

Editor: **V.S. Subrahmanian,** University of Maryland, vs@cs.umd.edu

# Model-Based Forecasting of Significant Societal Events

**Naren Ramakrishnan, Chang-Tien Lu, Madhav Marathe, Achla Marathe, Anil Vullikanti, Stephen Eubank, Scotland Leman, and Michael Roan,** *Virginia Tech*
**John S. Brownstein,** *Harvard Medical School*
**Kristen Summers,** *IBM*
**Lise Getoor,** *University of California, Santa Cruz*
**Aravind Srinivasan,** *University of Maryland, College Park*
**Tanzeem Choudhury,** *Cornell University*
**Dipak Gupta,** *San Diego State University*
**David Mares,** *University of California, San Diego*

Intelligence analysts today are faced with many challenges, chief among them being the need to fuse disparate streams of data and rapidly arrive at analytical decisions and quantitative predictions for use by policy makers. A forecasting tool to anticipate key events of interest is an invaluable aid in helping analysts cut through the chatter.

Our team is a university–industry partnership developing advanced forecasting algorithms for significant societal events such as disease outbreaks,[1,2] elections, domestic political crises, and civil unrest incidents.[3] Forecasting disease outbreaks spans both counts of influenza-like illnesses[1] and discrete incidents of rare diseases such as Hantavirus.[2] Civil unrest forecasting spans protests, strikes, and "occupy" events. Our system, Embers (Early Model-Based Event Recognition using Surrogates), is an automated environment to ingest myriad data streams and process them into alerts about population-level events of interest. The scope of Embers spans several countries of Latin America—namely, Argentina, Bolivia, Brazil, Chile, Costa Rica, Colombia, Ecuador, El Salvador, French Guiana, Guatemala, Honduras, Mexico, Nicaragua, Paraguay, Panama, Peru, Uruguay, and Venezuela. Our team includes researchers in data mining, machine learning, natural language processing, network dynamics, computational epidemiology, political science, systems integration, and Latin American studies.

The Embers-generated forecasts (also called alerts) are fine-grained in that they qualify the who, why, where, and when of an event. For instance, "Teachers will protest for wage-related reasons in the city of Curitaba, Brazil, this coming Wednesday" is an example of an alert. Forecasting the dates, locations, and participating populations in this manner can offer situational awareness into unfolding events. In addition, aggregating this information and the data that supports it can offer insights into the broader sociocultural environment. For example, an analyst who sees an increase in protests in a given population might examine the source data and find that certain ongoing issues, such as crime rates, are starting to produce more specific unrest than in the past, which in turn would spur analysis and insights of the factors affecting the events.

Embers has been generating alerts continuously since November 2012 without a human in the loop, as is the requirement of the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program supporting the development of Embers. Unlike retrospective studies of predictability, alerts generated by Embers are emailed in real time to IARPA and must precede the event being forecast to count as a prediction. The received alerts are evaluated monthly by an independent test and evaluation team (MITRE). Analysts and subject matter experts at MITRE survey international and domestic newspapers of record in each country that Embers studies and catalog a master set of events in these countries, known as the gold standard report (GSR).

Our goal in this article is to outline some salient aspects of Embers through its design considerations, system architecture, and user interface.



1541-1672/15/$31.00 © 2015 IEEE
*Published by the IEEE Computer Society*

## Predictive Analysis Using Surrogates

A central approach to forecasting in Embers is predictive analysis using surrogates. Surrogates are indirect indicators that are correlated with an event of interest. For instance, the idea of tracking flu activity using search query data, as done in Google Flu Trends, is a modern example of forecasting using surrogates. A second example is using nighttime luminosity data to quantify economic output of countries. Identifying and computing such surrogates is a key research topic in Embers.

Embers uses a range of surrogates for specific event classes of interest. For instance, in forecasting disease outbreaks, we use surrogates ranging from social media (Twitter activity) to physical indicators (such as humidity or vegetation index). We also explored unconventional surrogates, such as restaurant reservation and availability information from OpenTable[4] and parking lot imagery from hospitals.[5] Monitoring changes in restaurant use could potentially serve as a leading indicator of disruption; in particular, a decrease in restaurant use could serve as an early indicator of a disease-related event. Similarly, we showed that spikes in parking lot fill rate in general care hospitals can be used in a regression model to forecast influenza-like illness counts.[5] For forecasting civil unrest,[3] we explored a range of data sources, such as news, blogs, Twitter, Facebook, Wikipedia edits, and economic data. Two unconventional sources we explored include TOR (The Onion Router) routing statistics (that is, counts of users who opt for anonymous communication on the Internet) and the percentage of smiles in photos shared over social media.

## Model-Based Forecasting

A second theme in Embers is the integration of model-based approaches to forecasting with phenomenological or statistical methods. For instance, in forecasting influenza-like illness case counts, we use compartmental epidemiological models to develop not just point estimates of weekly case counts but also forecasts of the long-term characteristics of the epidemic curve—for example, the start and end of the flu season, the time it would take for the season to peak, and the total number of infections to occur in the season. Such long-term forecasts are critical in informing public health responses by policy makers.

Likewise, in the case of civil unrest, we use a suite of six forecasting algorithms (see Figure 1) that posit different approaches to modeling civil unrest. The planned protest model aims to identify incidents of organized and preannounced protests from news and Twitter using language-processing techniques. A second model uses spatial scan statistics to identify geolocated clusters of tweets enriched with a defined vocabulary of keywords (identified by our team's social scientists). These geolocated clusters are tracked over time, and their characteristics (such as density and growth) are used to issue a forecast. The cascade regression model recognizes situations where social media, such as Twitter, is utilized as the staging ground for galvanizing support for protests via online recruitment to the underlying causes. The dynamic query expansion model is similar to the spatial scan model but aims to learn new emerging keywords, unlike the static vocabulary used by the spatial scan model. (In 2013, there were a series of protests in Venezuela due to a shortage of toilet paper, a novel circumstance that was uncovered using this model.) The volume-based Least Absolute Shrinkage and Selection Operator (Lasso) regression model considers every possible data source in Embers (news, tweets, blogs, economic indicators, TOR, and smiles) to forecast the imminence of protests in the next day or two. Finally, the maximum likelihood estimator (MLE) model aims to identify regularities in the GSR and provides a baseline performance level. Each of these models is tuned for high

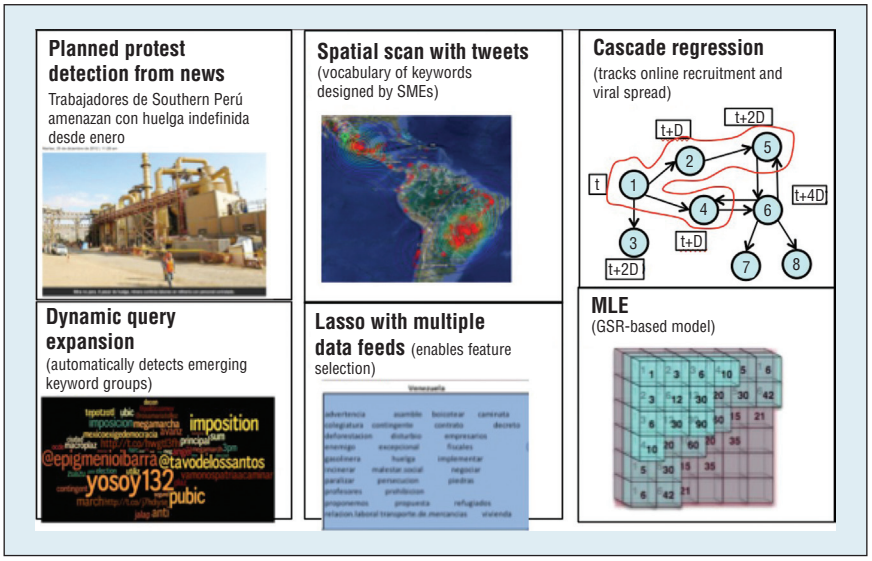


**Figure 1. Embers's suite of six forecasting approaches for civil unrest. Each model uses a different set of data sources and makes specific assumptions about conditions under which unrest happens. Results from these models are then fused into a final set of alerts. (Lasso: Least Absolute Shrinkage and Selection Operator; MLE: maximum likelihood estimator.)**

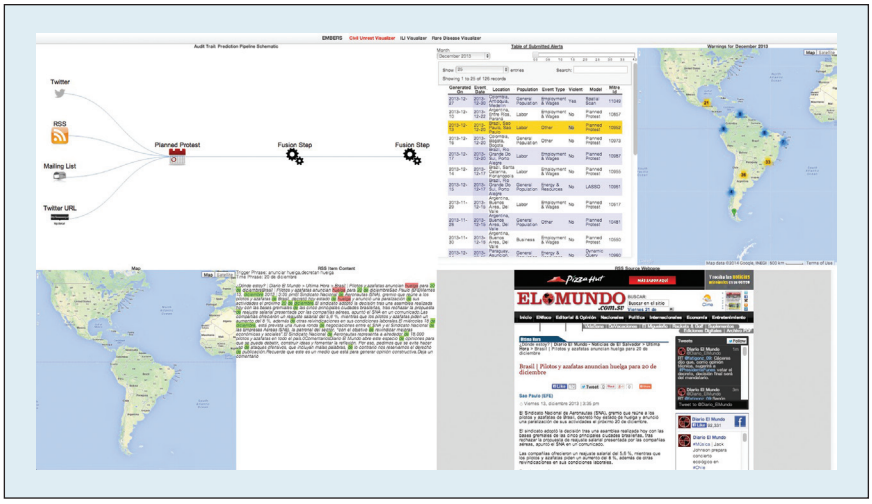


**Figure 2. Example Embers audit trail view. The top left panel shows the logical transduction of data sources into an alert. The top right panel depicts the selected alert for which the audit trail is shown. The bottom panels provide drill-down detail into the data sources used for alert generation.**

**Table 1. Embers performance metrics.**

| Program metric | Program target (at 24 months) | Embers score |
|---|---|---|
| Lead time | 3 days | 7.54 days |
| Quality score | 3.25 | 3.09 |
| Precision | 0.65 | 0.94 |
| Recall | 0.65 | 0.65 |
| Probability score | 0.7 | 0.89 |

precision, and their fusion aims to achieve high recall.

## Probabilistic Collective Reasoning

Several aspects of Embers require probabilistic collective reasoning, and we leverage the framework of probabilistic soft logic (PSL).[6] For instance, identifying locations from texts and tweets and tracking geolocated sentiment provides a useful view of the human terrain in an area. Embers uses a PSL program to perform collective reasoning over the underlying network of articles, mentions, and hyperlinks and to obtain from this surrounding context greater specificity into geolocation. Similarly, we also use PSL in Embers to discover evolving political vocabulary by harnessing the network of relationships underlying hashtags, Twitter users, and retweeting relationships.[7]

## Information Fusion under Uncertainty

A final theme in Embers is the development of a fusion engine that integrates initial alerts from different models to yield the final set of generated alerts. This can be viewed as a Bayesian system integration problem: given an alert from a model, should we suppress it, issue it, or merge it with a previously issued alert? The problem is nontrivial because models are trained to generate high-quality alerts in the individual, but the overall system performance depends on the final set of issued alerts. Statistical decision theory is used wherein loss functions associate costs with each type of decision. The resulting framework enables the tuning of precision and recall by enabling the analyst to issue a smaller or greater number of alerts as appropriate. Greater recall would be useful for an analyst who thinks of using Embers in an analytic triage—that is, a system to produce initial alerts for further human review. Greater precision would be appropriate for an analyst exploring a hypothesis about a specific population group, for instance.

## System Architecture

The Embers architecture provides a platform for continuous enrichment and interpretation of incoming data sources. It implements a share-nothing, message-based streaming architecture using 0MQ as the underlying method of data transport. Processing components are distributed among virtual machines. The system's highly modular and loosely coupled architecture allows for the ready incorporation of updates and new modules and ingestion of new data sources, while enabling seamless interaction among the components. It runs in the Amazon Cloud, and the current production cluster consists of 12 Elastic Compute Cloud (EC2) instances, with two dedicated to ingest processing, four dedicated to message enrichment, four dedicated to predictive modeling and warning generation, and one each dedicated to archiving and system monitoring. Deployments and updates are completely automated. Embers provides audit trail capabilities to inspect the rationale for an alert as well as ablation tools to pose what-if questions about the addition and removal of data sources (to investigate how such alterations affect the issuance of alerts; see Figure 2).

## Evaluation

Monthly scoring reports evaluate Embers's forecasts against the GSR by conducting a bipartite matching between events and alerts, with inclusion criteria for matching events to alerts. The bipartite matching aims to optimize the quality score, a measure of the match

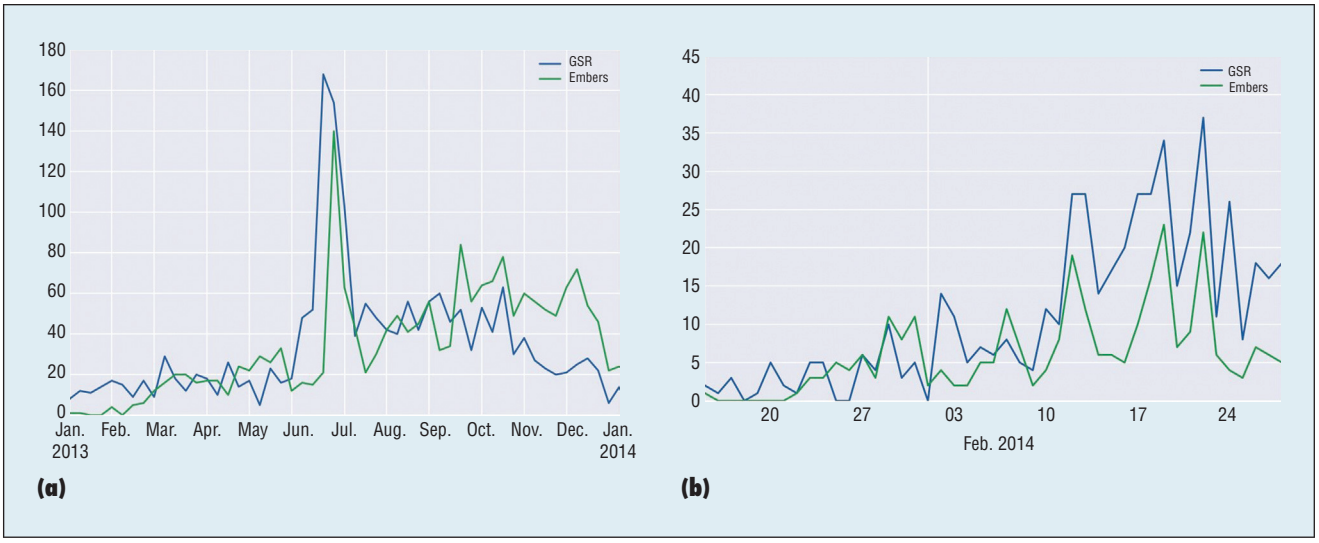


**Figure 3. Embers civil unrest performance during recent significant periods of unrest in (a) Brazil and (b) Venezuela. The curves are aligned with the date of the (actual or predicted) protest, but note that forecasts are made with a lead time of at least a day. Embers accurately forecast the order-of-magnitude increase in the number of protests during the "Brazilian Spring" (2013) as well as the multiple upticks of unrest during student-led protests in Venezuela (2014).**

between alerts and events according to the four dimensions (what, where, who, and why). Because alerts and events are time stamped, causal consistency imposes an interesting "noncrossing" constraint on the matchings produced, which our algorithms handle. Once a maximum bipartite matching is computed, several overall criteria can be read from such a matching—such as precision, recall, lead time, average quality score, and average probability score. Precision and recall have their accepted interpretations. Lead time refers to the average number of days that a warning is issued ahead of the event that it matches. The average quality score is simply the average of the quality scores across matched (warning, event) edge pairs. The average probability score is a measure of the system confidence computed as the Brier score.

Table 1 gives Embers's performance metrics over civil unrest events for a recent month (and program targets). Figure 3 shows an aggregated view of our forecasting performance; we present our forecasts of civil unrest in Brazil and Venezuela in recent periods compared against the GSR. Embers forecast the upticks of unrest in both countries.

The use of open source indicators to forecast significant societal events is fast becoming an established methodology in anticipatory intelligence (for example, see Steven Banaszak and colleagues for an approach to forecasting domestic stabilities of countries[8]). Our future work is organized around the tighter integration of social science theory-based reasoning with data-driven forecasting algorithms and the continued identification of novel surrogates. We are also interested in generalizing Embers's scope to other population-level events such as mass migrations.

We anticipate that systems like Embers will continue to mature and become an integral part of analysts' workflows. ■

## Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337. The US Government is authorized to reproduce and distribute reprints of this work for governmental purposes notwithstanding any copyright annotation thereon. 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 US Government.

## References

1. P. Chakraborty et al., "Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions," *Proc. SIAM Int'l Conf. Data Mining*, 2014; doi:10.1137/1.9781611973440.30.
2. T. Rekatsinas et al., "SourceSeer: Forecasting Rare Disease Outbreaks using Multiple Data Sources," *Proc. SIAM Int'l Conf. Data Mining*, 2015; http://linqs.cs.umd.edu/basilic/web/Publications/2015/rekatsinas:sdm15.
3. N. Ramakrishnan et al., "Beating the News with EMBERS: Forecasting Civil Unrest using Open Source Indicators," *Proc. 20th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2014, pp. 1799–1808.
4. E.O. Nsoesie, D.L. Buckeridge, and J.S. Brownstein, "Guess Who's Not Coming to Dinner? Evaluating Online Restaurant Reservations for Disease Surveillance," *J. Medical Internet Research*, vol. 16, no. 1, 2014, p. e22.
5. E.O. Nsoesie et al., "Modeling Disease Trends Using High Resolution Satellite Imagery: A Feasibility Study," *Scientific Reports*, vol. 5, Mar. 2015, article 9112.
6. A. Kimmig et al., "A Short Introduction to Probabilistic Soft Logic," *NIPS*

*Workshop Probabilistic Programming: Foundations and Applications*, 2012.
7. B. Huang et al., "Social Group Modeling with Probabilistic Soft Logic," *NIPS Workshop Social Network and Social Media Analysis: Methods, Models, and Applications*, 2012.
8. S. Banaszak et al., "Forecasting Country Stability in North Africa," *Proc. IEEE Joint Intelligence and Security Informatics Conf.*, 2014, pp. 304–307.

**Naren Ramakrishnan** is the Thomas L. Phillips Professor of Engineering and director of the Discovery Analytics Center at Virginia Tech. Contact him at naren@cs.vt.edu.

**Chang-Tien Lu** is an associate professor of computer science at Virginia Tech. Contact him at ctlu@vt.edu.

**Madhav Marathe** is the director of the Network Dynamics and Simulation Science Laboratory and a professor at the Virginia Bioinformatics Institute and the Department of Computer Science at Virginia Tech. Contact him at mmarathe@vbi.vt.edu.

**Achla Marathe** is a professor at the Virginia Bioinformatics Institute and in the Department of Agricultural and Applied Economics at Virginia Tech. Contact her at amarathe@vbi.vt.edu.

**Anil Vullikanti** is an associate professor at the Virginia Bioinformatics Institute and in the Department of Computer Science at Virginia Tech. Contact him at akumar@vbi.vt.edu.

**Stephen Eubank** is deputy director of the Network Dynamics and Simulation Science Laboratory, a professor in the Department of Population Health Sciences, and an adjunct professor in the Department of Physics at Virginia Tech. Contact him at seubank@vbi.vt.edu.

**Scotland Leman** is an associate professor of statistics at Virginia Tech. Contact him at leman@vt.edu.

**Michael Roan** is an associate professor in and associate head of the Mechanical Engineering Department at Virginia Tech. Contact him at mroan@vt.edu.

**John S. Brownstein** is Chief Innovation Officer at Boston Children's Hospital and an associate professor at Harvard Medical School. Contact him at john.brownstein@childrens.harvard.edu.

**Kristen Summers** is a senior managing consultant in IBM's Watson group. Contact her at kmsummer@us.ibm.com.

**Lise Getoor** is a professor of computer science at the University of California, Santa Cruz. Contact her at getoor@soe.ucsc.edu.

**Aravind Srinivasan** is a professor of computer science at the University of Maryland, College Park. Contact him at srin@cs.umd.edu.

**Tanzeem Choudhury** is an associate professor in the department of information science at Cornell University. Contact her at tanzeem.choudhury@cornell.edu.

**Dipak Gupta** is Distinguished Professor Emeritus in the Department of Political Science at San Diego State University. Contact him at dgupta@mail.sdsu.edu.

**David Mares** holds the Institute of the Americas Chair for Inter-American Affairs at the University of California, San Diego. Contact him at dmares@ucsd.edu.