Dr. Twitter: The Logistics of Practical Disease Surveillance using Social Media

Shannon P. Quinn, Arvind Ramanathan, Laura L. Pullum, and Chad A. Steed

***Abstract*—The confluence of increasing computational capacity, cheaper digital storage, and high-volume sources of digital health-related information has made large-scale biosurveillance frameworks for predicting threats to public health increasingly practical. These frameworks consume vast quantities of digital information from traditional (e.g., CDC) and so-called non-traditional (e.g., social media) sources to make rapid disease forecasts. However, many of these frameworks have technical or logistical problems that preclude their use on a large scale; these range from insufficiently expressive models to lack of practical scalability. The Oak Ridge Biosurveillance Toolkit (ORBiT) was built to address many of these technical and logistical shortcomings, but many more remain to be tackled. In particular, we'll discuss the modeling limitations of previous biosurveillance frameworks (e.g. Google Flu Trends) and what more robust models incorporate; we'll investigate common sources of noise in social media datasets and how they can be mitigated; finally, we'll consider the practicalities of scaling up such frameworks out of the research lab and into the real-time public health space.**

I. Introduction

The imminent threats from novel and emerging air-, water- and food-borne diseases that can potentially have devastating social and economic impact on widespread geographic regions within a short period of time underscores the importance for developing effective early-warning/forecasting systems that can enable rapid identification, analysis, and detection of these diseases. Official indicators of public health related data sources, including the national and international surveillance systems and other data repositories, often suffer from long turnaround times and low spatiotemporal resolution. Crowd-sourcing this information through social media platforms offers an attractive option, as it involves monitoring information from diverse, potentially high-volume, noisy data sources to identify emerging threats [1]. The collective intelligence and social power of individuals augmented with information from public health agencies promises to provide actionable insights on emerging health threats.

However, these "non-traditional" information sources require significant preprocessing and sophisticated modeling techniques before gaining insights into emerging disease outbreaks. We hypothesize that analysis of data from social media, intelligently aggregated with trusted and more "traditional" sources of information (e.g. emergency room visits at hospitals and clinics, prescription sales data, etc.), can provide an improved, effective, and reliable early warning system and situational predictor. Google Flu Trends (GFT) is perhaps one of the most well-known biosurveillance framework, but its limitations were evident and likely contributed to its shuttering earlier this year [2]. GFT appeared to use a simplistic linear regressor that resulted in vast over- and under-estimates; researchers were able to use similar data with more expressive predictive models to obtain much more reliable and robust flu predictions over the same periods [3], [4].

Here, we discuss the technical and practical improvements ORBiT and other successful biosurveillance frameworks have implemented. For instance, ORBiT integrates multiple heterogeneous health care datasets to provide more robust estimates of ongoing and emerging threats to public health [5]. Diseases that have co-occurring incidence patterns, such as influenza and asthma, can be modeled jointly to provide a more holistic spatial and temporal model of the evolution of these conditions [6]. Sophisticated epidemiological models can be implemented to robustly estimate disease incidence under highly noisy datasets [7]. Finally, we examine some of the practical technologies and design paradigms that are required to bring these frameworks and models into large-scale use.

S. P. Quinn is with the Departments of Computer Science and Cellular Biology, University of Georgia, Athens, GA 30602 USA (phone: 706-542-4661; e-mail: squinn@cs.uga.edu).

A. Ramanathan, L. L. Pullum, and C. A. Steed are with the Computational Science and Engineering Division, Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (phone: 865-576-7266; e-mail: {ramanathana,pullumll,steedca}@ornl.gov).

References

[1] F. Gesualdo, G. Stilo, A. D'Ambrosio, E. Carloni, E. Pandolfi, P. Velardi, A. Fiocchi, and A. E. Tozzi, "Can twitter be a source of information on allergy? correlation of pollen counts with tweets reporting symptoms of allergic rhinoconjunctivitis and names of antihistamine drugs," *PloS one*, vol. 10, no. 7, p. e0133706, 2015.

[2] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 14 March, 2014.

[3] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen, "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific reports*, vol. 5, 2015.

[4] S. Yang, M. Santillana, and S. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14 473–14 478, 2015.

[5] A. Ramanathan, L. Pullum, C. A. Steed, S. S. Quinn, C. S. Chennubhotla, and T. Parker, "Integrating heterogeneous health care datasets and visual analytics for disease bio-surveillance and dynamics," in *4th Wkshp on Integrative Text & Visual Analytics, IEEE Conf on Visual Analytics (VAST), Atlanta*, 2013.

[6] A. Ramanathan, L. L. Pullum, T. C. Hobson, C. G. Stahl, C. A. Steed, S. P. Quinn, C. S. Chennubhotla, and S. Valkova, "Discovering multi-scale co-occurrence patterns of asthma and influenza with oak ridge biosurveillance toolkit," *Frontiers in public health*, vol. 3, 2015.

[7] A. O'Hare, R. Orton, P. R. Bessell, and R. R. Kao, "Estimating epidemiological parameters for bovine tuberculosis in british cattle using a bayesian partial-likelihood approach," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 281, no. 1783, p. 20140248, 2014.