# A Spatial Point Process Model for Viral Infections

Murali Haran

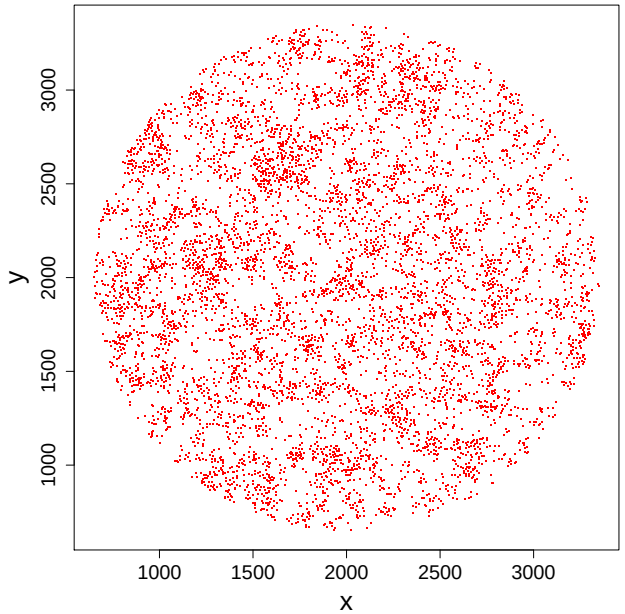
Department of Statistics, Pennsylvania State University

March 5, 2014

(joint work with Josh Goldstein, Ivan Simeonov, John Fricks, and Francesca Chiaromonte)

# Spatial point process modeling of virus infections

- Biologists are often interested in investigating the progression of viral infections
- Our goal: use data from imaging of cell cultures to study the spatial structure of an infection
- An *in vitro* cell culture study identifies and locates cells infected with two strains of the human respiratory syncytial virus (RSV-A and RSV-B)




Points represent locations of cells infected with RSV.

*Question:*
How does the presence of an infected cell impact infections in neighboring cells?

# Outline

- Spatial point processes in the plane provide a natural framework here
- Each point represents the 2D coordinates of an infected cell
- Goal: Infer spatial interaction among cells

**Contributions of this work:**

- A new spatial point process model for the RSV data
- Inferential methods for this computationally challenging problem
- Draw scientific conclusions from fitted model

# Spatial point processes

A spatial point process is a stochastic process, a realization of which consists of a set of points $X = (x_1, ..., x_n)$ in a bounded region $W \subseteq \mathbb{R}^d$.

Some SPPs can be used to model interactions:

$$f(X|\Theta) = \lambda^n \prod_{i \neq j} \phi(x_i, x_j)$$

where $\phi(x_i, x_j)$ is the *interaction function* between points $i$ and $j$.
In the homogeneous case, $\phi(x_i, x_j) = \phi(\|x_i - x_j\|) = \phi(r)$

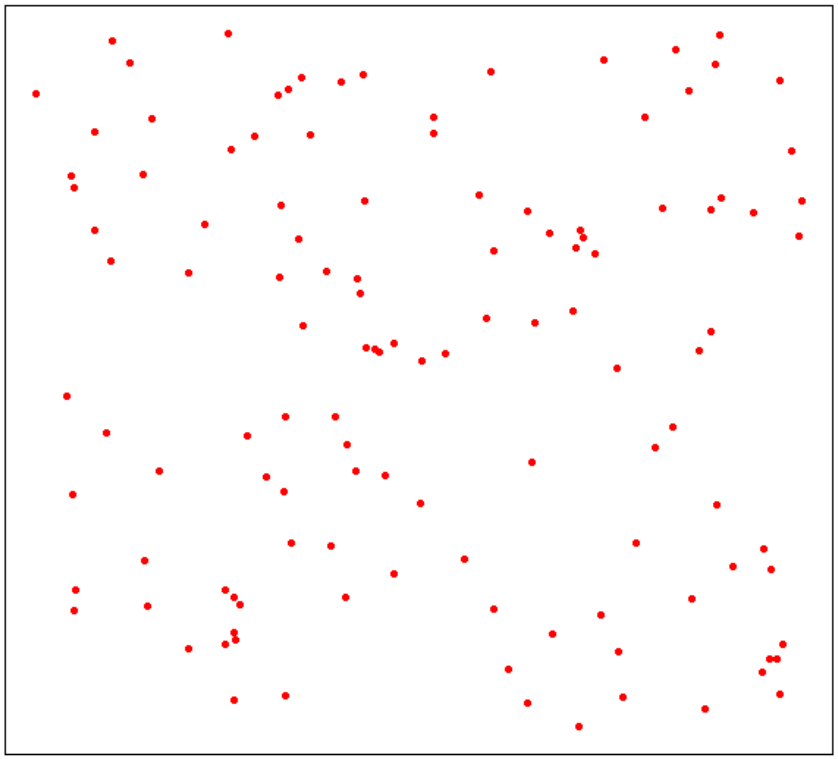
The Strauss process is a simple example,

$$\phi(r) = \begin{cases} \gamma, & 0 < r \leq R \\ 1, & r > R \end{cases}$$

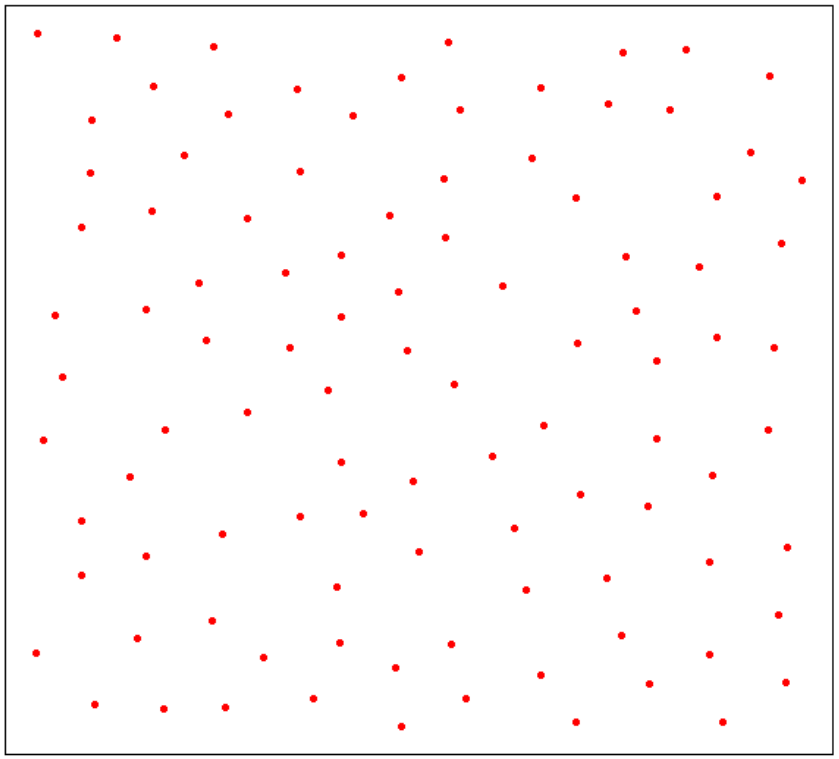
for $0 \leq \gamma \leq 1$. Since $\phi(r) \leq 1$ this is a repulsion point process.

# Poisson process vs. Strauss process

**Realization of Poisson Process**



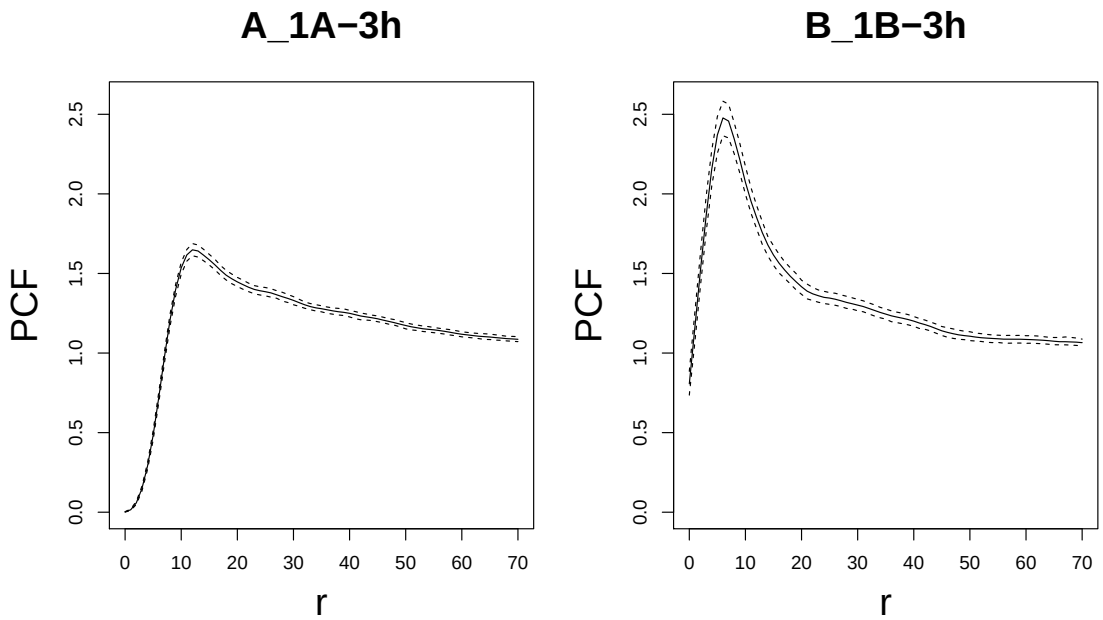
**Realization of Strauss Process**



# Exploratory analysis of RSV data: Need for a new model

The pair correlation function (PCF) $g(r)$ is an exploratory summary statistic that tells us the attraction-repulsion behavior of points separated by distance $r$ in a spatial point process.

- A value of $g(r) > 1$ indicates attraction, a tendency for points to cluster at distance $r$. Similarly, $g(r) < 1$ indicates repulsion at distance $r$. For our data:




Observed attraction-repulsion in RSV data varies smoothly in $r$.

# New attraction-repulsion model: Interaction function

Goal: Allow attraction-repulsion to vary smoothly with distance to model observed RSV behavior.

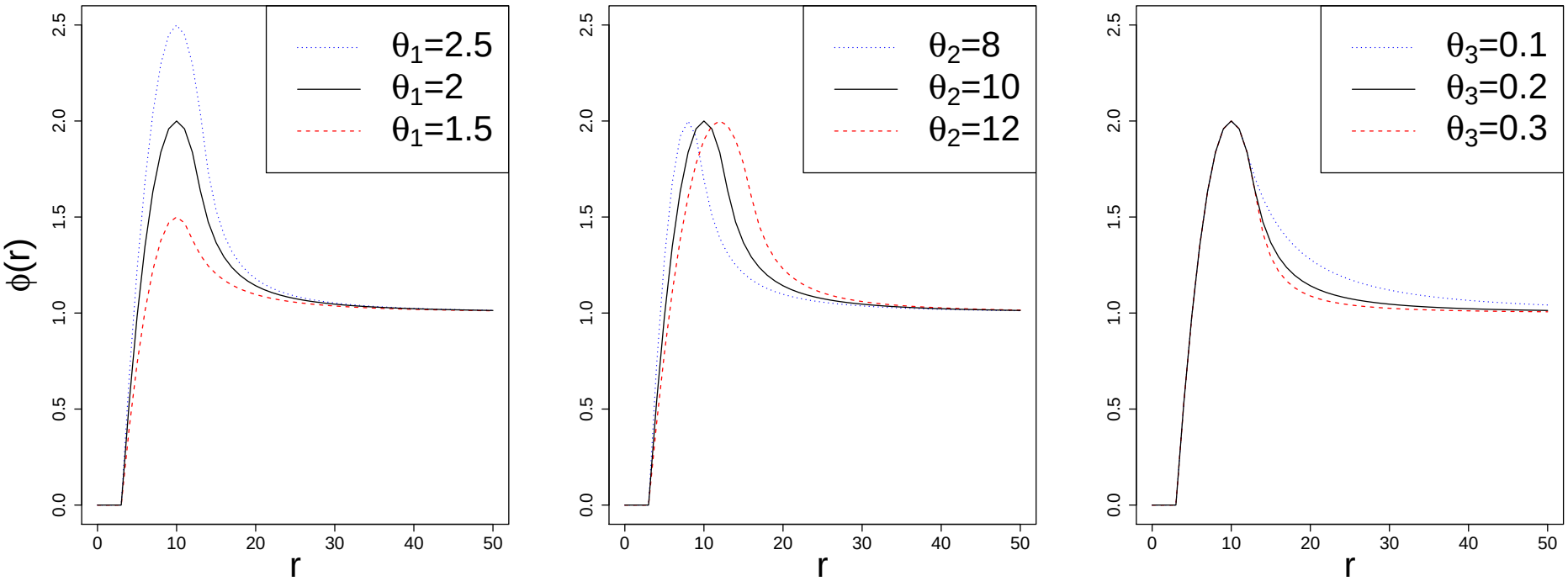
The interaction function $\phi(r)$ is defined piecewise,

$$\phi(r) = \begin{cases} 0, & 0 \leq r \leq R \\ \theta_1 - \left( \frac{\sqrt{\theta_1}}{\theta_2 - R}(r - \theta_2) \right)^2, & R < r \leq r_1 \\ 1 + \frac{1}{(\theta_3(r - r_2))^2}, & r > r_1 \end{cases}$$

where

- $\theta_1$: value of $\phi(\cdot)$ at the peak
- $\theta_2$: the value of $r$ at the peak
- $\theta_3$: rate of descent after the peak
- $R$: minimum allowable distance between points
- $r_1$, $r_2$: chosen to ensure $\phi(\cdot)$ is continuously differentiable.

# New attraction-repulsion model: Interaction function




- $\phi(r) > 1$: attraction, points tend to cluster at distance $r$
- $\phi(r) < 1$: repulsion

# Attraction-repulsion model

The likelihood can be written as

$$\mathcal{L}(X|\Theta) = \frac{f(X|\Theta)}{c(\Theta)}, f(X|\Theta) = \lambda^n \Big[ \prod_{i=1}^{n} e^{\min\left[\sum_{i \neq j} log(\phi(x_i, x_j)), k\right]} \Big]$$

**Model parameters:**

- $\lambda$ is the intensity of the process
- $\theta_1, \theta_2, \theta_3$ control the shape of $\phi(r)$.
- R is the minimum distance allowed between points
- k is a truncation constant necessary to prevent "clumping" behavior

**Important:** $c(\Theta)$ is intractable. This makes computing very challenging.

# Inference

- Let $\Theta = (\lambda, k, \theta_1, \theta_2, \theta_3)$. Likelihood $\mathcal{L}(X|\Theta)$.
- Bayesian inference for $\Theta$ is based on the posterior distribution

$$\pi(\Theta|X) \propto \mathcal{L}(X|\Theta)p(\Theta) = \frac{f(X|\Theta)p(\Theta)}{c(\Theta)}$$

- Markov chain Monte Carlo (MCMC) is a convenient approach to learning about $\pi(\Theta|X)$.
- Choose a gamma prior on k, prior for remaining parameters are uniform over a plausible range

# MCMC

- Construct a Markov chain with stationary distribution $\pi(\Theta|X)$.
- In MCMC, propose $\Theta'$ from $q(\Theta, \Theta')$ and calculate the following acceptance probablity:

$$\alpha = \min\left(1, \frac{p(\Theta')q(\Theta', \Theta)f(X|\Theta')}{p(\Theta)q(\Theta, \Theta')f(X|\Theta)} \frac{c(\Theta)}{c(\Theta')}\right)$$

- The intractable normalizing constant $c(\Theta)$ does not cancel. Traditional MCMC methods cannot be applied.

**Solution:** Introduce an auxiliary variable.

- Double Metropolis-Hastings algorithm of Liang (2010).
- Two nested MCMC samplers; the "inner sampler" generates an auxiliary point pattern at each step of the "outer" sampler.

# Exchange algorithm

The exchange algorithm (Møller et al, 2006; Murray et al, 2006):

1. If we are currently in state $\Theta$, propose a transition to $\Theta'$ from proposal $q(\Theta, \Theta')$.
2. Generate auxiliary point pattern $Y \sim \frac{1}{c(\Theta')} f(\cdot|\Theta')$ (perfect sampling)
3. Accept $\Theta'$ with probability

$$\alpha = \min\left(1, \frac{p(\Theta')q(\Theta', \Theta)f(X|\Theta')f(Y|\Theta)}{p(\Theta)q(\Theta, \Theta')f(Y|\Theta')f(X|\Theta)} \times \frac{c(\Theta)c(\Theta')}{c(\Theta)c(\Theta')}\right)$$

The normalizing constants cancel.

# Double Metropolis-Hastings algorithm

**Problem:** Exchange algorithm does not apply here; we do not have a perfect sample for this complex model.

- Approximate the perfect sample $Y$ by the last value from a long MCMC sampler.
- This is known as Double Metropolis-Hastings due to Liang (2010). It is an approximate version of the exchange algorithm.
- Double MH uses two nested MCMC samplers; the "inner" sampler generates an auxiliary point pattern at each step of the "outer" sampler.
- Asymptotically equivalent to exchange algorithm.

# Computational challenges

The largest datasets consist of 13,000-14,000 spatial locations. For data this large, the nested samplers are expensive; the inner sampler must be run for thousands of iterations at each step of the outer sampler.

- Inner sampler updates fast in practice since we only propose to add or remove a single point (birth-death sampler)
- R too slow, code in C and optimize.
- Greatly reduce computing by truncating the interaction function for large values of $r$ (evaluate $\phi(r)$ to 1 when $r > R_{max}$).
- Inference for three replicates of the largest dataset takes a few days on the Lion-X cluster.

# Conclusions

Can make meaningful scientific conclusions as a result of inference on model parameters across multiple RSV experiments (e.g. RSV-B infected cells have a higher propensity to lump together than RSV-A; suggests RSV-B induces stronger increase in susceptibility to infection).

Advantages of our method:

- Our model captures the complex scale-varying attraction and repulsion behavior observed in the RSV dataset; this flexibility is not available in existing models
- Parametric specification of the interaction function lets us draw meaningful conclusions based on parameter inference
- Inference works well for simulated examples – we can recover the truth. We also have a method using PCFs to verify the goodness of fit for our model.