Social data mining and seasonal influenza forecasts: the FluOutlook platform

Qian Zhang[1], Corrado Gioannini[2] and Daniela Paolotti[2] and Nicola Perra[1], Daniela Perrotta[2], Marco Quaggiotto[2], Michele Tizzoni[2], and Alessandro Vespignani[1,2]

[1] MOBS, Northeastern University, Boston, MA, USA
{qi.zhang,n.perra,a.vespignani}@neu.edu
[2] ISI Foundation, Turin, Italy
{corrado.gioannini,daniela.paolotti,daniela.perrotta,marco.quaggiotto,michele.tizzoni}@isi.it

Abstract. FluOutlook is an online platform where multiple data sources are integrated to initialize and train a portfolio of epidemic models for influenza forecast. During the 2014/15 season, the system has been used to provide real-time forecasts for 7 countries in North America and Europe.

Keywords: real-time forecasting, epidemic modeling, data mining

1 Introduction

The real-time monitoring and modeling of infectious disease is being redefined by the novel availability of large scale social media and digital surveillance data. Several methods use social data, like search engine queries and tweets, as inputs for time series analysis; Google Flu Trends (GFT) [1] being probably the most known example. Unfortunately, most of the current approaches are unable to capture the disease transmission dynamics and its long-term trends, and suffer from several issues related to biases and statistical sampling [2]. Here we present FluOutlook (`http://fluoutlook.org/`), an online platform exposing real-time seasonal influenza forecasts. It integrates current and historical surveillance data, social data mining and several forecast models. Along with standard regression statistical models, FluOutlook includes stochastic generative models simulating the disease progression at the level of single individuals. The platform reports in real-time the influenza intensity with a lead time of up to four weeks, as well as main indicators of the epidemic season at its early stages. FluOutlook provides a description of the seasonal influenza that could be used by public health agency to guide their decision making process, as well as to compare and assess the performance of different forecast approaches.

2 Methodology

The FluOutlook platform consists of two parts: a computational framework that provides predictions and a user-friendly website that provides their visualiza-

tion. The system architecture, shown in Fig. 1, is made by three main components. The first component mines and assimilates the social and surveillance data needed to initialize the modeling approaches. The second component is the computational system that generates the numerical output of the modeling approaches. The third component is the statistical pipeline that compares the models' output with the current ground truth, available to define the forecast ensemble that is eventually exposed on the platform. The website of the platform runs as a Python Flask application with a PostgreSQL database, served through the Apache web service. In the landing page, maps show the current influenza activity level in each country and indicate the observed trend. The forecasting page provides more detailed predictions for each country.

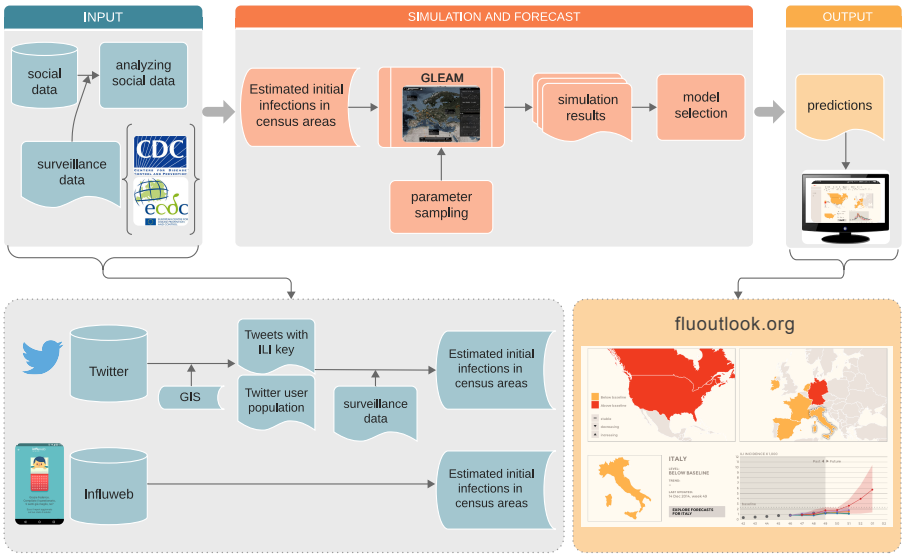

Fig. 1. The FluOutlook system architecture and the landing page of the Fluoutlook website.

2.1 Simulation engine

The FluOutlook platform has at its core a computational modeling and simulation engine. We use different forecast methodologies based on statistical regression approaches and the GLEAM (GLobal Epidemic And Mobility model) [3, 4]. GLEAM is a spatial, stochastic and individual based epidemic model based on three layers.The population layer is based on the high-resolution population database of the Gridded Population of the World project by the Socio-Economic Data and Applications Center (SEDAC) that estimates population with a granularity given by a lattice of cells covering the whole planet at a resolution of

15×15 minutes of arc. The Mobility layer integrates short-range and long-range transportation data. Long-range air travel mobility is based on travel flow data obtained from the Official Airline Guide (OAG) databases. The model identifies $3,362$ subpopulations in 220 different countries. The model simulates the short and long-range mobility of individuals between these subpopulations using a stochastic procedure. The infection dynamics takes place within each subpopulation and considers different infectious disease dynamic and the intervention measures being considered. GLEAM has been successfully used during the 2009 H1N1 pandemic to provide short and long term predictions of its course [5, 6].

2.2 Tracking seasonal flu with social data mining

One of the key and novel components in FluOutlook is the estimation of the initial cases necessary to run the GLEAM model (see Fig. 1). While many different sources of data can be integrated, currently FluOutlook adopts two sources of geolocalized data: Twitter and a participatory system for digital surveillance called Influweb [8].

Inferring initial infections from Twitter. By filtering geolocalized tweets with a set of ILI-related keywords, we can obtain spatial and temporal information about ILI. We use 40-50 keywords for each given language. Not all ILI keywords have the same relevance. Moreover, tweets containing an ILI keyword may not necessarily contain information related with influenza. Although it is still challenging to filter noisy information, we simplify this process by ranking the ILI keywords with consideration of the correlation between the time series of the surveillance data and the volume of tweets for a given key word. In each country, the volume of geolocalized tweets, Twitter user population, and actual population allow the estimation of the relative number of initial infections for a given week in each subpopulation area.

Inferring initial infections from Influweb. Open source indicators, such as Twitter, search engine queries, Wikipedia, may be mixed with irrelevant information and over-represent some particular demographic [2]. Novel data sources can overcome these issues. For instance, online self-reporting platforms are designed to provide more accurate in-time indicators of disease activity. Influenzanet [7] is one of such online platforms. It monitors ILI activities using self-reported information coming from volunteers across several countries in Europe. Influweb, a part of Influenzanet project, is a system to collect information on influenza-like-illness in Italy [8]. Voluntary participants across the country register on the website and submit to the system information about their locations, demographic and influenza-related health status. To better monitor the ILI activity, the volunteers are invited to weekly update their health conditions. The high quality and reliability of such data allow to infer initial infections in a given week in any census area directly. In FluOutlook, we use Influweb data to initialize GLEAM simulations in Italy.

2.3 Generative model selection and forecast output.

The simulation module in the platform performs a Latin hypercube sampling of a parameter space used in the GLEAM model, and generates for each sampled point P a statistical ensemble of the epidemic profiles. From each statistical ensemble, the model selection module estimates the likelihood function $L(P|X)$, where $X = x_0, x_1, ..., x_{N-1}$ indicates the ILI surveillance data in a given fitting window of length N. By considering the likelihood region defined by relative likelihood function in defining the parameters' range, the module selects a set of models. The selected models provide both long-term predictions for epidemic peak time and intensity, and short-term predictions for the epidemic profiles in the future four weeks. In Fig. 1, we show predicted epidemic profiles for Italy in week 2, 2015 with initial infections inferred from Twitter and Influweb, as well as other time-series predicting methods. In the forecasting page, the predicted statistical confidence intervals for peak time and intensity are given on the bottom of the web page.

3 Conclusion

Since the fall 2014 FluOutlook platform has provided real-time forecast of seasonal flu for the United States, Canada, Italy, France, Netherlands, Spain and Ireland The platform can incorporate surveillance data from any other countries or regions and is able to provide forecast of seasonal influenza for countries with high quality social data.

References

1. Ginsberg J, et al. Detecting influenza epidemics using search engine query data. *Nature* 457:1012-4. (2009)
2. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 343:1203-5. (2014)
3. Balcan D, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci.* 106:21484–21489. (2009)
4. Balcan D, et al. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *J. Comput. Sci.* 1:132–145. (2010)
5. Balcan D, et al. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine* 7:45. (2009)
6. Tizzoni M, et al. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine* 10:165. (2012)
7. `https://www.influenzanet.eu/`
8. `https://www.influweb.it/`