SDSC3016

SOCIAL NETWORK ANALYSIS

GROUP PROJECT - 2

Social Network Analysis of Tennis Network

INSTRUCTOR - Dr. QING Ke

MALHOTRA Akul (56698828)

BUDIMAN Darren Nathanael (56672641)

City University of Hong Kong

Semester B 2022/2023

Table Of Contents

1) INTRODUCTION

Tennis has been one of the most popular sports over the past century and is famous across different regions and continents like Europe, Asia and America. There are a number of tennis tournaments conducted every year which are followed by millions across the world. In the United States alone, in a single year, tennis participation grew by over 1 million with 23.6 million people playing the sport *(Tennis.com, n.d.)* Apart from being an extremely popular and competitive sport, tennis also offers many health advantages for those who play it regularly. Tennis players usually live an average of 9 years more than sedentary individuals and tennis reduces the risk of heart disease significantly. Because of its rich history, competitive tournaments and health benefits, tennis has become extremely popular and its popularity is growing each year. (*Spring et al, 2020*)

Alongside tennis, one thing which has also grown and developed significantly over the past few years is the use of technology for sport related statistics and analysis. All major sports today are heavily reliant on statistical analysis and in depth analysis of past data related to it's players. Sports across the globe are utilizing modern technologies to analyze as well as predict future matches and games. Machine Learning and AI tools have become mainstream.

While many different sports have brought in statistical and machine learning based tools for analysis, the number of studies relating to the application of Social Network Analysis tools to sports has been comparatively limited. Social network analysis involves an in depth study of relations of entities in a social network and also analyzes the social structures that can be formed from them. The use of Network Analysis can be extremely useful in a sport setting, wherein it can be used to examine and analyze a number of properties relating to sportsmen and the tournaments they participate in *(Social Network Analysis - an Overview | ScienceDirect Topics, n.d.).*

The goal of this specific project is quite similar, the project focuses on conducting a social network analysis of ATP Tour matches played by different players. It shall firstly analyze

the tournaments and the players who participated in them. The project shall focus on analyzing the patterns behind how tournaments have invited players and also analyze which tournaments the players have been participating in. The project shall be utilizing specifically social network analysis tools for the analysis.

A link prediction shall also be conducted in order to forecast which tournaments will host which players in the future . Further Machine Learning models can also be implemented as an extension to this project which focus on predicting the winners of the tournaments on the basis of the participating players, which are predicted by this study.

2) METHODOLOGY

The team followed a number of steps for analysis of the tennis dataset starting from data collection, data preprocessing, calculation of metrics, community detection, further analysis and link prediction. A number of different tools were used for the project including Python libraries like Pandas for data preprocessing and preparation and networkx library for analysis. Gephi was used for visualizing the social network as well as calculating various metrics related to the network. Algorithms to generate the communities were written in Python, after which Gephi was used to visualize them. Python libraries including networkx and scikitlearn were used for link prediction.

2.1) Dataset

The team created a network dataset on its own by transforming a normal dataset through a number of preparation and preprocessing steps. The dataset, namely ‘Association of Tennis Professional Results from 2000 to 2017’ was originally taken from Kaggle, and it contained 49 columns and 3364 rows. Many of the columns were not useful for our project and thus we only used the following columns:

- 'Tourney_name' : Tournament name
- 'Winner_name' : Winner’s name
- 'Winner_ioc' : Winner’s country
- 'Loser_name' : Loser’s name
- 'Loser_ioc' : Loser’s country

2.2) Data preparation & Preprocessing

The dataset preparation and preprocessing was done using Python libraries like Pandas. The dataset in the raw format could not be used to make a network, thus the dataset had to go through a number of different steps. The code for all the preprocessing steps has been attached as well.

Table 1.

Tournament to Player Data

	tourney_name	players
0	's-Hertogenbosch	Magnus Gustafsson, Jan Boruszewski, Jan Siemer...
1	Adelaide	Thomas Enqvist, Arnaud Clement, Roger Federer,...
2	Amsterdam	Christian Ruud, Nicolas Lapentti, Federico Bro...
3	Atlanta	Andre Agassi, Xavier Malisse, Jiri Vanek, Just...
4	Auckland	Tommy Haas, Jeff Tarango, Juan Balcells, Franc...

The goal of creating the network was to visualize a network which connects the tournaments to players. The reason behind it was to analyze the patterns in which tournaments are inviting which players and also to see which players are playing in which specific tournaments. This can help us understand underlying structure and patterns behind players and tournaments they play in.

The dataset after the initial preprocessing was put in the Gephi visualization tool to create the visualizations. The details of the visualizations and their analysis has been covered in the Data Analysis section.

2.3) Metrics Calculations

After preprocessing the data and also visualizing our result with Gephi, the team went on to calculate different metrics based on the networks.

The following metrics were calculated by the team:

a) Avg. Degree & Highest Degree - To analyze which nodes have the highest degrees and what is the average degree of the network. This helps in understanding the average number of tournaments a player participates in and also the average number of participants that are invited to a tournament.
b) Degree Distribution - To analyze if the network follows any degree distribution such as log-normal or power law. This helps in understanding the structure of the network and if the data follows any probability distribution.
c) Average Path Length - To analyze the average of the shortest paths in the network and if the network is a small world network or not. It describes the average number of steps going from one node to another in the network.
d) Graph density - To analyze if the number of links in the network is equal to the maximum number of nodes in the network. It describes if the network is strongly interconnected or sparse.
e) Centrality measures including Betweenness, Closeness & Eigenvector- All these measures can be used to find out the most influential nodes of the network and which nodes add most value and influence to the network. The different measures calculate centrality in different ways and allow for analysis of the network in different ways.

The details of metrics results and their analysis has been explained in the Data Analysis section.

2.4) Community Detection Algorithms

The team wished to conduct an analysis of the communities that exist within the network. Different Algorithms were utilized to detect the communities that exist in the network, the algorithms used for community detection are as follows:

a) Modularity - In order to analyze the communities detected by the modularity density. The communities were detected using the Modularity partitioning tool in Gephi.
b) Louvain Algorithm - The algorithm was again implemented using Python and the results were put to Gephi for visualization.
c) Label Propagation Method - In order to analyze the different communities as figured out by the Label Propagation Algorithm. The algorithm was implemented using Python and once the communities were detected, the results were fed to Gephi to visualize the communities.

The visualizations and analysis of our communities shall be done in the Data Analysis section.

2.5) Link Prediction

In order to extend the analysis of this project into future use, the team also made a model utilizing Python libraries like networkx and scikitlearn to conduct Link Prediction based on the network. This link prediction is really helpful as it can be used to predict and forecast the future. It can be used to forecast which tournaments shall invite which players in the future. This can help us know the likelihood of the tournaments inviting players in the future.

For these analysis, we created a dataset consists of all possibility of edges occur between two nodes. If there occur real edges between them, the column "Connection" will show binary number "1", and if there is no edges, it will show "0". The dataset looks like this:

Table 2.

Tournament to Player Data

	Source	Target	Connection
0	Rome Masters	Slava Dosedel	1
1	Rome Masters	Mounir El Aarej	0
2	Rome Masters	Tomas Behrend	0
3	Rome Masters	Fernando Meligeni	1
4	Rome Masters	Paul Goldstein	0
...	...	...	...

In this social network we conduct link prediction using three different methods: Random Forest, Rooted Pagerank, and SimRank. Other Machine Learning models can also be developed as an extension to our project which can predict the winner of the tournaments based on the players participating, which is predicted from our model.

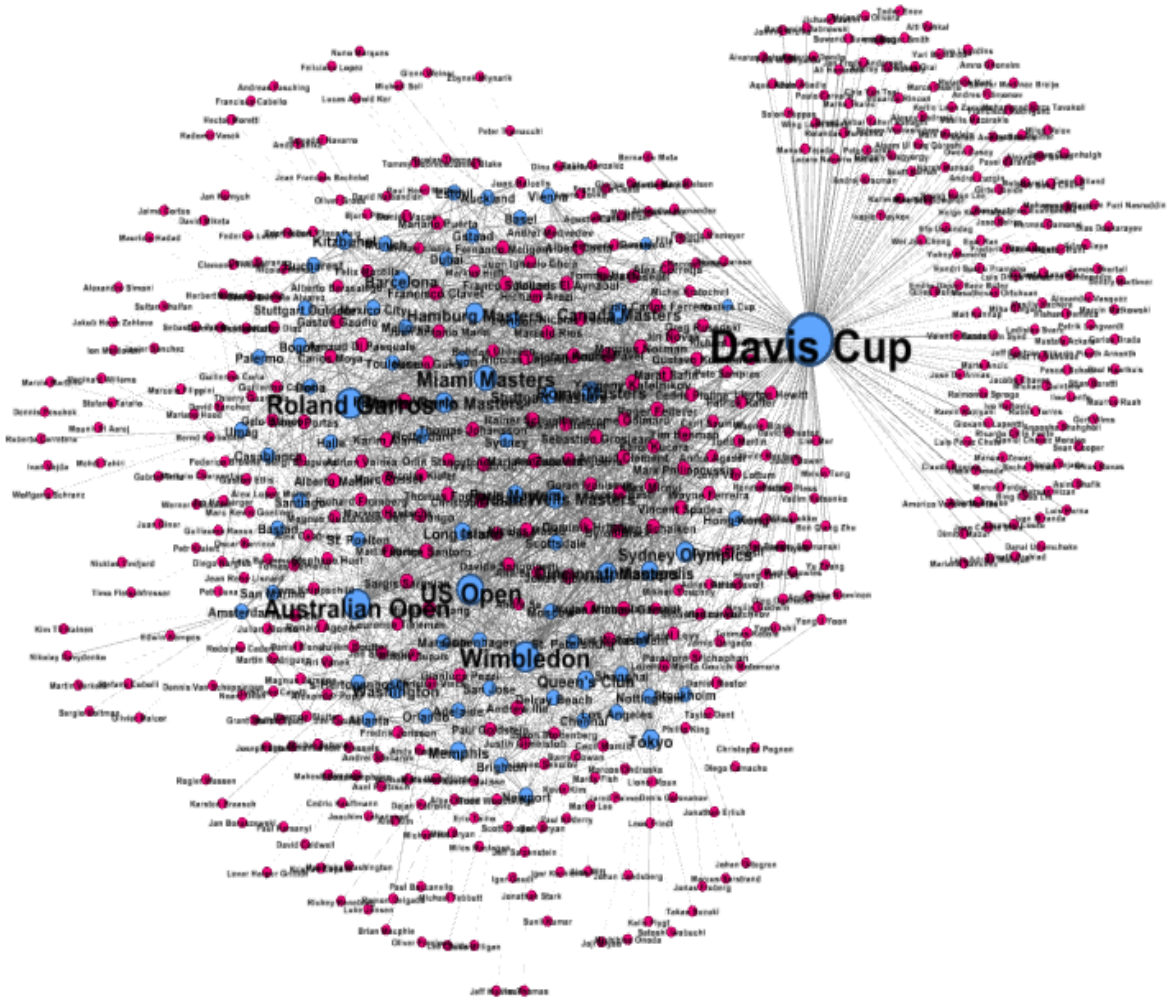
A list of all Python-and-gephi-related code can be found here:

https://github.com/darrenbudiman/SDSC3016-SNA-TENNIS-NETWORK-ANALYSIS

3) Data Analysis

Figure 1.

Tournament to Player Network.



This is an undirected network with blue nodes representing the name of the tournaments and the red nodes representing the name of the players. A link between 2 nodes representing that the player has played in the specific tournament. A larger size of the node in the network represents a higher degree for the node.

There are in total 558 nodes and 3367 edges in the network. There are in total 72 blue nodes representing 72 tournaments in this network and 486 players in the network.

3.1) Graph density analysis & Average Path Length Analysis

The graph density for the graph is very low which is 0.022, which is because it is a graph only from tournament to players and we are not linking the player to other players and thus because players are not connected to each other, the no. of links each player has is only to tournaments and thus possible no of links is very small compared to max no. of links that can exist in this network. The low graph density also suggests that tennis tournaments focus on inviting a specific group of players each time and not the entire group of all players for a single tournament. For eg. Grand Slams focus on inviting only the top players whereas some other tournaments focus on inviting only some lower amateur level players.

The average path length for this network is 3.002, indicating that average number of steps taken to go from one node to another in the network is quite low.

3.2) Degree Analysis and Degree Distribution Analysis

The Average degree for the tournaments is 46.09, indicating that on average a tournament invites close to 46 players.

The highest degrees for the tournaments are as follows:

Id	Label	Frequency	Type	Degree	Weighted Degree
Davis Cup	Davis Cup	1	tourney_name	266	650.0
Roland Garros	Roland Garros	1	tourney_name	128	254.0
US Open	US Open	1	tourney_name	128	254.0
Wimbledon	Wimbledon	1	tourney_name	128	254.0
Australian Open	Australian	1	tourney_name	128	254.0

	Open				
Miami Masters	Miami Masters	1	tourney_name	96	190.0

As it is visible Davis Cup has an extremely high degree compared to all other tournaments that are conducted.

Some interesting Observations based on metrics based on previous analysis:

Q) What do the average degree of tournaments and graph density tell us?

A) The graph density is low for the network which indicates that the number of links in the network is much less than the maximum number of links that can exist in the network. This is because it is a tournament to player network and we do not connect players to other players. This also tells us that tournaments focus on inviting a specific group of players based on their budget and location and do not invite a lot of players but 46 on average, thus the degree is much less than the maximum value, the number (avg. tournament degree) is inflated because there are some tournaments like Davis Cup which are team tournaments and thus invite many players, however other tournaments invite much less number of players based on various factors including their budget, location and other variables. The presence of tournaments like the US Open and Davis Cup inflate the average degree for all tournaments. The average degree for participants is not high which is close to 7, which indicates that a player usually plays in only a limited number of tournaments and thus the degree is much lower than the max degree. This also contributes to the low density of the network.

Q) Why does the Davis Cup have such a high degree compared to other tournaments?

A) As we see the tournament ‘Davis Cup’ has the highest degree out of all tournaments which refers to how it has invited the most number of players. This is also because Davis Cup is an amateur tournament which is held usually for countries against other countries and thus the number of players invited are very high leading to a high degree for it. The other tournaments are usually individual player events and not team events, and thus

invite less number of players. Thus the degree of Davis Cup is much more than the average degree for all the other tournaments.

Q) Why is the average path length so low for the network?

A) The presence of nodes which have extremely high degrees such as Davis Cup contributes to the small average path length for the network. These nodes can act as intermediate nodes between many nodes and thus the shortest path tends to be low. In addition to this there are players who play in the same tournaments often, as a regular group of players get invited to similar tournaments. For eg. usually all top players get invited to Grand Slams, and because of this again the shortest path between tem tends to be very low in this network.

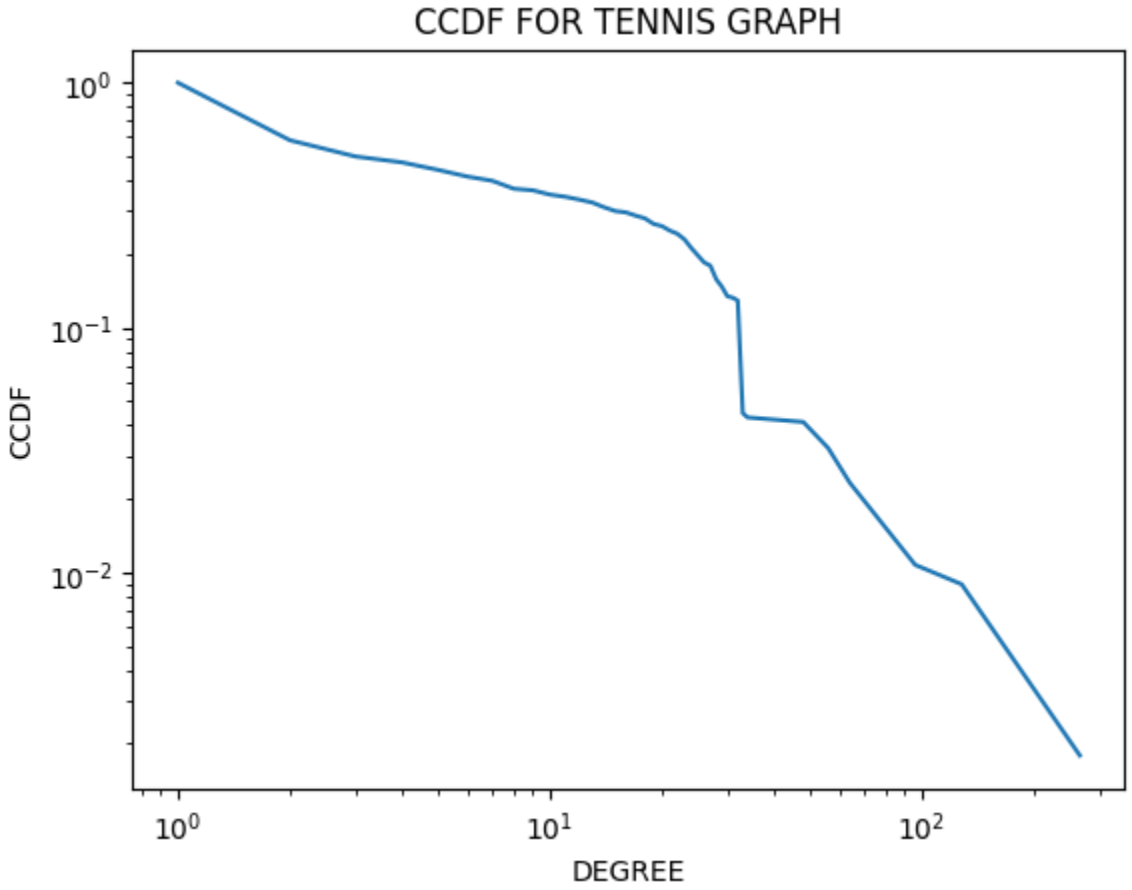
Now analyzing the average degree for players, the average degree for players is 6.9 which indicates that on average a player has participated in 6.9 tournaments. This indicated that a player usually participates in a limited number of tournaments during his/her career which is 6.9.

The highest degrees for the players are as follows:

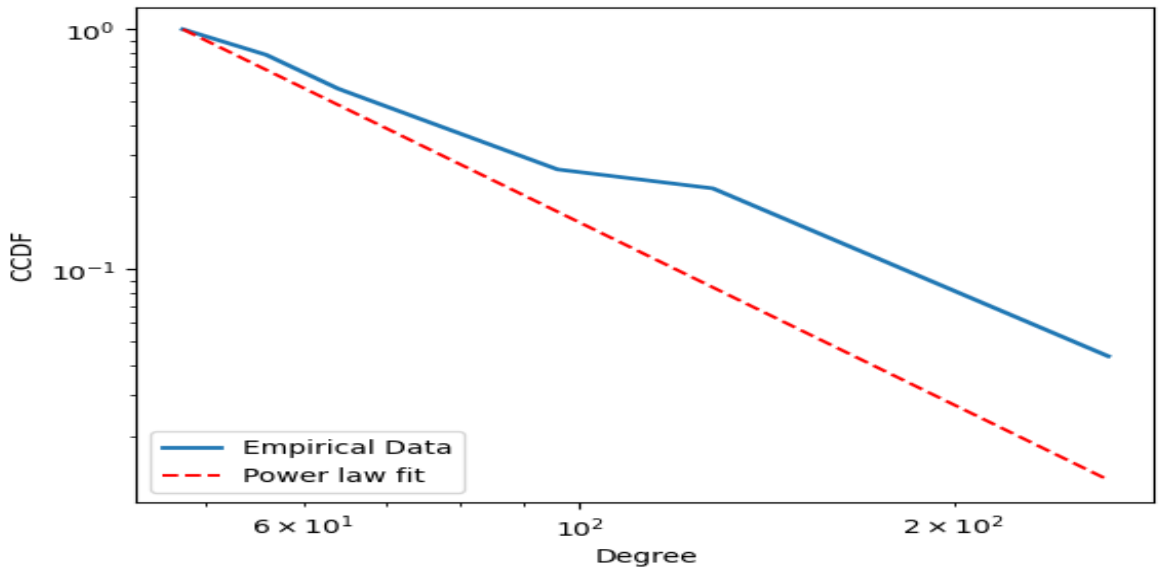
Name	**Label**	**Frequency**	**Type**	**Degree**
Yevgeny Kafelnikov	Yevgeny Kafelnikov	102	players	34
Fernando Vicente	Fernando Vicente	56	players	33
Marat Safin	Marat Safin	93	players	31
Francisco Clavet	Francisco Clavet	59	players	31

Thus on the basis of the highest degree we can say that, Yevgeny Kalfnikov has participated in the most number of tournaments out of everyone. The Russian Tennis player has participated in 34 tennis tournaments. This is much more than the average degree for the players.

The CCDF for our network is as follows:



The visualizations for the degree distribution that were tried to fit in are as follows:



The team tried a variety of different distributions including Power Law, Truncated Power Law, Exponential Function, however non of the distributions gave a good fit to the data. The p value that was calculated was always more than 0.05, thus it was never indicative of which distribution fits the data the best. Log Normal distribution could not be calculated as according to Networkx the optimal fit values were really extreme and it lacked the numerical precision to calculate it.

3.3) Centrality Measures

A. Closeness Centrality

The top nodes while taking into account closeness centrality are as follows:

Tournament	Closeness Centrality
Davis Cup	0.5215355805243446
Christophe Rochus	0.4434713375796178
Yevgeny Kafelnikov	0.44276629570747217
Rainer Schuettler	0.44276629570747217
Francisco Clavet	0.44206349206349205
Ivan Ljubicic	0.4392744479495268

As we see Davis Cup again is turning out to be an important node. This is because of its extremely high degree because of which it is very close to a large number of nodes and shortest path between it and many other nodes is very low.

It is extremely interesting that on number 2 is a player, Christophe Rochus, this is extremely important because even though the degree of Christophe Roches is not so high, i.e. 23. This indicates that maybe the nodes Christophe is connected to have high degrees because of which the shortest paths from it to other nodes is not so high.

B. Betweenness Centrality

The top nodes for betweenness centrality are:

Name	Betweenness Centrality
Jared Palmer	0.9815169510061637
Guillaume Raoux	0.9687283358757203
Oliver Gross	0.6784404642890304
Tommy Robredo	0.5589040829296619
Nicolas Thomann	0.47089795274415175

It is extremely interesting to see that while analyzing betweenness centrality, the top 5 nodes are all player's and not the tournaments. This is indicating many shortest paths pass through them and they add value to flow of information through the network, it is also because of their high degree they often come in shortest paths for many pairs of nodes and this also indicates how they might have participated in a lot of tournaments.

C. Eigenvector Centrality

The top 6 nodes with highest eigenvector centrality are as follows:

Name	Eigenvector Centrality

Roland Garros	1.0
US Open	0.9802298327495086
Wimbledon	0.9697802110371271
Davis Cup	0.9626685526230178
Australian Open	0.9268082794047706
Miami Masters	0.8330054309081909
Hamburg Masters	0.6233046326022827

It is interesting to see that all the top 6 names in the highest eigenvector centralities are tournaments with Roland Gross having exactly 1 and US Open and Wimbledon having almost close to 1. The indicates that they are connected to a number of important nodes in the network, this can also be explained because Rolands Garros (French Open), US Open and Wimbledon are one of the most major tournaments across the globe and thus they host one of the most important players in there events and thus are connected to important nodes.

Other Interesting Observations & Questions Based on Centrality Measures:

Q) Why do extremely popular and top tennis tournaments like Wimbledon and US Open have extremely high eigenvector centrality but not such high betweenness centrality?

A) The way the two centrality measures are calculated in a very different way from each other. Eigenvector centrality takes into account the quality of connections a node has in the network while betweenness centrality analyzes the number of shortest paths passing through the node.

The reason why important tournaments like the US Open and Wimbledon have high eigenvector centrality is because they are connected to top players, which are influential on the network, thus leading to high eigenvector centrality. However, since these nodes are only connected to top players they are not so influential in

linking other nodes of the network. The number of shortest paths passing through them is lower, thus leading to low betweenness centrality.

Q) While analyzing betweenness centrality why are all top 5 nodes players and not tournaments?

A) The tournaments which have high eigenvector centralities only invite top players and thus are connected to a limited segment of the nodes because of which they have low betweenness centrality, while players have higher centrality values because they have played in tournaments that connect different parts of the network together. These players act as bridges and intermediates often connecting shortest paths for different nodes across the network and thus have higher betweenness centrality.

Q) While we calculate Closeness Centrality, the top 5 nodes have one tournament (Davis Cup) which has the highest centrality, while the rest 4 are players, why is that no other tournament is having extremely high closeness centrality?

A) The major reason the Davis Cup has extremely has closeness centrality is because of its extremely high degree, which is highest in the network. Davis Cup is an amateur tournament held between different countries and thus the number of players invited are very high thus the degree is high and it has a very low shortest path to many other nodes. However when we consider other tournaments, they focus on inviting a limited group of players based on their budget, location and other factors. Thus they do not have such high degrees and thus closeness centrality can be low. Such tournaments have short oaths to the players they invite but the shortest paths can be more to other players.

Q) According to the above analysis, what are the most important nodes?

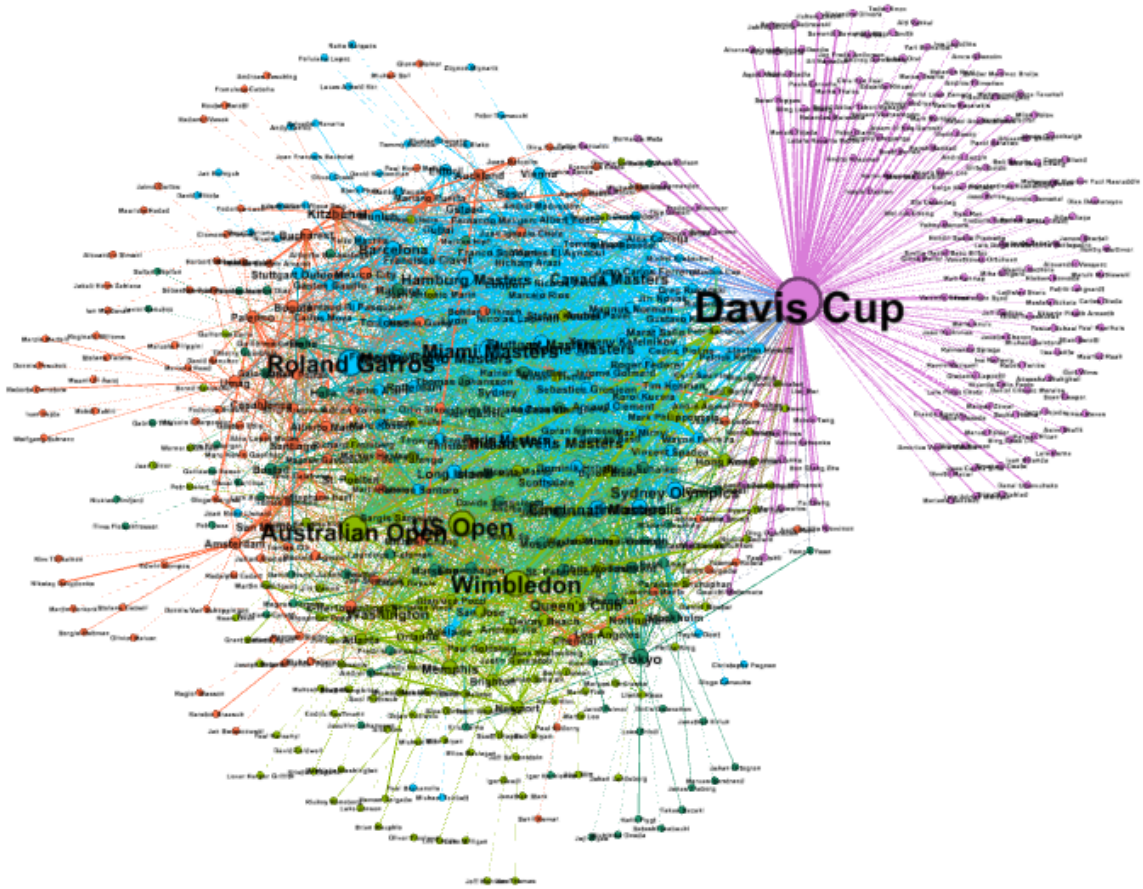
A) Based on the above analysis, “Davis Cup” is undoubtedly the most important node, Davis Cup has the highest degree, highest closeness centrality and also a very high eigenvector centrality. Thus it is connected to the most number of nodes and its nodes add influence to the network. Also, it has an extremely short distance to a number of nodes. Many other nodes including Wimbledon and US Open can also be said to be important as they have extremely high eigenvector centrality as well as a reasonably high degree. Among the players, ‘Jared Palmer’ and ‘Yevgeny Kafelnikov’ can be said to be important as well because of their high centrality values.

3.4) Community Detection

In Tennis, there are different categories of tournaments and different players participate in different kinds of tournaments based on their strength. Thus many communities should exist in the network naturally. To analyze if community detection algorithms work well the team decided to implement 3 community detection algorithms explained below.

A. Modularity Metrics

The first way we detected communities was using modularity. The graph obtained is as follows:



The table above shows the community detection using modularity. In general, modularity community detection is a measure of the way networks or graphs are divided into communities based on their structure. The modularity score of a network indicates that the nodes within modules are densely connected, but the connections between modules are sparse. In our network, there occur five communities that occur because of modularity scores. Based on our research from the data we managed to collect, the majority of the tennis competitions in the blue community are tennis masters cups. Most of the light green nodes represent Grand Slams cups. Purples show Davis Cup, dark greens include most of the WTA and ATP tour competitions, while the orange represents the ATP challengers.

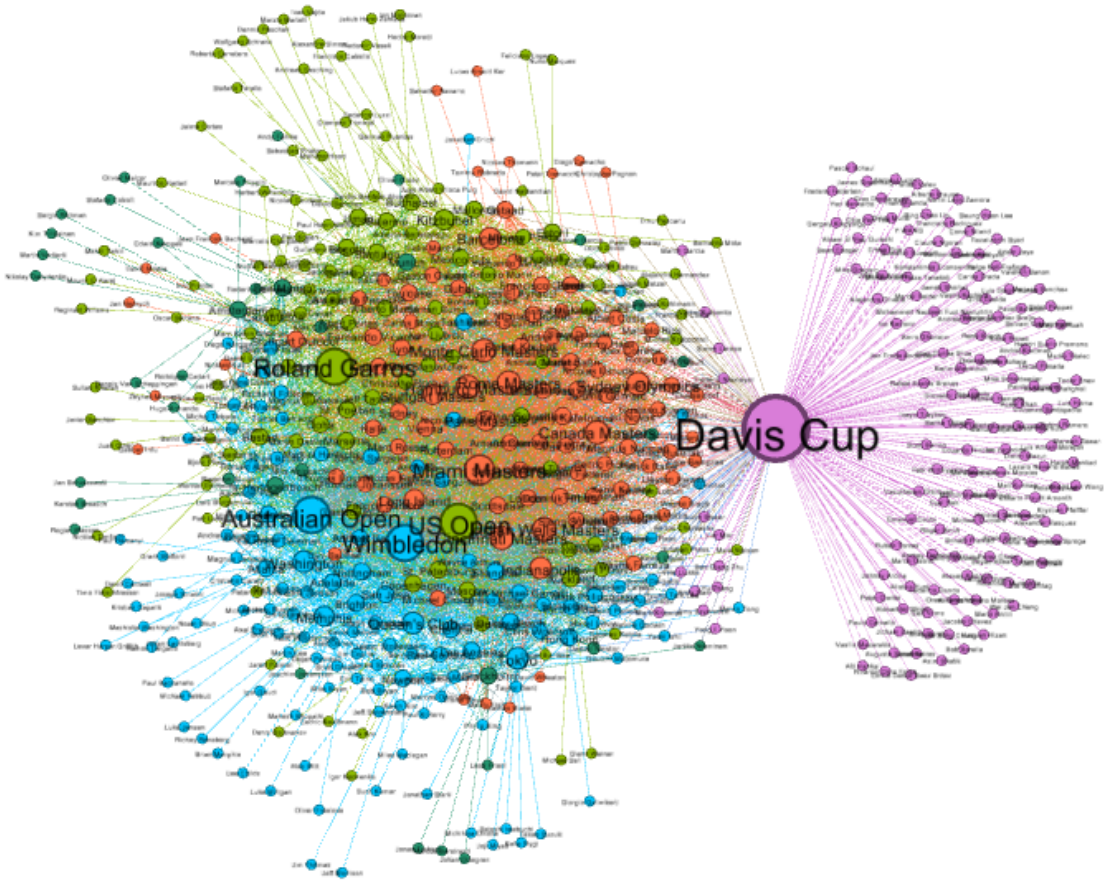
Generally different sets of players get invited to all these tournaments because of different levels of the tournaments. Davis Cup usually targets amateur players while Grand Slams for instance targets specifically top players. Thus in practice, different categories of players participate in different types of events. Fortunately, Modularity is able to detect this, and thereby creates different classes.

Players' node colors are affected the most by the competition they join most often. Even though some players join lots of competitions with different levels, their node's color is determined based on their most active participation in those competitions.

The modularity metric for community detection fits our data best as it can classify the groups of nodes based on the level of competition.

B. Louvain Algorithm

The second method utilized for community detection was the Louvain method, the graph obtained from it is as follows:



This community detection method focuses on assigning each node to the neighbor community that yields the largest modularity increase with respect to the current partitions. Over time, all nodes are revisited until they can no longer be moved to a different community to increase modularity. These procedures will keep looping until no further grouping of the clusters in the current partition increases the modularity.

In our data, there occur four big communities, colored pink, light green, blue, and orange. Actually, there occur a very small number of dark green nodes which we can ignore on further analysis.

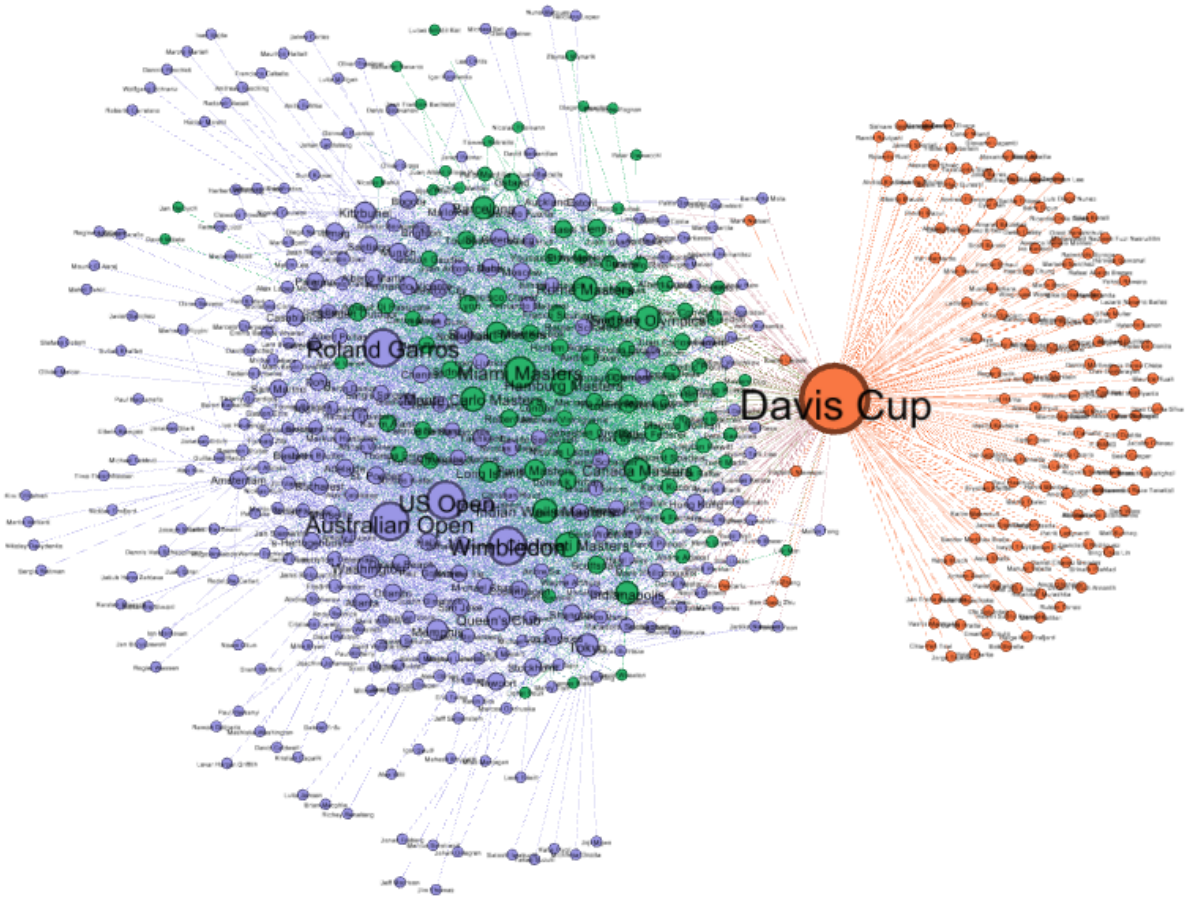
If we compared the Louvain algorithm and other information about tennis, we can assume that the pink nodes focus on players that mainly participate in the Davis Cup. Nodes in orange represent the Master's Cups that are held around the world, such as the Miami Masters, the Monte Carlo Masters, and many others. However, we cannot infer

any conclusion on the green and blue nodes, since they mixed Grand Slams, ATP, and WTP Tour randomly.

Thus, the Lovain method doesn't work as well as the Modularity method as it has mixed up players belonging to different categories of events, which actually should be in different categories.

C. Label Propagation Method

The third way used to analyze and detect community is the label propagation method.



Label propagation algorithm works on the principle that neighbors of a node usually belong to the same community. They work by assigning each node to different communities. Each node is labeled and then each of them took the label shared by the

majority of its neighbors. The loop ends when every node has the majority label of its neighbors.

From the graph above, we can see that the orange nodes cover mostly the players that participated in the Davis Cup. Besides, the green nodes cover the master cups and players correlate with them. The difference between Label propagation and the Lovain algorithm lies in the blue nodes which now cover all the Grand Slams, ATP, and WTA Tours competitions together.

This is also a good measure as many ATP tours feature top players and Grand Slams also feature top players. Thus the network has created a separate class for these players. Thus we can say that the Label Propagation method also performs well.

Other Interesting Observations & Questions Based on Community Detection

Q) What is the strongest community we have on our tennis network?

A) The Davis Cup can be said to be the strongest community we have on our network. Through three different methods of community detection, we can see that almost nothing changed in the Davis Cup community. It always consistently grouped with the participants of the competition and did not mix up with the other community groups.

This case can occur since half of the Davis Cup are amateur players which rarely participate in other professional competitions like the Grand Slams, Masters, ATP, or WTA tours. The position of the Davis Cup itself has been a local bridge that connects amateurs and professional players in our graph.

Masters Cup can also considerably be a strong community in our network. During the three community network tests we have done, the Master Cups has always

created its own groups consisting of players who actively participate in the competition. However, there still occur across Master Cups which sometimes get mixed with the other type of community.

Grand Slams, ATP, and WTA tours can be said to have a quite sparse community in our network. Only the modularity algorithm can identify the community that occurs across them pretty well. This happened because most of the players participated in one of those three events and also actively participated in the other two events since they all are prestigious events in tennis sports.

Q) What is the best community detection model for our tennis network?

A) The modularity algorithm suits our network best. It can identify the difference between Grand Slams and ATP / WTA Tour, which can not be done by the other models we used in this social network analysis.

Even though many top players participate in both Grand Slams and ATP tours, There are some top players that participate only in Grand Slams. Thus modularity does a very good job identifying these players.

Modularity community detection can fits our network well since our network is relatively well-defined and show a clear distinction between communities. This situation helps modularity community detection to identify the community structure within the network better.

The Label Propagation method also performs decently well but not as good as modularity. On the other hand, the Lovain Algorithm method performs the weakest out of the three algorithms.

In the real-world case, tennis players usually only played competitions based on their rank. High-ranked players usually participate in high-level competitions like Grand Slams, followed by Masters tournaments. The ATP Tour and WTA Tour are also highly ranked but not as prestigious as both tournaments mentioned before.

The Davis Cup is a team competition, and while it is still significant, it is generally considered to be of a lower level compared to individual tournaments like the Grand Slams and Masters.

This situation can create communities in the network, separated by the players that join the tournaments most often. Furthermore, community detection also can identify the level of each player, just based on which nodes (competitions) they participated in.

3.4) Link Prediction

In this analysis, we use three different link prediction models; Random Forest, Rooted PageRank, and SimRank. The results is shown in the table below:

Models	Random Forest	Rooted PageRank	SimRank
MSE	0.0944	0.0959	0.0960
Accuracy	0.9056	0.9041	0.9039

A) Random Forest Model

Random forest machine learning model is a combination of multiple decision tree models. This model is able to improve the accuracy and reduce the overfitting of the model, while also perform well in handling noisy data and missing values.

Since our dataset for this network is based on binary data type, we are able to do a link prediction using this classification-type model. However it performs best compared to the other method we use, with a slightly better performance than the Rooted PageRank and SimRank.

B) Rooted PageRank

By comparing the PageRank scores for each node in the network, PageRank algorithm is able to predict the probability of an edge between two nodes using the Rooted PageRank scores, before converting it to a binary prediction.

The Rooted PageRank Algorithm performs very well in this data since it shows a 90.41% accuracy, slightly better than SimRank but still not as good as the Random Forest Algorithm.

C) SimRank

SimRank link prediction method is a similarity-based algorithm that computes the structural similarity between nodes in a graph based on the similarity of their neighbors. In this case, the algorithm can be used to predict whether two nodes will form a link based on their structural similarity. The simrank similarity function from NetworkX library is used to compare the similarity scores between every pair of nodes in the network. The scores later are used to predict the likelihood of a link forming between two nodes.

Although it already performs well, giving very good results in link prediction with an accuracy of 90.39%, this method still performs slightly worse than the other two methods used in this network.

4) Results

Thus a number of important results can be gained from our Data Analysis, a summary of them is as follows:

1. The graph density is very low, which is because it is a network that only links tournaments and participants and not the players among themselves. Besides, tournaments tend to invite specific players repeatedly and not all the players leading to low density.
2. The average path length for the network is very low, which is 3, this is indicative of the fact that very few steps are required to go from one node to another. This is mainly because of certain nodes having extremely high degrees and thus acting as a bridge frequently between 2 nodes and reducing the overall avg. path length. Also tennis players participate in a number of tournaments, thus their degree is also not low and thereby avg. the shortest path is low for the network.
3. The Davis Cup has the highest degree among all tournaments, and its degree is substantially more than all other tournaments. This is mainly because it is a team tournament and thus a lot of players are invited to the tournament. Thus leading to a high degree for Davis Cup
4. Out of all the players in the network, Yevgeny Kafelnikov, has the highest degree, which is 34 which represents that he has participated in the most number of tournaments which is 34.
5. On average a tournament invites about 46 players and on average a player usually attends close to 7 tournaments
6. The CCDF was plotted for the graph and different degree distributions were tried for the graph, however none of the degree distributions were able to fit the graph well and the p value calculated was always significantly more than 0.05.
7. Various numbers of centrality measures were used to analyze the graph.
8. While calculating betweenness centrality, interestingly the player, “Jared Palmer” had the highest value indicating that many shortest paths passed through, and it acted as a bridge connecting different nodes on many occasions.

9. In Eigenvector centrality, Roland Gross, also called the French Open, and 5 other tournaments including Wimbledon and US Open, had the highest values. This is largely because these are extremely elite events and only invite top players. Thus their neighbors are a lot of top players who add valuable information to the network, thus leading to high eigenvector centrality values.
10. While calculating Closeness Centrality interestingly there were all players in the top 5 nodes having highest values of Closeness centrality.
11. The top tournaments like US Open and Wimbledon have high eigenvector centrality but relatively low betweenness centrality, this is largely because these nodes are connected to extremely top players who add influence to the network. However, these tournaments only invite top players and thus are connected to a limited number of nodes because of which they have low betweenness centrality, while players have higher centrality values because they have played in tournaments that connect different parts of the network together.
12. Davis Cup is the most important node because of its highest degree, closeness centrality as well as extremely high eigenvector centrality.
13. Different community detection algorithms were applied to our network and the modularity community detection method is the method that fits our data best. Its ability to handle a less-complex data fits our simple network of tennis players and the tournaments they participated in. Also, the Davis Cup has the strongest community compared to the other communities in our network.
14. Link prediction was also performed for the network, 3 models were implemented including Rooted Pagerank, Random Forests & SimRank. All of the models performed well giving a high accuracy.

5) Conclusion and Discussion

Over the past years, Sport Analytics has become mainstream, with many different sports heavily reliant on player statistics and Machine Learning for their analysis. However, the number of studies relating to application of Social Network Analysis has been relatively limited. The aim of the project has been to tackle the lack of the number of studies utilizing social network analysis tools, by conducting a social network analysis of tennis competitions and players participating in it. Different kinds of networks could be created from the dataset, however the team chose to create a tournament to player network so as to analyze what kinds of players or players are participating in what events and which tournaments are focusing on inviting which players. A co-existing network could also have been created from the dataset which is a player-player network indicating that 2 players have played against each other.

The team calculated several metrics to analyze the given network. Many interesting observations were made from the analysis. It was interesting to see that the network had a very short average path length indicating very few steps were required to go from one node to another. It was largely due to the presence of some nodes with extremely high degree which brought down the average path length for the network. The graph density was also very low. A degree analysis was also conducted for the network and it was important to note that none of the degree distributions were able to fit the network well. The team also went on to conduct centrality analysis utilizing many different centrality measures including Closeness, betweenness and Eigenvector. Each of the measures gave interesting results, with top tennis tournaments having the highest eigenvector centrality since they are connected to top tennis players while some tennis players having higher betweenness centrality measures. Also the closeness centrality had Davis Cup having the highest value while all other 4 nodes were players not tournaments, largely because tournaments usually invite certain segments of players and have shorter distance to such nodes whereas end up having really high distance to other players and tournaments. Davis cup was estimated to be the most important node in the network based on the metrics but other nodes including US Open, Wimbledon were also considered to be important for the network.

Different tennis players participate in different kinds of tournaments based on their strength. Thus naturally communities should exist in the network, in order to analyze if community detection algorithms could perform well in detecting the communities for the network, the team applied 3 community detection algorithms. These include Label propagation, Modularity and Lovain method. The modularity performed the best out of all algorithms and was able to detect different communities based on the strength of players. It was able to segregate top tournaments and other lower level tournaments.

In addition to this, the team wished to conduct predictive analysis so that the research can be used for forecasting as well. A link prediction was conducted as well, to predict which tournaments will invite which players in the future. 3 algorithms were applied for this including Sim Rank, Rooted PageRank and Random Forests. All algorithms performed very well giving an accuracy of about 90%.

It is hoped that our predictive analysis can also be used by other researchers across, who can conduct further predictive analysis focusing on the winners of tournaments based on its participants forecasted by this project. More importantly, the team hopes to influence many other researchers to apply network analysis to sports and thereby create a stimulus and a ripple effect to have many more similar studies.

REFERENCES

SPRING, K. E., HOLMES, M. E., & SMITH, J. W. (2020). Long-term Tennis Participation and Health Outcomes: An Investigation of "Lifetime" Activities. *International Journal of Exercise Science*, *13*(7), 1251–1261. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7523898/

Tennis.com. (n.d.). *USTA reports surge in tennis participation led by growth in ethnic diversity*. Tennis.com. Retrieved April 5, 2023, from https://www.tennis.com/baseline/articles/usta-reports-surge-in-tennis-participation-led-by-growth-in-ethnic-diversity

Social Network Analysis - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. https://www.sciencedirect.com/topics/social-sciences/social-network-analysis