Network Science – Homework 2
Ranking and Communities of Men's Single Tennis Matches

Umberto Michieli ID: 1150780
umberto.michieli@studenti.unipd.it

I. Introduction and Related Works

The main interest in analysis of tennis networks is directed towards the implementation of new ranking techniques instead of the ATP ranking system, together with their relative predictive power when a new match is played. There are no contributions in literature, instead, for what concerns link prediction (e. g. who are the most probable players to play against given that they did not played against before?) and community detection.
Radicchi in [1] is the first who applied the PageRank (PR) algorithm to tennis network thus identifying *Jimmy Connors* as the most valuable player in the tennis history. Dingle et al. in [2] applied the previous work to ATP and WTA (*Women's Tennis Association*) matches and they also provided a simple comparison on the basis of predictive power. In [3] the authors proposed yet another ranking method applying PageRank to the subgraph of the Top-100 players. In [4] many ranking methods, both through network and Markov chains analysis, have been proposed and verified by means of prediction power.

Also few other non-network-related approaches have been proposed so far: for example in [5] statistical models have been tested to improve the current ranking system, or in [6] the authors applied a novel method exploiting neural networks based on 22 features and achieving a 75% benefit in prediction through those techniques.

In this paper we are going to apply some algorithms aiming at confirming the present literature on ranking methods, thus seeing how active tennis players have improved their overall prestige over the recent years. However, at the same time, we aim at providing some useful considerations about link prediction and communities detection.

II. Results and Discussion

This section sometimes assumes that the reader has a broad knowledge about tennis matches dynamics and points assignment; if it is not the case please refer to the first Homework.

A. Ranking Methods and Predictive Power

1) Preliminaries on Ranking Algorithms: The analysis shown in this section assumes the direct representation of the network where the loser player has an edge to the winner player and the weight corresponds to the number of matches won by the winner.
The link analysis methods we are going to investigate are: *Hubs and Authorities* (HITS algorithm, *Hyperlink-Induced Topic Search* discussed in [7]), simple PageRank and PageRank with teleportation (see [8] and [9] as references).

The idea on which HITS algorithm is based regards the definitions of hubs and authorities. Authorities are nodes with a high number of edges pointing to them, hubs are nodes which link to many authorities; in our scenario, intuitively, we expect that authorities are often associated with the most successful players (because they won against a wide gamma of players), while the hubs with *mediocre* players with a long career. More formally we can compute the authority-scores $\mathbf{a}$ and the hub-scores $\mathbf{h}$ respectively as:

$$\mathbf{a} = \frac{\mathbf{A}\mathbf{h}}{||\mathbf{A}\mathbf{h}||} \qquad \mathbf{h} = \frac{\mathbf{A^T}\mathbf{a}}{||\mathbf{A^T}\mathbf{a}||}$$

where we assumed the adjacency matrix to be the transposed version of the one presented in Homework 1 (i.e. here an entry $a_{ij} = 1$ means that there is a link from node j to node i). Notice that $a_i \geq 0$ and $h_i \geq 0\ \forall i$. The problem can be solved through power iteration with convergence parameter ϵ.

The rationale behind PageRank is that of a random walk along the graph and the *prestige* score $\mathbf{p}$ for each player is determined as the probability of being at that node in stationarity conditions. The t-th update of $\mathbf{p}$ goes as:

$$\mathbf{p}_t = \mathbf{M}\mathbf{p}_{t-1}$$

where $\mathbf{M}$ is the column stochastic adjacency matrix.

The simple PageRank algorithm is affected by some undesirable problems. For example it would end up in dead ends, although it is not the case because who win one match will surely lose one other (unless the player plays only tournaments winning all of them, which never happens); and there also might be periodic behavior looping in cycles, which is somehow reasonable to expect since we are considering very different tennis epochs. Thus we can add to the model the possibility of not to follow the behavior but to jump to a random node in the network with a probability $\alpha \in (0, 1)$. Hence the t-th update step of $\mathbf{p}$ becomes:

$$\mathbf{p}_t = \breve{\mathbf{M}}\mathbf{p}_{t-1} = (1-\alpha)\mathbf{q_1}\mathbf{1^T}\mathbf{p}_{t-1} + \alpha\mathbf{M}\mathbf{p}_{t-1}$$

where $\mathbf{q_1}$ is the stochastic teleportation vector and we assumed it to be $\mathbf{q_1} = \frac{1}{N}\mathbf{1}$ (equal probabilities), with $\mathbf{1}$ column vector of N ones; α is a damping factor typically set to 0.85 (this

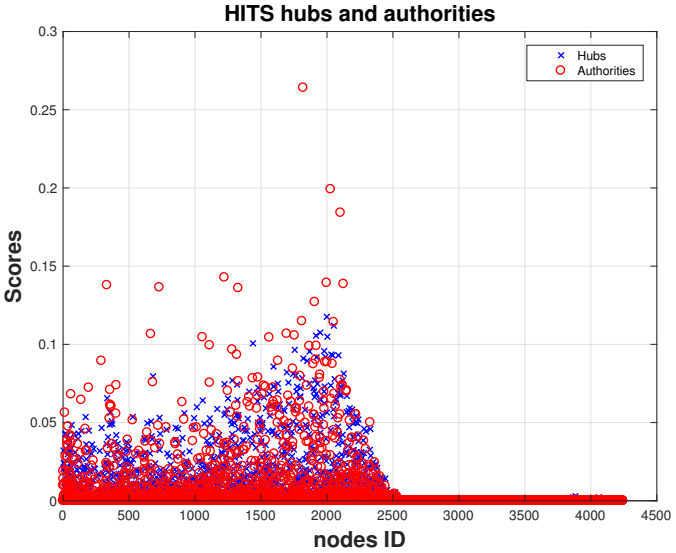

Figure 1. Hubs and authorities scores of HITS algorithm.

is due to historical reasons as proposed in the original paper [9] and for the sake of comparisons with other works). This considerations led us to a much simpler iteration procedure than the one proposed in [1] and [2], although they are equivalent. The simplifications are made possible thanks to the observation regarding the absence of sinks-like nodes and to a compact vectorial expression.

2) Discussion of results: Hubs and authorities scores are reported in Figure 1, where we can see that nodes ID corresponding to players who only have lost matches (the last ones) have zero authority score, because there are no links pointing to them, but possibly non-zero hub score.
The names of the Top-20 hubs and authorities are reported in the second and third columns of Table I for $\epsilon = 10^{-8}$; few changes happen varying this parameter and most of them not in the very first positions. Though there are no reference literature of HITS applied to tennis, nevertheless from the table we can confirm our previous intuition and also realize that those concepts are somehow similar to what we already discussed in the first homework talking about in-degree and out-degree hubs. Indeed, we can recognize that in-degree hubs and out-degree hubs are placed in the first positions respectively of authorities and of hubs, although not in the precise order. In the first homework we found as the five highest in-degree hubs ($\approx$ *authorities*): *Jimmy Connors, Roger Federer, Ivan Lendl, Guillermo Vilas* and *John McEnroe*. While the five highest out-degree hubs ($\approx$ *hubs*) were: *Fabrice Santoro, Feliciano Lopez, Mikhail Youzhny, Guillermo Vilas* and *John McEnroe*.

In terms of complexity we expect at most $t_{\max}$ iterations for the HITS algorithm to converge, where:

$$t_{\max} = \left\lceil -\frac{\ln(\epsilon) - \ln(\sqrt{N})}{2\ln(d_1/d_2)} \right\rceil$$

with d_1 and d_2 being the eigenvalues associated with the two highest eigenvectors of $\mathbf{M} = \mathbf{A}\mathbf{A}^{\mathbf{T}}$. Setting $\epsilon = 10^{-8}$ we find $t_{\max} = 180$ iterations, but in order to converge just $t = 100$ iterations are needed. In Table II is reported the computational time for the convergence of this algorithm.

Finally the predictive power of HITS based on authorities, defined as the percentage of times the higher ranked player will win, is reported in Table III for two different new dataset: first we considered the new matches played between September 2017 and November 2017 extremal included, which also concluded the 2017 tennis season, for a total number of 431 matches (those are independent data since are not considered in the training dataset); then we considered all the matches played in 2017 for a total number of 2633 matches. As *Modified ATP* we mean that the player who has obtained more ATP points in his career will win. As regards the smaller dataset, *HITS* behave well and similar to the *Modified ATP* system, while for the largest dataset the performances deteriorate.

The players prestige scores obtained through PageRank algorithms are plotted in Figure 3, where, similarly as before, we can see that the players who only lose matches have the same minimum value.
The Top-20 tennis players identified by those algorithm are reported in fourth and fifth columns of Table I. Without teleportation the podium remains the same as in the authorities of HITS algorithm, then there are many differences. With teleportation we are able to break the loops leading to the biggest authorities and achieve a fairer result.
Moreover those results are quite robust and they do not vary much by setting another value to α.
The fifth column of this table should confirm the goodness of our model being the results very similar to the ones reported in [1]. Actually this table can update the one shown in the mentioned paper where were used data up to 2010 and the resulting top-players were: *Jimmy Connors, Ivan Lendl, John McEnroe, Guillermo Vilas, Andre Agassi, Stefan Edberg, Roger Federer, Pete Sampras, Ilie Nastase, Bjorn Borg, Boris Becker, Arthur Ashe, Brian Gottfried, Stan Smith, Manuel Orantes, Michael Chang, Roscoe Tanner, Eddie Dibbs, Harold Solomon* and *Tom Okker*. Comparing those results with the fifth column of Table I we can appreciate how the players who are still in activity (*Roger Federer, Rafael Nadal, Novak Djokovic, Andy Murray* and *David Ferrer*) have gained some positions in the overall ranking. It should be stressed that those results are inherently biased toward already retired players, since still active players did not played all the matches of their career; this bias, however, could be removed, for example considering only matches played the same year, as done in [1]. For example, last year (2017) ranking comparisons are reported in the last three columns of Table I where we see that authorities and PageRank involve mostly the same players in slightly different orders, also with respect to the Official method.

Moreover, in Figure 2 probability distributions of prestige scores obtained through the proposed algorithms are shown. Notice that both $\sum_{i=1}^{N} \text{prestige}_i = 1$ and $\sum_{i=1}^{N} \text{P}[\text{prestige}_i] = 1$, but in this plot the prestige values are

Rank	Authorities	Hubs	Simple PR	PR with Teleport.	Authorities 2017	PR with Teleport. 2017	Official ATP 2017
1	***Roger Federer***	*David Ferrer*	***Roger Federer***	***Jimmy Connors***	***Rafael Nadal***	***Roger Federer***	***Rafael Nadal***
2	***Rafael Nadal***	*Tomas Berdych*	***Rafael Nadal***	***Ivan Lendl***	***Roger Federer***	***Rafael Nadal***	***Roger Federer***
3	***Novak Djokovic***	*Feliciano Lopez*	***Novak Djokovic***	***Roger Federer***	*Alexander Zverev*	*Alexander Zverev*	*Grigor Dimitrov*
4	***Andre Agassi***	*Mikhail Youzhny*	***Ivan Lendl***	***John McEnroe***	*Grigor Dimitrov*	*David Goffin*	*Alexander Zverev*
5	*David Ferrer*	*Fernando Verdasco*	***Andre Agassi***	***Rafael Nadal***	*David Goffin*	*Grigor Dimitrov*	*Dominic Thiem*
6	***Andy Murray***	*Fabrice Santoro*	***Pete Sampras***	***Novak Djokovic***	*Dominic Thiem*	*J. M. Del Potro*	*Marin Cilic*
7	***Jimmy Connors***	*Tommy Haas*	***Andy Murray***	*Guillermo Vilas*	*Marin Cilic*	*Dominic Thiem*	*David Goffin*
8	***Ivan Lendl***	*Jarkko Nieminen*	***Jimmy Connors***	***Ilie Nastase***	*Jack Sock*	*Jack Sock*	*Jack Sock*
9	***Pete Sampras***	*Tommy Robredo*	*David Ferrer*	***Andre Agassi***	*Roberto B. Agut*	*Nick Kyrgios*	*Stan Wawrinka*
10	***Andy Roddick***	*Philipp Kohlschreiber*	***Stefan Edberg***	***Bjorn Borg***	*J. M. Del Potro*	*Marin Cilic*	*Pablo C. Busta*
11	***Lleyton Hewitt***	*Andreas Seppi*	***Boris Becker***	***Stefan Edberg***	*Pablo C. Busta*	*Sam Querrey*	*J. M. Del Potro*
12	*Tomas Berdych*	*Stanislas Wawrinka*	***Andy Roddick***	***Pete Sampras***	*Diego Schwartzman*	*Roberto B. Agut*	***Novak Djokovic***
13	***Carlos Moya***	*Richard Gasquet*	***John McEnroe***	***Arthur Ashe***	*Lucas Pouille*	*Jo-Wilfried Tsonga*	*Sam Querrey*
14	***John McEnroe***	*Nikolaj Davydenko*	***Lleyton Hewitt***	***Boris Becker***	*Tomas Berdych*	*Giles Muller*	*Kevin Anderson*
15	*Tommy Haas*	***Roger Federer***	*Tomas Berdych*	*Stan Smith*	*Jo-Wilfried Tsonga*	***Novak Djokovic***	*Jo-Wilfried Tsonga*
16	***Stefan Edberg***	*Radek Stepanek*	*Michael Chang*	*Brian Gottfried*	***Novak Djokovic***	*Tomas Berdych*	***Andy Murray***
17	***Yevgeny Kafelnikov***	*Jonas Bjorkman*	***Yevgeny Kafelnikov***	*Manuel Orantes*	*Milos Raonic*	*Milos Raonic*	*John Isner*
18	***Boris Becker***	***Carlos Moya***	*Goran Ivanisevic*	***Andy Murray***	*Philipp Kohlschreiber*	*Kevin Anderson*	*Lucas Pouille*
19	*Nikolaj Davydenko*	***Andy Murray***	***Carlos Moya***	*David Ferrer*	*Kevin Anderson*	*Damir Dzumhur*	*Tomas Berdych*
20	*Tommy Robredo*	*Ivan Ljubicic*	*Tommy Haas*	*Roscoe Tanner*	*John Isner*	*Alberto R. Vinolas*	*Roberto B. Agut*

Table I
RANKING METHODS OUTCOMES; THE BOLD NAMES ARE PLAYERS WHO HAVE BEEN AT THE FIRST ATP POSITION DURING THEIR CAREER. PLAYERS LIKE *Manuel Orantes, Guillermo Vilas* AND *David Ferrer* ARE OFTEN REFERRED TO AS ETERNAL SECOND BEST AND IN THE COLLECTIVE IMAGINATION THEY DESERVED TO BE NUMBER ONE OF THE RANKING. UNDERLINED NAMES IN THE LAST COLUMNS ARE THE ONES RANKED IN THE SAME POSITION AS IN OFFICIAL ATP RANKING.

Algorithm	# of Iterations	Time [ms]
HITS	120	56
Simple PageRank	185	180
PageRank with Teleportation	53	164

Table II
NUMBER OF ITERATIONS AND TIME FOR CONVERGENCE OF THE PROPOSED RANKING ALGORITHMS WITH $\epsilon = 10^{-8}$.

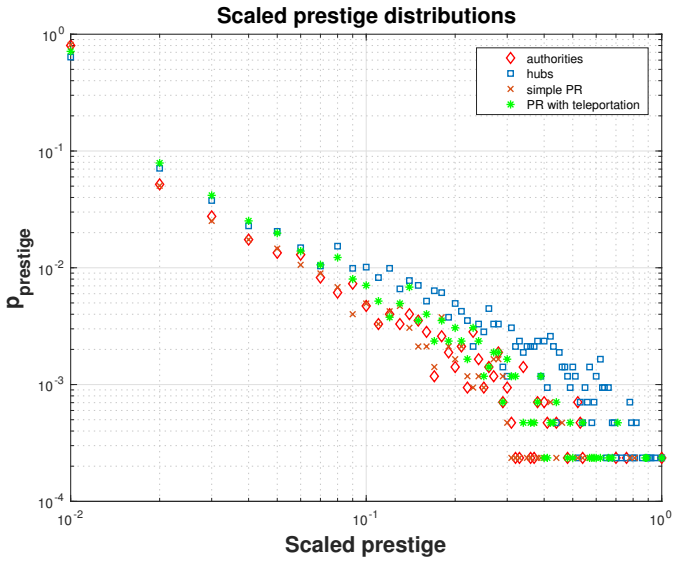

Figure 2. Scaled version of prestige scores distributions for the proposed algorithms in log-log plot.

reported in a common scale in order to compare the behaviors.

We can see that all the discussed ranking methods behave in a similar way: they have a lot of occurrences of small prestige nodes and a decreasing number of even more prestigious players, where the concept of *prestige* is defined by the specific algorithm. However the probability of highly prestigious players is not negligible since the behaviors follow heavy-tailed distributions.

The computational demand of the proposed algorithms using $\epsilon = 10^{-8}$ is reported in Table II and we ascertain that there is no need of speeding-up techniques for our purpose since N is not too large.

The predictive power of those algorithms is shown in Table III. In our analysis PageRank and *Modified ATP* ranking behave similarly and larger test sets are needed to investigate better the results. As order of magnitude the obtained results are consistent with the ones shown in [2] but we achieved a more robust *ATP* estimator by considering all the points gained by a player (we called it *Modified ATP*) and not the ATP ranking at the exact time of the match, which is done by the Official ATP estimator, but it has already been proven to achieve worst estimates than e.g. PageRank [2].

Finally, in Figure 4 is shown a comparison of the complexity of the proposed algorithms by varying the convergence param-

	New Data: from 01/09/17 to 30/11/17				New Data: all 2017			
# of Matches	431				2633			
	Modified ATP	**Authorities HITS**	**Simple PR**	**PR with teleportation**	**Modified ATP**	**Authorities HITS**	**Simple PR**	**PR with teleportation**
Right prediction %	59.53%	59.53%	60.70%	58.84%	60.92%	60.08%	60.46%	60.27%

Table III
PREDICTIVE POWER OF THE PROPOSED RANKING ALGORITHMS ON TWO DIFFERENT TEST SETS.

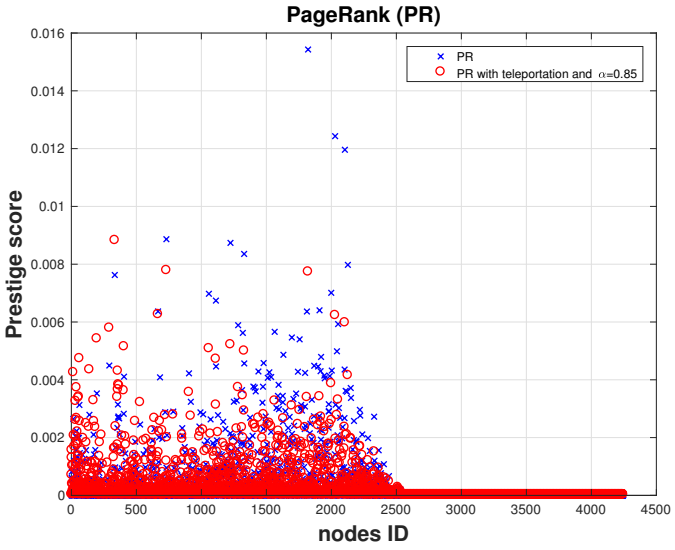

Figure 3. Prestige scores of PageRank algorithms.

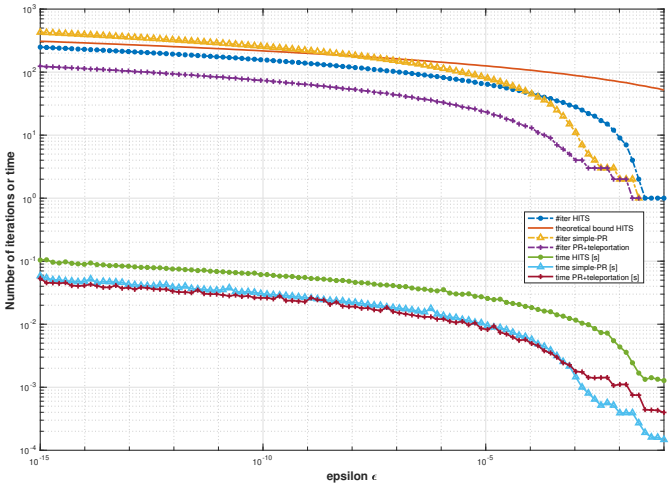

Figure 4. Number of iterations and amounts of time needed in order to run the proposed algorithms.

eter ϵ, both in terms of number of iterations and elapsed time. We can appreciate that even though the number of iterations needed by PageRank with teleportation is smaller than the others, the update step is more complex thus resulting in a computational time similar to the simple PageRank. Also, HITS algorithm performs worst than the others in terms of time needed and we can notice that the theoretical bound on its number of iterations is quite strict for small values of ϵ.

B. Link Prediction

In this section we are going to briefly investigate the players who are likely to play against in future, given that they have never played against before. This is actually a problem of link prediction and we can also consider the undirect and unweighted network's representation because we are looking for predictions at link level, not at the specific outcome of the match. The idea is that *similar* nodes are likely to build a link between them. As similarity metric we used the idea of *Common Neighbors* (CN) defined as: $S_{CN}(i,j) = |\mathcal{N}_i \cap \mathcal{N}_j|$ where $\mathcal{N}_x$ is the set of neighbors of node x. Actually, for undirect networks a simple expression holds: $\mathbf{S}_{CN} = \mathbf{A}^2$. Moreover we need to restrict our attention to active players, thus they could effectively play a match in future.

Applying all those considerations above we found that the six most likely matches to be drawn are: *Victor Troicki - Ivo Karlovic*, *Rafael Nadal - Yen Hsun Lu*, *Teymuraz Gabashvili - Gael Monfils*, *Marin Cilic - Dmitry Tursunov*, *Nicolas Mahut - Marcos Baghdatis* and *Fabio Fognini - Nicolas Mahut*.

The complete list is saved in the variable named `names_CN` of the attached file `tennis_wins.m`.

The complete task as presented takes about 6 seconds for the entire network. Many improvements could be brought by e.g. restricting the search to active players only.

C. Communities Detection

We now want to partition the graph in k disjoint groups, *communities*, through *spectral clustering* technique defined in [10]. A community is a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities. Intuitively, one should expect that k communities will appear, each containing players of the same era; but we may be interested in how to find the best partition such that minimizes the connections among the k groups.

For simplicity let's consider the case of $k = 2$, any other choice is a straightforward extension. We consider the normalized Laplacian matrix $\check{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D} = diag(\mathbf{d})$ and $\mathbf{d} = \mathbf{A1}$. The normalization makes the Laplacian matrix more *stable* in the sense that the produced eigenvectors are less noisy. Then we find the second largest eigenvalue λ_{N-1} and its eigenvector $\mathbf{v}_{N-1}$ respectively called *algebraic connectivity* and *Fiedler vector* from [11]: hence in order to find the two communities we can simply look at the sign of the Fiedler vector and assign indices corresponding to positive values to one community and vice versa.

Figure 5 reports the Fiedler vectors $\mathbf{v_{N-1}}$ and all the eigenvalues λ_i for both the direct and undirect representations.

First of all we can confirm that $\lambda_N = 0$ and $\lambda_1 < 2$ as we expect from the theory. Then we can notice that only two or three eigenvalues can be considered small and the eigengap between them is still quite large; hence a partitioning in two or three communities is *a posteriori* sensible. Moreover defining a conductance measure $h_G = \min_A \frac{\text{cut}(A,A^C)}{\min(\text{assoc}(A),\text{assoc}(A^C))}$, the Cheegar's inequality $\frac{1}{2}\lambda_{N-1} \leq h_G \leq \sqrt{2\lambda_{N-1}}$ helps in measuring the quality of spectral clustering: more specifically a low value of h_G means that the partitioning is good. We found $0.0321 \leq h_g \leq 0.3585$ where the upper bound is not very small, thus the partition will not be very accurate, because we need to divide the careers of many peer players, thus many links will exist between the communities.

From the sign of the Fiedler vector we can see that the previous intuition was correct and we can identify 1988 as the year of transition (i.e. around player ID 1400). That year is not at all the half of the considered period, which goes from 1968 to 2017; it indicates, instead, the year of a seminal moment in ATP history, because in 1988 "The Parking Lot Press Conference" [12] took place, which states the beginning of the ATP Tour era. From there onward tennis match schedules are similar to what we are used to nowadays while before the tennis circuit was very different.

This analysis also bring to light few mistakes (about 10) due to homonymy in the dataset: we can observe this by looking at the Fiedler vector where some players of the first years considered (IDs from 0 to 1200) are identified as belonging to the community of more recent players. Those inaccuracies, however, do not heavily influence the previous results since the players involved are not very *significant*.

In Figure 6 is shown the plot of the two partitions. In particular we can see in the center of this plot the aforementioned errors: the most evident is the case of player ID 682. We can confirm that the partitions will never be very accurate since many connections (i.e. matches played) exist across them. Notice also that the algorithm does not classify accurately the players who only have lost matches (players with high ID numbers) since there is too much noise; a better classification could be achieved by simply categorize them as belonging to the community of the players who beat them.
However, even though the quality as communities is not very good, still the partitions found are satisfactory and reflect our starting idea.

III. Conclusions and Future Work

In this paper we have performed a joint analysis of few different ranking techniques and we have evaluated them showing analogies and differences, also comparing and extending the results already present in literature. We have shown that *Jimmy Connors* is still the best player in tennis history up to 2017 according to the PageRank with teleportation algorithm, but actually *Roger Federer* is approaching the top position, indeed it is at about the same value of *Ivan Lendl*.

An interesting aspect of the proposed ranking systems is that they do not require any arbitrary introduction of external criteria for the evaluation of the quality of players and tournaments.

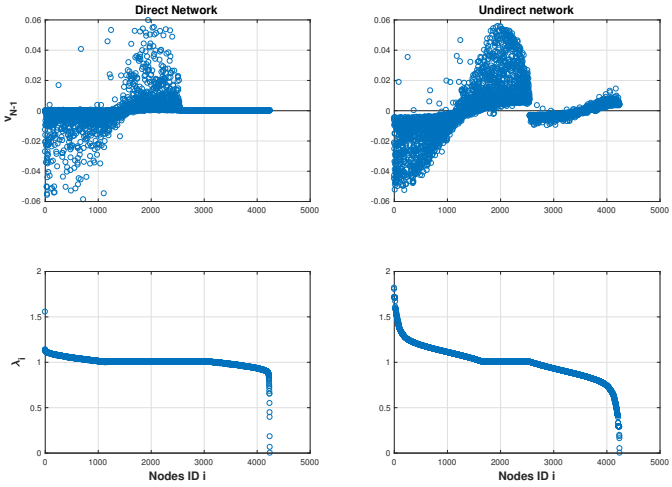

Figure 5. Fiedler vector $\mathbf{v_{N-1}}$ and eigenvalues λ_i for direct and undirect network.

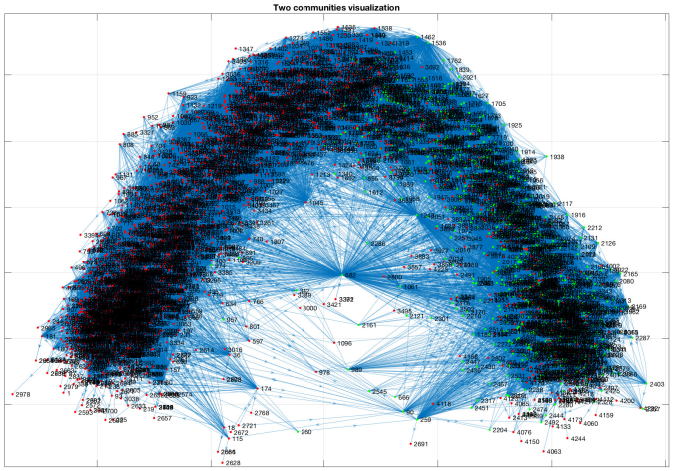

Figure 6. 2D network visualization identifying two communities. First partition in green, second in red.

Players' *prestige* is in fact self-determined by the network structure. The proposals achieve also similar predictive power to the modified ATP ranking and defeat the official one.

Those considerations on predictive power should be reinforced in the near future by choosing an enlarged test set. In future, for example, we would like to include in the statistic some matches played in 2018, in order to have independent data, and also include and evaluate other modifications to PageRank algorithm.

Moreover we have briefly discussed about an easy method of links prediction exploiting common neighbors as similarity metric. Many other metrics can be taken into account, e.g. *Adamic-Adar* or *resource allocation* and so on; the results can be compared but we do not expect large variations hence it is not worth to present them here.

Then we have seen an interesting and powerful application of spectral clustering for graph partitioning and we have recognized a promising result. We can further investigate how the partitions will change by increasing the number of cluster or by using a different communities detection algorithm.

REFERENCES

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

[2] N. J. Dingle, W. J. Knottenbelt, and D. Spanias, "On the (Page) Ranking of Professional Tennis Players.," in *EPEW/UKPEW*, pp. 237–247, Springer, 2012.

[3] D. Aparício, P. Ribeiro, and F. Silva, "A subgraph-based ranking system for professional tennis players," in *Complex Networks VII*, pp. 159–171, Springer, 2016.

[4] A. D. Spanias and B. W. Knottenbelt, "Tennis player ranking using quantitative models," 2013.

[5] D. J. Irons, S. Buckley, and T. Paulden, "Developing an improved tennis ranking system," *Journal of Quantitative Analysis in Sports*, vol. 10, no. 2, pp. 109–118, 2014.

[6] M. Sipko and W. Knottenbelt, "Machine Learning for the Prediction of Professional Tennis Matches," 2015.

[7] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[9] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[10] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[11] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak mathematical journal*, vol. 23, no. 2, pp. 298–305, 1973.

[12] J. Buddell, "The Tour Born In A Parking Lot: `http://www.atpworldtour.com/en/news/heritage-1988-parking-lot-press-conference-part-i`," 2013.