Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Hongwen Song December 1, 2021

Dynamic Network Anomaly Modeling of Cell-Phone Call Detail Records for Infectious Disease Surveillance

By

Hongwen Song

Carl Yang, Ph.D.
Advisor

Quantitative Theory and Methods

Carl Yang, Ph.D.
Advisor

Weihua An, Ph.D.
Committee Chair

Ymir Vigfusson, Ph.D.
Committee Member

2021

Dynamic Network Anomaly Modeling of Cell-Phone Call Detail Records for Infectious Disease Surveillance

By

Hongwen Song

Carl Yang, Ph.D.
Advisor

An abstract of
A thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2021

Abstract

Dynamic Network Anomaly Modeling of Cell-Phone
Call Detail Records for Infectious Disease Surveillance
By Hongwen Song

Global monitoring of novel diseases and outbreaks is crucial for pandemic prevention. To this end, movement data from cell-phones is already used to augment epidemiological models. Recent work has posed individual cell-phone metadata as a universal data source for syndromic surveillance for two key reasons: (1) these records are already collected for billing purposes in virtually every country and (2) they could allow deviations from people's routine behaviors, both in terms of mobility and social interactions such as during illness, to be detected. In this paper, we develop the necessary models to conduct population-level infectious disease surveillance by using cell-phone metadata individually linked with health outcomes. Specifically, we propose GraphDNA—a model that builds Graph neural networks (GNNs) into Dynamic Network Anomaly detection. Using cell-phone call records (CDR) linked with diagnostic information from Iceland during the H1N1v influenza outbreak, we show that GraphDNA outperforms state-of-the-art baselines on individual Date-of-Disease (DoD) prediction, while tracking the epidemic signal in the overall population. Our results suggest that proper modeling of the universal CDR data could inform public health officials and bolster epidemic preparedness measures.

Dynamic Network Anomaly Modeling of Cell-Phone Call Detail Records for Infectious Disease Surveillance

By

Hongwen Song

Carl Yang, Ph.D.
Advisor

A thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2021

Acknowledgments

I have been extremely fortunate to have had the support of the department family and friends near and far. I would like to thank Dr. Carl Yang and Dr. Ymir Vigfusson for their consistent support since the start of the honors program and their countless hours dedicated to this thesis. I would also like to thank Dr. Weihua An for being in my honors committee and teaching me wonderful course materials. What's more, I am grateful for the constant love and support of my family and my boyfriend, Ziwei Dong. Without their support, this work would not be possible.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The COVID-19 pandemic underscores the need for early outbreak detection and infectious disease surveillance. In normal times, public health officials continuously monitor emerging pathogens and smaller epidemics to mitigate the chances for any of these turning into a global pandemic. Without a crystal ball to know exactly how many people are infected at each moment, these efforts include *syndromic surveillance* where multiple data sources, such as hospital records, cross-sectional surveys, or even search-engine queries are searched for clusters of symptoms that warrant further scrutiny. For diseases where symptoms coincide with the infectious period, such as most influenza variants, such symptomatic surveillance can further track the progression of an epidemic and provide direct feedback for mitigation strategies, such as quarantines, lock-downs, or vaccination campaigns.

Recent efforts have advanced cell-phone metadata, such as the *call-detail records* (CDR), as a potential universal data source to augment symptomatic surveillance [32, 8, 51], for four key reasons. First, CDR data include the (anonymized) caller and recipient numbers, a timestamp of the call or text, and the GPS-coordinates of the cellular tower through which the call was routed; the cellular tower and timestamp of data access is also recorded. They thus provide time-series for individual mobility

and social interactions—behaviors that may differ when the person is ill. Second, in contrast with aggregated mobility models [14], CDR data may be linked with health data at the *individual* level while accommodating privacy concerns [51], allowing deviations from individual routines—such as staying home when ill—to be detected. Such localized signals can provide crucial information early in an epidemic. Third, CDR data are already recorded by virtually every mobile-network provider for billing purposes within an established regulatory and privacy framework. Deploying disease monitoring on top of an existing data source, such as CDR logs, is easier and cheaper than alternatives, while also allowing important ethical and privacy concerns to be addressed (cf. Appendix A). Finally, cell-phone use is ubiquitous (105 mobile subscriptions per 100 inhabitants; 97% of the world population covered by a mobile network) whereas Internet access is less pervasive (57% of the world population) and heavily skewed towards affluent regions (19% of individuals in the least developed countries (LDC) have Internet access), according to 2020 estimates [29]. Many lower and middle-income countries lack resources for public health measures and monitoring, and thus would stand to benefit the most from inexpensive disease monitors.

In 2010, a research has been made on monitoring disease infection in social network by simply monitoring the friends of randomly selected individuals [16], but this method did not consider the global network structure and requires additional efforts in friendship nomination. In 2016, another study proposed a pandemic modeling method using network data, specifically, monitoring and controlling influenza A (H1N1) using social network analysis and cloud computing [45], yet the proposed architecture fails to capture individual routines and is hard to be generalized to CDR data.

Key technical challenges must be resolved, however, to make individual CDR models practical for epidemiological surveillance. Using linked health and CDR data from the H1N1v epidemic in Iceland in 2009, Vigfusson *et al.* [51] showed that indi-

vidual mobility is reduced around the day of influenza-like illness diagnosis. While it is interesting that infection produces behavioral changes that are measurable from very sparse CDR data, the all-important question for the epidemiologist is whether *measured behavioral changes imply infection*, which would allow her to estimate how many people are taken ill at the given moment (prevalence) and tune forecasting model parameters. This direction is hard for several reasons, including networked signals (involving individuals' behaviors regarding both themselves and their social contacts), temporal routines (requiring the capture of dynamic behavioral patterns), and weak supervision (because disease labels are sparse and only weakly correlated with the behavioral anomalies).

In this paper, we work towards the vision of estimating population-level disease prevalence. Confronting the problem as an individual disease prediction task, we augment the existing individual-level features [51] with a social context to capture regular contacts and group interactions to better distill routine social interaction patterns. Central to our approach are graph neural networks (GNNs) [46, 31, 25] that have recently been adapted to model dynamic and temporal networks. We found that existing research into dynamic GNNs has predominantly been focused on modeling network formation and evolution in contexts such as link prediction [40, 55, 41, 15], but such GNNs are not yet suited to tracking dynamic social behaviors of individuals and their routines. On the other hand, several traditional (non-GNN-based) dynamic and temporal network models have been designed to capture emergent patterns during network evolution, and to identify abnormal individuals or subgraphs [3, 52, 39]. Yet these approaches were also unsuitable, since they cannot incorporate node features or be trained for specific tasks, such as disease prediction.

In this paper, we propose a novel integrated GNN for *Dynamic Network Anomaly* modeling (GraphDNA) to meet the goal of detecting deviations from an individual's routine social behaviors for predicting disease onset. Broadly, GraphDNA combines

two key modules concerning dynamic social behavior prediction and anomaly-based disease prediction. The former module employs a graph convolutional neural network (GCN) model [31] to capture individuals' social behaviors and builds it into a long-short term memory (LSTM) model [27] to record the dynamic patterns of such behaviors. The latter module then combines a data-driven learnable logistic regression (LR) model [28] and a temporal-pattern-oriented statistical Gaussian tail probability (GTP) model [1] to predict disease diagnosis from anomalies in the social behavior dynamics.

In our experiments on the same labeled dataset from the H1N1v epidemic in Iceland [51], we evaluate GraphDNA by comparison to the most relevant baselines from state-of-the-art including dynamic GNNs and other temporal or networked anomaly detection models. With a focus on estimating the Date-of-Diagnosis (DoD) of diagnosed individuals, we demonstrate the advantages of our GraphDNA method on the generic task of supervised dynamic network anomaly detection. We also apply the individual inference of GraphDNA to the larger population, tracking the epidemic curve within the diagnosed population and, further, finding an illness-associated behavioral change signal in the whole population. Finally, we analyze key design decisions, hyper-parameter settings, and provide an efficiency study of GraphDNA.

Chapter 2

Background and Related Work

2.1 Syndromic Surveillance

Keeping with technological developments and new data sources, *syndromic surveillance systems* emerged in the 2000s to "*seek to use existing health data in real time to provide immediate analysis and feedback to those charged with investigation and follow-up of potential outbreaks*" [26]. In 2009, Google Flu ushered in the era of big data syndromic surveillance through passively collected data sources by using aggregated search engine queries for flu-like symptoms to estimate regional influenza levels with a lag of only one day [22]. Google Flu's approach, however, was later found to have been fundamentally flawed, missing non-seasonal influenza outbreaks and frequently overestimating disease burden, and was shut down in 2015. Prominent researchers characterized the project's indifference to supplementing the existing body of science and instead seeking to replace it with black-box models as an example of "big data hubris" [35]. A flurry of research into other data sources for use in syndromic surveillance, such as social media, have followed [42, 44], usually built around technologies used primarily in high-income countries.

Aggregated CDR data, such as rates of population movement between cell-phone

towers [32], have informed epidemiological models for cholera [6], dengue fever [54], malaria [53, 9], Ebola [33], influenza [50], and recently SARS-CoV-2—the pathogen that causes COVID-19 [14]. Because these models lack linkage at the individual level, they rely on correlations between the aggregated data and other datasets, thereby limiting their statistical power and generality [19]. Individual CDR data were used during COVID-19 to infer likely contacts of infection in Israel, with staunch privacy objections [24] (cf. Appendix A). The aforementioned dataset from Iceland is the first large-scale study where individual CDR-data were linked to health outcomes, specifically influenza-like illness (ILI) diagnosis.

2.2 Dynamic Network Anomaly Modeling

Anomaly detection refers to the data mining process that measures the deviations of objects of interest from the majority group [12, 2]. One of the most common scenarios of anomaly detection is on sequential data (*e.g.*, time-series), where the algorithm is often composed by a sequence modeling part and a deviation scoring part [13]. For instance, [38, 1, 18] employ sequential neural networks such as LSTM and HTM (hierarchical temporal memory) to model sequential records and then access the likelihood of anomalies based on the models' predictions. Recent studies for many emerging real-world applications concern the more complicated problem of anomaly detection on graph data [36]. For example, [3, 52, 39] detect abnormal nodes in graphs based on their deviations from normal node clusters without supervision. [52] combines one-class classification with GNNs for graph anomaly detection in a supervised manner. However, these methods are designed only for static graphs.

Real-world networks can be modeled as dynamic graphs to represent the evolving objects and relationships among them [36, 7]. Attempts in modeling dynamic networks have been made in statistics, where the stochastic actor-oriented model

(SAOM) [48] is commonly used to model dynamic networks. Yet SAOM is based on Markov evolution and its decision making is based on full information of the network, so it is not suitable for modeling long-term dependent dynamics on sparse dataset. Aside from SAOM, the temporal exponential random graph models (TERGMs) [23] proposed a more flexible method in modeling dynamic network, yet it suffers from network size limitation and degeneracy. The computational intensiveness limits its practical data size within tens or a few hundreds of nodes, and the degeneracy means that for certain combinations of parameters the Markov chain Monte Carlo estimation rarely or never converges. On the other hand, extensive research has been done into dynamic network modeling in deep learning field, including tasks such as temporal link prediction [40, 55, 41, 15] and efficient graph streaming [5, 21, 17]—none of which encompass anomaly detection. AddGraph [58] and NetWalk [56] are two methods that are closest to our setting of dynamic network anomaly modeling. AddGraph employs temporal GCN to detect anomalous edges but cannot trivially detect anomalous nodes, whereas NetWalk leverages a DeepWalk-based framework to detect both anomalous nodes and edges, but cannot readily incorporate node attributes or task-specific supervision.

Chapter 3

The GraphDNA Framework

3.1 Data Analysis

3.1.1 Description

The data set from Iceland contains CDR data for 93,409 people (about a quarter of the Icelandic population) over a 3-year period beginning in February 2009, with 87,773 individuals making calls during the 1-year period beginning in February 2009 when the H1N1v epidemic occurred. The CDR records are linked with influenza-like illness (ILI) diagnosis data for 1,434 individuals who provide a spatially representative sample ($r > 0.86$) of the homogeneous Icelandic population [51] (we focus only on an individual's first ILI diagnosis). Each record contains the encrypted source and destination numbers for a call placed over a cell-phone tower, the GPS coordinates of the cell-phone tower, a timestamp, and the duration of call; similar metadata for text messages (SMS) are also included in the CDR data. No content of calls or text messages are included. The linked health dataset includes the encrypted number and the Date of Diagnosis (DoD) of ILI by any health provider in Iceland for the owner of that number.

The CDR data reveals rich movement and social patterns. Common contacts

and their own interactions give a proxy for daily communication networks. The GPS location of a call gives a proxy for a person's location; a series of such locations provides a proxy for movement; and a series of movements can act as proxy for routine patterns, such as weekday commute to and from work. Existing studies identified that the movement patterns were different on the day before the DoD and up to three days after were significantly different from regular days, specifically that 1.1–1.4 fewer unique tower locations were visited on average [51]. They also found that significantly fewer calls were placed but that calls were longer on the day following diagnosis. Prior work did not consider more advanced movement, social features, or dynamics.

3.1.2 Node features

We conducted principled analysis of the many node features that can be constructed from the CDR data, including location_num (number of unique tower recorded), avg_len (average call length), tot_len (total call length), call_cnt (call count), degree (number of contacts), clus_coeff (cluster coefficient), abg_lon (average longitude), avg_lat (average latitude), all of which are varying by day. Intuitively, multiple features may potentially indicate disease onset or diagnosis. We studied feature correlations based on the days from DoD to quantify such potential. Specifically, since these predominantly ordinal attributes usually did not follow normal distributions, we correlated features using the Spearman's Correlation Coefficient (SCC) [49].

3.1.3 Link features

To account for people's connections in the phone call network, we conduct the node feature analysis jointly with the network links. Specifically, we consider the *social behaviors* of every individual, that includes their own behaviors (node features) together with those of their neighbors in the phone call network. For simplicity, we

simplify social behaviors here as the features of a node together with the aggregate of the node features in its direct neighbors. In addition to binary indicators of whether two people were in contact during a day, the CDR data further allow us to extract various link features, such as call counts and (total) call durations. Before designing more complicated models beyond elementary GCN [31], we extend our data analysis over the correlations with days from DoD to study the potential impact of such link features in disease prediction.

3.1.4 Diagnostic features

Figure 3.1 demonstrates the results of our node and link feature analysis based on their correlations with days from diagnosis based on the training data. Even though the exact values on y-axis are hidden for credential reasons, all of the correlations are statistically significant at the 0.01 significance level. Based on the correlation scores, we set an empirical threshold to select the top five node features as a trade-off between model capacity and simplicity. For the link features, we found that unweighted links already encompass the strongest signal towards the DoD, obviating the need for more complicated GCN designs to model link features.

Combining nodes and links, in Figure 3.2, we visualize the dynamic social behaviors of individuals via three prominent node features aggregated through the (weighted/unweighted) links in the direct neighborhoods, where deviations are clearly observed around the DoD. Such observations motivate our goal of predicting disease infections based on dynamic network anomalies in the CDR data.

3.1.5 Other features

While we rely on real data analysis, both to identify the node and link features and to justify the design of our models, we underscore the "greedy" nature of such analysis and the potential over-simplification of the problem. However, the focus of our work

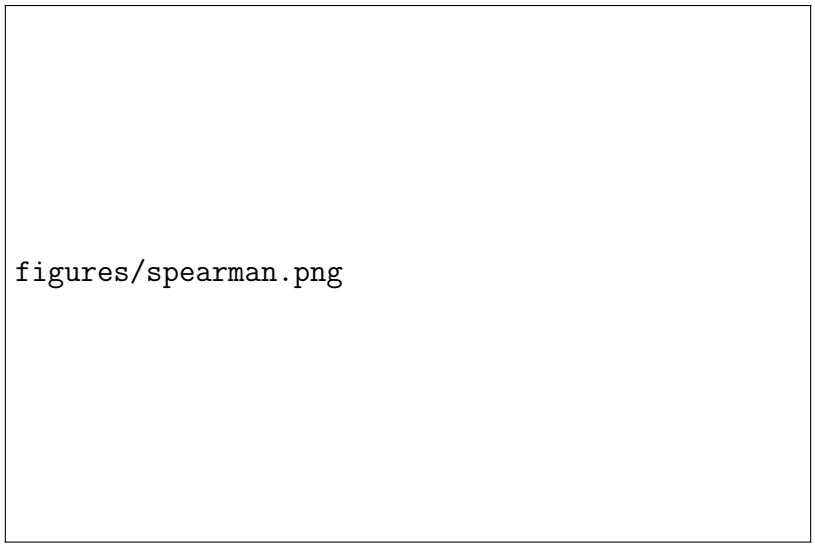


Figure 3.1: *Node and link feature analysis: Spearman's CC between social behaviors and days from diagnosis. We set an empirical threshold (dashed line) to choose relevant node features for inclusion. Unweighted links—links without additional features—were found to be the most useful.*

is to provide the first fundamental framework of disease prediction based on dynamic network anomalies in CDR data, and thus we believe model simplicity is crucial. We thus defer more advanced analysis and complex models to follow-up work.

3.2 Problem Formulation

3.2.1 Input

From the CDR data, we construct daily snapshots of the cell-phone call network as graphs $\mathcal{G} = \{G^{(t)} = (V, E^{(t)}, F^{(t)})\}_{t=1}^{T}$. Here, V is the set of all vertices (individuals) who have at least one call record in the entire dataset, $E^{(t)}$ is the set of unweighted directed links at timestamp (day) t, *i.e.*, $e_{ij}^{(t)} = 1$ if there is at least one call from v_i to v_j on day t, and 0 otherwise, and $F^{(t)}$ denotes the behavioral features at timestamp t, *i.e.*, $f_i^{(t)} \in \mathbb{R}^D$ denotes the individual behavioral features of v_i on day t. We model a full year of data between 02/01/2009 to 02/01/2010 to capture the entire Fall 2009 H1N1v outbreak in Iceland, and use $t \in \{0, 1, \ldots, T = 364\}$ to denote the relative

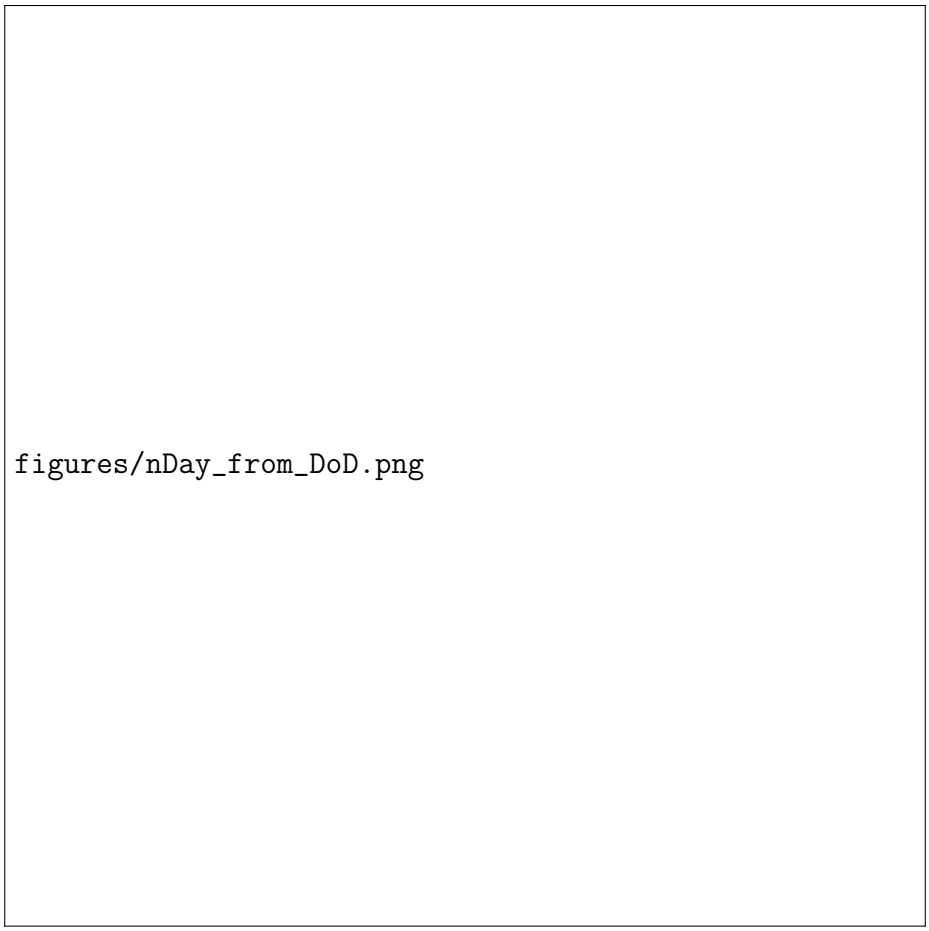


Figure 3.2: *Dynamic social behaviors of diagnosed people vs. days from DoD: We observe clear deviations of social behaviors around the DoD. The shaded interval marks the period between days -1 to +3 days from DoD when the largest deviations are observed.*

days within that time frame.

Within V, we pay special attention to the subset $V' \subset V$ who had a record for an influenza-like illness (ILI) diagnosis during the one year period. $Y \in \mathbb{R}^{|V'| \times |T|}$ stores the day of ILI diagnosis (DoD) labels of people in V' ($y_i^{(t)} = 1$ if v_i has a positive diagnosis on day t, and 0 otherwise). Recovery from an influenza illness may take several days and anomalous behavior is often observed in the several days surrounding the DoD [51]. We thus follow common practice [10] and prior work to define the extended DoD labels $\tilde{Y}$, where $\tilde{y}_i^{(t)} = 1$ if $y_i^{(t')} = 1$ and $t \in [t' - 1, t' + 3]$, and 0 otherwise.

3.2.2 Output

The primary goal of our work is to predict the DoD of $v_i \in V'$, through modeling the connection between people's dynamic social behaviors and disease diagnoses based on the phone call graphs $\mathcal{G}$ given above. Beyond V', the model should also generalize to the larger population V, where much of the diagnosis labels are unavailable, and yet provide disease prediction—whether and when an individual gets infected. Public health officials could use such estimates to monitor the effective disease burden of a population during an epidemic, and even segment estimates by groups, age, region, sub-populations, or others as needed [51].

3.3 Model Overview

The main aim of our work is to thoroughly model people's behaviors in $\mathcal{G}$ from the CDR data, and measure deviations from their behavior routines to facilitate disease surveillance. To meet this goal, we design a two-stage framework: (1) dynamic network behavior prediction, and (2) anomaly-based disease prediction, which can be further integrated through iterative training.

We present an overview of our proposed GraphDNA framework in Figure 3.3. In the first stage, a sequential graph representation learning module is designed to capture people's daily behaviors in the phone call graphs $\mathcal{G}$ and then make consecutive predictions on their next-day behaviors. In the second stage, an anomaly detection module is designed to compare the predicted behaviors with the true behaviors on each day and make predictions about whether a person might have fallen ill with H1N1v on that day.

We use a subset of people with diagnosis labels $V^1_{\text{train}} \subseteq V'$ and the entire set of non-diagnosed people $V - V'$ to train the dynamic social behavior prediction module in stage one. We then use a disjoint subset of people with diagnosis labels $V^2_{\text{train}} \subseteq V'$

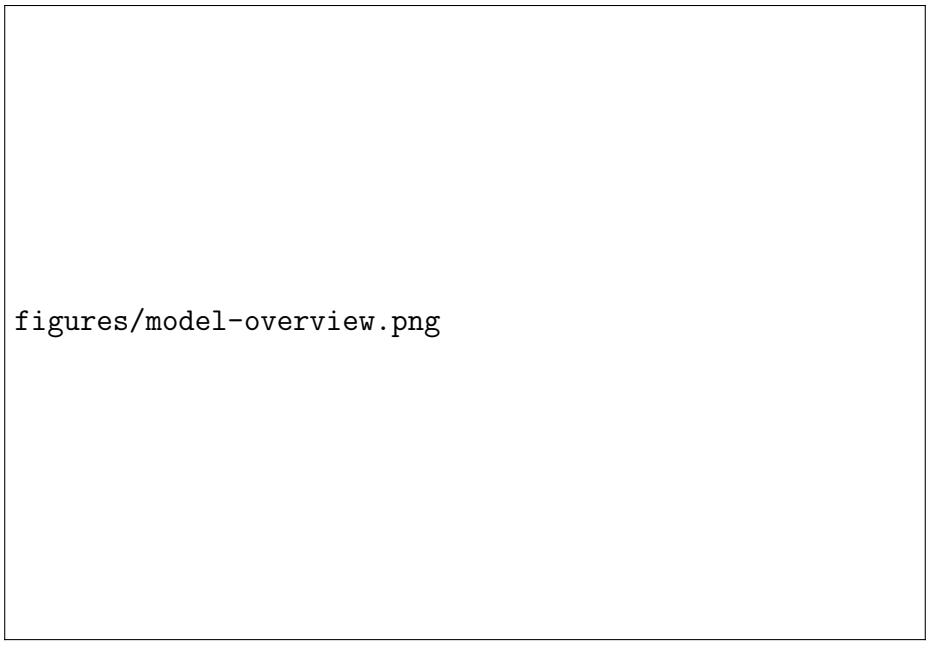


Figure 3.3: *An overview of our* Dynamic Network Anomaly *modeling (*GraphDNA*).*

to train the anomaly-based disease prediction module in stage two. Another disjoint set $V_{\text{val}} \subseteq V'$ is used to iteratively validate and improve the model design as well as tune the model hyper-parameters, and the final disjoint set $V_{\text{test}} \subseteq V'$ is held out until the final testing and reporting of the results.

3.4 Dynamic Social Behavior Prediction Module

To model people's routine behavior over time in the cell-phone call graphs, we design a dynamic graph model to predict people's behaviors at each day (*i.e.*, $f_i^{(t)}, \forall v_i \in V, t \in \{1, 2, \ldots, T\}$) based on their own past behaviors (*i.e.*, $\{f_i^{(t')} \mid t' = 0, \ldots, t-1\}$) and the past behaviors of their neighbors (*i.e.*, $\{f_j^{(t')} \mid t' = 0, \ldots, t-1; v_j \in \mathcal{N}(v_i, t', K)\}$, where $\mathcal{N}(v_i, t', K)$ denotes the K-hop neighborhood of v_i in graph $G^{(t')}$). To efficiently encode such dynamic social behaviors, we design an integrated model of GCN [31] and LSTM [43] that we train on the node set $\bar{V} = V^1_{\text{train}} \cup V - V'$.

3.4.1 Social behavior modeling

Motivated by recent advances in GCNs for node representation learning in content-rich networks [31], we employ GCN for the modeling of the static social behaviors of individuals based on the neighborhood of each node on each day in the cell-phone call graphs (*i.e.*, $\left\{f_i^{(t)}, f_j^{(t)} \mid j \in \mathcal{N}(v_i, t, K)\right\}, \forall v_i \in \bar{V}, t \in \{0, 1, \ldots, T\}$). We encode this information into the representation vectors $h_i^{(t)(K)}$ through the following recursive operations

$$H^{(t)(k)} = \phi\left(A^{(t)} H^{(t)(k-1)} W^{(k)} + b^{(k)}\right), \tag{3.1}$$

where $A^{(t)}$ is the normalized adjacency matrix with self-loop on day t, $W^{(k)}$ and $b^{(k)}$ are the learnable parameters of the GCN model, ϕ is a non-linear activation function such as LeakyReLU, and $k \in \{1, 2, \ldots, K\}$. $H^{(t)(0)} = F^{(t)}$ is the feature matrix on day t. Based on our data analysis in Section 3.1, we used the binary directed adjacency matrix $A^{(t)} \in \{0, 1\}^{N \times N}$ and real-valued feature matrix $F^{(t)} \in \mathbb{R}^{N \times D}$ of selected node features. The number of GCN layers K (also denoted as L_1) is a tunable hyper-parameter. To capture common patterns, we share and train the same GCN model across all nodes $v_i \in \bar{V}$ and all days $t \in \{0, 1, \ldots, T\}$.

3.4.2 Dynamic social behavior modeling

To integrate the history of past behaviors and model the dynamics of social behaviors, we further employ an LSTM model [43] based on the outputs of the GCN model. Specifically, given the sequence of representation vectors as the outputs of the GCN model (*i.e.*, $\{h_i^{(t)} \mid t = 0, \ldots, T-1\}, \forall v_i \in \bar{V}$), the LSTM model seeks to predict the node features of the next days (*i.e.*, $\{f_i^{(t)} | t = 1, \ldots, T\}, \forall v_i \in \bar{V}$), which is computed

through the following recursive operations

$$
\begin{aligned}
q^{(t)} &= \sigma\left(W_{qh}h_i^{(t)} + W_{qr}r^{(t-1)} + W_{qc}r^{(t-1)} + b_q\right), \\
e^{(t)} &= \sigma\left(W_{eh}h_i^{(t)} + W_{er}r^{(t-1)} + W_{ec}c^{(t-1)} + b_e\right), \\
c^{(t)} &= e^{(t)} \odot c^{(t-1)} + q^{(t)} \odot \tanh\left(W_{ch}h_i^{(t)} + W_{cr}r^{(t-1)} + b_c\right), \\
o^{(t)} &= \sigma\left(W_{oh}h_i^{(t)} + W_{or}r^{(t-1)} + W_{oc}c^{(t)} + b_o\right), \\
r^{(t)} &= o^{(t)} \odot \tanh(c^{(t)}), \\
\hat{f}_i^{(t+1)} &= \phi\left(W_{fr}r^{(t)} + b_f\right),
\end{aligned}
\tag{3.2}
$$

where $r^{(-1)}$ and $c^{(-1)}$ are initialized as all-zero vectors, $h_i^{(t)}$ is the sequential input to LSTM, W's and b's denote the learnable parameters of different LSTM cells, *i.e.*, the input gate q, forget gate e, output gate o, and activation vectors c. All cells are of the same size as the output activation vector r, and $\odot$ is the Hadamard product. Eq. (3.2) represents a single-layer LSTM model. In a multi-layer LSTM unit model, the input of the first layer is $h^{(t)}$, and the input of the l-th layer ($l \geq 2$) is the output activation vector of the previous layer, *i.e.*, $r^{(t)(l-1)}$. The number of LSTM layers L_2 is a tunable hyper-parameter. Given an input behavior representation of a node v_i on day t (*i.e.*, $h_i^{(t)}$), the final output of the LSTM model is the predicted behavior (node feature) of v_i on day $t+1$ (*i.e.*, $f_i^{(t+1)}$).

To capture the common patterns, we share and train the same LSTM model across the representation and feature sequences of all nodes $v_i \in \bar{V}$, which we do in an end-to-end fashion jointly with the GCN model through the following objective function:

$$
\min_{\Theta_1,\Theta_2} \sum_{v_i \in V} \sum_{t=1}^{T} \mathcal{L}_1\left(f_i^{(t)}, \hat{f}_i^{(t)}\right), \tag{3.3}
$$

where Θ_1 and Θ_2 denote the parameters of the GCN model and LSTM model, respectively. Here, $\mathcal{L}_1$ is a loss function such as MSE. We detail the training process in

Algorithm 1.

3.5 Anomaly-based Disease Prediction Module

We focus on the task of DoD prediction not only because we only have positive labels of diagnosed people in the dataset but also due to the crucial impact of accurate detection of patient DoD on disease transmission control. Following past studies [51] and our data analysis in Section 3.1, our central hypothesis is that the DoD labels may be predicted to an extent based on people's deviations from their routine behaviors (*i.e.*, anomalies) as captured in the cell-phone call graphs.

To detect anomalies, we first compute the deviation scores between the predicted behaviors and real behaviors for all people in another training set V^2_{train} that is disjoint from $\bar{V}$

$$s_i^{(t)} = \hat{f}_i^{(t)} - f_i^{(t)}, \forall v_i \in V^2_{\text{train}}, t \in \{1, 2, \ldots, T\}. \tag{3.4}$$

Every individual $v_i \in V^2_{\text{train}}$ is associated with a sequence of $T-1$ D-dimensional vectors $\{s_i^{(t)} | t = 1, 2, \ldots, T\}$, from which we will seek to predict the extended DoD labels $\{\tilde{y}_i^{(t)} | t = 1, 2, \ldots, T\}$.

We design and experiment with three representative types of anomaly detection models based on the output of our dynamic social behavior prediction stage: (1) a deep learning model based on logistic regression (LR) [28], (2) a statistical model based on Gaussian tail probabilities (GTP) [1], and (3) a hybrid model that integrates the first two.

3.5.1 Deep learning model

Since the DoD labels $\tilde{Y}$ are binary, we devise a LR model for binary classification [28]. To mitigate noise and asynchronous anomalies across different features, we smooth

the input sequences over a rolling window. We have

$$y_i^{(t)} = \sigma(MLP(\tilde{s}_i^{(t)})), \forall v_i \in V^2_{\text{train}}, t \in \{1, 2, \ldots, T\}, \tag{3.5}$$

where $\tilde{s}_i^{(t)} = \text{mean}(\ldots, s_i^{(t-1)}, s_i^{(t)}, s_i^{(t+1)}, \ldots)$. We pad both ends of the sequence with zeroes. Here, the window size Ω_0 is a tunable hyper-parameter, σ is the sigmoid function, MLP is the multilayer perceptron with LeakyReLU activation, and the number of layers L_3 is another tunable hyper-parameter.

The LR model is trained with the following objective function

$$\min_{\Theta_3} \sum_{v_i \in V^2_{\text{train}}} \sum_{t=1}^{T} \mathcal{L}_2 \left(\tilde{y}_i^{(t)}, \hat{y}_i^{(t)} \right), \tag{3.6}$$

where $\mathcal{L}_2$ is a loss function such as cross-entropy. Given the propensity of LR to simply predict the majority class when the class labels are imbalanced, we employ a top-k selection mechanism during testing where we predict the top k $\hat{y}_i^t$'s as 1 (illness) for each $v_i \in V_{\text{val}} \cup V_{\text{test}}$, and simply set k to 5 since the largest interval of concern around the DoD is 5 days ([t-1, t+3]). We pre-specified the diagnosis window to be [-1, +3] both because this is the average asymptotic range for H1N1v patients [57], and because we observed the most significant change in individual behaviors in this range. Future study is warranted to experiment with more variations of diagnosis windows.

3.5.2 Statistical model

While LR provides an effective way of searching the feature space and finding the inductive bias with the help of training data, it ignores dynamic contexts and is not designed to capture temporal anomalies. On the other hand, anomaly detection has been explored in temporal settings through statistical models such as the Gaussian

Tail Probability (GTP) model [1]. Following their design, to effectively detect temporal anomalies from the D-dimensional time-series data of the deviation scores of each individual v_i (*i.e.*, $\{s_i^{(t)} \mid t = 1, 2, \ldots, T\}$), we first apply two rolling windows W_1 and W_2 of sizes Ω_1 and Ω_2 as follows

$$\begin{aligned} W_1 &= [\max(0, t - \Omega_1/2), \max(0, t - \Omega_1/2) + \Omega_1 - 1] \\ W_2 &= [\max(0, t - \Omega_2/2), \max(0, t - \Omega_2/2) + \Omega_2 - 1], \end{aligned} \tag{3.7}$$

where $\Omega_1 > \Omega_2$ are two tunable hyper-parameters. We then model the values in W_1 as normal distributions, and use values in W_2 to compute the recent short-term average. An anomaly likelihood of $s_i^{(t)}$ based on the GTP is computed as

$$p_i^{(t)} = 1 - Q\left(\frac{\text{mean}\left(s_i^{(t)} \mid t \in W_2\right) - \text{mean}\left(s_i^{(t)} \mid t \in W_1\right)}{\text{std}\left(s_i^{(t)} \mid t \in W_1\right)}\right), \tag{3.8}$$

where Q represents the Gaussian tail probability approximation function [30]. The total anomaly probability of v_i on day t is computed as $\hat{p}_i^{(t)} = \prod_{d=1}^{D} p_i^{(t)(d)}$, which is directly used for the prediction of $\tilde{y}_i$ with the same top-k selection mechanism.

3.5.3 Hybrid model

The GTP model adds temporal context to the deviation scores and are thus more suitable for anomaly detection in the dynamic social behavior data. However, the multi-dimensional behavioral features are not parameterized for the task of disease (DoD) prediction. To this end, we propose a novel hybrid model that combines the power of both worlds—by simply replacing the $\tilde{s}_i^{(t)}$ in Eq. (3.5) with $p_i^{(t)}$ in Eq. (3.8). The sequence smoothing with Ω_0 is no longer needed due to the sliding windows W_1 and W_2.

3.6 Training Algorithms

The detailed training algorithms of the two modules are outlined in Algorithms 1 and 2. We note that our GraphDNA framework does not rely on more hyper-parameters than the basic ones for classic GCN, LSTM, LR, and GTP models. In this work, we train the two stages separately and achieve promising results for disease prediction in the end. Potentially, the two stages can also be trained jointly (iterative or end-to-end), which we leave as an interesting direction for future work.

Algorithm 1: Dynamic Social Behavior Prediction

Input: $\{G^{(t)} \mid t = 0, \ldots, T\}$, $\bar{V} = V^1_{\text{train}} \cup V - V'$, # GCN layers L_1, # LSTM layers L_2, hidden layer sizes H

Output: $f_i^{(t)}, \forall v_i \in V, t \in \{1, 2, \ldots, T\}$

```
while not converged do
    for t ← 0 to (T − 1) do
        H^(t) ← GCN(G^(t); L_1, H)
    end
    for v_i ∈ V̄ do
        for t ← 0 to (T − 1) do
            f̂_i^(t+1) ← LSTM(h^(t); L_2, H)
            loss ← L_1({f_i^(t)}, {f̂_i^(t)})
            Update the GCN and LSTM model parameters Θ_1 and Θ_2
             according to the loss
        end
    end
end
```

3.6.1 Complexity analysis

The training of the GCN model in stage one takes $O(N_1^2 T L_1 D H)$ time; the training of the LSTM model takes $O(N_1 T L_2 D H)$ time in each epoch, where $N_1 = |\bar{V}| \ll N = |V|$. In stage two, $O(N_2 T(\Omega_1 + \Omega_2))$ time is taken to calculate the GTP, and $O(N_2 L_3 H^2)$ time is taken to train the LR model, where $N_2 = |V^2_{\text{train}}| \ll N = |V|$. $T, L_1, L_2, L_3, D, H, \Omega_1, \Omega_2$ are all constant numbers—T is 364, and all others are

Algorithm 2: Anomaly-based Disease Prediction (Hybrid)

Input: $\{G^{(t)}|t=0,\ldots,T\}$, $\{\hat{f}_i^{(t)}|v_i \in V, t=1,2,\ldots,T\}$, V^2_{train}, # LR layers L_3, hidden layer sizes H, GTP window sizes Ω_1 and Ω_2

Output: $\{\hat{y}_i^{(t)}, \forall v_i \in V, t \in \{1,2,\ldots,T\}$

while *not converged* **do**
  for $v_i \in V^2_{train}$ **do**
    for $t \leftarrow 0$ *to* $(T-1)$ **do**
      $p_i^{(t)} \leftarrow \text{GTP}\left(s_i^{(t)}; \Omega_1, \Omega_2\right)$
      $\hat{y}_i \leftarrow \text{LR}\left(p_i^{(t)}; L_3\right)$
      $\text{loss} \leftarrow \mathcal{L}_2\left(\tilde{y}_i^{(t)}, \hat{y}_i^{(t)}\right)$
      Update the LR model parameters Θ_3 according to the loss
    end
  end
end

smaller than 100.

Chapter 4

Experiments

In this section, we evaluate GraphDNA by conducting extensive experiments on the CDR dataset, with a focus on the following research questions (RQs).

- **RQ1:** How does GraphDNA perform compared to closest baselines from state-of-the-art on DoD prediction?
- **RQ2:** Does GraphDNA have the potential to be generalized for disease prediction in the larger population?
- **RQ3:** How does each major component of GraphDNA contribute to the overall performance?
- **RQ4:** What are the effects of different tunable model hyper-parameters on GraphDNA?
- **RQ5:** Is the running time of GraphDNA practical?

4.1 Experimental Settings

4.1.1 Dataset

The Iceland CDR dataset has a total of 87,773 distinct nodes, and an average of 54,867 nodes and 30,451 links across the 365 graph snapshots. The nodes comprise

Table 4.1: *Anomaly detection performance comparison.*

Model	Metrics				
	Micro Precision	Micro Recall	Micro AUC	Micro F1	Macro Accuracy
NetWalk	0.0529 ± 0.0019	0.1599 ± 0.0028	0.5025 ± 0.0005	0.0773 ± 0.0004	0.1672 ± 0.0007
LSTM-AD	0.0386 ± 0.0035	0.2836 ± 0.0047	0.4995 ± 0.0003	0.0667 ± 0.0003	0.3016 ± 0.0014
OddBall	0.0362 ± 0.0001	0.3530 ± 0.0001	0.4988 ± 0.0001	0.0648 ± 0.0001	0.3578 ± 0.0001
OCGNN	0.1754 ± 0.0009	0.5491 ± 0.0073	0.5043 ± 0.0033	0.2586 ± 0.0043	0.5749 ± 0.0046
GraphDNA-w/o-GCN	0.0490 ± 0.0018	0.1356 ± 0.0006	0.0694 ± 0.0005	0.0693 ± 0.0006	0.1441 ± 0.0013
GraphDNA-w/o-LSTM	0.2326 ± 0.0063	0.6792 ± 0.0034	0.5855 ± 0.0040	0.3333 ± 0.0017	0.6871 ± 0.0061
GraphDNA-w/o-LR	0.2138 ± 0.0036	0.4652 ± 0.0063	0.5728 ± 0.0037	0.2807 ± 0.0012	0.4723 ± 0.0029
GraphDNA-w/o-GTP	0.0871 ± 0.0005	0.2356 ± 0.0016	0.5167 ± 0.0022	0.1222 ± 0.0015	0.2372 ± 0.0018
GraphDNA	**0.2344 ± 0.0106**	**0.6986 ± 0.0054**	**0.5895 ± 0.0019**	**0.3384 ± 0.0019**	**0.7005 ± 0.0087**

two types: the 1,414 diagnosed nodes V' and the remaining non-diagnosed nodes $V - V'$. There are DoD labels for diagnosed nodes, but we do not know if any individuals in the non-diagnosed set were infected or not. We divide the diagnosed nodes V' into V^1_{train}, V^2_{train}, V_{val}, and V_{test} as discussed in Section 3.3 with a ratio of 3:3:2:2. We use V^1_{train} and V^2_{train} to train the two stages of our model, respectively. We use V_{val} to improve the model design and tune the hyper-parameters, and only use V_{test} to report the experimental results.

4.1.2 Baselines

We adapted the following state-of-the-art algorithms for our task of DoD prediction based on the dynamic cell-phone call graphs constructed from the CDR dataset.

- NetWalk[56]: an anomalous node detection method that is closest to our dynamic network setting. It learns and dynamically updates the representations of non-attributed networks as they evolve in an unsupervised manner.

- LSTM-AD[38]: an algorithm using stacked LSTM networks for anomaly detection in multi-variate time-series data. Since it cannot model network data, we provide it only with dynamic node features.

- OddBall[3]: an unsupervised method to detect abnormal nodes in static networks. Since it cannot handle dynamic networks, we compute a separate model of it for every timestamp.

- OCGNN[52]: a one-class classification framework that combines GNN with the one-class objective for attributed network anomaly detection in a supervised manner. Since it cannot handle dynamic networks, we compute a separate model for every timestamp.

For the supervised baselines, the same $V^1_{\text{train}} \cup V^2_{\text{train}} \cup V - V'$ is used for training, V_{val} is used for hyper-parameter tuning, and V_{test} is used for performance reporting. The unsupervised baselines are run on the whole V and tested on V_{test}. When making predictions on V_{test}, the same top-k selection mechanism is used to predict k positive DoDs for each individual.

4.1.3 Ablations

We performed a comprehensive ablation study, removing each component from GraphDNA one at a time, to evaluate the effectiveness of the main components in GraphDNA.

4.1.4 Evaluation metrics

Based on the predicted DoD labels $\hat{Y}$ and extended true DoD labels $\tilde{Y}$, we compute the following metrics adopted from the standard evaluation of group classifications.

- Micro Precision, Micro Recall, Micro AUC, and Micro F1, which represent the Precision, Recall, AUC and F1 scores averaged across all the testing individuals in V_{test}.
- Macro Accuracy, which is the percentage of testing individuals in V_{test} who have at least one correct DoD prediction.

The suite of metrics compares the prediction results with ground-truth from different perspectives, thus comprehensively comparing the performance of the evaluated algorithms.

4.1.5 Parameter settings

We tune and set the hyper-parameters of GraphDNA as the following default values: we set the number of GCN layers L_1 to 2, LSTM layers L_2 to 1, and LR layers L_3 to 2; we set the embedding size H of all layers in all models to 16; the sizes of rolling windows in GTP are set to $\Omega_1 = 100$ and $\Omega_2 = 3$. To ensure fair comparison, we use the same hyper-parameters for all of our model ablations. For the baselines, we also optimize their hyper-parameters on V_{val}.

We implemented GraphDNA with PyTorch 1.7.1. All code will be released upon the acceptance of this work; the derived dataset will be released after further sanitization.

4.2 DoD Prediction Comparison (RQ1)

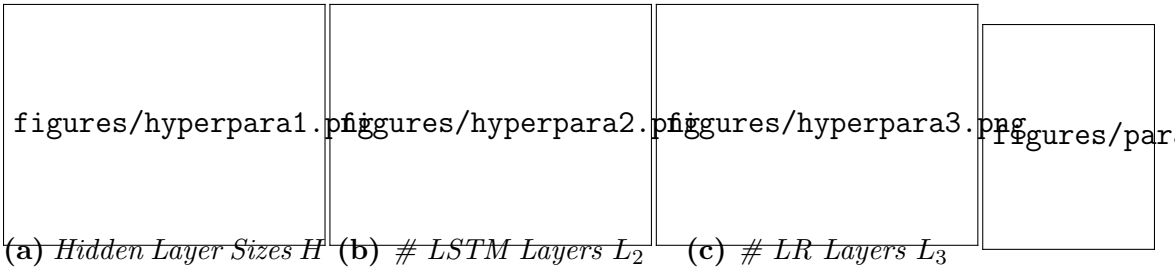
Table 4.1 shows that GraphDNA achieves the best performance across all metrics in the scenario of CDR-based DoD prediction. We highlight the following detailed observations.

- While not being fully consistent across the baselines, the multiple metrics we use demonstrate the same significant improvements of GraphDNA. Specifically, GraphDNA achieves 16.9%-33.6% relative gains over the strongest baseline across all metrics, indicating its superiority in the task of CDR-based DoD prediction.

- Although we have included the most relevant algorithms as baselines, none of them can properly integrate all important signals in our scenario, thus leading to unsatisfactory results across all metrics.

- Compared with LSTM-AD and OddBall, NetWalk focuses on structural anomalies and make cautious predictions, thus achieving better precision but worse recall.

Figure 4.1: *Average disease scores (ADS) of diagnosed group and whole population vs. daily diagnosed number (DDN) in the period of 2009 H1N1v outbreak in Iceland. Thin lines denote the medians/values of the ADS/DDN, thick lines indicate the smoothed medians/values, and shading delineates the 1^{st}–3^{rd} quantiles of the ADS.*



(a) *Hidden Layer Sizes H* **(b)** *# LSTM Layers L_2* **(c)** *# LR Layers L_3*

Figure 4.2: *Performance of* GraphDNA *with varying hyper-parameters. The dashed lines denote the performance of the best baseline (OCGNN).*

- OCGNN is the strongest baseline, likely due to its proper leverage of imbalanced task supervision, which indicates the importance of available DoD labels from the CDR data.

4.3 Anomaly Curve during Epidemic (RQ2)

Beyond predicting the DoD of diagnosed people, we examine the potential of GraphDNA to estimate wider disease infection among the entire population. In Figure 4.1, we

visualize the average disease score (ADS) in the diagnosed group (V_{test}) and the whole population (V) predicted by GraphDNA, versus the diagnosed number (DDN) in the ground-truth of V'. The main peak of the GraphDNA ADS estimate among the diagnosed group coincides with the ground-truth peak of H1N1v outbreak in Iceland around October 2009, suggesting that the ADS model captures behavioral anomalies associated with illness. The model also picks up anomalies during the winter holidays in December 2009. Interestingly, a small but significant anomaly signal also arises in the whole population during the epidemic (green curve). Notably, the model was not picking up time-of-year related artifacts, as evidenced by the baseline (orange curve)—showing ADS inference by the same model trained on a control group in which we matched an undiagnosed person with each diagnosed person 1:1 at random and assigned them the latter's DoD. We thus believe the model is discerning illness-specific anomalies in the whole population—a promising data source. We caution, however, that further research is warranted for predictive epidemic estimation (since the ADS scores in our model are based on training data from the entire 1-year period) and analyzing the disease transition among patients across the timeline.

4.4 In-depth Model Analysis (RQ3-5)

4.4.1 Ablation analysis (RQ3)

Table 4.1 also shows that each constituent part of GraphDNA contributes significantly to its overall performance. We further summarize several key observations as follows.

- Removing the GCN model causes the most significant performance drop, demonstrating the importance of modeling the neighborhood behaviors for DoD prediction—the key difference from our work to previous studies on the same CDR dataset [51].

- Surprisingly, removing the LSTM model actually does not significantly degrade performance—consistent with the reasonable performance of OCGNN. Perhaps evolutionary patterns are not be a key factor for DoD prediction, or perhaps LSTM is not the best model to capture such network evolution.

- Both the LR and GTP models are indispensable to GraphDNA, supporting our design principle of integrating the effective data-driven learning ability of LR with the anomaly-based feature engineering of GTP.

4.4.2 Hyper-parameter analysis (RQ4)

We varied important model hyper-parameters in Figure 4.2 to understand their effects.

- The hyper-parameters we tested have minimal impact on the performance of GraphDNA, maintaining significant margins from the best baseline across a vast range of values.

- Larger embedding sizes, fewer LSTM layers, and fewer LR layers generally improve results due to different trade-offs between model capacity and overfitting.

- The standard deviations across different settings remain in an acceptable range, indicating that GraphDNA's hyper-parameter are robust.

Due to the difficulty in implementing and running deep GCNs, we have not studied the performance of GraphDNA with the number of GCN layers L_1 greater than 2. While having significantly larger training and testing times, we have observed the performance of GCN with $L_1 = 2$ to be only slightly better than that with $L_1 = 1$, and thus lack compelling need to grow L_1 beyond 2 at the moment.

4.4.3 Efficiency analysis (RQ5)

We observe the computational cost of GraphDNA to be similar to those of OCGNN, which is slightly larger than those of LSTM-AD and NetWalk, yet within the same

order of magnitude (detailed results and analysis in Appendix B).

Chapter 5

Applications

The GraphDNA model estimates population-level disease prevalence based on deviations from behavioral patterns. After overcoming the data and technical challenges we tackled in the main text of the paper, the next question is how the model could be applied in practice.

5.1 Continuous forecasting models

Epidemiological models forecast the spread of a novel disease in or across societies to evaluate the impact of different public health policy options. While the most important driver for such models is the *transmission* of disease between people, disease transmission is also the most difficult component of the model to observe empirically, particularly at scale. The opaqueness of transmission causes models to instead rely on lagged or biased proxy measures, such as on-the-ground diagnoses or deaths. Instead, our GraphDNA model relays transmission dynamics by identifying behavioral changes consistent with disease symptoms of the entire population. The epidemiologist could thus employ the GraphDNA data source as a live stream of aggregate behavioral indicators to parameterize and improve their higher-layer forecasting models of the infectious disease dynamics, allowing for faster evaluation of

potential public-health measures and interventions during policy-making.

5.2 Web-based dashboard

The GraphDNA data source can inform population-level predictions for an epidemic, but it needs not treat the entire population as a monolith. Instead, we envision a web dashboard where the public health officials can break down the aggregates from the model by spatial (county-level) or demographic (race, sex, age) characteristics to understand how the disease is affecting different groups. The dashboard features should be designed with differential privacy in mind to protect the privacy of smaller groups [20, 51].

5.3 Broader connection with the web

Our approach for infectious disease surveillance, when deployed, is emblematic of the range of applications enabled by and emerging from the Web of Things: with the web acting as the narrow-waist protocol for information interchange, the collection of mobile phone data, the processing and parameterization of models, and interaction with the public health officials and epidemiologists all take place within the expanse of the World Wide Web protocols.

Chapter 6

Conclusion

Disease outbreak detection is difficult: population surveys are slow and skewed, and traditional syndromic surveillance requires the integration of a health-care data collection system with a responsive public health body to function adequately. However, monitoring behavioral anomalies through cell-phone metadata, as discussed here, offer a passive and universal approach to surveil an epidemic. Using real-world linked cell-phone and health data from the H1N1v pandemic in Iceland in 2009, we showed how GraphDNA identified individual behaviors indicative of disease and found evidence of illness-related anomalies the entire population that could potentially be used to assess the epidemic disease burden. Such estimates could inform both models and policy choices (*e.g.*, targeted lockdowns, quarantines, or vaccination campaigns) towards epidemic prevention—an important tool in the arsenal of public health defenders in our era of extreme connectedness.

We are aware of the limitations of the study, including the need of interpretation of disease-related signal captured in GraphDNA and the obstacle in retrieving phone-call data linked with health data in future applications. In this case, further study is warranted to distill the disease-related signal in the CDR-data picked up on by GraphDNA to provide a better understanding of the utility, generality,

and limitations of the proposed framework towards informing decision making in infectious disease surveillance—bolstering an active subject discussed by epidemiologists and policy makers [32, 8]. A parallel follow-up direction is to study how well GraphDNA generalizes to other datasets and tasks whose underlying challenges require appropriate dynamic network anomaly modeling.

Appendix A

Ethical & Privacy Considerations

While the COVID-19 pandemic has underscored the importance of disease surveillance to inform policy choices and prevent further transmission, it has also spurred important conversations about the interplay between individual privacy and the societal need for detailed information about infection and disease progression to curb an outbreak. For example, in response to the highly-contagious airborne nature of SARS-CoV-2, Google and Apple implemented mechanisms for allowing users of their mobile devices to opt-in to automatic and pseudononymous contact tracing of other Bluetooth devices lingering in their immediate vicinity to facilitate *contact tracing*—the discovery of potentially infectious transmission when the owner of any such device is known to have been contagious [47]. Large multi-country surveys across Western nations found strong public support for such contact tracing apps to fight COVID-19, with main reservations surrounding security, privacy, and the trust in government [4], only to then find lukewarm reception when the technology finally rolled out [11], ultimately limiting its significance for making public health strategy [37]. In lieu of an opt-in strategy, some nations took instead to scrutinizing CDR data, like we analyze here, to perform contact tracing. Such efforts, particularly in Israel, were panned in Western media due to their coercive nature and opaque enforcement [24] that is

misaligned with the ethical principles of effective contact tracing [34].

The approach for CDR data taken here, in contrast, concerns constructing *aggregates* to inform epidemiological policy rather than subjecting individuals to scrutiny by officials. We believe the use of aggregate CDR data sidesteps the false dilemma between health and privacy, offering an intriguing compromise to meet the ethical, privacy, and public health rigor needed to swiftly counter tomorrow's epidemics without sacrificing individual liberties in the process. To this end, we adopt the privacy-preserving framework of Vigfusson *et al.* [51] for gathering, managing, and consuming the sensitive data without placing undue trust on any stakeholder. Specifically, a neutral third-party organization receives deidentified CDR data from mobile-network operators, as well as deidentified health data from public health officials that uses the same anonymous individual identifier, and trains and uses the proposed models to produce aggregate information about the progression of the epidemic. The public health officials and epidemiologists consume the model predictions through an interface provided by the neutral third party (cf. chapter 5). The data sharing protocol ensures that the third party does not learn the original identities of the individuals in the data, that the mobile-network providers does not learn about health issues for their customers, and that the public health officials do not learn about individual mobile behavior or contacts. This distribution of trust through minimal privilege reduces further concentration of power within the already powerful corporate (mobile network operator) and government (disease control) entities through the disease monitoring technology, reducing chances for abuse of the technology so long as public trust in the neutral third-party and the ensuing policy actions by officials can be maintained.

The privacy-preserving approach taken by Vigfusson *et al.*, and which was used to generate the dataset we analyzed here, was vetted by the appropriate IRB board, specifically the national Icelandic Bioethics Commission, under approval #VSNb2010050012. The specific data analysis protocols we used in this paper were also IRB approved.

Appendix B

Time Complexity Experiments

We measure the running time of all compared algorithms on three dedicated GNU/Linux servers with 24 2.3 GHz Intel Xeon E5-2670v3 processors and 512 GiB of DRAM. Figure B.1 demonstrates the running times of baseline models, the ablations of GraphDNA, and the full GraphDNA. Since Oddball is fully unsupervised and disregards network evolution, its running time is significantly shorter than the other methods. All supervised baseline methods incur comparable running times with GraphDNA. Dropping the GCN or LSTM modules decreases the running time more than dropping LR or GTP modules in GraphDNA, which implies that the first stage of GraphDNA is more computationally intensive than its second stage—but the usage of GNNs, even for each timestamp, does not produce excessive execution times. This is mainly because despite the whole graphs being large, the training of GraphDNA is mostly done on the smaller set of diagnosed people, whereas the inference on the larger populations is only done when necessary, allowing for further parallelism. The biggest scalability bottleneck is the number of GCN layers L_1, which we observe to yield satisfactory results with small numbers (*e.g.*, 1, 2).

Figure B.1: *Running times of algorithms compared.*

Bibliography

[1] Subutai Ahmad and Scott Purdy. Real-time anomaly detection for streaming analytics. *arXiv preprint arXiv:1607.02480*, 2016.

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016.

[3] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*, 2010.

[4] Samuel Altmann, Luke Milsom, Hannah Zillessen, Raffaele Blasone, Frederic Gerdon, Ruben Bach, Frauke Kreuter, Daniele Nosenzo, Séverine Toussaert, Johannes Abeler, et al. Acceptability of app-based contact tracing for COVID-19: Cross-country survey study. *JMIR mHealth and uHealth*, 8(8):e19857, 2020.

[5] Khaled Ammar. Techniques and systems for large dynamic graphs. In *SIGMOD*, 2016.

[6] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.*, 5:8923, 2015.

[7] Maciej Besta, Marc Fischer, Vasiliki Kalavri, Michael Kapralov, and Torsten

Hoefler. Practice of streaming processing of dynamic graphs: Concepts, models, and systems. *arXiv preprint arXiv:1912.12740*, 2019.

[8] Nita Bharti. Linking human behaviors and infectious diseases. *PNAS*, 118(11), 2021.

[9] Caroline O Buckee, Amy Wesolowski, Nathan N Eagle, Elsa Hansen, and Robert W Snow. Mobile phones and malaria: modeling human and parasite travel. *Travel Med Infect Dis*, 11(1):15–22, 2013.

[10] Edenilson E Calore, David E Uip, and Nilda M Perez. Pathology of the swine-origin influenza a (h1n1) flu. *JPRP*, 207(2):86–90, 2011.

[11] Manuel Cebrian. The past, present and future of digital contact tracing. *Nature Electronics*, 4(1):2–4, 2021.

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[13] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *TKDE*, 24(5):823–839, 2010.

[14] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[15] Jinyin Chen, Xuanheng Xu, Yangyang Wu, and Haibin Zheng. Gc-lstm: Graph convolution embedded lstm for dynamic link prediction. *arXiv preprint arXiv:1812.04206*, 2018.

[16] Nicholas A Christakis and James H Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

[17] Laxman Dhulipala, Guy E Blelloch, and Julian Shun. Low-latency graph streaming using compressed purely-functional trees. In *PLDI*, 2019.

[18] Tolga Ergen and Suleyman Serdar Kozat. Unsupervised anomaly detection with lstm neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(8):3127–3141, 2019.

[19] Susan L. Erikson. Cell phones $\neq$ self and other problems with big data detection and containment during epidemics. *MAQ*, 32(3):315–339, 2018.

[20] Liyue Fan and Li Xiong. An adaptive approach to real-time aggregate monitoring with differential privacy. *IEEE Transactions on knowledge and data engineering*, 26(9):2094–2106, 2013.

[21] Guoyao Feng, Xiao Meng, and Khaled Ammar. Distinger: A distributed graph data structure for massive dynamic graph processing. In *IEEE BigData*, pages 1814–1822, 2015.

[22] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[23] Fan Guo, Steve Hanneke, Wenjie Fu, and Eric P Xing. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328, 2007.

[24] David M. Halbfinger, Isabel Kershner, and Ronen Bergman. To track Coronavirus, Israel moves to tap secret trove of cellphone data. *The New York Times*, March 2020.

[25] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

[26] Kelly J Henning. What is syndromic surveillance? *MMWR*, pages 7–11, 2004.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[28] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. 2013.

[29] International Telecommunication Union. Measuring digital development: Facts and figures 2020, Nov 2020.

[30] George K Karagiannidis and Athanasios S Lioumpas. An improved approximation for the gaussian q-function. *IEEE COMML*, 11(8):644–646, 2007.

[31] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[32] Nishant Kishore, Mathew V Kiang, Kenth Engø-Monsen, Navin Vembar, Andrew Schroeder, Satchit Balsari, and Caroline O Buckee. Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis. *The Lancet Digital Health*, 2020.

[33] Andrew M Kramer, J Tomlin Pulliam, Laura W Alexander, Andrew W Park, Pejman Rohani, and John M Drake. Spatial spread of the West Africa Ebola epidemic. *Royal Society Open Science*, 3(8):160294, 2016.

[34] Marjolein Lanzing. Contact tracing apps: an ethical roadmap. *Ethics and Information Technology*, pages 1–4, 2020.

[35] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[36] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Quan Z Sheng, and Hui Xiong. A comprehensive survey on graph anomaly detection with deep learning. *arXiv preprint arXiv:2106.07178*, 2021.

[37] Leonardo Maccari and Valeria Cagno. Do we need a contact tracing app? *Computer Communications*, 166:9–18, 2021.

[38] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *ESANN*, 2015.

[39] Emmanuel Müller, Patricia Iglesias Sánchez, Yvonne Mülle, and Klemens Böhm. Ranking outlier nodes in subspaces of attributed graphs. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013.

[40] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *AAAI*, 2020.

[41] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

[42] Adam Sadilek, Stephanie Caty, Lauren DiPrete, Raed Mansour, Tom Schenk, Mark Bergtholdt, Ashish Jha, Prem Ramaswami, and Evgeniy Gabrilovich. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *NPJ digital medicine*, 1(1):1–7, 2018.

[43] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, 2014.

[44] Loukas Samaras, Elena García-Barriocanal, and Miguel-Angel Sicilia. Syndromic surveillance using web data: a systematic review. *J Innov Health Inform*, pages 39–77, 2020.

[45] Rajinder Sandhu, Harsuminder K Gill, and Sandeep K Sood. Smart monitoring and controlling of pandemic influenza a (h1n1) using social network analysis and cloud computing. *Journal of Computational Science*, 12:11–22, 2016.

[46] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Netw. Learn. Syst.*, 20(1):61–80, 2008.

[47] Tamar Sharon. Blind-sided by privacy? digital contact tracing, the Apple/Google API and big tech's newfound role as global health policy makers. *Ethics and Information Technology*, pages 1–13, 2020.

[48] Tom AB Snijders. Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172, 1996.

[49] Charles Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 100(3/4):441–471, 1987.

[50] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, 10(7):e1003716, 2014.

[51] Ymir Vigfusson, Thorgeir A Karlsson, Derek Onken, Congzheng Song, Atli F Einarsson, Nishant Kishore, Rebecca M Mitchell, Ellen Brooks-Pollock, Gudrun Sigmundsdottir, et al. Cell-phone traces reveal infection-associated behavioral change. *PNAS*, 118(6), 2021.

[52] Xuhong Wang, Baihong Jin, Ying Du, Ping Cui, Yingshui Tan, and Yupu Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural. Comput. Appl.*, pages 1–13, 2021.

[53] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[54] Amy Wesolowski, Taimur Qureshi, Maciej F Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O Buckee. Impact of human mobility on the emergence of dengue epidemics in pakistan. *PNAS*, 112(38):11887–11892, 2015.

[55] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*, 2020.

[56] Wenchao Yu, Wei Cheng, Charu C Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In *SIGKDD*, 2018.

[57] Jin Zhang and Yanni Xiao. Modeling strategies for controlling h1n1 outbreaks in china. *International Journal of Biomathematics*, 5(04):1250017, 2012.

[58] Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. Addgraph: Anomaly detection in dynamic graph using attention-based temporal gcn. In *IJCAI*, 2019.