Small World Network of Athletes

Graph Representation of the World Professional Tennis Player

Hokky Situngkir
hs@compsoc.bandungfe.net
Dept. Computational Sociology Bandung Fe Institute
Research Fellow at Surya Research International

July 4th 2007

Abstract

The paper proposes an alternative way to observe and extract the multiple matches games of sports, *i.e.:* tennis tournament in the Athlete's Historical Relative Performance Index and its representation as graph. The finding of the small world topology is elaborated along with further statistical patterns in the fashion of the weighted and directed network. The explanation of the sport tournament system as a highly optimized system is hypothetically proposed. Finally, some elaborations regarding to further directions of the usability of the proposed methodology is discussed.

Keywords: small-world network, sports and tournaments, historical relative performance index of athletes.

1. Introduction

The role of network analysis has been somewhat becoming a very interesting topics, since this method allows us to see the circumstances of observed phenomena in a whole view (*cf.* Barabási, 2003). Network analysis has been employed to observe the interconnectedness of many things in our life where we can represent elements of the system as nodes and the connections between them as the edges. The growing works of complex network has discovered a lot of interesting patterns and statistical properties that invites major interdisciplinary researches (Albert & Barabási, 2002). It has been broadly accepted that network representation of system analysis and information extraction from wide varieties of data can be useful for further insights and to reveal complexities beneath the system. There have been a lot of application of this nature of analytical works observing social system, *e.g.*: stock market network visualizations (Situngkir, 2005), network of business and innovation (Khanafiah & Situngkir, 2006), political parties relations and inter-connection representation (Situngkir, 2004), and even more.

One of the most interesting fields with abundant data is the world of sports. Sports now cannot be separated with the modern life: not only for its use to keep us health and as good fields for learning about fairness and honesty, moreover, modern life has treated sports as a media of entertainment and even an important symbol of national pride of which citizens achieving high grade of international stage of sports.

The task represents in this report is an alternative to extract useful information in the world of sport with multiple matches and head-to-head game plays important role. The games in professional tennis is an interesting discourse with abundant availability of data and interesting statistics. In this short paper, we extract some of them[1], *i.e.:* the games of head-to-head games in Professional Tennis Tournament of Grand Slams held yearly with four games of tournaments (United States Open, Australian Open, Wimbledon, and French Open), all of them represents specifically different tennis court yard: the hard-court, grass-court, and clay-court respectively[2].

The paper is structured in the fashion of step-by-step observation of the matches in each tournaments in all of the sets of the games in the graph representation of un-weighted and un-directed, un-weighted and directed, and eventually the weighted and directed graphs. In the next section, we describe the complex network of the interconnectedness of the world tennis players that interestingly reveals the so-called small world topology. The next sections we analyze the statistical properties of the network and discuss things related to the tournament system and its highly optimizations that exhibits the power law distributions on some aspects in the professional tennis tournaments. The paper ends with some discussions on possible further directions to the analysis of games of sport and tournaments, and eventually to possible gain fruits from the proposed methodology.

---

[1] The data is publicly available online at the Wikipedia Internet Encyclopedia, URL: http://en.wikipedia.org/wiki/Grand_Slam_(tennis). We use the single male match statistics in all of the tournaments in the period of 1980-2006, and the single female matches in the period of 1995-2006 (both with exception of Australia Open which includes the result of 2007).

[2] We realize that there are some other important tennis tournaments, e.g.: the ATP and WTP tour, team cups (the Davis Cup, Fed Cups), as well as a lot more exhibition tournaments, but we have choose the tournaments in the Grand Slam for some resource-related constraints.

3. The Topology of the Network of Athletes

In each tournament played, professional tennis players meet in the head to head game in the nature of the applied competition system. There is somehow a kind of path must be won by an athlete to the top of the rank. The score of each set of game, however, reflect the dynamics within the game: the way in which each opposing sides try to beat each other. As it will be described later in the next section, we build the athlete's network by using the sum of all performance of all athletes in series of tournaments in the proposed historical performance index. The better a player perform relative to her opponents, the better values of the index she has in the global view. In this perspective, however, winning is not always representing the whole thing within a game: as a weaker player could resist in a tight game, it would interestingly resist her performance index relative to other athletes.

First of first, our network of athletes would be in the term of whether or not a player ever met with other players. A player is considered to be connected to other player if they have ever met in a game of tournament. The improvement and development of any tournament systems in the world of sports aims to preserve sets of the game that would only let the best player turns out to be the champion. There has been a lot of sports tournament system available and applicable today regarding to many constraints, *e.g.:* the length of the tournament season, etc (*cf.* Coakley, 1999). Most of the competitive system does not afford to let all players meet all other players, therefore from series of games in various tournaments, only the best players are connected to most numbers of other players.

For example, as the world professional tennis champion, Andre Agassi (USA) must have met large numbers of other players from around the world whose various skills and performances in order to be able to arrive at the grand final sessions and become the champion by winning the peak of the game in all of those tournaments. Thus, the more connected athletes are the most likely to be the best players. However, qualitatively speaking this is plausible for their performance in the large numbers of internationally recognized games consequently give positive feedback to enhance their maturity in other games. The famous term of "the richer gets richer" somehow entitles another meaning by representation of the word "rich" as the "richness in experience" by performance in the official games of sport. We realize this by understanding that game of sports is a complex thing in the perspective of athlete. The performance of sport is not merely calculable from the performance while in the training sessions but also dependent to their psychological circumstances as well as physiological ones. A good player is not only good at training field, but also in the middle of (must be faced) stress in the well-recognized sport events. The best players perform the best respect to all intertwining aspects. However, this grows the interesting part of the world sport tournaments and become the complex sources of the uncertainty of the game result.

Our observation to this kind of network structure could bring us to two analytical parts: the global topology of the emerging network and the representation of the nodes and links in the graph. If we consider the graph of athletes in a lot of tournaments, Ω_S, as defined by $\Omega_S = (A_S, E_S)$, where S', ($i = 1, 2, ..., N$) is the set of N players and $E = \{e_{ij}\}$ is the connections between players who has ever met in one of a game in the series of tournaments. Here, we have a certain function $e_{ij} = f(a_i, a_j)$, $f(a_i, a_j) \neq 0$ indicates that there is a link between player a_i and a_j, and *vice versa*. We will discuss later about this function since in this section, our focus right now is on the topology of the players' interconnectedness. In this simplistic case, the two connected players are adjacent and the degree of a given player is the number of edges that connects it other players. We do the observation on the head-to-head game among world professional tennis male and female players in the Grand Slam from the year of 1980-2006 and 1995-2006 respectively. Here, the Grand Slam world tennis championship are made by four once in a year tournaments, the Australia Open, Wimbledon Championship, French Open, and the US Open. Interestingly, from the glance view, we can see (fig. 1) that the yielded graph is highly clustered and likely short average paths between two players in the network. These two properties are well-known as the Small World graph topology.

Small world topology has been discovered in the representation of various complex network of the real world. Interdisciplinary works have revealed the similar network topology in biological and chemical networks (Alon, *et. al.,* 1999), human language and the use of words in sentences (Cancho & Solé, 2001), scientific collaborative networks (Newman, 2001), transportation networks (Li & Cai, 2007), power grids (Watts & Strogatz, 1998), the world wide web (Albert, *et. al.,* 1999), biological metabolism (Fell & Wagner, 2000), ecological and food-web networks (Dunne, *et. al.,* 2002), movie star collaborations (Amaral, *et. al.*, 2000), web of human sexual contacts (Liljeros, *et. al.* 2001), and even more networks representations of the real world systems (Barabási, 2003).

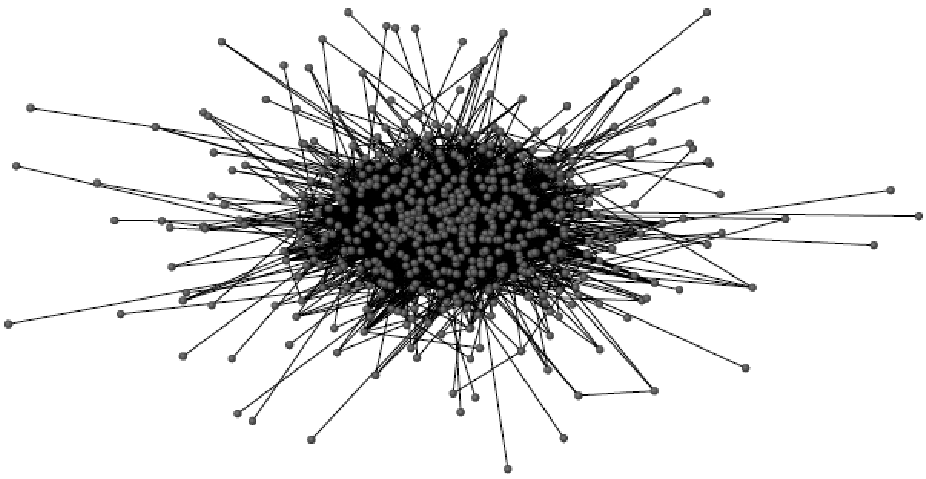

Figure 1. The small world topology of the interconnectedness among Female tennis players in Grand Slam Tournaments 1995-2006.

Small World Network can be recognized as a graph representation in which most vertices are not neighbors of one another, but most of them can be reached from every other by a small number of steps. As it showed by Watts & Strogatz (1988), the small world graph can be identified by observing the clustering coefficient (C) of the network and the average shortest path length between vertices ($\left[d\right]$) along with the the scaling properties in their degree distribution.

Topologically speaking, if we denote the set of links $e_{ij},(i,\, j=1,2,...,N)$ where $e_{ij}=1$ if an edge exists and $e_{ij}=0$ otherwise and that the average number of links per athlete is $\overline{k}$ and we can indicate by $\Gamma_i=\{j\,|\,e_{ij}=e_{ji}=1\}$ the set of immediate neighbors of an athlete

$a_i \in A$, $i = \{1, 2, ..., N\}$. Thus, here we want to observe two main properties of the small world network, *i.e.:*

- ✓ *Clustering coefficient,* defined as the proportion of links between vertices within its neighborhood divided by the number of links that could possibly exist between them. In other words, it measures the average fraction of pairs of neighbors of a node that are also neighbors of each other. Mathematically speaking, the clustering coefficient of each player i, thus can be written as

$$C_i = \frac{1}{|\Gamma_i|}\left(\sum_{j=1}^{N} e_{ij}\left[\sum_{k\in\Gamma_i; j<k} e_{jk}\right]\right) \quad (1)$$

where

$$|\Gamma_i| = C(k_i, 2) = \frac{k_i(k_i - 1)}{2} \quad (2)$$

since the topology represented here is an undirected graph of which identical e_{ij} and e_{ji}. Thus, the general clustering coefficient of the network can be written as the average over all players A_S,

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \quad (3)$$

- ✓ *Average shortest path length between vertices*. Let us denote $\delta(i, j) = \min(d_{ij})$ as the minimum path length that links two players $a_i, a_j \in A_S$, the average path length of a player to another can be written as

$$d_i = \frac{1}{N}\sum_{j=1}^{N} \delta_{\min}(i, j) \quad (4)$$

and thus the total average path length in the network,

$$d = \frac{1}{N}\sum_{i=1}^{N} d_i \quad (5)$$

- ✓ *Degree Distribution*. Some of the small world graphs exhibit the scaling properties in their degree distribution. In other words, the probability of a node with degree k scales over the power law

$$P(k) \approx k^{-\alpha} \quad (6)$$

The first two properties described above are the standard parameters to indicate the small world network. The small world structures are highly structured (indicated by $C >>$) but the interconnectedness of the vertices are relatively high (indicated by $d <<$) both relatively to the random graph with the similar structural properties. However, in our observation to both networks of male and female single head-to-head in each tournament within the Professional Tennis Grand Slams, the three properties are interestingly exhibited. Table 1 summarizes the result of our observation.

Table 1.
The exhibited pattern in Tennis Player Network from the Grand Slam Tournaments
(Australia Open, US Open, French Open, and Wimbledon)

graph	C	C_{random}	d	d_{random}	α
Ω_S (male)	0.1423	0.013	3.16	2.78	2.09
Ω_S (female)	0.1717	0.014	2.67	3.13	1.73

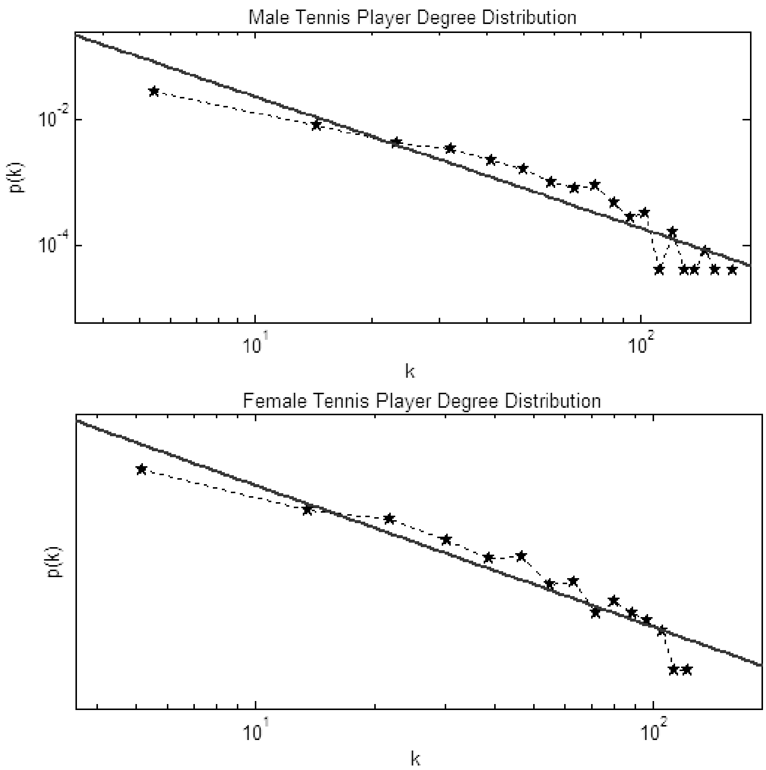

Figure 2
The power laws of the degree distributions in the network of world professional tennis players as extracted from the Grand Slam Tournaments with R=0.93687 (*left*) and R=0.94994 (*right*).

From table 1 it is obvious that relative to the similar random network, the clustering coefficients of both male and female players are higher up to more than ten times $C > C_{random}$ while the average degrees are not really that far $d \approx d_{random}$. This indicates the small world topology of the world tennis players over the grand slams. Furthermore, figure 2 shows the power laws of the degree distributions $P(k)$ over the network of male players and female players in the tennis yards of the grand slams. This power laws reflect that the scaling behavior over the numbers of athletes stay connected to most players – the new players tend to connected to the existing player with a probability proportional to the degree of such a node. In the terms of sports, probably we can say that the more an athlete meets other athlete in the head to head game, the more of her chances becoming the top of the world players and the tournament systems can now be seen as a highly optimized system that only let this happen to the best players overall the tournaments.

The presence of the small world topology and the scaling behavior are somehow reflects the nature of the tournament system applied in the competitions. The small amount of average distances among athlete shows that most of the players are connected not very far one another while the high clustering coefficients qualitatively reflect the classes of the athletes based on their respective historical performances. However, only the best players might be able to perform at the highest frequency of which opens the bigger possibilities (and opportunities) to her for meeting other players in the head-to-head game.

Such power-law distributions – have been widely understood – arise at self-organizing critical states, to which lots of systems would evolve in a consequence of consecutive introducing random local perturbations, and in a consequence of spreading the perturbations over those systems regarding to the simple dynamic rules in them (*cf.* Bak, 1996). Interestingly, we can interpret this understanding in such ways that the competition among the world tennis players is in such states of homeostatic and robustness that any perturbations came from newcomers, *etc.*, do not easily to change its global structure (*cf.* Watts & Strogatz, 1998, Barabási & Albert, 1999). However, as it has been noted in Jeźewski (2004), there is another scenario that possibly emerging such power-law behavior: it appears when systems with certain distributions of initiating events undergo optimization and resource constraints. Here, the highly optimized on the tournament and competitive system exhibits the scaling patterns of the connectiveness between athletes. In such ways, both explanations could not become counterintuitive.

4. The weighted and directed Athlete's Performance Graph

The weight of the matrix is constructed as follows. Imagine we have some historical tournaments with their respective results for each set of games. For example, we have two competing teams, *A* and *B*, and each team would be represented by $\{A_1, A_2, A_3\}$ and $\{B_1, B_2, B_3\}$ respectively. A game i could be comprised of a number of sets (τ_i), and the relative strength of player x to player y can be calculated by the fraction of scores gained by player x from the total scores hit throughout the game. However, in one tournament, player x and player y might be met more than once. Thus, in a single tournament, we can build an adjacency matrix of players (K) which elements are the aggregate fractions of the scores gained by players,

$$k_{xy} = \frac{1}{N}\sum_{i}^{N}\left[\frac{1}{\tau^{(j)}}\sum_{j}^{\tau^{(j)}}\frac{k_x^{(j)}}{k_x^{(j)} + k_y^{(j)}}\right],\ k_{xy} \in K \qquad (7)$$

where k_x denotes the total score of player-*x,* N denotes how many times the two players meets in the single tournament. The value of k_{xy} would be in [0,1]. As the $k_{xy} \to 1$, the game is relatively more easier for player x. It obvious that player x won most sets of the game when $k_{xy} > 0.5$ and vice versa.

Furthermore, from T series of historical tournaments, we have the relative strength index between player and player y. Thus, we now have, the Historical Athlete's Relative Index between player x and player y over T tournaments that is calculated as,

$$\alpha_{xy} = \sum_{t=1}^{T} \xi^{(t)} k_{xy}^{(t)} \tag{8}$$

and

$$\sum_{t=1}^{T} \xi^{(t)} = 1 \tag{9}$$

where $\xi^{(t)}$ is the weight factor of each tournament we put into account which is event-dependent in the horizon of historical games[3]. Since, the Historical Athlete's Relative Index becomes the elements of an adjacency matrix ($\alpha_{xy} \in \mathrm{A}$) in the athlete's network, any new tournaments, with some possibilities of the newcomers in the event the new tournaments may update the adjacency matrix A' as,

$$\alpha'_{xy} = \frac{\alpha_{xy} + \xi' k_{xy}'}{1 + \xi'} \tag{10}$$

This updating rule from any new tournament in the history of the corresponding sports gives us possibility to measure the relative strength of an athlete over time in our athlete's network. Here lies the interesting part of the dynamic Historical Athlete's Relative Performance Index (HRPI) with some further applications as described in Khanafiah, *et. al.* (2007).

As it has also been shown in the historical performance of the proposed index, we realize that any head-to-head game will always eventually yield a winner and a looser in a game. The more sets won by a player when meet another one, should consequently raise her performance index and this could be interpreted that she is relatively stronger. It is tempting to represent this by using the network representation and observe the emerging statistical facts. The small world analysis on directed graph has also been widely used, for instance in the network structure of world wide web, *etc*. (see Almaas & Barabási (2006) for more).

We apply the similar methodology of the directed graph model to our tennis player network similar to the one used in representing the internetworking graph model (Broder, *et. al.,* 2000). In this directed graph, the outward arcs from a node represent the mostly won games with players with the inward ones. As we discover that player i won most sets of games to player j, then we draw an arc of $i \to j$; it is apparently calculable from the historical performance index, as $k_{ij} > 0.5$, and vice versa.

---

[3] Obviously not all sport tournaments can be considered to be the same. For instance, the meeting between player x and player y in an Olympic round should be considered more important in the index relative to the one in a regional based tournament. The more important one game, the greater the value of $\xi^{(t)}$.

It is interesting that both the outward and inward interconnections among the tennis players are also exhibiting the power-laws as shown in figure 3 and table 2.

Table 2.
The inward and outward degree distributions in Tennis Player Directed Network

graph	inward degree		outward degree	
	α	R	α	R
Ω_S **(male)**	1.7676	0.90389	2.234	0.9566
Ω_S **(female)**	1.389	0.86134	1.737	0.96774

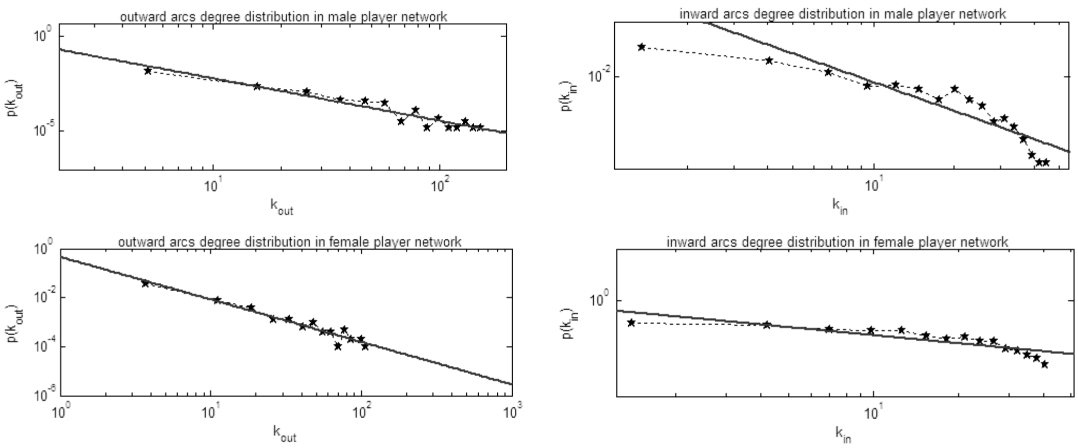

Figure 3
The power laws of the degree distributions tennis player directed graph representation.

From table 2, we could see that the inward arcs degree distribution of our directed graph exhibiting the steeper distribution relative to the outward ones for both female and male competitive worlds. This naturally reflects the situation in the tennis yards: as most tennis players come and go in the tournaments and not all of them but only some of the best persistently attend the exclusive professional competitions. Once again, we can witness how the tournament system actually optimized the network in such a way for the best players of the world.

In order to see more closely to the competitive system, we calculate the strength of the vertices in our graph by summing up the weights of the arcs connecting players (Yook, *et. al.,* 2001 and Barthélemy, *et. al.,* 2005),

$$s_i = \sum_{j \in N} e_{ij} \qquad (11)$$

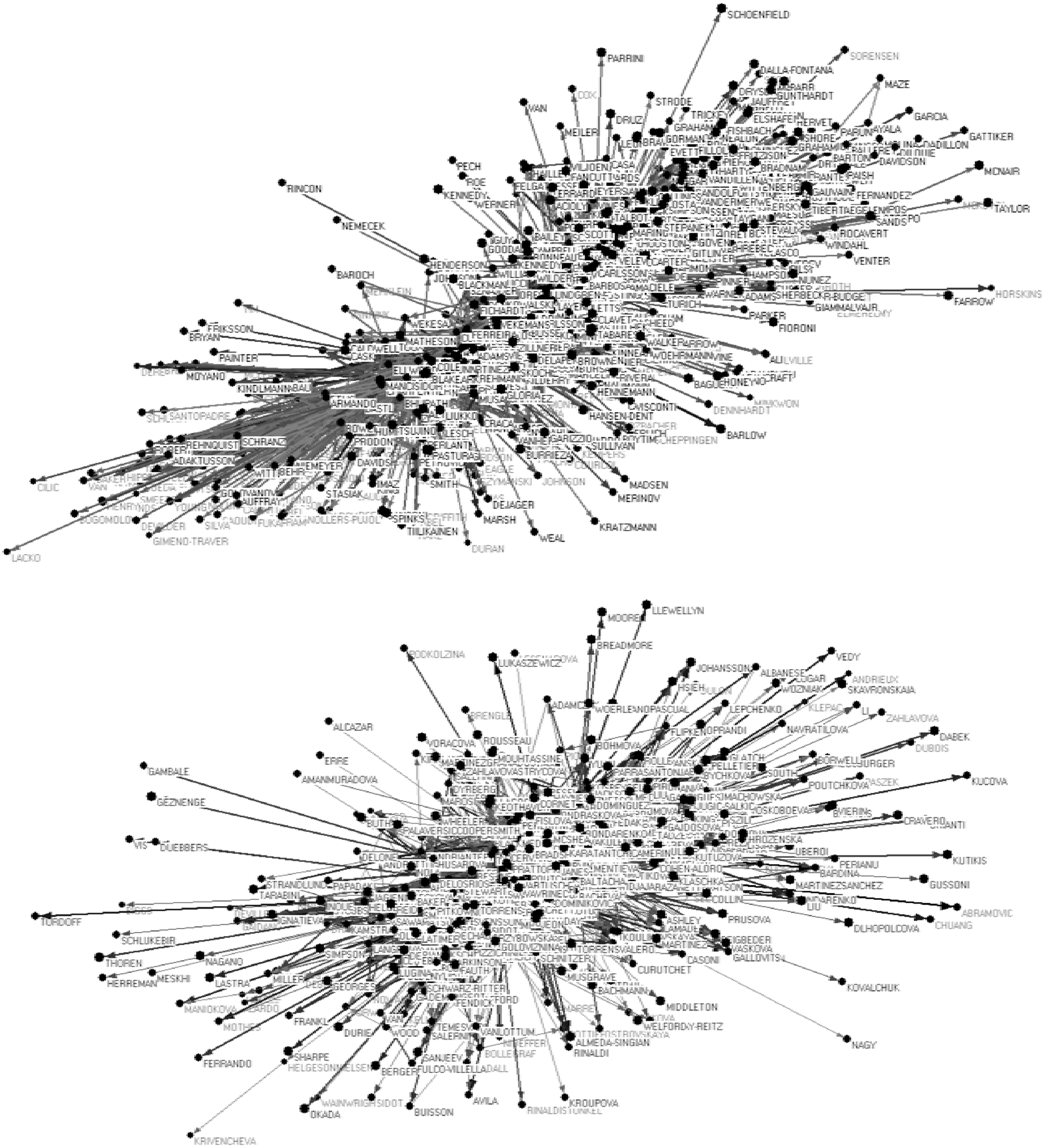

Figure 4

The network of world professional tennis players of single male (*above*) and single female (*below*). The inward arrows to a node represent relatively weaker performance from the historical data of the grand slams.

Interestingly, as it is shown in table 3 as well as figure 5, we can see that if we simply interpret the total sums of all weights for each vertex as the relative strength of tennis players, the distribution of the relative skill of tennis players in the yards of the grand slams are also power laws. Only small parts of the athletes can perform the world-champion class of tennis playing by beating as many as possible other players throughout many series of games and tournaments and this obviously reflects the optimization of the sport systems.

Table 3.
The power law distributions of node's strength in Tennis Player Directed Network

graph	*P(s)*	
	α	R
Ω_S **(male)**	2.4093	0.96767
Ω_S **(female)**	1.7813	0.97389

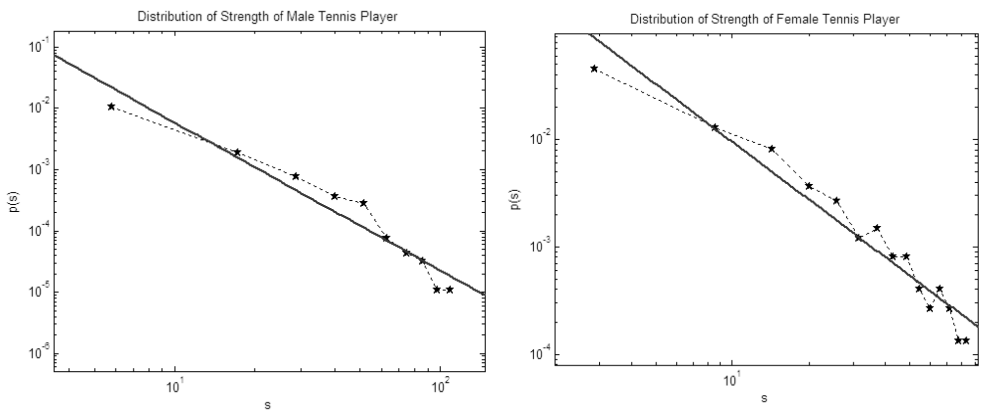

Figure 5
The power laws of the degree distributions of world tennis player directed graph representation.

From table 3, we could see the difference between the sloppiness of the distributions. Possibly it came from the different sizes of the data used between the female and male single games, since if it is not, it apparently showed the distinct competitiveness in the single female and male games of the grand slams and probably in general professional tennis competition.

6. Concluding Remarks

We represent an alternative way to extract the historical data of the sports tournament and use the historical head-to-head game of the world series of professional tennis Grand Slams. The graph representation is proposed by analytically observed the statistical properties of the yielded network. The small world patterns and the well known power law distributions of tennis player's degree (in un-weighted and un-directed graphs), the inward and outward degree distributions (in un-weighted and directed graphs), and the strength of the node (in weighted and directed graphs) are shown to be persistent in the graph representation. Along with the model is the Historical Performance Index as the way to measure the relative strength of each players to be represented as vertices and edges in the network.

Hypothetically, the power-law distributions are exhibited as logical consequence of the highly optimized tournaments system applied in the world tournament series as only the best players could meet more other players in the historical data. Newcomers could eventually become winner as well but most of them seated in the outer nodes in the network for a losing player may not proceed to the further stages of the tournaments while eventually reducing its connectivity with other players.

The methodology might be useful and applicable for further analysis of other historical data from other sports of which the head-to-head games or multiple matches incorporated in the tournament series. This proposal might be a first step for further observations in the world of sports for the purpose of evaluation, tournament system engineering, and even for the predictions of the results of sport games. This is an open and analytically challenging directions for future work as well as possibility for expanding the observational analysis on more data of other available tennis tournaments other than the Grand Slams.

Acknowledgement
I thank Deni Khanafiah and Rolan Mauludy Dahlan for assistance on discussions on the rough draft of the paper along with the data extracting and processing, Yohanes Surya and Tan Djoe Hok for introducing the problem.

Works Cited

Albert, R. & Barabási. A-L. (2002). “Statistical Mechanics of Complex Networks”. *Rev. Mod. Phys*. 74: 47-97.

Albert, R., Jeong, H., Barabasi, A-L. (1999). "Diameter of the World Wide Web". *Nature* 401: 130-1.

Almaas, E. & Barabási, A-L. (2006). "The Architecture of Complexity: From WWW to Cellular Metabolism". in Skjeltorp, A. T. & Belushkin, A. V. (eds.). *Dynamics of Complex Interconnected Systems: Networks and Bioprocesses*. p.107-25. Springer.

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). "Robustness in Bacterial Chemotaxis". *Nature* 397: 168-171.

Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. (2000). “Classes of Small-world Networks”. *PNAS* 97 (21): 11149-52.

Bak, P. (1996). *How Nature Works*. Springer-Verlag.

Barabási, A-L. (2003). *Linked: How Everything is Connected to Everything Else and What It Means for Bussiness, Science, and Everyday Life*. Plume.

Barabási, A-L. & Albert, R. (1999). “Emergence of Scaling in Random Networks”. *Science* 286: 509-11.

Barthélemy, M., Barrat, A., Pastor-Satoras, R., Vespignani, A. (2005). “Characterization and Modeling of Weighted Networks”. *Physica A* 346: 34-43.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). “Graph Structure in the Web”. *Comp. Net.* 33: 309-20.

Cancho. R. F., Solé, R. V. (2001). “The Small World of Human Language”. *Proc. R. Soc. Land. B* 268: 2261-5.

Coakley, J. (1999). *Inside Sports.* Routledge.

Dunne, J. A., Williams, R. J., Martinez, N. D. (2002). "Food-web Structure and Network Theory: the Role of Connectance and Size". *PNAS* 99: 12917-22.

Fell, D. A. & Wagner, A. (2000). "The Small World of Metabolism". *Nature Biotechnology* 18: 1121-2.

Jeźewski, W. (2004). "Scaling in Weighted Networks and Complex Systems". *Physica A* 337: 336-56.

Khanafiah, D. & Situngkir, H. (2006). "Visualizing the Phylomemetic Tree: Innovation as Evolutionary Process". *Journal of Social Complexity* 2(2): 20-30.

Khanafiah, D., Mauludy, R., & Situngkir. H. (2007). "Historical Relative Performance Index over Interconnectedness of Badminton Athletes". *BFI Working Paper Series* WPN2007.

Li, W. & Cai, X. (2007). "Empirical Analysis of a Scale-Free Railway Network in China". *Physica A* 382: 693-703.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., & Åberg, Y. (2001). "The Web of Human Sexual Contact". *Nature* 411: 907-8.

Newman, M. E. J. (2001). "The Structure of Scientific Collaboration Networks". *PNAS* 98 (2): 404-9.

Newman, M. E. J., Strogatz, S. H., Watts, D. J. (2001). "Random Graphs with Arbitrary Degree Distributions and their Applications". *Phys. Rev. E* 64: 269-85.

Park, J. & Newman, M. E. J. (2004). "The Statistical Mechanics of Networks". *Phys. Rev. E* 70 066117.

Situngkir, H. & Surya, Y. (2004). "The Political Robustness in Indonesia". *BFI Working Paper Series* WPM2004.

Situngkir, H. & Surya, Y. (2005). "On Stock Market Dynamics through Ultrametricity of Minimum Spanning Tree". *BFI Working Paper Series* WPH2005.

Watts, D. J.; Strogatz, S. H. (1998). "Collective dynamics of 'small-world' networks". *Nature* 393: 440-442.

Wu, Z-X., Xu, X-J., & Wang, Y-H. (2005). "Properties of Weighted Structured Scale-Free Networks". *Eur. Phys. J. B*. 45: 385-390.

Yook, S.-H., Jeong, H., Barabási, A.-L. & Tu, Y. (2001). Weighted evolving networks. *Phys. Rev. Lett*., 86: 5835-38.