Characterizing and Aggregating Agent Estimates

H. Van Dyke Parunak
Soar Technology
3600 Green Court
Suite 600
Ann Arbor, MI 48105
+1 734 887 7643
van.parunak@soartech.com

Sven A. Brueckner
Soar Technology
3600 Green Court
Suite 600
+1 734 887 7642
sven.brueckner@soartech.com

Lu Hong
Loyola University Chicago
1 E. Pearson
Suite 204
Chicago, IL 60611
+1 312 915 7067
lhong@luc.edu

Scott Page
University of Michigan
317 West Hall
Ann Arbor, MI 48106
+1 734 647-9193
spage@umich.edu

Richard Rohwer
SRI
9988 Hilbert St
Suite 203
San Diego, CA 92131
+1 858 527-1406
richard.rohwer@sri.com

ABSTRACT

In many applications, agents (whether human or computational) provide estimates that must be combined at a higher level. Recent research distinguishes two kinds of such estimates: interpreted and generated data. These two kinds of data require different kinds of aggregation processes, which behave differently from an information geometric perspective: interpreted estimates require methods such as voting that can leave the convex hull of the individual estimates, while the optimal aggregation for generated estimates lies within the convex hull and thus is accessible by methods such as weighted averages. We motivate our analysis in the context of a crowdsourced forecasting application, demonstrate the central insights theoretically, and show how these insights manifest themselves in actual data.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence]:** Distributed Artificial Intelligence – *Multiagent Systems*
H.3 [**Information Storage and Retrieval**]: Miscellaneous
J.4 [**Social and Behavioral Sciences**]: *Psychology*

General Terms

Algorithms, Experimentation, Theory

Keywords

Information fusion, agent characterization

1. INTRODUCTION

In many applications, agents (human, software, or hybrid) provide estimates that are combined in various ways to yield a common outcome. Examples include electronic markets, voting systems, product reviews, and crowdsourcing applications in general.

Techniques to understand such estimates are of increasing interest at AAMAS. In 2012, for example, [1] developed metrics to characterize the IQ of a crowd of humans. [13] studied ways to route a prediction task among human experts. [7] showed how human and machine reasoners could be combined in crowdsourcing. [8] described how to incentivize people to provide truthful answers in crowdsourcing. Our research contributes to this growing area of interest to the AAMAS community.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

Agents can produce their estimates in various ways. Two main approaches have been distinguished [6]. An estimate is said to be "generated" if the agent samples it from a distribution, perhaps with the addition of idiosyncratic error. The classic example of a generated estimate is the report of temperature provided by a thermocouple. An estimate is said to be "interpreted" if the agent derives the estimate from (a subset of) attributes associated with the object or event on which an estimate is being solicited. A rule-based agent would produce an interpreted signal, since it responds only to attributes of the problem represented on the left-hand side of its rules, and different estimates reflect agents with different sets of rules, attending to different attributes. One can imagine algorithms for both humans and computational agents that fall into either of these categories.

Estimates might combine features of both kinds of process. Our results suggest statistical indicators that reflect whether estimates are mostly generated or mostly interpreted. For clarity, we focus on purely generated and purely interpreted estimates.

The distributions of estimates that result from these two kinds of processes have very different statistical properties. This paper discusses a characterization that has not been observed previously. For simplicity, assume that the estimates being solicited are multinomial probability distributions. Each estimate is a point on the simplex of the appropriate dimension. For example, Figure 1 shows three forecasts against a three-outcome question. The dashed line is the simplex, which in this case is the space of all triples (p_1, p_2, p_3) such that $p_i \in [0, 1]$ and $\sum p_i = 1$. *a* assigns 100% to the first outcome and nothing to the other two. *b* assigns 50% each to the first and third outcomes, and nothing to the second. *c* is the uniform forecast, assigning 33.3% to each outcome. These three forecasts define a convex hull, shown as a solid line.

Our central result is that, under fairly benign constraints, the best aggregate summary of a set of generated estimates lies within the convex hull of the individual estimates, while the best aggregate summary of a set of interpreted estimates can leave the convex hull. The most common aggregation methods (such as a weighted average) are constrained to the convex hull, and so are suboptimal for interpreted estimates.

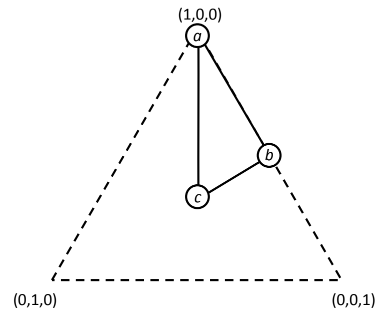

Figure 1: Three forecasts on the trinomial simplex.

Section 2 summarizes the specific application that motivates this research, and in which it is being applied. Section 3 demonstrates our central result, by example and proof, and provides technical qualification of our claim. Section 4 discusses the applicability of this result to real data (presumably interpreted) from our application, and tests our conclusions against synthetic, generated data. Section 5 discusses directions for future research. Section 6 concludes.

2. APPLICATION CONTEXT

This research, under the ongoing IARPA ACE program,[1] is motivated by the problem of how to aggregate forecasts about world events provided by multiple forecasters. Each forecaster reports her estimate of the relative probability of alternative outcomes for a number of questions about international affairs. Sample questions include:

1. Will Bashar al-Assad remain President of Syria through 31 January 2012? (a binomial question, with possible outcomes "Yes" and "No")
2. Will a run-off be required in the 2012 Russian presidential election? (a binomial question, with possible outcomes "Yes" and "No")
3. Who will be inaugurated as President of Russia in 2012? (a multinomial question, with possible outcomes "Putin," "Medvedev," and "Neither")

Our research focuses on how to aggregate multiple forecasts against such a question to increase the accuracy of the aggregate forecast over many such questions (over 150 so far).

The most obvious way to combine estimates from different forecasters is to average them together. The process of giving each forecast equal weight is called an "Unweighted Linear Opinion Pool," or ULinOP. In terms of the geometry of Figure 1, the ULinOP is the centroid of the forecasts, and necessarily falls within their convex hull.

The next level of refinement, and perhaps the most common in the forecasting community, is to apply a different weight to each forecast in seeking a more accurate aggregation. These weights are functions of various features, which may be features of the particular question, the particular forecaster, or the specific forecast from the forecaster that is being weighted. As long as the weights are positive, weighted averages are also constrained to the convex hull of the forecasts.

Another approach, and a motivator for this research, is to allow forecasts to vote for outcomes. Let *i* be the outcome favored by a given forecast. Then that forecast casts a vote for outcome *i*. The vote may be constant across all forecasts, or it may be a function of features. The accumulated votes for each outcome are then normalized by the total votes across all outcomes to yield the aggregated forecast. Unlike averages (whether weighted or not), this algorithm can yield an aggregate outside of the convex hull.

We evaluate a forecast by its Brier score [3]. Let f_{ai} be a forecast from aggregation method *a* on outcome *i* of a given question, and let the actual outcome of the event among *N* possible outcomes be outcome *j*. The Brier score assigned to method *a* is

Equation 1 $$b_a = \sum_{i=1}^{N}\left(f_{ai} - \delta_{ij}\right)^2$$

where δ_{ij}, the Kronecker delta, has value 1 if $i = j$ and 0 otherwise. The Brier score lies between 0 and 2, and its value for a uniform distribution, $b_u = (N - 1)/N$, varies with *N*. We prefer to work with a normalized and centered inverse Brier score ("accuracy"),

Equation 2 $$\beta_a = \begin{bmatrix} \mathrm{IF}(b_a \le b_u): & \frac{b_u - \frac{b_a}{2}}{b_u} \\ \mathrm{ELSE}: & \frac{b_a - 2}{2b_u - 4} \end{bmatrix} \in [0,1]$$

which assigns 1 to a perfect forecast, 0 to a complete miss, and takes the value 0.5 for a uniform distribution, independent of the number of possible outcomes.

3. THEORY

In this section, we give both simple examples and proofs distinguishing the aggregation process for interpreted and generated estimates.

3.1 Interpreted Estimates

Consider the following forecasting environment with interpreted signals. Assume there exists a dependent variable *y* that equals the sum of $2N + 1$ independent attributes,

Equation 3 $$y = \sum_{i=1}^{2N+1} x_i$$

Let each attribute $x_i \in \{0, 1\}$ with equal probability. Let the outcome be good if $y \ge N + 1$ and bad otherwise. Partition the $2N + 1$ attributes into two sets *S* and *U*, where those attributes in *S* can be seen and those in *U* cannot be seen. Let $M = |U|$. Assume that an agent *j* sees some subset of attributes $K_j \subset S$ and makes its prediction based on the values of those attributes.

In this section, we show that in some cases, the optimal aggregate prediction will not lie in the convex hull of the individual prediction. This result may seem surprising, but recall [6] that interpreted signals will be negatively correlated when there exists no overlap in the sets of attributes they include. That negative correlation will prove sufficiently large to drive the result.

We first consider a simple example where $N = 2$, so that there exist five attributes, and the outcome is good if at least three attributes are 1. Let $S = \{x_1, x_2, x_3\}$. Assume that $x_1 = x_2 = 1$ and $x_3 = 0$. For the outcome to be bad, both attributes in *U* must be 0, an event with probability ½ * ½ = ¼, so the true probability of a good outcome is 1 – ¼ = ¾. Suppose that there exist three agents indexed by *i* and that agent *i* looks at attribute x_i. Agents 1 and 2 each predict a good outcome with the probability that at least two of the attributes that they cannot see are 1. The probability of any given configuration of the other four attributes is $(1/2)^4 = 1/16$, and the number of configurations that would yield at least two additional 1's is $\binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 11$, so forecasters 1 and 2 predict a good outcome with probability 11/16 < 3/4. Forecaster 3 needs at least three of the four attributes that she cannot see to be 1 to predict a good outcome, to which she assigns probability $\left(\binom{4}{3} + \binom{4}{4}\right)/16 = 5/16 < 3/4$. The optimal forecast of ¾ lies outside the convex hull of the predictions. If we extend the example and consider a large number of forecasters, we still obtain that each forecaster will make a prediction of either 11/16 or 5/16, and the optimal prediction lies outside the convex hull.

We now state some preliminary results for this base model. They almost surely extend to cover more general cases.

For our first claim, we show that if forecasters look at single attributes, then this result always holds.

[1] http://www.iarpa.gov/solicitations_ace.html

***Claim 1*:** Suppose that each forecaster randomly chooses a single attribute from *S*. Then the optimal prediction lies outside the convex hull of the forecasts.

Proof: Without loss of generality, assume that the probability that the outcome is good exceeds 50% given information about all attributes in *S*. It suffices to show that all forecasters predict good outcomes with less than the probability conditional on knowing the value of all attributes in *S*. Any forecaster who sees an attribute with a value 0 predicts a good outcome with probability less than one half (*N*+1 of the remaining 2*N* attributes would have to be 1), and trivially satisfies the condition. Next consider a forecaster who sees an attribute with value 1. That forecaster predicts a good outcome with probability equal to the probability that at least *N* other attributes have value 1:

Equation 4 $P_1 = \sum_{i=N}^{2N} \binom{2N}{i} \left(\frac{1}{2}\right)^{2N}$

By assumption, *at least* one more attribute in *S* takes on value 1 than takes on value 0, so the probability P_S of a good outcome conditional on knowing the values in *S* is at least as large as when *exactly* one more attribute takes on value 1 than takes on value 0. That lower bound requires that at least half of the *M* = |*U*| unseen attributes take value 1.

Equation 5 $P_S \geq \sum_{i=\frac{M}{2}}^{M} \binom{M}{i} \left(\frac{1}{2}\right)^{M}$

Given *M* < 2*N*, $P_S > P_1$, and the result follows.[2]

We can now state a corollary to this claim.

Corollary 1: Suppose that each forecaster randomly chooses *k* attributes from *S*, and that *S* has at least *k* more attributes with value 1 than with value 0. Then the optimal prediction lies outside the convex hull of the forecasts.

The proof follows identically to the previous case.

Let's look more carefully at the restriction as it provides insights into both why the best aggregate forecast should leave the convex hull and why, when it doesn't, the forecast should still be biased toward higher probabilities. Let *D*(*S*) equal the number of attributes in *S* with value 1 minus the number of attributes in *S* with value 0. We can now state the following claim.

Claim 2: Let $R = \big(M + 1 - D(s)\big)/2$. Conditional on *S*, the probability of a good outcome equals

Equation 6 $P_S = \sum_{i=R}^{M} \binom{M}{i} \left(\frac{1}{2}\right)^{M}$

Note that if *D*(*S*) > *M*, then the optimal forecast equals good with probability one.

From this claim, one can show that even if the restriction on the number of attributes of value 1 in *S* does not hold, the optimal forecast might still lie outside the convex hull. Let *S*=11, with eight attributes of value 1 and three of value 0. Then *D*(*S*) =5. Assume that *M* = 6 so that *D*(*S*) < *M* and the chance of a good outcome is only 63/64.

Suppose that individuals see *k* attributes in *S*. Suppose that *k* = 3. The most optimistic forecast will come from a forecaster who sees only attributes of value 1. Such a forecaster would predict a good outcome with a probability equal to the probability that at least six of the other fourteen attributes (five in *S* and six in *U*) would have value 1. This equals

Equation 7 $\sum_{i=6}^{14} \binom{14}{i} \left(\frac{1}{2}\right)^{14} = 0.788025$

which is less than 63/64 = 0.984375.

3.2 Generated Estimates

As a comparison, consider the corresponding problem for generated signals. Assume that there exists a variable *y,*

Equation 8 $y = x + \theta$

where θ is a random variable with mean 0. Assume further that there exists a threshold *T* such that the outcome will be good if *y* > *T* and bad otherwise. Thus if *T* = *x,* the outcome will be good with probability 0.5. Assume that agent *i* gets a generated signal s_i that equals *x* plus an identically distributed idiosyncratic error term with mean 0:

Equation 9 $s_i = x + \epsilon_i$

It follows that agent *i* predicts a good outcome with probability

Equation 10 $P(s_i) = prob(x + \epsilon_i + \theta > T)$

The optimal aggregation of these predictions depends on assumptions on the distributions of ϵ_i and θ. Let's first assume that θ is uniformly distributed in [-*a, a*] and that *a* is sufficiently large given the ϵ_i that all forecasters predict probabilities in the open interval (0, 1). In other words, no forecaster knows the outcome with certainty.

The true probability of a good outcome equals

Equation 11 $prob(x + \theta > T) = 0.5 + \frac{x-T}{2a}$

The best possible collective forecast P^* would be to estimate *x* using the average of the s_i's. Denote this by $\bar{s}$.

Equation 12 $P^* = 0.5 + \frac{\bar{s}-T}{2a}$

But note that

Equation 13 $\frac{\bar{s}-T}{2a} = \frac{\frac{\sum_{i=1}^{N} s_i}{N} - T}{2a} = \frac{1}{N}\sum_{i=1}^{N} \frac{s_i - T}{2a}$

Therefore, the best possible collective forecast equals the mean of the individual forecasts:

Equation 14 $P^* = \frac{1}{N}\sum_{i=1}^{N} P(s_i)$

The fact that the optimal forecast is a simple average is an artifact of the uniform distribution. Consider the following counterexample with a non-uniform distribution. Suppose that θ has the density function $f(\theta) = 1/\theta^2$ with $\theta \in [1, \infty]$. Let *T*=4. Consider two forecasters with $s_1 = 2$ and $s_2 = 0$. $P(s_1) = prob(\theta > 2) = \frac{1}{2}$. $P(s_2) = prob(\theta > 4) = \frac{1}{4}$. Based on these two signals, the best estimate for *x* will be 1, which gives a best collective forecast of 1/3. However, a simple average of the two forecasts gives an estate of 3/8.

However, the optimal forecast still lies in the convex hull of the predictions. This will hold generally. Assuming generated signals of the form described above, the optimal collective forecast can be written as $P^* = P(\bar{s})$, the prediction that would be made if a single forecaster saw the average signal. By construction this will lie in the convex hull.

[2] The last step in the proof follows from the fact that the probability of exactly one half of the attributes taking value 1 decreases as the number of attributes increases.

The bottom line of this analysis is that, within the constraints of our examples and theorems, the best aggregation of generated estimates will lie within the convex hull of the individual estimates, while the best aggregation of interpreted estimates may escape the hull, and in any case will be biased toward the edge of the hull.

4. EMPIRICAL DATA

In this section, we examine actual data from the project described in Section 2 in the light of the insights developed in Section 3. The forecasters are anonymous. All of them are human; it would be interesting to apply similar analysis to data produced by artificial agents. We first outline our data and describe our aggregation methods, then report analyses in the light of Section 3, and finally offer comparable results on synthetic data that is generated rather than interpreted.

4.1 Data and Aggregation Methods

Our data consists of forecasts produced by these agents in response to 99 questions.[3] We produced aggregations once a day for each question, using the most recent forecast reported by each forecaster by the time of the aggregation. Questions were open for a variety of durations, from 1 to 263 days, yielding 7038 aggregation events. The total number of forecasters involved was about 165. Forecasters could respond repeatedly to a single question, and not every forecaster responded to every question.

This paper focuses on two categories of aggregation: **voting** and **averaging**. There are others as well [5,9], but these two relate most directly to the insights of Section 3.

In general, both voting and averaging can make use of scores assigned to individual forecasts, based on characteristics found to be associated with accurate and inaccurate forecasts. Such characteristics include the certainty of the forecaster (reflected in the entropy of a forecast), the recency of the forecast relative to the life-time of the question thus far, the past accuracy of the forecaster, and demographic features such as whether the forecaster has experience in intelligence analysis. Our scores are non-negative, computed as the product of factors each in [0, 1].

Let s_i be the score associated with the ith forecast be s_i. Let f_{ij} be the forecast from forecaster i on outcome j of the question.

The **averaging** aggregate for outcome j is

Equation 15 $$\bar{f}_j = \frac{\sum_i w_i f_{ij}}{\sum_i w_i}$$

For **voting**, let $F_j = \left\{i: \underset{k}{argmax}(f_{ik}) = j\right\}$, the set of forecasters who assign their largest probability to outcome j. Then the voting aggregate is[4]

Equation 16 $$\hat{f}_j = \frac{\sum_{i \in F_j} w_i}{\sum_i w_i}$$

The voting aggregate $\hat{f}_j$, unlike the averaging aggregate $\bar{f}_j$, can go outside the convex hull of the f_{ij}. It thus has the potential to be the more accurate aggregator on estimates that come from interpretation of internal models rather than sampling from distributions.

Of course, the theoretical possibility that the optimal aggregate of interpreted estimates leaves their convex hull does not mean that any algorithm that leaves the convex hull is automatically a better aggregator than one that cannot. In fact, as we will see, voting applied to generated data can yield an aggregate outside of the hull. However, the theoretical result does emphasize the importance of aggregation methods that can leave the hull, and urges us to pay attention to this aspect of an aggregator's behavior.

With real data, even for interpreted estimates, we may not achieve an aggregation that leaves the hull, for at least three reasons.

First, the population of agents may include both interpreting and generating agents. In an experiment fitting time series of forecasts to event-based domain models [10], some humans generated series of forecasts that were indistinguishable from random guesses, following the generated rather than the interpreted model. In such a composite population, the estimates for the generating agents will pull the ideal aggregate for their contributions toward the interior of their convex hull, making it more difficult for the overall aggregation to escape the hull. One would like to identify and eliminate the guessing agents, but at present the most robust signatures we have for interpreted estimates are aggregate (discussed below), not individual.

A second, related reason is that it is possible that a single agent's estimation process may include both generative and interpretive processes, so that even at the individual level the distinction is not useful. The success of voting methods in our research suggests that our data are in fact interpreted estimates, but in general we should be prepared for contaminated signals.

A third reason that an aggregation of interpreted estimates may not leave the hull is the existence of extreme forecasts. Empirically, we observe that as forecasts accumulate for a question, it becomes increasingly likely that one or another forecast will assign certainty to one or another outcome. It only takes a single forecaster committing to each outcome to saturate the simplex (expanding the convex hull to cover the complete simplex), and any aggregation of necessity falls within, or at most on the edge, of the convex hull.

4.2 Real Data

With these caveats in mind, let's examine some data from our experiments.

Figure 2 compares the results of averaging and voting aggregation. To allow comparison with synthetic data, we use fixed scores ($\forall i$: $s_i = 1$). This Figure follows these conventions:

- Each mark is a single aggregation event. Blue (round) marks indicate that voting selects the correct outcome, and red

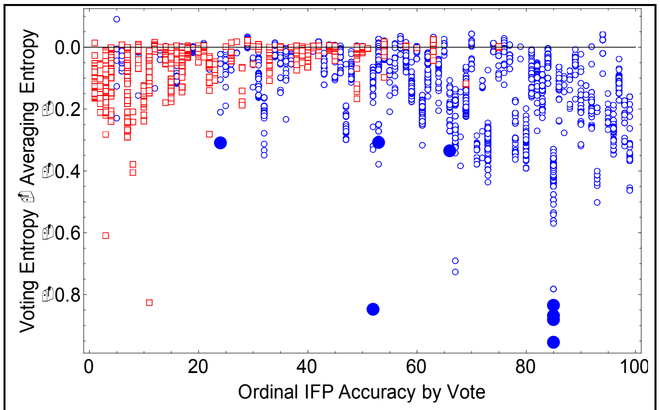

Figure 2: Aggregation Events.—Large filled marks lie outside the convex hull. See text for details.

[3] Collected via our website at https://ace-informed.net

[4] For questions with more than two outcomes, alternative voting methods (Section 5) can yield different winners [2].

(square) marks, the incorrect outcome. Large filled marks are cases where voting leaves the convex hull

- Each column contains the aggregation events belonging to a single question. Columns are ordered along the abscissa from lowest to highest accuracy (the average accuracy of all aggregations belonging to a single question) of the voting aggregation.
- A mark's ordinate is the difference between the entropies produced by voting and averaging aggregation. An aggregate's entropy measures how much it differentiates the various outcomes. Negative values indicate that voting produced a more extreme estimate than averaging.

Several features invite discussion.

- Almost every mark falls below the equal-entropy line, indicating that the voting aggregator is more extreme than the averaging aggregator.
- Events where voting actually leaves the convex hull (the large marks) are quite rare (only 8 out of 7038 events). However, in none of these cases did voting select the wrong outcome. Overall, the probability that an aggregation is in error is 0.26, so the probability of getting all eight out-of-hull forecasts correct strictly by chance is $(1 – 0.26)^8 = 0.1$, not impossible but unlikely. In general, a voting aggregation that leaves the convex hull is correct, which suggests that our voting mechanism is indeed detecting the theoretical characteristics of interpreted estimates discussed in Section 3. The higher entropy of these correct voting aggregations compared with averaging aggregations indicates that they are assigning a higher probability to the correct outcome, and thus providing a more accurate forecast.
- Examination of the detailed data behind this plot confirms that these excursions from the convex hull occur early in a question's lifetime (while the hull is still a strict subset of the simplex). All things being equal, early warnings are more valuable than later ones, and a voting aggregation that leaves the hull is a promising indication that the outcome it favors is in fact correct.
- We might hope that a large difference between voting and averaging would be a signature that voting is correct. Up to a point, this observation is correct. As one moves across Figure 2 from right to left (toward decreasing accuracy), the amplitude of the maximum difference decreases. However, frustratingly, it increases again for the questions on which voting gives the worst outcome. Figure 3 highlights this effect by plotting moving averages of 5 of the lower quartiles of each column in Figure 2. This same bimodality appears in other aggregation methods as well. We discuss its significance in Section 4.3.

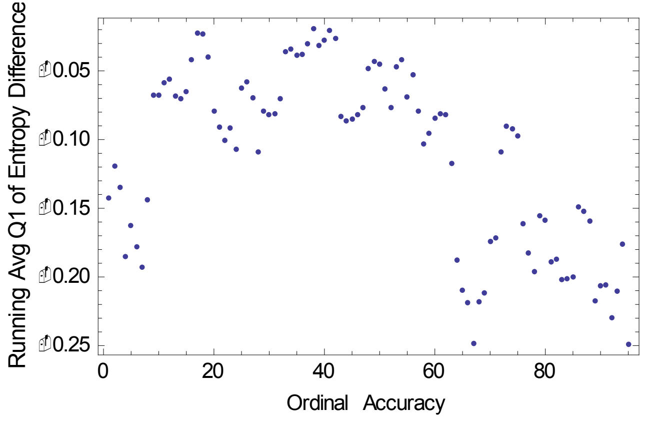

Figure 3: Trends in Entropy Difference.--Moving averages of 5 of the lower quartile of the entropy difference for each accuracy level.

Another way to look at this data is to ask when voting and averaging select qualitatively different outcomes. Figure 4 contains the same aggregation events as Figure 2, plotted against the same axes. However, this time the large filled marks are cases in which voting and averaging select different outcomes, which happens in 12% of our aggregations. When the two approaches differ qualitatively, the difference between their entropies is usually small, suggesting that the forecasts themselves offer little discriminating information.

When the approaches disagree, neither has an advantage: in one half of the cases of disagreement, voting gives the correct answer, and in the other half, averaging wins. (The preponderance of large red squares over large blue circles in Figure 4 is an artifact of the order in which marks are printed.) If one is interested only in selecting the most likely outcome, rather than obtaining the most accurate quantitative estimate of the probability assigned to that outcome, the venerable weighted average is quite serviceable.

4.3 Synthetic Generated Data

The central thesis of this paper is that interpreted data (such as one might expect from forecasters thinking about complex problems) requires different aggregation procedures than generated data. The observations in the previous section will be more meaningful if we compare them with the behavior of our aggregation methods on generated data.

To allow such a comparison, we construct a synthetic data set with gross characteristics similar to those of our real data. We begin with a set of 85 binary questions with 2347 aggregation events.

For each aggregation event on those questions, we extract from the real data the number of forecasters, and the mean and variance of their forecasts. We then average the means and variances of all aggregations for a single question to obtain an overall mean and variance, and use those values to fit a Beta distribution representing the question.

Each agent simulating a forecaster queries the Beta distribution (simulating Equation 8), and adds a uniform random error in [-ε, ε], where $\varepsilon \in [0, 0.5]$ is a characteristic of the agent (according to Equation 9). If the sum exceeds [0, 1], it is trimmed to this range.

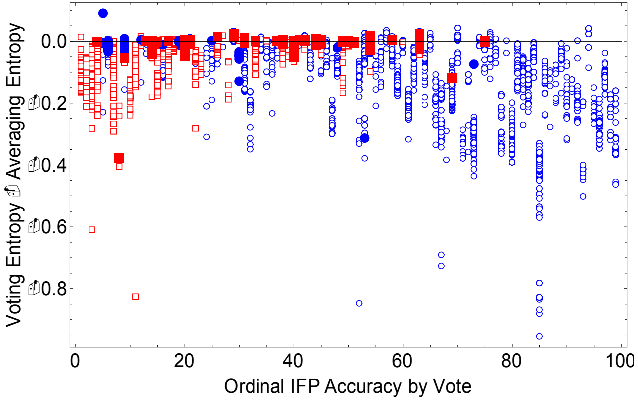

Figure 4: Aggregation Events.—Large filled marks indicate disagreement between voting and averaging.

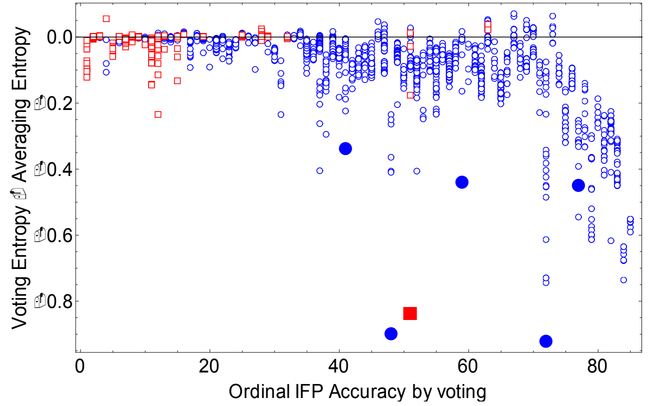

Figure 5: Synthetic Generated Data.—Large filled marks indicate out-of-hull aggregations

Each agent only responds once to each question, on the same simulated date on which the real forecaster first responded. The growth in the number of forecasts for a given question over time is the same as in the real data.

Contrast this algorithm with what we hypothesize is happening in the real data. Analysis of forecasters using other tools [9,11] suggests that they have idiosyncratic internal models [10] that are modulated by external events that they observe in the real world. Different internal models focus a forecaster's attention on different real-world events, fitting the general schema of Section 3.1. By contrast, our simulated forecasters all attend to the same data, a common Beta distribution. Their forecasts differ only on the deviation from the mean of the distribution inherent in the sampling, and their idiosyncratic error in [-*ε*, *ε*].

In terms of the analysis of Section 3.2, the best aggregated answer for a question would be the mean of the Beta distribution. However, for consistency with the analysis of our real data, we present the results in the same way, scoring success not by a theoretical "best achievable estimate," but by outcome of our questions in the real world.

As noted previously, voting has the potential to generate a result outside of the convex hull of the forecasts regardless of the process by which the forecasts are created. Figure 5 is the synthetic counterpart to Figure 2, and shows six cases where voting generates an out-of-hull aggregation. This time, one of them (16.7%) is incorrect, a larger percentage than the overall error rate (10%). With due caution because of the small numbers involved, the out-of-hull aggregations generated by voting with real data offers have a higher probability of being correct than the overall population of aggregations, while out-of-hull aggregations from voting on generated data have a higher probability of being wrong than the overall set of aggregations.

Another suggestive difference between Figure 2 and Figure 5 is the absence in generated data of the high entropy differences for low-accuracy aggregations seen with interpreted data. This difference can be traced directly to the difference between generated and interpreted data. Examination of low-accuracy aggregations with high entropy differences show that they all concern questions of the same type. Questions can broadly be divided into "by-date" questions ("Will event X happen **by date Y**?") and "on-date" questions ("Will proposition X be true **on date Y**?"). In the example questions in Section 2, question 1 is a by-date question, while the other two are on-date questions (the dates implicitly being the date of the Russian election for question 2 and inauguration day for question 3). One can also describe a by-date question as a "status quo" question, since it is asking whether the status quo will persist through the specified date, or whether something will change before then. A distinctive feature of by-date questions is that the question can be resolved before its expiration date by an event in the world. In our real data, low-accuracy aggregations with high entropy differences are all by-date questions that expired early (that is, in which the status quo was not preserved). Only forecasters whose mental models lead them to attend to precursors of the actual event that closes a by-date question early will give the correct forecast, while others will see no reason to update their initial forecasts. Voting aggregation amplifies this divergence, leading to high entropy low-accuracy aggregation events. The lack of such events with generated data is consistent with the lack of an underlying model.

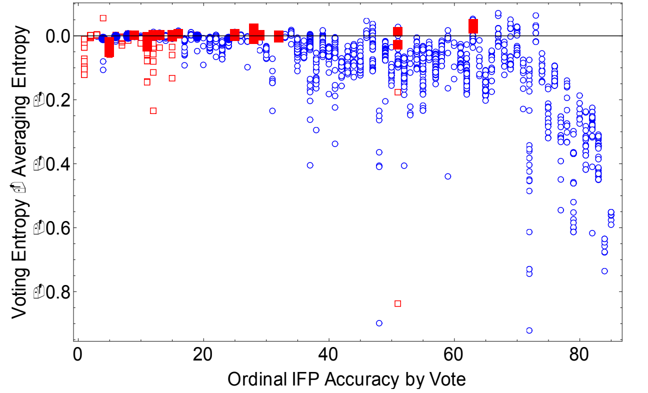

Figure 6: Synthetic Generated Data.—Large filled marks indicate disagreement between voting and averaging.

Figure 6 is the generated data counterpart to Figure 4. As with interpreted data, so with generated data, when voting and averaging reach different conclusions, the entropy of their distributions are almost identical. Again, neither approach offers a preponderance of correct answers.

5. FUTURE WORK

Our results suggest a number of lines of further work.

For questions with more than two outcomes, alternative voting methods can yield different winners [2]. Our voting method is a version of Condorcet voting: each agent indicates only its most favored outcome, discarding information about preferences among other outcomes. One could construct an analog to Borda voting, in which an agent's vote is distributed across its outcomes, discounted for the rank of each outcome. For example (and other formulations are possible), assume a question has *M* outcomes, indexed by *j*. Each agent *i* assigns a rank r_{ij} to each outcome, assigning rank *M* - 1 to its favored outcome, *M* – 2 to the next, and so forth, down to 0 for the least favored. It then votes

Equation 17 $$w_{ij} = \frac{r_{ij} w_i}{M(M-1)/2}$$

for each outcome. Note that $\sum_j w_{ij} = w_i$. Then Equation 16 becomes

Equation 18 $$\hat{f}_j = \frac{\sum_i w_{ij}}{\sum_i w_i}$$

where we no longer require the restriction to F_j. Because most of the questions in this report (and all of our synthetic data) are binary, the two voting approaches yield the same result, but Borda voting may offer useful benefits for multinomial questions.

For simplicity, and because of the data at our disposal, we have focused on aggregating estimates that consist of points in a multinomial simplex. Estimates come in other forms as well, including votes (which may be considered probabilistic estimates at the corners of the simplex), rankings, and bids in a market. It seems reasonable to expect that estimates expressed in these forms will also differ in appropriate aggregation methods, given appropriate analogies to the notions of "simplex" and "convex hull."

In particular, in some tasks, agents are asked to provide qualitative estimates. Such data could be either generated (if agents sample a multinomial distribution and report only the label of the winning outcome) or interpreted. In either case, a labeled estimate can be considered a quantitative one that assigns unit probability to one outcome, and then averaging gives the same result as voting. The agent, by reporting only its most favored outcome, has effectively cast a vote. In such cases, we can estimate whether agents are using the same underlying model or not by statistical measures involving either estimates from the same agents on different problems, or multiple forecasts by two agents on the same problem [12]. These methods lie beyond the scope of the current paper.

We have noted that the distinction between generated and interpreted estimates is sometimes fuzzy. By definition, aggregation is required just because we have multiple sources of estimates, and they may not all use the same reasoning mechanisms. In addition, one can imagine ways in which a single agent's reasoning may combine elements of interpretation and generation. For example, it is plausible that limbic responses such as emotion may have a large generated component, and if a problem evokes an emotional as well as a rational response from an agent, the two kinds of signals could be mixed. Aggregation research will be advanced if we can detect the nature of the process leading to an individual estimate, rather than just characterize the predominant underlying process based on the accuracy of various aggregations.

The notion of a "convex hull" in a probabilistic space is itself more nuanced than we have so far suggested. Briefly, there are infinitely many well-defined geometries that one can apply to a probabilistic space, none intrinsically more natural than any other. The geodesics defined by these geometries, and thus the convex hulls that they define given a set of points, differ from one another. Our results are appropriate to one such geometry, and it is an open question whether working in a different geometry would yield insightful or useful results.

To make this issue more precise, let's begin with the notion of a "well-defined geometry." A probability distribution $p_X(x)$, over, for example, a real variable x, can be represented in terms of a different variable y related to x by an invertible, differentiable function *y(x)*: $p_Y(y) = p_X(x)/|dy/dx|$ by introducing the Jacobian factor $|dy/dx|$ to compensate for the stretching and shrinking performed by the transformation. (For a discrete distribution, the analogous argument involves meaninglessly splitting one categorical outcome into two or more, spreading the density formerly in the old category across the new categories arbitrarily.) Any such change of variables changes the representation of the distribution but not the distribution itself, so for a relationship *between distributions* to be well defined, it must be *invariant* with respect to such transformations. This turns out to be a rather stringent condition, allowing essentially just a 1-parameter family of geometries [4]. These geometries are characterized by the information divergence formula [14]:

Equation 19 $$D_\delta(P,Q) = \frac{1-\sum_i p_i^\delta q_i^{1-\delta}}{\delta(1-\delta)}$$

where δ is the parameter identifying the geometry. Well-known examples of this function are the Hellinger divergence (for δ = 0.5) and the Kullback-Liebler divergence (or relative entropy, for δ = 1). Certain second derivatives of the divergence define a *metric*, the Fisher information matrix, that does not depend on δ, and certain third derivatives define *connections*, which do differ in the different geometries. The metric defines distance between infinitesimally separated distributions, and the connection defines parallel transport: how to carry a vector from one point on the manifold to another. In particular, the connection tells how to move a vector parallel to itself, which defines *straightness*. It turns out that in the δ-geometry, the points x on a straight line (*geodesic*) between distributions p and q satisfy $x^\delta = ap^\delta + (1-a)q^\delta$ for a in [0, 1].

While the form of $D_\delta(P, Q)$ is privileged in its invariance to transformations of the domain of a distribution, there is no "correct" value of δ. All of our results are stated for the geometry of δ = 1. However, as Figure 7 illustrates for the case of the trinomial simplex, the area of the simplex that is included in the convex hull of a set of points can vary, increasing as δ exceeds 1 and shrinking as it becomes less than 1. (The one case where this ambiguity is not present is in binary problems, which fortunately are extremely common.) Generalizing our claims to address the full family of δ-geometries is an important area of future work.

6. CONCLUSION

Agents reason in different ways in reaching estimates. The statistical properties of those estimates vary systematically with different kinds of reasoning, and these differences have implications for the best ways to aggregate them. Roughly, estimates resulting from an agent's sampling an underlying distribution are best aggregated by methods, such as weighted averages, that remain within the convex hull of the individual estimates, while estimates resulting from agents' interpretation of individual models are best aggregated by methods such as voting or latent variable graphical models that can move outside the convex hull. Weighted averages

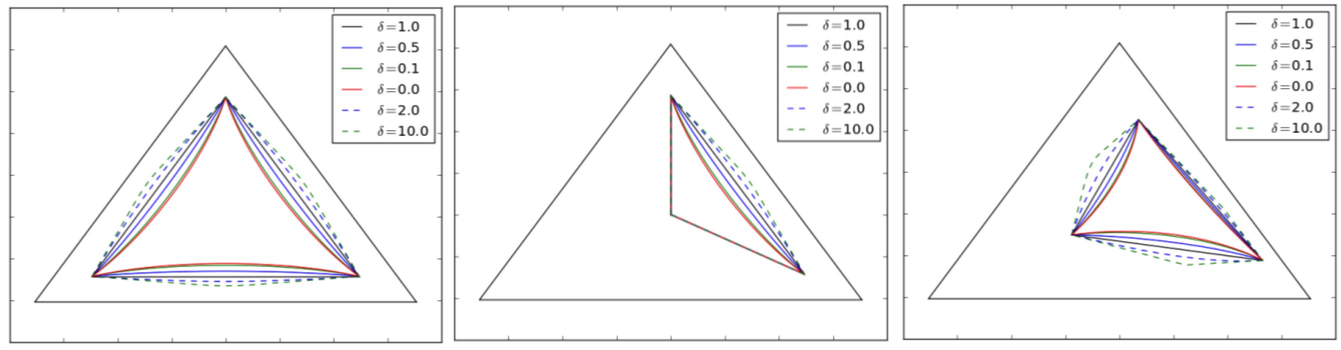

Figure 7: Variation of convex hull with δ

are widely used in practice, but handicapped by being constrained to the hull. Our results commend the decision of [1] to use voting to aggregate human data.

To focus the discussion, we have distinguished between purely generated and purely interpreted estimates. As we note, mixtures can occur, either because our population includes both interpreting and generating agents, or because individual agents combine generative and interpretive processes. In addition, in some cases we may not know what kind of process generates the estimates. Our results are potentially valuable in estimating the internal reasoning methods of collections of unknown agents. We have described two signatures of agents that are sources of interpreted rather than generated signals. For such agents, voting aggregation systematically gives more accurate results than averaging aggregation, and (for appropriate kinds of questions) low-accuracy questions can exhibit high differences between the entropies of voting and averaging aggregation.

7. ACKNOWLEDGMENTS

We thank the reviewers for several provocative observations that have improved the paper.

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20060. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

8. REFERENCES

[1]Y. Bachrach, T. Graepel, et al. Crowd IQ - Aggregating Opinions to Boost Performance. In *Proc. Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2012)*, IFAAMAS, 2012.

[2]S.J. Brams, P.C. Fishburn. Voting Procedures. In K. J. Arrow, A. K. Sen, and K. Suzumura, Editors, *Handbook of Social Choice and Welfare*, vol. 1, Elsevier Science, 2002.

[3]G.W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(2), 1950.

[4]N.N. Cencov. *Statistical Decision Rules and Optimal Inference*. Rhode Island, American Mathematical Society, 1982.

[5]C.-C. Cheng, R. Sasseen, et al. Factor Based Regression Models for Forecasting. In *Proc. Workshop on Machine Learning in Human Computation & Crowdsourcing at ICML 2012*, pages (forthcoming), 2012.

[6]L. Hong, S.E. Page. Interpreted and Generated Signals. *Journal of Economic Theory*, 144:2174-2196, 2009.

[7]E. Kamar, S. Hacker, et al. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *Proc. Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2012)*, IFAAMAS, 2012.

[8]E. Kamar, E. Horvitz. Incentives for Truthful Reporting in Crowdsourcing (Extended Abstract). In *Proc. Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2012)*, IFAAMAS, 2012.

[9]H.V.D. Parunak. Cluster-Weighted Aggregation. In *Proc. AAAI Fall Symposium: Machine Aggregation of Human Judgment (MAGG 2012)*, 2012.

[10]H.V.D. Parunak, S. Brueckner, et al. Swarming Estimation of Realistic Mental Models. In *Proc. Thirteenth Workshop on Multi-Agent Based Simulation (MABS 2012, at AAMAS 2012)*, pages (forthcoming), Springer, 2012.

[11]H.V.D. Parunak, E. Downs. Estimating Diversity among Forecaster Models In *Proc. AAAI Fall Symposium: Machine Aggregation of Human Judgment (MAGG 2012)*, 2012.

[12]H.V.D. Parunak, E. Downs. Exploiting User Model Diversity in Forecast Aggregation. In *Proc. International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP13)*, pages (forthcoming), 2013.

[13]H. Zhang, E. Horvitz, et al. Task Routing for Prediction Tasks. In *Proc. Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2012)*, IFAAMAS, 2012.

[14]H. Zhu, R. Rohwer. Measurements of Generalisation based on Information Geometry. In S. W. Ellacott, J. C. Mason, and I. J. Anderson, Editors, *Mathematics of Neural Networks: Models, Algorithms and Applications*, Kluwer, 1997.