

Napovedovanje izidov teniških dvobojev s pomočjo omrežja

Analiza omrežij

Andraž Krašček
ak7274@student.uni-lj.si

Bisera Miloshekska
bm1697@student.uni-lj.si

Žiga Zupanec
zz9698@student.uni-lj.si

ABSTRACT

V tem članku je predstavljen način, kako preko povezav v omrežju in s pomočjo napovedi stavnice točneje napovedati zmagovalca teniškega dvoboja. Obravnavali smo nekaj članov na to temo, ki predstavljajo naše izhodišče in nas motivirajo pri iskanju boljših rešitev od dosedaj znanih pristopov. V članku podrobno predstavimo dva algoritma. Prvi, Last Best Score (v nadaljevanju LBS), pogleda možne poti med dvema igralcema in izbere najnovejšo po času nastanka. Če je le-ta v prid favoritu se oceni stavnice prišteje določen procent. V nasprotem primeru, se ta ocena zmanjša za enak procent. Drugi, Časovno utežen Prestige Score (ČUPS v nadaljevanju) temelji na algoritmu PageRank. Algoritem vsakemu igralcu izračuna vrednost na podlagi izhodnih povezav, ki predstavljajo njegove zmage in Prestige Score vrednostjo poraženca v dvoboju. Vsaka povezava je utežena glede na preteklo število dni od dvoboja. Ocena stavnice se popravi za razliko med Prestige Scorom nasprotnikov. S predstavljenimi algoritmi smo dosegli profit v izmeri 291 enot, če bi stavili na vsak dvoboj od leta 2006 naprej.

Keywords

Networks, BFS, Bets, Tennis, PageRank

1. UVOD

Vse bolj pogosta je analiza socialnih sistemov zaradi vse prisotnosti raznih digitalnih podatkov ki povezujejo ljudi in njihovo obnašanje. V nekaterih primerih se analize razvijajo tudi v napovedovalnih matematičnih modelih na podlagi katerih lahko vnaprej določamo nadaljnje obnašanje nekega sistema. Realne socialne sisteme lahko predstavljamo v obliki matematičnih modelov s pomočjo uporabe grafov oziroma omrežja v katerih vozlišče predstavlja človeka ali skupino, povezava med dvemi vozlišči pa predstavlja odnos oziroma dogodek v katerem obe vozlišči sodelujeta. Takšna predstava socialnega sistema je zelo primerna predvsem za športne igre, kjer lahko analiziramo tekmovanja med različnimi igralci oziroma ekipami, ter njihove napredke in padce

v določenih časovnih obdobjih. Tako je tudi v tenisu, na primeru katerega bomo predstavili našo analizo. Kot v ostalih športih tudi v tenisu igralci z večjim številom zmag pridobijo višji status na skupni lestvici.

2. DOSEDANJE DELO NA TEM PODROČJU

2.1 Learning Betting Tips from Users' Bet Selections [3]

Članek preučuje možnost izboljšanja napovedi, ki jo ponujajo stavnice s pomočjo nasvetov uporabnikov na forumu. Avtorji favorizirajo napovedi tistih napovedovalcev, katerih dosedanji izidi so bili nadpovprečni, tj. imajo profitabilen izkupiček med vplačanimi stavami in izplačanimi dobički na dolgi rok. Da dobijo najboljše napovedovalce, uporablja in med seboj primerjajo različne metode. Te so:

- Metoda prvega pristopa, kjer izberejo tiste stave, na katere je na enak izid stavilo največ ljudi.
- Metoda probit, kjer z regresijskim modelom ocenjujejo vrednost posameznih stav.
- Metoda k-najbližjih sosedov, kjer pogledamo, katere stave so si med seboj podobne bodisi po tipu športa, časovni razlike in/ali količini dodatnih informacij o stavi. Kakovost stave se oceni koliko izbrana stava izstopa od srednje vrednosti drugih stav v okolici izbrane stave. Druga ocena meri kakovost uporabnikov, ki so izbrali določeno stavu in je definirana kot razlika med oceno verjetnosti izida uporabnika in oceno izida, kot jo po-nuja stavnica. Tisti uporabniki, ki se bolj približajo napovedi stavnice imajo boljšo oceno.

Končna ugotovitev je, da že z uporabo metode prvega pristopa, torej stavimo le na tiste izide, katere je izbralo največ uporabnikov, dosežemo boljše rezultate, kot če bi naključno stavili na vse izmed ponujenih izidov. Modela k-najbližjih sosedov še nadalje izboljšata rezultate. Metoda probit daje najboljše rezultate, če je število namigov katero stavu izbrati majhno.

Menimo, da je glavna pomankljivost članka Learning Betting Tips from Users' Bet Selections, da se zanaša na napovedi anonimnih uporabnikov. Te informacije so nezanesljive in lahko sčasoma postanejo namenoma zavajajoče. Tudi do podatkov o zgodovini uporabnikovih odločitev, ki je podlaga za oceno kakovosti stave, je zelo težko ali pa kar nemogoče priti. Menimo še, da uporabniki ne vidijo informacij, ki se skrivajo v posrednih poteh med dvema vozliščema in primerjajo zgolj posamezne igralce med seboj.

2.2 Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis [2]

Članek se osredotoča na določanje rank-a igralcem med leti 1968 in 2010. Določen rank primerja z obstoječimi sistemi na tem področju. Igralce razporedi v usmerjen graf, kjer dvoboj med dvema igralcema predstavlja povezava j-i. Povezave so utežene glede na število zmag igralca j proti igralcu i. Tako omrežje ima vse lastnosti socialnih omrežij opisanih v literaturi. Ena takšnih lastnosti je tudi razmerje med število vhodnih in izhodnih povezav iz katerega lahko določimo ali ima vozlišče večji ali manjši središčni vpliv kot ostala vozlišča. Na tej lastnosti temelji metoda PageRank, ki jo avtor vzame kot osnovo za svojo metodo računanja rank-a teniških igralcev „prestige score“. Prestige score se izračuna iz treh delov. Prvi del določa seštevek moči izhodnih povezav glede na utez povezave in prestige score-a sosednjih vozlišč. Drugi del določa enakomerno razporeditev prestige score-a vsem igralcem. Zadnji del je popravek za vozlišča brez izhodnih povezav in preprečuje, da bi takšna vozlišča delovala kot ponor. Dodatno avtor metodo razvije na grafu posameznega teniškega turnirja, kjer se prestige score izračuna na število tekem in zmag, ki jih igralec dosegel. Prav tako avtor pokaže da je mogoče razvito metodo uporabiti za določanje najboljših igralcev v posameznem letu ali na turnirjih z enako podlago, itd. Rezultati razvite metode so boljši, kot tisti uveljavljenih sistemov za uvrščanje igralcev. Vendar pa metoda ne upošteva par ključnih lastnosti, predvsem časovne komponente. Metoda za izračun rank-a upošteva predvsem število zmag zmagovalca in ne toliko proti komu so bile zmage dosegene in kdaj. Tako so na vrhu lestvice igralci z že zaključenimi karierami, ki so mlajše nasprotnike premagovali v najboljši formi. Ti mlajši nasprotniki imajo sedaj proti njim negativno povezano čeprav bi jih sedaj najverjetnejše premagali. Z algoritmom ČUPS vsako povezavo utežimo glede na čas nastanka povezave. Starejše povezave imajo manj vpliva na končno vrednost posameznega igralca.

3. OPIS PROBLEMA

Predlagamo, da kot v članku Learning Betting Tips from Users' Bet Selections, za privzeto vrednost izida uporabimo verjetnost, ki jo ponujajo stavnice. To verjetnost bomo korigirali z našo oceno, kar je bolj podrobno predstavljeno v naslednjih poglavjih.

Trdimo, da lahko z ugotavljanjem poti med dvema igralcema, v nadaljevanju igralca poimenujemo s p_1 in p_2 , v omrežju izboljšamo napoved končnega rezultata. Iz neposrednih povezav med dvema igralcema lahko ocenimo kateri igralec je boljši tudi če se igralca medsebojno nista še nikoli pomerila. Poleg števila skokov med dvema igralcama je pomemben tudi čas ko je bil dvoboj odigran. Starejši dvoboji ne prikažejo trenutnega stanja pripravljenosti igralca. Prvi pristem pristopu z iskanjem najkrajše in najmlajše poti ocenimo zmagovalca, medtem ko pri drugem izračunamo PageRank vsem igralcem in glede na njihovo oceno popravimo napoved. Naš pristop tudi ne potrebuje nobenih notranjih informacij. Podatki, ki jih potrebujemo za naš napovedni model so prosti dostopni in javni ter jih ni mogoče prirediti ali pa zapreti vir preko katerih prihajajo, saj bi to pomenilo izločitev javnosti (gledalcev), kar je nepredstavljivo.

4. OPIS PODATKOV

4.1 Podatki

Za izgradnjo omrežja bomo uporabili podatke o teniških dvobojih dobljenih na spletni strani (<http://www.tennis-data.co.uk/alldata.php>). Posamezna vrstica predstavlja dvoboj dveh igralcev. V posamezni vrstici je vrsta turnirja, lokacija, naziv turnirja, podlaga igrišča, ali se igra na prostem ali v dvorani, ime zmagovalca, ime poraženca, rang zmagovalca, rang poraženca, število točk zmagovalca, število točk poraženca, rezultati po posameznih nizih in verjetnosti izidov na izbranih športnih stavnicah.

4.2 Omrežje

Vsi igralec je v omrežju predstavljen kot posamezno vozlišče. Vsak dvoboj je predstavljen kot povezava med dvema igralcema. Smer povezave določa zmagovalca. Npr. povezava, ki gre iz p_1 v p_2 pomeni, da je zmagal p_1 . Dva igralca imata lahko med seboj več povezav v obe smeri.

V tabeli 1 so prikazane lastnosti prvotnega grafa G in pa lastnosti grafov, ki smo jim odstranili povezavo med p_1 in p_2 . Za graf \bar{H} so prikazane povprečne vrednosti iteracije preko 2605 grafov, ki predstavljajo 2605 napovedi (bolj podrobno je ta del predstavljen v poglavju Metode in tehnike). Pri grafu \bar{T} smo upoštevali samo tiste grafe, ki se razlikujejo od grafa G . Takih grafov je 637 (četrtina). Iz tabele je razvidno, da se ključne lastnosti grafov skoraj ne spreminja. Slika 1 je vizualizacija grafa teniških dvobojev v letu 2013. Kaže porezdelitev zmag in porazov desetih najboljših igralcev na lestvici ATP. Vizualizacija dvobojev za leta od 2008 do 2012 je v prilogi pod poglavjem Grafi in Krivulje ROC.

Table 1: Ključne lastnosti grafov.

	G	\bar{H}^1	\bar{T}^2
število povezav:	2675	2674.66	2673.60
nakopičenost:	0.15739	0.15736	0.15727
premer:	6	5.61	5.61
trojica:	6029	6025.84	6016.06

5. METODE IN TEHNIKE

Predlagana algoritma v precejšnji meri posnemata t.i. Teorijo statusa. Bistvo algoritmov je v napovedovanju povezave med p_1 in p_2 . Od Teorije status se razlikuje po tem, da ne gleda nujno trojice, ampak uporabi tudi druge poti med tema vozliščema (slika 2). Vsaka vrednost (napoved $P'(x)$) je sestavljena iz napovedi stavnice $P(x)$, ki je korigirana z delto ($\Delta p(x)$), le-ta pa je rezultat analize povezav na poti med dvema vozliščema in je specifična glede na uporabljen algoritmom (enačba 1).

$$P'(x) = P(x) + \Delta p(x). \quad (1)$$

5.1 Last Best Score

Osnovni cilj algoritma je izboljšati napoved izida med dvema igralcema, pri čemer je tudi uspešen. Za izboljšanje rezultatov algoritma smo pri testiranju uporabili več variacij

¹Povprečje grafov po odstranitvi povezave (normaliziran).

²Povprečje grafov po odstranitvi povezave (filtriran, normaliziran) - upoštevajo se samo tisti grafi, ki se razlikujejo od grafa G.

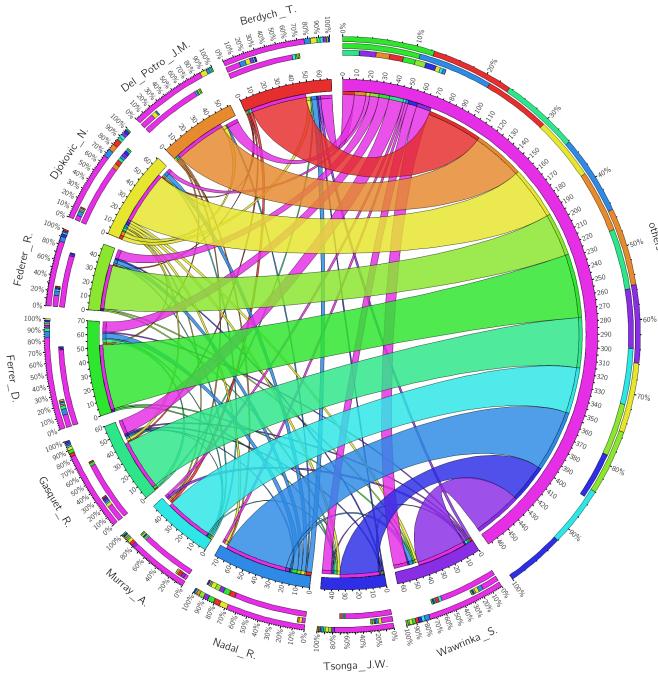


Figure 1: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2013.

osnovne metode. Osnovna metoda abstrahira vse posredne povezave med dvema igralcema z BFS preiskovanjem učnega grafa. Potem izloči eno posredno povezavo v kateri je favorit zmagal ter eno posredno povezavo, v kateri je slabši igralec zmagal. Nato ustrezno oceni te dve napovedi oz. povezave ter se na podlagi te odloči, kako napovedati izid naslednjega dvoboja med igralcem. V ovisnosti od tega na kakšen način izbere posredno povezavo med dvema igralcema ter kako jo ocenjuje, smo razvili več variacij osnovne metode, ki jih bomo podali v nadaljevanju.

Z BFS preiskovanjem učnega grafa dobimo vse posredne povezave, ki obstajajo med dvema igralcema. Izmed vseh povezav izberemo tisto povezavo, ki je najnovejša po času in primerjamo povezavo, v kateri je zmagal favorit, s povezavo, v kateri je bil poražen, na podlagi časa nastanka teh dveh povezav. Novejše povezave so bolj pomembne kot starejše, ker lahko bolj natančno napovedujejo izide. Starost povezave določamo tako, da števamo čas, ki je potekel od dvoba do danes za vsako posamezno neposredno povezavo, ki je del posredne povezave(*path*). Ta ocena starosti vrne najbolj verodostojne vrednosti. Če bi vzeli najdaljši čas, je lahko povezava, ki je sestavljena iz samih novejših povezav in samo ene starejše v primerjavi z ostalim, ocenjena kot manj vredna. Na ta način upoštevamo tudi dejstvo, da so daljše poti manj vredne kot krajše, saj bodo imele večji časovni seštevek preko vseh neposrednih povezav, ki jih sestavljajo. Torej je ključna formula, na kateri temeljijo napovedi, naslednja:

$$r_i = \min_{p \in P} \left(\sum_{e \in p} (T_{curr} - T_e) \right) \quad (2)$$

kjer je P množica vseh indirektnih poti med začetnim in končnim vozliščem, p predstavlja eno pot iz te množice, T_{curr} je trenutni čas, T_e pa čas ko se je zgodil konkretni

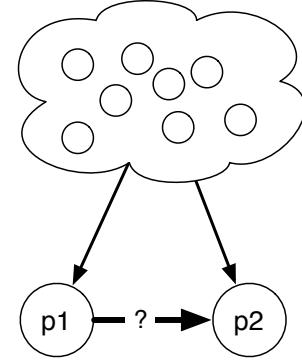


Figure 2: Napoved povezave med p_1 in p_2 preko n vozlišč.

dvoboja e ki je del indirektnih poti p .

Po zgoraj navedeni formuli izračunamo vrednost povezave v kateri je i zmagovalec, r_w , ter vrednost obrnjene povezave v kateri je i poraženec, r_l . Na koncu še izračunamo naš popravek k originalnim rezultatom.

$$\Delta p = (r_w - r_l) \quad (3)$$

5.2 Časovno utežen Prestige Score

ČUPS temelji na algoritmu PageRank [4]. Osnovna ideja algoritma je da oceni posamezno vozlišče na osnovi povezav, ki kažejo nanj in vrednosti sosednjih vozlišč. Osnovni algoritmom smo nadgradili tako da ne upošteva samo PageRank povezanih vozlišč in števila izhodnih povezav posameznega vozlišča ampak tudi čas ko je povezava nastala. V našem primeru to pomeni, da so novejši dvoboji med dvema igralcema bolj pomembni kot tisti iz prejšnje sezone ali še starejši. Ne smemo pa starejših dvobojev kar zanemariti saj veliko dobroih rezultatov še vedno pomeni da gre za dobrega igralca kljub kakšnemu presenetljivemu porazu v zadnjem času.

$$r_j = \sum_{j \rightarrow i} \beta \frac{w_{ji} r_i}{\sum w_{in}} + (1 - \beta) \frac{1}{n} \quad (4)$$

Rank igralca j , r_j , je vsota njegovih izhodnih povezav (zmag). Vsaka izhodna povezava ima utež w_{ji} . Utež povezave izračunamo glede na število dni, ki je preteklo od dneva vzpostavitve povezave, normalizirano s številom dni, ki so pretekli od najstarejše povezave v omrežju. Produkt w_{ji} in Rank-a vozlišča i je normaliziran z vsoto uteži vseh vhodnih povezav v vozlišče i . Drugi del $(1 - \beta) \frac{1}{n}$ preprečuje napake v izračunu ko pridemo do vozlišč brez izhodnih povezav. Takšna vozlišča predstavljajo igralci z majhnim številom dvobojev in brez zmag. Za parameter β smo izbrali vrednost 0.85.

Izračunane PageRank-e vsakega vozlišča uporabimo za izračun parametra Δp , kjer je r_w PageRank najverjetnejšega zmagovalca in r_l PageRank najverjetnejšega poraženca (enačba 3).

6. REZULTATI

6.1 Testno okolje

Algoritma (klasifikatorja) se učita na učni množici U_i in napovedujeta rezultate teniških dvobojev na množici T_{i+1} , pri čemer indeks i predstavlja koledarsko leto začenši z letom 2005 za učno množico. Klasifikator vrne vektor vrednosti.

6.2 Ocena rezultatov

Rezultate smo ocenili s krivuljo ROC (Receiver operating characteristic), ki analizira razmerje med senzitivnostjo in specifičnostjo [1]. Vodoravna os predstavlja 1 - specifičnost (False positive rate), navpična pa senzitivnost (True positive rate). Kakovost klasifikatorja predstavlja površina pod krivuljo ROC - AUC (Area under curve). Klasifikator je uporabnejši, če je površina pod krivuljo kar se da stran od 0.5 na intervalu od $[0, 1.0]$. Naše izhodišče s katerim se primerjamo je napoved stavnice (modra krivulja ROC na sliki 3), naš cilj pa je zgraditi model, ki ima površino pod krivuljo večjo od 0.69. Razred, ki ga napovedujemo zavzema vrednosti 0, kar pomeni, da je zmagal p_1 oziroma 1, če zmaga p_2 .

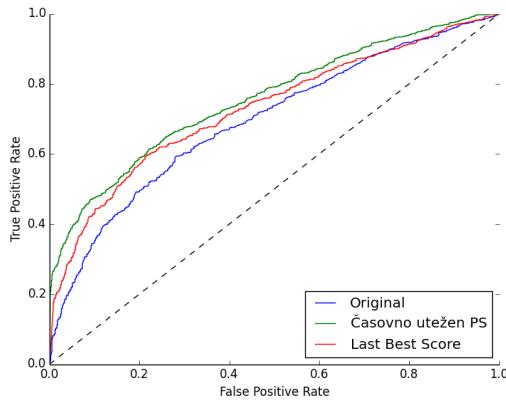


Figure 3: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2013.

6.3 Razprava

Če uporabimo algoritem LBS lahko dosežemo do 5% ($AUC_{stavnice} = 0.69$, $AUC_{LBS} = 0.73$) izboljšavo originalnega pristopa, ki je bila izmerjena na primeru napovedovanja rezultatov za leto 2013, na podlagi učne množice podatkov za leto 2012. Z istim algoritmom in uporabo učne množice podatkov za leto 2011 smo napovedali rezultate z 1.4% ($AUC_{stavnice} = 0.69$, $AUC_{LBS} = 0.70$) izboljšavo originalnega pristopa za napovedovanje. In za zadnji primer 2011/2010 smo izmerili 3.6% ($AUC_{stavnice} = 0.67$, $AUC_{LBS} = 0.70$) izboljšavo.

Mogoče bi ta algoritem deloval bolje, če bi vzeli rezultate za vsa pretekla leta in bi imeli tako večjo učno množico podatkov. Vendar takšen popravek ne bo veliko vplival na rezultat, saj je algoritem osredotočen na krajše ter mlajše povezave. Lahko torej rečemo, da je naš algoritem optimiziran glede na svojo osnovno idejo in cilj, ki ga želimo doseči.

ČUPS doseže 9% ($AUC_{stavnice} = 0.69$, $AUC_{CUPS} = 0.75$) izboljšavo napovedi stavnice.

Iz dobljenih rezultatov obih predstavljenih algoritmov (Tabela 2) lahko ugotovimo, da smo izide posameznih dvobojev napovedali natančneje kot stavnice. Tu gre predvsem za

boljše napovedi v primerih ko veliko boljši igralec premaga veliko slabšega igralca po ATP lestvici in lahko z veliko govorstvo trdimo da bo favorit zmagal z verjetnostjo 1. Takšni dvoboje so značilni za prve kroge turnirjev. Morda v teh primerih želijo stavnice sprožiti špekulacije o zmagovalcu in s tem povečati vplačila na slabšega igralca.

Krivulje ROC za leta od 2007 do 2012 so v prilogi pod poglavjem Grafi in Krivulje ROC.

7 REFERENCES

- [1] Kononenko, Igor, and Marko Robnik Šikonja. "Inteligentni sistemi." Založba FE in FRI, 2010.
- [2] Radicchi, Filippo. "Who is the best player ever? A complex network analysis of the history of professional tennis." PloS one 6.2 (2011): e17249.
- [3] Štrumbelj, Erik, Marko Robnik Šikonja, and Igor Kononenko. "Learning Betting Tips from Users' Bet Selections." Springer Berlin Heidelberg, 2009.
- [4] Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: bringing order to the web." (1999).

Table 2: Rezultati.

leto	št. dvobojev	pravilno napovedani dvoboji				Priigran denar			AUC		
		stavnice	LBS	ČUPS	stavnice	LBS	ČUPS	stavnice	LBS	ČUPS	
2006	2865	1975	1951	1981	-145	30	-117	0,67	0,67	0,68	
2007	2768	1982	1920	1988	-60	20	-18	0,66	0,67	0,69	
2008	2679	1890	1843	1918	-86	14	55	0,66	0,67	0,72	
2009	2711	1906	1886	1935	-150	11	-18	0,69	0,75	0,70	
2010	1720	1199	1173	1234	-81	-18	64	0,67	0,75	0,68	
2011	2664	1923	1920	1931	-65	129	-3	0,67	0,70	0,72	
2012	2605	1870	1839	1878	-99	39	-61	0,69	0,70	0,73	
2013	2424	1700	1703	1728	-136	66	76	0,69	0,73	0,75	

APPENDIX

A. GRAFI IN KRIVULJE ROC

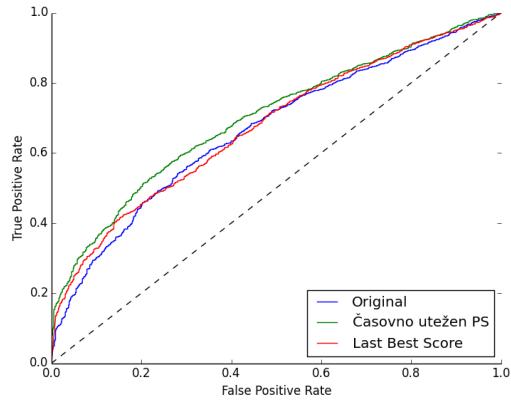


Figure 5: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2007.

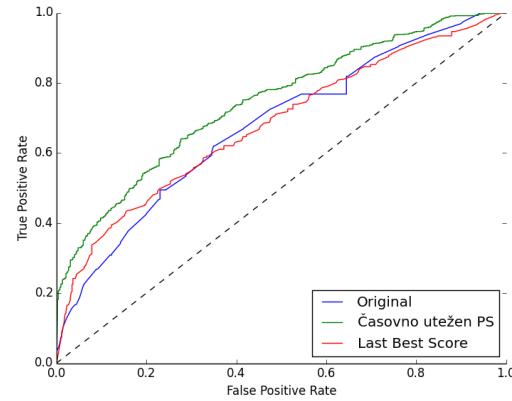


Figure 8: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2010.

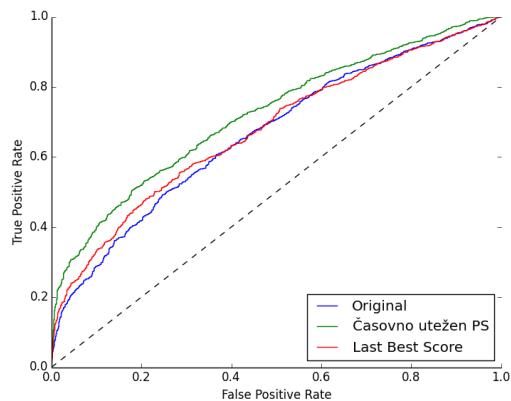


Figure 6: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2008.

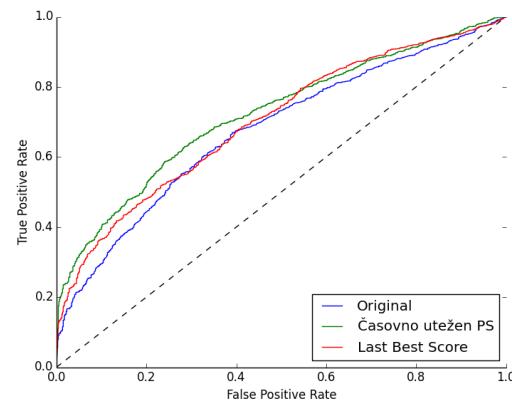


Figure 9: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2011.

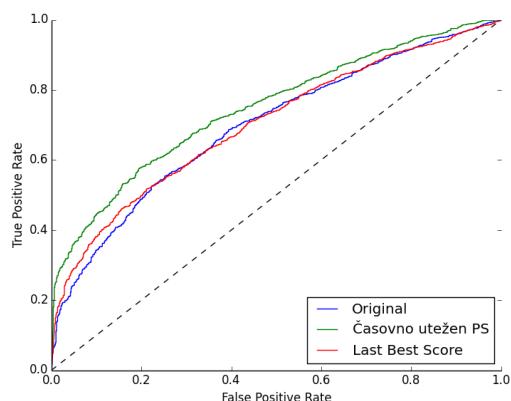


Figure 7: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2009.

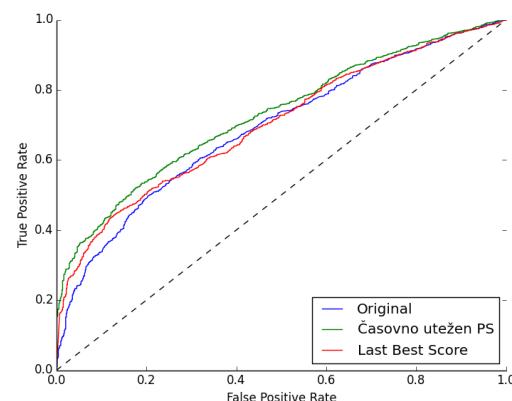


Figure 10: Krivulje ROC. Točnost napovedovanja algoritmov za leto 2012.

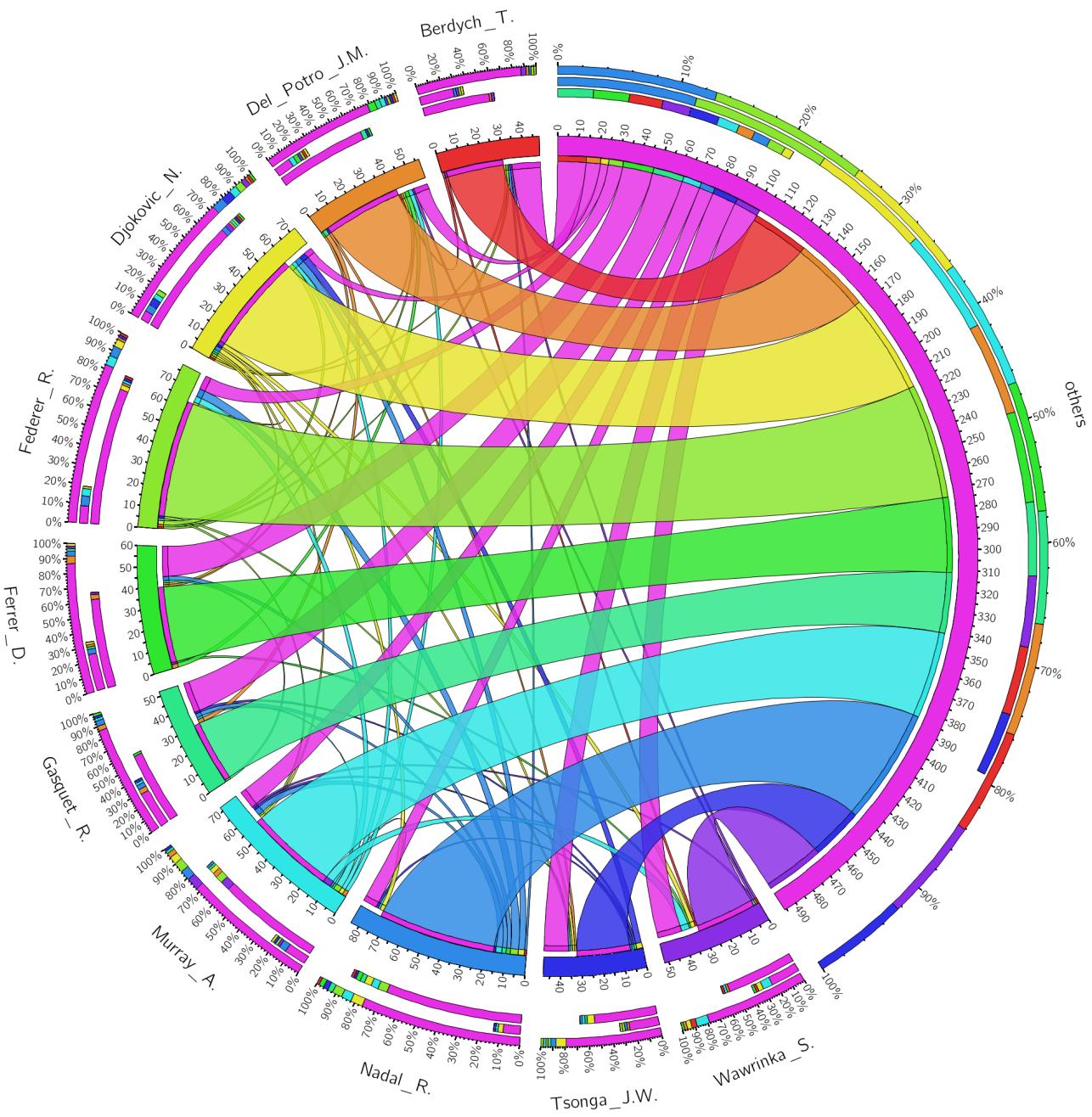


Figure 10: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2008.

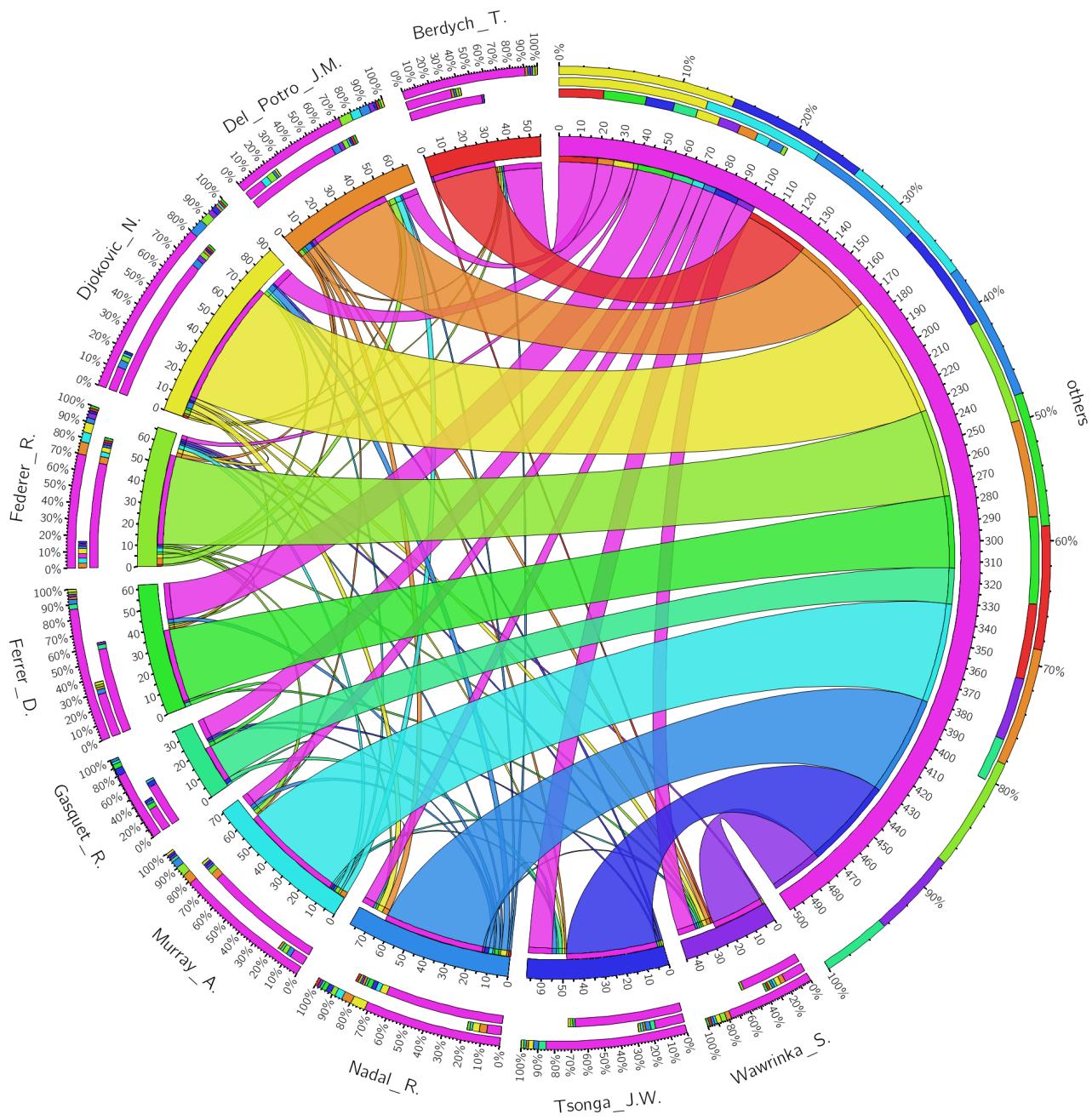


Figure 11: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2009.

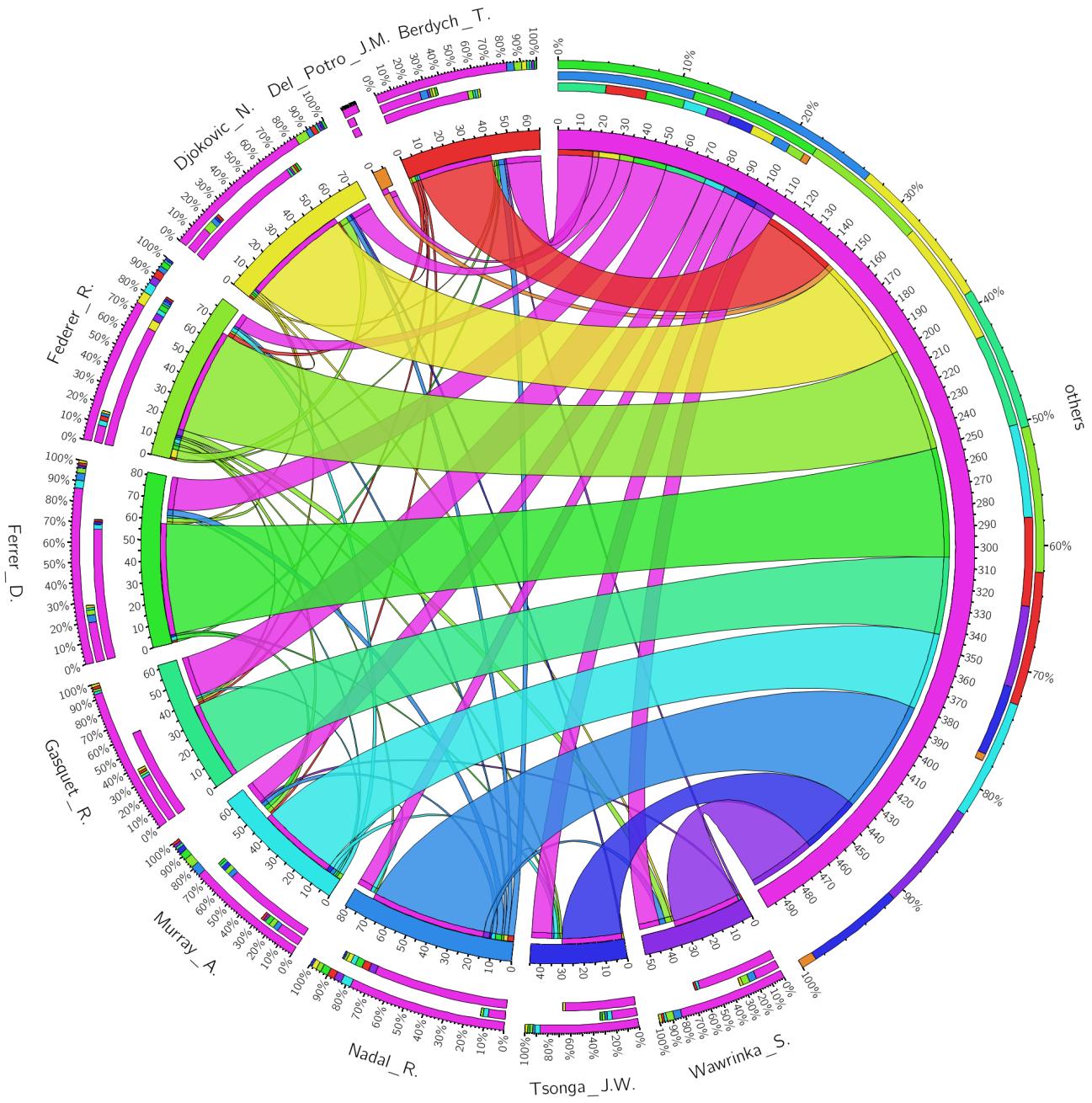


Figure 12: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2010.

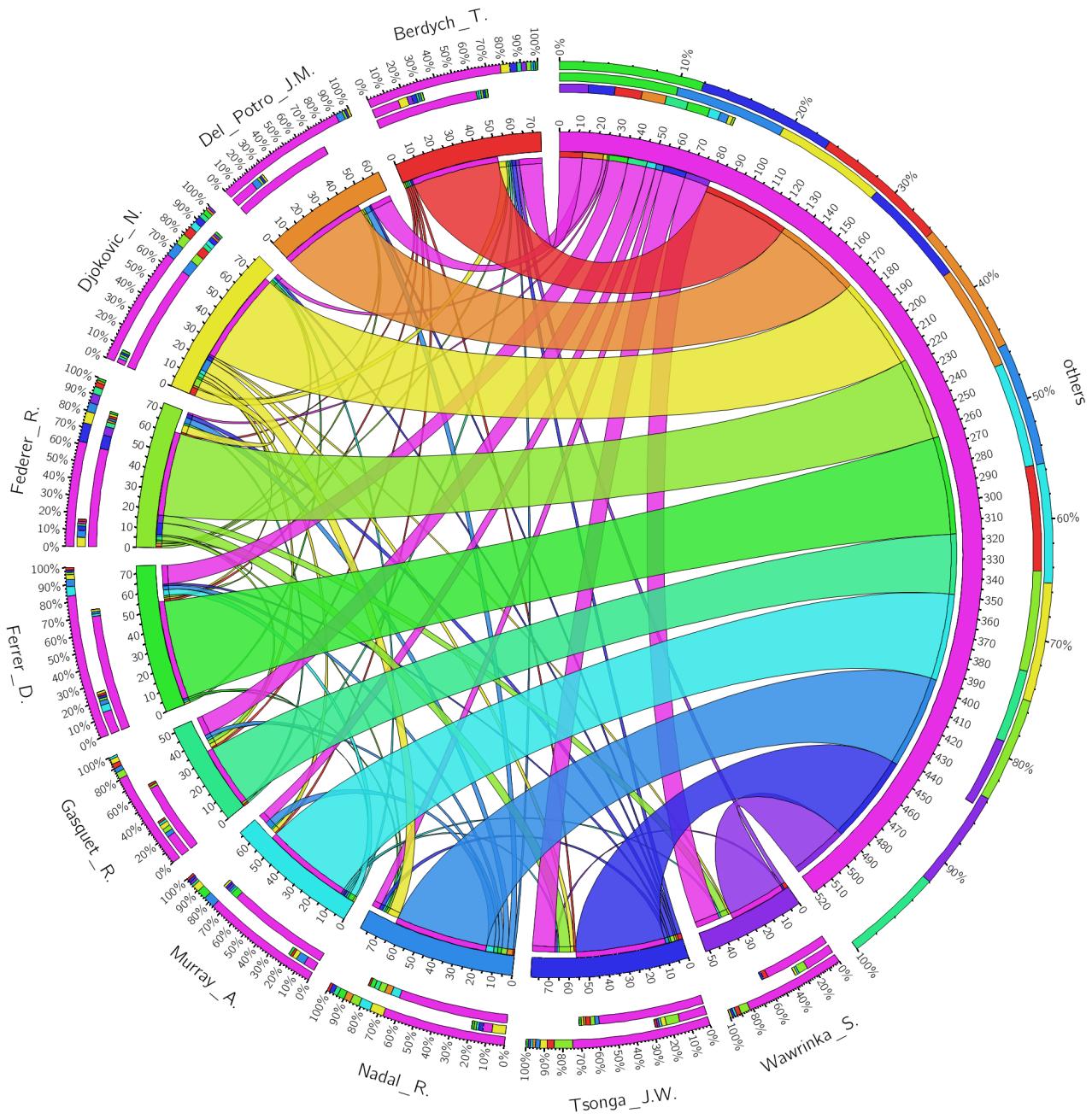


Figure 13: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2011.

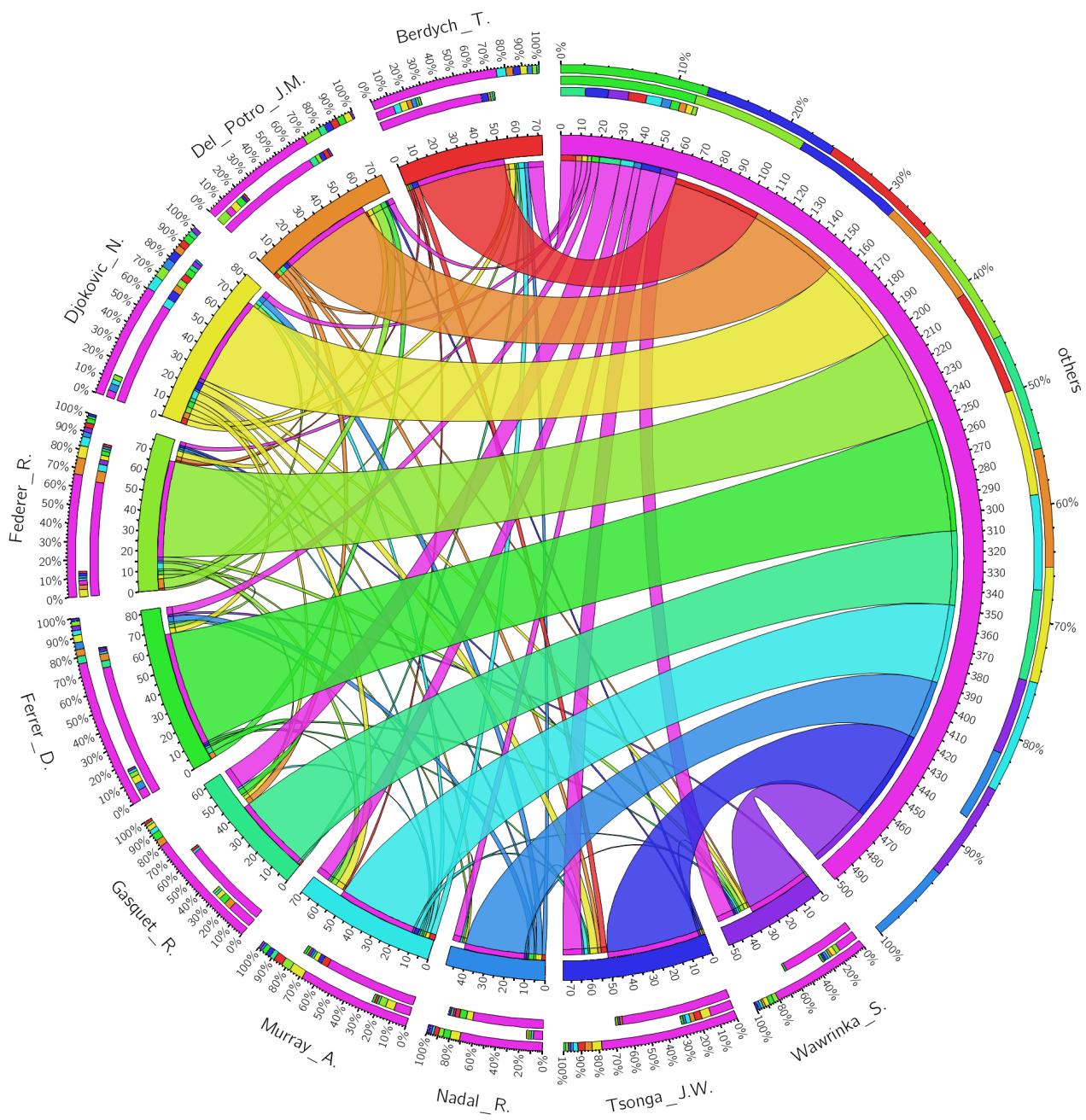


Figure 14: Porazdelitev zmag najboljših deset igralcev na lestvici ATP v letu 2012.