

# INFECTIOUS DISEASE MODELLING IN THE TIME OF COVID

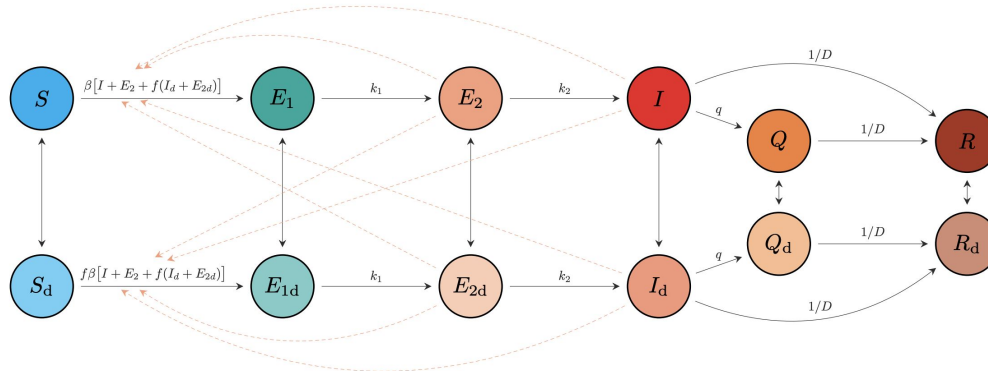
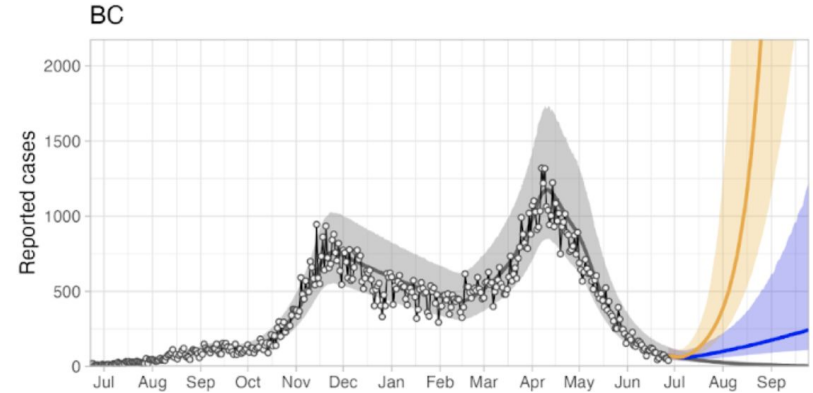
**Caroline Colijn**

Collaborations: Paul Tupper; Shraddha Pai (U of T)  
COVID Schools Canada; Pengyu Liu

**PRIMA CONGRESS 2022**

# A CLASSIC MATHEMATICAL BIOLOGY SETUP

- Mathematical model describes system dynamics in time (e.g. with an ODE model)
- Contrast: statistical model
- Explore model dynamics!
- Compare the model to data.
- Project forward in time



Schematic: ODE model with susceptible, exposed, infectious, quarantined, recovered individuals.

Distancing individuals (d) have less contact than others.

Model helps estimate impact of distancing measures, and makes projections.

# THIS TALK: AN OUTLINE

In the time of COVID, we need **different kinds of mathematical models**.

**Part 1** -- combining mathematical and statistical models helps us understand transmission using very limited data:

- **eventR**, a new R number we can use to model transmission at your event.
- **School transmission**: learning from readily available cluster size data.

**Part 2** -- mathematical frameworks using trees and tree metrics bring together very rich data and mechanistic models.

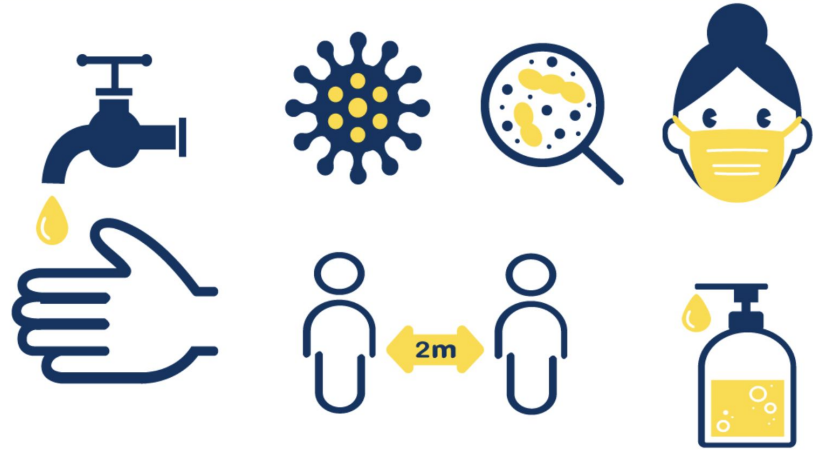
- **polynomials** uniquely define **unlabelled phylogenetic trees**
- we can use the polynomials to create **metrics** on these trees
- and we can use the metrics to compare mechanistic simulations to data

# PART 1. TRANSMISSION, LIMITED DATA, WHAT TO DO

It was summer, 2020. Vaccines were not available. The economy was reopening anyway. COVID-19 transmission was a problem.

People proposed plexiglass shields, handwashing, masks, “social bubbles”, distancing guidelines, capacity reductions.

Which ones should *you* use for *your* event, workspace? Why?



<https://campnurse.org/2021/03/25/npis/>

# FAMOUS COVID-19 OUTBREAKS

Morbidity and Mortality Weekly Report (MMWR)

CDC

## High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice — Skagit County, Washington, March 2020

Following a 2.5-hour choir practice attended by 61 persons, including a symptomatic index patient, 32 confirmed and 20 probable secondary COVID-19 cases occurred (attack rate = 53.3% to 86.7%); three patients were hospitalized, and two died. Transmission was likely facilitated by close proximity (within 6 feet) during practice and augmented by the act of singing.



News



## Almost an entire class of students caught coronavirus at a Trois-Rivières school

*Outbreak amid preventive measures illustrates the difficulty of controlling the spread of COVID-19 in a classroom of young children.*

Matthew Lapierre • Local Journalism Initiative Reporter

## *Days After a Wedding, a Dead Groom and Dozens of Coronavirus Cases*

Officials in India have opened an investigation into the gathering after more than 100 wedding guests tested positive for the coronavirus.

News | Coronavirus pandemic

## After one infected 16 at Berlin nightclub, coronavirus fears grow

*Panic spreads across Germany, with officials calling on people to avoid concerts, nightclubs and football games.*



# A SIMPLE MODEL FOR EVENT TRANSMISSION

Consider an event that lasts a total time  $T$ .

- If an infectious individual attends and is in contact with a single susceptible individual for a time  $\tau$  with a constant per unit time probability of transmission  $\beta$ , then the probability that the susceptible individual becomes infected is  $(1 - e^{-\beta\tau})$ .
- If they contact  $k$  others, the expected number of new infections is  $k(1 - e^{-\beta\tau})$ .
- If they mix around and contact  $T/\tau$  groups of  $k$  people, the expected number is

$$R_{\text{event}} = \frac{kT}{\tau} (1 - e^{-\beta\tau})$$

# REDUCING TRANSMISSION

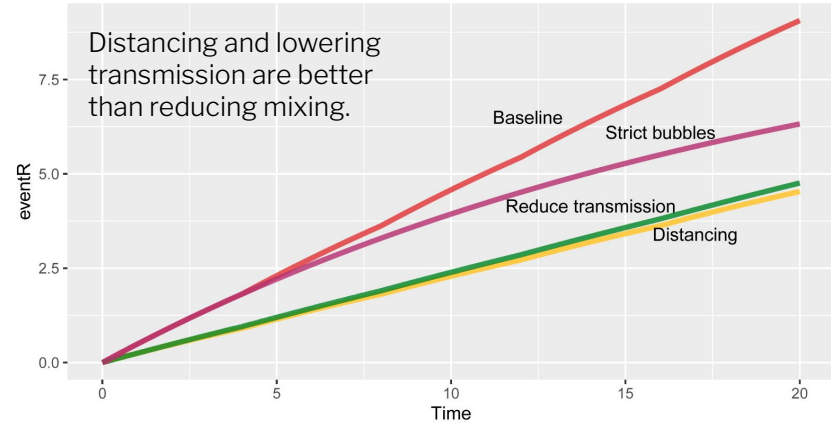
We can reduce  $R_{event}$  by changing:

- the crowding ( $k$ ): reduce capacity or density
- the level of “mixing”  $T/\tau$ : use “social bubbles”; people circulate less. In the model this means increasing  $\tau$ .  $\tau$  occurs twice in the equation.
- the duration of the event
- the per-person, per-unit-time, transmission rate: e.g. with masking

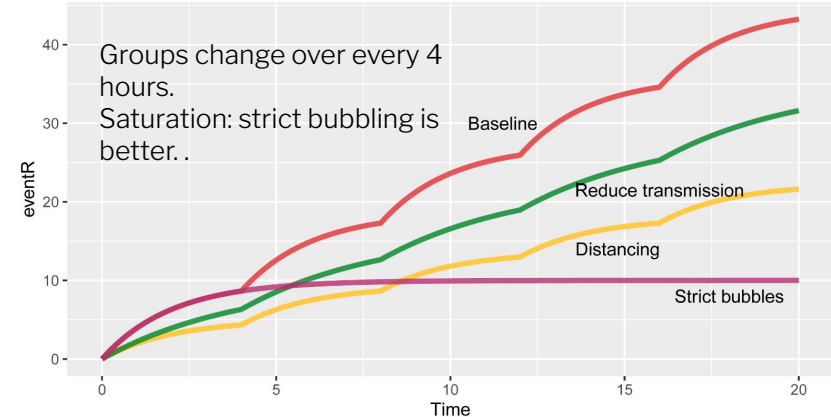
$$R_{event} = \frac{kT}{\tau} (1 - e^{-\beta\tau})$$

**Which of these interventions is better, when, and why?**

Low  $\beta$ : linear regime.



High  $\beta$ : saturating regime



# WE DO NOT KNOW THE TRANSMISSION RATE

We might know something about  $k$ ,  $T$  and  $\tau$ ,  
but we don't know the transmission rate,  $\beta$ .

We developed a method to estimate  $\beta$  from 15  
published outbreaks at events where there  
was some information for  $k$ ,  $T$  and  $\tau$ .

- 52 of 60 singers became infected after a choir rehearsal in Washington
- 5 of 39 passengers were infected in China when a man took a 2-h bus ride without a mask, whereas none of 14 passengers on his next 50-min bus journey were infected when he wore a mask
- 19 people were infected by a single individual in a nightclub outbreak.





# ESTIMATING THE TRANSMISSION RATE

$n_{inf}$  is the number of direct infections from the one person, at the event.

$R$  is its expectation.

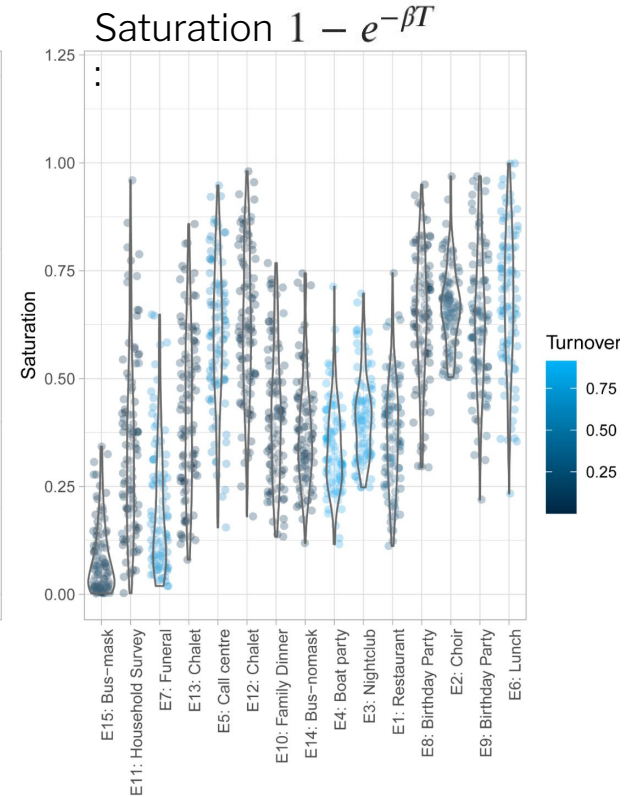
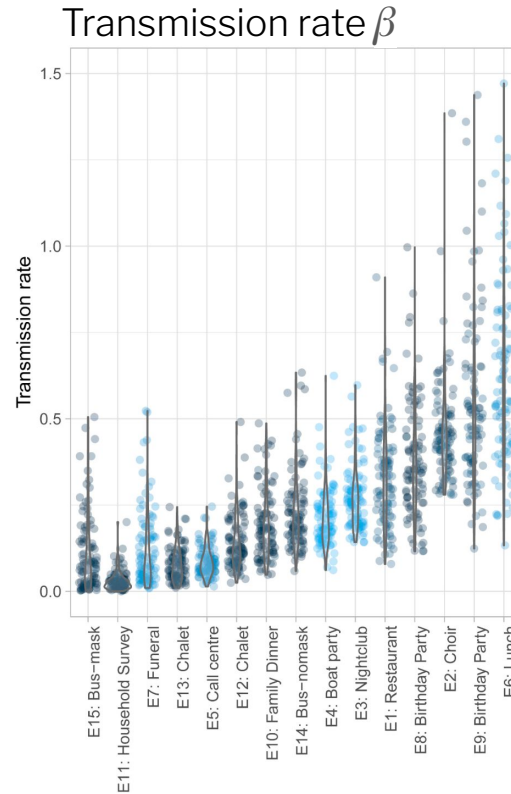
1. Fix  $T$ . Sample from appropriate normal distributions for  $k$  and  $\tau$ .
2.  $n_{inf}$  is an observation of a binomial random variable  $X$ :
3.  $p = (1 - \exp(-\beta\tau))$ ,  $n = kT/\tau$
4. We have  $\mathcal{P}(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$
5. Take a Bayesian approach, uniform prior: observe  $i = n_{inf}$ ; the likelihood of  $p$  is proportional to  $p^i (1 - p)^{n-i} = p^{n_{inf}} (1 - p)^{n - n_{inf}}$
6. This is a Beta distribution. Sample from it, and invert to get  $\beta$

# TRANSMISSION RATES AND SATURATION

Low range of transmission rate: 0.02-0.05 per contact per hour. Household, funeral.

High range: 0.5 p. c. p. hour (!). Choir, party, lunch.

Events with speaking, singing, and eating (parties, meals) generally higher than those without these.



IF YOU WANT TO READ MORE ABOUT eventR

RESEARCH ARTICLE | APPLIED MATHEMATICS | 

**PNAS**

# Event-specific interventions to minimize COVID-19 transmission

[Paul Tupper](#) , [Himani Boury](#) , [Madi Yerlanov](#) , and [Caroline Colijn](#)  [Authors Info & Affiliations](#)

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved November 2, 2020 (received for review September 15, 2020)

**November 19, 2020** | 117 (50) 32038-32045 | <https://doi.org/10.1073/pnas.2019324117>

<https://www.pnas.org/doi/10.1073/pnas.2019324117>

# BIAS

**Newsworthy outbreaks are larger than typical.**

Even if the transmission rate were the same for all similar events, **a larger observation is more likely to be reported.** (We found that this bias was likely to have been small).

And different choir practices (and index cases) will have **different transmission rates.** The transmission rate range for *this* choir practice is presumably larger than typical.

The value range we identified is possible, so it's relevant for understanding transmission. But it can't help predict what would happen in a *typical* choir practice.

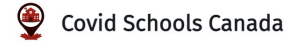
Adjusting for this would require data on exposures that did not lead to outbreaks, and on smaller outbreaks, systematically collected.

**FAST-FORWARD MORE THAN A YEAR...**

# SCHOOLS IN THE TIME OF COVID-19

Volunteers collected data: all confirmed COVID-19 cases and outbreaks in schools across Canada.

- Each report is linked to supporting evidence, such as a letter from a school board to parents, report on the school board or public health website, Provincial report, and/or media updates.
- Aim: increase transparency about the risk posed by COVID-19 in schools and to communities
- This includes exposures that did not lead to onward transmission, and those that did (outbreaks, or clusters in BC).

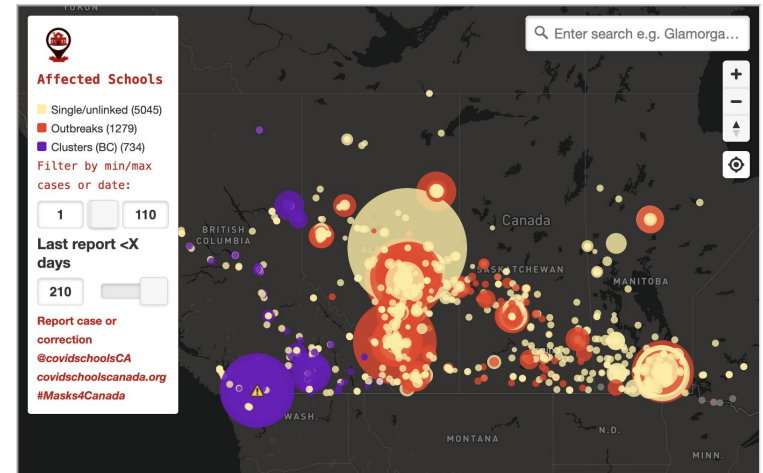


A Project of Masks4Canada

## Canada COVID-19 School Map

Move slides to change minimum number of cases and recency of reports. Circle size indicates total number of cases in the school

[Open in full screen](#)



<https://covidsschoolscanada.org/faq.html>

# THE SCHOOL DEBATE IN BRIEF (in 2020-2021)

## Don't close schools:

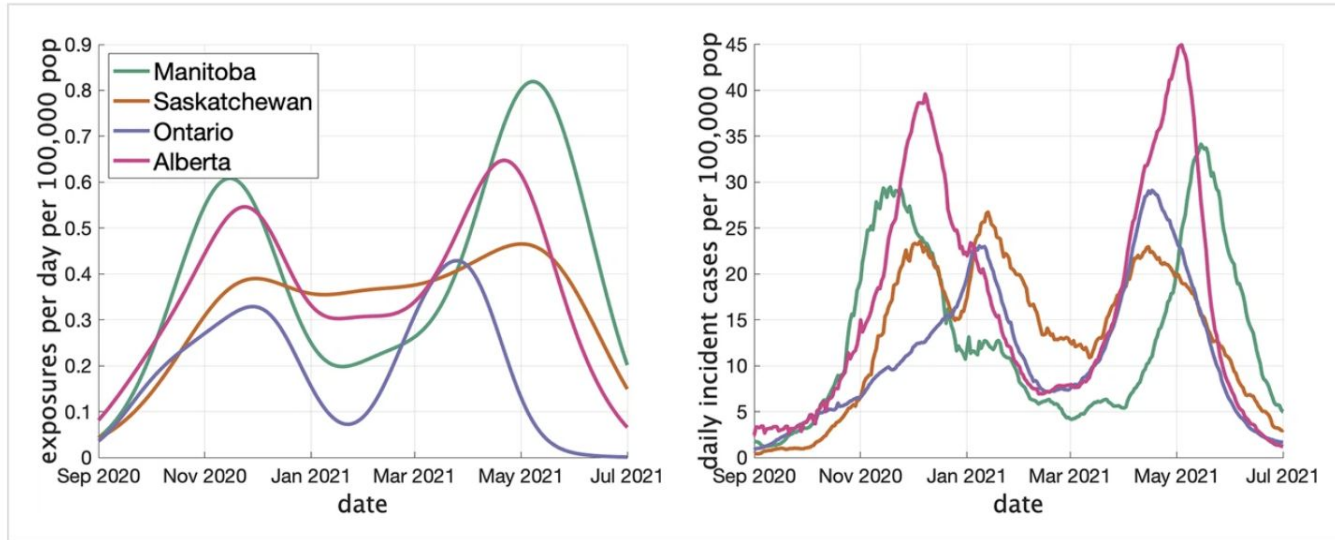
- Rates were not elevated (and often lower) among children compared to adults
- Children have less severe COVID-19, with lower hospitalization risk, more often asymptomatic
- Contact studies found many instances where children attended while positive with no onward transmission
- Schools are important and closing them has high social, economic and psychological costs

## Consider closing schools:

- Some outbreaks in schools and among children were very large
- Schools are part of our communities, and they contribute to transmission
- Children are impacted when their parents, grandparents, carers get COVID-19
- Children *can* get severe COVID-19, long COVID, MIS-C and other serious impacts

Studies on the impact of school closures on overall transmission have mixed results.

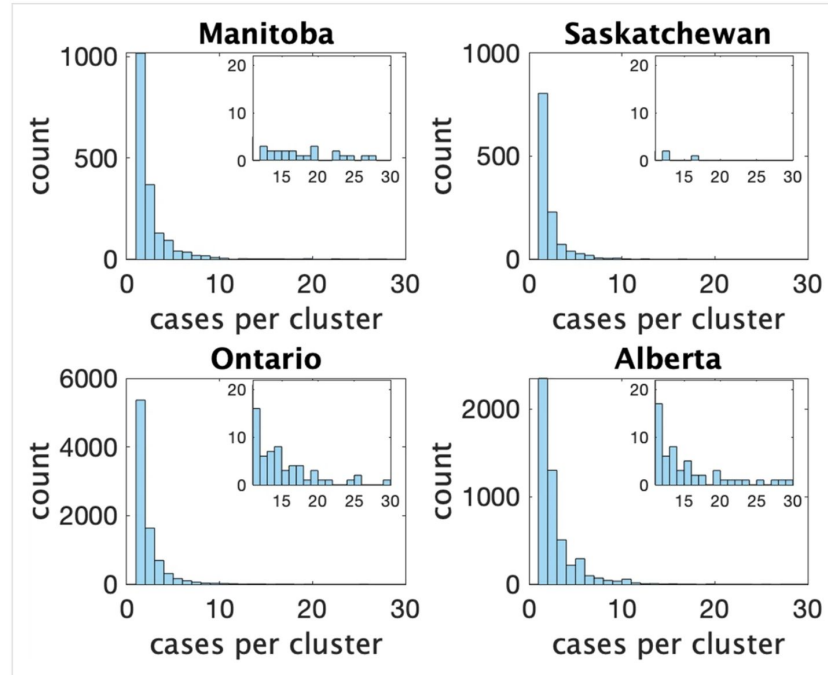
# NUMBERS OF CLUSTERS, NUMBERS OF CASES



**Two indicators of COVID prevalence over time in the four Canadian provinces.**

(Left) Estimates of the rate of new clusters (per 100,000 population) as a function of time in each province. (Right) Incident cases per day (per 100,000 population) in the same province over the corresponding time interval. Case counts are averaged over a 2-week window.

# DISTRIBUTION OF CLUSTER SIZES



Histograms of observed cluster sizes in four Canadian provinces.

Inset histograms only show clusters of size 11 or more on a different scale. Each dot represents a single cluster of size 11 or larger, and indicates the presence of (more rare) larger clusters.



# DISTRIBUTION OF CLUSTER SIZES

- Exposure events occur according to a Poisson process with variable rate.
- Independently of this process, once an exposure event occurs at a school, we say  $Z$  additional people are infected by the index case, for a total of  $Z+1$  individuals in the cluster.
- $Z$  includes individuals directly infected by the index case, as well as any subsequent infected individuals that are included in the same cluster.
- we model  $Z$  as a Poisson random variable with parameter  $\nu$ , where  $\nu$  itself is a Gamma-distributed random variable
- $Z$  is then a negative binomial random variable, with a mean  $R_c$  and a dispersion  $k$
- The cluster size is  $Z+1$

# DISTRIBUTION OF CLUSTER SIZES

The number of the  $Z+1$  cases that are actually *observed*,  $X$ , depends on testing. In our model: each case is observed with probability  $q$ . The observed cluster size  $X$  (conditioned on  $Z$ ) is binomial with parameters  $n=Z+1$  and probability  $q$ .

The probability that in this model a cluster has  $i$  individuals is

$$W_i = \sum_j \binom{j+1}{i} P(j|R_c, k) q^i (1 - q)^{j+1-i}$$

where  $P(j)$  is negative binomial. We cannot observe 0 individuals, so we condition on observing at least 1 to get  $P(i|i \geq 1)$

The log likelihood is

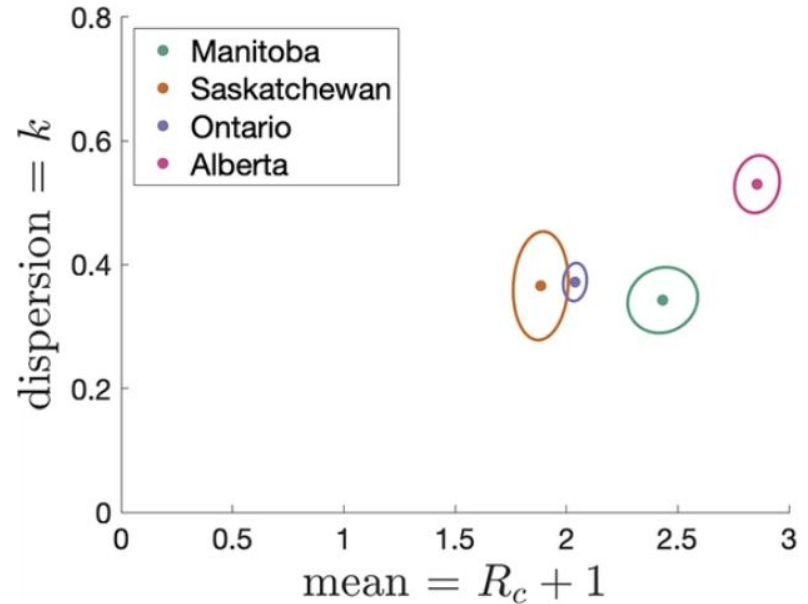
$$\prod_{\text{clusters}} P(i_c | i \geq 1)$$

# CLUSTER SIZES AND WHAT THEY MEAN

There are 1-2 new infections per exposure if we are finding 75% of them.

If we are finding 50%, it's 2-5 new infections per exposure, on average.

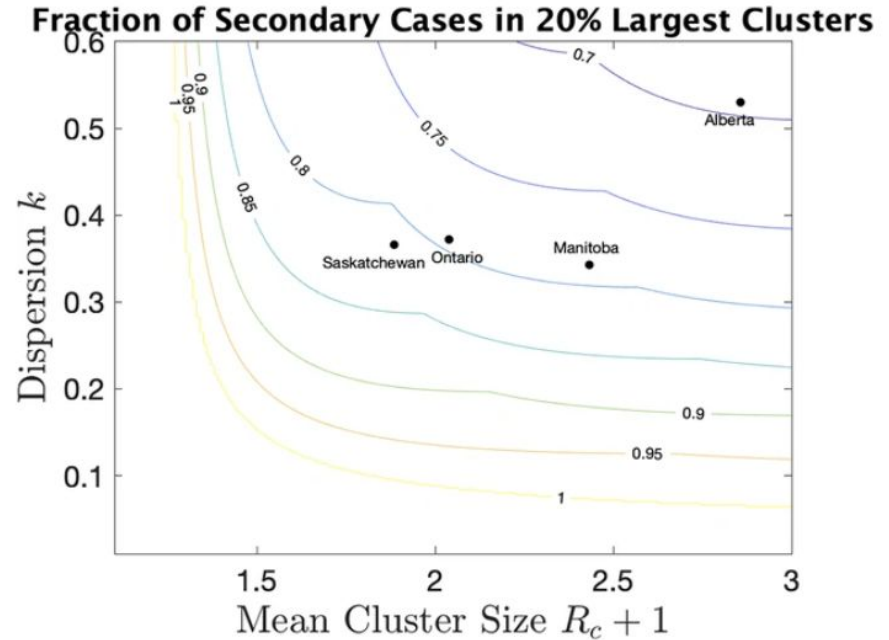
But there are plenty of exposures that lead to no onward transmission.



# FRACTION OF CASES IN THE LARGEST CLUSTERS

80% of the secondary cases occur in the largest 20% of clusters.

If we were finding 50% of the cases, we would infer that it's even more skewed, with 85-90% of secondary infections in the largest 20% of clusters.



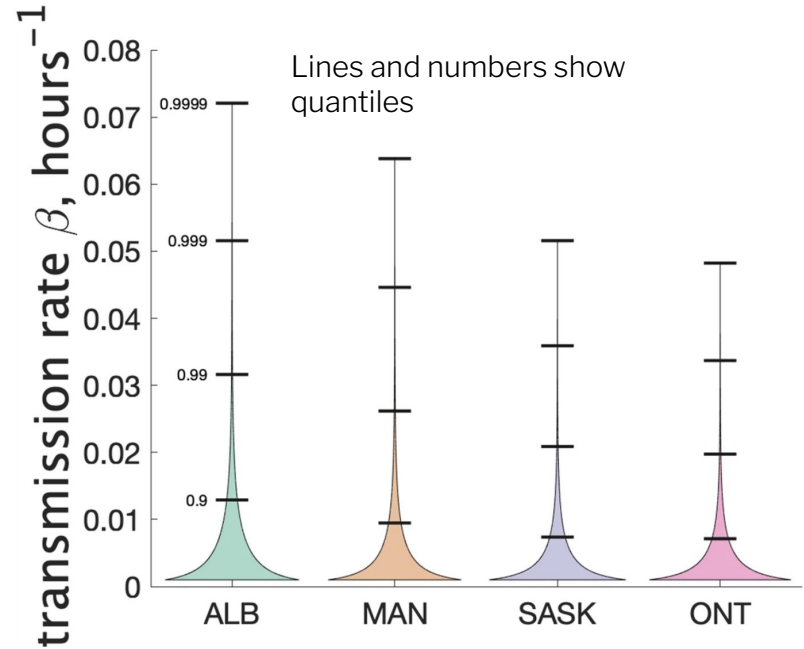
# TRANSMISSION RATES

We estimated the Poisson parameter  $\nu$  in the cluster distribution.

To estimate the transmission rate, assume 25 contacts per classroom and 12 hours of exposure before an infectious individual is identified/removed.

Transmission rates are highly variable.

Medians are 0.00074 - 0.002 per contact per hour. The highest median is in Alberta.

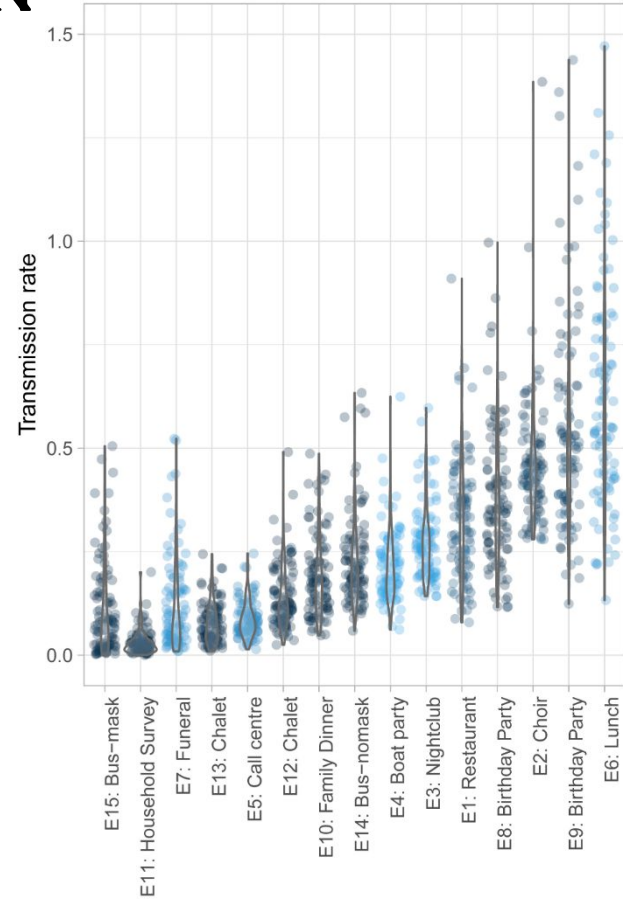
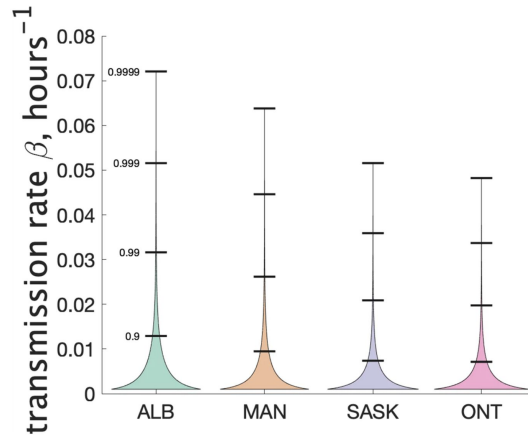


# TRANSMISSION RATE COMPARISON

Can both be true? I think so.

In the famous superspreader outbreaks: adults; no mitigations in place; newsworthy.

In the school clusters: children; mitigations in place (reduced hours, distancing measures, reduced risky activities); not necessarily newsworthy!



# IF YOU WANT TO READ MORE ABOUT SCHOOLS

Research Article

Epidemiology and Global Health, Microbiology and Infectious Disease



## COVID-19 cluster size and transmission rates in schools from crowdsourced case reports

Paul Tupper , Shraddha Pai, COVID Schools Canada, Caroline Colijn 

Department of Mathematics, Simon Fraser University, Canada;  
Institute for Cancer Research, Canada

**PLOS** COMPUTATIONAL BIOLOGY

<https://elifesciences.org/articles/76174#digest>

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

### COVID-19 in schools: Mitigating classroom clusters in the context of variable transmission

Paul Tupper , Caroline Colijn

Published: July 8, 2021 • <https://doi.org/10.1371/journal.pcbi.1009120>

# SUMMARY OF PART 1

We used novel combinations of simple statistical and mathematical models to learn about transmission.

We estimated a quantity that is almost never available, yet fundamental for modelling: the rate of transmission per contact per unit time.

We did this with readily available, and very limited, data: cluster sizes, instead of individual-level data or cases through time.





## PART 2: THE OTHER DATA EXTREME

Viruses, bacteria and other pathogens have the potential to change the world.

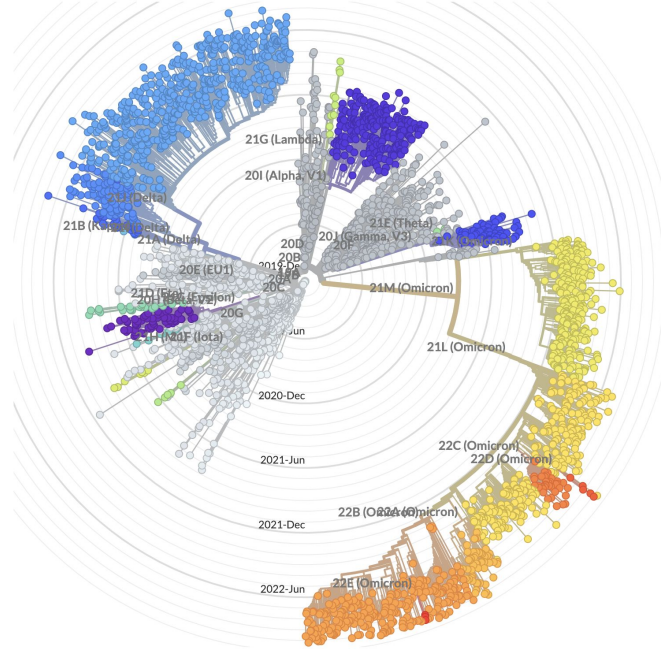
We have new data to learn about them at very high resolution:

# SEQUENCE DATA



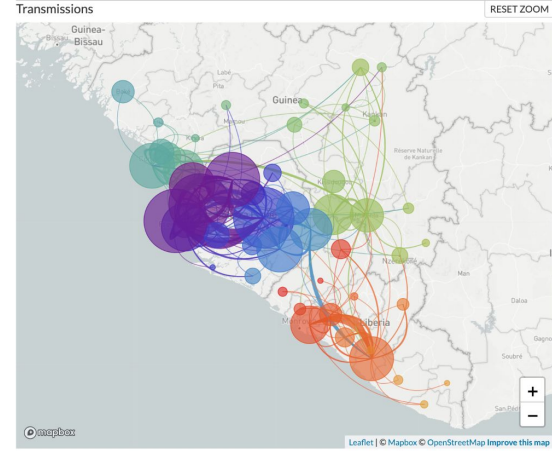
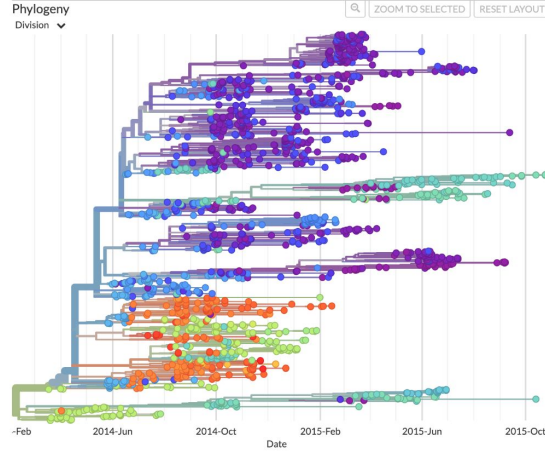
# SEQUENCE DATA AND VIRUSES

COVID-19



## Genomic epidemiology of the 2013-2016 West African Ebola epidemic

Built with [nextstrain/ebola](#). Maintained by [Trevor Bedford](#).  
Showing 1493 of 1493 genomes sampled between Mar 2014 and Oct 2015.



hCoV-19 data sharing via GISAID

**13,597,843**  
genome sequence submissions

# NEED FOR NEW MATHEMATICAL IDEAS AND TOOLS

Why?

- Describe evolution
- Forecasting! What's going to happen?
- How is diversity maintained?

Inform decisions:

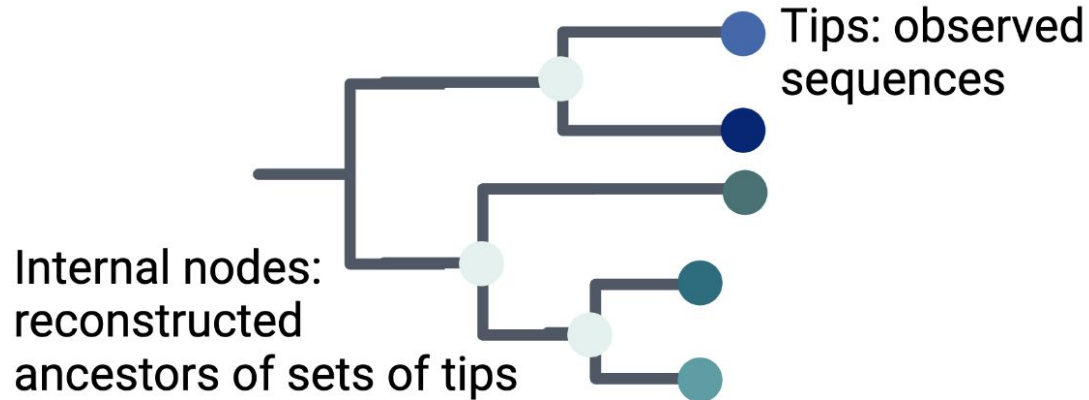
**How to manage antibiotic resistance? Future burden of disease? Vaccine escape?  
Improve outbreak control. Impact of border and travel measures.**



# DISCRETE STRUCTURES: PHYLOGENIES

For this talk: a phylogeny is a rooted binary tree. It shows patterns of ancestry of a set of organisms. Edges may have lengths (time, genetic distance)

Phylogenies are the main tool to study sets of sequences.



# Where do phylogenies come from? (This is not the focus here.)



**Data**

```
W K V D E V A E L T E K L K  
W K I E E V K E L E Q K L R  
W K L E E V K E L T E L I K  
W K T L M L R E L E E L F S  
R K V K I V S E A T E L L Q  
W K K D E I E N I K E L I Q  
W K K D E I E N I K E L I Q  
Y K V R A V E E I K R M I S  
W K R R E V K E L K E L M D  
W K K K E V Q E L H D L I K
```



**Trees**

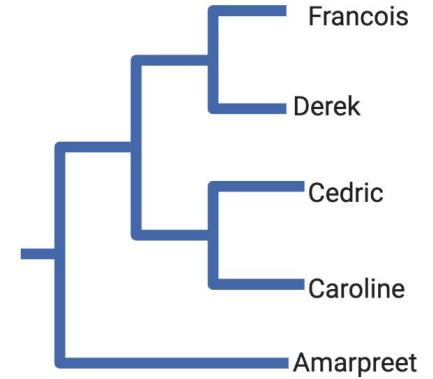


Maximum likelihood tree reconstruction.  
Bayesian tree reconstruction.  
Methods: developing since the 1970s.

# PHYLOGENIES AND SOME APPLICATIONS

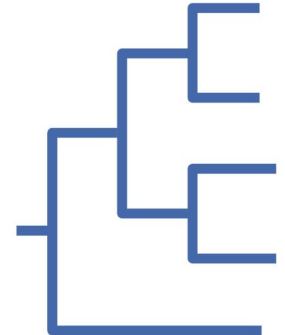
## Labelled phylogeny:

- tips have labels
- internal nodes do not have labels
- Applications: represent and compare patterns of ancestry for this *one* set of organisms



## Unlabelled phylogeny:

- tips and nodes do not have labels
- **Applications:** compare models to data, compare datasets to each other, *different* sets of organisms



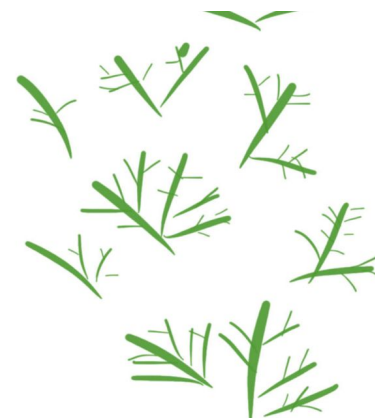
# TOOLS TO COMPARE UNLABELLED TREES HAVE BEEN LIMITED

People use summary statistics:

- Imbalance - how asymmetric is the tree?
- Numbers of features (two tips joined in a “cherry” configuration, clades of size  $k$ )
- Network science: betweenness, centrality, spectral methods -- limited for binary trees
- Height and width



PLACE 1

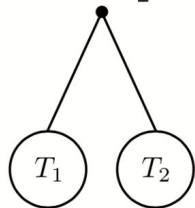


PLACE 2

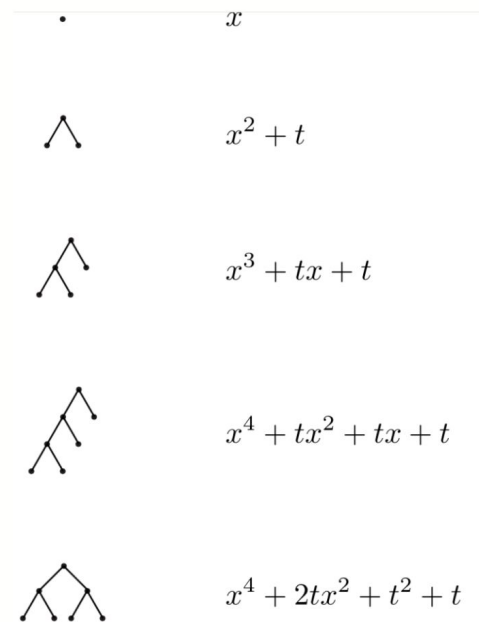
# POLYNOMIALS TO CHARACTERIZE UNLABELLED TREES

- ▶ Let the single node  $\bullet$  have polynomial  $p(\bullet) = x$ .
- ▶ For  $T = T_1 \wedge T_2$ , let  $p(T) = p(T_1)p(T_2) + t$

The tree  $T_1 \wedge T_2$ :



- ▶  $p(T) \in \mathbb{Z}[x, t]$ : a unique factorization domain



Some examples

Work by Pengyu Liu





# THE POLYNOMIAL UNIQUELY DEFINES THE TREE

The polynomial  $p(T)$  is irreducible (by Eisenstein's criterion)

Take two trees,  $T_1$  with polynomial  $p(T_1)$ , and  $T_2$  with polynomial  $p(T_2)$ .

$$p(T_1) = p(L_1 \wedge R_1) = p(L_1)p(R_1) + t$$

$$p(T_2) = p(L_2 \wedge R_2) = p(L_2)p(R_2) + t$$

If  $p(T_1) = p(T_2)$ , then  $p(L_1)p(R_1) = p(L_2)p(R_2)$ .

Since the polynomials are irreducible, the only way this can happen is if  $L_1$  is isomorphic to one of  $L_2, R_2$ . Then  $R_1$  is isomorphic to the other.

In this case,  $T_1$  is isomorphic to  $T_2$ .

# SOME COEFFICIENTS IN THE POLYNOMIAL HAVE NICE INTERPRETATIONS

- ▶ The **number of primary subtrees of  $T$**  is the sum of all of the coefficients in  $p(T)$ , ie  $\sum_{i,j} a_{i,j}$
- ▶ Let  $c$  be the number of cherries in  $T$ . For each cherry there is precisely one primary subtree that has all leaves of  $T$  except the cherry. Each contributes a  $x^{n-2}t$  to  $p(T)$ .
- ▶ The **number of cherries** is the coefficient of  $x^{n-2}t$
- ▶ The **number of clades of size  $k$**  is the coefficient of  $x^{n-k}t$
- ▶ Higher order terms: more complex; related to number of ways to choose clades of total size

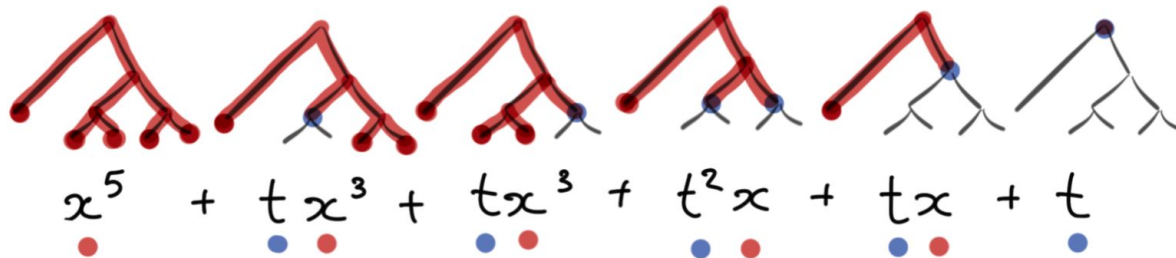
Primary subtree: rooted (full) binary subtree of  $T$  sharing  $T$ 's root

If  $S$  is a primary subtree of  $T$  with leaves that are:

- ▶  $k$  leaves of  $T$
- ▶  $\ell$  internal nodes of  $T$

then give  $S$  the term  $q(S) = x^k t^\ell$ .

Then  $p(T) = \sum_S q(S)$  (easy by induction)



Because the polynomial uniquely defines the tree, we can use it to make metrics:

$$d(T_1, T_2) = \sum_{i,j} \gamma(a_{ij}^1 - a_{ij}^2)$$

where  $a_{ij}^1$  are the coefficients in  $p(T_1)$  and similarly for  $T_2$ .

$$\text{Here, } \gamma(c_1, c_2) = \begin{cases} |c_1 - c_2| / (c_1 + c_2) & \text{if } c_1, c_2 \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

There are other natural metrics - we could choose weights to emphasize the  $x^k y$  terms (for the clade size distribution), use the  $L^2$  or another norm, etc.

# APPLICATION: TREES FROM DIFFERENT SIMULATION MODELS

Yule: pure birth process, no death; run to fixed size

Proportional to Distinguishable Arrangements (PDA): each tree sampled with probability proportional to the number of distinct labelled trees with its shape

Aldous: asymmetric split distribution, parameterized by  $\beta$

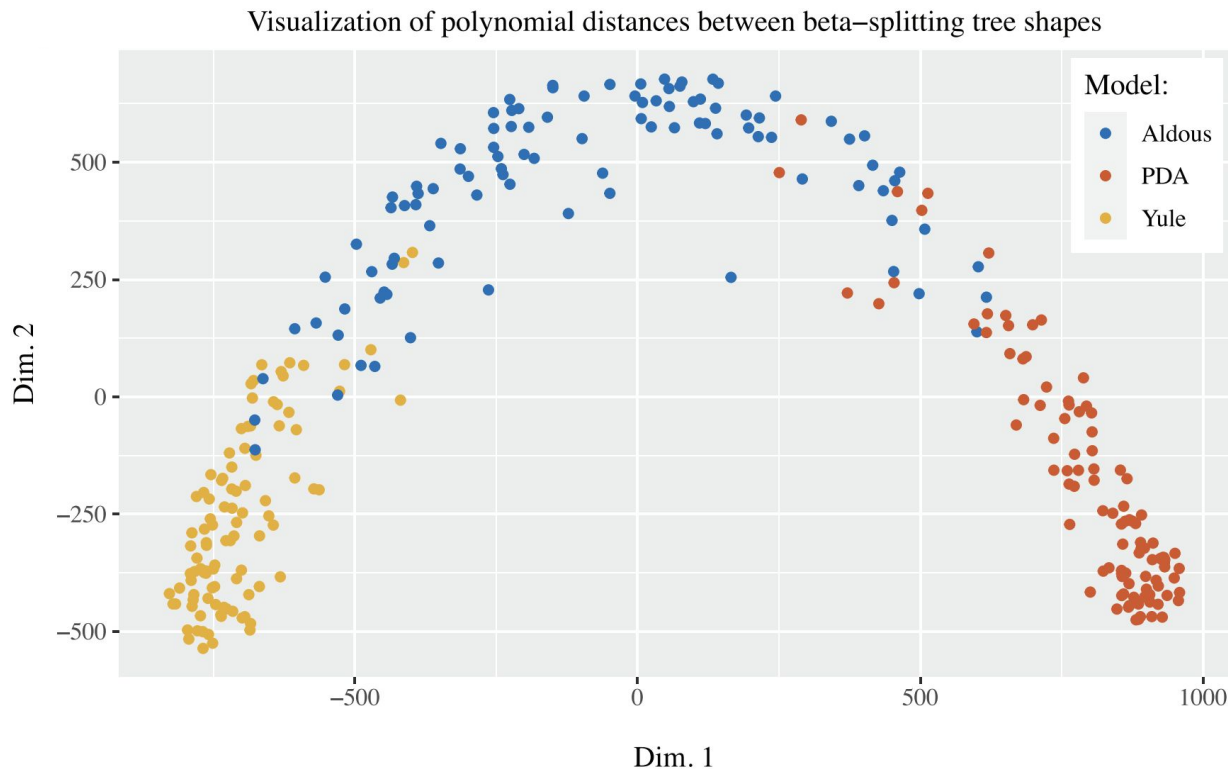
Do simulated trees from different processes look different, in the metric?  
Can the metric help us detect which process generated a tree?

# VISUALIZATION OF THE TREE-TREE DISTANCES IN THE METRIC

Make the distance matrix.

Make a multi-dimensional scaling (MDS) plot (top eigenvectors; projection).

Colour according to the underlying process that created each tree.



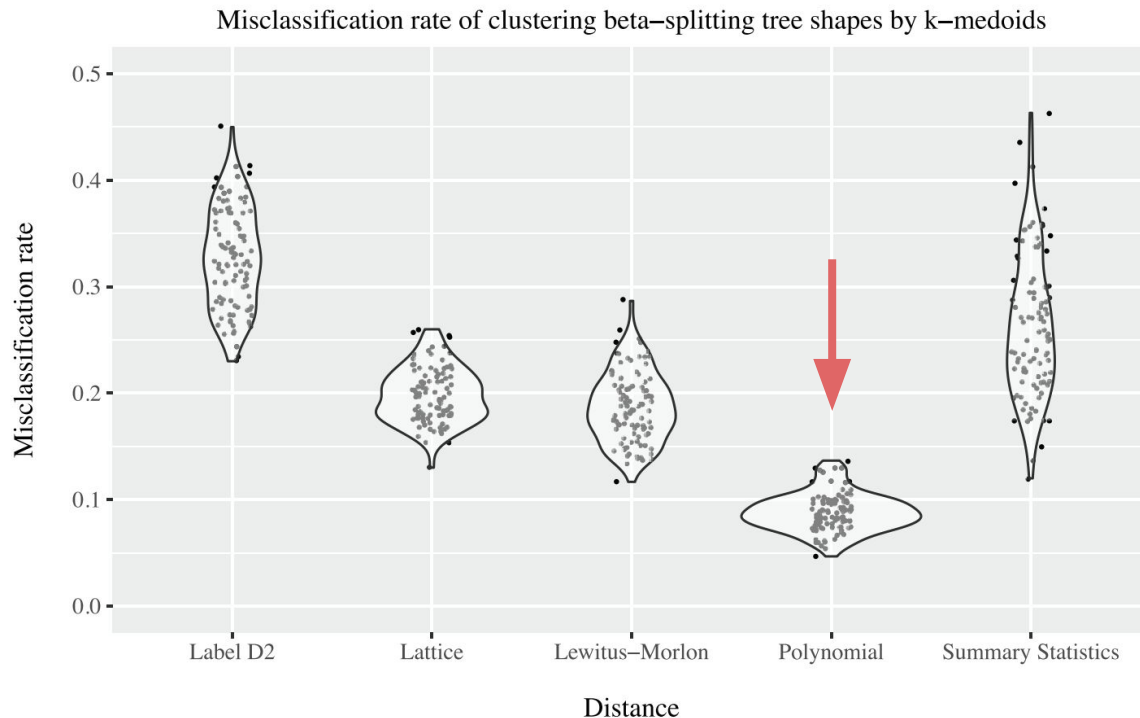
# CAN WE GUESS WHICH MODEL CREATED A TREE, USING THE METRIC?

Cluster the trees using the distances (k-medoids; unsupervised).

Label each cluster PDA, Yule, Aldous, by majority.

Classify a new tree as PDA, Yule, Aldous -- which cluster does the new tree land in?

Compare our metric to other ways to compare unlabelled trees.



# IF YOU WANT TO READ MORE ABOUT TREES AND METRICS:



Discrete Applied Mathematics

Volume 288, 15 January 2021, Pages 1-8

## A tree distinguishing polynomial

Pengyu Liu

> *Syst Biol.* 2022 Oct 12;71(6):1378-1390. doi: 10.1093/sysbio/syac008.

## Analyzing Phylogenetic Trees with a Tree Lattice Coordinate System and a Graph Polynomial

Pengyu Liu <sup>1</sup>, Priscila Biller <sup>1</sup>, Matthew Gould <sup>1</sup>, Caroline Colijn <sup>1</sup>

JOURNAL ARTICLE

## A Metric on Phylogenetic Tree Shapes

C. Colijn , G. Plazzotta

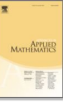
*Systematic Biology*, Volume 67, Issue 1, January 2018, Pages 113-126,  
<https://doi.org/10.1093/sysbio/syx046>

Published: 02 May 2017 **Article history** ▾



Advances in Applied Mathematics

Volume 27, Issue 4, November 2001, Pages 733-767



Regular Article

## Geometry of the Space of Phylogenetic Trees

Louis J. Billera <sup>a</sup>, Susan P. Holmes <sup>b, c</sup>, Karen Vogtmann <sup>a</sup>

[Submitted on 31 May 2018 (v1), last revised 29 Jun 2022 (this version, v8)]

## Tropical Geometry of Phylogenetic Tree Space: A Statistical Perspective

Anthea Monod, Bo Lin, Ruriko Yoshida, Qiwen Kang



# CONCLUSIONS

Infectious diseases -- and pandemics! -- bring new mathematical challenges.

They go beyond the “classic mode” of mathematical biology.

New innovations bring in applied probability, statistics, genomics, discrete mathematics and more.



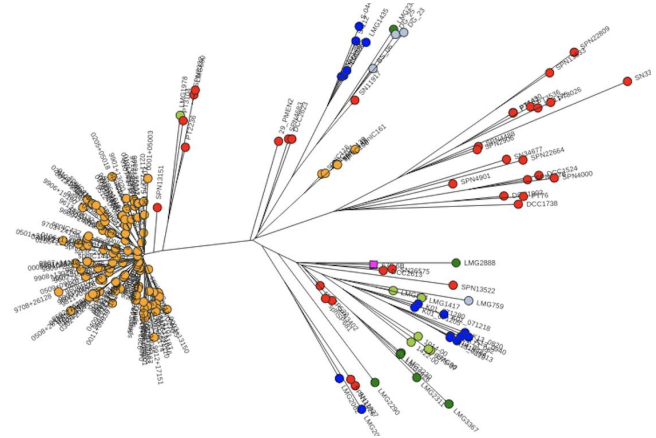
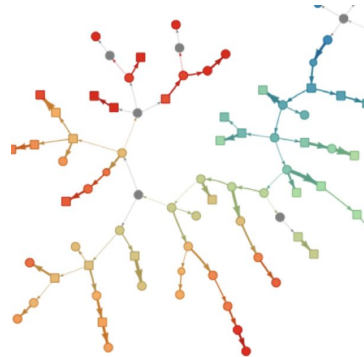
Face Masks



Hand Sanitizer



Plexiglass Barrier



# THANK YOU FOR COMING

## Acknowledgements:

- Paul Tupper
- Shraddha Pai, COVID Schools Canada
- Pengyu Liu, Matthew Gould, Priscila Biller
- MAGPIE research group, Department of Mathematics

Funders: Canada 150 Research Program, NSERC, Michael Smith Health Research BC, Genome BC

## CURRENT OPPORTUNITIES



### PHD RESEARCH SOFTWARE DEVELOPER/ SENIOR SCIENTIST, MAGPIE GROUP

Begin ASAP for a 3-year term. (The start date is negotiable)

We will begin to review applications on November 15, 2022.

### GRADUATE STUDENTSHIPS IN THE MAGPIE GROUP

Seeking applicants for MAGPIE's new Studentship Program situated in Masters or PhD programs in Mathematics at SFU. Competitive graduate student funding is available.

**SFU**

SIMON FRASER  
UNIVERSITY

**MATHEMATICS, GENOMICS AND PREDICTION IN INFECTION AND EVOLUTION**