Exploring a Chess dataset - Finding Garry Kasparov

Henrique Branquinho
Faculdade de Ciências - Universidade do Porto
Porto,Portugal
up201804341@fc.up.pt

ABSTRACT

Chess is a very competitive sport. Many skilled players duel to become the one world champion and go down in chess history. In this paper, we analysed a chess network, containing match results from 1998 to mid 2006 between high rate players. With tools such as degree analysis, PageRank and motif analysis, we extracted some interesting properties from the network. We were able to successfully identify some of the top players by their game results, using an heuristic to calculate ELO ratings that we validated with PageRank, and to understand how top players behave. We were also able to find out what are the most common tournament types in high competition chess, and to classify this network by comparing it to other networks with motif analysis.

1. INTRODUCTION

Chess if one of the world's oldest and most challenging games. Through history, chess has always been seen as an intellectual battlefield between two individuals, where one tries to outsmart the other to successfully trap the enemy's king. To this day, the number of different games and moves that are possible in a chess game is still unknown. Clause Shannon presented a lower bound for this number to be 10^{120} [6]. The estimated number of atoms in the observable universe is roughly 10^{80}. Therefore, it is more easy to understand the universe than to fully master the art of chess (or so I believe, being a low rated player). In current days, it is a very competitive sport, where very few can try to contest the current world champion and set their own reign.
In this article, we will explore a chess dataset that contains chess results between top tier players from 1998 to mid 2006. The objective of this paper is to do an exploratory analysis of the network to try and get some meaningful conclusions about the network's structure and to try to rank top players according to match results. The players are only identified by a number, so there is no way to associate these results with the actual players. We will try to identify the highest rated player in that period of time through the network's structure, and try to get a sense of how these matches were organized.

1.1 Motivation

Chess is my favorite sport to study and to play. I decided to use this network as it is not a very large network so I can easily study it and because I found the topic interesting.

1.2 Organization

On section 2 we present some terminology and concepts that will be used while exploring the network. On section 3 we present the datasets used on this paper. On section 4 we present the exploratory analysis made on the network. On section 5 we summarize some important conclusions taken from the analysis.

2. BACKGROUND

In this section we introduce the terminology used throughout the paper along with some network science concepts used to analyze the network. We also provide some background into the world of chess and chess ratings.

2.1 Terminology

A graph G is defined as a tuple (V, E) where $V = \{v_1, v_2, ...\}$ is a set of vertices (or nodes) and E is a set of edges $e_1 = (v_1, v_2)$ that connect two vertices of the graph. The size of the graph is given by $k = |V|$ and G is said to be a k-graph. Edges can have direction. In this case, we say the graph is *directed* and that it has an edge $e_1 = (v_1, v_2)$ from v_1 to v_2. Edges can also be annotated with a value w. w is said to be the edge's *weight* and the graph is said to be *weighted.* A graph is said to be *simple* if it has no *self-loops* (edges connecting one vertex to itself) and no multiple edges (more that one edge connecting the same pair of vertices).

The *degree* of a vertex $d(v)$ is the number of edges that have v as one of its endpoints. In the case of a directed graph, the *in-degree* of a vertex v is the number of edges to v, and the *out-degree* is the number of edges from v. The *weighted* degree of a vertex is the sum of the weights of the edges taken into account to calculate the vertex degree.

A graph is said to be a *temporal* graph if it takes time into account. Edges can be labelled with a timestamp t, meaning they were added to the graph at that time.

A *path* between two vertices v_1 and v_2, $p(v_1, v_2)$, is a set of edges that allow us to traverse from v_1 to v_2. We say that v_1 is at distance $d(v_1, v_2) = |p(v_1, v_2)|$ from v_2. The *shortest path* between two vertices v_1, v_2 is the path that minimizes $|p(v_1, v_2)|$.

A *subgraph* G_k of a graph G is a k-graph in which $V(G_k) \subseteq V(G)$ and $E(G_k) \subseteq E(G)$. A subgraph is said to be *induced* if $\forall v, w \in V(G_k) : (v, w) \in E(G) \Rightarrow (v, w) \in E(G_k)$.

A *connected component* is a connected induced subgraph of G, such that no node in the component is connected to nodes outside of the component. The biggest component of the network is referred to as the *giant component.*

A graph is *connected* if there is a path connecting every pair of vertices of the graph, meaning it only has one connected component, which is the graph itself.

2.2 Centrality Measures

In this subsection we present the definition of some centrality measures, that allow us to evaluate the structural role of a vertex in a graph.

Betweenness Centrality captures how a vertex i is between other vertices in the network, though shortest paths. It calculates the number of shortest paths between all pairs of vertices (j, k) that pass through i. It is calculated as:

$$C_b(i) = \sum_{j<k} \frac{g_{j,k}(i)}{g_{j,k}} \tag{1}$$

where $g_{j,k}$ is the number of shortest paths connecting j and k, and $g_{j,k}(i)$ is the number of shortest paths between j and k that pass through i.

Closeness Centrality is based on how close a node is to all of the other nodes in the network, through the average shortest path length between the node and the other nodes. It is calculated as:

$$C_c(i) = [\sum_{j=1}^{N} d(i,j)]^{-1} \tag{2}$$

2.3 Communities

A *community* can be defined as a subset of nodes of a graph that have many internal connections and few external ones (to the rest of the network). Community discovery can be done by maximizing a parameter that measures how well the network is divided into communities. This is usually done by maximizing *modularity.* Modularity can be calculated as:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \sigma(c_i, c_j) \tag{3}$$

where: A_{ij} represents the edge weight between nodes i and j (for unweighted, it is simply 1 if they are connected or 0 otherwise), k_i and k_j are the sum of the weights of the edges attached to i and j, $2m$ is the sum of the weights of all edges in the graph (number of edges in unweighted case), c_i and c_j are the communities of the nodes i and j , and σ is an indicator function. (Definitions taken from the class slides).

2.4 PageRank

PageRank is an algorithm used by Google's search engine that measures the importance of web pages. PageRank can be used to study a network and to find the most important nodes in that network. Each node is attributed an initial PageRank value (usually , $\frac{1}{N}$ where N is the size of the graph. Then, each node's PageRank is updated according to previous PageRank values. This algorithm repeats until converging (the difference between PageRank values from one iteration to the next is less than a pre-determined value ϵ). For a given node, PageRank is calculated as follows:

$$PR(u) = (1 - \beta) + \beta \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{4}$$

$B(u)$ is the set of nodes that have an edge directed to u. $PR(v)$ is the PageRank value of node v in the previous iteration. N_v is the number of outgoing edges of node v. β is the damping factor, which is a probability of teleportation to another node without following edge direction.

2.5 Motifs

A *motif* was first described by Milo et al. [4] as a subgraph that appears more frequently in a network than in other randomly generated networks with same degree distribution, meaning it is over-represented. A motif is therefore a statistically significant subgraph. The finding of motifs can help us learn more about network's structure and can lead to network classification based on their motif fingerprint. This means we can look at subgraphs as building blocks of a network. G-Tries were used to analyse motifs [5].

2.6 Chess

A typical chess game is played between two players, each with 16 pieces. One player plays with white pieces, and the opponent playes with black pieces. The white pieces always move first. I will not go into detail on the rules of the game as they are not necessary for the understanding of the paper. A chess player can either win, lose of draw a game. Typically, a win counts as 1 point and a draw counts as a half point. Each player can be given a certain amount of time to make moves. Depending on the total time given to each player, a chess game can be classified as classical, rapid or a blitz game, in decreasing time available. World champions are decided on classical games, although there are world titles for rapid and blitz games. Chess players are rated according to the FIDE rating. FIDE stands for Fédération Internationale des Échecs , or International Chess Federation. The FIDE ratings of players are updated on FIDE tournaments or matches, according to the rating differences between the players. This means that the rating is balanced so that a high rated player does not win much rating by defeating a low rated player. Chess players are also awarded with titles according to their skill, the most notorious one being the Grandmaster (GM) title. All the top tier players are Grandmasters. To become a GM, among other requirements, a player needs to have a FIDE rating of at least 2500. The record for highest FIDE rating in classical chess belongs to Magnus Carlsen, with 2882 FIDE. He broke the world record in 2010 held by Garry Kasparov in 2010 when he reached a FIDE rating of 2863. In the time this dataset regards, world organization of chess championships were a mess. There were conflicts between FIDE and other institutions, so there was more than one world champion at the time. This weakens our analysis, as it is hard to get a reliable source of top players in this period. Therefore, our analysis will be very simplistic. There are usually three ways to organize a chess tournament:

1. Knockout tournament - Players are eliminated from the tournament when they lose a pre-determined number of games. The remaining players play with each other until there is only one player left.

2. Round-robin tournament - A player must play with all other players who are in the tournament. The winner is decided by total number of points at the end of a tournament.

3. Swiss-system tournament - When there are many players in a tournament, Round-robin becomes infeasible as it takes too long. In a Swiss tournament, players are matched with other players that have a similar performance on that particular tournament. This means that players who win more games will be matched against players who also won games, and players who lose games will be matched with players who also lost some games. This way, they don't have to play everyone.

2.7 Tools used

To analyse the networks, we used Gephi, some Python3 scripts we made, and gtrieScanner [5] for motif analysis.

3. DATASETS

In this section we present the datasets used throughout the paper. The networks we created were inspired by Michieli [3], who analysed tennis networks to create new ranking systems (this paper was presented in the Network Science class by João Carvalho).

The dataset used was taken from a Kaggle competition [7]. It comprises 65,053 games results among 8,631 of the world's top 13,000 chess players, played between 1998 and 2010. Of all the games they played, only 70% of these games are present in the dataset. The original network is a temporal weighted multi-edge directed graph. The graph has an edge from v_1 to v_2 if players 1 and 2 played a game. The direction of the edge indicates who played with the white pieces. The edge goes from the white pieces player to the black pieces player. Edges are annotated with a timestamp $t \in [1, 100]$, that represent different months. Edges are weighted with the game result $w \in \{0, 0.5, 1\}$. The result regards the white player, meaning that if an edge $e = (v_1, v_2)$ has $w = 1$ and $t = 3$, player 1 played with the white pieces against player 2 in month 3 and won. Of course this means that player 2 was awarded 0 points for this game. In the case of a draw, both players are awarded 0.5 points. From this network, we derived some other networks that were used for different purposes. Following is a list of these networks:

- **Network 1** - The original network
- **Network 2** - A directed weighted simple graph. This network is obtained by considering the scores each player obtained versus every other player in every game they played. Network 2 will have an edge from v_1 to v_2 if player 2 scored more than 0 points against player 1. The weight of this edge will be the total number of points player 2 scored against player 1. There can also be a symmetric edge with the same principle (imagine they only play one game and it's a draw. There will be 2 symmetric edges between them, each with 0.5 weight).
- **Network 3** - An undirected weighted simple graph. In this network, there is an edge between v_1 and v_2 if players 1 and 2 played at least one game. The weight of the edge is the total number of games they played against each other.

Networks 1 and 2 will be mostly used to evaluate a player's overall ranking. Network 3 will be used to study the overall network structure and how players relate to other players.

4. EXPLORATORY ANALYSIS

In section 4.1, we analyse Network 1, mainly focusing on the degree of the network's nodes, and with some early community exploration. In section 4.2, we analyse Network 3, mainly focusing on degree, edge weight analysis, and PageRank analysis. In section 4.3, we analyse win rates for the top players. In section 4.4, we propose an heuristic to rank players and present some results obtained from it. In section 4.5, we analyse Network 2, mainly focusing on a PageRank analysis to validate our heuristic. In section 4.6, we derive some conclusions regarding how chess players interact with each other. In section 4.7 we associate nodes in the network with famous chess players. In section 4.8, we perform a motif analysis on the network to study tournament organization.

4.1 Analysing Network 1

The first thing we see when analyzing Network 1 is that some nodes are missing, meaning that not all 8,631 players chosen have recorded matches in this dataset. This is because the original dataset was built as a training set for a classification task, and the missing players are in the test dataset. Network 1 has 7301 nodes.

White victories	Black victories	Draws
32.53%	23.4%	44.07%

Table 1: Distribution of wins or draws according to piece color

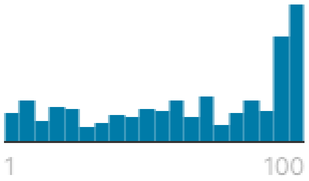

Figure 1: Distribution of timestamps

In table 1 we can already see a curious aspect of high performance chess. Most games between high rated players end in draws, sometimes very early on in the game. This is because chess players have a very solid understanding of positions and possible derivations, and already know some common responses to the opponent's moves, so they both understand that there is no way they can defeat each other, agreeing to a draw. It is also clear that the white pieces have advantage by making the first move, as they have a higher win rate than black pieces. In figure 1 we can see that the games are mostly evenly distributed through time, with a higher number of games being played in the final period the dataset comprises.

By the degree distribution in figure 2, we can see that the network appears to follow a power law distribution with an exponential cutoff. From this we can already conclude that most players in the dataset played few games, and that few players participated in the majority of the games. This could

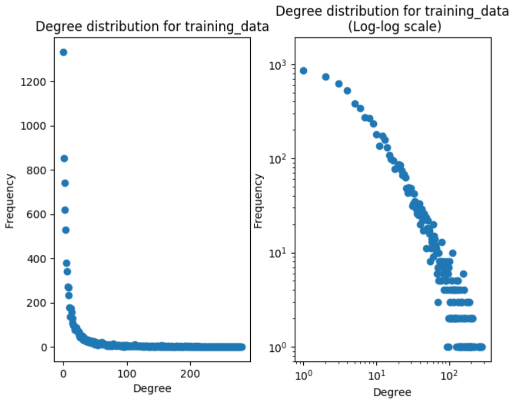

Figure 2: Degree distribution for network 1

Player ID	# of games	Player ID	# of games
4850	280	1512	257
1594	276	391	255
4037	272	1098	252
1286	270	4171	232
64	258	1397	221

Table 2: Top 10 players according to total games played (node degree)

be indicative that the highest rated players could be the ones who played more games. The average degree of this network is 8.91 (meaning every node has on average 8.91 in-degree and 8.91 out-degree), which means that each player played on average 17 games in the time comprised in the dataset. Just by looking at the nodes with higher degree in table 2 we can already get some clue about who the highest rated players could be, since they tend to participate and win at a lot of tournaments, meaning they end up playing a lot of games. The in-degree and out-degree of all nodes is pretty balanced, because when two players play more than one game, it is common that they switch color to even out the odds of each player winning. We analysed betweenness centrality and closeness centrality but since the top nodes coincided with the high degree nodes, we decided to not include these results in the article for the sake of simplification of information. The average path length in this network is 4.01 and the average clustering coefficient is 0.2, which means that there are many players with no games between them, but with many common opponents due to the relative small average path length.

Network 1 has 77 connected components, and the giant component of the network contains 7115 nodes (97.45% of the original network) and 64926 games (99.8% of all games). The nodes that are not in the giant component are probably weak players who only participated in few tournaments, retired players who played their last matches in the time corresponding to the beginning to the dataset, and new players who only started participating in the end of the dataset. In any case, we can surely discard these players as the highest rated players.

Considering only the giant component, we ran a community discovery algorithm in Gephi, without considering the edges weight. The modularity of the network was 0.502, which means the network presents a strong community structure. 17 different communities were found.

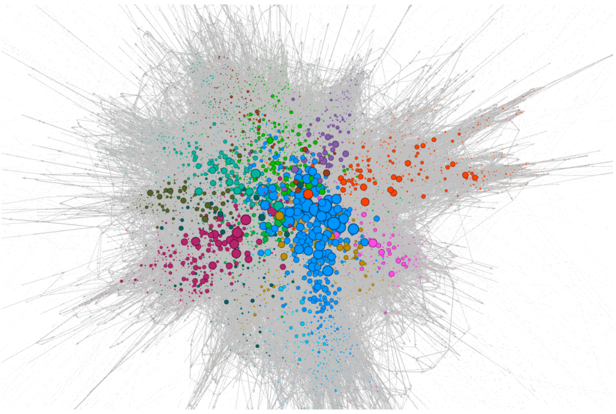

Figure 3: Main communities in Network 1

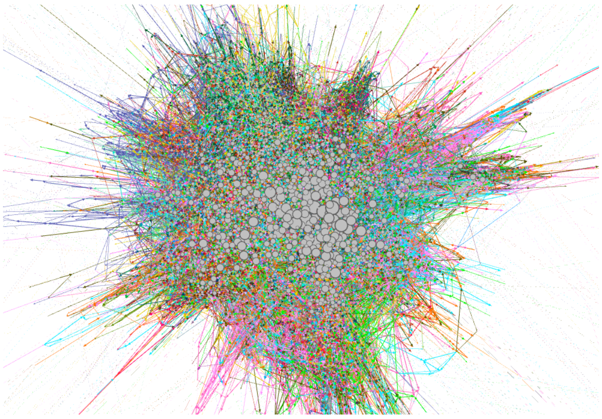

Figure 4: Temporal relationship between players in Network 1

In figure 3, communities are represented by different colors, and the size of each node reflects its degree (higher degree, bigger the node). Our first assumption was that each community regarded different time periods. However, when coloring the edges according to the timestamp in figure 4 but maintaining node position, we realized that this is not true, as the structure is very different. We can see that players in the same community play at several different timestamps. Having this in mind, we concluded the communities are probably a reflection of nationalities and the division in the chess scene in that period of time. There were several chess organizations running at the same time, and more than one world champion, before the FIDE reunification in 2006. However, the community in blue has the higher degree nodes in it, so we can assume it is the community that will likely have the higher rated players, as they tend to play with each other.

We will now use Network 3 to further analyse the network structure.

4.2 Analysing Network 3

Network 3 is an undirected weighted graph where each edge connects two players if they played at least once. The weight of the edge is the number of times they played. It has 55899 edges, which is about 86% of the original network's edges. This means that there are not many games between the same two players.

Player ID	# of opponents	Player ID	# of opponents
1512	181	1397	163
4037	181	1286	160
4850	177	5187	158
1098	166	391	155
6270	165	6586	154

Table 3: Top 10 players according to number of opponents (node degree)

In table 3 we see 3 new players that were not present in table 2: players 6270, 5187 and 6586. It is hard to give meaning to this for now.

Player #1 ID	Player #2 ID	# of games
158	8287	15
158	7715	14
699	2581	13
64	4194	12
4194	7953	12
1594	8267	11
64	7953	10
391	**4171**	10
7715	8287	10
64	7848	10

Table 4: Top 10 most popular match ups

In table 4 we can see the top 10 most popular matchups. In bold we have players that were in the top 10 players with more games played. It is curious to see lots of new players show up here. Our assumption so far was that higher rated players will have more games. We were also assuming that similarly rated players tend to play more often together. From this we derive two hypothesis: either the new players that showed up here are low rated players who play together, or maybe they are the high rated players we are looking for, and our initial assumption was wrong. A player that shows up a lot in this table is player 64. Player 64 has a very strong game history with 3 other players. This may be indicative that player 64 is a very prominent player. There is also a strong triangle of games between players 64, 4194 and 7953. More information can help us understand this relation better. We will now use PageRank to figure out who are the most important players in this network.

4.2.1 *PageRank Analysis*

We calculated PageRank in Network 3 with $\beta = 0.85$ and $\epsilon = 0.001$. We did two runs of the PageRank algorithm, one considering the weights in the network (number of times two players played), and one without considering the weights, to get a purely structural view on the network.

In table 5 we can see that the nodes with higher PageRank values are also present in tables 2 and 3. In bold are players that were not in the previous tables. This is expected as PageRank values are associated with higher degree, but for now this strengthens our belief that these high degree and PageRank nodes correspond to the highest rated players. So far, we have been only analysing players according to the number of games they played and the number of different opponents they faced. However, this does not tell us anything concrete about their performance. We also need to take into account how well these players did in all of these games. To achieve this, we will now switch to Network 2 to get a better understanding of the players' performance.

PageRank (weighted)	PageRank (unweighted)
4850	1512
4037	6270
1512	4037
1594	4850
1286	**4732**
1098	5187
391	6586
6270	1098
2599	1397
4171	**2599**

Table 5: Top 10 players ranked by PageRank in Network 3

4.3 Win Rates

Using Network 1, we calculated every player's win rate. The win rate is calculated summing all the game scores of each player and dividing by the total number of games they played:

$$WinRate(v) = \frac{\sum Points(v)}{degree(v)} \tag{5}$$

This calculation presents a problem for players with low degree. Some players only played a few games in the entire time span of the dataset, which results in inflated win rates. To combat this, we added a minimum degree as threshold for players to appear in this ranking. We tried out several minimum degrees to see which degree served as good threshold (not allowing many low degree nodes, but not including only high degree ones) and we settled for a minimum degree of 100 (which averages to one game per month for each player). Since we are looking for the most high rated players in this time span, we want to look for players who were somewhat active during these years.

Player ID	Win Rate	# of games
2479	66.97%	165
6946	66.159%	164
6212	65.748%	127
6586	65.707%	191
8121	65.584%	154
5942	65.445%	191
3271	64.851%	101
3344	64.75%	200
5050	64.583%	216
2599	64.423%	208

Table 6: Top 10 players ranked by Win Rate with minimum degree of 100

In tabel 6 there are a lot of new players showing up. They all have similar win rates, which we suspect is due to the high number of draws in high competition chess. If a player draws every game he plays, his win rate will be 50%, so we assume that draws have a heavy role in this calculation. However, this does not invalidate the fact that these players also have a high number of victories. In bold we have two players that had showed up before. Player 6586 was on the top ten players according to number of opponents in table 3. Both players 6586 and 2599 were featured in the top 10 players according to PageRank (in Network 3), in table 5. From this we can already see that these players are surely influential players in the chess world, as they have high win rates, they play consistently and they are ranked as important nodes by PageRank. However, in chess, it is more important who you defeat than how many times you win. True skill is shown by defeating high ELO players. Having this in mind, we created an heuristic for ELO calculations.

4.4 ELO ratings

Initially, we set all player's ELO rating as 2500, which is the minimum rating for a player to become a GM. Then, for each game that was played, we update both player's rating according to the game output. This is done by first calculating what is the expected score of each player, and then comparing that to the actual score. This is used by FIDE to prevent low rated players to lose a lot of rating by losing against high rated players and to prevent high rated players to quickly gain rating by beating low rated players. This means that the expected score of a high rated player, when playing against a low rated player, will be high. If he wins, he does not gain much rating, but if he loses he is penalized and his low rated opponent gets some much deserved ELO. The way we calculate this is based on a simplification of FIDE's calculations, based on a Wikipedia article [8].

The expected scores of player A and player B are calculated as follows:

$$\begin{aligned} E_A &= \frac{1}{1+10^{(R_B-R_A)/400}} \\ E_B &= \frac{1}{1+10^{(R_A-R_B)/400}} \end{aligned} \tag{6}$$

R_A and R_B are the current ELO ratings of player A and B. After the game, their ratings are updated as follows:

$$\begin{aligned} R'_A &= R_A + K(S_A - E_A) \\ R'_B &= R_B + K(S_B - E_B) \end{aligned} \tag{7}$$

S_A and S_B are the scores of players A and B (1 for the winner and 0 for the loser, or 0.5 to both in case of a draw). K is an attenuation factor to control big changes in ELO. For GMs, it is usually set as 10, and that was the value we used. We collected the top 10 players according to their ELO rating at the start of each year represented in the dataset. In order to give the heuristic some time to converge to some more meaningful values, we ignored earlier years, and we present the results in January for years 2004, 2005 and 2006, and also for April, 2006 (the end of the data set).

We will now analyse these results, present in table 7. At this point we ask the reader to bare with us and to trust in our ability to analyse all this poorly presented information. There is a lot of information we can get from this table. The first we notice is the intersection between this table and the win rate table 6. Only players 6212 and 3271 from the top 10 win rate are not present in the top 10 ELO rating for any year. The rest of the top 10 win rate players were in the top 10 ELO rating at some point. The other thing we notice is the number of games played by these high ELO players. Our initial assumption was that players with many games were probably high rated players. This assumption has proven to be wrong. Most of the high rated players were not present in tables 2,3 and 5 (most games, most opponents and PageRank on Network 3, respectively). The intersection between those tables and this are players:

- **64**, who is a very active player (top 5 in number of games);
- **1286**, who is in the top 4 in number of games, top 7 in number of opponents, is represented as an important player by PageRank on Network 3;
- **2599**, who is represented as an important player by PageRank on Network 3 and is on the top 10 highest win rate;
- **6586**, who is on the top 10 in number of opponents, is rated as important by PageRank on Network 3 and is top 5 in win rate.

Before we take any conclusions from this information, we will validate this heuristic by using PageRank on Network 2, and hopefully get some similar results in terms of player's importance.

4.5 Analysing Network 2

In Network 2, direction of edges translates to points a player scored over another (from the loser to the winner). This is helpful for PageRank, because a node's importance is influenced by incoming edges and their weights. With Network 2, we hoped that a player's importance was influenced not only by how many games they played, like in Network 3, but also by their performance. We hoped this structure could form a chain of directed edges from worst players to high rated players, assuming that they win over weaker opponents. Network 2 has 82026 edges. The maximum weight in this network is 8.5, corresponding to games between 8287 and 158 (where 158 won 8.5 points over 8287). These players are the most popular match up in table 4.

We calculated PageRank in Network 2 with $\beta = 0.85$ and $\epsilon = 0.001$. We did two runs of the PageRank algorithm, one considering the weights in the network, and one without considering the weights, to get a purely structural view on the network.

In black text, we have players who appeared previously either in the top 10 with more games, top 10 with more opponents, or top PageRank for Network 3. In blue text, we have players who appeared either in the top 10 win rate, or in our top ELO players heuristic. In red, we have players who appeared in both these contexts.

Even though this analysis is still heavily influenced by the number of games a player has, it is clear that incorporating scores in the network helped highlighting some players by their performance. Player **64** is considered the most relevant player, which we believe is a junction of both his high number of games played, and his good performance in early years as was shown by our ELO heuristic. The top ELO players at the end of the time span of the dataset (**7848** and **3344**) also show up in this PageRank analysis, along with other players who also made an appearance in the top ELO

2004			**2005**		
Player ID	Elo Rating	# of games	**Player ID**	Elo Rating	# of games
3344	2643	134	**3344**	2670	162
8121	2636	139	**8121**	2658	152
64	2634	175	**158**	2647	122
1286	2629	199	**6946**	2638	121
158	2629	111	**4082**	2635	121
2599	2627	128	**6586**	2632	146
4082	2623	98	**8287**	2630	156
5050	2617	161	**1286**	2630	221
1782	2616	120	**5050**	2630	187
5942	2616	112	**5942**	2628	143
2006			**04/2006**		
Player ID	Elo Rating	# of games	**Player ID**	Elo Rating	# of games
3344	2684	190	**7848**	2682	185
7848	2671	178	**3344**	2676	200
8121	2659	154	**2479**	2659	165
158	2656	142	**8121**	2659	154
2479	2653	159	**158**	2655	148
5942	2651	185	**2295**	2649	162
2295	2645	157	**5942**	2648	191
64	2644	246	**2581**	2647	158
1286	2644	257	**8287**	2645	192
2283	2638	204	**2283**	2643	211

Table 7: Top 10 players ranked by ELO rating

PageRank (weighted)	PageRank (unweighted)
64	4037
1594	4850
1286	1397
391	1512
4037	391
3344	1098
7848	1286
1098	5050
4850	2599
7953	1594

Table 8: Top 10 players ranked by PageRank in Network 2

ratings, from which we conclude that our ELO heuristic is a good heuristic and it is suitable to evaluate a player's importance in the chess competitive world. We are now ready to try to take some conclusions about properties of the chess world and to identify the most relevant players from this network.

4.6 Player behaviour

Throughout this paper, we made some assumptions to try to interpret the results that we were obtaining. The main assumptions we made were:

- Higher rated players are the ones who play more games
- Similarly rated players tend to play together
- High rated players have high win rates

Regarding the first assumption, we conclude that it is not entirely true. If we look at our ELO ratings, we see that all the players in there have a significant number of games played (the minimum is 98 and the maximum is 257). However, only 4 of these players were also present in our initial analysis, like we mentioned before in section 4.4. From this, we conclude that high rated players tend to play many games, but are not the ones who play more games. This can be due to them being more selective of tournaments they participate in, so that they can conserve their high ratings. Regarding the second assumption, we can not conclude with certain if it is true or not. In table 4, we do have some match ups between high rated players, like players 158 and 8287, and players 64 and 7848 (both very strong at some point in the data set). However, the remaining match ups are all either between players who are in the top 10 games played but with low ratings, or between a high rated player and a low rated player (like player 64 and 7953, who is low rated). Since this information is mainly dominated by players who have the higher number of games, we don't feel certainty to claim our initial assumption was right. Regarding the third assumption, we conclude that it is true, as almost all players in the top 10 win ratio were classified as high ELO players by our heuristic.

4.7 Who is who?

We will now play a game of guessing a player's identity based on information extracted from the network. Our main goal here is to identify Garry Kasparov, the chess mastermind of the era this data set registered, and one of the greatest chess players ever to exist. To do this, we fetched some real world information first.

2004	2005	2006	04/2006
Kasparov	Kasparov	Kasparov	Topalov
Kramnik	Anand	Topalov	Anand
Anand	Topalov	Anand	Aronian
Svidler	Kramnik	Svidler	Svidler
Shirov	Leko	Aronian	Leko

Table 9: Top 5 players ranked by FIDE

By matching table 9 (taken from [2]) with table 7, we can already guess one player. That player is Viswanathan "Vishy" Anand, FIDE World Chess Champion from 2000 to 2002 and undisputed World Champion from 2007 to 2014, losing

Player	# games	Birth Year
Korchnoi	4641	1931
Farago	3890	1946
Ivanchuk	3883	1969
Karpov	3777	1951
Anand	3732	1969

Table 10: Top 5 players with more games (to 2020)

the title to the chess Mozart of our current days, Magnus Carlsen. Anand corresponds to player ID **3344**. Anand was in the top 3 highest FIDE rated player in 2004, 2005, 2006 and April 2006. Player 3344 is the highest rated player in 2004, 2005 and 2006, and second highest rated in April of 2006 in our ELO ratings. What made us believe this player was Anand was his consistency in maintaining itself with the highest rating, along with player 8121. Also, player 3344 has many games played (200) when compared with the remaining players, and Anand is featured in the top 5 players with more games by FIDE in table 10 (taken from [1]).

Figure 5: Player **3344** - "Vishy" Anand

However, the reader may notice that player 3344 surpassed player 8121 in every year, and that in the FIDE ratings, Kasparov is number one, always ahead of Anand. So why didn't we chose 3344 as Kasparov instead? The answer lies in the number of games they played (the degree of the player's vertex). Kasparov retired from competitions around 2005. Player 8121 was second in ELO rating in 2004 and 2005. In 2006, his rating only changed by a point, and his number of games only increased by two (meaning that he only had 2 games in 2005 and he probably won them). From the beginning of 2006 to April, player 8121 did not play any games. Anand, however, still plays today, and player 3344 kept playing after 2005. From this, we concluded that player **8121** is **Garry Kasparov**.

Figure 6: Player **8121** - Garry Kasparov

Regarding player **7848**, it is harder to identify him, as he makes a sudden appearance in the top ELO players in 2006. In any case, we identified him as **Topalov**, because of the way Topalov climbed the top 3 FIDE positions, much like player **7848**

There are other two players who made regular appearances throughout the paper and that we believe we can identify. These players are players 64 and 1286.

We believe player **1286** is **Vladimir Kramnik**. Player 1286 is considered a relevant player in the network across this paper, and makes an appearance in our top ELO rating table. In 2004, he appears in 4th, and from there he enters a slow descent, losing position to other higher rated players, much like Kramnik did in table 9.

Player 64 is the 5th player with more games in this network. Plays often with some players, including player 7848 which we identified as Topalov. Is considered the most important player in Network 2 by PageRank, and appears as one of the players with higher ELO in 2004 and 2006. We concluded player 64 is **Vassily Ivanchuk**.

Figure 7: Player **64** - Vassily Ivanchuk

Ivanchuck is known to be a chess addict, and a player with erratic play, who has skill but ends up losing many games. Ivanchuck is featured as the 3rd player with more games played in table 10. This is consistent with player 64 in this network, as he plays many games, some against high rated players, had a good ELO rating, but never managed to keep himself in the top.

Of course, these interpretations will most likely not match reality, but we saw this as a fun game to play to add some color to this article. We could try to confirm this by matching edges with real life games but we feel that is beyond our desired scope of the article.

4.8 Motif Analysis

For the motif analysis, we used gtrieScanner [5], a tool that creates a trie in a prefix tree manner for subgraphs, and that matches subgraphs to their isomorphic class. Then, it generates random networks with same degree distribution and calculates the significance of each subgraph found, when compared to the random networks. We analysed subgraphs of size 4.

Analysing the most common subgraphs may give us some insight on the most popular tournament types described in section 2.6. In competitive chess, most of the games played are in tournaments. It is very rare for players to have single matches. We will now try to see what are the most popular

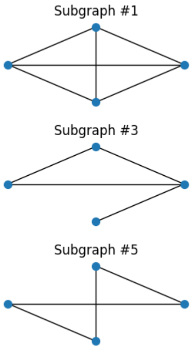

Figure 8: Labels for subgraphs (size 4)

Subgraph ID	Frequency
Subgraph #2	102392079
Subgraph #4	47977971
Subgraph #5	17074671
Subgraph #6	1473001
Subgraph #3	1331150
Subgraph #1	151024

Table 11: Size 4 subgraphs and their frequencies in Network 3

tournament types. The first thing we can see is that **round-robin** tournaments are the least popular tournaments. The subgraph that would fit a round-robint tournament is subgraph #1 (in round-robin tournaments, every player plays every other player in the tournament), because it is the only one where all 4 players play with the other 3 players. The most popular subgraph is subgraph # 2. This could either correspond to a knockout tournament (each player with degree 2 eliminates one opponent of degree 1, and then they play between themselves, or it is a succession of players who won one game and got eliminated in the next, if we look at it in a "path" point of view), or a swiss tournament (players with degree 2 play one game between them, and then each of them plays another game with another player who has a similar tournament rating). The second most popular subgraph is subgraph # 4. We would argue this is clearly a representation of a knockout tournament, as the player with degree 3 plays a single game with 3 other players, eliminating them. Since none of the other 3 players have games between themselves, we believe that it is not a representation of a swiss tournament.

From this analysis, we conclude that the most popular tournament types are **knockout** and **swiss**. The least popular is round-robin, which makes sense because it is the tournament style with longest duration of the three (if N players participate, it would require N*N(-1) games minimum, assuming they only play one game).

However, just by looking at the frequencies of the subgraphs, we cannot retrieve any information on the structure of the network. We need to compare it with other random networks, so that we can see which subgraphs are more significant.

We can see from figure 9 that the most significant subgraph is actually the less frequent one: subgraph # 1 (a clique). This means that this subgraph is over-represented in this network. To get a better sense of what kind of network this is, we decided to compare the significance profiles of this network with profiles of some other networks from homework 2. We selected the following networks:

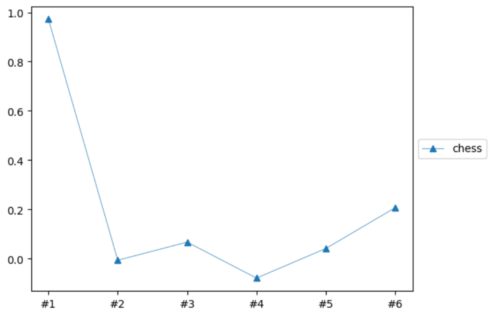

Figure 9: Significance profile for motifs (size 4) in chess network

- yeast - a transcriptional gene regulation network
- residence - a network of human friendships
- highschool - a network of human friendships
- circuit1 - a network of of digital fractional multipliers electronic circuits

Running simulations for all of these resulted in new subgraph labeling, in figure 10.

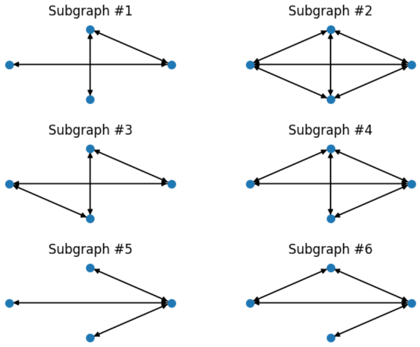

Figure 10: Labels for subgraphs (size 4)

From figures 11 and 12, we can see that our chess network has a significance profile very similar to the high school and residence networks. We were already expecting these results, as all of these three networks represent **social interactions** of some kind. Chess may be out of this world, but humans are the ones who are playing it, so it is natural that the structure of the network follows the same patterns as other social networks.

5. CONCLUSIONS

We were successful in identifying important players through PageRank, degree analysis and through a ELO heuristic we created. We also concluded that high rated players tend to play many games but are not the ones who play the most games, and that higher rated players tend to have higher

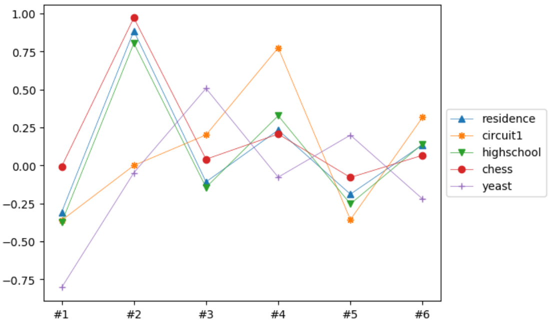

Figure 11: Significance profile for motifs (size 4) with various networks

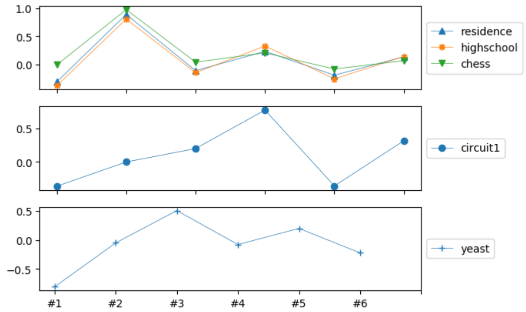

Figure 12: Classifying networks by their significance profile

win rates. Regarding tournament styles, we concluded that knockout and swiss tournaments are the most common. We also confirmed that the network presents a structure similar to other social networks. The quality of the dataset however was not the best, and it would have been better for this article if we already had the player's identities, as we could have validated our conclusions much easier.

5.1 Future Work

Edge matching could be made to try and identify important tournaments and match ups. Link prediction and future scores are also possible tasks for this dataset (originally, the data set was from a Kaggle contest where the objective was to predict game outcomes). I could have also used HITS to rank nodes, but after a few trials I realised it was not worthy to focus on that. I could have created a snapshot network of each year and try to identify different trends and important players taking only that year into account.

6. REFERENCES

[1] 365Chess. Chess players by number of games. `https://www.365chess.com/top-chess-players-games.php`, 2020.

[2] FIDE. Fide statistics. `https://ratings.fide.com/toplist.phtml`, 2020.

[3] U. Michieli. Complex Network Analysis of Men Single ATP Tennis Matches. Technical report.

[4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, oct 2002.

[5] P. Ribeiro and F. Silva. *G-Tries: an efficient data structure for discovering network motifs*. 2010.

[6] C. Shannon. Programming a computer for playing chess. *Philosophical Magazine*, (41):314, 1950.

[7] Unknown. Kaggle chess dataset. `https://www.kaggle.com/c/chess/data`, 2020.

[8] Wikipedia. Elo rating system. `https://en.wikipedia.org/wiki/Elo_rating_system`, 2020.