Study on the Network Structure of the Lichess Games Database

Afonso Vaz - 64505
fc64505@alunos.fc.ul.pt
Almada, Portugal

Gonçalo Moreno - 49783
fc49783@alunos.fc.ul.pt
Montijo, Portugal

Sebastião Cancela - 58282
fc58282@alunos.fc.ul.pt
Lisboa, Portugal

Abstract

This project was a study on the the network dynamics and characteristics of the users and games from the Lichess website. For such, we not only had to analyze the common graph statistics such as centrality measures and distribution of nodes, but also if there are community-like structures in our network. Finally, we compute distance metrics to analyze how close together different players from different rating brackets are.[1]

Keywords

Network, Graph, Degree, Elo, Node, Edge

1 Introduction

The dataset used in this study is derived from the games played on Lichess, a popular online chess platform, during the month of April 2016. With 5,922,667 of games recorded, this dataset provides an introspective view into competitive chess interactions across a diverse spectrum of player skill levels. As fans of chess, this dataset was chosen because it offers a large-scale and rich network representation of competitive chess players. Such a dataset is theoretically ideal for exploring network structures, community detection and player influence within the chess community. [1]

In this project, the primary objective is to investigate the structural properties of the Lichess competitive network. By utilizing network analysis methodologies, we aim to study the following:

- **Structural Analysis**: The structural properties of the network, to understand how players interact and form competitive hierarchies.
- **Node Centrality**: This analysis will reveal the most "influential" players and their potential roles in shaping the dynamics of the network
- **Community Structure**: The community structure of the network, which will provide insights into potential player groupings formed by competitive relationships.
- **Distance and Connectivity**: Analyze metrics that will evaluate the cohesion of the player network.

In the context of network representations, the nodes in the graph represent individual players, while the directed edges denote victory relationships, where an edge from node A to B indicates that player A won against player B. This directed network provides additional attributes for each node, including player titles (e.g. Grandmaster, International Master) and ELO ratings[2], thus, enriching the network with meaningful metadata for deeper analysis.

By conducting this analysis, the project aims to provide an understanding of how players relate to each other within the Lichess platform, highlighting competitive relationships and the formation of communities. Ultimately, this study contributes to a broader application of network science in modeling competitive chess systems.

1.1 Data Cleaning and Transformation

To ensure analytical accuracy and data integrity, the raw dataset underwent a pre-processing pipeline that addressed noise, redundancy, and missing information. The primary steps were as follows:

1.1.1 Exclusion of Non-Decisive Games:

- Games that ended in draws were excluded because they do not establish a winner-loser relationship, leaving only games with decisive outcomes.

1.1.2 Standardization of Player Strength Metrics:

- ELO ratings, a universally recognized measure of chess skill, were converted to numeric values. This ensured compatibility with subsequent analysis and facilitated classification of players according to their skill levels.

1.1.3 Node and Edge Construction:

- Separate dataframes for nodes and edges were created. The nodes dataframe contains information about all players, including their titles and ELO ratings, and contains 117,150 nodes.
- The edges dataframe aggregates the number of wins between specific pairs of players to account for multiple games played between the same opponents and contains 4,339,980 edges.

1.1.4 Graph Representation:

- A directed graph G was constructed using the processed data, where each node represents a player and each weighted edge captures the outcome of games between two players. The weight attribute signifies the frequency of victories against another specific player.

1.2 Relevance and Scope

The transformation of raw data into a structured, analyzable network allows for the application of sophisticated graph-based techniques to explore matters of centrality, community structure, and competitive dynamics.

2 Network Structural Analysis

In this section, we perform an in-depth analysis of the structural properties of the directed network. This analysis uncovers insights into matchmaking tendencies and the network's overall characteristics.

[1] Github repository for the project **here**.

[2] Lichess uses the Glicko-2 rating system, [2], however, for the sake of simplicity, we will be referring it to as ELO, since it is a universal metric.

2.1 Degree Distribution

The degree distribution captures the proportion of nodes (players) with a given degree, where the degree of a node represents the number of distinct opponents a player has competed against.

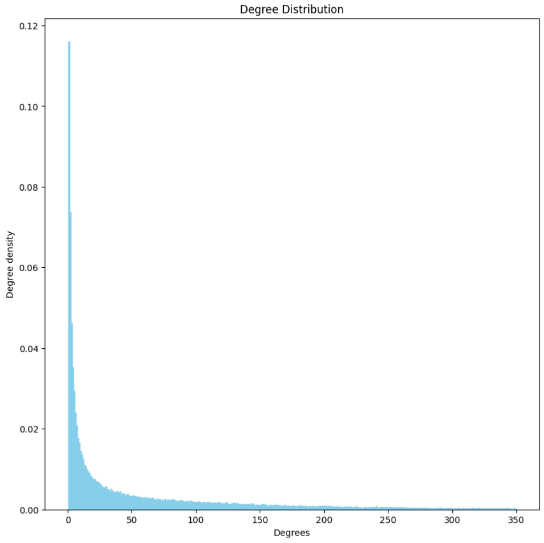


Figure 1: Bar Plot representing Degree Distribution

As seen in *Figure 1*, the degree distribution is heavily skewed, with a large proportion of nodes concentrated at lower degrees. Approximately 12% of players played against only one opponent, and the density decreases sharply as degree increases.

This skewness suggests that the majority of players interacted minimally with the chess platform, likely reflecting a casual player base. The rapid decline in density for higher degrees indicates that only a small fraction of users engaged in a substantial number of games or played competitively over extended periods. [3]

The long tail of the distribution, while thin, represents non-active players who participated in very few matches. On platforms like Lichess, this is often a result of the dual nature of user engagement: casual users who play sporadically and elite players or dedicated enthusiasts who are significantly more active.

It's also worth noting that the degree distribution was plotted only up to 350 degrees for better visualization, considering the max degree was 2500, and that at a certain point the degree density would be so small that it wouldn't be seen in the plot.

2.2 Assortativity

2.2.1 *Degree Assortativity.*

The degree assortativity coefficent, calculated as 0.162, quantifies the tendency of nodes to connect with others that have a similar degree. This positive but modest value indicates that players with similar levels of activity (in terms of opponents faced) are somewhat more likely to compete against each other. This reflects a degree of homophily, where casual players tend to play other casual players, and highly active players are more likely to interact with other highly active players.

2.2.2 *ELO Assortativity.*

Similarly to Degree Assortativity, ELO Assortativity calculates how likely nodes are to be connected to nodes with similar ELOs. The numeric assortativity coefficient for ELO, calculated as 0.536, shows a moderate to strong positive assortativity based on player ratings. This means that players with similar ELO ratings are more likely to play against one another. Such a pattern aligns with how chess platforms typically pair opponents to ensure games are competitive. Higher-rated players tend to play each other, while lower-rated players cluster together in matches, maintaining a balanced competitive environment.

2.3 Graph Density

The graph density, calculated as 0.000316, is extremely low, indicating that the network is very sparse. This is expected given the scale of the dataset and the nature of online chess. While there are many potential interactions between players (given the large number of users), only a tiny fraction of those potential connections are realized as actual games. [4]

In practical terms, this sparsity highlights that the majority of players have limited interactions within the network. For competitive platforms, this could reflect a wide user base where only a small subset of users actively engage in repeated interactions, in this case, games.

2.4 Connected Components

2.4.1 *Strongly Connected Components.*

Strongly connected components (SCCs) are subgraphs where every node can reach every other node in the component via directed paths. The network contains 24,237 strongly connected components, many of which are single nodes. These single-node SCCs represent players who did not have reciprocal interactions with others - they either won or lost all their games without reversing the outcome against the same opponents, which is to be expected, since finding the same opponent multiple times is only something that happens occasionally in high ranking games. [5]

The biggest strongly connected component is comprised of 92,374 nodes and has an average rating of 1584, which makes sense, since this component has 78.9% of the players of the network and considering that when a new player creates an account it starts with 1500 rating. [2]

2.4.2 *Weakly Connected Components.*

Weakly connected components (WCCs), calculated by ignoring edge directions, show a much more interconnected structure. With 730 weakly connected components, we can observe that many players are part of clusters where paths exist between all nodes if the graph is treated as undirected, indicating that most players are indirectly connected through chains of matches, reflecting the overall interconnectedness of the user base despite the sparsity of individual player interactions.

The disparity between the number of SCCs and WCCs highlights the directed nature of the network. While the dataset reflects a vast number of one-sided relationships, the community as a whole remains loosely connected when directionality is ignored.

3 Node Centrality

The concept of node centrality is integral to understanding the roles and relative importance of individual nodes within a network. In the context of this study, centrality metrics highlight the relevance of players in the chess games network, revealing their "influence" and connectivity. This analysis focused on three core centrality measures: Degree Centrality, Closeness Centrality, and Betweenness Centrality, each providing unique insights into player interactions. Also important to note that, while Degree Centrality was computed using the entire network, Closeness and Betweenness Centrality were calculated using a subset compromising players with ELO ratings above 2300[3], having a total of 1172 nodes and 19993 edges. This subset was chosen to ensure computational feasibility given the intensive amount of total nodes and edges in the original dataset.

3.1 Degree Centrality

Degree centrality measures the importance of a node based on the number of direct connections it has in the network. For a chess network, this metric quantifies the influence of a player by the number of opponents they played against.[4] Degree centrality is further decomposed into:

3.1.1 In-Degree Centrality.

- This measures the proportion of nodes with edges directed toward a specific node. For a chess player, this corresponds to the number of games they lost to opponents.

  In-degree centrality of node $v = \frac{\text{in-degree of } v}{N-1}$

3.1.2 Out-Degree Centrality.

- This measures the proportion of nodes a specific node has directed edges toward. For a player, this reflects the number of games they won.

  Out-degree centrality of node $v = \frac{\text{out-degree of } v}{N-1}$

3.1.3 Overall Degree Centrality.

- This is the sum of the in-degree and the out-degree centrality, representing a player's total engagement in the network.

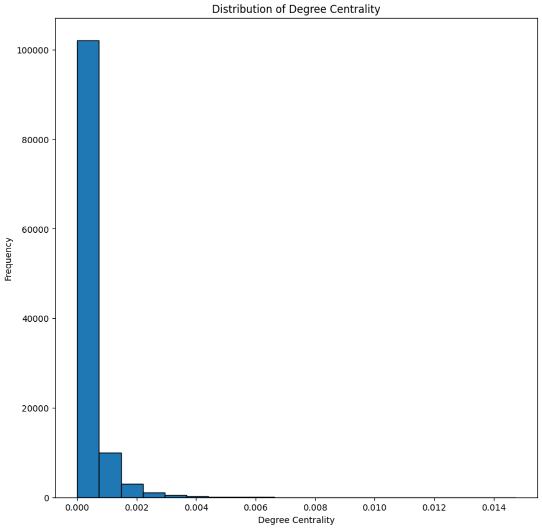


Figure 2: Histogram representing the Distribution of Overall Degree Centrality

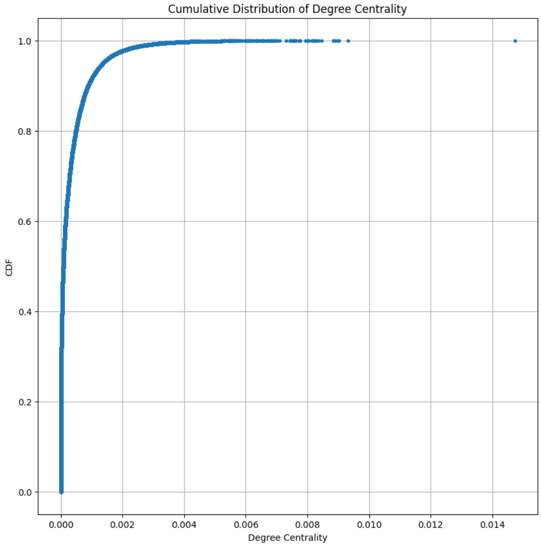


Figure 3: Function representing the Cumulative Distribution of Overall Degree Centrality

The results indicate a highly skewed distribution for all three metrics. Most players exhibit low centrality values, engaging with a limited number of opponents. However, a small number of players stand out as hubs with significantly higher centrality, representing those who played against many opponents and, consequently, have a disproportionate influence within the network.

Visualizing the degree centrality distributions through histograms and cumulative distribution functions can provide further insights into the structural properties of the network:

- **Distribution of Degree Centrality (Histogram)**: As shown in Figure 2, the majority of the nodes (players) are concentrated at the lowest range of degree centrality values.
- **Cumulative Distribution of Degree Centrality**: Figure 3 illustrates that a vast majority of nodes possess degree centrality values close to zero. The curve rapidly approaches a plateau near 1, confirming that the bulk of the network consists of players with minimal connections, while only a few nodes maintain disproportionately high centrality values.

These patterns are consistent across the in-degree, out-degree, and overall degree centrality distributions, thus indicating similar structural characteristics regardless of directionality.

3.2 Closeness Centrality

Closeness centrality measures the efficiency with which a node can access all other nodes in the network. It is computed as the reciprocal of the sum of the shortest path length from the node to all other nodes in the graph, $C(v) = \frac{N-1}{\sum_{u \neq v} d(u,v)}$, where $d(u,v)$ is the shortest path distance between nodes u and v.

In the context of chess games, nodes with high closeness centrality correspond to players who can, in theory, "reach" others with minimal intermediary steps. This measure is meaningful for identifying central figures in a network where efficient reachability matters. [6]

[3] The subset was created for the top 1% of players, which corresponded to about 2300 ELO and above.

[4] For the sake of context, every player has played, on average, against 74 unique opponents, with the exception of an outlier that played against 2648.

To manage computational constraints, closeness centrality was calculated on players from the previously mentioned smaller subset. This subset was chosen because highly rated players are more likely to have meaningful connections in competitive play, as in, play with other players more than once.

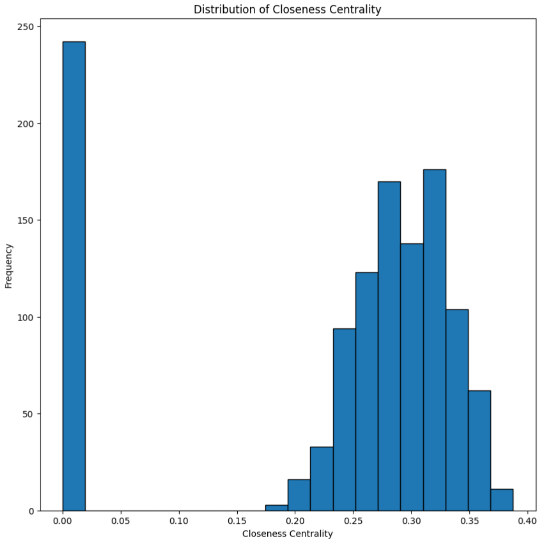


Figure 4: Histogram representing the Distribution of Closeness Centrality

The closeness centrality results reveal that a significant amount of nodes have relatively low or zero closeness centrality values. This reflects the fragmented nature of the network, because the low scores for these nodes suggest that their reachability within the network is minimal. They are mostly isolated nodes or part of disconnected components. Then we have the rest of the nodes with values of 0.175 and above. These players are part of the network's main connected component and they can reach other players in the network relatively efficiently. This goes accordingly to what we've seen in subsection 2.4, for the Strongly Connected Components.

3.3 Betweenness Centrality

Betweenness centrality quantifies the extent to which a node acts as a bridge between other nodes by appearing on their shortest paths. It is defined as, $B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} represents the total number of shortest paths between nodes s and t, and $\sigma_{st}(v)$ is the number of those paths that pass through v.

In the context of a chess network, nodes with high betweenness centrality would represent players who act as intermediaries, facilitating connections between otherwise disparate parts of the network. Such players are critical for the overall connectivity of the network.

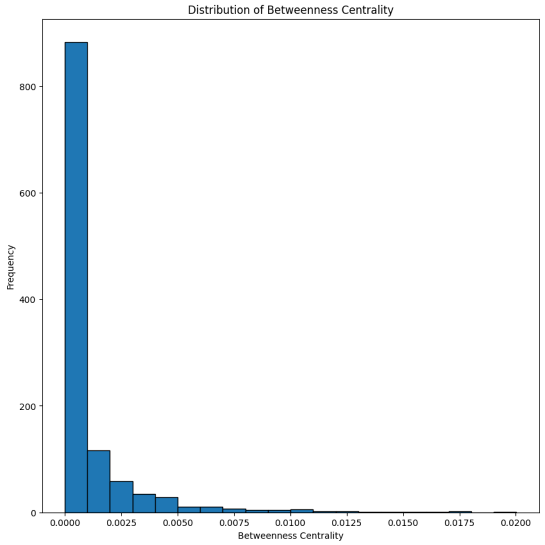


Figure 5: Histogram representing the Distribution of Betweeness Centrality

Like closeness centrality, betweenness centrality was calculated for the ELO > 2300 subset to ensure computational feasibility. Results showed that the majority of players exhibited low or zero betweenness centrality values. This suggest that most players are not functioning as significant bridges within the network, instead, they are primarily connected directly to their opponents or within smaller subgroups.

Key Observations:

- **Localized and Modular Network**:
  - The results reveal that the chess games network is likely highly modular in nature, meaning players compete within distinct clusters, in this case, the specific ELO rankings > 2300.
- **Lack of Intermediaries**:
  - No players exhibited exceptionally high betweenness centrality values, which suggests an absence of key "connecting" nodes that bridge large sections of the network.
- **Minor Connectors**:
  - A few players had slightly higher betweenness centrality values, indicating that they might serve as modest bridges between certain clusters. However, their roles as intermediaries are not particularly pronounced in the overall network structure.

4 Community Structure

Community detection is an essential part of network analysis, where we try to find groups of nodes that are more densely connected with each other than with the rest of the network. When attempting to use community-finding algorithms on the entire network, we quickly found that, like before, it was computationally impossible. As such, we once again decided to utilize the subset consisting of the top players.

There were many ways we could've on chosen how to filter our graph, but we ended up selecting only the 99th percentile of players by elo. We decided on this because the size of the resulting network was of decent proportion for computation, but also because as we have seen, the higher ranked players are more likely to play against each other than in the lower ranks, and so more communities are likely to form and be visible.

For assistance with this, we chose to use the application Gephi, which allows for great visualisation of our network and its different node properties, alongside being able to calculate some important algorithms that we'll use further ahead.

Firstly, we examined the graph according to the elo of the players, grouping by the nearest one hundred.

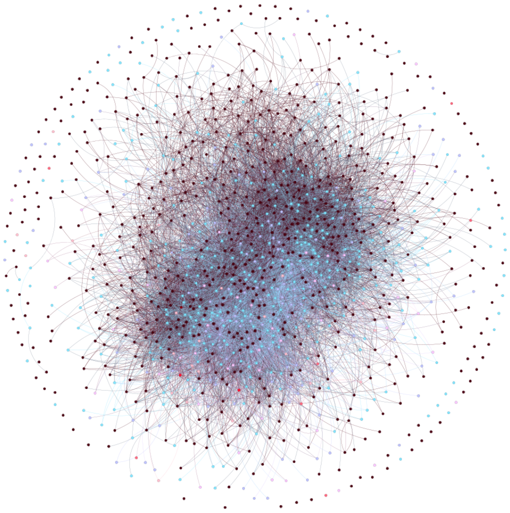
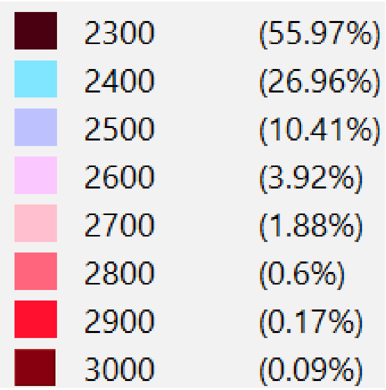


Figure 6: Network visualization according to elo



🟫	2300	(55.97%)
🟦	2400	(26.96%)
🟪	2500	(10.41%)
🟪	2600	(3.92%)
🟥	2700	(1.88%)
🟥	2800	(0.6%)
🟥	2900	(0.17%)
🟥	3000	(0.09%)

Figure 7: Elo colour labels

This view allows us to see there's a vast predominance of the lower end of the ranks, and that they're mostly clumped with each other. It is now interesting to compare to one of the community generating algorithms (Louvain) [7]

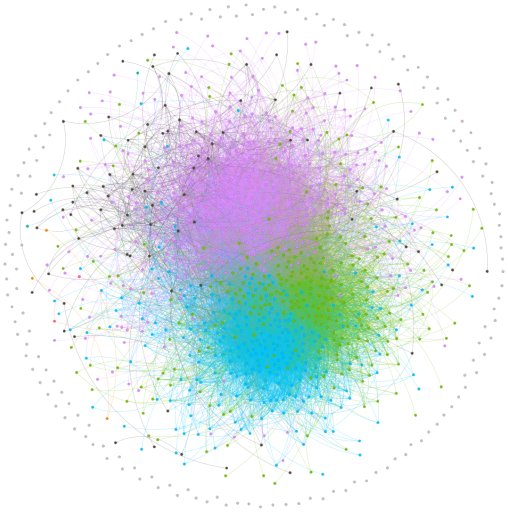


Figure 8: Louvain communities

Here, we observe that there are clear communities created, with some overlap between them. It's not so visible due to the size of the larger communities, but in reality there are 165 communities. With a corresponding modularity score of 0.336, we can conclude that although it looks visually pleasing, the algorithm does not produce communities of optimal significance.

5 Distance Metrics and Connectivity

In this section, we discuss the concepts of average distance, graph diameter, and path length within the overall dataset. These metrics offer insights into the structural properties of the network and how interconnected players are in terms of sequential victories [5].

5.1 Average Distance

The average distance between pairs of nodes reflects the typical number of steps or "wins" required to connect two players in the network. For our dataset, the computed average distance was 3.68, indicating that, on average, a player can connect to another in approximately four steps through the "won against" relationship. We also computed the average distance for the subset composed of the top 1% of players and found that the result was 2.778, indicating an even more compact network.

These results highlights a network that is relatively well-connected, where players can be reached through only a few sequential victories. The low average distance reflects a competitive environment characterized by frequent and widespread interactions between players.

5.2 Graph Diameter

The diameter of the overall network represents the longest shortest path between any two players within this subset. For our dataset, the diameter was calculated as 9. This means that the furthest two

[5] The results indicated in this section only represent an approximation and not the real values.

players in this strongly connected component are separated by 9 victories.

The diameter provides a sense of the network's "depth" or maximum reach. This indicates a more centralized and compact network, where all players are relatively close in terms of sequential wins. [8]

5.3 Path Length Distribution

In this section, we decided against calculating the path length for the strongly connected component of the 2300+ ELO subset, as such an analysis would have been less meaningful, as it would merely reveal the shortest path length within an already tight-knit and highly competitive community.

Instead, we opted for a more interesting approach: calculating the shortest path length between the player with the lowest ELO rating and the player with the highest ELO rating across the entire network.

The result was remarkable. Only 4 victories separate the "worst" player from the "best". While surprising, this finding aligns with the overall structure of the dataset, where the average shortest path length is 3.68. This underscores the interconnected nature of the player network and highlights how even the most skill-diverse participants are, in some sense, closely linked.

6 Conclusion

This study analyzed the network structure of Lichess games, uncovering relationships between players and how these interact with each other.

Firstly, the network is characterized by a high degree of sparsity, which reflects the casual nature of most users' participation. While the majority of players engage in a limited number of matches, a smaller group (comprised of highly active or even competitive players) stands out as key contributors to the network's connectivity.

Secondly, the analysis of centrality measures revealed that "influential" players are often those with higher ELO ratings, which makes sense given their reduced number of opponents and increased activity levels between them. These players act as hubs within the network, connecting different parts and playing a strong role in shaping the structure of the communities they find themselves in. However, the majority of players exhibit low centrality scores, thus reflecting less prominent roles within the network.

The community detection analysis also provided meaningful insights, demonstrating that players with similar ELO ratings naturally match together. This clustering is a direct consequence of the matchmaking system, which pairs players with comparable skill levels to ensure fairness in competitive games.

Finally, the distance metrics and connectivity analysis painted a picture of a network that is surprisingly strongly connected. The average path length of nearly 4 and the relatively small diameter highlight a tightly-knit network, where each player can reach another in only a small amount of victories, even taking into account players with extremely diverse ELO ratings. This was definitely not to be expected of, since we initially believed that the network would be more constrained between ELO brackets.

Overall, this study not only gives us a glimpse into the dynamics of competitive chess on Lichess, but also demonstrates the broader applicability of network science in understanding large-scale interactions in digital competitive environments.

Acknowledgments

Our thanks go out to OpenAI, the application Gephi, and our professor André Falcão.

References

[1] Predrag Obradović and Marko Mišić. "Network dynamics of the online chess platform Lichess: A social network analysis case study". In: ().

[2] Lichess. *Chess rating systems*. https://lichess.org/page/rating-systems.

[3] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[4] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[5] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[6] Maarten Van Steen. "Graph theory and complex networks". In: *An introduction* 144.1 (2010).

[7] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[8] Kun Zhang et al. "Towards identifying influential nodes in complex networks using semi-local centrality metrics". In: *Journal of King Saud University-Computer and Information Sciences* 35.10 (2023), p. 101798.

Annex

A In-Degree Centrality Charts

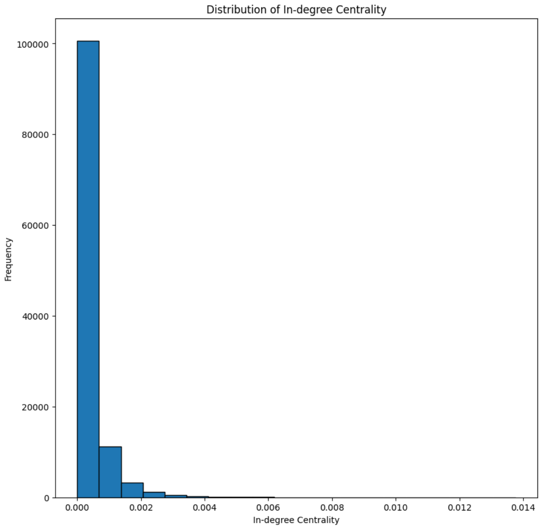


Figure 9: Histogram representing the Distribution of In-Degree Centrality

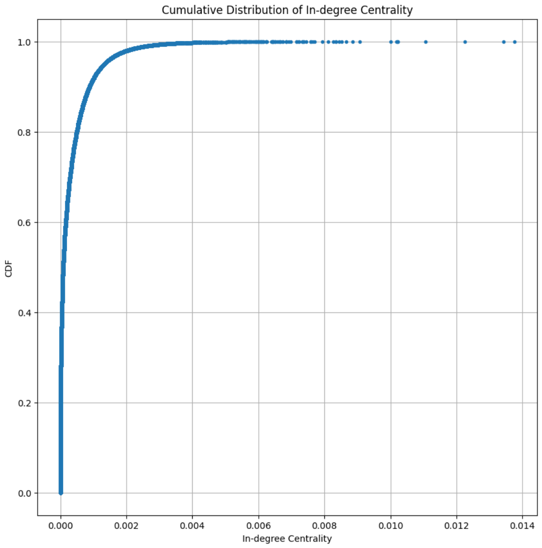


Figure 10: Function representing the Cumulative Distribution of In-Degree Centrality

B Out-Degree Centrality Charts

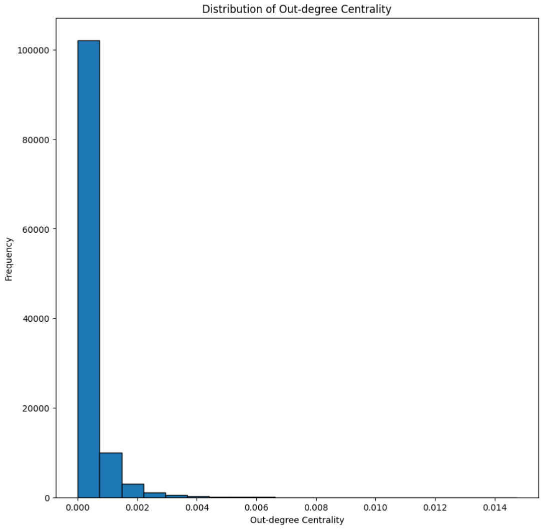


Figure 11: Histogram representing the Distribution of Out-Degree Centrality

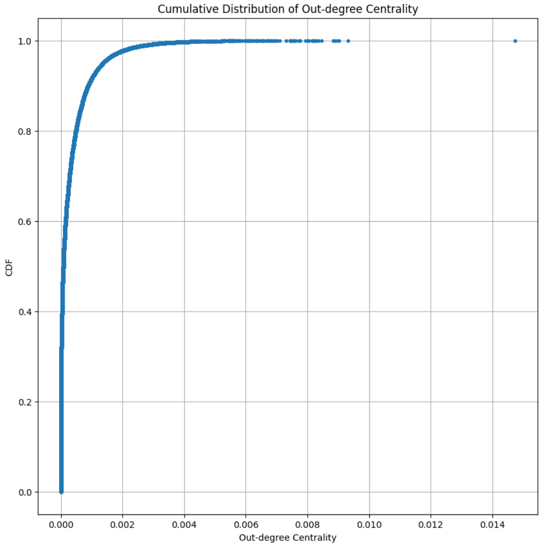


Figure 12: Function representing the Cumulative Distribution of Out-Degree Centrality

C Closeness Centrality Cumulative Distribution

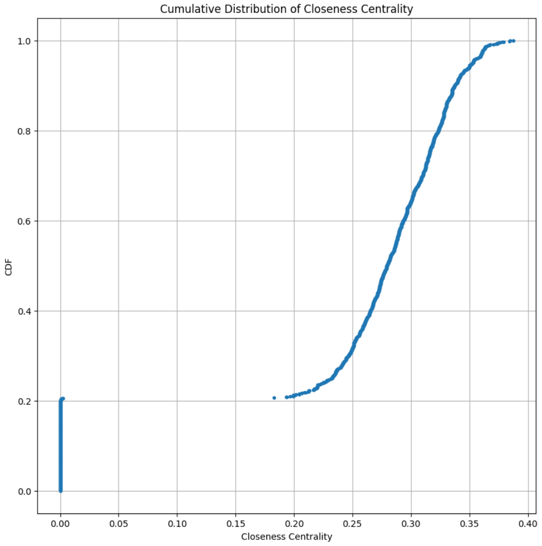


Figure 13: Function representing the Cumulative Distribution of Closeness Centrality

D Betweeness Centrality Cumulative Distribution

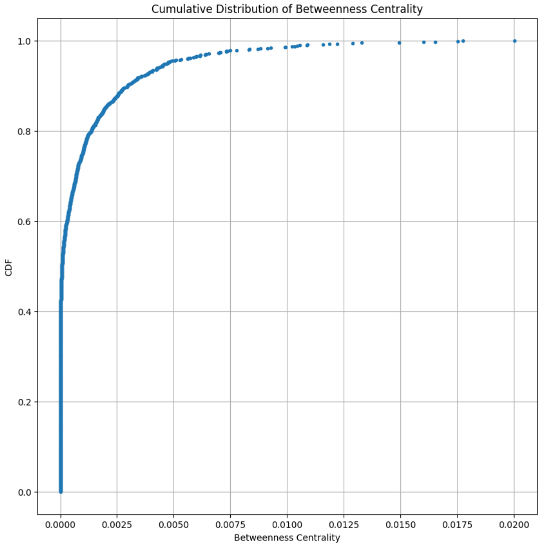


Figure 14: Function representing the Cumulative Distribution of Betweeness Centrality

E Edge Betweeness Centrality Charts

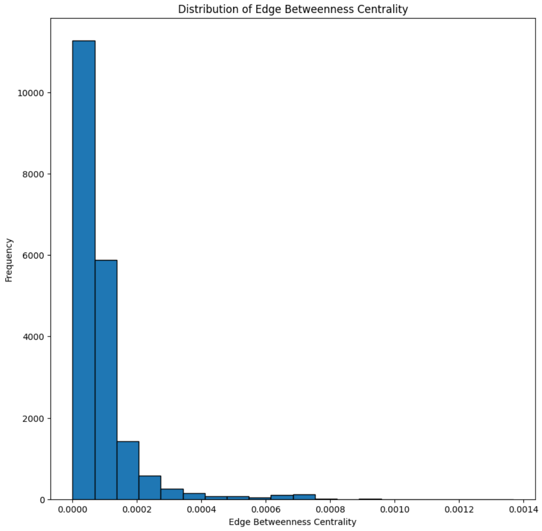


Figure 15: Histogram representing the Distribution of Edge Betweeness Centrality

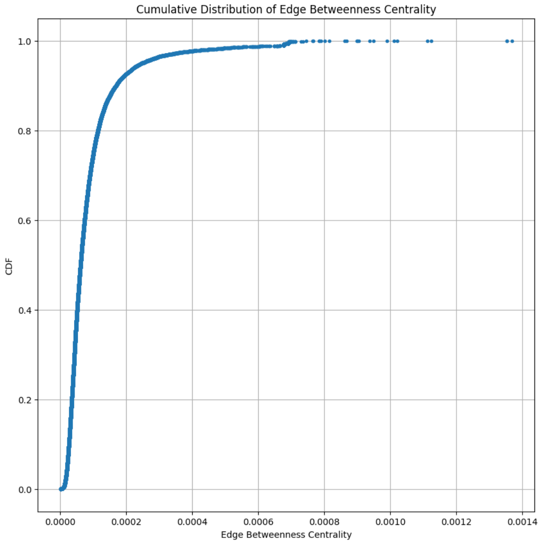


Figure 16: Function representing the Cumulative Distribution of Edge Betweeness Centrality