



# Early career wins and tournament prestige characterize tennis players' trajectories

Chiara Zappalà<sup>1,2,3,4\*</sup> , Sandro Sousa<sup>3,4†</sup> , Tiago Cunha<sup>3†</sup> , Alessandro Pluchino<sup>2</sup> ,  
Andrea Rapisarda<sup>2,5</sup>  and Roberta Sinatra<sup>4,3,5,6,7</sup> 

\*Correspondence:

[chiara.zappala@uni-corvinus.hu](mailto:chiara.zappala@uni-corvinus.hu)

<sup>1</sup>Center for Collective Learning, Corvinus Institute for Advanced Studies (CIAS), Corvinus University, Budapest, 1093, Hungary

<sup>2</sup>Department of Physics and Astronomy, University of Catania and INFN sezione di Catania, 95123, Catania, Italy

Full list of author information is available at the end of the article

<sup>†</sup>Equal contributors

## Abstract

Success in sports is a complex phenomenon that has only garnered limited research attention. In particular, we lack a deep scientific understanding of success in sports like tennis and the factors that contribute to it. Here, we study the unfolding of tennis players' careers to understand the role of early career stages and the impact of specific tournaments on players' trajectories. We employ a comprehensive approach combining network science and analysis of the Association of Tennis Professionals (ATP) tournament data and introduce a novel method to quantify tournament prestige based on the eigenvector centrality of the co-attendance network of tournaments. Focusing on the interplay between participation in central tournaments and players' performance, we find that the level of the tournament where players achieve their first win is associated with becoming a top player. This work sheds light on the critical role of the initial stages in the progression of players' careers, offering valuable insights into the dynamics of success in tennis.

**Keywords:** Network science applications; Success; Sports analytics

## 1 Introduction

Understanding the complex mechanisms behind the origin of success is a challenging task that has captured the attention of researchers in recent years, as it encompasses a wide range of systems. To mention some examples, paper citations [1–3] and funding [4] in science, start-ups [5], show business [6], art [7] and cryptoart [8, 9] ecosystems, music [10–12], and other creative careers [13, 14], have been investigated to date.

A common issue in these systems is to unambiguously distinguish between performance and success [15]. Whereas performance refers to objectively measurable accomplishments in a particular field [16], such as the publication record of a scientist [17], success represents the reward of a given level of performance [18], intended as its collective recognition, such as the citation impact in science [1, 18] or prize and awards in fields like music [11]. However, assessing the impact of creative work only through prizes and fame might fail to consider the abilities of the individuals involved, that is, to disentangle performance from success [15].

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Sports allow us to overcome this issue. First, they offer objective metrics for evaluating performance, like the winning record of an athlete or a team [19]. Most importantly, successful players are identifiable by the reward system of the sport itself, i.e. rankings based on score systems, especially in individual sports based on knockout tournaments. In fact, sports rankings depend on criteria that are external to the athletes' performance, i.e., the quality of a tournament (also called *tourney*) has an *a priori* fixed value and points are distributed accordingly to the round reached in it [20]. Therefore, unlike the previous literature on sports [16], here we consider the ranking as a metric of success that is inherently provided by sports rules, not determined by the popularity of players.

Although few works have analyzed sports disciplines from a complex systems perspective [16, 21–24], the determinants of successful careers in sports remain elusive. Particularly, we lack a systematic analysis of the impact of early career stages on players' future achievements, despite the proven importance of these initial stages in many different kinds of careers [25]. Often, we imagine the top players as predestined champions who need to be extraordinarily talented and hard-working to get to the top [26]. Yet, evidence suggests that a combination of talent and effort does not guarantee success [27, 28]. Rather, some initial fortuitous events might play a role in shaping the evolution of top players' careers, as shown for individual sports [23, 24, 26]. The role of chance at the early stages can be later amplified by a cumulative advantage dynamic [29]. The aforementioned elements provide compelling reasons to delve into the trajectories of players' careers, i.e., the temporal sequences of the competitions they attended.

Here, we focus on tennis and aim to analyze the key factors behind top players' success at the beginning of their careers. Specifically, we analyze the career progressions of professional male players between 2000 and 2019. We collect data from the official rankings of the Association of Tennis Professionals (ATP) [30], along with the results of matches from various tournaments [31]. The top tennis athletes are identified by their career peak, which is determined by the highest number of ranking points they have achieved in the ATP rankings. Through our analysis, we observe distinctive characteristics among accomplished players compared to others, including longer career spans and a pattern of consistently higher ranking points throughout their career's initial stages.

We hypothesize that the rise of top players in tennis is associated with their performance in high-level competitions early on in their careers. This phenomenon is akin to the success of well-known artists who gain recognition from showcasing their initial work at prestigious institutions [7]. In fact, the prestige of the first venues in which artists perform is crucial to their future success, as the same artwork may receive different responses from the audience based on the prestige of the institution where it is first exhibited. Similarly, players with comparable performance in more (less) relevant events may get more (less) attention from the rest of the tennis community. To test our hypothesis, we introduce a novel approach to quantify the level of ATP tournaments, which not only includes their historical prestige but also takes into account the participation of players. This method, based on network science principles, presents a contribution that, to our knowledge, has not been explored in the existing literature on tennis. Previous studies have used networks solely to analyze match relationships [20, 26, 32, 33]. We expand upon this by constructing a network of co-attendance among tennis tournaments, where nodes represent *tourneys*, and links are created based on players' trajectories, that is, a link connects two *tourneys* if there is at least one player who competed in these two *tourneys* during his career. Conse-

quently, we can establish connections between competitions that may be geographically distant or temporally separated. By leveraging this co-attendance network, we derive a measure of tournament prestige using eigenvector centrality [34], following the methodology of Ref. [7].

In the following sections, we will show that the level of the tourney where players secure their first match win allows us to characterize future successful players. First, we group career trajectories and analyze the difference between bottom, middle, and top players, focusing on the initial stages of their careers. Second, we identify highly central tournaments using the constructed co-attendance network. We then associate the tourney level with players' performance, assessed by their first match win, and we find a relation with top players' trajectories. Finally, we check the robustness of our findings using two distinct null models.

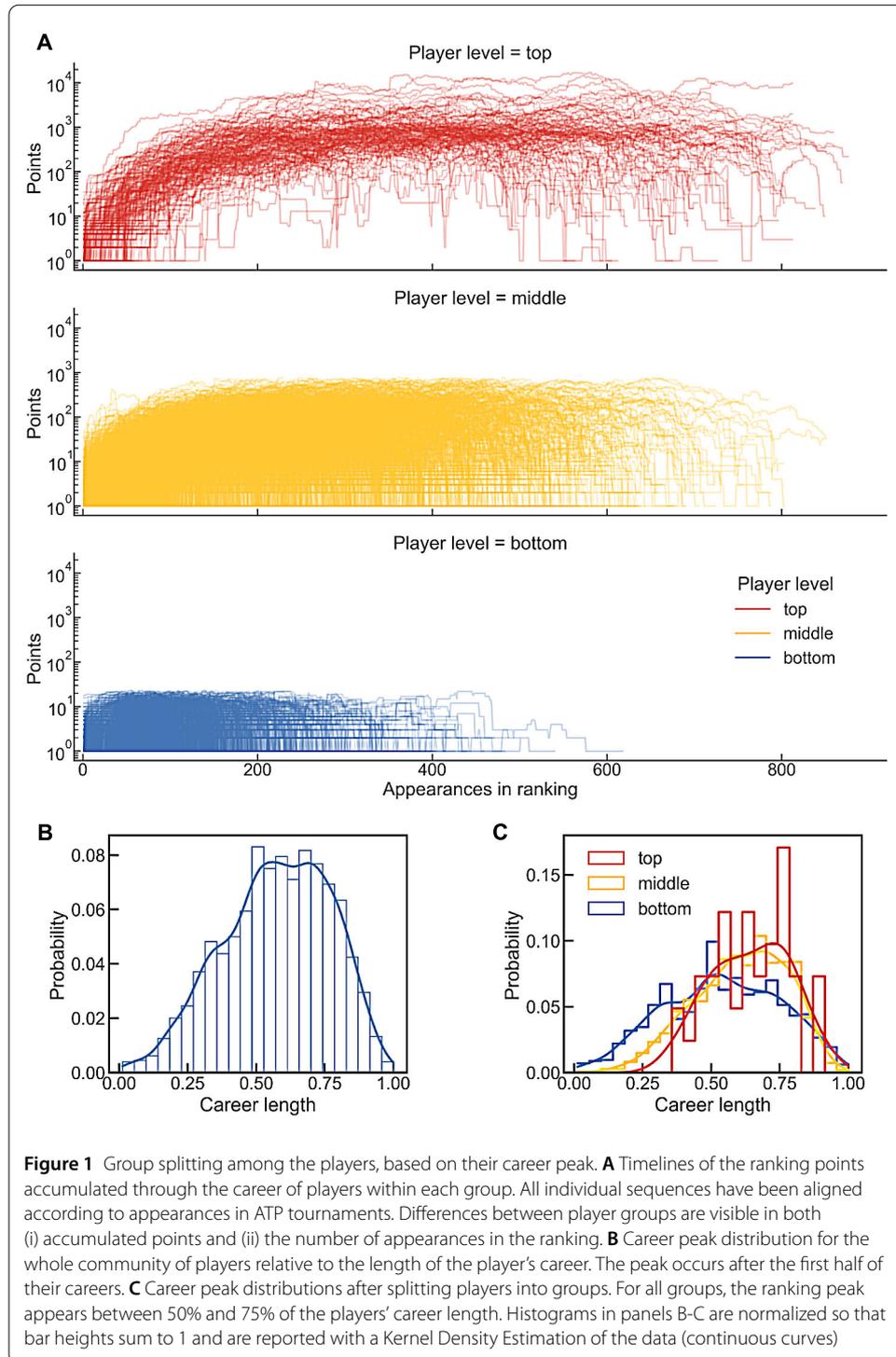
## 2 Results

### 2.1 Characterizing patterns in tennis careers

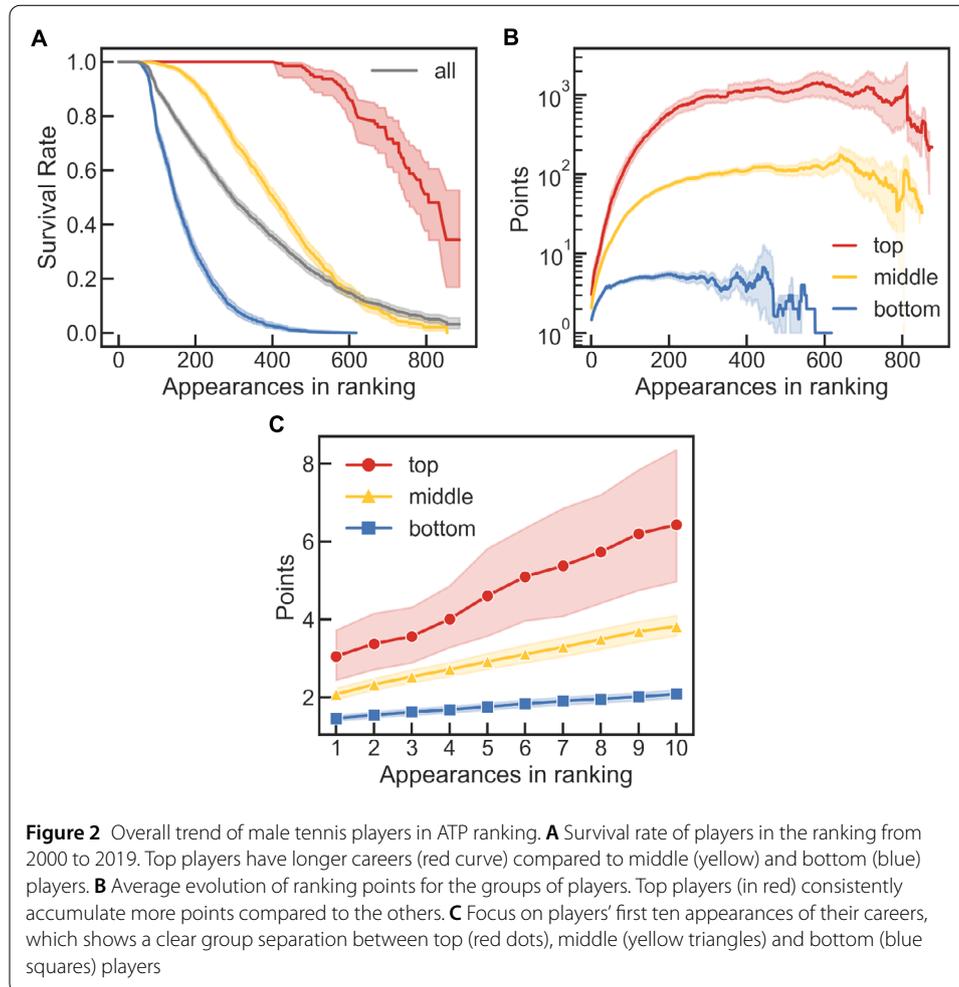
We study the evolution of the careers of male professional tennis players from 2000 to 2019. We obtained data from the official ranking of the ATP [30, 31], together with the match results of different tournaments: Grand Slam (the competitions with the highest value in terms of winner points), Masters 1000, ATP 500 and 250, Challenger (the competitions with the lowest value in our dataset) [31]. We selected players who started their careers in the timespan of our dataset and had at least two years of activity. We consider 3455 players and 651 tourneys, specifically 4 Grand Slams, 11 Masters 1000, 98 ATP 500 and 250, and 538 Challengers.

To distinguish between top and less successful tennis players, we group them according to the maximum amount of points they reached in the ATP ranking, which ranks players based on the score points they accumulate during a season [30]. We can conceive the ATP ranking as a first proxy of success, as it might weigh similar outcomes of players' performance, which would be winning or losing one or more matches, in very different ways. For instance, winning a match in the round of 32 awards 5 points in a Challenger and 90 points in a Grand Slam. Thus, rather than relying on popularity to quantify success in sports [16], we explore the dynamics of success embedded in tennis rules, neglecting the influence of subsequent elements such as prize money, income, popularity, sponsors, etc. Moreover, we use the highest number of points players reach in the ATP ranking (namely, their career peak) instead of ranking placements. The reason is that the point totals of players with consecutive ranks can vary significantly. For example, consider three players ranked 1, 2, and 3, with point totals of 12,000, 10,000, and 9995, respectively. Although players ranked 1 and 2 are only one position apart from each other, as well as players ranked 2 and 3, there is a greater point difference (2000 points) between players ranked 1 and 2 than between players ranked 2 and 3 (5 points). Therefore, using point totals instead of placements allows us to assess differences between players more accurately. Also, it lets us compare rankings with varying numbers of players over the years.

We split male tennis players into three groups: Top players are defined by those with a career peak above the 95th percentile (top 5%); bottom players are within the 40th percentile (bottom 40%); the 55% left composes the middle group. Individual timelines of players within each of these three groups and their respective ranking points are reported in panel A of Fig. 1. Because players can start their careers at different times, we aligned

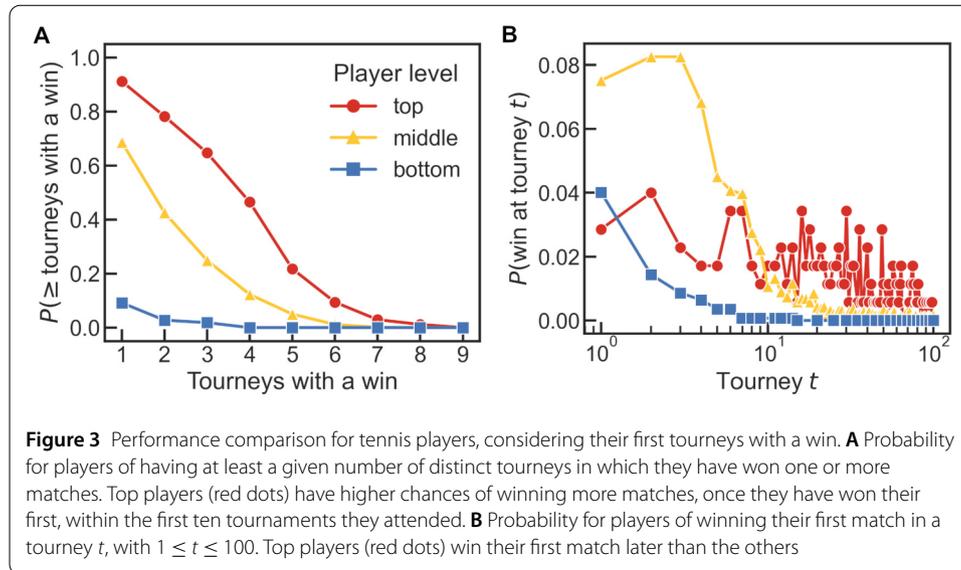


their trajectories by the first appearance in the ATP rankings. Thus, a player's trajectory is a time series composed of the sequence of ATP ranking updates. These updates occur weekly, with the exception of the weeks of the Grand Slams and few other tournaments which last longer (consequently, the ranking is run approximately 45 times a year) [30]. As long as a player is still active (i.e., an ATP professional player), he will appear in the ranking.



Our analysis is based on the peak of professional tennis players' careers, prompting the question of whether this peak is obtained at a consistent time across all individuals within our dataset. To answer this question, we look at the time distribution of career peak, first considering the aggregated data, then each group separately (respectively, panels B and C of Fig. 1). To avoid right-censoring bias [35], we exclude active players from Fig. 1B-C (more details are provided in the Supplementary Information, where we report the case without the right-censoring correction in Figure S1, see Additional file 1). To deal with different career lengths, we normalize the time of the career peak of each player according to their career duration. We find a common tendency for the peak to occur after the first half of players' careers in all three groups. This result, previously observed only for the top players [36, 37], suggests that peak time is not closely related to individual success.

Observing the individual sequences of the three groups in Fig. 1A, we notice marked differences between them, both in ranking appearances and accumulated points. In particular, the bottom players have shorter careers compared to the other groups. We further investigate this by looking at the survival rate [38] of tennis players in our dataset, bearing in mind that in this context "surviving" at time  $t$  means still playing or, in other words, being in the ATP ranking. The results are shown in panel A of Fig. 2. The bottom players' survival curve (in blue) is the shortest and goes rapidly to zero (decay starts before 100 appearances), followed by the middle players' (in yellow), which starts decaying a few



rank updates later but at a slower rate. Conversely, the top players' curve (in red) starts falling much later (at around 400 rank updates) and at a slower decay rate, meaning that they have longer professional careers, in line with previous work [39]. We also reported the survival rate of all players (in gray) for reference.

To highlight when the group differences in accumulated points appear, we take the average of the sequences shown in Fig. 1A, which results in the trend of Fig. 2B: We can observe that the top players have more ranking points compared to the others. Such a discrepancy in the number of points could be interpreted as a mere artifact of our definition of top/middle/bottom players. Yet, this difference emerges from the beginning, as indicated in panel C of Fig. 2, which zooms in on the points cumulated in only the first ten appearances of a player in the ATP ranking. Note that here we consider all players, thus including active players. See Figure S2 of the SI for an analysis that considers only those players who started and ended their careers in the dataset, checking for the eventual effects of right-censored data on Fig. 2B.

The initial gap in the average amount of points between the top players and the others may arise from different mechanisms. A first explanation for such a gap may lie in the differences in players' performance. That is, top players may win more matches from the early stages of their careers, leading to the gap forming. To compare performance across the groups, we first consider the number of competitions in which a player wins at least one match. Panel A of Fig. 3 shows the probability  $P$  that a player, with at least one match won, reports a win in more than a given number of tournaments, within the first ten (see Methods for the mathematical definition). Top players (red dots) have higher chances of winning more matches, once they have won their first, at the beginning of their career. However, if we look at the probability  $P$  of players winning their first match in a certain tournament  $t$  after they turned professional (Fig. 3B, see Methods for the mathematical formulation), we do not see top players emerge. On the contrary, top players tend to win their first match later than the others.

The results of Fig. 3 show that, even though top players achieve more victories after their initial one, they have difficulty in winning their very first match during the early stages of their career. This counterintuitive behavior points out that players' performance is not

enough to explain the formation of the gap between top and less accomplished athletes. Hence, other mechanisms might be at play. For instance, our analysis so far has not taken into account the prestige of the different tournaments that players can attend at the beginning of their careers. Therefore, having illustrated the scenario of the initial stages of players' careers in men's professional tennis, we investigate the influence of the first tournaments they can access, together with their results. We aim to untangle the importance of early participation in more prestigious competitions from how players perform in those competitions.

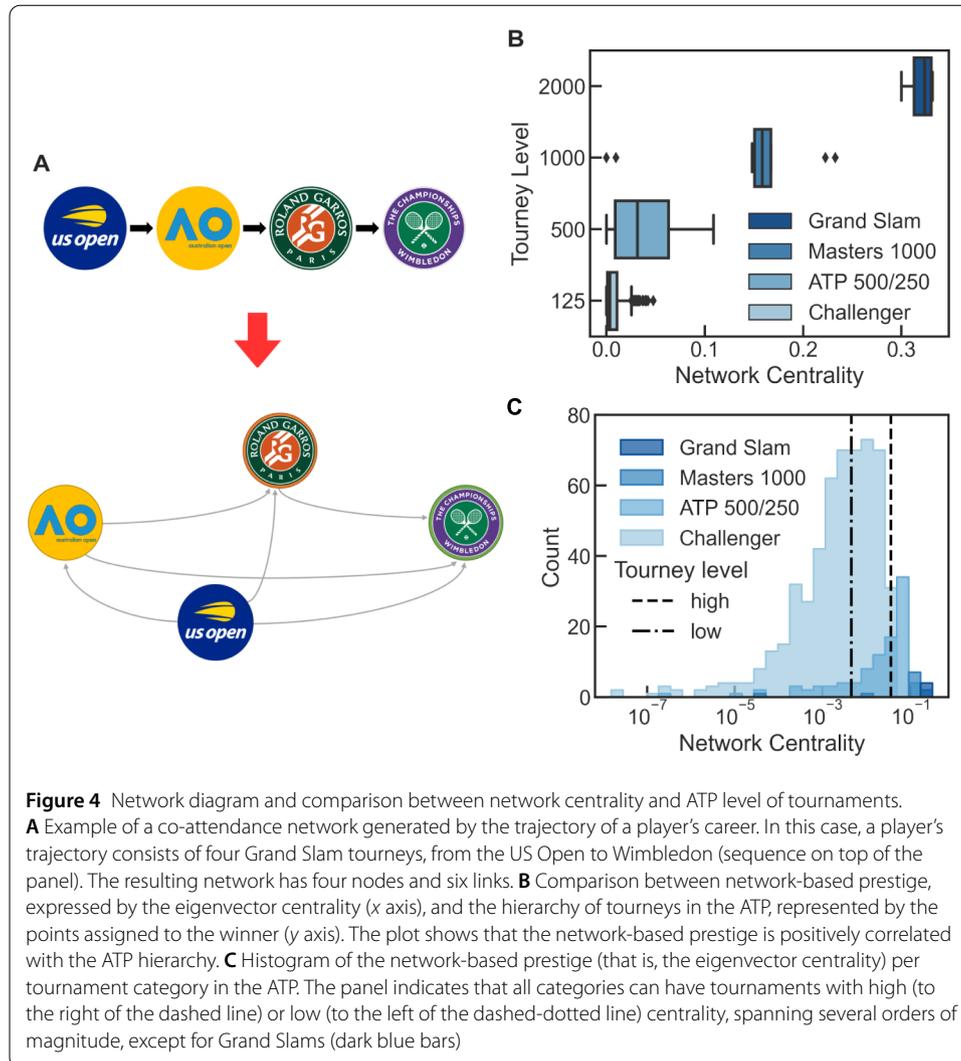
## 2.2 The co-attendance network

Early access to prestigious tournaments could affect the career trajectories of players in the ATP circuit in a non-trivial way. Those trajectories then create complex relationships between players and tourneys. Indeed, players do not have the possibility to participate in all available tournaments and choose which tourney to sign up for based not only on their own skills but also on the characteristics of tourneys themselves (e.g., the court), their relevance during a season (preceding or succeeding famous events, for example), and their prestige. One could quantify tournament prestige from their prizes in terms of awarded points. Yet, assessing the tourney level based only on prizes does not capture the prestige perceived by the players and determined by their choices. Following Ref. [7], we propose a new method to assess the level of a certain ATP tournament, which not only captures its historical prestige but also includes the reciprocal influence of players and the reputation of a given competition. More importantly, this method does not require any previous knowledge about the tournaments or their prize points. We define a network where nodes are ATP tourneys and links depend on players' careers. A directed link  $(i, j)$  is created when a player first competes in tourney  $i$ , then in tourney  $j$ , and is weighted by the number of players who have the same attendance sequence [7] (see panel A of Fig. 4 for an example). Note that we connect tourney  $i$  to all consecutive tournaments attended by a player and not only to the tourney attended immediately after  $i$ . Thus, we consider the effects of all the competition choices that players made during their careers in the data. In this way, the movements of players link competitions far in space and time, and recurrent movements signal that those competitions tend to co-occur in players' careers.

The resulting network has 659 nodes and 255,055 edges. We focus on the largest strongly connected component of the original network, which has 651 nodes and 254,583 links; from now on, we refer to the largest component as our network (see Table S1 in the SI for more details on the features of the network, such as its density and clustering coefficient).

From the co-attendance network, we can extract a measure of tourney prestige that correlates with the importance of tournaments in terms of their points. The prestige of a tournament can be derived from the topology of the network, using the eigenvector centrality [7, 40] (see also Methods for the mathematical definition we used). This definition captures the historical level of the competitions (Fig. 4B), expressed by the maximum number of points assigned to each tournament category (the allocation of points awarded per tournament is explained in the Methods section and summarized in Table S2 of the SI). Note that identifying tournaments based on their awarded points still remains a categorical definition, as tourneys belonging to the same categories award the same points in each round, in general (see Table S2 of the SI).

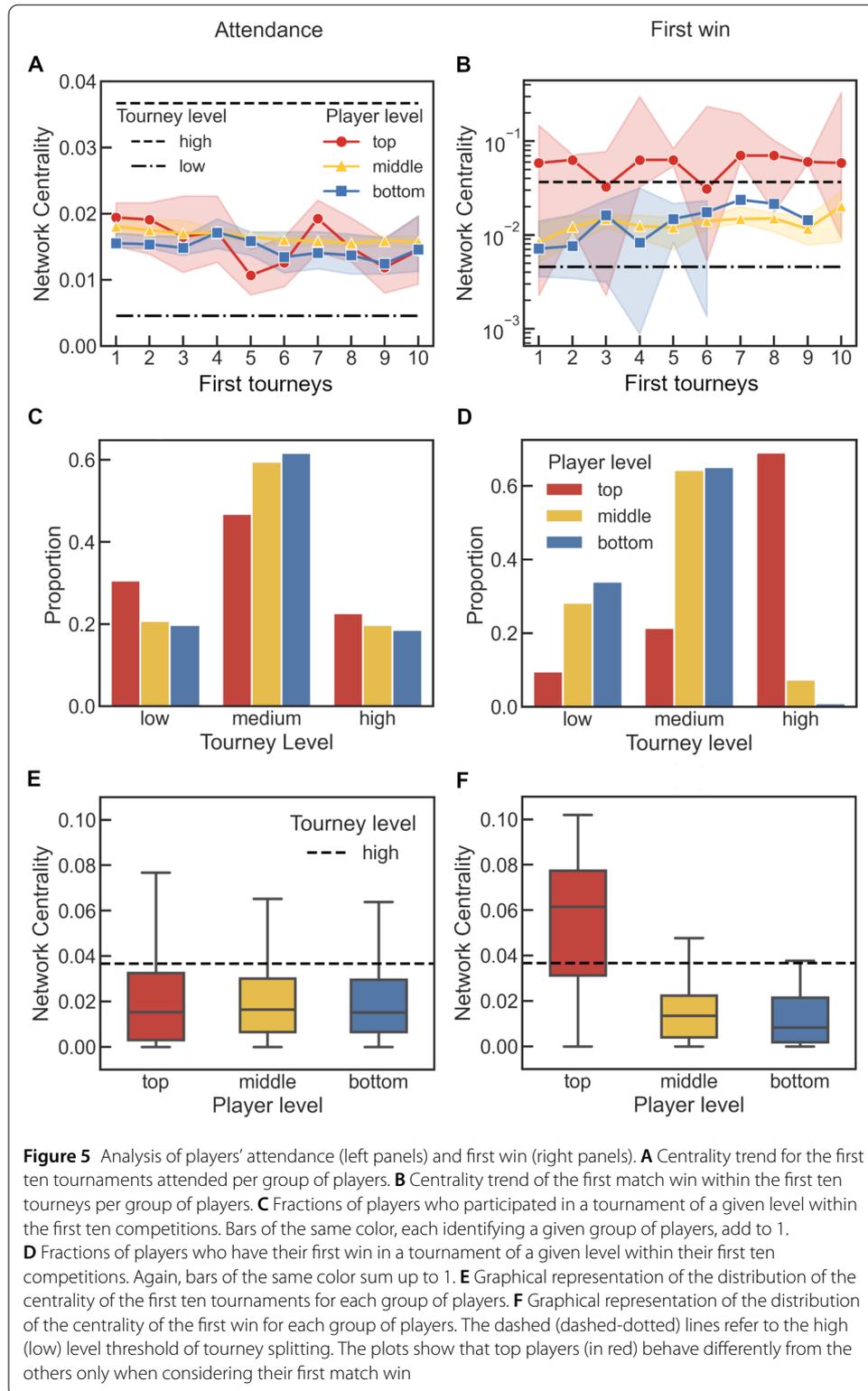
We divide competitions into three groups based on their centrality: The most prestigious tournaments are in the top 10% (above the 90th percentile), and we refer to them as



high-level tournaments; the bottom 50% (below the 50th percentile) of the competitions are labeled as low-level tournaments; the others fall into the medium-level group of events. Tournaments belonging to the same category in the ATP can have vastly different centralities, suggesting that the network topology allows a fine-grained distinction between them that cannot be obtained by looking at categories only (see Fig. 4C). Having defined the level of the tournaments, we now focus on the possible connections between those levels and the success of players in the ATP circuit. These connections are crucial to determine whether the opportunity of competing in a given tournament could be more relevant than, or at least as relevant as, players' abilities.

### 2.3 Early access to prestigious tournaments and the impact of the first win

We check the possible differences in tournament attendance within the first ten tournaments of athletes' professional careers on the ATP circuit (left panels of Fig. 5). First, in Fig. 5A, we observe the eigenvector centrality of the first ten tournaments for each group of players based on their career peak. We find that, at the beginning of their career, players attend competitions with comparable centrality, having median values in between the thresholds (dashed lines) of tournament splitting (we consider the median due to the



asymmetric distribution of the centrality in our network, as shown in Figure S3 of the SI). Only after a considerable number of tournaments do top players attend only high-level tournaments, which means that they consistently participate in events having a central position in the co-attendance network (see Figure S4 in the SI). Then, we inspect the fraction of

players who enter a certain tourney of a given level at the beginning of their career (panel C of Fig. 5). We do not observe a pronounced prevalence of future top players in high-level competitions (red bars in Fig. 5C).

Interestingly, we find no significant differences in the prestige of tournaments players can access when their careers start. One might argue that the seasonality of tournaments plays a role, hence affecting the centrality of players' first competitions: In other words, if the centrality of the first tourneys of the season is around the median value, we should expect the trend observed in panel A of Fig. 5. Nonetheless, professional players can start their career on the ATP tournament circuit at any time during a competitive season, coincident with the calendar year. Therefore, the centrality of the opening tournaments of the season (in other words, the tournaments organized in January/February) does not determine the entire trend of Fig. 5A. Consequently, players' first attended tournaments can vary widely from athlete to athlete. It is also worth mentioning that we neglect the influence of the junior circuit on players' professional development. According to some studies [41–43], the youth career could impact the future success of an athlete in tennis. Even if that impact were not a prerequisite for professional success [37, 44], young players who performed well at the junior level could be favored to access more prestigious ATP venues. However, such an effect, if present, does not create a significant gap among players in terms of the level of the first attended tourneys. We specify that we do not differentiate players by age or other factors like country of origin or physical characteristics (e.g., height, left-handed or right-handed, etc.).

Since no patterns emerge when looking at tournament attendance, one might ask if there are differences related to performance. In our framework, tennis performance is expressed by the outcome of matches. Thus, we check for patterns linking the victory of matches and the start of players' professional careers in the ATP circuit. To do so, we focus on the first win of a match at the beginning of tennis players' careers. Specifically, we are interested in the first victory in the main draw (i.e., the starting lineup of a tourney after the qualification rounds) of the first ten tournaments they attended. We consider the first match win because it allows us to directly compare the outcome of players' performance for all the competitors. To visualize the relationship between the first win and the tourney level, in terms of centrality, we refer to the right panels of Fig. 5. In Fig. 5B, the eigenvector centrality of the top players is the only one above the threshold of high-level competitions. Note that the trend remains constant irrespective of the time of the first win, indicating that there is no distinction between winning earlier (within the first five tourneys) or later (after the sixth tourney). We find that most of the top players (around 70%) have their first win in the main draw of a high-level tournament (Fig. 5D). Furthermore, only top players can be identified by looking at the prestige of their first match win. Both middle and bottom players have similar behavior (Fig. 5D), and their first victory rarely occurs in high-level competitions.

To better understand the discrepancy in the behavior of top players, either when we consider only their attendance or when we add their first win, we compare the boxplots of the centrality of the players' first tournaments with that of their first win (see panels E and F of Fig. 5, respectively, while a fine-grained visualization is available in the SI, Figure S5). In this way, we observe a clear difference between the two situations. In Fig. 5E, there is no distinction between the top, middle, and bottom players, with respect to the network-based prestige of the first tournaments they attended. Panel F of Fig. 5, instead, shows that

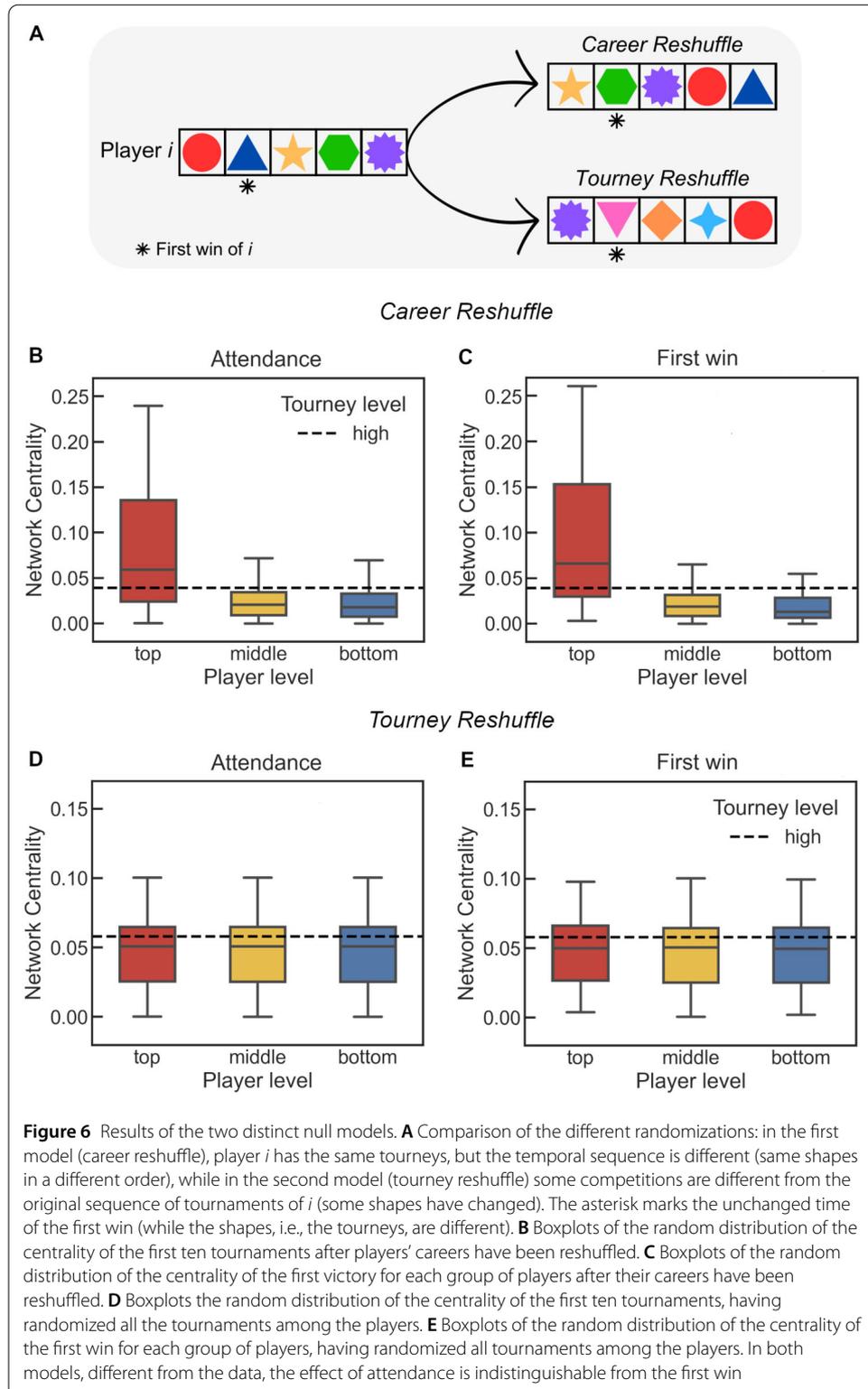
the average centrality of the first win of the top players crosses the threshold of high-level tournaments. In particular, the top players' boxplot is the only one that changes from panel E to F, which means that the higher prestige of the top players' first win cannot be explained by the average level of the first attended tournaments. Panels E and F of Fig. 5 highlight the relationship between the prestige of the tournament in which the top players win their first match in a main draw and their future success. To further validate our analysis, we computed the correlation between (1) the player's career peak and the median centrality of the first tournaments they played; and (2) the player's career peak and the centrality of the tourney where they got their first win in a main draw. We use Spearman's correlation coefficient and find that  $r_{s,1} = -0.008^{\text{ns}}$  and  $r_{s,2} = 0.47^{***}$ , where <sup>ns</sup> and <sup>\*\*\*</sup> indicate the level of significance of the two values, that is, the p-value is greater than 0.05 (not significant) and less than 0.001 (significant), respectively.

Whether comparing players in groups or directly through their maximum number of points, we conclude that the prestige of the tournament where they first win a match in the main draw is a revealing factor for the future career of male tennis players. In the Supplementary Information, we show two examples of individual career progression before and after their first win, each time comparing a middle and a top player having won their first match in a tournament of the same ATP category but different centrality level (Figure S6). It should be noted that taking the qualification rounds into account does not appreciably change our findings (see Figure S7 of the SI). Furthermore, we do not assume that players should attend at least ten tournaments to be in the dataset, and we do not exclude active players, but adding these constraints does not significantly alter our results (see Figures S8-S9 in the SI). Lastly, we check the eventual relationship between ranking points and tournament centrality. We observe that the increase in ranking points that players had a week after winning their first match is only weakly correlated either with the centrality of the tourney in which they had their first win or with their future success (see Figure S10 in the SI).

## 2.4 Significance of the results

To assess the significance of our findings, we build two distinct null models for the network of co-attendance of ATP tournaments. Building on previous work [1, 7], we proceed as follows (Fig. 6A): In the first model, we reshuffle the careers of each player individually so that the events they play are the same but have a different temporal order; in the second model, we reshuffle all the competitions attended by the players, so that each player's career has the same number of events, but it can consist of different tournaments. In both cases, all temporal correlations are destroyed. We choose these two randomizations because they focus on different aspects: The first randomization preserves not only the number of competitions but also the actual events players attended; the second randomization preserves the number of tournaments of each player and the distribution of competitions among all players, destroying, however, the possible player-tourney correlations.

We repeat these two randomizations multiple times to create an ensemble of 500 random realizations. We specify that in both configurations we preserve the information about the time of the first win as given by the data. Therefore, the randomizations would only change the corresponding tournament in the sequence, but not *when* the first win of a player occurred (as illustrated by the asterisk in Fig. 6 A). For each realization, we build the correspondent co-attendance network and evaluate tournament centrality, considering the prestige of competitions as done in the data.



We analyze the average distributions of tournament centrality per group of players, thus keeping the possible relationships between players' success and prestige of their initial ten competitions. We follow the order of tourney attendance and define tourney levels based on their importance in the null models. Panels B to E of Fig. 6 show that the null

**Table 1** Spearman's correlation coefficients for the null models, related to players' participation and first win, within the initial ten tournaments. We also report the correlation coefficients found in the data for comparison

Null model	Attendance	First win
	$\bar{r}_{s,1} \pm \sigma_{r_{s,1}}$	$\bar{r}_{s,2} \pm \sigma_{r_{s,2}}$
Career reshuffle	0.182 ± 0.005	0.14 ± 0.03
Tourney reshuffle	0.001 ± 0.007	0.01 ± 0.03
Data	-0.008 <sup>ns</sup>	0.47***

models cannot reproduce at the same time both the prestige of the tournaments attended and that of the first match win in the early stage of top players' professional career in tennis (Fig. 5E-F).

The reshuffle of the individual sequences of tournaments per player increases the gap between top and middle-bottom players: given the cyclic nature of individual sports, where competitions repeat themselves every year around the same week, players are encouraged to attend the same tourneys season after season, to preserve or improve their amount of points. It follows that reshuffling the careers of top players only anticipates those tournaments they start to play once they have already reached the top. In contrast to empirical data, consequently, top players tend to compete more in high-level tournaments from the beginning of their professional careers, so that they are more likely to win their first match in highly central competitions (panels B-C of Fig. 6).

On the other hand, the randomization of all tourneys destroys the cyclic trend of sports based on seasonal tournaments. Thus, we do not expect significant differences in the level of competition among players or any eventual correlation between their career peak and their results. Indeed, we observe in Fig. 6D-E that there is no distinction between the groups and no patterns emerge in the prestige of their tournaments.

We also compare the mean value of Spearman's correlation coefficients for both null models over all configurations. As described for the data, we computed the correlation between (1) the player's career peak and the centrality of the first tournaments they played, and (2) the player's career peak and the centrality of the tourney where they got their first win in a main draw. The results of both randomizations are summarized in Table 1. When reshuffling individual sequences, we observe that the athlete's career peak is slightly positively correlated with both the centrality of the first tournaments attended and the centrality of the first win. Instead, when we randomize all possible tournaments among the players, we find almost zero correlation in both cases.

Whereas we find a significant difference between these two correlation coefficients in the data, we observe that such a discrepancy is not significantly different from zero for both null models. This means that the behavior we observe in the data cannot happen by chance, i.e., the discrepancy between the centrality of the first tourneys top players attend compared to where they first win a match in their professional career is not random. Thus, the prestige of the tournament where male tennis players have their first win is a predictor of their future careers.

### 3 Discussion

In this work, we analyze the career evolution of tennis players to uncover the key features that characterize top players and their future achievements. To do so, we introduce

a network-based ranking of tournaments that captures the underlying connections created by players' movements in the ATP circuit according to their attendance. Our focus is on the early stages of tennis players' career and we look at the level of tourneys they attend upon entering the ATP circuit. We find that participation in tournaments of different levels is not a good predictor of athletes' success. Instead, we find that the level of the tourney where players win their first match allows us to identify the top players. We conclude that the first match win in highly central tournaments is a revealing factor for the future success of male tennis players.

We can speculate on possible explanations for this relationship between the prestige of the first win and the success in tennis. Up-and-coming players who win at a central venue might have their visibility boosted, attracting the attention of the rest of the tennis community, especially talent scouts and tournament organizers. The former could bring motivation, new staff, and perhaps even fans, ultimately reaching a broader audience through the media. The latter could award promising players with a wild card, which would allow them to access more relevant tournaments without the required ranking (wild cards are awarded at the discretion of the organizers) [30]. These circumstances would boost players' confidence in the management of highly demanding matches, both physically and mentally. Also, players with comparable performance, but in a less prestigious event, receive fewer ranking points. Therefore, a first win in a prestigious competition paves the way for accessing more and more important tournaments. Additionally, the economic benefits of winning in tennis (partially due to the prize money of the tourneys themselves, more commonly related to sponsorship and advertising) could play a role in shaping players' careers.

Our findings highlight the impact of the initial stages of players' careers, as a single match win can affect their future trajectories. Furthermore, they advocate for a deeper investigation of the economic implications that follow relevant sports results and might influence the professional development of players.

## 4 Methods

### 4.1 Ranking point scale

Professional male tennis players accumulate points in the ATP ranking during a season (52 weeks). Any new result cancels out the corresponding one from the previous year, if present, so the rankings are updated approximately every week [30]. Tournaments of different ATP categories award different point scales. Within each category, players generally compete for the same (fixed) number of points in each round. Among the competitions we considered in this work, Challengers are the less prestigious, as awarded points vary from 3 (to the loser of the first round of qualifications) up to 125 (winner of the tourney), whereas Grand Slams are the most prestigious, as players' awards range from 8 to 2000 points. The other tourneys fall in between: Masters 1000 points vary from 8 to 1000; ATP 500 points scale from 4 to 500; ATP 250 points range from 3 to 250. Detailed scales per tournament are available in the SI (Table S2).

### 4.2 Statistics of match wins

In Fig. 3A, we show that players belonging to a group  $i$  have a probability  $P_i(T \geq t)$  of attending at least  $t$  tourneys where they win a match, within the first ten tournaments of their career in the ATP. This probability results from the cumulative distribution of the

function  $p_i(t)$ :

$$P_i(T \geq t) = \int_t^{\infty} p_i(t') dt'. \quad (1)$$

Where  $p_i(t)$  is the fraction of players of a group  $i = \{\text{top, middle, bottom}\}$  with a win  $w$  in exactly  $t$  attended tourneys, with  $1 \leq t \leq 10$ , namely:

$$p_i(t) = \frac{N_i(w=t)}{\sum_{s=1}^{10} N_i(w=s)}. \quad (2)$$

We also consider the probability  $P$  that players have won their first match since becoming professional in a given tourney  $t$  (Fig. 3B). Equation (3) reports the fraction of players who have their first win at time  $t$ , that is, at the tournament  $1 \leq t \leq 100$ , grouped by their career peak. Given the players in the group  $i$  with their first win  $w^*$  at the time  $t$ , we can write the following:

$$P_i(t, w^*) = \frac{N_i(t, w^*)}{N_i(t)}. \quad (3)$$

Where  $P_i(t, w^*)$  is the fraction of players with their first win  $w^*$  at time  $t$ , that is the ratio of the number of athletes  $N_i(t, w^*)$  with their first win  $w^*$  at time  $t$ , divided by the number of players who competed at time  $t$ ,  $N_i(t)$ .

### 4.3 Network centrality

The co-attendance network of tennis tournaments is based on the career trajectories of the players in our data. This results in a weighted directed network, where nodes are tourneys and links  $(i, j)$  are created when players first attend tournament  $i$ , then  $j$ . Link weights are obtained by the number of times different players generate the same link. Specifically, every link  $(i, j)$  has a weight  $\tilde{\omega}_{ij} = \frac{\omega_{ij}}{\omega_{\max}}$ , normalized to the maximum possible weight found in the network, i.e.,  $\omega_{\max} = \max(\omega_{ij})$ . We use the topology of the co-attendance network to assess the prestige of tourneys. Specifically, we rely on the eigenvector centrality  $x_i$  [34], defined for a node  $i$  in a directed network as proportional to the centralities of the nodes that point to  $i$  [40]:

$$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j. \quad (4)$$

Where the term  $\kappa_1$  represents the largest eigenvalue of the adjacency matrix  $\mathbf{A}$  whose elements are  $A_{ij}$ .

#### Abbreviations

ATP, Association of Tennis Professionals; SI, Supplementary Information.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-024-00472-3>.

**Additional file 1.** See the attached file (PDF 1.3 MB)

### Acknowledgements

CZ would like to thank L. Gallo for valuable discussions and comments.

### Author contributions

CZ and TC conceived the research. CZ performed the analysis and wrote the first draft of the manuscript. SS provided methodological insights. RS provided the interpretation of some results and the null model formulation. TC, SS, and RS supervised the study. AP and AR helped supervise the project. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

### Funding

Open access funding provided by Corvinus University of Budapest. CZ acknowledges the support of 101086712-LearnData-HORIZON-WIDERA-2022-TALENTS-01 financed by EUROPEAN RESEARCH EXECUTIVE AGENCY (REA) (<https://cordis.europa.eu/project/id/101086712>), the Erasmus Mobility Network, and the Danish Data Science Academy (DDSA) for funding her visits to the research group of RS. AP and AR acknowledge partial financial support of PRIN 2017WZFTZP *Stochastic forecasting in complex systems*. RS and SS acknowledge support from Villum Fonden through the Villum Young Investigator program (project number: 00037394).

### Data availability

The datasets analyzed during the current study are available at the GitHub repository:

[https://github.com/JeffSackmann/tennis\\_atp](https://github.com/JeffSackmann/tennis_atp). The data generated from the analysis are available from the corresponding author upon reasonable request.

### Declarations

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Center for Collective Learning, Corvinus Institute for Advanced Studies (CIAS), Corvinus University, Budapest, 1093, Hungary. <sup>2</sup>Department of Physics and Astronomy, University of Catania and INFN sezione di Catania, 95123, Catania, Italy. <sup>3</sup>NETworks, Data, and Society (NERDS), Computer Science Department, IT University of Copenhagen, 2300, Copenhagen, Denmark. <sup>4</sup>Center for Social Data Science (SODAS), University of Copenhagen, 1353, Copenhagen, Denmark. <sup>5</sup>Complexity Science Hub, 1080, Vienna, Austria. <sup>6</sup>ISI Foundation, 10126, Turin, Italy. <sup>7</sup>Pioneer Centre for AI (P1), 1350, Copenhagen, Denmark.

Received: 11 January 2024 Accepted: 6 April 2024 Published online: 19 April 2024

### References

1. Sinatra R, Wang D, Deville P, Song C, Barabási AL (2016) Quantifying the evolution of individual scientific impact. *Science* 354(6312):aaf5239. <https://doi.org/10.1126/science.aaf5239>
2. Clauset A, Larremore DB, Sinatra R (2017) Data-driven predictions in the science of science. *Science* 355(6324):477–480. <https://doi.org/10.1126/science.aal4217>. <https://www.science.org/doi/abs/10.1126/science.aal4217>
3. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási AL (2018) Science of science. *Science* 359(6379):eaa0185. <https://doi.org/10.1126/science.aao0185>. <https://www.science.org/doi/abs/10.1126/science.aao0185>
4. Bol T, De Vaan M, van de Rijt A (2018) The Matthew effect in science funding. *Proc Natl Acad Sci* 115(19):4887–4890
5. Bonaventura M, Ciotti V, Panzarasa P, Liverani S, Lacasa L, Latora V (2020) Predicting success in the worldwide start-up network. *Sci Rep* 10(1):345. <https://doi.org/10.1038/s41598-019-57209-w>. <https://www.nature.com/articles/s41598-019-57209-w>
6. Williams OE, Lacasa L, Latora V (2019) Quantifying and predicting success in show business. *Nat Commun* 10(1):1–8
7. Fraiberger SP, Sinatra R, Resch M, Riedl C, Barabási AL (2018) Quantifying reputation and success in art. *Science* 362(6416):825–829. <https://doi.org/10.1126/science.aau7224>
8. Nadini M, Alessandretti L, Di Giacinto F, Martino M, Aiello LM, Baronchelli A (2021) Mapping the NFT revolution: market trends, trade networks, and visual features. *Sci Rep* 11(1):20902
9. Vasan K, Janosov M, Barabási AL (2022) Quantifying NFT-driven networks in crypto art. *Sci Rep* 12(1):2769
10. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856. <https://doi.org/10.1126/science.1121066>. <https://www.science.org/doi/abs/10.1126/science.1121066>
11. Janosov M, Musciotto F, Battiston F, Iñiguez G (2020) Elites, communities and the limited benefits of mentorship in electronic music. *Sci Rep* 10(1):3136. <https://doi.org/10.1038/s41598-020-60055-w>
12. Kang I, Mandulak M, Szymanski BK (2022) Analyzing and predicting success of professional musicians. *Sci Rep* 12(1):21838. <https://doi.org/10.1038/s41598-022-25430-9>. <https://www.nature.com/articles/s41598-022-25430-9>
13. Wang X, Yuceosoy B, Varol O, Eliassi-Rad T, Barabási AL (2019) Success in books: predicting book sales before publication. *EPJ Data Sci* 8(1):31. <https://doi.org/10.1140/epjds/s13688-019-0208-6>
14. Janosov M, Battiston F, Sinatra R (2020) Success and luck in creative careers. *EPJ Data Sci* 9(1):9. <https://doi.org/10.1140/epjds/s13688-020-00227-w>. arXiv:1909.07956
15. Murray CA (1950) *Human accomplishment: the pursuit of excellence in the arts and sciences, 800 BC to 1950*. Harper Collins, New York
16. Yuceosoy B, Barabási AL (2016) Untangling performance from success. *EPJ Data Sci* 5(1):17. <https://doi.org/10.1140/epjds/s13688-016-0079-z>. arXiv:1512.00894

17. Lehmann S, Jackson AD, Lautrup BE (2006) Measures for measures. *Nature* 444(7122):1003–1004
18. Barabási AL (2018) *The formula: the five laws behind why people succeed*. Pan Macmillan, London
19. Radicchi F (2012) Universality, limits and predictability of gold-medal performances at the Olympic games. *PLoS ONE* 7(7):e40335. <https://doi.org/10.1371/journal.pone.0040335>. arXiv:1203.3058
20. Radicchi F (2011) Who is the best player ever? A complex network analysis of the history of professional tennis. *PLoS ONE* 6(2):e17249. <https://doi.org/10.1371/journal.pone.0017249>. arXiv:1101.4028
21. Tennant AG, Ahmad N, Derrible S (2017) Complexity analysis in the sport of boxing. *J Complex Netw* 5(6):953–963. <https://doi.org/10.1093/comnet/cnx010>. <https://academic.oup.com/comnet/article/5/6/953/3897367>
22. Pappalardo L, Cintia P (2018) Quantifying the relation between performance and success in soccer. *Adv Complex Syst* 21(03n04):1750014. <https://doi.org/10.1142/S021952591750014X>. <https://www.worldscientific.com/doi/abs/10.1142/S021952591750014X>
23. Sobkowicz P, Frank RH, Biondo AE, Pluchino A, Rapisarda A (2020) Inequalities, chance and success in sport competitions: simulations vs empirical data. *Phys A, Stat Mech Appl* 557:124899. <https://doi.org/10.1016/j.physa.2020.124899>
24. Zappalà C, Pluchino A, Rapisarda A, Biondo AE, Sobkowicz P (2022) On the role of chance in fencing tournaments: an agent-based approach. *PLoS ONE* 17(5):1–17. <https://doi.org/10.1371/journal.pone.0267541>
25. Petersen AM, Jung WS, Yang JS, Stanley HE (2011) Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc Natl Acad Sci USA* 108(1):18–23. <https://doi.org/10.1073/pnas.1016733108>. arXiv:0806.1224
26. Zappalà C, Biondo AE, Pluchino A, Rapisarda A (2023) The paradox of talent: how chance affects success in tennis tournaments. *Chaos Solitons Fractals* 176:114088. <https://doi.org/10.1016/j.chaos.2023.114088>. <https://www.sciencedirect.com/science/article/pii/S096007792300989X>
27. Mauboussin MJ (2012) *The success equation: untangling skill and luck in business, sports, and investing*. Harvard Business Review Press
28. Frank RH (2016) *Success and luck: good fortune and the myth of meritocracy*. Princeton University Press, Princeton, p 208
29. Perc M (2014) The Matthew effect in empirical data. *J R Soc Interface* 11(98):20140378
30. Official site of men's professional tennis ATP tour. <https://www.atptour.com/>
31. Tennis data repository. [https://github.com/JeffSackmann/tennis\\_atp](https://github.com/JeffSackmann/tennis_atp)
32. Breznik K (2015) Revealing the best doubles teams and players in tennis history. *Int J Perform Anal Sport* 15(3):1213–1226. <https://doi.org/10.1080/24748668.2015.11868863>
33. Aparício D, Ribeiro P, Silva F (2016) A subgraph-based ranking system for professional tennis players. *Stud Comput Intell* 644:159–171. [https://doi.org/10.1007/978-3-319-30569-1\\_12](https://doi.org/10.1007/978-3-319-30569-1_12)
34. Bonacich P (1987) Power and centrality: a family of measures. *Am J Sociol* 92(5):1170–1182
35. Wu MC, Carroll RJ (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44(1):175–188
36. Schulz R, Curnow CS (1988) Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. *J Gerontol* 43(5):P113–20
37. Guillaume M, Len S, Tafflet M, Quinquin L, Montalvan B, Schaal K, Nassif H, Desgorces FD, Toussaint JF (2011) Success and decline. *Med Sci Sports Exerc* 43(11):2148–2154. <https://doi.org/10.1249/MSS.0b013e31821eb533>. <https://journals.lww.com/00005768-2011111000-00017>
38. Davidson-Pilon C (2019) Lifelines: survival analysis in Python. *J Open Sour Softw* 4(40):1317. <https://doi.org/10.21105/joss.01317>
39. Baker J, Koz D, Kungl AM, Fraser-Thomas J, Schorer J (2013) Staying at the top: playing position and performance affect career length in professional sport. *High Abil Stud* 24(1):63–76. <https://doi.org/10.1080/13598139.2012.738325>
40. Newman M (2018) *Networks*. Oxford University Press, London
41. Reid M, Crespo M, Santilli L, Miley D, Dimmock J (2007) The importance of the international tennis federation's junior boys' circuit in the development of professional tennis players. *J Sports Sci* 25(6):667–672. <https://doi.org/10.1080/02640410600811932>. PMID: 17454534
42. Kovalchik SA, Bane MK, Reid M (2017) Getting to the top: an analysis of 25 years of career rankings trajectories for professional women's tennis. *J Sports Sci* 35(19):1904–1910. <https://doi.org/10.1080/02640414.2016.1241419>. <https://www.tandfonline.com/doi/full/10.1080/02640414.2016.1241419>
43. Li P, Bosscher VD, Weissensteiner JR (2018) The journey to elite success: a thirty-year longitudinal study of the career trajectories of top professional tennis players. *Int J Perform Anal Sport* 18(6):961–972. <https://doi.org/10.1080/24748668.2018.1534197>
44. Brouwers J, De Bosscher V, Sotiriadou P (2012) An examination of the importance of performances in youth and junior competition as an indicator of later success in tennis. *Sport Manag Rev* 15(4):461–475. <https://doi.org/10.1016/j.smr.2012.05.002>. <https://www.sciencedirect.com/science/article/pii/S1441352312000496>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.