CS224W Final Report
A Network Analysis of Professional Tennis

James Sun

November 17, 2016

1 Introduction

Professional sports analyses has long fascinated people. Among those sports, tennis is uniquely suited for network modeling. Compared to team sports, tennis interactions are between two individuals rather than groups, so participants and matches are much more numerous. Also, the tennis network evolves extremely naturally because players elect to enter a tournament rather than being assigned by a governing body. Finally, tennis's long history provides a wealth of data to validate analyses. For these reasons, tennis presents a special opportunity to use well-studied network techniques to gain insights in the underlying backbone of the sport.

In this project, I specifically investigate the subnetworks within tennis. At a high level, the Association of Tennis Professionals (ATP) covers the male tennis players while the Women's Tennis Association (WTA) covers the female players. Below that, the ATP is divided into 3 tiers; in decreasing order of prestige, they are the World Tour, the Challenger Tour, and the Futures Circuit. The WTA is divided into only 2 tiers; in order of decreasing prestige, they are the Women's Tour and the Pro Circuit (also known as the ITF). For the remainder of this report, I refer to the Women's Tour as the WTA and the Women's Pro Circuit as the ITF.

By building static and year-to-year dynamic network models, I attempt to quantify the salient similarities and differences between the Men's and Women's networks as well as the characteristics of the different tiers as a consequence.

2 Related Work

Radicchi's work on analyzing tennis matches in a graph framework provided the inspiration for this research [1]. However, Radicchi only considered ATP World Tour matches before 2010 focused only on developing a player ranking. I include a much larger dataset and use network analysis to identifying characteristics of groups within tennis.

Leskovec et al. and Kumar et al. investigated the time evolution of network graphs. The former empirically demonstrated that networks densify and shrink over time [2], and the latter found similar findings (though with notably different elements) [3]. These counter-intuitive traits were previously undiscovered because of the rarity of long-term graph data. However, my data covers over 40 years and was well-suited to this analysis, and comparison to these traditional networks helped identify distinguishing traits of some of the tennis subnetworks.

Finally, the main area of interest in the tennis graph is the interactions between players which is captured by the edges and node degrees. These quantities are associated with power laws, and Clauset et al. give a very useful framework to properly analyze and fit power law distributions.

3 Formulation

I formulate two types of graphs:

1. An undirected graph where each node represents a player, and an edge between two nodes means the respective players have played a match at least once. There are no multiedges.

2. A directed graph where each node represents a player, and edges represent matches. Each edge points from loser to victor, in accordance with the status theory of social science applied to networks [5]. Multiedges are allowed.

Both graph types are valid network representations and give complementary insights. I formulate these graphs for each of the following types of matches: ATP World, ATP Challenger, ATP Futures, WTA Tour, and ITF. I use the SNAP library and the Gephi tool to perform graph analyses [6, 7].

4 Data

Both the ATP and WTA have official publicly available archives, but only the ATP is amenable to systematic data collection. Instead, I used data from www.tennisabstract.com which provides both Men's and Women's data in analysis-friendly formats.

The Men's data consists of ATP matches between 1968-2015, Challenger matches between 1991-2015, and Futures matches between 1991-2015. The Women's data consists of Tour matches between 1968-2015 and ITF matches between 1969-2015. The data is described in Table 1.

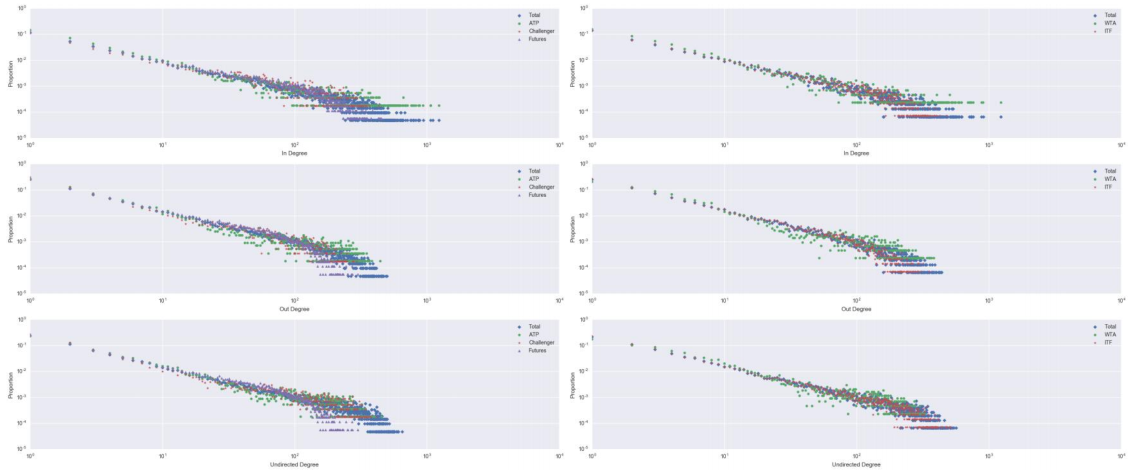
Table 1: Data Statistics

Type	# Players	# Matches
Men's World	5544	160902
Men's Challenger	5702	118472
Men's Futures	17724	356545
Total Men's	**20923**	**635919**
Women's WTA	4272	103032
Women's ITF	13964	274108
Total Women's	**15204**	**377140**

5 Methods and Results

5.1 Degree Distribution

I examined the in-, out-, and undirected degree distributions for each of the graphs listed in Section 3. The in- and out-degree distributions considered multi-edges while the undirected degree distributions did not. Visually, all graph networks produced similar distributions with slight differences between degree types as shown in Figure 1. For example, the in-degree distributions had noticeably longer tails than the out-degree distributions, meaning players with an extremely large number of wins are more likely to exist than players with an extremely large number of losses. This is intuitive because 1. players can have multiple wins per tournament but only one loss and 2. players likely do not stay in the network long enough to amass an extreme loss count.



(a) Men

(b) Women

Figure 1: Degree Distributions

All datasets superficial appeared to follow a power law with exponent ≈ 1, which is impossible in real networks because of an infinite first moment. To investigate this, I followed the method described in [4]. I chose 5 candidate distributions: pure power law, log-normal positive, exponential, stretched exponential, and truncated power law (power law with exponential cut-off). First, I fit the degree data on each candidate distribution using Maximum Likelihood Estimation (MLE). Then, I computed the log-likelihood ratio (LR) between every pair of fits along with the corresponding p-value. A positive LR meant that the first distribution fits the data better than the second. Results are given in the following subsection.

5.1.1 Degree Distribution Results

An example using the undirected ATP World data is shown in Table 2. The row corresponding to the truncated power law (the Cutoff row) has only positive LR values associated with ≈ 0 p-values. Thus, the truncated power law was a statistically better fit than all other candidate distributions.

Notably, this does not preclude an untested distribution from being more appropriate, but overlaying the truncated power fit on top of the pdf in Figure 2 does indeed look like a good fit. A truncated power law distribution obeys the following equation:

$$f_{trunc}(x) = C(\alpha, \lambda, x_{min})x^{-\alpha}e^{-\lambda x}$$

	Power		Log-Normal		Exp		Cutoff		Stretched	
	LR	p-val	LR	p-val	LR	p-val	LR	p-val	LR	p-val
Power	-	-	-279.93	0.00	1611.25	0.00	-427.77	0.00	-334.13	0.00
Log-Normal	279.93	0.00	-	-	1891.18	0.00	-147.83	0.00	-54.20	0.00
Exp	-1611.25	0.00	-1891.18	0.00	-	-	-2039.01	0.00	-1945.38	0.00
Cutoff	427.77	0.00	147.83	0.00	2039.01	0.00	-	-	93.64	0.00
Stretched	334.13	0.00	54.20	0.00	1945.38	0.00	-93.64	0.00	-	-

Table 2: *Each row represents a candidate distribution. Each column represents a comparison distribution. LR> 0 means the candidate is a better fit than the comparison*

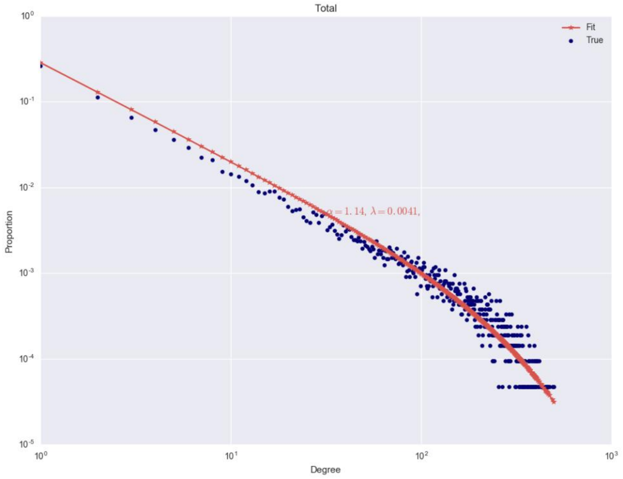


Figure 2: Fitting a truncated power law to the ATP degree distribution

The exponent was fit as 1.14, much less than seen in most real networks [8, 9]. However, the cutoff prevents the first moment from diverging, and cutoff power distributions with small exponents have been observed in real networks [10]. Furthermore, both the cutoff and low exponent have intuitive explanations:

- Cutoff: Only a small number of uncommonly talented players have long enough careers to play a very large number of matches. Most of these top players exit the network at around the same time in their careers, when old age prevents them from competing
- Small Exponent: The lower tiers do not attract the most experienced and gifted players, allowing the newer and less skilled players to compete without experiencing too much crushing discouragement from facing the elite professionals.

In fact, the truncated exponential was the unambiguous best fit for all undirected graphs with fit results given in Table 3. From the fits, it seems the ITF is more analogous to the Futures Circuit then the Challengers Tour. The ITF and the Futures networks have similar fit parameters, and both seem to influence the overall networks more than the other tiers.

Network	α	λ
ATP	1.14	0.0036
Challenger	1.17	0.0032
Future	1.07	0.0048
Overall Men	1.09	0.003
WTA	1.12	0.0052
ITF	1.09	0.0045
Overall Women	1.09	0.0038

Table 3: Truncated Power Law Fits

The directed networks had identical results though with different fit parameters. Also, the directed Challengers network had a relatively high p-value of 0.075 when concluding that the truncated distribution was a better fit than the stretched exponential. These two distributions are very similar, and I did not investigate further.

5.2 Time Evolution

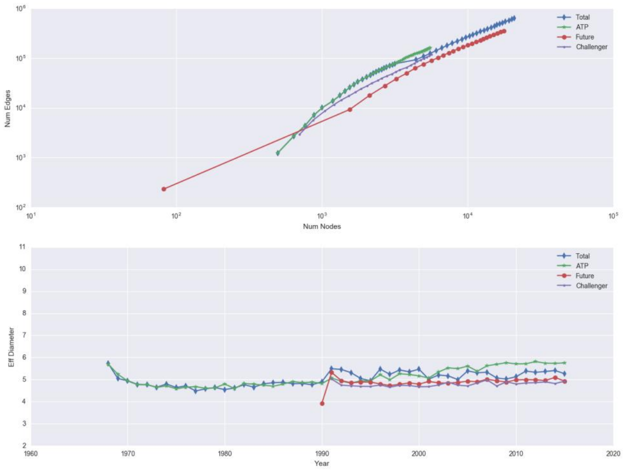
I also investigated the time evolution of the tennis networks by successively adding each year's data to the graphs. First, I investigated network density and diameters as in [2], with results for the undirected graph given in Figure 3. The number of nodes and edges was indeed consistent with the expected densification results. I also found that the Futures and ITF tiers were again similar in that they were the fastest growing tiers, as the number of Futures/ITF players and matches significantly outnumber those of the others.

However, I did not find increasing diameters. Instead, the effective diameters either remained constant or trended slightly upwards, with the ATP network exhibiting the most notable upward trend. The exception was the ITF network in which the diameter trended downward after reaching a very high value in the 70s. Numerical diameters over aggregated year data is presented in Table 4.

Graph	Undirected Diameter	Directed Diameter
ATP	5.64	5.78
Challenger	4.86	4.81
Futures	5.03	5.03
All Men	5.43	5.31
WTA	5.86	5.79
ITF	5.69	5.68
All Women	5.64	5.55

Table 4: 90th Percentile Diameters

The lack of shrinking diameters may be explained again by the fact that the tier system in tennis actively encourages newer and less skilled players to stay in the profession by separating them from the veteran players. The increasing ATP diameter might then point to a growth in either the number of extremely skilled or number of newly elite players. Conversely, the relatively constant WTA diameter may indicate a stagnation in the upper and lower echelons of the highest tier of Women's players.

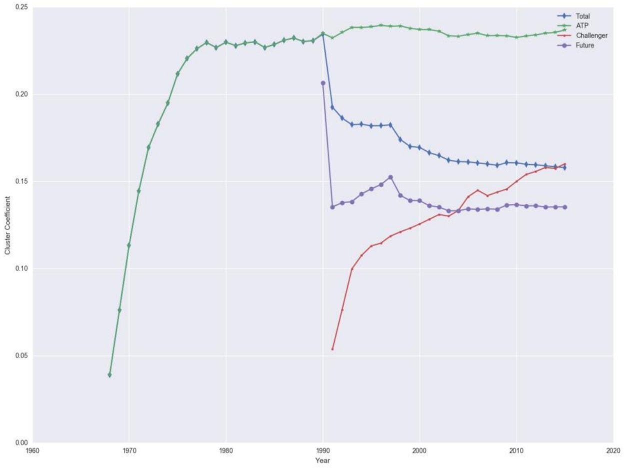


(a) Men

(b) Women

Figure 3: Densification and Diameter Evolution for Undirected Graphs

For further insight, I also calculated the evolution of the average global clustering coefficients for each graph shown in Figure 4. For the Men, the clustering coefficient has remained constant since 2005 in all networks except the Challenger's Tour in which it is trending upwards. In fact, the Challenger Network has surpassed the overall Men's network in Clustering Coefficient as of 2015. For the Women, the clustering coefficient for the overall network seems to have converged since 2005, but both the ITF and the WTA subnetworks seem to have a very slight upward trend, and the ITF network has a substantially lower clustering coefficient than the overall Women's network. Figure 4 emphasizes that Women do not have a middle tier that is analogous to Challenger Tour.



(a) Men

(b) Women

Figure 4: Evolution of the average Clustering Coefficient for Undirected Graphs

5.3 Communities

For a final holistic comparison between the Men's and Women's networks, I performed community detection. I limited the networks to data between 2005-2015 for computational reasons and to focus on the modern era. I used the directed, weighted representations and filtered out the players with fewer than 10 matches. I also filtered out was a tiny disconnected component of 2 players in the Women's data.

In both the Men's and Women's networks, I detected 5 communities using the Louvain method implemented in Gephi [7, 11]. Visualizations of the communities are shown in Figures 5 and 6. Nodes are color coded by community, edges are color coded by source node community, and node sizes correspond to PageRank. For each community member node, I aggregated the number of matches of each separate tier in which the corresponding player appeared. Afterwards, I calculated the proportion of match types represented by each community. These proportions along with the aggregated PageRanks significantly differentiated the communities. The results are summarized in Tables 5 and 6.

Community	**Size**	**PageRank Sum**	**ATP Prop**	**Challenger Prop**	**Future Prop**
Blue	2185	0.110332	0.024870	0.174352	0.800778
Orange	1510	0.089454	0.089454	0.174352	0.800778
Green	3455	0.17999	0.015752	0.139413	0.844834
Purple	3313	0.204719	0.007460	0.111247	0.881286
Red	901	0.408461	0.353309	0.488333	0.158359

Table 5: Men's Communities

Community	**Size**	**PageRank Sum**	**WTA Prop**	**ITF Prop**
Blue	678	0.093389	0.030491	0.969509
Orange	408	0.051863	0.009760	0.990240
Green	588	0.092855	0.036237	0.963763
Purple	1706	0.246948	0.016969	0.983031
Red	493	0.415444	0.401022	0.598978

Table 6: Women's Communities

The communities between the Men's and Women's networks are strikingly similar. The tier proportion breakdown again indicates the correspondence between the ITF and the Futures, but it also suggests that a middle tier in the Women's network would draw players from both the ITF and the WTA. As an aside, the visualization also shows that PageRank is much more distributed in the Women's game than the Men's game, revealing that there is a small group of very dominant Male players but no corresponding group of Female players.

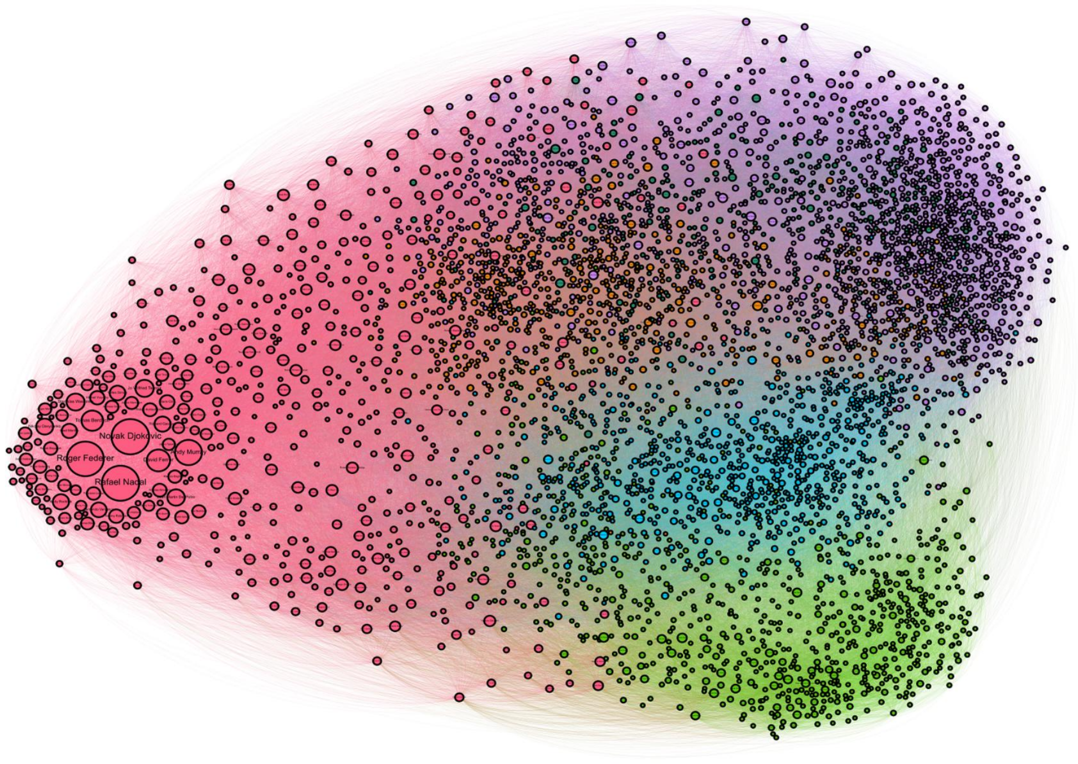


Figure 5: Men's Network

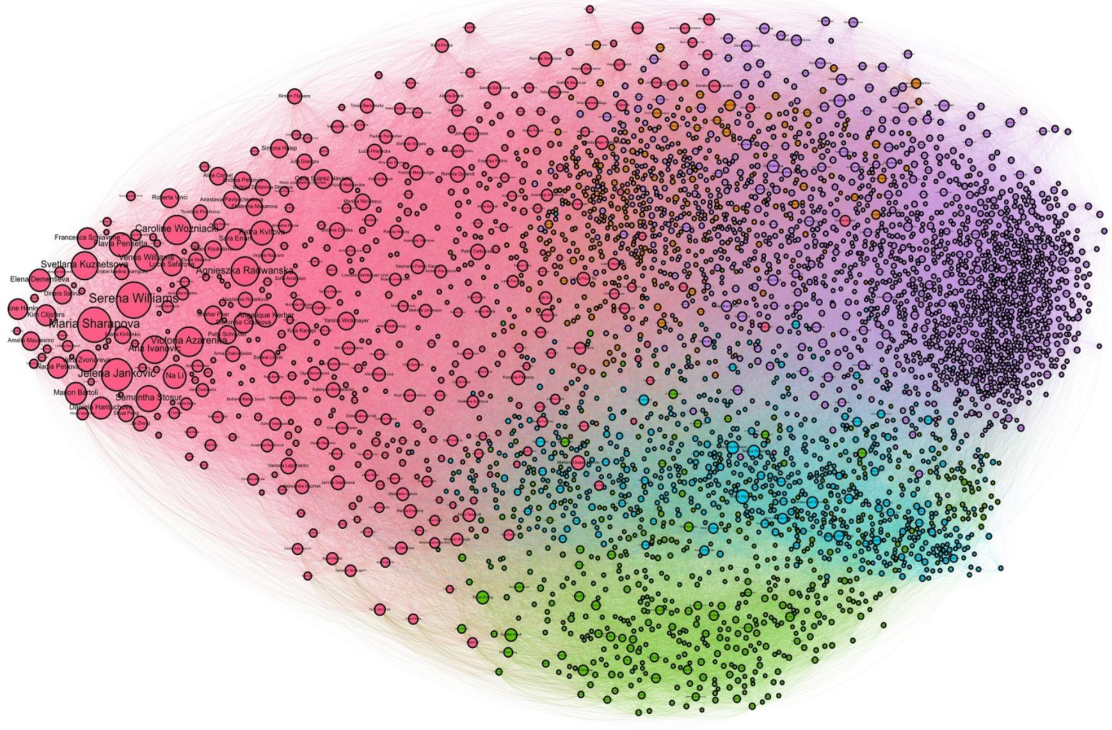


Figure 6: Women's Network

6 Conclusion and Future Work

This project presents multiple findings that differentiate the Male and Female tennis network. The degree distribution is one aspect in which the Women and Men networks are extremely similar. However, looking at the time evolution of the networks reveals that one of the major differences in the two networks arises in the lack of a middle tier for the Women. Looking at community structure reveals that the two networks are structured exactly the same and that the equivalent of a Women's middle tier is currently divided between the lower and upper tiers of competition.

Future work would further investigate whether this omission has a substantial effect on the Women's network, as hinted by the stagnant WTA diameter referenced in Section 5.2. Further work could also be done to investigate the different PageRank profiles between the two networks. Early analysis suggests that PageRank is more concentrated in the Men's than in the Women's, but further consideration must be done before drawing any conclusions.

References

[1] F. Radicchi, "Who is the best player ever? a complex network analysis of the history of professional tennis," *PloS one*, vol. 6, no. 2, p. e17249, 2011.

[2] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.* ACM, 2005, pp. 177–187.

[3] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link mining: models, algorithms, and applications.* Springer, 2010, pp. 337–357.

[4] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, 2010, pp. 1361–1370.

[6] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 1, 2016.

[7] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," 2009. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[8] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[9] S. Redner, "How popular is your paper? an empirical study of the citation distribution," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 4, no. 2, pp. 131–134, 1998.

[10] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proceedings of the 17th international conference on World Wide Web.* ACM, 2008, pp. 915–924.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.