SI. Early Career Wins and Tournament Prestige Characterize Tennis Players' Trajectories

Chiara Zappalà[1,2,3,4], Sandro Sousa[3,4], Tiago Cunha[3], Alessandro Pluchino[2], Andrea Rapisarda[2,5], and Roberta Sinatra[4,3,5,6]

[1]Center for Collective Learning, Corvinus Institute for Advanced Studies (CIAS), Corvinus University, 1093 Budapest, Hungary
[2]Department of Physics and Astronomy, University of Catania and INFN sezione di Catania, 95123 Catania, Italy
[3]NEtwoRks, Data, and Society (NERDS), Computer Science Department, IT University of Copenhagen, 2300 Copenhagen, Denmark
[4]Center for Social Data Science (SODAS), University of Copenhagen, 1353 Copenhagen, Denmark
[5]Complexity Science Hub, 1080 Vienna, Austria
[6]ISI Foundation, 10126 Turin, Italy

SUPPLEMENTARY INFORMATION

The effect of active players

The ATP dataset we analyzed consists of players who have appeared for at least two years in the official ranking. We only consider players that started their careers within our dataset, so from 2000 on. However, some of those players could still be active at the end of 2019, which is the upper bound of our dataset. In the main text, we controlled for right-censored data, which occur when the time of observation ends before a certain event, when we examined the time of the peak along a career, in terms of ranking points. In panels B-C of Fig. 1, we considered a subset of 2,262 players who started and ended their careers within our observation time. Here, we show the time of the career peak if we include active players (Fig. S1). In this case, we see that top players (red line) are the only ones with different behavior, most likely because their careers tend to last longer; thus, they either have not reached their peak yet or their amount of points is stable.

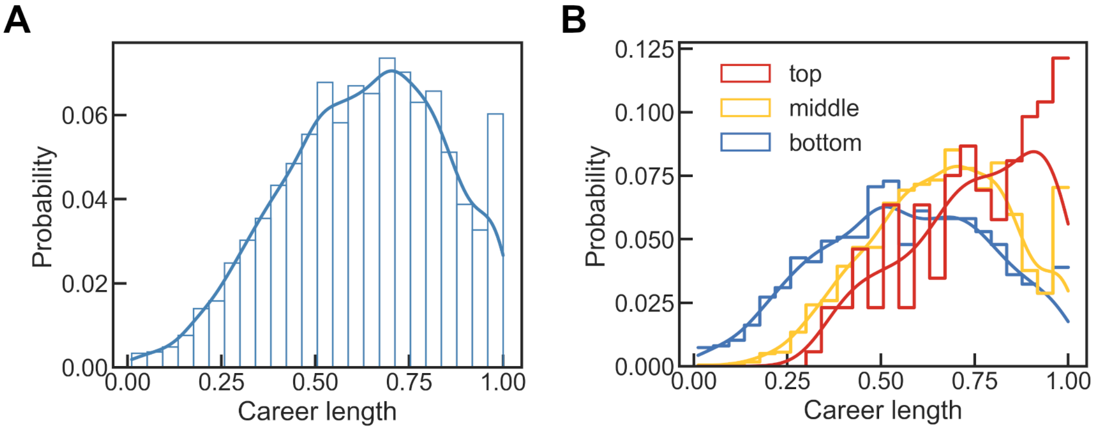

Figure S1. Career peak distributions, active players included. **A** Distribution for the whole community of players. **B** Distribution after splitting players into groups. Top players (in red) have different behavior, most likely due to their longer career at the professional level. Histograms are normalized so that bar heights sum to 1, and are reported with a Kernel Density Estimation of the data (continuous curves).

We specify that players might reach their maximum number of points more than once in their professional career. We tackle the possibility of multiple peaks by choosing the time of their peak at random. It is worth mentioning that the continuous curves of Fig. S1 derive from a Kernel Density Estimation of the data [1].

The presence of active players in the data also impacts the evolution of their ranking points over time. Indeed, the curves in panel B of Fig. 2 all seem to decline as they approach the end of the observation time. This could suggest that players experience an overall decrease in their points before the end of their career due to a reduction in the number of tournaments played or an increase in poorer performances [2]. However, we can appreciate the decline in the tail of these timelines only if we disentangle the contribution of active players. Therefore, in Fig. S2, we show the average trend of points in terms of ranking appearances for players who started and ended their careers within the dataset. We observe a steep decrease in the ranking points of the top players, more evident than in the other groups.

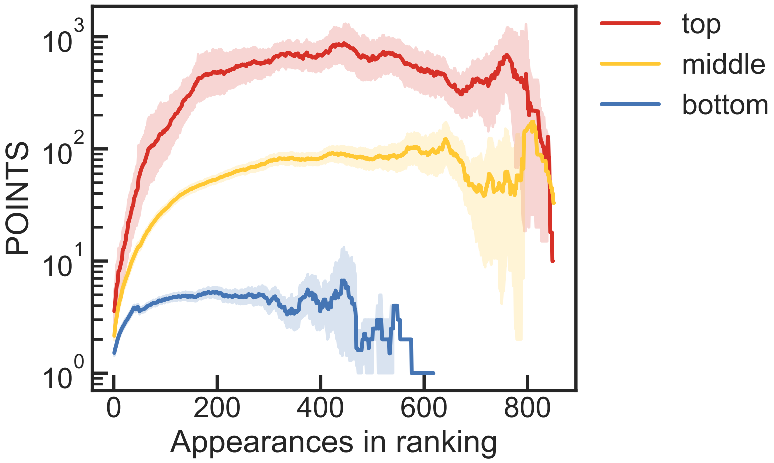

Figure S2. Average trend of male tennis players in ATP ranking, active players excluded. The tail of the evolution of points for the top players (red curve) suddenly drops, while the decline is smoother for the middle (yellow) and bottom (blue) players.

Network features

As explained in the main text, we built the co-attendance network of tennis tournaments based on the trajectories of players along their careers. This results in a weighted directed network, where nodes are tourneys and links (i, j) are created when players first attend tournament i, then j. In Table S1 we summarize the main characteristics of this network, which is a dense and highly clustered graph. These characteristics could come from the seasonality of the ATP tour, which "forces" athletes to repeat their trajectories to preserve their ranking points from the previous year, if not improve them.

Feature	Value
Nodes	651
Links	254583
Density	0.60
Clustering	0.77

Table S1. Summary of the features of the co-attendance network.

In detail, the density of the network can be calculated as the ratio:

$$\frac{m}{n(n-1)} \tag{S1}$$

Where m is the number of links and n the number of nodes in the network. The average clustering coefficient is defined as follows:

$$C = \frac{1}{n}\sum_i \frac{2t_i}{k_i(k_i-1)} \tag{S2}$$

Where n is the number of nodes, k_i is the degree of node i and t_i the number of triangles having node i as one of the vertices [3]. For simplicity, only in the case of Eq. (S2) we assume an undirected and unweighted network.

We use the topology of the co-attendance network to assess the prestige of tourneys. Specifically, we rely on the eigenvector centrality [4, 5]. For this network, the distribution of the eigenvector centrality is asymmetric, as shown in Fig. S3.

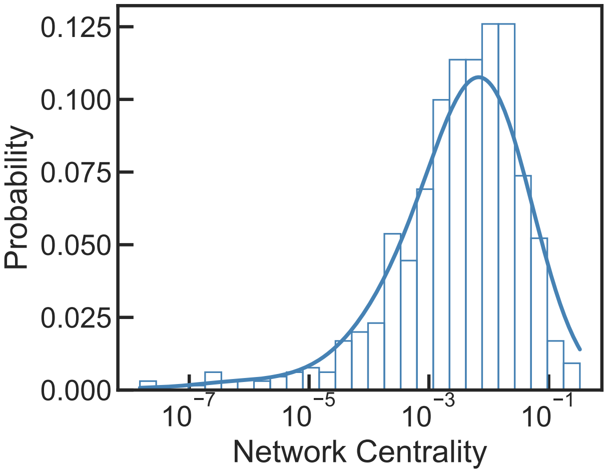

Figure S3. Distribution of the eigenvector centrality for the tourneys (i.e., the nodes) of the co-attendance network. It is asymmetric around its peak. Bar heights sum to 1, the curve shows a Kernel Density Estimation of the data.

Tournaments and their impact on players' careers

The centrality of the competitions in the network is in agreement with the historical level of the tournaments (see Fig. 4B), expressed by the different point scales that each category of tourney can award [6], as summarized below (Table S2). However, tournaments belonging to the same ATP category can have vastly different centralities (see Fig. 4C), suggesting that network topology offers unique lens to the distinctiveness of each competition. It is worth mentioning that here we refer to the ATP rulebook of 2019. Since then, newer versions of the rulebook have been adopted.

	Grand Slam	Masters 1000	ATP 500	ATP 250	Challenger
Winner	2000	1000	500	250	125
Final	1200	600	300	150	75
Semifinal	720	360	180	90	45
Quarter-final	360	180	90	45	25
Round-of-16	180	90	45	20	10
Round-of-32	90	45	20	10	5
Round-of-64	45	25	-	-	-
Round-of-128	10	10	-	-	-
Qualif-1st	25	16	10	5	-
Qualif-2nd	16	8	4	3	-
Qualif-3rd	8	-	-	-	-

Table S2. Allocation of points per tournament and round. Points are assigned to the losers of the indicated round. Please note that draws do not have a fixed length (except for Grand Slams, which always have a 128-draw). Here, we show the points players can receive in tournaments with the maximum possible draw per type.

In the main text, we showed that the centrality of the first ten tournaments players attend is not associated with players' future success. As we can see in Fig. S4, the top players can be distinguished from the others by the prestige of the tourneys attended only after 40 competitions, and they consistently compete in high-level venues around the 60th tournament.

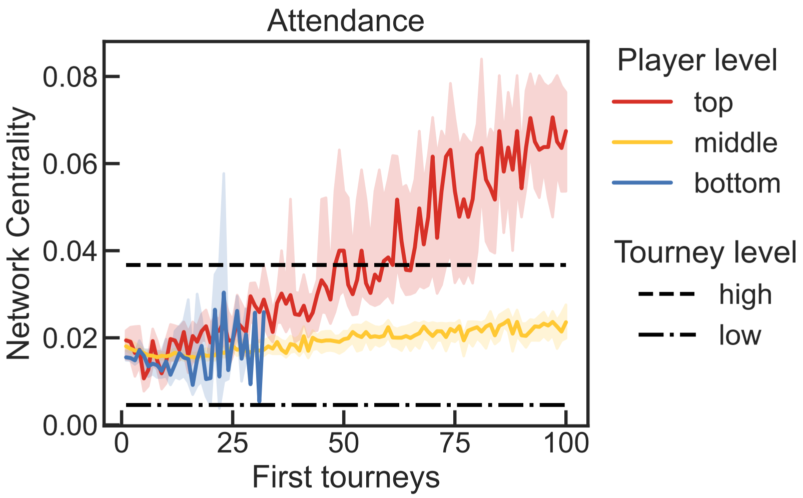

Figure S4. Centrality for the first 100 attended tournaments per group of players. The solid lines show the median trend and the shadows represent their confidence interval. The dashed (dashed-dotted) lines refer to the high (low) level threshold of tourney splitting. Top players (red dots) consistently compete in high-level tournaments only after 60 attendances.

In addition, we focus on the centrality distribution for each of the tournaments we consider (from tournament 1 to tournament 10 per player level). Again, we observe a common trend among the groups of players, even in this fine-grained visualization (see Fig. 5E of the main text for the aggregated version), when we only consider participation (Fig. S5A). In other words, there is no appreciable distinction in the average level of the first ten tourneys that players attend. However, a difference emerges if we consider the level of the tournament in which the players won a match for the first time, as shown in Fig. S5B: The centrality distributions are often above the high-level threshold for the top players (red boxplots), confirming the robustness of the results (see Fig. 5F of the main text for an aggregated visualization of the distributions).

In Fig. S6 we show the weekly evolution of the ranking points for two distinct pairs of players, each consisting of a top (in red) and a middle (in yellow), before and after they won a match in a tournament belonging to the same ATP category (ATP 250) but with different eigenvector centrality (a star marks the week of their first match win, red for top and yellow for middle, and the respective network-based tourney level is annotated next to it). The weeks during which they participated in a tournament in our dataset are highlighted by the white-edged red dots (top) and white-edged yellow triangles (middle). Although all players won their first match in an ATP 250 tournament, the two top players won in a high-level tourney, while the two middle players won in a medium-level tourney. In both panels, we see that the middle players were not able to rise in the ATP ranking, not even when their first win happened much earlier than the top player's one (panel A). Also, in both cases the middle player had more ranking points than the top player at the time of their first win (respectively, 52 and 20 in panel A, 75 and 50 in panel B). Notice that in panel A we do not observe any decreasing trend in the number of points because they are both still active players, while we observe a decrease in panel B because both players ended their career within the dataset.

Robustness of the impact of the first win

We check that the predictive power of the first win still holds if we add (or remove) some constraints. In detail, we verified that the first win within the first ten tournaments of players at the beginning of their career identifies the top players in the following cases: if we count the qualification rounds (Fig. S7); if we consider only players who attended more than ten competitions in our dataset (Fig. S8); if we exclude active players (Fig. S9). In all those scenarios, we can still observe that the top players act differently

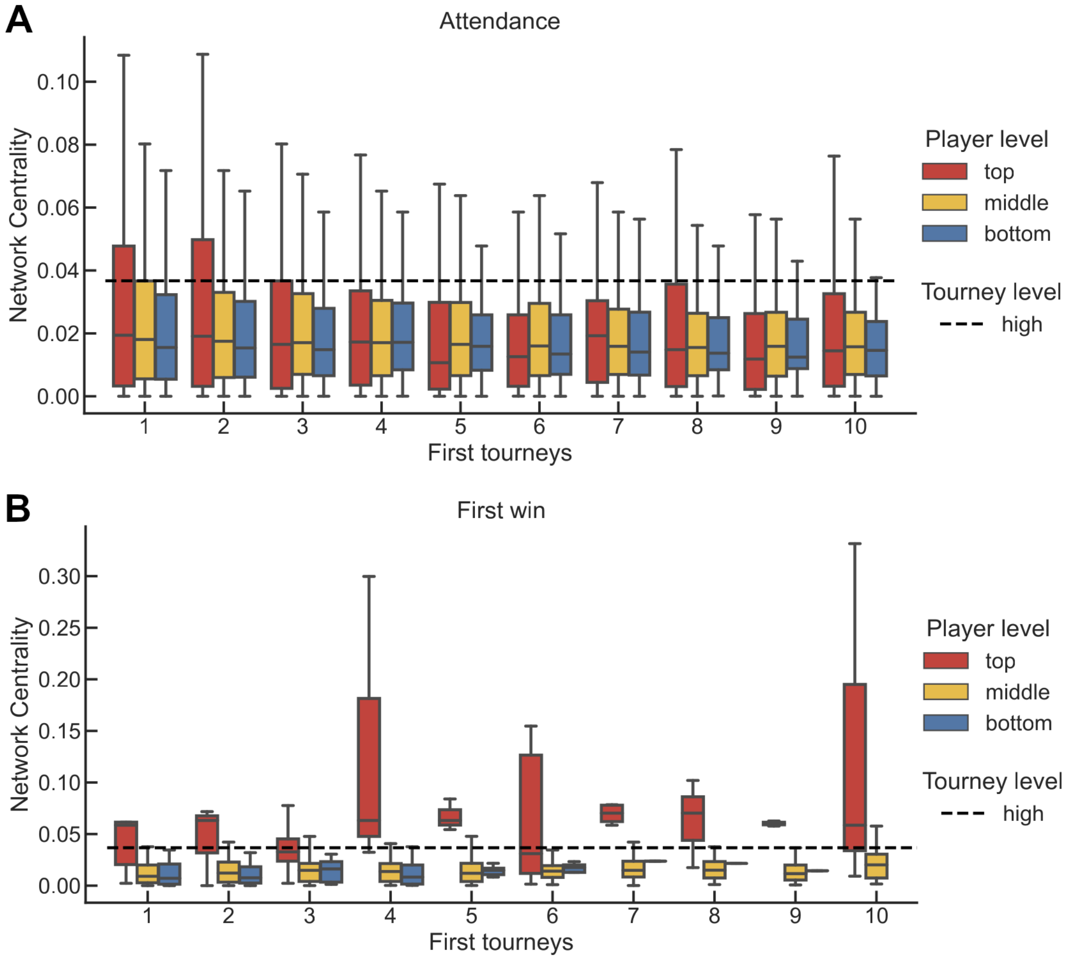

Figure S5. Centrality distribution for each of the first ten tournaments, divided by player level. **A** Boxplots of the eigenvector centrality of the tourneys that players attend (grouped by their career peak). The average level of tourneys is below the high centrality threshold (dashed line). **B** Boxplots of the eigenvector centrality referred to players' first match win. Only top players consistently cross the high-level threshold (dashed line).

from the middle/bottom players. Note that in Fig. S7 we did not report the level of tournaments attended because it does not change, compared to the one shown in the main text (Figs. 5C-E).

Lastly, we check whether a relationship exists between the increment of points owned by players after they won their first match and the centrality of the tournament in which it happened. In Fig. S10 we observe that no clear trend emerges when we consider the players' points by the time they won their first match, while the eigenvector centrality allows us to clearly distinguish top players (the majority of whom are above the high-level threshold, highlighted by the dashed line) between both middle (only a small number is above the high-level threshold) and bottom players (almost none of them overcome the threshold). Even if we compute the Spearman's coefficient we find a very weak correlation between the tournament centrality and players' increase in points after they won their first match ($r_s = 0.25^{***}$).

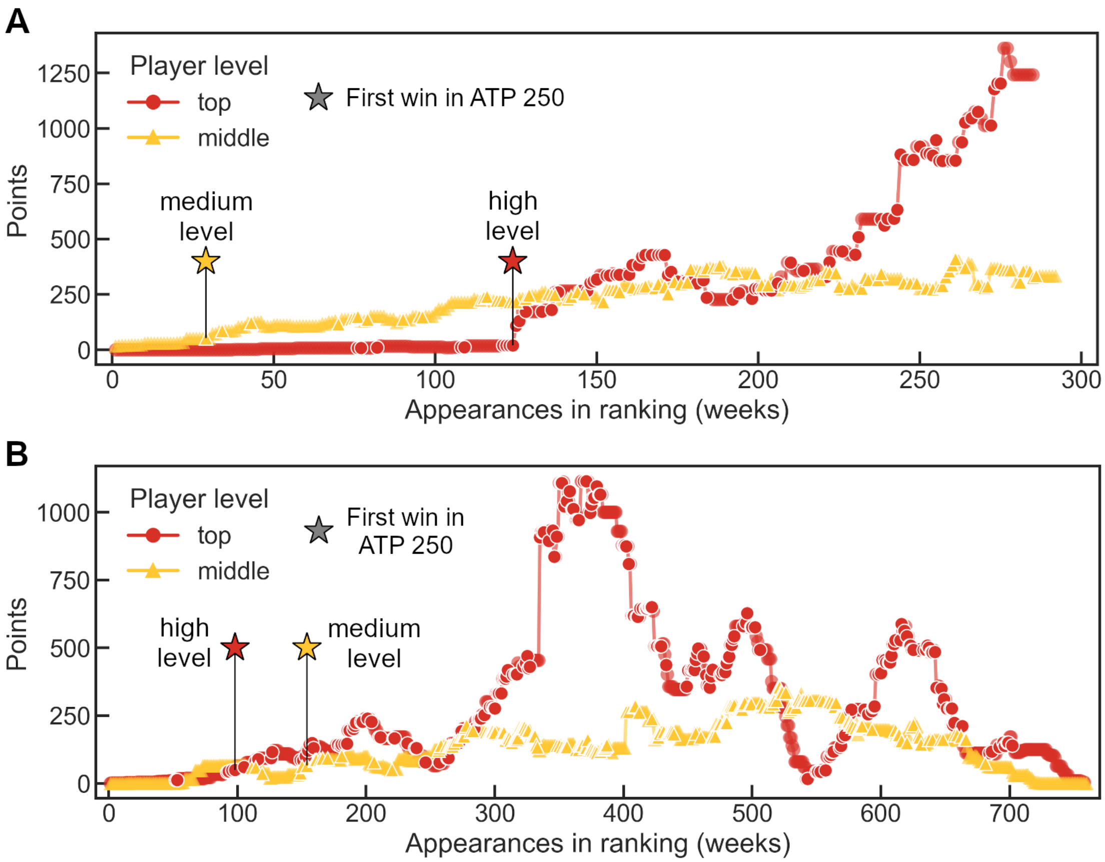

Figure S6. Career progression for two distinct pairs of top/middle players after their first win in high/medium level tourneys of the same ATP category (ATP 250). **A** The careers of two still active players, Nikola Milojevic (middle, yellow triangles) and Reilly Opelka (top, red dots). Those weeks during which they attended a tournament present in our data are highlighted by the white-edged red dots (top) and white-edged yellow triangles (middle); a star marks the week of their first match win (red for top and yellow for middle), and the corresponding (centrality-based) tourney level is annotated next to it. Notice that Nicola won his first match much earlier than Reilly, and had more ranking points by then (respectively, 52 and 20). **B** The complete careers of two players, Sam Warburg (middle, yellow triangles) and Andreas Beck (top, red dots). Those weeks during which they attended a tournament present in our data are highlighted by the white-edged red dots (top) and white-edged yellow triangles (middle), and a star marks the week of their first match win (red for top and yellow for middle). Looking at their appearances, Sam won his first match later than Andreas, but he had more points in the ranking (75 and 50, respectively).

REFERENCES

[1] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[2] Marion Guillaume, Stephane Len, Muriel Tafflet, Laurent Quinquis, Bernard Montalvan, Karine Schaal, Hala Nassif, François Denis Desgorces, and Jean-François Toussaint. Success and Decline. *Medicine & Science in Sports & Exercise*, 43(11):2148–2154, nov 2011.

[3] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.

[4] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[5] Mark Newman. *Networks*. Oxford university press, 2018.

[6] Official site of men's professional tennis - ATP tour. `https://www.atptour.com/`.

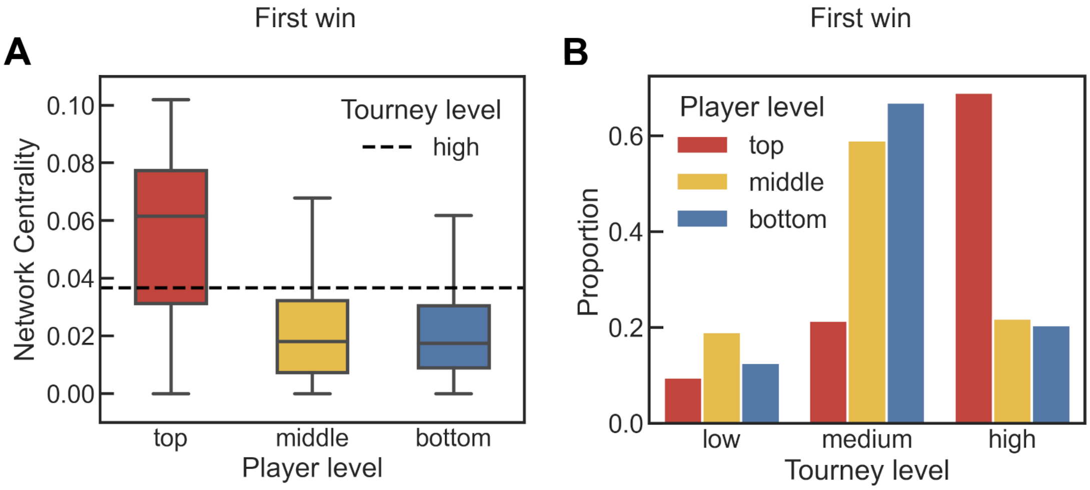

Figure S7. First win with qualification rounds. Both panels are based on the level of tournaments in which players have their first match win within the first ten attended competitions, and the top players (red) show distinct behavior compared to the others. **A** Distribution of the centrality of the first win for each group of players. The dashed line identify the threshold of high-level tourneys. **B** Fractions of players who have their first win in a tournament of a given level. Bars of the same color, each identifying a given group of players, add to 1.

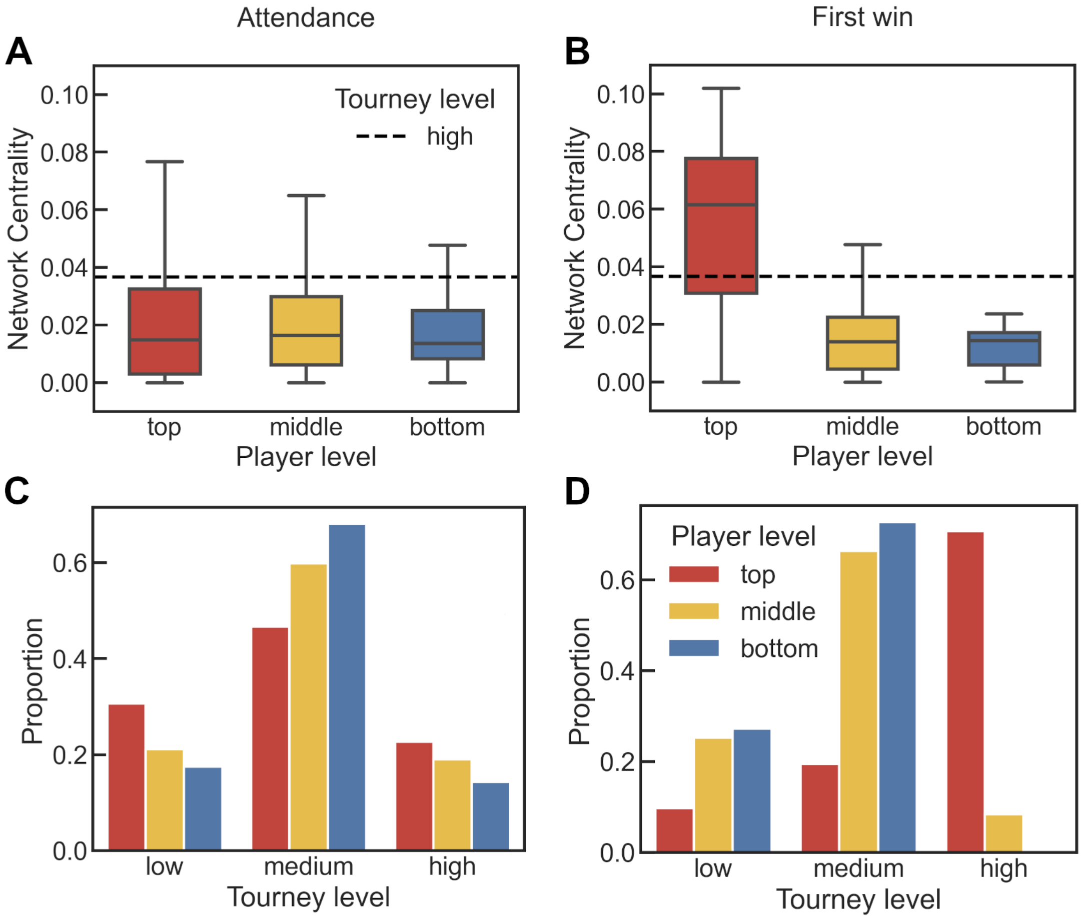

Figure S8. Attendance and first win, considering players with more than ten tournaments in our dataset. The panels on the left are based on the level of tourneys that the players attend, and no significant differences emerge among the groups. The panels on the right are based on the level of tourneys where players have their first match win, and the top players show distinct behavior compared to the others. **A** Distribution of the centrality of the first ten tournaments for each group of players. **B** Distribution of the centrality of the first win for each group of players. In panels A-B, the dashed line identifies the threshold of high-level tourneys. **C** Fractions of players who participated in a tournament of a given level within the first ten competitions. **D** Fractions of players who have their first win in a tournament of a given level within their initial ten competitions. In panels C-D, bars of the same color, each identifying a given group of players, add to 1.

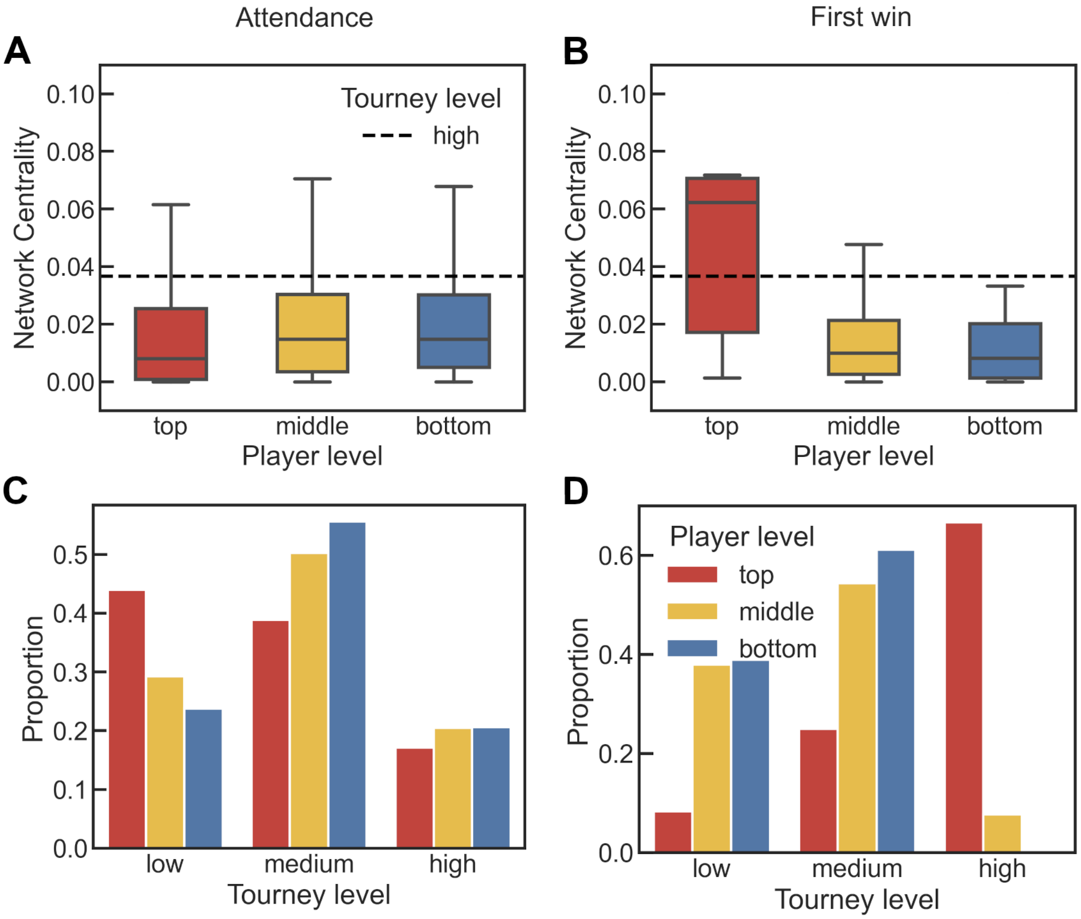

Figure S9. Attendance and first win, excluding players that were still active at the end of our dataset. The panels on the left are based on the level of tourneys that the players attend, and no significant differences emerge among the groups. The panels on the right are based on the level of tourneys where players have their first match win, and the top players show distinct behavior compared to the others. **A** Distribution of the centrality of the first ten tournaments for each group of players. **B** Distribution of the centrality of the first win for each group of players. In panels A-B, the dashed line indicates the threshold of high-level tourneys. **C** Fractions of players who participated in a tournament of a given level within the first ten competitions. **D** Fractions of players who have their first win in a tournament of a given level within their initial ten competitions. In panels C-D, bars of the same color, each identifying a given group of players, add to 1.

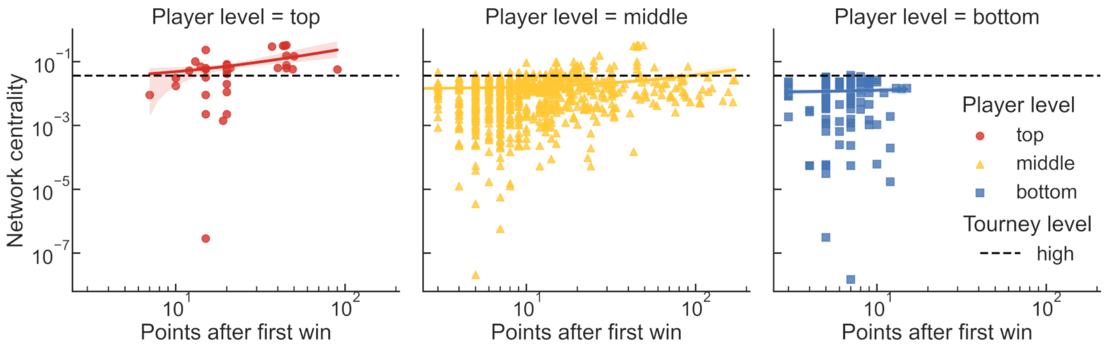

Figure S10. Scatterplot of players' ranking points and the centrality of their first win, split by players' group (red dots for top, yellow triangles for middle, blue squares for bottom). In each subpanel, a regression line is reported for every group (Spearman's coefficients: $r_{\text{top}} = 0.48^{*}$, $r_{\text{middle}} = 0.21^{***}$, $r_{\text{bottom}} = 0.08^{\text{ns}}$,), while the dashed line indicates the threshold of highly central tournaments (based on the tourney splitting explained in the manuscript). Overall, no clear trend emerges when we consider players' increase in points after they won their first match, except for a weak correlation for top players. Instead, network centrality allows us to distinguish top players (the majority of whom are above the high-level threshold, highlighted by the dashed line) between both middle (only a small number is above the high-level threshold) and bottom players (almost none of them overcome the threshold).