Math 525 Final Project: Historical Match-ups of Competitive Chess Players

Andrew Belz

December 9, 2025

1 Introduction

1.1 About the Data

The network I chose to analyze for this project was historical game match-ups of competitive chess players. The data set I used is called Lumbra's Gigabase[1], downloaded as a SQLite file. From this database I extracted game data, joined it with player information, and saved the result as a .csv file.

Note that FIDE is the international governing body of chess, and they assign ratings to players according to playing strength. The data is filtered to games played between players with FIDE ratings greater than or equal to 1800, which corresponds to roughly the top 30% of players. From FIDE's website, I was able to download an up-to-date list of player titles, which I was able to match to the game data by player name. Titles are awarded to players based on rating as well as achievement.[2]

Going into this analysis, I hypothesized that the structure would be highly reminiscent of a typical real-world network.

1.2 Graph Construction

The network was then constructed using the list of games as an edge list. Players without a game in this list are excluded. The resulting model is an unweighted multigraph $G = (V, E)$, where V is the set of all players with a recorded game and E is the set of all games in the database. One edge exists between players u and v for each recorded game that u and v have played against each other. For many measures recorded in this data set, multi-edges are either discarded or treated as a single edge with weight equal to its multiplicity. The graph has 526,519 vertices and 9,570,564 edges. Because of the large size of the graph, the C-optimized Python package **igraph** was used for many computationally intensive tasks, while **networkx** was used for the rest of the calculations. I used **matplotlib** for the visualizations in this report.

1.3 Notes About the Data

It is important to note the bias introduced by filtering the data to only include games played between players rated 1800 and above. Lumbra's Gigabase was created for the purpose of providing a large database of high quality games played between players of high caliber for study by serious competitive chess players, not for analyzing the network of interactions between competitive chess players. Besides the bias caused by omission, most of the other problems encountered in this data set are likely related to missing information or improper data aggregation. However, these errors are seemingly minimal, and didn't drastically impact the utility of Lumbra's Gigabase for this analysis.

The size of the data set did provide significant challenges in computing network measures exactly, so for many of the computed measures in this report are approximations or heuristics.

[1]Downloaded from encroissant.org. To learn more about Lumbra's Gigabase, visit lumbrasgigabase.com

[2]For more information about FIDE titles, visit https://www.chess.com/terms/chess-titles.

2 Centrality Measures

I used three primary measures of centrality for measuring player (node) importance:

1. **Degree Centrality.** For this graph, the degree of a node can be directly interpreted as the number of games played by the associated player that are listed in the edge list.

2. **Eigenvector Centrality.** To compute this measure, I computed eigenvector centrality on the weighted simple graph form of G. Rather than a multigraph, I chose to keep just one edge per player pair, and weight that edge by the multiplicity of the node pairs in the edge list.

3. **Betweenness Centrality.** This measure was also computed on the multigraph G. A representative simple random sample of nodes ($n = 1000$) was chosen to estimate edge-betweenness for each node.

The results were interesting. There wasn't very high correlation between betweenness centrality and eigenvector centrality ($R = 0.13$) or degree centrality ($R = 0.27$), but eigenvector centrality and degree centrality exhibited mild correlation ($R = 0.5$). The results are displayed in Table 1.

Table 1: Top 5 Players By Centrality

Rank	**Degree**	**Eigenvector**	**Betweenness**
1	Viktor Korchnoi	Viswanathan Anand	Levon Aronian
2	Russell Sherwood	Levon Aronian	Jaan Ehlvest
3	Vasyl Ivanchuk	Vasyl Ivanchuk	Lopez
4	Alexei Shirov	Vladimir Kramnik	Gawain Jones
5	Ivan Farago	Boris Gelfand	Efstratios Grivas

Degree centrality is, again, very easy to interpret. The players with highest degree centrality played the most chess games in their careers. For example, Viktor Korchnoi was a Soviet grandmaster whose competitive chess career spanned 70 years. He played several thousands of games. All top players by degree centrality share this trait. Each had a long, illustrious playing career.

Eigenvector centrality, while correlated with degree centrality, exposes a different yet equally insightful characterization of importance in this network. Indeed, it does a better job than both degree and betweenness centrality at underscoring the historical nature of the data. While players with high eigenvector centrality do tend to exhibit high degree as well, they sit at a unique place in chess history as well. These players were top players in the 90s and 2000s, a time when chess was rapidly changing. They had the unique opportunity of facing off against some of the greatest players of the late 20th century such as Anatoly Karpov, Gary Kasparov, and more, as well as being the world's strongest players in the emerging era of online chess, where players could play against each other via the Internet. They have also played against the greatest players of the 21st century, such as Magnus Carlsen, Fabiano Caruana, and others. Anand and Kramnik themselves were back-to-back Chess World Champions. Thus, eigenvector is the most historically insightful of the 3 measures, in my opinion.

Betweenness centrality seems to identify players who were well-traveled, and made competitive chess their day jobs. These were players that often moved or traveled to different countries and played regularly in a wide variety of tournaments. This led to these players being extremely well-connected in the chess community.

It is important to note that Lopez is ranked 3rd by betweenness centrality, but the lack of a recorded full-name does raise some concerns about the validity of this data point. It's possible the games of different players were aggregated by the same last name, so it may be necessary to treat this observation with a grain of salt. However, the rest of the results are consistent with clean, accurate data.

3 Community Structures

3.1 Assortativity

I decided to measure assortativity based on two primary node attributes: degree and title. I hypothesized that the network would be assortative by both degree and title.

Indeed, players are assortative by degree, with an assortativity coefficient of 0.39. More specifically, players with higher degree tend to play other players with higher degree. Figure 1 shows a heat map of games played plotted by the degrees of each player.

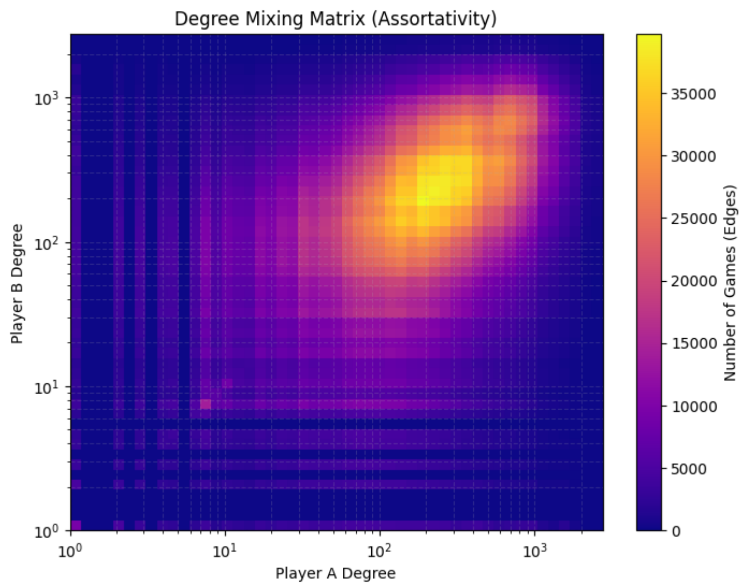

Figure 1: Assortativity by degree.

However, players do not tend to assort by title. The assortativity coefficient of this data set with respect to title is only about 0.003, which is nearly zero. Thus players do not prefer to play against players with the same title as themselves.

3.2 Clustering

The global clustering coefficient of the network was about 0.09. While this is seemingly not very high, it is much higher than what we would expect for a random graph. This suggests that to some extent, players do tend to play other players who have played each other as well.

3.3 K-Cores and Modularity

A natural follow-up question to the question of assortativity is regarding modularity. How modular is the network with respect to title? The modularity score with respect to title for this network was about 0.00027, which is nearly zero. This suggest that players of the same title do not form dense isolated communities.

In fact, the network exhibits a clear onion core-periphery structure. Figure 2 shows the number of nodes belonging to the k-core ranging across k. We see that the number of nodes belonging to the k-core quickly drops off as k increases, then begins to level off. This suggests a small, highly connected inner core of chess players that play against each other often.

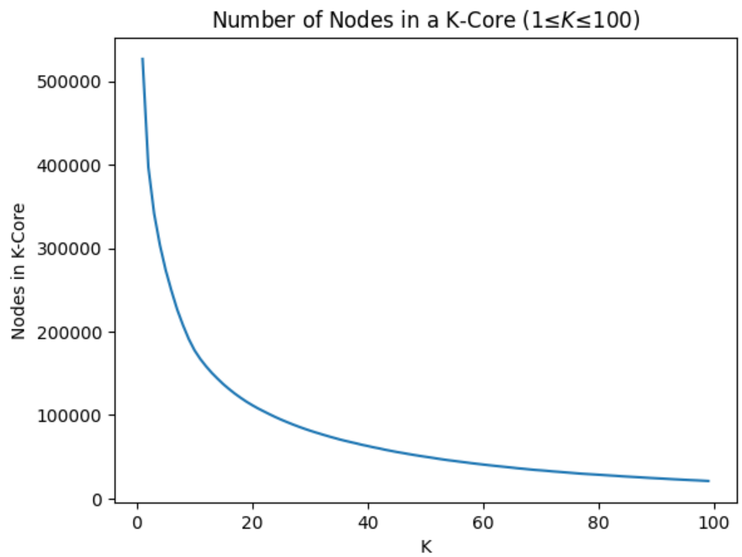

Figure 2: k-core sizes.

4 Real-World Properties

4.1 The Giant Component

The graph G is comprised of 5,267 connected components of varying sizes. The graph does have a giant component containing 514,305 nodes, or about 97.6% of players in the data set. The second largest component contains about 5,000 times fewer nodes than the giant component. There are several thousand components with extremely small size, and just a handful of components with size greater than 10. In general, the distribution of component sizes follows a heavy-tailed distribution, where small components are extremely common and large components are rare. This is a trait shared by many real-world networks. Figure 3 shows the distribution of these component sizes.

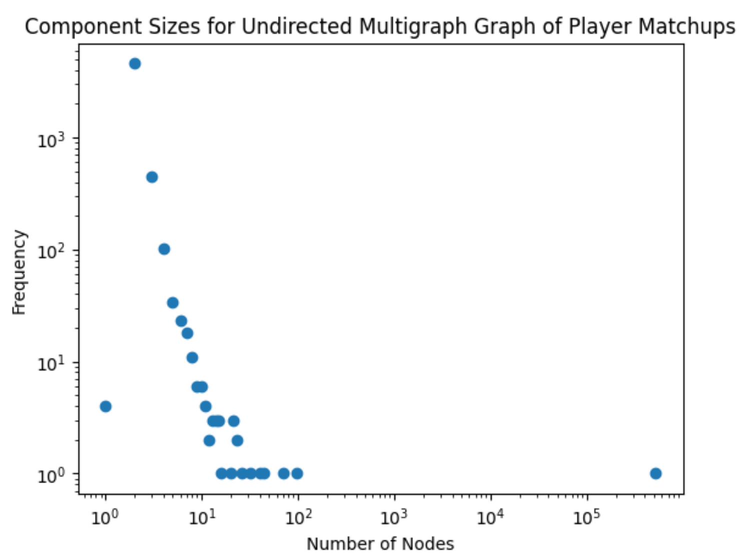

Figure 3: Frequency of component sizes in G.

4.2 Degree Distribution

This network exhibits a scale-free-like degree distribution, with some important nuances. Figure 4 shows the degree distribution of G.

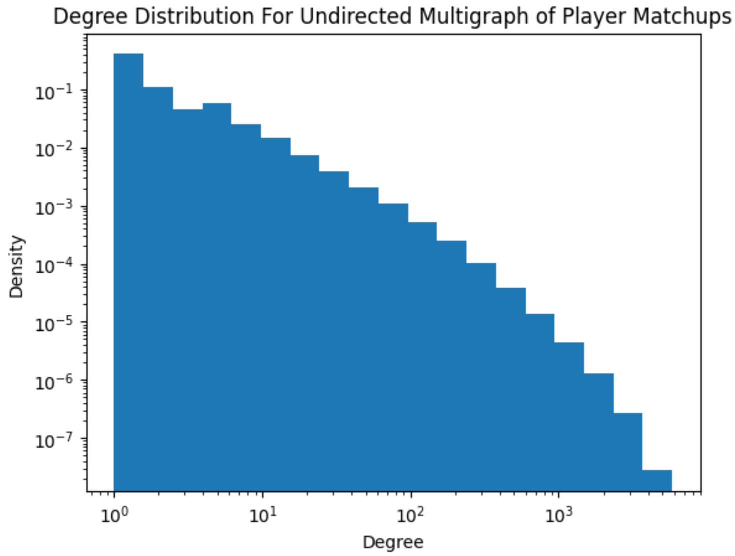

Figure 4: Degree distribution of G (log-log scale).

As seen in the figure, the degree distribution is power-law-like at first but with a truncated tail, differing from a theoretical scale-free network. Indeed, we find that in fitting a power-law to the distribution with $k_{\min} = 20$, we obtain $\alpha \approx 1.781$, with $\sigma \approx 0.002$. This indicates that a power law is actually a decent fit for the distribution. Visually, we do see this steep drop-off in the tail. This can easily be explained by the physical limit to how many games a chess player can play in their lifetime.

4.3 Small-World Properties

This network has striking small-world-like properties, such as high global clustering and short geodesics. Recall that earlier we identified the global clustering coefficient to be about 0.09, which is higher than what we would expect for a random graph. Moreover, the average geodesic length in the graph is about 4.5, which is right on par with what we would expect to see in a small-world network. The Watts-Strogatz model, with intermediate rewiring probability, may provide a reasonable description for the generation of this network, at least to the extent of describing general matching dynamics in competitive chess, but there are some important distinctions that ultimately make this an imperfect fit. For example, the generation of our network is temporally-dependent: there are certain players who could never possibly have played each other. Additionally, this model fails to explain the empirical degree distribution that we see here.

A generally better alternative is the Barabási-Albert model. This model would do a better job at explaining some of the temporal dependencies, the power law distribution, and the observed preferential attachment by degree, but it fails to explain the structure of disconnected components. Additionally we run into the same problem, where the issue of the impossibility of certain match-ups would require an adjustment of the rules for connecting new nodes.

5 Conclusion

This network of chess player match-ups does exhibit a wide variety of different properties exhibited by real-world networks. It was surprising to see just how much insight into the world of competitive chess could be derived just from analyzing this network.