GM Nakamura's 45 wins - was it interesting?

Frederik Harly Baumgarten

On November 20th, 2023, a former chess World champion drew attention to a streak of 45 wins and 1 draw by GM Hikaru Nakamura, stating that "I believe everyone would find this interesting". This was by some seen as an accusation of cheating. Widespread focus and commentary from public chess figures ensued about whether GM Nakamura should be able to achieve such streaks using only his own human capabilities. But does the number of wins in such a streak warrant suspicion of cheating? I simulated thousands of games using win-probabilities derived from FIDE and Chess.com elo ratings of GM Nakamura and his streak opponents and counted the streaks which occurred. I also explored several scenarios which differed in how charitable they were to the strength of GM Nakamura's streak opponents. In conclusion, I find no evidence that the size of the streak in question should warrant suspicion of cheating.

1. Introduction

On November 20th, 2023, former World chess champion, Grandmaster (GM) Vladimir Kramnik, made a post on www.Chess.com titled "INFORMATIONAL" (Kramnik, 2023). In it, GM Kramnik highlighted a streak of 45 wins and 1 draw by an unnamed player. GM Kramnik concluded his post by stating that "I believe everyone would find this interesting". This was by some seen as a call for scrutiny and for others an accusation of cheating. Although GM Kramnik did not name the player in question, the streak details were like breadcrumbs leading back to GM Hikaru Nakamura - the second highest rated classical chess player in the world (as of this writing).

2. My aim with this work

My goal is to explore whether the streak score achieved by GM Nakamura should merit suspicion of cheating.

3. Winning streak details

I show the players involved in the 46 span of games in Table 1, along with their peak Chess.com and FIDE blitz ratings.

Table 1: Players involved in the 46 game winning streak by GM Nakamura and their peak Chess.com blitz ratings as well as their peak FIDE blitz ratings (as of June 27, 2024)

Player	Number of games in streak	Peak blitz rating	
		Chess.com[1]	FIDE[2]
Hikaru Nakamura	46	3405	2934
Levan Bregadze	16	3026	2478
Liam Putnam	9	2990	2260
Artin Ashraf	8	3132	2279
Artur Davtyan	8	2977	2417
Yagiz Kaan Erdogmus	4	3104	2379
Aram Hakobyan	1	3140	2584

[1] ratings were retrieved from Chess.com (2024)
[2] ratings were retrieved from FIDE (2024)

4. Methodology

4.1. Simulations

GM Nakamura has played tens of thousands of blitz games on Chess.com. The question really comes down to how reasonable it is for GM Nakamura to achieve a streak of ≥ 45 wins (allowing for a single draw), in the subset of games where his competition was comparable to his streak opponents. I term this subset the "streak-set". I therefore simulated each game of a streak-set and counted the sizes of the streaks which occurred. Each game of the streak-set can result in a win, a draw or a loss for GM Nakamura. I use win/draw/loss probabilities from the "Elo Win Probability Calculator" (Labelle, 2024). Fundamentally, win and draw probabilities derive from the nature of the elo rating system as well as the modelling of material odds per elo difference at the relevant elos. See the work of Labelle (2024) for more details. In order for me to estimate outcome probabilities, I require a rating for GM Nakamura and a rating for the representative simulation opponent to estimate probabilities. Simulating all games of the streak-set once will show one way in which the games might play out. GM Nakamura might have been "lucky" or "unlucky" in that particular iteration. So in order to understand what streaks sizes are likely to emerge, I repeated the simulation 1000 times. In how many of these simulations did GM Nakamura achieve similar streaks?

4.2. Simulated scenarios

All parameters that go into the simulation are arguable (number of games in streak-set, playing strength of GM Nakamura and opponents). I detail pros and cons as I see them for both FIDE and Chess.com ratings in more detail in section 7.1 and 7.2. I do to some degree favor FIDE-ratings for this particular study. This is mostly because the Chess.com rating of GM Nakamura perhaps should be disqualified in exploring plausible win streaks, since the validity of this rating is indirectly the object of study. But also due to the fact that farming mechanics and casual gameplay might obfuscate the true playing strength differences between GM Nakamura and his streak opponents. I see no obvious way to overcome these challenges. A main challenge of FIDE ratings is that streak opponents might be stronger online relative to GM Nakamura and that the FIDE ratings of streak-opponents might not have caught up to each player's actual playing strength. Both points would affect the rating difference between GM Nakamura and his opponents, and do so in the same direction. I may then explore the "wriggle room" of my results by constructing three different scenarios, which differ in how charitable they are to the streak opponents. In so doing, I am hoping that the unknown real scenario lies somewhere in-between. The three scenarios are:

- Scenario I: I use 25.000 games played per simulation. I use a GM Nakamura rating of 2934

(peak FIDE blitz rating, Table 1). I use a simulated opponent rating of 2384 (the mean peak FIDE blitz rating of streak-opponents, weighted by frequency of streak games, Table 1).

- Scenario II: I use 20.000 games played per simulation. I use a GM Nakamura rating of 2934 (peak FIDE blitz rating, Table 1). I use a simulated opponent rating of 2484 (midway between weighted average peak FIDE blitz rating of streak-opponents and maximum streak opponent rating, Table 1).
- Scenario III: I use 15.000 games played per simulation. I use a GM Nakamura rating of 2934 (peak FIDE blitz rating, Table 1). I use a simulated opponent rating of 2584 (max peak FIDE blitz rating of streak-opponents, Table 1).

It is still worth exploring a fourth scenario centered around peak Chess.com blitz ratings, as the streak in question did happen on Chess.com:

- Scenario IV: I use 20.000 games played per simulation. I use a GM Nakamura rating of 3405 (peak Chess.com blitz rating, Table 1). I use a simulated opponent rating of 3037 (the mean peak Chess.com blitz rating of streak-opponents weighted by frequency of streak games, Table 1).

Table 2 shows win/draw/loss probabilities of GM Nakamura given my chosen ratings within each scenario.

Table 2: Win/draw/loss probabilities per wismuth "Elo Win Probability Calculator" CITE given chosen ratings in Scenarios I, II, III and IV.

Scenario	GM Nakamura rating	Sim-opponent rating	Win prop. (%)	Draw prop. (%)	Loss prop. (%)
I	2934	2384	95.0	4.6	0.4
II	2934	2484	89.4	9.6	0.9
III	2934	2584	80.0	17.9	2.1
IV	3405	3037	80.8	18.7	0.5

I go into more detail about the inner workings of the simulations in section 7.3.

4.3. Software

I built the simulator and processed results using the following Python modules: Pandas, NumPy and Random (Wes McKinney, 2010; Harris et al., 2020; Van Rossum, 2020). I created all illustrations using the R-package "ggplot2" (Wickham, 2016).

5. Results

5.1. How the first simulation played out

Figure 1 shows the sizes of each >1 win-streak achieved during the very first simulation of each scenario.

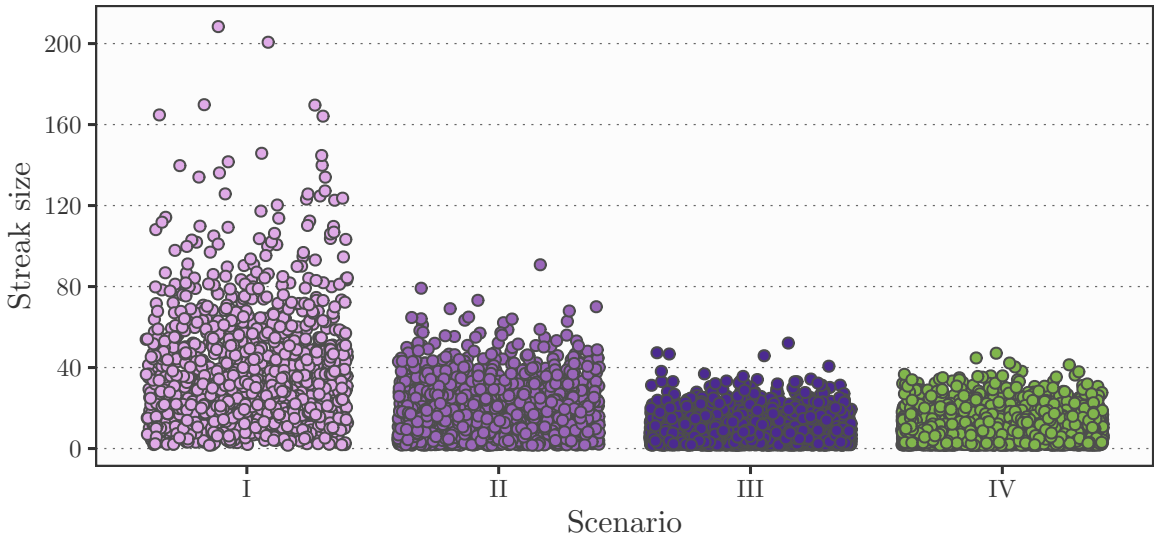

Figure 1: Sizes of win-streaks (number of consecutive wins, allowing for a single draw within any one streak) achieved by GM Nakamura in each scenario. The spacing of points in the x-direction within each scenario (jittering) is only to make more individual points discernable and conveys no information itself.

Keep in mind that since a single draw is allowed, a string of wins will be counted as part of streaks on either side of adjacent draws. Consider these outcomes:

$$w - w - d - w - w - w - d - w - l$$

Here, there are both one 5-game win-streak and one 4-game win-streak. The implication of this is that a pair of large streaks might be driven by the same string of wins (I refer to these as "overlapping games" between two streaks). It is also why the sum of streak-sizes is far greater than the number of games played (nearly double).

From Figure 1, it is clear that Scenario I yielded numerous streaks above 40 games. In fact, in this particular simulation, there were even two >200 streaks. And from inspection of the data, these streaks were not derived from overlapping games. For Scenario II the streaks were predictably less impressive, but still showed multiple >40 streaks, as well as a streak of 91 games. Scenario III also showed several streaks of 40 games or more (2 of these streaks shared overlapping games) with a maximum streak of 52 games. Finally, the Scenario IV simulation proved very similar to that of Scenario III.

5.2. How all simulations played out

The real interest of this study is to see in how many simulations that streaks of 45 games or more occurred (within each scenario).

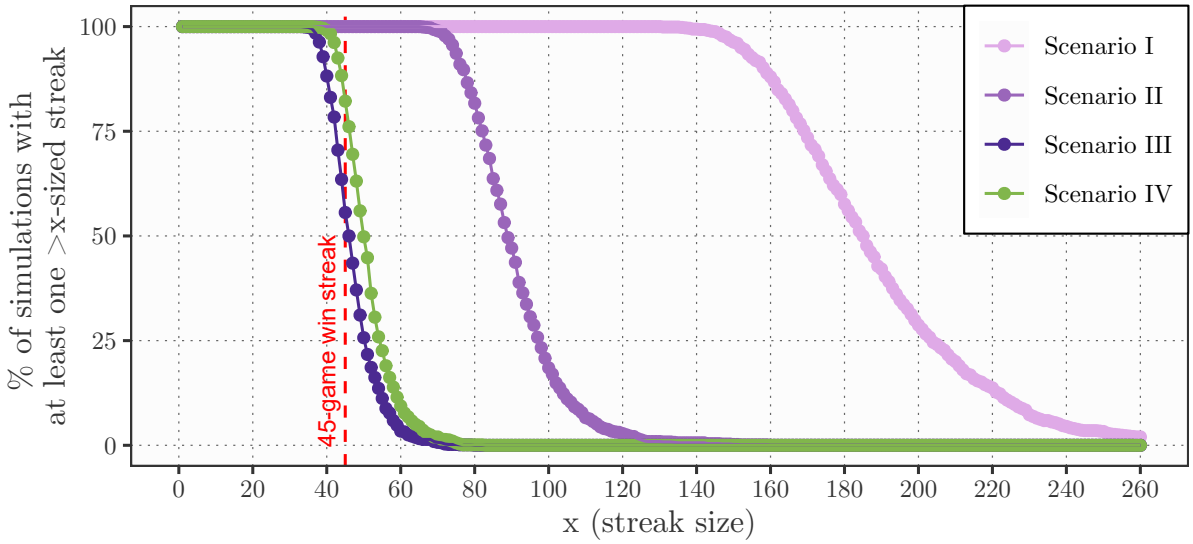

Figure 2: Percentage of 1000 simulations with at least one streak of >x games. The dashed red line indicates a streak of 45 games.

As foretold by Figure 1, a 45-game streak is trivial in Scenario I (see Figure 2). 100% of Scenario I simulations contained at least one >130-game streak. For this scenario, more than half of simulations contained >180-game streaks. Likewise, 100% of Scenario II simulations contained >45-game streaks. Only streaks upwards of 70 games stopped occuring in some simulations. Interestingly, for the least charitable scenario I investigated, Scenario III, about half of simulations contained at least one 45-game streak. It is also observable that the maximum streak achieved in Scenario III across the 1000 simulations counted about 70 games. For Scenario IV, more than 75% of simulations produced at least one > 45 streak.

6. Discussion & Conclusions

The question is: is it plausible for GM Nakamura to achieve streaks of 45 games or more (allowing for one draw)? To answer this question, I need here only focus on Scenario III. This is because Scenario III is the least charitable scenario to GM Nakamura (featuring a very optimistic view on streak-opponent true playing strength), and even here streaks of >45 games were likely to occur (about half of all simulations). So even at this extreme, it should still be very achievable for GM Nakamura to match (or even surpass) the actual streak that happened on Chess.com. Should the reader of this text take a view that listed FIDE ratings for each player are accurate or at least in the ballpark, my results would suggest that the question should rather be if GM Nakamura should be able to achieve >100-game streaks or >150-game streaks. These simulations are not real life. They do not account for the myriad of variables that exist in reality. One might for example imagine a fatal complacency sweeping in after winning the 29th game in a row, or tilt lowering the quality of play when losing a 10th time in a row. But what these results show is, that at a baseline mathematical level, a streak of 45 wins for GM Nakamura should not be an impossible event given his play activity against players like those he faced in the infamous streak - even when being charitable to the playing strength of the streak-opponents. One might argue that these results can only pertain to OTB chess, because I mostly employ OTB chess ratings and because online chess is not comparable to OTB chess. But if GM Nakamura indeed should be able to produce these streaks OTB, I would at a minimum consider using this knowledge as a prior for predicting and evaluating online performances.

In conclusion, I do not find evidence to support the argument that the size of the streak achieved by GM Nakamura of 45 wins and 1 draw should warrant suspicion of cheating.

7. Supplementary Information

7.1. Indicator of playing strength: FIDE vs Chess.com

It is not a straight forward task to choose between FIDE or Chess.com blitz-ratings as the measure of playing strength to use in the simulations.

Firstly, the streak in question happened on Chess.com - an online chess platform. Online chess is interesting in that it may incorporate and reward skillsets not present OTB such as "mouse skills" (the ability to move and play quickly using a computer mouse and quickly deploying pre-move strategies etc.). Players can (and most likely often do) play from a place of comfort such as at home, without the stress and nerves of OTB tournament play. OTB chess and online chess may therefore not be directly comparable - one could conceive of a player who is much stronger online than OTB. Another advantage of using Chess.com ratings to assess the plausibility of a Chess.com event is that Chess.com ratings are backed by a large dataset, usually thousands of games. OTB FIDE-rated events are more rare and are a lot less accessible. And as the streak-opponents are generally much younger than GM Nakamura, with relatively few recorded FIDE-rated blitz games, FIDE ratings may not have caught up with their actual playing strength.

But it is also the case that players on Chess.com have the ability to select opponents and play them any number of times - at any time during the day. This creates a way for players to try and "synthetically" maximize elo: a player could conceivably try to get only matches against stylistically favorable opponents (positional/solid/tactical/open/slow/fast/etc. players). They could also try and get matches against opponents they perceive to be "overrated" (to maximize the ratio between win-probability and elo-gain or the ratio of loss-probability and elo-loss). It is even possible to target only players coming off bad losing streaks (and who are thus potentially "tilting") or target players from vastly different timezones. There are probably many such ways to try and gain small advantages, which together create the phenomenon of elo-"farming". Additionally, players can play as many games in this casual environment as time permits, with shifting levels of concentration, ambition and commitment (which goes for opponents as well). And even if an individual player does not consciously engage in "farming" practices at all, the elo ecosystem might still be affected by the effects of farming and casual gameplay. Consequently, I suspect that online chess ratings experience much larger ebbs and flows when compared to OTB ratings. And this could to some degree obfuscate the actual playing strength difference between GM Nakamura and his streak-opponents. I also have to consider that if GM Nakamura is in fact cheating online, then it would not be sensible to use his online chess rating to predict outcomes - this rating could be tainted in a way that would work towards self-fulfilling the prophecy that he should win with a certain frequency. Conversely, OTB chess is less likely to be affected by cheating, as OTB cheating is considerably harder to achieve.

Ultimately, I chose to explore outcomes using both rating-systems.

7.2. Why "peak" ratings?

The rating of a player is not a static entity. It goes up and down from game to game. Even just a handful of games may cause a dramatic change in a player's rating. It quickly became clear to me that it was not so easy to choose a rating metric to compare players. If I for example used the current Chess.com ratings by the time of the streak, I might wonder whether each player's rating was at a low or high and I would worry that the playing strength difference between GM Nakamura and each streak-opponent would be obscured by the happenings of just those few days leading up to the event. Another option would be to go for the mean rating held by each player. But over what interval of time should I derive this mean? It is not only the rating that may shift over time, true playing strength might also change. Perhaps a player improves his or her game? Perhaps "player A" is a player who is sometimes very focused on playing good

chess, while at other times not so much. These, and other irregular performance patterns might create a multi-modal rating distribution for a player in which the mean might not be such an effective measure of "usual playing strength", but also not a good point of comparison between players. I chose peak ratings, because here each player was conceivably at his or her strongest, most focused and most "lucky" - a comparable circumstance between players. But admittedly, I would then be comparing the players at their best - a circumstance which is not likely to have been the case at the time of the streak. There are also some caveats to highlight here in terms of OTB ratings. Since players have not played thousands of OTB blitz games, there is a chance that this "peak" situation has not come about for any of the players. I do however still prefer peak ratings for OTB. This is because the young streak opponents may not even have reached their stationary level, or perhaps reached it only recently - a difficult situation in which to select an interval of time within which to calculate a meaningful mean-rating. In these situations, it seems to me that the peak ratings will at worst produce the same potentially exaggerated gap between the streak opponents and GM Nakamura as when using the mean, and at best reduce this exaggeration - especially because GM Nakamura's FIDE blitz ratings varies little over time.

7.3. Simulation algorithm description

In this section I detail the algorithm I constructed to simulate and count streaks over a specified set of games. At its core, the algorithm makes use of two streak counters that are offset.

Simulation algorithm

Input:

- `simID`: The identifier for the current simulation.
- n_{games}: The number of games to be simulated.
- p_{win}: The probability of winning a game.
- p_{draw}: The probability of drawing a game.

1. Initialize:

$$streakCounter = 0$$
$$drawCounter = 0$$
$$streakCounter\_alt = 0$$
$$drawCounter\_alt = 0$$
$$p_{\text{loss}} = 1 - p_{\text{win}} - p_{\text{draw}}$$
$$data = empty\ list$$

2. Game Simulation Loop:

for *i = 1 to* n_{games} **do**
  Determine Game Outcome from $p_{\text{win}}, p_{\text{draw}}, p_{\text{loss}}$:
  $$G_i \in \{\text{win}, \text{draw}, \text{loss}\}$$
  if G_i = *win* **then**
    if *drawCounter = 0 and* $i < n_{games}$ **then**
      streakCounter = streakCounter + 1
    if *drawCounter = 1 and* $i < n_{games}$ **then**
      streakCounter = streakCounter + 1
      streakCounter_alt = streakCounter_alt + 1
    if *drawCounter_alt = 1 and* $i < n_{games}$ **then**
      streakCounter_alt = streakCounter_alt + 1
    if *drawCounter = 0 and* $i = n_{games}$ **then**
      streakCounter = streakCounter + 1
      Append (simID, streakCounter) to data
    if *drawCounter = 1 and* $i = n_{games}$ **then**
      streakCounter = streakCounter + 1
      streakCounter_alt = streakCounter_alt + 1
      Append (simID, streakCounter) to data
      Append (simID, streakCounter_alt) to data
  if G_i = *draw* **then**
    if $i = n_{games}$ **then**
      Append (simID, streakCounter) to data
      Append (simID, streakCounter_alt) to data
    if $i < n_{games}$ **then**
      drawCounter = drawCounter + 1
      drawCounter_alt = drawCounter_alt + 1
      if *streakCounter = 0 and streakCounter_alt = 0 and drawCounter = 1 and drawCounter_alt = 1* **then**
        Reset drawCounter
        Reset drawCounter_alt
      if *drawCounter = 1 and drawCounter_alt = 1* **then**
        drawCounter_alt = drawCounter_alt - 1
      if *drawCounter = 2* **then**
        Append (simID, streakCounter) to data
        Reset streakCounter
        Reset drawCounter
      if *drawCounter_alt = 2* **then**
        Append (simID, streakCounter_alt) to data
        Reset streakCounter_alt
        Reset drawCounter_alt
  if G_i = *loss* **then**
    Append (simID, streakCounter) to data
    Append (simID, streakCounter_alt) to data
    Reset streakCounter
    Reset streakCounter_alt
    Reset drawCounter
    Reset drawCounter_alt
end

The astute observer may wonder why games are simulated one at a time and why streaks are updated after each game. Surely it would be more effective to play out all games first and then scan the whole block of games for streaks? I built the simulator this way, because I wanted to allow the possibility of win probabilities to dynamically change due to what was happening game to game. For example, I explored modelling "tilt", when losses accumulated or if win streaks grew and grew (however, I did not include these in this work).

Bibliography

Chess.com (2024). URL: https://chess.com. Accessed: 2024-06-27.

FIDE (2024). URL: https://ratings.fide.com. Accessed: 2024-06-27.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Kramnik, V. (2023). INFORMATIONAL. URL: https://www.chess.com/blog/VladimirKramnik/informational Accessed: 2024-06-27.

Labelle, F. (2024). Elo win probability calculator. URL: https://wismuth.com/elo/calculator.html. Accessed: 2024-06-27.

Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2.* Python Software Foundation.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.