A subgraph-based ranking system for professional tennis players

David Aparício, Pedro Ribeiro and Fernando Silva

Abstract This paper introduces a novel ranking system for competitive sports based around the notion of subgraphs. Although the system is targeted specifically to professional tennis it could be applied to any dominance network due to its generality. The results of about 140,000 tennis matches played between Top-100 players are used to create a colored directed network where colors represent different surfaces and edge direction depends on head-to-read results between players. The main contribution of this work is a ranking system which relies on the occurrences of 4-node directed subgraphs and the positions (or orbits) where the players appear on them. Since the concept of orbit is intrinsically connected with node dominance, appearing frequently in dominant orbits indicates that the player himself is dominant. Even in a very sparse network and without any background knowledge on the tournaments or stages of the matches, our proposal is able to extract meaningful rankings which capture the intricate competitive relationships between players from different eras.

1 Introduction

Debating who is the best player (or team) is one of the most discussed topics in any competitive sport and it can stir heated arguments between fans. Objectively quantifying player achievements is not straightforward, even when personal preferences are set aside, since multiple criteria can be used to compare players and the sports themselves evolve throughout the years. Nevertheless, competitive sports require a system that is able to rank players (or teams) according to their performance.

Most existing ranking systems focus on some set of numerical features, with different weights and time spans used depending on the sport under consideration [11]. Professional tennis in particular is governed by the Association of Tennis Professionals (ATP) which ranks players based on their results in official ATP tournaments. The ATP ranking is updated on a weekly basis and aggregates the results from the previous 52 weeks. Points are awarded to players according to the round of the tournament that they reach and the ranking of the tournament itself. Recently, with the emergence of network science, node centrality metrics have been applied to sports datasets in order to derive rankings [5, 7, 9, 10]. The vast majority of these ranking methods are adaptations of the PageRank algorithm [2]. In this work we take a different perspective by instead considering the role of small subgraphs. Subgraph-based metrics have been used to evaluate node importance in other fields such as

David Aparício · Pedro Ribeiro · Fernando Silva
CRACS & INESC-TEC, DCC-FCUP, Universidade do Porto, Portugal
e-mail: {daparicio, pribeiro, fds}@dcc.fc.up.pt

biology [12]. Our goal is to provide a ranking system that truly captures the dynamics of the network. For that purpose, we devise a ranking mechanism that considers not only the subgraphs themselves but also the position (or orbit) of the players in the subgraphs. Orbit information allows us to discover indirect dominance while at the same time weighting both inward and outward edges. This method contrasts with PageRank which essentially considers only one of the two possible edge directions, giving importance to wins and almost disregarding losses, or vice-versa.

Our approach was tested on one of the most popular individual sports: men's professional tennis. Our results show that, even without any kind of prior knowledge, the methodology put forward is able to produce consistent and meaningful results using only the topology of the dominance network .

2 Network Description

In order to construct the dominance network we first collected the names of all tennis players that have been ranked in the Top-100 of the ATP year-end rankings from 1974 until 2015 and then extracted their match information from Tennis Abstract[1]. Going beyond the Top-100 introduces noise in the data and is not necessary for our purposes since players below the Top-100 only enter a few major tournaments. A total of 856 tennis players have been in the Top-100 throughout the years and they have played about 140,000 matches between themselves. The amount of matches played annually on each surface is presented in Figure 1 as well as the total number (dotted line). This number increased significantly in the 1990s but has dropped in recent years mostly due to changes in the ranking system that encourage players to only participate in the most prestigious tournaments and also thanks to an increased awareness of the sport's physical demands. Nowadays, most tennis tournaments are contested on either clay or hard courts, with only a handful of matches played on grass each year. Carpet was a popular surface until the mid-1990s but it was discontinued from the ATP Tour in the late 2000s. The surface characteristics affect the pace of the game, favouring different playing styles. Usually, grass is the *fastest* surface to play on, followed by carpet, hard and finally clay.

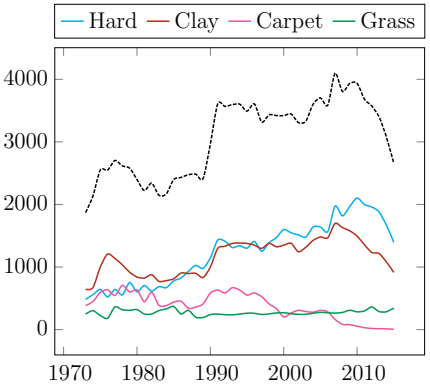


Fig. 1: Matches played by year on each surface.

Table 1: Global network statistics of the dominance networks, discriminated by surface.

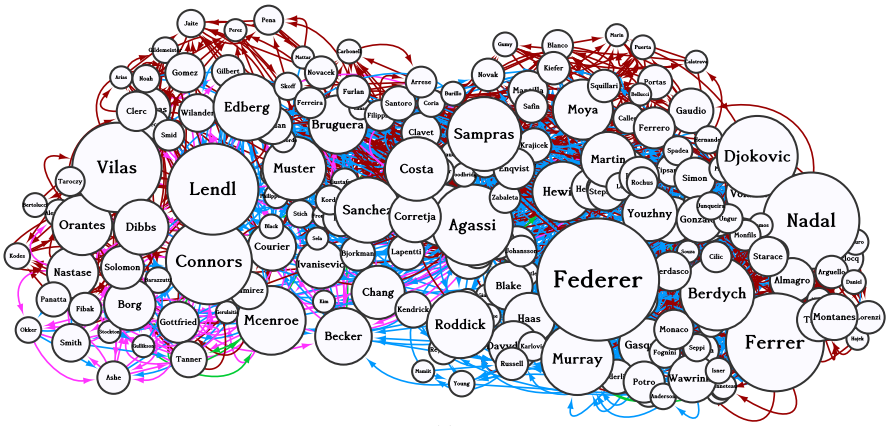
Surface	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	$\frac{\|\mathbf{E}\|}{\|\mathbf{V}\|}$	$\frac{\|\mathbf{E}_{\Rightarrow}\|}{\|\mathbf{E}\|}$
Hard	301	868	2.88	0.64
Clay	289	793	2.74	0.65
Grass	140	173	1.24	0.90
Carpet	97	188	1.94	0.72
Overall	585	3279	5.61	0.68

Following data extraction, the information is processed in order to construct 6-tuples of the form $(Player_1, Player_2, Surface, Year, Matches, WinPercentage)$ for each pair of players. Besides creating a tuple for each surface, an additional 6-tuple is necessary to account for overall head-to-head. Using this data, a *dominance network* is created where nodes are players and the orientation of the *colored directed edges* between two players depends on their head-to-head on a given surface. Consider tuple $(p_i, p_j, s, t, m_{ij}^{s,t}, w_{ij}^{s,t})$ and parameters δ and ϕ: a colored directed edge (p_i, p_j) is created if player p_i won at least $\delta\%$ of the matches against p_j on surface s in a given year t (Equation 1) and they played a minimum ϕ matches in surface s during their careers (Equation 2). Our networks were built with $\delta = \frac{2}{3}$, meaning that one player only *dominates* another if he has defeated him in more than 66% of the matches. A minimum of 3 matches ($\phi = 3$) is required to establish a dominance relation between two players on grass courts, and 5 for the other surfaces. An overall dominance relation that disregards playing surface also requires at least 5 matches.

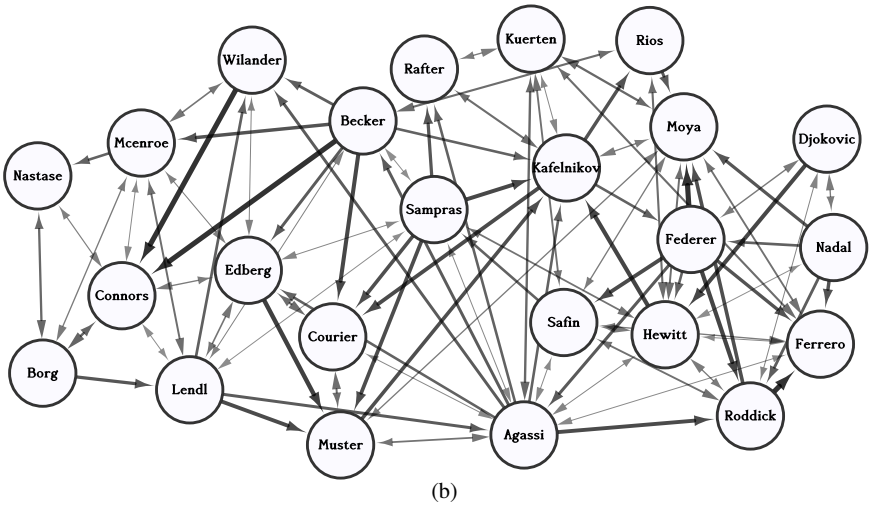
$$w_{ij}^{s,t} \geq \delta\% \qquad (1) \qquad\qquad \left(\sum_{t=1974}^{2015} m_{ij}^{s,t} \right) \geq \phi \qquad (2)$$

An *aggregated* (or *career*) dominance network is assembled by calculating dominances using the career win-percentage, instead of yearly results. The resulting network has 585 vertices and 5,301 directed edges with 5 possible labels (or *colors*): hard, clay, grass, carpet or overall. The number of *overall* edges is not simply the sum of the edges from all surfaces since an overall dominance is established by playing a minimum ϕ matches on any surface (for instance, one player can dominate another in overall matches without having ϕ encounters with him in any particular surface). Notice that only 585 of the original 856 players are represented in the network since the others did not play the required ϕ matches against any other Top-100 player, and consequently have no edges. Requiring the win-percentage to be above a certain threshold δ for a dominance relation to be established results in the creation of bidirectional (or *reciprocal*) edges, meaning that two players met in at least ϕ matches but neither one dominates the other. The *dominant* (unidirectional) edges and *non-dominant* (bidirectional) edges are henceforth represented as $E_{\Rightarrow}$ and $E_{\Leftrightarrow}$, respectively. Table 1 summarizes the networks' global statistics. In regard to individual players, Jimmy Connors dominates the most other players (63), followed by Roger Federer (60) and Ivan Lendl (59). On hard courts Roger Federer leads with 46 out-edges, Guillermo Vilas on clay with 37, John McEnroe on carpet with 23 and Roger Federer on grass with 17 out-edges.

Two visual representations of the network are presented in Figure 2. The giant component of the aggregate network is shown in Figure 2 (a). Each edge color matches a surface: blue for hard, brown for clay, green for grass, pink for carpet and black for overall dominances. Node size depends on the number of out-edges; Roger Federer corresponds to the largest node since he has the most out-edges (132). Figure 2 (b) shows the relations between all 25 players that have been ranked as the ATP Top-1 player. Edges are only relative to overall dominance and the line thickness reflects how unbalanced the relation is. It is interesting to notice that Jimmy Connors,



(a)



(b)

Fig. 2: Player dominance networks: in (a) blue edges are drawn for dominances in hard courts, brown for clay, green for grass and pink for carpet. The nodes' size increases proportionally with their out-degree. (b) shows the relations between all ATP Top-1 players, disregarding surface.

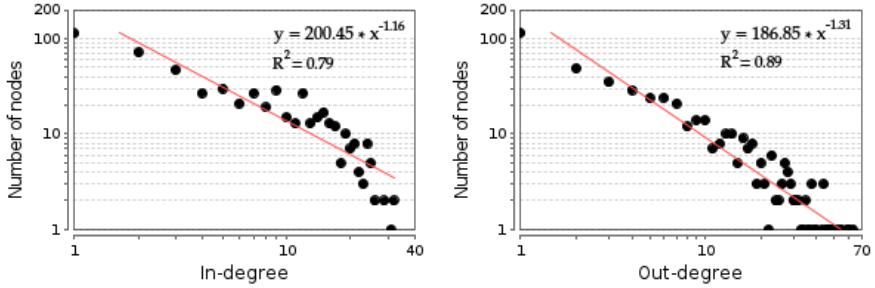


Fig. 3: In- and out-degree distribution of tennis players dominance networks (overall matches).

one of the players with most out-edges, does not dominate any Top-1 player. The fact that he faced the others when they were closer to their prime than himself might be the main reason for this. It seems reasonable to expect younger players, which are at their peak, to dominate players declining in form. However, that is generally only the case for players of the same level: very good young players tend to dominate very good older players, but average young players do not usually win against very good older players. Furthermore, considering the players' full history allows us to capture the various stages of their careers. Comparing players from different eras might seem unfair if one inspects only individual relations but what really makes a player *dominant* is the global aspect of his career and the head-to-head results that he had against players from his own era, players from the era preceding his and players from the subsequent era. Therefore, it is difficult to infer that a player p_i from one era dominates another player p_j from a different era, however it is possible to say that p_i is *generally more dominant* than p_j, and those are the relations that we intend to capture using our ranking mechanism.

Figure 3 shows that the networks' in- and out-degrees follow a power law. Results are only presented for overall matches but the surface networks are also scale-free.

3 Subgraph-based Ranking System

Related Work

The discussion of who is the best tennis player of all-time is open for debate and multiple criteria can be used. Ranking players simply by the number of matches that they won unduly favours players that had very long careers, such as Jimmy Connors, and ranking players by their win-ratio excessively benefits those that, like Björn Borg, retired at the peak of their careers. Furthermore, these possibilities do not take into account the intricate relations between the players. Grand Slam tournament victories (or *grand slams* for short) are often used to compare tennis players; however, before the 1990s several top-ranked players willingly skipped some of the annual Grand Slam tournaments since it was not yet the norm to evaluate players by their number of grand slams.

The work by Radicchi et al. [9] proposes a PageRank-like ranking system for male tennis players. Dingle et al [3] also used Radicchi's ranking system to produce a more up-to-date ranking of both male and female tennis players. The network that Raddicchi et al. built is different from our own since a) their edges are *weighted* (w_{ij}: number of times that p_i beats p_j) while ours are *simple directed edges* reflecting win-percentages, b) they used match information from 1969 until 2010 whilst our networks are relative to matches from 1974 to 2015 and c) they only considered matches played on either Grand Slam tournaments or ATP Masters 1000 whereas we use information from all official ATP tournaments. Traditional PageRank does not decrease the node's rank with respect to its out-edges (in this case, meaning *loses against*) and is therefore not suitable to determine player dominance relations. The prestige score presented in [9] lowers the w_{ij} according to p_j's out-degree (the

number of times p_j loses against someone); therefore, dominating a *dominated* node gives less prestige than dominating a more *dominant* player. However, the prestige score is not decreased according to p_i's out-edges, which may result in *dominated* players having a high score as long as they dominate a few *dominant* players. Our scoring system increases the players' score in respect to the players that they dominate and, likewise, decreases their score when they are themselves dominated. Another approach was followed by Motegi and Masuda [7] where they use a dynamic win-loss score that takes into account temporal information and fluctuations in the ranking. They not only consider direct wins and losses but also indirect ones, namely those corresponding to directed paths of size 2. Our work differs because we use subgraphs of size 4, which encapsulate more information than paths of size 2. Furthermore, we consider global dominance relations to obtain an earned ranking, while their work focuses on obtaining a temporal snapshot for a particular point in time and use it for prediction purposes.

Methodology

A simple way to assess node dominance is to compare its out-degree (*dominant*) with its in-degree (*dominated*). However, tennis players face a limited set of opponents due to their ranking (higher ranked players seldom play against lower ranked players) and career span (players from different periods never face each other). Moreover, requiring at least ϕ matches to be played for a relation to be established further decreases the amount of direct relations, resulting in very sparse networks $\left(\frac{|E|}{|N|^2} \leq 0.05\right)$. Therefore, comparing players only by degree is not sufficient.

Another option is to consider richer structural units: *subgraphs*. Actually, the degree of a vertex $v \in V(G)$ can be regarded as a 2-node subgraph where v occupies one of its two possible positions. In this work, instead of looking only at the directed degree (or subgraphs $\{v \to u\}$, $\{u \to v\}$ and $\{v \leftrightarrow u\}$), we analyse slightly larger subgraphs and observe at which position vertex v appears in each occurrence. As illustrated in Figure 4, this allows not only for direct *dominances* ($a \to b$) or *equivalences* ($a \leftrightarrow b$) to be captured but also for indirect dominances ($a \to b \leftrightarrow c$, therefore $a \dashrightarrow c$) and super dominances ($a \to b \to c$, therefore $a \twoheadrightarrow c$) due to graph transitivity. This is particularly useful in the tennis players network since, as discussed previously, players have a very limited number of edges (direct dominances). Another advantage lies in the fact that it enables dominance relations to be established between players of different eras by following the path of the subgraph, such as $\{Federer \to Agassi \to Becker \to Connors\}$, which leads to the conclusion that $\{Federer \twoheadrightarrow Connors\}$. However, there are many other possible paths from Federer to Connors and in some of them Connors may actually indirectly dominate Federer. Therefore, all paths from one player to another must be enumerated in order to assess indirect dominances.

Graphlets [8] are subgraphs that take the node position of the subgraph (or *orbit*) into account. Graphlet usage is often restricted to analyzing only the set of 30 undirected graphs of up to five nodes due to computational limitations. Using undirected subgraphs would not produce meaningful results in dominance networks since edge

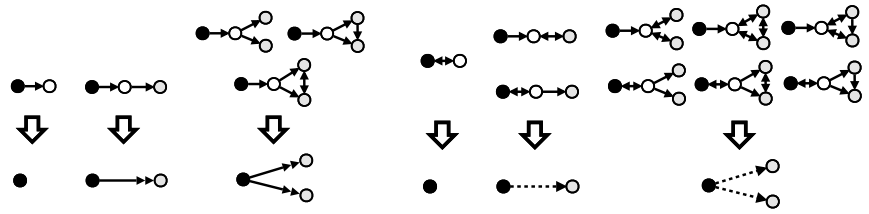


Fig. 4: Graph transitivity translated to *direct dominance* ($a \rightarrow b$), *super dominance* ($a \twoheadrightarrow b$) and *indirect dominance* ($a \dashrightarrow b$).

direction is crucial. An extension of graphlets to directed networks was proposed in [1]. Graphlets can be used, for instance, to compare the topology of networks [4, 8] or nodes [6]. The key idea is to compute how many times a given node appears in an orbit and repeat that process for all possible orbits. Two nodes are more or less alike depending on how similar their orbit frequencies are. For instance, two nodes present at the center of multiple stars are more similar to each other than to another node that appears more frequently at the stars' periphery. Usually, graphlet computation is not concerned with specific types of subgraphs (such as chains, stars or cliques), but instead with all possible subgraphs of a given size. The results presented here are relative to all 199 possible directed subgraphs with 4 nodes.

In a first step, our subgraph-based ranking system receives as input a set of graphlets and assigns scores to their orbits. Then, during subgraph enumeration, the player's score is increased or decreased according to the orbits that he appears in. Orbit scores are calculated using the transitivity closure of the subgraph, as shown in Figure 5, where d_{ij} is the path length between node n_i and node n_j. Notice that different nodes of a subgraph may be in the same orbit, and will always have the same score (see orbit e from subgraph G_B for instance). Looking at G_B we identify orbit f as *dominant* since it has 3 out-edges and no in-edges, while orbit e has no out-edges and 2 in-edges, representing a *dominated* orbit. Orbits a, b, c and d from G_A con-

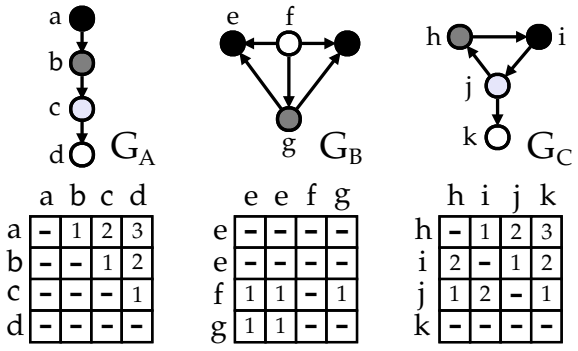


Fig. 5: Graph transitivity of 3 subgraphs. Nodes with the same shade are in the same orbit. Orbit scores are assessed using the transitivity matrix: row values are *positive* points while column values are *negative*. Higher cell values mean that the connection is less direct.

stitute a 4-node chain where the orbits at its start are more dominant than the ones at its end since they indirectly dominate more orbits. Orbits h, i and j of G_C form a cycle and are therefore equivalent. However, orbit j dominates k directly while h and i dominate k indirectly. Also, orbit i dominates orbit k more directly than orbit h does. These considerations are taken into account by our scoring mechanism.

Orbit scores are calculated as shown in Equation 3. The main idea is to subtract the negative points $\left(\sum_{j=0}^{|\mathcal{S}_o|}\beta^{k-d(o_j,o)}\right)$ from the positive ones $\left(\sum_{i=0}^{|\mathcal{I}_o|}\beta^{k-d(o,o_i)}\right)$. Set $\mathcal{I}_o$ is formed by the orbits *inferior* to the orbit being computed while $\mathcal{S}_o$ is the set of orbits *superior* to it. The *distance* between o_i and o_j is given by $d(o_i,o_j)$ and it can be at most $k-1$, where k is the size of the subgraph. Basically, *direct dominant connections* give more points than indirected ones and, conversely, *direct dominated connections* take more points away. Parameter β controls the relative importance of the *directedness*, i.e. a small β (closer to 1) means that direct and indirect dominances give roughly the same points while an high β means that direct dominances are more important. Notice that if β is too big the score becomes almost equivalent to the degree. A parameter $\lambda \in [0,1]$ is also inserted to control the influence of *dominating* versus *being dominated*. Using $\lambda \approx 1$ means that a player is mostly evaluated by how many players he dominates (out-edges) while the amount of times that he himself is dominated (in-edges) does not have a big impact in the rankings, and vice-versa when $\lambda \approx 0$, i.e. the player is ranked higher if he is dominated by few players. These considerations produce a flexible scoring mechanism with just two parameters. The score of a player p_i is obtained by summing his occurrences in all orbits and multiplying them by their score, as shown in Equation 4. Finally, players are ordered from the lowest to the highest score to produce the ranking.

$$S(o) = \left(\lambda \times \sum_{i=0}^{|\mathcal{I}_o|}\beta^{k-d(o,o_i)}\right) - \left((1-\lambda)\times \sum_{j=0}^{|\mathcal{S}_o|}\beta^{k-d(o_j,o)}\right) \tag{3}$$

$$S(p_i) = \sum_{o=0}^{|\mathcal{O}|} Fr(p_i,o)\times S(o) \tag{4}$$

4 Results

Table 2 presents the 15 players with the highest scores depending on λ. In the middle column λ is $\frac{1}{2}$, meaning that *dominating* and *not being dominated* is equally important for the players' scores, and this value is used for comparison. When $\lambda < \frac{1}{2}$ the ranking mechanism gives more importance to *not being dominated* that to *dominating* other players. Players such as Björn Borg and Gustavo Kuerten benefit from this parameter choice whereas Guillermo Vilas is penalized. If ones keeps decreasing λ, Rafael Nadal eventually tops the ranking because very few players have a positive win-loss ratio against him. However, making λ too small results in a meaningless ranking since players that have few out-edges unrealistically climb in the

rankings as long as they have very few (or none) in-edges. By contrast, when $\lambda > \frac{1}{2}$, players such as Carlos Moya and Guillermo Vilas climb in the rankings while Björn Borg and Novak Djokovic drop some positions. Having $\lambda \approx 1$ still produces meaningful results since the ranking eventually stabilizes and ranks very highly players that dominate many others. Nevertheless, it does not seem fitting to completely disregard the *dominated* edges of the players when building a dominance based ranking system. In the remaining results λ is set to $\frac{1}{3}$, hence giving a slight edge to players that are not dominated by many others while still producing meaningful results.

Rank	Player	⇕	Player	⇕	Player	Player	⇕	Player	⇕
1	I. Lendl	1 ▲	I. Lendl	1 ▲	R. Federer	R. Federer		R. Federer	
2	R. Federer	1 ▼	R. Federer	1 ▼	I. Lendl	I. Lendl		I. Lendl	
3	J. Connors		J. Connors		J. Connors	J. Connors		J. Connors	
4	R. Nadal	1 ▲	A. Agassi		A. Agassi	A. Agassi		A. Agassi	
5	N. Djokovic	1 ▲	R. Nadal		R. Nadal	R. Nadal		R. Nadal	
6	B. Becker		N. Djokovic	1 ▲	B. Becker	B. Becker		B. Becker	
7	A. Agassi	3 ▼	B. Becker	1 ▼	N. Djokovic	S. Edberg	1 ▲	G. Vilas	2 ▲
8	B. Borg	5 ▲	S. Edberg		S. Edberg	N. Djokovic	1 ▼	S. Edberg	
9	S. Edberg	1 ▼	J. McEnroe	1 ▲	G. Vilas	G. Vilas		N. Djokovic	2 ▼
10	P. Sampras	2 ▲	G. Vilas	1 ▼	J. McEnroe	J. McEnroe		J. McEnroe	
11	J. McEnroe	1 ▼	L. Hewitt		L. Hewitt	L. Hewitt		L. Hewitt	
12	A. Murray	4 ▲	P. Sampras		P. Sampras	P. Sampras		Y. Kafelnikov	2 ▲
13	L. Hewitt	2 ▼	B. Borg		B. Borg	Y. Kafelnikov	1 ▲	P. Sampras	1 ▼
14	G. Kuerten	7 ▲	A. Murray	2 ▲	Y. Kafelnikov	C. Moya	3 ▲	C. Moya	3 ▲
15	G. Vilas	6 ▼	A. Roddick		A. Roddick	B. Borg	2 ▼	D. Ferrer	2 ▲
	$\lambda = \frac{1}{6}$		$\lambda = \frac{1}{3}$		$\lambda = \frac{1}{2}$	$\lambda = \frac{2}{3}$		$\lambda = \frac{5}{6}$	

Table 2: Ranking obtained by varying λ: the relative weight between dominating (out-edges) and being dominated (in-edges).

Table 3 shows the career rankings with $\beta = 1.5$. To illustrate the effect of β take graph G_A from Figure 5 as an example: if $\beta = 1$, $S(a) = 1^{(4-1)} + 1^{(4-2)} + 1^{(4-3)} = 3$, $S(b) = 1$, $S(c) = -1$ and $S(d) = -3$; if $\beta = 2$, $S(a) = 2^{(4-1)} + 2^{(4-2)} + 2^{(4-3)} = 14$, $S(b) = 4$, $S(c) = -4$ and $S(d) = -14$. In pratice, this means that orbits a and b, for instance, are much more alike when $\beta = 1$ than when $\beta = 2$. A low β (≈ 1) does not distinguish direct from indirect relations while a high β (≈ 2) penalizes indirect ones too heavily, therefore an intermediate value for β (1.5) was chosen.

Roger Federer is the most dominant player since 1974 according to our ranking system, followed by Jimmy Connors and Ivan Lendl. Evaluating if the results are correct is not straightforward and highly subjective. Nonetheless, one of the most commonly used criteria to judge the quality of a tennis player is the number of grand slams that he won during his career. From Table 3 (a) it can be observed that winning grand slams is correlated with a higher position in our ranking. From the Top-25 players only David Ferrer, Tim Henman and Robin Soderling failed to win any grand slams. Table 3 (b) shows the Top-10 by surface and also the number of grand slams contested on that surface that they won. Roger Federer is the most dominant

player both on grass and hard courts, Guillermo Vilas is the best player on clay and McEnroe is ranked first in carpet courts. Again, the number of grand slam victories is correlated with the ranking. We point out that no grand slam tournament was ever contested on carpet. Table 3 (c) gives a more in-depth look at all 25 players that have been the Top-1 player in the ATP rankings from 1974 until 2015. A dash (–) means that the player does not have a single connection on that particular surface, i.e. he did not play the minimum ϕ matches against anyone. The position of the player is presented in bold-face only if our system ranks him among the Top-25 of that particular surface. As can be observed, most (76%) ATP Top-1 players are also ranked as one of the Top-25 most dominant players by our system. The exceptions are John Newcombe, Mats Wilander, Jim Courier, Marcelo Rios, Patrick Rafter and Marat Safin. Patrick Rafter is a notable outlier since he is ranked at the bottom half of the table (381th out of 585 players). Notice however that he was only ranked as the ATP Top-1 for one week. Our ranking also detects surface specialists (such as Wilander, Muster, Rios, Kuerten and Ferrero on clay, Courier, Agassi and Hewitt on hard courts, and Newcombe and Rafter on grass), all-round players (such as Năstase,

Rank	Player
1	R. Federer[17]
2	J. Connors[8]
3	I. Lendl[8]
4	A. Agassi[8]
5	R. Nadal[14]
6	J. McEnroe[7]
7	G. Vilas[4]
8	N. Djokovic[10]
9	B. Becker[6]
10	P. Sampras[14]
11	S. Edberg[6]
12	A. Roddick[1]
13	A. Murray[2]
14	L. Hewitt[2]
15	B. Borg[11]
16	T. Muster[1]
17	C. Moya[1]
18	I. Năstase[2]
19	D. Ferrer*
20	G. Kuerten[3]
21	Y. Kafelnikov[2]
22	A. Ashe[3]
23	JC. Ferrero[1]
24	T. Henman*
25	R. Soderling*

(a)

Rank	Player	Player
1	R. Federer[9]	G. Vilas[2]
2	N. Djokovic[7]	R. Nadal[9]
3	A. Agassi[6]	T. Muster[1]
4	A. Murray[1]	S. Bruguera[2]
5	A. Roddick[1]	G. Kuerten[3]
6	R. Nadal[3]	M. Orantes[1]
7	P. Sampras[7]	B. Borg[6]
8	L. Hewitt[1]	M. Wilander[3]
9	T. Berdych	I. Năstase[1]
10	I. Lendl[5]	I. Lendl[3]
	Hard	**Clay**
Rank	**Player**	**Player**
1	R. Federer[7]	J. McEnroe
2	J. Connors[4]	B. Becker
3	Edmondson[1]	I. Lendl
4	J. McEnroe[3]	J. Connors
5	R. Tanner[1]	G. Ivanisevic
6	B. Becker[3]	P. Sampras
7	S. Edberg[4]	B. Borg
8	N. Djokovic[3]	A. Ashe
9	P. Cash[1]	K. Rosewall
10	P. Sampras[7]	B. Walts
	Grass	**Carpet**

(b)

Player	Overall	Hard	Clay	Grass	Carpet
I. Nătase	**18**	26	**9**	–	**18**
J. Newcombe	38	–	–	128	38
J. Connors	**2**	**11**	27	**2**	**4**
B. Borg	**15**	31	**7**	**12**	**7**
J. McEnroe	**6**	27	297	**4**	**1**
I. Lendl	**3**	**10**	**10**	90	**3**
M. Wilander	27	234	**8**	96	78
S. Edberg	**11**	**12**	118	**7**	70
B. Becker	**9**	192	55	**6**	**2**
J. Courier	41	**13**	37	32	73
P. Sampras	**10**	**7**	66	**11**	**6**
A. Agassi	**4**	**3**	63	28	41
T. Muster	**16**	178	**3**	–	–
M. Rios	33	252	**20**	–	–
C. Moya	**17**	190	149	–	–
Y. Kafelnikov	**21**	269	321	**22**	49
P. Rafter	381	243	193	**20**	–
M. Safin	46	298	31	167	–
G. Kuerten	**20**	**21**	**5**	–	–
L. Hewitt	**14**	**8**	177	496	–
JC. Ferrero	**23**	231	**16**	–	–
A. Roddick	**12**	**5**	76	63	–
R. Federer	**1**	**1**	**14**	**1**	–
R. Nadal	**5**	**6**	**2**	118	–
N. Djokovic	**8**	**2**	35	**8**	–

(c)

Table 3: Ranking of tennis players with $\lambda = \frac{1}{3}$ and $\beta = 2$: (a) Top-25 players, (b) Top-10 players by surface and (c) our rankings for all players ranked as Top-1 by the ATP.

Connors and Federer) and players with an Achilles-heel on a specific surface (such as Sampras and Djokovic on clay, Borg on hard, and Lendl and Nadal on grass). We should note that, for instance, Rafael Nadal has a very low score on grass despite having a $\approx 79\%$ win-loss ratio in that surface and winning two grand slams on grass. His very low score comes primarily from the fact that he is dominated by Roger Federer on that surface and, because Federer is a hub-like node in grass, Nadal ends up appearing in many different subgraphs with Federer and the other players that Federer dominates. Since Nadal occupies a negative orbit in those subgraphs his score is continuously decreased. This negative effect is primarily felt on small and sparse networks such as the grass network where even a single connection has a very high impact. A possible solution to reduce the influence of hubs would be to ensure that each player only decreases the score of another player once.

5 Conclusion

The first contribution of this work is the distribution[2] of a network summarizing the complete match history between all male Top-100 ATP players since 1974. The data is discriminated by year as well as playing surface. The constructed dominance network models the relations between players: if a player wins against another one more than ϕ times and wins at least $\delta\%$ of the matches, a directed connection is drawn between them. An exploratory analysis was performed in order to verify that these choices are adequate and produce a meaningful representation. It was also observed that, like many real-world networks, both its in- and out-degree distributions follow a power-law, meaning that there are few very dominant players, few very dominated players and many average players.

We present a ranking system based on the subgraph topology of the dominance network that offers a different view than past approaches based on the PageRank algorithm. A complete subgraph enumeration is performed in the original network in order to compute the ranking. During the enumeration process, the position that the player appears in the subgraph is stored and his score is updated: if the player appears in a dominant orbit his ranking is increased, while if he appears in a dominated orbit his ranking is decreased. The ranking system does not require any meta-information about the network such as the tournament or the round that the players faced each other to produce meaningful results, however it could easily be extended to support it by adjusting the edge weights. The system is also flexible since it is possible to control i) λ the importance of *being dominant* versus *being dominated* and ii) β the importance of *direct* versus *indirect* dominances.

We assess which values of λ and β are better-suited for this particular tennis network and present rankings for the best overall players since 1974 and the most dominant players by surface. Our ranking system produces results that agree with the ATP ranking while at same time offering a different perspective since wins are

[2] http://www.dcc.fc.up.pt/~daparicio/networks

not discriminated by tournaments (which some are more valuable than others) nor rounds (where a win in a later round gives more ATP points) and the intricate relations between players are also captured. This approach gives a better idea of actual player dominance which is valuable when trying to assess who are the best tennis players. Using our ranking system it was possible, for instance, to i) observe that player performance is heavily influenced by the playing surface and ii) discover which former ATP World Top-1 players were actually dominant players and which ones were not. We also performed a yearly ranking not included here for space concerns where we i) observed that the most dominant players are usually the ones that reach more tournament finals, semi-finals and quarter-finals but they are not necessarily the ones that win more tournaments due to the unbalanced nature of ATP ranking system, ii) identified which seasons were *most dominated* by a single (or a few) player(s), iii) pinpointed tennis transition-eras (1987-1989 and 1999-2003) and iv) noticed that it is rare for a player be very dominant both on fast (hard or grass) and slow courts (clay).

Acknowledgements: This work is partially funded by FCT (Portuguese Foundation for Science and Technology) within project UID/EEA/50014/2013. David Aparício is supported by a FCT/MAP-i PhD research grant (PD/BD/105801/2014).

References

1. Aparício, D., Ribeiro, P., Silva, F.: Network comparison using directed graphlets. arXiv preprint arXiv:1511.01964 (2015)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference (WWW 1998) (1998)
3. Dingle, N., Knottenbelt, W., Spanias, D.: On the (page) ranking of professional tennis players. In: Computer Performance Engineering, pp. 237–247. Springer (2013)
4. Hayes, W., Sun, K., Pržulj, N.: Graphlet-based measures are suitable for biological network comparison. Bioinformatics 29(4), 483–491 (2013)
5. London, A., Németh, J., Németh, T.: Time-dependent network algorithm for ranking in sports. Acta Cybernetica 21(3), 495–506 (2014)
6. Milenković, T., Ng, W.L., Hayes, W., Pržulj, N.: Optimal network alignment with graphlet degree vectors. Cancer informatics 9, 121 (2010)
7. Motegi, S., Masuda, N.: A network-based dynamical ranking system for competitive sports. Scientific Reports 2(904) (2012)
8. Pržulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics 23, 177–183 (2007)
9. Radicchi, F., Perc, M.: Who is the best player ever? a complex network analysis of the history of professional tennis. PloS one 6(2), e17249 (2011)
10. Shan, Z., Li, S., Dai, Y.: Gamerank: Ranking and analyzing baseball network. In: Social Informatics. pp. 244–251. IEEE Computer Society (2012)
11. Stefani, R.: Survey of the major world sports rating systems. Journal of Applied Statistics 24(6), 635–646 (1997)
12. Wang, P., L, J., Yu, X.: Identification of important nodes in directed biological networks: A network motif approach. PLoS ONE 9(8), e106132 (08 2014)