The Wisdom of Smaller, Smarter Crowds

DANIEL G. GOLDSTEIN, Microsoft Research
R. PRESTON MCAFEE, Google Strategic Technologies
SIDDHARTH SURI, Microsoft Research

The "wisdom of crowds" refers to the phenomenon that aggregated predictions from a large group of people can rival or even beat the accuracy of experts. In domains with substantial stochastic elements, such as stock picking, crowd strategies (e.g. indexing) are difficult to beat. However, in domains in which some crowd members have demonstrably more skill than others, smart sub-crowds could possibly outperform the whole. The central question this work addresses is whether such smart subsets of a crowd can be identified *a priori* in a large-scale prediction contest that has substantial skill and luck components. We study this question with data obtained from fantasy soccer, a game in which millions of people choose professional players from the English Premier League to be on their fantasy soccer teams. The better the professional players do in real life games, the more points fantasy teams earn. Fantasy soccer is ideally suited to this investigation because it comprises millions of individual-level, within-subject predictions, past performance indicators, and the ability to test the effectiveness of arbitrary player-selection strategies. We find that smaller, smarter crowds can be identified in advance and that they beat the wisdom of the larger crowd. We also show that many players would do better by simply imitating the strategy of a player who has done well in the past. Finally, we provide a theoretical model that explains the results we see from our empirical analyses.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

Additional Key Words and Phrases: wisdom of crowds; crowdsourcing

1. INTRODUCTION

The "wisdom of crowds" refers to the phenomenon in which aggregated predictions from a large group of people can be more accurate than most individual judgments and can rival or even beat the accuracy of subject matter experts [Surowiecki 2005]. The seminal work on this this topic is by Galton [1907b], who attended a regional fair at which 800 people each guessed the weight of an ox. Galton observed that the average guess of 1,197 lbs. was just 1 lb. away from the ox's true weight of 1,198 lbs. [Galton 1907a]. Later, Knight [1921] had students estimate the temperature of a classroom. The average estimate was just 0.4 degrees off the correct answer and was closer than 80% of the individual estimates. Treynor [1987] also replicated this phenomenon in an experiment in which when he asked his students to guess the number of jelly beans in a jar. In one experiment, the mean guess of 871 was closer to the actual number of 850 than all but 1 of the 56 guesses. There have been numerous replications of this phenomenon across a variety of different academic fields and across a variety of different problem domains [Surowiecki 2005; Lorge et al. 1958].

The literature gives some insight into the underpinnings of this phenomenon. Each judgment can be modeled as consisting of two components: information and error [Surowiecki 2005; Simmons et al. 2011]. Intuitively, if the judgments are unbiased

Author's addresses: D. G. Goldstein and S. Suri, Microsoft Research, New York City, 641 Avenue of the Americas, 7th Floor, New York, NY 10011, {dgg, suri}@microsoft.com; R. P. McAfee, Google Strategic Technologies, 1600 Amphitheatre Parkway, Mountain View, CA 94043, preston@mcafee.cc

EC'14, June 8–12, 2014, Stanford, CA, USA. ACM 978-1-4503-2565-3/14/06.
http://dx.doi.org/10.1145/2600057.2602886

and independent, the errors (deviations from the ground truth) will largely cancel out through averaging. However, for this to happen, there are a few requirements on the crowd and its judgments. First, members of the crowd should have some information on the judgment in question. Second, members of the crowd should be motivated to give accurate judgments. These two requirements help ensure that there is at least some information in the judgments reported by each of the crowd members. Third, in order for errors to cancel out, the judgments should be somewhat independent. Diversity of experience of judges is thought to prevent "group think" phenomena [Surowiecki 2005]. Lastly, there should be no systematic bias in the judgments of the individuals (for example, each judge being off by a constant amount) as this can severely impact the accuracy of the aggregated estimate [Simmons et al. 2011; Lorenz et al. 2011; Muchnik et al. 2013].

Aggregating over a larger crowd helps ensure that for any individual's error there is another individual with roughly an equal and opposite error. Moreover, if a rare piece of information is relevant to the judgment task, large crowds also increase the probability that some member has that relevant bit of information. This might imply that aggregating over a larger crowd would result in better aggregated judgments. On the other hand, there might be a sub-crowd in which each member has small error. Given a crowd of people C we define a sub-crowd S to simply be a subset of the crowd, $S \subseteq C$. Expert sub-crowds are those with an advantage at the estimation task, such as some prior expertise or familiarity with a certain type of judgment. Aggregating over a smaller, smarter crowd might result in smaller overall error. The central question this paper addresses is: does one get more accurate aggregations if one maximizes the number of judgments aggregated or if one finds a smaller crowd within the larger crowd in which the members have smaller individual errors?

There are two primary motives for pursuing smarter sub-crowds: efficiency and accuracy. If collecting judgments is costly, then alternatives that require fewer judgments are more efficient [Herzog and Hertwig 2009]. And if smarter sub-crowds exist, it may be possible to attain higher accuracy than is possible with conventional wisdom-of-the-crowd aggregation techniques. Efficient, accurate crowdsourcing of judgments should be welcomed in the fields of online polling, prediction, and forecasting.

Our domain of exploration of this question will be the actions of players in a fantasy soccer league from which we collected millions of individual-level player histories. In fantasy soccer, every person who plays this game (henceforth "manager") manages a nominal ("fantasy") team comprising professional soccer players. Week to week performance of the real-life players is tied to the performance of the same players on the fantasy team. The better the players perform in real life games, the more points the fantasy team earns. Thus choosing a real-life player to be on one's fantasy team can be considered a vote that that specific player will do well in the future. Fantasy soccer is ideally suited to this investigation because it comprises millions of individual-level, within-subject predictions, past performance indicators, and the ability to test the effectiveness of arbitrary player selection strategies. The first step in our analysis will be to predict managers' future performance based on their past performance, in order to assess whether there is a skill component to the game. The second step will be to use a variety of different methods to construct crowd strategies based on imitation heuristics or sub-crowds of managers of different predicted skill levels. Our final step will be to evaluate and compare these methods to each other and the whole crowd.

2. RELATED WORK

There is an old literature on finding wiser, small crowds in larger crowds. See Lorge et al. [1958] for a survey. Generally these studies work in similar ways. The experimenter gets students to make estimates on things like the temperature of a room [Knight 1921], the ranking or estimation of weights [Gordon 1924; Bruce 1936], or the number of objects in a bottle [Klugman 1945]. Afterwards the experimenter aggregates different size groups of these estimates and compares them to the best estimate and the aggregate of the entire group. Mannes et al. [2013] recently and independently revisited the same idea. They analyze two data sets of numerical estimates. The first consists of estimates of every day measures such as temperatures, distances, prices, etc. The second consists of forecasts of economic indicators gathered in the *Quarterly Survey of Professional Forecasters*. Generally all of these papers show that there exists a small sub-crowd with estimates that outperform the overwhelming majority of individuals and the entire group.

We build on and extend this literature in a number of ways. First, these authors only consider sub-crowds of sizes roughly at most 50. Since, our data set is at least three orders of magnitude larger than any of the above, we can consider sub-crowds ranging from singletons into the tens of thousands. This enables us to examine the tradeoff between more concentrated expertise and more diverse opinions across a much wider range of sizes.

Second, since the prior work focuses on numerical estimates, it only considers averaging as a method of aggregation. In our domain, a manager selecting a player counts as a vote for that player and we aggregate by choosing players with the most votes. Thus we consider the wisdom of the crowd's strategies as opposed to the wisdom of the crowd's estimates.

A third contribution of our work regards the type of data we analyze. Most of the prior works [Bruce 1936; Gordon 1924; Klugman 1945; Knight 1921; Lorge et al. 1958; Mannes et al. 2013] analyze data sets carefully designed to experimentally test the existence of small, smart crowds. The internet age has made it much easier to find data sets that are logs of some type of human behavior. These data sets are often large but often do not have data that is perfectly suited to the research question. For example, in our case managers have to pick teams subject to budget constraints, constraints on the number and cost of transfers that can be made, constraints on the positions of the players that can be picked, and subject to the constraint that a maximum of 3 players can come from any English Premier League team. (We will describe these constraints in more detail in the next section.) Moreover, since there is no limit on how many managers can own a player, managers have an incentive to look for lesser owned players to differentiate their teams. While we do believe fantasy soccer is a good venue for answering our research question, all of these constraints make this data far from the perfectly designed data constructed in much prior work. As a result, our results have applicability and generalizability beyond carefully designed lab experiments.

Our final contribution is that we will look for smart sub-crowds in a predictive manner. That is, we will predict a fantasy soccer manager's skill level using past performance variables. Then we will see how strategies based on sub-crowds of managers with different predicted performance. We use this approach so to show that our methods will useful for predicting future events as opposed to much easier tasks of "predicting the past" or "predicting the present".

3. FANTASY SOCCER

Next we discuss the English Premier League, the rules of the Fantasy Premier League and why these data are appropriate to study the wisdom of crowds.

The English Premier League consists of 20 professional soccer teams throughout England and Wales. In a season, every team plays every other team once at home and once away. Accordingly, each team plays 38 games. Roughly speaking, games are held on the weekends and all teams play each weekend. Each of these weekends is called a gameweek. The series of 38 gameweeks is the soccer season.

The Fantasy Premier League[1] is a game where anyone can sign up to manage a fantasy soccer team. A fantasy soccer team consists of 15 players: 5 defenders, 5 midfielders, 3 strikers and 2 goalies, all of which are chosen from the professional players in the English Premier League. A fantasy team is restricted to have at most 3 players from the same Premier League team. Figure 9 (Appendix) shows a screenshot of a current fantasy team. At a high level, the better a player does in real life, the more points that player earns in the fantasy game. The object of the game is to earn the largest amount of total points by the end of the season.

In finer detail, strikers gain 4 points for every goal they score, midfielders gain 5 points for every goal they score and defenders and goalies gain 6 points for every goal they score. An assist (pass given to a teammate who then scores) earns a player 3 points. If a defender or goalie earns a “clean sheet” or a “shutout” (their Premier League Team prevents their opponent from scoring) and they play at least 60 minutes they earn 4 points. Each player earns an additional 2 points for playing at least 60 minutes. The league also grants 1–3 bonus points to players who played exceptionally well in a game. These are the most common ways for a player to earn points in the Fantasy Premier League. We do not list all the rules here because the exact mechanics of fantasy soccer are not relevant for our investigation. In general, managers attempt to maximize the sum over all of their players and all of the 38 gameweeks of the total points earned by their players.

Each gameweek, a manager chooses 11 of his/her 15 players to start. The rules say that a manager must start 1 goalkeeper, at least 3 defenders and at least 1 forward. If a starting player plays in the real life game that gameweek, his points are determined by his performance as described above. For each starting player that does not play in the real life game that gameweek, bench players are automatically substituted instead. Every week, each manager chooses a captain of his team. The captain gets double points for his performance that gameweek. Managers are free to change their captain choice every gameweek if they would like to do so. The only restriction on the captain choice is rather obvious: a manager has to own the player to make him the captain. The choice of captain will be the main choice we study in this work.

Players have prices and before the season starts all managers pick teams subject to the same budget constraint. The prices of the players rise and fall based on the players’ popularity among the fantasy managers. In between gameweeks managers are allowed to make transfers, that is, they can sell one player and buy another player, again subject to the budget the fantasy manager has on hand. If a manager wants to transfer more than one player he pays 4 points for each additional transfer.

There are no restrictions on how many managers can own a player. This is crucial for this study and it makes the Fantasy Premier League unlike many other fantasy sports. Since every manager could, in theory, own a certain superstar player, the Fantasy Premier League is not a true prediction market. Thus, every gameweek a manager chooses a player we count as a vote for that player.

3.1. Data

At the time of the writing of this paper the 2013/2014 season is roughly half over, thus the 2012/2013 season is the most current complete season we have data for. There were roughly 2.5 million fantasy managers during the 2012/2013 season. We have for each of these managers how many points their team earned during each gameweek of the 2012/2013 season and how many points they earned in total for each of the previous six seasons they might have participated in. (Premier League Fantasy began in

[1] http://fantasy.premierleague.com/

the 2006/2007 season.) This required scraping 2.5 million URLs. In addition, we have all 15 player choices and captain designations, i.e. the entire team composition, for all 38 gameweeks of the 2012/2013 season of 100,000 managers sampled uniformly at random yielding a total of 57 million judgments. This involved scraping another 3.8 million URLs. All of the data scraped is available publicly. We also only analyzed anonymized team identifiers and did not connect any of our analyses with any personally identifiable information. Finally, we have the actual performance of all 706 English Premier League players allowing us to evaluate the performance of any team we construct.

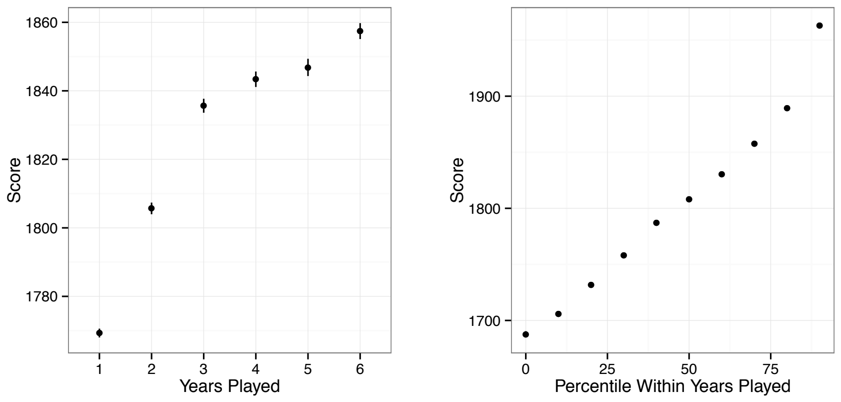
4. LUCK VS. SKILL

Before we begin to look for smart sub-crowds in the larger crowd of fantasy soccer managers, we first address the degree to which success in fantasy soccer can be predicted, or the relative contributions of luck and skill in this domain. The luck/skill distinction is important in this context because if success at fantasy soccer is as random as roulette, then aggregating the player choices of those who did well in the past should be uninformative for choosing players in the future. If, on the other hand, there is a substantial skill component to the game, it may (but not necessarily) be possible to identify managers who are likely to do well, and exploit the wisdom of these smaller, smarter crowds through various strategies. We say it is not necessarily the case because it is not clear that predictably high-scoring managers' decisions can be aggregated intelligently due to the budget and transfer constraints managers face in choosing their players (as described in Section 3).

The luck/skill relationship has received attention in the literature especially as it relates to the legality of gambling [Levitt et al. 2012], and to the question of whether financial portfolio managers actually perform better than chance [Fama and French 2010]. Levitt et al. [2012] put forth properties of a game of pure chance (p. 584), including that "payoffs do not vary systematically with the observable characteristics of players" and "a player's past success (or failure) does not predict his future likelihood of success (or failure)". They state that when games do not have such properties, players' skill causally affects their outcomes. In this section, we show that fantasy soccer seems to have a detectable skill component—based on observables some managers do score predictably higher than others. In particular, past experience (having played the last several years in a row) and past percentile (percentile rank among managers who have completed in the same seasons' contests) are strong predictors of future scores.

Recall that we collected the histories for all of the roughly 2.5 million fantasy managers who played in the 2012/13 season. In Figure 1(a) we plot the average of the total number of points earned by those managers who have played in each of the previous $1, 2, \ldots, 6$ seasons. (The Fantasy Premier League started 6 years prior.) One year's experience in the past three years corresponds to scoring 30 to 35 points better in the future, with diminished returns of 5 to 10 points per year for each season of experience beyond that. Past experience is therefore one observable that predicts future scores, suggesting there is some skill involved in doing well in fantasy soccer.

In Figure 1(b) we plot past percentile: the percentile rank of each managers' past scores in relative to other managers who played in the same exact years. We bin managers by deciles and plot their mean 2012/13 scores. Figure 1(b) shows that each decile increase in the distribution of comparable managers is associated with roughly 25 more points in the 2012/13 season. However, membership in the highest decile is associated with more than 50 additional points in the future, suggesting that managers in the highest percentiles may be using different or more sophisticated strategies than those in the larger population. Were fantasy soccer a game of pure luck, past percentile rank would not correlate with future scores. Furthermore, were fantasy soccer a game



(a)

(b)

Fig. 1. Predictors of 2012/13 score. First, past experience expressed as consecutive years of play prior to 2012/13. Second, past performance, expressed as decile rank among managers who have played the same seasons.

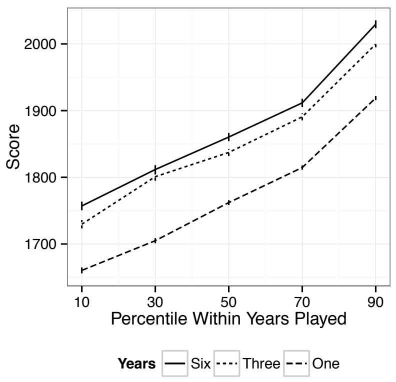


Fig. 2. The interaction of past experience and past performance on 2012/13 score, with a selection of levels shown for clarity.

of skill and luck, but with a uniform level of skill distributed through the population, then we would not expect to see additional marginal score increases for those in the highest decile.

Figure 2 shows the interaction of the above-mentioned past experience and past percentile variables as they related to 2012/13 score. Here we see that as the experience levels increase the curves shift upwards indicating higher scores. Also, as the percentile rank increases across all experience levels the curves slope up and to the right indicating that average scores increase as well.

Taken as a whole these results indicate there is predictive power of past experience and past percentile rank on future performance. Next we combine these features and their interaction in a regression model shown as Model 1 in Table I. This model predicts the 2012/13 score as a function of the number of years played, mean percentile rank over years played, and their interaction. We do not add higher order terms and dummy variables in search of the "best" model but rather stick to simple predictors to

Table I. Models predicting future score based the number of years played and the mean percentile rank across all years played. Model 1 is fit to 6 years of past experience data; it predicts outcomes in 2012/13 based on scores in years 2006/7 through 2011/12 and includes the holdout set of managers. For an out of sample test, Model 2 is trained on past data only (predicting outcomes in 2011/12 based on scores in years 2006/07 through 2010/11) and excludes managers in the holdout set we will later use to test crowd strategies. As reflected in the estimated coefficients, the models are quite similar. Indeed, predicted scores on the holdout set are highly correlated between the two models with Spearman and Pearson correlations of .99.

	Model 1	Model 2
(Intercept)	1598.90^{***}	1652.17^{***}
	(0.61)	(0.83)
Years Played	-6.84^{***}	-15.72^{***}
	(0.27)	(0.36)
Percentile	336.00^{***}	309.79^{***}
	(1.17)	(1.59)
Years Played:Experience	32.12^{***}	31.48^{***}
	(0.47)	(0.63)
R^2	0.23	0.21
Adj. R^2	0.23	0.21
Num. obs.	1309085	772355

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

make for a conservative test. That is, if we identify smart sub-crowds with a simple model of managerial skill, then results at least as good should be possible with more specialized techniques. Note that in the Figures 1(b) and 2 above we plot percentile rank relative to managers who played the same seasons for visual clarity while in Table I we simply model the mean percentile rank from past seasons played (not relative to managers who played the same seasons) because the interaction term accounts for the relationship between the two variables.

Table I shows that past experience, past percentile and their interaction are significant predictors of future scores. Furthermore, the R^2 of this model shows that 23% of the variance in 2012/13 scores can be explained simply by the number of years a manager has played and his or her past percentile. Were fantasy soccer a game of pure luck, no manager attribute would predict future scores and a model based on manager attributes would not explain any of the variance in future scores. These results hold promise for the larger goal of boosting the wisdom of the crowd by tapping into smaller sub-crowds, a topic we turn to next.

5. CAPTAIN CHOICES AND THEIR AGGREGATION

In this section we first describe the judgments of the managers and then how we aggregate them. Recall from Section 3 that each week every manager chooses one player on his team to be the captain. The captain earns double points for that week for that managers team. For example, if a player earns 2 points for playing over 60 minutes and 4 points for scoring a goal, managers who captained that player earned 12 points for their team whereas managers who did not captain that player only earned 6 points for their team.

We view choosing a captain as a revealed preference. That is, when a manager designates a player as a captain for a gameweek he is revealing that he thinks that player will earn more points than the other 14 players on his team. Recall that there are 706 players in the Fantasy Premier League to choose from, but the captain choice only reveals that a manager thinks a given player will outperform the other 14 players on that managers team. Since captains earn double points, managers do have an incentive to own the highest performing players and make them their captain. Also, managers who choose not to have a player on their team are expressing a weak prediction that those on their team will outperform those players not on their team. We say "weak" prediction because managers may also not own a player due to budget or trading constraints.

If we were to design the perfect data set for our analyses we would ask every manager, every week who he thinks will be the top performing player. Because we are using a data set that we scraped from the Web, we have slightly different data: each manager's choices about who will score the most among the players on his or her roster. Thus we are aggregating judgments of over a subset of players into a judgment over the whole set of players.

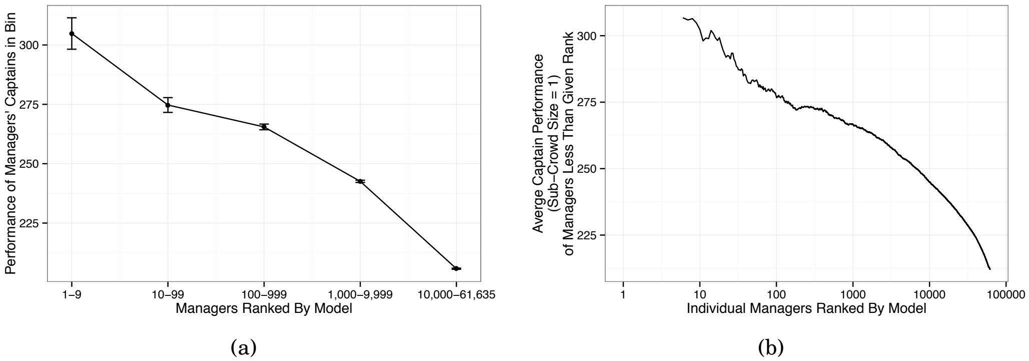
Next we describe how we aggregate the judgments of the managers. Let S be a set of fantasy soccer managers. The captain choices are aggregated from S independently for each gameweek, so let $i \in \{1, 2, \ldots, 38\}$ be one of the gameweeks. We count how many managers in S chose each captain during gameweek i. The most popular captain choice among S is the choice of the subcrowd S for gameweek i. The performance of S for gameweek i is how many points the most popular captain in S earned during gameweek i. The *performance* of S for the whole season is the sum of the performance of S over each gameweek, and this is the quantity that most of our analysis will focus on. The performance of S is how a crowd strategy based on the captain choices of the managers in S would perform. We distinguish the performance of S from the *score* of each of the managers captains in S. Given any manager $s \in S$ the score of that manager's captain is how many total points his or her captain choices actually earned in a given season (we will examine the 2012/13 season). We can aggregate the scores of the managers by averaging them or taking the median for example. But this is different than aggregating the captain choices of these managers which is what the performance of S captures.

5.1. Predicting performance from estimated manager quality

To begin our analyses, we predict the quality of individual managers and relate those to performance of the crowd-chosen captains' performance. We conduct our analyses from the perspective of the beginning of the 2012/13 season. Accordingly, and in the spirit of "predicting the future" instead of "predicting the present", we train a linear model (Model 2 in Table I) only on data from 2011/12 and before. We also exclude a holdout set of individuals from the model's training data so that we are not fitting and testing the model on the same individuals. Recall that we sampled 100,000 managers uniformly at random from the millions who played the 2012/13 season and scraped their week-by-week player rosters. Of these, 38,365 were playing for the first year in 2012/13 and were not used for a lack of past performance data, leaving 61,635 managers, which constitute the "holdout set" we use for testing.

Applying the model to the holdout set results in 61,635 individual predictions of how managers will score in the 2012/13 season. Because we are interested in captain performance and not manager score (as explained in the previous section), we use these predictions simply to rank order the managers from first (i.e., highest predicted "manager quality") to last. We are unable to rank managers by the performance of their captain choices from previous years because the Fantasy Premier League does not make this data available. Moreover, overall scores could be a better, more representative measure of overall manager quality than captain performance.

Figure 3(a) shows the relationship of estimated manager quality rank to 2012/13 captain score averaged over the whole season. Each point corresponds to the average performance of all the managers' captains in a bin. The model, which was only trained



(a)

(b)

Fig. 3. In the left panel, manager quality rank relates to captain performance in the 2012/13 season. Each point reflects the mean captain performance of all managers' captains in its respective bin. Error bars are ±1 standard error. In the right panel, mean score achieved by imitating the captain choice of a randomly chosen manager of a given rank or less.

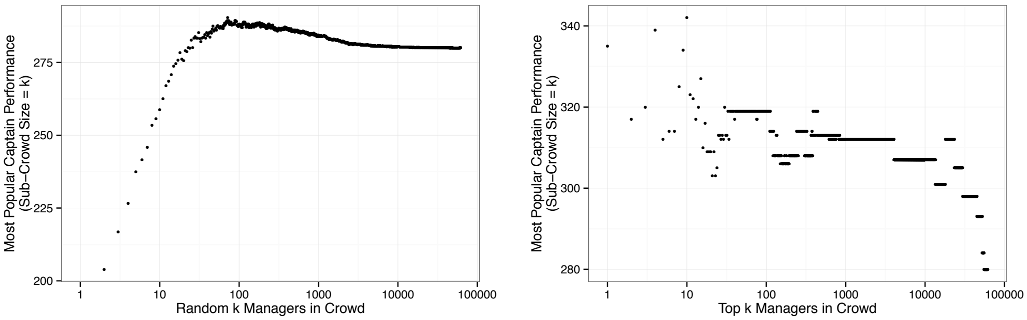
on scores (not captain performance), and on pre-2012/13 data seems to reasonably capture the idea that some managers are better at picking captains than others. Steep drops in performance are observed near the highest-ranked managers. For instance, performance drops about 60 points over the most skilled one-sixth of the distribution (form the 1st to the 9,999th manager) and then drops around only 40 more points over the remaining five-sixths of the distribution (from about the 10,000th to the 61,635th manager).

In the next sections, we model strategies for choosing captains from simple crowd choices, to more complex schemes, and examine their performance relative to one another.

5.2. Strategy: Imitation

Perhaps the simplest crowd-based strategy a manager could undertake is an imitation heuristic [Gigerenzer 2008], which could be operationalized as simply copying the captain choice of another manager (who is willing to share this information). A number of bloggers publish their their fantasy teams, making the imitation heuristic an implementable strategy. In Figure 3(a), the mean performance of all individual managers' captains in the figure was 212 points. This means that copying a manager's captain at random would earn 212 points on average. This mediocrity comes at the cost of variance, however: the 95% confidence interval of scores using this strategy runs from 106 to 318, making the heuristic's performance far from a sure thing.

A natural improvement upon this strategy would be imitating a good manager. Suppose it were possible to put an upper bound on a friend's quality rank (recall that a lower rank implies a better manager). A second simple strategy is to imitate a random manager of a given rank or lower, based on their predicted quality. Figure 3(b) shows the average performance that would be achieved by imitating such a random individual. At the extreme right, we see the familiar 212 performance obtained by imitating any manager's captain choice at random. The improvements over this naive strategy can be considerable. Imitating a random top 10,000 manager gives 245 points in expectation, and a random top 100 manager yields 277 points, but with considerable variation (standard deviations of 44 and 30 points, respectively). If one is fortunate enough to copy a random top 10 manager, performance of over 302 points seems possible, with a standard deviation of 20 points.



(a) Ordering managers randomly

(b) Ordering managers by predicted quality

Fig. 4. In the left panel, we show the performance of choosing a captain based on popularity among k random managers. In the right panel, we show the performance choosing a captain based on popularity among the top k managers, as predicted by Model 2 in Table I.

Imitating a skilled manager leads to increased performance, but with considerable uncertainty in outcomes. The focal strategy of our paper is the wisdom of the crowd, aggregating opinions of managers to make a crowd-based choice of captain. The logic underlying this strategy is that in general, popular captains (as judged by membership on managers' teams) should yield more points.

5.3. Strategy: The wisdom of sub-crowds

A voting-based wisdom of the crowd strategy would entail tallying all managers' captain choices and choosing the most popular one. If collecting additional rosters is costly, a cheaper strategy would be simply picking k managers at random and choosing the most popular captain in this random sub-crowd. In Figure 4(a), we show the results of the random sub-crowd strategy for all values of k ranging from one manager to all 61,635 managers. Each point in the figure is the average of 100 runs. The crowd's choice of captain exhibits better and better performance as votes are added, reflecting the classic wisdom of the crowd. There is a slight decrease in performance as sub-crowd size increases from 100 to around 3,000. A consistent pattern of plots of this type is that they will all converge to the same point. Figure 4(a), at the far right, shows that carrying out a wisdom of the crowds strategy with all 61,635 managers creates a crowd team that achieves a performance of roughly 280. To put this in perspective, recall that imitating a random manager yielded a mean 212 points, a top 100 manager 277 points and only managers in the top 10 had performance of 302 that exceeds the performance of this random crowd strategy.

Can we do better? Figure 3(a) shows that some managers are predictably better than others, and therefore it might be possible to make better manager choices based on the top managers according to a simple model. Next, we create such sub-crowds consisting of the top k managers in terms of predicted quality (via Model 2 in Table I) where we vary k from $1, 2, \ldots, 61,635$ and compute the most popular captain choice for each value of k, breaking ties at random. At one extreme, we will have a crowd of size one, corresponding to the top-ranked manager. At the other extreme, we will have the crowd of all 61,635 managers. Figure 4(b) shows the result. Most striking is the overall high performance observed for the top sub-crowds. While no other strategy yields reliably more than 300 points, in Figure 4(b) we see that all sub-crowds of size 10,000 or less resulted in captains that scored more than 300 points. Surprisingly, 93% of the top 10,000 ranked managers failed to choose captains who scored more than 300 points. That is, the crowd choices beat 93% of the individuals whose votes determined the crowd choices. Smaller crowds have considerable variation in captain performance as marginal managers are added to the sub-crowd. At the same time, larger crowds give more stable performance as managers are added but the cost of the additional, less-informed opinions leads to steady decreases in performance. At the extreme right, we see the familiar value of about 280, which could also be attained by averaging the choices of 30 or more random managers (Figure 4(a)), or by imitating the choice of one randomly-chosen top 50 manager (Figure 3(b)).

In Section A.2 of the Appendix we explore how one could weight managers captain choices with the aim of getting even better crowd performance. We test two different, standard weighting schemes and show that the gain in performance is slight.

Two points stand out from our empirical analyses. First, small crowds make noisier predictions. This stands to reason because small crowds are swayed by individual votes. The second is one of the key conclusions of this work: smaller, more expert crowds perform better than larger crowds. In the next section we consider the effects of differential expertise on the performance of crowd opinion.

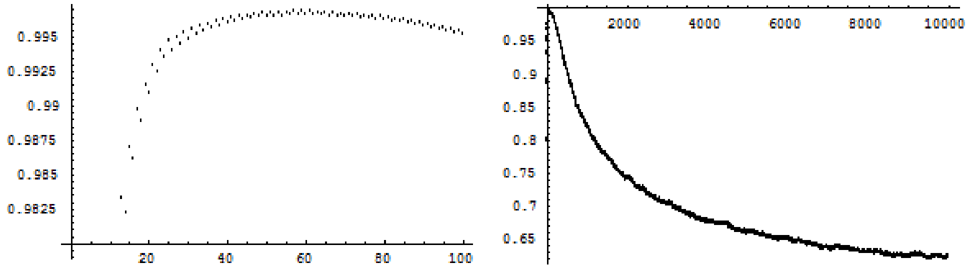
6. THEORETICAL MODELING

Some managers have better, more accurate, more up-to-date, more comprehensive information than others. We investigate the effects of differential information quality in two simple, highly stylized models. The first model considers a captain choice, and two potential candidate players. The second model is a better match to guessing the weight of ox—the standard wisdom of crowds setting—than to fantasy soccer, but with two managers.

Suppose there are two players, A and B. A manager chooses one of them, and wins if that player is the better player. Any given manager m will have a probability of being right, p_m, and we order the managers from highest to lowest. Note that such an ordering could be constructed using historical performance, as we found empirically. We assume that the managers choices' are independently distributed, and the probability of being right is independent of the identity of the better player. Consider the model with

$$p_m = \frac{1}{2} + 0.3 \times 0.98^{m-1}. \tag{1}$$

In this case, the best manager ($m = 1$) is right 80% of the time, and all managers are better than 50%, though most are just barely better. Here manager 60 is right 59% of the time, and manager 100 picks the better captain with probability 54.5%. We pick a consensus captain for the top k managers by picking player A when a majority of the top k managers picked player A, by picking randomly when there is a tie among the top k managers (only possible when k is even), and otherwise picking B. How well does the consensus captain do? Figure 5(a) shows the probability that the consensus captain gets the right answer as a function of k. It is computed by repeatedly letting managers choose independently and get the right choice with the probability specified in (1). There are several points worth noting in this figure. First, the performance of the consensus captain is maximized at 61 managers; add more and the performance drops off. Thus, even though additional managers have information, they do not have good enough information to be worthwhile. A smaller, smarter crowd has greater wisdom than any individual manager, but also greater wisdom than a big crowd. Second, the performance of consensus captains with an even number of managers is worse than



(a)

(b)

Fig. 5. Probability that Crowd is Correct, as a function of Crowd Size. Figure 5(a) shows the probability (y-axis) for k (x-axis) ranging from 1 to 100, while 5(b) shows the probability (y-axis) as k (x-axis) ranges up to 10,000.

the complementary consensus captains. This is a consequence of ties, which are only possible with an even number of managers.[2]

The finding of an interior maximum in crowd size represents a challenge to the conventional wisdom of crowds analysis. While the crowd dominates any single manager, with differences in estimation ability, a big crowd is not very good either. Instead, the optimum involves a relatively modest crowd, at least in this example. Increasing the number of managers can have a substantially deleterious effect on performance. Figure 5(b) shows the extension of Figure 5(a) beyond 100 managers; performance falls off quite dramatically, eventually being worse than the best manager at 80%.

This model is related to a literature on voting; see e.g. [McMurray 2013]. The major difference with that paper is that the present model contains an ordering of expertise, while individual voters in McMurray [2013] estimate their position based on the quality of information.

As the odd-even performance shows, the binomial case is poorly behaved. To gain more insight, consider grouping sets of managers, e.g. the first fifty, the second fifty, and so on. Then it becomes reasonable to treat each group of managers as a random draw from a normal distribution, specifically that each group i contributes the best player X_i times, with X_i a drawn from $N(\tilde{\mu}_i, \sigma_i^2)$. The value $\tilde{\mu}_i$ should exceed ½ of the group size. This approximation is justified for large groups, but we will set the group size equal to one. Such a approximation continues to be valid provided the total number of managers is large enough to be reasonably approximated by normal distributions, since only the union of the groups is considered. Approximating with normal distributions also captures the "guess the weight of the ox" model, which provides an alternate justification for considering it. Moreover, the theorem applies to that situation as well. Given that the first k manager groups are included in our consensus, the correct captain is chosen by the consensus whenever more than half the groups choose the correct player. If the number of managers choosing correctly is $X \sim N\left(\sum_{i=1}^{k} \tilde{\mu}_i, \sum_{i=1}^{k} \sigma_i^2\right)$, then

[2] Without the declining probabilities of success embodied in Equation 1, the differential effects of even and odd persist. In particular, suppose every manager has a probability $p > 1/2$ of being right. Then if n is an odd number of managers, the likelihood that a majority favor the correct answer is the same as with $n + 1$.

the correct player is chosen with probability

$$P\left(X > {}^{k}\!/_{2}\right) = P\left(\frac{X - \sum_{i=1}^{k}\tilde{\mu}_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}} > \frac{{}^{k}\!/_{2} - \sum_{i=1}^{k}\tilde{\mu}_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}\right) \tag{2}$$

$$= P\left(Z > -\frac{\sum_{i=1}^{k}\left(\tilde{\mu}_i - \frac{1}{2}\right)}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}\right) = P\left(Z > -\frac{\sum_{i=1}^{k}\mu_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}\right) \tag{3}$$

where $\mu_i = \tilde{\mu}_i - \frac{1}{2}$. Thus, the best consensus captain arises from the top k managers where k maximizes $\frac{\sum_{i=1}^{k}\mu_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}$. This analysis shows that the optimal consensus captain maximizes the coefficient of variation of the estimate. It is not the case that adding additional managers is necessarily good; adding managers improves consensus teams only when the incremental manager has sufficient accuracy to outweigh the noise they add.

The next theorem provides conditions characterizing when adding managers improves the consensus. It uses the notation

$$\bar{\mu}_k = \frac{1}{k}\sum_{i=1}^{k}\mu_i, \text{ and } \bar{\sigma}_k^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2 \tag{4}$$

THEOREM 6.1. *An additional manager $k+1$ improves the consensus performance if and only if*

$$\frac{\mu_{k+1}}{\bar{\mu}_k} \geq \sqrt{k^2 + k\frac{\sigma_{k+1}^2}{\bar{\sigma}_k^2}} - k \underset{k\to\infty}{\to} \frac{1}{2}\frac{\sigma_{k+1}^2}{\bar{\sigma}_k^2}. \tag{5}$$

Theorem 6.1 provides several insights. First, it provides an exact formula for when adding an additional manager will improve consensus. This formula shows that there is a tradeoff between signal and noise: a higher mean is needed to justify a higher variance, because the right hand side of the inequality is increasing in the variance. This tradeoff demonstrates that the wisdom of a large crowd may be dominated by the wisdom of a smaller, more expert crowd.

Second, when the variances are the same, the ordering of managers is unambiguous: higher mean entails higher accuracy. Theorem 6.1 shows that an additional manager is helpful to the consensus when the manager is at least as good as $\sqrt{k^2+k} - k$ times the average performance of the existing managers in the consensus. Moreover, $\sqrt{k^2+k} - k$ is in this instance approximately ½. Colloquially, it is worth adding a manager to the consensus view provided that manager is not "half bad," relative to the average performance of managers in the consensus.

Third, for large k, it is worth adding a manager when $\frac{\mu_{k+1}}{\bar{\mu}_k} \geq \frac{1}{2}\frac{\sigma_{k+1}^2}{\bar{\sigma}_k^2}$, which may be rearranged to require

$$\frac{\mu_{k+1}}{\sigma_{k+1}^2} \geq \frac{1}{2}\frac{\bar{\mu}_k}{\bar{\sigma}_k^2}. \tag{6}$$

The inequality (6) provides an ordering on the managers, for large k: order by mean over variance. While this ordering is not exact, it holds for large values of k. (It is shown in the proof that the rate of convergence is ${}^{1}\!/_{k}$.) This approximation also reinforces the

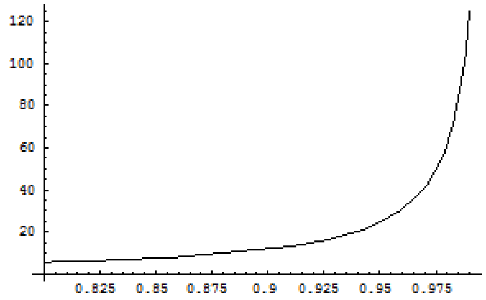


Fig. 6. The optimal k (y-axis) as a function of the depreciation rate b (x-axis). The figure shows the variance-minimizing size of the crowd, as a function of the rate b at which successive managers' predictions worsen, for $b < 1$.

"add a manager who is not half bad" conclusion of common variances, because the critical value is half the average mean over the average variance.

The best consensus may arise from less than the wisdom of the full crowd, as in the example at the beginning of the section. It is not optimal to add managers who are significantly worse than half the average performance of the consensus. Recall that μ_i is the probability of being correct minus ½. Thus, for example, if the average accuracy of the group is 70%, adding a manager with accuracy 60% improves the consensus, while adding a manager with accuracy 59% will not when $k \geq 3$.

To complete this analysis, we return to the first example, where $p_m = 1/2 + a \times b^m$. In this case, $\mu_i = p_i - 1/2 = ab^i$. When the variance be approximately constant, the optimal number of managers in the consensus is determined by maximizing

$$\frac{\sum_{i=1}^{k} \mu_i}{\sqrt{k}} = \frac{ab}{1-b} \frac{1-b^k}{\sqrt{k}}$$

Figure 6 shows the accuracy-maximizing value of k as a function of the deterioration rate b of the managers. The maximum has an approximation

$$k* \cong \frac{0.284668}{\log(b)}.$$

which is exact except that the numerator is approximated. Note that for $b = .98$, we obtain the solution of $k = 62$, which is similar to the value 61 that we found by direct computation.

6.1. Modeling Information

How much weight does a manager with access to better estimates put on well-known information?

As shown above, managers can add more variance to the consensus than their information justifies, decreasing the performance of the consensus captain choice. The consensus captain might want to ignore the choices of worse-performing managers for a completely different reason: if the information of the worse managers has already been incorporated into the better manager's estimates.

We switch the context of the model to estimating a quantity observed with error, such as the weight of an ox or the number of jelly beans in a jar. This alternate model makes more sense with small numbers of managers, and is widely applicable in wisdom of crowds settings.

We consider two managers. Manager 1 has one signal. Manager 2 has access to two signals: an idiosyncratic signal, and a noisy version of manager 1's signal. Manager 2 will incorporate information about manager 1's signal in his estimate, perhaps enough that manager 2's estimate is a sufficient statistic for both managers' estimates. That the incorporation of manager 1's information into manager 2's estimate provides a distinct explanation of why a manager may not be helpful for a consensus captain choice will follow from the fact that the optimal weight to place on manager 1's signal may be negative, because manager 2 places too much weight on manager 1's signal.

Suppose the true value being estimated—the actual weight of the ox—is μ. We let μ have a diffuse prior, so that the Bayes update given any signal will just be the signal. Let manager 1 observe $\mu + \epsilon$ and manager 2 observes both a signal $\mu + u$ as well as a noisy copy of manager 1's signal, $\mu + \epsilon + v$. We restrict manager 2 to a convex combination of the two signals, that is, $\beta(\mu + u) + (1 - \beta)(\mu + \varepsilon + v)$. The random variables u, v and ϵ are all assumed to have mean zero, be independently distributed, and we denote the variance of any variable x by σ_x^2, the CDF by F_x and the PDF by f_x.

THEOREM 6.2. *The probability that manager 1 wins is*

$$\int_0^\infty f_z(z)\left(1 - F_\varepsilon\left(\frac{z}{\beta}\right) + F_\varepsilon\left(\frac{-z}{2-\beta}\right)\right)dz + \int_{-\infty}^0 f_z(z)\left(1 - F_\varepsilon\left(\frac{-z}{2-\beta}\right) + F_\varepsilon\left(\frac{z}{\beta}\right)\right)dz$$

If the distributions are normal, this expression is

$$1 - \frac{1}{\pi}\left(ArcCot\left(\frac{\beta\sigma_\varepsilon}{\sqrt{\beta^2\sigma_u^2 + (1-\beta)^2\sigma_v^2)}}\right) + ArcCot\left(\frac{(2-\beta)\sigma_\varepsilon}{\sqrt{\beta^2\sigma_u^2 + (1-\beta)^2\sigma_v^2)}}\right)\right)$$

To illustrate the theorem, we consider the case $\sigma_u^2 = \sigma_v^2 = 1$. This means that the prediction errors made by manager 2 are the same on both his observation of manager 1's signal, and his own signal. Such a situation might arise if the source of the errors were transcription errors. How much weight would minimize the variance of the manager's estimates? This variance is $\beta^2\sigma_u^2 + (1 - \beta)^2(\sigma_v^2 + \sigma_\varepsilon^2) = \beta^2 + (1 - \beta)^2(1 + \sigma_\varepsilon^2)$, which is minimized at $\beta = \frac{1+\sigma_\varepsilon^2}{2+\sigma_\varepsilon^2}$. This formula provides a benchmark, because it is the level a manager would choose to maximize accuracy. Note that if manager 2 chooses that level, much of the desired weight on manager 1's signal is already incorporated into manager 2's estimate.

How much does a manager actually choose when the manager maximizes the probability of winning? In Figure 6.1, we plot the value of β that maximizes manager 2's probability of winning as a function of σ_ε, and compare it to the weight that minimizes variance of manager 2's estimate. Interestingly, taking into account the competitive nature of the manager's problem can either increase or decrease the amount of weight placed on manager 1's signal relative to the variance-minimizing weight, although it only decreases it when manager 2's signal is more precise than manager 1's signal.

When manager 1 is very accurate (σ_ε^2 is small), manager 2 puts less weight on his estimate of manager 1's signal and more on his private knowledge than would maximize manager 2's accuracy. This is because manager 1 has the advantage over manager 2 (a better estimate than is available to manager 1) and manager 2 essentially bets on his private signal. In contrast, when manager 2 has a better estimate available (specifically, when $1 = \sigma_u^2 \leq \sigma_\varepsilon^2$), manager 2 puts less weight on his own signal than would minimize squared error.

For the more plausible case of large values of σ_ε^2, manager 2's estimates put too much weight on old information. Thus, to maximize overall accuracy, manager 1's signal should probably enter negatively in a crowd-wisdom average, in order to undercut the overweighting of this information in manager 2's signal. Indeed, the most weight

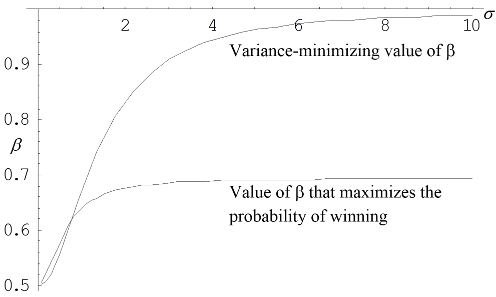


Fig. 7. The weight placed on own information. The upper curve shows the weight which minimizes variance, while the mostly lower curve shows the amount that maximizes the probability of winning. This shows that for most values, the manager underweights his own information and overweights the noisy signal of the rival's information.

placed on the manager's own signal, even as the variance of manager 1's signal diverges, is 69.4% (the actual value is a weird number involving $\sqrt{114}$).

The interesting finding in this section is that if better managers have access to the information available to lesser managers, the better managers will not only incorporate that information, but will put too much weight on the information available to lesser managers. Thus, when using the predictions of the better managers, the information of the lesser managers has already been incorporated, and then some. Consequently, if the predictions of the lesser managers were to be incorporated, these predictions would enter negatively, to undo some of the better managers' overweighting of the information. The desire to win requires the better managers to blunt the value of information available to lesser managers, resulting in the overweighting of this information. Wise crowd members negate the value of less wise crowd members.

We have seen that large crowds make worse predictions than smaller, more expert crowds for two quite distinct reasons. First, differences in prediction ability mean that adding poor predictors could lower overall predictive performance. In particular we showed a trade off between the content of additional predictions and the noise that is added. Approximately speaking it is worth adding an extra manager when that manager is at least half as good as the group average. The second reason for eliminating weaker managers is that their information may already be incorporated in the predictions of the stronger managers.

7. CONCLUSION AND FUTURE WORK

In this paper, we have empirically assessed three main methods for extracting wisdom from the crowd: Imitating a random manager of a given rank or less, aggregating the choices of randomly-chosen managers, and aggregating the choices of top-ranking managers. To evaluate these strategies relative to one another, Figure 8 summarizes these strategies' performance at selected crowd sizes.

The relative success of imitation to aggregating random opinions depends on the number of experts available and what is known about their skill. It turned out to be better to imitate a manager known to be in the top 30 than to aggregate the choices of 30 random people, but the average top 100 manager does worse than aggregating 100 opinions of randomly chosen people. One point our analysis and 8 makes clear is that tapping the wisdom of small, smart crowds can beat both imitation and random crowd strategies by a large margin. In games with at least some element of skill it makes sense that one may be able to predict who may have more skill based on past behavior. The main conclusion of this work is that one can find smarter, smaller crowds within a larger crowd by aggregating only those with higher predicted skill levels.

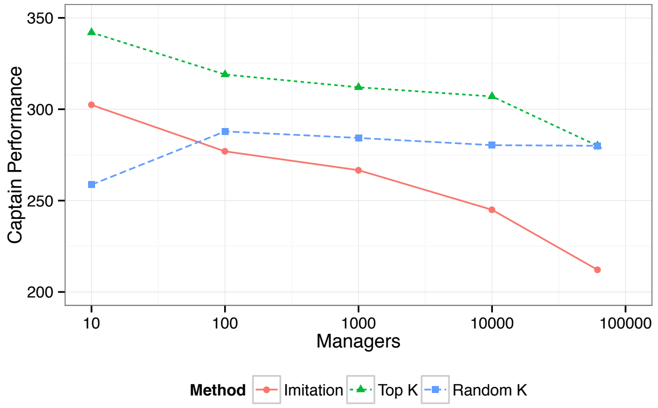


Fig. 8. Comparison of three main strategies at selected points from Figures 3(b), 4(a) and 4(b).

When expertise is not evenly spread throughout the crowd, it is better to focus on the concentration of the expertise as opposed to diluting it with experts of a lower quality. As a result, the wisdom of the experts in the crowd can beat the wisdom of the whole crowd.

We showed theoretically that that there is a trade-off between adding an informative manager to a crowd and the variance that manager brings. When the predictions are normally distributed, the criterion for improving the crowd wisdom takes a particularly simple form. The rule is approximately to add a manager when that manager is at least half as good as the crowd average, that is, they are not half bad. In addition we show the managers with access to better information may over-weight the information of rivals. In particular the information held by lesser managers is already incorporated into the estimates created by better managers. This incorporation arises through competitive effects, as managers try to mitigate the influence of information held by others. Consequently it may be optimal to ignore, or even negatively weight the information held by less expert managers.

Moreover, aggregating many of those predicted to be experts resulted in better predictions than simply imitating the best predicted experts. (Compare the leftmost point in Figure 3(a) to the leftmost points in Figure 4(b)). These overall effects are robust to different weighting schemes. Furthermore, we showed that crowd strategies are effective in settings beyond the tradition one of averaging numerical estimates.

The wisdom of crowds phenomenon underlies many prediction mechanisms such as prediction markets, risk assessments, and economic forecasts. Our method shows that improved predictions may be obtained by giving high weights to the opinions of those who have demonstrated skill in the past. Because data sets with individual-level records of human performance are becoming increasingly available, the future may hold more occasions on which to improve upon crowd predictions by identifying and tapping into the wisdom of smart sub-crowds.

Prediction markets are an interesting case of crowd wisdom. Since they have an endogenous weights, experts bet more. However, the size of the bets is also determined by wealth, risk tolerance, enjoyment of the game, as well as information. Our results suggest that it may be desirable to limit participation in prediction markets to those with proven expertise. Moreover, it would be very interesting to study, both theoretically and empirically, whether participants in prediction markets overweight information held by others.

There are a number of directions one could take this research in the future. For example, one could analyze more complex weighting schemes or apply more sophisticated models to better identify highly skilled managers. One could also try to make new teams based on popular players, or analyze the performance of the most common entire teams. Finally, one could also investigate if the smart-crowd techniques examined here can boost the predictive abilities of polls, expert forecasts, and prediction markets.

REFERENCES

BRUCE, R. S. 1935–1936. Group judgements in the fields of lifted weights and visual discrimination. *Journal of Psychology 1*, 117–121.

FAMA, E. F. AND FRENCH, K. R. 2010. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance LXV,* 5, 1915–1947.

GALTON, F. 1907a. Letters to the editor. *Nature 75,* 1952.

GALTON, F. 1907b. Vox populi. *Nature 75,* 1949, 450–451.

GIGERENZER, G. 2008. Why heuristics work. *Perspectives on Psychological Science 3,* 1, 20–29.

GORDON, K. 1924. Group judgements in the field of lifted weights. *Journal of Experimental Psychology 7*, 389–400.

HERZOG, S. M. AND HERTWIG, R. 2009. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science 20*, 231–237.

KLUGMAN, S. F. 1945. Group judgement for familiar and unfamiliar materials. *Journal of Genetic Psychology 32*, 103–110.

KNIGHT, H. 1921. A comparison of the reliability of group and individual judgements. M.S. thesis, Columbia University. unpublished.

LEVITT, S. D., MILES, T. J., AND ROSENFIELD, A. M. 2012. Is texas hold'em a game of chance? a legal and economic analysis. *The Georgetown Law Journal 101*, 581–636.

LORENZ, J., RAUHUT, H., SCHWEITZER, F., AND HELBING, D. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences 108,* 22, 9020–9025.

LORGE, I., FOX, D., DAVITZ, J., AND BRENNER, M. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin 55,* 6, 337–372.

MANNES, A. E., SOLL, J. B., AND LARRICK, R. P. 2013. The wisdom of small crowds. Manuscript in Preparation.

MCMURRAY, J. C. 2013. Aggregating information by voting: The wisdom of the experts versus the wisdom of the masses. *The Review of Economic Studies 80,* 1, 277–312.

MUCHNIK, L., ARAL, S., AND TAYLOR, S. J. 2013. Social influence bias: A randomized experiment. *Science 341,* 6146, 647–651.

SIMMONS, J. P., NELSON, L. D., GALAK, J., AND FREDERICK, S. 2011. Intuitive biases in choice vs. estimation: Implications for the wisdom of crowds. *Journal of Consumer Research 38,* 1, 1–15.

SUROWIECKI, J. 2005. *The Wisdom of Crowds*. Anchor.

TREYNOR, J. L. 1987. Market efficiency and the bean jar experiment. *Financial Analysts Journal 43,* 3, 50–53.