GuessTheKarma: A Game to Assess Social Rating Systems

MARIA GLENSKI, University of Notre Dame, United States

GREG STODDARD, University of Chicago Crime Lab, United States

PAUL RESNICK, University of Michigan, United States

TIM WENINGER, University of Notre Dame, United States

Popularity systems, like Twitter retweets, Reddit upvotes, and Pinterest pins have the potential to guide people toward posts that others liked. That, however, creates a feedback loop that reduces their informativeness: items marked as more popular get more attention, so that additional upvotes and retweets may simply reflect the increased attention and not independent information about the fraction of people that like the items. How much information remains? For example, how confident can we be that more people prefer item A to item B if item A had hundreds of upvotes on Reddit and item B had only a few? We investigate using an Internet game called GuessTheKarma that collects independent preference judgments ($N=20,674$) for 400 pairs of images, approximately 50 per pair. Unlike the rating systems that dominate social media services, GuessTheKarma is devoid of social and ranking effects that influence ratings. Overall, Reddit scores were not very good predictors of the true population preferences for items as measured by GuessTheKarma: the image with higher score was preferred by a majority of independent raters only 68% of the time. However, when one image had a low score and the other was one of the highest scoring in its subreddit, the higher scoring image was preferred nearly 90% of the time by the majority of independent raters. Similarly, Imgur view counts for the images were poor predictors except when there were orders of magnitude differences between the pairs. We conclude that popularity systems marked by feedback loops may convey a strong signal about population preferences, but only when comparing items that received vastly different popularity scores.

59

CCS Concepts: • **Information systems** → **Social recommendation**; • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; *Empirical studies in collaborative and social computing*;

ACM Reference Format:

Maria Glenski, Greg Stoddard, Paul Resnick, and Tim Weninger. 2018. GuessTheKarma: A Game to Assess Social Rating Systems. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 59 (November 2018), 15 pages. <https://doi.org/10.1145/3274328>

People increasingly rely on social media as a major source of news, information, and entertainment. Although social media platforms differ in the way that information is curated and delivered, user ratings (e.g., likes, upvotes, pins) play a significant role in determining what is considered to be popular or trending. Tasked with curating an ever-increasing amount of content, providers leverage the collective ratings of the crowd, and measures of collective attention, to identify which content to show users. Unfortunately, peer recommendation can result in irrational herding, where upvotes

Authors' addresses: Maria Glenski, University of Notre Dame, Notre Dame, United States, mglenski@nd.edu; Greg Stoddard, University of Chicago Crime Lab, New York City, United States, stoddardg@gmail.com; Paul Resnick, University of Michigan, Ann Arbor, United States, presnick@nd.edu; Tim Weninger, University of Notre Dame, Notre Dame, United States, tweninge@nd.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery. 2573-0142/2018/11-ART59 \$15.00
<https://doi.org/10.1145/3274328>

beget more upvotes and downvotes may cause comments or posts to go unnoticed by the crowd [16] or where identical content receives widely different attention [8, 15].

Social recommendation systems are built on the assumption that collective opinions result in high quality judgments, even better than the judgments of experts [28]. However, the so-called wisdom of the crowd tends to work best when ratings are performed independently [14]. When ratings are not performed independently, as in, for example, social media systems, social influence dynamics such as *herding* [11, 21, 24], *information cascades* [6], and the *ranking bias effect* [17] significantly bias individual judgment [18, 30]. Likewise, the way an item is described has a tremendous impact on its popularity within a community [15, 29]. Because individuals' perceptions of quality follow the behavior of a group [13], the content made popular by online rating systems may be only weakly correlated with its intrinsic quality [2, 3, 29].

Other results on modern social rating systems such as Facebook, Reddit, and Imgur [27] as well as in online education systems [4] argue that the popularity of a post is a relatively strong reflection of its intrinsic quality. Figueiredo et al, in a study somewhat analogous to ours, had mixed results [7]. They asked 6-10 MTurk workers to pick which of two YouTube videos they preferred. YouTube view counts were predictive of reported preferences when there was a clear consensus among the Turkers. However, a large discrepancy in YouTube view counts for a pair of videos was not predictive of whether there would be consensus: in other words, high popularity was not a reliable indicator of consensus preference. So what do we make of these contradictory reports? Do social influence dynamics fatally distort the relationship between measured popularity and true preferences in social media?

It is easy to tell when something is popular: we simply count the number of views, votes, mentions, etc. On the other hand, it is difficult (perhaps impossible) to determine the objective quality of some content – perhaps no such thing exists [1]. Our goal is to remove the social influence dynamics as well as the algorithmic and design biases that exist in most social rating systems and solicit independent judgments in the absence of such factors – *i.e.*, true preferences. We examine the relationship between path-independent judgments of user preference and the respective path-dependent popularity on live social rating systems. Similar outcomes would be evidence that social rating systems promote the content that users collectively prefer. Dissimilar outcomes would be evidence to the contrary.

Methods to solicit ratings from social media users vary significantly. For instance, online purchases and ride-sharing interactions are often measured on a scale from 0 to 5-stars; however, comparing judgments across items and users can be problematic because each person has a different idea of what constitutes a 5-star rating [19]. To relieve this so-called calibration problem, many social surveys ask users to pick the better of a pair of items or the 'best' out of a handful of options.

People who care more about an item, positively or negatively, will be more drawn to rating it. This creates a selection bias. At the extreme of selection bias, popularity estimates can be manipulated by actively recruiting participants to rate particular items. Selection bias can be reduced by having the system assign things for people to rate rather than letting people choose for themselves [25].

While assigning items to raters reduces opportunities for deliberate manipulation or inadvertent selection bias, it may reduce motivation for people to participate. It also does nothing to provide incentives for effort in evaluating items or honest reporting of answers. Games have been developed that randomly pair two users and award points if they both pick the same object as the 'best' [12, 31]. This motivates participation because of the challenge and collaboration involved and encourages users to take the task seriously because they are only awarded points if they choose accurately.

Unfortunately, simply rewarding users for matching responses of others creates an incentive to report "focal points" that may not be very informative; if you suspect that others are unlikely to notice a detail about an item, you have an incentive not to base your rating on that detail. Peer

prediction techniques respond to this problem by converting people's ratings into predictions about the distribution of ratings that peers will provide and awarding points based on the match of that implied distribution to actual peer ratings rather than based on simply matching the reports [20] or by rewarding matches on individual ratings but penalizing for blind agreement across items [5].

Another approach to encourage honest reporting is the Bayesian Truth Serum (BTS): each user is asked to report their own rating and to predict the ratings of others concurrently; a scoring rule awards points in a way that rewards honest reporting of both values [22]. These surveys have been shown to return high-quality ratings [26]. A study comparing product ratings collected traditionally versus using a BTS approach found that the BTS-ratings converged to the traditional ratings but did so faster and with far less variance than the traditional survey [9].

In the present work we introduce and present the results of an Internet game called *Guess-TheKarma* that asked users to select, from a pair of images, which they personally prefer and to guess which image had received more votes when previously posted in an image-based subreddit on Reddit. The personal preference reports provide a proxy for the groundtruth popularity, the aggregate preference of all potential viewers. To prevent selection effects, the system chose pairs to present rather than allowing users to choose. We did not use a Peer Prediction or BTS scoring system to incentivize honest reporting because we lacked the mechanism to solicit peer ratings. The primary role of asking users to predict which image scored best on Reddit was to make participation fun; once they were participating, players had no reason not to honestly report which images they preferred.

We use the groundtruth determined by the *GuessTheKarma* player's reported preferences to assess the informativeness of other available popularity measures based on self-selected raters and user attention. One such popularity measure is the Reddit vote-score, essentially the number of upvotes minus the number of downvotes an item received from Reddit users. The score of a post on Reddit can be influenced by many factors other than the population's true preference for the item. A post's submission time, the number of early votes, its relative position, the subreddit to which it is submitted, and other factors are all important in determining a post's fate. *GuessTheKarma* eliminates these social, visual, and algorithmic biases and instead asks human judges to focus solely on comparing the content of the paired images.

A second, more passive, measure of popularity is viewership. Many images posted to Reddit are hosted on Imgur, which reports viewership statistics of each hosted image. We can employ Imgur viewership statistics as another popularity predictor. Because of its algorithmic and presentation design choices, Imgur is likely to have different types of socio-technical biases, which can provide a different perspective on the same content.

The goal of this work is to assess the ability of social media popularity metrics to highlight informative or interesting content. This assessment is performed by using the results of the *Guess-TheKarma* game as groundtruth for content preference, and treating the social media outcomes, *i.e.*, Imgur views and Reddit scores, as predictors which come from a live, influence-rich environment. Prediction accuracy can then be measured against the *GuessTheKarma* groundtruth and used as evidence of the ability of social media platforms to identify preferred content.

Our results show that Reddit scores and Imgur views are surprisingly poor predictors of user preference. Each image pair was evaluated by about 50 *GuessTheKarma* players. The image with a higher Reddit score or more Imgur views was preferred by the majority of the players around two-thirds of the time. Conversely, *GuessTheKarma* players were able to predict which image was preferred only about 60% of the time, and self-reported Reddit-powerusers were no better than non-users. Additional analyses investigated the effect of player agreement and subreddit size. From this analysis we were able to distill circumstances that resulted in good predictive performance.

Fig. 1. Example of a single round of GuessTheKarma with main design aspects highlighted wherein the change in survey question after a selection is made is illustrated after “– Click –”.

In summary, our contributions are two-fold:

- (1) GuessTheKarma, a game-like survey tool that can be used to explore the performance of social rating systems, and
- (2) Analysis of whether Reddit vote-scores and Imgur view-counts are predictors of the true majority opinion, when that opinion is gathered without social influence.

1 METHODOLOGY

1.1 GuessTheKarma Design

GuessTheKarma is a single-player online game that asks players which of a pair of image-posts they prefer, *i.e.*, find more interesting or informative. The game is straightforward: immediately upon entering the Web page, the player is randomly assigned one of eight subreddits (*r/aww*, *r/pics*, *r/funny*, *r/OldSchoolCool*, */r/photocritique*, */r/CrappyDesign*, */r/itookapicture*, or */r/EarthPorn*) from which images will be selected. For example, if the subreddit */r/funny* is selected, then all images in the game would be drawn from image-posts submitted to */r/funny*. Then, without further instruction, each player is presented with two image-posts and the survey question(s). Figure 1 illustrates a single round of the game, wherein the survey-text changes when a player makes a choice, denoted by – Click –. These eight subreddits were chosen because they primarily host images and because their subscriber count, *i.e.*, viewership, varies from very large to very small. Previous work has found that post titles have a dramatic effect on the final popularity of Reddit images [15]. The post title is integral to the content of the post, so we include it above the image.

The player may choose a different subreddit manually at anytime but, if that occurs, the game will restart using the new subreddit.

Upon making their selection(s), the final scores of both posts are displayed (*i.e.*, the number of upvotes minus the number of downvotes received), the area around the decision-buttons turns green, and the button-text changes to “Correct!” if the player guesses correctly. If the player guesses incorrectly, the area around the decision-buttons turns red and the button-text changes to “Wrong!”. Figure 1 illustrates the result of an incorrect choice.

After displaying the score totals for 3 seconds, the game automatically progresses to the next pair of images. The counter at the top of the page increments with every selection until 10 image-pairs have been judged by the player. At any time the user may manually select a different subreddit, but this restarts the game. Partially completed games are not recorded. After the final image-pair, but before we showed a player their accuracy, we asked each player to answer a short usage questionnaire shown in Tab. 2

There is an important distinction between asking “what do you prefer?” and asking “what do you think others prefer?” Hacker and Von Ahn compare this distinction to the problem of election polling [12], where a similar distinction exists in the questions: “who will you vote for?” versus “who do you think will win?” respectively. It is unclear which question better elicits information about an image post but a study on political polling showed that asking voters who they think will win, rather than who they will vote for, is a better predictor of the actual election winner [23]. Nevertheless, the goal of this work is to assess how well social media systems aggregate ratings from social media users, so we use the preference question to formulate a groundtruth of user preferences. In our study, the prediction question served mainly to motivate participation: it was challenging and fun to try to predict which image got more votes on Reddit.

Despite the wording of the prediction-judgment survey question, which was chosen based on feedback from beta-testers of the game, we judge correctness based on the *score* of the post not the number of upvotes. Unfortunately, it is impossible to know the actual number of upvotes on any Reddit post through Reddit’s API. We chose to use the word “upvote” instead of “score” because 1) asking for the number of upvotes is more clear – many game players do not understand the details of how the score of a post is calculated, and 2) there is little distinction anyways – the post with the most upvotes will normally also have the higher score.

We collected image-posts for each subreddit from the Reddit BigQuery dataset¹, which contained posts submitted between January 1, 2008 and August 31, 2015. Rather than randomly pairing images, we separated image posts into very high-scoring (VH), high-scoring (H), medium-scoring (M), and low-scoring (L) bins, which correspond to posts with a final score > 95%, between 75-95%, 50-75%, and < 50% of all posts within their respective subreddits. Then, when we created the sets of image-pairs to display to players, we randomly sampled image-pairs from 6 bin-pair-permutations to create games using only VH-VH, H-H, VH-H, H-M, H-L, and VH-L pairings. Images were placed on the left or right randomly. Reposts of the same image and post-title are possible resulting in different scores for the same data. In these cases, we select the highest scoring post.

Of the 4^2 possible combinations, we purposely limited the image pairings to those that contained at least one high- or very high-scoring image. In beta-tests, players complained that L-L, M-L, etc. pairings were too difficult and unfair. These complaints were reasonable; an image-pair with a score of 5 versus 4 is essentially a toss up. Furthermore, the goal of the game is to assess whether those images which became popular on social media match a path-independent groundtruth based on GuessTheKarma’s survey of user preference. We used 400 total combinations of 325 image-posts and added safeguards to the survey so that a player does not see the same image-pair twice. Limiting the

¹ Available at https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts

Table 1. Guess The Karma Dataset.

Subreddit	# Judgements	# Image Pairs	# Images
/r/funny	3,090	50	42
/r/OldSchoolCool	3,460	50	42
/r/aww	3,334	50	41
/r/pics	3,280	50	42
/r/photocritique	1,130	50	42
/r/CrappyDesign	1,726	50	35
/r/itookapicture	3,170	50	42
/r/EarthPorn	1,484	50	39
Total	20,674	400	325

Table 2. Summary of the responses to the Reddit usage questionnaire.

Question	Distribution of Responses		
Describe your level of Reddit use	Heavy 1294 (59.9%)	Casual 724 (33.5%)	Don't use Reddit 144 (6.7%)
How long have you used Reddit?	Over a year 1815 (84.0%)	0-12 months 202 (9.3%)	Don't use Reddit 145 (6.7%)
Do you pay attention to r/x?	Yes 606 (28.0%)	No 1227 (56.8%)	Don't use Reddit 329 (15.2%)
Do you vote on posts?	Yes 1371 (63.4%)	No 643 (29.7%)	Don't use Reddit 148 (6.8%)
Do you vote on posts in reddit.com/r/new?	Yes 279 (12.9%)	No 1674 (77.4%)	Don't use Reddit 209 (9.7%)

number of image-posts resulted in each pair receiving judgments from multiple players. Multiple judgments on the same image-pair permit an assessment using a crowd-sourced majority-opinion – our groundtruth.

The GuessTheKarma methodology was approved by the University of Notre Dame's Institutional Review Board (#17-06-3941).

1.2 Data Collection

On February 29, 2017, we opened the Web site and recruited players. We recruited game players via posts to Reddit, Digg, Twitter, and various other social networks. It is difficult to determine precisely, but we estimate that the majority of the players arrived via Reddit. By March 7, 2017, 2,660 people had played the game and provided 20,674 preference judgments. We only recorded votes after the player made judgments for all 10 pairs. Table 1 shows the distribution of judgments and distinct images used across subreddits. Of the 2,660 total players, 2,083 (78.3%) completed the usage questionnaire at the end of the survey. Their responses are displayed in Table 2.

Sampling bias is a fundamental limitation of many Web-based studies; GuessTheKarma is no exception. We attempted to limit this issue by directly recruiting participants from Reddit (rather than Mechanical Turk) but our particular sample of users may not be completely representative of all social media users or all persons. For example, many of our participants came from a post submitted to the /r/webGames subreddit, but it is unclear how the demographics of the /r/webGames subreddit reflect the general demographics on Reddit or the demographics of the users of the image-subreddits used to populate the games.

1.3 Analysis

In this section, we outline how we compare accuracy of real-world, socially-aggregated outcomes (Reddit vote-scores and Imgur view-counts) against GuessTheKarma preferences. The GuessTheKarma outcome of an image-pair is measured by the majority preference of GuessTheKarma players. As discussed earlier, we view the majority preference as a kind of unbiased groundtruth to which other outcomes are compared. Before we examine accuracy against the groundtruth measure, we first examine the agreement among the GuessTheKarma players.

We use Fleiss' Kappa (κ) to measure the agreement of judgments for an image-pair where 1 is complete agreement and 0 is agreement indistinguishable from random. Negative κ scores, therefore, can be interpreted as being worse than random agreement.

Next, we examine the overall accuracy of our two predictors. First, we want to know: how accurately do social aggregators like Reddit and Imgur predict the majority opinions of “the crowd”? These predictors rely on judgments collected from an environment with social influence effects, as well as rank, design, and other algorithmic-biases; poor prediction accuracy may therefore be attributed to the presence of these biases. We compare the accuracy and 95% confidence intervals for each predictor and perform tests of statistical significance. We also look for correlations between the predictor accuracy and groundtruth agreement, the choice of subreddit, the Reddit-score-percentile differences within image-pairs, and player expertise.

Agreement Effect Accuracy may be affected by the level of agreement for the pair of image-posts. We hypothesize that image-pairs with complete agreement ($\kappa = 1$) would be *easier* to predict than those with low ($\kappa \approx 0$) agreement. To answer this question, we plot the accuracy of each image-pair as a function of their agreement. We then perform statistical tests to analyze what correlation, if any, exists.

Subreddit Effect The choice of subreddit may also affect the accuracy of popularity predictions. Across the eight subreddits, the size of the community, *i.e.*, the number of subscribers to the subreddit, varies widely; for example, as of February 23, 2017, /r/pics had 18.4 million subscribers while /r/photocritique had only 36 thousand subscribers. The /r/photocritique, /r/CrappyDesign, /r/itookapicture, and /r/EarthPorn subreddits also have a much narrower focus for the images submitted than r/aww, r/pics, r/funny, or r/OldSchoolCool. We compare the accuracy of platform predictors for each subreddit.

Image-pair Score Balance Recall that images were labeled as being VH, H, M, and L scoring based on their subreddit-conditioned score percentile. Images were purposefully paired according to their labels. We call image-pairs comprised of similar scoring posts balanced, and image-pairs with dissimilar scores unbalanced. We expect that it is easier to predict the correct image from an unbalanced image-pairing (*i.e.*, where one image received far more attention than the other), than from a balanced image-pair (where the two images receive about the same amount of attention). We test this hypothesis by plotting platform accuracy with respect to the score difference of the image-pair.

Data, the GuessTheKarma source code, and the complete statistical analysis scripts are available online at <https://github.com/nddsg/GTK-paper>.

2 RESULTS

First, we examine the overall accuracy of our platform predictors against the majority preference of the GuessTheKarma players. Accuracy is measured by comparing the predictor choice, *i.e.*, which post received the higher Reddit-score or more Imgur-views, against the GuessTheKarma majority preference.

Fig. 2. Accuracy of predictor as a function of agreement (κ), with logistic regression lines plotted. R^2 and p -values are listed below each title. Results are 0 or 1 but jitter is added to the y-axis for a more comprehensive illustration.

Fig. 3. Accuracy for each subreddit as a function of the number of subscribers. Linear regression lines are plotted with R^2 and p -values below the titles. Accuracy is negatively correlated with subreddit size ($p < 0.05$).

The score of Reddit posts and the number of views on Imgur was highly correlated ($R^2 = 0.80$, $p < 0.001$); of the 400 image-pairs in our data set, the image with the higher Imgur-popularity matched the post with the higher Reddit score 86.3% of the time. Thus, the informativeness of both measures was similar.

We report 95% confidence intervals around the accuracy estimates. Reddit vote-scores and Imgur view-counts had accuracies of $68.0\% \pm 4.6\%$ and $64.7\% \pm 4.7\%$, respectively.

Agreement Effect Here we analyze the effect of agreement on the accuracy of each predictor. We measured agreement among GuessTheKarma players by the Fleiss κ score. We plot the accuracy as a function of the agreement in Fig. 2 along with a logistic regression line and its associated coefficient of determination and p -value. We find a weak, but statistically significant correlation; accuracy of the platform predictor as indicator of majority vote of the GuessTheKarma players increases from about 60% in the low agreement range to about 75% in the high agreement range.

Subreddit Effect Next, we examine whether the accuracy of our predictors is affected by the choice of subreddit. Figure 3 shows that accuracy did vary by subreddit. It appears to be negatively correlated with size of the subreddits. The effect is statistically significant ($p < 0.05$) but with so few subreddits we should be wary of reading too much into this correlation.

Image-pair Score Balance Here we compare platform accuracy with respect to the image-pair's score balance. We expect that if an image-pair is severely unbalanced (e.g., VH-L), then it is more likely that the groundtruth preference of GuessTheKarma players is aligned in favor of the VH post. Figure 4 shows the accuracy of the platform predictors according to their pairing balance. As expected, image-pairs with the largest imbalance resulted in significantly higher accuracy from the Reddit predictor ($p < 0.05$) than the pairs where both images were very high-scoring.

Our primary concern with the previous image-pair balance analysis is that image scores within labels are not compared. So we instead compare our platform predictor as a function of the difference

Fig. 4. Accuracy for each pairing type ordered left to right by the size of score differentials. The distribution of image-pairs within each pairing type are plotted on bottom.

Fig. 5. Correctness of prediction for each image-pair plotted by the Reddit score-percentile difference (Δ) with logistic regression lines plotted. Accuracy is weakly, positively correlated with Reddit score-percentile difference. R^2 and p -values are listed below each title. Results are 0 or 1 but jitter is added to the y-axis for a more comprehensive illustration.

between the images' score-percentiles. For example, given a pair composed of images A and B, where A has a higher score than B; if A has a score that is higher than 95% of the other images in the subreddit (i.e., a score-percentile of 95%), and B has a score-percentile of 60%, then the score-percentile Δ is $95\% - 60\% = 35\%$. The score-percentile Δ is therefore a more fine-grained measure of the balance of an image-pair.

Figure 5 shows that the score-percentile Δ of an image pair is significantly ($p < 0.05$) correlated with accuracy for both predictors. This indicates, as expected, that image-pairs that received similar amounts of attention are more difficult to predict than unbalanced image-pairs. Further, we see that once the balance ratio reaches a threshold of 60%, the platform predictors were able to accurately predict the images that the majority of our respondents preferred.

The image-pair balance results show that Reddit and to a lesser extent Imgur are able to accurately make binary predictions of user preference, but only when the image-pairs are severely unbalanced. It is important to note that most of the image-pairs are concentrated within the lower range of score-percentile Δ , i.e., when the score-percentile Δ is less than 20%. Overall, the predictive ability of these social platforms seems to be limited to only a few rare instances.

2.1 Individual Player Ratings

The previous analysis and results determined how accurately Reddit vote-scores and Imgur view-counts predict the majority preferences of GuessTheKarma players. The majority preferences were

used as the groundtruth because they were collected without social influence and algorithmic-bias effects that are often found in social media systems. Next we analyzed how well individual players predict the platform outcome. The results of this task were only slightly better than random guess: personal preference accuracy was 54.0% with a 95% confidence interval of $\pm 0.8\%$ and the player prediction accuracy was $60.6\% \pm 0.6\%$. We again emphasize that the player accuracy is measured differently than in the previous groundtruth analysis. Here, we measure the player accuracy as their individual ability to predict the winning image from Reddit scores or Imgur view-counts.

Expertise Effect Do those who use Reddit frequently perform better than those who use Reddit casually or not at all? To answer this question we performed a battery of correlation and statistical tests comparing the accuracies of players grouped by their questionnaire responses. We used the one-tailed two valued t -test; for each question, the null hypothesis is that players who self report that they do not use Reddit have the same accuracy as those who answer otherwise. For example, we compare the accuracy distributions of players who self-reported heavy Reddit usage to those that do not use Reddit at all, and those who use Reddit casually to those that do not use Reddit at all, and likewise for the four other questions. We also performed the one-tailed two valued t -test where, for each question, the alternate hypothesis was that players who indicated heavy use had higher accuracy as players that indicated casual use.

There are three responses for each of five questions, resulting in dozens of statistical comparisons. Here we also employed Bonferroni correction to mitigate problems that arise when performing multiple statistical comparisons. No statistically significant correlations were found. In particular no difference was observed between the accuracy of users who were active within a subreddit (even a small subreddit) and non-users with mean accuracies of $62.8\% \pm 4.2$ and $63.5\% \pm 2.6$, respectively.

Next we ask: do self-reported *powerusers*, *i.e.*, those that indicated frequent Reddit use on all survey questions, perform better than non-powerusers? To answer this question we again use the one-tailed two value t -test with Bonferroni correction to correct for the multiple tests. We considered two null hypotheses: 1) that powerusers have the same accuracy as non-powerusers, and 2) that powerusers have the same accuracy as those players who do not use Reddit, *i.e.*, players who answered “Don’t use Reddit” to all questions. We also repeat this analysis for powerusers that answered Yes or No to the subreddit-use survey question.

Of the players who completed the survey, only 174 qualified to be called powerusers. Within this set of powerusers, 60 also reported that they use the subreddit from which their image-pairs were drawn. No statistically significant differences were found. Overall, we see no evidence of an expertise effect, lending confidence to our inclusion of raters who were unfamiliar with the particular subreddits.

Effort Effect Finally, we consider response times (delays between when a player is shown an image-pair and when they made a selection) as a proxy for player effort. We compare the response time distributions for correct and incorrect judgments to determine if incorrect judgments are correlated with player effort as measured by player response time.

Using the two valued two-tailed t -test, we found that the response time distributions were similar for correct and incorrect predictions ($p = 0.419$ with means of 22 and 26 seconds, respectively). Moreover, player response times, organized into one-second-sized bins and shown in Fig. 6 were not significantly correlated with prediction accuracy ($R^2 = 0.18$, $p = 0.572$). Overall we find that there is little effect of player effort on prediction accuracy.

3 DISCUSSION

Several methods of analysis were performed. Altogether, the totality of the results leads us to conclude that, except in specific circumstances where image-pairs are highly unbalanced and

Fig. 6. Response time (seconds) for correct and incorrect player predictions. Distributions are not significantly different ($p = 0.419$).

players are in strong agreement, Reddit vote-scores and Imgur view-counts are relatively poor predictors of user preference.

Our results are consistent with the finding of Salganik, Dodds, and Watts, in their study of artificial cultural markets [24]. They simulated a music marketplace with social influence feedback loops; songs were shown in sorted order based on previous downloads. They ran eight versions of the marketplace, and one with random ordering to avoid social influence and thus measure true population preference. They found that the truly most preferred songs in the independent information condition rarely fell to the bottom and that the truly least preferred songs rarely stayed at the top but beyond that the ordering in any of the marketplaces was unpredictable from the true popularity. This matches our finding that GuessTheKarma players generally preferred a very high scoring image over a paired low scoring image.

Our results are only partially consistent with those of Stoddard [27]. He concluded that vote counts in several subreddits were quite informative for items that got at least a few votes, whereas we find that they are only good at distinguishing the highest scoring from the lowest. Stoddard's method was to fit a Poisson model, then infer an underlying quality attribute of each item as a prediction of its votes. In the absence of ground truth values, however, it is not clear how good the model fit has to be in order to conclude that the social media vote counts are informative about user preferences.

Our results also differ somewhat from those of Figueiredo et al [7]. Their findings for YouTube suggested that a huge difference in views was not a reliable indicator of consensus preference for one video over another. By contrast, we find that a huge difference in Reddit score or Imgur views is predictive of the majority's true preference (Fig. 4, Fig. 5 and Fig. 7). It's not clear why the results differ. While our game attempts to make voluntary participation more attractive, the structure of the task itself was similar between our studies: subjects offered an opinion about which of a pair of items they preferred. One possible explanation of the different results is that there are algorithmic and design differences between YouTube and Reddit, with extreme differences in popularity of scores for images on Reddit being more informative of popular opinion. Another possibility is differences in the items being evaluated: there may be taste variation for the YouTube videos assessed, such that topical preferences outweighed any assessment of production quality. By contrast, even though some people may prefer cat images to dog images, it may be that most people appreciate any really high quality image. This is a clear avenue for additional research. Additionally, Figueiredo et al also found that, contingent on one video being strongly preferred over another, the preferred one almost always had more views. Our results show a similar finding for $\kappa \geq .40$, with the added complication that only 10% of the image-pairs had an agreement larger than 0.40. Even then, near perfect agreement still resulted in incorrect predictions some of the time. These findings are illustrated in Figure 7, which plots the score-percentile Δ as a function of the agreement level with correct and incorrect predictions in green $\circ$ and red $\square$, respectively.

Fig. 7. Reddit score-percentile Δ of image-pairs as a function of groundtruth agreement (κ). Pairs for which Reddit vote-scores correctly predicted the majority opinion are denoted with green $\circ$ with incorrect predictions represented with red $\square$.

The Reddit scores, it seems, provide a very useful signal for the casual users who only look at the top few images in a subreddit. They are reliably better (i.e., preferred by more people) than low scoring images. Differences within the top quartile of scores, however, are not meaningful. This suggests two potential design implications. First, if score differences between adjacent posts are not meaningful, it might be better not to show them at all. Second, it might be valuable to partially randomize the order so as to dampen social influence effects.

We found that image-pairs sourced from smaller subreddits were better predictors of Guess-TheKarma majority vote than image-pairs sourced from larger subreddits. One plausible explanation is that, because of their size and the large volume of posts, large subreddits are more prone to algorithmic and social biases. Larger subreddits rely heavily on Reddit's algorithm to surface the most informative or interesting posts. This may create a ranking bias effect. A post that garnered few votes and little attention within a large subreddit will be crowded out by more popular posts and is not likely to be seen by many users. However, in a small subreddit with little volume or competition, ranking bias will have a far smaller effect. In addition, larger subreddits' posts are more frequently included on Reddit's frontpage, which may create a strong herding effect. Further study is needed to thoroughly understand the relationship between the size of an online community and its ability to solicit user preferences.

One should interpret the findings and design recommendations with some caution. First, while we had many pairs of images with similar Reddit scores, we had few pairs with very different Reddit scores, the VH-L and H-L combinations. In a larger sample of image pairs, the correlation of Reddit score difference with true population preference might not hold up. Second, our ground truth proxy from the GuessTheKarma players may not always reflect the true preference of the majority of readers of the particular subreddits. We tried to recruit Redditors, but many of them came from a games subreddit. We deliberately chose image subreddits where specialized knowledge was not important, but our players' tastes may not have matched readers of the image subreddits. In addition, game players judged images months, sometimes years, after they were originally posted on Reddit. Tastes could have changed in the interim; this seems unlikely for dog and cat pictures, but might be the case in, say, /r/CrappyDesign. Third, it is possible that a user disregards the post image and title and instead perceives quality from the social engagement of the comment section. However, recent reports suggest that a post is rarely up/down-voted after reading the comments section [10].

A critical feature of GuessTheKarma as a method of collecting data is that the game had to be fun in order to attract voluntary participation. We think two things made this game fun. First, the challenge level was just right, neither too easy nor too hard. Indeed, in pilot tests people complained about some pairs that were too hard to judge, and these turned out to be pairs where both were in the bottom half of Reddit scores. Once we removed L-L comparisons, we got fewer complaints. Second, the images were generally interesting or entertaining to look at. A nice side effect of eliminating the L-L comparisons was that at least one of every pair had received many upvotes and thus was probably not terrible. Thus, participating in the game was in part a discovery activity comparable to browsing one of the image subreddits.

This, however, suggests an important limitation of GuessTheKarma as a data collector. It may work only for pairs of items that are a) interesting to interact with in their own right, and b) somewhat but not too challenging to guess the popular score for. If the information signal from a social popularity metric is too low, then trying to guess the metric will be too challenging, and people won't participate. If the information signal from a social popularity metric is too high, then trying to guess the metric will be too easy, and people won't participate. Thus, our finding that the Reddit scores provide only coarse-grained information but not fine-grained information about true popularity may be the only state of the world where we could have gathered data using GuessTheKarma. If, in fact, Reddit scores provided fine-grained information or no information at all, we might not have been able to attract participants to play the game voluntarily.

It is important to note that the predictions collected in the experiment described above are binary predictions; participants either chose image A or image B as their preference. As a result, a player's confidence in their choice is not indicated in their judgment. Future work on this task should explore alternative designs that allow more information to be collected from players. These designs would allow players to include a degree of confidence in their choices, to estimate the difference between image scores, or to "pass" on image-pairs that they are not confident in. Future designs might also be adapted to solicit a measure of quality to allow direct comparisons between quality and popularity. Finally, it may be possible to allow players to win small cash prizes and even wager their earnings to increase their reward.

4 CONCLUSIONS

In this work we presented GuessTheKarma, a survey tool that can be used to explore the performance of social rating systems, and an analysis of two social media rating systems, Reddit votes and Imgur views. Through its tournament style setup, the GuessTheKarma game removed many of the social influence effects and algorithmic, design, and ranking biases, that have been previously reported in social rating systems research. Therefore, discrepancies between GuessTheKarma-judgments and Reddit scores (or Imgur view counts) can be primarily explained by the presence of socio-technical influence dynamics in the the Reddit scores and Imgur view counts.

The results suggest that Reddit vote-scores and Imgur view-counts are relatively poor predictors of user preference except when one image has a much higher score than the other. For example, images with scores in the top 5% of scores within their subreddit paired with an image whose score falls within the bottom half of all scores were preferred by GuessTheKarma players almost 90% of the time. Unsurprisingly, we found that individual players were less accurate predictors, performing slightly better than random. However, we also did not find significant differences between players of varying expertise or familiarity with the platforms or communities we compared. Finally, we found no difference in the accuracy of active members of a subreddit, even small subreddits, compared to those respondents that did not use Reddit.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support by the Defense Advanced Research Projects Agency (DARPA #W911NF-17-C-0094). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

- [1] Ioannis Arapakis, B Barla Cambazoglu, and Mounia Lalmas. 2014. On the Feasibility of Predicting News Popularity at Cold Start. In *Social Informatics*. Springer, 290–299.
- [2] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* (1992), 992–1026.
- [3] Keith Burghardt, Emanuel F Alsina, Michelle Girvan, William M Rand, and Kristina Lerman. 2016. The Myopia of Crowds: A Study of Collective Evaluation on Stack Exchange. *Robert H. Smith School Research Paper No. RHS 2736568* (2016).
- [4] L Elisa Celis, Peter M Krafft, and Nathan Kobe. 2016. Sequential Voting Promotes Collective Discovery in Social Recommendation Systems. *arXiv preprint arXiv:1603.04466* (2016).
- [5] Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 319–330. <https://doi.org/10.1145/2488388.2488417>
- [6] P Alex Dow, Lada A Adamic, and Adrien Friggeri. 2013. The Anatomy of Large Facebook Cascades.. In *ICWSM*.
- [7] Flavio Figueiredo, Jussara M. Almeida, Fabricio Benevenuto, and Krishna P. Gummadi. 2014. Does Content Determine Information Popularity in Social Media?: A Case Study of Youtube Videos' Content and Their Popularity. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 979–982. <https://doi.org/10.1145/2556288.2557285>
- [8] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 803–808.
- [9] Eric Gilbert. 2014. What if we ask a different question?: social inferences create product ratings faster. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2759–2762.
- [10] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206.
- [11] Maria Glenski and Tim Weninger. 2016. Rating Effects on Social News Posts and Comments. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 2 (2016), 16.
- [12] Severin Hacker and Luis Von Ahn. 2009. Matchin: eliciting user preferences with an online game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1207–1216.
- [13] Irving Lester Janis. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes*. Vol. 349. Houghton Mifflin Boston.
- [14] Krishna K Ladha. 1995. Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization* 26, 3 (1995), 353–372.
- [15] Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *ICWSM* 1, 2 (2013), 3.
- [16] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 543–550. <https://doi.org/10.1145/985692.985761>
- [17] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PloS one* 9, 6 (2014), e98914.
- [18] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.
- [19] Judith Masthoff. 2003. Modeling the multiple people that are me. In *International Conference on User Modeling*. Springer, 258–262.
- [20] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359–1373. <https://doi.org/10.1287/mnsc.1050.0379> arXiv:<https://doi.org/10.1287/mnsc.1050.0379>
- [21] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

- [22] Dražen Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695 (2004), 462–466. <https://doi.org/10.1126/science.1102081> arXiv:<http://science.sciencemag.org/content/306/5695/462.full.pdf>
- [23] David M Rothschild and Justin Wolfers. 2011. Forecasting elections: Voter intentions versus expectations. *Available at SSRN 1884644* (2011).
- [24] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.
- [25] Matthew J. Salganik and Karen E. C. Levy. 2015. Wiki Surveys: Open and Quantifiable Social Data Collection. *PLOS ONE* 10, 5 (05 2015), 1–17. <https://doi.org/10.1371/journal.pone.0123483>
- [26] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 275–284.
- [27] Greg Stoddard. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *AAAI Conference on Web and Social Media (ICWSM)*.
- [28] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.
- [29] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of ACL*.
- [30] Arnout van de Rijt, Soong Moon Kang, Michael Restivo, and Akshay Patil. 2014. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* 111, 19 (2014), 6934–6939.
- [31] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

Received April 2018; revised July 2018; accepted September 2018