Expert Stock Picker: The Wisdom of (Experts in) Crowds

Shawndra Hill and Noah Ready-Campbell

ABSTRACT: The phrase "the wisdom of crowds" suggests that good verdicts can be achieved by averaging the opinions and insights of large, diverse groups of people who possess varied types of information. Online user-generated content enables researchers to view the opinions of large numbers of users publicly. These opinions, in the form of reviews and votes, can be used to automatically generate remarkably accurate verdicts—collective estimations of future performance—about companies, products, and people on the Web to resolve very tough problems. The wealth and richness of user-generated content may enable firms and individuals to aggregate consumer-think for better business understanding. Our main contribution, here applied to user-generated stock pick votes from a widely used online financial newsletter, is a genetic algorithm approach that can be used to identify the appropriate vote weights for users based on their prior individual voting success. Our method allows us to identify and rank "experts" within the crowd, enabling better stock pick decisions than the S&P 500. We show that the online crowd performs better, on average, than the S&P 500 for two test time periods, 2008 and 2009, in terms of both overall returns and risk-adjusted returns, as measured by the Sharpe ratio. Furthermore, we show that giving more weight to the votes of the experts in the crowds increases the accuracy of the verdicts, yielding an even greater return in the same time periods. We test our approach by utilizing more than three years of publicly available stock pick data. We compare our method to approaches derived from both the computer science and finance literature. We believe that our approach can be generalized to other domains where user opinions are publicly available early and where those opinions can be evaluated. For example, YouTube video ratings may be used to predict downloads, or online reviewer ratings on Digg may be used to predict the success or popularity of a story.

KEY WORDS AND PHRASES: data mining, prediction markets, social media, user-generated content, wisdom of crowds.

In this paper we show that user-generated content (UGC) is an acceptable theater in which crowd wisdom can be used to identify good verdicts—in this case, accurate stock picks. Furthermore, we show that when we identify, or at least reveal, experts and weight their votes accordingly, we perform more accurately than when we use everyone in the crowd to vote for stocks. Our contribution is that we provide a method based on a genetic algorithm (GA) to learn the appropriate contributions of independent users through the use of observed past individual performance. We compare and evaluate our approach in the context of criteria used in past research to generate stock portfolios.

The authors thank Anthony Crawford for research assistance, Theodoros Evgeniou, Steve Kimbrough, and Vasant Dhar for valuable comments, and Motley Fool CAPS for allowing us to use their valuable data.

International Journal of Electronic Commerce / Spring 2011, Vol. 15, No. 3, pp. 73–102.
Copyright © 2011 M.E. Sharpe, Inc. All rights reserved.
1086-4415/2011 $9.50 + 0.00.
DOI 10.2753/JEC1086-4415150304

04 hill.indd 73 2/21/2011 11:23:14 PM

In prior research, stock market analysts that worked in financial firms were evaluated to find "start analysts" [7]—the top performing analysts, to make stock market predictions. In addition, people that make stock market picks, as a leisure activity, online [11], have been evaluated in aggregate to find good stock portfolios. Gu et al. [11] weigh the different posts by the author's credibility based on the accuracy of the author's past post and most credible authors are considered experts. Likewise, we use a mixture of experts approach [11] informed by the method of Jordan and Jacobs [14] In our work, we use historical individual level stock pick data from online users to identify users that have been successful at the task of picking stocks in the past—we call these successful users "experts." The ability to identify experts as part of the crowd enables us to take better advantage of the "wisdom of crowds" [22] by restricting the crowd to a set of experts.

The purpose of this research is to implement a stock-trading strategy using the publicly available Motley Fool CAPS data (http://caps.fool.com). If the trading strategy proves to be modestly successful, it could be of broad interest to investment managers looking for alternative investment strategies, at least in the short term. In addition, the improved scoring system could be of considerable interest to the CAPS team and other firms making stock voting data available. Most important, however, showing that the wisdom of crowds is effective for decision making may have implications for how firms and social systems should be organized around group voting for tough decision making.

There were approximately 116 million consumers of UGC and 82.5 million content creators in February 2009, according to market research and analysis firm eMarketer [24]. The bottom line: groups and crowds are contributing their opinions online in public venues at a spectacular rate. In this research we take advantage of publicly available UGC for decision making—specifically, stock picks.

There are many sources of online UGC submitted by millions of creators; for example, social networking, blogs, online reviews, question-answer, pictures, video, and wikis. In addition, votes and aggregate opinion are available from voting and information/prediction market sites, which are most often used to predict financial, election, and sports outcomes. UGC has been used in aggregate to predict recommendation system ratings [15], music sales [6], and blockbuster performance [8]. User-generated text has also been used to predict stock market performance [2, 18, 25].

The types of UGC sites we are interested in are the many online prediction and voting markets, such as BetFair, NewsFutures, Hollywood Stock Exchange, and Popular Science Predictions Exchange. As their names suggest, these sites enable users to bet on and make predictions about the outcomes of future events. Some use virtual money, and others use real money on their exchanges, with varying missions from profit to philanthropy.

Most relevant to our research are the major players in the stock voting game: Piqqem, Cake Financial, Covestor, Predictify, and the Motley Fool CAPS. These sites fall into two categories—quasi-prediction markets (where "quasi" means data are aggregated from disparate sources on the Web, but explicit voting did not necessarily occur) and prediction markets (where explicit voting did

occur). In both cases, the sites offer solutions to help aid in financial decision making. To our knowledge no one has applied the wisdom of (expert) crowds theory to a large-scale stock pick data set.

The challenge in successfully picking stocks, however, is that, obviously, it is difficult. Even though there are challenges to the efficient market hypothesis [12], methods to challenge the hypothesis historically required in-house experts or proprietary models based on financial indicators and environmental factors to identify good stock picks, and even the most sophisticated resources and tools are unreliable. In this paper we propose to augment existing approaches with UGC for this problem. Using aggregate-level expert votes is not new—in fact, it has been found that in aggregate expert financial analysts tend to perform better than they do alone [9, 13, 19, 20]. But using large-scale user data from online sources is new—especially if it is not known whether the users are experts or not.

We believe the results of this research to be of considerable intellectual and practical value not only to the financial discipline but also to domains where problems are hard and voting on the solutions is possible. In fact, because of our findings in this study, we advocate for reputation mechanisms for all UGC to enable firms and individuals to identify experts and therefore make more accurate predictions and decisions. Identifying the experts in the crowd, or the wisdom of the few, has already been shown to be useful in domains outside finance.

For example, in collaborative filtering, a nearest neighbor collaboration approach was augmented with external expert validation data in order to identify the users that should be considered for the nearest neighbor approach. The approach filters expert users based on their expertise in making accurate recommendations for users [1]. In addition, harnessing the wisdom of the few in Wikipedia has shown to be useful [16, 17]. For Wikipedia, increasing the number of users beyond a threshold is costly because it leads to noisy posts and high coordination needs among users. Finally, most recently, researchers have integrated social network data to study influence in the context of different assumptions about trust of network neighbors on the network [10]. To date, mechanisms for voting and identifying trust in different contexts are few. The landscape of UGC contributions is changing, however.

In the field of investment management and quantitative investing, for example, there is no known prior strategy using broad stock voting data. Until recently, such data were impossible to acquire, because the online voting systems, like CAPS, are unlike any financial voting system previously created. These systems, for the first time, allow us to measure expertise externally. Additionally, this study builds on and extends the work of others that have applied machine learning techniques to portfolio management [5].

We pursue two hypotheses in this work. First we test the hypothesis that using the stock picks of a large sample of online users from Motley Fool CAPS enables us to outperform the S&P 500. Second, we explore the hypothesis that applying our approach to identify experts in the crowds of online stock pickers on the Motley Fool CAPS site will help us do better than the baseline of the S&P 500 for stock price as well as better than letting the entire online crowd vote.

04 hill.indd 75 2/21/2011 11:23:14 PM

This research aims to capitalize on the tension between needing the crowd and needing expert decisions for prediction. Specifically, we evaluate different methods to rank these online experts based on our custom approach against variants of two established methods in the literature: (1) a mixture of experts' strategy [14] that was used on message board posts and the experts' associated sentiments [11] and (2) a method proposed by Fang and Yasuda [7] that was used on real trading data from stock picks of real-world star analysts.

The remainder of the paper is organized as follows: In the next section we discuss our testbed. In the third we discuss our method of utilizing the large-scale data to reach a verdict in reference to which stocks to pick. We discuss our results in the fourth section, and in the fifth we conclude with a discussion of future work.

Testbed

The Motley Fool is a well-respected financial newsletter publisher with a strong online presence. The firm created a new service in 2006 called CAPS, a stock voting system whereby each user can make predictions about the performance of stocks—namely, whether they will under- or outperform the market. Users are ranked according to the accuracy of their predictions, and stocks are ranked according to the quantity and quality of the users voting for and against them. In this way, each stock is ranked on a five-star system, with the theory being that stocks with five stars will perform better than stocks with one star. The exact rating equation used by the CAPS site is not public. However, the ranking system described in general terms on the CAPS site is the inspiration for our approach to identify experts from their prior stock pick performance. Our user ranking system is described in detail in the Method section. In the section below, we describe the data used from CAPS to identify the experts.

Data Acquisition

We sourced the publicly available votes directly from the CAPS Web site.[1] The data stored do not contain any identifying information on voters, nor are they used for our analysis. We were able to track the votes from January 2007 through December 2009. Altogether, we use over 2 million stock picks in our analysis.

We combined the CAPS data with stock price data from the Center for Research in Securities Prices (CRSP), which was downloaded through Wharton Research Data Services (WRDS). These data were used to calculate returns for stocks (and hence scores for users). CRSP was preferable to other price providers, as it has a history of reliability, and it also provides a "holding period return" value for each stock and trading day. This number differs from a simple ratio of prices in that it takes into account splits, dividends, and other pricing anomalies. The CRSP data also provided S&P 500 prices, which were used for evaluating our overall method as well as for evaluating the expert voters in our data set.

04 hill.indd 76 2/21/2011 11:23:14 PM

Stock Pick Data

By the end of 2008, there were at least 773,861 registered users. We determined that the number of picks per month appears to be increasing. For each pick, we can collect a number of attributes. An example of user stock picks is shown in Table 1. The picks data were saved in .csv files. The column identities were as follows: the date of the pick; whether the pick was added or removed by the user on this date; the ticker symbol of the applicable stock; whether the pick predicted under- or outperformance; the predicted time horizon in which the under- or outperformance would be realized; the price of the stock when the pick was made; and the hashed ID of the user who made the pick. There were approximately 2 million user-generated stock picks in our data set.

Descriptive Statistics

Over the first testing period (June 2, 2008 through December 24, 2008), we took a simple approach and just let the entire crowd vote on the stock picks. The whole-crowd approach fared very poorly on overall return, –23.2 percent. However, the S&P 500 total return over the same period was even worse, –35.5 percent. The difference suggests that we can learn something from the CAPS data.

Users participated in voting at different rates. A significant number of people made few picks, and others made a significant number: the minimum number of picks was 0; the maximum was 13,104. The distribution of number of stock picks is shown in Figure 1.

In Figure 2 (left) we show the distribution of average performance of users with respect to their prediction accuracy. In this plot a user gets a prediction right if the user says that a stock will outperform the market and the stock price for that stock goes up the next day. Likewise, if the user predicts underperformance and the stock goes down, we count it as an accurate prediction in Figure 2. For an individual user we take the number of correct picks divided by the total number of picks. On average, users are 49.1 percent correct, by this definition of correct, over the entire test data set, January 2007 through December 2008. On the surface the plots look as if the stock picks of the users are just a coin flip—the users get the stock movement direction right 50 percent of the time.

In Figure 2 we see that there are outliers at 0 and 100 percent correct. This is due in part to a significant number of people making only one pick. This one pick is either right or wrong, leading to the outliers. With Laplace correction, we are just advocating that one should take the number of picks a user makes when assigning a probability to how right the user might be in the future based on the user's picks.

If we apply Laplace correction to adjust for the variation in the number of picks, we get the distribution of "probability estimates"—the likelihood the user is correct—corrected for the number of picks the user made shown in Figure 2 (right). The plot indicates that indeed some people perform better than others with respect to just getting the direction of stock movement right.

04 hill.indd 78

Table 1. Stock Pick Data Examples.

Time	Add	Ticker	Out	Holding period	Price	ID
01/03/2007	Added	NVEC	U	5Y	$31.30	1
01/03/2007	Added	DEBS	DL	3W	$26.78	2
01/03/2007	Added	EDU	O	5Y	$35.17	3

Notes: **<<ALL ABBREVIATIONS NEED TO BE DEFINE / NVEC, DEBS, EDU, U, DL, O, Y, W>>**

2/21/2011 11:23:14 PM

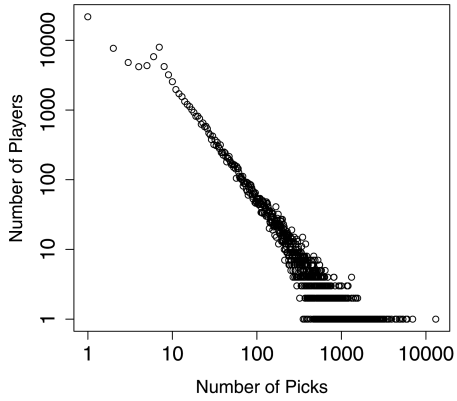

Figure 1. Number of Picks Distribution

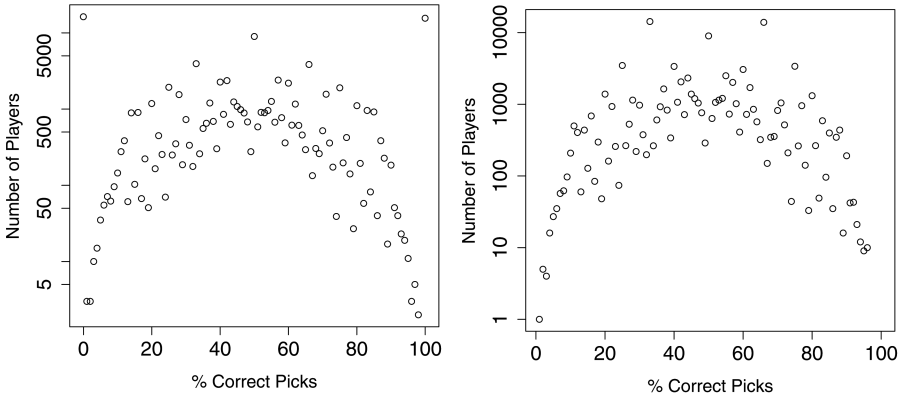

Figure 2. Distribution of Picking Accuracy (left) and Distribution of Picking Accuracy After Laplace Correction (right)

Figure 2 indicates nothing about future returns, however. We will explore returns further when we get to the "Methods" section. In the future, we can use these probability estimates to rank users, and thus their associated stock picks, thereby using individuals as probability estimators. Our method, described in the third section, will rank the best users by their expertise to help decide whose votes should count when picking stock portfolios.

Data Preprocessing

Several preprocessing steps were performed before beginning data mining. The data set proved to be too large for time-efficient computations, so a 25 percent random sample was taken for each model built and for each data set used to test the model. Many picks were made on nontrading days (e.g., weekends and holidays). To enable efficient calculations, each pick made on

a nontrading day was bumped to the closest later trading day. For example, a pick made on Saturday was processed such that its associated date became the following Monday (assuming Monday was a trading day). Sometimes a given ticker and date pair was unable to result in an accurate return. These errors were generally the result of clerical errors, such as changing ticker symbols, incompatibilities in ticker formatting (e.g., BRK-A vs. BRK/A), and the like. The errors were simply caught and ignored, as they accounted for less than 3 percent of returns calculations. All of the preprocessing steps were compared against the raw data (e.g., 25 percent sample vs. 100 percent sample) to make sure we did not introduce any obvious bias.

Expert Stock Picker Approach

We used a GA to find the final trading strategy. As in all GAs, we optimized a fitness function by crossing (mating) and mutating "organisms"—that is, collections of input parameter values that represent potential solutions—and selecting the best solutions based on their "fitness." The fitness function and its input parameters are discussed in following sections. In this project each gene was simply set to mutate at the same rate, 0.01. The crossover rate was set to 0.01 as well. Thus, with each generation there is a 1 percent chance that a given gene will randomly change and a 1 percent chance of crossover.

GAs are commonly used in financial settings, and they offer several advantages [5] (of which we took advantage for this project) over other competing classification techniques. For example, they are designed for very large solution spaces, where a simpler, brute-force optimization technique would be impractical. In this project, assuming ranges and steps similar to those used in the GA, a brute-force optimization consisting of testing every possible permutation of fitness function inputs would result in approximately 2 billion runs. At approximately 2 minutes for each run, that is more than 7,000 years of testing.

Unlike neural networks, also used often in finance applications, GAs deliver understandable and communicable results. This is very important in many financial settings, particularly with faith in "blackbox—only inputs and outputs of the model can be observed as opposed to details of the model," investing strategies ebbing. GAs also allow for more fluid optimization objectives than those afforded by standard classification methods like classification trees, naive Bayes and logistic regression where classification accuracy is typically the primary objective or fitness function. GAs on the other hand can be used with virtually any fitness function that can be articulated. In our study, the ability to work with a strategy-based fitness function was very valuable: a significant experimental redesign would have been required if GAs had been forgone.

Method

To test our two hypotheses—(1) stock picks based on the entire crowd–based wisdom will outperform the S&P 500 and (2) stock picks based on the picks

of experts, identified by our expert ranking approach, will outperform using the entire crowd to vote—we used a GA to identify an investment strategy relying on a blend of expert-based wisdom and crowd-based wisdom. We then compared the performance of this strategy to two baseline strategies, the first driven solely by the crowd—by the S&P 500—and the second ranking people only on their past two stock picks. In addition to the baselines we used for comparison, we ran two very important sanity checks, which we discuss later. Finally, we compared our strategy to two methods discussed in the literature, one from computer science and one from finance. Our approach includes both a ranking component to identify experts and an investment component. From now on, we will refer to these two components either in part or together as the custom strategy. In broad terms, our custom-custom-blended strategy was designed as follows.

Our custom stock-picking strategy consisted of several parts. First, we ranked all the users (according to a metric described later in this paper). Then, we took the top A expert users and invested in them all equally. (It is important to note that this investment only occurred after a delay of one day, in order to ensure that a real-world implementation would be able to acquire the necessary picks before the market opened the following day.) That is, we invested 1/A of the portfolio in each user's picks, with the portion of the portfolio assigned to a given user being distributed evenly through all of the stocks picked by that user. We re-ranked the users every B days; then, we reinvested the portfolio according to the new ranking.

Our user-ranking methodology scored users based on the performance of their picked stocks. We looked at all of the picks made by a user over the past C days, then looked at the return of those stocks over a holding period of D days. A user's score was the product of the returns of the user's picked stocks over the C-day period, with the inverse return being used for stocks picked to underperform. Thus, there were four parameters to our strategy, which are summarized as follows: A = Number of experts—the number of top-ranked users whose picks we considered; B = Portfolio holding period—the period of portfolio re-balancing and user re-ranking; C = User test period—users' picks during this period influenced their score; D = User test holding period—users' picked stock returns were calculated over this time period. When we learn a set of parameters (A,B,C,D) on a training data set, we call those optimized parameters *opt*-CAPS for that training set.

To better understand our methodology, consider the following example: Suppose our parameter values are A = 5, B = 3, C = 10, D = 2. Every 3 days (the portfolio holding period), we rank all the users, then look at the top 5 (the number of experts). We then invest 20 percent of the portfolio in each of the users. Suppose, for example, that one day the top user picked 5 stocks, and the second-best user picked 10. In this case, we would invest 4 percent of the portfolio in each of the 5 stocks picked by the top user and 2 percent of the portfolio in each of the 10 stocks picked by the second-best user. If the two users had a pick in common, then we would invest 6 percent of the portfolio in that position. Also, we note that if a pick expects a stock to underperform, then we would take a short position on that stock.

With regard to ranking users in our example, we would consider a user's picks over the last 10 trading days (the user test period). Over the course of those 10 days, we would tally the returns of the user's picked stocks using holding periods of 2 days (the user test holding period). The user's score would then be the product of these returns—using inverse returns for underperforming picks. Naturally, we would not consider the stocks picked by a user yesterday, because, with a user test holding period of 2, the results of those picks could not be known until tomorrow.

We used a GA to learn the values for these four parameters, on our training period data, to maximize our fitness function—namely, the Sharpe ratio of an investment strategy over the training period. The Sharpe ratio is defined as

$$S = \frac{R - R_f}{\sigma} = \frac{E\left[R - R_f\right]}{\sqrt{\text{var}\left[R - R_f\right]}},$$

where R is the asset return, R_f is the return on a benchmark asset (we use the risk-free rate of return [3 percent annualized, which is a close approximation of the long-term mean risk-free rate—the risk-free interest rate for investing money]), $E[R - R_f]$ is the expected value of the excess of the asset return over the benchmark return, and σ is the standard deviation of the asset excess return [21]. Once the parameters are learned on the training period data set in one time period, we apply the learned model to test data drawn from a future time period.

Discussion of Competing Investment Strategies Found in the Literature

In the previous section we described our proposed approach to generate portfolios that can be evaluated by the Sharpe ratio. In this subsection, we describe additional strategies for portfolio selection, discussed in the research literature, that we will adapt to use as baselines for comparison to our approach, using the Sharpe ratio. In the next subsection, we will detail how we informed our baseline approaches from the papers discussed.

First, we draw on a mixture of experts [14] strategy utilized by Gu et al. [11] for stock prediction using message board posts. These authors focus on extracting sentiment data from online message board postings on Yahoo! Finance. These message boards are typically open to the public, and those considered here focus on 71 specific equities. Typical content consists of advice, predictions, and opinions on a given message board's assigned stock. Additionally, each post can be self-labeled as positive or negative; this is the source of sentiment data used in this investigation, not the actual textual content. Predictions are synthesized using a mixture of experts' frameworks, in which each posting is considered to be a single expert prediction, and, using a weighted average, all are aggregated to form a single prediction for the given stock. Weights for each expert are updated according to an exponential averaging technique. Gu

04 hill.indd 82 2/21/2011 11:23:16 PM

et al. [11], therefore, propose a strategy for ranking users. From now on, we will refer to their *ranking* of experts strategy as the *mixexp* strategy.

The Gu et al. [11] paper's second focus is creating a trading strategy that leverages these predictions to make them profitable in the real world. The strategy focuses on trades varying in length between 1 and 50 days. Some ancillary costs, such as commissions and fees, are taken into account. However, others, such as borrowing costs and taxes (which are important given the short trading period), are not. The authors demonstrate the successful extraction of sentiment data from an online community using a fully automated crawling system. More significant, they demonstrate that financially profitable information is contained within this data, at least with regard to the given stocks and within the given time period. The paper does not offer a full analysis of the economic profitability of the message board information, as it fails to take into account some of the aforementioned trading costs. This Gu et al. [11] paper is particularly relevant to our paper because of the short time window and the fact that the message board data, without textual analysis, are similar to the data available from CAPS. In the present work, we expand on the approach used in the paper by Gu et al. [11], using many more stocks, a longer window, and more varied market conditions. In particular, we combine a mixture of experts' [14] strategy first proposed by Jordan and Jacobs with a strategy proposed by Fang and Yusuda [11], utilizing the rankings to invest in stocks.

The stars' opinion work by Fang and Yusuda [7] analyzes the relation between the reputation of stock analysts and the performance of their recommended positions (naively, one would expect a strong correlation). The efficacy of analyst recommendations, in general, is also considered, given the great importance of analyst ratings in the financial community. Fang and Yusuda [7] uses *Institutional Investor* magazine's All-American awards to provide an objective assessment of each analyst's reputation. These awards are based on surveys sent to all the top institutional investors in the nation. Each analyst can be first, second, or third place, as well as runner-up, or not ranked by the awards. The top-ranked analysts make up only 2 percent of the analyst population and they are considered stars.

Fang and Yusuda [7] are able to successfully catalog the relation between an analyst's reputation and the profitability of the analyst's recommendations. In particular, a simple strategy, based on the first- and second-place analysts' recommendations, is modeled, showing statistically significant risk-adjusted outperformance. Some transaction costs are taken into account, which, combined with the simplicity of the strategy, indicates that some analysts' recommendations contain economically profitable information. However, low-ranked and unranked analysts' recommendations did not exhibit statistically significant profitability.

Fang and Yusuda's [7] work is relevant to our paper because profitable predictions may be concentrated in only the very best pickers. It focuses on data over a short time period, from 1994 through 2002, with 10,000 analysts and 250,000 recommendations. We will call the investment strategy of Fang and Yasuda [7] the *starsop* investment strategy.

Investment Strategies

In all of our strategies, we need the ability to first rank users and second take positions (or invest) on a simulated portfolio based on the highest ranked users' stock picks. Our work discussed in the Method section until now (discussed in Section 3.2) will be called the custom-custom strategy, in which we have defined both our custom ranking strategy and our custom investment strategy. In addition, we develop baseline strategies informed by the work discussed in the previous section. We use the relevant literature above to inform both an additional expert ranking strategy—mixexp and an additional investment strategy—starsop. The names of these additional strategies were picked to reflect the titles of the papers that informed them. Stars are no different then experts, in that both terms stars and experts refer to the highest ranked users ranked by past performance.

In the next subsections, we will discuss how we combine the two different ranking strategies (custom, mixexp) with the two different investment strategies (custom, starsop) into trading algorithms, incorporated and compare the existing trading strategies [7, 11], adapted for this study, with our custom-custom algorithm. The flow of each algorithm is shown in Figure 3.

Mixture of Experts Ranking and Star's Investment—Mixexp-starsop

This algorithm is divided into two stages. The first stage is based on a mixture of experts approach [11, 14] and consists of ranking the users based on their past individual performance. The second investment stage is based on the stars' opinion paper [7] and uses the users rankings to take positions in a simulated portfolio.

For a given time T, the first stage—ranking—takes place over the year preceding T. Typically, T is chosen to coincide with the beginning of a calendar year, thus mimicking the *Institutional Investor* All-American awards cycle. Users are ranked according to a coefficient that represents the performance of their picks. This coefficient is the mean of a series of flags from throughout the ranking period. Each user's flags correspond to the user's picks. A flag is 0 if the pick is incorrect and 1 if it is correct. A pick is correct if the predicted stock movement (outperform or underperform) actually takes place X trading days later. X can vary, with the typical tested value being approximately 40 days, or 2 months.

The second stage—investing—takes place over the year following time T. Essentially, each pick a user makes is coupled with the user's ranking from the first stage, thereby generating a buy or sell signal for the corresponding stock, as well as determining the magnitude of said signal (high for top-ranked users, low for bottom-ranked). The validity of this approach was examined by considering the performance of the investments chosen by each percentile of users. Performance was calculated by simulating a stock portfolio's return through the end of the year. The portfolio was created by shorting underperform picks and buying overperform picks. Duplicate picks made by different users resulted in simply increasing the portfolio's stake in the corresponding

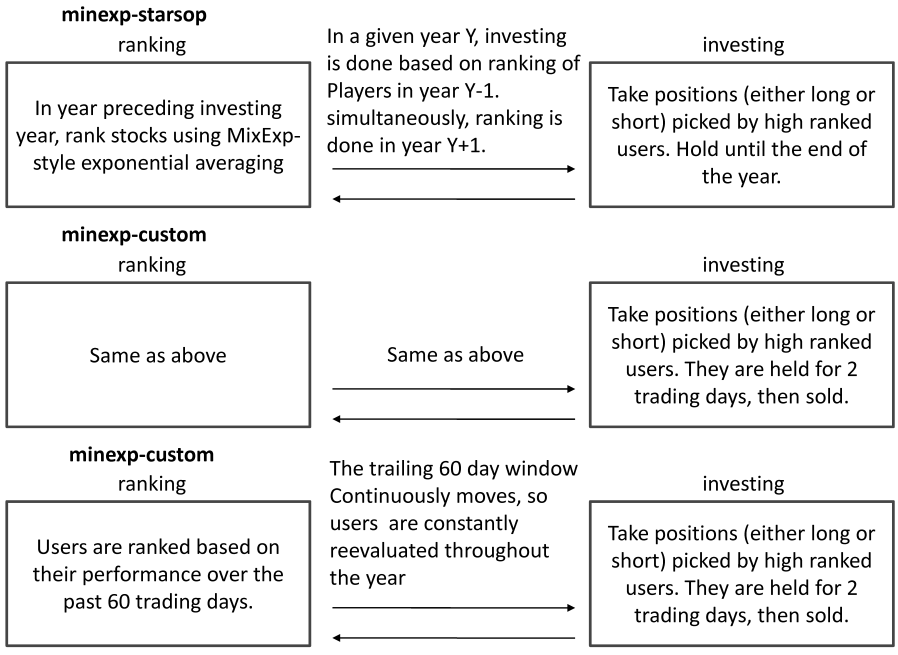

Figure 3. Flow of Three Algorithms

Notes: mixesp-starsop, mixexp-custom, and custom-custom that combine ranking strategies (mixexp and custom) with investment strategies (starsop and custom)

position (e.g., doubling for two identical picks). No trading costs were taken into account.

Custom CAPS Algorithm for Ranking and Investment—Custom-Custom

As discussed above in the Method section, in our custom algorithm, ranking and investing are temporally near, but are staggered relative to time *T*. Users again receive scores corresponding to their skill at picking positions. The score is a moving geometric mean that varies over the course of the testing period. The mean is composed of the returns of outperform picks and the inverse of returns for underperform picks, calculated over the course of two trading days directly preceding *T*.

A portfolio is simulated to gauge the effectiveness of these scores, where positions are taken at time *T* and held for one day. The positions taken correspond to the picks made on the day preceding *T* by the top *X* users, again with short positions corresponding to underperform picks and long positions corresponding to outperform picks. *X* can vary from 1 to 100. Duplicate picks again simply result in greater exposure to that position for the portfolio.

The fixed numerical values (2, 1) in this algorithm can theoretically vary. The specific values used were chosen for their simplicity, but higher performing values are likely possible.

Mixture of Experts Ranking and Custom CAPS Algorithm Investment-Mixexp-custom

This algorithm, a hybrid of the mixexp-starsop algorithm, uses the same ranking system as the Mixexp-starsop algorithm: users are ranked over the year prior to *T* using a mean of 0/1 flags. The algorithm then uses the custom investment strategy of the second algorithm, custom-custom, to best utilize these rankings. In particular, positions are taken for one day following each pick being made by a user, rather than through the end of the year like with starsop investment strategy. The results of the combined strategy are again gauged by the performance of the portfolio.

Custom CAPS Algorithm and Stars' Opinion Investment

This algorithm—whereby the custom ranking algorithm is combined with the stars' opinion investing algorithm—was not tested, as these two components are inherently incompatible. The stars' opinion paper's [7] investment strategy is based on a ranking methodology whose results are constant for the duration of the investment period. Additionally, its mean investment period is at least six months. The custom ranking algorithm, in contrast, is designed to run in staggered parallel with the investment algorithm, with both using relatively short duration cycles. The rankings change frequently, and each position is taken based on the best available rankings at the time. Thus, the integration of these two components is not feasible.

Results

We first trained our GA on picks from the period between August 1, 2007, and December 31, 2007, then validated and tweaked on January 1, 2008, and May 31, 2008, and the resulting strategies were tested between June 1, 2008 and December 31, 2008. This time-separated training/validation/testing split ensured the validity of the project's results—a measure of particular importance given the interactions between financial returns over different time periods. Because CAPS is a new product, the number of picks available increased greatly over this time period. August 2007 was chosen as the start date because the fitness functions required several months of prior data, and the body of picks was too small prior to mid-2007. During our test period, the S&P 500 was trending downward (see Figure 4). The S&P 500 total return over the time period (June 2, 2008, to December 24, 2008) was –35.5 percent with a Sharpe ratio of –0.077.

There were approximately 2 million user picks, which was sampled down to 591,581 (approximately 25 percent) per sample. In the test period when the market was trending downward, although it is counterintuitive, there were only 21 percent underperforming picks to 78 percent outperforming picks. Nonetheless, our final optimized strategy relied on both types of picks

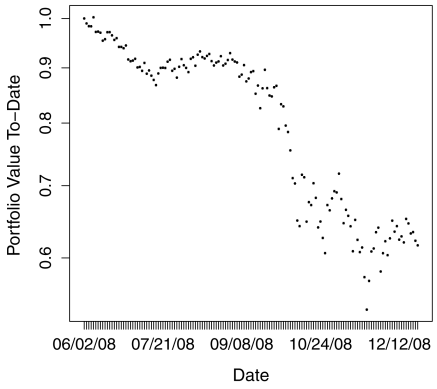

Figure 4. S&P 500 Trend for the 2008 Test Period

in about that same fraction. There were 84,917 picks made during the testing period by 25,364 unique users.

The remainder of this section proceeds as follows. We score the users based on past performance. We then rank the users. We build a model on the training data set that finds the best parameter values for the number of experts, portfolio holding period, user test period, and user test holding period. Note that we plot the Sharpe ratio for a different number of experts for a given parameter setting (to see how the number of experts might affect performance—even if that value is not selected as the optimal). We find that there is a bump in these charts, indicating that there is an optimal number of experts (indicated by a star for the optimal solution on training set) on which to base predictions—too many "experts" and the chart trends downward; too few and we do not have enough confidence in the stock votes. These trends are apparent in all data sets (see Figures 5 and 6).

Once we have the learned parameter set on the training data that includes the sweet spot for the number of experts, we apply it to the validation set to make sure we are not overfitting. We tweak the parameter values slightly, if necessary, on the validation set to perform well on both the training and validation sets. We then apply our model with the applied parameter values to the test set. We find that the optimized set of parameters outperforms our baselines. The three baselines we consider are (1) the performance of the S&P 500, (2) the performance we would get from letting the entire crowd vote (note that often the entire crowd performs poorly, but using a crowd of about 250 always does better than the S&P 500 in our tests), and (3) the performance we would get if we just used the last two picks of the users to assess their expert score (the idea being that we want to learn whether more history—enabling us to assess their expertise—is valuable). We pick the best number of experts on the training data and then apply that parameter set to the test data. We find that the optimized parameter sets significantly outperform our baselines so much that we run a significant number of sanity checks. Two worth noting are the following: (1) instead of ranking the users by their expert scores, we pick the users (experts) at random instead of ranking them by their past performance,

04 hill.indd 87 2/21/2011 11:23:18 PM

04 hill.indd 88

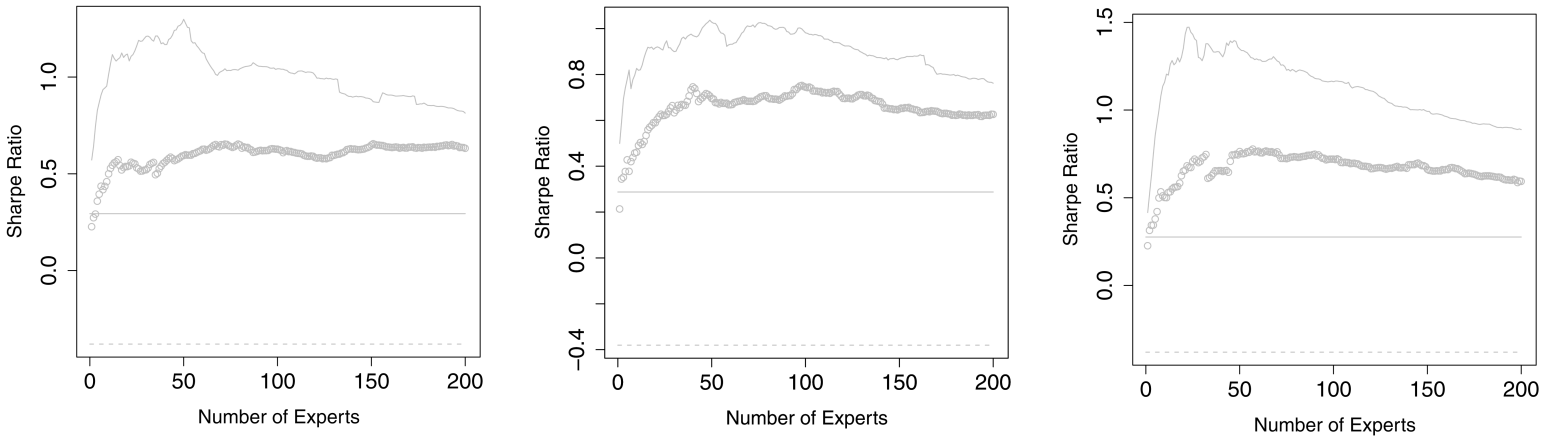

Figure 5. Number of Experts Versus Baselines for Train Data for 3 Samples (Validation Results not Shown).

Notes: The gray solid line at the top indicates the customized opt-CAPS parameter set using the custom-custom strategy on the training data. In each plot, the top solid gray line is for optimized parameter setting found for the parameters (A,B,C,D), then we vary the number of experts A. The circle points curve is for baseline parameter setting (A,B=2,C=1,D=1) where we vary the number of experts A. The solid horizontal straight line refers to the entire crowd model score baseline, and the dotted horizontal line corresponds to the S&P performance baseline.

2/21/2011 11:23:18 PM

04 hill.indd 89

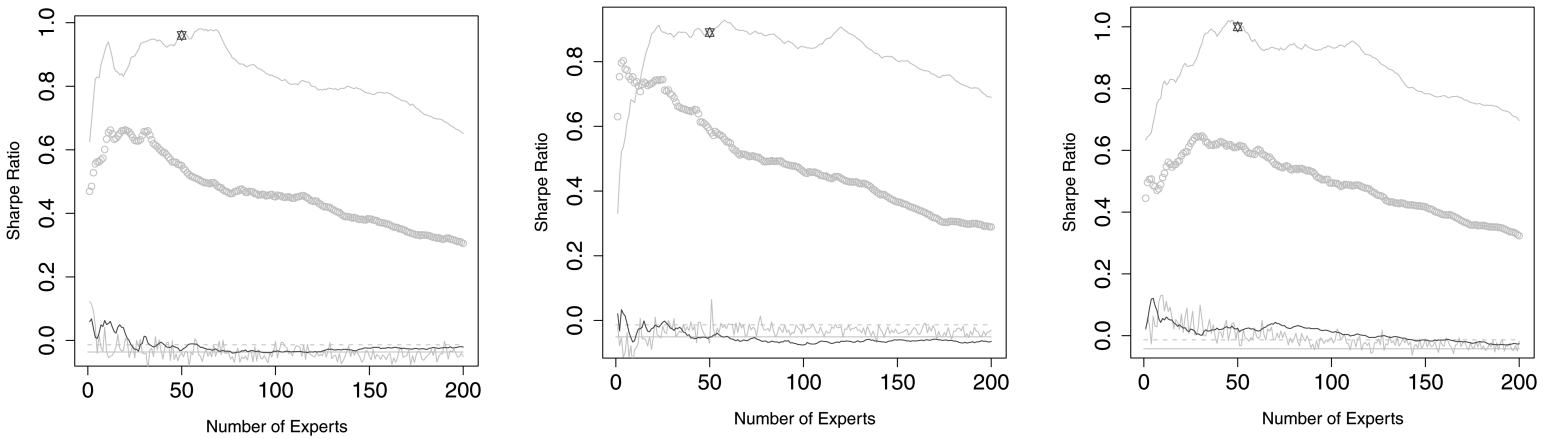

Figure 6. Opt-CAPS and Sanity Checks on the Test Set

Notes: The star on the solid curved line corresponds to the optimal parameter value solution, opt-CAPS, found using the custom-custom strategy. The remainder of the solid curved line is the result ofhen varying the number of experts, A. The circle points curve show results for the baseline parameter setting (A,B=2,C=1,D=1) and then varying the number of experts A. The solid horizontal line is the crowd score baseline and the S&P performance baseline during the time period is indicated by the dotted horizontal line. The dark jagged line at the bottom, refers to the baseline where we reassign all stock picks to users randomly. The gray jagged line at bottom, selecting the users at random (as opposed to ranking by the custom expert ranking strategy).

2/21/2011 11:23:19 PM

Table 2. Sharpe Ratio Statistics for the Investing Strategies and the S&P 500 over Three Samples During the Testing Period.

	Naive 2-1-1	Opt-CAPS	S&P	Crowd
Valid	0.88	1.24	0.02	0.23
Test	0.48	0.87	0.08	0.04

and (2) we randomly reassign all of the stocks to different user IDs to see if we still get the bump in the plot that shows there is a sweet spot for experts when we rely on their past performance. Both of these sanity checks (results are shown in Figure 6) perform dismally; furthermore, there is absolutely no pattern or shape to the plots, indicating again that ranking the users by their score has significant prediction value.

Baselines: "Let the Whole Crowd Vote" and "Random"

The whole-crowd approach fares poorly: –23.2 percent return and –0.101 Sharpe. However, the crowd outperforms the S&P 500 during both the training and validation periods and performs about the same as the crowd in the test period, suggesting that there is some information in the crowd. Note that if we use only a large subset of the crowd—about 250 users—we significantly outperform the S&P 500 but do not do as well as when we learn the best number of experts. For the expert baselines, we look at the scores of all the users making picks that day, then invest equally in all the stocks picked by the top N. We pick the best N on the training data and apply it to the validation and test. If we were to pick the average number of experts from the validation period and apply that to the three test periods, we would get an average Sharpe ratio of 0.48 (Table 2). We expect our optimized method needs to beat this naive voting strategy. Instead of ranking the experts, we pick the number of "experts" at random (as opposed to ranking them by prior performance), which is indicated by the dark jagged solid line at the bottom of the plots in Figure 6. In addition, we randomly reassign all of the stock picks to different users (indicated by the light gray jagged line in Figure 6). These sanity checks enable us to see that there is indeed some value both in the true expert rankings and in using the experts together as the crowd.

Custom Expert Stock Picker: Optimized Results

By our definition, experts are users who are scored highly with respect to the number of correct predictions in the past. Users are scored by looking at the past C trading days' worth of picks, and then getting the product of all their picked stocks' returns during that period, with each return calculated using a 1-day holding period. (If a pick expects a stock to underperform, the inverse of the stock's return is used in the product calculation instead.) Thus, 50 experts

are just the 50 users with the highest scores. This scoring methodology favors prolific scorers. We are investigating changing the scoring slightly so that it uses the average return over the past *C* days instead. Note that the top expert is the top expert for each day, not for all time, so it is possible that we are following a different user every day (we have noticed some cases in which one user has a few multiday streaks, however).

For a set of parameters (A,B,C,D), those found by our optimized approach, we invest equally in every stock picked that day (long if the stock was predicted to outperform, short if predicted to underperform). If there are 2+ picks for a given ticker, we weight the portfolio. The GA settles on a set of input parameters (average across three training samples) to the fitness function as follows: A = number of experts: 42.76; B = portfolio holding period: 1.89; C = user test period: 60.748; D = user test holding period: 1.

These parameters ultimately proved to be quite successful during the testing period. The optimized investment strategy had an average Sharpe ratio of 0.87, which contrasts very favorably with the S&P 500's average of 0.08 across the test samples. The statistics are shown in Table 2 for the crowd model and the naive expert model as well.

In addition to the test results reported above, we made predictions for 2009 to make sure we could outperform the S&P 500 on another test time period. In Figure 7, we plot the Sharpe ratios and returns for 2008 and 2009. In 2009, the economy began to rebound, with the S&P 500 gaining about 23%. However, this also makes it a fairly unusual year, relative to more typical market conditions. Still, our approach outperforms the S&P 500.

In effect, the custom-custom strategy was able to deliver both higher returns and lower risk, thereby creating the elusive "alpha." In finance terms, this is profitability above and beyond what could be expected given the riskiness of the project. The concept is important because, as the popular capital asset pricing model (CAPM), used to calculate the required rate of a return on an asset given its risk, hypothesizes and empirical research supports, simply allocating resources to exceptionally risky projects also can result in high mean profitability. However, in those cases, the high profitability is simply the reward you are given for undertaking such a risky and unsavory project. Over this time period, one or more of our strategies has an alpha of 28.12 percent annualized, which compares very well to top hedge fund alpha values. In the next section we present results comparing our results to the baselines identified in the literature and find we compare well.

More Baselines: Custom Expert Stock Picker Compared with Competing Methods

In Figure 8 we present the performance of the custom-custom approach for 2008, 2009, and the combined time period 2008–9, as well as the performance of all of the other strategies. The bar chart indicates that the way we rank experts makes a tremendous difference in terms of the performance of our stock picks. Indeed, the custom-custom approach offers significant and profitable outperformance. The S&P 500 numbers are also included for reference.

04 hill.indd 92

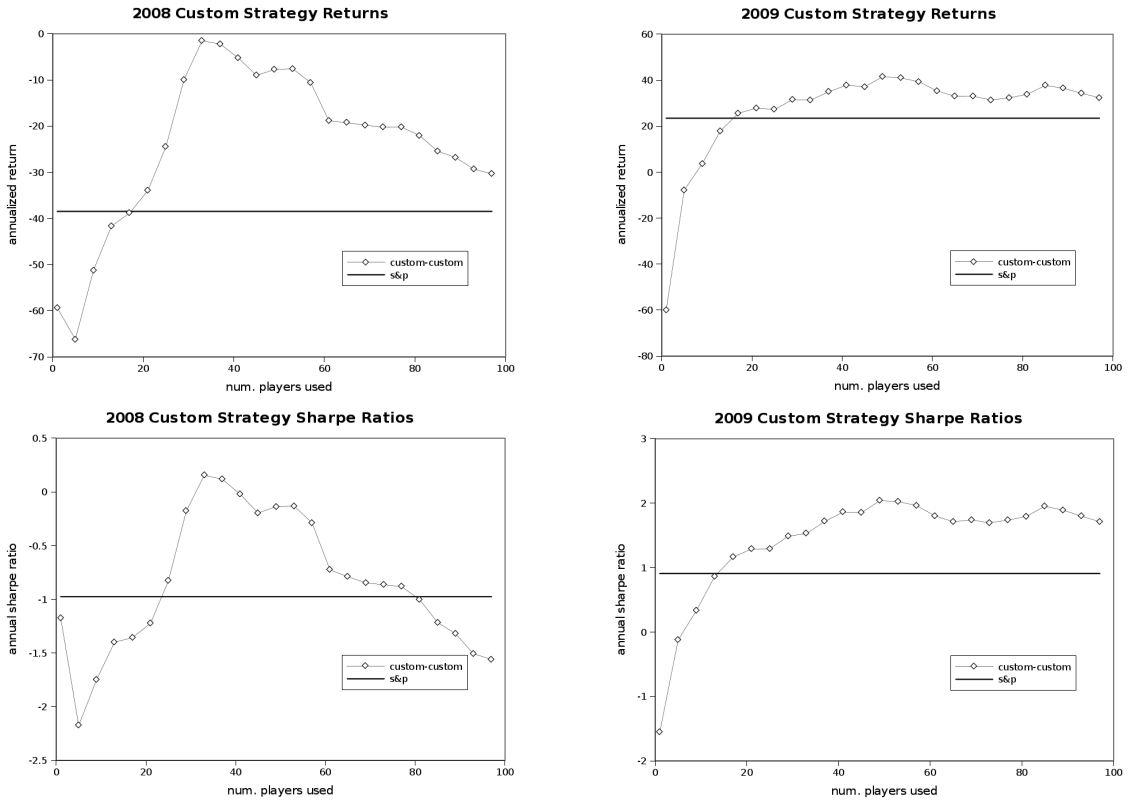

Figure 7. Comparison of Our Custom-Custom Approach to S&P 500 for Both 2008 and 2009 for Both Annualized Returns and Sharpe Ratios

Note: In the limit, the ranked experts converge to approximately the S&P 500.

2/21/2011 11:23:19 PM

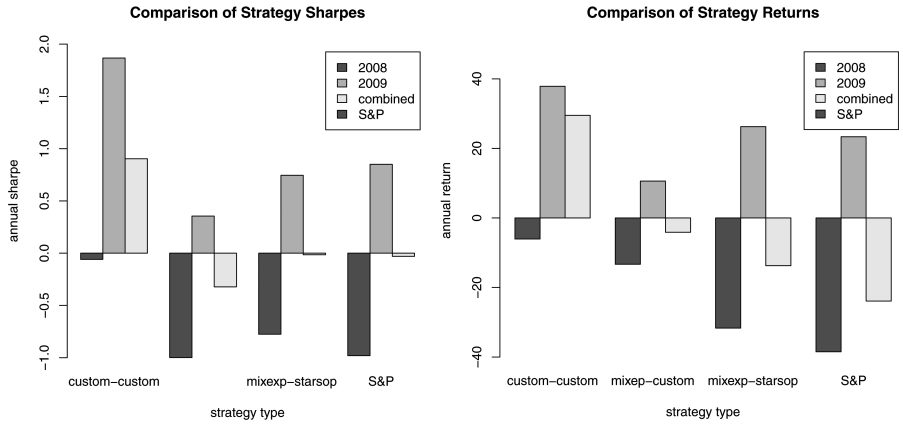

Figure 8. Comparison of Custom-Custom, Mixexp-Custom, Mixexp-Starsop Strategies, as well as the S&P 500, for Both Annual Sharpe (left) and Annualized Return (right)

Notes: The custom-custom strategy uses the average of several different strains of the genetic algorithm, with the number of users included varying from 39 to 47, as was found optimal in our earlier testing. The mixexp strategies use only the top 2% of users, as was found optimal when examining analysts [7].

Mixexp-starsop. This approach generally was not able to generate substantial outperformance. There may have been a slightly stronger signal for the short positions, a trend supported by research in the stars' opinion paper [7] as well. There was also substantial variation in the returns for each percentile of users. However, given that there was no noticeable upward trend as the hypothetical skill of users increased, these variations can be attributed to noise.

Mixexp-custom. This algorithm uses the same ranking system as the mixexp-starsop algorithm: users are ranked over the year prior to T using a mean of 0/1 flags. The algorithm then uses the custom investment strategy to best utilize these rankings. In particular, positions are taken for one day following each pick being made by a user, rather than through the end of the year. The results of the combined strategy are again gauged by the performance of the portfolio.

The results of this approach were generally less promising than those of the custom-custom algorithm, but more than those of the mixture of experts' and stars' opinion approach (mixexp-starsop). In addition, the noise visible in the mixexp-starsop algorithm's results was reduced. However, the additional gains from the incorporation of the custom investment approach may be unrealistic, given the higher trading frequency and the fact that trading costs were not built into the model.

In the mixture of experts' and stars' opinion derivative, the experts are chosen too long ago for their picks to make much difference. Similarly, the mean holding period of their picks (6 months) is so great that the noise overcomes the signal. If we model each day's return as an independent, identically

distributed random variable, the standard deviation (SD) of the cumulative return of a sequence will increase as the sequence gets longer—specifically, it will increase with the square root of the length. Thus, the SD for the 6-month holding period (and 12-month delay) is much greater than the SD for the several-day sequences relevant in our old methodology. And the high SD causes the returns for the stars' opinion method to seem essentially random, with respect to our calculated expertness of the users.

This could explain the disparity between the approaches derived from the Star's opinion [7] and Mixture of Experts [11, 14] papers and the purely custom approach, but it leaves open the question of why the stars' opinion investment technique [7] works with real-world analysts and not with online users voting on stocks at leisure. Perhaps the analysts are simply much better than the users, and if we had used the custom methodology with analysts, we would have seen even more of a hump. The users might also, for some reason, be using a shorter time horizon for their picks.

In addition, the custom approach is gauged by examining performance when using a variable number of users, from the single best up to the top 100. Percentiles, in contrast, were used for testing the derivative users. Given that several hundred players were being considered at most times, the use of percentiles forces a much coarser analysis. Further research will examine this possibility.

Investigation of the Experts. The results presented in the preceding section indicate that our approach outperforms the S&P 500 and two approaches informed by the extant literature. In this section we examine the frequency and consistency over time of the experts selected by our approach.

We looked at experts who were picked over the course of 2008–2009 combined, using 50 experts per day, as this was around the "sweet spot" that we found to be the "optimal" set of experts in our study. If we rank the experts by how often they appear in the experts list, and then plot the frequency of occurrence on the horizontal axis versus the rank, we get a distribution that resembles a power law distribution. If we plot on log-log scale, we get something close to a straight line, with slope of –0.8 (Figure 9).

In addition to the highly ranked experts being ranked high in many time periods, the highly ranked experts make a lot of picks. Specifically, in our data set, the experts make almost 10 times as many picks as nonexperts—on average 302.217 picks per user versus 31.737 picks for nonexperts.

If we were to take all of the experts identified by our algorithm and rank them in descending order by the number of times they appeared as an expert, we find that the experts who appear more often (the more expert experts) also make more picks than the less frequently appearing ones, on average (see Figure 10). On visual inspection, the shape of the distribution looks like a very noisy power law distribution.

Intuitively, we think we would like our "expertness" division to be clean—people are good, or bad, with no variation. Thus, we would like all of our experts to be picked roughly the same number of times—a distribution with no head and a really fat tail. More concretely, if experts appeared randomly in our ranking (i.e., for each expert there is a k/n chance of a user appearing

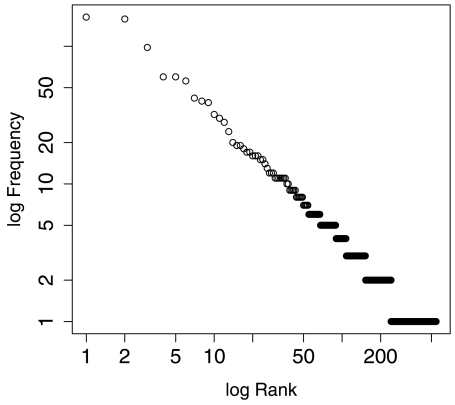

Figure 9. Experts Ranked by the Number of Times They Are Identified as an Expert by Our Custom-Custom Algorithm in 2008–2009

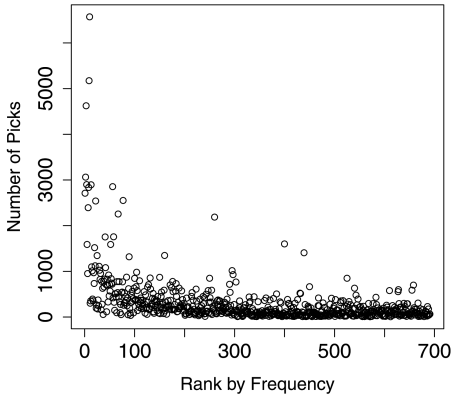

Figure 10. Experts Ranked by Number of Picks

on a given day, where n is the total number of experts and k is the number of experts chosen that day), then overall we would expect to see a normal distribution for the frequency of experts being chosen for a given day. The fact that we see a distribution that follows a power law means the distribution of expertise is not random.

Given that the highest ranked experts seem to be making more picks, we wanted to make sure that it is not just that the most prolific users are doing the best. A strategy of just looking at the top 50 most prolific users does not perform very well compared with either our approach or the S&P 500. We based our results on the top 50 users ranked by overall number of picks, with a holding period of 3 days (very similar to previous settings). There are a number of additional tests we could perform to try to understand the experts. For example, we did not observe any malicious spamming activity. But perhaps removing spamming would increase results. We also assume that users are

independent. But this is certainly not the case, as users have the ability to view one another's picks. Looking at the extent to which users appear to influence one another may be explored in the future.

Discussion

In this research, when testing our first hypothesis that using the stock picks of a large sample of online users from Motley Fool CAPS enables us to outperform the S&P 500, we find that letting large samples of users selected from the entire crowd vote often outperforms the S&P 500 in our test data. For example, in the second half of 2008, the overall crowd return was –23.2 percent compared to –35.5 percent for the S&p500. We find similar results for 2009. When we tested our second hypothesis, that our expert ranking algorithm will enable us to find better stock portfolios than the S&P 500 and entire crowd, however, we found that our approach for ranking experts in online user generated stock votes helped to identify better stock portfolios than the S&P500 as well as the portfolios based on the entire crowds picks 100 percent of the time.

The punch line: the "wisdom of crowds" seems to be proven in this data set; however, identifying a subset of experts in the crowd enables us to perform much better. Our approach to rank individuals to identify the few experts to use for prediction is what separates our work from that explained in a working paper we found that considers a very similar data set sourced from CAPS [3]. In our work, we focus on picking the best subset of voters as opposed to only analyzing what happens when the entire crowd votes and the aggregated proprietary star ratings of stocks from CAPS are used. The star ratings indicate stocks that voters in aggregate have picked to outperform the market. In prior work, the authors find the CAPS proprietary star ratings are also quite successful at identifying a good set of stocks [3] and therefore their results are consistent with our findings. Our method allows us to gain confidence in an individual experts' votes by evaluating their historical vote performance, as well as confidence in the stock picks by having many people vote with more experts—thereby enabling us to pick the optimal crowd.

In Figure 11 we see that when we use only 1 expert, relatively few stocks are voted on more than once, in contrast to when we use 250 experts. There is a tension between the two—experts and crowd—however, because ultimately even though a crowd of size 250 outperforms the S&P 500 but performs less well than the optimized strategy, the entire crowd should perform only about as well as the S&P 500. We plan to explore this tension with more experiments and with simulation. In addition, we would still like to rework the user scoring functionality once we have a better definition of expert—currently, we take the product of the voter's performance over time, which has the effect of greatly favoring the most prolific pickers.

In a real-world setting, this strategy likely would be integrated with other quantitative investing strategies. Given that this is applied machine learning research, it is important to consider the results in a business context. To that end, several important factors related to how these results would translate into real money are summarized below. We invest in a large number of stocks,

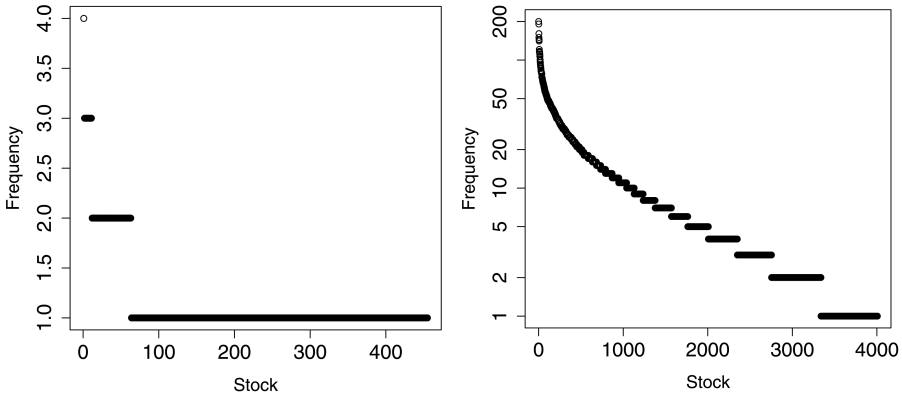

Figure 11. Frequency of Votes per Stock When 1 (left) Versus 250 (Right) Experts Are Used to Vote on a (1, 2, 1, 1) Strategy and (250, 2, 1, 1) Strategy, Respectively

frequently in the range of 50–100. The long-short split is close to the overall over/under split, which is to say that about one-fourth to one-fifth are long and the rest are short. The combination of a moderately large short component and a large number of stocks with a very short rebalancing period (typically 1 day, of course, and sometimes up to 3 or 4) means that transaction costs would be significant in practice.

As the holding period is fairly short, there are also tax implications. A significant component of the strategy is short, which likely will result in borrowing fees from the brokerage. There may be insufficient liquidity or volume in some of the securities traded. This strategy is tested in a universe of 7,142 stocks. Given that the S&P 500 often is considered a good approximation of the market, this universe clearly includes a large number of microcap stocks. Depending on the size of the portfolio and the specific positions selected, the markets for some of these securities may not be large enough to support the level of trading that the strategy relies on. These caveats certainly do not invalidate the results shown; they merely serve to remind us that there may be important differences between theory and practice.

Evaluation of Results

Even more fundamental than implementation costs, it is important to consider whether the research is actually valid. For example, overfitting is a constant concern in statistics and data mining, and it is common—although often forgotten—knowledge that "past performance is no guarantee of future returns." Some of the areas in which this project is weak are as follows: the time period covered by the data set is very short. Analysis is done only on approximately 16 months of pick data, with testing comprising only 5 months. Any financial strategy should be back-tested far more—ideally 5 or 10 years, at least. Similarly, during our test period, the market was abnormal. In hindsight, 2007–9

were very unusual years, both in terms of volatility and actual returns. Thus, a strategy's outcome during this period may not be indicative of its future performance. Despite the limitations, we took care to spend a substantial portion of the project attempting to ensure that the results are valid.

Why does this approach work? Do users have more information or expertise than investors in general? One explanation could be that experts go to the Motley Fool CAPS site. To find evidence of this, we looked at the extent to which CAPS visitors are different (in terms of demographics) from the average Internet user (of course, a better comparison would be the average investor). We cannot claim that the visitors are experts. But they are different from the average Internet user on many demographic dimensions, as shown in Figure 12.

Future Work

As is to be expected, in many ways this research prompts more questions than it answers. Most critically, will the results stand up with more data? Although every effort has been made to validate these results, their weakest point is the short time period in which they were created. Because testing stopped in December 2009, there are now additional data with which to further test the GA's results. What would be the effect of some type of rolling optimization? For example, what if, rather than using the training period to create one canonical set of parameter values, the system instead trained over months 1 and 2, and then used the results of the training in month 3? There is some evidence that the closer temporal proximity between the training and testing periods results in better returns, and this technique potentially takes advantage of this effect.

Would weighting the positions improve returns? It reasonably may be assumed that the algorithm is more "sure" that the first stock is a better long bet than the fifteenth stock, although both positions will be taken. Thus, it may make more sense to place a greater fraction of the portfolio on the 1st stock's position. Is momentum in the number of picks at all significant? Although the total number of picks from day to day does not change rapidly, the number of picks on any given stock does vary substantially. It is possible that there are additional data contained in this behavior that the strategy may be able to take advantage of. Naturally, it would be advantageous to consider incorporating other variables, such as stock fundamentals or price momentum, as well. We plan to incorporate traditional variables in future work. We also plan to apply this approach to other data sets (Digg and YouTube) to predict hits. Finally, we cannot ignore the behavioral aspects of this study—we need a better understanding of what motivates people to use the Motley Fool CAPS site and try to get the picks right. In addition, we will consider that the votes of one user may influence the votes of others. The votes not only may be used as online word of mouth [4] but also may influence [23] trust in a particular stock.

04 hill.indd 98 2/21/2011 11:23:21 PM

Figure 12. Statistics on Motley Fool CAPS Users' Education, Gender, and Income Taken from Quantcast

Note: Statistics for caps.fool.com are taken directly from Quantcast (www.quantcast.com/caps.fool.com#demographics).

04 hill.indd 99
2/21/2011 11:23:21 PM

NOTE

1. We contacted the copyright office at Motley Fool to verify the CAPS data could be used for academic research.

REFERENCES

1. Amatriain, X.; Lathia, N.; Pujol, J.M.; Kwak, H.; and Oliver, N. The wisdom of the few: A collaborative filtering approach based on expert opinions from the Web. In *SIGIR'09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* Boston: ACM Press, 2009, pp. 532–539.
2. Antweiler, W., and Frank, M.Z. Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance, 59,* 3 (2004), 1259–1294.
3. Avery, C.; Chevalier, J.; and Zeckhauser, R. The "CAPS" prediction system and stock market returns. HKS Faculty Research Working Papers Series RWP09-011, John F. Kennedy School of Government, Harvard University, 2009.
4. Cheung, M.Y.; Luo, C.; Sia, C.L.; and Chen, H. Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce, 13,* 4 (2009), 9–38.
4. Dhar, V.; Chou D.; and Provost, F. Discovering interesting patterns for investment decision making with GLOWER—A genetic learner overlaid with entropy reduction. *Data Mining and Knowledge Discovery, 4,* 4 (2000), 251–280.
6. Dhar, V., III, and Chang, E. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing, 23,* 4 (2009), 300–307.
7. Fang, L.H., and Yasuda, A. Are stars' opinions worth more? The relation between analyst reputation and recommendation values. Working paper, University of Pennsylvania, Wharton School, INSEAD, 2009.
8. Foutz, N.Z., and Jank, W. The wisdom of crowds: Pre-release Forecasting via functional shape analysis of the online virtual stock market. Paper presented at the Marketing Science Conference, Singapore, 2007.
9. Givoly, D., and Lakonishok, J. The quality of analysts' forecasts of earnings. *Financial Analysts Journal, 40,* 5 (1984), 40–47.
10. Golub, B., and Jackson, M.O. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics, 2,* 1 (2010), 112–149.
11. Gu, B.; Konana, P.; Liu, A.; Rajagopalan, B.; and Ghosh, J. Predictive value of stock message board sentiments. McCombs Research Paper No. IROM-11-06, University of Texas at Austin, 2006.
12. Hendricks, K.B., and Singhal, V.R. The long-run stock price performance of firms with effective TQM programs. *Management Science, 47,* 3 (2001), 359–368.

13. Hirschey, M.; Richardson, V.J.; and Scholz, S. Stock-price effects of internet buy-sell recommendations: The Motley Fool case. *Financial Review, 35,* 2 (2000), 147–174.
14. Jordan, M.I., and Jacobs, R.A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation, 6,* 2 (1994), 181–214.
15. Julià, C.; Sappa, A.D.; Lumbreras, F.; Serrat, J.; and López, A. Predicting missing ratings in recommender systems: Adapted factorization approach. *International Journal of Electronic Commerce, 14,* 2 (2009), 89–108.
16. Kittur, A.; Chi, E.; Pendleton, B.A.; Suh, B.; and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web Internet and Web Information Systems, 1,* 2 (2007), 1–9.
17. Kittur, A., and Kraut, R.E. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work.* New York: ACM Press, 2008, pp. 37–46.
18. Schumaker, R., and Chen, H. Textual analysis of stock market prediction using breaking financial news: The AZFIN text system. *ACM Transactions on Information Systems, 27,* 2 (2009), 1–19.
19. Seyhun, H.N. *Investment Intelligence from Insider Trading.* Cambridge: MIT Press, 1998.
20. Seyhun, H.N. Why does aggregate insider trading predict future stock returns? *Quarterly Journal of Economics, 107,* 4 (1992), 1303–1331.
21. Sharpe, W.F. The Sharpe ratio. *Journal of Portfolio Management, 21,* 1 (1994), 49–58.
22. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* New York: Doubleday, 2004.
23. Utz, S.; Matzat, U.; and Snijders, C. On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *International Journal of Electronic Commerce, 13,* 3 (2009), 95–118.
24. Verna, P. User-generated content: More popular than profitable. *eMarketer.com* (January 2009) (available at www.emarketer.com/Report.aspx?code=emarketer_2000549).
25. Wuthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; and Zhang, J. Daily stock market forecast from textual Web data. In *1998 IEEE International Conference on Systems, Man, and Cybernetics*. Los Alamitos, CA: IEEE Computer Society Press, pp. 2720–2725.

SHAWNDRA HILL (shawndra@wharton.upenn.edu) is an assistant professor of operations and information management at the Wharton School of the University of Pennsylvania. Her research is funded in part by the Office of Naval Research, WPP and Google, and the National Institutes of Health (NIH). Dr. Hill holds a B.S. in mathematics from Spelman College, a BEE from the Georgia Institute of Technology, and a Ph.D. in information systems from NYU's Stern School of Business. Her recent work appears in *IEEE Transactions on Data and Knowledge Engineering, Journal of Computational and Graphical Statistics, SIGKDD Explorations,* and *Statistical Science.*

NOAH READY-CAMPBELL (noah.readcamp@gmail.com) is an associate product manager at Google, working in the Local Ads group. He recently graduated from the

University of Pennsylvania, with MSE and BSE degrees in computer science and a B.S. in economics. He has won several awards, including the 2010 Wharton Undergraduate Research Award and one of the 2009 Wharton Venture Awards.