Public health in genetic spaces:

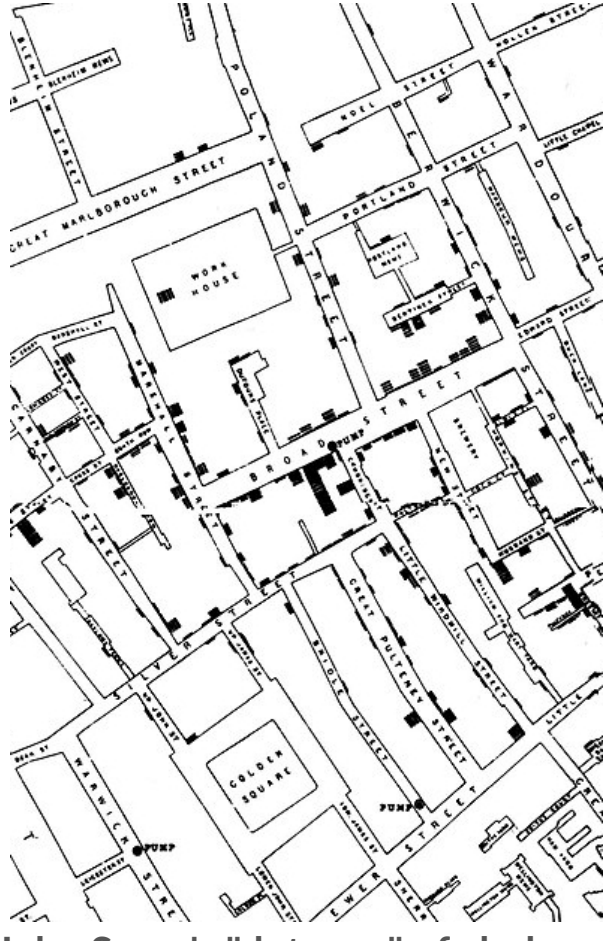
A new framework for HIV cluster forecasting

Art Poon
Western University

Departments of Pathology & Laboratory Medicine; Applied Mathematics; Microbiology and Immunology

Cluster detection

- Spatiotemporal clusters are an essential public health tool to detect outbreaks
- Cases that cluster in space and time may have a common risk factor, *e.g.*, outbreaks of bacterial infection in a hospital unit
- This is more difficult for pathogens that establish **chronic infections with a long asymptomatic period**, *e.g.*, HIV-1, hepatitis C virus, *M. tuberculosis*



John Snow's "dot map" of cholera cases on Broad Street, London.

Clusters in genetic space

- Spatial clustering may be less useful for pathogens that are transmitted by intimate contact.
- HIV-1 evolves so rapidly that the virus infection becomes unique to each person within months.
- A group of people diagnosed with HIV-1 that are genetically similar may have been affected by an outbreak.
- Routine genotyping presents an opportunity to detect outbreaks in "near real time".

Protect people

- Genetic clustering has also been misused as forensic evidence in HIV-1 transmission cases.

HIV is not a crime.

- Clustering must be done in a way that **does not add risk to individuals**, while providing a measurable benefit for public health.
- Reach out to groups without targeting the individual.

Are genetic clusters helpful?

- If we are going to use genetic clustering to guide public health decisions in "near real time"...
- ... we need to be able to evaluate whether genetic clusters can help us make better decisions.

> *Our aim is to develop a new statistical framework where it is possible to evaluate a clustering method against any other method of prioritizing groups.*

Pairwise clustering

- We start by analyzing the simplest method.
- **Pairwise distance clustering** is a popular approach that builds up clusters from pairs of sequences (*e.g.*, HIV-TRACE).
- This threshold is often entirely arbitrary!
- (Some studies have used **longitudinal HIV-1 divergence** — how far a patient's own virus population may evolve — to select a threshold, *e.g.*, B.C.)

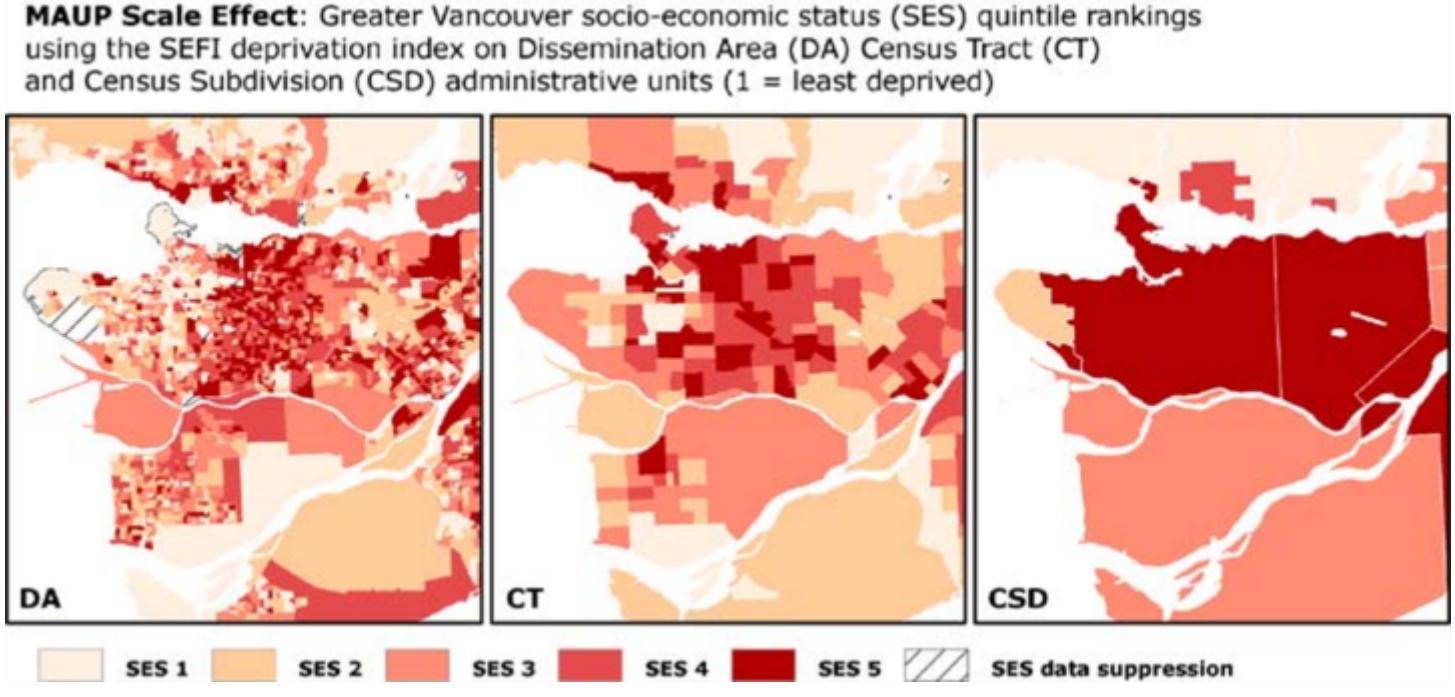
Optimizing the cutoff

- The genetic distance is affected as much by when a person's infection is **diagnosed** as when they were **infected**[1,2].
- There is no such thing as a "gold standard" cutoff for all settings.
- If the cutoff is too relaxed (high), then nearly all HIV-1 sequences will appear in a single giant cluster.
- Too strict (low), then every sequence is a cluster of one.

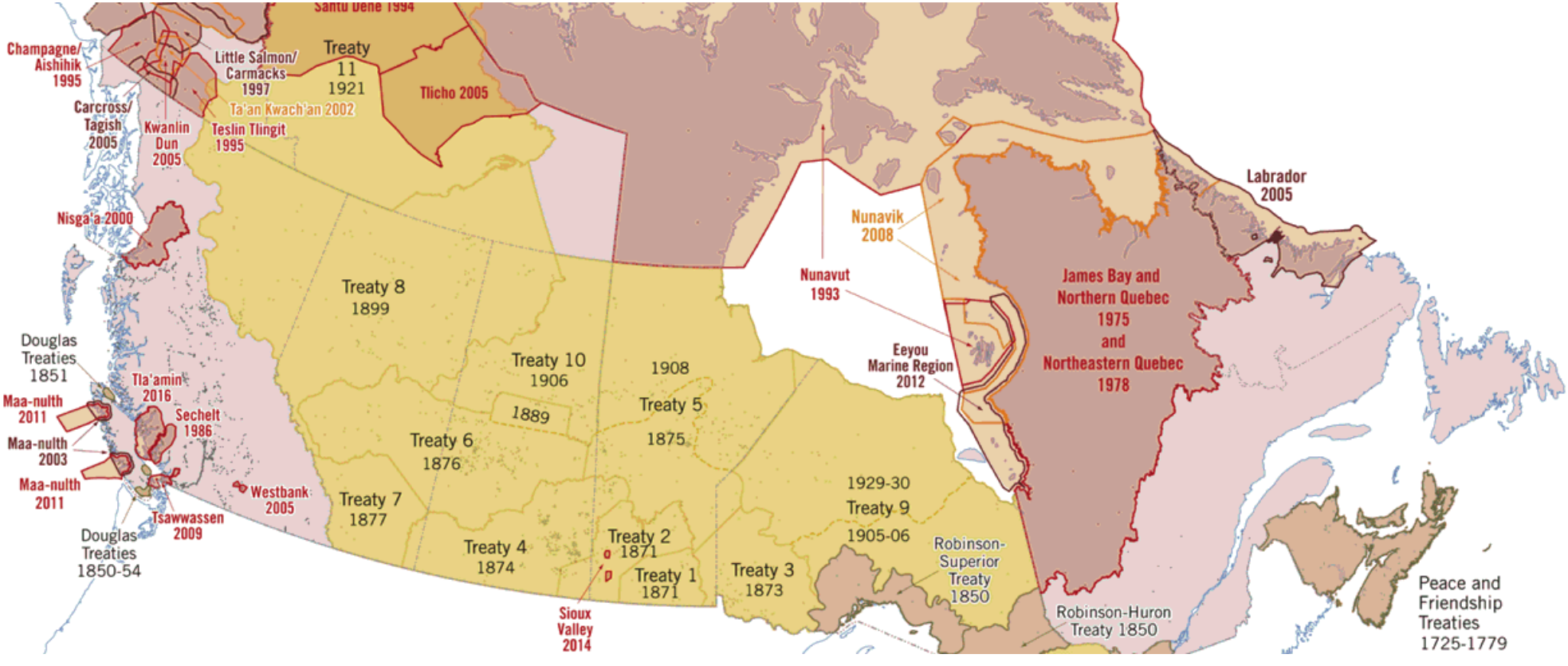
[1] Volz EM *et al.* (2012) PLOS Comput Biol 8:e1002552; [2] Poon AFY (2016) Virus Evol 2: vew031.

The modifiable areal unit problem (MAUP)

- This is a common problem in spatial statistics.
- Statistical associations depend on how you partition geographic space into units of observation.



Example of MAUP from N Schuurman *et al.* (2007) J Urban Health 84:591.



Map produced by GIS and Cartography Office, Department of Geography, University of Toronto for the exhibit Canada By Treaty: Negotiating Histories, co-curated by Heidi Bohaker, James Bird and Laurie Bertram. (c)2017.

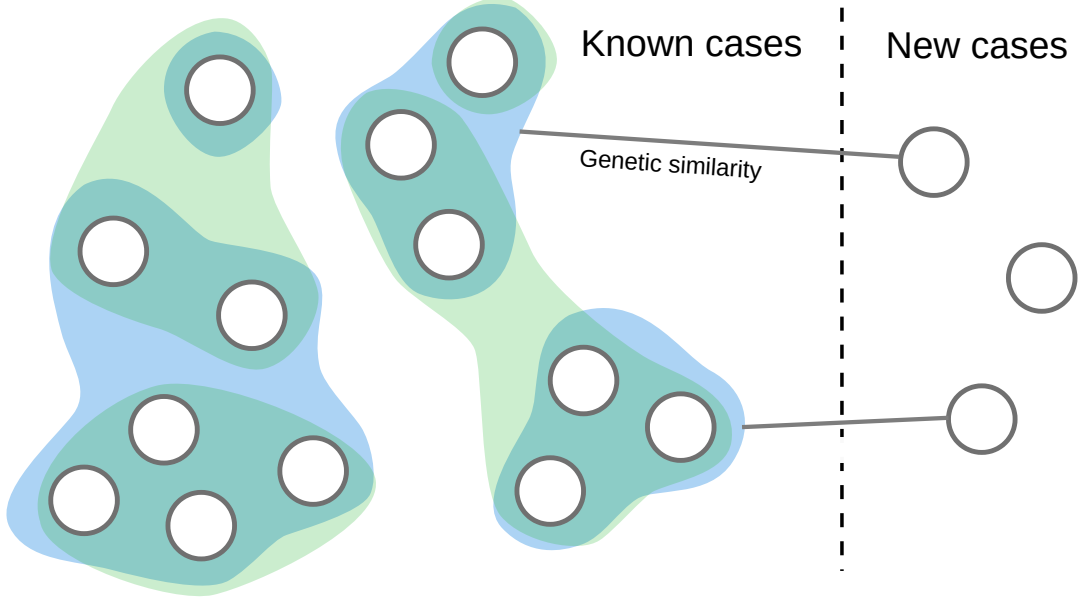
MAUP and HIV clustering

- We propose to draw an analogy between geography and **genetic space**.
- Naïvely using some cutoff for HIV clustering is equivalent to ignoring how different boundaries can be drawn on a map.
- How can we justify using one set of boundaries instead of another?

> *We propose that the best criteria to select a clustering method is our ability to predict where the next cases will occur.*

Prediction models

- A cutoff defines a partition of known cases into clusters C_i
- The number of new cases with sufficient similarity to C_i is a Poisson-distributed outcome.



Similarity does not imply transmission.

Prediction models

- Let the predicted number of new cases adjacent (similar) to the i-th cluster be:

$$\hat{R}(C_i) = \exp\left(\alpha + \beta \sum_{v \in C_i} \rho(v)\right)$$

- In other words, the predicted "growth" of a cluster obtained by summing over all members of the group.
- $\rho(v)$ is the *weight* of an individual vertex representing a known case
- α and β are parameters to be estimated by Poisson regression

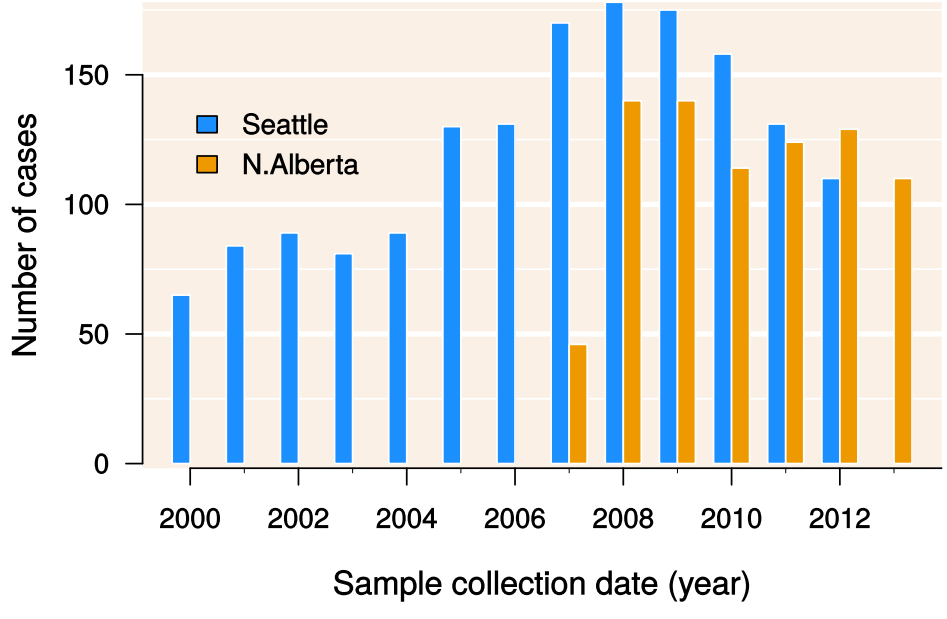
Similarity as a function of case recency

- The weight of a known case ρ can be any linear combination of individual attributes $A(v)$
- For demonstration, we use a single attribute based on the sample collection date $t(v)$.
- If new cases appear in time interval $t_{\max}$, then we calculate the expected weight of known case v from its recency as follows:

$$\log\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right) = \alpha + \beta_0\left(t(v) - t_{\max}\right)$$

Datasets

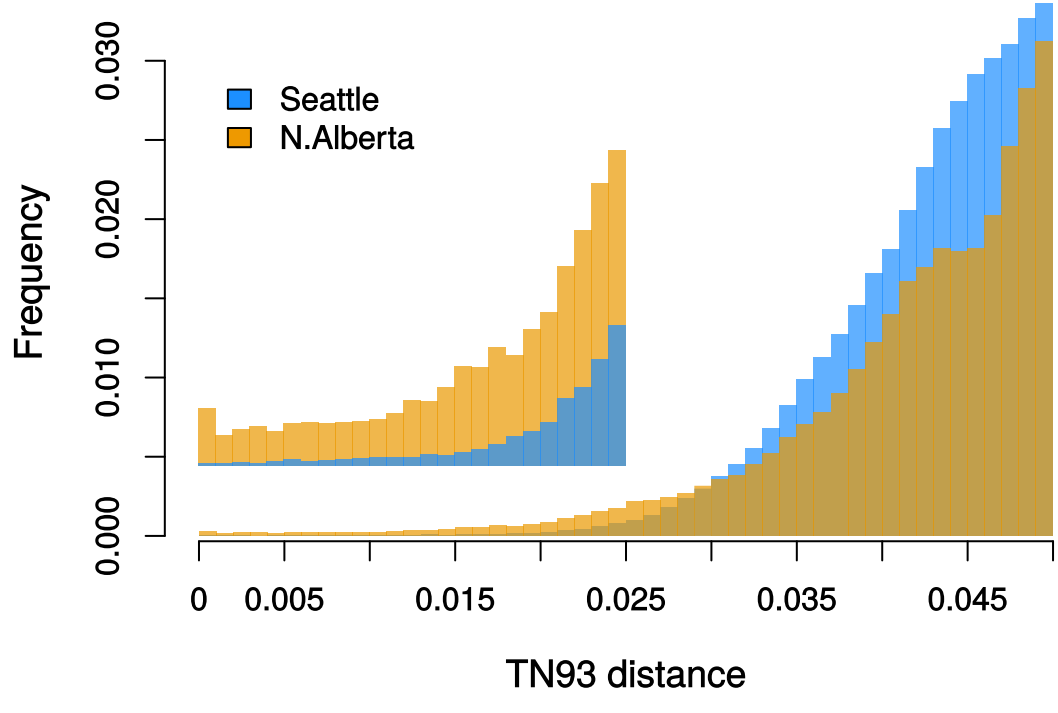
- We obtained two published data sets of anonymized HIV-1 *pol* sequences collected in Seattle[1] ($n = 1,591$) and Northern Alberta[2] ($n = 803$).
- $n = 110$ "new cases" in last year of sampling (2012, 2013).



[1] Wolf E *et al.* (2017) ARHR 33:318; [2] Vrancken B *et al.* (2017) Inf Genet Evol 52:100.

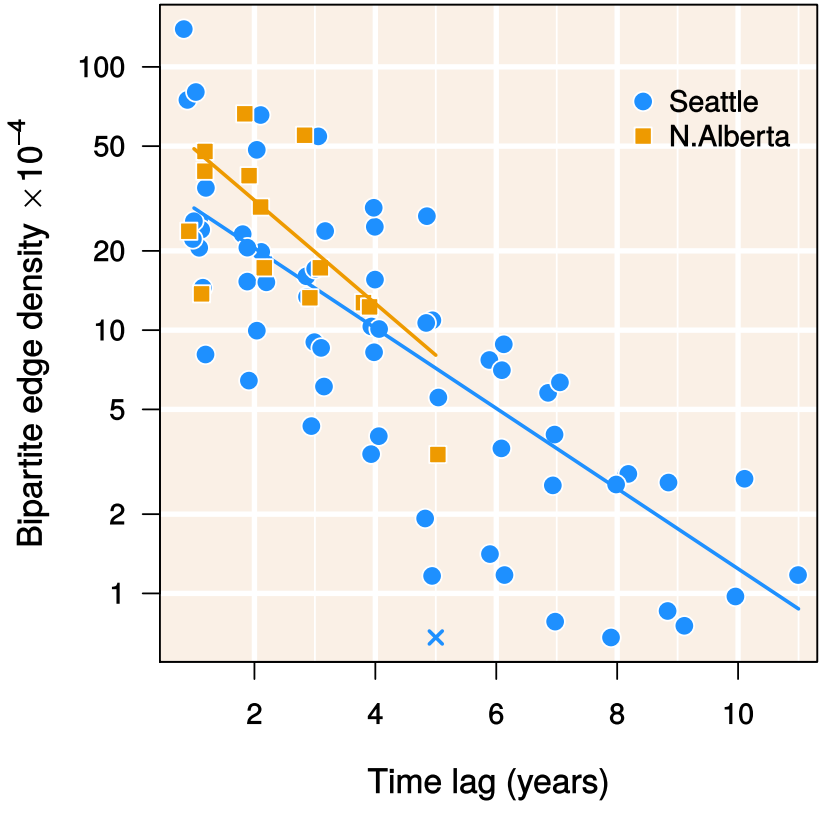
Distance distributions

- Calculated Tamura-Nei (TN93) pairwise distances using C program http://github.com/veg/tn93
- Distances in Northern Alberta data set tended to be shorter than Seattle:



Effect of case recency

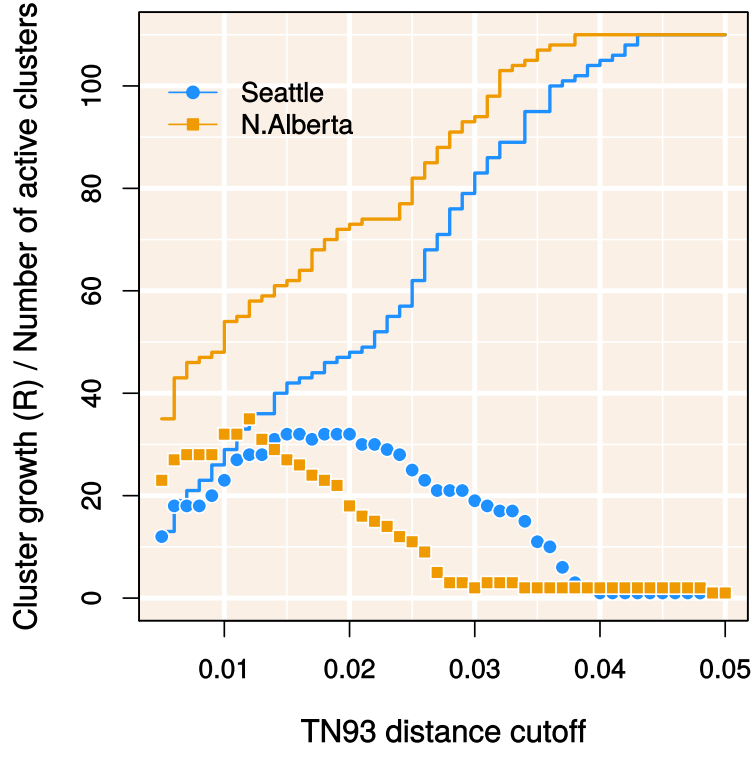
- At a given cutoff, the observed densities of edges to new cases increase with recency of known cases (lower time lag).
- In other words, the odds that a known case has genetic similarity to a new case decays exponentially over time.
- Confirms predictive value of case recency[1,2]



[1] McVea D *et al.* (2017) Top Antivir Med 25:360s; [2] Wertheim J *et al.* (2018) J Infect Dis 218:1943.

The effect of varying cutoffs

- The number of new cases adjacent to clusters declines with lower cutoffs.
- Represents upper limit to how many clusters gain a new case.
- The number of these "active clusters" has an intermediate optimum.
- At relaxed cutoffs, all known cases are grouped into one giant cluster.



Model comparison

- We optimize the distance cutoff by comparing our model against a null model:

$$\hat{R}_0(C_i) = \exp\left(\frac{|C_i|}{|V^c|}R\right)$$

- $|C_i|$ is the number of known cases in the i-th cluster
- $|V^c|$ is the total number of known cases
- R is the total number of new cases that are adjacent to clusters
- This null model assumes that cluster growth is proportionate to cluster size.

An information criterion

- Nakaya[1] describes a "generalized Akaike information criterion" (GAIC) to select the optimum partition of a region into districts.
- It is simply the difference in AIC:

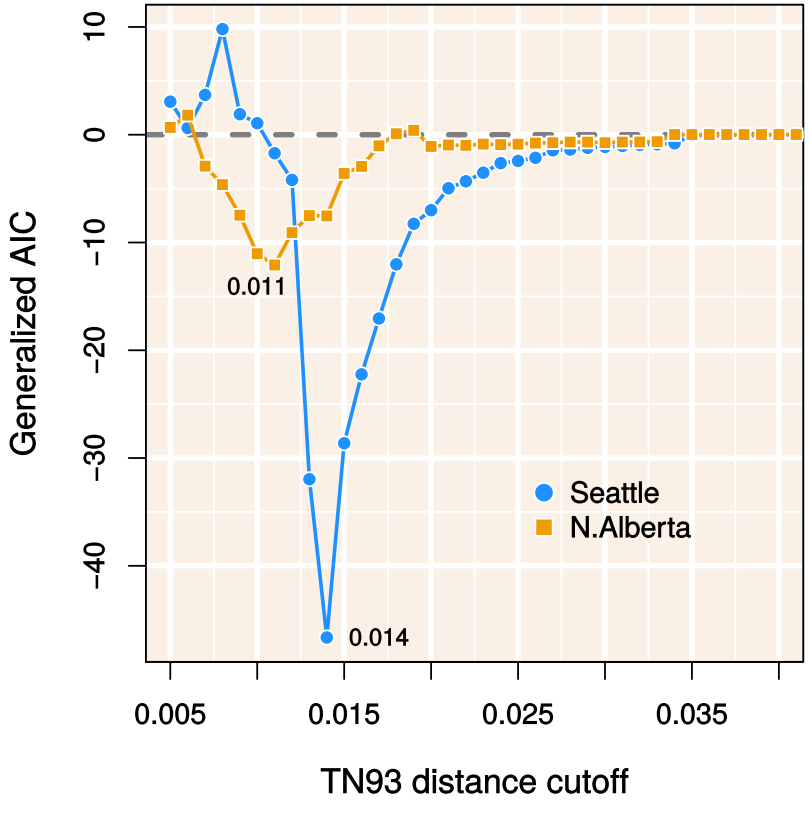
$$\begin{aligned}\mathrm{GAIC} &= \mathrm{AIC}(\hat{R}) - \mathrm{AIC}(\hat{R}_0) \\ &= 2(k - k_0) - 2\left(\log L(\hat{R}) - \log L(\hat{R}_0)\right)\end{aligned}$$

- The optimal cutoff minimizes the GAIC.

[1] Nakaya T (2000) *An information statistical approach to the modifiable areal unit problem in incidence rate maps*. Environment and Planning A 32:91.

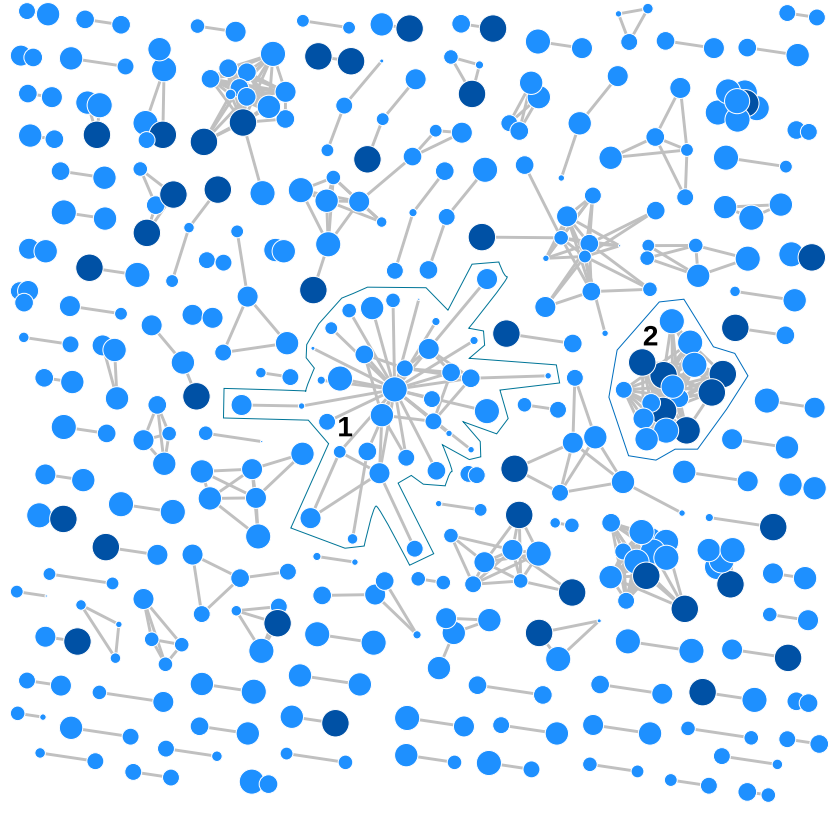
Results

- When cutoff is high, most known cases are in one large cluster; the weighted model is no different from the null model (GAIC near 0)
- When cutoff is too low, we have random unpredictable growth of small clusters and the weighted model is worse.
- Optimal cutoffs were similar between data sets (0.014 and 0.011).



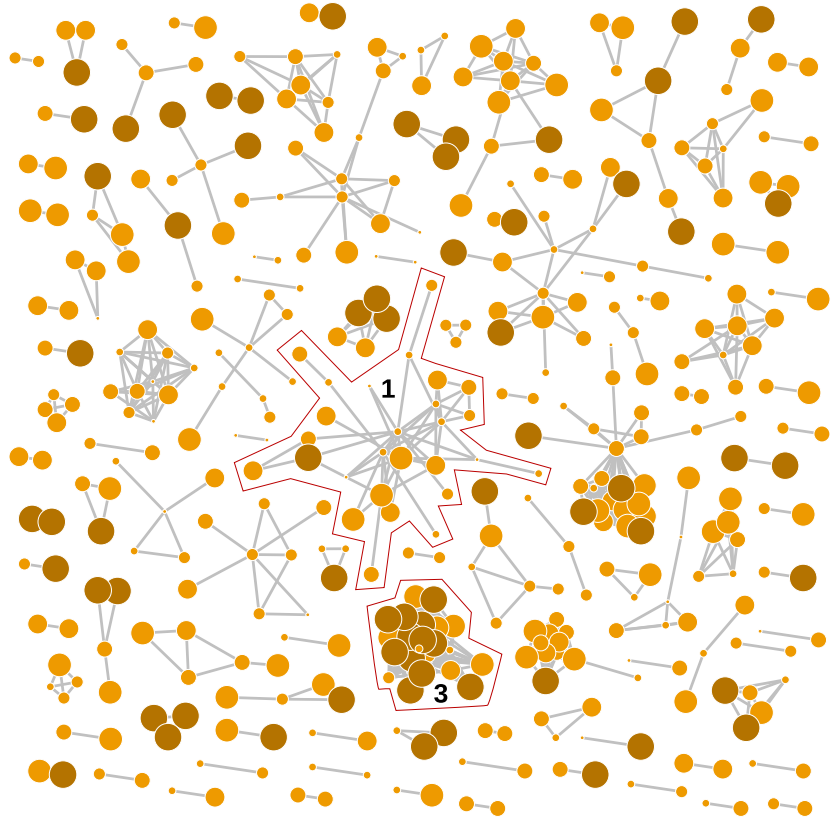
Seattle

- Generated graph at optimal cutoff for Seattle data set.
- Vertex size scaled to case recency (more recent cases are larger).
- Largest cluster (**1**, 35 known cases) was not adjacent to any new cases; older sampling dates.
- Second-largest cluster (**2**, 10 known cases) more recent, grew by 6 new cases.



Northern Alberta

- Largest cluster (**1**, 26 known cases) also older, grew by 1 new case only.
- Third-largest cluster (**3**, 10 known cases) more recent, grew by 12 new cases.
- In both data sets, **roughly half of all new cases were not adjacent to any cluster** at the optimized cutoffs (67% Seattle, 42% Northern Alberta).



Closing remarks

- Our framework provides an objective criterion for defining clusters that maximize public health value.
- Using stricter distance cutoffs can subvert the public health value of clusters.
- The GAIC framework is not limited to case recency; the Poisson model can incorporate **any number of predictors**.
- It is not limited to pairwise distance clustering; it can evaluate **any partition** of known cases.

Acknowledgements

This framework was implemented in *R* by my first-year MSc student, **Connor Chato**, shown below doing a podcast.

Funding

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada