New Data Paradigms: From the Crowd and Back

Rumi Chunara
Computer Science & Engineering and College of Global Public Health
New York University
New York, USA
rumi.chunara@nyu.edu

***Abstract*—Knowledge generation from citizens is becoming both more feasible as well as important. Data directly from individuals can be critical as it can add information beyond what is available otherwise. Crowdsourced data also is very amenable in open data efforts given the nature of its generation. In this talk I will describe several efforts in which we are generating crowdsourced knowledge from open data and using it to more readily improve knowledge in public health.**

***Keywords*-Crowdsourcing; inference algorithms; public healthcare;**

I. Introduction

Public health has had a culture of information-sourcing directly from individuals, whether through surveys or other means, and thus citizen science has naturally become intertwined in many public health efforts. For example, syndromic reporting for acute respiratory infections (flus and colds) is now in place in 11 countries in Europe (Influenzanet), Australia (Flutracking) and the United States and Canada (Flu Near You). The types of data being garnered from these growing systems is also expanding; spanning biological specimens (nasal, saliva, fecal) [1], [2], physiological data [3] and text-based data [4]. Many of these systems aggregate and share data in various forms. In particular we expand on a particular effort in sourcing molecular data directly from individuals as well as our development of methods to use open-data in disease models.

II. Crowdsourcing Molecular Disease Data

Participatory systems have been shown useful to address the need for real-time information about infections circulating in the community, however they have been limited to self-reported syndromic information only [4], [5]. Though important for early outbreak detection and epidemic monitoring, symptom information on it's own presents an incomplete picture of disease incidence. Without information about the viral etiology of the illness, viral burden and prevalence cannot be accurately estimated, and selection of appropriate treatment and prevention measures is limited. To address this challenge, we developed the GoViral platform [1]. The platform was intended to help gather data and evaluate whether a cohort of lay volunteers could, and would find it useful to, contribute the syndromic data as well as specimens for molecular analysis. Volunteers are recruited, given a kit (collection materials and customized instructions), instructed to report their symptoms weekly, and when sick with cold or flu-like symptoms, requested to collect specimens (saliva and nasal swab). In this talk we will discuss the system in more detail, as well as how the data is collected and shared.

III. Mobility Prediction from Open Data

In public health efforts, modeling of disease is an important challenge and as part of this, predicting individual-level mobility patterns is vital. While data sources such as GPS trackers or Call Data Records are temporally-rich, they are expensive, often not publicly available, or garnered only in select locations, restricting their wide use [6], [7]. Conversely, geo-located social media data are publicly and freely available, but present challenges due to their sparse nature. Further, much existing work has focused on predicting next location only, though knowledge of an entire movement timeline is relevant for emerging applications. Accordingly, we have developed a stochastic framework, Intermediate Location Computing (ILC) which combines approaches from several existing mobility prediction methods, alongside community behavior, to predict every missing location from an individuals social media timeline [8]. We compare ILC with several state-of-the-art approaches. For three major cities, ILC predicts at 1 and 2-hour resolution with up to 86% accuracy, 19% better than compared methods. We show how amount of community data improves prediction, and that community movement improves prediction of an individuals movement more on weekends versus weekdays. Overall this work presents a new algorithm to predict practical and continuous individual-level mobility patterns with sparse but a readily available, open data source.

Acknowledgment

Support for this work is from NSF grants 1643576, 1551036 and 1737987 and NIH R21AA023901 and R21MH110190.

REFERENCES

[1] J. Goff, A. Rowe, J. S. Brownstein, and R. Chunara, "Surveillance of acute respiratory infections using community-submitted symptoms and specimens for molecular diagnostic testing," *PLoS currents*, vol. 7, 2015.

[2] J. W. Debelius, Y. Vázquez-Baeza, D. McDonald, Z. Xu, E. Wolfe, and R. Knight, "Turning participatory microbiome research into usable data: lessons from the american gut project," *Journal of microbiology & biology education*, vol. 17, no. 1, p. 46, 2016.

[3] E. H. Chung and K. D. Guise, "Qtc intervals can be assessed with the alivecor heart monitor in patients on dofetilide for atrial fibrillation," *Journal of electrocardiology*, vol. 48, no. 1, pp. 8–9, 2015.

[4] R. Chunara, S. Aman, M. Smolinski, and J. S. Brownstein, "Flu near you: an online self-reported influenza surveillance system in the usa," *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.

[5] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin *et al.*, "Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience," *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 17–21, 2014.

[6] C. O. Buckee, A. Wesolowski, N. N. Eagle, E. Hansen, and R. W. Snow, "Mobile phones and malaria: modeling human and parasite travel," *Travel medicine and infectious disease*, vol. 11, no. 1, pp. 15–22, 2013.

[7] G. M. Vazquez-Prokopec, D. Bisanzio, S. T. Stoddard, V. Paz-Soldan, A. C. Morrison, J. P. Elder, J. Ramirez-Paredes, E. S. Halsey, T. J. Kochel, T. W. Scott *et al.*, "Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment," *PloS one*, vol. 8, no. 4, p. e58802, 2013.

[8] N. A. Rehman, K. Relia, and R. Chunara, "Using sparse digital traces to fill in individual level mobility timelines," *arXiv preprint arXiv:1710.02475*, 2017.