AAAI Technical Report FS-12-06
Machine Aggregation of Human Judgment

Estimating Diversity among Forecaster Models

H. Van Dyke Parunak and Elizabeth Downs

Jacobs Technology Group
3520 Green Court, Suite 250
Ann Arbor, MI 48105
{van.parunak, liz.downs}@jacobs.com

Abstract

There is strong theoretical evidence that aggregation of human judgments should not simply average multiple forecasts together (the unweighted linear opinion pool, or ULinOP), but weight them in such a way as to insure representation of a maximally diverse set of models among the experts from whom they are elicited. Explicitly eliciting these models places a major burden on the experts. We report on a variety of approaches to estimating these models, or at least the diversity among them, with minimal explicit input from the experts.

Introduction

In many domains (e.g., intelligence analysis, business planning, and economic forecasting), human judgments are the most accessible data on which to base decisions. An aggregation of opinions of many people is often more accurate than that of a few (Surowiecki 2004). One theory explaining this observation is that different people have different mental models of the domain (Hong and Page 2009), and the aggregation should insure the representation of as many models as possible (Page 2007).

In the nature of the case, forecaster models are internal, and may not even be consciously articulated by the forecasters. So we need to estimate them on the basis of external signals. Fortunately, we don't have to recover the actual models. It is sufficient if we can estimate the diversity among models held by various forecasters. We would like to be able to assign a number $\delta_{ij} \in [0,1]$ to each pair (i, j) of forecasters, where 0 means that we can discern no difference between their models and 1 is maximal distinction over the population of forecasters. Diversity among models captures a relevant characteristic of diversity among forecasters, which we ought to take into account in aggregating forecasts.

This paper reviews a number of approaches that we have explored to estimating the diversity of forecaster models. First we discuss what we mean by a mental model. Next, we summarize the information that we elicit. Then we describe a series of diversity measures, and offer discussion and next steps.

What is a Mental Model?

The question of just what constitutes a "mental model" is non-trivial. We are interested in how a forecaster reasons her way to a future outcome, so we seek a formal structure that can describe how the world evolves. One can imagine a variety of forms that such a model might take, including differential or difference equations over *state variables*, Markov processes over *states*, or conditional probability dependencies among *statements*. Elsewhere (Parunak et al. 2012) we describe these alternatives in more detail, examine psychological evidence for and against each of them, and argue for a weighted directed graph, the Narrative Generator (NG), that applies the probabilistic transitions of a Markov process to statements rather than states. Unlike states, statements are not mutually exclusive and collectively exhaustive, so the full Markovian apparatus is not applicable, and we develop a Monte Carlo sampling mechanism, PENS (Probability Estimation in Narrative Spaces), to fit the transition probabilities to the forecasts that we observe. Each node of a NG corresponds to statements about the domain of the IFP. Each arc indicates the coherence of a narrative beginning with the node at the tail of the arc and moving to the node at the head of the arc. The weight on an arc is the probability that the forecaster, having constructed a narrative that reaches the tail of the arc, will follow that arc for the next statement in the narrative

In the most concrete terms, when we speak of "diversity among forecaster models," we have in mind some measure of the difference among the forecasters' NGs. Of course, we do not have access to these NGs, only to the forecasts

that they produce. The various measures that we discuss draw on different kinds of information that plausibly reflect some aspects of the underlying NGs. Thus the diversities estimated by the different measures, like shadows of an object cast from different angles, are not directly commensurable, but we will argue that they all reflect meaningful differences among the underlying mental models.

Elicited Information

The methods in this paper draw on three kinds of elicited information. All of them require forecasts, two of them require a log of world events, and two require a graphical model linking those events. In our current system, forecasts are elicited from a large population of experts, while the event log and graphical model are elicited from a small team of subject-matter experts (SMEs).

Forecasts

The forecasts used in this research concern questions

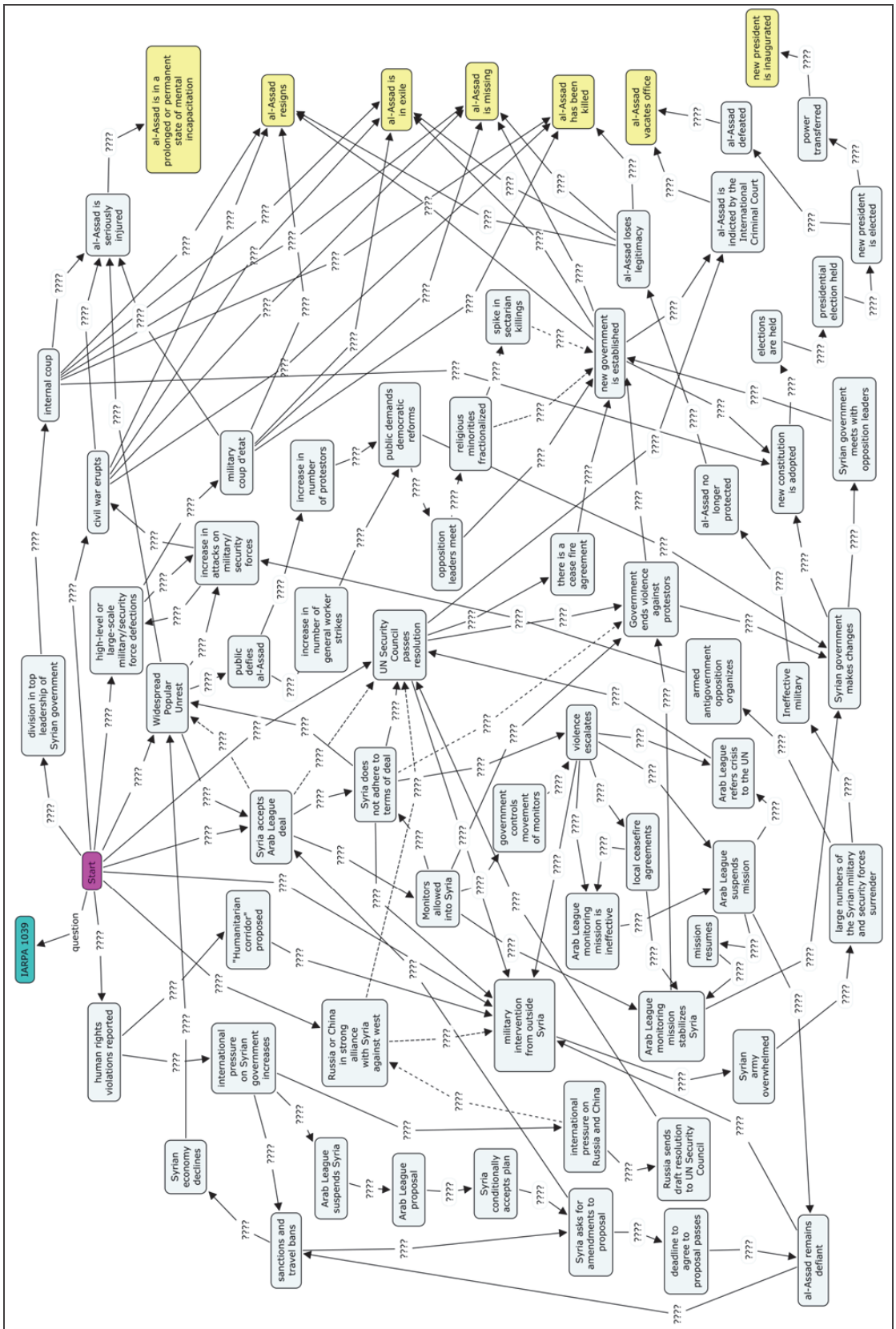

Figure 1: Example NSM for al-Assad IFP. Any path from "Start" to a yellow node is a narrative supporting "No." Edge weights (transition probabilities) are fit to forecasts by PENS algorithm.

about international affairs, of the form, "Will Bashar al-Assad remain President of Syria through 31 January 2012? (Yes/No)." Each such question is an "Individual Forecasting Problem," or IFP, with a discrete unordered set of outcomes (in this example, two, though some problems offer more). IFPs are introduced centrally at a specific time, and close either when the focal event takes place, or when a specified time window expires. A forecast consists of a probability distribution over possible responses. We elicit forecasts through a web-based interface that forecasters can access at any time, and allow (in fact, encourage) forecasters to update their forecasts over time. Forecasts reflect the underlying NGs because they are generated by following a trajectory through the NG to a statement (a node in the NG) that asserts one of the outcomes of the IFP.

Narrative Space Model

A NG consists of a directed graph of statements with weights on its edges. We estimate a forecaster's NG by fitting transition probabilities to a preexisting graph, elicited from a small team of SMEs for each IFP. We call this graph a "Narrative Space Model," or NSM, because it represents the space of possible narratives that would explain the IFP. The objective of the NSM is to support as wide a range of coherent narratives about how the forecasted event might occur as possible. Each NSM has a unique start node, and as many end nodes as there are outcomes to the IFP. A trajectory from the start node to one of the outcome notes generates a sequence of statements (the nodes traversed by the trajectory) that constitutes a narrative explaining how the outcome might arise. At this point, we rely on the expertise of our SMEs to construct NSMs that are comprehensive enough to contain forecasters' individual NGs as subgraphs.

Figure 1 is an example of a NSM for the al-Assad problem. The "Start" node is at the top left. The node marked "IARPA-1039" contains documentation on the IFP, and is not part of the structure of the model. The yellow nodes along the right edge are outcomes that could lead to a "No" answer. All other nodes have an implicit link to a "Yes" answer, which will be the answer if none of the "No" nodes is reached.

Event Log

Table 1: Estimation Procedures.

Procedure	Required Information
Transition Spectra	Forecasts, Event list with NSM node IDs, NSM links
Event Spectra	Forecasts, Event list with NSM node IDs
Event-Forecast Correlation	Forecasts, Event dates
Forecast Divergences	Forecasts

Each node in a NSM is a statement about the world. Our intuition is that the probability that a forecaster will move to a statement depends on events relevant to the IFP that the forecaster observes. Thus, correlations between forecasts and visible events should provide information about a forecaster's movement over her (internal) NG.

Each day, a SME who is acquainted with the current open set of IFPs reviews a number of news outlets, such as Reuters, BBC, and CNN, and notes events that might be relevant to each IFP. Each such record includes up to four elements: the date of the reported event, the IFP identifier, a free text description of the event and its relevance to the IFP, and (if an NSM exists for the IFP) the node identifier in the NSM for which the event is relevant. More than one event may attest to the same node, and a single event may attest to more than one node.

Estimation Procedures

We describe four different ways to estimate diversity among forecasters' mental models from the elicited information, in order of decreasing amounts of required information (Table 1).

Transition Spectra

The PENS algorithm integrates information about events relevant to the IFP with the dated forecasts issued by a forecaster to assign a probability to each edge of the NSM such that random sampling of the resulting NG according to the probabilities will generate outcomes in the proportions indicated by a forecaster's latest forecast. We call this list of probabilities, in an arbitrary but fixed ordering of the edges, the NG's *transition spectrum.* Clearly, this spectrum reflects the underlying NG, and thus (*ex hypothesi*) the forecaster's mental model. The transition probabilities from each node sum to 1, and the NSM has a fixed number of nodes, so the sum of transition probabilities for any NG fit to the same NSM will be the same, and we can treat the entire spectrum as a probability distribution by normalizing by the number of nodes.

The only way to compare two probability distributions that is invariant under transformations of their domain is the family of delta divergences (Zhu and Rohwer 1995):

$$D^{(\delta)}(p||q) = \frac{1 - \int p(x)^{\delta} q(x)^{1-\delta} dx}{\delta\,(1-\delta)}$$

$D^{(1)}(p||q)$ is the Kullback-Leibler (KL) divergence, while $D^{(.5)}(p||q)$ is four times the square of the Hellinger divergence. The KL divergence is unbounded, but for $\delta < 1$, the divergence is bounded. In deriving NSM spectra, we use a symmetric version of $D^{(.95)}(p||q)$ scaled to [0, 1].

Figure 2 shows the differences between successive spectra generated by a forecaster on successive elicitations of the binary question about al-Assad given above. The top

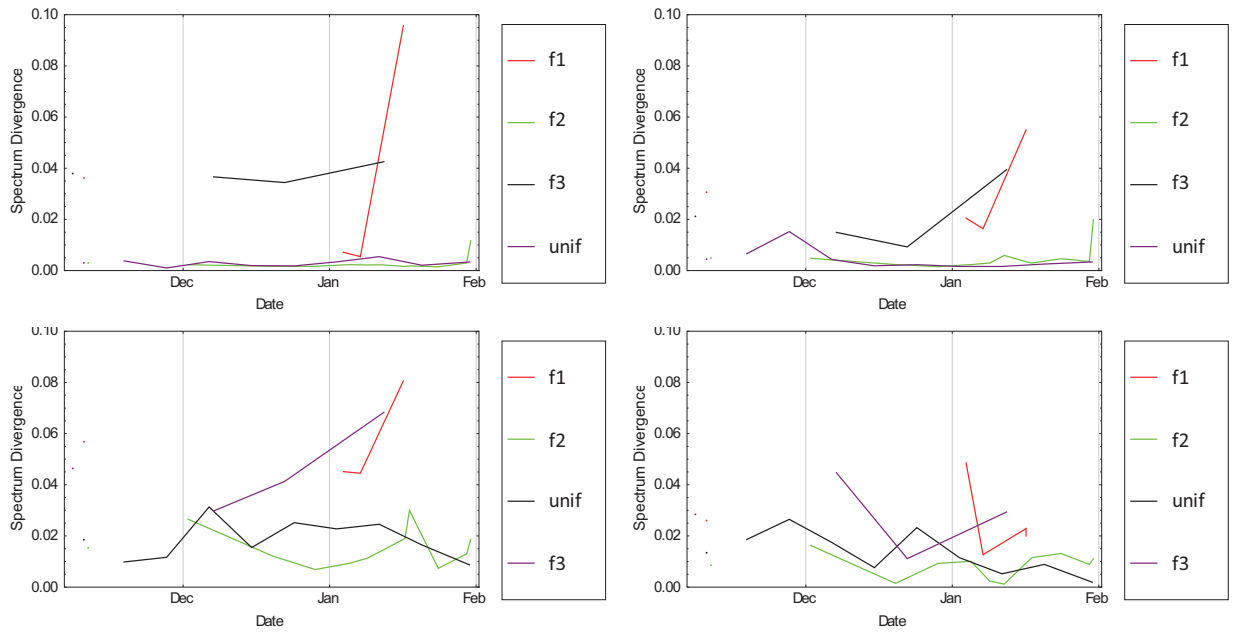

Figure 2: Evolution of the Transition Spectrum. Top: two fittings of the same data, the al-Assad IFP, to a SME-generated NSM. Bottom: fittings of the same data to two different random graphs.

two plots show spectra from a SME-based NSM. The bottom two plots show spectra from random graphs, discussed below. The plots share the several features:

- We plot forecasts from three users "f1," "f2," and "f3", and a uniform (0.5, 0.5) forecast "unif" as a baseline.
- The lines trace the divergence of the spectrum for a forecast at the given date compared with the previous forecast. At the far left of each plot, small dots show the date of the first forecast.
- PENS is stochastic, thus the same series of forecasts does not generate the same sequence of spectra, accounting for the differences between the two plots in a single row.
- PENS is influenced by events that have happened prior to the date of the forecast. Thus even the uniform series of forecasts does not yield the same spectrum for each forecast.

In the top row, we observe:

- The series of forecasts by f2 (green line) is remarkably similar to the uniform series, even though the forecasts themselves are not uniform. However, they do change very slowly, and in ways that do not stimulate any major shifts in the spectrum.
- The uniform series shows very little change in the spectrum from one forecast to the next, much less than the level of change characteristic of the forecasters other than f2.

How do the patterns in Figure 2 depend on the semantic structure built into the NSM by the SME? To explore this question, we construct a random NSM that preserves the number of nodes, the node names, and the number of nodes at each minimum distance from the "Start" node in the SME graph (Downs and Parunak 2012). It does not preserve either the detailed connectivity of the NSM, or the logical sequence of nodes. We feed events to the nodes with which they are labeled, and process the same series of forecasts against the random graph. The bottom row of Figure 2 shows NGs fitted to two different random graphs derived from the al-Assad graph.

Qualitatively, the random graph shows much less difference between the uniform spectral changes, on the one hand, and the spectral changes in forecasters f1 and f3, on the other, than does the SME graph. PENS is able to integrate meaningful structure in the SME graph with forecasts made by people, and distinguish those forecasts from a series of uniform forecasts. The role played by the SME graph in this process is substantiated by the lack of a clear distinction when the same forecasts are fit to a random graph.

Event Spectra

Transition spectra require a SME to identify both a limited set of nodes (to which events can be mapped) and a narrative structure among those nodes. What could we do in the absence of the structure? Our event log notes the date on which each node is activated. Forecasters make forecasts at points in time. It's reasonable that one factor determining the date of a forecast would be the occurrence of an event that the forecaster deems relevant. If an event happens that one forecaster deems relevant and another does not, the first becomes more likely to make a forecast than the second. Thus comparing correlations between forecasts and events is likely to reflect a forecaster's mental model, and

differences among those patterns of correlation would indicate diversity among forecasters. This approach has two clear benefits.

First, if we can extract a useful signal about model content from the temporal correlation between event postings and forecasts, we could estimate forecaster model diversity without the need to construct a full NSM.

Second, focusing just on nodes, without the need for edges, permits crowdsourcing the required semantic information. It is unrealistic to ask forecasters for complete narratives about how an IFP might resolve. But we can elicit short statements about what influenced their latest forecast. Our SME could curate these statements to provide the nodes against which to index events, greatly reducing the overhead of managing an IFP.

We visualize this effect as an *event spectrum*, the set of NSM nodes for which the event log shows events on a given day. Figure 3 is a plot of NSM nodes (rows) against days on which events are noted (columns). The intensity of a dot reflects the number of events on the column's day that attest to the node associated with the column's row.

Not all nodes are equally attested, and those that are attested most are not attested evenly over the time period in question. We surmise that the day on which a user makes a forecast is characterized by the events in the recent past, say n days (including the day of the forecast). We call the sum of these spectra the n-day event spectrum associated with a forecast.

Figure 4 shows some examples for $n = 2$. The axes are reversed from the previous figure: now each row is a successive forecast, and each column is a topic (an NSM node). The darkness of a cell in the plot is the number of events associated with the column that occurred either on the day of the forecast, or the day before. These spectra show distinctions in the events to which the forecaster appears to be attending. For example:

- The third spectrum routinely ignores nodes in the 4-13 and 24-31 range, which all the others hit
- The last two routinely hit node 3, which the first two hit only once
- The last three routinely hit 38, which then first forecaster hits only once

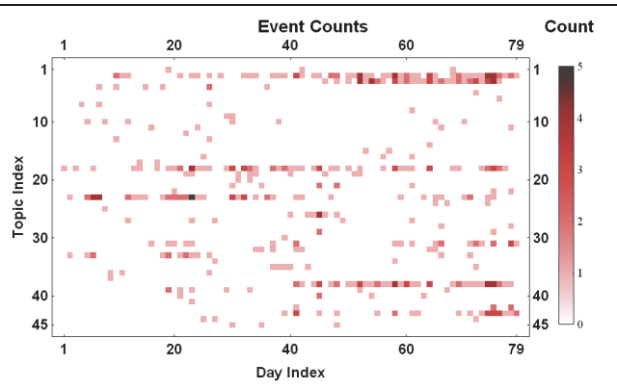

Figure 3: Event Counts per Day. Each row is a node in the NSM; each column is a day; darkness of a cell corresponds to number of events that attest that node on that day.

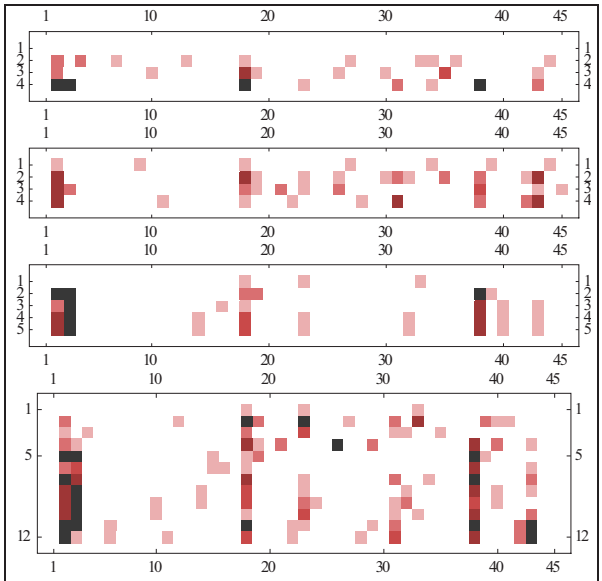

Figure 4: Example Two-Day Event Spectra for different forecasters on the al-Assad IFP. Each row is a day; each topic is an NSM node; darkness of a cell reflects the number of attestations of that node in a two-day window including the day of the forecast and the previous day.

We reduce each such collection of event spectra to a distribution over NSM nodes, then use a delta divergence to estimate the dissimilarity of each pair of forecasters (for example, summing event counts across all forecasts of a given forecaster, then normalizing). Forecasters with empty event spectra (two out of 61 forecasters for the al-Assad IFP) are assigned a uniform distribution over events. This gives us a dissimilarity matrix among forecasters.

Figure 5 shows the result when we apply average-linkage hierarchical clustering to the 61 forecasters for the al-Assad IFP. There is considerable structure in this data; many clusters form at fairly low separations, then combine at much higher levels. The cophenetic correlation coeffi-

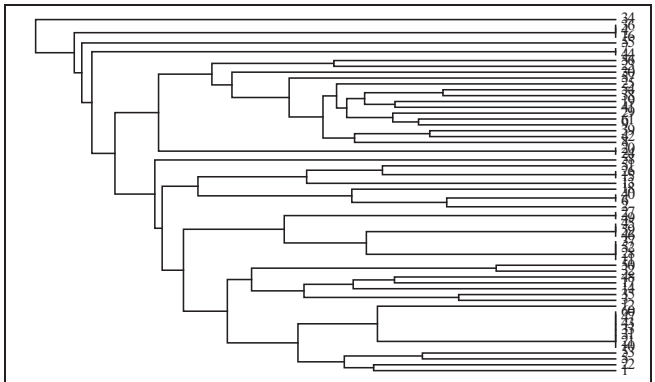

Figure 5: Average-Linkage Clustering of Diversities from Event Data. CCC = 0.86, highest value (2σ) expected from random data is 0.7.

cient (CCC) is 0.86, higher than one would expect from random data (where the largest value at two standard deviations is 0.7, (Rohlf and Fisher 1968)). Thus looking at which event types occur in a forecast's recent past does seem to partition the users non-trivially.

Event-Forecast Correlation

So far, we have abstracted away the structure of the NSM to focus just on the nodes. The method outlined in the previous section still requires matching event reports to NSM nodes. What we can do with unclassified event reports? Is there a correlation between the dates of events, and the dates on which forecasts are made, regardless of the nature of the events? Differences in such correlations across forecasters could reflect differences in their underlying models.

For each IFP, we define two vectors, each indexed by successive days. The event vector counts the number of events relevant to that IFP reported on that day, and the forecast vector reports the number of forecasts made on that IFP on that day. We normalize both vectors.

We then compute the dot product between both vectors, at a variety of offsets. Negative offsets align events with forecasts that follow them; positive offsets align events with forecasts that precede them. If people are attending to events in timing their forecasts, we ought to see higher correlations for negative offsets than for positive ones.

Figure 6 shows representative results from this computation. There does seem to be a difference between positive and negative offsets, but in the wrong direction! Forecasts are correlated with events that *follow* them in time, rather than preceding them.

To understand this anomaly, consider Figure 7, a spreadsheet of forecast counts per day (row) and IFP (column). The cells are color-coded: pink cells have no forecasts, yellow cells have one, green cells have two, and blue cells have more than two.

Our project releases IFPs to forecasters in batches, giving Figure 7 a stairstep appearance. The top of each blue step marks a date on which a new batch of forecasts appears. Forecasts are most common on the day an IFP is

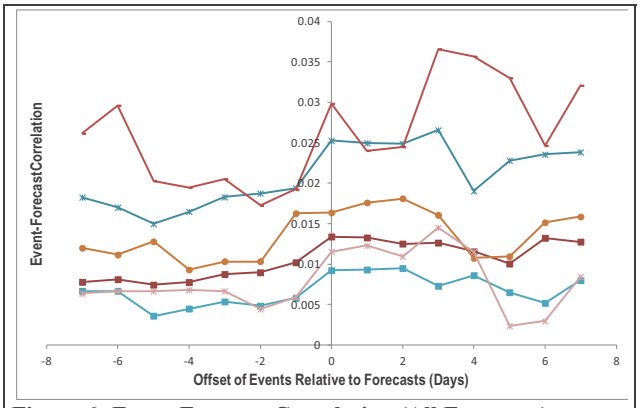

Figure 6: Event-Forecast Correlation (All Forecasts)

Figure 7: Forecast counts per day (row) and IFP (column)

released, and the two or three following days, as seen by the concentration of blue at the top of each step. In addition, note the line of blue extending across previous problems from the top of each step. When forecasters enter the system to consider new problems, they routinely update earlier problems as well.

This dynamic accounts for the anomalous correlations in Figure 6. When an IFP is initially released, it naturally has no events in the event log. When the first events are added a day or two later, we see a jump in correlation between forecasts and events.

Figure 8 repeats the analysis of Figure 6, omitting each forecaster's initial forecast on each problem. Now there is a discernible, if slight, negative slope, evidence that occurrence of an event is correlated with subsequent occurrence of a forecast. Unlike the analysis of events linked to NSM nodes, this analysis does not attempt to distinguish different patterns of correlation for different users. Making such a distinction would strengthen the effect.

Forecast Divergences

What could we tell about forecaster diversity using only the collection of forecasts, without SME input on events or NSMs? A forecast is a probability distribution over outcomes, so we can compare two forecasts using a delta divergence (in the examples given here, a symmetrized KL divergence). The diversity of two forecasters (i,j) on the

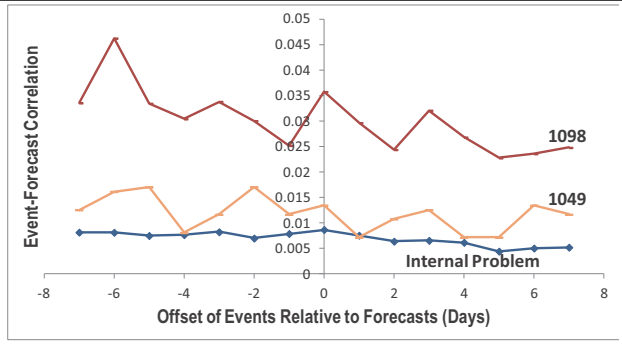

Figure 8: Event-Forecast Correlations (non-initial forecasts)

same IFP is their within-IFP diversity w_{ij}, while the average of their diversities across all IFPs on which they both issue forecasts is their cross-IFP diversity c_{ij}. On one hand, if two forecasters tend to give the same forecasts across many IFPs (indicated by a low c_{ij}), they are likely thinking about them in the same way, and attending to the same kinds of evidence. On the other hand, if their forecasts tend to differ across many IFPs (high c_{ij}), they are probably thinking about them in different ways. Thus c_{ij} is arguably correlated with forecaster model diversity.

We can learn a good deal by combining c_{ij} and w_{ij}. Consider their ratio

$$\gamma_{ij} = \frac{(c_{ij} + \zeta)}{(w_{ij} + \zeta)}$$

where the small constant ζ (= 0.1 in our current work) avoids singularities. γ_{ij} ~1 indicates the unremarkable circumstance that similar models yield similar forecasts, and dissimilar models yield dissimilar ones. However, when $\gamma_{ij} > 1$, dissimilar models are yielding similar forecasts, surely meriting additional attention to their common estimates. Conversely, $\gamma_{ij} < 1$ indicates that even though forecasters share a common model, they disagree on a given problem, suggesting that their shared model offers no purchase on this particular question, and encouraging us to discount their forecasts.

Figure 9 supports this analysis. High Brier scores tend to be concentrated among forecasts on the back walls of the plot (low average c_{ij} and high average w_{ij}), while the forecasts with the best scores have high average c_{ij} and low average w_{ij}.

c_{ij} and w_{ij} can be applied in a number of ways to aggregate forecasts. We have found that it is important not to average out the pairwise divergences between forecasters

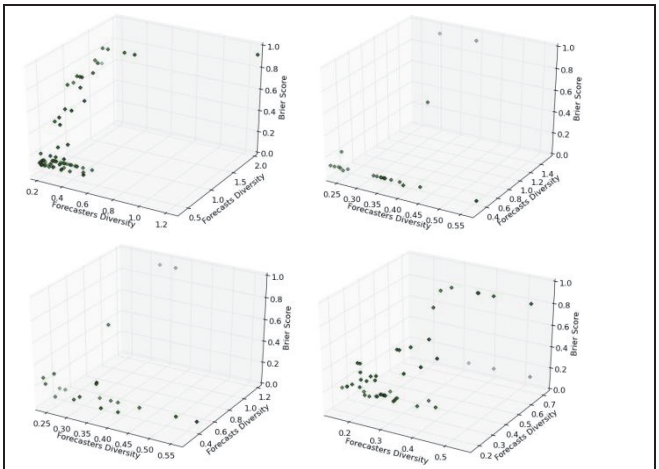

Figure 9: Interaction of average c_{ij} ("forecasters diversity") and average w_{ij} ("forecasts diversity") on four IFPs. The axis across the left front of each plot is forecaster diversity. The axis along the right is forecast diversity. The vertical axis records the Brier score. Each point represents the most recent forecast from one forecaster.

prematurely. We use cluster-weighted aggregation (CWA) (Parunak 2012a; b), first clustering the forecasters by a weighted average of c_{ij} and w_{ij}, then applying γ_{ij} incrementally as we climb the tree from individual forecasts to the root. This approach yields a lift of 17% over ULinOP when applied to the last forecast in each of the 77 closed IFPs in our test corpus. The mean number of forecasters on each IFP is 60, and the mean number of forecasts per forecasters is 29.

Discussion and Next Steps

The results reported in this paper demonstrate that diversity among forecaster models can be estimated from the forecasts themselves, in some cases with the addition of information from SMEs. The most broadly applicable method (because it does not require SME input) is the forecast divergences method, which also yields impressive gains on test data under the CWA algorithm.

Success in deriving quantitative estimates of diversity across forecaster models is affected by some features of the operational environment, quite apart from the actual elicited forecasts themselves. These include:

- **Forecast timing**.—The noise generated by grouped release of IFPs to forecasters overwhelms the important signal of correlation between news events and whether or not a forecaster issues a forecast. This grouping may be more prominent in our experimental framework than in an operating analytic environment, but in any case it is worthwhile to explore ways to reduce this effect.
- **Forecaster retention**.—The forecast divergences method requires a pool of shared forecasts to derive c_{ij}. In spite of its theoretical elegance and evidence of practical success, it is not applicable when forecasters intersect only on single IFPs. Use of this method places a premium on keeping forecasters engaged across many IFPs.
- **Repeat forecasts**.—The two spectral approaches are more effective when each forecaster offers repeated forecasts on each IFP, so that we can track the dependency between their forecasts and news events. The event-forecast correlation method also is much stronger when we have repeated forecasts, since the timing of the first forecast is usually driven by the release of the problem rather than IFP-specific news events.

These results lead to several directions for future research.

- We plan to offer a forecasting app for mobile devices, to encourage repeat forecasts and updating of forecasts the moment a forecaster notices a relevant event.
- Spectral methods will only be practical when we can reduce the cost and time needed to construct NSMs. We are planning to crowdsource at least the nodes of NSMs, a process that will also reduce the potential bias of NSMs formulated entirely by a restricted set of SMEs.
- One may reasonably ask whether these various measures are in fact measuring the same thing, or whether they are

detecting different facets of forecaster model diversity. We plan to study correlations among them.

- At present, only the forecast divergences method is mature enough to evaluate for its impact on aggregation accuracy. We will be refining the other methods to evaluate their performance in aggregation as well.

Acknowledgements

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, Dol/NBC, or the U.S. Government.

References

Downs, E. A. and Parunak, H. V. D. 2012. How to Create a Random NSM Graph. *INFORMED Working Papers*. Ann Arbor, MI: Jacobs Technology.

Hong, L. and Page, S. E. 2009. "Interpreted and Generated Signals." *Journal of Economic Theory* 144: 2174-2196. http://www.cscs.umich.edu/~spage/signals.pdf.

Page, S. E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ: Princeton University Press.

Parunak, H. V. D. 2012a. Cluster-Weighted Aggregation. *INFORMED Working Papers*. Ann Arbor, MI: Jacobs Technology.

Parunak, H. V. D. 2012b. Implementing Cluster-Weighted Aggregation. *INFORMED Working Papers*. Ann Arbor, MI: Jacobs Technology.

Parunak, H. V. D., Brueckner, S., Downs, L. and Sappelsa, L. 2012. Swarming Estimation of Realistic Mental Models. *Thirteenth Workshop on Multi-Agent Based Simulation (MABS 2012, at AAMAS 2012)*. Valencia, Spain: Springer**:** (forthcoming).

Rohlf, F. J. and Fisher, D. R. 1968. "Test for hierarchical structure in random data sets." *Systematic Zoology* 17: 407-412. http://life.bio.sunysb.edu/ee/rohlf/reprints/RohlfFisher_1968.pdf.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York, NY: Doubleday.

Zhu, H. and Rohwer, R. 1995. Information Geometric Measurements of Generalization. Birmingham, UK: Aston University, Dept of Computer Science and Advanced Mathematics, Neural Computing Research Group. http://eprints.aston.ac.uk/507/1/NCRG_95_005.pdf.