

Improving Forecasting Accuracy Using Bayesian Network Decomposition in Prediction Markets

Anamaria Berea

George Mason University
Fairfax, VA
aberea@gmu.edu

Daniel Maxwell

KaDSci LLC
dmaxwell@kadsci.com

Charles Twardy

George Mason University
ctwardy@gmu.edu

Abstract

We propose to improve the accuracy of prediction market forecasts by using Bayesian networks to constrain probabilities among related questions. Prediction markets are already known to increase forecast accuracy compared to single best estimates. Our own flat prediction market substantially beat a baseline linear opinion pool during the first year. One way to improve performance is by expressing relationships among the questions. Elsewhere we describe work on combinatorial markets. Here we show how to use Bayesian networks within a flat market.

The general approach is to decompose a target question (hypothesis) into a set of related variables (causal factors and evidence), when the relationship among the variables is known with some confidence. Then the marginal probabilities for the variables in the Bayes net are updated using the market estimates, with the Bayes net enforcing coherence. This paper describes the overall concept, shows the results for a particular model of the potential Greek exit from the European Union, and describes the team's future research plan.

Introduction

Both the business and national security communities invest substantial resources in forecasting. But forecasting is hard, and the best practice is still to average the forecasts from a diverse crowd with some knowledge of the question of interest – a practice now known as “the wisdom of the crowds” (Surowiecki 2005). In most cases, crowd-sourcing outperforms both individuals and small groups of experts that are using traditional elicitation. DAGGRE is a research project at George Mason University sponsored by the Aggregative Contingent Estimation (ACE) program at the Intelligence Advanced Research Project Activity (IARPA). DAGGRE, which is short for Decomposition-based elicitation and AGGREGation, involves five universities, two commercial research and software companies, and most importantly, over

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

1,000 volunteer forecasters. The experiment described here is part of a much broader forecasting research agenda.

While prediction markets are a proven forecasting technique (Hanson 2003, 2007), our specific hypothesis is that the performance of a prediction market can be improved by complementing the market with a set of Bayesian Network models that decompose a target question (the hypothesis) into likely causes and indicators (i.e. evidence) that are also separately forecasted in the market.

We focus on forecasting world events: usually questions with extended time horizons and significant irreducible uncertainty. These are the types of questions that have historically been the most vexing to intelligence analysts, economists and others. As Cameron asserted, “Not everything that counts can be counted and not everything that can be counted, counts” (Cameron 1963). We aim to add factors that “count” for the target question. We report the forecast probability to IARPA every day, and are scored using the average Brier score (Brier 1950) over the period of time that the target question (hypothesis) is active. This approach has the benefit of rewarding forecasts that identify and trend toward the correct outcome early during the period of time the question is being forecasted.

Methodology

The overall approach to Bayesian decomposition parallels most Bayes Net modeling and analysis projects. In general, there are three steps involved in Bayesian decomposition: 1) the model structure is elicited, 2) initial probabilities are elicited, and 3) the model is updated using available evidence.

The approach begins by eliciting from domain experts a set of variables that might provide information about the target. Further elicitation provides the structure of a Bayesian Network, with some variables directly informing the target question and others influencing the target only by informing intermediate variables. This explicit representation of indirect relationships can be used to help communicate to leaders the reasoning behind the forecasts.

As we create more models, we expect two kinds of patterns to emerge: first, model templates that substantially simplify the modeling process, and second, independence structures such as Noisy-OR (Pearl 1988) that simplify the elicitation of probabilities within the model structure.

Once the model structure is developed, along with its conditional probabilities, the researchers can provide initial marginal probabilities for the key variables. Another approach being considered is initializing the model with a flat prior distribution and allowing the market to dynamically inform and update this probability distribution in real time. (The combinatorial market allows participants to edit the conditional probabilities as well (Sun et al. 2012).)

Because this is an experimental setting, most of the model development effort to date has been accomplished using student researchers and members of the DAGGRE research team. However, there is significant evidence that intelligence analysts with proper training can successfully develop these types of models and that they are perceived as having utility for supporting analysis (Sticha, Buede and Rees 2006). Additionally, a related effort supported by the DAGGRE project (but outside the scope of this paper) is exploring more efficient approaches for eliciting Bayesian Networks. Consequently, we have high confidence that should our approach to improve forecasting accuracy be replicated, it is feasible to deploy it in operational settings.

Once the model structure has been elicited, the related questions are posted on the DAGGRE prediction market and the participants provide probability estimates for the questions. We are exploring two variations of this approach: one in which the target question and the supporting questions are all updated via the market and another one where the probability of the target question is updated using only the related questions via the Bayes Net.

The Generalized Model

To demonstrate the concept, we explain a simple model that implements a Bayesian network for forecasting regime change.

A common target question in the prediction market has this form: *Will Leader X of Country Y remain in power continuously until Date Z?* This is a reasonably straightforward question, and there are a number of reasons why a leader might leave or lose power. He or she could be incapacitated due to natural causes, lose an election, or might even be overthrown by revolt. These causes have two properties that lend themselves to the decomposition approach we are exploring. First, the causes are relatively independent. This supports the use of an independence structure, simplifying model development and explanation. Second, the sources of expertise for each of these causes are different. Consequently, it is likely that different people will choose and estimate the supporting questions heterogeneously. This active specialization distinguishes our approach – and prediction markets generally – from the averaging normally referred to by the “wisdom of the crowd”.

The generalized model has a hypothesis or target node and a number of parent and/or evidential nodes. Given a Bayesian network, $BN = (Q_i, Q_{ij})$, where Q_i = target node

and Q_{ij} = causal or evidential nodes), we compute $\pi(Q_i | Q_{ij})$, where π is the probability of the target question given the estimates of the other nodes in the BN. We also read $\pi'(Q_i)$, where π' is the probability of the core question that is given by the market.

The Case Study Model: Will Greece exit the EU?

For example, assume we have such a question Q_i on the DAGGRE prediction market. We decompose it into causal nodes and add these supporting questions Q_{ij} on the market as well. The algorithm reads the estimates/probabilities for Q_{ij} , and replaces the current Q_{ij} distributions in the BN with those from the market. Technically, this updating conforms to Jeffrey’s rule; we implement it by creating temporary evidence nodes Q'_{ij} and calculating the likelihood ratios Q'_{ij}/Q_{ij} that will produce the desired distribution on each Q_{ij} , and observing the new Q'_{ij} . The temporary variables are then absorbed at their current value.

Figure 1 depicts such a model implemented in UnBBayes.

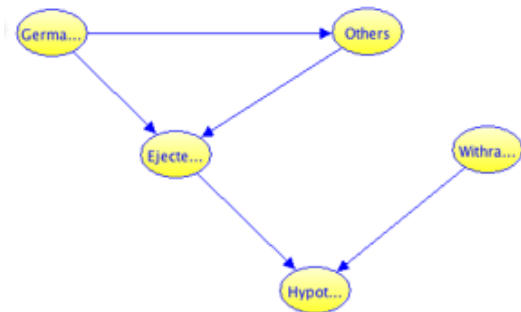


Figure 1: The Bayesian Network decomposition of the Greek Exit case study.

The nodes in this decomposition are:

Q_1 = Will Greece exit the European Union by June 1st, 2012? We decomposed this node, Q_1 , into the following decision nodes:

- Q_{11} = Will Greece be ejected from the EU before June 1st, 2012?
- Q_{12} = Will Greece withdraw from the EU before June 1st, 2012?

Additionally, Q_{11} was decomposed into the following parent (causal) nodes:

- Q_{111} = Will Germany vote to reject Greece from the EU before June 1st, 2012?
- Q_{112} = Will the other EU members vote to reject Greece from the EU before June 1st, 2012?

All the nodes in this BN are binary (yes/no). The founding Maastricht Treaty requires a unanimous vote from all the other EU member countries to eject an EU member. Past votes in the EU showed that Germany’s strong strategic and

financial position in the EU is reflected in their influence over the other EU members. Therefore the modelers gave a probability of 90% that the other members of the EU will follow Germany's vote. We also assigned an initial probability of 80% that Germany would not propose to eject Greece from the EU. Although these probabilities are based on human judgement, the modelers were confident based on expert analysis that these values are close enough to depict the voting behavior in the EU.

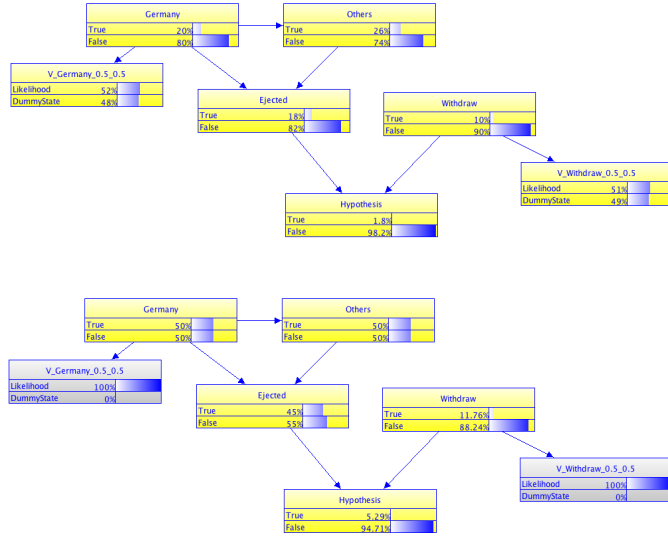


Figure 2: The likelihood updating in the Greek Exit case study. Top: Evidence nodes Q_{111} and Q_{112} are continuously updated with the estimates from the prediction market; Bottom: Updating the hypothesis estimate through BN propagation.

Only three of these nodes were posed in prediction market: Q_1 , Q_{12} and Q_{111} . The others remained “hidden variables” in the model. Therefore our procedure was, for each edit on Q_{11} or Q_{12} :

1. Read $p \leftarrow \pi(Q_1)$, the raw market output.
2. Create Q'_{111} and/or Q'_{112} with appropriate likelihood ratios.
3. Calculate $q \leftarrow \pi(Q_1|Q'_{111}Q'_{112})$.
4. Record p, q .
5. Absorb Q'_{111} and Q'_{112} .

As you can see, by March 29th, the date when we added the auxiliary questions, the market had already settled on less than a 10% chance of a Greek exit, and in fact it did not happen by the deadline. The all-time Brier Score for (Q_1) was 0.10632, but the Brier Score from March 29th onwards was 0.002455.

As we will see, we also have current versions of these questions on the market, and continue to collect forecasts on the continued possibility of a Greek exit.

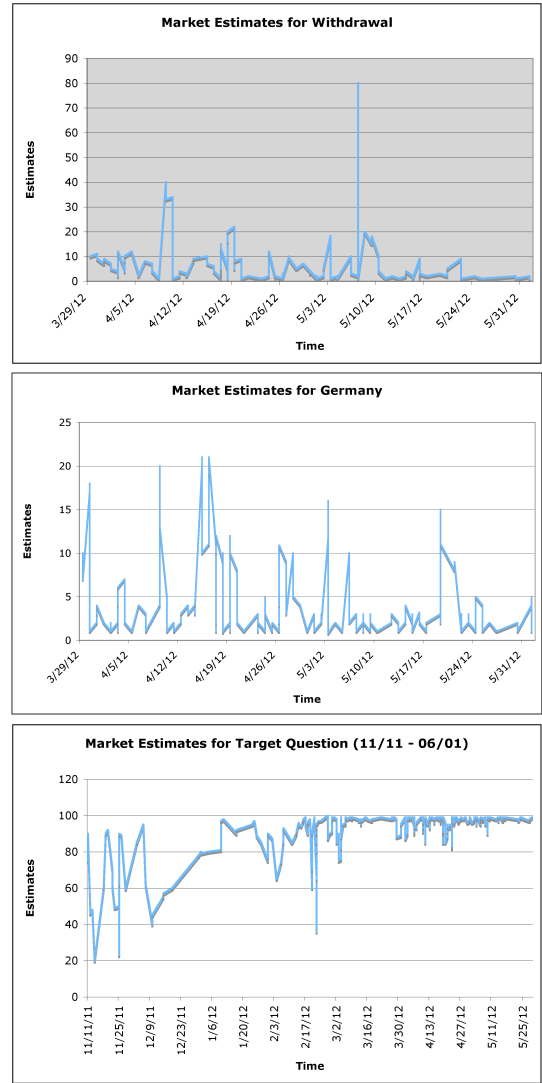


Figure 3: The prediction market estimates for the 3 questions in the Greek Exit case study, starting March 29th when we added the auxiliary questions. Top: Q_{12} = “Withdrawal”; Middle: Q_{111} = “Germany”; Bottom: Q_1 = core hypothesis, “Gexit”.

Preliminary Results

Figure 4 shows the differences between the market $\pi(Q_1)$ and the BN $\pi'(Q_1)$ from March 29th. Even though the market had largely settled on the correct answer, the BN can be seen to smooth the estimates, and in this case to improve them.

In this test case, forecasting a question based on causal or evidential nodes as a result of a Bayesian decomposition is better than directly estimating the hypothesis on the market. Figure 5 shows that the fluctuation in estimation converges towards zero over time. This is common, but not universal.

As shown in the table below, the average per-edit Brier Score for the BN is two orders of magnitude smaller (better) than the prediction market. However, that is because it could

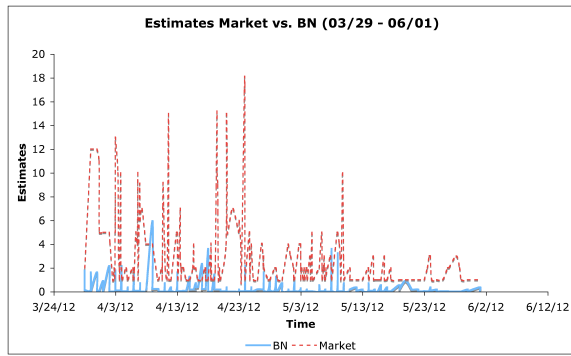


Figure 4: The prediction market estimates $\pi(Q_1)$ and Bayesian network estimates $\pi'(Q_1)$ of the Greek Exit case study.

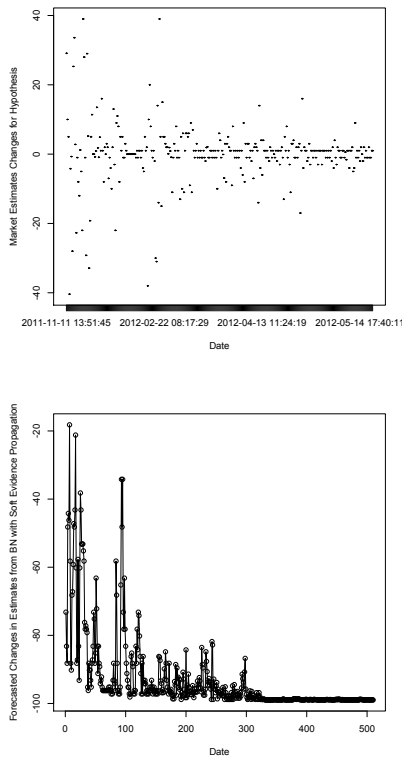


Figure 5: Changes in estimates for the target question. Top: the prediction market, $\Delta\pi'(Q_1)$; Bottom: the difference between the BN estimations and the market estimations ($\pi(Q_1) - \pi'(Q_1)$).

be far more extreme than market forecasters who were limited to values no lower than 1%. But if we similarly clip the BN, its score is still an order of magnitude better.

Brier Scores for March 29 – June 1

Prediction Market	0.002455	(10^{-3})
Actual BN	0.000032	(10^{-5})
BN Limited to 1%	0.000124	(10^{-4})

Ongoing Forecasts

While this particular question expired and thereby resolved “no”, the larger question about whether Greece will leave the Eurozone or the EU remains. We have now re-issued the question with an expiration date of 1 April 2013, and we have responded by creating the two auxiliary questions again. Figure 6 compares the market forecast on the new target question with the BN-generated estimates derived from market values of the auxiliary questions.

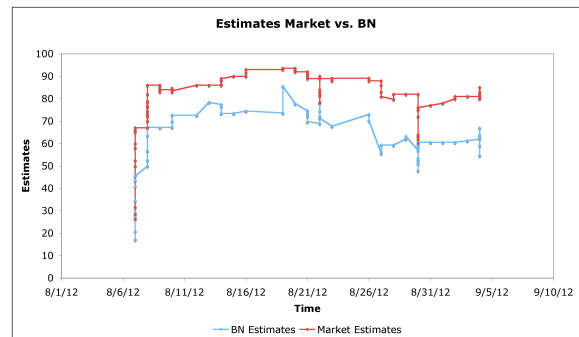


Figure 6: Market vs BN for a new Greek exit question expiring 1 April 2013. (Snapshot as of September 4, 2012.)

This time, the market values are mostly below 90%, and the BN is being more cautious than the participants. Possibly that is due to flaws in the model. Possibly it is due to relative neglect of the auxiliary questions in favor of the target question. But possibly, it’s because the forecasters are not keeping their own beliefs coherent.

We will soon be applying this modeling approach to other questions and topic areas. In order to do that, there are several steps that need to be taken in the future research.

Future Work

One of these steps is to allow for the conditional probabilities to be informed by the market as well instead of keeping them constant as originally defined by human judgment. This capability is provided by our combinatorial market. Another step is to replicate and improve this model with respect to several countries in the EU, since the question of a EU breakdown is becoming more acute in the news and mass media. Another future research direction is to parametrize the “noise” node with respect to voting/decision power in star type networks (as currently is EU). Another possible extension of this application is to modify this template in order to accommodate for other types of decision networks.

As indicated earlier in the paper, we are near the beginning of a multiyear campaign of experimentation. That said, we are already gleaning some insights from the work that increases our confidence in the usefulness of this approach. There are very positive indicators in two key areas: first,

there is strong evidence that model templates are possible and that they will be useful as anticipated and the templates for regime change and election outcomes have already been developed. They will be tested in the market during this year's experimental campaign.

Perhaps more important than the efficiencies that appear possible through the use of the templates is the potential for faster belief revision. We know from experimental psychology, that people generally revise their beliefs too slowly in the face of new evidence, compared to a Bayesian ideal. That may reflect judicious skepticism about their own mental (or statistical) model of the situation, but it could also be plain bias. If the analyst or analysts express their likelihoods for evidence under alternate hypotheses, and then observe the evidence, Bayesian updating will often lead to faster revisions.

For example, this question resolved as “no”: Greece was still a member of the EU on June 1st 2012. But, during the time period leading up to the country's elections held in the spring, the outcome of the election was very uncertain. That election outcome likely would have significantly influenced the probability that Greece would have voted to withdraw from the EU. The uncertainty surrounding the election that indirectly related to the target question was very real at the time and two things are likely true: 1) a different outcome on the election could have caused the target question to resolve differently and 2) this model would have responded quickly to such changes.

There a number of open research questions requiring further exploration. Of course the first is the principal hypothesis of the research: does target question decomposition and the use of Bayesian Networks improve the state of the art in forecasting? A key emerging question is: “What does better mean?”. For these longer-term questions where initial probabilities are very uncertain, it may be the case that movement in the probability is more significant than the actual value. This is especially true if the goal is to provide indications and warning for decision makers on low probability events.

Over the next year, our campaign of experiments will include multiple sets of Bayesian Networks interacting with the market to address longer term questions. Additionally, we expect to continue the work on model elicitation techniques, continuing to seek more efficient approaches to developing the initial models for placement in the market.

Over the longer term, for the next two to three years, our goal is to explore how to best integrate a BN-based algorithm with the other classes of automated agents and techniques being developed by DAGGRE. Our overall goal is to maximize the accuracy, timeliness, and value of intelligence forecasts.

Acknowledgments

The authors are very grateful for the software and coding support provided by Shou Matsumoto. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20062. 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.

References

- Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 75:1-3.
- Jeffreys, H. 1946. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*. 186 (1007): 453–461.
- Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers.
- Matsumoto, S., Carvalho, R.N., Ladeira, M., da Costa, P.C.G., Santos, L.L., Silva, D., Onishi, M., Machado, E., and Cai, K. 2011. UnBBayes: a Java Framework for Probabilistic Models in AI. In *Java in Academia and Research*. iConcept Press. <http://unbbayes.sourceforge.net>
- Cameron, W. B. 1963. *Informal Sociology: A Casual Introduction to Sociological Thinking*. Random House.
- Hanson, R. 2003. Combinatorial information market design. *Information Systems Frontiers*, 5(1), 107119.
- Hanson, R. 2007. Logarithmic market scoring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1), 315.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
- Sticha, P. J., Buede, D. M., and Rees, R. L. 2006. It's the people, stupid: the role of personality and situational variables in predicting decisionmaker behavior. HumRRO Technical Report based on their presentation at the 2005 Proceedings of the Military Operations Research Society Symposium (73rd MORSS).
- Sun, W., Hanson, R., Laskey, K.B. and Twardy, C.R. 2012. Probability and Asset Updating Using Bayesian Networks for Combinatorial Prediction Markets. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*. Catalina, CA.
- Surowiecki, J. 2005. *The Wisdom of Crowds*. Anchor Books.