The 4th International Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK 4.0 @ KDD2021)

Bijaya Adhikari
University of Iowa

Ajitesh Srivastava
University of Southern California

Sen Pei
Columbia University

Sarah Kefayati
IBM

Rose Yu
University of California San Diego

Amulya Yadav
Pennsylvania State University

Alexander Rodríguez
Georgia Institute of Technology

Arvind Ramanathan
Argonne National Laboratory

Anil Vullikanti
University of Virginia

B. Aditya Prakash
Georgia Institute of Technology

ABSTRACT

The 4th epiDAMIK@SIGKDD workshop is a forum to discuss new insights into how data mining can play a bigger role in epidemiology and public health research. While the integration of data science methods into epidemiology has significant potential, it remains under studied. We aim to raise the profile of this emerging research area of data-driven and computational epidemiology, and create a venue for presenting state-of-the-art and in-progress results—in particular, results that would otherwise be difficult to present at a major data mining conference, including lessons learnt in the 'trenches'. The current COVID-19 pandemic has only showcased the urgency and importance of this area. Our target audience consists of data mining and machine learning researchers from both academia and industry who are interested in epidemiological and public-health applications of their work, and practitioners from the areas of mathematical epidemiology and public health.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning**; • **Applied computing** → Epidemiology;

KEYWORDS

epidemiology, public health, forecasting, AI for good

ACM Reference Format:
Bijaya Adhikari, Ajitesh Srivastava, Sen Pei, Sarah Kefayati, Rose Yu, Amulya Yadav, Alexander Rodríguez, Arvind Ramanathan, Anil Vullikanti, and B. Aditya Prakash. 2021. The 4th International Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK 4.0 @KDD2021). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Min-ing (KDD 2021), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3469475

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3469475

1 INTRODUCTION

The devastating impact of the currently unfolding COVID19 outbreak and those of the H1N1, Zika, SARS, MERS, and Ebola outbreaks over the past decade has sharply illustrated our enormous vulnerability to emerging infectious diseases. 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, CDC COVID-19 Forecasting Hub etc., 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 leveraging novel datasets. We believe the next stage of advances will result from closer collaborations between these two communities—the main objective of epiDAMIK.

Our target audience consists of data mining and machine learning researchers from both academia and industry who are interested in epidemiological and public-health applications of their work. Additionally, we are aiming to attract researchers and practitioners from the areas of mathematical epidemiology and public health, who are increasingly dealing with more complex models and novel data sources––these problems bring up novel challenges from a data mining and machine learning perspective.

2 WORKSHOP RELEVANCE

The current COVID-19 outbreak has already infected over 114 million people across six continents and has already impacted almost all aspects of modern life: the majority of countries are in recession, world economies are still struggling with unemployment (e.g. in the US increased from 3.7% in 2019 to 8.9% in 2020), and 2.5 million

people have perished due to the novel virus. Furthermore, COVID-19 is not the only disease outbreak to have occurred in the last decade which also saw other outbreaks including that of Ebola in 2013 and MERS in 2012. Moreover, seasonal influenza also has the potential to cause deleterious impacts on our society. For example, the 2017-18's flu season was severe (by some accounts, the worst since 2009), resulting in many lives lost and economic damage, which also coincided with the 100th year anniversary of the worst pandemic in human history (the 1918 influenza pandemic). Hence the impact of infectious disease epidemics has been in sharp focus worldwide with an intense interest in real rapid progress in this area from the public, government and academic stakeholders. State and federal public health agencies are becoming invested in data-driven approaches for infectious disease surveillance, forecasting and control, and funders are committing more resources to this important area. With an increasing focus on health in the data science community as well, we are well-positioned to further attract many researchers, including those who may not attend SIGKDD otherwise.

Relevance to SIGKDD Our workshop's relevance to the KDD community is significant and very timely, as 'data mining for public good' has been a major goal of data mining. We believe that a ACM SIGKDD workshop on this topic is needed because, while the integration of data mining into epidemiology has a large potential, so far it has been remarkably under studied; with this workshop we want to spark more attention on this topic. In fact, we were the first workshop to focus on bridging data mining and knowledge discovery and epidemiology. The ACM SIGKDD is the premier data mining conference, and hence a natural 'home' for the first such workshop of its kind.

3 TOPICS OF INTEREST

Topics of interest of epiDAMIK include, but are not limited to:

- Epidemiologically-relevant data collection
- Advances in modeling, simulation and calibration of disease spread models
- Syndromic surveillance using social media, search and other data sources
- Challenges in model validation against ground truth
- Outbreak detection and inference
- Visualization of epidemiological data
- Planning for public health policy
- Role of open source data and community in epidemiology
- Data-driven advances in control and optimization (like immunization)
- Forecasting disease outcomes including COVID-19 projections
- Graph mining and network science approaches to epidemiology
- Crowdsourced methods for detection and forecasting
- Use of novel datasets for prediction and analysis (including EHR records)
- Data mining data for hospital-acquired infections like C.diff, MRSA etc.
- Identifying health behaviors
- Handling missing and noisy data
- Disease forecasting challenge (like the CDC FluSight) experiences
- Any late-breaking work on the COVID-19 epidemic
- Interpretable and expert-driven AI for public health
- Experiences of real-time forecasting

4 PARTICIPATION AND REVIEW PROCESS

Each submitted paper received at least two blind reviews. The final acceptance/rejection decision is made by program chairs based on the reviews each paper received. Final decision also includes the information on whether each paper is accepted as a long paper, a short paper, or a poster. The accepted papers must be formatted according to the ACM SIG Proceedings template. All accepted papers will be featured in the workshop's website.

At least one author from each accepted paper must register, attend, and present their work at the workshop. The long papers will receive 20 minutes time to present their work and the short papers will receive 10 minutes. Additionally, there will also be a virtual poster session.

5 PROGRAM OUTLINE

epiDAMIK 4.0 is a full-day workshop. We are aiming to present a balanced program consisting of the following elements, allowing for ample opportunity for discussion and networking. We assume a 8hr workshop including a one-hour lunch break and two 30 minute coffee-breaks:

- Invited Talks (about 4, 45 minutes each, including questions)
- Contributed Papers (15 minutes for presentation, 5 for discussion)
- Poster Papers (45 minute poster and networking session)

As in our previous editions, we welcomed position papers and reports on high quality work-in-progress. In general, these lead to interesting and useful discussions, which in our case are specifically important in bringing the fields of data mining, machine learning and epidemiology together.

6 ORGANIZING COMMITTEE

The workshop organizing committee composed of the following:

Program Committee Chairs

- Bijaya Adhikari, UIowa
- Ajitesh Srivastava, USC
- Sen Pei, Columbia
- Sarah Kefayati, IBM
- Rose Yu, UCSD
- Amulya Yadav, Penn State
- Alexander Rodríguez, Georgia Tech

Steering Committee

- B Aditya Prakash, Georgia Tech
- Anil Vullikanti, UVA
- Arvind Ramanathan, Argonne National Laboratory