Whale of a Crowd: Quantifying the Effectiveness of Crowd-Sourced Serious Games

Umit Tellioglu*, Geoffrey G. Xie†, Justin P. Rohrer‡ and Charles Prince§
Graduate School of Operational and Information Sciences
Naval Postgraduate School
Monterey, California 93943
Email: *utelliog@nps.edu, †xie@nps.edu, ‡jprohrer@nps.edu, §cdprince@nps.edu

***Abstract*—In this paper we analyze several Crowd-Sourced Serious Games (CSSGs), a new genre focused on advancing widely respected causes such as social equality and science. We observe that the general effectiveness of these games has remained largely unknown. Existing performance analyses have been limited to documenting experiences with individual systems. More importantly, existing game analytics approaches are designed for games that provide personal experience and entertainment. In contrast, CSSGs attract participants by evoking their sense of social responsibility and sympathy for others. Intuitively, social awareness and sympathy alone may not result in the same level of consistent participation as personal achievement, or fun. Consequently, the success of a CSSG may be more tightly linked to the contributions of few highly-dedicated players (whales).**

***Keywords*: Crowd-Sourced Serious Games, Game Analytics, Whale Effect Graph, Player Engagement Rate**

I. Introduction

The increasingly pervasive Internet provides a platform for effective group communications at a global scale, even among otherwise strangers living in different continents. This transformation in communication has led people to envision crowd-sourcing as a potentially cost-effective method for tackling tasks that previously can only be performed by domain experts. Two highly-publicized executions of this vision are the Duolingo portal [1] and the EyeWire project [2]. The ultimate goal behind the free-of-charge Duolingo portal is to translate the web into all major languages, and the "crowd" is made of people who desire both to learn a foreign language and to support the cause of making useful web content universally accessible. Most of the exercises and exams completed via the Duolingo portal are in fact translating fragments of some real-world web pages from one language to another. The underlying purpose of the EyeWire project is to decipher the structure of the human brain at the neuron level. The researchers set up a web front-end in the form of a virtual I-spy game to recruit a crowd of volunteers to accelerate the process of mapping 2-D images of brain slices into 3-D neuron connectivity patterns.

More recently, the concept of crowd sourcing is also being explored in the highly specialized field of formal software verification [3]. A collection of puzzle style games, called VeriGames, has been created and hosted publicly on the Internet. Each instance of a game level corresponds to an attempt to assert some properties about a code segment. A backend verification engine then combines the assertions produced from all related game instances and tries to obtain conditions that can rule out certain types of bugs in that code segment.

In this paper, we broadly classify such crowd-sourcing efforts into a new genre called Crowd-Sourced Serious Games (CSSGs) as their primary focus is to advance widely respected causes such as social equality (in the case of Duolingo) and science (in the cases of EyeWire and VeriGames). Furthermore, we observe that the general effectiveness of these games is largely unknown. The few performance analyses in current literature are limited to documenting experiences with individual systems. More importantly, existing game analytics approaches are designed for games that provide *personal* experience and entertainment. In contrast, CSSGs attract participants by evoking their sense of social responsibility and sympathy for others. Intuitively, social awareness and sympathy alone may not result in the same level of consistent participation as personal achievement, or "fun". Consequently, the success of a CSSG may be more tightly linked to the contributions of few highly-dedicated players (commonly referred to as whales in the current literature).

It is important to develop a systematic methodology to accurately characterize the performance of CSSGs in order to identify the best practices for improving them as a genre. As a first step to this end, we collect data from a sample set of CSSGs and evaluate the following two hypotheses.

1) Player retention is more challenging for crowd-sourced serious games (CSSGs) than for traditional games (whether leisure or serious games).

2) The difference in achievement levels between whales and typical players is bigger with CSSGs than the traditional games. In other words, it might be more critical for CSSGs to not just recruit new players, but retain highly-productive players, and at the same time incentivize existing players to elevate their productivity.

The results show that CSSGs have smaller audience and higher fraction of non-returning players when compared with traditional games. Moreover, existing loyal CSSG players play games infrequently, as a result of low engagement. Lastly, there is a small group of players who account for the majority of the productivity both in CSSGs and traditional games. This small audience is very important for CSSGs to attract and maintain, given their small audience overall.

TABLE I. AVERAGE DAU AND MAU FOR SELECTED MOBILE AND SOCIAL ONLINE GAMES

	DAU	MAU
Zyanga*	11.1M	292M
Storm8*	4M	-
Glu Mobile*	3.4M	29M
Angry Birds	20M	200M
Temple Run	7M	-
Stardom	74K	-
Deer Hunter	271K	-
Junkies	114K	-
Triple Town	-	160K
Parallel Kingdom	-	50K
DeNa*	-	16.9M
GREE*	-	13.9M

*A collection of games from the named game developer/operator.

TABLE II. DAU OF SAMPLE VERIGAMES

	DAU				
	Min	Max	Mean	Median	StDev
VeriGame A	2	872	71.5	23	158.2
VeriGame B	1	887	64.5	16	135.5

II. RELATED WORK

CSSG developers have often provided the total number of registered participants as an indicator of game success. For example, in a November 2013 press release, Duolingo claimed 14 million registered users. EyeWire researchers stated in a recent paper [2] that more than 100,000 registered players from more than 130 countries had contributed to their experiment.

Other CSSG developers have used a measure of work performed to assess the contributions of their crowd toward the motivating cause. The creator of Phylo, a CSSG whose players solve puzzles to help find solutions to genetic disorders, reported obtaining a total of 254,485 completed puzzles (generated by $\sim$12,000 registered players) in the first seven months of deployment [4]. The Malaria Training Game (MTG), created for advancing the concept of tele-diagnosis of diseases, was able to screen more than 1.5 million red blood cell images for malaria infection in less than 4 months, with the help of 2,150 people from 77 countries [5]. Comparative studies are also applicable in some cases. One such study concluded that Duolingo is more effective than Rosetta Stone or college classes in helping people to learn a foreign language [6].

Finally, the literature on CSSGs repeatedly describes the presence of and the key roles played by a few whales in the crowd. For example, according to one study [4], top 10% Phylo players (in terms of their skills of solving puzzles) participated in nearly 80% of the completed puzzles.

Common to all these studies is that their data and conclusions are specific to an individual game. The general effectiveness of CSSGs and methodologies for applying the classic commercial game analytics to this new genre have not been examined. This observation is not unexpected, given the relatively short history of CSSGs.

TABLE III. MAU OF SAMPLE VERIGAMES

	MAU				
Month	Dec	Jan	Feb	Mar	Apr
VeriGame A	7555	957	615	415	460
VeriGame B	5000	504	244	-	-

TABLE IV. SUMMARY OF CSSG DATA USED

	Collection Period	Total Users
VeriGame A	1 Dec 2013 - 9 May 2014	1475 Reg. 8399 Anon.
VeriGame B	1 Dec 2013 - 17 Mar 2014	717 Reg. 7029 Anon.
EyeWire	Since December 2012	Over 100K
Foldit	Since May 2008	Over 500K
Phylo	Dec 2010 - Jun 2011	Over 12K
MTG	May 2012 - Aug 2012	Over 2K

III. METHODOLOGY

A. Data Sets

In our research we refer to datasets belonging to two game types. The first dataset, from gamesbrief.com [7], includes Daily Active Users (DAU) and Monthly Active Users (MAU) [8], [9] statistics for mobile and social online games which have been compiled from various resources. The data for each game (Table I) includes averages calculated during several months within 2011 and 2012.

The second dataset consists of players' session and productivity data for two games from verigames.com (referred to as VeriGame A and VeriGame B in the rest of the paper). This data was obtained directly from game developers. VeriGame A data has information for more than 30K player sessions, while VeriGame B has more than 100K sessions. Our analysis included data not only for registered players, but also for anonymous players because both games support anonymous play. The DAU and MAU statistics for these two games are shown in Tables II and III.

We also gathered data from the literature about several other CSSGs including EyeWire [2], Phylo [4], Foldit [10], and Malaria Training Game (MTG) [5]. The sizes of these additional data sets are shown in Table IV, and we will refer to this data in Section IV.

B. Metrics

It is relatively simple to measure productivity of retail electronic games: Count DVDs/CDs sold, multiply with sell price, and compare with the cost of producing the game. Productivity in the commercial online gaming market (with a similar ecosystem to that of CSSGs) is a much more complex function of purchase price ($0 in many cases) along with in-game purchasing and subscriptions. Theoretically, a player can spend zero to infinity dollars. In other words, while players traditionally spent a constant amount for a retail game, their spending can significantly exceed that amount for free-to-play games [11]. Due to these new pricing paradigms, not only maximizing the number of players, but also transforming free-players into paying-players are important issues for online games.

Because of fluctuations in player spending over time, it is vital that game developers track players' attitudes towards particular games. Two of the most common metrics to measure players attitudes towards games are daily active users (DAU) and monthly active users (MAU) [12], [13]. According to Fields [12], DAU is the count of unique players in a day, and MAU records either unique or non-unique players in a calendar month. In our research we counted unique users for both DAU and MAU. In addition, we used weekly active users (WAU) to count unique users in a seven day period.

Engagement Rate: Although DAU and MAU are very useful metrics, as independent values they are insufficient to represent games' potential because they count all players, including non-returning one-time players, without capturing user engagement [12]. By examining the relationship between DAU and MAU we are able to quantify the engagement rate (ER) of players. Formally, we define ER as the DAU to MAU ratio:

$$ER = (DAU/MAU) * 100 \quad (1)$$

This metric represents a game's "stickiness", which also roughly expresses its ability of retaining players. In addition, ER may provide an indicator about the long term success of a game [12], [13].

Whale Effect Graph: As was shown by Pareto's 80-20 rule, which basically claims there is an unbalanced situation between input and output, players' spending is not uniformly distributed in free-to-play online games [14]. A small subset of players called *whales* (a term borrowed from the casino gambling industry) far outspend average players. Jesse Divnich defined whales as the top 5% of spenders [15]. He considered whales to be players that spend more than ten dollars monthly for online mobile games. While that doesn't sound very impressive, it constitutes a large percentage of the total revenue of online games. For example, a director of *Clash of Dragons* declared 40% of the in-game purchases were made by only 2% of players [16]. A recent report about monetization in mobile games also shows that 50% of revenue comes from 0.15% of players [17]. At this point a standardized definition of a "whale" has not been established, and each game determines which players are whales based on a different standard.

To study the effects of whales on the VeriGames and CSSGs in general, we propose to use a Whale Effect Graph (WEG), an example of which is shown in Figure 1. In this graph the x-axis shows the cumulative percentile of players sorted by productivity and the y-axis shows the cumulative percentile of overall game productivity. In other words, any point on the curve shows the percentage of contribution to the overall productivity produced by the selected fraction of the most effective players. Therefore, in contrast to focusing on either an arbitrary fraction of top players or the cumulative distribution of players based on their productivity, a WEG provides a *complete view* of how players of different productive levels contribute to the *overall productivity* of the game.

In the case of VeriGames, the goal is not monetization, so in order to measure productivity we were required to choose metrics other than money. Based on advice from the developers of the two games we choose to quantify productivity using the assertion count for VeriGame A, and the game score for VeriGame B. Since these two metrics are measured on different scales, we normalize the values and present the results in Section IV using percentile graphs.

TABLE V. THIS TABLE SHOWS ER OF GAME A AND GAME IN THE MONTHLY AVERAGE BASIS

	Monthly Engagement Rate (%)					
Month	Dec	Jan	Feb	Mar	Apr	**Avg**
VeriGame A	3.41	3.85	4.36	4.10	3.59	3.86
VeriGame B	3.27	3.39	3.71	-	-	3.45

TABLE VI. THIS TABLE SHOWS ER OF SOME MOBILE AND ONLINE SOCIAL GAMES AND DEVELOPERS

	Average Engagement Rate (%)
Angry Birds	10
Parallel Kingdoms	30
Glu Mobile*	11.7
Zynga*	22.5
Scrabble	30
Bejeweled Blitz	27
Pet Society	14

Total Session Times and Counts: The ER metric as defined above has a limitation in that it cannot capture the magnitude of total player activities. For example, ER = 1 even if only five players remains for a game, as long as they are active every day of the month. Therefore, we also use the aggregate session time (ST) and session count (SC) metrics to analyze CSSGs, as done in prior work [18]. ST is the amount of time a player interacts with a game until leaving. We counted ST as hours in our research. SC shows how many times a game is played. We measure ST and SC over different time intervals such as weekly (WST, WSC) and monthly (MST, MSC).

These game play metrics are closely related to the game productivity and whale effect. Recent research shows that while paying and non-paying players have an average WST of about four hours, whales typically spend close to twelve hours each week [15].

IV. EVALUATION

In this research we use data from both registered and anonymous players to derive metrics. Extended information about players for both VeriGame A and VeriGame B is shown in Figure 5.

Engagement Rates: We have obtained the monthly ERs and the averages for VeriGame A and B using the data of Tables II and III. The results are shown in Table V. Similarly, we have derived the average ERs for some of the sample commercial online games listed in Table I. The results along with data available from the literature [7] are shown in Table VI. We observe that the ERs of VeriGame A and B are less than 5%, much lower than the ERs of the commercial online games, which are between 10% and 30%. In particular, the ERs of the VeriGames are the lowest in the first month of deployment, although the number of players recorded (MAU) is the highest for that month. We attribute the high drop off rate of MAU primarily to having low engagement rates, caused by non-returning players. The ERs of the VeriGames tend to increase

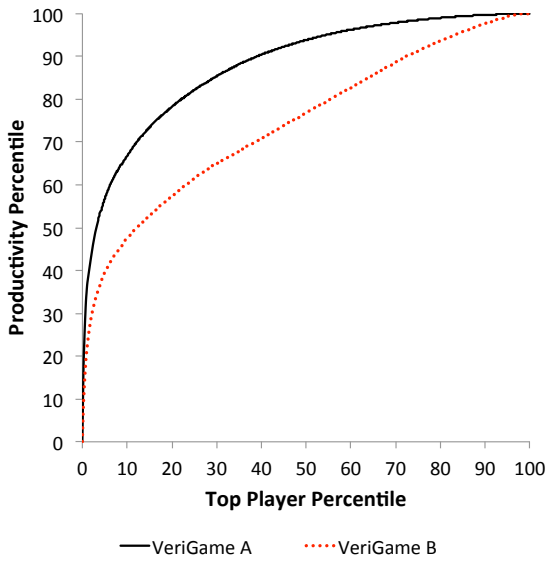


Fig. 1. Productivity percentiles of productive registered players

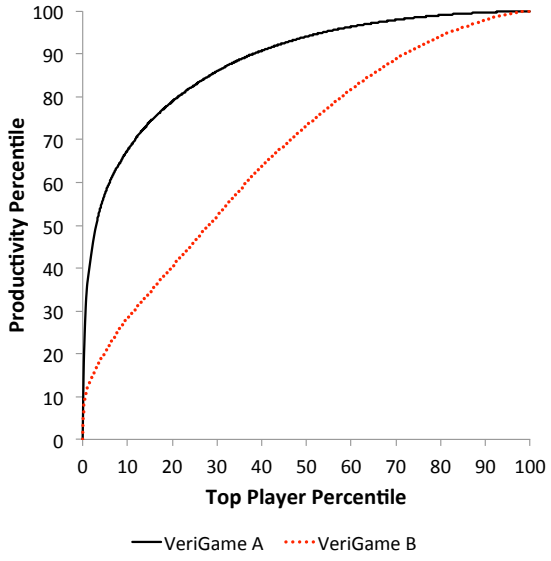


Fig. 2. Productivity percentiles of productive registered & anonymous players

monthly while MAU is steadily decreasing over the first three months. This may show that the VeriGames obtained a core set of loyal people who keep playing.

When examining other CSSGs, we saw that they also have low engagement rates and high drop off rates. For example, Phylo, which requires players to solve puzzles to assist in finding a solution for genetic disorders, had around 12,000 registered players seven months after release, but only 23% of those players returned one more time to play the game [4]. 42% of acquired Phylo players gave up playing without completing a single puzzle [4]. FoldIt has more than 500K players on its soloist leaderboard [10], but about 80% of those players have not scored any points, which also indicates a high drop off rate and possible low engagement rate.

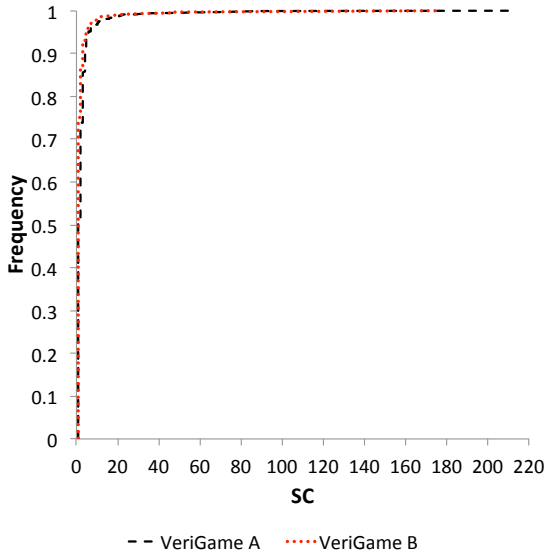
Whale Effect: We emphasized the importance of whales in Section III. When we examined VeriGames A and B, we saw that in both cases there is a small group of whales performing significantly better than the other players. Figure 1 shows the whale effect graph (WEG) for the registered users of these games. The rapid increase in productivity percentile over the first few percent of the players on the WEG shows the effectiveness of the whales. For VeriGame A, over 60% of the productivity is attributable to less than 10% of the players. For VeriGame B the top 10% of players produce more than 40% of the overall productivity. The steeper curve of VeriGame A clearly indicates that the whales of VeriGame A are more productive than those of VeriGame B. In other words, VeriGame A relies more on whales than VeriGame B.

Figure 2 shows the WEG after including data from all players, including those that do not choose to register. The WEG curve of VeriGame A has the same shape as before, while the slope of the curve for VeriGame B is more linear, possibly resulting from distinctive game mechanisms. Particularly, VeriGame B allows non-registered players to accumulate scores while VeriGame A does not.

When examining other CSSGs, we saw that whales are also important for them, and the fractions of whale are low compared to commercial games. 90% of registered 12K Phylo players finished less than 25 puzzles while the top 10% of players participated in nearly 80% of all solutions produced by registered players and the top 20 players solved more than 700 puzzles each [4]. On FoldIt's soloist leaderboard, three players have more than 40K points each, eight players have between 40K-30K each, 27 players fall between 30K-20K, and 64 players are between 10K-20K [10]. This indicates a similar WEG curve for Foldit players. EyeWire also relies heavily on whales [2]. Kim et al. stated that more than 100K registered non-expert players from more than 130 countries have contributed to the experiment, however the 100 most productive players generated almost half of the production.

Session Counts and Session Times: The cumulative distributions of the session time (ST) and session count (SC) metrics for the registered players of the two VeriGames are shown in Figure 3. The registered players of VeriGame A spent 1236 hours in total, and the average is one hour per player. We observe that the order of the players by their session time (ST) is identical to the order of their productivity for the top ten players except one. For VeriGame B, the registered players spent 558 hours in total, and the average is again close to one hour per player. Eight of the ten most productive players are also in the top 20 in terms of session time. In addition, for both games, each of the top 20 most productive players played more than ten hours. In other words, the ratios of STs between whales and average players are about 10 to 1, much higher than the 3 to 1 ratio previously reported for social mobile games [15].

Figure 4 shows both the productivity and ST percentiles in one WEG. The ST curves of VeriGame A and B have similar slopes to those of the productivity curves, indicating that the whales of these games tend to spend more time playing than others. Furthermore, unlike the CDF plots, the WEG exposes a drastic difference between the two games. For VeriGame A, the ST curve is below the productivity curve, meaning that the whales for this game are more productive per unit of time than



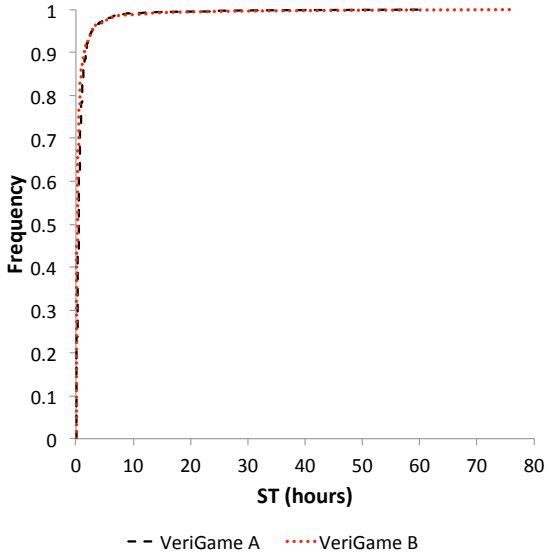


Fig. 3. CDFs of SC and ST for registered players of the two VeriGames

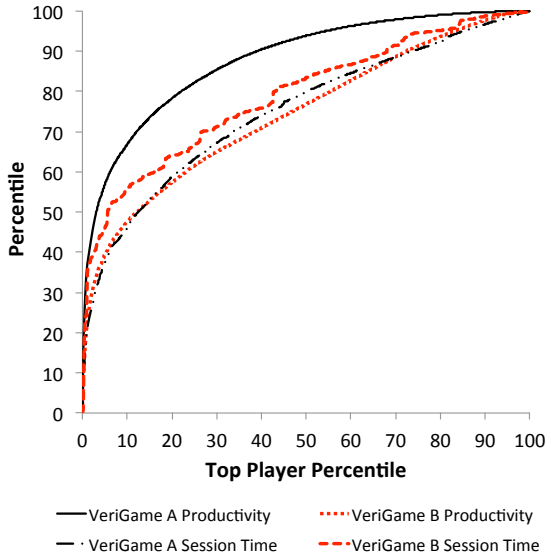


Fig. 4. Productivity and ST percentiles of productive registered players

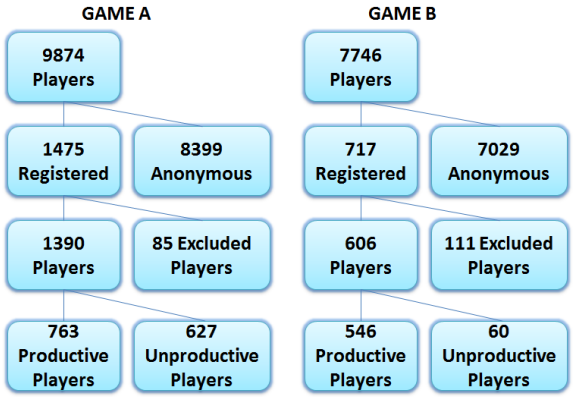


Fig. 5. Player profiles of VeriGames A and B

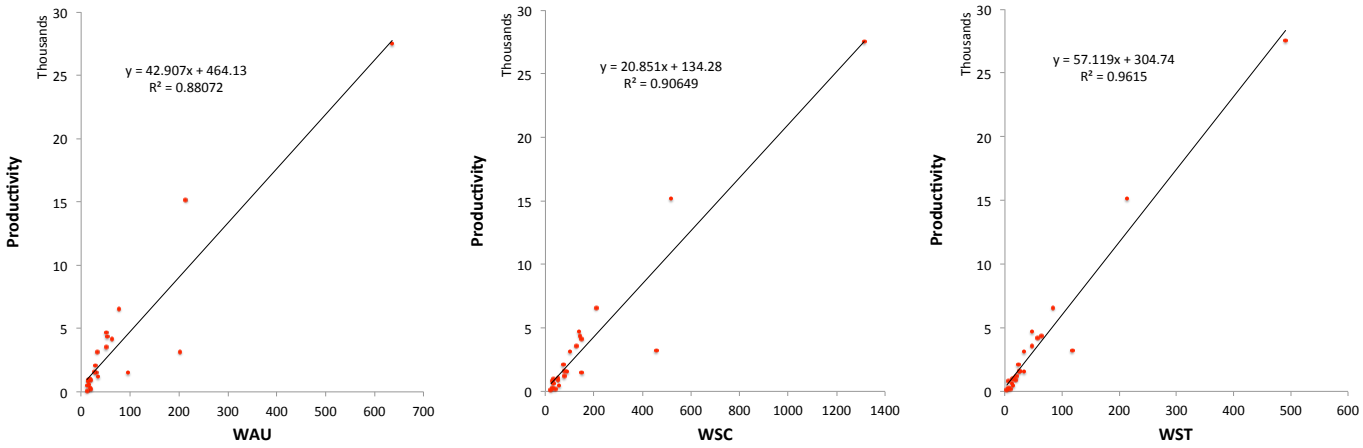
an average player. This is an expected outcome as a player's game skills should improve with more playing time. However, for VeriGame B, the situation is opposite: the ST curve is *above* the productivity curve, meaning that a player produces less per unit of time when spending more time with the game. This indicates a potential deficiency in VeriGame B's scoring system or game design.

V. Discussion

We perform additional analysis using the more detailed VeriGames datasets, seeking to further explain some of the results presented in the previous sections.

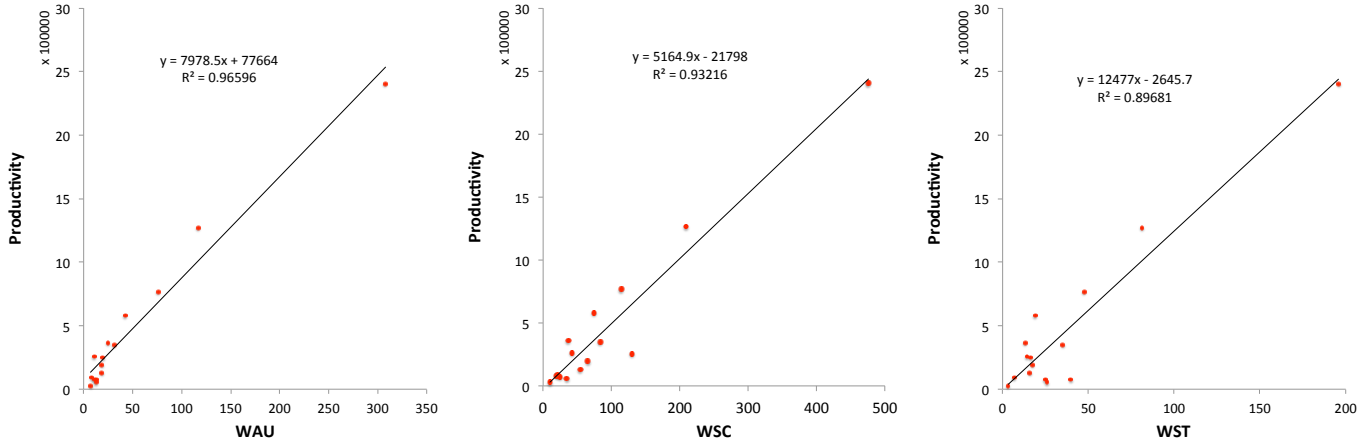
First, we analyze the player attrition pattern going through the registration and tutorial phases of each game, for comparison with a prior study of Duolingo player attrition [6]. The results are presented in Figure 5. The patterns are very similar in both VeriGames. Most players did not maintain their interest after initially trying out the games. Only 10-15% of the players completed the registration process. After filtering out erroneous registrations, game development team members, and unproductive players (who completed the tutorials, but did not complete any game levels), one can conclude that fewer than $\sim$8% of the total players recorded in our VeriGame datasets are productive players. Given such a low fraction of productive players to start with, the long-tail whale effect graphs presented in the last section are easily understood.

Second, we perform a linear regression analysis to determine the best aggregate game play metric for predicting the total productivity over a period of time. We evaluate three such game play metrics: total active users, total session counts, and total session time. Each of the three metrics was evaluated over two different time intervals: per day and per week. The



(a) VeriGame A WAU fitted line plot (b) VeriGame A WSC fitted line plot (c) VeriGame A WST fitted line plot

Fig. 6. Fitted line plots of VeriGame A game play metrics



(a) VeriGame B WAU fitted line plot (b) VeriGame B WSC fitted line plot (c) VeriGame B WST fitted line plot

Fig. 7. Fitted line plots of VeriGame B game play metrics

TABLE VII. R^2 AND p VALUES FOR REGRESSION ANALYSIS

	VeriGame A	**VeriGame B**
R^2	0.99	0.97
	p **values**	
WAU	0.077	0.004
WSC	0.047	0.540
WST	2.217E-10	0.173

results are similar for the two time intervals, therefore we show plots only for the weekly statistics. Our weekly fitted regression line plots are shown in Figures 6a– 7c. We observe that all three metrics are good indicators for game productivity. However, upon inspection of the p-values obtained (Table VII) when all three metrics are jointly considered in a multiple linear analysis, we conclude that the total session time is best for predicting the productivity of VeriGame A while the total active users is best for VeriGame B. This result is consistent with the observation we made about Figure 4 in the end of last Section.

VI. CONCLUSION

From the data available to us, it appears that CSSGs have lower engagement rates than traditional games. Low engagement rate can be a significant obstacle in the path of CSSGs making a significant impact and accomplishing their ultimate purpose. CSSGs in general have not wielded a level of intrinsic attraction sufficient to attract and retain high numbers of long-term players. Given that situation, if the existing players only play the games occasionally, CSSGs face a serious productivity problem. Both VeriGames and other CSSGs examined in this paper have a high proportion of non-returning players and relatively low ER. There may be several reasons for that such as CSSGs' purpose-driven game mechanisms which do not directly target players' personal entertainment, and relatively low game-development budgets.

All of this leads us to focus on the contribution that whales make to the productivity of CSSGs. CSSGs benefit from whales as do commercial games. Vulnerability caused by low ER and non-returning players can be partially mitigated by focusing on attracting new whales to CSSGs who are

ideologically supportive of the games' underlying purpose. While the specific threshold for differentiating whales from other players varies from game to game, and will likely always do so, the Whale Effect Graph allows us to quickly evaluate the extent to which a particular game relies on whales' productivity, as well as qualitatively comparing their impact across multiple games. Unfortunately we do not have sufficient data from traditional games to create WEGs for them, which would allow us to state conclusively whether whales are more significant to CSSGs than to traditional games. This is an area for future research.

Acknowledgments: This research is sponsored by DARPA, under the Crowd Sourced Formal Verification (CSFV) program. We thank the developers of the two VeriGames for providing us the game data used in this paper. We also thank Mehmet Yilmaz for contributing to the initial discussion of this research.

Disclaimer: The views expressed in this document are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

REFERENCES

[1] Duolingo web portal. [Online]. Available: http://www.duolingo.com

[2] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi *et al.*, "Space-time wiring specificity supports direction selectivity in the retina," *Nature*, vol. 509, no. 7500, pp. 331–336, 2014.

[3] DARPA, "Crowd sourced formal verification (csfv) program." [Online]. Available: http://www.darpa.mil/Our_Work/I2O/Programs/Crowd_Sourced_Formal_Verification_(CSFV).aspx

[4] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, and J. Waldispühl, "Phylo: A citizen science approach for improving multiple sequence alignment," *PloS ONE*, vol. 7, no. 3, p. e31362, 2012.

[5] S. Mavandadi, S. Feng, F. Yu, S. Dimitrov, R. Yu, and A. Ozcan, "Biogames: A platform for crowd-sourced biomedical image analysis and telediagnosis," *GAMES FOR HEALTH: Research, Development, and Clinical Applications*, vol. 1, no. 5, pp. 373–376, 2012.

[6] R. Vesselinov and J. Grego, "Duolingo effectiveness study: Final report," Dec. 2012. [Online]. Available: http://static.duolingo.com/s3/DuolingoReport_Final.pdf

[7] N. Lovell, "Dau mau engagement." [Online]. Available: http://www.gamesbrief.com/2011/10/daumau-engagement/

[8] ——, "Daily active users." [Online]. Available: http://www.gamesbrief.com/2011/10/daily-active-users-daus/

[9] ——, "Monthly active users." [Online]. Available: http://www.gamesbrief.com/2011/10/monthly-active-users/

[10] Foldit Web Portal. [Online]. Available: http://fold.it/portal/players/s_ever

[11] M. Rose, "Chasing the whale: Examining the ethics of free-to-play games," Jul. 2013. [Online]. Available: http://www.gamasutra.com/view/feature/195806/

[12] T. V. Fields, "Game industry metrics terminology and analytics case study," in *Game Analytics*. Springer, 2013, pp. 53–71.

[13] M. Carpenter, "How to measure social games success," Jan. 2011. [Online]. Available: http://www.gamesindustry.biz/articles/2011-01-24-how-popcap-measures-social-games-success-editorial

[14] A. W. Hafner, "Pareto's principle: The 80-20 rule," Mar. 2001. [Online]. Available: http://www.bsu.edu/libraries/ahafner/awh-th-math-pareto.html

[15] E. Kain, "Mobile gaming's whales overwhelmingly male, spend big on all types of video games," Aug. 2013. [Online]. Available: http://www.forbes.com/sites/erikkain/archive/2013/08/

[16] S. Carmichael, "What it means to be a whale and why social gamers are just gamers," Mar. 2013. [Online]. Available: http://venturebeat.com/2013/03/14/whales-and-why-social-gamers-are-just-gamers

[17] D. Takahashi, "Only 0.15 percent of mobile gamers account for 50 percent of all in-game revenue (exclusive)," Feb. 2014. [Online]. Available: http://venturebeat.com/2014/02/26/

[18] A. Drachen, M. S. El-Nasr, and A. Canossa, "Game analytics–the basics," in *Game Analytics*. Springer, 2013, pp. 13–40.