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

Prediction Using Propagation: From Flu Trends to Cybersecurity

B. Aditya Prakash, *Virginia Tech*

Graphs or networks are ubiquitous, from online social networks, communication networks, hospital networks, and gene-regulatory networks to router graphs. Many processes and situations in real life, such as social systems, cybersecurity, epidemiology, and biology, can be modeled using them. They effectively model many phenomena because they simultaneously expose local dependencies and capture large-scale structure. Additionally, propagation (diffusion) processes—those in which an agent's state (or action) depends on its neighbors' states (or actions)—over networks can give rise to a wide array of macroscopic behavior, leading to challenging and exciting research problems. How do contagions like Ebola and influenza spread in population networks?[1] Which group should we market to for maximizing product penetration? How do we place sensors in a water-distribution network? How do rumors spread on Twitter or Facebook? All of these questions are related to propagation phenomena on networks.

Social network websites like Facebook count millions in users and revenue. Hospital-acquired infections take thousands of lives and cost billions of dollars per year. The societal impacts of networked collaboration during political events such as the Arab Spring have been well documented, too. Cybersecurity is also a serious national economic issue right now. Hence, research in this area, helping us answer questions like how information spreads through social media[2] and how to distribute antibiotics across hospitals,[3,4] holds great scientific, social, and commercial value.

This article will examine recent efforts at utilizing propagation-based concepts for predicting flu trends using public Twitter data.[5,6] In addition, we will also briefly discuss leveraging propagation for malware count prediction[7,8] using extensive field datasets.

Syndromic Surveillance of Flu

Machine learning techniques for "nowcasting" the flu have made significant inroads into correlating social media trends to case counts and the prevalence of epidemics in a population. Web searches and social media sources such as Twitter and Facebook have emerged as surrogate data sources for monitoring and forecasting the rise of public health epidemics. The celebrated example of such surrogate sources is arguably Google Flu Trends (GFT), which harnessed user query volume for a handcrafted vocabulary of keywords in order to yield estimates of flu case counts. Such surrogates thus provide an easy to observe, indirect approach to understanding population-level health events. However, recent research has noted GFT's lackluster performance,[9] which could be attributed to it not accounting for the evolving nature of the vocabulary, and a lack of transparency about which keywords are used, which affects verification.

Motivated by such considerations, we aim to better bridge the gap between syndromic surveillance strategies and contagion-based epidemiological modeling. We focus on Twitter data from 15 South American countries for this purpose. Diseases such as the flu have been traditionally modeled as a propagation process on population contact networks using models such as SI (Susceptible-Infected) and SEIS (Susceptible-Exposed-Infected-Susceptible).[1] Current methods do not use this observation for prediction. Using just keywords to track infected users on Twitter cannot distinguish between users belonging to these different epidemiological phases. For example, tweets such as "Down with flu. Not going to school." and "Recovered from flu after 5 days, now going to the beach" denote the users'

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

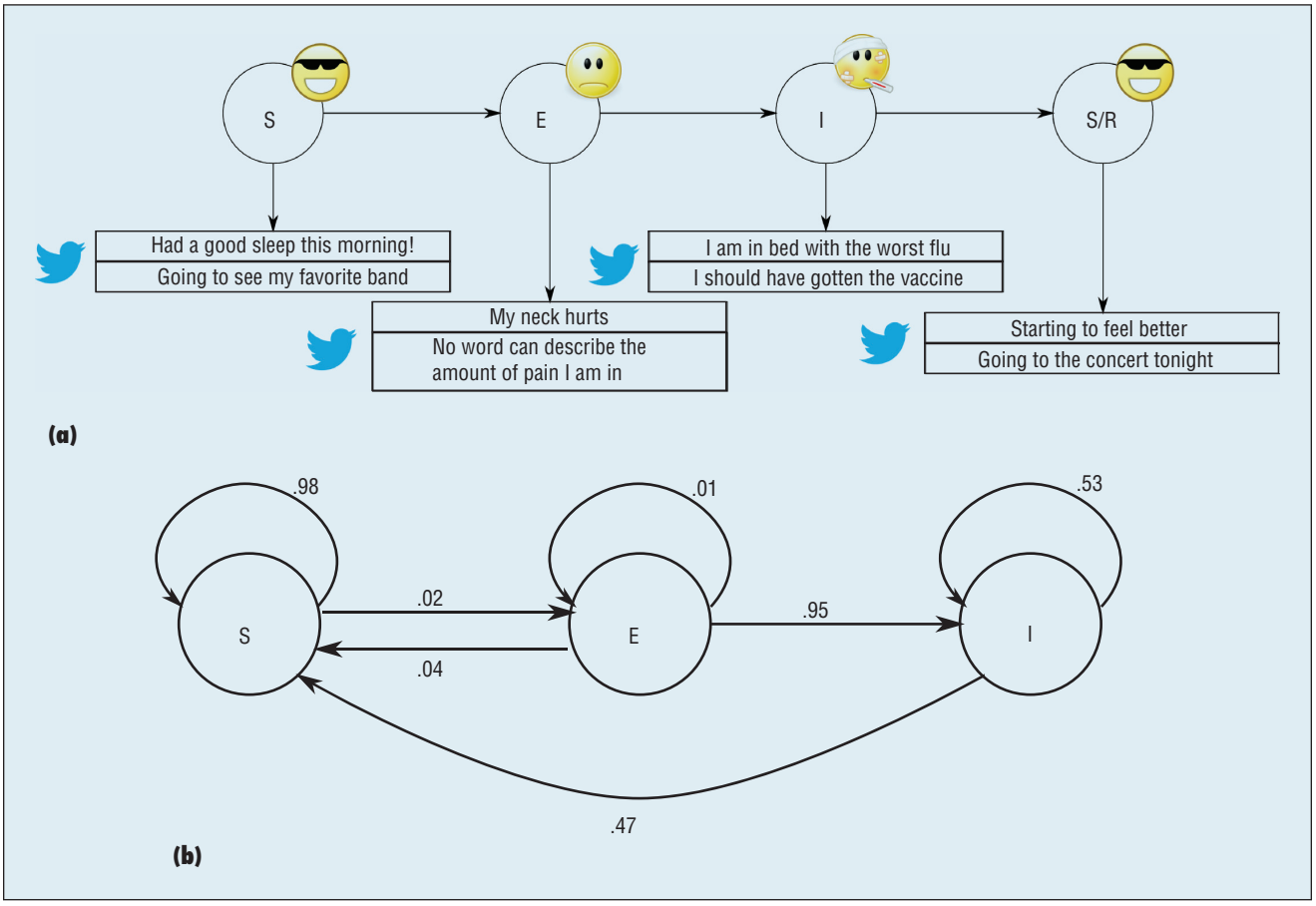

Figure 1. Comparison between expected state transition and the state transitions learned by our model. (a) A toy example showing possible user states and a tweet associated with each state. (b) State transition probabilities learned by our method.[5] Note that the state transition probabilities learned by our method match with the expected epidemiological SEIS model.

different epidemiological states (see also Figure 1a).

We show that we can separate out these states from the tweets using a temporal topic model. This not only helps in interpretability, but it also leads to more accurate predictions of flu-case counts robust to noisy vocabularies. The key idea is to assume different generating topic distributions for users in each epidemiological phase, and then assume Markov-chain-style transitions between the states. We also assume the presence of background topics and non-flu-related topics that do not denote any flu-related state. We can then fit this model to training data using standard methods (we used Expectation-maximization; others, such as Gibbs-sampling, could also be used). We show the state transition learned by our model HFSTM (Hidden Flu State from Tweet Model) automatically on the real data in Figure 1b; it matches well with the standard SEIS model.

Figure 2 shows the most frequent words for each learned state distribution via a word cloud. Again, the words meaningfully correspond to the states. In addition, thanks to the finer-grained modeling, our approach gets better predictions of the incidence of flu cases than direct keyword counting and also sometimes gets better predictions of flu peaks than sophisticated methods such as GFT (see Figure 3). Our original model used unsupervised topic modeling, so it needed an initial clean flu-related vocabulary. However, we extended it by using semisupervised models in which words in the vocabulary can have different aspects (for example, flu or non-flu related). Intuitively, this way words get a soft assignment instead of the hard assignment we had originally. As a result, we could robustly learn states and topics even with an enlarged and noisier vocabulary,[6] which will also help mitigate the effects of the changing nature of the vocabulary in real deployment.

Malware Surveillance

Similarly, propagation-based concepts can also play an important role in cybersecurity. In the security sphere, such problems include understanding

Figure 2. The translated word cloud for the most probable words in the (a) S, (b) E, and (c) I state-topic distributions, as learned by our method. Words are originally learned and inferred in Spanish; we then translate the result using Google Translate for ease of understanding. The size of a word is proportional to its probability in the corresponding topic distribution. Our model can tease out the differences in the word distributions between them.

the propagation of malware (such as estimating the number of machines infected) or characteristics of benign files. These questions have numerous implications for cybersecurity, from designing better antivirus software to designing and implementing targeted patches to more accurately measuring the economic impact of breaches. These problems are compounded by the fact that, as externals, we can only detect a fraction of actual malware infections.

To answer such problems, security researchers and analysts are increasingly using comprehensive, field-gathered data that highlights the current trends in the cyber-threat landscape. We have been working on Symantec's Worldwide Intelligence Network Environment (WINE) data for precisely this purpose. This data is collected from real-world hosts running their consumer antivirus software. Users of Symantec's consumer product line can opt in to report telemetry about the security events (for example, executable file downloads or virus detections) that occur on their hosts. The events included in WINE are representative of events that Symantec observes around the world, and they do not include personally identifiable information.

Our approach has been to leverage generative propagation-based models, sometimes in conjunction with careful feature engineering, to better predict trends and actual estimates of malware infections.[7,8] As the models are generative, their parameters can also serve as features for further analytics tasks such as anomaly detection. Our ideas included having specific phases matching domain-based constraints (for example, having different "infected" versus "detected" versus "patched" states), or exploring nonexponential residence times in each state. After building the model, we fit it by minimizing the least-square errors using standard nonlinear numerical methods (see Figure 4). One lesson we learned was

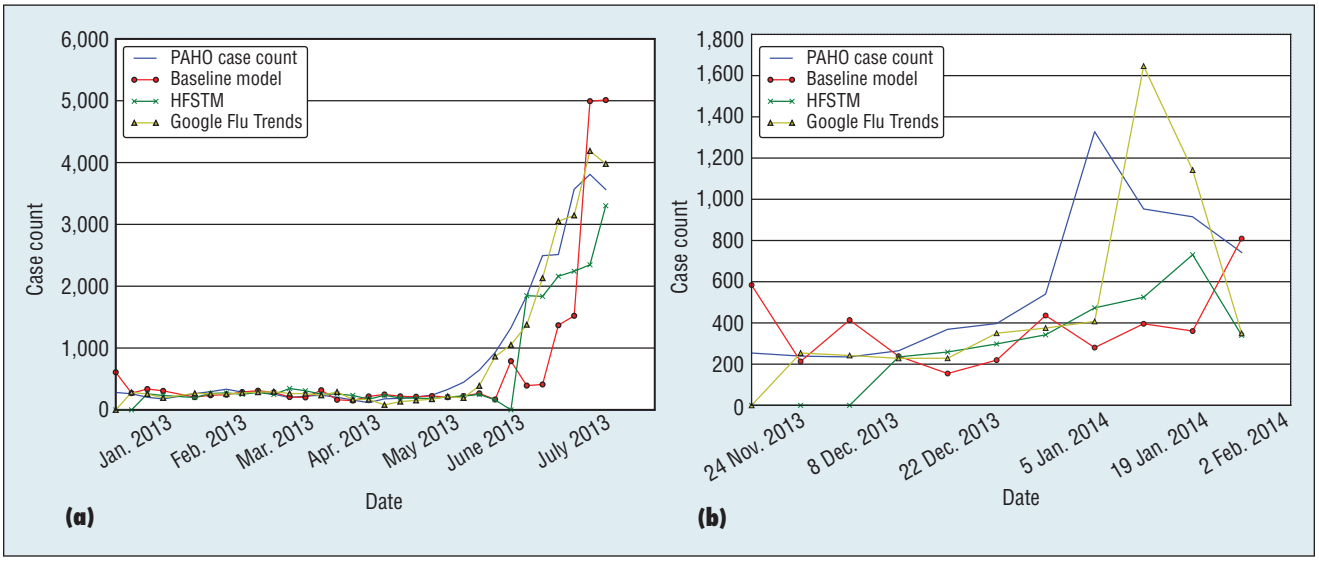

Figure 3. Evaluation for the two test scenarios: (a) test period 1 and (b) test period 2. Comparison of the week-to-week predictions against ground truth Pan American Health Organization (PAHO) case counts using the three models: a baseline model, which does simple keyword counting; our method, HFSTM; and Google Flu Trends (GFT). Our model outperforms the baseline and is comparable to GFT, beating it in test period 2. GFT overestimates the peak in both test periods.

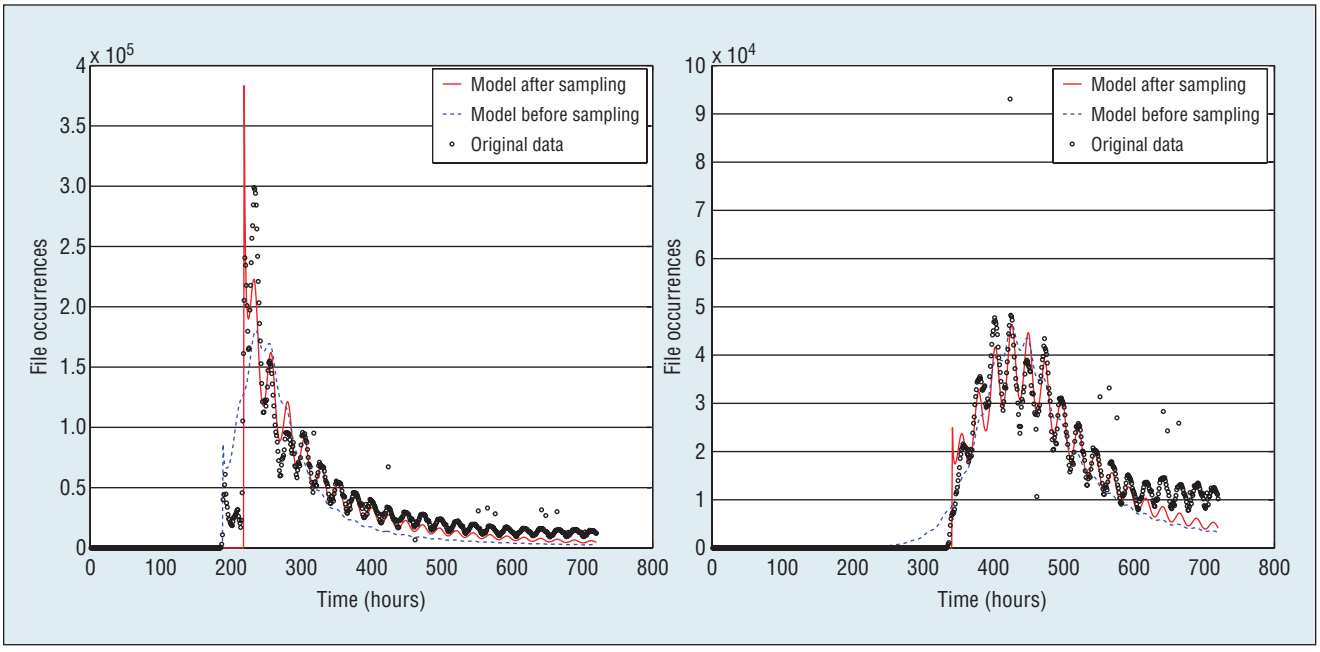

Figure 4. Our propagation-based model[7] fits real data from Symantec's Worldwide Intelligence Network Environment (WINE) database about malware infections per unit time very well, both before and after sampling. The median relative standard error in this case was 0.0741.

that such models typically work for high-volume files (that is, files that have enough samples such that any form of meaningful modeling is possible). For low-prevalence files, feature-based approaches tend to give low prediction errors. Thus, we further improved our predictions and made them more robust by building ensemble methods that combine the best of both generative and feature-based models.[8]

This is a diverse area, because propagation and networks occur in many different applications. The recent explosion in the availability of large-scale datasets presents a unique opportunity to conduct large-scale predictive studies

using these models. There are many open problems: for example, in the online sphere, similar questions can be posed about predicting how memes spread over blogs and websites. Here, too, propagation-inspired models tailored to the application (for example, by incorporating the human response-time distributions)[2] can be useful in outperforming other standard time-series analysis tools. Overall, there is rich overlap of propagation with many areas in data mining, and we envision many more use cases for such models in the future. ■

Acknowledgments

We thank Symantec for providing access to the WINE platform. This article is based on work partially supported by the National Science Foundation under grant number IIS-1353346, the Maryland Procurement Office under contract H98230-14-C-0127, the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, a Facebook Faculty Gift, and the VT College of Engineering. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the respective funding agencies.

References

1. R.M. Anderson and R.M. May, *Infectious Diseases of Humans*, Oxford Univ. Press, 1991.
2. Y. Matsubara et al., "Rise and Fall Patterns of Information Diffusion: Model and Implications," *Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2012, pp. 6–14.
3. B.A. Prakash et al., "Fractional Immunization over Large Networks," *Proc. SIAM Data Mining Conf.*, 2013, pp. 659–667.
4. Y. Zhang and B.A. Prakash, "Data-Aware Vaccine Allocation over Large Networks," *ACM Trans. Knowledge Discovery from Data*, vol. 10, no. 2, 2015, pp. 20:1–20:32.
5. L. Chen et al., "Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models," *Proc. IEEE Int'l Conf. Data Mining*, 2014, pp. 755–760.
6. L. Chen et al., "Syndromic Surveillance of Flu on Twitter Using Weakly Supervised Temporal Topic Models," *Data Mining and Knowledge Discovery*, 2015, pp. 1–30.
7. E.E. Papalexakis et al., "Spatio-temporal Mining of Software Adoption & Penetration," *Proc. IEEE/ACM Int'l Conf. Advances in Social Networks Analysis and Mining*, 2013, pp. 878–885.
8. C. Kang et al., "Ensemble Models for Data-Driven Prediction of Malware Infections," to be published in *Proc. Int'l Conf. Web Search and Data Mining*, 2016.
9. D.M. Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, vol. 343, no. 6176, 2014, pp. 1203–1205.

B. Aditya Prakash is an assistant professor in the Computer Science Department at Virginia Tech. Contact him at badityap@cs.vt.edu.