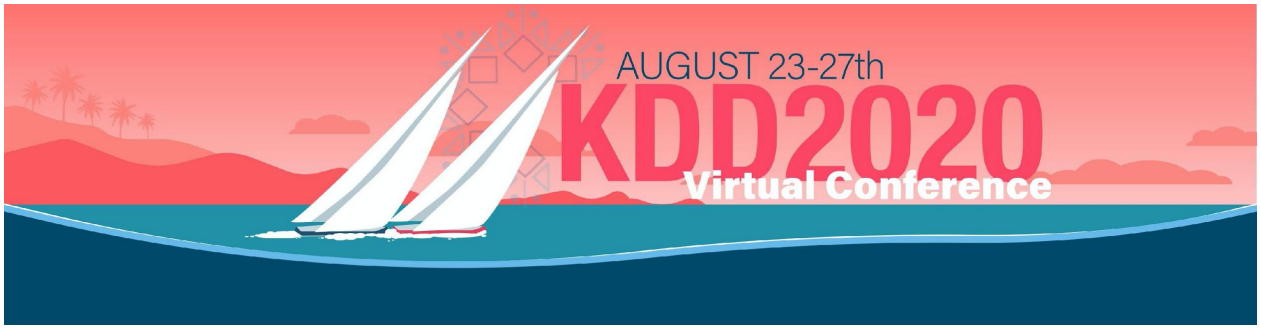


***epiDAMIK 3.0*: The 3rd International workshop on Epidemiology meets Data Mining and Knowledge discovery**

held in conjuction with ACM SIGKDD 2020
Virtual Conference.
8am - 5pm, August 24, 2020

Workshop Proceedings

Editors: B. Aditya Prakash, Anil Vullikanti, Shweta Bansal, Adam Sadilek
Mauricio Santillana, Srinivasan Venkatramanan, Bijaya Adhikari

Proceedings of the 3rd ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK 3.0)

These proceedings are not included in the ACM Digital Library.

epiDAMIK'20, August 24, 2020, Held Virtually.

ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)

Organizers:

B. Aditya Prakash (Georgia Institute of Technology)
Anil Vullikanti (University of Virginia)
Shweta Bansal (Georgetown University)
Adam Sadelik (Google)
Mauricio Santillana (Harvard University)
Srinivasan Venkatramanan (University of Virginia)
Bijaya Adhikari (University of Iowa)

Webmaster:

Bijaya Adhikari (University of Iowa)

Preface

The devastating impact of the currently unfolding global COVID19 pandemic and those of the Zika, SARS, MERS, and Ebola outbreaks over the past decade has sharply illustrated our enormous vulnerability to emerging infectious diseases. We are living in an era during which human activity is the dominant influence on climate and the environment. With escalating globalization, urbanization, and ecological pressures, the threat of a global pandemic has become more pronounced. There is an urgent need to develop sound theoretical principles and transformative computational approaches that will allow us to address the escalating threat of current and future pandemics. Data mining and Knowledge discovery have an important role to play in this regard. Different aspects of infectious disease modeling, analysis and control have traditionally been studied within the confines of individual disciplines, such as mathematical epidemiology and public health, and data mining and machine learning. Coupled with increasing data generation across multiple domains (like electronic medical records and social media), there is a clear need for analyzing them to inform public health policies and outcomes. Recent advances in disease surveillance and forecasting, and initiatives such as the CDC Flu Challenge, have brought these disciplines closer––public health practitioners seek to use novel datasets and techniques whereas researchers from data mining and machine learning develop novel tools for solving many fundamental problems in the public health policy planning process. We believe the next stage of advances will result from closer collaborations between these two communities---the main objective of epiDAMIK.

COVID-19 pandemic has highlighted, like never before, the importance of integrating large datasets spanning various domains such as infectious disease surveillance, human mobility, sociodemographic factors and public policy. It has also revealed the need for novel methods that can work with streaming, noisy datasets to support pandemic response in real-time.

The main program of epiDAMIK'20 consists of seven papers that cover various aspects of data mining and public health. In addition there were four keynotes. Five were presented orally, and six additional papers presented during the interactive poster session. These papers were selected after a thorough reviewing process. We sincerely thank the authors of the submissions and the attendees of the workshop. We also wish to thank the members of our program committee for their help in selecting a set of high-quality papers. Furthermore, we are very grateful to Milind Tambe, Amy Wesolowski, Adam Sadilek, and Sara del Valle for engaging keynote presentations.

B. Aditya Prakash
Anil Vullikanti
Shweta Bansal
Adam Sadilek
Mauricio Santillana
Srinivasan Venkatramanan
Bijaya Adhikari

August 2020

Table of Contents

Invited Talk

AI for Public Health: Learning and Planning in the Data-to-Deployment Pipeline

Milind Tambe
Gordon McKay Professor of Computer Science, Harvard University
Director "AI for Social Good", Google Research India
milindtambe@google.com

Abstract:
With the maturing of AI and multiagent systems research, we have a tremendous opportunity to direct these advances towards addressing complex societal problems. In this talk, we focus on public health challenges such as HIV prevention and TB prevention, and present research advances in multiagent systems to address one key cross-cutting challenge: how to effectively deploy our limited intervention resources in these problem domains. We present results from large-scale studies and deployments, as well as lessons learned that we hope are of use to researchers who are interested in AI for Social Impact. Achieving social impact in these domains often requires methodological advances; we will highlight key research advances in topics such as influence maximization in social networks, multi-armed bandits and agent-based modeling for addressing challenges in public health. In pushing this research agenda, we believe AI can indeed play an important role in fighting social injustice and improving society.

Bio:
Milind Tambe is Gordon McKay Professor of Computer Science and Director of Center for Research in Computation and Society at Harvard University; concurrently, he is also Director "AI for Social Good" at Google Research India. He is a recipient of the IJCAI John McCarthy Award, ACM/SIGAI Autonomous Agents Research Award from AAMAS, AAAI Robert S Engelmore Memorial Lecture award, INFORMS Wagner prize, Rist Prize of the Military Operations Research Society, the Christopher Columbus Fellowship Foundation Homeland security award, AAMAS influential paper award, best paper awards at conferences including AAMAS, IJCAI, IVA. He has also received meritorious commendations and letters of appreciation from the US Coast Guard, Los Angeles Airport, and the US Federal Air Marshals Service. Prof. Tambe is a fellow of AAAI and ACM.

Invited Talk

Use of novel data sets to understand the spatial spread of infectious diseases and allocation of public health interventions

Amy Wesolowski
Assistant Professor, Epidemiology
Johns Hopkins Bloomberg School of Public Health
awesolowski@jhu.edu

Abstract:
Increasingly novel data sets are being used to inform our understanding of infectious disease epidemiology and the spatial spread of these pathogens. One clear example has been the use of data to quantify and model human travel patterns that has broad implications for predicting the spatial spread and populations at risk for disease outbreaks. Here, we will review the use of these types of data to understand human behavior and the implications for control programs covering a wide range of applications including malaria control and elimination, dengue surveillance and preparedness, and SARS-CoV-2 transmission models. We will highlight how and where these data may be integrated in public health and used to better inform models of disease transmission.

Bio:
Amy Wesolowski is an Assistant Professor in Epidemiology at the Johns Hopkins Bloomberg School of Public Health. She received her PhD from Carnegie Mellon University in Engineering and Public Policy. She then completed postdoctoral fellowships at Harvard TH Chan School of Public Health in the Center for Communicable Disease Dynamics and Princeton University in the Department of Ecology and Evolutionary Biology. Her group's research focuses on using novel data sets, particularly mobile phone data, to quantify human behavior and use this information to inform our understanding of infectious disease epidemiology. They work on a wide range of pathogens including malaria, dengue, measles, and rubella. This research is primarily focused in low and middle-income settings including a number of field projects in Kenya, Madagascar, Zambia, and Sri Lanka.

Invited Talk

Machine-Learned Epidemiology

Adam Sadilek
Senior Engineer, Google Research
sadilekadam@google.com

Abstract:
Work in computational epidemiology to date has been limited by coarseness and lack of timeliness of observational data. Most existing models are based on hand-curated statistics that are often delayed, expensive to collect, and cover only limited jurisdictions. Our goal is to lift the state of the art in epidemiology to a new qualitative state, where real-time health predictions become feasible and actionable. We do this at scale by applying federated machine learning and secure aggregation to online data to infer what likely contributed to the contagion. In this talk, I will sample current projects at Google focusing on privacy-first epidemiology research and recent publications.
[1] nature.com/articles/s41467-019-12809-y
[2] nature.com/articles/s41746-018-0045-1
[3] science.sciencemag.org/content/sci/early/2020/07/16/science.abc5096
[4] nature.com/articles/s41562-020-0875-0

Bio:
Adam Sadilek focuses on large-scale machine learning applied to health and ecology at Google Research. Before that, he worked on speech understanding at Google[x]. Prior to joining Google, Adam was a co-founder of Fount.in, a machine learning startup providing automated text understanding.

Invited Talk

Real-time Data Fusion to Guide Disease Forecasting Models

Sara del Valle
Deputy Group leader, Information Systems and Modeling Group
Los Alamos National Laboratory
sdelvall@lanl.gov

Abstract:
Globalization has created complex problems that can no longer be adequately understood and mitigated using traditional data analysis techniques and data sources. As such, there is a need for the integration of nontraditional data streams and approaches such as social media and machine learning to address these new challenges. In this talk, I will discuss how our team is applying approaches from the weather forecasting community including data collection, assimilating heterogeneous data streams into models, and quantifying uncertainty to forecast infectious diseases. In addition, I will demonstrate that although epidemic forecasting is still in its infancy, it's a growing field with great potential and mathematical modeling will play a key role in making this happen.

Bio:
Sara Del Valle is a scientist and Deputy Group leader for the Information Systems and Modeling Group at Los Alamos National Laboratory, where she works on the development of mathematical and computational models for infectious diseases. Her research focuses on using mathematical and computational models to improve our understanding of human behavior and the spread of infectious diseases. She has developed epidemiological models for many diseases including smallpox, anthrax, HIV, pertussis, MERS-CoV, malaria, dengue, influenza, Ebola, zika, chikungunya, and COVID-19. She has also worked on investigating the role of Internet data streams on monitoring emergent behavior during outbreaks and forecasting infectious diseases. Most recently, her team is investigating the role of large-scale data analytics such as satellite imagery, Internet data, climate, and census data on detecting, monitoring, and forecasting infectious diseases.

A Data-driven Approach to Identifying Asymptomatic *C. diff* Cases

Hankyu Jang
hankyu-jang@uiowa.edu
Dept of Computer Science
The University of Iowa

Philip M. Polgreen
philip-polgreen@uiowa.edu
Dept of Internal Medicine
The University of Iowa

Alberto M. Segre
alberto-segre@uiowa.edu
Dept of Computer Science
The University of Iowa

Daniel K. Sewell
daniel-sewell@uiowa.edu
Dept of Biostatistics
The University of Iowa

Sriram V. Pemmaraju
sriram-pemmaraju@uiowa.edu
Dept of Computer Science
The University of Iowa

*For the CDC MInD-Healthcare Group

ABSTRACT

Asymptomatic carriers of an infection make it more challenging to understand the characteristics of that infection (e.g., parameters such as R_0) and to design, implement, and evaluate interventions. Asymptomatic carriers are usually not tested, which also means we do not have "ground truth" labels for these cases in our data. In this paper, we propose a 2-stage classification model for inferring asymptomatic carriers of *Clostridioides difficile* (*C. diff*) infections (CDI), a common healthcare-associated infection that causes almost half a million illnesses in the US each year. Guided by hypotheses derived from literature on risk factors for C. diff carriers, we design a Stage 1 model for detecting *asymptomatic* C. diff carriers that is trained on *symptomatic* CDI cases. We evaluate the performance of this Stage 1 model by designing a Stage 2 model to predict CDI incidence that uses among its inputs exposure to asymptomatic C. diff carriers inferred by our Stage 1 model. Results from this evaluation lead to two findings. First, our results show that the best performing Stage 1 model depends on all of the standard risk factors for CDI except for high-risk antibiotics. This is an intriguing finding that highlights an important difference between the risk profile of CDI patients and C. diff carriers. Second, we show that adding exposure to asymptomatic cases as an input to the Stage 2 CDI classification model leads to better performance. This result implies that asymptomatic C. diff carriers do in fact contribute to CDI spread, confirming an important conjecture from the CDI literature.

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning**; • **Applied computing → Health informatics**.

KEYWORDS

Asymptomatic carrier detection, Clostridioides difficile, Colonization pressure, High-risk antibiotics, Spatio-temporal clustering

ACM Reference Format:
Hankyu Jang, Philip M. Polgreen, Alberto M. Segre, Daniel K. Sewell, and Sriram V. Pemmaraju. 2020. A Data-driven Approach to Identifying Asymptomatic *C. diff* Cases. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 8 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

For many infections, *asymptomatic* cases present a major obstacle to understanding precisely how the infection is spread, and they make implementing effective interventions that much more challenging. Indeed, asymptomatic cases are widely believed to play a substantial role in the spread of COVID-19 [3, 21] and asymptomatic transmission of SARS-CoV-2 has been called the "Achilles' heel" of control strategies for COVID-19 [13].

The focus of this paper is on inferring asymptomatic cases of a common *healthcare-associated infection* (HAI) known as *C. diff infection*, or CDI. An HAI is an infection that a patient acquires in a healthcare facility while being treated for another condition. At any given time, 1 in 25 patients in the US has an HAI [23]. CDI is caused by the bacterium *Clostridioides difficile*, and is characterized by diarrhea and inflammation of the colon: there are almost half a million cases of CDI in the US each year [12]. CDI, and HAIs in general, pose a major challenge to healthcare systems worldwide, especially because some of these infections are becoming resistant to antibiotics, the primary treatment used to address these infections.

There is evidence that a substantial fraction of patients admitted to a healthcare facility are *asymptomatic* C. diff carriers [18, 19]. One particular study [19] found that up to 10% of patients admitted

to a tertiary hospital in Minnesota during March-April 2009 were in fact asymptomatic C. diff carriers. Yet the role of asymptomatic cases in the spread of CDI within healthcare facilities is largely unexplored [1], though there is accumulating indirect evidence that this role is substantial. For example, another study [10] found that 45% of CDI cases originated from sources other than symptomatic cases, suggesting a significant role for asymptomatic persons; a still more recent study [29], found that only 17% of CDI cases in a hospital ward had direct contact with other symptomatic patients, also suggesting that the pathogen had been acquired from other, presumably asymptomatic, sources.

Understanding the role of asymptomatic C. diff carriers is a critical element in designing effective interventions. Our paper presents a data-driven approach to identifying and understanding the role of asymptomatic C. diff carriers on the diffusion of CDI in a healthcare setting.

1.1 Results and Approach

Guided by literature on risk factors for being an asymptomatic C. diff carrier [16], we evaluate multiple data-driven models for inferring if a patient is an asymptomatic C. diff carrier. Our evaluation is based on retrospective data, for the time period 2007-2011, from the University of Iowa Hospitals and Clinics (UIHC), containing about 154K patient visits and associated demographic fields and rich spatio-temporal information on procedures, antibiotics, comorbidities, within-hospital transfers, etc. It is known that risk factors for (symptomatic) CDI include age, length of hospital stay, recent prior hospital admission, use of certain antibiotics considered high-risk for CDI, use of proton pump inhibitors, and severity of other comorbidities [9]. However, much less is known about the risk factors for asymptomatic C. diff carriage. Our first finding is that a predictive model for inferring asymptomatic C. diff carriage that uses all the (above mentioned) features that are risk factors for symptomatic CDI, *except for high-risk antibiotics* has good performance, relative to other models we consider. Specifically, excluding antibiotics as a risk factor seems to lead to a model with better performance than the model obtained by including antibiotics as a risk factor. This is an intriguing data-driven finding that is consistent with [16], where antibiotic use is not listed as a risk factor for asymptomatic C. diff carriage. However, as mentioned earlier, there is a lot unknown about risk factors for asymptomatic C. diff carriage and in other literature (e.g., [8]) the Cephalosporin class of antibiotics were found to be a risk factor for asymptomatic C. diff carriage.

The key difficulty in training and testing a predictive model for asymptomatic C. diff carriage is that we do not have any "ground truth" data, i.e., we have no labels identifying certain patients as being asymptomatic C. diff carriers. Our data – like most large-scale inpatient data from hospitals – only contain information on patients who tested positive for CDI, and these tests are invariably administered to patients who show symptoms. We overcome this missing label problem in two ways. First, we consider two alternative hypotheses on the relationship between CDI and asymptomatic C. diff carriage and use these hypotheses to generate a number of different prediction models for asymptomatic C. diff carriage. Second, we test out models indirectly by viewing these models for predicting asymptomatic C. diff carriage as constituting the first stage in a 2-stage model. We design the Stage 2 model for predicting *symptomatic* CDI cases. Inspired by the approach in [7, 9, 30], we use measures of exposure to asymptomatic C. diff carriers identified by the Stage 1 model as features in the Stage 2 model. Our second finding is that a model that includes exposure to asymptomatic C. diff carriers outperforms models that don't include this exposure. This finding simultaneously shows two things. First, it reveals the predictive power of our Stage 1 models and identifies Stage 1 models that outperform other models (e.g., the Stage 1 model that uses all CDI risk factors except for antibiotics). Second, it shows that exposure to asymptomatic C. diff carriers is a salient risk factor for CDI, something that has been conjectured widely in CDI literature [10, 29].

Additionally, we also investigate spatio-temporal clustering of the cases inferred to be C. diff carriers by our model. In prior work [27], we have shown that CDI cases at the UIHC exhibit spatio-temporal clustering. Using similar statistical tests, we show here that the observed CDI cases along with the inferred asymptomatic C. diff carriers also exhibit spatio-temporal clustering. This finding provides additional indirect evidence that in-hospital exposure to asymptomatic C. diff carriers may be playing a role in the spread of CDI in the hospital.

1.2 Other Related Work

Besides the papers cited earlier, there are two computational approaches to the problem of inferring asymptomatic cases, that are worth mentioning here. Makar et al. [20] define a generative probabilistic model for problem of inferring asymptomatic cases and their impact on other agents via exposure. Their main contribution is a computational method for solving for the parameters of the model. A different strand of research uses [28, 31, 32] the Steiner tree problem as a model for the problem where some nodes in a contact network are observably infected (i.e., symptomatic) and the infection status of other nodes is latent.

2 THE STAGE 1 MODEL: INFERRING ASYMPTOMATIC C. DIFF CARRIERS

2.1 The UIHC DataSet

The data used in this paper consist of anonymized electronic medical records (EMR) and admission-discharge-transfer records (ADT) for patient visits at the UIHC for the period 2007–2011. The 154,230 patient visits in the data are divided into two groups: (i) $visit_{CDI}$, visits during which patients tested positive for CDI and (ii) $visit_{CDIx}$, the rest of the visits. As in [11, 24, 30], we exclude short visits in both $visit_{CDI}$ and $visit_{CDIx}$, where patients are discharged within 48 hours of admission. The reason for excluding short visits from $visit_{CDI}$ is that such CDI cases are unlikely to be hospital-associated and the reason for then excluding short visits from $visit_{CDIx}$ is that otherwise the length of a visit field might end up being a prominent artificial signal of a non-CDI visit. For each visit in $visit_{CDIx}$, we generate one *instance* per day (*CDIx instances*) from the admission date to discharge date for that visit. Similarly, we generate daily instances (*CDI instances*) for each visit in $visit_{CDI}$, starting from the admission date, but only until three days before the CDI positive test date [22]. We exclude instances for the last three days

before a positive CDI test because there could be modifications to patient treatment during this period that could be in response to potential CDI. This process results in 8,946 CDI instances from 750 visits in $visit_{CDI}$ and 988,780 CDIx instances from 115,271 visits in $visit_{CDIx}$.

2.1.1 Individual risk factors for CDI. We include in each instance, 25 features extracted from the EMR and ADT data, which are considered risk factors for CDI in literature [7, 9]: length of stay of the visit until the date of the instance (*LOS*), *age*, *gender*, previous UIHC visit within 60 days (*PV*), the number of high-risk antibiotics prescribed (*ABX*s) and the number of gastric acid suppressors prescribed (*GAS*s) during the visit. Guided by literature on antibiotics that are considered high risk for CDI [25], we use the following five *ABX*s as features: (i) Amoxillin or Ampicillin (ABX 1), (ii) Clindamycin (ABX 2), (iii) Third generation Cephalosporin (ABX 3), (iv) Fourth genenration Cephalosporin (ABX 4), and (v) Fluoroquinolone (ABX 5). Similarly, guided by literature on risk factors for CDI [6], we use the following two *GAS*s as features: (i) H2-receptor antagonists (GAS 1), and (ii) proton pump inhibitors (GAS 2). We generate three features each for the seven medications (*ABX*s and *GAS*s): (i) prescription ($P_{medication}$), a binary feature, indicating if the medication was prescribed on the date of the instance, (ii) sum prescription count ($SP_{medication}$), number of days where the medication was prescribed to the patient, and (iii) mean prescription count ($MP_{medication} = \frac{SP_{medication}}{LOS}$) of the medication. We use ABX_x for $x \in \{1, 2, 3, 4, 5\}$ to denote the tuple $(P_{ABX_x}, SP_{ABX_x}, MP_{ABX_x})$ correspoding to ABX x. Similarly, we use GAS_x, for $x \in \{1, 2\}$ to denote the tuple $(P_{GAS_x}, SP_{GAS_x}, MP_{GAS_x})$.

2.1.2 Exposure risk factors for CDI. Colonization pressure is a measure of the proportion of patients infected or colonized with a specific pathogen in a specific physical area (e.g., a hospital ward or a geographic region) over a specified period of time [2]. Colonization pressure serves as a proxy measure for exposure, and the notion of colonization pressure has also been applied to CDI, albeit only those patients who have tested positive for CDI are included in the pressure calculation [7, 30]. Colonized patients who are asymptomatic are typically undetected and are usually excluded from pressure calculations. As has been done in other studies [7, 30], we compute this modified measure of colonization pressure, which we call *CDI pressure* and use it as an exposure risk factor for CDI.

We assume that CDI patients are infectious 3 days before the positive result and up to 14 days after the test date. For each visit in $visit_{CDI}$ and $visit_{CDIx}$, we keep track of the number of infectious CDI patients in the same room or unit, daily. From these counts, we generate the following four features:

- *Unit sum CDI pressure* (SCP_{unit}): cumulative daily number of infectious CDI patients in the same unit, from admission date up to the date of the instance
- *Room sum CDI pressure* (SCP_{room}): cumulative daily number of infectious CDI patients in the same room from admission date up to the date of the instance
- *Unit mean CDI pressure* (MCP_{unit}): $\frac{SCP_{unit}}{LOS}$
- *Room mean CDI pressure* (MCP_{room}): $\frac{SCP_{room}}{LOS}$

Table 1 summarizes basic statistics of these features for CDI visits and CDIx visits.

2.2 Training the Stage 1 Model

The goal of our Stage 1 model is to predict the likelihood of an individual being an asymptomatic C. diff carrier, as a function of certain hand-curated risk factors. As mentioned earlier, the fundamental obstacle to training this model is the fact that our data lacks "ground truth" labels. So the training of our Stage 1 model depends on hypotheses we make regarding how asymptomatic C. diff carriers relate to patients who have tested positive for CDI. The first hypothesis we consider is the following.

> **Hypothesis 1**: Asymptomatic C. diff carriers and CDI cases have similar risk profiles.

This hypothesis is not necessarily backed by studies in the literature; as mentioned earlier, the risk factors for asymptomatic C. diff carriage and the progression from C. diff carriage to CDI is not well understood. We propose this as a simple, reasonable hypothesis that allows us to train C. diff carriage prediction models that we can then evaluate. If we assume this hypothesis, we can train our Stage 1 model using CDI cases as instance labels. Then, patients who are assigned a high probability by a model trained in this manner, but are not CDI cases, are inferred to be asymptomatic C. diff carriers. Variants of this Stage 1 model can be obtained by using different subsets of features. More specifically, we partition the set of features into three groups: (i) *baseline* feature set B, consisting of *LOS*, *age*, *gender*, *PV*, GAS_1, and GAS_2, (ii) *colonization pressure* feature set CP, consisting of SCP_{unit}, SCP_{room}, MCP_{unit}, and MCP_{room}, and (iii) *ABX* feature set ABX, consisting of the 5 high-risk antibiotic feature tuples described earlier. For a subset

Table 1: Basic statistics of features. The values denote mean over each visit in $visit_{CDI}$ or $visit_{CDIx}$ and values in the bracket denote std. dev. For most of the features the values for $visit_{CDI}$ are much larger than the corresponding values for $visit_{CDIx}$ (e.g., LOS: 10.93 vs 7.58).

Feature	$visit_{CDI}$	$visit_{CDIx}$
LOS	10.93 (23.09)	7.58 (11.14)
age	53.5 (23.23)	44.23 (24.9)
$gender$	0.55 (0.5)	0.48 (0.5)
PV	0.35 (0.48)	0.19 (0.39)
SP_{GAS1}	1.71 (5.54)	0.92 (3.37)
SP_{GAS2}	5.81 (13.98)	2.98 (6.37)
MP_{GAS1}	0.17 (0.34)	0.11 (0.28)
MP_{GAS2}	0.42 (0.41)	0.33 (0.4)
SP_{ABX1}	0.46 (2.37)	0.48 (2.37)
SP_{ABX2}	0.1 (1)	0.05 (0.57)
SP_{ABX3}	0.39 (2.07)	0.21 (1.23)
SP_{ABX4}	1.2 (3.55)	0.24 (1.58)
SP_{ABX5}	1.58 (4.68)	0.73 (2.47)
MP_{ABX1}	0.04 (0.15)	0.05 (0.19)
MP_{ABX2}	0.01 (0.06)	0 (0.04)
MP_{ABX3}	0.04 (0.16)	0.03 (0.13)
MP_{ABX4}	0.1 (0.26)	0.02 (0.13)
MP_{ABX5}	0.11 (0.23)	0.08 (0.21)
SCP_{unit}	1.47 (3.18)	2.16 (4.84)
SCP_{room}	0.03 (0.23)	0.08 (0.75)
MCP_{unit}	0.23 (0.46)	0.26 (0.47)
MCP_{room}	0.01 (0.1)	0.01 (0.06)

$S \subseteq \{B, CP, ABX\}$, let D^S denote the dataset with every CDI and CDIx instance consisting of features from S. We train 4 different Stage 1 models using datasets $D^B, D^{B,CP}, D^{B,ABX}, D^{B,CP,ABX}$.

We train additional Stage 1 models on the basis of the following hypothesis.

> **Hypothesis 2**: The mechanism for acquiring (symptomatic) CDI consists of the patient first being an asymptomatic C. diff carrier and then being prescribed high-risk antibiotics.

Again, this hypothesis is not necessarily backed by medical studies, though mechanistic models for CDI (e.g., [33]) often attribute the transition from C. diff carriage to CDI to the use of additional high-risk antibiotics. This hypothesis has the following useful implication. Suppose A is the subset of patients who were prescribed high-risk antibiotics during their visit. Then, the subset $A_{CDI} \subseteq A$, consisting of patients who tested positive for CDI is exactly identical to the subset of A of patients who were asymptomatic C. diff carriers (prior to receiving antibiotics) and $A \setminus A_{CDI}$ is exactly the subset of A of patients who are not asymptomatic C. diff carriers. This motivates the restriction of our data set to just those daily instances where patients are prescribed to at least one ABX since admission. When a model is trained on this subset of data, the instances in $visit_{CDIx}$ that the model assigns the `True` label are inferred to be asymptomatic C. diff carriers. 5,483 CDI instances out of 359 visits from $visit_{CDI}$ and 374,821 CDIx instances out of 35,002 visits from $visit_{CDIx}$ result from this restriction. Using this restricted data set, we train 4 additional Stage 1 models using datasets $D^B_{ABX>0}, D^{B,CP}_{ABX>0}, D^{B,ABX}_{ABX>0}, D^{B,CP,ABX}_{ABX>0}$ that are obtained by considering different subsets of features.

2.2.1 Model training. Each dataset of instances mentioned in the previous section contains timestamped instances for the 5-year period 2007–2011. For each dataset, we build five prediction models, each model obtained by training on a 4-year subset, with one year excluded. Recall that the labels in our datasets correspond to a positive CDI test, whereas our goal for each model is to predict the likelihood of a patient being an asymptomatic C. diff carrier. For each dataset, a multi-layer perceptron model (MLP) is trained on the instances in 4 years (we use 20% of instances as a validation set, not used in training), and tested on the instances in the remaining year. We train a two-layer MLP, with a hidden layer size of 16, ReLU activation, and drop out of 0.5 using the Adam optimizer with a learning rate of 0.01 and maximum training for 200 epochs, but with an early stopping if the validation loss does not decrease for 3 consecutive epochs.

After the training and testing of the five models is completed, for each instance (a day during a patient visit), we have a probability that we interpret to be the likelihood of that patient being an asymptomatic C. diff carrier on that day. We now assign to each visit in $visit_{CDIx}$, the maximum probability of all the instances from the visit. We interpret this probability as the likelihood that the patient was a C. diff carrier during this visit. Our next step is to use these probabilities to mark a subset of the visits as being C. diff carrier visits. According to a survey [14] of studies on the prevalence of C. diff carriage, 0–17.5% of healthy adults were carriers of C. diff strains without clinical signs of CDI. Keeping this range in mind, we separately select the top 10%, top 5%, and top 3% of the visits in $visit_{CDIx}$ by probability and designate these sets of visits as $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$, respectively. Note that we have 8 different Stage 1 models, which means we have 8 different sets of $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$ as a result.

3 EVALUATING ASYMPTOMATIC C. DIFF CARRIER PREDICTIONS

The output of the Stage 1 Model is a subset of patient visits that are marked with the patient being an asymptomatic C. diff carrier during the visit. Note that the patients do not have a positive CDI test during these visits. As mentioned earlier, the key difficulty in evaluating this inference is that we do not have "ground truth" labels for asymptomatic C. diff carriers. We propose two indirect ways of validating and evaluating our Stage 1 model predictions.

(i) We design a 2-stage model for predicting symptomatic CDI cases that uses, in addition to standard risk factors of CDI, features that measure exposure to asymptomatic C. diff carriers (as predicted by our Stage 1 model). We investigate if this 2-stage model has improved performance due to inclusion of these additional exposure features. Furthermore, this framework also allows us to indirectly compare different Stage 1 models, by virtue of how well the 2-Stage model using that particular Stage 1 model performed.

(ii) We perform statistical tests to determine if the collection of CDI cases and asymptomatic C. diff carriers (as inferred by our Stage 1 model) exhibit spatio-temporal clustering. In our prior work [27], we observed statistically significant spatio-temporal interaction and clustering of CDI cases at the UIHC. Note that these were just the cases with a positive CDI test. We interpreted this finding as providing evidence of the within-hospital spread of CDI. A similar result for the collection of cases that additionally includes asymptomatic C. diff carriers will provide evidence that asymptomatic C. diff carriers also have a role to play in the within-hospital spread of CDI.

3.1 Training the Stage 2 Model

We now design a CDI prediction model that includes exposure to asymptomatic C. diff carriers (as predicted by our Stage 1 model) as features. We investigate the question of whether including these exposure features improves the CDI model prediction.

In Section 2.1.2 we defined 4 different measures, called CDI pressures, of exposure to CDI cases. In a similar manner, we define 4 measures of exposure to asymptomatic C. diff carriers. We start by assuming that any patient designated to be an asymptomatic C. diff carrier during a visit is infectious throughout the visit. This assumption leads to the following definition of *asymptomatic C. diff carrier pressures AP*, consisting of SAP_{unit}, SAP_{room}, MAP_{unit}, and MAP_{room}.

- Unit sum asymptomatic C. diff pressure (SAP_{unit}): cumulative daily exposure to asymptomatic C. diff carriers detected in the Stage 1 model in the same unit from admission date up to the date of the instance

- Room sum asymptomatic C. diff pressure (SAP_{room}): cumulative daily exposure to asymptomatic C. diff carriers in the same room from admission date up to the date of the instance
- Unit mean asymptomatic C. diff pressure (MAP_{unit}): $\frac{SAP_{unit}}{LOS}$
- Room mean asymptomatic C. diff pressure (MAP_{room}): $\frac{SAP_{room}}{LOS}$

Figure 1 shows the interaction between Stage 1 and Stage 2 models.

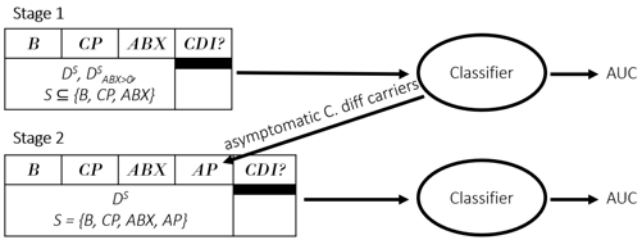


Figure 1: Diagram of the 2-stage model

In Section 2, we defined 8 different models for predicting asymptomatic C. diff carriers, 4 for each of the two hypotheses. From each of these 8 models, we get a different set of 4 exposure features, representing exposure to asymptomatic C. diff carriers. As a result, we evaluate 8 different Stage 2 models (Table 3) and for comparison we also evaluate one Stage 1 model (Table 2) without any feature corresponding to exposure to asymptomatic C. diff carriers.

3.2 Spatio-temporal Clustering of Symptomatic and Asymptomatic CDI Cases

In the previous work, we created a hospital graph of UIHC using room and spaces in a corridor as nodes (19K) and direct passage between node pairs in the 5-6m distance as edges (47K) [5]. We associate with each CDI case a timestamp (date of positive CDI test) and a location (room the patient was in at the time of positive CDI test). Two CDI cases are said to be in *spatio-temporal proximity* if the two cases occurred within 14 days of each other in rooms that are (roughly) within 30 m apart from each other, which is within 5 hop distance in the hospital graph [26]. This notion is conveniently described to be a *CDI case proximity graph* $G_{CDI} = (V_{CDI}, E_{CDI})$, where V_{CDI} is the set of CDI cases at the UIHC during the period 2007–2011 and E_{CDI} is the edges that connects pairs of CDI cases in spatio-temporal proximity. Note that CDI cases that tested positive for CDI within 48 hours of admission are not included in V_{CDI} because these cases are unlikely to be acquired during the hospital visit. We can generalize the notion of CDI case proximity graph in a natural way to include asymptomatic C. diff carriers. With each patient visit marked as an asymptomatic C. diff carrier case by our Stage 1 model, we associate a date, which is the date during the visit that was assigned the highest probability of being an asymptomatic C. diff carrier. Once a date is assigned to a visit, we can also associate a location to the visit, which is the room occupied by the patient on that date. For $x \in \{3, 5, 10\}$, let Let $G_{RCDIx\%} = (V_{RCDIx\%}, E_{RCDIx\%})$ denote the *revealed CDI case proximity graph*. Here $V_{RCDIx\%}$ is the union of the set of CDI cases and the set of asymptomatic C. diff cases output by our Stage 1 model when it was required to mark $x\%$ of visits in $visit_{CDIx}$ as asymptomatic C. diff carrier visits. Among the 8 sets of asymptomatic C. diff cases from 8 different Stage 1 models, we select the set of cases where adding the exposure features from these cases yields the *best* performance on the Stage 2 model. $E_{RCDIx\%}$ is the set of edges connecting pairs of nodes in $V_{RCDIx\%}$ that are in spatio-temporal proximity.

We compute a number of basic network statistics of $G_{RCDIx\%}, x \in \{3, 5, 10\}$ and compare these with corresponding statistics for G_{CDI} (Table 6). We then compute specific measures of network density and make a similar comparison (Table 7). Finally, we perform statistical tests on $G_{RCDIx\%}, x \in \{3, 5, 10\}$ (e.g., Knox test [17]) for testing if the union of the set of CDI cases and the set of asymptomatic C. diff cases exhibit spatio-temporal clustering. The results from these computations are described in Section 4.

4 RESULTS

4.1 Stage 1 Model

Table 2 summarizes the performance of the 8 Stage 1 Models, 4 models derived from each hypothesis (see Section 2). Recall that even though the purpose of these models is to predict asymptomatic C. diff carriers, they are trained on labeled data, where the labels indicate CDI. Table 2 shows how well these models are able to predict CDI. As an evaluation measure, we report AUC, the area under the receiver operating characteristic (ROC) curve, as the evaluation metric for our models since AUC is widely used as an evaluation metric for an imbalanced dataset. Note that our datasets are highly imbalanced: the imbalance ratio of datasets D^S and $D^S_{ABX>0}$, $S \subseteq \{B, CP, ABX\}$ is 111:1 and 68:1, respectively, which makes model training challenging. The AUCs reported in Table 2 are the test AUCs averaged over five years of training and testing on each dataset; this procedure is similar to k-fold cross-validation, but each fold corresponds to the instances in the same year. The 8 columns on the right of the table correspond to the 8 different models, as indicated by the column labels. The best performing Stage 1 Model is the one trained on $D^{B,ABX,CP}$, with a mean AUC of 0.719. This is not surprising because this model uses features of all the standard risk factors for symptomatic CDI. The next best model is the one trained on $D^{B,CP}$, with a mean AUC of 0.704. This result shows that *ABX* helps the prediction of symptomatic CDI. Again this is not surprising because high-risk antibiotics play an important role in predictive models for CDI. Exposure to CDI patients consistently help the prediction, as revealed by pairwise comparisons of models trained on features that use CDI pressures vs. those that do not use CDI pressures, e.g. $D^{B,ABX,CP}$ and $D^{B,ABX}$. The overall AUCs from the models trained on $D^S_{ABX>0}, S \subseteq \{B, CP, ABX\}$ is smaller compared to those trained on $D^S, S \subseteq \{B, CP, ABX\}$, though this comparison may not be fair since $D^S_{ABX>0}$ has a smaller set of instances compared to D^S.

Table 2: AUC on Stage 1 models

	D^B	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^B_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
AUC	0.676	0.635	0.704	*0.719	0.594	0.584	0.672	0.648

[a] AUC with asterisk denote best performer for $D^S, S \subseteq \{B, CP, ABX\}$

4.2 Stage 2 Model

The results of Stage 2 models are shown in Table 3. Each AUC in the table corresponds to the mean test AUC averaged over five years of training and testing on each dataset. As denoted by the label at the top (in the first row), every Stage 2 model evaluated here is trained on $D^{B,ABX,CP,AP}$, i.e., the dataset consisting of all risk factors for symptomatic CDI (*B*, *ABX*, and *CP*) along with *asymptomatic pressures AP*. The 24 models shown in this table differ in how asymptomatic C. diff carriers are identified in Stage 1. The 8 column labels in the second row on the right of the table correspond to the 8 different models on which the $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$ (bottom 3 rows in the table) are detected, as indicated by the column labels. The most important takeaway from this table is that using $D^{B,CP}$ as the dataset during Stage 1 consistently leads to the best performance. In other words, a model that uses baseline features (*B*) and colonization pressure features (*CP*), but not high-risk antibiotic features (*ABX*) to identify asymptomatic C. diff carriers, seems to most accurately identify C. diff carriers. This intriguing finding that is consistent with [16], seems to indicate that antibiotics that are risk factors for CDI are not associated with asymptomatic C.diff carriage.

The three Stage 2 models corresponding to $D^{B,CP}$ (AUC: 0.733, 0.729, 0.727) outperform the best performing Stage 1 model ($D^{B,ABX,CP}$, AUC: 0.719), clearly indicating that exposure to asymptomatic C. diff carriers impacts the spread of CDI. Most of the remaining Stage 2 models perform even worse than the Stage 1 model using $D^{B,ABX,CP}$. In other words, using exposure to asymptomatic C. diff carriers is worse than not using such exposure features, if asymptomatic C. diff carriers are detected poorly. Table 5 shows the *AP* of $visit_{CDI}$ and $visit_{CDIx}$ that is computed from asymptomatic C. diff carriers which are detected in Stage 1 Model on $D^{B,CP}$.

As a sensitivity test of our Stage 2 models, we train models on additional datasets that contain as features, exposure to *randomly* selected visits in $visit_{CDIx}$, instead of *AP*. We randomly select 10% of the visits in $visit_{CDIx}$, and generate 4 exposure features from these visits (*RP*) in the same manner as the *AP* features were generated. We repeat this five times to generate five different sets of random exposure features (*RP*s), namely $Random_i, i \in \{1 \cdots 5\}$. The results are in Table 4. The mean AUCs on these Stage 2 models are all worse than the AUCs obtained just by using the Stage 1 model on $D^{B,ABX,CP}$. This result shows that adding pressure features from a random subset of visits does not improve the CDI prediction.

4.3 Spatio-temporal Clustering

Table 6 shows the network statistics of G_{CDI} and revealed CDI case proximity graphs $G_{RCDIx\%} = (V_{RCDIx\%}, E_{RCDIx\%})$. Here $V_{RCDIx\%}$ is the union of the set of CDI cases and the set of asymptomatic C. diff cases output by our best-performing Stage 1 model ($D^{B,CP}$), with the requirement that $x\%$ of the visits from $visit_{CDIx}$ are marked as asymptomatic C. diff carrier visits. The number of nodes and edges ($|V|$, $|E|$), average, max and std dev of degrees ($\langle k \rangle$, k_{max}, and std), the clustering coefficient (cc), the average size of connected components $avg(|E_{cpnt}|)$, and the number of nodes and edges of the giant component ($|V_{giant}|$, $|E_{giant}|$), all increase as we add more asymptomatic C. diff cases to the graph.

Figure 2 shows a connected component of $G_{RCDI10\%}$ that contains 7 CDI cases and 48 asymptomatic C. diff carriers over 3 months period (March 21 - July 6 2011). The CDI case (July 6) in the bottom of the graph is only connected to an asymptomatic C. diff carrier (July 1) who has connections to CDI cases (June 18, June 19). This asymptomatic carrier may be attributable to the CDI case that is not directly connected with other CDI cases.

We compared the G_{CDI} and $G_{RCDIx\%}, x \in \{3, 5, 10\}$ on the four different measures of density: (1) $\frac{|E|}{|E*|}$, number of edges / number of possible edges, (2) $\frac{|E|}{|V|}$, number of edges / number of nodes, (3) $\frac{|E_{giant}|}{|V|}$, the size of the giant component / number of nodes, and (4) $\frac{avg(|E_{cpnt}|)}{|V|}$, average size of connected components / number of nodes. All of the density measures were larger in the revealed CDI case proximity graphs compared to those in the CDI case proximity graph. Furthermore, all four density measures of $G_{RCDI10\%}$ were the largest, followed by $G_{RCDI5\%}$ and $G_{RCDI3\%}$, as shown in Table 7.

Table 3: AUC on Stage 2 models

	$D^{B,ABX,CP,AP}$							
AP	D^{B}	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^{B}_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
10%	0.712	0.687	*0.733	0.710	0.700	0.724	0.697	0.703
5%	0.701	0.690	*0.727	0.685	0.693	0.714	0.689	0.702
3%	0.689	0.698	*0.729	0.690	0.710	0.704	0.686	0.711

[a]AUC with asterisk denote best performer

Table 4: AUC on Stage 2 models (pressures are computed from random selection of 10% of the visits in $visit_{CDIx}$)

	$D^{B,ABX,CP,RP}$				
RP	Random1	Random2	Random3	Random4	Random5
AUC	0.703	0.709	0.684	*0.711	0.696

[a]AUC with asterisk denote best performer

Table 5: Statistics of *AP* computed from $visit_{ACDI10\%}$ detected in Stage 1 model trained on $D^{B,CP}$

Feature	$visit_{CDI}$	$visit_{CDIx}$
SAP_{unit}	35.74 (83.24)	34.33 (64.16)
SAP_{room}	2.15 (18.26)	2.91 (16.02)
MAP_{unit}	2.99 (2.9)	3.93 (4.02)
MAP_{room}	0.15 (0.34)	0.27 (0.55)

Table 6: Network statistics

	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\|V\|$	783	4241	6546	12310
$\|E\|$	120	4150	10630	37842
$\langle k \rangle$	0.307	1.957	3.248	6.148
k_{max}	4	18	31	47
std	0.581	2.095	3.145	5.195
cc	0.013	0.306	0.443	0.561
$avg(\|E_{cpnt}\|)$	0.179	2.262	5.502	21.141
$\|V_{giant}\|$	8	118	245	1239
$\|E_{giant}\|$	10	232	738	6393

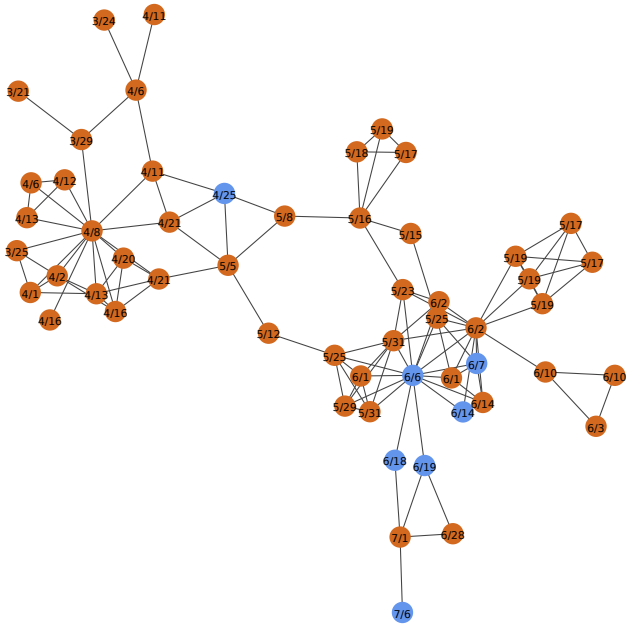


Figure 2: A connected component in $G_{RCDI10\%}$ composed of 7 CDI cases, in blue, and 48 asymptomatic C. diff carriers, in orange, over 3 months period (March 21 - July 6 2011). The CDI case (July 6) on the bottom-most of the graph is not connected to other CDI cases directly, but an asymptomatic C. diff carrier (July 1) connects them to CDI cases (June 18, June 19).

Additionally, we performed statistical tests on $G_{RCDIx\%}, x \in \{3, 5, 10\}$ to test if the union of the set of CDI cases and the set of asymptomatic C. diff cases exhibits spatio-temporal clustering. For each revealed CDI case proximity graph, we performed the Knox test by comparing the number of edges in $G_{RCDIx\%}$ with the distribution of the number of edges in the graphs that are obtained by permuting the timestamp of the cases in $G_{RCDIx\%}$ for random 100 permutations. Similarly, we test the statistical significance of the average size of the largest component ($avg(|E_{cpnt}|)$) and the size of the largest component ($|E_{giant}|$). In Table 8, the p-value of the Knox test and the average size of the largest component was 0 for all of the revealed CDI case proximity graphs that indicate spatio-temporal clustering of the cases. However, we observed that the size of the $|E_{giant}|$ in the permuted graphs is mostly larger than the revealed case proximity graphs of $G_{RCDI5\%}$ and $G_{RCDI10\%}$. Our conjecture regarding this last result is that the time interval of 5 years is not long enough to scatter the timestamps of cases far away from each other.

5 DISCUSSION AND FUTURE WORK

Our results point to several avenues for future work that involve gathering prospective clinical data. Our Stage 1 model for identifying patients who are likely to be asymptomatic C. diff carriers needs to be clinically tested. Designing low-cost clinical protocols for gathering these data and performing appropriate statistical tests is critical in order to have confidence in our results. One of our findings suggests that risk factors for asymptomatic C. diff carriage include most of the standard risk factors, with the exception of high-risk antibiotics. This finding needs to be made more precise and also tested by gathering prospective clinical data.

The datasets that are used in this paper are highly imbalanced with the imbalance ratio of 111:1 and 68:1 for $D^S_{ABX>0}$ and D^S, $S \subseteq \{B, CP, ABX, AP\}$, respectively, that makes the classification problem extremely difficult. To combat its extreme imbalance, we explored undersampling the majority instances in the training set during the training procedures of Stage 1 models; we gained some improvement in the training AUCs, but there was not much of a difference in the testing set AUCs, as we maintained the imbalance in the test set. We aim to explore oversampling strategies such as SMOTE [4] in our future work to improve the overall performance of our classifiers.

In this paper, we only consider the possibility of CDI cases being exposed to asymptomatic C. diff carriers. We do not consider more complicated chains of exposure involving sequences of asymptomatic C. diff carriers. Combining more complicated exposure chains with individual risk models is another avenue for future work. It seems possible to use formulations that involve the Steiner tree problem [28, 31, 32] for this purpose.

Another direction of the future work is using deep embedding approaches, such as Graph Convolutional Networks (GCN) [15] where we let the deep neural network to learn from individual risk factors in the EMR and their exposure to other patients that are captured in the ADT data.

Our asymptomatic C. diff carrier detection method can be applied in other infectious diseases where exposure plays an important role in disease diffusion. It is usually unknown if people we come in contact with are asymptomatic carriers of an infectious diseases. However, if data on an individual's risk factors to an infectious disease, contact information between these individuals, and a subset

Table 7: Network density

	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\frac{\|E\|}{\|E*\|}$	0.000392	0.000462	0.000496	0.000499
$\frac{\|E\|}{\|V\|}$	0.153257	0.978543	1.623892	3.074086
$\frac{\|E_{giant}\|}{\|V\|}$	0.012771	0.054704	0.112741	0.519334
$\frac{avg(\|E_{cpnt}\|)}{\|V\|}$	0.000229	0.000533	0.000841	0.001717

Table 8: Statistical test results on $G_{RCDIx\%}$ and the mean values of the statistics on the permuted graphs. Values in brackets denote std. dev.

		$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
p-value	$\|E\|$, Knox test	0	0	0
	$avg(\|E_{cpnt}\|)$	0	0	0
	$\|E_{giant}\|$	0.37	0.99	0.77
statistics	$\|E\|$	3650 (58)	9213 (115)	33790 (223)
	$avg(\|E_{cpnt}\|)$	1.87 (0.04)	4.53 (0.09)	18.56 (0.28)
	$\|E_{giant}\|$	228 (75)	1091 (142)	6620 (325)

of individuals' infectious state is available, then our model would be able to detect the latent spreaders.

6 ACKNOWLEDGMENTS

This project is funded by CDC MInD-Healthcare via CDC cooperative agreement U01CK000531. The authors acknowledge feedback from other University of Iowa CompEpi group members.

REFERENCES

[1] Faisal Alasmari, Sondra M. Seiler, Tiffany Hink, Carey-Ann D. Burnham, and Erik R. Dubberke. 2014. Prevalence and Risk Factors for Asymptomatic Clostridium difficile Carriage. *Clinical Infectious Diseases* 59, 2 (04 2014), 216–222. https://doi.org/10.1093/cid/ciu258

[2] M. J. Bonten, S. Slaughter, A. W. Ambergen, M. K. Hayden, J. van Voorhis, C. Nathan, and R. A. Weinstein. 1998. The role of "colonization pressure" in the spread of Vancomycin-resistant Enterococci: an important infection control variable. *Arch. Intern. Med.* 158, 10 (May 1998), 1127–1132.

[3] Diana C Buitrago-Garcia, Dianne Egli-Gany, Michel J Counotte, Stefanie Hossmann, Hira Imeri, Aziz Mert Ipekci, Georgia Salanti, and Nicola Low. 2020. The role of asymptomatic SARS-CoV-2 infections: rapid living systematic review and meta-analysis. *medRxiv* (2020). https://doi.org/10.1101/2020.04.25.20079103

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] D.E. Curtis, C.S. Hlady, G. Kanade, S.V. Pemmaraju, P.M. Polgreen, and A.M. Segre. 2013. Healthcare Worker Contact Networks and the Prevention of Hospital-Acquired Infections. *PLOS One* (December 2013). https://doi.org/doi:10.1371/journal.pone.0079906

[6] Sandra Dial, JAC Delaney, Alan N Barkun, and Samy Suissa. 2005. Use of gastric acid–suppressive agents and the risk of community-acquired Clostridium difficile–associated disease. *Jama* 294, 23 (2005), 2989–2995.

[7] E. R. Dubberke, K. A. Reske, M. A. Olsen, K. M. McMullen, J. L. Mayfield, L. C. McDonald, and V. J. Fraser. 2007. Evaluation of Clostridium difficile-Associated Disease Pressure as a Risk Factor for C difficile-Associated Disease. *Archives of Internal Medicine* 167, 10 (05 2007), 1092–1097. https://doi.org/10.1001/archinte.167.10.1092

[8] Erik R. Dubberke, Kimberly A. Reske, Sondra Seiler, Tiffany Hink, Jennie H. Kwon, and Carey-Ann D. Burnham. 2015. Risk Factors for Acquisition and Loss of Clostridium difficile Colonization in Hospitalized Patients. *Antimicrobial Agents and Chemotherapy* 59, 8 (2015), 4533–4543. https://doi.org/10.1128/AAC.00642-15

[9] Erik R Dubberke, Yan Yan, Kimberly A Reske, Anne M Butler, Joshua Doherty, Victor Pham, and Victoria J Fraser. 2011. Development and validation of a Clostridium difficile infection risk prediction model. *Infection Control & Hospital Epidemiology* 32, 4 (2011), 360–366.

[10] David W. Eyre, Madeleine L. Cule, Daniel J. Wilson, David Griffiths, Alison Vaughan, Lily O'Connor, Camilla L.C. Ip, Tanya Golubchik, Elizabeth M. Batty, John M. Finney, David H. Wyllie, Xavier Didelot, Paolo Piazza, Rory Bowden, Kate E. Dingle, Rosalind M. Harding, Derrick W. Crook, Mark H. Wilcox, Tim E.A. Peto, and A. Sarah Walker. 2013. Diverse Sources of C. difficile Infection Identified on Whole-Genome Sequencing. *New England Journal of Medicine* 369, 13 (2013), 1195–1205. https://doi.org/10.1056/NEJMoa1216064 PMID: 24066741.

[11] Centers for Disease Control and Prevention. Jan, 2020 (accessed June 11, 2020). *Identifying Healthcare-associated Infections (HAI) for NHSN Surveillance.* https://www.cdc.gov/nhsn/pdfs/pscmanual/2psc_identifyinghais_nhsncurrent.pdf

[12] Centers for Disease Control and Prevention. Reviewed Nov 13, 2019 (accessed June 10, 2020). *Clostridioides difficile Infection.* https://www.cdc.gov/hai/organisms/cdiff/cdiff_infect.html

[13] Monica Gandhi, Deborah S. Yokoe, and Diane V. Havlir. 2020. Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *New England Journal of Medicine* 382, 22 (2020), 2158–2160. https://doi.org/10.1056/NEJMe2009758

[14] Schäffler H. and Breitrück A. 2018. Clostridium difficile - From Colonization to Infection. *Frontiers in Microbiology* 9, 646 (2018). https://doi.org/10.3389/fmicb.2018.00646

[15] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* (2016).

[16] Ling Yuan Kong, Nandini Dendukuri, Ian Schiller, Anne-Marie Bourgault, Paul Brassard, Louise Poirier, François Lamothe, Claire Béliveau, Sophie Michaud, Nathalie Turgeon, et al. 2015. Predictors of asymptomatic Clostridium difficile colonization on hospital admission. *American journal of infection control* 43, 3 (2015), 248–253.

[17] Martin Kulldorff and Ulf Hjalmars. 1999. The Knox Method and Other Tests for Space-Time Interaction. *Biometrics* 55, 2 (1999), 544–552. http://www.jstor.org/stable/2533804

[18] Lorraine Kyne, Michel Warny, Amir Qamar, and Ciarán P Kelly. 2000. Asymptomatic carriage of Clostridium difficile and serum levels of IgG antibody against toxin A. *New England Journal of Medicine* 342, 6 (2000), 390–397.

[19] Surbhi Leekha, Kimberly C Aronhalt, Lynne M Sloan, Robin Patel, and Robert Orenstein. 2013. Asymptomatic Clostridium difficile colonization in a tertiary care hospital: admission prevalence and risk factors. *American journal of infection control* 41, 5 (May 2013), 390–393. https://doi.org/10.1016/j.ajic.2012.09.023

[20] Maggie Makar, John V. Guttag, and Jenna Wiens. 2018. Learning the Probability of Activation in the Presence of Latent Spreaders. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018.* 134–141. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16980

[21] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. 2020. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* 25, 10, Article 2000180 (2020). https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180

[22] M.N. Monsalve, S.V. Pemmaraju, S. Johnson, and P.M. Polgreen. 2015. Improving Risk Prediction of Clostridium Difficile Infection Using Temporal Event-Pairs. In *2015 International Conference on Healthcare Informatics.* 140–149. https://doi.org/10.1109/ICHI.2015.24

[23] U.S. Department of Health and Human Services. Jan 15, 2020 (accessed June 10, 2020). *Health Care-Associated Infections.* https://health.gov/our-work/health-care-quality/health-care-associated-infections

[24] Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E Ryan, Laraine Washer, Lauren R West, Vincent B Young, John Guttag, et al. 2018. A generalizable, data-driven approach to predict daily risk of Clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology* 39, 4 (2018), 425–433.

[25] Robert C Owens Jr, Curtis J Donskey, Robert P Gaynes, Vivian G Loo, and Carlene A Muto. 2008. Antimicrobial-associated risk factors for Clostridium difficile infection. *Clinical Infectious Diseases* 46, Supplement_1 (2008), S19–S31.

[26] S. Pai, S.V. Pemmaraju, P.M. Polgreen, A.M. Segre, and D.K. Sewell. In press. Spatiotemporal Clustering of In-Hospital *Clostridioides difficile* Infection (CDI). *Infection Control and Hospital Epidemiology* (June In press).

[27] Shreyas Pai, Philip M. Polgreen, Alberto Maria Segre, Daniel K. Sewell, and Sriram V. Pemmaraju. 2020. Spatiotemporal clustering of in-hospital Clostridioides difficile infection. *Infection Control & Hospital Epidemiology* 41, 4 (2020), 418–424. https://doi.org/10.1017/ice.2019.350

[28] Polina Rozenshtein, Aristides Gionis, B. Aditya Prakash, and Jilles Vreeken. 2016. Reconstructing an Epidemic Over Time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* Association for Computing Machinery, New York, NY, USA, 1835–1844. https://doi.org/10.1145/2939672.2939865

[29] García-Fernández S, Frentrup M, Steglich M, Gonzaga A, Cobo M, López-Fresnena N, Cobo J, Morosini MI, Cantón R, Del Campo R, and Nübel U. 2019. Whole-genome sequencing reveals nosocomial Clostridioides difficile transmission and a previously unsuspected epidemic scenario. *Sci Rep.* 9 (May 2019). Issue 1.

[30] N. Khan P.M. Polgeen A.M. Segre D.K. Sewell T. Riaz, A. Kharkar and S.V. Pemmaraju. 2020. Highly Local CDI Pressures As Risk Factors for CDI. To appear in Decennial 2020: 6th International Conference on Healthcare Associated Infections, Atlanta, GA.

[31] Han Xiao, Çigdem Aslay, and Aristides Gionis. 2018. Robust Cascade Reconstruction by Steiner Tree Sampling. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018.* IEEE Computer Society, 637–646. https://doi.org/10.1109/ICDM.2018.00079

[32] Han Xiao, Polina Rozenshtein, Nikolaj Tatti, and Aristides Gionis. 2018. Reconstructing a cascade from temporal observations. In *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*, Martin Ester and Dino Pedreschi (Eds.). SIAM, 666–674. https://doi.org/10.1137/1.9781611975321.75

[33] Laith Yakob, Thomas V. Riley, David L. Paterson, and Archie C.A. Clements. 2013. Clostridium difficile exposure as an insidious source of infection in healthcare settings: an epidemiological model. *BMC Infectious Diseases* 13, 376 (2013). https://doi.org/10.1186/1471-2334-13-376

Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks

Amol Kapoor*, Xue Ben*, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, Shawn O'Banion
[ajkapoor,sherryben,luyangliu,hubris,mattbarnes,blais,obanion]@google.com
Google Research

ABSTRACT

In this work, we examine a novel forecasting approach for COVID-19 case prediction that uses Graph Neural Networks and mobility data. In contrast to existing time series forecasting models, the proposed approach learns from a single large-scale spatio-temporal graph, where nodes represent the region-level human mobility, spatial edges represent the human mobility based inter-region connectivity, and temporal edges represent node features through time. We evaluate this approach on the US county level COVID-19 dataset, and demonstrate that the rich spatial and temporal information leveraged by the graph neural network allows the model to learn complex dynamics. We show a 6% reduction of RMSLE and an absolute Pearson Correlation improvement from 0.9978 to 0.998 compared to the best performing baseline models. This novel source of information combined with graph based deep learning approaches can be a powerful tool to understand the spread and evolution of COVID-19. We encourage others to further develop a novel modeling paradigm for infectious disease based on GNNs and high resolution mobility data.

1 INTRODUCTION

From late 2019 to early 2020, COVID-19 went from a local outbreak to a worldwide pandemic, one that has infected over 6.67M people and resulted in over 391K deaths worldwide [29]. Between large-scale country-wide quarantines and 'lockdowns', COVID-19 is responsible for an estimated 3-10 trillion dollars in economic damage to the global economy [21]. In a state of pandemic, the ability to accurately forecast caseload is extremely important to help inform policymakers on how to provision limited healthcare resources, rapidly control outbreaks, and ensure the safety of the general public.

In order to prepare, understand, and control the spread of the disease, researchers worldwide have come together in a collaborative effort to model and forecast COVID-19. Based on our review of the literature, there are two popular approaches for such epidemiological modelling. One is the mechanistic approach – for example, compartmental and agent based models that hard-code predefined disease transmission dynamics at either the population level [24, 34] or the individual level [8]. The other is the time series learning approach – for example, applying curve-fitting [20], Autoregression (AR) [12], or deep learning [34] on time series data.

These approaches often assume a relatively closed-system, where forecasts for a given location are dependent only on information from that location or some observed patterns from other locations. In practice, we intuit that infection data on inter-regional interactions provides a unique and highly meaningful avenue for modelling forecasts. In other words, it is reasonable that a regionâĂŹs future disease cases are dependent on its own historical information as well as other regions', people traveling to/out of this region and regions with similar epidemic patterns, etc. Based on this insight, we believe we can improve forecast accuracy by 1) utilizing more accurate real-time data that can describe the inter-region interactions and region-level mobility and 2) developing a unifying approach that can encompass both the temporal and spatial interactions for infectious disease modeling. Historically, this kind of regional movement is difficult to capture. However, researchers have correctly noticed that the widespread use of GPS enabled mobile devices provides a novel and highly accurate source of mobility data, and have called upon the epidemiological community to make ample use of this powerful new data source [7, 23].

In this work, we focus on the problem of forecasting COVID-19 at the county level in the United States. We propose a spatio-temporal graph neural network that can learn the complex dynamics inherent to disease modeling, and use this model to make forecasts on COVID-19 daily new cases from fine-grained mobility data. We run several experiments showing the power of novel mobility data within the GNN framework, and conclude with an analysis of mobility data and its potential in tracking disease spread.

2 BACKGROUND

2.1 Mobility Data in Graphs

Obtaining fine-grained human mobility data that can effectively capture the inter- and intra-region flows of human activity has become significantly more feasible in the last decade. In addition to being vital for accurately modeling disease spread, these data sources are especially important to understand the efficacy of non-pharmaceutical interventions (NPI) against COVID-19, such as social distancing, shelter-in-place, and the shut-down of interstate and international travel.

The rapid work of the epidemiological academic community was vital for understanding the role of international flights in the early spread of COVID-19 to different countries [1, 34], while epidemic curve fitting analysis for COVID-19 on the SafeGraph dataset [31] helped to better model the effects and efficacy of social distancing. We build on those efforts by examining and utilizing two Google mobility datasets, which offer a global and comprehensive view

* Both authors contributed equally to this work.

Conference'20, August 2020, San Diego, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

of inter- and intra-region human mobility. These datasets are described in more detail in subsection 4.1.

2.2 Spatio-Temporal Graph Neural Networks

Graphs are natural representations for a wide variety of real-life data in social, biological, financial, and many other fields. Recently, graph neural network (GNN) based deep learning methods [4, 6, 32, 37, 38] have shown superior performance on several tasks, including semi-supervised node classification [14, 16, 28], link prediction [5, 17, 36], community detection [9, 15, 26], graph classification [13, 22, 33], and recommendations [19, 35].

Spatio-temporal graphs are a kind of graph that model connections between nodes as a function of time and space, and have found uses in a wide variety of fields [25]. GNNs have been successfully applied to spatio-temporal traffic graphs [11] and (especially relevant to this work) spatio-temporal influenza forecasting [10]. In these latter two cases, temporal dependencies were primarily incorporated at the model level, either through decomposition of a dynamic Laplacian matrix or through a recurrent neural net.

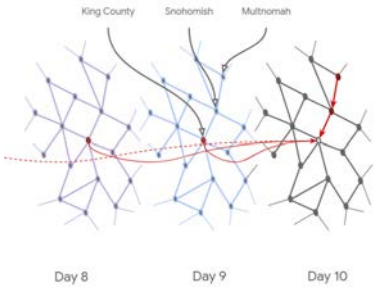


Figure 1: A slice of the COVID-19 graph showing spatial and temporal edges (highlighted in red) across three days. Each slice represents spatial connections between counties, while the connections between slices represent temporal relationships. For clarity, only temporal edges to the center node are shown; in practice, every node in the graph has direct temporal edges to nodes in d previous days.

3 METHOD

3.1 Graph Neural Networks

The core insight behind graph neural network models is that the transformation of the input node's signal can be coupled with the propagation of information from a node's neighbors in order to better inform the future hidden state of the original input. This is most evident in the message-passing framework proposed by Gilmer et al. [13], which unifies many previously proposed methods. In such approaches, the update at layer $(l+1)$ is:

$$\boldsymbol{m}_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \mathcal{F}^{(l)}\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}\right), \quad \boldsymbol{h}_i^{(l+1)} = \mathcal{G}^{(l)}\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{m}_i^{(l+1)}\right) \quad (1)$$

where $\mathcal{F}^{(l)}$ and $\mathcal{G}^{(l)}$ are learned message functions and node update functions respectively, $\boldsymbol{m}^{(l)}$ are the messages passed between nodes, and $\boldsymbol{h}_i^{(l)}$ are the node representations. The computation is carried out in two phases: first, messages are propagated along the neighbors; and second, the messages are aggregated to obtain the updated representations.

3.2 Modelling the COVID-19 Graph

In infectious disease modeling, we usually have multiple time-series sequences that represent the observables of transmission dynamics in each location. The prediction problem is usually formulated as a regression learning task that takes in a certain time series $t-k, \ldots, t-1, t$ and outputs a single value $t+1$ or future time series $t+1, t+2, \ldots$ as forecasted values. However, time series make a poor fit for modeling human mobility across locations. Mobility data is naturally represented as a spatial-graph, where any individual node represents a location i that is connected to an arbitrary number of other nodes $j, l, m, \ldots$, and where edge-weights correspond to measures of human mobility between the nodes.

In order to model spatial and temporal dependencies, we create a graph with different edge types. In the spatial domain, edges represent direct location-to-location movement and are weighted based on mobility flows normalized against the intra-flow (in other words, the amount of flow internal to the location). In the temporal domain, edges simply represent binary connections to past days. The graph manifests as 100 stacked layers. Each layer represents the county connectivity graph for that day, with the bottom layer representing Feb 22nd, 2020 (when cases began appearing in earnest in the US), and the top layer representing May 31st, 2020. Each node within each layer has direct edges to the 7 nodes directly before it in time, i.e. a week's worth of temporal information. We provide a visual of a part of the graph in Figure 1.

3.3 Skip-Connections Model

For our graph convolutions, we use a version of the spectral graph convolution model proposed by Kipf and Welling [16], modified with skip-connections between layers to avoid diluting the self-node feature state. Specifically, the output of each layer is concatenated with a learned embedding from the temporal node features. The model prediction P can be represented as:

$$\mathbf{H_0} = \text{mlp}(\mathbf{x_t}|\mathbf{x_{t-1}}|...|\mathbf{x_{t-d}}) \quad (2)$$

$$\mathbf{H_{l+1}} = \sigma(\hat{\mathbf{A}}\mathbf{H_l}\mathbf{W_l}) \mid \mathbf{H_0} \quad (3)$$

$$\mathbf{P} = \text{mlp}(\mathbf{H_s}) \quad (4)$$

where H represents the hidden state at layer l, $\hat{A}$ is the spectral normalized adjacency matrix, W is the learned weight matrix at layer l, $|$ is the concat operator, and σ is a nonlinearity (in our case, a relu). See Figure 2 for a visual representation. The first embedding, H_0, is simply the output of an mlp over the node's temporal features x at time t reaching back d days, while the final prediction is the output of an mlp over s spatial hops.

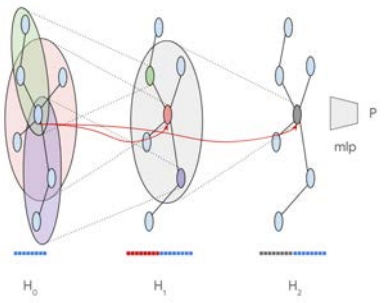


Figure 2: A visualization of a 2-hop Skip-Connection model. Multiple layers of spatial aggregations are used on temporal embedding vectors. At each layer, the embedding of the seed-node (represented in blue) is concatenated and propagated up to the next embedding layer. The final embedding is passed through an MLP and used to predict P.

4 EXPERIMENTS

4.1 Data

We make use of three datasets: the New York Times (NYT) COVID-19 dataset[1], the Google COVID-19 Aggregated Mobility Research Dataset, and the Google Community Mobility Reports[2]. The Aggregated Mobility Research Dataset helps us understand the quantity of movement, while the Community Mobility Reports helps us understand the dynamics of various types of movement. Together, these datasets add significant lift to the standard node features provided by the NYT.

4.1.1 Common Node Features. Each node contains features for state, county, day, past cases, and past deaths. The latter two are represented as normalized vectors that stretch back d days. We use COVID-19 case and death count numbers published by the New York Times [27], which includes daily reports of new infections and deaths at both state and county level in US.

4.1.2 Aggregated Mobility Research Dataset. The Google COVID-19 Aggregated Mobility Research Dataset aggregates weekly flows of users from region to region, where the region is at a resolution of 5km^2. The flows can be further aggregated to obtain inter-county flows and intra-county flows(source and destination regions are in the same county) to build our proposed graph network. This information is useful for understanding how people move before and during the pandemic – for example, Figure 3 shows the reduction in inter-county flows in US counties in April, compared to a January baseline. Figure 4 illustrates the change in mobility to King County, Washington, where mobility dropped by nearly 100% from distant counties, likely due to reductions in air travel. By comparison, reductions are less strong from nearby counties, e.g. 64% reduction from Snohomish County, Washington. For a full description of how the Aggregated Mobility Research Dataset is created, see (Appendix) 6.1.

[1]https://github.com/nytimes/covid-19-data
[2]https://www.google.com/covid19/mobility/

4.1.3 Community Mobility Reports. The Community Mobility Reports summarize mobility trends at various categories of places that are aggregated at the county level. The categories include: grocery and pharmacy, parks, transit stations, workplaces, residential, and retail and recreation. The dataset was normalized to have 0 as the 'normal' mobility based on median value for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020 [18], and deviations are measured as the relative changes in mobility from the baseline. A value of -0.25 under transit stations therefore represents a 25% reduction in visits to public transit stations compared against baseline. Figure 5 provides a visual example of the daily mobility changes in King County, Washington for each category in Google's Community Mobility Reports.

4.1.4 Limitations of Data Sources. These results should be interpreted in light of several important limitations. First, the Google mobility data is limited to smartphone users who have opted in to Google's Location History feature, which is off by default. These data may not be representative of the population as whole, and furthermore their representativeness may vary by location. Importantly, these limited data are only viewed through the lens of differential privacy algorithms, specifically designed to protect user anonymity and obscure fine detail. Moreover, comparisons across rather than within locations are only descriptive since these regions can differ in substantial ways. This data can be viewed as similar to the data used to show how busy certain types of places are in Google Maps — for example, helping identify when a local business tends to be the most crowded.

We also note that there are significant other factors not captured in any of these datasets, such as the increased prevalence of wearing masks or changes in the weather. These factors, combined with increased awareness, can effectively reduce the transmission even when mobility remains unchanged. We encourage future work that explores the addition of these external features.

4.2 Hyperparameters, Architectures, and Splits

Unless explicitly stated otherwise, for all of our GNN experiments, we use a 7 day (i.e. one week) time horizon and look over 2 hops of spatial data (using the 32 neighbors with the highest edge weight for each hop). GNN models were implemented in Tensorflow. We utilize an ADAM optimizer with learning rate set to 1e-5. We use a two hop spatial model with a single layer MLP on either side. Therefore, we have four hidden layers – an initial embedding layer, the two hops of spatial aggregation, and the final prediction layer. The hidden layer architecture for W_0, W_1, W_2, and W_3 are [64, 32, 32, 32], respectively. Each layer has a dropout rate of 0.5, and a l2 regularization term of 5e-4. GNN models were trained for 1M steps with a MSLE regression loss.

All models were trained to predict the change in the number of cases on day $t + 1$, given previous information. We have data from January 1st onwards; however, we do not observe cases in the US until late February. As a result, we use data from days 59-120 (roughly, March and April, 2020) for training, and data from days 120 to 150 (roughly, May, 2020) was used for testing. For each model, we look at the top 20 counties by population. The reported values are averaged across all counties for all thirty days of inference.

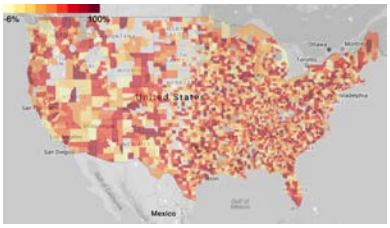


Figure 3: The reduction of inter-county mobility flow for US counties, comparing flows in April to baseline values in the first 6 weeks of 2020.

Figure 4: The reduction of inflow for King county from various US counties. Note that because King county has an airport, it has direct edges to US counties that may be physically distant.

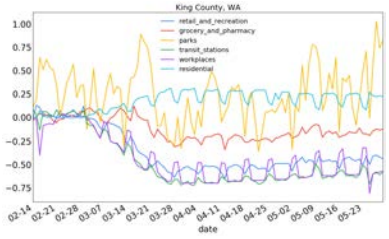


Figure 5: The mobility trends for King county. There are dramatic reductions in many of the mobility categories in late March due to non-pharmaceutical interventions like social distancing and quarantine.

4.3 Baselines

To evaluate the benefits of the GNN framework, we compare against a range of popular methods as baselines. For all of our baselines, we examine how region-level mobility features, such as aggregated flows and place visit trends, affect our results. 'No Mob' versions of our baselines indicate that these baselines do not utilize any mobility information.

4.3.1 Previous Day. Next day case prediction is highly correlated with features from the previous day. We use two previous day baselines. For Previous Delta, we predict that the delta in the number of cases will be the same as the delta from the previous day. For Previous Cases, we predict that the delta in the number of cases will be 0 (and that the actual number of cases will be the same as the previous day). These baselines help us understand what lift, if any, our models are able to extract from the rest of the provided features; however, we do not treat these as 'model' baselines in our analysis.

4.3.2 ARIMA. We utilize a univariate ARIMA model that treats the time dependent daily new cases as a univariate time series that follows a fixed dynamic. Each day's new case count is dependent on the previous p days of observations and the previous q days of estimation errors. We selected the order of the ARIMA model using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to balance model complexity and generalization, we minimize parameters by using a constant trend with ARIMA(7, 1, 3).

4.3.3 LSTM and Seq2Seq. Our LSTM baseline contains a stack of two LSTM layers (with 32, 16 units respectively) and a final dense layer. The LSTM layers encode sequential information from input through the recurrent network. The dense connected layer takes the final output from the second LSTM layer and outputs a vector of size four, which is equal to the number of steps ahead predictions needed.

The Seq2Seq model has an encoder-decoder architecture, where the encoder is composed of a fully connected dense layer and a GRU layer that can learn from sequential input and return a sequence of encoded outputs in a final hidden state. The decoder is an inverse of the encoder. The dense layer is 16 units and the GRU layer is 32 units. To match common practice, we apply Bahdanau attention [2] on the sequence of encoder outputs at each decoding step to make next step prediction. Both the LSTM and Seq2Seq models, we use a Huber loss, an Adam optimizer with a learning rate of 0.02, and a dropout rate of 0.2 for training, which works best in our experiments. During inference, both models observe data from the previous 10 days in order to make a prediction about the next day in the sequence.

4.4 Case Prediction Performance

In Table 1, we compare the forecasting performance of the spatio-temporal GNN with a range of baseline models. We report the RMSLE and Pearson Correlation for the predicted caseload (RMSLE, Corr), calculated as the sum of the predicted delta and the previous day's cases. We aggregate the performance metrics from top 20 populated counties in US. We note that we can trivially achieve a high correlation because the problem framing naturally relies on the general trend of the data from time t – in fact, the Previous Cases baseline achieves the highest case correlation overall. To account for this, we also report the RMSLE and Pearson Correlations for the case deltas (Δ RMSLE, Δ Corr), even though we expect the ground truth values to be confounded by unaccounted variables like the availability of testing centers or whether it is a workday.

We find that the GNN successfully outperforms our baselines, achieving either best or second-best score on each evaluation metric. Further, we note that for all of our deep models, introducing additional mobility data improves results. Interestingly, introducing mobility data resulted in worse performance for the ARIMA baseline. ARIMA assumes fixed dynamics and a linear dependence on the county-level mobility – while this helps the ARIMA model in the early stages of the epidemic, when there was a strong positive correlation between reduced mobility and daily new cases, it may cause the model to under-perform with the increase of mobility in late May.

5 CONCLUSION

In this work we developed a graph neural network based approach for COVID-19 forecasting with spatio-temporal mobility signals. This modeling framework can be readily extended to regression problems with large scale spatio-temporal data – in particular for our case, disease status reports and human mobility patterns at various temporal and geographical scales. In comparison to previous mechanistic or autoregressive approaches, our model does not

Model	RMSLE	Corr	Δ RMSLE	Δ Corr
Previous Cases	0.0226	**0.9981**	4.7879	NaN
Previous Delta	0.0127	0.9965	0.9697	0.1854
No Mob ARIMA	0.0124	0.9968	0.9217	0.1449
ARIMA	0.0144	0.9952	0.9624	0.0966
No Mob LSTM	0.0125	0.9978	0.9172	0.1540
LSTM	0.0121	0.9978	0.9163	0.1863
No Mob Seq2Seq	0.0118	0.9976	0.8467	0.1020
Seq2Seq	0.0116	0.9977	0.8634	**0.2802**
GNN	**0.0109**	0.9980	**0.7983**	0.2230

Table 1: Summary of model performances.

rely on assumptions of the underlying disease dynamics and can learn from a variety of data, including inter-region interaction and region-level features.

There is still much to be done, both for COVID-19 and for modeling infectious disease in general; we hope that this paper sparks an increased focus on leveraging this powerful new source of mobility information through novel techniques in graph learning. Future work can expand on these results by incorporating new features, expanding the time horizon for long term predictions, and experimenting on epidemiological mobility data in other parts of the world.

REFERENCES

[1] Aniruddha Adiga, Srinivasan Venkatramanan, James Schlitt, Akhil Peddireddy, Allan Dickerman, Andrei Bura, Andrew Warren, Brian D Klahn, Chunhong Mao, Dawen Xie, Dustin Machi, Erin Raymond, Fanchao Meng, Golda Barrow, Henning Mortveit, Jiangzhuo Chen, Jim Walke, Joshua Goldstein, Mandy L Wilson, Mark Orr, Przemyslaw Porebski, Pyrros A Telionis, Richard Beckman, Stefan Hoops, Stephen Eubank, Young Yun Baek, Bryan Lewis, Madhav Marathe, and Chris Barrett. 2020. Evaluating the impact of international airline suspensions on the early global spread of COVID-19. *medRxiv* (2020). https://doi.org/10.1101/2020.02.20.20025882 arXiv:https://www.medrxiv.org/content/early/2020/03/02/2020.02.20.20025882.full.pdf

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[3] Aleix Bassolas, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, Bryant Gipson, Surendra A Hazarie, Henry Kautz, et al. 2019. Hierarchical organization of urban mobility and its connection with city livability. *Nature communications* 10, 1 (2019), 1–10.

[4] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).

[5] Aleksandar Bojchevski and Stephan Günnemann. 2018. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. (2018).

[6] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.

[7] Caroline O Buckee, Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, Urs Gasser, Yonatan H Grad, Bryan Grenfell, M Elizabeth Halloran, Moritz UG Kraemer, et al. 2020. Aggregated mobility data could help fight COVID-19. *Science (New York, NY)* 368, 6487 (2020), 145.

[8] Sheryl L Chang, Nathan Harding, Cameron Zachreson, Oliver M Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. *arXiv preprint arXiv:2003.10218* (2020).

[9] Zhengdao Chen, Lisha Li, and Joan Bruna. 2018. Supervised Community Detection with Line Graph Neural Networks. (2018).

[10] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. 2019. Graph Message Passing with Cross-location Attentions for Long-term ILI Prediction. *arXiv preprint arXiv:1912.10202* (2019).

[11] Zulong Diao, Xin Wang, Dafang Zhang, Yingru Liu, Kun Xie, and Shaoyao He. 2019. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 890–897.

[12] James Durbin and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Methods: Second Edition* (2nd ed.). Oxford University Press.

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212* (2017).

[14] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[15] Tatsuro Kawamoto, Masashi Tsubaki, and Tomoyuki Obuchi. 2018. Mean-field theory of graph neural networks in graph partitioning. In *NeurIPS*.

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

[17] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[18] Google LLC. 2020. Google COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/ (visited on 6/4/2020). (2020).

[19] Federico Monti, Michael M. Bronstein, and Xavier Bresson. 2017. Geometric Matrix Completion with Recurrent Multi-Graph Neural Networks. In *NIPS*.

[20] CJ Murray et al. 2020. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. (2020).

[21] United Nations. 2020. COVID-19 to slash global economic output by 8.5 trillion over next two years. https://www.un.org/development/desa/en/news/policy/wesp-mid-2020-report.html (visited on 6/4/2020). (2020).

[22] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*. 2014–2023.

[23] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. (2020).

[24] Sen Pei and Jeffrey Shaman. 2020. Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US. *medRxiv* (2020). https://doi.org/10.1101/2020.03.21.20040303 arXiv:https://www.medrxiv.org/content/early/2020/03/27/2020.03.21.20040303.full.pdf

[25] Alex Reinhart et al. 2018. A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* 33, 3 (2018), 299–318.

[26] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. 2018. SpectralNet: Spectral Clustering using Deep Neural Networks. *arXiv preprint arXiv:1801.01587* (2018).

[27] The New York Times. 2020. The New York Times COVID-19 Tracking Page. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html. (2020).

[28] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* 1, 2 (2017).

[29] WHO. 2020. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/ (visited on 6/4/2020). (2020).

[30] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. 2020. Differentially private sql with bounded user contribution. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 230–250.

[31] Spencer Woody, Mauricio Garcia Tec, Maytal Dahan, Kelly Gaither, Michael Lachmann, Spencer Fox, Lauren Ancel Meyers, and James G Scott. 2020. Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones. *medRxiv* (2020).

[32] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* (2019).

[33] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How Powerful are Graph Neural Networks? *arXiv preprint arXiv:1810.00826* (2018).

[34] Zifeng Yang, Zhiqi Zeng, Ke Wang, SS Wong, W Liang, M Zanin, P Liu, X Cao, Z Gao, Z Mai, et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12, 2 (2020).

[35] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. *arXiv preprint arXiv:1806.01973* (2018).

[36] Muhan Zhang and Yixin Chen. 2018. Link Prediction Based on Graph Neural Networks. *arXiv preprint arXiv:1802.09691* (2018).

[37] Ziwei Zhang, Peng Cui, and Wenwu Zhu. 2018. Deep Learning on Graphs: A Survey. *CoRR* abs/1812.04202 (2018).

[38] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Graph Neural Networks: A Review of Methods and Applications. *arXiv preprint arXiv:1812.08434* (2018).

6 APPENDIX

6.1 Google COVID-19 Aggregated Mobility Research Dataset

The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google LLC. The Dataset contains anonymized mobility flows aggregated over users who have turned on the Location History setting, which is off by default. This is similar to the data used to show how busy certain types of places are in Google Maps — helping identify when a local business tends to be the most crowded. The dataset aggregates flows of people from region to region, which is further aggregated at the level of US county, weekly in this study.

To produce this dataset, machine learning is applied to logs data to automatically segment it into semantic trips [3]. To provide strong privacy guarantees, all trips were anonymized and aggregated using a differentially private mechanism [30] to aggregate flows over time[3]. This research is done on the resulting heavily aggregated and differentially private data. No individual user data was ever manually inspected, only heavily aggregated flows of large populations were handled.

All anonymized trips are processed in aggregate to extract their origin and destination location and time. For example, if users traveled from location a to location b within time interval t, the corresponding cell (a, b, t) in the tensor would be $n \pm err$, where err is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero mean Laplace distribution and yields (ϵ, δ)-differential privacy guarantee of $\epsilon = 0.66$ and $\delta = 2.1 \times 10^{-29}$ per metric. Specifically, for each week W and each location pair (A, B), we compute the number of unique users who took a trip from location A to location B during week W. To each of these metrics, we add Laplace noise from a zero-mean distribution of scale $\frac{1}{0.66}$. We then remove all metrics for which the noisy number of users is lower than 100, following the process described in https://research.google/pubs/pub48778/, and publish the rest. This yields that each metric we publish satisfies (ϵ, γ)-differential privacy with values defined above. The parameter ϵ controls the noise intensity in terms of its variance, while γ represents the deviation from pure ϵ-privacy. The closer they are to zero, the stronger the privacy guarantees.

[3] See https://policies.google.com/technologies/anonymization for more.

Effectiveness and Compliance to Social Distancing During COVID-19

Kristi Bushman
Dpt. of Computer Science
University of Pittsburgh
k.bushman@pitt.edu

Konstantinos Pelechrinis
Dpt. of Informatics & Networked Systems
University of Pittsburgh
kpele@pitt.edu

Alexandros Labrinidis
Dpt. of Computer Science
University of Pittsburgh
labrinid@cs.pitt.edu

ABSTRACT

In the absence of pharmaceutical interventions to curb the spread of COVID-19, countries relied on a number of nonpharmaceutical interventions to fight the first wave of the pandemic. The most prevalent one has been stay-at-home orders, whose the goal is to limit the physical contact between people, which consequently will reduce the number of secondary infections generated. In this work, we use a detailed set of mobility data to evaluate the impact that these interventions had on alleviating the spread of the virus in the US as measured through the COVID-19-related deaths. To establish this impact, we use the notion of Granger causality between two time-series. We show that there is a unidirectional Granger causality, from the median percentage of time spent daily at home to the daily number of COVID-19-related deaths with a lag of 2 weeks. We further analyze the mobility patterns at the census block level to identify which parts of the population might encounter difficulties in adhering and complying with social distancing measures. This information is important, since it can consequently drive interventions that aim at helping these parts of the population.

ACM Reference Format:
Kristi Bushman, Konstantinos Pelechrinis, and Alexandros Labrinidis. 2020. Effectiveness and Compliance to Social Distancing During COVID-19. In *The 3rd Workshop on Epidemiology meets Data Mining and Knowledge discovery, August 24, 2020, San Diego, CA, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330719

1 INTRODUCTION

Since the first reported case of COVID-19 in early December 2019 in Wuhan, China, the world has experienced dramatic changes in an effort for societies to deal with the pandemic. Given the absence of pharmaceutical interventions (i.e., a medical treatment or a vaccine), governments and health officials have relied on non-pharmaceutical interventions, including *shelter-at-home* orders, contact tracing and volume testing. The reasoning behind shelter-at-home interventions is to limit the physical contacts between people, which furthermore limits the transmission of the virus. Given the absence of a vaccine, this does not mean that the virus will be eradicated but rather, limiting people's mobility will allow the health systems to operate under capacity and be as effective as possible, consequently limiting the number of fatalities.

Of course, these measures have not been without controversy. Hence, it is important to examine whether they are effective in achieving their goal. Using a granular mobility dataset for the US obtained from SafeGraphs (details provided in the following sections) and COVID-19-related fatalities we show that average fraction of people staying home weekly *Granger-causes* the number of COVID-19-related fatalities with a 3-weeks lag. We also examined for and did not find any evidence of bidirectional Granger causality, i.e., feedback effects of people altering their mobility as a response to the change of the number of fatalities (e.g., as a reaction to the news). We also provide a short-term prediction model for the number of COVID-19 related fatalities in US one-week out, using only information about population-level mobility behavior and fatalities over the past three weeks.

Given the effectiveness of these measures it is critical to understand who in the population complies and to what extent. Differences in compliance levels are not necessarily by choice. For example, many people are essential workers and hence, need to spend time outside of their home. Others might not have to physically be at work, but they have to take care of family members living in other households etc. Identifying these parts of the population can provide critical information on possible policies/interventions that could further increase compliance, without compromising the needs of people. Therefore, in this work we build a framework using a beta regression model to predict the percentage of time spent daily at home within a census block based on demographic characteristics. Using these models we can then examine various hypotheses on whether specific demographics of interest are associated with a change in mobility above and beyond of what was expected from the mobility patterns prior to stay-at-home orders. We focus on two particular demographics, age and race, and show that show that minorities and older people, while significantly increasing their stay at home, this increase is smaller compared to that white and younger people. We further provide some possible mechanisms that lead to this observation and show that income disparities can explain a sizable part of this difference. The main contributions of our work can be summarized as follows:

- Provide a Granger-causality analysis on the impact of stay-at-home orders on COVID-19-related fatalities
- Design a framework for quantifying adherence to social distancing according to various demographics
- Design a dynamic dashboard to visualize both the *raw* mobility data as well as, the results from our analysis.

epiDAMIK '20, August 24, 2020, San Diego, CA, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330719

We believe that our work can provide critical information to local officials and policy makers. The rest of the paper is organized as follows. Section 2 provides a description of the data we used for our analysis, as well as, a brief review on related to our study literature. Section 3 provides our Granger-causality analysis, while Section 4 introduces our framework for identifying the relationship between social distancing compliance and demographics. We conclude our work and discuss its limitations and directions for future work in Section 5.

2 DATA AND RELATED WORK

In this section we describe the dataset we use for our analysis, as well as, relevant to our study literature. The code for the analysis presented in the paper can be found on our github repository: https://github.com/kpelechrinis/epiDAMIK20-COVID.

SafeGraph data: SafeGraph has released a detailed mobility dataset based on the locations of about 18 million mobile phones across the US. This information is obtained through various mobile applications that partner with SafeGraph. This provides diverse population coverage, while the data are provided in an aggregated manner, with steps taken towards satisfying differential privacy requirements. While a detailed description can be found on SafeGraph's COVID-19 data consortium page [25], the main information that we will use is the daily mobility patterns for census block groups (CBG). In particular, for each day and each census block group since 01/01/2020 we obtain - among other - the following daily information:

- `completely_home_device_count`: This is the number of devices within the CBG of interest that did not leave their home.
- `distance_traveled_from_home`: This is the median distance traveled during the day from all the devices whose home is within the CBG of interest
- `median_percentage_time_home`: This is the median percentage of time spent at home during the day from devices whose home is within the CBG of interest
- `destination_cbgs`: This is the CBGs that were visited during the day from devices whose home is within the CBG of interest. Each destination block is also associated with the number of devices in the SafeGraph dataset that performed this transition.

COVID-19 data: In order to evaluate any (Granger causal) impact between mobility and COVID-19-related fatalities we need to utilize data related to the number of confirmed cases and deaths. While an accurate number for the daily number of infections would be the most appropriate variable for this analysis, it is widely known that the reported numbers are a severe undercount of the actual number of infections. On the other hand the number of fatalities is also inaccurate but it is considered a more robust signal for the prevalence of the disease. Albeit it is a lagged signal, with an average of 15-20 days delay [15]. We obtain our data from the NY Times github repository [28].

COVID-19 and mobility: Excluding clinical interventions (potential treatments, vaccine, etc.), limiting mobility and inter-personal contacts has been the most central intervention in an effort to contain the pandemic. As such, several studies have analysed the changes in human mobility across various regions using granular mobility data (e.g., [4, 9, 13, 23] with the list being non-exhaustive). Aleta *et al.* [2] further utilize these mobility information to drive agent-based simulators in order to understand the impact of contact tracing and testing on a possible second wave of the disease. Zhang *et al.* [31] have further analyzed contact surveys from the early epidemic stage in China and built transmission models to quantify the impact of social distancing and school closures. This line of research is of course still developing as restrictions are lifted, new measures potentially coming in the possibility of a second wave etc.

Public health non-pharmaceutical evaluation: Of course, similar non-pharmaceutical interventions have been applied in the past as well and there is a volume of research that evaluates their impact. For example, Ahmed *et al.* [1] provide a review study on social distancing measures in workplace. Their review includes both epidemiological as well as, modeling studies and they concluded that overall workplace social distancing reduced the influenza attack rate approximately 23%. Similarly, Rashid *et al.* [24] reviewed studies that evaluated various measures (school closings, work-from-home etc.) for dealing with the 2009 influenza pandemic. They identified that workplace interventions provide moderate reduction in transmissions (20-30%). Other non-pharmaceutical interventions include the banning of mass events. While intuitively this seems to be a particularly effective strategy, prior literature has shown that this is true only in combination with other interventions [12, 16]. One of the reasons for this is the contact time at such events is relatively small compared to the time spent in schools, workplaces, or other community locations [7]. The literature aforementioned is not exhaustive. However, to the best of our knowledge, there is no study that uses the notion of Granger causality for non-pharmaceutical interventions. Contrary to the majority of existing studies that rely on large-scale simulation models, or, analyzing a small case (e.g., a restaurant, a specific workplace etc.), we take a macroscopic approach, looking at the aggregate adherence to these interventions and the macroscopic results (e.g., total fatalities).

3 EVALUATING SOCIAL DISTANCING

In this section we will begin by introducing the notion of Granger causality between two time series and then we will see how it applies to our case.

3.1 Granger Causality

Granger causality is a statistical test that aims at identifying whether a time-series $x(t)$ provides useful information in predicting time-series $y(t)$ [10]. It is eminent to understand that Granger causality is what Granger himself described, "temporally related" or "predictive causality", rather than the traditional notion of causality. Simply put, $x(t)$ is said to Granger-cause $y(t)$ if it precedes in time and is able to improve the predictions of $y(t)$ beyond auto-regressive models. While this might not be a useful notion for what is needed in areas like clinical treatments, it is particularly useful and has been extensively used in econometrics, public policy etc. (e.g., [3, 5, 6, 11, 18] with the list being non-exhaustive).

Formally, the examination of whether $x(t)$ Granger-causes (G-causes for short) $y(t)$ one needs to build the following two models:

$$M_0: \quad y(t) = a_{00} + \sum_{i=1}^{m} a_{0i} y(t-i) + \epsilon_0 \tag{1}$$

$$M_1: \quad y(t) = a_{10} + \sum_{i=1}^{m} a_{1i} y(t-i) + \sum_{i=1}^{p} b_i x(t-i) + \epsilon_1 \tag{2}$$

The first model (Eq. 1) is essentially a pure auto-regressive model on y up to lag m (called the restricted model), while the second one includes lagged terms from the time-series $x(t)$ to be explored as a potential Granger cause (called the unrestricted model). Given this setting the following null hypothesis is examined: via conducting an F-test:

$$H_0: \quad b_1 = b_2 = \cdots = b_p = 0 \tag{3}$$

The null here is the hypothesis that no explanatory power is jointly added from the lags of x. So eventually, we retain all the lagged values of x that are individually statistically significant (t-statistic), but in order to reject H_0 that x does not G-cause y, all these lags need to add explanatory power (as compared to the restricted model). We would like to note here that the time series need to be stationary before performing the Granger test. Hence, if the original data are not stationary they should be transformed to eliminate the possibility of autocorrelation (e.g., through differentiation).

3.2 Shelter-at-home and COVID-19 fatalities

We are interested in examining whether the mobility of people in the US G-causes the number of fatalities from COVID-19. Here, we would like to emphasize that for the latter, we are using the number of COVID-19 deaths ϕ reported from health authorities as discussed in Section 2. We do not make use of any information related to excess fatalities, or any attempt to estimate the under-reporting factor in fatalities. For the G-cause variable, we first obtain the fraction of devices in each census block group b that stayed exclusively at home daily[1] h_b. We then obtain a weighted average value over all the CBGs, $h_{US}(t)$, for each day t, where the weights are the sample size in each block. We further aggregate the data weekly, since there are known inconsistencies and delays in reporting cases and deaths. Weekly aggregation should remove some of the associated noise with COVID-19 daily reports.

Figure 1 shows the two weekly time-series of interest for the period between 01/21/2020 (when the COVID-19 cases started being recorded) and 07/03/2020. We apply the Kwiatkowski–Phillips–Schmidt–Shin test [14] and we identify that these time-series are not stationary. However, differentiating both time-series will lead to stationarity. Running the Granger causality test for lags up to 6 weeks (given the length of our time-series longer lags cannot be tested), we obtain the results in Table 1. As we can see, there is evidence that mobility G-causes COVID-19-related fatalities at a lag of about 2 weeks. We also examined for bidirectional G-causality, i.e., people listening to the news and number of fatalities, and reacting with changes in their mobility. However, we did not find any supporting evidence.

Given the results from our Granger causality analysis we can build a time-series prediction model for estimating the weekly number of fatalities in the near-future (e.g., one week ahead). We

[1]We also examined the median percentage of time spent at home, with similar results. In fact, the median percentage of time spent at home and the fraction of people staying completely at home daily are highly correlated, with a correlation coefficient $\rho = 0.93$.

	Lag					
	1	**2**	**3**	**4**	**5**	**6**
b_1	263.4*	156.0	236.9*	230.9*	344.4**	289.5*
b_2	-	401.4**	432.7**	539.6**	447.8**	574.9*
b_3	-	-	348.1*	516.2*	675.9**	760.1*
b_4	-	-	-	186.7	−18.1	65.1
b_5	-	-	-	-	145.6	−64.3
b_6	-	-	-	-	-	−10.3
Adj R^2	0.46	0.69	0.77	0.79	0.91	0.9
F-test	5.08	10.3	11.1	18.9	8.5	12.6
(p-val)	(0.03)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)

$^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 1: Individual coefficients' significance and F-test result for various lags.

experiment with two different models, namely, a Vector AutoRegression (VAR) and a Long-Short Term Memory neural network. The VAR model is essentially the unrestricted model in the Granger-causality test (Equation 2), where $m = p = 3$ Table 2 shows the corresponding model. As we can see, increased fraction of people staying home will result in a reduction in the predicted number of fatalities 3 weeks ahead. We also examined a stacked LSTM architecture, with 2 layers with 50 hidden units each, followed by a dense layer with ReLU activation. We use again a sequence of size of 3 and train the model over multiple epochs using early stopping. The results from our two models are presented in Figure 2. In particular, we provide predictions for the last 5 weeks (as of this writing) and we train each model using all the data up to the week of interest. Consequently we make our out-of-sample predictions with each model which are overlaid with the actual fatalities. Overall, both models perform relatively well, especially given the short span of the time-series, as well as, the simplicity of the models in terms of input features. We would like to note here that these models are not appropriate for longer term predictions (e.g., fatality count in 4 months), which is the focus of most of the fatality-related prediction models developed (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html).

4 QUANTIFYING SOCIAL DISTANCING BEHAVIOR PER DEMOGRAPHICS

In the previous section, we saw that there is strong evidence that limiting mobility Granger-causes fewer fatalities from COVID-19. Therefore, it is important to understand if and which parts of the population are not able to *adhere* to the guidelines. This information is critical to be communicated to health officials and policy makers, since it can drive interventions that will help everyone follow the recommendations to the extent possible. In this section, we present a framework based on a beta regression model from the daily percentage of time spent home and the difference-in-differences method that can identify the relationship between demographics of interest and the way they relate to social distancing behavior.

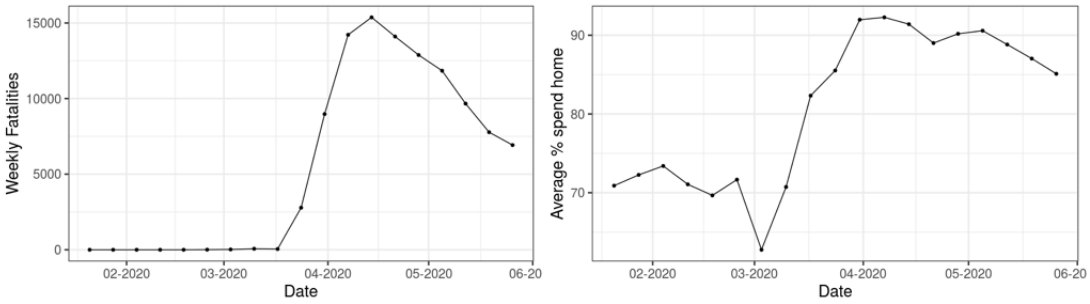


Figure 1: Weekly time series for COVID-19-related fatalities (left) and percentage of people staying at home (right).

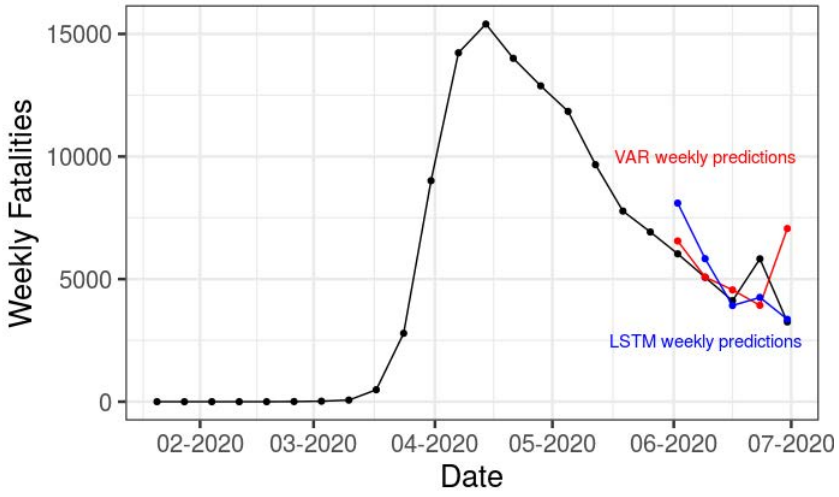


Figure 2: Both models perform reasonably out-of-sample, relative to the amount of data available for learning and the simplicity of the input features.

Variable	Coefficient	p-val
$h_{US}(t-1)$	137.9	0.21
$h_{US}(t-2)$	252.4	0.15
$\mathbf{h_{US}(t-3)}$	-384.7	< 0.01
y(t-1)	1.35	< 0.01
y(t-2)	-0.59	0.16
y(t-3)	0.19	0.40
R^2	0.86	
SE_{res}	1250	

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: VAR model for predicting weekly fatalities one-week-out.

4.1 Beta regression model

Our goal is to model the percentage of time $h_{\mathcal{P}}$ that a specific population $\mathcal{P}$ spends home daily. Given that our dependent variable $h_{\mathcal{P}}$ is real-valued, bounded in the unit interval a linear regression is not an appropriate model. Hence, we choose to use a beta regression model [8], where essentially the data are assumed to follow a beta distribution. A useful parametrization of the beta distribution for this type of models is given by:

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\cdot\Gamma((1-\mu)(\phi))} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad (4)$$

where μ is the mean of the beta distribution and ϕ is a parameter called precision. ϕ *controls* the variance of the distribution; for a fixed μ, higher precision leads to smaller variance. With this setting the beta regression model for $h_{\mathcal{P}}$ is:

$$g(\overline{h}_{\mathcal{P}}) = \mathbf{x}_{\mathcal{P}}^T \cdot \mathbf{b} + \epsilon \quad (5)$$

where $\overline{h}_{\mathcal{P}}$ is the average daily fraction of time spent home for $\mathcal{P}$, $\mathbf{x}_{\mathcal{P}}$ is the vector of the model's covariates, $\mathbf{b}$ is the vector of the regression's coefficients and $g(\cdot)$ is a link function (strictly increasing and twice differentiable). This model is very similar to a generalized linear model (e.g., logistic, Poisson or negative

binomial regression) and it is solved through a Maximum Likelihood Estimation (MLE). The MLE identifies the coefficients $\mathbf{b}$, but also the precision parameter ϕ, which is a constant and not a function of $\mathbf{x}_{\mathcal{P}}$[2].

4.2 Demographics Analysis

In this section we will begin by modeling the fraction of time spent at home daily in each census block as a function of specific demographics of the population. We start with race, where census data provide information on the percentage of people within each census block that belong to the following categories: White, Black, Hispanic, Asian, American Indian or Native Alaskan, and Other races[3]. Since we want to estimate the relationship between these demographics and the changes observed after the social distancing recommendations, we build two separate models; one that captures the mobility prior to stay-at-home orders (M_{pre}) and one that captures mobility after these orders were put in place (M_{post}). One of the problems is that different parts of the country put these measures in place in different times through the course of the pandemic. Given that the majority of the orders were put in place sometime within March 2020, we build M_{pre} using data from February 2020, and M_{post} using data from April 2020. Table 3 presents the results of these regressions. Using these results we can start examining the average percentage of time spent daily at home by the population of a hypothetic census block group (HCBG) with a specific racial demographic composition. For example, Figure 3 presents the beta distribution for racially homogeneous (hypothetical) census block groups. As we can see, there are differences across these hypothetical census block groups, both for the same time period, as well as, their shift as the stay-at-home orders were put in place. More specifically, Table 4 presents the average stay home percentage for each of the hypothetical blocks.

Variable	M_{pre}	M_{post}
White%	-0.45^{***}	-0.48^{***}
Black%	-0.27^{***}	-0.56^{***}
Hispanic%	0.29^{***}	0.39^{***}
Asian%	-0.40^{***}	1.87^{***}
Natives+Others%	-0.51^{***}	-0.93^{***}
constant	1.39^{***}	2.5^{***}
ϕ	14.5	5.8
N	201,917	201,917

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Beta regression model for the average daily percentage of time of stay home at a census block group before (02/20) and after (04/20) stay-at-home orders.

Table 4, while providing us with a quick view of how specific parts of the population might comply with the social distancing recommendations (in terms of staying home), it does not provide the whole picture. In particular, we can see that different demographics are *associated* with different levels of mobility outside of their home even before the stay-at-home order. So any change observed after the orders were put in place, they need to be compared with the original difference. This process is visualized in Figure 5, where we see two populations $\mathcal{P}_1$ and $\mathcal{P}_2$, with their pre-lockdown daily percentage of staying home, as well as, their post-lockdown daily percentage of staying home. While Δ_2 provides us with information about what is happening in the two populations after the stay-at-home orders were put in place, it does not adjust for the *behavior* of the two populations prior to the intervention, and the difference $\delta(\mathcal{P}_1, \mathcal{P}_2) = \Delta_2 - \Delta_1$ is more informative. Hence, in order to identify demographic discrepancies between two populations, $\mathcal{P}_1$ and $\mathcal{P}_2$, in complying with stay-at-home orders, we performed the following hypothesis test:

Hypothetical Block	Pre	Post
White	71.8%	89.6%
Black	75.6%	88.6%
Hispanic	84.4%	94.9%
Asian	72.9%	98.9%
Natives+Others	70.7%	84.8%

Table 4: Percentage of time spent home daily for hypothetical racially homogeneous census block groups based on the beta regression models from Table 3.

$$H_0: \quad \delta(\mathcal{P}_1, \mathcal{P}_2) = 0 \tag{6}$$

$$H_1: \quad \delta(\mathcal{P}_1, \mathcal{P}_2) \neq 0 \tag{7}$$

In order to perform this test, we use the full beta distribution for each population-time combination and repeatedly sample them to build the distribution of $\delta(\mathcal{P}_1, \mathcal{P}_2)$. Then we can perform the above hypothesis test. Table 5 presents the results for the various comparisons between the minority HCBG and the white one. As we can see all minority HCBG - except the Asian one - exhibit a smaller increase as what was expected based on their pre-intervention patterns. Particularly interesting is the case of the Hispanic HCBG, which even though exhibits the second highest daily percentage of staying home after the stay-at-home orders, the observed increase is smaller as compared to the white HCBG. Furthermore, it is interesting that the Asian HCBG exhibits a 7.5% higher *compliance* as compared to the white HCBG. While the reasons for this are not clear - and we cannot identify them through the data we have - there are a few reasons that are plausible, including the increase of racist attacks targeting Asians in the US at the wake of the pandemic [19–22, 29, 30].

While for the Asian population, staying at home more might also be a way of avoiding racist attacks, the question remains, why are there discrepancies for the rest of the minorities as compared to the white HCBG ? One plausible explanation is that a large fraction of these minorities are essential workers and while overall they increase their stay at home, they really need to go to their work. Another possible reason is that minorities live in inner cities and as such they are close to their families. Furthermore, these minorities

[2]There are extensions of this model [26] that models the precision as a function of a set of regressions $\mathbf{z}$, i.e., $g'(\phi) = \mathbf{z}^T \cdot \mathbf{c} + \epsilon$.

[3]For the purposes of our analysis we merge the American Indian and Native Alaskan category with the Other races category.

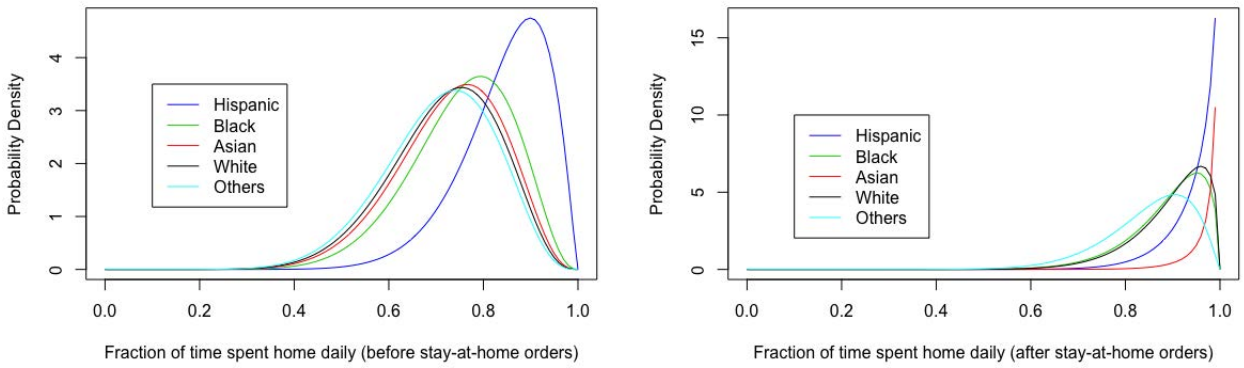


Figure 3: Beta distribution for hypothetical - racially homogeneous - census block group before (left) and after (right) stay-at-home orders across the US.

$\mathcal{P}_1$	$\mathcal{P}_2$	$\delta(\mathcal{P}_1, \mathcal{P}_2)$
Black	White	$-4.8\%^{***}$
Hispanic	White	$-6.2\%^{***}$
Asian	White	$7.5\%^{***}$
Natives+Others	White	$-3.6\%^{***}$

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 5: Minority HCBGs exhibit lower percentage of stay-at-home, as compared to white HCBGs.

Variable	M_{pre}	M_{post}
White%	-0.43^{***}	-0.61^{***}
Black%	-0.29^{***}	-0.3^{***}
Hispanic%	0.27^{***}	0.7^{***}
Asian%	-0.29^{***}	0.87^{***}
Natives+Others%	-0.52^{***}	-0.79^{***}
Median Income	$-9.9 \cdot 10^{-7}$ ***	$9.9 \cdot 10^{-6}$ ***
constant	1.43^{***}	2.13^{***}
ϕ	14.6	6.34
N	201,917	201,917

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Beta regression model for the average daily percentage of time of stay home at a census block group before (02/20) and after (04/20) stay-at-home orders adjusting for median income (expressed in thousands of $s) in the CBGs.

$\mathcal{P}_1$	$\mathcal{P}_2$	$\delta(\mathcal{P}_1, \mathcal{P}_2)$
Black	White	$6 \cdot 10^{-4}\%$
Hispanic	White	$-4.3\%^{***}$
Asian	White	$5.7\%^{***}$
Natives+Others	White	$-5 \cdot 10^{-3}\%$

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7: When adjusting for income a large percentage of the mobility differences between HCBGs during stay-at-home orders is explained.

have come to rely and support their extended families [27] and hence, they might be providing them with help (e.g., childcare support for essential workers etc.) during this time, leading to higher mobility outside the home. Other plausible reasons include the relationship between these groups and technology. In particular, ethnic minorities have traditionally been slower in adopting new technology for a variety of reasons [17] and this could mean in a situation like the current pandemic, their inability or unwillingness to use online platforms for essential errands such as grocery shopping. While we cannot show with our current data whether any of these plausible reasons are in play, we can examine one additional factor that is relevant to all of the above possibilities; their median income. Low income families typically live in inner-city and are of ethnic minorities, they have issues with accessing and adopting technology, while many of the essential workers are low-paid employees (e.g., grocery store workers, delivery, etc.). Tables 6 and 7 present the same results when we adjust for the median income of an HCBG. As we can see, the mobility differences between black and white HCBGs, as well as native and other minorities and white HCBGs, disappears, while for Hispanic and Asian HCBGs the differences are reduced.

We also examined another demographic attribute, namely, age. While census provides a breakdown of the age of a census block group in several age brackets, we aggregated them into two bins; younger or older than 50 year old[4]. Again, we build a beta regression model with the same dependent variable as before but the

[4]Obviously one can repeat the analysis with more bins, but we wanted to keep things simpler mainly for presentation purposes.

independent variable is the percentage of the population in the CBG that is older than 50 years old. The results are presented in Table 8, where as we can see the older population is associated with a reduced stay-at-home daily time as compared to younger population (less than 50). Figure 4 further visualizes the beta distributions for hypothetical CBGs with only population older or younger than 50 years old. Furthermore, by performing a similar hypothesis test as in Eq. (6)-(7), we find that the HCBG with population older than 50 years old stays at home 2.6% (p-val < 0.01) **less** time at home on average as compare to younger population and based on their pre-intervention patterns. In contrast to the race case, when adjusting for the median income, the difference remains (-2.5%, p-val < 0.01). A potential reason for this difference between population in the opposite side of the 50 years old mark, can be their *technology fluency*. Younger people that are avid users of (mobile) technology can take advantage of various services that can help people complete their errands (e.g., grocery shopping), while staying at home. This might not be the case for older people (at least to the same extent). Again, while this is a plausible mechanism that can drive the observed difference, the data in our disposal does not allow us to further examine this.

Variable	M_{pre}	M_{post}
Older50%	0.09^{***}	-0.28^{***}
constant	0.98^{***}	2.39^{***}
ϕ	14.2	5.5
N	201,917	201,917

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 8: Beta regression model for the average daily percentage of time of stay home at a census block group before (02/20) and after (04/20) stay-at-home orders based on the percentage of the population that are older than 50 years old.

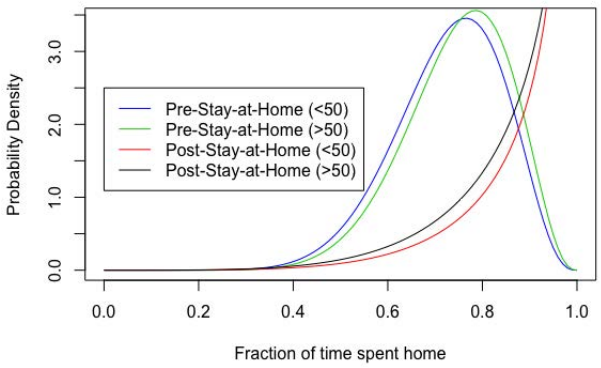


Figure 4: Beta distributions for the daily fraction of time spent at home for population older and younger than 50.

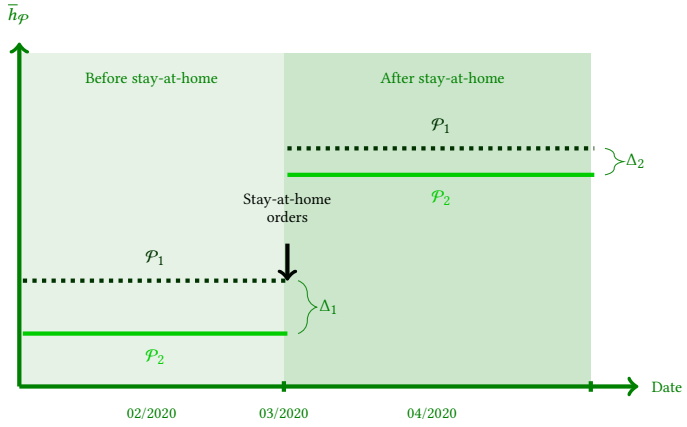


Figure 5: When comparing the mobility post-lockdown for different populations, we need to consider the pre-lockdown mobility as well.

4.3 Dashboard

We have also created a dashboard to visualize this mobility information in an interactive manner[5]. Figure 6 presents a screenshot from the dashboard that depicts the census block tracts of Allegheny County on the left half. The user can choose a tract (the selected tract will be colored red as in the figure) and information about the outgoing mobility (i.e., movements of people whose home CBG is the selected one) and incoming mobility (i.e., movements from people whose home CBG is not the selected one but they visited it) associated with it is presented. The choice between outgoing and incoming mobility can be made through the control buttons above the map. For example, in Figure 6 outgoing mobility information for people whose home CBG is the selected origin CBG (420035231001) is presented on the map. The color for each census block group tract i represents the fraction of the total foot traffic from the residents of the origin CBG, over the period selected from the user[6], that visited CBG i. On the right half of the figure, there are two time-series depicted that provide temporal information for the CBG that the user is currently hovering over (say CBG_h). In the specific situation depicted here, this is CBG 420035231001. The top time series provides the daily number of visits in CBG_h from the origin CBG, while the bottom time series represents the fraction of time residents of CBG_h spent at home. It is interesting here to note that if we hover over the origin CGB, i.e., CBG_h is the selected CBG, then the top time-series represents *self-loops*. That is, traffic from residents of the CBG that was destined to other venues/points of interest within the CBG. Finally, we also present a table with some basic demographic information about the origin CBG related to our analysis, such as racial and age composition of the population, median income and total population.

We would also like to note here that this dashboard is still *work-in-progress* in the sense that new features are being added prior

[5]http://mobility.pittsmartliving.com/.
[6]The user can select the time period through the slider under the map.

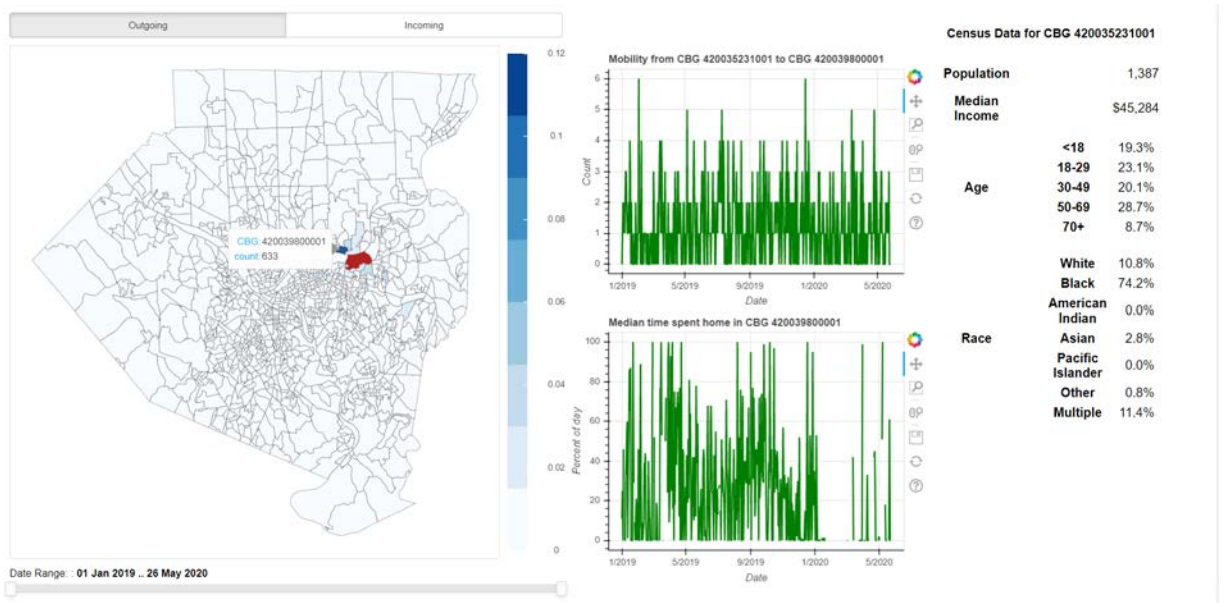


Figure 6: Our dashboard for Allegheny County showing the outgoing mobility from the selected CBG (red) alongside with demographic information. The two time-series plots further provide information related to the interaction between the selected CBG and another CBG that the user hovers over.

to going publicly live. For example, our immediate future plan is to visualize information about specific businesses and their geographical reach (i.e., where do customers of different establishments come from?). This information can be very helpful for local health authorities when identifying a plan for interventions and the corresponding protocols.

5 CONCLUSIONS AND DISCUSSION

In this study we perform a macroscopic analysis of the effectiveness of social distancing measures in the US during the COVID-19 pandemic using the notion of Granger causality. Our analysis indicate that the average daily fraction of population staying completely at home Granger-causes the number of COVID-19 fatalities in a 3-week period. We further examine the presence of bidirectional Granger causality and we do not find any supporting evidence. Using this observation, we also build two simple prediction models for weekly COVID-19-related fatalities, using auto-regressive and mobility features. We further provide a framework to identify the relationship between demographics and social distancing behavior. While this analysis does not provide causal relationships, it can certainly provide important information for policy makers while thinking of ways to increase *compliance*. Finally, we provide a visualization dashboard with the raw data as well as, the results from our analysis. This dashboard is constantly being updated with new results and data.

We would like to emphasize here that even though we have included a prediction model in our analysis, this is only to showcase in practise the conclusions from the Granger causality analysis[7].

[7]Furthermore, there are several well-performing prediction models in the public sphere - tracked by CDC as well - and our goal is certainly not to add yet another model.

Furthermore, while the model performs well out-of-sample, several improvements can be achieved by including even more informative features. For instance, just an aggregate number of how many hours a person spends out of their home does not capture factors important for the prediction of infections. Was this movement to a high-risk location (e.g., a grocery store) or was it for a stroll around the neighborhood? Disentangling this is certainly not trivial and we are working in methods for identifying the number of potential contacts a person from a specific CBG is expected to have based on their mobility and the POI foot traffic data. Furthermore, it will be particularly useful to extend our analysis to a more (spatially) fine granularity, focusing on a microscopic analysis (e.g., at the county, or city, level). This will allow us to identify the exact time points of interventions and possibly attempt to extract causal relationships using quasi-experimental methods, such as instrumental variables and difference-in-differences.

ACKNOWLEDGMENTS

This work partly supported by NSF awards CNS-1739413 and CNS-2034625.

REFERENCES

[1] Faruque Ahmed, Nicole Zviedrite, and Amra Uzicanin. 2018. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. *BMC public health* 18, 1 (2018), 518.

[2] Alberto Aleta, David Martin-Corral, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E Dean, M Elizabeth Halloran, Ira M Longini, Stefano Merler, et al. 2020. Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the COVID-19 epidemic. *medRxiv* (2020).

[3] Mak B Arvin, Rudra P Pradhan, and Neville R Norman. 2015. Transportation intensity, urbanization, economic growth, and CO2 emissions in the G-20 countries. *Utilities Policy* 35 (2015), 50–66.

[4] J. Bayham, J. Adams, D. Ghosh, and P. Jackson. 2020. Colorado Mobility Patterns During the COVID-19 Response. *Technical Report* (2020).

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[6] Willem H Buiter. 1984. Granger-causality and policy effectiveness. *Economica* 51, 202 (1984), 151–162.

[7] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. 2020. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. (2020).

[8] Silvia Ferrari and Francisco Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of applied statistics* 31, 7 (2004), 799–815.

[9] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *Available at SSRN 3570145* (2020).

[10] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[11] MV Hood, Quentin Kidd, and Irwin L Morris. 2008. Two sides of the same coin? Employing Granger causality tests in a time series cross-section framework. *Political Analysis* 16, 3 (2008), 324–344.

[12] David A Ishola and Nick Phin. 2011. Could influenza transmission be reduced by restricting mass gatherings? Towards an evidence-based policy framework. *Journal of epidemiology and global health* 1, 1 (2011), 33–60.

[13] Brennan Klein, T LaRocky, S McCabey, L Torresy, Filippo Privitera, Brennan Lake, Moritz UG Kraemer, John S Brownstein, David Lazer, Tina Eliassi-Rad, et al. 2020. Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak.

[14] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, Yongcheol Shin, et al. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of econometrics* 54, 1-3 (1992), 159–178.

[15] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. 2020. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine* 172, 9 (2020), 577–582.

[16] Howard Markel, Harvey B Lipman, J Alexander Navarro, Alexandra Sloan, Joseph R Michalsen, Alexandra Minna Stern, and Martin S Cetron. 2007. Nonpharmaceutical interventions implemented by US cities during the 1918-1919 influenza pandemic. *Jama* 298, 6 (2007), 644–654.

[17] Karen Mossberger, Caroline J Tolbert, Daniel Bowen, and Benedict Jimenez. 2012. Unraveling different barriers to Internet use: Urban residents and neighborhood effects. *Urban Affairs Review* 48, 6 (2012), 771–810.

[18] Paresh Kumar Narayan and Stephan Popp. 2012. The energy consumption-real GDP nexus revisited: Empirical evidence from 93 countries. *Economic Modelling* 29, 2 (2012), 303–308.

[19] NBC News. 2020. Smashed windows and racist graffiti: Vandals target Asian Americans amid coronavirus. https://www.nbcnews.com/news/asian-america/smashed-windows-racist-graffiti-vandals-target-asian-americans-amid-coronavirus-n1180556

[20] NPR. 2020. New Site Collects Reports Of Racism Against Asian Americans Amid Coronavirus Pandemic. https://www.npr.org/sections/coronavirus-live-updates/2020/03/27/822187627/new-site-collects-reports-of-anti-asian-american-sentiment-amid-coronavirus-pand

[21] PBS. 2020. Asian Americans describe ǵut punchóf racist attacks during coronavirus pandemic. https://www.pbs.org/newshour/nation/asian-americans-describe-gut-punch-of-racist-attacks-during-coronavirus-pandemic

[22] The Washington Post. 2020. As the coronavirus spreads, so does online racism targeting Asians, new research shows. https://www.washingtonpost.com/technology/2020/04/08/coronavirus-spreads-so-does-online-racism-targeting-asians-new-research-shows/

[23] Giulia Pullano, Eugenio Valdano, Nicola Scarpa, Stefania Rubrichi, and Vittoria Colizza. 2020. Population mobility reductions during COVID-19 epidemic in France under lockdown. *Epicx-lab Technical Report* (2020).

[24] Harunor Rashid, Iman Ridda, Catherine King, Matthew Begun, Hatice Tekin, James G Wood, and Robert Booy. 2015. Evidence compendium and advice on social distancing and other related measures for response to an influenza pandemic. *Paediatric respiratory reviews* 16, 2 (2015), 119–126.

[25] SafeGraph. 2020. SafeGraph COVID-19 Data Consortium. https://www.safegraph.com/covid-19-data-consortium

[26] Alexandre B Simas, Wagner Barreto-Souza, and Andréa V Rocha. 2010. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis* 54, 2 (2010), 348–366.

[27] Robert Joseph Taylor, Linda M Chatters, Amanda Toler Woodward, and Edna Brown. 2013. Racial and ethnic differences in extended family, friendship, fictive kin, and congregational informal support networks. *Family relations* 62, 4 (2013), 609–624.

[28] The New York Times. 2020. COVID-19 Data. https://github.com/nytimes/covid-19-data

[29] The New York Times. 2020. How Asian-American Leaders Are Grappling With Xenophobia Amid Coronavirus. https://www.nytimes.com/2020/03/29/us/politics/coronavirus-asian-americans.html

[30] USA Today. 2020. Ẃe just want to be safe: Hate crimes, harassment of Asian Americans rise amid coronavirus pandemic. https://www.usatoday.com/story/news/politics/2020/05/20/coronavirus-hate-crimes-against-asian-americans-continue-rise/5212123002/

[31] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cécile Viboud, Alessandro Vespignani, et al. 2020. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* (2020).

Neural Networks for Pulmonary Disease Diagnosis using Auditory and Demographic Information

Morteza Hosseini
University of Maryland
Baltimore County

Haoran Ren
University of Maryland
Baltimore County

Hasib-Al Rashid
University of Maryland
Baltimore County

Arnab Neelim Mazumder
University of Maryland
Baltimore County

Bharat Prakash
University of Maryland
Baltimore County

Tinoosh Mohsenin
University of Maryland
Baltimore County

ABSTRACT

Pulmonary diseases impact millions of lives globally and annually. The recent outbreak of the pandemic of the COVID-19, a novel pulmonary infection, has more than ever brought the attention of the research community to the machine-aided diagnosis of respiratory problems. This paper is thus an effort to exploit machine learning for classification of respiratory problems and proposes a framework that employs as much correlated information (auditory and demographic information in this work) as a dataset provides to increase the sensitivity and specificity of a diagnosing system. First, we use deep convolutional neural networks (DCNNs) to process and classify a publicly released pulmonary auditory dataset, and then we take advantage of the existing demographic information within the dataset and show that the accuracy of the pulmonary classification increases by 5% when trained on the auditory information in conjunction with the demographic information. Since the demographic data can be extracted using computer vision, we suggest using another parallel DCNN to estimate the demographic information of the subject under test visioned by the processing computer. Lastly, as a proposition to bring the healthcare system to users' fingertips, we measure deployment characteristics of the auditory DCNN model onto processing components of an NVIDIA TX2 development board.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Neural networks*; • **Computer systems organization** → *Sensor networks*.

KEYWORDS

respiratory sounds dataset, demographic feature extraction, deep convolutional neural networks, embedded devices, early-stage diagnosis, public health

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

ACM Reference Format:
Morteza Hosseini, Haoran Ren, Hasib-Al Rashid, Arnab Neelim Mazumder, Bharat Prakash, and Tinoosh Mohsenin. 2020. Neural Networks for Pulmonary Disease Diagnosis using Auditory and Demographic Information. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 5 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

In 2016, pulmonary diseases were among the top 10 causes of death: ranked 1 for low-income and ranked 5 for high-income countries [1]. Recently, with the outbreak of the COVID-19 as a novel pulmonary infection, a tremendous amount of attention has been directed to control the pandemic crisis about which extreme measures are taken by countries to diagnose the infected patients. Measures such as extensive testing and early-stage diagnosis help to locate and contain the infection, and are reportedly the most effective preventive actions to control the contagion in a pandemic.

Pulmonary problems encompass a wide range of chronic and infectious diseases, and because of the common organ, lung, that they affect, they develop respiratory symptoms whose auditory signals recorded from various medical devices are among the first to be scrutinized by a medical expert. As an example, COVID-19 develops symptoms such as dry cough, fever, fatigue, dyspnea, and shortness of breath that vary in severity at different stages of the disease progression, and correlate with certain ethnicity, gender, and age groups differently [11]. More than 70% of the confirmed COVID-19 patients have reported fever in tandem with a dry cough [24]. Meanwhile, clinical case records indicate that the young population is less likely to develop COVID-19 relevant symptoms, contrary to the elderly that is the most vulnerable group [10].

Traditionally, when someone feels symptoms, they either call a doctor or have themselves seen/scrutinized by medical experts at walk-in clinics, where extensive use of vital signs, visual and auditory information are applied to make diagnostic decisions. Such practice during a pandemic or in remote locations is unsuitable/impractical as a result of the limited capacity of existing facilities and human resources at health centers, and, ironically, can expedite spreading the infection. On the other hand, calls are made by governments/organizations during the pandemic for people to stay at home that, by itself, has caused a state of confusion and has made another barrier. Thus, early-stage and clinic independent machine assistance is critical for the initial diagnosis of the disease and/or for evaluating/assessing its severity.

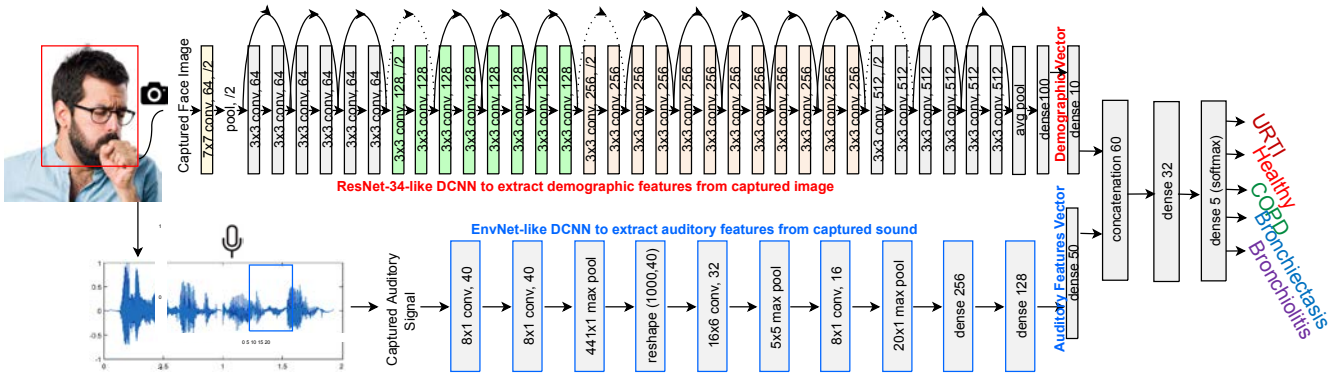


Figure 1: The proposed framework to classify respiratory problem has two DCNN components that process data from a user under test. Part of the information is auditory, such as the audio sound recorded from a medical electronic device like a microphone or a stethoscope, and part of that information is the demographic information, such as age, gender, and ethnicity, that can either be estimated using a computer vision algorithm or inserted manually. The framework is flexible and scalable in the sense that it can incorporate new sensors easily, allowing the system to be tailored to a variety of kinds of situations, such as in-home consultations, clinical visits, or even symptom detection in public milieus using non-contact sensors.

Our goal in this research is to allow machine learning algorithms running on general computing processors (e.g., those in cell-phones and tablets) to assess patients similar to what doctors do at triage and telemedicine, using passively recorded audio and/or video and self-declared information, to bring proactive healthcare to users' fingertips and to estimate the urgency/necessity of whether they need to attend clinics and have themselves further examined with the use of more specialized test-kits or facilities. Our vision is to provide a detection framework that can provide early detection for anyone and anywhere. We develop our work on a publicly released respiratory sound database that includes both auditory and demographic information recorded from 126 subjects covering 7 pulmonary diseases including healthy condition, and with two sets of annotations. More specifically, since the respiratory sound dataset includes the two types of information per patient, we examine how the lack/existence of the demographic information impacts the total accuracy of the model. For further compilation, we develop another deep neural network that estimates demographic information including the age and the gender of the captured images and to correlate them with the auditory signals recorded from the subject under test in order to assess a higher sensitivity and specificity rate of diagnosis. The main contributions of this work include:

- Statistically analyze the information in a public respiratory sound database, to justify extracting a reasonably balanced dataset out of it.
- Train a DCNN on the extracted auditory dataset without considering the demographic information.
- Train another DCNN model on a face images dataset annotated with age, gender, and ethnicity so to estimate/extract demographic information of a subject visioned by a computer.
- Train the auditory dataset in conjunction with the demographic information.
- Deploy the first DCNN model to TX2 embedded system and measure its implementation characteristics for CPU and GPU.

2 RELATED WORK

With the advancement of machine learning and deep learning algorithms, audio-based biomedical diagnosis and anomaly detection have recently become an active area of research. Some important aspects of audio-based diagnosis using deep learning include detection of sleep apnea, recognition of cough tone, and classification of heart sound, to name a few. Early research [8] shows that machine learning (ML) tools on a limited unpublished dataset can distinguish solely between coughs from COVID-19 patients and those who are healthy or with upper-respiratory coughs with high accuracy of 96.8%. [7] introduces End to End convolutional neural networks for cough and dyspnea detection. Authors in [3] used both DCNNs and recurrent neural networks (RNNs) to classify cough sound that they collected using chest-mounted sensors. Authors in [14] used deep learning to detect sleep apnea. Classification of heart sound into normal and abnormal classes was conducted in [20] using DCNNs. Authors in [4] and [16] used DCNNs and RNNs to classify lung sounds respectively. Most of these works report high levels of accuracy on unpublished datasets that are accessible by the research community. The 2017 International Conference on Biomedical Health Informatics (ICBHI) [18] issued a benchmark dataset of respiratory sound to facilitate researching on respiratory sound classification. Since then, researchers proposed various algorithms [5, 12, 15, 17] using different deep learning techniques to classify respiratory cycle anomalies such as the precise locations of wheezes and crackles within the cycle of each respiratory sound recording. In [13] authors showed innovativeness by proposing a digital stethoscope to provide an immediate diagnosis of respiratory diseases. They developed a modified bi-ResNet architecture using STFT and wavelet feature extraction. Log quantized deep CNN-RNN based model for respiratory sound classification was proposed in [2] for memory limited wearable devices.

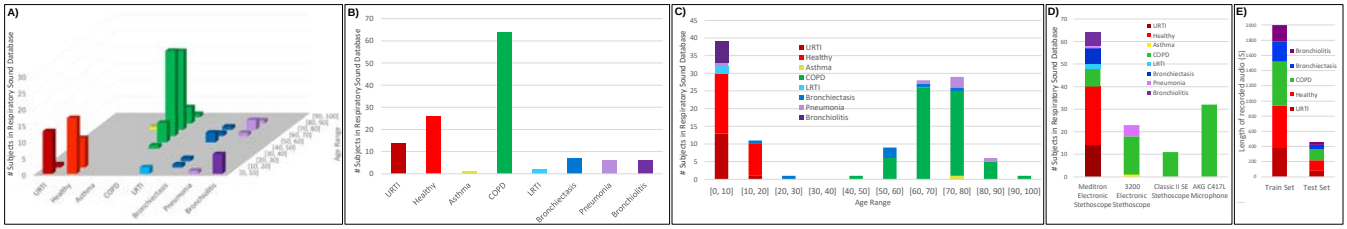


Figure 2: Statistics of the respiratory sound database that contains auditory samples from 126 patients, A) a 2D histogram of 7 pulmonary classes with respect to 10 age groups, B) Break-down of the pulmonary classes in the dataset, C) Break-down of each pulmonary class with respect to age groups, D) Break-down of each pulmonary class with respect to the four recording medical devices, E) Our selection of 52 and 11 individual subjects for train and test datasets respectively that cover 5 pulmonary classes recorded with Meditron Electronic Stethoscope.

3 PROPOSED METHOD

The framework, depicted in Fig. 1, leverages audio/video to extract necessary and medically relevant information and combines the extracted features with other inserted/self-declared patient data. The audio processing incorporates an ML approach such as a DCNN that extracts symptomatic features like crackles and wheezes of lung sounds from a given window of recorded sound of a subject under test. At the video processing path, the captured RGB images are given to a ResNet-34 DCNN to process and estimate the user's other demographic and symptomatic features such as age and gender. The extracted audio/video features along with the other relevant inserted data are concatenated towards the final layers and with the addition of a few more neural network layers or an ensemble of classifiers in the last layer, a probability vector of diagnosis for the user under test is reported. Both audio/video data can extend the scope of the clinical-reported symptoms to more diverse features that may be invisible to a human's perception. For example, when listened by a trained machine, the features extracted from the sound of a patient's cough can include more useful features beyond terms like "dry" or "productive" that are commonly reported in clinical case records.

3.1 Datasets

3.1.1 Respiratory Sound Database. For the auditory dataset, we used a public respiratory sound database [19], which includes 920 recordings acquired from 126 participants annotated with 8 types of respiratory conditions including URTI, Healthy, Asthma, COPD, LRTI, Bronchiectasis, Pneumonia, and Bronchiolitis. The recordings were collected using four types of medical equipment including AKG C417L Microphone, 3M Littmann Classic II SE Stethoscope, 3M Littmann 3200 Electronic Stethoscope, and Welch Allyn Meditron Master Elite Electronic Stethoscope. The duration of each recording range from 10 to 90 seconds mostly dominated with 20s samples. Fig. 2-A, B, and C plot the distribution of the subjects with respect to their diagnosed disease and the age groups they impact, and Fig. 2-D shows the contribution of each of the 4 medical devices for recordings from participants. Among the four recording devices, the Meditron Electronic Stethoscope is the only device that encompasses the 8 pulmonary conditions except for the Asthma, and is used for 63 out of the 126 participants. The recordings from the other 3 devices are majorly taken from COPD-diagnosed participants. By eliminating Asthma, Pneumonia, and LRTI that have little or no samples within the Meditron recordings, we extracted a random subset encompassing all 63 participants and split it into a semi-balanced train and a test set of 52 and 11 participants that include 5 types of pulmonary classes. Fig. 2-E shows a plot of the selected train/test dataset based on the total duration of each class. The database is meanwhile provided with demographic information of the 126 participants and another annotation that marks begin/end of respiratory cycles and the precise locations of events of crackles and wheezes per recording. Based on the second annotation, we counted the total number of respiratory cycles to estimate the slowest and average respiratory cycles within the dataset and to decide on a window size to cut the recordings into smaller frames. Table 1 summarized the number of subjects, duration of recordings, and the number of respiratory cycles per pulmonary class within both train and test datasets.

Table 1: Selected semi-balanced Dataset out of the respiratory sound database. Meditron recordings from 61 patients that include 5 pulmonary classes were selected and the 20s sounds were chopped into overlapping frames of 5s. The total dataset of frames includes 1968 samples that were split into mutually-subject-exclusive between train (81%) and test (19%) subsets.

Selected Dataset		URTI	Healthy	COPD	Bronchiec.	Bronchiol.	Total
Train Set	# Subject	12	24	6	5	5	52
	Duration (S)	380	560	580	260	220	2000
	# Respiratory Cycles	207	257	406	88	141	1099
	# Augmented Frames	304	448	464	208	176	1600
Test Set	# Subject	2	4	2	2	1	9
	Duration (S)	80	140	140	60	40	460
	# Respiratory Cycles	26	48	119	17	16	226
	# Augmented Frames	64	112	112	48	32	368

3.1.2 Face Images Database. UTKFace dataset [23] is a large-scale dataset consisting of over 20,000 face images with annotations of age (ranging from 0 to 116 years old) gender, and ethnicity. In [9], the UTKFace dataset is trained on a ResNet-34 [6], and we reproduced the results of training over ResNet-34, and report the accuracy it gives for precise age as well as for the age group of a random split of 20% test data.

Table 2: The classification accuracy and the model complexity of a ResNet-34 DCNN that extracts demographic information on the UTKface test dataset.

DCNN characteristics			Test accuracy			
Model	#params	FLOPs	Age±0	Age±5	Age±10	Gender
ResNet-34	21M	3.6B	19.6%	65.5%	87.1%	90.3%

3.2 Data Pre-processing and Augmentation

For data augmentation of the respiratory sound database, every recorded audio sample is cut into frames with a duration of 5s and with a stride of 1s, which means every two adjacent frames overlap a duration of 4s, and every 20s recorded sample results in 16 5s frames. Therefore the total 2000 seconds of the training dataset generates 1600 frames, and the total 460s testing data generates 368 frames of 5s samples. The choice of the 5s frames is inferred empirically by experiencing frames ranging from 1s to 10s.

For the data augmentation of the UTKface images, we use common image augmentation techniques such as flipping, shifting and resizing the images within the dataset.

4 EXPERIMENTAL SETUP

We used a ResNet-34 DCNN [6] for the UTKface RGB images of size 200×200, and an EnvNet-like [22] DCNN for the respiratory sound frames of size 1×220500. For the EnvNet-like DCNN, the input from the audio recordings is a one-dimensional vector where the size depends on the window selected for the framework. To best utilize the one-dimensional input, we use two one dimensional convolution layers to extract relevant features with a follow up of non-overlapping max-pooling operation to downsample the feature map. The subsequent layers include two-dimensional convolutional layers with max-pooling layers in between for efficient classification of the diseases. Finally, the fully connected layers summarize the required feature information and feed it to the extended model to generate a generalized output that classifies 5 types of pulmonary conditions as in our extracted dataset.

4.1 Demographic Classification

4.1.1 ResNet-34 for UTKFace.

The classification accuracy of age and gender estimation of ResNet-34 is reported in Table 2. Although the DCNN model does not precisely classify the age within the test dataset, it is able to classify the gender and estimate the age groups when the range of the groups expands. This is in correspondence to combining the auditory data with the age group, as conducted and reported in the next subsection where we combine the auditory information with the age group of the subjects, rather than the precise age of each participant.

4.2 Auditory Classification

4.2.1 EnvNet for Respiratory Sound and Demographic Information.

We first conduct a set of experiments to explore the best DCNN configuration based on the EnvNet DCNN that achieves the highest accuracy. Then, we combined the audio dataset with the age groups they are recorded from as depicted in Fig. 1. Table 3 compares the two sets of experiments, indicating that the COPD and healthy conditions are diagnosed with higher accuracy and resulting in a total test accuracy increase by 5% when the demographic information is taken into account.

Table 3: Respiratory sound classification accuracy and model complexity with and without taking the demographic information into account.

DCNN characteristics			Sensitivity					Accuracy
Model	#params	FLOPs	URTI	Healthy	COPD	Bronchiec.	Bronchiol.	Test
EnvNet-like on Sound w/o Demographic Info	320k	0.194B	21%	68%	96%	88%	4%	78%
EnvNet-like on Sound with Demographic Info	320k	0.194B	16%	72%	100%	88%	15%	83%

Table 4: Deploying the EnvNet model to commercial off-the-shelf devices including a dual-core Denver CPU, a quad-core ARM A57 CPU, and a combination of ARM CPU + Pascal GPU from the NVIDIA TX2 board.

Configuration	CPU Freq. (MHz)	GPU Freq. (MHz)	Power (mW)	Latency (S)	Performance (GFLOP/S)	Energy (J)	Energy Eff (GFLOPS/W)
Denver CPU	345	-	881	10.0	0.019	8.81	0.021
	2035	-	3170	0.9	0.215	2.85	0.068
ARM A57 CPU	345	-	1168	3.7	0.052	4.32	0.045
	2035	-	4425	0.6	0.322	2.66	0.073
TX2 CPU+GPU	2035	1300.5	9106	0.1	1.935	0.91	0.210

5 COMMERCIAL OFF-THE-SHELF DEVICE DEPLOYMENT

The framework is intended to be flexibly deployable for general-purpose devices where the developed ML models trained on the framework can be deployed onto processing machines that may range from front-end edge devices to back-end computer servers. Trading off between the computation complexity and the classification accuracy, trained ML models can be deployed to edge devices (e.g. a cell-phone, tablet) to process data locally if the information privacy is a concern, or otherwise to the cloud servers that can process data with more elaborate up-to-date models that yield higher quality metrics.

All of the DCNN models are attributed to at least two hardware-level characteristics: the model size and the number of computer operations per inference, both of which are upper-bounded by the platform resources that they deploy to, or by the inference deadline. When putting all the components of the framework together, both the hardware resource constraints and the diagnosis latency should meet the application goals. Having set the batch-size equal to 1, the trained models obtained from the previous Section are deployed on two mobile CPUs including Denver (dual-core) and ARM-Cortex A57 (quad-core) as well as an embedded CPU+GPU implementation with different frequency settings. All of the settings were performed on the TX2 development board that provides precise on-board power measurement. Table 4 summarizes the implementation, indicating that, provided a 5s frame of recording to the memory, the least power dissipating implementation (Denver with a low frequency) takes 10 seconds to classify one frame whereas the most energy-efficient implementation (CPU+GPU) dissipates approximately 10× more power to classify the same frame within 0.1 seconds.

Table 5: Comparison to the Related Work

Related Work	#Augmented Audio Samples within the Dataset								Test
	URTI	Healthy	COPD	Bronchiec.	Bronchiol.	Asthma	LRTI	Pneum.	Accuracy
[21]	403	455	10,205	377	-	13	26	481	97%
This Work	370	560	567	256	208	-	-	-	83%

6 COMPARISON

The most related work to ours that has developed a DCNN on the same respiratory sound database is the work in [21] that reports an overall accuracy of 97%. The main difference between the two works is that our model uses additional information in tandem with the audio data and proposes a framework that suggests combining as much existing correlated information within the dataset as possible to rectify and increase the diagnosis accuracy. The other difference is that our selected dataset is semi-balanced among 5 classes of respiratory sounds recorded from one unique medical device that has been indistinguishably utilized for 61 subjects diagnosed with 7 out of 8 classes within the database, whereas the dataset selection in [21] is excessively dominated with COPD recordings, a major portion of which, as depicted in Fig. 2-D, are recorded by two medical devices that have been used to merely sample from COPD-diagnosed participants. Table 5 provides a comparison and a summary of the total number of augmented samples per class within the two works.

7 CONCLUSION

In an attempt to exploit machine learning algorithms to classify respiratory problems, we proposed a framework that employs as much correlated information as a dataset provides and showed that with combining both auditory and demographic information for a selection of reasonably balanced dataset out of a publicly released respiratory sound database the diagnosis accuracy of the trained deep convolutional neural networks (DCNNs) increases by 5%. Since the demographic data can be extracted and estimated using computer vision, we suggest using another DCNN that works in parallel to the auditory signal processing DCNN to estimate the demographic information of the subject under test. Lastly, we deploy our DCNN models on a dual-core Denver CPU, a quad-core ARM Cortex A57, and a heterogeneous implementation of CPU+GPU from the NVIDIA TX2 development board to measure hardware characteristics when deploying the model to an embedded device.

REFERENCES

[1] 2018 (accessed June, 2020). The Top 10 Causes of Death. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Jyotibdha Acharya and Arindam Basu. 2020. Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning. *IEEE transactions on biomedical circuits and systems* 14, 3 (2020), 535–544.

[3] Justice Amoh and Kofi Odame. 2016. Deep neural networks for identifying cough sounds. *IEEE transactions on biomedical circuits and systems* 10, 5 (2016), 1003–1011.

[4] Murat Aykanat, Özkan Kılıç, Bahar Kurt, and Sevgi Saryal. 2017. Classification of lung sounds using convolutional neural networks. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 65.

[5] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj. 2020. Convolutional neural networks based efficient approach for classification of lung diseases. *Health Information Science and Systems* 8, 1 (2020), 4.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] H.Ren et al. 2020, in press. End-to-end Scalable and Low Power Multi-modal CNN for Respiratory-related Symptoms Detection. In *2020 IEEE 33rd International System-on-Chip Conference (SOCC) (SOCC 2020)*.

[8] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275* (2020).

[9] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913* (2019).

[10] Ping-Ing Lee, Ya-Li Hu, Po-Yen Chen, Yhu-Chering Huang, and Po-Ren Hsueh. 2020. Are children less susceptible to COVID-19? *Journal of Microbiology, Immunology, and Infection* (2020).

[11] Kai Liu, Ying Chen, Ruzheng Lin, and Kunyuan Han. 2020. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection* (2020).

[12] Renyu Liu, Shengsheng Cai, Kexin Zhang, and Nan Hu. 2019. Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network. In *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. IEEE, 298–303.

[13] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang. 2019. LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.

[14] Hiroshi Nakano, Tomokazu Furukawa, and Takeshi Tanigawa. 2019. Tracheal sound analysis using a deep neural network to detect sleep apnea. *Journal of Clinical Sleep Medicine* 15, 8 (2019), 1125–1133.

[15] Diego Perna. 2018. Convolutional neural networks learning from respiratory data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2109–2113.

[16] Diego Perna and Andrea Tagarelli. 2019. Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 50–55.

[17] Lam Pham, Ian McLoughlin, Huy Phan, Minh Tran, Truc Nguyen, and Ramaswamy Palaniappan. 2020. Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases. *arXiv preprint arXiv:2002.03894* (2020).

[18] BM Rocha, D Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, A Oliveira, C Jácome, A Marques, et al. 2017. A respiratory sound database for the development of automated classification. In *International Conference on Biomedical and Health Informatics*. Springer, 33–37.

[19] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement* 40, 3 (2019), 035001.

[20] Heechang Ryu, Jinkyoo Park, and Hayong Shin. 2016. Classification of heart sound recordings using convolution neural network. In *2016 Computing in Cardiology Conference (CinC)*. IEEE, 1153–1156.

[21] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. 2019. Lung Disease Classification using Deep Convolutional Neural Network. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 732–735.

[22] Yuji Tokozume and Tatsuya Harada. 2017. Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2721–2725.

[23] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[24] Xianxian Zhao, Bili Zhang, Pan Li, Chaoqun Ma, Jiawei Gu, Pan Hou, Zhifu Guo, Hong Wu, and Yuan Bai. 2020. Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. *MedRxiv* (2020).

On Machine Learning-Based Short-Term Adjustment of Epidemiological Projections of COVID-19 in US

Machine Learning Adjustment of COVID-19 Epidemiological Projections

Sarah Kefayati*
IBM Watson Health, Cambridge, MA, USA
Sarah.kefayati@ibm.com

Hu Huang
IBM Watson Health, Cambridge, MA, USA
hu.huang@ibm.com

Prithwish Chakraborty
IBM Research, Yorktown Heights, NY, USA
prithwish.chakraborty@ibm.com

Fred Roberts
IBM Watson Health, Cambridge, MA, USA
robertsf@us.ibm.com

Vishrawas Gopalakrishnan
IBM Watson Health, Cambridge, MA, USA
vishrawas.gopalakrishnan1@ibm.com

Raman Srinivasan
IBM Watson Health, Cambridge, MA, USA
rsrin@us.ibm.com

Sayali Pethe
IBM Watson Health, Cambridge, MA, USA
sayali.pethe@ibm.com

Piyush Madan
IBM Research, Yorktown Heights, NY, USA
piyush.madan1@ibm.com

Ajay Deshpande
IBM Watson Health, Cambridge, MA, USA
ajayd@us.ibm.com

Xuan Liu
IBM Watson Health, Cambridge, MA, USA
xuanliu@us.ibm.com

Jianying Hu
IBM Research, Yorktown Heights, NY, USA
jyhu@us.ibm.com

Gretchen Jackson
IBM Watson Health, Cambridge, MA, USA
gretchen.jackson@ibm.com

ABSTRACT

Epidemiological models have provided valuable information for the outlook of COVID-19 pandemic and relative impact of different mitigation scenarios. However, more accurate forecasts are often needed at near term for planning and staffing. We present our early results from a systemic analysis of short-term adjustment of epidemiological modeling of COVID 19 pandemic in US during March-April 2020. Our analysis includes the importance of various types of features for short term adjustment of the predictions. In addition, we explore the potential of data augmentation to address the data limitation for an emerging pandemic. Following published literature, we employ data augmentation via clustering of regions and evaluate a number of clustering strategies to identify early patterns from the data.

From our early analysis, we used CovidActNow as our underlying epidemiological model and found that the most impactful features for the one-day prediction horizon are population density, workers in commuting flow, number of deaths in the day prior to prediction date, and the autoregressive features of new COVID-19 cases from three previous dates of the prediction. Interestingly, we also found that counties clustered with New York County resulted in best preforming model with maximum of R^2= 0.90 and minimum of R^2= 0.85 for state-based and COVID-based clustering strategy, respectively.

KEYWORDS

COVID-19, epidemiological projections, machine learning, clustering, US demographic

ACM Reference format:

Sarah Kefayati, Hu Huang, Prithwish Chakraborty, Fred Roberts, Vishrawas Gopalakrishnan, Raman Srinivasan, Sayali Pethe, Piyush Madan, Ajay Deshpande, Xuan Liu, Jianying Hu, Gretchen Jackson 2020. On Machine Learning-Based Short-Term Adjustment of Epidemiological Projections of COVID-19 in US. *KDD epiDAMIK 3.0, August 2020, San Diego, CA, USA*

INTRODUCTION

The novel Coronavirus, SARS-CoV-2, was first detected in Wuhan, China on December 31, 2019 and by early January 2020, it spread to 21 countries. The first case in the United States (US) was reported on January 21, 2020 in Snohomish County in the state

* To whom correspondence should be addressed.

of Washington [1]. By the middle of March 2020, the number of COVID-19 infected cases started to peak up in the US [2], inducing panic about potential shortages of hospital capacities and supplies.

As more cases were detected in the US, many epidemiological models, which had been mainly developed for other pandemics such as influenza and season-long forecasting, started to adjust to and project COVID-19 growth patterns. Due to the critical role of forecasting COVID-19 in helping policymaker for future planning, the Center of Disease Control and Prevention (CDC) has initiated an effort for forecasting in form of challenge with the participation of several of the epidemiological models [3].

Although overall epidemiological models have been useful for understanding the future outlook of COVID-19 pandemic, they often have been criticized for overestimating the projections and inducing uncertainties [4]. The limitations in predicting future cases of COVID-19 stems from a combination of limited understanding and a lack of data about its infectious spread pattern due to its novel nature. Beside data limitations, region-specific factors were often not accounted-for in many published epidemiological models of the COVID-19 pandemic. Factors such as age distribution, comorbidities, and pre-excising conditions, as well as socioeconomic factors, are expected to play role in the severity and duration of this disease.

Given the importance of COVID-19 forecasting and considering the majority of epidemiological models are best suited for long-term projections, in this work rather than developing a new forecasting model, we aim at improving the overestimated projected numbers from epidemiological models on a short-term basis based on machine learning. In particular, we conducted a systematic analysis to study the importance of various data elements for our short-term prediction. Our predictive modeling included a carefully selected 40 geo-specific features for the US counties.

A previous study by Liu et al. on machine learning-based predictions of COVID-19 outbreak in China, showed that modeling for the regions clustered together based on geo-specific similarities resulted in improved predictive performance for majority of China provinces [5]. In this work, we studied four different clustering strategies, primarily based on demographic similarities, COVID activity trends, state boundaries, as well as a national cluster, to examine the effects of these boundary conditions in predictive power of our model.

Our main contribution is investigating the role of machine learning to adjust the short-term epidemiological projections with the ultimate aim of helping counties and hospitals better plan their resources. We are also investigating which additional region-specific features play a role in short-term adjustments of the COVID-19 projections.

1 RESULTS

For historical epidemiological projection, we used the CovidActNow model as one of the early epidemiological models that became opensource [6]. For the data augmentation purpose, we considered four clustering strategies: one national cluster including all the counties in our training set, second clustering based on counties state boundary, and another two clustering strategies were based on counties similarities in COVID-19 spread characteristics and demographic information as described in the Methods section.

The epidemiological historical projections were one-day ahead on any given day during March 23rd-April 20th, 2020 period. We trained an individual date-dependent models for each of the clusters within each clustering strategy. The out-of-sample test sets were evaluated from April 6th to April 19th, 2020 individually each with the date-appropriate trained model (the total of 14 date-dependent models for each cluster). The comparison and the impact of the four different clustering boundary conditions is shown in Table 1.

Regardless of the clustering strategy, our one-day-ahead prediction showed ~97% uplift compared to the epidemic projection with the same one-day-ahead horizon. Overall, the performance of the models based on each clustering strategy was similar.

Table 1: Comparison of four clustering strategies. *RMSE* values are derived by comparing our model predictions to the ActNow model ground truth values. The uplift percentage is derived based on ActNow projections (*RMSE* = 6427.27)

Clustering boundary condition	National *tot. 1*	Demographic *tot. 6*	COVID *tot. 12*	State *tot. 51*
RMSE	193.29	194.10	192.49	196.94
Overall R^2	0.883	0.875	0.875	0.865
% uplift	96.99%	96.98%	97.00%	96.93%

Also consistent across all four different clustering strategies, the best model performance was achieved for the cluster in which New York County was included. Figure 1 shows the comparison of our model output to the CovidActNow projected “all infected” cases both compared with the ground truth (i.e., estimated all infected) for three example counties which clustered together with New York County for two different clustering strategies. Both COVID-based and demographic-based clustering achieved 97% uplift compared to the epidemiological projections for the cluster model that included New York County.

Closer assessment of two major counties of New York and Los Angeles further revealed the similarities of four different clustering strategies, particularly for New York County as shown in Figure 2. The best performing cluster model for New York County was found to be state-based clustering with $R^2 = 0.90$. Los Angeles County was obtained from the COVID-based clustering method, which follows similar trajectory as the national clustering with comparable *RMSE* (676.34 and 695.61, respectively). State-clustered model showed highest deviation with *RMSE* = 1249.78, followed by second highest *RMSE* = 723.75 from the demographic-based clustering model. We found that the level of COVID-19 inactivity or low activity in the cluster impacted the performance for the corresponding model. For example, comparing two of the

clusters containing Los Angeles County from two different clustering strategies, the state-based cluster contained 36% zero cases as opposed to only 2% zero cases in the COVID-based cluster.

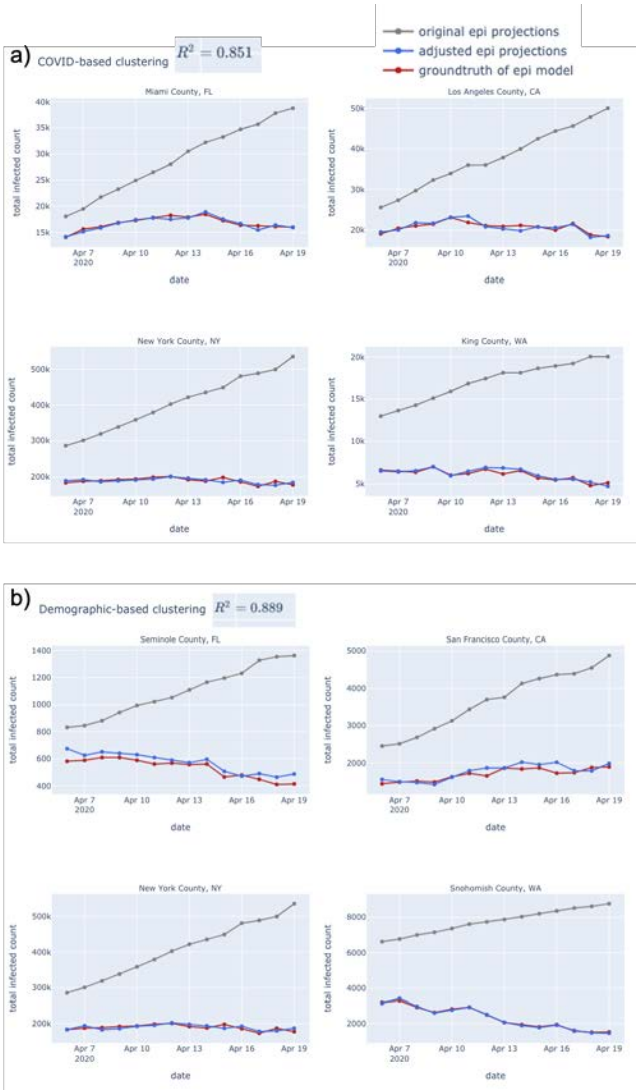


Figure 1: One-day-horizon predictions from our machine learning model and from the CovidActNow epidemiological model both compared to the ground truth for two different clustering strategies: a) one of the 12 clusters of the COVID-based clustering with 49 counties in the cluster, and b) one of the 6 clusters of the demographic-based clustering with 259 counties in the cluster.

As LASSO gives spare weights by driving small weights to zero, we can detect the most predictive features of our model. As walk-forward split expands by every date and more training data gets included in the split, fewer features become prominent. For the national model, as shown in Figure 3, the most impactful features for the last trained model were population density, and the autoregressive features of new COVID cases from three previous date of the prediction. In the first trained model (with the least training data), in addition to population density and autoregressive features, workers in commuting flow and cumulative number of cases had positive predictive effects on predicting the next day's infected cases. In contrast, population and next day projected cases from the CovidActNow output had negative predictive power. We speculate the negative predictive power of CovidActNow is due to its increasing trend as opposed to ground truth that has downward trend generally. It is also important to note for the first trained model (first expanding window iteration) with limited training data, LASSO attempts to distribute the weights to more features. As the training data expands in the last trained window, the effect of CovidActNow projections is eliminated. Other features had small weights (either negative or positive) and eventually weighted to zero for the last trained model. Similarly, we found that population density and three autoregressive features were the most important (non-zero and positive) features for the majority including best performing clusters, for other clustering strategies. For the demographic-based clustering due to homogeneous distribution of the, regardless of the date of trained model, none of the demographic features carried any weight with the exception of the population density and workers in commuting flow. Overall, no particular pattern was detected for the role of race and sex with few exceptions. For instance, the proportion of non-Hispanic black population showed positive predictive impact for the State of Wisconsin with the state-based clustering approach. For feature importance map of other three clustering strategy please refer to APPENDIX II.

2 Methods

For the purpose of data augmentation, we considered grouping counties four different ways to ensure the that the scarcity of available training data during early period of COVID-19 pandemic (window considered March 23rd-April 20th, 2020) did not impact the model performance. One grouping included counties of each of the US states, yielding 51 individual trained models. Another national approach involved grouping all the counties together and training one model for all. The two other groupings were conducted by clustering the counties based on their similarities in COVID-19 spread characteristics and demographic information as described below.

2.1 Clustering Strategy

Clustering was conducted based on agglomerative hierarchal clustering algorithm (scikit-learn package), which is a bottom-up approach merging counties based on their similarities until reaching one big cluster. The optimal number of clusters was obtained by maximizing Calinski-Harabasz score.

The counties were clustered based on their demographic information for which the optimization yielded 6 clusters. Features included race, ethnicity, gender, elderly (age > 65 years) and young (age < 18 years) population, total and density of population, county traffic volume, county average commute flow, as well as Area

Deprivation Index, rankings of counties by socioeconomic status disadvantage [7].

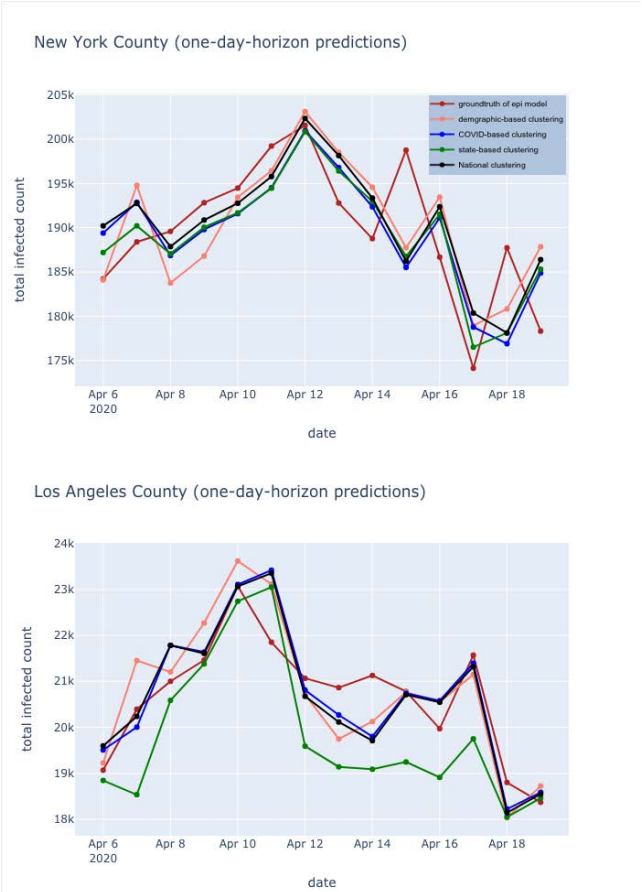


Figure 2: One-day-horizon predictions of our models trained individually according to the clustering strategy for two of the major counties.

Counties were also separately clustered based on time-varying COVID-19 characteristics, resulting in 12 dissimilar clusters. Features included reciprocal doubling time based on the growth during the selected time window, the pattern of the growth curve (logistic, exponential or none), cumulative cases on April 20th, 2020, and number of days stayed at home since the stay-at-home order for each county. Maps of the clustered US counties based on their demographic and COVID-19 activity similarities are available in APPENDIX I.

2.2 Epidemiological Projections

The Coronavirus Act Now model is SEIR model, one of the main groups of epidemiological models used by epidemiologists and researchers to project the evolution of a disease. The model works by categorizing the population at various states and modeling as the population moves through susceptible (S), individuals becoming exposed (E) to the virus, then infected (I), and infected population either recovering (R) or dying (D). To model hospitalization and need for ICU bed, the infected cases were categorized into three levels of disease severity: mild (no hospitalization), moderate (requiring hospitalization), or severe (requiring ICU bed)[1].

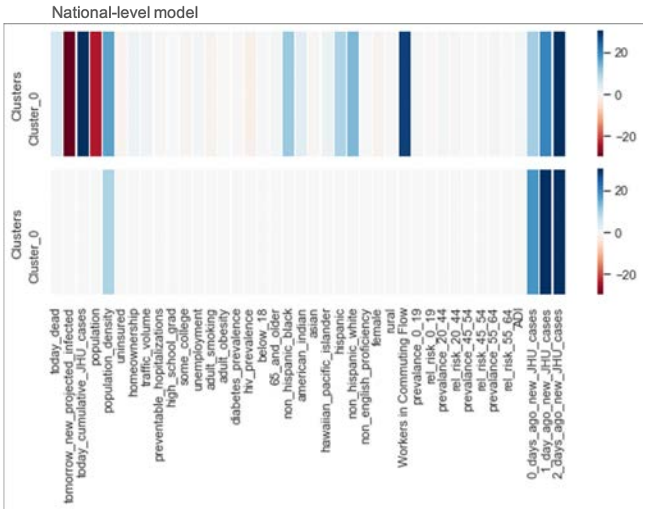


Figure 3: LASSO selected features for the first (top panel) and last (bottom panel) trained models (out of 14 date-dependents models) trained for all of the counties included in our model (total of 1578 counties). Although majority are small values, none of the features in the first trained model have zero coefficients. The last model has 36 features with coefficients value of zero displayed for better comparison.

For the earlier version of the ActNow model, caseload data (number of confirmed cases and death) were updated daily from the Johns Hopkins University (JHU) Center for Systems Science and Engineering's Coronavirus Tracking Dashboard [8] with the county-level data becoming available as of March 22nd, 2020.

We used "all infected" projection (including mild, moderate, and severe cases) output of the CovidActNow model to compare with our machine learning forecasting. In order to compare our model results with CovidActNow, we used the same ground truth as CovidActNow [6]. In the earlier version of the ActNow model[1], the ground truth for the "all infected" (i.e. estimated all infected) was derived as follow: "estimated recovered" was estimated from actual COVID cases report (JHU) shifted by 13 days. The "estimated recovered" cases together with total number of deceased populations on a given date gives an estimate of active cases. Due to lack of reported data on hospitalization at the county level, CovidActNow then estimates that a quarter of the "estimated active" cases are hospitalized. And finally, that "estimated hospitalization" is about 7.3% of "estimated all infected" (mild,

[1] As of April 12th, 2020, the ActNow model has gone through an update and "asymptomatic individuals" category has been added to the model. However, the projections used in this study are the outputs of the model prior to this change and thus do not include asymptomatic cases.

moderate, and severe) cases. Thus, we used "estimated all infected" as the ground truth to compare our forecasting to the ActNow model projections.

The one-day-horizon CovidActNow historical projections were obtained by running a model for every day during March 23rd-April 20th, 2020 period and selecting the next-day projections yielding 28 days datapoints. Since the ActNow model produced several intervention scenarios, for each date, we selected the output that matched the actual in-place intervention. We obtained the intervention policies for each state from The New York Times website. For some of the states, the policies were not held state-wide, thus scraped those individual county policies from various news outlets.

2.3 Predictive Model

For the predictive modeling, we fitted our data to a LASSO model, a multivariate linear model with L1-norm regularization (penalizing the absolute sum of the model coefficients). The feature vector included both time-dependent and static features for each county. The time-dependent were COVID-19 dependent features both from the epidemiological model output and from JHU reported parameters. The time-independent features were demographic and characteristics of each county. In total, static and time-varying, 40 features were included in our model. For the list of the features and their sources please refer to APPENDIX III.

The number of new confirmed COVID-19 cases for the next day was then predicted based on:

$$y_{t+1} = \sum_{i=0}^{2} \alpha_i y_{t-i} + \beta d_t + \gamma i_{t+1} + \delta C_t + \sum_{j=1}^{34} \mu_j S_j$$

Where y_{t+1} , y_t , y_{t-1}, and y_{t-2} are the new confirmed cases for the next day, same date, the day before, and two days prior of date t. d_t refers to the new death number at date t; i_{t+1} is the new "all infected" cases projected for next day at date t from the epidemiological model, and C_t is the cumulative number of COVID-19 cases on date t. S_j refers to collection of 34 static features including, population density, county commute flow, age, gender, race, socioeconomic features such as unemployment and Area Deprivation Index, as well as disease prevalence and comorbidities.

Due to time-series nature and daily update of the data, a walk-forward with expanding window validation fashion was considered resulting in 15 splits with one day horizon and initial training window of 10 days. Since the split was performed according to the date of data, each cluster yielded 15 sets consistently; however, depending on number of counties in each cluster, different train-test splits were expected for each cluster. The total of 1578 counties were included in this study each with feature vector of size 40. To compare the predicted outcome of our model with the epidemiological projected cases, the predicted new confirmed cases were then summed with the JHU-reported "cumulative number of cases" from the previous date, and then converted to "estimated all infected" according to the conversion used in the CovidActNow model.

The best model parameters, including LASSO alpha, were selected in a Grid search manner. To simulate the real-life scenario in which the last trained model needs to be updated daily as the new data comes in, the model trained on the first n-1 splits (n=2, 3, 4, …15), was evaluated on the test set of nth split (i.e. out-of-sample test set).

3 CONCLUSIONS

Our machine learning results showed significant improvement over the epidemiological projection with a one-day prediction horizon. Although small, the changes in model performance was detectable based on different clustering strategies. Assessing New York County, for example, we found state-based clustering achieved the highest performance for the state of New York (R^2 = 0.90). However, state-based clustering resulted in lowest performance for overall counties compared to other strategies (Table 1). This finding suggests an advantage to considering training individual models based on the geographic region.

The results shown for our model are from March-April 2020 period of time, a time of dramatic increases in COVID-19 spread in major cities with subsequent implementation of mitigation strategies, such as stay-at-home policies with California first to enforce such a policy on March 19th, 2020. Our findings of different clustering strategies and feature importance were likely due to the highly dynamic nature of infectious spread and local policies at this time. Moreover, as more data has become available over time, epidemiological models are also improving and more scenarios are considered in the modeling methods, such as the impact of asymptomatic cases in the recent version of the ActNow model.

It is also important to note that for this work, our prediction is short and limited to one-day horizon. Although, it is not expected that all of the 40 features that we have included have predictive power, we suspect with longer prediction horizon, demographic and geo-specific features will have important roles.

With initially flattening and now new rises in the pandemic curve, we plan to expand our modeling prediction horizon to as long as two weeks ahead. Some of the challenges that we foresee with longer prediction horizon is the policy changes as the states would go through different phases of reopening. We suspect for longer prediction horizon, more COVID-specific trends for each state as well as information about the upcoming policies can improve the predictive powers. Also, we intend to incorporate several epidemiological models' outputs to study their difference and impact in our modeling.

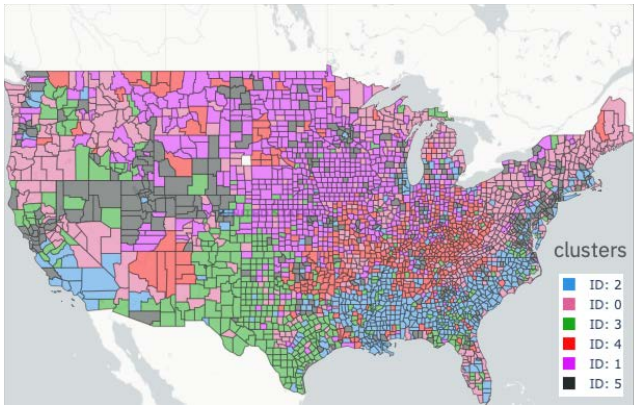
REFERENCES

[1] M. L. Holshue *et al.*, "First Case of 2019 Novel Coronavirus in the United States," *New England Journal of Medicine,* vol. 382, no. 10, pp. 929-936, 2020, doi: 10.1056/NEJMoa2001191.

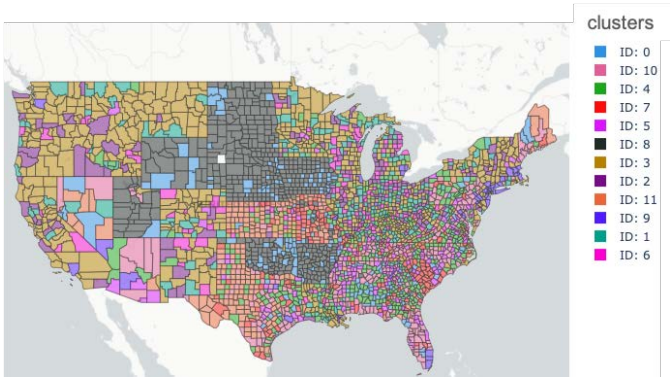
[2] "WHO report on US COVID-19 Cases." https://covid19.who.int/region/amro/country/us

[3] "CDC COVID-19 Forecasting Group Challenge." https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html
[4] N. P. Jewell, J. A. Lewnard, and B. L. Jewell, "Predictive Mathematical Models of the COVID-19 Pandemic: Underlying Principles and Value of Projections," *JAMA,* vol. 323, no. 19, pp. 1893-1894, 2020, doi: 10.1001/jama.2020.6585.
[5] D. Liu *et al.*, *A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models*. 2020.
[6] "Covid Act Now Model Reference/Assumptions." https://data.covidactnow.org/Covid_Act_Now_Model_References_and_Assumptions.pdf
[7] G. K. Singh, "Area deprivation and widening inequalities in US mortality, 1969-1998," (in eng), *Am J Public Health,* vol. 93, no. 7, pp. 1137-43, Jul 2003, doi: 10.2105/ajph.93.7.1137.
[8] "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)." https://coronavirus.jhu.edu/map.html

APPENDIX I: Clustering of US Counties

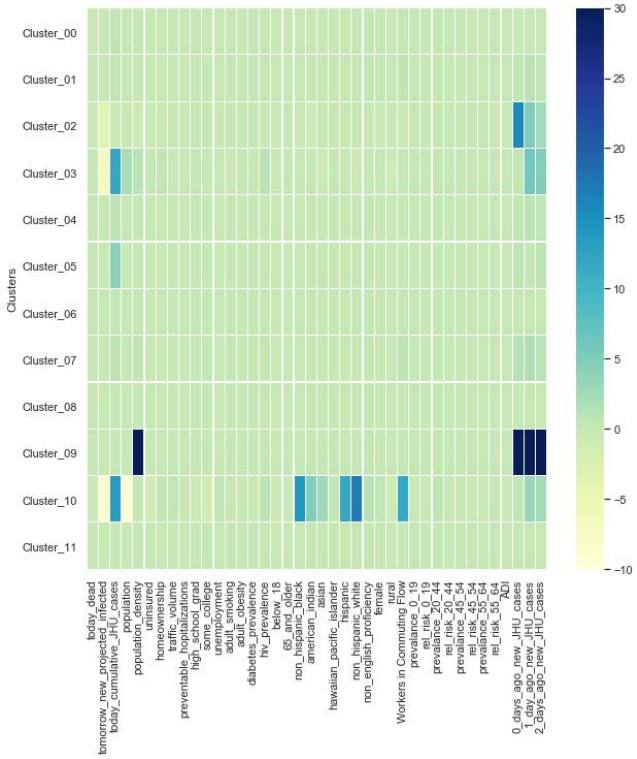


Supplementary Figure 1: US counties clustering based on their demographic similarities. For the reference, New York county is in cluster 5.

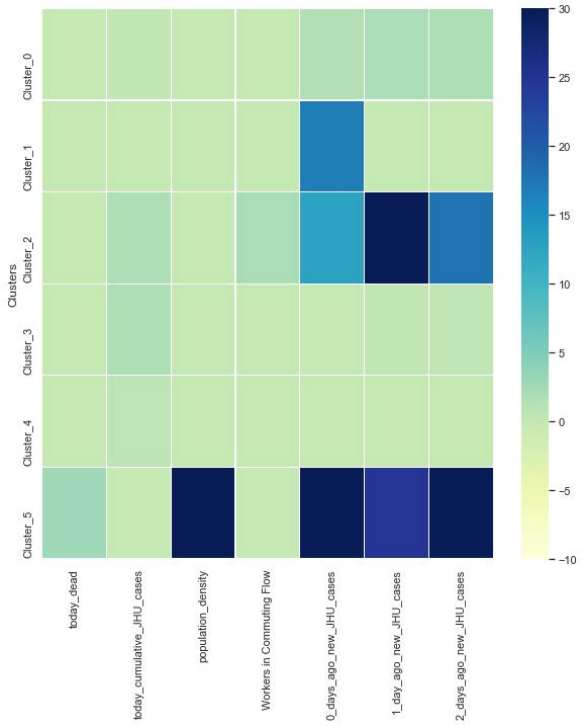


Supplementary Figure 2: US counties clustering based on their COVID-19 activity similarities during March 23rd-April 20th 2020. For the reference, New York county belongs to cluster 9.

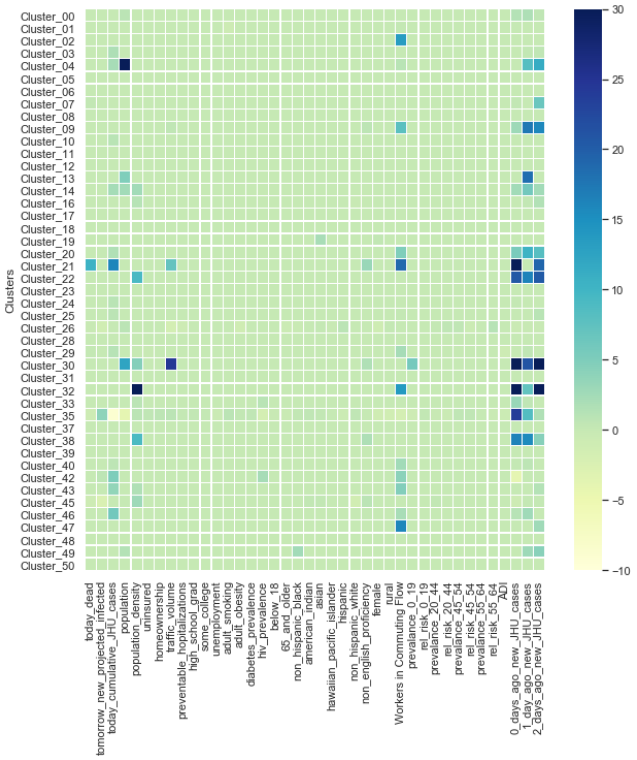
APPENDIX II: LASSO Selected Features



Supplementary Figure 3: LASSO selected features for the last trained model (out of 14 date-dependents models) of each cluster trained for the counties clustered together based on similarities of COVID characteristics. New York County is part of cluster 9, yielding best preforming model.



Supplementary Figure 4: LASSO selected features for the last trained model (out of 14 date-dependents models) of each cluster trained for the counties clustered together based on their demographic similarities. New York County is part of cluster 5, yielding best preforming model. The features with coefficients value of zero for across all the cluster-based models have been removed from the heatmap for better visual (33 zero-coefficient features)



Supplementary Figure 5: LASSO selected features for the last trained model (out of 14 date-dependents models) of each cluster trained for the counties clustered together based on their state boundary. New York County is part of cluster 32, yielding best preforming model.

APPENDIX III: Data sources

Feature	Source	Note
dead count	JHU	Daily report
New projected all infected	ActNow output for March-April version	Date t+1 values were extracted from model ran on date t
Cumulative COVID cases	JHU	Daily report
New COVID cases for date t-0	JHU	t-0 cumulative cases subtracted from day before
New COVID cases for date t-1	JHU	t-1 cumulative cases subtracted from day before
New COVID cases for date t-2	JHU	t-2 cumulative cases subtracted from day before
Area Deprivation Index	University of Wisconsin	Year 2018
Workers in commuting flow	United States Census Bureau	2011-2015 ACS commuting flows

Feature	Source	Note
- Prevalence of respiratory diseases per county broken down by age groups, - Relative risk scores reflective of comorbidities	IBM® Advantage Suite®	2018
- Population, - population density, - uninsured, - homeownership, - traffic-volume, - preventable hospitalization, - high-school grad, - some college, - unemployment, - adult smoking, - adult obesity, - diabetes prevalence, - HIV prevalence, - Age below 18, - Age above 65, - Non-Hispanic black, - American Indian, - Asian, - Hawaiian pacific islander - Hispanic, - Non-Hispanic white, - Non-English proficiency, - Female, - Rural	County Health Rankings and Roadmaps	- Data compiled from CDC and other public agency reports. Year 2018

Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs

Catherine Ordun
cordun1@umbc.edu
Univ. of Maryland, Baltimore County

Sanjay Purushotham
psanjay@umbc.edu
Univ. of Maryland, Baltimore County

Edward Raff
Raff_Edward@bah.com
Booz Allen Hamilton
eraff1@umbc.edu
Univ. of Maryland, Baltimore County

ABSTRACT

This paper illustrates five different techniques to assess the distinctiveness of topics, key terms and features, speed of information dissemination, and network behaviors for Covid19 tweets. First, we use pattern matching and second, topic modeling through Latent Dirichlet Allocation (LDA) to generate twenty different topics that discuss case spread, healthcare workers, and personal protective equipment (PPE). One topic specific to U.S. cases would start to uptick immediately after live White House Coronavirus Task Force briefings, implying that many Twitter users are paying attention to government announcements. We contribute machine learning methods not previously reported in the Covid19 Twitter literature. This includes our third method, Uniform Manifold Approximation and Projection (UMAP), that identifies unique clustering-behavior of distinct topics to improve our understanding of important themes in the corpus and help assess the quality of generated topics. Fourth, we calculated retweeting times to understand how fast information about Covid19 propagates on Twitter. Our analysis indicates that the median retweeting time of Covid19 for a sample corpus in March 2020 was 2.87 hours, approximately 50 minutes faster than repostings from Chinese social media about H7N9 in March 2013. Lastly, we sought to understand retweet cascades, by visualizing the connections of users over time from fast to slow retweeting. As the time to retweet increases, the density of connections also increase where in our sample, we found distinct users dominating the attention of Covid19 retweeters. One of the simplest highlights of this analysis is that early-stage descriptive methods like regular expressions can successfully identify high-level themes which were consistently verified as important through every subsequent analysis.

CCS CONCEPTS

• **Mathematics of computing** → **Statistical paradigms**; • **Information systems** → **Clustering and classification**; • **Computing methodologies** → *Machine learning approaches.*

KEYWORDS

network modeling, umap, topic modeling, social media

ACM Reference Format:
Catherine Ordun, Sanjay Purushotham, and Edward Raff. 2020. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 15 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

Twitter has been used as an early warning notifier, emergency communication channel, public perception monitor, and proxy public health surveillance data source in a variety of disaster and disease outbreaks from hurricanes[58], terrorist bombings [7], tsunamis [8], earthquakes [18], seasonal influenza [36], Swine flu [52], and Ebola [38]. In this paper, we conduct an exploratory analysis of topics and network dynamics of Covid19 tweets. Since January 2020, there have been a growing number of papers that analyze Twitter activity during the Covid19 pandemic in the United States.

Our contributions to the current Covid19 Twitter analyses since January 1, 2020 in Table 4, are applying machine learning methods not previously analyzed on Covid19 Twitter data, mainly Uniform Manifold Approximation and Projection (UMAP) to visualize LDA generated topics and directed graph visualizations of Covid19 retweet cascades. Topics generated by LDA can be difficult to interpret and while there exist coherence values [43] that are intended to score the interpretability of topics, they continue to be difficult to interpret and are subjective. As a result, we apply UMAP, a dimensionality reduction algorithm and visualization tool that "clusters" documents by topic. Vectorizing the tweets using term-frequency inverse-document-frequency (TF-IDF) and plotting a UMAP visualization with the assigned topics from LDA allowed us to identify strongly localized and distinct topics. We then visualized "retweet cascades", which describes how a social media network propagates information [22], through the use of graph models to understand how dense networks become over time and which users dominate the Covid19 conversations. This paper studies five research questions:

(1) What high-level trends can be inferred from Covid19 tweets?
(2) Are there any events that lead to spikes in Covid19 Twitter activity?
(3) Which topics are distinct from each other?
(4) How does the speed of retweeting in Covid19 compare to other emergencies, and especially similar infectious disease outbreaks?
(5) How do Covid19 networks behave as information spreads?

2 DATA COLLECTION

Similar to researchers in Table 4, we collected Twitter data by leveraging the free Streaming API. From March 24, 2020 to April 9, 2020, we collected 23,830,322 (173 GB) tweets. Note, in this paper, we refer to the Twitter data interchangeably as both "dataset" and "corpora" and refer to the posts as "tweets". Our dataset is a collection of tweets from different time periods shown in Table 5.

Table 1: Average Frequency of Keyword Tweets by Minute

Corpus	bed	hospital	mask	icu	help	nurse	doctors	vent	test_pos	serious_cond	exposure	cough	fever
3/24/2020	3.341	30.068	**38.295**	3.159	2.591	4.886	8.455	25.977	0.636	0.023	0.250	0.409	0.023
3/25/2020	3.117	33.021	**38.734**	2.819	3.181	3.745	8.064	24.691	1.298	0.043	0.277	0.372	0.106
3/28/2020	1.819	30.648	**34.352**	1.714	2.362	4.800	8.486	38.790	0.962	0.019	0.181	0.181	0.029
3/30/2020	2.783	40.957	**53.796**	2.311	3.287	6.996	13.009	24.887	1.111	0.025	0.215	0.296	0.043
3/31/2020	2.109	30.673	**72.877**	1.447	3.677	5.633	10.410	17.995	1.020	0.014	0.152	0.494	0.147
4/2/2020	2.065	29.410	**84.467**	1.474	3.164	6.147	10.450	23.424	0.814	0.018	0.192	0.357	0.045
4/5/2020	2.218	31.812	**62.786**	2.493	3.039	5.798	10.735	17.909	1.026	0.014	0.175	0.309	0.052
Mean	2.493	32.370	**55.044**	2.203	3.043	5.429	9.944	24.811	0.981	0.022	0.206	0.345	0.064

Using the Twitter API through tweepy, a Python Twitter mining and authentication API, we first queried the Twitter track on twelve query terms to capture a healthcare-focused dataset: 'ICU beds', 'ppe', 'masks', 'long hours', 'deaths', 'hospitalized', 'cases', 'ventilators', 'respiratory', 'hospitals', '#covid', and '#coronavirus'. For the keyword analysis, topic modeling, and UMAP tasks, we analyzed non-retweets that brought the corpus down to 5,506,223 tweets. In the Time-to-Retweet and Network Analysis, we included retweets but selected a sample out of the larger 23.8 million corpus of 736,561 tweets. Our preprocessing steps are described in the Data Analysis section that follows.

3 KEYWORD TREND ANALYSIS

Prior to applying keyword analysis, we first had to pre-process the corpus on the "text" field. First, we removed retweets using regular expressions, in order to focus the text on original tweets and authorship, as opposed to retweets that can inflate the number of messages in the corpus. We use no-retweeted corpora for both the keyword trend analysis and the topic modeling and UMAP analyses. Further we formatted datetime to UTC format, removed digits, short words less than 3 characters, extended the NLTK stopwords list to also exclude "coronavirus", "covid19", "19", "covid", removed "https:" hyperlinks, removed "@" signs for usernames, removed non-Latin characters such as Arabic or Chinese characters, and implemented lower-casing, stemming, and tokenization. Finally, using regular expressions, we extracted tweets that contained the following thirteen single terms: 'bed', 'hospital', 'mask', 'icu', 'help', 'nurse', 'doctors', 'vent', 'test_pos', 'serious_cond', 'exposure', 'cough', and 'fever', in order to gain insights about currently trending public concerns. We present values of the raw counts of the tweets in the Appendix under Table 6 and the frequencies of tweets per minute here in Table 1.

The greatest rate of tweets occurred for the tweets consisting of the term "mask" (mean 55.044) in Table 1, followed by "hospital" (mean 32.370) and "vent" (mean 24.811). Tweets of less than 1.0 mean tweets per minute, came from groups about testing positive, being in serious condition, exposure, cough, and fever. This may indicate that people are discussing the issues around Covid19 more frequently than symptoms and health conditions in this dataset. We will later find out that several themes consistent with these keyword findings are mentioned in topic modeling to include personal protective equipment (PPE) like ventilators and masks, and healthcare workers like nurses and doctors.

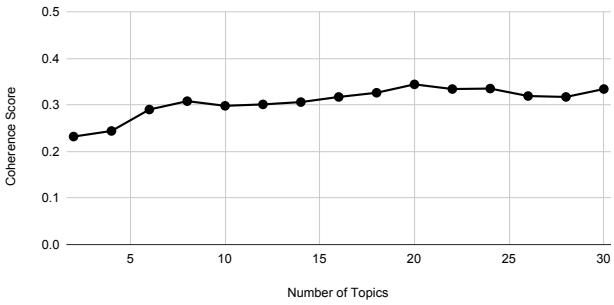


Figure 1: Coherence Scores by Number of Topics

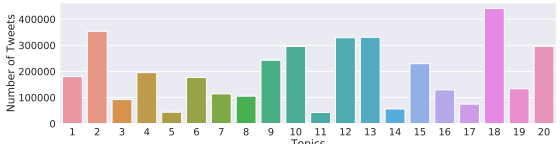


Figure 2: Distribution of 20 Topics in the Corpora

4 TOPIC MODELING

LDA are mixture models, meaning that documents can belong to multiple topics and membership is fractional [6]. Similar to methods described by Syed et al. [51], we ran 15 different LDA experiments varying the number of topics from 2 to 30, and selected the model with the highest coherence value score. We selected the LDA model that generated 20 topics, with a medium coherence value score of 0.344. Roder et al. [43] developed the coherence value as a metric that calculates the agreement of a set of pairs and word subsets and their associated word probabilities into a single score.

Our final model generated 20 topics using the default parameters of the Gensim LDA MultiCore model [1] with an overall coherence score of 0.428 after modifying the chunksize to 50,000. The topics are provided in Figure 2 and include the terms generated and each topic's coherence score measuring interpretability. Similar to the high-level trends inferred from extracting keywords, themes about PPE and healthcare workers dominate the nature of topics. The

[1] https://radimrehurek.com/gensim/models/ldamulticore.html

Table 2: 20 Topics Generated from LDA Model

Topic	C_V	Terms	Language
1	0.922	de, la, el, en, que, lo, por, del, para, se, es, con, un, al, est, una, su, ms, caso, todo	Spanish
2	0.241	like, look, work, dont, amp, peopl, time, read, support, respiratori, great, death, us, case, hospit, listen, im, presid, agre, way	English
3	0.222	hospit, realli, patient, johnson, bori, oh, shit, amp, peopl, make, death, e, blood, like, call, treat, human, trial, guy	English
4	0.171	china, thank, lockdown, viru, latest, corona, pandem, covid2019, us, lie, hai, ye, stayhom, trump, daili, way, social, quarantin, help, 5g	English
5	0.363	case, spread, help, slow, risk, symptom, daili, mask, identifi, sooner, asymptomat, us, test, market, selfreport, de, 2, 9, question, commun	English
6	0.413	day, case, week, news, ago, state, health, two, month, death, last, 15, us, delhi, hospit, one, 2, new, said, lockdown	English
7	0.287	test, case, hospit, posit, corona, dr, viru, kit, patient, ppe, doctor, data, govern, work, de, say, vaccin, death, drug, amp	English
8	0.173	die, world, peopl, case, us, death, der, tell, und, flu, corona, da, im, never, cant, fr, thousand, africa, help, ist	English
9	0.413	mask, face, wear, make, one, public, protect, cdc, peopl, dont, n95, recommend, us, viru, love, cloth, new, 0, trump, work	English
10	0.440	mask, home, stay, peopl, pleas, ppe, hospit, help, work, wear, amp, like, worker, care, nurs, safe, sure, dont, doctor, hand	English
11	0.296	hospit, nurs, le, case, de, ppe, work, new, doctor, go, pay, help, let, one, live, us, local, time, staff, lockdown	English
12	0.572	case, death, new, report, total, confirm, day, posit, number, york, us, state, 1, today, 2, 3, updat, test, peopl, rise	English
13	0.483	mask, ppe, ventil, hospit, medic, trump, suppli, donat, us, need, worker, state, china, n95, million, use, help, order, equip, amp	English
14	0.713	de, que, e, em, da, per, el, com, la, para, um, se, os, le, na, un, mai, brasil, dia, del	Portuguese
15	0.490	case, death, number, total, countri, updat, time, india, confirm, recov, china, corona, hour, last, us, news, peopl, new, activ, hospit	English
16	0.582	di, il, e, la, na, per, che, non, sa, al, si, un, da, del, ng, ang, le, ha, con, het	Italian
17	0.247	great, god, news, sad, shame, ppe, bless, hydroxychloroquin, hospit, de, death, ventil, stori, die, amp, hear, man, case, hong, holi	English
18	0.329	trump, peopl, death, american, live, stop, amp, us, let, hospit, time, viru, caus, like, one, dont, true, go, kill, media	English
19	0.904	de, le la, en, et, du, pour, un, pa, que, il, ce, au, qui, confin, dan, une, est, cest, sur	French
20	0.293	hospit, im, peopl, still, govern, dont, thing, amp, death, fuck, one, work, job, state, money, model, us, start, happen, ive	English

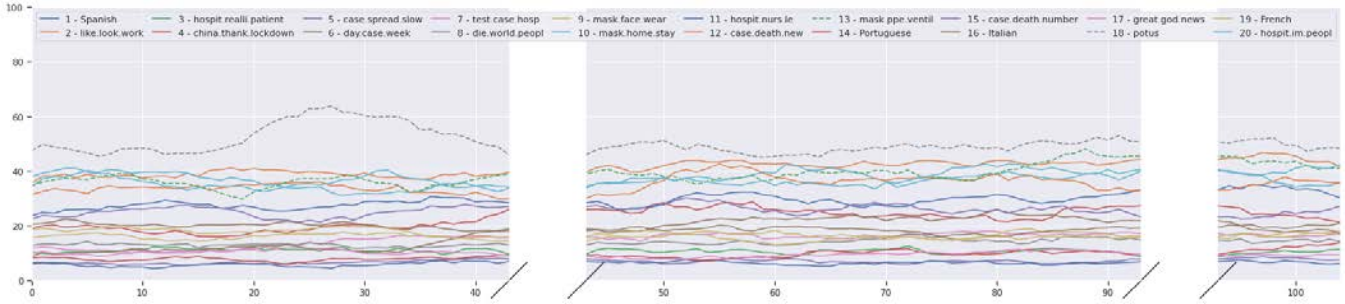


Figure 3: Trend of Topics over Time from March 24 to March 28, 2020

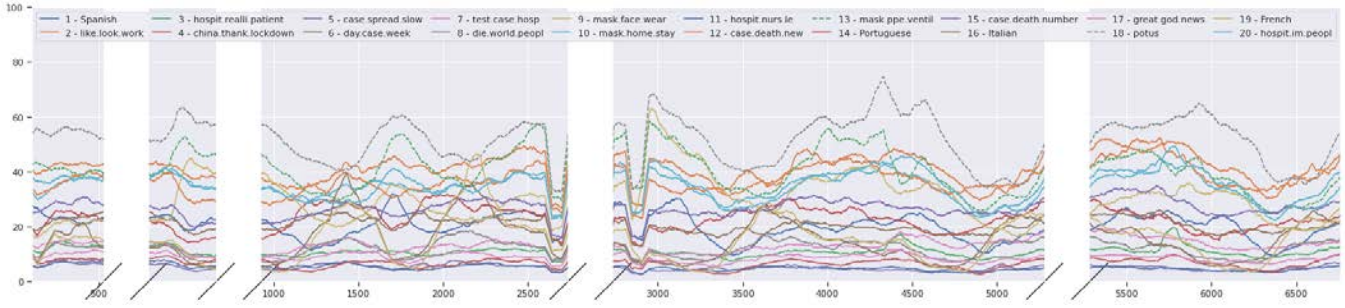


Figure 4: Trend of Topics over Time from March 30 to April 8, 2020

terms generated also indicate emerging words in public conversation including "hydroxychloroquine" and "asymptomatic".

Our results also show four topics that are in non-English languages. In our preprocessing, we removed non-Latin characters in order to filter out a high volume of Arabic and Chinese characters. We did not filter out for Latin characters, leading our topics to be a mix of English, Spanish, Italian, French, and Portuguese. This is consistent with what Singh et al. [46] reported as a variety of languages in Covid19 tweets upon analyzing over 2 million tweets. As a result, we labeled the four topics by the language of the terms in the respective topics: "Spanish" (Topic 1), "Portuguese" (Topic 14), "Italian" (Topic 16) and "French" (Topic 19). We used Google Translate to infer the language of the terms. This study may be strengthened by working with a native speaker of these languages

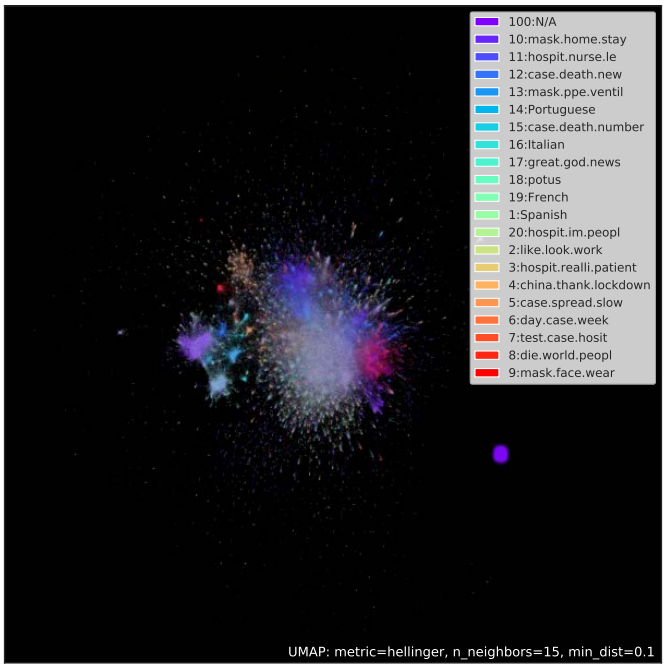


Figure 5: Visualization of One Million Tweets with Topic Labels

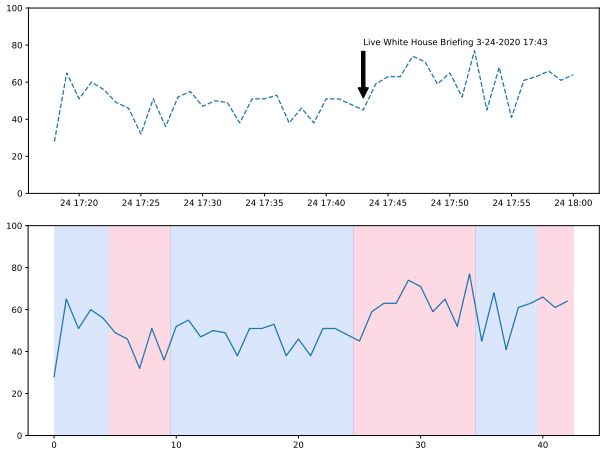


Figure 6: Change Point Detection using Binary Segmentation for March 24, 2020

to filter out stop words from these languages in order to improve the resolution and interpretability of foreign topics.

When examining the distribution of the 20 topics across the corpora in Figure 2, Topics 18 ("potus"), 12 ("case.death.new"), 13 ("mask.ppe.ventil"), and 2 ("like.look.work") were the top five in the entire corpora. For each plot, we labeled each topic with the first three terms of each topic for interpretability, for the exception of Topic 18. In our trend analysis, we summed the number of tweets per minute, and then applied a moving weighted average of 60 minutes for topics March 24 - March 28, and 60 minutes for topics March 30 to April 8th. The results plotted in figures Figure 3 and Figure 4 show similar trends on a time-series basis per minute across the entire corpora of 5,506,223 tweets. These plots are in a style of "broken axes" [2] to indicate that the corpora are not continuous periods of time, but discrete time frames, which we selected to plot on one axis for convenience and legibility. We direct the reader to Table 5 for reference on the start and end datetimes, which are in UTC format, so please adjust accordingly for time zone.

The x-axis denotes the number of minutes, where the entire corpora is 8463 total minutes of tweets. Figure 3 shows that for the corpora of March 24, 25, and 28, the topics (denoted in hash-marked lines) focused on Topic 18 "potus" and Topic 13 "mask.ppe.ventil" trended greatest. For the later time periods of March 30, March 31, April 4, 5 and 8 in Figure 4, Topic 18 "potus" and Topic 13 "mask.ppe.ventil" (also in hash-marked lines) continued to trended high. Our topic findings are consistent with the published analyses on Covid19 and Twitter, such as [46] who found major themes of healthcare and illness and international dialogue, as we noticed in our four non-English topics. They are also similar to by Thelwall et al. [55] who manually reviewed tweets from a corpus of 12 million tweets occurring earlier and overlapping our dataset (March 10 - 29). Similar topics from their findings to ours includes "lockdown life", "politics", "safety messages", "people with COVID-19", "support for key workers", "work", and "COVID-19 facts/news".

When examining the trend of the Topic 18 "potus" topic, we found that several live press briefings with the Coronavirus Task Force from @WhiteHouse would stimulate a spike in the Topic 18 topic 60 tweets per minute:

[2] https://github.com/bendichter/brokenaxes

- March 24, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:43 PM EST
- April 3, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:24 PM EST followed by a retweet from @WhiteHouse "Coronavirus—and we salute the great medical professionals on the front lines." at 5:59 PM EST
- April 4, 2020, LIVE: Press Briefing with Coronavirus Task Force at 4:13 PM EST
- April 5, 2020, LIVE: Press Briefing with Coronavirus Task Force at 6:53 PM EST
- April 6, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:41 PM EST
- April 8, 2020: LIVE: Press Briefing with Coronavirus Task Force at 5:46 PM EST

We applied change point detection in the time series of tweets per minute for Topic 18 in the datasets March 24, 2020, April 3 - 4, 2020, April 5 - 6, 2020, and April 8, 2020, to identify whether the live press briefings coincided with inflections in time. Using the ruptures Python package [56] containing a variety of change point detection methods, we used binary segmentation [25], a standard method for change point detection. Given a sequence of data $y_{1:n} = (y_1, ..., y_n)$ the model will have m changepoints with their positions $\tau_{1:m} = (\tau_1, ..., \tau_m)$. Each changepoint position is an integer between 1 and $n-1$. The m changepoints split the time series data into $m+1$ segments, with the ith segment containing $y_{(\tau_{i-1}+1):\tau_i}$. Changepoints are identified by minimizing a cost function, C for a given segment, where $\beta f(m)$ is a penalty to prevent overfitting.

$$\sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m)$$

where twice the negative log-likelihood is a commonly used cost function. Binary segmentation detects multiple changepoints across the time series by repeatedly testing on different subsets of the sequence. It checks to see if a τ exists that satisfies:

$$C(y_{1:\tau}) + C(y_{(\tau+1):n}) + \beta < C(y_{1:n})$$

If not, then no changepoint is detected and the method stops. But if a changepoint is detected, the data are split into two segments consisting of the time series before (Figure 6 blue) and after (Figure 6 pink) the changepoint. We can clearly see in Figure 6 that the timing of the White House briefing indicates a changepoint in time, giving us the intuition that this briefing influenced an uptick in the the number of tweets. We provide additional examples in the Appendix.

5 UNIFORM MANIFOLD APPROXIMATION AND PROJECTION

TF-IDF [40] is a weight that signifies how valuable a term is within a document in a corpus, and can be calculated at the n-gram level. With TF-IDF, unique words carry greater information and value than common, high frequency words across the corpus. TF-IDF has been widely applied for feature extraction on tweets used for text classification [29] [19], analyzing sentiment [5], and for text matching in political rumor detection [22]

Using the Scikit-Learn implementation of TfidfVectorizer and setting max_features to 10000, we transformed our corpus of 5,506,223 tweets into a $\mathbb{R}^{n \times k}$ sparse dimensional matrix of shape (5506223, 10000). Note, prior to fitting the vectorizer, our corpus of tweets was preprocessed during the keyword analysis stage. We chose to visualize how the 20 topics grouped together using Uniform Manifold Approximation and Projection (UMAP) [34]. UMAP is a dimension reduction algorithm that finds a low dimensional representation of data with similar topological properties as the high dimensional space. It measures the local distance of points across a neighborhood graph of the high dimensional data, capturing what is called a fuzzy topological representation of the data. Optimization is then used to find the closest fuzzy topological structure by first approximating nearest neighbors using the Nearest-Neighbor-Descent algorithm and then minimizing local distances of the approximate topology using stochastic gradient descent [33]. When compared to t-Distributed Stochastic Neighbor Embedding (t-SNE), UMAP has been observed to be faster [14] with clearer separation of groups.

Due to compute limitations in fitting the entire high dimensional vector of nearly 5.5M records, we randomly sampled one million records. We created an embedding of the vectors along two components to fit the UMAP model with the Hellinger metric which compares distances between probability distributions, as follows:

$$h(P, Q) = \frac{1}{\sqrt{2}} \cdot \left\| \left(\sqrt{P} - \sqrt{Q} \right) \right\|_2$$

We visualized the word vectors with their respective labels, which were the assigned topics generated from the LDA model. We used the default parameters of n_neighbors = 15 and min_dist = 0.1. Figure 5 presents the visualization of the TF-IDF word vectors for each of the 1 million tweets with their labeled topics. UMAP is supposed to preserve local and global structure of data, unlike t-SNE that separates groups but does not preserve global structure. As a result, UMAP visualizations intend to allow the reader to interpret distances between groups as meaningful. In Figure 5 each topic is color-coded by its respective topic.

The UMAP plots appear to provide further evidence of the quality and number of topics generated. Our observations is that many of these topic "clusters" appear to have a single dominant color indicating distinct grouping. There is strong local clustering for topics that were also prominent in the keyword analysis and topic modeling time series plots. A very distinct and separated mass of purple tweets represents the "100: N/A" topic which is an undefined topic. This means that the LDA model outputted equal scores across all 20 topics for any single tweet. As a result, we could not assign a topic to these tweets because they all had uniform scores. But this visualization informs us that the contents of these tweets were uniquely distinct from the others. Examples of tweets in this "100: N/A" cateogry include "See, #Democrats are always guilty of whatever", "Why are people still getting in cruise ships?!?", "Thank you Mike you are always helping others and sponsoring Anchors media shows.", "We cannot let this woman's brave and courageous actions go to waste! #ChinaLiedPeopleDied #Chinaneedstopay", "I wish people in this country would just stay the hell home instead of GOING TO THE BEACH". Other observations reveal that the mask-related topic 10 in purple, and potentially a combination of 8 and 9 in red are distinct from the mass of noisy topics in the center of the plot. We can also see distinct separation of aqua-colored topic 18 "potus" and potentially topics 5 and 6 in yellow.

We refer the reader to other examples where UMAP has been leveraged for Twitter analysis, to include Darwish et al. [16] for identifying clusters of Twitter users with controversial topic similarity, Vargas [57] for event detection, political polarization by Darwish et al. [16] and Stefanov for estimating political leaning of users by [48]. Future steps for this study include evaluating other dimensionality reduction techniques to include t-SNE such as the works of [4, 10, 15] and Principal Component Analysis (PCA) such as [26, 53] to discover feature correlation and localization.

6 TIME-TO-RETWEET ANALYSIS

A highly retweeted tweet might signal that an issue has attracted attention in the highly competitive Twitter environment, and may give insight about issues that resonate with the public [37]. We extracted metadata from our corpora for the Tweet, User, and Entities objects. Due to compute limitations, we selected a sample that consisted of 736,561 tweets that included retweets from the corpora of March 24 - 28, 2020. However, since we were only focused on retweets, out of the corpus of 736,561 tweets, we reduced it to 567,909 (77%) that were only retweets. We used the corpus of retweets and analyzed the time between the tweet created_at and the retweeted created_at.

$$time\_to\_rt = rt\_object - tw\_object$$

Here, the rt_object is the datetime in UTC format for when the message that was retweeted was originally posted. The tw_object is the datetime in UTC format when the current tweet was posted. This measures the time it took for the author of the current tweet to retweet the originating message.

Wang et al. [58] calls this "response time" and used it to measure response efficiency and speed of information dissemination during Hurricane Sandy. Wang analyzed 986,579 tweets and found that 67% of re-tweets occur within 1 h [58]. We also tried to identify the speed of retweeting in disasters and emergencies. For example, Earle [17] reported 19 seconds was the time it took to retweet following an earthquake. Kuang et al. [28] similarly defined response time of the retweet to be the time difference between the time of the first retweet and that of the origin tweet. Further, Spiro et al. [47] calls these "waiting times". The median time-to-retweet for our corpus was 2.87 hours meaning that half of the tweets occurred within this time (less than what Wang reported as 1.0 hour), and the mean was 12.3 hours. Figure 15 shows the histogram of the number of tweets by their time to retweet in seconds.

Further, we found that compared to the 2013 Avian Influenza outbreak (H7N9) in China described by Zhang et al. [63] Covid19 retweeters sent more messages earlier than H7N9. Zhang analyzed the log distribution of 61,024 H7N9-related posts during April 2013 and plotted reposting time of messages on Sina Weibo, a Chinese Twitter-like platform and one of the largest microblogging sites in China Figure 9. Zhang found that H7N9 reposting occurred with a median time of 222 minutes (i.e. 3.7 hours) and a mean of 8520 minutes (i.e. 142 hours). Compared to Zhang's study, we found our median retweet time to be 2.87 hours, about 50 minutes faster than the reposting time during H7N9 of 3.7 hours. When comparing Figure 8 and Figure 9, it appears that Covid19 retweeting does not completely slow down until 2.78 hours later (10^4 seconds). Whereas, for H7N9 it appears to slow down much earlier by 10 seconds. This may be a sign of the sustained global intensity of the volume and transmission of high volumes of Covid-related information across social media, when compared to H7N9 which may be more localized to a limited geography.

Table 3: Statistics about Each Network Community

Graphs	Ranking Speed	Time Point	Density	Nodes	1st	2nd	3rd
G1	1	19 sec	0.000428	1278	11	11	9
G2	2	328 sec (5.47 min)	0.000449	1248	17	8	8
G3	3	591 sec (9.85 min)	0.000450	1247	13	12	9
G4	4	885.6 sec (14.76 min)	0.000460	1234	17	10	10
G5	5	3600 sec (60 min)	0.000567	1110	41	27	20
G6	6	10000 sec (2.78 hrs)	0.000538	1139	18	15	15
G7	7	13,320 sec (3.7 hrs)	0.000540	1138	17	17	11
G8	8	86,400 sec (24 hrs)	0.000685	1005	63	43	26
G9	9	604,800 sec (1 week)	0.000598	1067	92	9	9

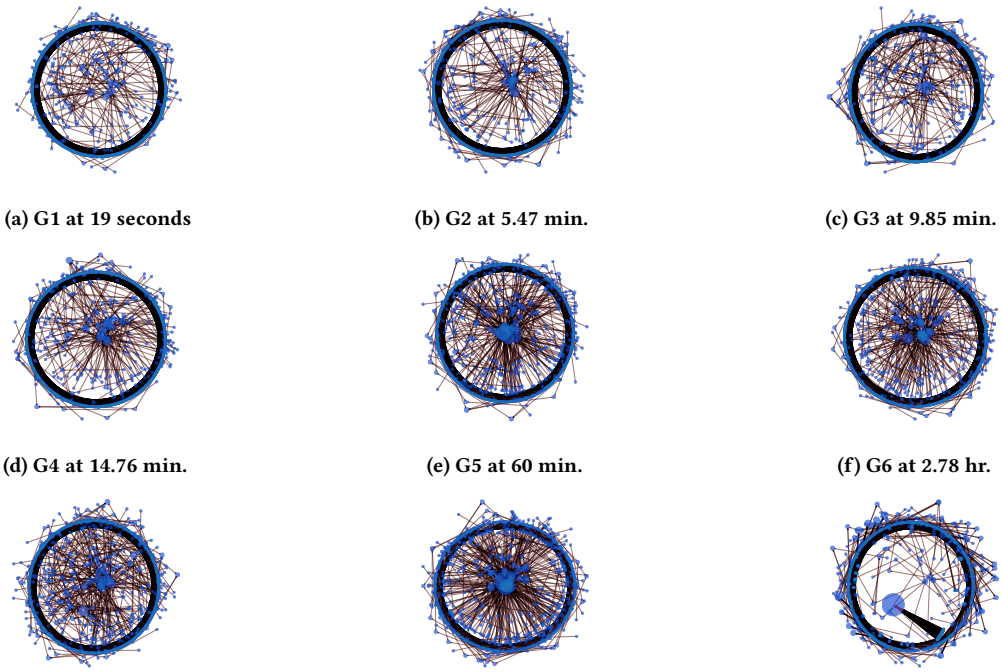
Unfortunately few studies appear to document retweeting times during infectious disease outbreaks which made it hard to compare how Covid19 retweeting behavior against similar situations. Further, the H7N9 outbreak in China occurred seven years ago and may not be a comparable set of data for numerous reasons. Chinese social media may not represent similar behaviors with American Twitter and this analysis does not take into account multiple factors that imply retweeting behavior to include the context, the user's position, and the time the tweet was posted [37].

6.1 TF-IDF Message and User Description Features of Rapid Retweeters

We also analyzed what rapid retweeters, or those retweeting messages even faster than the median, in less than 10,000 seconds were saying. In Figure 20 we plotted the top 50 TF-IDF features by their scores for the text of the retweets. It is intuitive to see that URLs are being retweeted quickly by the presence of "https" in the body of the retweeted text. This is also consistent with studies by Suh et al. [49] who indicated that tweets with URLs were a significant factor impacting retweetability. We found terms that were frequently mentioned during the early-stage keyword analysis and topic modeling mentioned again: "cases", "ventilators", "hospitals", "deaths", "masks", "test", "american", "cuomo", "york", "president", "china", and "news". When analyzing the descriptions of the users who were retweeted in Figure 20, we ran the TF-IDF vectorizer on bigrams in order to elicit more interpretable terms. User accounts whose tweets were rapidly retweeted, appeared to describe themselves as political, news-related, or some form of social media account, all of which are difficult to verify as real or fake.

7 NETWORK MODELING

We analyzed the network dynamics of nine different time periods within the March 24 - 28, 2020 Covid19 dataset, and visualized them based on their speed of retweeting. These types of graphs have been referred to as "retweet cascades" which describes how a social media network propagates information [22]. Similar methods have been applied for visualizing rumor propogation by Jin et al. [22] We wanted to analyze how Covid19 retweeting behaves at different time points. We used median retweeting times of disasters



(a) G1 at 19 seconds (b) G2 at 5.47 min. (c) G3 at 9.85 min.

(d) G4 at 14.76 min. (e) G5 at 60 min. (f) G6 at 2.78 hr.

(g) G7 at 3.7 hr. (h) G8 at 24 hr. (i) G9 at 168 hr.

Figure 7: Directed Graphs of Covid19 Retweeting Activity at Nine Different Points in Time (G1 - G9) between March 24 - March 28th using the Kamada Kawai Layout

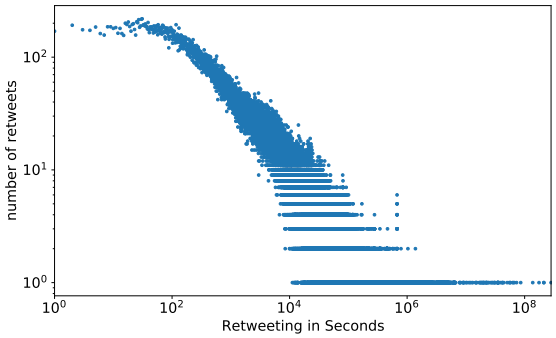


Figure 8: Log Distribution of Covid19 Retweets from March 24 - 28, 2020

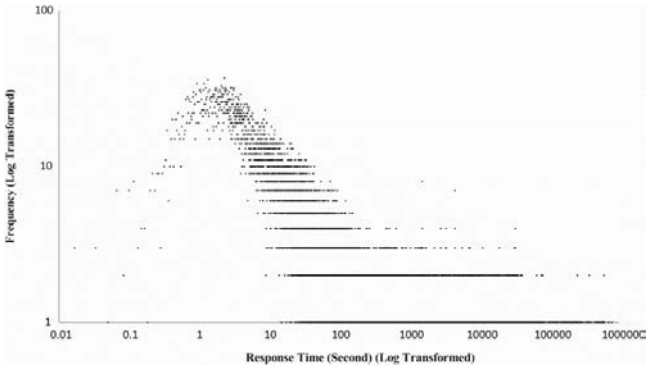


Figure 9: Log Distribution of H7N9-related messages on Sina Weibo, March 2013

and emergencies as a benchmark. These include Spiro et al. [47] for the time it took users to retweet messages based on hazardous keywords like "Funnel Cloud", "Aftershock", and "Mudslide". We also used the H7N9 reposting time which Zhang et al. [63] published of 3.7 hours.

We generated a Directed Graph for each of the nine time periods, where the network consisted of a source which was the author of the tweet (User object, the id_str) and a target which was the original retweeter shown in Table 3. The goal was to analyze how connections change as the retweeting speed increases. The nine networks are visualized in Figure 7. Graphs were plotted using networkx and drawn using the Kamada Kawai Layout[23], a force-directed algorithm. We modeled 700 users for each graph. We found that more nodes became too difficult to interpret. The size of the node indicates the number of degrees, or users that it is connected to. It means that the node has been retweeted by others several times.

The density of each network increases over time shown in Figure 7 and Figure 10 that shows the time point, density, number

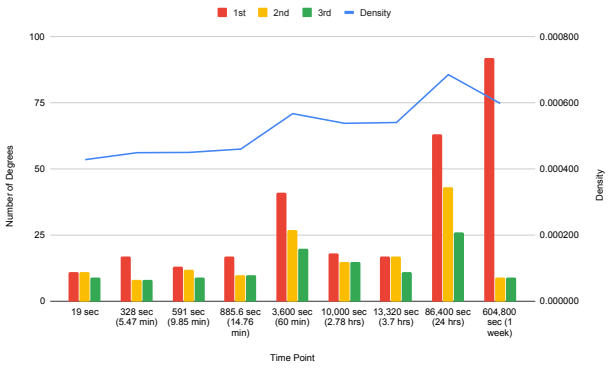


Figure 10: Increasing Density and Degree for Top 3 Users

of nodes, and number of degrees for the first, second, and third top user accounts in each network. Very rapid retweeters, in the time it takes to retweet after an earthquake, start off with a sparse network with a few nodes in the center being the focus of retweets in Figure 7a. By the time we reach Figure 7d, the retweeted users are much more clustered in the center and there are more connections and activity. The top retweeted user in our median time network Figure 7g, was a news network and tweeted *"The team took less than a week to take the ventilator from the drawing board to working prototype, so that it can"*. By 24 hours out in Figure 7h, we see a concentrated set of users being retweeted and by Figure 7i, one account appears to dominate the space being retweeted 92 times. This account was retweeting the following message several times *"She was doing #chemotherapy couldn't leave the house because of the threat of #coronavirus so her line sisters..."*. In addition, the number of nodes generally decreased from 1278 in "earthquake" time to 1067 in one week, and the density also generally increased, shown in Table 3.

These retweet cascade graphs provide only an exploratory analysis. Network structures like these have been used to predict virality of messages, for example memes over time as the message is diffused across networks [59]. But, analyzing them further could enable 1) an improved understanding about how Covid19 information diffusion is different than other outbreaks, or global events, 2) How information is transmitted differently from region to region across the world, and 3) What users and messages are being concentrated on over time. This would support strategies to improve government communications, emergency messaging, dispelling medical rumors, and tailoring public health announcements.

8 LIMITATIONS

There are several limitations with this study. First, our dataset is discontinuous and trends seen in Figure 3 and Figure 4 where there is an interruption in time should be taken with caution. Although there appears to be a trend between one discrete time and another, without the missing data, it is impossible to confirm this as a trend. Next, the corpus we analyzed was already pre-filtered with thirteen "track" terms from the Twitter Streaming API that focused the dataset towards healthcare related concerns. This may be the reason why the high level keywords extracted in the first round of analysis were consistently mentioned throughout the different stages of modeling.

Third, the users and conversations in Twitter are not a direct representation of the U.S. or global population. The Pew Research Foundation found that only 22% of American adults use Twitter [39] and that this group is different from the majority of U.S. adults, because they are on average younger, more likely to identify as Democrats, more highly educated and possess higher incomes [60]. The users were also not verified and should be considered as a possible mixture of human and bot accounts. Fourth, we reduced our corpus to remove retweets for the keyword and topic modeling analyses since retweets can obscure the message by introducing virality and altering the perception of the information [32]. As a result, this reduced the size of our corpus by nearly 77% from 23,820,322 tweets to 5,506,223 tweets. However, there appears to be variability in terms of consistent corpora sizes in the Twitter analysis literature both in Table 4 and other health-related studies [2, 20, 24, 30, 50, 64].

Fifth, our compute limitations prohibited us from analyzing a larger corpus for the UMAP, time-series, and network modeling. For the LDA models we leveraged the gensim MulticoreLDA model that allowed us to leverage multiprocessing across 20 workers. But for UMAP and the network modeling, we were constrained to use a CPU. Applying our methods across the entire 23.8 million corpora for UMAP and the network models may yield more meaningful results. Sixth, we were only able to iterate over 15 different LDA models based on changing the number of topics, whereas Syed et al. [51] iterated on 480 models to select coherent models. We believe that applying a manual gridsearch of the LDA parameters such as iterations, alpha, gamma threshold, chunksize, and number of passes would lead to a more diverse representation of LDA models and possibly more coherent topics. This study could also be strengthened by implementing spatio-temporal modeling of LDA topics by distinct geographies such as U.S. cities and states which might reveal insights about retweeting behavior and local topics of interest. For example, Cheng et al. [11] conducted event detection by analyzing clusters of topics using LDA by the "geo" Twitter metatag for tweets in London between January 14 and 18, 2013 for a helicopter crash disaster. He was able to plot hourly and daily clusters of topic activity on a map of London, offering insight about sustained or fleeting interest of the topics to the local population.

Seven, it was challenging to identify papers that analyzed Twitter networks according to their speed of retweets for public health emergencies and disease outbreaks similar to our methods. Zhang et al. [63] points out that there are not enough studies of temporal measurement of public response to health emergencies. We were lucky to find papers by Zhang et al. [63] and Spiro et al. [47] who published on disaster waiting times. Although other researchers have published a variety of studies on Twitter during public health emergencies [12, 52, 54], it was difficult to compare our results directly with other disease outbreaks for retweet cascade times and network models.

9 CONCLUSION

We answered five research questions about Covid19 tweets during March 24, 2020 - April 8, 2020. First, we found high-level trends that could be inferred from keyword analysis. Second, we found that live White House Coronavirus Briefings led to spikes in Topic 18 ("potus"). Third, using UMAP, we found strong local "clustering" of topics representing PPE, healthcare workers, and government concerns. Fourth, we used retweets to calculate the speed of retweeting. We found that the median retweeting time was 2.87 hours. Fifth, using directed graphs we plotted the networks of Covid19 retweeting communities from rapid to longer retweeting times.

ACKNOWLEDGMENT

The authors would like to acknowledge John Larson and Steve Escaravage from Booz Allen Hamilton for their support and review of this article.

REFERENCES

[1] Alshaabi, T., Minot, J., Arnold, M., Adams, J. L., Dewhurst, D. R., Reagan, A. J., Muhamad, R., Danforth, C. M., and Dodds, P. S. How the world's collective attention is being paid to a pandemic: Covid-19 related 1-gram time series for 24 languages on twitter. *arXiv preprint arXiv:2003.12614* (2020).

[2] Alvarez-Melis, D., and Saveski, M. Topic modeling in twitter: Aggregating tweets by conversations. In *Tenth international AAAI conference on web and social media* (2016).

[3] Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., and Chowell, G. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688* (2020).

[4] Banerjee, I., Madhavan, S., Goldman, R. E., and Rubin, D. L. Intelligent word embeddings of free-text radiology reports. In *AMIA Annual Symposium Proceedings* (2017), vol. 2017, American Medical Informatics Association, p. 411.

[5] Barnaghi, P., Ghaffari, P., and Breslin, J. G. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)* (2016), IEEE, pp. 52–57.

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022.

[7] Buntain, C., Golbeck, J., Liu, B., and LaFree, G. Evaluating public response to the boston marathon bombing and other acts of terrorism through twitter. In *Tenth International AAAI Conference on Web and Social Media* (2016).

[8] Chatfield, A., and Brajawidagda, U. Twitter tsunami early warning network: a social network analysis of twitter information flows.

[9] Chen, E., Lerman, K., and Ferrara, E. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372* (2020).

[10] Chen, Y., Yuan, J., You, Q., and Luo, J. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th ACM international conference on Multimedia* (2018), pp. 117–125.

[11] Cheng, T., and Wicks, T. Event detection using twitter: a spatio-temporal approach. *PloS one 9*, 6 (2014), e97807.

[12] Chew, C., and Eysenbach, G. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one 5*, 11 (2010).

[13] Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004* (2020).

[14] Coenen, A., and Pearce, A. Understanding umap.

[15] Dai, X., Bikdash, M., and Meyer, B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017* (2017), IEEE, pp. 1–7.

[16] Darwish, K., Stefanov, P., Aupetit, M. J., and Nakov, P. Unsupervised user stance detection on twitter. *arXiv preprint arXiv:1904.02000* (2019).

[17] Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters 81*, 2 (2010), 246–251.

[18] Earle, P. S., Bowden, D. C., and Guy, M. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics 54*, 6 (2012).

[19] Hong, L., Dan, O., and Davison, B. D. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web* (2011), pp. 57–58.

[20] Hong, L., and Davison, B. D. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (2010), pp. 80–88.

[21] Jahanbin, K., and Rahmanian, V. Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine* (2020), 13.

[22] Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., and Luo, J. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (2017), Springer, pp. 14–24.

[23] Kamada, T., Kawai, S., et al. An algorithm for drawing general undirected graphs. *Information processing letters 31*, 1 (1989), 7–15.

[24] Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., and Shaw Jr, G. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management 38*, 1 (2018), 1–6.

[25] Killick, R., Fearnhead, P., and Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*, 500 (2012), 1590–1598.

[26] Kondor, D., Csabai, I., Dobos, L., Szüle, J., Barankai, N., Hanyecz, T., Sebők, T., Kallus, Z., and Vattay, G. Using robust pca to estimate regional characteristics of language use from geo-tagged twitter messages. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* (2013), IEEE, pp. 393–398.

[27] Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., and Baddour, K. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus 12*, 3 (2020).

[28] Kuang, L., Tang, X., and Guo, K. Predicting the times of retweeting in microblogs. *Mathematical Problems in Engineering 2014* (2014).

[29] Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., and Choudhary, A. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops* (2011), IEEE, pp. 251–258.

[30] Lim, K. W., Chen, C., and Buntine, W. Twitter-network topic model: A full bayesian treatment for social network and text modeling. *arXiv preprint arXiv:1609.06791* (2016).

[31] Lopez, C. E., Vasu, M., and Gallemore, C. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359* (2020).

[32] Madrigal, A. C. Retweets are trash, Mar 2018.

[33] McInnes, L. How umap works¶.

[34] McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[35] Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., and Lehmann, C. U. An" infodemic": Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak. *medRxiv* (2020).

[36] Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., and Brownstein, J. S. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research 16*, 10 (2014), e236.

[37] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference* (2011), pp. 1–7.

[38] Odlum, M., and Yoon, S. What can we learn about the ebola outbreak from tweets? *American journal of infection control 43*, 6 (2015), 563–571.

[39] Perrin, A., and Anderson, M. Share of u.s. adults using social media, including facebook, is mostly unchanged since 2018, Apr 2019.

[40] Ramos, J., et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (2003), vol. 242, Piscataway, NJ, pp. 133–142.

[41] Rehurek, R. Lda model parameters, Jul 2014.

[42] Řehůřek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 2010), ELRA, pp. 45–50. http://is.muni.cz/publication/884893/en.

[43] Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (2015), pp. 399–408.

[44] Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., and Zannettou, S. " go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19. *arXiv preprint arXiv:2004.04046* (2020).

[45] Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., and Liu, Y. Coronavirus on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309* (2020).

[46] Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., and Wang, Y. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907* (2020).

[47] Spiro, E., Irvine, C., DuBois, C., and Butts, C. Waiting for a retweet: modeling waiting times in information propagation. In *2012 NIPS workshop of social networks and social media conference. http://snap. stanford. edu/social2012/papers/spiro-dubois-butts. pdf. Accessed* (2012), vol. 12.

[48] Stefanov, P., Darwish, K., and Nakov, P. Predicting the topical stance of media and popular twitter users. *arXiv preprint arXiv:1907.01260* (2019).

[49] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing* (2010), IEEE, pp. 177–184.

[50] Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., and Dunn, A. G. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. *Journal of medical Internet research 18*, 8 (2016), e232.

[51] Syed, S., and Spruit, M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)* (2017), IEEE, pp. 165–174.

[52] Szomszor, M., Kostkova, P., and De Quincey, E. # swineflu: Twitter predicts swine flu outbreak in 2009. In *International conference on electronic healthcare* (2010), Springer, pp. 18–26.

[53] Tan, L., Zhang, H., Clarke, C., and Smucker, M. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (2015), pp. 657–661.

[54] Tang, L., Bie, B., and Zhi, D. Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease. *American journal of infection control 46*, 12 (2018), 1375–1380.

[55] Thelwall, M., and Thelwall, S. Retweeting for covid-19: Consensus building, information sharing, dissent, and lockdown life. *arXiv preprint arXiv:2004.02793* (2020).

[56] Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing 167* (2020), 107299.

[57] Vargas-Calderón, V., Camargo, J. E., Vinck-Posada, H., et al. Event detection in colombian security twitter news using fine-grained latent topic analysis. *arXiv preprint arXiv:1911.08370* (2019).

[58] Wang, B., and Zhuang, J. Crisis information distribution on twitter: a content analysis of tweets during hurricane sandy. *Natural hazards 89*, 1 (2017), 161–181.

[59] Weng, L., Menczer, F., and Ahn, Y.-Y. Virality prediction and community structure in social networks. *Scientific reports 3* (2013), 2522.

[60] Wojcik, S., and Hughes, A. How twitter users compare to the general public, Jan 2020.

[61] Yang, K.-C., Torres-Lugo, C., and Menczer, F. Prevalence of low-credibility information on twitter during the covid-19 outbreak. *arXiv preprint arXiv:2004.14484* (2020).

[62] Yasin Kabir, M., and Madria, S. Coronavis: A real-time covid-19 tweets analyzer. *arXiv* (2020), arXiv–2004.

[63] Zhang, L., Xu, L., and Zhang, W. Social media as amplification station: factors that influence the speed of online public response to health emergencies. *Asian Journal of Communication 27*, 3 (2017), 322–338.

[64] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (2011), Springer, pp. 338–349.

A COVID19 TWITTER PAPERS

Table 4: Papers published on Covid19 Twitter Analysis since January 2020

Author	Number Tweets	Time Period	Keywords	Feature Analysis	Geospatial	Topic Modeling	Sentiment	Transmission	Network Models	UMAP
Jahanbin [21], et al.	364,080	Dec. 31 2019 - Feb. 6 2020			x					
Banda, et al.[3]	30,990,645	Jan. 1 - Apr 4, 2020	x							
Medford, et al. [35]	126,049	Jan. 14 - Jan. 28, 2020	x	x		x	x			
Singh, et al.[46]	2,792,513	Jan. 16, 2020 - Mar. 15, 2020	x	x	x	x				
Lopez, et al. [31]	6,468,526	Jan. 22 - Mar. 13, 2020	x	x	x					
Cinelli, et al. [13]	1,187,482	Jan. 27 - Feb. 14, 2020		x		x		x		
Kouzy, et al. [27]	673	Feb 27, 2020	x	x						
Alshaabi, et al. [1]	Unknown	Mar. 1 - Mar 21, 2020	x	x						
Sharma, et al. [45]	30,800,000	Mar. 1, 2020 - Mar. 30, 2020	x	x	x	x	x	x	x	
Chen, et al. [9]	8,919,411	Mar. 5, 2020 - Mar. 12, 2020	x							
Schild [44]	222,212,841	Nov. 1, 2019 - Mar. 22, 2020	x	x		x			x	
Yang, et al.[61]	Unknown	Mar. 9, 2020 - Mar. 29, 2020	x						x	
Ours	**23,830,322**	**Mar. 24 - Apr. 9, 2020**	**x**	**x**		**x**			**x**	**x**
Yasin-Kabir, et al.[62]	100,000,000	Mar. 5, 2020 - Apr. 24, 2020	x	x	x		x			

B TWITTER DATASET IN UTC TIME

Table 5: Twitter Data Sets March 24, 2020 - April 8, 2020

Corpus	Time Start	Time End	Total Minutes	Size, GB	Total Tweets	No Retweets	Perc No Retweets
3/24/2020	2020-03-24 21:17:27+00:00	2020-03-24 22:00:48+00:00	44	1	132,658	27,374	20.64%
3/25/2020	2020-03-25 14:45:12+00:00	2020-03-25 16:18:47+00:00	94	2	286,405	63,649	22.22%
3/28/2020	2020-03-28 00:17:20+00:00	2020-03-28 02:01:08+00:00	105	2.3	317,498	61,933	19.51%
3/30/2020	2020-03-30 12:55:38+00:00	2020-03-30 21:44:35+00:00	530	11.5	1,618,620	365,808	22.60%
3/31/2020	2020-03-30 21:47:53+00:00	2020-03-31 13:15:36+00:00	929	20.3	2,802,069	576,741	20.58%
4/4/2020	2020-04-03 00:29:11+00:00	2020-04-04 22:05:12+00:00	2737	56.2	7,755,704	1,795,912	23.16%
4/5/2020	2020-04-05 20:41:43+00:00	2020-04-07 15:07:11+00:00	2547	49.4	6,810,216	1,599,455	23.49%
4/8/2020	2020-04-08 13:54:33+0000	2020-04-09 14:30:54+0000	1477	30.4	4,107,152	1,015,351	24.72%
Total			**8463**	**173.1**	**23,830,322**	**5,506,223**	**23.11%**

Table 6: Keyword Raw Counts

Corpus	bed	hospital	mask	icu	help	nurse	doctors	vent	test_pos	serious_cond	exposure	cough	fever
3/24/2020	147	1,323	1,685	139	114	215	372	1,143	28	1	11	18	1
3/25/2020	293	3,104	3,641	265	299	352	758	2,321	122	4	26	35	10
3/28/2020	191	3,218	3,607	180	248	504	891	4,073	101	2	19	19	3
3/30/2020	1,475	21,707	28,512	1,225	1,742	3,708	6,895	13,190	589	13	114	157	23
3/31/2020	1,959	28,495	67,703	1,344	3,416	5,233	9,671	16,717	948	13	141	459	137
4/2/2020	5,652	80,495	231,185	4,034	8,661	16,823	28,603	64,112	2,228	48	525	977	122
4/5/2020	5,648	81,025	159,915	6,350	7,741	14,767	27,341	45,614	2,612	36	445	786	133
Total	15,365	219,367	496,248	13,537	22,221	41,602	74,531	147,170	6,628	117	1,281	688	429

C TOPIC MODELING IMPLEMENTATION DETAILS

For the LDA topic modeling, we used the gensim Python library [41, 42]. It provides four different coherence metrics. We used the "c_v" metric for coherence developed by Roder[43]. Coherence metrics are used to rate the quality and human interpretability of a topic generated. All models were run with the default parameters using a LdaMulticore model parallel computing on 20 workers, default gamma threshhold of 0.001, chunksize of 10,000, 100 iterations, 2 passes.

D LIVE PRESS BRIEFINGS AND TOPIC TIME SERIES

Note - Sudden decreases in Figure 12 signal may be due to temporary internet disconnection.

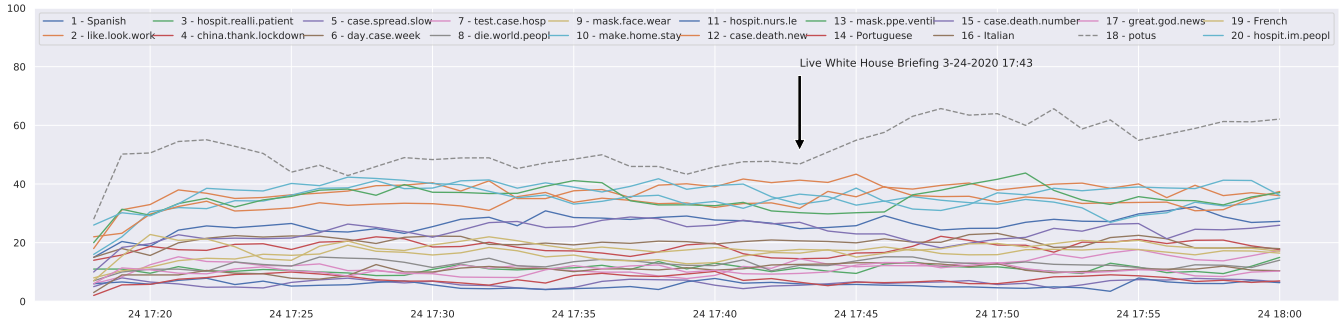


Figure 11: March 24 5:17 PM to 6:00 PM EST Topics Time Series

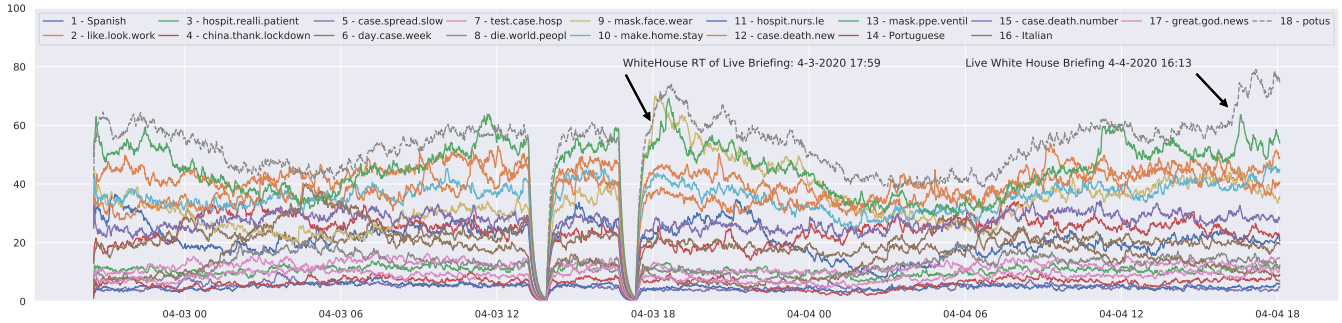


Figure 12: April 3 8:29 PM EST to April 4 6:05 PM EST Topics Time Series

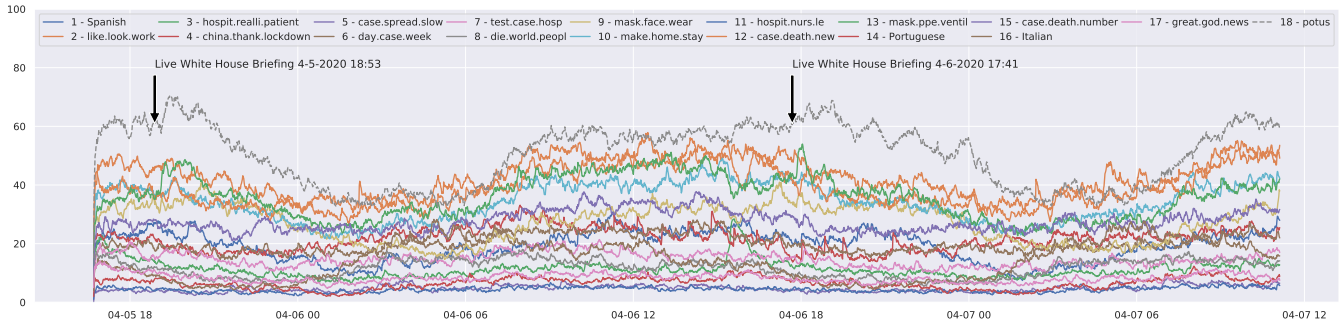


Figure 13: April 5 4:41 PM EST to April 7 11:07 AM EST Topics Time Series

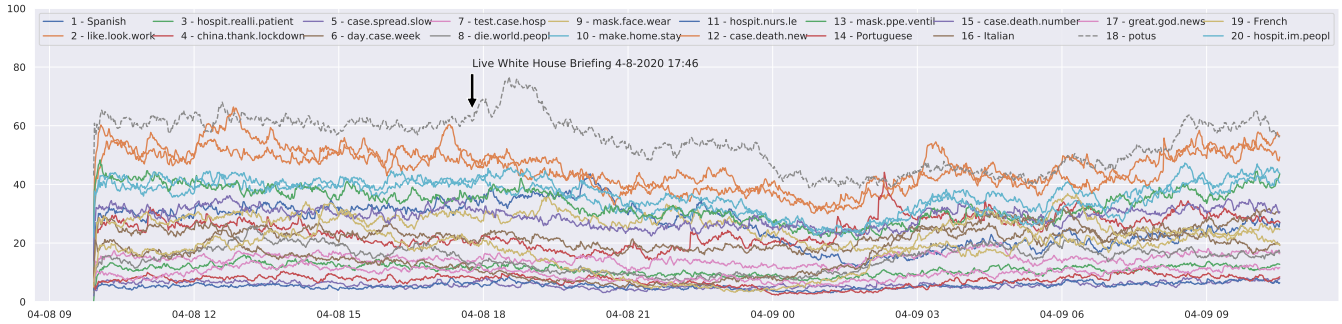


Figure 14: April 8 9:54 AM EST to April 9 10:30 AM EST Topics Time Series

E SECONDS TO RETWEET, MARCH 24 - 28TH CORPORA

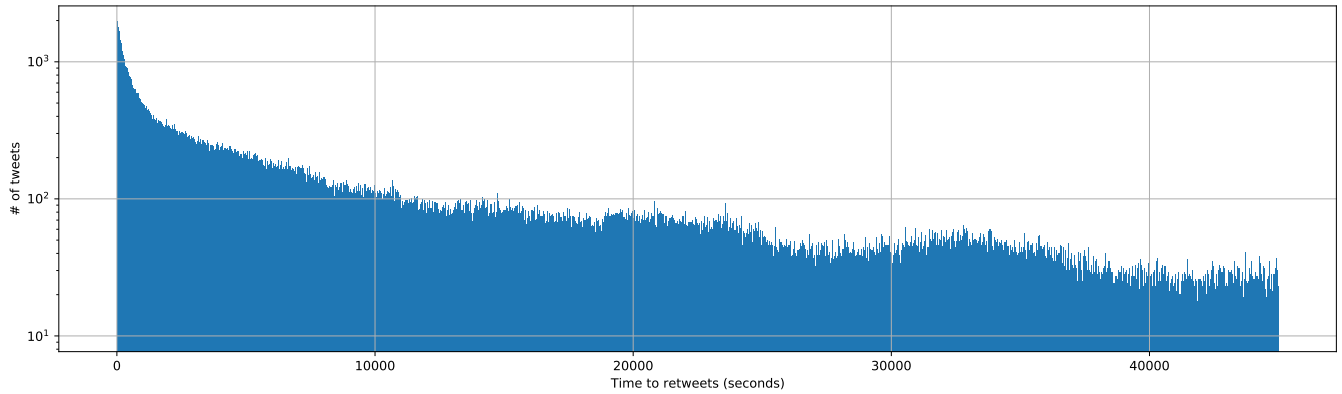


Figure 15: Seconds to Retweet, March 24 - 28th Corpora

F CHANGE POINT DETECTION TIME SERIES

Models were calculated using the ruptures Python package. We also applied exponential weighted moving average using the ewm pandas function. We applied a span of 5 for March 24, 2020 and a span of 20 for April 3 - 4 datasets, April 5 - 6 datasets, and April 8 - 9 datasets. Our parameters for binary segmentation included selecting the "l2" model to fit the points for Topic 18, using 10 n_bkps (breakpoints).

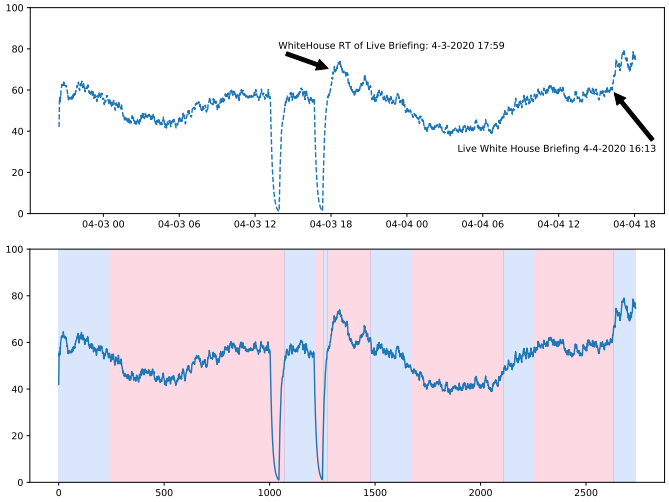


Figure 16: Change Point Detection using Binary Segmentation for April 3 - 4, 2020

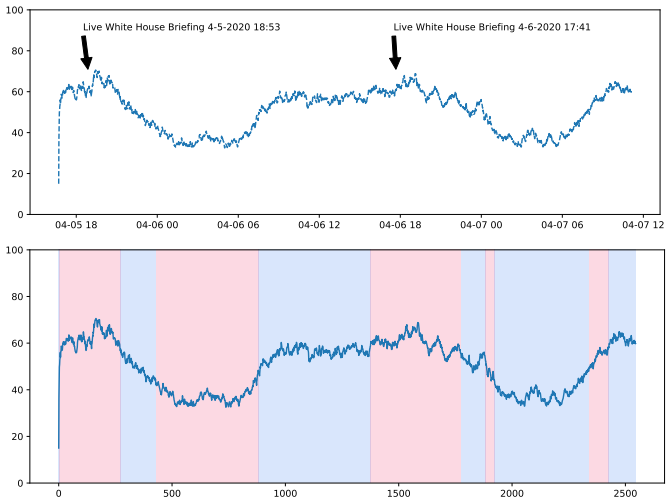


Figure 17: Change Point Detection using Binary Segmentation for April 5, 2020

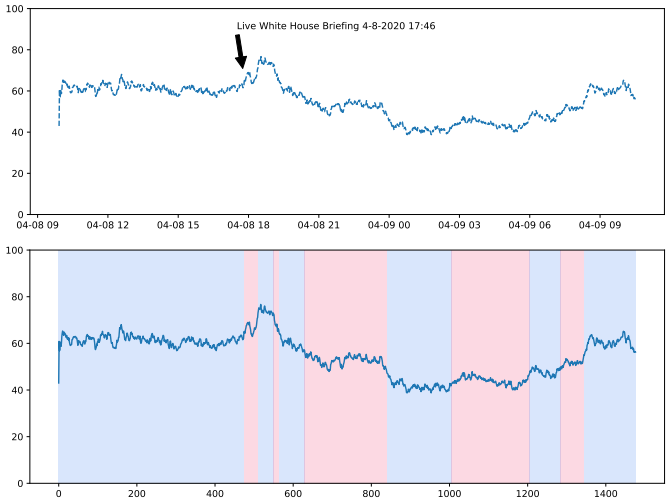


Figure 18: Change Point Detection using Binary Segmentation for April 8, 2020

G TF-IDF FREQUENCIES OF TWEETS RAPIDLY RETWEETED

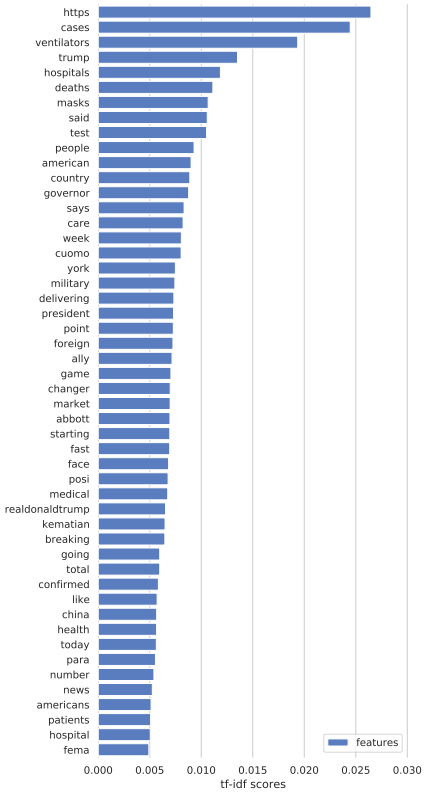


Figure 19: TF-IDF Scores for Rapidly Retweeted Messages

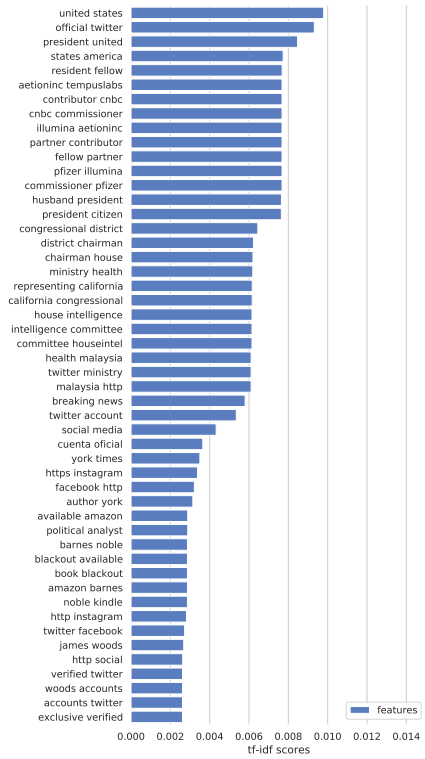


Figure 20: TF-IDF Scores for Descriptions of Retweeted Users

CovEx: An Exploratory Search System for COVID-19 Scientific Literature

Behnam Rahdari
behnam.r@pitt.edu
University of Pittsburgh
Pittsburgh, PA

Peter Brusilovsky
peterb@pitt.edu
University of Pittsburgh
Pittsburgh, PA

Khushboo Thaker
k.thaker@pitt.edu
University of Pittsburgh
Pittsburgh, PA

Hung Kim Chau
hkc6@pitt.edu
University of Pittsburgh
Pittsburgh, PA

ABSTRACT

This paper presents our attempt to create an exploratory search system CovEx for a collection of academic papers related to COVID-19. CovEx uses concept extraction, knowledge graphs, and user-controlled recommendation to assist users with various levels of domain expertise in their information needs.

CCS CONCEPTS

• **Information systems → Graph-based database models**; **Web searching and information discovery**.

KEYWORDS

COVID-19, Knowledge Graph, Exploratory Search Systems

ACM Reference Format:
Behnam Rahdari, Peter Brusilovsky, Khushboo Thaker, and Hung Kim Chau. 2020. CovEx: An Exploratory Search System for COVID-19 Scientific Literature. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 4 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

Exploratory search systems form an increasingly popular category of information access and exploration tools. These systems creatively combined search, browsing, and information analysis steps shifting user efforts from recall (formulating a query) to recognition (i.e.,selecting a link) and helping them to gradually learn more about the explored domain [23]. In this paper we presenting our attempt to augment the set of search systems focused on COVID-19 research literature [25] with a personalized exploratory search system COVID Explorer (CovEx [1]). We hope that CovEx ability to support information discovery, learning-while-searching, and personalization, the system could help a broader set of users to benefit from the assembled collection of COVID-19 resources [22].

We start the paper with the presentation of CovEx interface and follow with the details on concept extraction, knowledge graph organization, and recommendation that enable the work of this interface.

2 RELATED WORK

The *CovEx* system presented in this paper combines the ideas of exploratory search with an important stream of research on personalization, user control, and transparency, It attempts to help researchers discover their *interest profiles* [10], which, in turn, are used to find relevant publications with matching concepts.

2.1 Exploratory Search

A number of real-life search tasks require a considerable amount of learning during the search process to achieve adequate results. These tasks are known as *exploratory search* tasks [4, 16]. Since simple search systems are usually not efficient in supporting exploratory search, a range of advanced exploratory systems have been developed and evaluated [13, 24]. More recently, few projects in this area demonstrated that the effectiveness of exploratory search could be improved by using a personalized system, which builds a profile of user interests and adapts to the individual user [5, 11, 19]. The work presented in this paper investigates the ideas of profile-based exploratory search in the context of finding research publications related to Covid-19 pandemic.

2.2 Controllability

User controllability has been recognized as a valuable component of advanced information access interfaces. This research was made popular by a stream of work on user controllable recommender systems [14, 17]. However the value of extended user control has been also demonstrated in the area of exploratory search. For example, NameSieve [1] presented a summary of search results in the form of entity clouds, which a controllable filtering and exploration of results. PeopleExplorer [12] offered users an option to re-sort people search results based on multiple user-related factors. uRank [8] introduced a controllable interface for refining and reorganizing search results. An extension of this work [7] integrated a controllable *social search* into an exploratory search system.

[1] http://scythian.exp.sis.pitt.edu/covex/

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

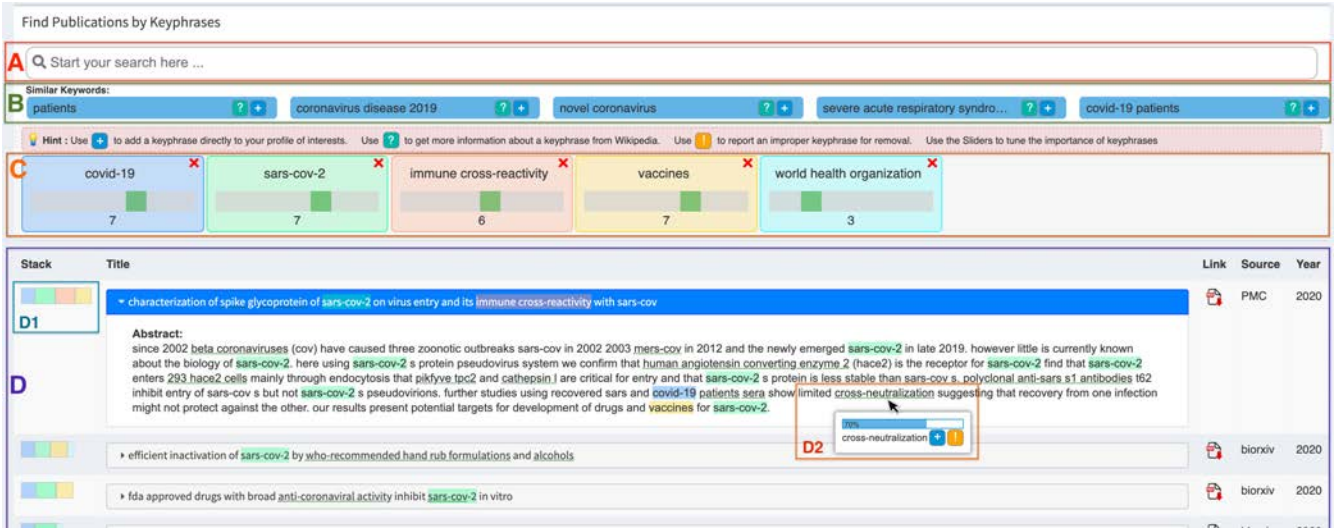


Figure 1: Interface Design of Covex representing different parts of the system.

2.3 Open User Profile

The idea to apply open user profiles (also known as open user models) to better support personalized information access was among the early ideas explored in this field. Open user profiles allow users to examine and possibly change the content of their interest profiles, which are used to personalize their search or browsing process. Since the open user profiles increase interactivity, transparency, and controllability of the information exploration process, their application was a good match to the nature of exploratory search. While first attempts to introduce “bag-of-words" open user profiles had mixed success [2], more recent work focused on semantic level user profiles demonstrated its potential for personalized exploratory search [5, 19, 20].

3 THE INTERFACE OF COVEX

Personalized information exploration in CovEx is centered around user interest profile[18] - a collection of keyphrases (keywords) that express user search interests. Unlike traditional search that requires users to specify all keyphrases in a query, CovEx supports users in the process of gradual discovery and refinement of their interests. It also allows the users to control the importance of each keyphrase in recommending relevant results. CovEx interface consists of the following main sections.

Instant Search Box. The search box (Figure 1A) is the gateway to the system. Using an instant search approach, it allows users to discover relevant topics without a fully formulated query. When a user starts typing a query, a series of frequent similar keywords appears, which helps the user to discover a range of matching topics (e.g., cell culture and infected cells). When an item is selected from the list, it will automatically added to the slider area (Figure 1B). at the same time, an updated list of search results will be presented to the user.

Similar Keywords. When at least one keyword is added to the user’s profile, a series of five semantically similar topics appear in the *Similar Keywords* area of the interface (Figure 1B). Users can add recommended keywords to their interest profiles by clicking on the plus button to the right of each keyword. As the user’s profile grows and refines, the set of recommended keywords is updated since the system recommends instances similar to all keywords in the user’s profile. Each recommended keyword also provides users with a short description of the topic. Clicking on the question mark button next to the add button, opens up a separate window containing the abstract of that keyword’s Wikipedia entry. This information is crucial when the user is not familiar with the recommended keyword and needs more knowledge to decide whether the keyword must be added to the interest profile.

Slider Area. The slider area (Figure 1C) displays the current interest profile of the user. CovEx implements a content-based recommendation approach, which generates the list of recommended results (Figure 1D) using the interest profile. To support transparency and controllability of this process, the interest profile is visible and directly editable by the end users. To build the profile the user can add relevant topics as explained above as well as remove less relevant keywords (using the red x) as they discover more relevant topics or explore different interests. Sliders associated with each keyword enable users to control the relative importance of a topic compared to others in their profile, ranging from 1 (least important) to 10 (most important). The use of sliders for fine-tuning of user profile was motivated by keyword tuning approach in uRank design [9], which was confirmed as a user-friendly and efficient in an exploratory search context. The initial value of the sliders is set to five but can be changed at any time. All actions within the profile (adding, removing, or adjusting sliders) immediately affects the search results list.

Search Results. As soon as the user adds the first keyword to the interest profile, a table of the 20 most relevant publications is generated (Figure 1:D). The first column of the table visualizes the combined relevance between keyphrases in the user interest profile

and each result. The colors in the stacked-bar (Figure 1:D1) are matched with the color of slider in the profile and the size and opacity of each bar expresses the relevance of the result to each profile keyphrase.The second column of table lists the titles of relevant publications. Clicking on each title expands a window that holds the abstract of the paper. The mentioned keyphrases are highlighted with corresponding colors. The opacity of the colors reflect the relevance of a keyphrase to the paper and the current value of slider for that keyphrase. To further assist the users, CovEx underlines all available keyphrases in the text (both in title and abstract). Hovering over the underlined portion of the text opens a popup window (Figure 1:D2) that enable user to (1) see the relevance of the keyphrase to the text in a form of a vertical bar-chart, (2) add the keyphrase directly to the interest profile, and (3) report the improper keyphrases to the administrator for removal. The latter helps us to improve the quality of extracted keywords and eliminate the occasional errors in the process of extraction. Finally, last three columns provide a link to the content of the paper, source and year of publication.

4 THE KNOWLEDGE GRAPH

The knowledge graph consists of three main entities - publications, authors, keyphrases and their relationships - extracted from our data set and hosted in a native graph database Neo4j[2]. Figure 2 presents the schematic representation of the knowledge graph. Authors are interconnected by the relation *Co-Author* (based on co-authorship) and connected to papers by the relation *Published*.Papers connected to keyphrases usign the *Has-Key* relationship. The latter carries a weight that determines the strength of the relationship between each keyphrase and the publication.

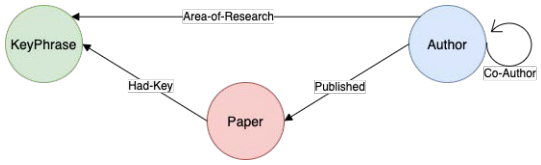


Figure 2: Graph Schema representing the entities of the knowledge graph and the relationship between them

4.1 Data Source and Graph Statistics

We used COVID-19 Open Research Dataset Challenge (CORD-19)[3] as the main source of data to build the knowledge graph and extract the keyphrases. The dataset contains 51078 document, out of which 48251 documents contain either title or abstract.

Using this dataset and the concept extraction explained below, we generated the knowledge graph covering 48251 publications related to COVID-19 research that have been authored by 157589 researchers. 211862 keyphrases were extracted from titles and abstracts of these publications. Table 1 shows the basic statistics of our knowledge graph.

[2]https://en.wikipedia.org/wiki/Neo4j
[3]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Labels	No. Nodes	Avg. Properties	Avg. Relations
[Keyword]	211862	3	11.02
[Paper]	48251	12	12.91
[Author]	157589	1	3.65

Table 1: Graph Statistics

4.2 Keyphrase Extraction and Weighting

We approach the keyphrase extraction problem as a sequence labeling task. We apply a Bi-LSTM-CRF architecture to perform this task, which has been shown to achieve the best performance across several public datasets [3]. The standard Bi-LSTM-CRF model consists of three main components, the Embedding layer, the Bi-LSTM layer and the CRF layer. Our implementation of the model is based on the version presented in [15][4]. We obtain the character embeddings of 30 dimensions by training additional Bi-LSTM networks along with the main model. We use the Glove pre-trained word embeddings of 100-dimensions[5]. A 300-dimension hidden layer of LSTM units is used for both the character-level embedding model and the main model. The models are trained using mini-batch stochastic gradient descent with momentum. The batch size, learning rate and decay ratio are set to 10, 0.015 and 0.05, respectively. We also apply dropout to avoid over-fitting and gradient clipping of 5.0 to increase the model's stability.

We train the model with the GENIA dataset[6]: includes 2000 titles and abstracts of scientific articles from Medline database. GENIA is an fully annotated dataset, in which the annotated technical terms cover the identification of physical biological entities (e.g., proteins, cell types) as well as other important terms. We randomly select 300 articles for evaluating and our model achieves 82% of F1-score.

To assign weight for each keyphrase extracted from the the document we found the distance of the keyphrase from the document in embedding space [6]. For training the embedding for concepts extracted from CORD documents we utilized keyphrase embedding [21] and trained the embedding with context extracted from CORD dataset. EmbedRank [6] is used to assign weight to each keyphrase based on the cosine similarity between keyphrase embedding and document embedding.

5 PROFILE-BASED SEARCH

We deploy a two-phase search process to produce the most relevant results based on user interest profile. In the first phase a primary list of candidate have been selected from the graph and the second phase assure that the results are presented to the user in the right order based on their relevancy to the query. We describe these to phases in more details in the following:

Candidate selection. We used the Cypher Querying Language to generate the initial list of candidate publications. At each instance of user interaction with the system (e.g., adding/removing keywords or tuning the sliders), the system considers all publications connected to at least one of the topics of interest in the user profile.

[4]github.com/LiyuanLucasLiu/LM-LSTM-CRF
[5]https://nlp.stanford.edu/projects/glove/
[6]http://www.geniaproject.org/genia-corpus/term-corpus

If the number of candidates are less that 20, the system uses similar keyphrases to populate the candidate list. The process of finding similar keyphrases is explained below.

Reordering the results. After generating the list of candidate results, the system rearranges the results in a way that the most relevant results appear at the top of the list.In order to do that, first a complete list of keyphrases that appear in the text (title and abstract) of each publication, alongside with their relevancy score (weight) is being generated. Then for every keyphrase that exist in the user interest profile, we multiplied it's weight with the value of corresponding slider. Finally, the relevance score is assigned to each candidate considering candidate's similarity to each of profile topics and the value of the sliders (Equation 1).

$$RelevanceScore_{(f,A)} = \sum_{i=0}^{|A|} Sim_{(a_i,f)} * w_i \quad (1)$$

1: Calculation of relevance score for each candidate publication

In equation 1, A is a set of tuples $\{(a_1, w_1), (a_2, w_2), ...(a_n, w_n)\}$ that represent the current state of the user's profile (topics and weights) and f is a given publication in the graph. a_i and w_i correspond for i^{th} keyword and its slider value at the moment. $Sim_{(a_i,f)}$ shows the value of relevance between a given keyword and a candidate publication in our knowledge graph that has been described in section 4.2

Keyphrase Recommendations. To generate recommended keywords for the current set of keywords in the interest profile, the system generates two sets of candidate keywords using the co-occurrence of seed keyphrase with publications and authors (using collaborative filtering. Then, the system combines the number of co-occurred keyphrases in both sets and uses it as a ranking mechanism. The system presents the top five results to the user.

6 EXPERIENCE AND FUTURE WORK

CovEx system has been deployed online and also demonstrated to several target users. The early results indicate that the success of the system to a considerable extent depends on the quality of keyphrase extraction. Moreover, the nature of exploratory search calls for special extraction approaches. While we used a relatively powerful approach, it was trained to model gold standard annotation of individual documents in GENIA dataset. We believe, however, that keyphrase extraction has to consider the collection as a whole increasing user chances to discover keyphrases that could lead to other papers. We are interested to collaborate with experts on keyphrase extraction to develop approaches optimized for exploratory search.

REFERENCES

[1] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Radu Florian. 2010. Semantic annotation based exploratory search for information analysts. *Information processing & management* 46, 4 (2010), 383–402.

[2] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. 2007. Open user profiles for adaptive news systems: help or harm?. In *the 16th international conference on World Wide Web, WWW '07*. ACM, 11–20.

[3] Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. 2019. Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. In *The World Wide Web Conference*. 2551–2557.

[4] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[5] Fedor Bakalov, Birgitta König-Ries, Andreas Nauerz, and Martin Welsch. 2010. IntrospectiveViews: An Interface for Scrutinizing Semantic User Models. In *18th International Conference on User Modeling, Adaptation, and Personalization*. Springer, 219–230.

[6] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 221–229.

[7] Cecilia di Sciascio, Peter Brusilovsky, and Eduardo Veas. 2018. A study on user-controllable social exploratory search. In *23rd International Conference on Intelligent User Interfaces*. ACM, 353–364.

[8] Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank As You Go: User-Driven Exploration of Search Results. In *the 21st International Conference on Intelligent User Interfaces (IUI'16)*. ACM, 118–129.

[9] Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank as you go: User-driven exploration of search results. In *21st International Conference on Intelligent User Interfaces*. 118–129.

[10] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. 2007. User Profiles for Personalized Information Access. In *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer-Verlag, Berlin Heidelberg New York, 54–89.

[11] Amanda Gonçalves Dias, Evangelos E. Milios, and Maria Cristina Ferreira de Oliveira. 2019. TRIVIR: A Visualization System to Support Document Retrieval with High Recall. In *ACM Symposium on Document Engineering*. Article 10.

[12] Shuguang Han, Daqing He, Jiepu Jiang, and Zhen Yue. 2013. Supporting exploratory people search: a study of factor transparency and user control. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 449–458.

[13] Tingting Jiang. 2014. Exploratory search: a critical analysis of the theoretical foundations, system features, and research trends. In *Library and Information Sciences: Trends and Research*. Springer, Berlin, Heidelberg, 79–103.

[14] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and Control in Social Recommenders. In *6th ACM Conference on Recommender Systems*. 43–50.

[15] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank F. Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *32nd AAAI Conference on Artificial Intelligence*. 5253–5260.

[16] Garry Marchionini. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[17] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM, 1085–1088. http://dx.doi.org/10.1145/1357054.1357222

[18] Behnam Rahdari, Peter Brusilovsky, and Dmitriy Babichenko. 2020. Personalizing Information Exploration with an Open User Model. In *31st ACM Conference on Hypertext and Social Media (HT '20)*. Association for Computing Machinery, New York, NY, USA, 0. https://doi.org/10.1145/3372923.3404797

[19] Tuukka Ruotsalo, Kumaripaba Athukorala, Dorota Glowacka, Ksenia Konyushkova, Antti Oulasvirta, Samuli Kaipiainen, Samuel Kaski, and Giulio Jacucci. 2013. Supporting exploratory search tasks with interactive user modeling. In *2013 Annual Meeting of American Society for Information Science and Technology*, Vol. 50. Wiley, 1–10.

[20] Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2019. Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement. *Journal of the Association for Information Science and Technology* (2019), In Press.

[21] Khushboo Maulikmihir Thaker, Peter Brusilovsky, and Daqing He. 2018. Concept Enhanced Content Representation for Linking Educational Resources. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence*. 413–420.

[22] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. arXiv:cs.IR/2005.04474

[23] Ryen W White, Bill Kules, Steven M Drucker, et al. 2006. Supporting exploratory search. *Commun. ACM* 49, 4 (2006), 36–39.

[24] R. W. White and R. A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan and Claypool.

[25] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125* (2020).

A Canine Census to Influence Public Policy

Matias Apa, Maria Cecilia Faini
{matias_apa,mcfaini}@yahoo.com.ar
Facultad de Ciencias Veterinarias
Universidad Nacional de Rosario
Casilda, Santa Fe, Argentina

Mohammad Aliannejadi
m.aliannejadi@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Maria Soledad Pera
solepera@boisestate.edu
People and Information Research
Team (PIReT), Boise State University
Boise, Idaho, USA

ABSTRACT

The potential threat that domestic animals pose to the health of human populations tends to be overlooked. We posit that positive steps forward can be made in this area, via suitable state-wide public policy. In this paper, we describe the data collection process that took place in Casilda (a city in Argentina), in the context of a canine census. We outline preliminary findings emerging from the data, based on a number of perspectives, along with implications of these findings in terms of informing public policy.

CCS CONCEPTS

• **Mathematics of computing** → **Exploratory data analysis**; • **Applied computing** → *Life and medical sciences.*

KEYWORDS

Census, public policy, visualization, canine, epidemiology

ACM Reference Format:
Matias Apa, Maria Cecilia Faini, Mohammad Aliannejadi, and Maria Soledad Pera. 2020. A Canine Census to Influence Public Policy. In *epiDAMIK 2020: 3rd epiDAMIK SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 4 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

Using the epidemiology of urban health as a lens, we can study the environment and context of a region to understand (i) the ties and relationships of species among themselves and with the environment, (ii) the complexity of the urban context, and (iii) the consequences that result from these complex interactions and the social determinants of health [9]. Ecosystems and human health are deep-rooted on biological processes that are socially defined. The fact that social mandates influence health-related determinations posits a dialectical perspective to explore "social-biological" and "society-nature" interactions, both of which contribute towards the phenomenology of health. The transformation patterns observed between society and the environment are continually evolving; yet, social determinations are hierarchically imposed and are the ones that most prominently prevail in nature [4].

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

Domestic animals are a clear example of social constructs prevailing over biological ones. In their case, dynamics within the population are defined by social norms and standards, as well as political and cultural practices. The number of animals in a given region depends on the availability of resources (food, water, shelter) and human acceptance of the particular population. This is the reason why canine ecology is deeply interconnected with human-related activities [7]. Ongoing development of regions results in changes in habits and behavior of their inhabitants, such as the increase in the number of companion animals that are now part of households, especially dogs and cats. The bond between humans and companion animals has both positive and negative effects on health. Examples of the latter are zoonoses, animal bites, and pollution. It is worth noting that all the concerns become more critical when these animals have access to public roads [8].

Policy and campaigning messages to promote a healthy human-animal coexistence depend on a better understanding of the companion animals' social placement and dynamics in a region. This can be achieved by collecting representative data and analyzing the demographic characteristics of animal populations, local traits, and natural human-animal interactions [3]. Our study focuses on the data collected in Casilda (Santa Fe), Argentina. For decades, the local community has demanded that the city council address concerns related to dog ownership and welfare. In fact, the city council introduced an ordinance concerning the canine census in 2008 [5]. However, there has been a long delay before we conducted the first census in 2018 due to a lack of study protocols.

In this paper, we present the results of Casilda's first-ever domestic canine census. Doing so involved an interdisciplinary group of faculty and students in epidemiology, public health, ethics, and legislation for veterinary sciences, statistics, and computer science, who were mindful of the economic and political constraints inherent of the region and designed a protocol for data collection and conducted the associated analysis. The main objectives guiding our work include:

- Providing an exhaustive description of Casilda's domestic canine population. To do so, we conducted an empirical exploration of census data; this revealed scientific indicators that allow assessment of the quality of life of canines, along with the quality of their interactions with humans.
- Identifying new problems that emerge from the surveyed canine population can be addressed by new public policies, thus responding to the demands of the region and its inhabitants [1].

Our work lays the grounds for future work in this area by introducing the necessary protocols that could be used in similar studies. Moreover, it aids the local government in making informed decisions in response to the existing canine-human problems.

2 PROTOCOL FOR DATA COLLECTION

We collected census data and conducted systematic probabilistic sampling by areas. In establishing these areas, we considered different traits (social, environmental, and economic) that characterize Casilda (population 37,441). This resulted in the 3 geographical areas (Figure 1): **Area 1** ($5.22km^2$), upper/upper middle class; **Area 2** ($3.82km^2$), middle class; **Area 3** ($1.50km^2$), working/lower class.

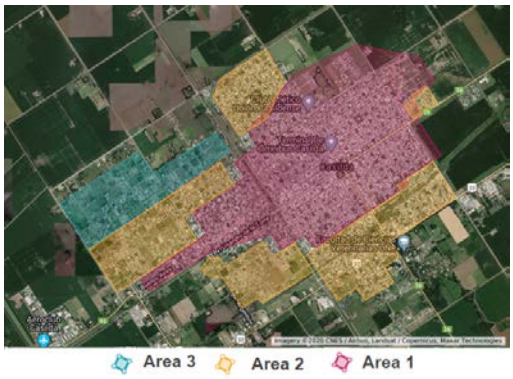


Figure 1: Geographical areas considered in the census.

Data collection took place in June 2018; involving a team of 80 students and 18 faculty in the Epidemiology Department at Universidad Nacional de Rosario [1]. The surveyed area included 60 blocks (1,189 households that resulted in 486 voluntary responses). The team reported a general low predisposition on behalf of household occupants in taking part in the census. This rendered the sample insufficient for statistical inference. To address this limitation, the team sub-sampled 125 new households to survey [2].

For data collection, the team created a dynamic form with response-depended questions using Google Forms. The questionnaire included 26 questions (Table 1), some closed-ended and others multiple-choice; grouped by data related to *households*, *household occupants*, *canines*, *responsible ownership*, and *general*. Google Forms was chosen as it is a free tool that eases immediate digitization of the collected data, reducing operational costs and the use of paper – these are constraints that influenced data collection decisions, given that resources at public universities in Argentina are scarce.

3 ANALYSIS AND DISCUSSION

Below we summarize general observations that emerged from collected data; these are meant to offer context of the geographical area and human and canine populations considered in the census. Thereafter, we present detailed findings from census data, along with their implications for public policy.

3.1 A General Description of the Population

Based on collected responses, we analyze the data of 841 dogs, uniformly distributed across gender. We summarize sanitary conditions and sterilization in Table 2. Other insights include:

- **Breed**: 33% were pure-breeds, the rest mongrels.

[1]Training for data collection is part of the curriculum for one of the epidemiology-related classes offered at Universidad Nacional de Rosario [6].

Table 1: Questionnaire used for data collection purposes.

Type	ID	Question
Household	1	Area
	2	Address
	3	Household type
	4	Services (e.g., gas, water, etc.)
Household occupants	5	Are they in and willing to answer questionnaire?
	6	How many people live in the household?
Canines (repeated for each dog in household)	7	Breed
	8	Gender
	9	Age
	10	Size
	11	Origin (e.g., adopted, found, etc.)
	12	Sterilized?
	13	Where did the sterilization take place?
	14	If not, why not?
Responsible ownership	15	Where does your dog live? (patio, indoors, etc.)
	16	How often in your dog on public roads?
	17	If veterinary services are required, where do you go?
	18	How often do you deworm your dog? (internally)
	19	How often do you deworm your dog (externally)
	20	In the last year, have you vaccinated your dog for rabies? Where?
	21	In the last year, have you vaccinated your dog for Leptospirosis? Where?
General	22	Are there any other animals in the household. Elaborate.
	23	In the last year, have you experienced any of the following: bites, dog involved in accident, chasing bicycles and/or people walking, saw canine mistreatment, etc.
	24	Do you know your neighbourhood's health center?
	25	Regarding your neighbourhood's health center
	26	For your own health-related matters, where do you go?

- **Size**: close to 50% were small breeds (e.g., Beagle, Poodle Toy), 33 % medium (e.g., French bulldog), and the remaining, larger breeds (e.g., Golden retriever).
- **Origin**: 74% were either adopted or found, 20% were purchased, and for the remaining ones, survey respondents did not recall.
- **Age**: 498 were adults between 1 and 7 years old, 154 were puppies (i.e., less than 12 months), and the remaining 189 were seniors (i.e., 8 years or more).
- **Area**: 233 in Area 1, 401 in Area 2, and 207 in Area 3.
- **Inference from sampling**: 13,557 dogs in households, 4,863 strays [6].

Table 2: Overview of surveyed canine population

Canines	Total	Male	Female
Surveyed	841	422	419
Sterilized	318	67	251
Internal deworming	692	344	348
External deworming	728	364	364
Rabies vaccination	440	219	221
Leptospirosis vaccination	299	137	162

3.2 Findings and Implications

To further characterize Casilda's domestic canine population, and more importantly, identify issues directly related to this population, we further examined census responses from various perspectives.

3.2.1 Sterilization. We see a statistical significant correlation between gender and age, when it comes to sterilization (Chi-square: 24.85; p-value= 5.38e-05). As reported in Table 2, close to 40% of

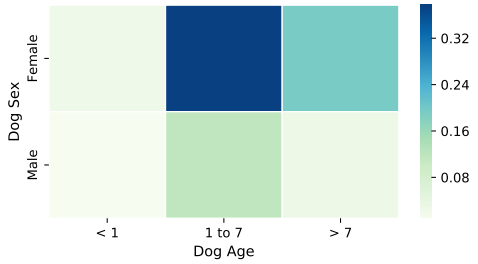


Figure 2: Correlation with respect to sterilization.

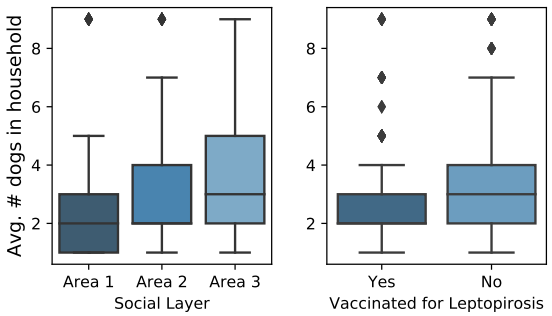
domestic canines have been sterilized; for the most part, females. It is also apparent from Figure 2 that sterilization rarely occurs on canines less than 12 months old; the majority of sterilizations happening on adult specimens (i.e., aged 1 to 7). As for why owners bypass sterilization (question 14 in Table 1), close to 30% *"do not think it is necessary"* and 3.4% *"disagrees with the premise of sterilization"*. It is of note that 13% of the owners *"plan sterilization in the future"* and 1.3% have yet to do so *"due to economic impediments"*.

Female sterilization is a positive discovery, especially when considering that it occurs at an age range that correlates with the highest fertility peaks. Unfortunately, lack of sterilization in males counteracts intended population control. Further, the high proportion of unsterilized males is a definite concern that must be addressed. Their social behavior entails wandering and territoriality, often resulting in dog fights, bites of people, the transmission of diseases, and traffic accidents. Owners' views against sterilization reflect that population control policy must be thought of as a comprehensive scheme. The system must ensure the economic and geographical accessibility to an operating room. It must also include educational strategies that raise awareness of the negative consequences of non-sterilization.

3.2.2 Sanitary Conditions. We examine the degree of influence, or lack thereof, that the number of dogs per household has on traits related to responsible ownership practises (questions 18-21 in Table 1). We find that *deworming* (internal or external) and *vaccinations* for *rabies* are not conditioned by the number of dogs in a household. There is a statistical significant correlation between *vaccination* for *Leptospirosis* and number of dogs per households, where more dogs implies a higher likelihood of overlooking this type of vaccination (ANOVA, p-value = 0.001; Figure 3b).

These results show the broad access that the local population has to the rabies vaccine. In Argentina, Law No. 22953 establishes this vaccine as mandatory. The city sponsors free vaccination campaigns, together with the application of dewormers. Further, deworming is a low-cost procedure when performed at private veterinary clinics. On the other hand, the Leptospirosis vaccine is not part of sponsored campaigns, and Leptospirosis vaccination at private clinics is very expensive. Therefore, state policy responding to this concern should include targeted campaigns on high-risk areas.

3.2.3 Socio-economic Influence. When using socio-economic factors as lenses to drive exploration, census data reveals a correlation between geographic areas and number of dogs per households (Figure 3a Chi-square: 22.87; p-value= 0.00013). Upon deeper inspection,



(a) Dogs per household **(b) Vaccination per household**

Figure 3: Average number of dogs per household distributed by areas (a) and the ones vaccinated for Leptospirosis (b).

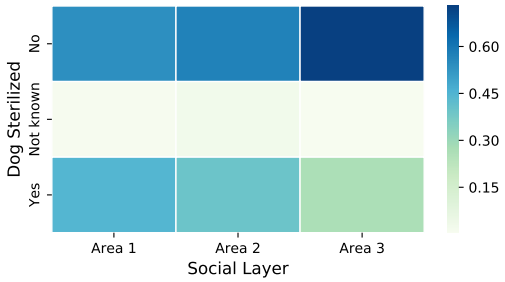


Figure 4: Canine sterilization across areas. "Not known" refer to cases when the dog owner was not present to answer.

we see that the highest percentage of unsterilized dogs come from households in Area 3, whereas most sterilized dogs come from households in Area 1 (Figure 4). Based on Pearson's correlation among area and reasons given by household owners to justify they do not favor sterilziation (Figure 5), we see that the reason that yields the highest correlation for households in the least affluent area (i.e., Area 3) is *"Lack of time"*, followed by *"I will do so in the future"*. On the other spectrum, households in more affluent areas (i.e., Area 1 and Area 2) justify not sterilizing their dogs since they *"Lives inside"* and *"Would like to breed in the future"*, respectively.

The results above evidence the fact that low-income regions should be the focus of attention for public policy related to responsible pet ownership. In these regions, it is imperative to ensure economic and geographical assistance by performing State-sponsored (i.e., free) sterilization in peripheral neighborhoods. Despite the fact that lack of time is not an impediment for sterilization in Areas 1 and 2, sterilization rates are not 100% in these areas (Figure 4); on Area 3, lack of time is the main issue hindering sterilization procedures. These findings could suggest that education and awareness campaigns would be more effective in Areas 1 and 2, whereas those mentioned earlier sponsored and geographically-targeted sterilization procedures could be more effective for Area 3.

3.2.4 Humans. Canines with frequent access to public roads pose a risk to human health. We identified 687 dogs that have access

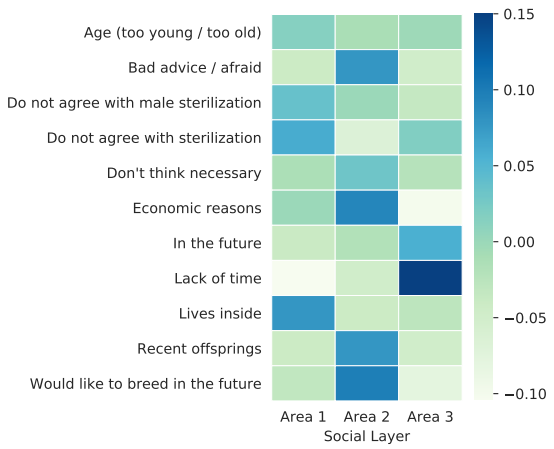


Figure 5: Pearson's correlation for area with respect to reasons why household owners chose to avoid sterilization.

to public roads; 362 males, 325 females (question 16 in Table 1). As reported in Table 3, only 50% of these dogs are vaccinated for rabies–a low percentage when considering that this vaccination is mandatory. The percentage decreases even further for Leptospirosis (∼ 30%). Compared to vaccinations, the percentage of frequently-dewormed dogs with access to public roads is much higher (∼ 75%).

The high proportion of dogs with access to public roads is a threat to public health. Because of unvaccinated dogs, the risk of exposure to diseases increases. Leptospirosis is an endemic zoonotic disease in Casilda. Thus actions by the State to address the low vaccination coverage are a must. Rabies-related concerns are much more worrisome: given that in addition to being a lethal zoonosis, there is evidence of the circulation of this virus in Casilda, vaccinations rates reach 100%. On the upside, the high proportion of dewormed dogs is positive for health care, as it prevents disease spread to other dogs and humans, which can be done via contaminated dog feces or ticks, to name a few.

3.2.5 Overpopulation. Dogs with access to public roads may cause an unexpected increase in canine populations: specimens that have not been sterilized, yet have access to public roads are bound to become a link in a chain of unplanned litters. As shown in Table 3, 60% of females with access to public roads are sterilized, a percentage that drastically decreases among males (∼ 14%).

When campaigns fostering sterilization are not prominent, dog population growth rates remain high. Given the significant proportion of unsterilized females with access to public roads, compounded by the very high percentage of unsterilized males, breeding likelihood is high. That is why sterilization mechanism should be intensified, with a greater emphasis on males and social sectors with economic difficulties (i.e., low-income areas). These actions should be supplemented with an educational policy that emphasizes the importance of long-term behavior change regarding responsible pet ownership, specifically adopting new habits that foster health care for dogs and their environment.

4 CONCLUSIONS

We have presented the analysis results we conducted on data collected in response to a domestic canine population census.

Table 3: Canine population that has access (on its own, leashed and/or unleashed) to public roads (n=687), based on vaccination, deworming, and sterilization perspectives.

		Canine Population		
		Total	Male	Female
Vaccination	Rabies	0.52	0.49	0.54
	Leptospirosis	0.33	0.30	0.35
Deworming	Internal	0.74	0.72	0.76
	External	0.75	0.74	0.75
Sterilization		0.36	0.14	0.60

Outcomes from our empirical exploration reveal representative traits of Casilda's canine population, which till now were unavailable. We were also able to recognize potential risks originated from the population under study, mainly the transmission of zoonosis and uncontrolled breeding. At the same time, we identified geographic areas and social stratum that should be of primary concern to the city council when it comes to implementing immediate actions regarding sterilization, improvement of sanitary conditions, and education related to responsible pet ownership. This study serves as preliminary evidence on the importance of generating information on canine demography and its link with humans and the environment at the national level. An adapted version of the proposed data collection/analysis protocol – based on lessons learned and limitations we observed – could be included as part of the national population census, which takes place every ten years.

ACKNOWLEDGMENTS

We appreciate Ion Madrazo Azpiazu's feedback on data collection and Federico Abud's work on statistical inference.

REFERENCES

[1] M. Apa, G. Uranga, M. Gay, A. Alfieri, M. Lopez Hiriart, P. Cucchiari, D. Federici, E. Perazo, D. Frati, F. Guzman, L. Bittel, C. Dieguez, N. Quaglia, F. Abud, and M.C. Faini. 2019. Censo canino de la ciudad de Casilda. Año 2018. VII Jornada Latinoamericana. V Jornadas de Ciencia y Tecnología. Facultad de Ciencias Agrarias. IV Reunión Transdisciplinaria en Ciencias Agropecuarias, UNR.

[2] F. Azorín and J.L. Sánchez-Crespo. 1994. *Métodos y aplicaciones del muestreo.* Alianza Madrid.

[3] M Bovisio, MC Fracuelli, B González, O Lencinas, N Mestres, A Varela, and E Marcos. 2004. Características de la convivencia humano-animal en la ciudad de Buenos Aires y su relación con la prevención de zoonosis. *Trabajo original. Instituto de Zoonosis Luis Pasteur* (2004).

[4] J Breilh. 2010. La epidemiología crítica: una nueva forma de mirar la salud en el espacio urbano. *Salud colectiva* 6 (2010), 83–101.

[5] Casilda City Council. 2008. Ordenanza Nnumero 1669. http://concejocasilda.com.ar/digesto/doc01113.pdf.

[6] M.C. Faini, G. Green, F. Abud, D. Frati, F. Guzmán, D. Sisofo, A. Alfieri, and M. Apa. 2018. Diseno de un estudio para la caracterización de la población de perros de Casilda. Libro de resumenes de la XX Congreso XXXVIII Reunión Anual de la Sociedad de Biología de Rosario 2018.

[7] OIE World Organisation for Animal Health. 2010. Stray Dog Population Control. In *Terrestrial Animal Health Code.* Chapter 7.7, 382–396. https://www.oie.int/index.php?id=169&L=0&htmfile=chapitre_aw_stray_dog.htm.

[8] A Loza. 2014. Caracterización de la población canina y felina en Santa Cruz de la Sierra. *Concurso de Ciencias Agropecuarias y Salud Animal. Santa Cruz de la Sierra: Universidad Autónoma "Gabriel Rene Moreno"* (2014), 57.

[9] J. Spiegel, J. Breilh, and A. Yassi. 2015. Why language matters: insights and challenges in applying a social determination of health approach in a North-South collaborative research program. *Globalization and health* 11, 1 (2015), 9.

A Deep Learning Approach for COVID-19 Trend Prediction

Tong Yang*
Department of Physics
Boston College
Chestnut Hill, Massachusetts, USA

Long Sha*
Department of Computer Science
Brandeis University
Waltham, Massachusetts, USA

Justin Li
Del Norte High School
San Diego, California, USA

Pengyu Hong
Department of Computer Science
Brandeis University
Waltham, Massachusetts, USA

ABSTRACT

The ongoing COVID-19 pandemic, due to the novel coronavirus SARS-CoV-2, has affected not only the healthcare system but the whole society worldwide. While a large number of medical works and researchers are battling the pandemic crisis on the front line, with large amount of accessible epidemic information, data-driven research and learning based approaches could provide rich insights about the challenge on the population and society level. In this work, we apply a recurrent-network based model to study the epidemic data in the United States. By incorporating both the epidemic time series and socioeconomic characteristic data, our model provides both a promising predictive power in forecasting the trend of new confirmed cases, and an illustrative description about the interplay between the local epidemic evolution and demographic features.

ACM Reference Format:
Tong Yang, Long Sha, Justin Li, and Pengyu Hong. 2020. A Deep Learning Approach for COVID-19 Trend Prediction. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 7 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

In late 2019, the COVID-19 outbreak initially detected and reported in Wuhan (Hubei, China) due to the *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), spread rapidly, firstly across regions in China and east-Asian countries, and then, since late February, to nearly all continents in the world. As of June 15, 2020, there have been more than 7.91 million cases confirmed across 225 countries and regions, associated with 433 thousands deaths [22, 26]. Declared as a pandemic by the World Health Organization on March 11, the COVID-19 outbreak has brought severe challenges to not only local healthcare systems (especially in underdeveloped areas) but our society as a whole. At the same time, a large number of related research have been emerging recently in various subjects and fields, attempting to contribute to the battle against COVID-19. Beside pharmacologic and genomics studies on the SARS-CoV-2 virus, data-driven research both on the spread of COVID-19 among local population and on general social impacts brought by the pandemic have been providing valuable insights especially for local policy makers.

On the one hand, various types of sequential models have been implemented to study the spreading behavior of COVID-19 in a generic population, including compartmental model based approaches [3, 5, 12, 18], which are motivated by conventional dynamical models in epidemiology, and deep learning based nonparametric approaches [19, 24]. While the second class, i.e., artificial-learning base approaches, might produce better predictions on disease related statistics, it lacks interpretability for the most part due to the black-box nature of neural network estimators.

On the other hand, the interaction between social environments and the local COVID-19 outbreak remains an important topic. The identification of highly related exogenous factors that impact the local epidemic evolution significantly, e.g., the local population density and the local age structures, is of great importance. In addition to understanding dominant environmental factors that govern the outbreak, the relation between the epidemic evolution and socioeconomic characteristics could potentially also reveal the inverse impact of the COVID-19 on a local community [1].

In the current work, we apply a learning-based approach both to produce an accurate prediction about a near future, and to reveal the interplay between environmental factors and the epidemic evolution. To accomplish this, we implement a neural-network based sequential model, and integrate the time-varying epidemic information, i.e., related statistics including confirmed cases and deceased records, with environmental factors including both dynamical ones, e.g., local restriction policies, and static ones, such as demographic features. Environmental factors enter the model via an "kick-start" mechanism [1], which, after being fixed through training, offers a smoking-gun for the relevance of different factors to the epidemic evolution.

The rest of the paper is organized as follows. Section 2 introduces the ongoing pandemic situation and mentions some related works which motivate our study. In section 3, we enumerate several candidate factors that could potentially affect the epidemic evolution process significantly and discuss the potential epidemiological

*Equal Contribution

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

[1] See Section 4 for details.

dependence, along with other important socioeconomic characteristics, which could be used to analyze the inverse social impact of the ongoing public health crisis. Section 4 elaborates our application approach in details, explaining both the information flow in the model and the way to extract the relevance of environmental factors to the local epidemic outbreak. Section 5 explains model training, including data sources, model structures, and training results. In section 6, we firstly demonstrate the prediction power of the trained model, and then discuss the relevance of different environmental factors to the epidemic emergence extracted from the trained representation. Finally, we summarize our work in Section 7 and discuss potential directions for further investigations with more data and complex models.

2 THE ONGOING PANDEMIC AND RELATED WORKS

As of June 15, 2020, the COVID-19 has hit countries worldwide. We summarize the latest pandemic situation in Table 1.

Category	Statistics
Number of countries reporting COVID-19 cases	225
Number of confirmed cases reported worldwide	7,912,426
Number of deceased cases reported worldwide	433,391
Number of recovered cases reported worldwide	377,131
Currently estimated fatality rate	5.5%

Table 1: Summary statistics of the ongoing COVID-19 pandemic situation worldwide. Data is from the CovidNet project [26].

Governments and organizations across the world have been taking measures at different levels in response to this pandemic crisis. Extreme measures were adopted by the Chinese government in Wuhan, where a complete lock-down of the whole city was implemented. This has been proved later to be very effective to slow down the spread of SARS-CoV-2, the contagious level of which was later revealed to be much higher than two previous deadly viruses, i.e., MERS and SARS. The response in Wuhan then inspired other countries and regions, including South Korea, Thailand, Italy and etc., addressing the importance of social distancing. In spite of the fact that extreme measures have been proven to be effective in fighting against the coronavirus spread, regional lock-down remains a difficult decision to be made for any local government taking into account the economic expense. Therefore, it is of extreme importance to provide policy makers necessary tools to both predict the future trend and understand the social impacts brought by the public health emergency [7, 13, 20].

On the prediction side, conventional *Susceptible Infectious Recovered* [2] (*SIR*) model based approaches have been widely implemented [3, 12, 18], with model parameters estimated from regional epidemic data. Motivated by the data-driven estimation process, deep-learning models have also been hybridized into prediction models [5]. While the *SIR* model and its variants indeed could roughly capture the epidemic law of a generic disease spreading behavior, there are two major problems in practice:

- ODE systems capture continuous dynamics, howbeit real-life epidemic data is usually collected in discrete time. There are also delays in case reporting, which, even worse, never uniform in time [3]. Therefore, there exists a significant mismatch between the true ongoing epidemic process, which can be approximately described by ODEs, and the reported data, which highly depends on human-involved operations.
- *SIR* and its variants only take into account the lowest-order dynamics, which includes the linear terms describing population transitions between compartments and product terms describing interaction/contact between compartments, while keeping transition parameters constants. In reality, however, human responses would also evolve along with the epidemic evolution, which, reversely, could remarkably affect the transmission (due to restriction policies and change in crowd behaviors) and the fatality (due to the improved medical response) of the disease.

Above problems, especially the first one due to human operations, make the task of prediction with compartmental models impractical.

Beside the predictive power, another drawback of ODE based compartmental models is the absence of environmental factors. The trend prediction alone is not enough for designing policy. Instead, understanding the interplay between environmental factors and the epidemic evolution could benefit local policy makers [7, 13, 20], and the dependence of the local outbreak on demographic features and transportation data is essential to calibrate restriction/reopening strategies [6]. At the same time, it is also of great importance to examine the social impact of the public health emergence on the local community, especially on different population groups characterized by genders, races, and ages. Most recent works only provide either qualitative arguments [21] or simple statistical analysis, e.g., the linear regression [1], which has limited modeling capability.

Compared with above methods, our current approach attempts to incorporate environmental factors into the prediction module directly. The explicit factor-dependence and therefore interpretability of the model are available via a proper analysis on the learnt representation. Details of our modeling and training can be found in Section 4 and 5.

3 THE INTERPLAY BETWEEN ENVIRONMENTAL FACTORS AND THE COVID-19 EPIDEMIC

Before introducing specific model structures and training designs, it is necessary to discuss and distinguish candidate environment factors, which either directly govern the epidemic evolution process, a.k.a. *exogenous factors*, or reveal the social impact of COVID-19 from essential perspectives.

3.1 Dominant factors of the transmission

In reality, many exogenous factors would affect the transmission beside disease characteristics. Most intuitively, there is a higher

[2] Or *Susceptible Infectious Removed in some literature, which also considered deceased cases.*

[3] For example, the obvious periodic (week-wise) pattern in the U.S. death data is due to the reporting schedule of the official departments.

probability in regions with denser population distributions that a fast outbreak would emerge. New York City, being the most densely populated county-equivalence in the United States, would serve as a typical example. The outbreak in NYC evolved rather rapidly from the beginning and, as of May 30, 2020, NYC has accounted for than 55 percents of cumulatively confirmed cases across the whole state.

Another dominating factor for transmission would be the restrictive order issued by the local government. Restrictions have been implemented onto various industry/business activities as well as local residents' daily life. While industry/business restrictions differ region by region, and are usually difficult to study quantitatively, in the present work we instead use the restriction on local residents' daily life, i.e. *the stay-at-home order or its equivalencies*, as an aggregated representation of the *local restriction level* to capture the overall impact of local policies[4]. Importantly, the change of the restrictive level, from the no-restriction stage, to the restrictive-order stage, and finally to the reopen stage, results into a time-varying transmission behavior of COVID-19, which is in contrast to ODE-based compartmental models that assume a constant transmission factor β.

More generally, the interaction within a local population contributes significantly to the transmission. We therefore adopt the average annual enplanements per capital [16] to capture the active level of human interactions.

3.2 Vulnerable Population Groups and Descriptive Factors

It has been confirmed by data from multiple countries and regions [9, 14, 17] that the patient's age is highly related to the development of severe pneumonia symptoms. Aged people are in general more vulnerable to the virus. We therefore incorporate the age structure as an important demographic category.

More generally, the physical condition of individuals would result into differences in the probability of infection. For instance, people with poor respiratory condition experience higher risk of being infected. We therefore include *the population with high risk* [10] as an input feature of modeling.

3.3 Smoking-gun factors for social impact analysis

While above mentioned environmental factors focus more on the biological aspects of the epidemic evolution, there are also other socioeconomic characteristics, which, although not biologically relevant, could be potentially correlated with the evolution behavior.

For example, a study [1] focusing on the New York data has shown that a higher probability of positive testing rate is in poorer neighborhoods, in neighborhoods where large numbers of people reside together, and in neighborhoods with a large black or immigrant population. At the same time, however, people residing in poorer or immigrant neighborhoods were less likely to be tested. As a result, an understanding of which types of neighborhoods are disproportionately affected by the pandemic requires an examination of how socioeconomic characteristics correlate with different epidemic statistics.

Motivated by the above discussion, we selected several socioeconomic characteristics, which are not only related to the epidemic statistics from the pure data perspective, but also important in revealing potential disproportions of the COVID-19 impacts, including local gross domestic product (GDP) per capita and local race compositions.

4 A LEARNING BASED APPROACH

Now we introduce our learning based approach which, in a nutshell, implements a recurrent-neural-network model for the trend prediction along with an embedding of environmental factors to extract relevant information. There are two classes of inputs: the epidemic time series, including both confirmed cases and deceased cases, and environmental socioeconomic factors.

The epidemic time series data enters the learning module through a stacked Gated Recurrent Unit (GRU) model, which is well-known for both its power in dealing with sequential data by incorporating history information properly, and its efficiently simplified structure. We cast a fixed-length (L) sequence with recent history information to predict each subsequent data point, by implementing a sliding window on the full time series. Importantly, we would like to address that this GRU-based model structure is powerful in the following sense:

(1) Firstly, this GRU model structure, at least, is capable of capturing the dynamics of compartmental models in the discrete-time regime of compartmental models, where the value of the next time-step only relies on the current state. For instance, this can be easily shown through the following set of *difference equations*:

$$\begin{aligned} s_{t+1} - s_t &= -\beta \cdot s_t \cdot i_t, \\ i_{t+1} - i_t &= \beta \cdot s_t \cdot i_t - \gamma \cdot i_t, \\ r_{t+1} - r_t &= \gamma \cdot i_t, \end{aligned} \tag{1}$$

where $\{s_t, i_t, r_t\}$ represent susceptible, infectious, and removed population fractions respectively, and $\{\beta, \gamma\}$ describe transmission rate and removal rate of the disease. This set of equations can be viewed as a discrete version of a *SIR* model [2], and they clearly only depend on the current snapshot of epidemic statistics; The above dynamics, in the ideal scenario, can be captured by a GRU model with only $L = 1$ sliding-window length.

(2) Mathematically, the above difference equations in Eqs.1 are not always a legitimate format. Indeed, a rigorous transformation from the continuous time regime to the discrete one requires much more caution, and the exact solution form would bring in more complicated terms with both longer history dependence and higher-order polynomial terms. To confront this difficulty, we apply both a sliding window with longer length $L > 1$, and more layers in the GRU module to capture complicated nonlinear terms.

Another input feature with time-varying values is the status of the local restriction policy, i.e., the stay-at-home policy or similar

[4]Practically, this is also the only consistent data category accessible to the general public

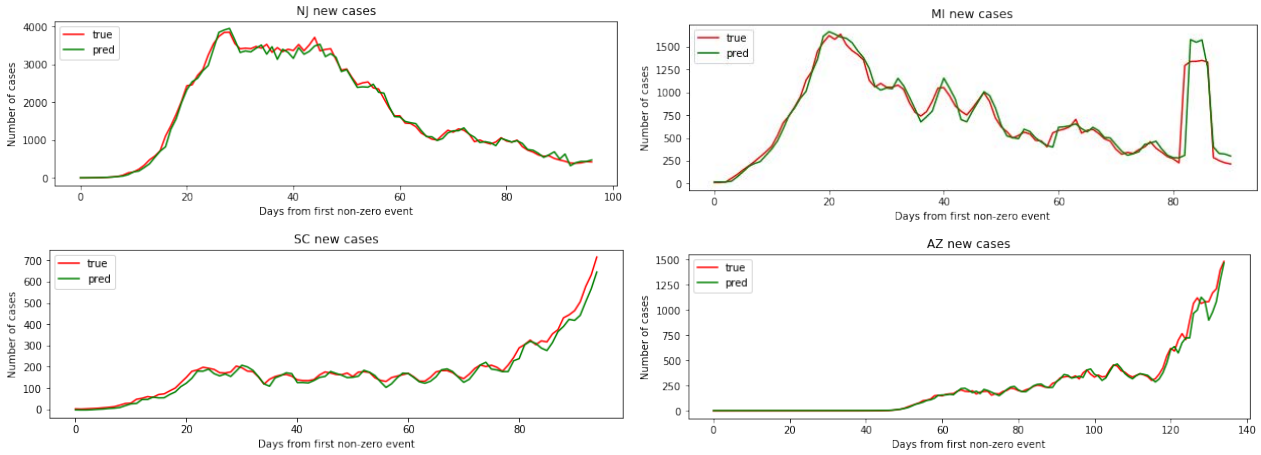


Figure 1: The *1-step-ahead prediction* task on testing states: NJ, MI, SC, AZ, the data from which has never been accessible to the model. The predicted curves follow the true values closely.

measures [5], which, different from numeric data, is categorical with a binary status: "stay-at-home" or "reopen". It is naturally expected that the epidemic evolution would be different under different restriction statuses. Therefore, we apply a "double-channel" structure in the GRU module: a sequential data point would enter channel-1 if there is a restriction policy on the corresponding date, and would enter channel-2 otherwise.

In addition to time-series data, we have also integrated the following list of environmental factors as static input features:

- local population density;
- local GDP per capita;
- local age structure (fractions of 6 non-overlapping age groups);
- local race structure (fraction of 7 different race categories);
- high risk population;
- local annual enplanements per capital;
- local restrictive order level;

where all factors are summarized and represented on state level. Different from the sequential data of epidemic statistics which enters the model via a black-box, although reasonable, as explained above, model structure, we would like to investigate the interplay between the epidemic evolution and various socioeconomic characteristics. Therefore, when we incorporate the input of environmental factors, we apply a linear embedding, which produces interpretable weights on each input dimension. Technically, the embedded representation of environmental factors is taken as *the initialization of the hidden state in the GRU model*, which we call as a "kick-start" mechanism.

Through the above design of the information flow, we implicitly construct a desired interaction between environmental factors and the epidemic time series data: these two input-categories are conducted to interact with each other via various gates in the GRU module. From an epidemic perspective, within the GRU structure, the hidden state could be regarded as an evolving "environment", whose initial status, i.e., before the first infectious case emerges, only depends on exogenous demographic factors of the local community.

[5]This includes the stay-at-home advisory issued in Massachusetts, the curfew issued in Puerto Rico, and so on

5 MODEL TRAINING

In this section, we elaborate the practice of model implementation, including data sources, hyperparameters of implemented model, and details of the training procedure.

Sources of Different Data Categories.

- **COVID-19 case data:** Case data is from the CovidNet project [26], including confirmed and deceased counts of 50 U.S. states and the District of Columbia, ranging from January 21, 2020, to June 14, 2020.
- **State restriction policy:** Restriction policy information is collected from "The Coronavirus Outbreak" forum on the New York Times [23].
- **Population and density:** We have used population data from the U.S. Census Bureau [25].
- **Population with higher risk:** Population in each state with higher risk to develop severe symptoms are estimated in [10], and used as an exogenous factor in our application.
- **Age structure data:** We have used the age structure dataset built by the Kaiser Family Foundation [11].
- **Race structure data:** Race structure data is collected from the COVID Tracking Project [4].
- **Annual enplanements data:** We collected the data of annual enplanements per capital in each states (not including D.C.) from the U.S. Department of Transportation [16].
- **Gross domestic product per capita:** We collect the data from the United States Census Bureau [25]

Hyper-parameters of the Implemented Model. Our model is implemented with an embedding module, recurrent module and output module. The embedding module sparsely encodes the 21-dimension state-specific demographic vector into a 100-dimension vector; the recurrent module is using 3 stacked GRU layers with 100-dimension hidden states, the recurrent module takes the previous

embedding result as its first latent state and the windowed state total confirmed cases and new cases as inputs; a dense layer is used for the output layer in order to predict the target. We trained the model using Adam optimizer [15] with 1e-4 learning rate and discounted the learning rate with a factor of 0.3 if the training loss didn't decrease over 20 epochs. The model is trained on an MacBook Pro with 6-Core CPU.

Details on the Training Design and Process. The detailed model architecture is demonstrated in Figure. 3. We use a five-day window time-series historical data and use recurrent module for prediction. The input data for recurrent module are total confirmed cases (*cc*) and new confirmed cases (*dc*), we pro-process each of them into two series: one contains value only when there is restriction policy undergoing (*cc_res* and *dc_res*) and another only contains value when there is no restriction policy (*cc_nores* and *dc_nores*) as shown in the model architecture. The processed state demographic data feeds into the state demographic embedding layer, and we use Sigmoid activation function inside the layer. All three layers of GRU receive the embedding output as its first hidden state h_0. Root mean squared error(RMSE) is chosen as the loss function. We separate our data firstly withholding 5 states as test data. The others are processed as windowed input-output pairs and separated into a proportion of 80% for training and 20% for evaluation. The model is learned via back-propagation utill convergence.

6 RESULT ANALYSIS

As mentioned earlier, the current work targets both a prediction of the epidemic evolution, and an understanding of the interplay between environmental factors and the local epidemic outbreak. In this section, we discuss the two aspects with the trained model.

6.1 Prediction of the Epidemic Evolution

As we have applied the random shuffling during training[6] among the training data, which is transformed into *sequence-to-point* pairs, we would demonstrate the prediction power in two ways:

(1) "1-step-ahead prediction" on testing states: during the training stage, we have randomly eliminated several states from the complete dataset [7]. We would test the performance of the trained model on these states, whose history records have never been acknowledged by the model;
(2) Long-term prediction from an auto-regressive process: the model was trained for 1-step-ahead prediction only during the training stage, therefore long term prediction would be a non-trivial demonstration for the model's capability of capturing the true dynamics.

Figure 1 shows the performance of the model on the *1-step-ahead prediction task*. Clearly, even though data from testing states have never been accessible to the model, the trained model can still predict the future value very well. It is therefore reasonable to state that, rather than over-fitting the given data during the training stage, the model instead capture the general law of the epidemic evolution in a generic population. The existence of such a law is not a surprise, and has already been hypothesised in conventional compartmental-model methods. However, the general law could easily become intractable from real data due to the human operation in reporting schemes [8]. By applying the proposed learning-based method, we extract this general law from the noisy real-life data.

Compared with the *1-step-ahead prediction* task, the *long-term prediction* is much more challenging, in the sense that the task nature has deviated from the training stage. The performance of long-term predictions is shown in Figure 2. While in some states, the deviation from the true data become visible, the overall trend has still been well captured by the auto-regressive process, except noisy fluctuations. In practice, the long-term prediction could provide more timely information to policy makers, and hence is more valuable than *1-step-ahead predictions*.

6.2 Relevance of Environmental Factors

To reveal the interplay between environmental factors and the epidemic evolution, we start from an analysis on the relevance of each input feature to the dynamics. As introduced in Section 4, static environmental factors, after being embedded through a linear transformation, enter the model via the "kick-start" mechanism. Due to the simple structure of this embedding module, we could easily identify the relevance of input features by examining the Frobenius norm of each embedding vector.

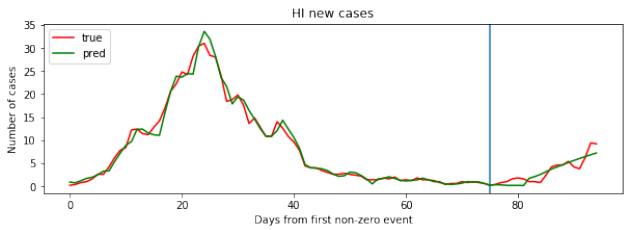
Firstly, there are two classes of population structure data: the age structure and the race structure. Figure 4(a) and 4(b) show the relevance of different age groups and race groups respectively.

In the age group relevance chart, it is clear that the two young-age groups, i.e., *age from 19 to 25* and *26 to 34*, show the highest relevance. This is consistent with the demographic report released by CDC on age distribution [8], where the age group $18-44$ contributes the largest portion to the confirmed cases.

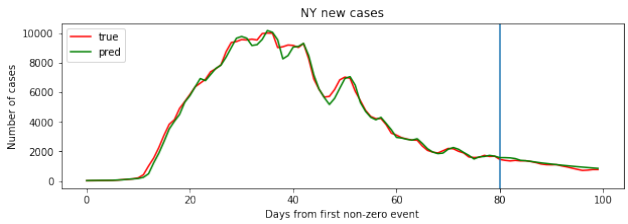
Among all race groups, the two groups, *Asians* and *Black or African American alone*, appear to be more relevant in epidemic dynamics, while both *White (non-Hispanic)* and *Hispanic/Latino* show lower relevance. On the other hand, it is interesting to note that, according to the report released by CDC [8], *White (non-Hispanic)* and *Hispanic/Latino* contribute largest portion in the confirmed cases. While the share in confirmed cases may be more related to the absolute population size of different race groups, our relevance analysis, instead, focuses on the fraction of each race group in a certain state. The above mismatch between the results obtained via the two descriptive perspectives suggests a potentially existing disproportional impacts of the COVID-19 on different groups. While the above argument does not provide a rigorous causal analysis, it illustrates the importance of diversity of perspectives when studying the social impact of COVID-19.

Beside the above two types of population structures, we also notice a high relevance (1.6810614) of enplanements data to the epidemic dynamics. This confirms our earlier hypothesis that the enplanements data could be used as a nice indicator for the active level of local socioeconomic activities.

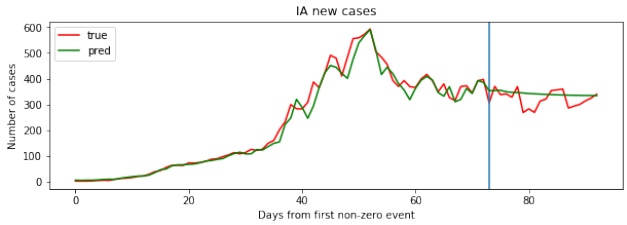
[6] See details explained in Section 5
[7] In our practice, we have randomly selected 4 states for testing purpose: AZ, MI,NJ, SC.
[8] See Section 2 for detailed discussions.



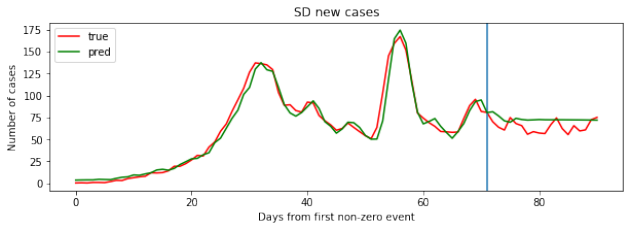
(a) Long-term prediction of Hawaii data.



(b) Long-term prediction of New York data.



(c) Long-term prediction of Iowa data.



(d) Long-term prediction of South Dakota.

Figure 2: The *long-term prediction* task with 4 instantiating states. The solid vertical blue line represents the starting point of the auto-regressive running.

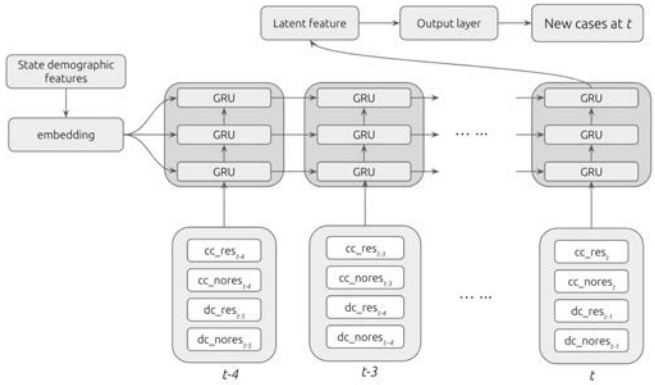
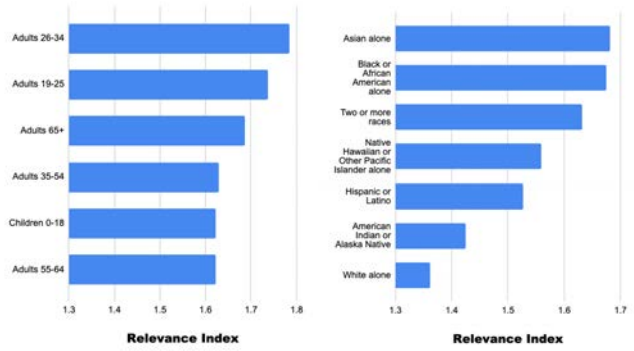


Figure 3: Model architecture



(a) Age-group *relevance*. (b) Race-group *relevance*.

Figure 4: Relevance of different age-group fractions 4(a) and race-group fractions 4(b). The *relevance index* is defined as the Frobenius norm of the embedding vector of each input feature.

7 CONCLUSIONS AND DISCUSSIONS

We have demonstrated the predictive power of the proposed recurrent-network based model, and discussed the relevance of different environmental factors by studying the embedding vector of each socioeconomic characteristic.

On the prediction side, the proposed model performs well in both *1-step-ahead prediction on new states* and *long-term prediction* tasks. One could conclude that the recurrent structure has successfully extracted and captured a general law of the epidemic evolution in a generic population, from the real-life noisy data.

On the other hand, studying the relevance of environmental factors to the epidemic dynamics enables us both to identify potential factors that contribute most to the disease spreading, and to understand the social impact of COVID-19 on the local community. More specifically, we noticed that young age groups and average emplacements are highly relevant to the dynamics, verifying the fact that socioeconomic activities contribute significantly to the disease spread; besides, there might exist a disproportion of the social impact on different race groups brought by the COVID-19.

In general, one could expect that more insights about the ongoing public health crisis could be gained through data-driven research. Besides medical and clinical studies that directly battle the COVID-19 emergence, it is also important to obtain a more complete understanding about general social impacts of the pandemic on the population and society level. This does not only assist local policy makers in decision making, but also helps the whole society to confront the challenge together.

8 ACKNOWLEDGEMENTS

Funding for the shared GPU-computing facility used in this research was provided by NSF OAC 1920147.

REFERENCES

[1] George J. Borjas. 2020. Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods. (2020). https://www.nber.org/papers/w26952

[2] Fred Brauer. 2008. *Compartmental Models in Epidemiology.* Springer Berlin Heidelberg, Berlin, Heidelberg, 19–79. https://doi.org/10.1007/978-3-540-78911-6_2

[3] Ben-Hur Francisco Cardoso and Sebasti'an Gonccalves. 2020. Urban Scaling of COVID-19 epidemics. *arXiv: Populations and Evolution* (2020).

[4] Covid Tracking Project Team. 2020. Covid Tracking Project. https://covidtracking.com/

[5] Raj Dandekar and George Barbastathis. 2020. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. *medRxiv* (2020). https://doi.org/10.1101/2020.04.03.20052084

[6] Jennifer Beam Dowd, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, and Melinda C. Mills. 2020. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences* 117, 18 (2020), 9696–9698. https://doi.org/10.1073/pnas.2004911117

[7] Samuel Engle, John Stromme, and Anson Zhou. 2020. Staying at Home: Mobility Effects of COVID-19. (2020). https://doi.org/10.2139/ssrn.3565703

[8] Center for Diseases Control and Prevention. 2020. Coronavirus Disease 2019 (COVID-19): Cases in U.S. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[9] National Center for Health Statistics. 2020. Weekly Updates by Select Demographic and Geographic Characteristics. https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm#AgeAndSex

[10] KAISER FAMILY FOUNDATION. 2018. *KFF analysis of 2018 Behavioral Risk Factor Surveillance System.* https://www.kff.org/other/state-indicator/adults-at-higher-risk-of-serious-illness-if-infected-with-coronavirus/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D#notes

[11] KAISER FAMILY FOUNDATION. 2018. *Population Distribution by Age.* https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D#notes

[12] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. 2020. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine* (2020), 1–6.

[13] Lawrence O. Gostin and Lindsay F. Wiley. 2020. Governmental Public Health Powers During the COVID-19 Pandemic: Stay-at-home Orders, Business Closures, and Travel Restrictions. *JAMA* 323, 21 (06 2020), 2137–2138. https://doi.org/10.1001/jama.2020.5460

[14] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David S.C. Hui, Bin Du, Lan-juan Li, Guang Zeng, Kwok-Yung Yuen, Ru-chong Chen, Chun-li Tang, Tao Wang, Ping-yan Chen, Jie Xiang, Shi-yue Li, Jin-lin Wang, Zi-jing Liang, Yi-xiang Peng, Li Wei, Yong Liu, Ya-hua Hu, Peng Peng, Jian-ming Wang, Ji-yang Liu, Zhong Chen, Gang Li, Zhi-jian Zheng, Shao-qin Qiu, Jie Luo, Chang-jiang Ye, Shao-yong Zhu, and Nan-shan Zhong. 2020. Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* 382, 18 (2020), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] U.S. Department of Transportation. 2019. *Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports.* https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/

[17] Graziano Onder, Giovanni Rezza, and Silvio Brusaferro. 2020. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* 323, 18 (05 2020), 1775–1776. https://doi.org/10.1001/jama.2020.4683

[18] Nicola Picchiotti, Monica Salvioli, Elena Zanardini, and Francesco Missale. 2020. COVID-19 pandemic: a mobility-dependent SEIR model with undetected cases in Italy, Europe and US. *arXiv: Populations and Evolution* (2020).

[19] Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2020. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. *medRxiv* (2020). https://doi.org/10.1101/2020.04.08.20057679

[20] Brendon Sen-Crowe, Mark McKenney, and Adel Elkbuli. 2020. Social distancing during the COVID-19 pandemic: Staying home save lives. *American Journal of Emergency Medicine* (2020). https://doi.org/10.1016/j.ajem.2020.03.063

[21] Kaiyuan Sun, Jenny Chen, and Cécile Viboud. 2020. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet Digital Health* 2, 4 (2020), e201 – e208. https://doi.org/10.1016/S2589-7500(20)30026-1

[22] 1Point3Acres COVID-19 Tracker Team. 2020. 1Point3Acres COVID-19 Tracker. https://coronavirus.1point3acres.com/

[23] The New York Times. 2020. The Coronavirus Outbreak: See How All 50 States Are Reopening. https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html

[24] Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gill. 2020. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things* 11 (2020), 100222. https://doi.org/10.1016/j.iot.2020.100222

[25] U.S. Census Bureau. 2019. State Population Totals and Components of Change: 2010-2019. https://www.census.gov/

[26] Tong Yang, Kai Shen, Sixuan He, Enyu Li, Peter Sun, Lin Zuo, Jiayue Hu, Yiwen Mo, Weiwei Zhang, Ping-Ying Chen, Haonan Zhang, Jingxue Chen, and Yu Guo. 2020. CovidNet: To Bring the Data Transparency in Era of COVID-19. *ArXiv* abs/2005.10948 (2020).

What is government supposed to do in the epidemic: from the perspective of Chinese stock market and disinformation

Yufei Wu
Rensselaer Polytechnic Institute
Troy, New York, USA
wuy31@rpi.edu

ABSTRACT

Based on the four data sets including the epidemic situation data set, SSE Composite index data set, primary sector indices dataset and rumor dataset, this paper studies the impact of COVID-19 on them and discussed what action the government should take in the further epidemic. For the stock market, we use the VIF and ANOVA to screen the independent variables and adopted the linear regression to judge the impact of the global and Chinese epidemic on the Chinese stock market as a whole and various sectors in stock market. In terms of rumor, TF-IDF is calculated to generate a word cloud, and the LDA model is used to find the keywords of different topics.

In the results of data analysis, we can see that the number of people daily cured in China and the global daily recovery rate has a positive impact on the stock market, no matter index or volume, while the global daily confirmed number hurts the stock market. Also, different industries have different effects on the stock index. For example, when the number of people recovered in China increased, the SSE telecommunication services sector index increased rapidly, but the SSE health care sector index decreased. Because of the influence on the stock market, the relevant macro-control policies issued by the state are necessary. In terms of the analysis result of rumors, people are mainly concerned about the spread of the epidemic, including the use of protective equipmentdisinfection tools and how long the virus will last in different fabrics. At the same time, due to the lockdown of Wuhan, people were also anxious about the situation in Wuhan. This is also an entry point for the government to formulate relevant publicity policies. It requires the government to promptly announce the epidemic situation, especially in Wuhan, a relatively severe area, organize competent medical experts to publicize virus-related knowledge on various platforms, and improve the productivity of relevant protective products, such as masks, to reduce people's anxiety

CCS CONCEPTS

• **Mathematics of computing** → **Probability and statistics**; **Multivariate statistics**; • **Social and professional topics**;

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

KEYWORDS

COVID-19,stock market, rumor, government, data-driven analysis

ACM Reference Format:
Yufei Wu. 2020. What is government supposed to do in the epidemic: from the perspective of Chinese stock market and disinformation. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 6 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

When the outbreak occurs, the economy of certain country and the world will be affected. Just like the warning of SARS to us[5], when a new epidemic happens, we can predict the impact of the outbreak on the economy and build a macroeconomic model to ensure the security of the country to a certain extent. At the same time, rumors will cause panic[9], so the government needs to take relevant measures to solve the disinformation in time. In 2020, we are facing the global impact of COVID-19. The economy and rumors are also problems that need to be addressed in the COVID-19 epidemic. This paper's core is to explore the impact of the COVID-19 on the two areas and prepare for the further epidemic. To delve into these two issues, we chose the two angles, which are stock markets and rumors, because they will give us a view from a statistical perspective and meet our research demand.

From the economic aspect, we mainly focus on the daily confirmed cases, daily deaths, and daily recovered cases in China and the world. And how they influenced the SSE Composite Index, which is a significant indicator of the overall situation of the Chinese stock market, and the primary sector indices, giving people an overview of different industries in China. Based on all the information, we can have a comprehensive understanding of the overall economic situation and development of various sectors. Regarding specific methodslinear expression is used to judge the influence of COVID-19 on the economy. At the same time, we need to study whether the influence is different on distinct industries.

On the other hand, we are concerned about how important to open information in the epidemic, especially in the area of rumors. The rumor data is the original rumor and the clarifications of rumors. The primary method we implement is to recognize the rumor categories and visualize the keywords. To be more specific, the latent Dirichlet allocation model is mainly used to understand which aspects are more likely to generate rumor and determine the direction of the clarifications of disinformation.

2 DESCRIPTION OF DATASET

There are mainly four datasets used in the research. The first one is about the COVID-19 cases from JHU CSSE. It records the current

confirmed, dead, and recovered instances in different regions in the world. From this dataset, we can calculate the daily new cases in the world and China. The second one and the third one is about the Chinese stock market. The second one is about the SSE Composite index, while the third one is about primary sector indices. The SSE Composite Index, also known as SSE Index, is a stock market index of all stocks traded at the Shanghai Stock Exchange. It is one of the most important indicators of the Chinese stock market. The primary sector indices give the stock index of the different industries, including energy, materials, manufacturing, consumption, medicine, finance, information, telecommunications, public utilities, etc. From these datasets, We can have a general idea from a macro and micro perspective. The fourth one is rumor-refuting dataset. The source of the data is DXY, which is a critical channel for Chinese people to obtain relevant epidemic information.

Since this paper is mainly related to the epidemic situation in China, the time range of data we chose is from January 24 to May 15. In late January, reliable and clear data about the infection were available, and in mid-May, the COVID-19 was under control in China. On the whole, this period can represent the curve of epidemic development in China.

3 STOCK MARKET

Under the influence of the epidemic, the macro-economy of all countries in the world has been hit, and GDP has declined[2], which also has an impact on the stock market[7]. Therefore, it is necessary for the state to formulate relevant policies for the stock market. Because there are many attributes in the COVID-19 cases dataset, including daily confirmed cases, recovered cases, dead cases, recovered ratio(the ratio of everyday recovered cases to confirmed cases) in the world and in China. And it is obvious that all of them are closely tied. To select a high correlation and reduce the multicollinearity of all independent variables, we chose MANOVA[1] and variance inflation factor (VIF)[4] to select the suitable independent variables for our model. After getting all the elements whose P-value is less than 0.05, we also test the combination of factors to ensure that the VIF value is less than 5. Finally, the independent variables we chose are daily confirmed cases in the world, recovered instances in China, and improved ratio in the world. The result of VIF can be seen as Table1. All of them are less than 5, and the combination is relatively comprehensive of all the combination which meet the demand.

Table 1: the VIF value of the combination of confirmed cases in the world, recovered cases in the China, recovered ratio in the world

Attribute	VIF
confirmed cases in the world	1.986104
recovered cases in the China	3.114696
recovered ratio in the world	1.930471

After obtaining the combination of independent variables that will influence the stock market, we try to figure out to what degree it will affect the index and volume of the stock market. Therefore, we chose linear regression to evaluate them. The dependent variable can be the index of the SSE composite index, the volume of stocks trading on the Shanghai Stock Exchange and primary sectors indices.

3.1 SSE composite index

The SSE composite index is a statistical index reflecting the general trend of listed stocks in the Shanghai Stock Exchange. The total market value, circulation market value, quantity proportion, and transaction amount proportion of stocks in the Shanghai Stock Exchange are quite considerable in the Chinese stock market. Therefore, for this research, we mainly focus on two attributes in this dataset, which are volume and index, to evaluate the impact of COVID-19 on the Chinese stock market from a macro perspective. The coefficient of linear regression about the SSE composite index can be seen in Table 2, and the result about the volume of the SSE composite index can be seen in Table 3.

Table 2: the coefficient of linear regression about index

Attribute	Coefficient
confirmed cases in the world	-0.19899411
recovered cases in the China	0.34849017
recovered ratio in the world	0.32447878

Table 3: the coefficient of linear regression about volume

Attribute	Coefficient
confirmed cases in the world	-0.27922668
recovered cases in the China	0.34849017
recovered ratio in the world	0.32447878

When building the model, we divide the data into two parts according to the ratio of 9 to 1: training set and test set. The reason why we split data like this is that there is not a long time before the outbreak, so our data volume is relatively small. To validate our model's accuracy, Root Mean Square Error(RMSE)[3] is used. RMSE is around 0.19 in the model of regarding index as the dependent variable, so there is a forecast error. RMSE is 0.07 in the model regarding volume as the dependent variable. It shows that the overall accuracy of the model is relatively high. We will also consider these errors in the subsequent analysis.

According to the result above, even though there is a difference between the result of index and volume, all three factors have a particular impact on China's stock market. Among them, Chinese daily recovery cases and the world's daily recovery ratio have a positive effect on the stock market, while the world's confirmed cases hurt the stock market.

The reason why Chinese recovery cases will largely influence the stock market is that the Chinese stock market is deeply affected by the high proportion of individual investors. The massive recovery of the infected population has given them confidence in the market, raising the stock index and volume. It is believed that the domestic

epidemic control has achieved remarkable results. Besides, in order to support more enterprises to return to work and production, the government has issued a series of supportive policies and implemented them, such as providing tax relief and credit support for enterprises, especially small and medium-sized enterprises. At the same time, it offers more jobs for infrastructure construction and promotes the development of infrastructure-related industries. At the same time, due to the later stage of the epidemic, the world's epidemic center was transferred, and foreign commercial capital returned. The influx of money further promotes the development of related industries.

In the meanwhile, global situation of COVID-19 also has a meaningful impact on China's stock market. It is undeniable that China's economy is closely connected with the world in the context of economic globalization. In the later stage of COVID-19, due to the effect of the COVID-19, the economy has been dramatically impacted in many regions of the world[8], and it in t will influence China to some degree, which has close trade ties with them. The close relationship between China and the global economy is also why the global recovery ratio has a positive impact on China's stock market.

3.2 primary sectors indices

Then, we turn our attention to different areas of the stock market. We still used the linear regression method to study various industries in the world, only to change the data set from the SSE Composite index to primary sector indices. The coefficient of linear regression will help us to compare the impact of the epidemic on different industries. The results of linear regression are shown in the table 4 below.

Table 4: the coefficient of linear regression about primary sectors indices

Global confirmed cases	Recovered cases in China	Global recovered ratio	section code
0.578599	0.152707	0.166588	sh.000032
-0.133817	0.371597	0.380406	sh.000033
0.11084	0.583916	0.202579	sh.000034
-0.390236	0.233471	0.381991	sh.000035
0.56165	0.127547	0.273332	sh.000036
0.39067	-0.110289	0.407933	sh.000037
-0.35649	0.04493	0.381908	sh.000038
0.004645	0.41213	0.485089	sh.000039
0.109616	0.81435	0.036243	sh.000040
-0.081437	0.397544	0.206555	sh.000041

According to the result above, the most negatively affected by confirmed cases worldwide are sh.000035 and sh.000038. They represent the SSE Consumer Staples Sector index and SSE Financials Sector index, respectively. And there are also some indexes positively correlated with confirmed cases worldwide, which are sh.000032 and sh.000036. They represent the SSE Energy Sector index and SSE Consumer Discretionary Sector index. Although the SSE Consumer Discretionary Sector index and SSE Consumer Staples Sector index look similar, they are indeed about different areas of consumption. The former is the most basic and necessary consumer goods in our daily life, including agricultural, animal husbandry, fishery products, food, personal household products, etc. The latter is the consumption in addition to the necessary expenditure, including cars, clothing, media, etc. Because of the aggravation of the global epidemic situation and the stagnation of industries in various countries, China's production has also been affected. The Chinese financial sector has also been affected due to repeated meltdowns of the US stock market. At the same time, during the severe period of the global epidemic, the epidemic situation in China has been controlled. After long isolation, people's desire for essential consumption has rebounded. But given that the economy has been struck by the epidemic, many people don't choose luxury consumption. In the meantime, China, as an energy importing country, has benefited to some extent from the collapse of oil, which explained why the energy-related index will increase when the worldwide confirmed cases rise.

When the number of recovery cases in China increased, the most positively affected were SSE Telecommunication Services Sector index (sh.000040) and SSE Industrials Sector index (sh.000040). SSE Health Care Sector index (sh.000037) was the only one that decreased with recovery increase. Wuhan is a crucial photovoltaic base in China. Wuhan East Lake New Technology Development Zone has gathered many optoelectronic enterprises and has formed an industrial pattern led by the optoelectronic information industry. The control of China's disease shows that production can gradually start to recover, especially the industrial recovery in Wuhan areas has promoted the development of the telecommunications industry. 5G construction after the peak of COVID-19 is also one of the reasons. Also, it is clear that as the overall epidemic's impact on China has diminished, the medical-related index has fallen somewhat.

The impact of the recovery ratio of the global epidemic on various industries is also positive, especially for the SSE Information Technology Sector index (sh.000039) and SSE Health Care Sector index (sh.000037). That is because, in the middle and late stages of the epidemic, people chose to study and work at home, and the related information technology industry developed. At the same time, people had the need to go out to work with the awareness of self-protection in the epidemic, so the demand for PPE increased, which promoted the development of healthcare-related fields in China. Also, the need for vaccine development may be another reason.

4 RUMOR

For the rumor dataset, what we need to care about are doing Chinese word segmentation and computing TF-IDF[6] term frequency-inverse document frequency). Different from English, there is no word boundary in Chinese sentences, so when processing Chinese natural language, it is usually necessary to segment words first. In this article, we choose the Jieba module to do this task. TF-IDF is a statistical method to evaluate the importance of a single word in a document set or corpus. The value of TF-IDF will increase with the number of words appearing in the document and decrease with the number of words appearing in the corpus. We drew a word cloud plot based on the TF-IDF value below. And we set the number of

words in the clouds are one hundred. The word cloud can be seen as Figure1.

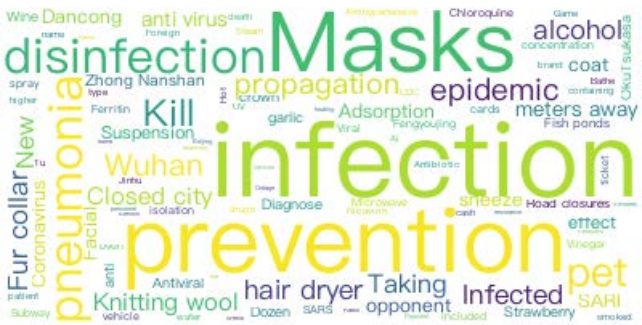


Figure 1: Word Cloud of rumors

From the image above, the larger font of the word is, the higher the weight of its corresponding TF-IDF value is. In this case, we can see that the rumor is mainly about how to reduce the spread of the virus to prevent infection. For example, in the word cloud plot, we can see the word Mask, which shows that due to the shortage of masks, how to use masks correctly to prolong the life of them has become a topic of concern. Also, we can see that heat, alcohol, hairdryer are also in the same category. People are concerned about whether heating and alcohol can effectively eradicate viruses. In addition, there are coat and fur collar. People are worried about how long the virus will last in different clothes. The rumor disseminator just uses people's psychology of being afraid of viruses to spread the related disinformation to cause panic or dump goods.

On the other hand, people are also very concerned about the living condition and medical conditions in Wuhan. Because the lockdown is really strict in Wuhan, information flow may not be timely. Then it is more likely to lead to the spread of rumors, so the relative proportion of fake news in Wuhan is very high. In this case, the government must organize not only relevant medical experts to refute rumors in time and guide people to prevent epidemic scientifically but also use various media to show people's life in Wuhan to rest assured the whole nation.

To better understand the keywords, we can apply LDA [10]. LDA can also be regarded as a Bayesian probability model with three levels of structure, including words, topics, and documents. We choose the attribute named body as text source because it contains not only the rumors and why it is wrong. For the LDA model, it is essential for us to pick a value k to determine how many topics we could get. To get the value k, we should compute perplexity based on different k to choose the optimal one. The graph of perplexity value based on different k can be seen as the Figure 2 below.

According to the image, it shows the value of perplexity from k=1 to k=50. Perplexity is a measurement method of information theory. It is the evaluation of the LDA model and the judgment of improved parameters. To determine the optimal k of our LDA topics, we have to select the value of k at the elbow of the line. Looking at the trend in the image, we could choose a value of k from 5 to 10. And since the amount of data is not large, if we want a larger k, the topics we finally got may not be representative. So that we ultimately choose 5 as the k value. After training our models

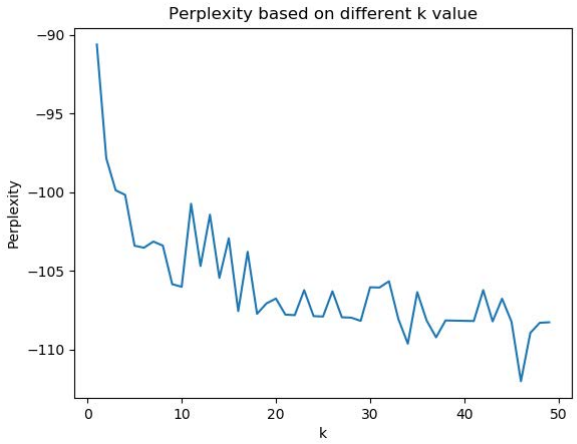


Figure 2: The value of perplexity based on different k-value

and translating the keywords in the different topics into English, our result can be seen in Table 4 below. And for each cluster, we show five most representative keywords to describe a specific topic. The result can be seen in Table 5.

Table 5: the coefficient of linear regression about primary sectors indices

Topic Number	keywords	corresponding probability
1	mask, country, over, immunization, medical	0.149, 0.045, 0.023, 0.021, 0.013
2	health, Center, Dog, Disease Control, Diagnosis	0.046, 0.031, 0.024, 0.023, 0.021
3	Wuhan, medical, epidemic, symptoms, pneumonia	0.061, 0.051, 0.045, 0.036, 0.035
4	effect report rumors isolation heat	0.087, 0.073, 0.046, 0.031, 0.020
5	respiratory, respiratory tract, droplets, spread, coronavirus	0.080, 0.038, 0.034, 0.034, 0.031

According to the table above, there are five topics and their corresponding keywords and probability. For the first topic, we could see that the keyword mask has a much higher value than other words. It said that the usage and effect of the mask, and how to identify the most effective mask are what people really care about, just like word cloud plot shows. Under this circumstance, especially in the early stage of China's epidemic, the production capacity o has not been improved, and medical masks and N95 masks are in short supply. So people will think of various ways to extend the validity of masks. To meet the needs of people, multiple rumors about prolonging the life of masks came into being. This is what the government needs to pay attention to. It is of considerable significance to guide people on how to choose and use masks correctly

for the prevention and control of the epidemic and to avoid greater social panic.

For the second topic, as the keywords described, it mainly focuses on the health care system and pets. As for the healthcare system, it is not the source of rumors, but the subject of refutation. After reading the original file of the rumors, we realized the theme of the groundless information is professionals from hospitals or health systems. As experts in the field of infectious diseases, their statements are more persuasive and credible than those of ordinary media or politicians and can achieve the effect of refuting rumors. For pets, people are concerned about whether diseases can spread between people and pets. Because the related research is not complete, especially in the early stage of the epidemic, the medical resources were limited, and the associated research was limited. Until now, since there is not enough evidence to show whether all the pets can infect people. What people need to do is to care about pets who have been exposed to the patients for isolation observation and related environmental disinfection, rather than spreading rumors to abandon pets.

The third topic, it mainly about the symptoms and the situation in Wuhan. It is quite obvious why these two topics are concerned. People are worried about what kind of symptoms can be detected in the hospital and the research of asymptomatic patients. One aspect of the rumor is how infectious asymptomatic patients are. At the same time, as the most severe epidemic area in China, whether local residents or other places are very concerned about the situation of the epidemic and related news in Wuhan. Therefore, people who have an abnormal mind are more likely to spread rumors in this area.

The fourth topic is mainly about virus information and virus treatment. Because there is a lot of data about rumors, We also need some professional medical knowledge to refute the rumors. In other words, people are also concerned about the methods of treatment and prevention. Because the virus is scary, some lawless people in order to dump drugs or health care products, spread the rumors that related drugs are useful for the prevention or cure of COVID-19. However, up to now, no specific drug has been proven effective for the prevention of this disease. The government and relevant medical institutions should increase the publicity on this aspect and crack down on the rumor mongers so that the general public can not go crazy to buy drugs, or even take some harmful drugs.

The fifth topic is mainly about the way of virus transmission because the virus is primarily transmitted through droplets rather than simply left in the air, which is a piece of strong refutation evidence for many ways of virus error prevention. That's why it appears many times.

5 CONCLUSION

To sum up, in the above analysis, we mainly analyze the influence of the COVID-19 epidemic on the Chinese economy and open governmental information. In the process of spreading, we should recognize what the most rumors are about and find out the keywords and angles. On the one hand, we can realize what we care about most in the epidemic. On the other hand, it is also for the government and medical professionals to publicize and refute rumors and find the right direction. to prevent social panic and protect people's safety. From an economic point of view, we choose the SSE Composite Index and primary sector indexes to study the whole market and different fields. We hope to see how the global epidemic and Chinese epidemic affect Chinese economy, respectively, and the extent of their influence and the reasons behind their differences. That's why we used linear regression and the LDA model. In this part, we will classify and summarize all the analysis results.

From the perspective of transmission, according to the above analysis, because the epidemic is highly contagious and relatively low mortality, people are most worried about the transfer of the virus. In order to slow down the spread of the virus, we should control the source of infection, cut off the route of transmission, and protect the susceptible population. Since all people are susceptible to infection, the focus is mainly on the source and way of transmission. We can get the following conclusions through word map and LDA topic analysis.

From the perspective of infectious sources, we are concerned about what individuals can become contagious sources. We have known in the early days that the virus can spread from person to person. So whether pets, especially mammal pets, such as cats and dogs, can also spread the virus is a topic of concern. So there are a lot of rumors that pets can carry a lot of viruses. But at present, there is not enough evidence that pets can infect the virus. The government should guide people to disinfect pets and their living environment and avoid them listening to rumors and discard them at will.

In terms of the transmission way of the virus, we need to consider in which transmission way the virus will survive for a longer time and in which environment it can be killed entirely. Because people live in fear of the virus. For example, they are curious about how long the virus can survive in the air and how long it can survive in different clothing materials. Some rumors exaggerate the time virus can stay, which aggravates people's panic. On the other hand, we attach great importance to how to isolate the transmission channels. For example, masks are a common tool used in epidemic situations. How to use masks correctly and how to prolong the life of masks are what people are concerned about. At the same time, high temperature is also one of the ways people thought is efficient. However, because high-temperature antivirus has a high requirement, it is impossible to achieve this effect by using a hair dryer for a short time, which is said useful in many rumors.

In addition to the virus itself, there are many rumors about Wuhan. On the one hand, Wuhan is the place with the most severe epidemic in China. Besides, due to strict measures to close the city, people in other provinces and cities are more curious about the situation in this area, which also aggravates the spread of rumors. The virus comes from the laboratory, or the first patient has been found, but these are not scientifically verified and supported. Then we are supposed to talk about what role the government should play. After identifying specific rumors, we can formulate relevant measures based on these rumors. The measures mainly come from two aspects: one is propaganda and denying the rumors, the other is to solve people's actual needs. We can see many rumors about how to disinfect and use masks in the above analysis. At this time, the government needs to invite professional medical experts to

help people to better understand the virus and how to do self-protection on various platforms to help people better use PPE. The authoritative statement can also relieve people's anxiety to some extent. At the same time, during the period of Wuhan's lockdown, it is also necessary to strengthen the reporting of the situation in Wuhan so that people who worried about Wuhan can quickly learn the news rather than believing in rumors. At the same time, if the people in Wuhan can share their lives on social media, it will be more conducive to resist rumors, and also make it easier for people in other places to help. At the same time, we can see that the demand for masks shows that the masks' production capacity was insufficient at that time, so we need to increase the production capacity of masks and make a can afford PPE. Rumors about pets and their origins suggest that more research about the epidemic should be done, and the government should introduce the result of research on social media.

From an economic point of view, the core is to determine the impact of the global epidemic and China's epidemic on Chinese stock markets. From China's perspective, because the epidemic will have a substantial impact on the overall economy in the short term. In this crucial period, the macro-control policy of the government over the economy is fundamental to the rough the support of enterprises in crisis and unemployed people, as well as strengthening the activity of the market to avoid a more significant economic turmoil or even social crisis. The government can learn from the previous experience in dealing with SARS and H1N1 to understand the financial measures during the epidemic period suitable for the country.

From the perspective of the whole world, all industries of various countries are closely linked because of economic globalization. When they benefit from globalization, they inevitably need to bear the huge risks in the global financial crisis. Therefore, how to balance the national economic security with the issues that need to be considered by all countries in the opening-up mode. At the same time, we also need to pay attention to the impact of the epidemic on the global economic system. We cannot deny the impact of the COVID-19 on the economy. As the global value chain has exposed defects and vulnerabilities in the epidemic, countries may increase their intervention in industrial layout. Whether we should be alert to the trend of globalization is also a consideration for policy-making. From the two aspects of the stock market and false information, we can summarize the measures that the government should take in the epidemic as follows:

1. Actively understand people's needs and panic.

2. Set up more production lines to improve the production efficiency of products people need, meet people's needs and control the price of products.

3. The government should promptly publish the epidemic situation on multiple platforms, including medical data, the movement track of the confirmed personnel, etc.

4. We should organize and arrange authoritative experts in relevant fields to explain the medical knowledge related to the epidemic situation to help you protect scientifically.

5. For the blocked areas or areas with particularly serious epidemic situations, we should increase the reports on the disease and people's daily lives so that people throughout the country can rest assured, and it will reduce the spread of rumors.

6. The government should consider the impact of the epidemic on the economy and adopt relevant macro-control policies, such as supporting small and medium-sized enterprises, increasing infrastructure to solve employment problems, and other measures.

7. The government should pay attention to the protection of economic security during the epidemic because the global epidemic hit the industries of various countries and caused financial turbulence. So we need to make policies to guarantee the national financial security and solve the global crisis like foreign exchange management.

In general, since all observations and conjectures are based on short-term analysis, more data accumulation is needed for longer-term analysis. At the same time, if we can make a horizontal comparison with the data of other countries or other epidemic periods, we will have a deeper understanding of the trend of the stock market.

REFERENCES

[1] Gregory Carey. 1998. Multivariate analysis of variance (MANOVA): I. Theory. *Retrieved May* 14 (1998), 2011.

[2] N Fernandes. [n.d.]. Economic effects of coronavirus outbreak (COVID-19) on the world economy. 2020. *Available at SSRN* ([n. d.]).

[3] Rob J Hyndman and Anne B Koehler. 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22, 4 (2006), 679–688.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

[5] Marcus Richard Keogh-Brown and Richard David Smith. 2008. The economic impact of SARS: how does the reality match the predictions? *Health policy* 88, 1 (2008), 110–120.

[6] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*. 61–68.

[7] Stefano Ramelli and Alexander Wagner. 2020. What the stock market tells us about the consequences of COVID-19. *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever* (2020), 63.

[8] Gerard Salton and Clement T Yu. 1973. On the construction of effective vocabularies for information retrieval. *Acm Sigplan Notices* 10, 1 (1973), 48–60.

[9] Dayong Zhang, Min Hu, and Qiang Ji. 2020. Financial markets under the global pandemic of COVID-19. *Finance Research Letters* (2020), 101528.

Incorporating Expert Guidance in Epidemic Forecasting

Alexander Rodriguez*, Bijaya Adhikari†, Naren Ramakrishnan+ and B. Aditya Prakash*
*College of Computing, Georgia Institute of Technology
†Department of Computer Science, University of Iowa
+Department of Computer Science, Virginia Tech
Email: {arodriguezc, badityap}@cc.gatech.edu, bijaya-adhikari@uiowa.edu, naren@cs.vt.edu

ABSTRACT

Forecasting influenza like illnesses (ILI) has rapidly progressed in recent years from an art to a science with a plethora of data-driven methods. While these methods have achieved qualified success, their applicability is limited due to their inability to incorporate expert feedback and guidance systematically into the forecasting framework. We propose a new approach leveraging the Seldonian optimization framework from AI safety and demonstrate how it can be adapted to epidemic forecasting. We study a specific of guidance-smoothness, and show that by its successful incorporation, we are able to not only bound the probability of undesirable behavior to happen, but also to reduce RMSE on test data by up to 17%.

ACM Reference Format:
Alexander Rodriguez*, Bijaya Adhikari†, Naren Ramakrishnan+ and B. Aditya Prakash*. 2020. Incorporating Expert Guidance in Epidemic Forecasting. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

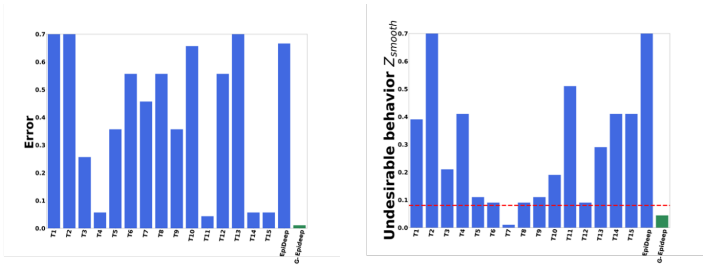
1 INTRODUCTION

Epidemic outbreaks incur heavy burden in terms of both health and economic costs (like the ongoing 2019-Covid corona virus epidemic). According to the world health organization (WHO), more than 15 thousand lives were lost due to the Ebola outbreak in West Africa between 2013 and 2016[1]. The economic cost of Ebola is estimated to be more than 53 billion dollars[2]. Timely forecasting of epidemic outbreaks is critical. Accurately forecasting various metrics of an epidemic outbreak informs practitioners and policymakers about impending scenarios and helps them devise strategic countermeasures, such as quarantining subpopulations, increasing vaccination availability, and school closures.

In this paper, we focus on influenza forecasting, motivated by the CDC FluSight prediction challenge [2] which seeks to predict the incidence of Influenza-like-Illnesses (ILI) in the US. Influenza is a major disease in the United States and beyond, causing thousands of fatalities every year. ILI is a symptomatic definition of illnesses that serves as a bellwether for real influenza incidence in a population. There has been a surge in recent research interest in influenza forecasting giving rise to a variety of mechanistic [22, 32] and statistical approaches [1, 5]. Mechanistic approaches predict influenza burden using simulation and aggregation of large epidemiological models. These models require a lot of calibration and hence are limited by their parameters to generalize well and fit the data [18]. Hence many researchers have begun exploring statistical approaches for this task, which train on historical ILI data and use the trained model to make forecasts for the current season.

Influenza seasons tend to be highly dynamic and have high variability due to numerous factors (e.g., weather, human mobility, virus strains circulating amongst the population) affecting the overall characteristics of the season. Moreover, different seasons and regions have different dominating influenza virus types. Further, the surveillance data collected (using ILINet) is a composite of multiple sources, is non-uniform, and is biased in many domain-specific ways. Hence while statistical approaches can frequently perform more accurate predictions than mechanistic models, they often show undesirable, unexplainable, or otherwise unexpected behavior.

For example, consider influenza incidence during the annual holiday season in the US. During this period, patients typically self-select and refrain from going to health providers, unless the situation is serious. This causes a temporary drop in recorded ILI incidence. However, as human mobility is high, flu activity rapidly



(a) Error (b) Guidance

Figure 1: Comparison of approaches in terms of (a) error in forecasting and (b) a guidance metric. In both plots lower is better. The red line in (b) is the threshold determined by guidance. T1 to T14 is the performance of teams participating in the 2015 FluSight challenge. Our method Guided-Epideep (GEpideep in the plot) is the only method which satisfies the guidance and gives the lowest prediction performance error.

[1] http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest
[2] https://www.reuters.com/

epiDAMIK 2020, Aug 24, 2020, San Diego, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

increases in the following weeks. This can not be modeled using standard mechanistic epidemiological models [19]. At the same time, statistical approaches 'over-correct' and exaggerate the temporary 'dip'. Hence if we can ensure that the forecasting model's predictions are reasonably 'smooth', such a behavior can be avoided. This 'smoothness' of the forecasts is well-motivated from other epidemiological considerations as well. As Figure 1 (b) shows, almost all the methods used in the 2015 CDC FluSight challenge show this 'lack of smoothness' behavior (lower is better).

To tackle such issues, in this paper, we propose incorporating *expert guidance* into statistical models for epidemic forecasting. In the case above, may be the expert can give the guidance that week-to-week forecast should be smooth, which can alleviate the over-correction problem. Indeed, incorporating this guidance helps our approach outperform the baselines while maintaining accuracy. Our approach 'Guided EpiDeep' is the only method to show desirable behavior (having guidance metric Z_{smooth} below the predefined threshold (red line)) while also getting the lowest errors (Figure 1).

To design a forecasting framework as envisioned here, there are several challenges. The first challenge is (a) how to design a general framework for any influenza statistical forecasting model to ingest and leverage expert guidance. Designing a general framework to incorporate guidance allows existing approaches to include expert guidance. The second challenge is (b) how to ensure that the framework is easy to use and generates useful feedback to the user. Moreover, the framework should communicate the extent to which the guidance was successfully incorporated and whether the guidance is helpful or not. Such a framework will aid in selecting guidance and make the forecast interpretable with respect to the guidance provided. None of the existing approaches is able to tackle these challenges.

In this paper, we leverage the Seldonian Optimization framework proposed in AI safety to enforce expert guidance (desired behavior) and prevent undesirable behavior. Our framework provides feedback to the user regarding the success or failure in the incorporating the expert insights. In case the framework fails to incorporate the insight, it communicates the failure to the user, who in turn can take steps to alter/improve the insight or change data or modify model hyper-parameters. Our contributions are as follows.

- **Novel method for incorporating expert guidance:** We explore a novel problem & adapt a successful framework to obtain domain-based consistency (and guidance), and perform extensive experiments to show properties of the framework.
- **Flexible user interaction framework:** The framework adapts to the user's requirements.
- **Real data case-study:** We present concrete case studies showing examples of expert guidance motivated by epidemiologist observation, and how our method helps to achieve experts requirements.

The rest of the paper is organized in the following way: we first motivate our problem, and formulate it. Then we present our method, and then empirical studies on real CDC data. We finally end with related work and conclusions.

2 PROBLEM FORMULATION

In this section, we introduce the novel problem of aiding statistical epidemic forecasting models with an expert's guidance. Before, we formalize our problem, we present the problem setting.

2.1 Epidemic Forecasting

Motivated by the setup of the CDC FluSight challenge, we study the epidemic forecasting problem from a temporal seasonality standpoint, such as in influenza. For this problem, we are given data $\mathcal{D}$ in the form of time series (e.g. the wILI burden per week for every season) and a predictive task $\mathcal{T}_w$, which sets what the target is and the time w (usually a week) when this prediction is to be made. Examples of targets are immediate-future incidences, peak intensity for season i, and the time when the peak value occurs.

The annual FluSight Challenge hosted by CDC asks to forecast metrics related to the current influenza season for the national and regional levels [2, 3]. The CDC releases influenza surveillance data, referred to as weighted Influenza-like Illness wILI, each week for every region. Given the latest partially observed influenza season, often represented as a time-series, the challenge asks to perform four different types of prediction tasks $\mathcal{T}_w$. They involve forecasting the incidence (wILI) value for the next four weeks, the onset of the season, the peak incidence value, and the time when the peak occurs. wILI incidence curves for each season since 1997/98 are publicly available[3].

2.2 Expert guidance

Expert guidance for epidemic forecasting is about leveraging multiple forms of domain knowledge and other preferences. An expert may want to guide a statistical model based on many considerations. Such considerations may include the epidemiology of the disease, characteristics of active virus strains (e.g. transmissability, reproduction rate), activity intensity in other other latitudes that dealt with the same virus strain, or efficacy of the vaccines to active strains of virus. It can also include some auxiliary knowledge. For example, it is well known that the Christmas holiday season in the US has specific impact on the flu spread which can not be captured by regular mechanistic epidemiological models [19]. It can include other public health policy considerations too, to ensure desirable behaviors like fairness in resource allocations.

As mentioned earlier, during the holiday season recorded epidemic activity temporarily drops due to patients' tendency to not seek healthcare. However, an expert notices that predictions of current statistical models are not 'smooth' i.e. they change a lot week-week and 'over-correct' during this time (a fact we demonstrate later in our experiments using the predictions of all the teams which participated in the CDC FluSight 2015 challenge). Hence s/he may want to more accurately forecast influenza incidence during the Holiday season incorporating the 'smoothness' property.

2.3 Desired Properties of Guidance

In this paper, we focus on incorporating such types of guidance into statistical epidemic forecasting models. Designing a framework which can incorporate guidance must be able to exhibit some ideal

[3] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

properties, especially as it is meant for experts who may not know the internals or any technical details of the statistical models.

- **P1.** Promote one or more desirable behaviors during training.
- **P2**. Have a mechanism to guarantee tolerance on deployment.
- **P3.** Be flexible to any generic ad-hoc guidance and be compatible with state of the art epidemic forecasting models.
- **P4.** Be easy to use for the user/expert.
- **P5.** Provide feedback to user if guidance could not be incorporated.

Let us discuss the underlying reasons behind the importance of each of the properties described above. P1 is an important facet of guidance: when training, the preference should be given to candidate models aligned with guidance goals. In addition, an ideal framework should be able to enforce more that one desirable behavior. Once the training completed, one can not expect that guidance will be met in unseen data at all times. Hence, it is natural to think about a probability of the trained model in meeting the guidance in unseen data. P2 sets guarantees on the expected probability of a model to exhibit the desirable behavior on unseen (test) data. P3 takes into consideration that experts' requirements may be related to any characteristic of the epidemic season. Furthermore, to leverage existing statistical forecasting models, the framework should provide a path to easily incorporate them. P4 aims to provide the user an easy interface to leverage the framework, treating the statistical model essentially as a black box. Finally, P5 importantly aspires to clearly communicate the result of attempting to incorporate the proposed guidance. Note that in our context, expert guidance can be motivated by practical considerations, and the data is really a composite signal (ILI cases rather than exactly flu cases). Hence sometimes expert guidance can indeed not be borne out by data, or be 'completely wrong' (unlike theory-guided data science [16], where scientific knowledge is considered ground truth) – so our framework needs a principled mechanism to signal this fact and provide feedback. The feedback provided opens possibilities to fruitful interactions as the expert may explore with different behaviors and tolerances to find the most suitable, and even test 'what-if' scenarios.

2.4 Definitions

Taking these properties in consideration, we make the following definitions to then state our problem.

Definition 2.1. Expert guidance: We represent expert guidance as a tuple $\langle g, \delta \rangle$, where $g : \Theta \rightarrow \mathbb{R}$ is a function that maps a candidate forecasting model $\theta \in \Theta$ to a measure of desirable or undesirable behavior of θ, and $\delta \in [0, 1]$ is a tolerance which constraints the probability of the model to exhibit this behavior.

Definition 2.2. Successful incorporation of guidance: We successfully incorporate guidance when we obtain a forecasting model θ for which our desired tolerance is met.

Note that our definition of expert guidance allows any framework which adopts it to exhibit all five desirable properties. Since the function g encodes one or more desirable behavior quantitatively, it can be used to enforce the behavior, satisfying P1. The parameter δ is the tolerance of undesirable behavior as mentioned in P2. Our definition of guidance is general enough to incorporate wide range of user insights to meet property P3. The only requirement is that the deviation from the desired behavior needs to be captured by the function g. Similarly, the user/expert do not need to be aware of underlying optimization framework and statistical model to incorporate the guidance as the function g is independent of both, satisfying P4. Similarly, if the value of the function g is greater than the threshold δ, the framework can communicate with the user regarding its inability to meet the guidance. We show how we can adapt our examples given before using our framework later (in Section 3.4).

2.5 Problem Statement

Having defined the notion of guidance that meets all the desired properties, we can state our problem as follows:

Guided Epidemic Forecasting: *Given a forecasting model which defines hypothesis space Θ, data $\mathcal{D}$, a predictive task $\mathcal{T}_w$, and expert guidance $\langle g(\theta), \delta \rangle$, we are required to return an optimal model θ, if found, that successfully incorporates expert guidance or return feedback that such θ could not be found.*

In this paper, the predictive task we consider is the future incidence forecasting. Our task $\mathcal{T}_w$ asks for influenza incidence at week $w + 1$ given that the incidence till week w is observed. And as the problem states, our goal is to enforce expert guidance, while solving for the predictive task. However our framework can easily handle other predictive tasks as well (like peak prediction etc).

3 OUR METHOD

As stated above, the Guided Epidemic Forecasting problem requires a base forecasting model upon which the guidance is enforced. To enforce the guidance, we need a framework which optimizes for performance with respect to the predictive task $\mathcal{T}_w$ as well as ensures that the constraint imposed by the guidance $\langle g(\theta), \delta \rangle$ is met. Here we leverage Seldonian Optimization which does this.

3.1 Seldonian Optimization

The Seldonian optimization framework [25] was recently proposed for Artificial Intelligence (AI) safety. It is designed to prevent AI models from showing undesirable behaviour such as gender or racial bias. Traditional AI algorithms optimize an objective function to select a model θ as a solution from the space of all possible models Θ. This framework precludes undesirable behaviour of AI model by enforcing behavioral constraints on the optimization objective. Hence, a probabilistic constraint is added to the optimization such that the probability that the value of a predefined undesirable behavior metric $g(\theta)$ is greater than 0. After training, to ensure the behavioral constraint will be met when the solution is deployed, this framework has a safety test mechanism, which is performed in unseen data. If the model meets the requirements of the safety test, the trained model is returned, else the framework returns no solution found (NSF).

A natural question that arises is what kind of base forecasting model (which is required by our problem) best works with the Seldonian optimization framework. Intuitively, models which learn/train by back-propagating errors are most suited for the Seldonian Framework as it learns through back-propagating as well. Hence, here we

chose a recently proposed deep learning based influenza forecasting model EpiDeep [1] as the base model upon which the guidance is to be enforced. We desribe Epideep briefly next. However we wish to emphasize that our framework is general.

3.2 EpiDeep

Epideep [1] is a recent deep neural architecture designed specifically for influenza forecasting. It exploits seasonal similarity between the current season and historical seasons via deep clustering [30]. The clustering module learns a latent low dimensional embedding of the seasons, such that the similarity between the seasons in the embedding space is meaningful for the task at hand. The clustering module in EpiDeep is designed such that it is possible to learn the embedding of the partially observed current season in the space of fully observed historical seasons. Epideep also uses long short-term memory (LSTM) [10] to encode in-season patterns of the current season. It then combines the embeddings from the clustering module and the LSTM and feeds the aggregated embedding to the decoder module, which make predictions for task $\mathcal{T}_w$. For the set of seasons $\mathcal{S}$ where each season $S \in \mathcal{S}$ is represented as a time series $S = s_1, s_2, \ldots, s_T$ in the training season, to predict the incidence observed in week w EpiDeep is trained with a loss function $\mathcal{L}(\theta) = \sum_{S \in \mathcal{S}} ||\hat{y} - s_w|| + \beta$, where $\theta \in \Theta$ is the trained model,$\hat{y}$ is the prediction made by θ and s_w is the observed incidence and β is the internal loss for Epideep not directly related to the task T_w. Note that while training only the weeks prior to week w is leveraged.

3.3 Expert-guided EpiDeep

The next natural question is how to adapt the Seldonian optimization framework to train EpiDeep with expert guidance. Before we answer that, let us define some notations. Let us have several different expert guidance to incorporate $\{\langle g_i, \delta_i \rangle\}_{i=1}^{n}$. We adopt the convention that if $g_i(\theta) \leq \epsilon$ for some small $0 \leq \epsilon$, the forecasting model θ does not exhibit undesirable behavior. Hence we impose probabilistic constraint on $g_i(\theta)$, on the model optimization. Hence, our updated optimization objective is as follows.

$$\arg\min_{\theta} \sum_{S \in \mathcal{S}} ||\hat{y} - s_w|| + \beta$$
$$s.t. \Pr\left(g_i(\theta) \leq \epsilon\right) \geq 1 - \delta_i, \forall i \in \{1, \ldots, n\} \quad (1)$$

Here, $\hat{y}$ is the prediction made by model θ for the prediction task $\mathcal{T}_w$. The objective above indicates that we want to ensure the probability that the desirable behavior (i.e, $g_i(\theta) \leq \epsilon$) occurs is greater than $1 - \delta_i$ for some small $0 \leq \delta \leq 1$, while the difference between predicted incidence value and the eventually observed value is minimized. Note that, we also want to ensure that the probability that the desirable behavior holds even in test/deployment stage.

Following [25], we ensure that our approach optimizes the objective while not violating the constraints and it generalizes to other unseen data with high confidence. It does so by dividing the given training data $\mathcal{D}$ into two partitions D_c and D_s. D_c partition is used for the model selection/optimization, while D_s partition is only used to verify that the guidance behavior is met in unobserved data. If the guidance behaviour is not met in D_s, then the framework ensures that no model is returned. In Algorithm 1, we leverage the Seldonian framework to design our algorithm Guided EpiDeep as follows.

Algorithm 1 Guided EpiDeep

1: **Input:** $\mathcal{D}$, $\langle g, \delta \rangle$, $U_{\mathcal{L}}$.
2: Partition $\mathcal{D}$ into D_c (for candidate selection) and D_s (for safety test)
3: $\theta_c \in \arg\min_{\theta \in \Theta}$ CandidateLossFunction$(D_c, \delta, \epsilon, U_{\mathcal{L}}, |D_s|)$
4: { Safety test using D_s }
5: **if** UpperBound$(\theta_c, D_s, \delta, U_{\mathcal{L}}) \leq \epsilon$ **then**
6: return θ_c
7: **else**
8: return No Solution Found (NSF)
9: **end if**

Here, the function UpperBound in line 5 measures if the behavior of candidate model θ_c is desirable as per the guidance provided. Based on predictions made by θ_c, variable Z is defined to quantify the deviation from the desirable behaviour for each prediction made. We discuss how variable Z is constructed for the guidance we consider in Section 3.4. Once Z is defined, we use compute UpperBound as suggested in [25]. We employ an empirical upper bound on the magnitude of $\mathcal{L}$, which is denoted as $U_{\mathcal{L}}$. This bound is necessary to prevent gradient explosion when switching losses in our CandidateLossFunction. Next we present the CandidateLossFunction subroutine in line 3 of Algorithm 2.

In the CandidateLossFunction subroutine, we use the D_c partition of the training data to train on both the objective with respect to the task $\mathcal{T}_w$ and to ensure that the returned model, θ is consistent with the guidance. To do so, variable $Z = \{Z_i | \forall i \in D_c\}$ is created using the predictions made by the model θ. Then the upper bound on variable Z is computed. If the upper bound computed is less than ϵ, indicating that the model is showing desirable behavior with respect to the guidance, then loss on $\mathcal{L}(\theta)$ is returned else, the loss on the bound is returned. Note that internally, $\lambda \in R_{>=0}$ balances the trade-off between loss and guidance.

Algorithm 2 CandidateLossFunction

1: **Input:** Candidate θ_c, D_c, $\langle g, \delta \rangle$, $U_{\mathcal{L}}$, $|D_s|$
2: Create an array of Z_i, where $i \in D_c$
3: $\hat{U}$ = PredictedBound$(Z_i, \delta, |D_s|)$
4: **if** $\hat{U} \leq \epsilon_i$ **then**
5: return $\sum_{S \in \mathcal{S}} ||\hat{y} - s_w|| + \beta + \lambda \frac{1}{|Z|} \sum_{i=1}^{|Z|} |Z_i|$
6: **end if**
7: return $U_{\mathcal{L}} + \hat{U} + (\lambda - 1)\epsilon$

Now, the question is how to define the variables Z for a given guidance. We discuss it next.

3.4 Constructing Behavioral Constraints

In this paper, we select two distinct expert guidance for seasonal influenza forecasting, namely smoothness and regional equity. In this section, we show the construction of constraint objectives for these expert guidance in the form of the Z variables. The guidance g and the variable Z are related to each other as follows:

$$g_i(\theta) = Z_i(\theta) - \epsilon \quad (2)$$

3.4.1 Smoothness. Mechanistic epidemiological models reveal that the epidemiological curves tend to be smooth with a single peak [23]. Hence, we expect epidemic influenza seasons to be generally smooth. In fact, we observe influenza incidence curve to be smooth with the consecutive values not changing drastically. Usually, incidence are low in the beginning of the season, they gradually rise till the peak is observed and then decline near monotonically. Hence, forecasting that the influenza incidence while ensuring that the predictions are smooth is a desirable property.

The smoothness also helps in correcting the drop observed in the influenza activity during the holiday season (discusses earlier), which arises due to the artifact of data collection. The existing approaches tend to overcompensate for the drop. Enforcing smoothness in forecasts prevents such undesirable behaviour. Here we describe smoothness as follows:

Definition 3.1. Smoothness is the max allowed difference ϵ between the predicted value and its predecessor.

We have a smoothness parameter ϵ, which is the maximum change allowed between current influenza incidence and the next incidence. In simple words, we want to ensure that the probability of smoothness function being greater than ϵ.

$$g(\theta) = E(|\hat{y}_{t+1} - Y_t|) - \epsilon \leq 0 \tag{3}$$

Here, $E(|\hat{y}_{t+1} - Y_t|)$ is the expected absolute difference between the predicted incidence $\hat{y}_{t+1}$ by the model θ and the last observed incidence Y_t. The guidance function $g(\theta)$ quantifies the smoothness by computing the difference between predicted and the previously observed value. Now, we have

$$E(|\hat{y}_{t+1} - Y_t|) \leq \epsilon \tag{4}$$

The equation above, highlights that the expected difference between the forecasted value and the previously observed value should be less than some constant ϵ ensuring that the forecast maintains week-to-week smoothness. Following this, we define the varaible Z for smoothness as follows:

$$Z_{\text{smooth}} = |\hat{y}_{t+1} - Y_t| \tag{5}$$

3.5 Expert Interaction

As mentioned in Section 2 two of the desirable properties of a framework to incorporate guidance is that it should be easy to use (P4) and should be able to provide feedback to user (P5). In this section, we present how our framework can be leveraged for exploration as well as demonstrate how an user might be able to interact with the framework.

From a user perspective, our framework provides three knobs: data, model, tolerance. An user is able to decide on how to partition the data into D_c for candidate model selection and D_s for safety test. Similarly, the user can decide on the base model suitable for the task at hand. The final knob corresponds to the tolerance with which the failure to incorporate the provided guidance is allowed. An expert/user can interact with the system by varying the values corresponding to the knobs.

Since our model has a mechanism for the safety test, it may return 'No Solution Found' (NSF) indicating that the guidance provided could not be met given the values of the knobs. If the model returns

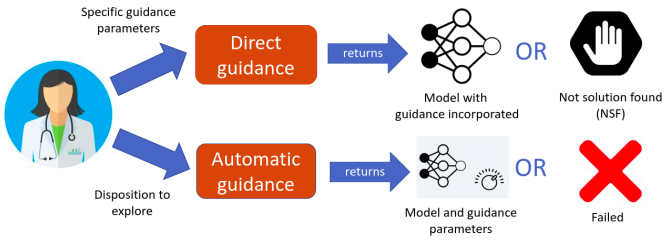


Figure 2: Flow diagram of expert interaction with Guided EpiDeep. Expert is given two modes: direct guidance and automatic guidance. The choice depends on the underlying motivation of the expert. Depending on the mode selected, the feedback is adapted to report success or failure.

NSF, it is an indication for the user to either consider the guidance provided or to vary the knobs. For example, if guidance related to smoothness is decided to be changed, this can be changed from $\epsilon = 0.5$ to $\epsilon = 1$. If tolerance is changed, confidence in guidance is changed. Hence, the model might be able to incorporate the guidance with a lower confidence on its generalizability. On the other hand, an expert can also change data by deciding to exclude some historical seasons that are preventing the guidance provided from being.

For ease of usage and interaction, our framework provides two modes of usages, namely Direct guidance and Automatic guidance, and depicted in Figure 2. We discuss each of the usages next.

3.5.1 Direct Guidance. In this mode, the user specifies guidance along with all the *all* parameters. Then our framework tries to incorporate the guidance within the constraints imposed by the parameters. If the framework fails to find a forecasting model which guarantees guidance incorporation, the the framework returns NSF. The direct guidance framework is presented in Algorithm 1.

3.5.2 Automatic Guidance. The user or epidemiological expert may not have data science/mathematical background to estimate the parameters with which the guidance can be incorporated and is willing to explore. Hence, in such cases, the framework tries to find the parameters which ensures that the guidance is met and the performance is maintained.

Our framework is able to provide such a exploration mode for users. Here the user may specify a *subset* of the parameters, and requirements in terms of performance. Our framework then explores the parameter space to find such a model. If none of the parameters explored is able to induce a model which satisfies the user requirements, the the framework returns NSF, indicating tha the guidance could not be incorporated. In this paper, as an example of automatic guidance, we ask our framework to explore the parameter ϵ_i such that no compromise is made in terms of RMSE.

4 EXPERIMENTS

4.1 Setup

We describe the experimental setup next. All experiments are conducted using a 4 Xeon E7-4850 CPU with 512GB of 1066 Mhz main memory. Our method is very fast, training for one prediction task

(on 1 week) in about 3 mins. We will release the code for academic purposes.

Data Here we use the weighted Influenza-like Illness (wILI) data released and updated by the CDC. CDC collects the wILI data through the Outpatient Influenza-like Illness Surveillance Network (ILINet) which consists of more than 3,500 outpatient healthcare providers all over the United States.CDC defines Influenza-like Illness (ILI) as "fever (temperature of 100◦F [37.8◦C] or greater) and a cough and/or a sore throat without a known cause other than influenza. Weekly wILI incidence curves for each season since 1997/98 are publicly available[4].

Research questions to address. In our experiments we want to compare the performance of our approach Guided-EpiDeep with the baselines for both the smoothness and regional equity guidance. We also want to evaluate the automatic guidance mode of Guided-EpiDeep. Specifically, we are interested in answering the following questions.

Q1. Is Guided-EpiDeep successful in incorporating guidance?
Q2. Does Guided-EpiDeep give feedback?
Q3. Is Guided-EpiDeep successful in realistic scenarios on real WILI data?

Evaluation. Here, we use the test data T, which is separated out during training to evaluate the performance of Guided-EpiDeep. Note that the test data is not used in either partition of the training data, namely D_c used for candidate model selection and D_s used for safety test.

To evaluate Guided-EpiDeep with respect to Q1, we will test if the guidance is incorporated in the test data. For Q1, we train the model on D_c and D_s to incorporate the guidance. Once the model is trained we evaluate whether the behavior of the model in forecasting influenza season in T is desirable with respect to the given guidance. To evaluate the degree to which the behaviour mandated by guidance is met in the test, we compute the probability that the behavior defined by the guidance $g(\theta)$, as defined in Section 3, falls outside the bounds. We name this metric as the failure rate of the model θ. Formally, we define the failure rate as $\Pr(g_i(\theta))$. To evaluate Guided-EpiDeep with respect to Q2 and Q3, we perform several case-studies.

Baselines. We use EpiDeep for performance and state of art baselines from the FluSight challenge for our case studies to show how they perform in a real-world scenario as posed by the CDC FluSight challenge. The complete list of teams we use is presented in the supplementary.

4.2 Direct Guidance

As mentioned earlier, in the direct guidance mode, the user/expert provides guidance as well as other parameters. Here, Guided-EpiDeep searches for the model which in able to incorporate the guidance within the constraints imposed by the parameters. For direct guidance, we evaluate Guided-EpiDeep in terms of Q1 and Q3.

4.2.1 Performance. Here, we want to quantify the rate at which Guided-EpiDeep is able to ensure that the behavior imposed by the guidance is met in the test set. To do so, here we split the historical seasonal influenza data into the training set D which consists 80% of the seasons and test set T, which has the remaining seasons. Guided-EpiDeep is trained on D with the smoothness constraint with a $\epsilon = 0.25$ and $\delta = 0.2$ to return a model θ. We repeat the experiment to make forecasts starting at week 40 of the epidemiological season till week 17. We then measure the failure rate, as defined earlier, of θ on the test set T for each week. Then we repeat the experiment with $\epsilon = 0.5$ and $\delta = 0.1$. The result is presented in Figure 3.

As seen in both Figure 3, for both settings, Guided-EpiDeep almost always ensures that the behavior imposed by the guidance is carried to the test data T. We observe that only 1 out of 80 observations, only one Guided-EpiDeep has a failure rate higher than the threshold δ. On the other hand, the baseline EpiDeep has significantly higher failure rate consistently, with the failure rate for many observations greater than δ. This experiment demonstrates that Guided-EpiDeep ensures that the desirable behavior is observed while forecasting on test data, while the baselines fail to do so.

Failure in Incorporation of Guidance. In the rare case when the Guided-EpiDeep fails to return a model (NSF) or the returned model does not ensure that the desirable behavior is observed in test data, as in week 52 in Figure 3 (left), the user is free to adjust one or more of the three knows our framework, data, model, and tolerance to allow the framework to search for a better forecasting model. For example, in the same example, setting a higher δ may ensure that the selected model satisfies the constraint in the test data as well.

4.3 Automatic Guidance

As mentioned earlier, in the automatic guidance mode, the user/expert provides the guidance. However, the other parameters are not known. Here, Guided-EpiDeep searches for the ideal parameter set which can incorporate the guidance. Here, for automatic guidance, we evaluate Guided-EpiDeep in terms of all the questions.

4.3.1 Performance. Here, we want to measure if Guided-EpiDeep can find a parameter which can satisfy the constraints imposed by the guidance provided by the user. Here we use the same setup as in Direct guidance. We split the data into training D and test T sets with 4:1 ratio. The expert's requirement here is to ensure that the performance of Guided-EpiDeep is better than the baseline model EpiDeep. We do so by enforcing that the ratio of RMSE of Guided-EpiDeep to the RMSE of Epideep is less than 1. And the parameter to explore/detect is ϵ.We repeated the experiment for each week in the influenza season. Figure 4 shows the result.

As seen in the figure, for most of the week Guided-EpiDeep is able to find an ϵ such that the constraint defined by the user is met. Among, 40 weeks Guided-EpiDeep fails to find ϵ in only 6 weeks, demonstrating that our framework is able to incorporate expert's guidance in the automatic guidance mode. For the weeks where ϵ could not be found, Guided-EpiDeep communicates its inability to find a solution to the user.

5 RELATED WORK

Epidemic Forecasting: Epidemic forecasting models and generally categorized into statistical [1, 6, 19, 26] and mechanistic based approaches [23, 32]. Ensemble of mechanistic and statistical

[4]https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

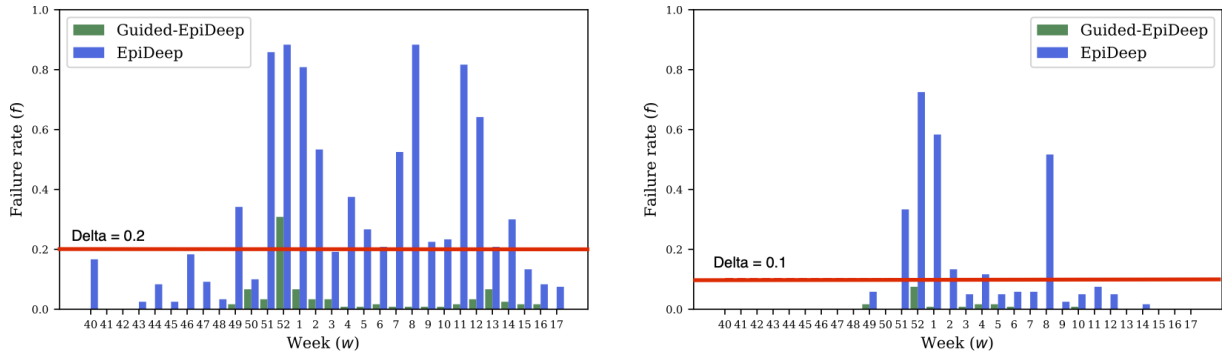


Figure 3: Performance of GUIDED-EPIDEEP in specific guidance. Figures show failure rate (f) for different combinations of ϵ and δ: (left) $\epsilon = 0.25$ and $\delta = 0.2$; (right) $\epsilon = 0.5$ and $\delta = 0.1$. Guided EpiDeep is successful incorporating expert guidance in epidemic task $\mathcal{T}_w$ for every week w in standard flu season as it is mostly within the bounds given by δ. Note that f in EpiDeep is higher than the required tolerance δ, but GUIDED-EPIDEEP is able to exhibit the desired behavior within the required tolerance.

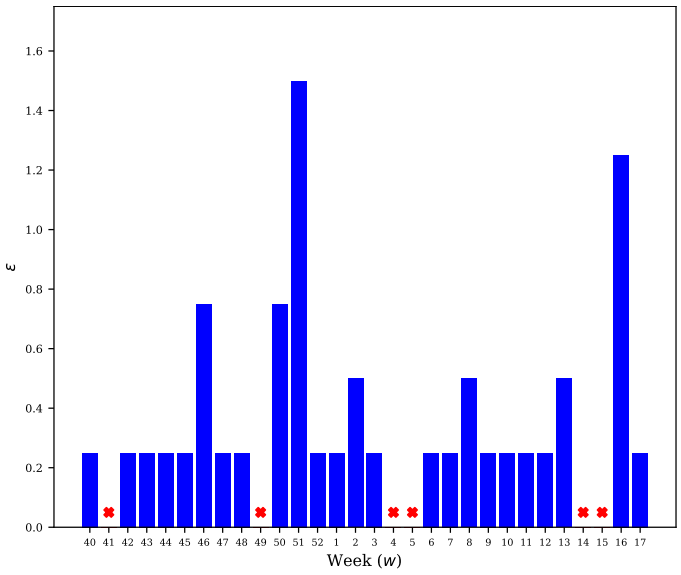


Figure 4: Automatic guidance over weeks. The y axis shows the value of ϵ found by GUIDED-EPIDEEP in automatic guidance mode. The red crosses represent the weeks where no suitable ϵ was found.

approaches too have been proposed [20]. There also has been interest in leveraging external data source in epidemic forecasting such as social media [8, 17], search engine [11, 31], environmental and weather reports [22, 24], and a combination of heterogeneous data [7].

Recently, there has been surging interest in leveraging deep learning for influenza forecasting.Adhikari et al. [1] proposed EpiDeep which leverages deep architecture to exploit seasonal similarity for epidemic forecasting. Similarly, Wang et al. proposed DEFSI [29] which exploits intra and inter seasonal data for forecasting. Other approaches like [27, 28] have limited use case (example, for military population) and/or require external data sources (example, twitter, weather). However, none of these approaches are able to incorporate expert guidance.

Time Series Analysis: A field related to o epidemic forecasting in data mining is Time Series Analysis. Several approaches have been proposed such as auto-regression, kalman-filters and groups/panels [4, 13, 21]. Several deep learning approaches have also been used for time series analysis [9, 12].

Guided prediction framework: The Seldonian optimization framework [25] discussed earlier presents a general framework for expert guided prediction framework. Based on the Seldonian framework, Metevier et al. proposed Robinhood, an algorithm for fairness in a bandit setting. Several other approaches have been proposed for specific fairness objectives as well [14, 15]. However, to the best of our knowledge, we are the first to introduce a guidance-based machine learning approach for epidemic forecasting.

6 CONCLUSIONS

In this paper, we study the novel general problem of incorporating expert guidance to statistical epidemic forecasting methods, using influenza prediction as an example. Leveraging the Seldonian optimization framework, we showcase a flexible, adaptable framework which can ensure that the given guidance can be incorporated subject to some probabilistic tolerance, whilst also maintaining performance accuracy. Additionally, our method also gives valuable feedback to the expert, if the guidance can not be successfully incorporated, to promote interactions. Via two natural guidance scenarios (smoothness and regional consistency) we show on real CDC surveillance data, that our method bounds the probability of undesirable behavior while also reducing RMSE by 17%. As future work, one can focus on extending this framework to more types of guidance, and also handling probabilistic predictions (as opposed to point predictions we considered here).

ACKNOWLEDGEMENTS

This paper is based on work partially supported by the National Science Foundation (Expeditions CCF-1918770, CAREER IIS-1750407, RAPID IIS-2027862, Medium IIS-1955883, DGE-1545362 IIS-1633363), ORNL and Georgia Tech.

REFERENCES

[1] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 577–586.

[2] Matthew Biggerstaff, David Alper, Mark Dredze, Spencer Fox, Isaac Chun-Hai Fung, Kyle S Hickmann, Bryan Lewis, Roni Rosenfeld, Jeffrey Shaman, Ming-Hsiang Tsou, et al. 2016. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC infectious diseases* 16, 1 (2016), 357.

[3] Matthew Biggerstaff, Michael Johansson, David Alper, Logan C Brooks, Prithwish Chakraborty, David C Farrow, Sangwon Hyun, Sasikiran Kandula, Craig McGowan, Naren Ramakrishnan, et al. 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* (2018).

[4] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

[5] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. 2015. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology* 11, 8 (2015), e1004382.

[6] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. 2018. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology* 14, 6 (2018), e1006134.

[7] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav V Marathe, et al. 2014. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining*. SIAM, 262–270.

[8] Liangzhe Chen, KSM Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B Aditya Prakash. 2016. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data mining and knowledge discovery* 30, 3 (2016), 681–710.

[9] Jerome T Connor, R Douglas Martin, and Les E Atlas. 1994. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks* 5, 2 (1994), 240–254.

[10] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

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

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

[13] Abhay Jha, Shubhankar Ray, Brian Seaman, and Inderjit S Dhillon. 2015. Clustering to forecast sparse time-series data. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 1388–1399.

[14] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2018. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 158–163.

[15] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

[16] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (Oct. 2017), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

[17] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1474–1477.

[18] Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses* 8, 3 (2014), 309–316.

[19] Dave Osthus, James Gattiker, Reid Priedhorsky, Sara Y Del Valle, et al. 2019. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Analysis* 14, 1 (2019), 261–312.

[20] Nicholas G Reich, Craig J McGowan, Teresa K Yamana, Abhinav Tushar, Evan L Ray, Dave Osthus, Sasikiran Kandula, Logan C Brooks, Willow Crawford-Crudell, Graham Casey Gibson, et al. 2019. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. *PLoS computational biology* 15, 11 (2019).

[21] Nicholas I Sapankevych and Ravi Sankar. 2009. Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine* 4, 2 (2009).

[22] Jeffrey Shaman, Edward Goldstein, and Marc Lipsitch. 2010. Absolute humidity and pandemic versus epidemic influenza. *American journal of epidemiology* 173, 2 (2010), 127–135.

[23] Jeffrey Shaman and Alicia Karspeck. 2012. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences* 109, 50 (2012), 20425–20430.

[24] James D Tamerius, Jeffrey Shaman, Wladmir J Alonso, Kimberly Bloom-Feshbach, Christopher K Uejio, Andrew Comrie, and Cécile Viboud. 2013. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens* 9, 3 (2013), e1003194.

[25] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.

[26] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine* 10, 1 (2012), 165.

[27] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony Maida, and Stephen Nichols. 2017. A novel data-driven model for real-time influenza forecasting. *bioRxiv* (2017), 185512.

[28] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12, 12 (2017), e0188941.

[29] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9607–9612.

[30] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. 478–487.

[31] Qingyu Yuan, Elaine O Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S Brownstein. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS one* 8, 5 (2013), e64323.

[32] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 311–319.