Supplementary Materials for
Untangling Performance from Success

Burcu Yucesoy and Albert-László Barabási

CONTENTS

S1. DATA

Performance: The ATP website [1] collects data about all matches played by professional male tennis players from 1973 to 2015. For each player this includes the name and starting date of each tournament he participated in, his ranking at the time of the tournament, all matches he played in that tournament and against whom (both name and ranking of the player at the time), and the total points he received from the event. From 1990 on the weekly scores of all players in top 1,500 at the time are also available.

Popularity: Since December 2007, the hourly Wikipedia page-views of each and every Wikipedia article is recorded and made available as raw data files at [2] and as analyzed Wikipedia article traffic statistics at [3]. From [2] we meticulously parsed the Wikipedia page-view statistics of all professional male tennis players' pages in all the available languages, between the dates of December 2007 and January 2015.

S2. WIKIPEDIA PAGE-VIEW DISTRIBUTION AND ATHLETE SCORES

We claim that even after retirement players maintain visibility. This fact is illustrated in Figure S1A where we show a comparison between the distribution of Wikipedia page-views for all tennis players, active and retired, that has ever been in the ATP rankings lists (red), and only for active players who were on the rankings list of December 31, 2012 (blue). The difference represents the retired players, indicating that players continue to receive a varying number of page-views even after they stop performing officially.

For those players who were on the ranking list of December 31, 2012, the distribution of the score points they collected during the year 2012 was shown in Fig. 1A. Here in Fig. S1B we show how those score points relate to the number of Wikipedia visits the same players received during the year. We see the same strong correlation and significant variability that we observed for the rankings in Fig. 1D.

S3. CALCULATING $W(t)$ AND PLAYER SELECTION

Figure S2 shows how we construct the observed Wikipedia page-views $W(t)$:

In A we compare Novak Djokovic's daily page-views and his $W(t)$ between 2008 and 2015. Note that the spikes in the daily page-views in Fig. S2A always come shortly after the beginning of a tournament, which matches with our intuition that tournaments drive page-views. This is also the case for all other players but for Djokovic the delay between a tournament start and the main peak can be more pronounced because he often participates in the final rounds. Considering a time unit a few days longer than the tournament duration ensures that we never miss those peaks, therefore we choose 17 days as our time unit, slightly longer than the average Grand Slam, which lasts about two weeks.

In B we zoom into the period marked by the black rectangle in A to explain the process. We first start by marking all t points which correspond to either at every tournament start Djokovic participated in or 17 days after the previous tournament start, whichever happens first. The red lines indicate the tournament starts and black lines are those placed in between tournaments after 17 days. We than add up all daily page-views between the dates t and $t+1$ marked by the grey box in the figure and consider that number our $W(t)$ value. These values are shown in B as dark blue dots.

The fact that some Δt are shorter than 17 days needs to be taken into account in the fitting process, therefore we do all the fittings on $W(t)/\Delta t$, then calculate $W_M(t)$ by multiplying the resulting fit value using the appropriate Δt. Consequently, sometimes both measured and predicted page-views are uncharacteristically low thanks to shorter than usual Δt periods, occurring when two consequent tournaments are only several days after one another.

We construct $W(t)$ for all ranked players for the whole duration of the time when they had both an active career and Wikipedia page-views. We define an active career as being included in the weekly ATP rankings lists. Not all players who have been in the rankings list between the years of 2008 and 2014 are considered for our analysis, however: Players at the bottom of the rankings list sometimes gain a small number of page-views without having an actual Wikipedia presence because of the way visits are recorded. Any existing link on Wikipedia can collect page-views even if it does not lead to a currently active page. Sometimes a link is created in case a page will be added in the future or an existing page is deleted by the editors leaving the link behind, and clicks on any of those links are recorded as page-views. We first exclude the players who did not have an established Wikipedia page during the considered period whose page-views came from the aforementioned links. For this we use the fact that only a small number of page-views can be collected via these links and enforce a threshold of reaching a minimum of $W(t) = 50$ visits at least 10 times (i.e. for 10 separate instances of t) between 2008 and 2014. From the 600 players who match both criteria we first excluded those who share the exact first and last name with another famous person. Then we looked carefully into players whose names can be spelled in a variety of ways. This variation is caused by letters or accents not recognized by the English language, and prompts Wikipedia to keep multiple page-view records for the same person. For an accurate count, all these separate records have to be added together. Consequently we had to exclude several players with such characters in their names, if we were not confident we detected all Wikipedia-recorded spelling versions of those names. Finally we ended up with 499 players, and for those the $W(t)$ data includes views of the players' Wikipedia pages in all languages that they exist in and all versions of the player's name.

S4. DEPENDENCY OF VARIABLES

We know that performance drives popularity through a combination of performance measures. But before combining these variables, we first need to check that none of these variables can be expressed in terms of the other. To do that we calculated the correlation matrices of Table S1 (and S2) to see if any of them (or their logarithm) is linearly correlated to another. The first matrix indicates a lack of significant correlations between all variables. The second shows some correlation, especially between $log(V(t))$ and $log(r(t))$, but a fit attempting to express one in terms of the other proved unsuccessful, i.e. $log(V(t)) = -\alpha\ log(r(t)) + const.$ results in a good fit with $R^2 = 0.5$ but $V(t) = Ar(t)^{-\alpha} + const.$ results in a poor fit with $R^2 = 0.03$.

S5. QUALITY ANALYSIS OF SUB-MODELS TO PROMO

The PROMO (2) has an R^2 value of 0.6 and an Akaike Information Criterion (AIC) value of $2.074e+06$ (in comparison, the sum of all variables with multipliers obtained via an ordinary least squares (OLS) fitting has $R^2 = 0.31$ and $AIC = 2.124e+06$), and as such is the best model for accurate prediction. To make sure that a similar predictive could not be obtained using less variables, we did a quality analysis of the sub-models, each containing one less variable than the PROMO. The following is the list of such sub-models, and the resulting analysis can be seen in Table S3.

$$W_{SM1}(t) = AY(t)V(t)n(t)\ e^{\Delta r(t)H(\Delta r)/r(t)} + CY(t). \tag{S1}$$

$$W_{SM2}(t) = A\frac{Y(t)}{r(t)}n(t)\ e^{\Delta r(t)H(\Delta r)/r(t)} + C\frac{Y(t)}{r(t)}. \tag{S2}$$

$$W_{SM3}(t) = A\frac{Y(t)}{r(t)}V(t)n(t) + C\frac{Y(t)}{r(t)}. \tag{S3}$$

$$W_{SM4}(t) = A\frac{Y(t)}{r(t)}V(t)\ e^{\Delta r(t)H(\Delta r)/r(t)} + C\frac{Y(t)}{r(t)}. \tag{S4}$$

$$W_{SM5}(t) = A\frac{1}{r(t)}V(t)n(t)\ e^{\Delta r(t)H(\Delta r)/r(t)} + C\frac{1}{r(t)}. \tag{S5}$$

S6. DIFFERENT TRAINING SETS

The PROMO (2) has two fitting parameters A and C. Throughout our research we used the values $A = 3.747$ and $C = 7929$ which we calculated from a training data set of the first two years, 2008 and 2009. Figure S3 shows how these coefficients A and C would change if less or more years of data were to be included in the test data. Here we see that any training set spanning longer than the first two years do not have a significant effect on A or C.

S7. COMPARING $W_M(t)$ AND $W(t)$

We claim that PROMO (2) accurately predicts the bulk of the real visibility data. In fact, it only deviates from the measured values for very young players and/or unremarkable performers. To illustrate that, in Figs. S4A and B we show how the accuracy of prediction is affected by the players' career lengths and rank. The accuracy of prediction as a function of a player's momentary rankings $r(t)$ is shown in A and the same for the number of years the player was active $Y(t)$ is shown in B. The histograms corresponding to various cross-sections of the data emphasize that PROMO (2) works very well for highly ranked players who have been active for at least several years.

S8. THE OUTLIERS

To understand if popularity can be induced by factors unrelated to performance, we inspected the outliers, athletes whose popularity ($\sum W(t)$) is significantly higher than

$\sum W_M(t)$ predicted by their performance, captured by a modified z-score higher than 3.5. These are given in detail in Table S4.

We consider a player inactive (retired) if he has not participated in a tournament for at least two months. We also notice several outliers among the retired players, whose popularity is calculated using model (3), given in Table S5 and shown in red in Fig. S5A. Here, Ryan Sweeting is not only the most outstanding outlier, he is also a special case in which his popularity is unrelated to anything he does on the court. In Fig. S5B we show his Wikipedia page-views time-line between 2008 and 2015. He had limited visibility when he was active and when he first stopped playing, his total page-views were accurately predicted by the model. But when he got engaged to actress Kaley Cuoco, her fans inundated his Wikipedia page, causing a spike in his page-views. Since then he participated only in one more tournament, but got married and his page-views remain elevated. All but last of the other outliers stopped playing in singles but continue their tennis career in doubles, the source of their continued visibility. The final outlier in Table S5, Ziadi, is a special case where the Wikipedia page dedicated to him was edited out of existence shortly after his last tournament and no more page-views could be recorded, hence his outlier status comes from the fact that his recorded popularity is much lower than the predicted value.

S9. PROJECTED POPULARITY

Observing long-term progress in popularity is not as straightforward as observing performance because the Wikipedia page-views data is recorded only since 2008. However, the model (2) allows us to infer a projected fame for the previous years, providing for each player its time-resolved popularity throughout their career based on their performance history. We show this projection for Nicolas Lapentti in Fig. S6, indicating that we can use model (2) to generate for each player its time-resolved popularity starting from the beginning of their career, relying on their known performance parameters as input. To find the projected daily average page-views of players at the time of each tournament (as used in the main research), we first divide all $W_{MP}(t)$ by the corresponding Δt, then take the average based on a moving sum of three months preceding each tournament.

S10. RIVALS

One of the terms we used to predict how much visibility an athlete gains was who he would play against. From all possible variations, the best ranked rival a player faces in a tournament has the most impact on the player's visibility, but only if the said rival is better ranked than the player himself. To see how often such a pairing can occur, we considered the collective statistics of rank difference pairings in Fig. S7A. Here we have the total number of pairings between players of ranks r_i and r_j versus the difference between those ranks, $\Delta r = r_i - r_j$, gathered from all recorded ATP tournaments between 1972 and 2014. Most pairings happen between players of very close rankings and the big-rank-difference wings are populated with dots representing the few pairings happening between the worst ranked players and those ranked in the middle. The two extremes (best-worst) almost never get to play against each other. The entire plot is symmetric as expected since whenever player r_i plays against r_j, r_j is playing against r_i as well.

Additionally, we can look at the collective statistics of winning probabilities (Fig. S7B). Since the majority of matches are happening between players of similar rank, we have excellent statistics and very little noise around zero. This changes once we reach a rank difference of ±60 where the noise starts. We can go a step further and visualize the winning ratios for all matches between players of rank 1 to 1,000 at the time of the pairing (match): Figure S8 shows the ratio of winning as a color scale where $p = 0$ (green) means always lose and $p = 1$ (dark red) means always win. Dark blue means there are no pairings between players of those ranks. Note that top rated players and players ranked worst hardly ever play against each other. The highest concentration happens near the 45 degree line and for ranks between 1 and 400, approximately. The concentration shifts after rank 400 and players start to play more against those of better ratings rather than of ratings similar to themselves. The 45 degree line itself is blank until rank 1000 since no two players of the same rank exist before. The winning ratios unsurprisingly reflect the fact that better ranked players would win against worse ranked players most of the time: The part above/under the 45 degree line is overwhelmingly green/red in the upper/lower regions. It is not a clear-cut phenomena, however, since we still see quite a few surprises in there, shown as unexpected green dots in lower and red dots in upper regions, reflecting upsets.

To better distinguish between colors, we coarse-grain and collect all ranks in bins of 10.

In Fig. S9 now each square represents statistics from 10 rankings, first from 1 to 10, second from 10 to 20, etc., allowing us to observe the color changes better. The few upsets we see, indicated by red squares in the upper triangle and green ones in the lower are mostly consequences of too few statistics, as evidenced by the disappearance of the color gradient that was very prominent in the lower left corner, in the upper right corner and surroundings.

[1] ATP World Tour - Official Site of Men's Professional Tennis, `http://www.atpworldtour.com/`, accessed: 2015-02-29

[2] Page view statistics for Wikimedia projects, `http://dumps.wikimedia.org/other/pagecounts-raw/`, accessed: 2015-02-01

[3] Wikipedia article traffic statistics, `http://stats.grok.se/`, accessed: 2015-02-01

TABLE S1. Pearson correlation coefficient matrix for all considered variables, obtained from the entire data set, indicating a lack of significant correlations between these variables.

	$Y(t)$	$V(t)$	$n(t)$	$\frac{\Delta r(t)H(\Delta r)}{r(t)}$	$1/r(t)$
$Y(t)$	1.0				
$V(t)$	0.20	1.0			
$n(t)$	-0.01	-0.13	1.0		
$\frac{\Delta r(t)H(\Delta r)}{r(t)}$	-0.02	0.18	0.10	1.0	
$1/r(t)$	0.12	0.20	0.15	-0.06	1.0

TABLE S2. Pearson correlation coefficients matrix for the *log* of all considered variables, obtained from the non-zero values of the data set. Here we see some correlation, especially between $log(V(t))$ and $log(r(t))$, but a fit attempting to express one in terms of the other proved unsuccessful, i.e. $log(V(t)) = -\alpha\ log(r(t)) + const.$ results in a good fit with $R^2 = 0.5$ but $V(t) = Ar(t)^{-\alpha} + const.$ results in a poor fit with $R^2 = 0.03$.

	$Y(t)$	$V(t)$	$n(t)$	$\frac{\Delta r(t)H(\Delta r)}{r(t)}$	$1/r(t)$
$Y(t)$	1.0				
$V(t)$	0.31	1.0			
$n(t)$	-0.09	-0.46	1.0		
$\frac{\Delta r(t)H(\Delta r)}{r(t)}$	0.01	0.18	0.20	1.0	
$r(t)$	-0.37	-0.71	0.10	0.03	1.0

TABLE S3. Quality analysis of sub-models to PROMO.

Model	R^2	AIC	ΔAIC
$PROMO$	0.6000	2.074e+06	0
$SM1$	0.1282	2.146e+06	0.72e+05
$SM2$	0.4702	2.100e+06	0.26e+05
$SM3$	0.5400	2.087e+06	0.13e+05
$SM4$	0.5075	2.093e+06	0.19e+05
$SM5$	0.5657	2.081e+06	0.07e+05

TABLE S4. Outliers, athletes whose popularity is significantly higher than expected based on their performance, listed in the order of their modified z-score.

Player	z-score	born	first senior tourn.
Marco Djokovic	4.98	1991	2007
Gianluigi Quinzi	4.76	1996	2010
Thanasi Kokkinakis	4.44	1996	2011
Filip Peliwo	4.10	1994	2010
Oliver Golding	4.03	1993	2009
Liam Broady	3.75	1994	2009
Kyle Edmund	3.72	1995	2011
Alexander Zverev	3.67	1997	2011

TABLE S5. Retired players who are outliers, as their enduring fame is significantly different from what is expected based on their past performance, listed from the highest to lowest in order of their modified z-score.

Player	z-score	best rank	active for
Ryan Sweeting	7.24	64	9 years
Marin Draganja	4.50	550	7 years
Henri Kontinen	4.17	220	7 years
Samuel Groth	4.15	157	10 years
Rohan Bopanna	3.85	213	13 years
Mehdi Ziadi	-3.99	531	12 years

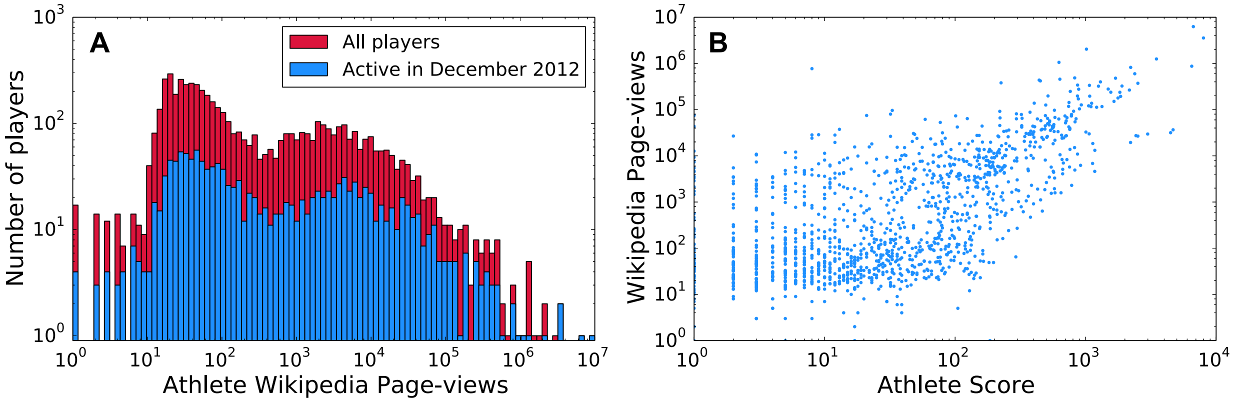

FIG. S1. (**A**) The distribution of the number of Wikipedia page-views for athletes during the year 2012, (red) for all players active and retired, (blue) for players in the rankings list of December 31, 2012, who accumulated page-views during 2012. (**B**) The number of Wikipedia page-views for all active players during the year 2012 shown in function of the number score points they accumulated during the year 2012.

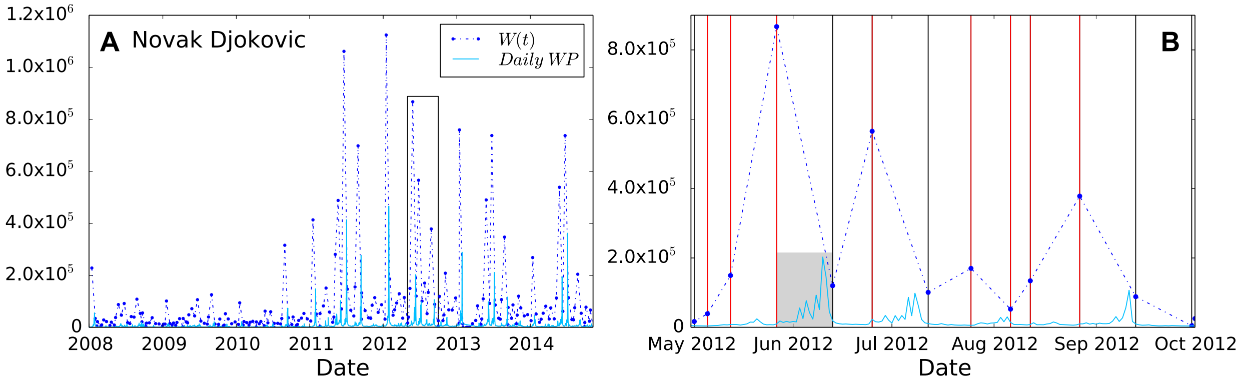

FIG. S2. Comparison between Novak Djokovic's collected Wikipedia page-views $W(t)$ (blue dots) and his daily page-views (solid line). Panel (**B**) zooms into the black rectangle in (**A**), showing the dates t as vertical lines. The dates t correspond to either the beginning of a tournament (red lines) or every 17 days (black lines), whichever happens first. All daily page-views between those lines are added together (such as the ones in the grey box) and these make up the $W(t)$ values.

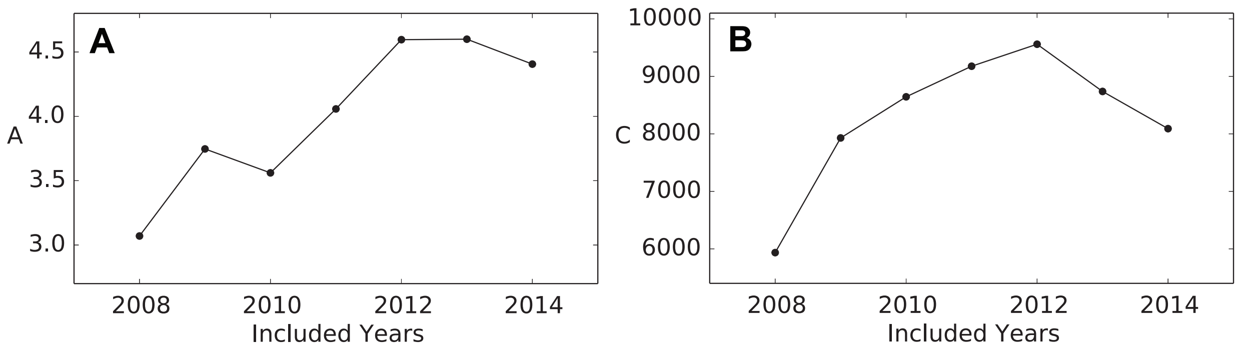

FIG. S3. The dependence of the coefficients (**A**) A and (**B**) C in PROMO (2) on the number of years included in the training data.

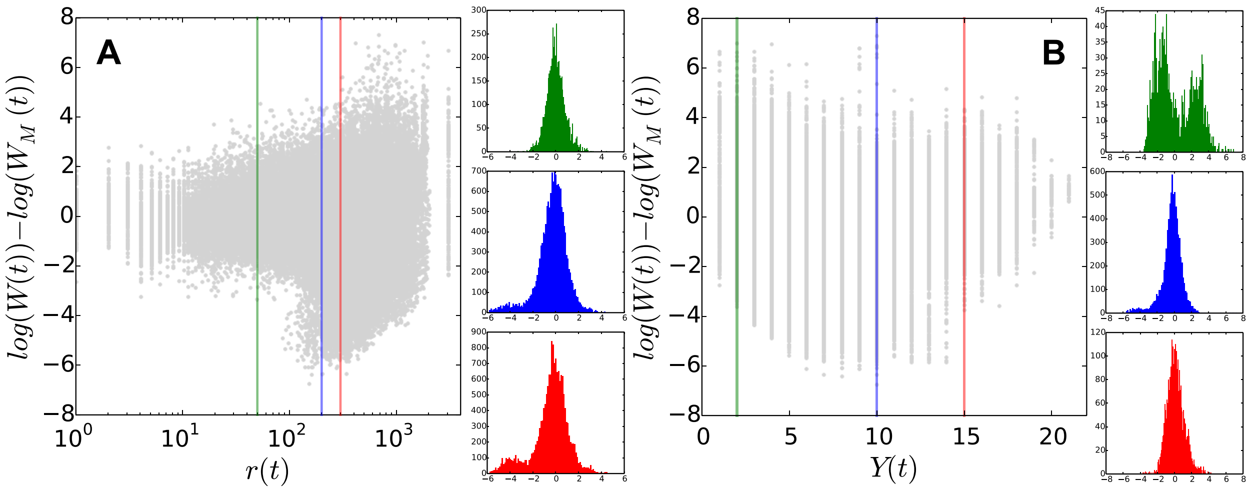

FIG. S4. The accuracy of prediction in function of (**A**) a player's momentary rankings $r(t)$ and (**B**) the number of years the player was active $Y(t)$. The colored lines show the portion of the data points histograms on the side represent. The model (2) predicts extremely well the visibility of highly ranked players who have been active for at least several years, only deviating from the measured values for very young players and/or unremarkable performers.

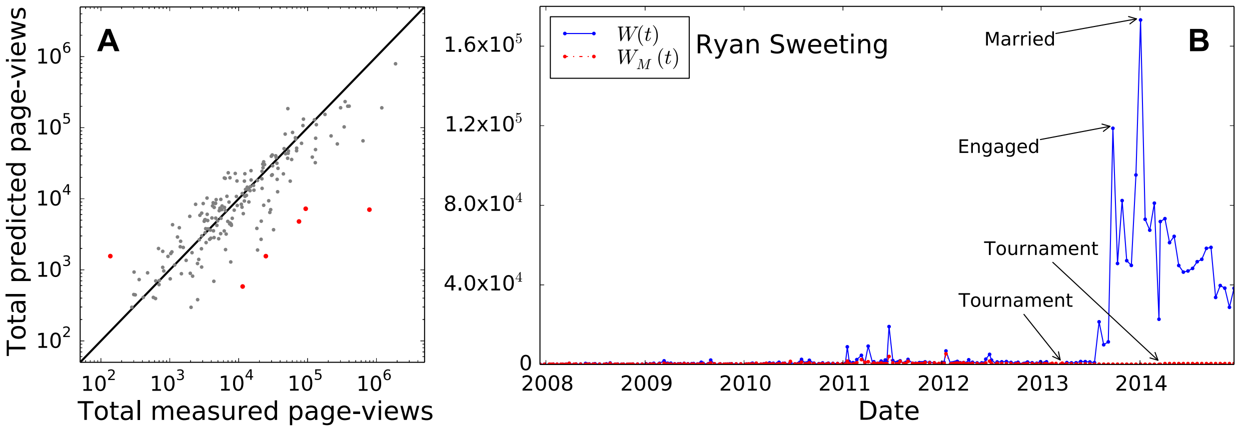

FIG. S5. (**A**) The total predicted Wikipedia page-views for each player compared to the actual total Wikipedia page-views they collected after they stopped participating in new tournaments. The dots are colored red for the outliers, whose modified z-value for the logarithmic distance between the prediction and reality exceeds 3.5. We consider a player inactive if he has not participated in a tournament for at least two months. (**B**) Comparisons between Ryan Sweeting's actual Wikipedia page-views $W(t)$ (blue) and his predicted page-views $W_M(t)$ both during his active career and after he stopped participating in new tournaments.

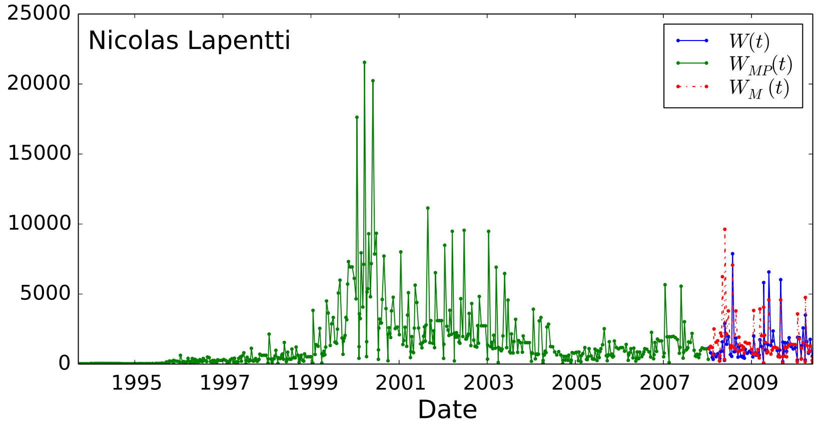

FIG. S6. The projected page-views $W_{MP}(t)$ for player Nicolas Lapentti. The blue and red lines are the actual page-views and the model predictions respectively, and the green line is a projection of his fame calculated using the same model on past performance data.

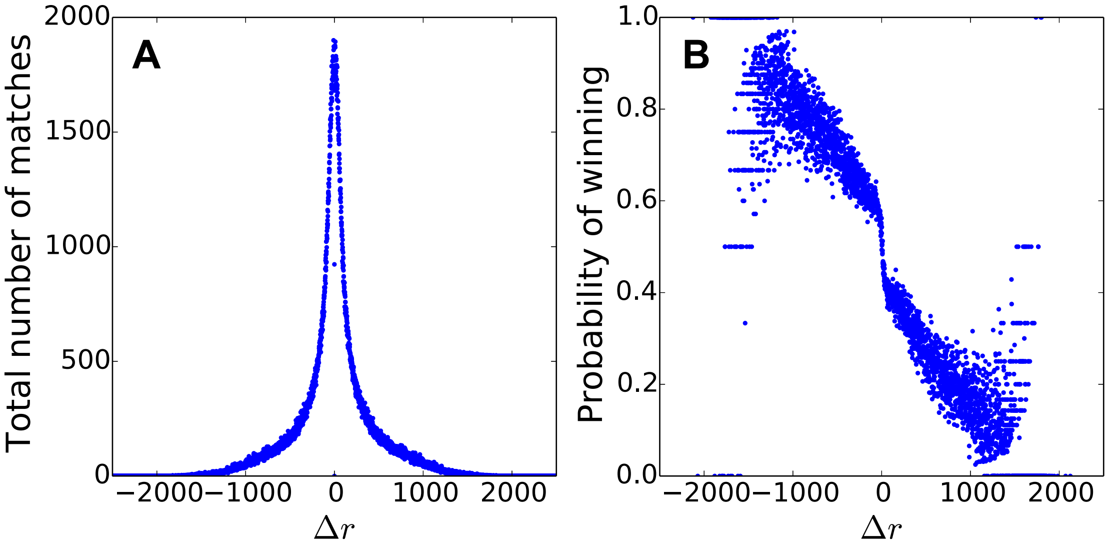

FIG. S7. (**A**) Total number of matches played between players r_i and r_j vs. the difference of their rankings $\Delta r = r_i - r_j$, and (**B**) the probability of winning of a player of rank r_i against a player of rank r_j vs the difference in their rankings Δr.

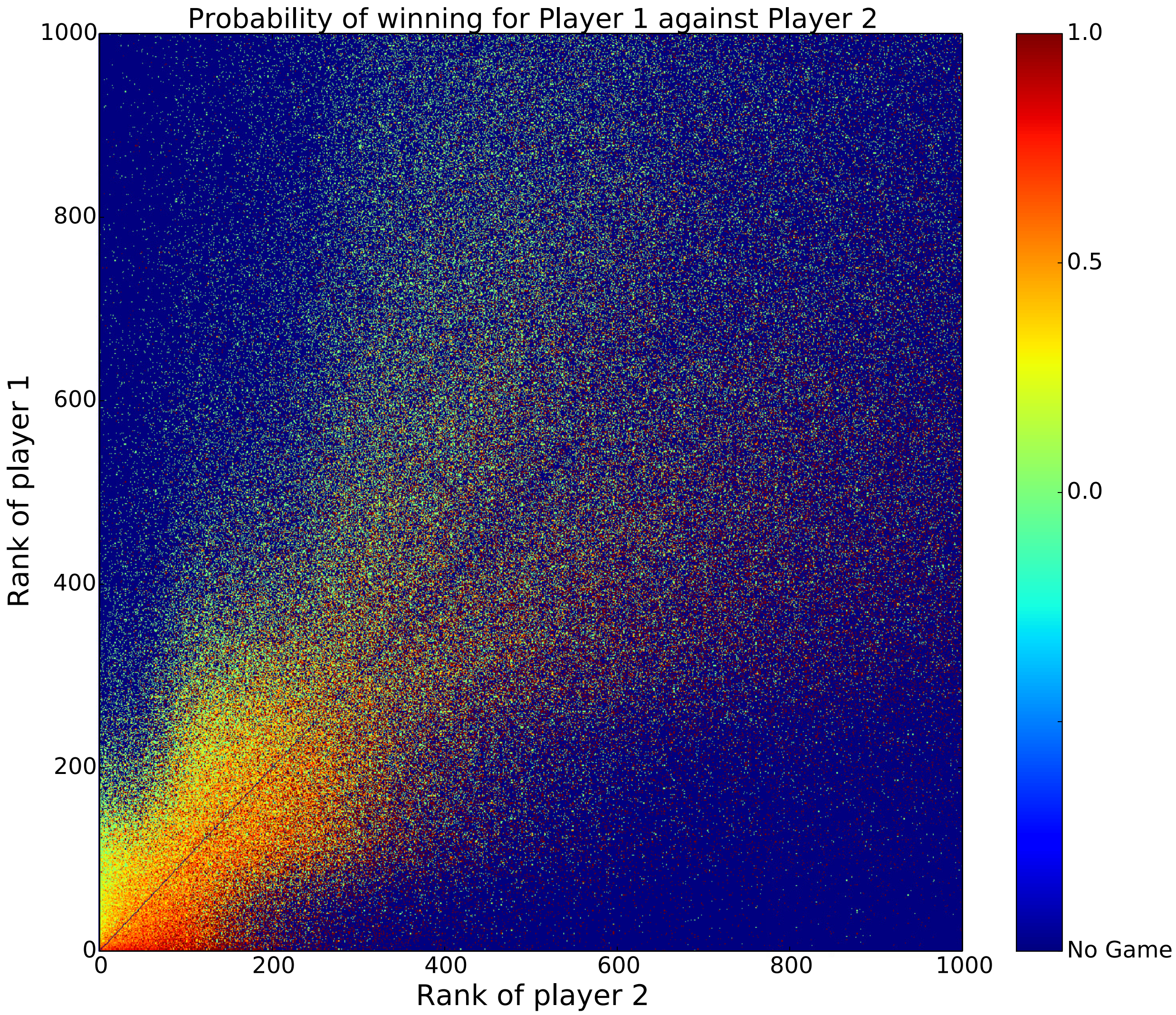

FIG. S8. Visualization of all games played between players of rank 1-1,000. The colors stand for winning probabilities, where dark red is probability 1.0 (100% winning record) and green is probability zero (all losses). Dark blue means that there has been no matches between players of those rankings.

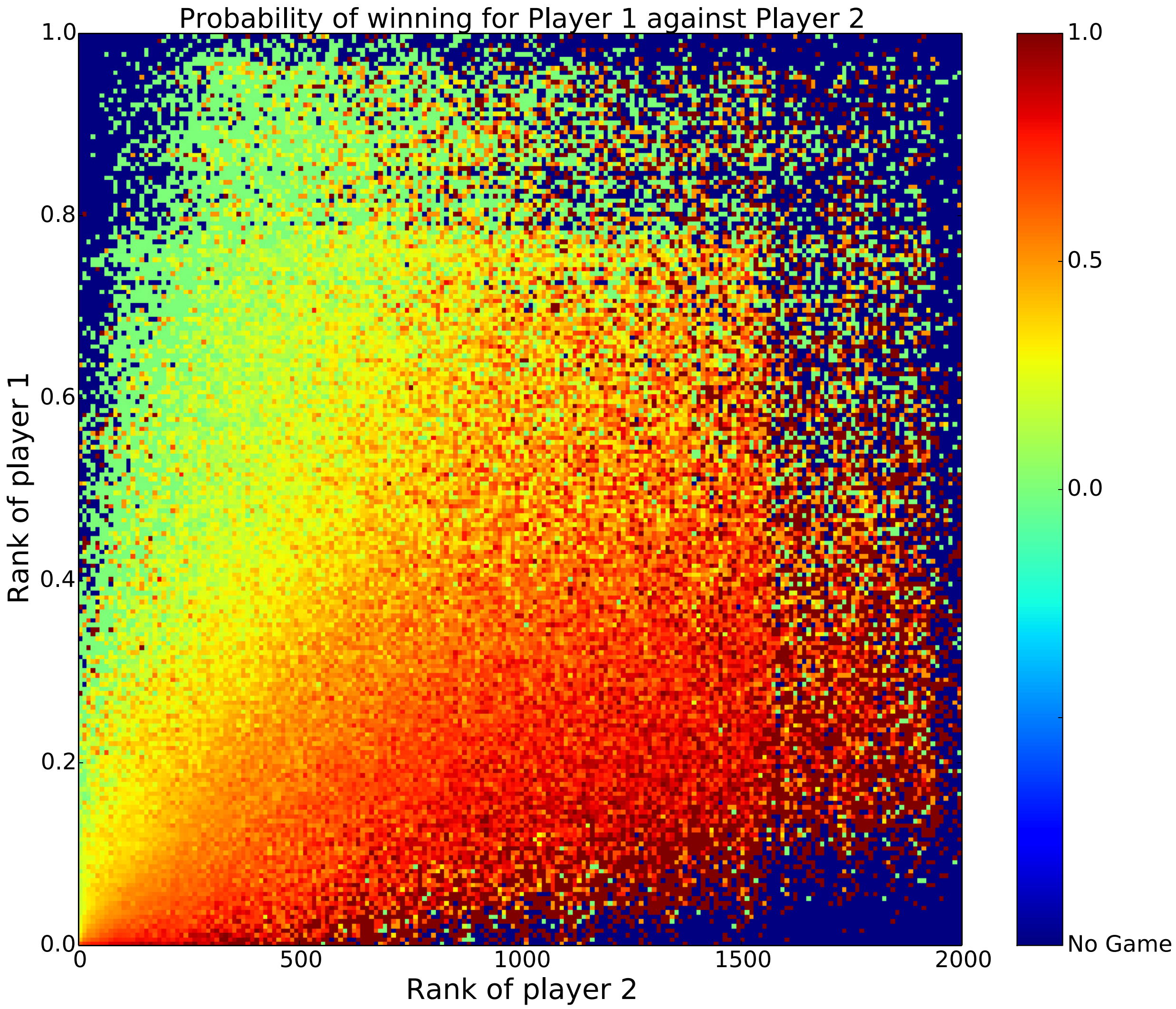

FIG. S9. Visualization of all games played between players of rank 1-2,000, averaged over bins of 10. The colors stand for winning probabilities, dark red is probability 1.0 (100% winning record) and green is probability zero (all losses). Dark blue means that there has been no matches between players of those rankings.