Forecasting Influenza in Senegal with Call Detail Records

Hao Wu*‡, Prithwish Chakraborty†‡, Saurav Ghosh†‡ and Naren Ramakrishnan†‡

*Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA
†Department of Computer Science, Virginia Tech, Arlington, VA 22203, USA
‡Discovery Analytics Center, Virginia Tech, Arlington, VA 22203, USA

Abstract—**As part of the D4D Senegal Challenge we describe the use of call detail records (CDRs) in seeding parameters for an epidemiological model around metapopulations. We apply this model to the study of influenza-like illnesses and validate model results against epidemiological surveillance data.**

I. Introduction

Epidemiological surveillance and forecasting has become an established discipline with the availability of myriad direct surveillance and surrogate data sources. Perhaps the most mature methods are available for forecasting influenza-like illnesses (ILI). Researchers have explored the integration of social and physical indicators [1]; while physical indicators (e.g., humidity, temperature, season) contribute the most to performance quality, social indicators (e.g., activity on Twitter) do provide a measurable improvement especially in identifying non-traditional progression of disease. Most recently, researchers have explored the possibility of non-traditional data sources such as restaurant reservations [2] and hospital parking lot imagery [3].

In this paper, we describe the use of mobile call detail records (CDRs) from the D4D Senegal Challenge in seeding parameters for an epidemiological model around metapopulations. We apply this model to the study of influenza-like illnesses and validate model results against epidemiological surveillance data. Related research in this space is discussed in greater detail toward the end of the paper.

II. Preliminaries and Problem Formulation

In this section, we will first introduce some notation and preliminaries to be used in the rest of the paper, followed by the formal definition of the problem studied here.

A. Preliminaries

a) Call Detail Records (CDR): A mobile call detail record (CDR) is a three element tuple (u, t, loc), where u specifies a mobile phone user (e.g., a numerical ID), t represents the time stamp that this call event happens, and loc denotes the location where this call is originated. The loc element can possibly be at different granularity levels, e.g. the arrondissement level or the mobile phone antenna tower level. When the loc is at the antenna tower level, further CDR metadata is also available in the format of a four element tuple: $(site, arr, lon, lat)$, where $site$ and arr represent the antenna tower ID and the arrondissement ID the tower located in, respectively, and lon and lat denote the longitude and latitude of the antenna tower.

Given a CDR tuple r, we will use $r[u], r[t]$ and $r[loc]$ to represent its mobile user ID, time stamp and location values, respectively. $date(r[t])$ is used to represent the date associated with the time stamp $r[t]$ of the CDR tuple r. When loc is at the antenna tower level, $r[loc][arr]$, $r[loc][lon]$, and $r[loc][lat]$ are used to represent the corresponding arrondissement area, tower longitude, and latitude values, respectively.

b) Network Notations: A weighted directed network G is defined as a ternary tuple $G = (V, E, W)$ where V is the vertex set, E is the edge set and W is the weight matrix for the edges in E. Given any two vertices $v_i, v_j \in V$, $e_{ij} = (v_i, v_j)$ represents the directed edge from vertex v_i to v_j, and $W(e_{ij})$ denotes the corresponding weight for edge e_{ij}.

c) Disease Spread Model: . To simulate a disease spread model we used a modified SIR model with interacting meta-populations in this work. Given any node (arondisment or tower) i and time t, we denote the number of susceptible people in the node by $S_i(t)$, number of infected people by $I_i(t)$ and the number of recovered people by $R_i(t)$. At any time point t, the summation over all I_i is used as the total predicted ILI case count in the country.

B. Problem Formulation

Given a CDR dataset $R = \{r_i\}$, the problem is to infer a weighted directed network $G = (V, E, W)$ which captures the mobile phone user mobility information in R, and forecast the spread of influenza by imposing a disease spread model over G.

III. Influenza Forecast Model

In this section, we describe our approach to forecasting influenzing from CDR datasets. We will begin by estimating the active mobile phone population at the arrondissement and mobile phone antenna tower levels, followed by a method to instantiate a disease propagation network.

A. Active Mobile Phone Population Estimation

Given CDR data $R = \{r_i\}$, we first classify R into two groups according to its location granularity, e.g. R_{arr} and

R_{tower}, where

$$R_{arr} = \{r_i \mid r_i[loc] \in ID_{arr}, r_i \in R\}$$
$$R_{tower} = \{r_i \mid r_i[loc] \in ID_{tower}, r_i \in R\}$$

Here ID_{arr} and ID_{tower} denote the set of all arrondissement IDs and mobile phone antenna tower IDs that appear in R, respectively.

Let N denote the total number of mobile phone users in Senegal [4], and let n_{arr} and n_{tower} represent the number of sampled mobile phone users in CDR set R_{arr} and R_{tower}, respectively. We estimate the active mobile phone population in each arrondissement area and each antenna tower region on a daily basis using the following approach.

For a given arrondissement area id_j^{arr} at a given date d_j, we find all the sampled mobile phone users that are involved in some mobile phone activity (e.g. make a call or send a text message), i.e.,

$$U_j^{arr} = \{r_i[u] \mid r_i \in R_{arr}, date(r_i[t]) = d_j, r_i[loc] = id_j^{arr}\}.$$

The estimate of the active mobile phone user in arrondissement area id_j^{arr} at date d_j is then given by:

$$N_j^{arr} = \frac{|U_j^{arr}|}{N_{R_{arr}}} N,$$

where $N_{R_{arr}}$ is the total number of sampled users in R_{arr}, and $|\cdot|$ represent the cardinality of the set.

Next, we will estimate the number of active mobile phone users in each antenna tower region in the given arrondissement area id_j^{arr} at date d_j. Similar to what we have done at the arrondissement level, for any mobile phone antenna tower id_k^{tower} located in id_j^{arr}, we find all the sampled users that are involved in at least one mobile phone activity in the covered region by antenna tower id_k^{tower} from R_{tower}, i.e.,

$$U_k^{tower} = \{r_i[u] \mid r_i \in R_{tower}, date(r_i[t]) = d_j,$$
$$r_i[loc] = id_k^{tower}\}.$$

The estimated active mobile phone population in antenna tower region id_k^{tower} is then given by:

$$N_k^{tower} = \frac{|U_k^{tower}|}{N_{id_j^{arr}}} N_j^{arr},$$

where $N_{id_j^{arr}}$ is the number of sampled mobile phone users who are active at date d_j in the arrondissement area id_j^{arr} in R_{tower}.

B. Propagation Network Estimation with CDR Data

We estimate the propagation network using the trajectories of mobile phone users at the antenna tower level, which could be computed from the CDR set R_{tower}. Thus, before we proceed to estimate the propagation network, we will first define the trajectories of mobile phone users at the antenna tower level, and describe how to compute the trajectories from the CDR set R_{tower}.

Algorithm 1: Construct propagation network

```
input : date $d_j$, antenna tower ID set $ID_{tower}$,
        trajectories $\mathcal{T}^{d_j} = \{T_{u_1}^{d_j}, T_{u_2}^{d_j}, \ldots, T_{u_{n_{tower}}}^{d_j}\}$.
output: network $G_{d_j}$
$V_{d_j} \leftarrow ID_{tower},\ E_{d_j} \leftarrow \emptyset,\ W_{d_j} \leftarrow \emptyset$;
for $T_{u_i}^{d_j} \in \mathcal{T}^{d_j}$ do
    for $k = 1$ to len($T_{u_i}^{d_j}$) $-1$ do
        if $r_k[loc] \neq r_{k+1}[loc]$ then
            $e_k \leftarrow (r_k[loc], r_{k+1}[loc])$;
            $E_{d_j} \leftarrow E_{d_j} \cup \{e_k\}$;
            $W_{d_j}(e_k) \leftarrow W_{d_j}(e_k) + 1$;
        end
    end
end
$G_{d_j} \leftarrow (V_{d_j}, E_{d_j}, W_{d_j})$;
return $G_{d_j}$;
```

A trajectory at the antenna tower level for a mobile phone user is an ordered sequence of antenna tower IDs that this mobile phone user appears in and participates in actively during a particular day. The sequence of the antenna tower IDs should be ordered by the corresponding CDR time stamps in increasing order (e.g. time stamp t_1 being smaller than time stamp t_2 means the CDR related to t_1 happens earlier than that related to t_2). To be more specific, a trajectory $T_{u_i}^{d_j}$ for mobile phone user u_i at date d_j is defined as:

$$T_{u_i}^{d_j} = (r_1[loc], r_2[loc], \ldots, r_l[loc]),$$
$$\text{where } r_k \in R_{tower}, r_k[u] = u_i, date(r_k[t]) = d_j, k = 1, \ldots, l$$
$$r_k[t] \leq r_{k+1}[t], k = 1, \ldots, l-1$$

If a location $r_k[loc]$ belongs to a trajectory $T_{u_i}^{d_j}$, we use $r_k[loc] \in T_{u_i}^{d_j}$ to represent this aspect. To compute the trajectories from the CDR set R_{tower}, we just follow the trajectory definition described above.

With calculated mobile phone user trajectories $\mathcal{T}^{d_j} = \{T_{u_1}^{d_j}, T_{u_2}^{d_j}, \ldots, T_{u_{n_{tower}}}^{d_j}\}$ at a particular date d_j, we will continue to construct the propagation network $G_{d_j} = (V_{d_j}, E_{d_j}, W_{d_j})$ for date d_j. The basic idea is to create the network G_{d_j} in such a way that it captures the mobile phone user flow information between any two of the antenna towers. Thus, we will use each antenna tower as a vertex in the network G_{d_j}, that is $V_{d_j} = ID_{tower}$. To construct the edge set E_{d_j}, for each trajectory $T_{u_i}^{d_j} \in \mathcal{T}^{d_j}$, if any pair of consecutive locations in $T_{u_i}^{d_j}$ is different from each other, e.g. $r_k[loc] \neq r_{k+1}[loc]$ for $r_k[loc], r_{k+1}[loc] \in T_{u_i}^{d_j}$, we add a directed edge $(r_k[loc], r_{k+1}[loc])$ into edge set E_{d_j}, and increase its weight by 1. Algorithm 1 illustrates this network construction procedure, where $len(T_{u_i}^{d_j})$ represents the number of locations in the trajectory $T_{u_i}^{d_j}$.

Finally, we use the average propagation network during a period with the edge weights normalized between 0 and 1 as our estimation of the influenza propagation network. To be more specific, the estimated influenza propagation network during the period from date d_1 to d_m would be:

$$\hat{G}^{d_1}_{d_m} = (V^{d_1}_{d_m}, E^{d_1}_{d_m}, W^{d_1}_{d_m})$$

$$\text{where } V^{d_1}_{d_m} = ID_{tower},\ E^{d_1}_{d_m} = \bigcup_{j=1}^{m} E_{d_j}$$

$$W^{d_1}_{d_m} = \frac{\bar{W}^{d_1}_{d_m}}{\max\left(\bar{W}^{d_1}_{d_m}\right)},\ \bar{W}^{d_1}_{d_m} = \frac{1}{m}\sum_{j=1}^{m} W_{d_j}$$

where $\max(\cdot)$ denotes the largest element of the given edge weight matrix.

C. *SIR Influenza Spread Model*

We used a discrete *SIR* meta-population model to capture the spread of influenza in the network. For each epidemiological week we consider each node (arrondissement level or tower level) to be a sub-population experiencing a single SIR dynamic process for influenza.

Under discrete approximations, for the node i the SIR evolution equation can be given as:

$$I_i(t+1) \sim NegBin\left(\lambda_i(t+1), I_i(t)\right) \tag{1}$$

where $NegBin$ signifies the negative binomial distribution and $\lambda_i(t+1)$ denotes the expected number of new infections in unit time in node i.

Following similar steps as in [5], we model this expected count as follows:

$$\lambda_i(t+1) = \frac{\beta(t) \times S_i(t) \times (I_i(t) + \mathcal{I}_i(t))^{\alpha}}{N_i(t)} \tag{2}$$

where $\beta(t)$ indicates the transmissibility of the disease which we model to be independent of spatial characteristics. $\mathcal{I}_i(t)$ captures the spatial spread from neighboring nodes and α is a factor used to correct for discrete assumptions (see [5]).

We use a gravity model to capture the spatial force $\mathcal{I}$. Since influenza data is reported weekly, we assumed the same population estimate $N_j(t)$ for node j for the full epidemiological week. Under this assumption, we model the spatial force as a Gamma process, i.e.,

$$\mathcal{I}_i(t) \sim Gamma(m_k(t), 1) \tag{3}$$

Here m_k signifies the spatial coupling. Under the generalized gravity model the coupling induced from node k to j can be given as :

$$m_{k\rightarrow j}(t) \propto I_k(t) \cdot N_j(t)/d_{kj}$$

d_{kj} here signifies the directed edge distance (inverse of edge weight) from node k to j. The overall spatial coupling for node j can then be given as:

$$m_j(t) = \theta N_j(t) \sum_{j \neq k} \frac{I_k(t)}{d_{kj}} \tag{4}$$

Here θ is a constant signifying the strength of spatial interactions.

TABLE I
COMPARISON OF ACCURACY FOR DIFFERENT METHODS

Method	Percentage Relative Accuracy
ARMA	43.25
National Level	67.50
Arrondissement Level	**80.25**
Tower Level	70.30

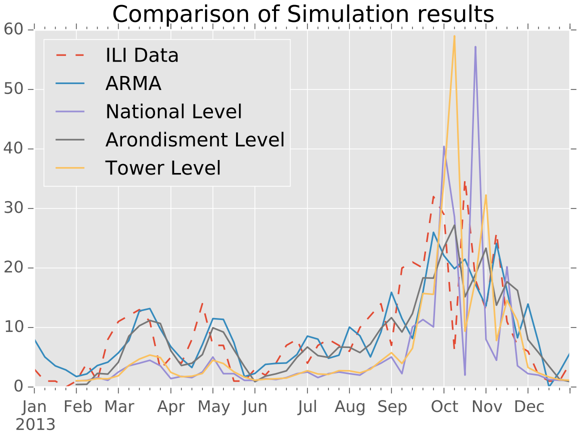

Fig. 1. Average Curves generated by different simulations

IV. SIMULATION

We used the discrete SIR meta population model as described in section III and ran multiple simulations over the network to compare the efficacy of using mobility as surrogates for influenza network spread. We randomly initialized a single node with an infected individual and ran the stochastic model described earlier. We used a static $\beta(t)$ for this work. Parameters of this process such as θ, α and β play a crucial part in setting up the model. We used a latin hyber-cube sampler to create a grid for the said parameters and found the best parameters through cross-validation. The cross-validation procedure used the accuracy measure defined as:

$$accuracy = \left(1 - \frac{|actual - predicted|}{\max(actual, predicted, 10)}\right) \times 100 \tag{5}$$

We ran the simulation with the tower graphs (i.e. where the nodes and hence the sub-population are at the tower level) and at the arrondissement level. We also ran a discrete SIR process without any spatial force using the national level data (i.e. no network structure) for comparison. Finally, we also implemented a simple ARMA model to compare against the epidemic models. All the influenza data used in the simulations are downloaded from WHO FluNet [6]. In Table I, we present the accuracy results (percentages) while predicting two weeks ahead and assuming full knowledge of the mobile network. As can be seen, the arrondissement level meta-population gives the best accuracy. Also, from ARMA to the arrondissement level we can see a 100% increase in accuracy. For better visual comparison, Figure 1 presents the average plots for each of the different methods.

V. Related Work

In this section, we provide a brief survey of related research. In particular, we discuss work related to estimating human mobility from CDR data and other type of mobile data, and modeling epidemics over networked metapopulations.

A. Human Mobility Modeling

Learning human mobility patterns provides insight into understanding and solving key problems in epidemiology and social science. Thus, concomitant with the development of advanced mobile technology, much active research has been conducted towards understanding human mobility with mobile data. In [7], the authors analyzed billions of anonymous CDR data to characterize the mobility patterns of thousands of people, and explored different aspects of human mobility, e.g. daily travel range and traffic volumes. The authors of [8] studied the travel patterns of 500,000 individuals in Cote d'Ivoire using mobile phone CDR datasets. Through considering both the uncertainty of movements and temporal correlations of individual trajectories, the authors performed a theoretical analysis of the limits of predictability in human mobility.

Beyond mobile phone CDR datamany, research has also explored other forms of mobile data in studying human mobility. In [9], the authors studied the trajectory data of 100,000 mobile phone users whose positions are tracked for a six-month period through their cellphones. They found that individual travel patterns could be collapsed into a single spatial probability distribution, indicating the inherent similarity of human travel patterns. In [10], the authors collected close proximity interactions (CPIs) data from 788 individuals in an American high school using wireless sensor network technology, and thus, inferred the human contact network for estimation of infectious disease transmissions. However, collecting such non-CDR data requires additional mobile devices or software, which may be inconvenient to apply to large scale populations.

Brennan et. al. [11] study the relationship between the human mobility and the spread of infectious disease at a *global* level. They aim to solve the task of predicting the prevalence of flu-like illness in a given city. The flows of individuals between cities are inferred with geo-tagged twitter status of travelers. The authors of [12] studied the global spread of smallpox after an intentional release event through the simulation over a large-scale structured metapopulation model considering human mobility.

B. Epidemic Modeling over Metapopulations

Balcan et. al. developed and presented the Global Epidemic and Mobility (GLEaM) model in [13], which integrates sociodemographic and population mobility data into a spatially structured stochastic disease approach. The flexible structure of GLEaM makes it suitable for computational modeling of global epidemic spread while considering population mobility at the same time. In [14], by considering three European countries and the corresponding commuting networks at different resolution scales, the authors explored the approach of using proxies for individual mobility to describe the commuting flows and predict the diffusion of an influenza-like illness epidemic. Goufo et. al. presented a fractional SEIR model over metapopulation system in [15] to study the spread of measles between four distinct cities. The condition for the stability of the disease-free equilibrium was discussed, and the numerical simulation showed that infection was proportional to the size of population in each city. Wang et. al. [16] provide a survey on the latest progresses on spatial epidemiology on networked metapopulation, in which empirical and theoretical findings that verify the validity of networked metapopulation modeling are discussed.

VI. Conclusion

Our initial exploration of CDRs shows promise in creating a synthetic model upon which we can impose and study different epidemiological scenarios. Future work will be aimed at capturing behavioral interventions as well as detecting significant shifts of population-level activity and studying their effects on (or influences by) disease progression.

Acknowledgments

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. Disclaimer: 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, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. V. Marathe, and N. Ramakrishnan, *Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions*, ch. 30, pp. 262–270.

[2] O. E. Nsoesie, L. D. Buckeridge, and S. J. Brownstein, "Guess who's not coming to dinner? evaluating online restaurant reservations for disease surveillance," *J Med Internet Res*, vol. 16, no. 1, p. e22, Jan 2014.

[3] P. Butler, N. Ramakrishnan, E. O. Nsoesie, and J. S. Brownstein, "Satellite imagery analysis: What can hospital parking lots tell us about a disease outbreak?" *Computer*, vol. 47, no. 4, pp. 94–97, 2014.

[4] World Bank Group, "Mobile cellular subscriptions," Dec. 2014. [Online]. Available: http://data.worldbank.org/indicator/IT.CEL.SETS.P2

[5] Y. Xia, O. N. Bjørnstad, and B. T. Grenfell, "Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics," *The American Naturalist*, vol. 164, no. 2, pp. 267–281, 2004.

[6] World Health Organization, "Flunet," 2014. [Online]. Available: http://www.who.int/influenza/gisrs_laboratory/flunet/en/

[7] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013.

[8] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, Oct. 2013.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[10] M. Salath, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, vol. 107, no. 51, pp. 22 020–22 025, 2010.

[11] S. Brennan, A. Sadilek, and H. Kautz, “Towards understanding global spread of disease from everyday interpersonal interactions,” in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI ’13. AAAI Press, 2013, pp. 2783–2789.

[12] B. Goncalves, D. Balcan, and A. Vespignani, “Human mobility and the worldwide impact of intentional localized highly pathogenic virus release,” *Scientific Reports*, vol. 3, Jul. 2013.

[13] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, “Modeling the spatial spread of infectious diseases: The GLobal epidemic and mobility computational model,” *Journal of Computational Science*, vol. 1, no. 3, pp. 132–145, Aug. 2010.

[14] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. Gonzlez, and V. Colizza, “On the use of human mobility proxies for modeling epidemics,” *PLoS Comput Biol*, vol. 10, no. 7, p. e1003716, 07 2014.

[15] E. F. D. Guofo, S. C. O. Noutchie, and S. Mugisha, “A fractional seir epidemic model for spatial and temporal spread of measles in metapopulations,” *Abstract and Applied Analysis*, vol. 2014, 2014.

[16] L. Wang and X. Li, “Spatial epidemiology of networked metapopulation: An overview,” *bioRxiv*, 2014.