JOURNAL OF
EMERGING INVESTIGATORS

Article

A comparative study of dynamic scoring formulas for capture-the-flag competitions

Ethan Ho[1], Michael Steele[1]
[1] Mathematics, Franklin High School, Elk Grove, California

SUMMARY

Recently, the use of gamification techniques to teach cybersecurity has increased, predominantly through capture-the-flag security competitions where players solve cybersecurity problems (challenges) to gain points on a scoreboard. To determine how point values are determined for each challenge, a formula is often used to assign point values based on the number of teams that have solved the challenge, tying point values to the difficulty of the challenge. There are many ways to create this formula; however, the effects that different scoring formulas can have on a competition is not clear. This is important because the way that scores are calculated can impact the motivations, morale, and overall experience of participants as the scores are reflected on scoreboards. In this study, we examine the effect that changing the scoring formula used in these competitions would have. We predicted that changing the scoring formulas would have a large effect on the distributions, with more gently sloped formulas moving the center of the scoring distribution to a higher number. This is because of the fundamental differences in the formulas' mathematical compositions. We found that while the overall distributions and median scores stay largely the same when the competitions were re-scored using different formulas, the variability in the scores differed. This is important because score variation and distribution during a competition can have serious implications in competitions that aim to find talent or introduce people to the field.

INTRODUCTION

In recent years, we have seen an increase in the use of gamification techniques in various areas of education. This use of scoreboards, systems for scoring points through reaching certain levels of achievement, and competitive environments have allowed for many to learn more effectively through pushing participants to set goals and strive to continue learning (1).

Being found in many large cybersecurity programs, including education programs like CyberStart and PicoCTF, as well as talent search programs such as Battelle's CTF, gamification has been utilized extensively in cybersecurity education (2). Especially in high school and college environments, capture-the-flag (CTF) competitions aim to introduce and test technical skills in computer security, often requiring participants to use their skills to find vulnerabilities in computer programs. In the most popular format for these competitions, participants solve individual challenges, which require them to reverse engineer, understand, and attack systems or applications in order to obtain a proof of completion, called a flag. This flag can be redeemed for points on a scoreboard over the course of the competition, meaning that the competition is decided by the number of flags that participants capture. Through turning cybersecurity into a series of challenges in a competition, participants become more engaged and learn more about the subject matters tested (3).

However, with increased gamification of cybersecurity comes vital decisions about how these scoring systems are utilized and configured. Optimizing the variables in how scoring systems work can potentially increase the effectiveness of these educational games, as they impact how people interact with the educational material. One important matter of focus is the structure of the scoreboard and point system itself. The way that challenges are scored directly influences what participants are rewarded for focusing on during competitions. Because of how scoring systems affect competition participants, research has been done to investigate the optimal methods and formulas for scoring (4).

The two major types of scoring for capture-the-flag competitions are static scoring and dynamic scoring. In a static scoring system, each challenge in the competition has a set number of points, assigned at the competition start by the organizers of the event. When a player or team obtains the flag for the challenges, their points on the scoreboard increases by a set amount, often set to a smaller value for an easier challenge or a larger value for a harder challenge (5). However, this method of scoring can often lead to inaccurate representations of challenge difficulty because without extensive testing, the true difficulty of a challenge can be hard to determine (6). To solve this problem, some competitions use a dynamic scoring system, which uses the number of solves for each challenge to calculate the point value of the challenge. At the beginning of the competition, each challenge is weighted equally. Each time a challenge is solved, the point value of the challenge is decreased, making the point values of each challenge at the end of the competition reflect how many people solved it (5). As more people solve easier challenges and less people solve harder challenges, this scoring method assigns difficulty measures using competition data.

In recent years, dynamic scoring has become the most popular method of scoring capture-the-flag competitions. This is because scoring challenges dynamically offers a clear advantage over pre-set point values subjectively by competition organizers beforehand and introduces an accurate way to gauge challenge difficulty (6). However, there is currently no standardized formula for the way dynamic scoring is implemented. The formula used to determine how point values decrease can be linear, logarithmic, or polynomial, which can

https://doi.org/10.59720/23-173

all impact the rate of point decay. Because of this variability in possible scoring methods, there may be lost potential for competition improvement, as changing the dynamic scoring formulas in a competition may be a tool to improve player experience.

Our study examines the effect of different scoring formulas on the final score distributions in capture-the-flag competitions. We predicted that different scoring formulas (based on polynomial, logarithmic, rational, or linear functions) would greatly change the score distributions, altering how scores are spread throughout the scoreboard.

To test this, we re-scored the final results of two capture-the-flag competitions (ImaginaryCTF 2021 and 2022) using several of the most widely used scoring formulas. We used the CTFd, Order of the Overflow (OOO), and Chaos Computer Club (CCC) formulas, as well as a static scoring formula for reference (with each challenge being worth the same number of points). Each formula uses a different mathematical function for calculating score decay as a challenge is solved by more teams. The CTFd formula utilizes a quadratic equation, the OOO formula utilizes a logarithmic function, and the CCC formula uses a rational function (**Table 1**) (5).

After re-scoring the results using each formula, we compared the final score distributions. Through this experiment, we found that although final scores are minimally changed by differences in scoring formulas, different formulas do influence the variability of the final scores. This can have implications for player morale and motivation, as looking at the scoreboards during the competition can be discouraging when the top competitors have much higher scores than everyone else. By creating competitions where scores for different teams are closer together, we can more effectively make educational competitions where players are encouraged to keep trying new challenges and learn about new topics. In the same way, by changing scoring formulas to make competitions with scores that are farther apart, we can also create competitive events that more effectively differentiate top players.

	Scoring Formula, a=500, b=100
CTFd formula	$points(x) = \frac{b-a}{5000^2}x^2 + a$
OOO formula	$points(x) = b + \frac{a-b}{a + 0.08x log_{10}x}$
CCC formula	$points(x) = b + \frac{a-b}{1 + (max\,(0, x-1)/11.92201)^{1.206069}}$
Static formula	$points(x) = 100$

Table 1: Mathematical formulas for selected CTF competition scoring formulas. Table showing the formulas behind the CTFd, OOO, CCC, and static formulas, which calculate point values for a challenge as a function of solve counts. The formulas were taken from the Order of the Overflow's scoring-playground project and configured with challenges starting at 500 points and decaying to a minimum of 100 points (4). In all formulas, a is the maximum number of points for a challenge, b is the minimum number of points for a challenge, and x is the number of solves on a challenge.

RESULTS

We compared the final scoreboards of CTF competitions (with 1018 data points for ImaginaryCTF 2021 and 813 data points for ImaginaryCTF 2022) when re-scored with multiple different scoring formulas in order to find the effect that different dynamic scoring formulas would have on final score distributions in capture-the-flag competitions. We picked the CTFd, OOO, CCC, and static formulas to re-score with because each formula uses a different mathematical function to calculate point values.

To process the scoring data from each competition, we stored scoring data from each competition in a database. This data included information about which teams solved which challenges during the competition. We used each of the four scoring formulas to re-assign point values to the challenges and calculated the final scores of each team based on these point values (**Figure 1**). After re-calculating the final scores for each competition under each formula, we compiled a list of each team in the competition with its corresponding score. We then recorded the median, standard deviation, minimum, and maximum of each score distribution.

Comparing the score distributions of the scoreboards of ImaginaryCTF 2021 and 2022 with each scoring formula, we found that there was little to no difference in the overall spread of the scores when using each formula. For all of the

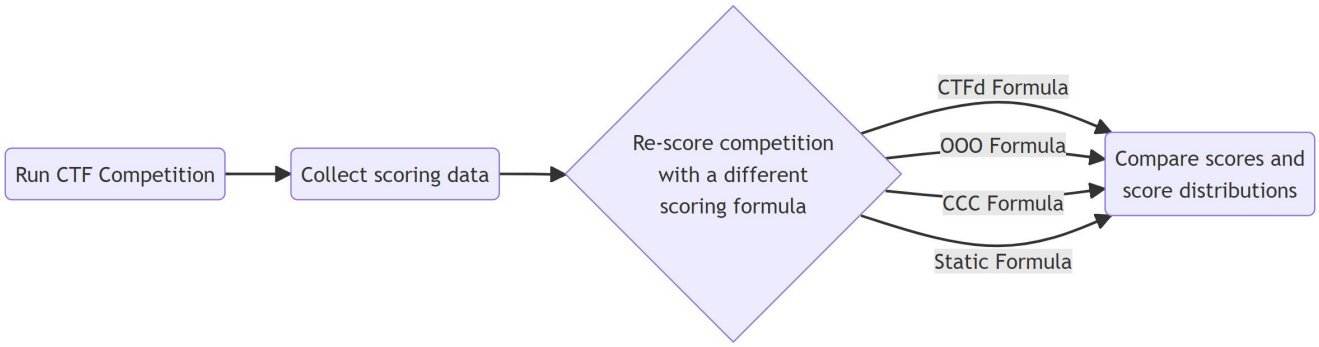

Figure 1: Competition re-scoring methodology showing how different scoring formulas were tested on the same data. Scoring data from the competitions was re-scored using different scoring formulas to test the effect of different formulas on final results. The process to test the effects of different scoring formulas is depicted. This begins with running a competition and collecting the data, and re-scoring each competition using the different scoring formulas used in this study. The re-scored competition data is then compared across the different formulas.

https://doi.org/10.59720/23-173

	Median	Standard deviation	Minimum score	Maximum Score
CTFd formula	600	1502	0	13371
OOO formula	604	877	0	7344
CCC formula	616	947	0	8821
Static formula	600	793	0	5500

Table 2: Median, standard deviation, minimum, and maximum scores in ImaginaryCTF 2021. Scoring formula variations create small changes in median scores while greatly changing standard deviations. Final scores from ImaginaryCTF 2021 were re-scored using four scoring formulas (n=2 repetitions).

	Median	Standard deviation	Minimum score	Maximum Score
CTFd formula	673	1646	0	16121
OOO formula	619	1048	0	8864
CCC formula	657	1217	0	10657
Static formula	600	855	0	5700

Table 3: Median, standard deviation, minimum, and maximum scores in ImaginaryCTF 2022. Scoring formula variations create small changes in median scores while greatly changing standard deviations. Final scores from ImaginaryCTF 2022 were re-scored using four scoring formulas (n=2 repetitions).

formulas, the amount of players in each range of points is similar (**Tables 2, 3**). The median scores for both 2021 and 2022 vary between 600 and 673 points with different scoring formulas, which is a very limited range (**Tables 2, 3**). The difference in total scores between using different formulas is less than the point value of a single challenge (100 to 500 points); this shows how different scoring formulas do not greatly impact the median scores in a competition (**Figure 2).** However, we found a difference in the distributions of the scores, as there is large difference in the standard deviations of the scores under different formulas. The CTFd scoring formula produced the greatest standard deviation (1502 and 1646 points for ImaginaryCTF 2021 and 2022, respectively), while the static scoring created the smallest standard deviation (793 and 855 points for ImaginaryCTF 2021 and 2022, respectively). The maximum scores for the CTFd formula are higher as well (13371 and 16121 points for ImaginaryCTF 2021 and 2022, respectively) than those from all other formulas (ranging from 5500 to 8821 and 5700 to 8864 for ImaginaryCTF 2021 and 2022, respectively) (**Tables 2, 3**). This demonstrates how the highest scores in the competition are higher when the CTFd formula is used.

DISCUSSION

In this study, we investigated how much the use of different scoring formulas in capture-the-flag competitions affects final scoring distributions. We predicted that different formulas would result in changes in both the median and the variability of the final scores. However, we found that changing the scoring formula does not greatly impact the median scores of the competition scoreboards but does impact the variability of the final scores.

We found that some formulas (the CTFd and CCC formulas) tend to make the highest scores in the competition much greater than the average scores. This is because the average rate of point decay for challenges as they are solved is much slower than in the other formulas, making more challenges worth a larger number of points. This makes the final score distribution more spread out, with the highest scores farther away from the average.

The presence of these disproportionately high values on competition scoreboards can influence how participants interact with the competitions. Having a large score disparity on a scoreboard can mean that the competition effectively differentiates the best players from the others. This would mean that differences in skill level from player to player will be amplified through the scoreboard and would be clearly seen through larger gaps in point values. However, this separation between the best players and the average player can also have discouraging effects on the average player, making it seem harder than it really is to reach higher levels of achievement during the competition and increase their position on the scoreboard.

This experiment suggests that the decision for what scoring formula to use in a competition is best decided based on the goal behind the event itself. For events that aim to find the most skilled participants in the field, such as qualifying events for teams (including qualification events for national competition teams such as those for the International Cybersecurity Challenge), a formula with a larger standard distribution like the CTFd formula or CCC formula are more suited (7). However, for events whose main focus is education or introducing beginners to new topics, a formula like the OOO formula with a lower standard distribution, or even a static scoring system may be more desirable. This can help to create a competition environment that, while remaining competitive, prevents participants from being discouraged seeing that the top teams or players have many times more points than the average player.

This experiment was limited in the amount of data involved. The data used for this study were drawn from only two competitions (ImaginaryCTF 2021 and 2022), which had similar challenge difficulties and competition format (8,9). Due to this, more replications of this study may be required with challenges from different authors and competitions with different average difficulties to produce more generalizable results. In addition to this, one factor that could have introduced bias into the results was the original scoring format of the competitions, showing a limitation in recalculating scoreboards from past competitions to test different scoring formulas. Because the competitions had live scoreboards, the original scoring formula (which was a modified version of the CTFd formula) for the scoreboard may have played a role in how the competition played out originally. The formula used to determine

https://doi.org/10.59720/23-173

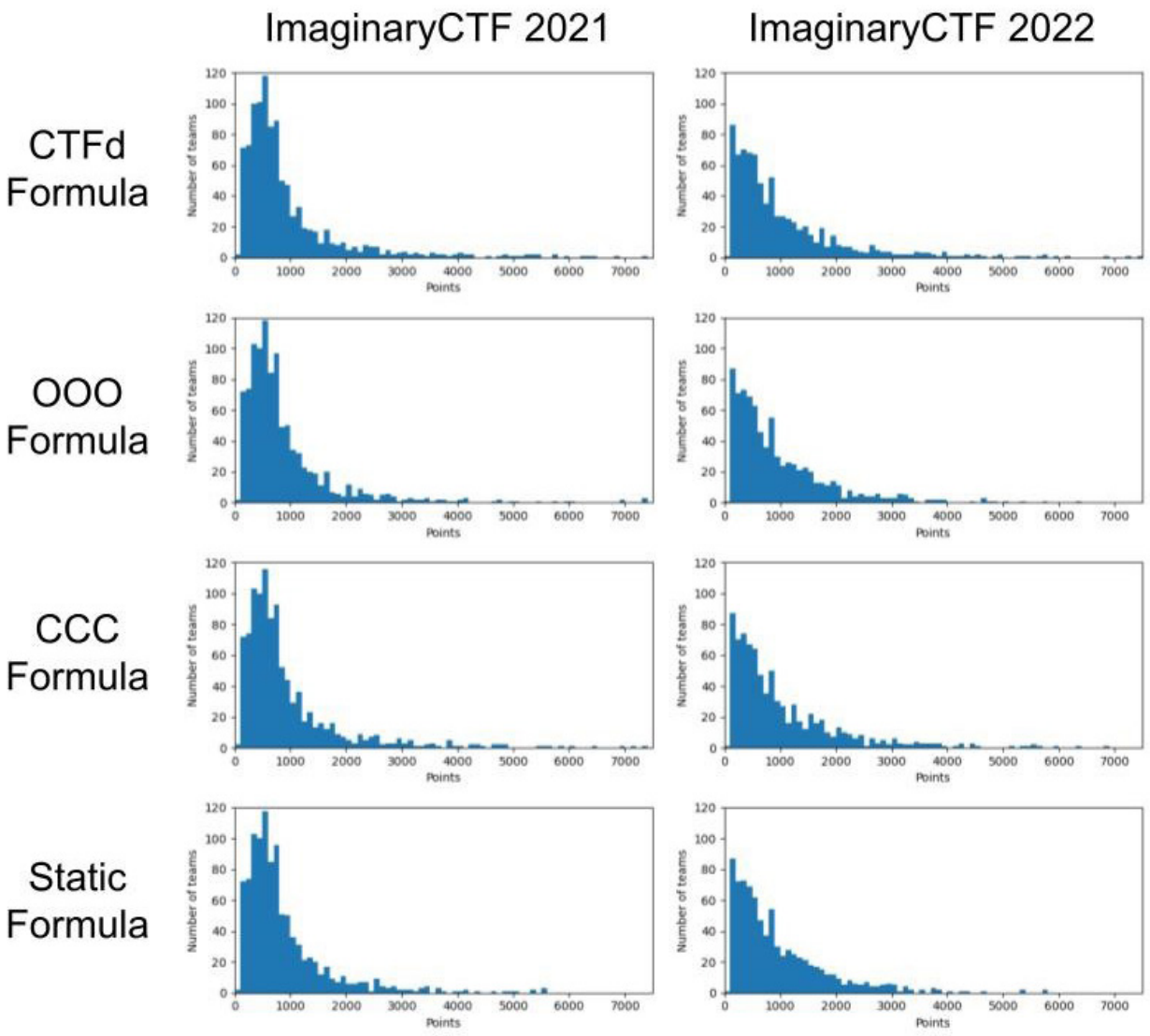

Figure 2: Minimal differences in scoring distributions when competitions are re-scored using different formulas. Histogram showing spread of scores for the studied competitions (with 1017 data points for ImaginaryCTF 2021 and 813 data points for ImaginaryCTF 2022) when re-scored post-competition with the CTFd, OOO, CCC, and static scoring formulas (n=2). Competition scoreboard data including the number of solves per challenge and challenges solved per player were inputted into each scoring formula, and the scoreboards for each competition were regenerated based on each formula used.

point values for dynamic scoring during the competition could have made players value some challenges more than others during the competition, affecting how they allocated their time and effort. To gain more accurate results, future experiments should use the selected scoring format during the competition itself, rather than re-scoring post-competition.

MATERIALS AND METHODS

Data - ImaginaryCTF 2021 and 2022

We experimented with the effects of different scoring formulas using data collected from capture-the-flag competitions, taken from ImaginaryCTF 2021 and 2022. The data was pulled from the competition MySQL database as well as the archived competition websites. This data included information about challenge solves per team, such as how many teams solved each challenge at the end of the competition. With this information, it was possible to reconstruct what the scoreboard and score distribution of the competition might have looked like at the end of the event.

ImaginaryCTF 2021 had 56 challenges, and 1017 participating teams. The most solved challenge, a "sanity check" designed to make sure players knew how the system worked, had 996 solves. The least solved challenge, a reverse engineering challenge, had 5 solves. In the end, three teams solved all the challenges, achieving the maximum score possible in the competition.

ImaginaryCTF 2022 had 58 challenges, and 813 participating teams. The most solved challenge was also a "sanity check," which had 616 solves. The least solved challenge, a systems security challenge, had 5 solves. Two teams solved all the challenges.

Data analysis

We utilized a custom Python script to analyze and re-

https://doi.org/10.59720/23-173

score the competition data. The script uses the Python pyplot library to create the graphs and process the data (**Figure 1**).

Received: July 26, 2023
Accepted: September 22, 2023
Published: August 30, 2024

REFERENCES

1. Landers, Richard N., et al. "Gamification of Task Performance with Leaderboards: A Goal Setting Experiment." Computers in Human Behavior, vol. 71, 22 Mar. 2017, pp. 508–515, https://doi.org/10.1016/j.chb.2015.08.008
2. McDaniel, Lucas, et al. "Capture the Flag as Cyber Security Introduction," 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 2016, pp. 5479-5486, https://doi.org/10.1109/HICSS.2016.677 Crawley, K. "What is CTF? an introduction to exciting hacking games." *Hack The Box*, 21 March 2022. Accessed 09 July 2023.
3. Sharma, Sugandha, et al. "Intensifying Practical Based Learning of Penetration Testing using CTF," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 1378-1381, https://doi.org/10.1109/ICAC3N53548.2021.9725762
4. Shoshitaishvili, Yan, et al. "O-O-Overflow/Scoring-Playground: Tool to Test Different CTF Scoring Algorithms on Real Data." *GitHub*, Order of the Overflow, 3 May 2021. Accessed 09 July 2023.
5. Chung, Kevin, et al. "Dynamic Value." CTFd Docs, CTFd LLC. Accessed 22 July 2023.
6. Gulick, Jessica, et al. "US Cyber Team - Event Schedule." US Cyber Games. Accessed 22 July 2023.
7. "CTFtime.Org / ImaginaryCTF 2021." *CTFtime.Org / All about CTF (Capture The Flag)*, May 2021. Accessed 08 Aug. 2024.
8. "CTFtime.Org / ImaginaryCTF 2022." *CTFtime.Org / All about CTF (Capture The Flag)*, May 2021. Accessed 08 Aug. 2024.

APPENDIX

The Python code used to re-score competition scoreboards can be found at https://gist.github.com/Eth007/eb230018fd3f3ed7203e0cf063772547.