Modeling the ATP tour matches: A social networks analysis approach

Balša Knežević, Miloš Obradović, Predrag Obradović, and Marko Mišić, *Member, IEEE*

***Abstract*—Professional men's tennis is a demanding sport which greatly benefits from various approaches to performance analysis. More specifically, a complex network theory can be used to model and explain the dynamics of players and tournaments, based on the recorded matches. In this paper, played matches are used to model a social interaction between players. Several undirected weighted networks are constructed to model the ATP tour matches from 2018 to 2020. Moreover, the three most dominant players on the tour (the "Big Three") were observed and analyzed using ego networks approach. The chosen time frame further allowed for the exploration of impact of COVID-19 on the dynamics of the ATP tour. Different network properties were explored, such as small world phenomenon, core-periphery model applicability, community structure, and the rich club phenomenon. Our results based on network theory approach showed that analyzed networks expose similar topological properties, despite the lower numbers of tournaments held in the year 2020.**

***Index Terms*—collaboration network analysis; community detection; ego networks; men's tennis; network modelling.**

I. Introduction

Computational analysis of the results of sports competitions, as well as the performance of teams and individual athletes, has long been present in various sports. The development of data science and artificial intelligence, as well as the possibility of processing large amounts of data, have enabled new approaches to analyze the performance of both teams and individual players. In addition to traditional statistical methods, new methods have been developed, such as collaboration analysis and various prediction techniques.

Several methods based on network science were successfully applied to the analysis of team performance in collective sports, such as football [1][2], basketball [3], and water polo [4]. Furthermore, applications in individual sports are known, such as men's [5][6][7] and women's tennis [8], boxing [9], chess [10], cricket [11], etc. The goal of this paper is to further explore the usage of complex network analysis methodology in the field of men's tennis.

Balša Knežević is with the University of Belgrade - School of Electrical Engineering, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: balsa.knezevic@etf.bg.ac.rs).

Miloš Obradović is with the University of Belgrade - School of Electrical Engineering, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: miobra@etf.bg.ac.rs).

Predrag Obradović is with the University of Belgrade - School of Electrical Engineering, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: predrag.obradovic@etf.bg.ac.rs).

Marko Mišić is with the University of Belgrade - School of Electrical Engineering, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: marko.misic@etf.bg.ac.rs).

The world of tennis tournaments is a complex system that consists mainly of players, the tournaments they play and the matches they have played in those tournaments. Inherently, such a system is very convenient to model with an appropriate collaboration network. Most often, such a system is modeled by players representing the nodes of the network, while the matches that the players play are in some way depicted by the edges in the network.

In this paper, the state of men's professional tennis in the three years from 2018 to 2020 is modeled and analyzed. Similar analyses have already been done in the past for men's tennis in singles [5][7] and doubles [6]. In the meantime, great changes have taken place in the world of tennis. That primarily refers to more than a decade of domination of tennis players from the so-called "Big Three" to which Roger Federer, Rafael Nadal, and Novak Djokovic belong. Moreover, the COVID-19 virus pandemic affected the holding of tournaments in 2020, while tennis tournaments in 2018 and 2019 took place regularly. This allowed for comparative analysis and additional remarks on the impact of the COVID-19 virus pandemic. Therefore, various research methods have been applied in the paper, such as quantitative and qualitative analysis of the collaboration network, community detection, analysis of ego networks of members of the "Big Three", data visualization, etc.

The paper is divided into several sections. The second section describes the studied data sets and provides an overview of the used methodology. The third section presents the results of the research which are then discussed. Appropriate quantitative and qualitative analyses of the data, as well as the produced visualizations, are given. The last section provides guidelines for future work and a brief conclusion.

II. Data sets and methodology of analysis

This section presents the primary dataset and transformations performed on it in order to construct the dataset used for analysis. Furthermore, this section contains the methodology of analysis.

A. Data sets

This paper analyzes the results of men's singles matches played on the ATP tour in the period from 2018 to 2020. Although data from earlier years are available, this timeframe was chosen with the intent to include years 2018 and 2019 which are two consecutive years with regularly held tournament seasons, and the year 2020 which was influenced by the epidemic of COVID-19 disease. Thus, it is possible to

determine the influence of pandemics, as the dataset includes both the influenced data points and the two years of regular tennis seasons used as reference points.

The match data was taken on 12/22/2020 from a repository maintained by Jeff Sackman [12] and forms the primary data set for analysis. At the time of analysis, the tennis season for 2020 was completed. The primary data set consists of files containing data on matches in singles competition in the specified period, a list of all players ever ranked on the ATP list, data on the ranking of active tennis players on the ATP list in the period from 2010 to 2020. Match data contains information about the tournament, players, match results with statistics, and the performance of both players during the match. According to the author, the primary data set is largely refined and complete, but there may be certain inconsistencies or incompleteness where the data was not available.

The primary data set contains data on 7117 matches in the specified period. In addition, the data set contains data on 54,975 players who at some point during the observed period had at least 1 point on the ATP list. If several players have the same number of points, then they can share the same ranking on the ATP list, depending on other parameters.

The secondary data set was formed based on the primary dataset, as a refined and cleaned version of it. The data cleaning was performed according to the needs and goals of the research. During the process of cleaning and refinement, some data not necessary for the research itself was intentionally omitted, such as data on players who did not play any matches in the observed period, certain contradictory data, as well as redundant information (columns) that were not used in the analysis. The final secondary data set included data on 7117 matches, as well as data on 581 players.

B. Methodology of analysis

Firstly, a thorough statistical analysis of the dataset was conducted. Analyzed properties include the average number of tennis matches in certain years, the average number of tournaments in which tennis players participate, and the ranking of tennis players depending on the number of matches or tournaments played. Most interesting results are presented in the following section. Following the statistical analysis, the refined data set of tennis players and their mutual matches were used to create multiple collaboration networks. These networks were then further studied using methods of complex network theory.

Tennis tournaments are grouped in a season that lasts for a whole year. Therefore, three independent networks were constructed, each holding data for a specific year (N-18, N-19, N-20). Additionally, to allow the analysis of the whole data set, the three networks were aggregated into N-T. The four networks together are referred to as N-series networks.

As per common practice in the field of social network analysis, the network is represented through a set of nodes that describe the actors within the social network and the edges that represent social relations. In the case of networks used in this paper, the nodes of the network are tennis players who played at least one official ATP match in the analyzed period. The two tennis players are connected if they have played at least one official ATP match. The weight of the edge represents the number of matches that the tennis players played with each other. The networks are undirected.

In addition to networks representing all players and matches, the ego networks of the members of the "Big Three" were constructed for each year. These consist of prominent ego nodes, their direct connections with the neighbors, as well as the mutual connections of the neighbors. Furthermore, these three ego networks were unified, and then aggregated They were used to analyze the core-periphery property and the topology of the core of the N-series networks.

Community detection was performed by the Louvain method over the entire network, as well as over the aggregated ego network. For this purpose, a set of filtered and reduced networks was constructed. Clustering strength was evaluated, and the rich club phenomenon was examined.

Python programming language was used to collect and refine data, model the network, and calculate specific metrics using the NetworkX package [13] for network analysis. Gephi [14] was used to visualize and determine network metrics.

III. RESULTS

This section presents the results of the research. The first subsection explores the basic properties of N-series networks, while the rest explores the derived ego networks and community detection.

A. Basic properties of N-series networks

A statistical analysis of networks N-18, N-19, N-20, and N-T was conducted. Basic quantifiable features of those networks are presented in Table 1. As expected, the number of tournaments and matches held in 2020 is significantly lower due to the pandemic. This is further reflected in the weighted and unweighted degrees of nodes. However, looking only at the statistical data does not give the whole picture, as it would lead one to believe that year 2020 was significantly different from the previous two years. Only after applying the complex network theory methods discussed below one can give a proper conclusion about the impact of COVID-19 on the dynamics of the observed data sets.

TABLE I
METRICS OF CONSTRUCTED N-SERIES NETWORKS

	N-18	**N-19**	**N-20**	**N-T**
Players (nodes)	419	364	345	581
Edges	2489	2378	1325	5330
Matches	2974	2696	1447	7117
Tournaments total	138	123	67	328
Tournaments hard surface	81	80	46	207
Tournaments clay surface	47	34	20	101
Tournaments grass surface	10	9	1	20
Avg. weighted degree	13.79	15.28	8.39	24.50
Avg. unweighted degree	11.88	13.07	7.68	18.35
Network density	0.03	0.04	0.02	0.03
Avg. shortest path length	3.13	3.04	3.18	3.23
Diameter	11	9	9	10
Avg. clustering coefficient	0.17	0.19	0.14	0.26

Networks N-18, N-19, and N-20 have an exceptionally low density and a relatively low average shortest path length. Given the low average local clustering coefficient, the networks do not express the small-world property. This is in contrast with previous works in the field [15], but the discrepancies come from a completely different network model. These observations also stand for the aggregated network N-T, as the aggregation does not significantly increase the density nor strengthen the clustering.

Another interesting observation can be made about the average weighted and unweighted degrees. As shown in Table 1, the relative difference between weighted and unweighted degrees is small for networks N-18, N-19, and N-20. This shows that an average pair of tennis players rarely meet more than one time per season. Similarly, in the aggregate network N-T, the annual expected number of matches played by a pair of players is lower than 2. Given the bracket organization of tennis tournaments and loser-go-home policy, only the best players are expected to play multiple matches in a tournament. This leads to the probability of two players meeting in a tournament being quite low, even if they both play in the tournament. In addition, a low annual number of tournaments leads to a low number of annual matches and further decreases the possibility of two players meeting.

A further discussion on this topic can be made when tournament seeding is taken into consideration. The probability of the first and second seed in a tournament gets further artificially lowered, as they are seeded in opposite sides of the bracket and are unable to meet before the finals. If a pair of players is consistently seeded with the top two seeds, this can lead to a measurable decrease in the weight of the edge connecting them.

B. Analysis of ego networks

Looking only at the average number of matches played does not show the whole picture and unravel the true topology of the constructed networks. Therefore, a distribution of the number of matches played during the observed period has been calculated and is shown in Fig. 1. As can be seen, many players have only played one or two matches and are thus very isolated, suggesting a core-periphery topology.

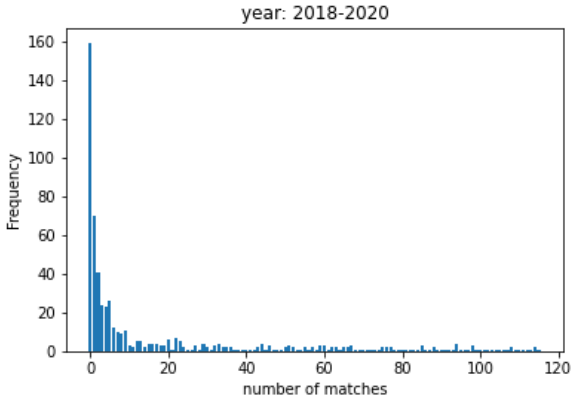

Fig. 1. Distribution of the number of matches played during the three years from 2018 to 2020. The distribution largely resembles a Pareto distribution.

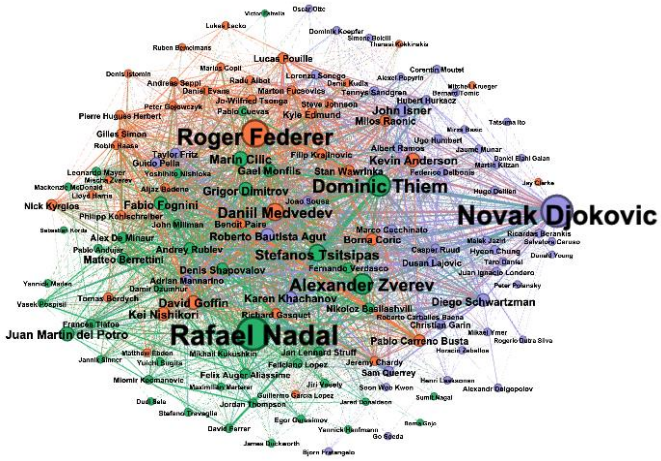

Fig. 2. EGO-T, a unified ego network of the "Big Three" for the period from 2018 to 2020. The size of the node represents weighted node degree and nodes are colored based on clustering.

As stated in the section about methodology, to check if N-18, N-19, N-20, and N-T networks follow the core-periphery model and unravel the topology of the cores of specified networks, several ego networks centered around the members of "Big Three" were constructed. Annual ego networks of Djokovic, Nadal, and Federer were then unified into EGO-18, EGO-19, and EGO-20. Together with these ego networks, their aggregated network EGO-T, shown in Fig. 2, was built.

Clustering the EGO-T network using the Louvain method [16] and tuning the resolution to give 3 clusters reveals a very interesting phenomenon. The original ego nodes bind stronger to some of the other nodes in the network than between themselves. This is in concert with the aforementioned observation about the bracket system and seeding principles influencing the edge weights on the very top of the ATP list.

Exploring the number of nodes and edges of N-series networks included in EGO-series networks can help us explore the properties of the core of N series networks. These statistics are therefore shown in Table 2.

TABLE II
METRICS OF EGO-SERIES NETWORKS

	EGO-18	EGO-19	EGO-20	EGO-T
Nodes	81	88	57	136
Nodes covered	19.33%	24.17%	16.50%	23.4%
Edges	691	744	202	2563
Edges covered	27.76%	31.28%	15.24%	48.08%

Given the percentage of all players and matches included in EGO-series networks, it is obvious that even EGO-T which aggregates other EGO networks and enhances the core property can not be considered a core by itself. Further exploring this topic, the Rombach core finding algorithm [17] was applied to find cores of N-18, N-19, N-20, and N-T, giving cores with 234, 200, 193, and 315 players, respectively. These cores are much larger than EGO series networks and include most of the players.

However, a remark has to be made about the EGO-T network and the percentage of matches included in it. Even though EGO-T is more than two times smaller than the core of N-T, it includes 48.08% of all matches recorded in N-T,

which is an astounding amount. This means that matches between the players from EGO-T represent nearly half of all the ATP matches played from 2018 and 2020 and could be used to study some phenomena on a smaller, but representative, group of players, without drastically compromising the number of matches included in the data.

C. Community detection and the rich club phenomenon

To discover a more fine-grained structure in the constructed networks, in addition to exploring network cores, the Louvain method was used once again to find communities in N-18, N-19, N-20, and N-T. Before running the Louvain method, all nodes with degrees lower than 3 were removed from N-18, N-19, and N-20 to avoid the formation of forced and unnatural clusters due to modularity optimization. Characteristics of these reduced networks, aptly named R-18, R-19, and R-20 (R standing for "reduced"), are shown in Table 3. Moreover, a similar procedure was applied to N-T, removing all players with less than 5 matches during the three years, giving us R-T, a reduced network of total aggregated data.

TABLE III
METRICS OF R SERIES (REDUCED) NETWORKS

	R-18	R-19	R-20	R-T
Nodes	244	203	181	287
Nodes retained	58.23%	55.77%	52.46%	49.40%
Edges	2292	2190	1159	4889
Edges retained	92.09%	92.09%	87.47%	91.73%
Communities	9	6	8	7
Avg. clustering coeff.	0.22	0.24	0.18	0.32

The process of node removal is validated by looking at the percentage of nodes and edges retained in the reduced networks. As we can see in Table 3, 49.40% of players played 91.73% of matches during the observed three-year period. This phenomenon can also be seen in Figure 1. As the distribution of the number of matches loosely follows a Pareto distribution, it is to be expected that a rich-club phenomenon can express itself when considering the number of matches as "wealth". This is somewhat validated by looking at EGO-T, as it consists of a small group of players which bind strongly to each other and monopolize the number of matches over the observed period.

Communities formed by the Louvain modularity clustering are grouped by average rating during the period. This is to be expected, as players of similar ratings choose to play and qualify for the same class of tournaments and are more likely to meet each other. However, the clustering is still not strongly expressed, as can be seen from the average local clustering coefficients.

IV. CONCLUSION

Studying interactions of men's tennis players proved to be interesting in several aspects. Motivated by the available data, several undirected weighted networks with node metadata were constructed, analyzed, and characterized and multiple common phenomena in the field of complex network theory were explored. Those include small world phenomenon, core-periphery model applicability, community detection, and the rich club phenomenon. In addition, the authors' own experience with the topic helped explain many of the observed properties and the given explanations are one of the biggest results of this paper, as they give a much better understanding of the dynamics of men's tennis and are a result of social network analysis and network theory approach to the problem.

In addition, provided network models clearly show an impact of the COVID-19 pandemic on the tennis world, through a smaller number of matches and participants. However, the network theory methodology applied in this paper also shows that the topological properties of the data (such as clustering properties, rich club and small-world phenomena, core-periphery property) stay largely the same, which could not be inhered by naive statistical analysis of the primary data set.

This paper and the constructed networks form a strong basis for further exploration of the topic, including the analysis of mixing patterns in the data depending on the ratings of players, geographical locations of tournaments, affiliations of players, etc. Furthermore, the data in network form is much more suitable for solving some regularly asked questions in the field, such as ranking and match outcome prediction using graph convolutional networks or graph attention models. Lastly, the provided networks are an ideal model for the problem of choosing the representatives of the international tennis community, touching upon the problem of choosing the dominating set of the graph.

ACKNOWLEDGMENT

This work has been partially funded by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. Grant numbers III44009 and TR32047.

REFERENCES

[1] T. Grund, "Network structure and team performance: The case of English Premier League soccer teams", *Social Networks*, 34(4), pp.682-690, 2012.

[2] J. Gama, M. Couceiro, G. Dias, V. Vaz, "Small-world networks in professional football: conceptual model and data", *European Journal of Human Movement*, 35, 85-113, 2015.

[3] F. M. Clemente, F. M. L. Martins, D. Kalamaras, R. S. Mendes. "Network analysis in basketball: Inspecting the prominent players using centrality metrics", *Journal of Physical Education and Sport*, 15(2), 212, 2015.

[4] P. Passos, K. Davids, D. Araújo, N. Paz, J. Minguéns, J. Mendes, "Networks as a novel tool for studying team ball sports as complex social systems", *Journal of Science and Medicine in Sport*, 14(2), 170-176, 2011.

[5] F. Radicchi, "Who is the best player ever? A complex network analysis of the history of professional tennis", *PloS one*, 6(2), e17249, 2011.

[6] K. Breznik, "Revealing the best doubles teams and players in tennis history", *International Journal of Performance Analysis in Sport*, 15(3), 1213-1226, 2015.

[7] U. Michieli, "Complex Network Analysis of Men Single ATP Tennis Matches", *arXiv preprint arXiv:1804.08138*, 2018.

[8] M. Kostić, D. Drašković, "Complex Network Analysis of Women's Singles Tennis Matches", Telecommunications Forum (TELFOR), Belgrade, Serbia pp. 1-4, IEEE, 2020.

[9] A. G. Tennant, C. M. Smith, J. E. Chen C, "Who was the greatest of all-time? A historical analysis by a complex network of professional boxing", *Journal of Complex Networks*, 8(1), cnaa009, 2020.

[10] N. Almeira, A. L. Schaigorodsky, J. I. Perotti, O. V. Billoni, “Structure constrained by metadata in networks of chess players”, *Scientific reports*, 7(1), 1-10, 2017.
[11] S. Mukherjee, “Identifying the greatest team and captain—A complex network approach to cricket matches”, *Physica A: Statistical Mechanics and its Applications*, 391(23), 6066-6076, 2012.
[12] J. Sackmann, Repository tennis_atp, available on: https://github.com/JeffSackmann, accessed: 22.12.2020.
[13] A. A. Hagberg, D. A. Schult, P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA, pp. 11–15, August, 2008.
[14] M. Bastian, S. Heymann, M. Jacomy, (2009, March). “Gephi: an open source software for exploring and manipulating networks”, Proceedings of the International AAAI Conference on Web and Social Media, vol. 3, no. 1, 2009.
[15] H. Situngkir, “Small world network of athletes: Graph representation of the world professional tennis player”, Available at SSRN 1001917, 2007.
[16] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, “Fast unfolding of communities in large networks”, *Journal of statistical mechanics: theory and experiment*, (10), P10008, 2008.
[17] M. P. Rombach, M. A. Porter, J. H. Fowler, P. J. Mucha, “Core-periphery structure in networks”, *SIAM Journal on Applied mathematics*, *74*(1), 167-190, 2014.