Accurately Estimating Unreported Infections using Information Theory

Jiaming Cui [1], Bijaya Adhikari [2], Arash Haddadan [3], A S M Ahsan-Ul Haque [3], Jilles Vreeken [4], Anil Vullikanti [3], B. Aditya Prakash [1]

[1]Georgia Institute of Technology
[2]University of Iowa
[3]University of Virginia
[4]CISPA Helmholtz Center for Information Security

May 3, 2025 @ SDM'25

Unreported infections

- One of the most significant challenges in combating against the spread of infectious diseases

Unreported infections

- One of the most significant challenges in combating against the spread of infectious diseases
- For example, large number of COVID-19 infections were unreported
  - Lack of testing
  - Asymptomatic infections

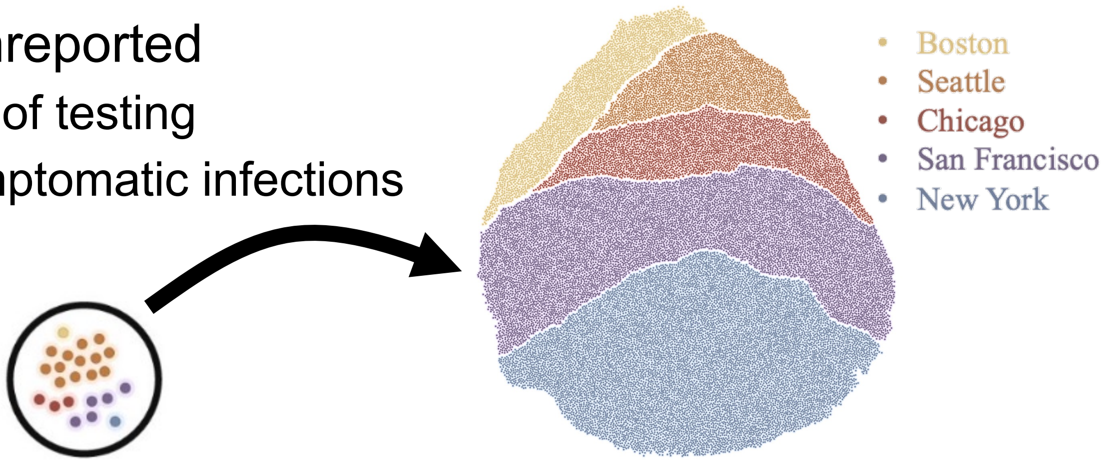
Unreported infections

- One of the most significant challenges in combating against the spread of infectious diseases
- For example, large number of COVID-19 infections were unreported
  - Lack of testing
  - Asymptomatic infections

Only 23 reported infections in 5 major U.S. cities by March 1, 2020.

Unreported infections

- One of the most significant challenges in combating against the spread of infectious diseases
- For example, large number of COVID-19 infections were unreported
  - Lack of testing
  - Asymptomatic infections



Only 23 reported infections in 5 major U.S. cities by March 1, 2020.

It was estimated that there were already more than 28000 infections.

Unreported infections

- One of the most significant challenges in combating against the spread of infectious diseases
- For example, large number of COVID-19 infections were unreported
  - Lack of testing
  - Asymptomatic infections

Inability in estimating the unreported infections allowed them to drive up disease spread in the U.S. and worldwide.

Only 23 reported infections in 5 major U.S. cities by March 1, 2020.

It was estimated that there were already more than 28000 infections.

Reported rate

- Epidemiologists use reported rate (α_{reported}) to capture total infections (i.e., both reported and unreported)

- Definition:

$$\alpha_{\text{reported}} = \frac{\text{reported infections}}{\text{total infections}}$$

Reported rate estimation methods

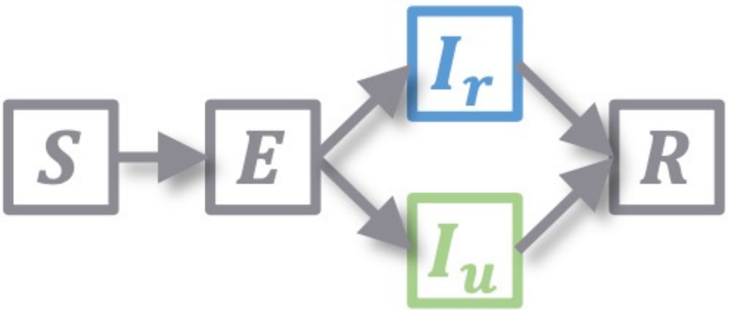
- Serological studies: Prevalence of antibiotics
  - Accurate, but expensive and have unavoidable delays
- Influenza surveillance systems
  - Suffer from ad-hoc corrections between COVID-19 and influenza
- Epidemiological models

Reported rate estimation methods

- Serological studies: Prevalence of antibiotics
  - Accurate, but expensive and have unavoidable delays
- Influenza surveillance systems
  - Suffer from ad-hoc corrections between COVID-19 and influenza
- **Epidemiological models**

Reported rate in epi models & estimation

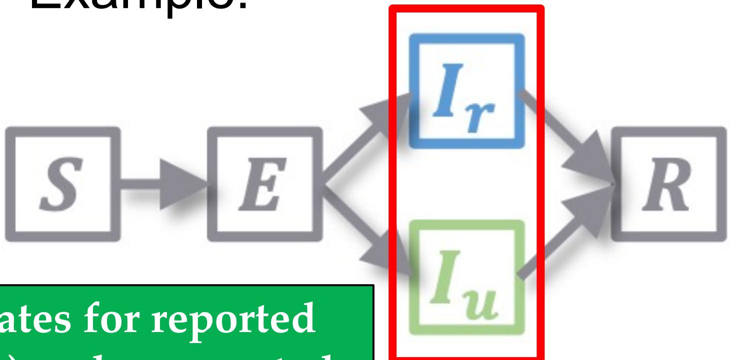
- Many epi models have reported rate as a parameter
- Example:



$$
\begin{cases}
\frac{dS}{dt} = -\beta \frac{SI}{N} \\
\frac{dE}{dt} = \beta \frac{SI}{N} - \gamma E \\
\frac{dI_r}{dt} = \alpha_{reported} \gamma E - \delta I_r \\
\frac{dI_u}{dt} = (1 - \alpha_{reported}) \gamma E - \delta I_u \\
\frac{dR}{dt} = \delta (I_r + I_u)
\end{cases}
$$

Reported rate in epi models & estimation

- Many epi models have reported rate as a parameter
- Example:

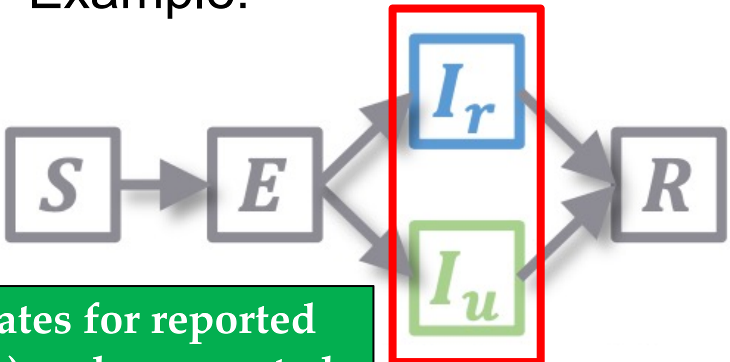


States for reported (I_r) and unreported cases (I_u)

$$
\begin{cases}
\frac{dS}{dt} = -\beta \frac{SI}{N} \\
\frac{dE}{dt} = \beta \frac{SI}{N} - \gamma E \\
\frac{dI_r}{dt} = \alpha_{reported} \gamma E - \delta I_r \\
\frac{dI_u}{dt} = (1 - \alpha_{reported}) \gamma E - \delta I_u \\
\frac{dR}{dt} = \delta (I_r + I_u)
\end{cases}
$$

Reported rate in epi models & estimation

- Many epi models have reported rate as a parameter
- Example:



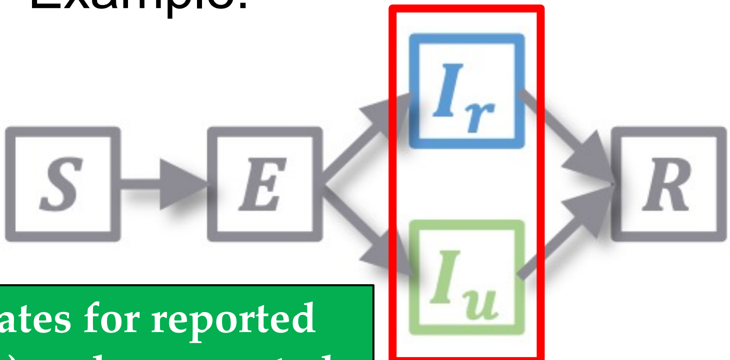
States for reported (I_r) and unreported cases (I_u)

$$
\begin{cases}
\frac{dS}{dt} = -\beta \frac{SI}{N} \\
\frac{dE}{dt} = \beta \frac{SI}{N} - \gamma E \\
\frac{dI_r}{dt} = \underline{\alpha_{reported}} \gamma E - \delta I_r \\
\frac{dI_u}{dt} = (1 - \underline{\alpha_{reported}}) \gamma E - \delta I_u \\
\frac{dR}{dt} = \delta (I_r + I_u)
\end{cases}
$$

$\alpha_{reported}$: reported rate

Reported rate in epi models & estimation

- Many epi models have reported rate as a parameter
- Example:



$$
\begin{cases}
\frac{dS}{dt} = -\beta \frac{SI}{N} \\
\frac{dE}{dt} = \beta \frac{SI}{N} - \gamma E \\
\frac{dI_r}{dt} = \underline{\alpha_{reported}}\gamma E - \delta I_r \\
\frac{dI_u}{dt} = (1 - \underline{\alpha_{reported}})\gamma E - \delta I_u \\
\frac{dR}{dt} = \delta(I_r + I_u)
\end{cases}
$$

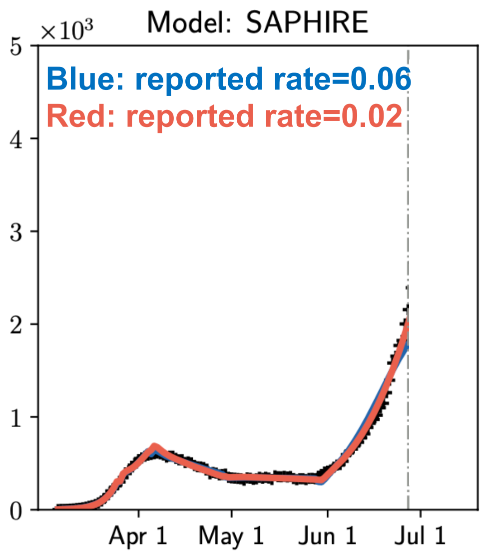
States for reported (I_r) and unreported cases (I_u)

$\alpha_{reported}$: reported rate

- Calibration: Fit I_r to reported cases to estimate the unknown parameters (including reported rate)

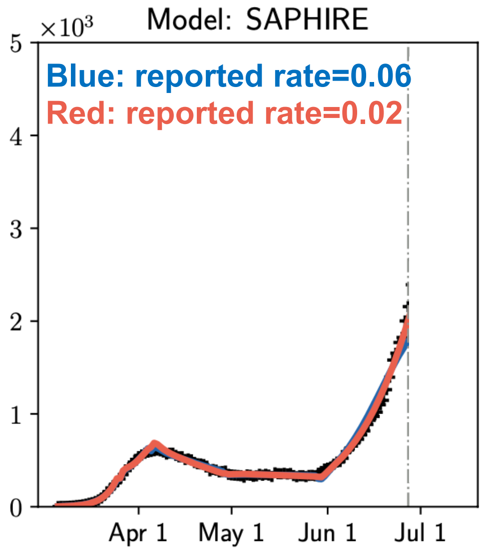
Sub-optimal calibration

- Parameter tuning is hard
  - Example: COVID-19 cases in Florida in 2020



Sub-optimal calibration

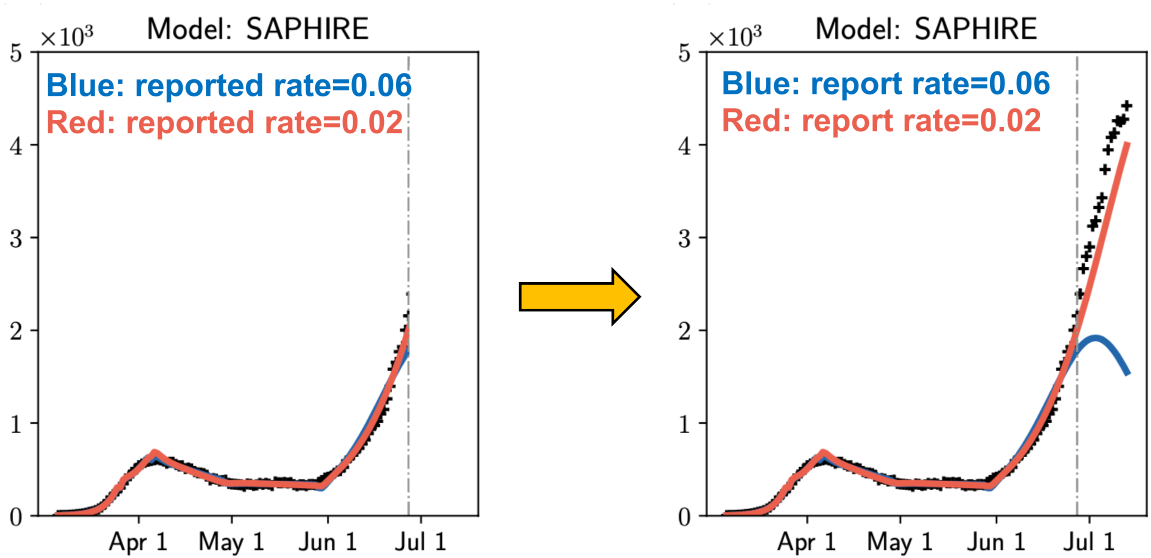
- Parameter tuning is hard
  - Example: COVID-19 cases in Florida in 2020



Both parameter fit the training data well.

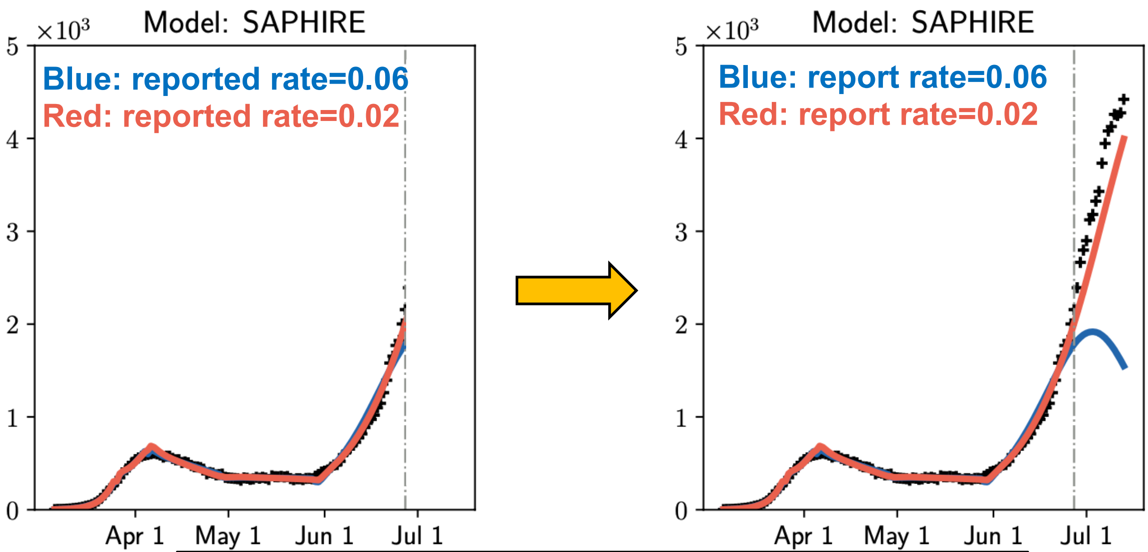
Sub-optimal calibration

- Parameter tuning is hard
  - Example: COVID-19 cases in Florida in 2020



Sub-optimal calibration

- Parameter tuning is hard
  - Example: COVID-19 cases in Florida in 2020



But their performance for forecasting are significantly different

Our goal

- Therefore, our goal is to estimate an **accurate** reported rate

Intuition

- We already know reported cases $D_{reported}$
- Imagine we are also given accurate values of total cases D
  - Then calibrating the model to $\left(D, D_{reported}\right)$ will lead to better fit of $D_{reported}$
  - We can also learn better parameters including $\alpha_{reported}$

- Hence, our problem can be stated as finding the D^* that fits the $D_{reported}$ best

Outline

- Introduction
- Two-part MDL: Sender-receiver framework
- MDLINFER: Information-theory based approach
- Experiment results
- Conclusions

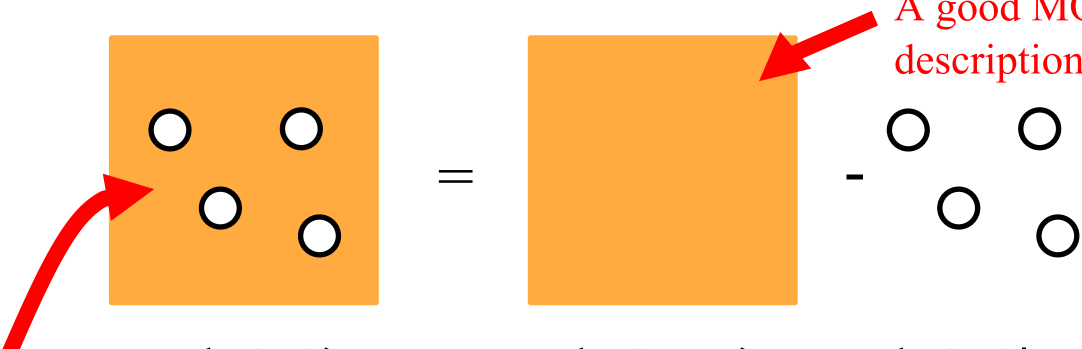
Two-part MDL: Sender-receiver framework

- Two hypothetical actors: Sender S and Receiver R
  - Sender S wants to send the “DATA” to Receiver R using a good “ MODEL”
  - To measure the “good fit”, use the number of bits to encode DATA: $\mathrm{L(DATA|MODEL)} + \mathrm{L(MODEL)}$

Two-part MDL: Sender-receiver framework

- Two hypothetical actors: Sender S and Receiver R
  - Sender S wants to send the "DATA" to Receiver R using a good " MODEL"
  - To measure the "good fit", use the number of bits to encode DATA: $\mathrm{L(DATA|MODEL) + L(MODEL)}$

A good MODEL makes description easier!



Describe it directly is hard!

$$\mathrm{L(DATA) = L(MODEL) + L(DATA|MODEL)}$$

Two-part MDL: Sender-receiver framework

- Two hypothetical actors: Sender S and Receiver R
  - Sender S wants to send the "DATA" to Receiver R using a good " MODEL"
  - To measure the "good fit", use the number of bits to encode DATA: $\mathrm{L(DATA|MODEL)} + \mathrm{L(MODEL)}$

A good MODEL makes description easier!

Sender S searches for the best possible MODEL, which minimizes the overall cost of encoding and transmitting both the MODEL and the DATA given the MODEL.

Describe it directly is hard!

$$\mathrm{L(DATA)} = \mathrm{L(MODEL)} + \mathrm{L(DATA|MODEL)}$$

Outline

- Introduction
- Two-part MDL: Sender-receiver framework
- **MDLINFER: Information-theory based approach**
- Experiment results
- Conclusions

MDLINFER: MODEL and DATA

- Recall that our problem can be stated as finding the D^* that fits the $D_{reported}$ best

- We also have

$$D_{\mathrm{reported}} = \alpha_{\mathrm{reported}} \times D$$

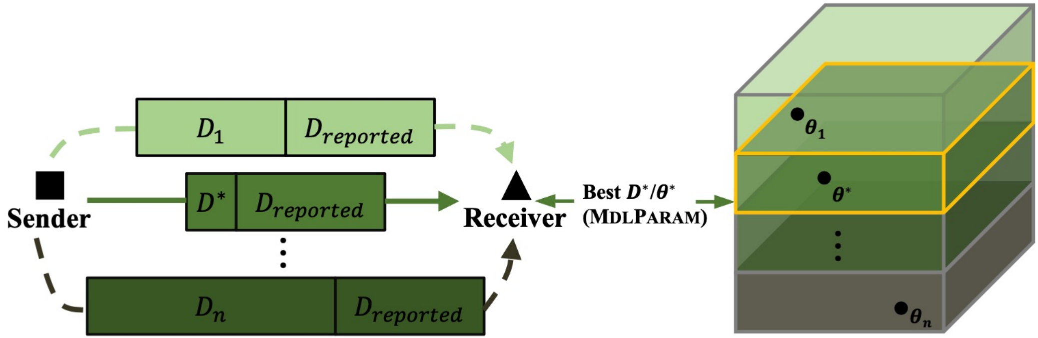
- Hence, we intuitively use
  - D_{reported} as DATA
  - D and $\alpha_{\mathrm{reported}}$ as MODEL

MDLINFER: Problem formulation

- Problem formulation

$$\boxed{D^* = \underset{D}{\operatorname{argmin}}\, L\left(D_{reported} \middle| D, \alpha_{\text{reported}}\right) + L(D, \alpha_{\text{reported}})}$$

- Here, $L(\cdot)$ denotes the number of bits for encoding



Use MDL Sender-Receiver framework to help find the latent variable D^* that explains $D_{reported}$ best.

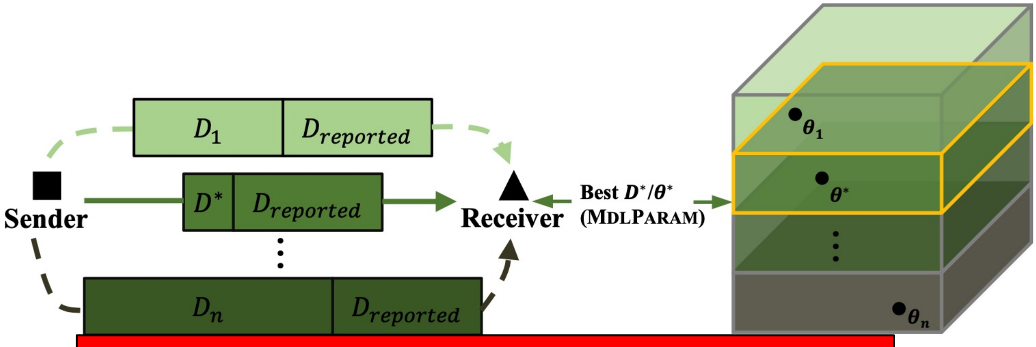
The latent variable D^* helps find more accurate θ^* for the epidemiological model.

MDLINFER: Problem formulation

- Problem formulation

$$D^* = \underset{D}{\operatorname{argmin}}\, L\left(D_{reported} \middle| D, \alpha_{\text{reported}}\right) + L(D, \alpha_{\text{reported}})$$

- Here, $L(\cdot)$ denotes the number of bits for encoding



Use M... the lat... ...le D^* helps ...te θ^* for the ...model.

Intuitively, by trying different D, we do fine-grain search in the search space.

Outline

- Introduction
- Two-part MDL: Sender-receiver framework
- MdlInfer: Information-theory based approach
- **Experiment results**
- Conclusions

Results

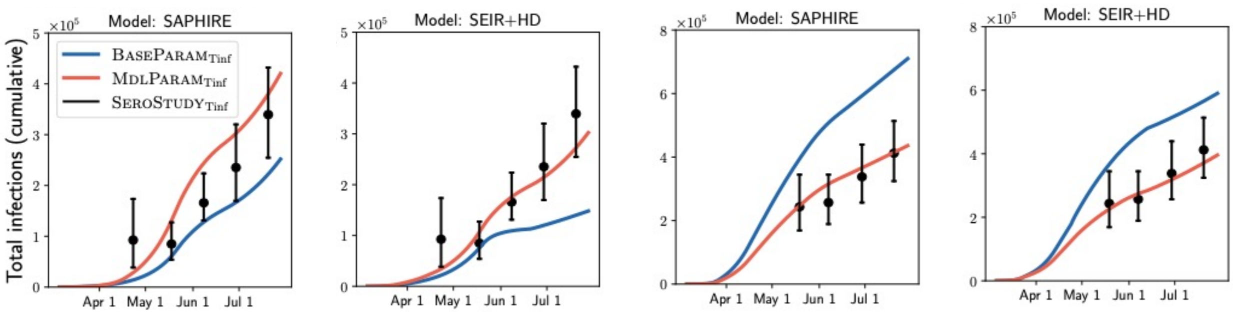
- More accurate estimation of total cases
  - Black: ground-truth total cases by serological studies [1]
  - Red for us, blue for current estimation methods

[1] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html

Results

- More accurate estimation of total cases
  - Black: ground-truth total cases by serological studies [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020 Philadelphia-Spring 2020

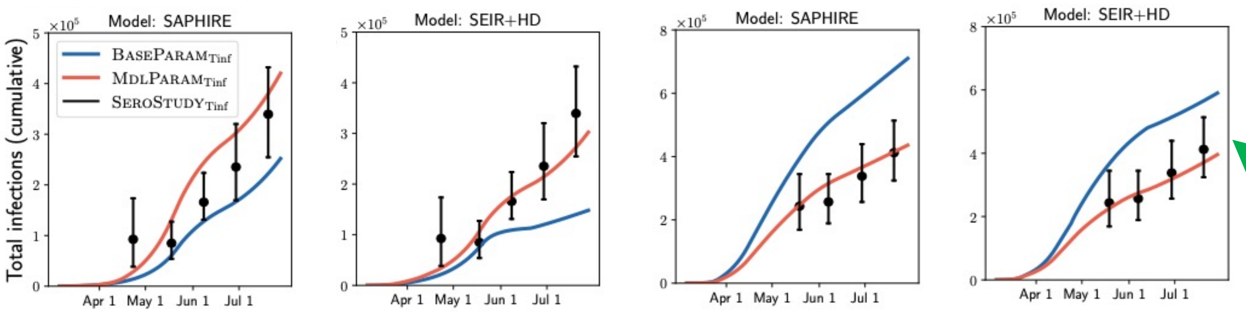


[1] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html

Results

- More accurate estimation of total cases
  - Black: ground-truth total cases by serological studies [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020 Philadelphia-Spring 2020



We show our performance on two different ODE models

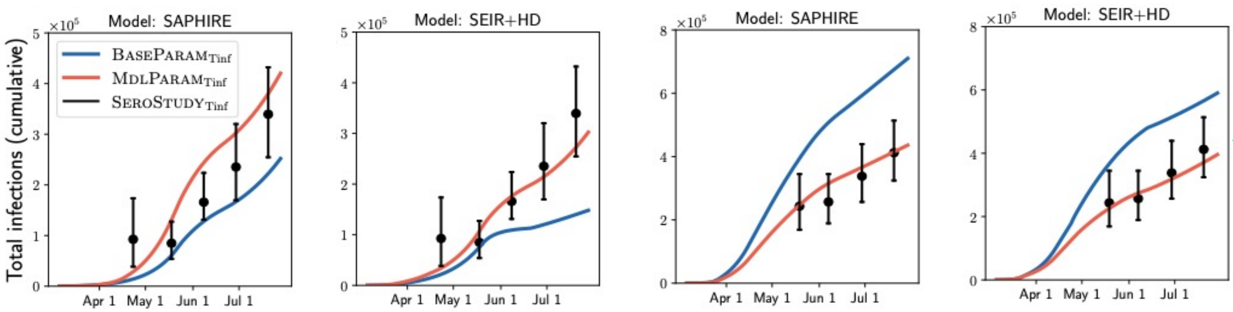
Our method (red) always fit the serological studies well

[1] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html

Results

- More accurate estimation of total cases
  - Black: ground-truth total cases by serological studies [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020　　　　Philadelphia-Spring 2020



We show our performance on two different ODE models

Our method (red) always fit the serological studies well

[1] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html

Results

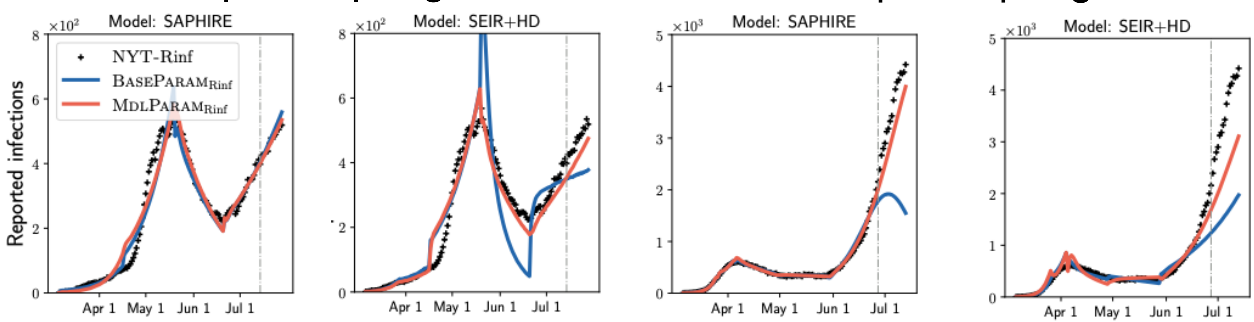
- Better fit of reported cases
  - Black for ground-truth reported cases by NY Times [1]
  - Red for us, blue for current estimation methods

The New York Times [1] https://www.nytimes.com/interactive/2021/us/covid-cases.html

Results

- Better fit of reported cases
  - Black for ground-truth reported cases by NY Times [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020 Philadelphia-Spring 2020

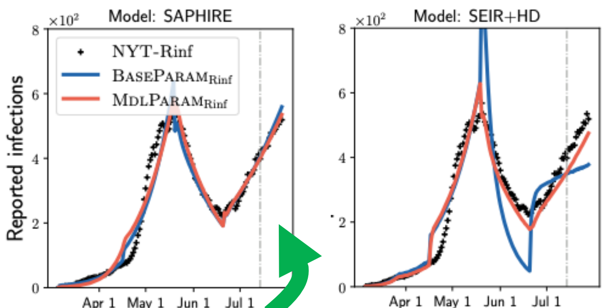


[1] https://www.nytimes.com/interactive/2021/us/covid-cases.html

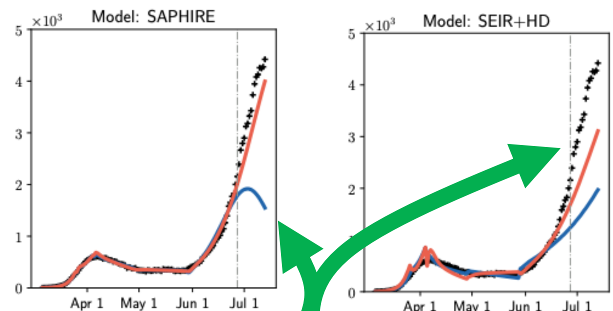
Results

- Better fit of reported cases
  - Black for ground-truth reported cases by NY Times [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020



Philadelphia-Spring 2020



Grey dash line divides observed period (left) and forecast period (right).

Good performance in predicting future!

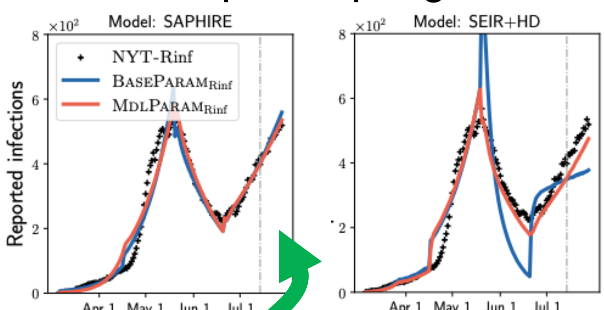
[1] https://www.nytimes.com/interactive/2021/us/covid-cases.html

More results for other locations in the paper

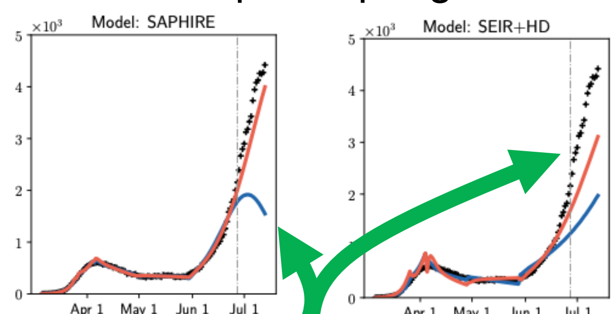
Results

- Better fit of reported cases
  - Black for ground-truth reported cases by NY Times [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring 2020



Philadelphia-Spring 2020



Grey dash line divides observed period (left) and forecast period (right).

Good performance in predicting future!

The New York Times

[1] https://www.nytimes.com/interactive/2021/us/covid-cases.html

Results

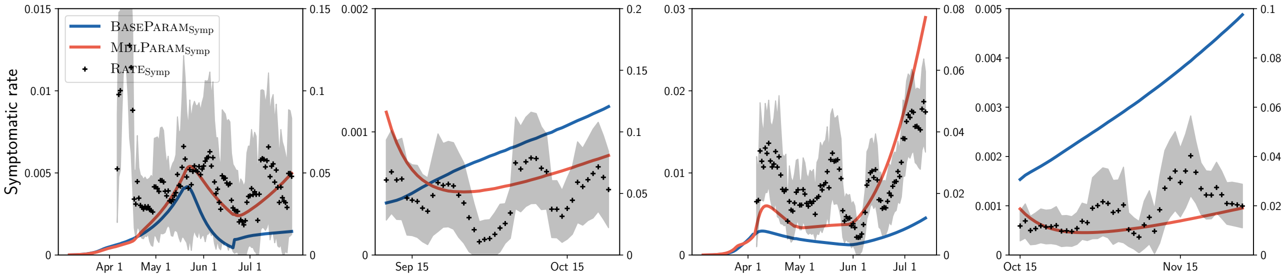
- Better estimate for symptomatic rate trends
  - Black for symptom trends by Facebook's survey [1]
  - Red for us, blue for current estimation methods

[1] Reinhart, Alex, et al. An open repository of real-time COVID-19 indicators. PNAS 2021.

Results

- Better estimate for symptomatic rate trends
  - Black for symptom trends by Facebook's survey [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring/Fall 2020 South Florida-Spring/Fall 2020

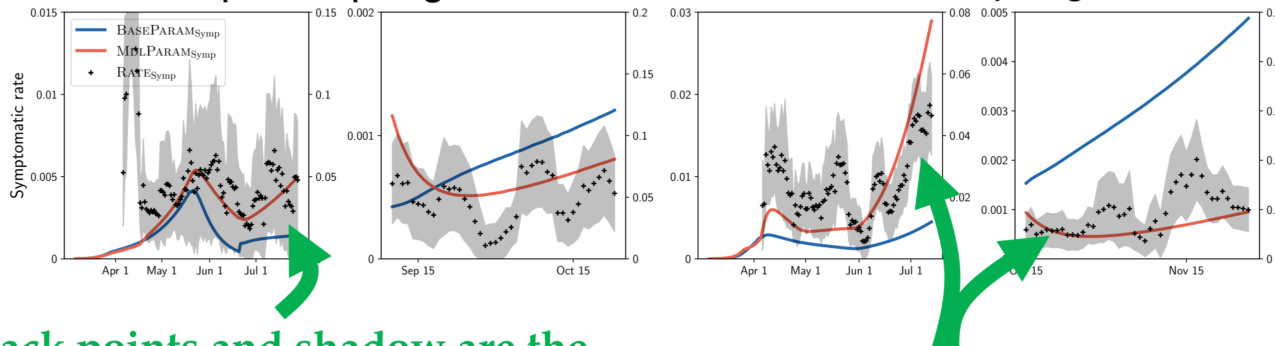


[1] Reinhart, Alex, et al. An open repository of real-time COVID-19 indicators. PNAS 2021.

Results

- Better estimate for symptomatic rate trends
  - Black for symptom trends by Facebook's survey [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring/Fall 2020 South Florida-Spring/Fall 2020



Black points and shadow are the point estimate and standard error

Good performance in capturing symptomatic rate trends

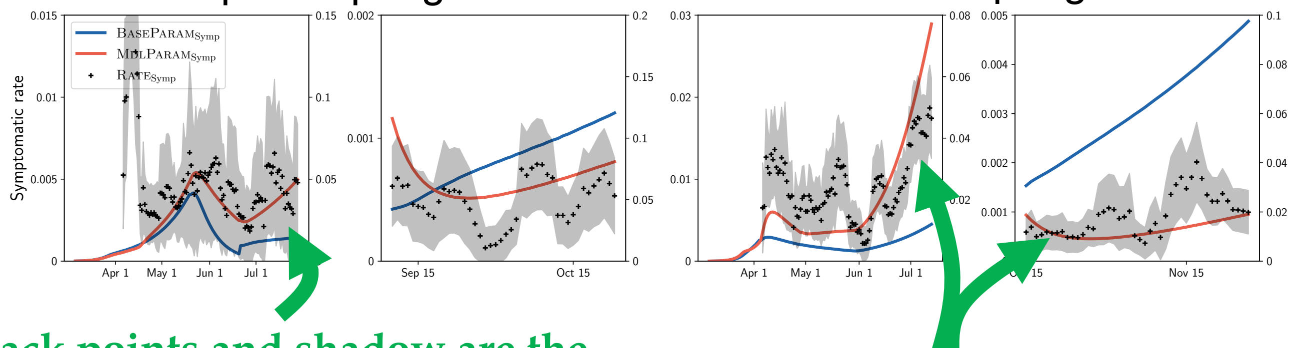
[1] Reinhart, Alex, et al. An open repository of real-time COVID-19 indicators. PNAS 2021.

More results for other locations in the paper

Results

- Better estimate for symptomatic rate trends
  - Black for symptom trends by Facebook's survey [1]
  - Red for us, blue for current estimation methods

Minneapolis-Spring/Fall 2020   South Florida-Spring/Fall 2020



Black points and shadow are the point estimate and standard error

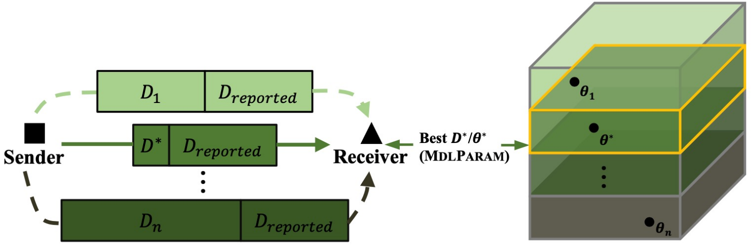
Good performance in capturing symptomatic rate trends

[1] Reinhart, Alex, et al. An open repository of real-time COVID-19 indicators. PNAS 2021.

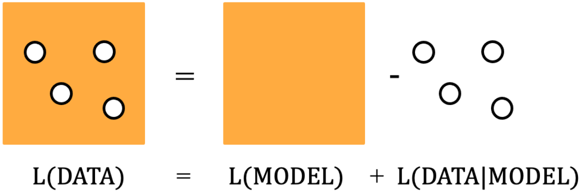
Outline

- Introduction
- Two-part MDL: Sender-receiver framework
- MDLINFER: Information-theory based approach
- Experiment results
- **Conclusions**

Conclusions

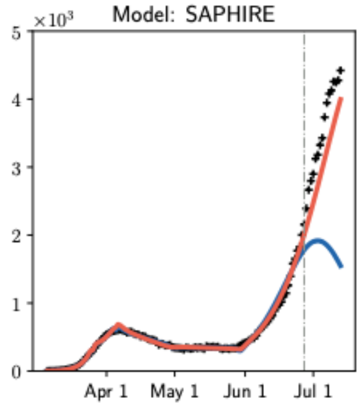


- We propose MDLINFER, a data-driven method to identify reported rate



- Leverage the information theory-based MDL framework

- Better performance in identifying total infections and forecasting future infections



Authors

Jiaming Cui

Bijaya Adhikari

Arash Haddadan

A S M Ahsan-Ul Haque

Jilles Vreeken

Anil Vullikanti

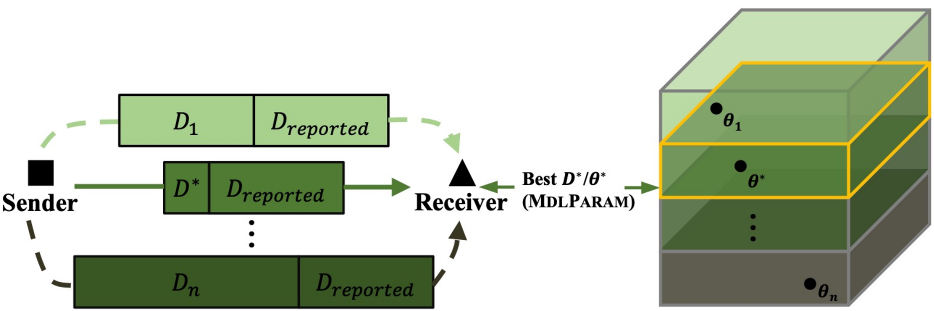
B. Aditya Prakash

: jiamingcui@vt.edu
: @jiamingcui1997

Thank you

- Code & papers available at: people.cs.vt.edu/jiamingcui/

Link to the paper:



Acknowledgements: NSF (Expeditions CCF-1918770 and CCF-1918656, CAREER IIS-2028586, RAPID IIS- 2027862, Medium IIS-1955883, Medium IIS-2106961, IIS-2403240, IIS-1931628, IIS-1955797, IIS-2027848 , IIS-2331315, PIPP CCF-2200269), NIH 2R01GM109718, CDC MInD program U01CK000589, ORNL, Dolby faculty research award, UVA GIDI, Georgia Tech.