reducing the graph, several centrality measurements were performed to understand how the nodes interacted with each other, which were the top 15 most important for each metric, and so on.

Here is the result of the calculated measurements:

Node	Betweenness Value	Node	Closeness Value
iama	0.13724	askreddit	0.51059
askreddit	0.13645	iama	0.49433
subreddidrama	0.10317	videos	0.45797
outoftheloop	0.06999	pics	0.45104
gaming	0.03613	funny	0.43858
leagueoflegends	0.03288	gaming	0.43750
writingprompts	0.03284	outoftheloop	0.43397
legaladvice	0.03031	worldnews	0.43426
explainlikeimfive	0.02220	technology	0.43205
techsupport	0.02115	news	0.42589
conspiracy	0.01981	explainlikeimfive	0.42459
games	0.01613	pcmasterrace	0.42424
dogecoin	0.01588	dataisbeautiful	0.41863
clashofclans	0.01481	science	0.41765
buildapc	0.01470	leagueoflegends	0.41652

Table 4: Top 15 Nodes by Betweenness and Closeness Centrality

Node	Hub Value	Node	Authority Value
subreddidrama	0.00740	askreddit	0.00960
copypasta	0.00497	iama	0.00845
outoftheloop	0.00483	pics	0.00655
circlebroke	0.00476	videos	0.00652
circlejerkcopypasta	0.00461	worldnews	0.00577
drama	0.00449	funny	0.00560
shitliberalssay	0.00437	news	0.00534
conspiracy	0.00418	explainlikeimfive	0.00464
justunsubbed	0.00411	gaming	0.00459
hailcorporate	0.00408	technology	0.00442
self	0.00402	outoftheloop	0.00440
askreddit	0.00393	science	0.00433
nostupidquestions	0.00387	showerthoughts	0.00403
bestofoutrageculture	0.00363	gifs	0.00392
karmacourt	0.00345	tifu	0.00387

Table 5: Top 15 Nodes by Hub and Authority Centrality using HITS

Subreddit	PageRank Score
iama	0.0148
askreddit	0.0143
pics	0.0086
videos	0.0076
videos_discussion	0.0065
outoftheloop	0.0058
gaming	0.0052
funny	0.0044
worldnews	0.0043
explainlikeimfive	0.0039
news	0.0038
pcmasterrace	0.0038
technology	0.0037
leagueoflegends	0.0036
movies	0.0034

Table 6: Top 15 PageRank Scores for Subreddits

AskReddit and **IAmA** consistently rank high across all metrics, highlighting their pivotal roles in the Reddit ecosystem. **AskReddit** facilitates wide-ranging discussions, as evidenced by its top positions in both betweenness and closeness centrality. This indicates that **AskReddit** acts as a central node, efficiently spreading information across the network due to its inclusive nature, where users ask diverse questions that generate extensive community interaction. Similarly, **IAmA** ranks highly in betweenness centrality and PageRank, underscoring its role in connecting users with notable individuals. This allows **IAmA** to act as a critical intermediary, bridging various user groups and ensuring high visibility and influence across Reddit.

SubredditDrama and **OutOfTheLoop** serve as connectors and aggregators, linking various parts of the Reddit community and providing context for ongoing discussions. **SubredditDrama's** high betweenness centrality shows its importance in connecting disparate discussions, often highlighting conflicts and significant events across Reddit. **OutOfTheLoop** also ranks high in betweenness and closeness centrality, indicating its role in helping users catch up on trending topics and events, effectively bridging knowledge gaps between various subreddits.

Copypasta plays a significant role in content aggregation, especially for viral text content. Its high hub score in the HITS algorithm reflects its influence in linking to a wide variety of other subreddits, thus distributing content that often becomes viral across Reddit.

Subreddits like **Videos**, **Pics**, **Funny**, and **Gaming** are central for content consumption. **Videos** and **Pics** rank highly in both closeness centrality and PageRank, indicating they are key nodes for multimedia content, making them easily accessible and highly influential within the network. **Funny** and **Gaming** also have high closeness centrality, showing their role as focal points for humor and gaming-related discussions, respectively. These subreddits serve as hubs where users frequently engage, share, and consume a wide array of content, making them central to user engagement and content dissemination.

Analysis of the undirected version of the graph

After running various metrics on the directed graph and realizing which nodes play a crucial role in Reddit's network, it was decided to analyse the graph from a different point of view, namely by making it undirected.

The first interesting thing that can be observed is the diameter, which is 9, which can lead us to think of the small-world phenomenon in that out of a total of 4623 nodes there are only 9 maximum hops between the most distant nodes (thus indicating a good connection). Calculating the assortativity (on the direct graph) instead yielded a value of -0.11, even if it's not a big value compared to 0 it may show the tendency of a dissortative mixing between nodes. It could makes sense because influence Subreddit (hub) is often linked by topic with less importance (node with low degree), but it's not always true.

After that we were interested in communities, in total 44 were found but only about 20 or so report relevant characteristics, the others are too small as they have less than 10 nodes with a number of arcs less than 5. The most interesting are the three largest ones as, although they consist of around 700-1600 nodes, it can be seen that most of the network is made up of these large communities:

Community	Node	Edges	Avg Clustering Coefficient	Average Degree	Avg Shortest Path
0	1667	11899	0.360325	20.9124	3.00067
1	1403	21698	0.391643	40.7904	2.44389
2	781	4019	0.44755	17.8617	3.23576

Table 7: Top 3 communities of Subreddits

As we can see in Figure 7, the nodes with the highest degree of these 3 communities are:

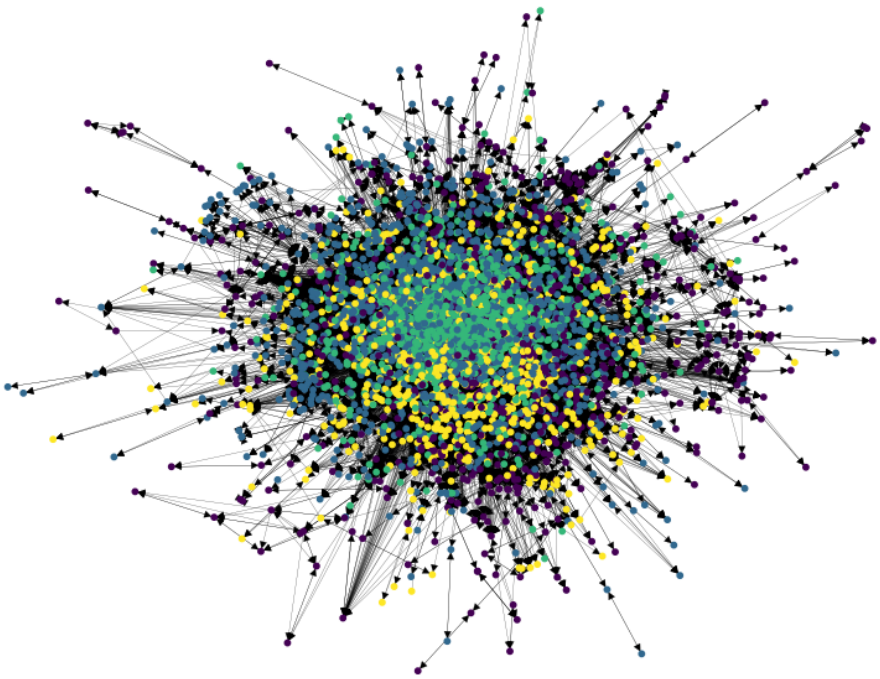

Figure 7: Pageranking reduced graph first, second and third bigger communities

More in details the three best nodes for degree for each of the chosen community are shown in Table from 8 to 10:

Node	Degree
gaming	423
games	318
techsupport	313

Table 8: Community 0

Node	Degree
askreddit	872
subredditdrama	669
conspiracy	374

Table 9: Community 1

Node	Degree
iama	255
nfl	142
soccer	115

Table 10: Community 2

These nodes are precisely the same as those analyzed above, emphasizing their centrality within Reddit's network. It can also be seen from here how these three communities are subdivided by topic, the first gaming and the other two based on Q/A with various links to sports rather than topic-specific topics.

Conclusion

As one would expect from a social network-related graph, one can find certain characteristics such as low density, the presence of large hubs, power law distribution, small-world and, given the nature of Reddit, the presence of many communities (44 only in this small portion of the original graph).

The initial graph, despite its high number of nodes and edges, is sparse, with many strongly and weakly connected components. The structure of the network, in particular its degree distribution, is consistent with the well-known properties of scale-free networks, dominated by a few highly connected hubs and many nodes with fewer connections.

Reducing the graph to its largest strongly connected component (SCC) allowed a more focused analysis while retaining the essential features of the network. Further reduction using PageRank provided insight into the core nodes, highlighting the importance of influential subreddits. Subreddits such as **AskReddit** and **IAmA** consistently ranked high in various centrality metrics, demonstrating their central role in content dissemination and user interaction within Reddit.

Community detection revealed that Reddit's network is made up of numerous tightly knit groups. The largest communities, centered around specific interests such as gaming and Q&A forums, highlight the platform's ability to target diverse groups of users while maintaining connectivity through key nodes.

Nodes such as **SubredditDrama** and **OutOfTheLoop** act as key connectors, facilitating the flow of information between different parts of the network. This underscores the importance of subreddits, which aggregate and contextualise discussions, increasing the overall connectivity and coherence of the Reddit network.

Reddit's Hyperlinks network shows how structure, centrality and community work together in a large social network. The analysis shows how the platform keeps a diverse and dynamic ecosystem going, with influential nodes and cohesive communities, so that information flows and users engage.