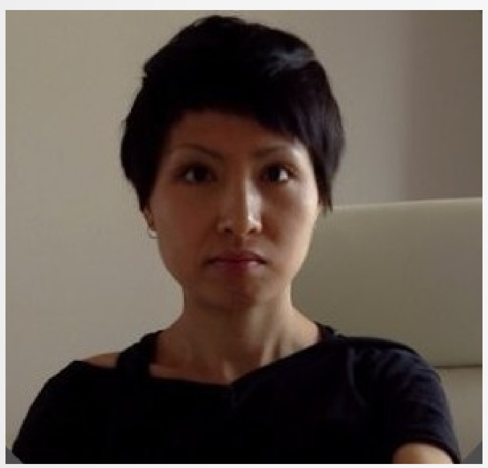
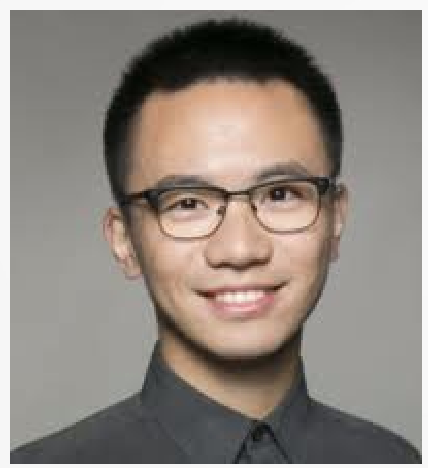
SIR-Hawkes: Linking Epidemic Models and Hawkes Processes

Marian-Andrei Rizoiu

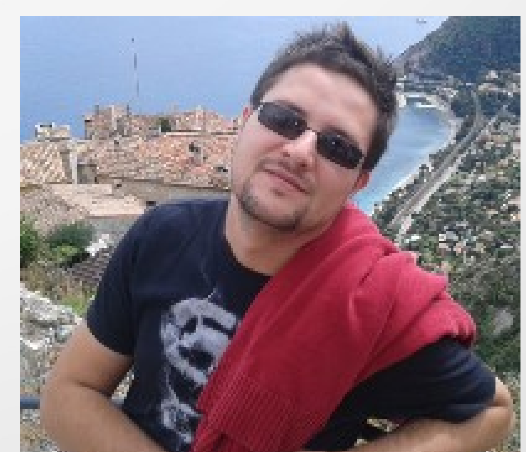
The research group

Behavioral
Data Science

1 research associate, 3 PhD students, 2 Honors students, 1 lecturer



Research income & grants

Behavioral Data Science

~$460k

2019 – current:	Crawford School of Public Policy grants, "**Evaluating democratic equity through analysing data around public donation to presidential candidates**", co-CI.
2019 – current:	UTS cross-faculty collaboration scheme, “**SocialSense: Making sense of the opinions and interactions of online users**”, CI.
2019 – current:	Data61 Challenge model grants, “**Adaptive skills taxonomy to enable labour market agility”**, CI.
2018	ANU Social Science Cross-College Grants, **“Advanced tools and methods for analysing the role and influence of bots in social media”**, CI.
2018	ANU Social Science Cross-College Grants, **“Identify Hate Speech and Predict Mass Atrocities”**, CI.

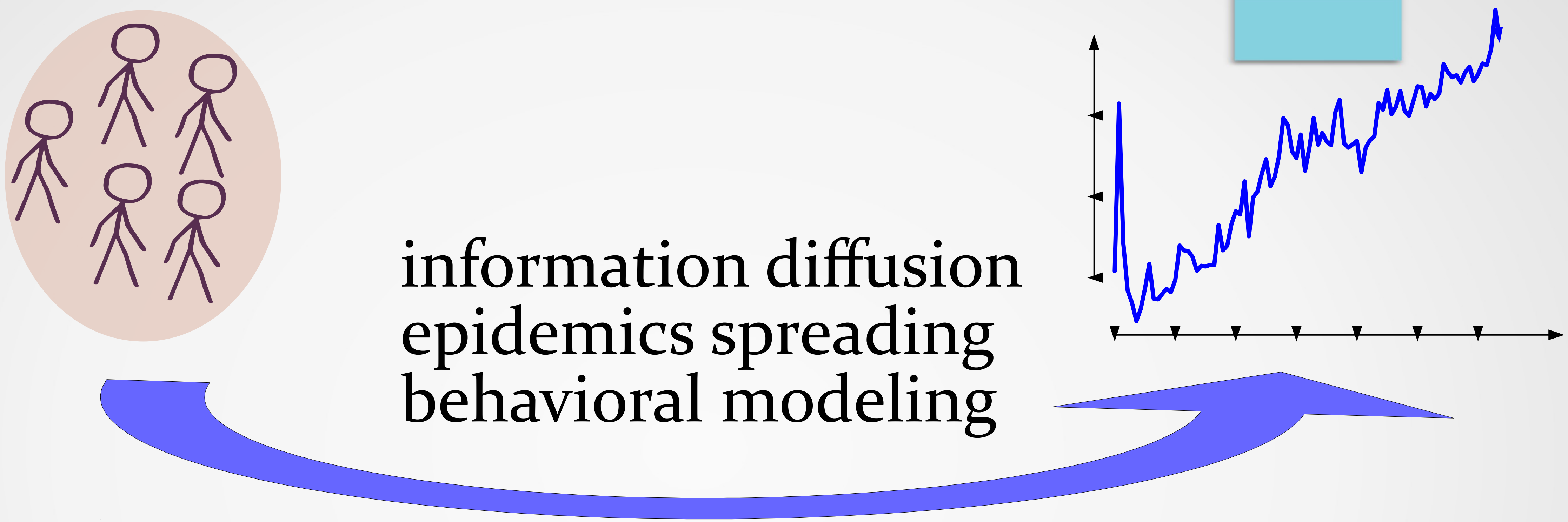
WIKIMEDIA FOUNDATION

NETFLIX

Research objectives

Behavioral Data Science

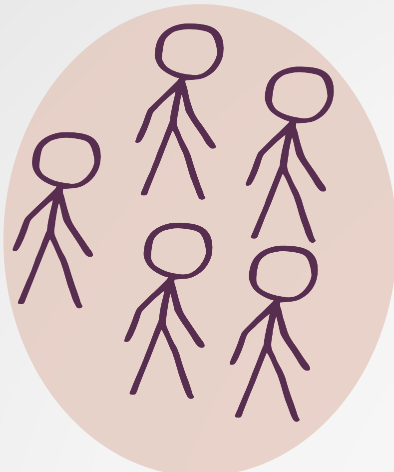
1.



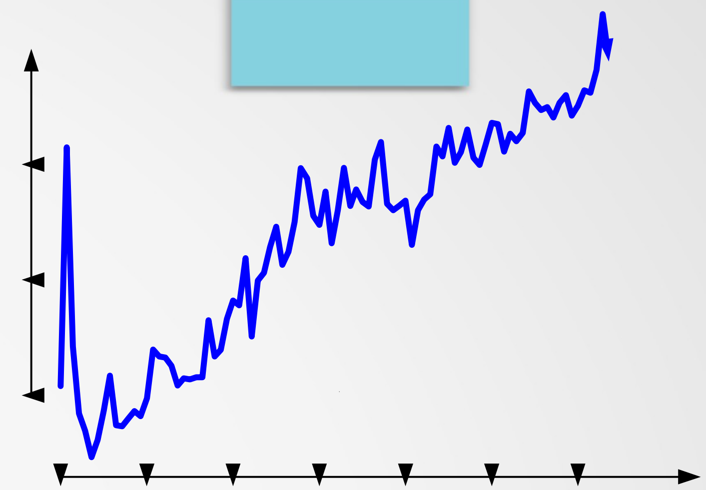
Research objectives

Behavioral Data Science

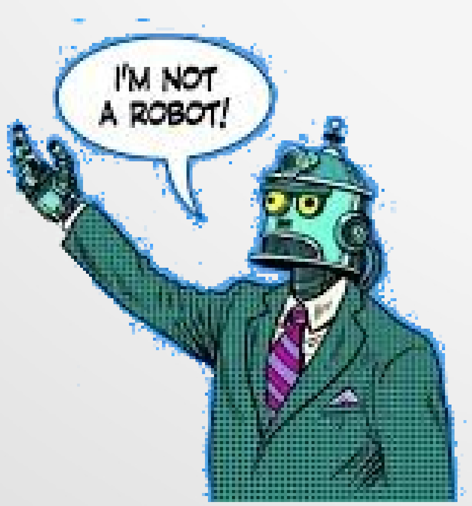
1.



information diffusion
epidemics spreading
behavioral modeling



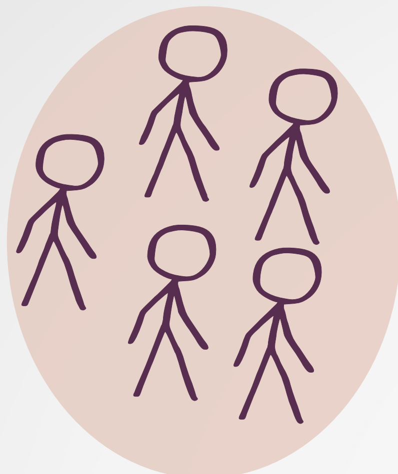
2.



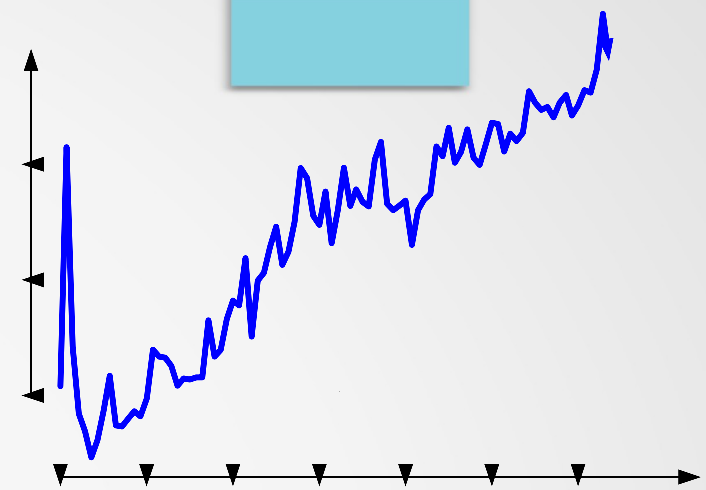
Research objectives

Behavioral Data Science

1.

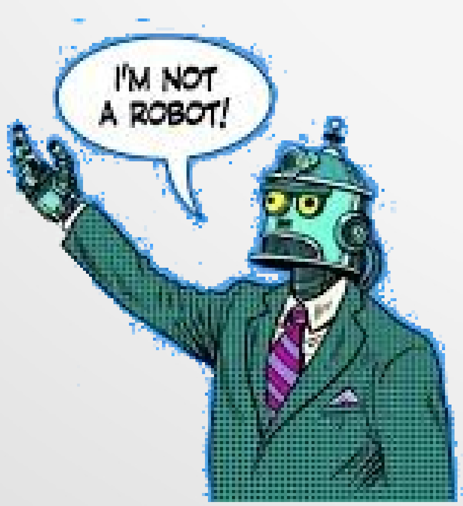


information diffusion
epidemics spreading
behavioral modeling



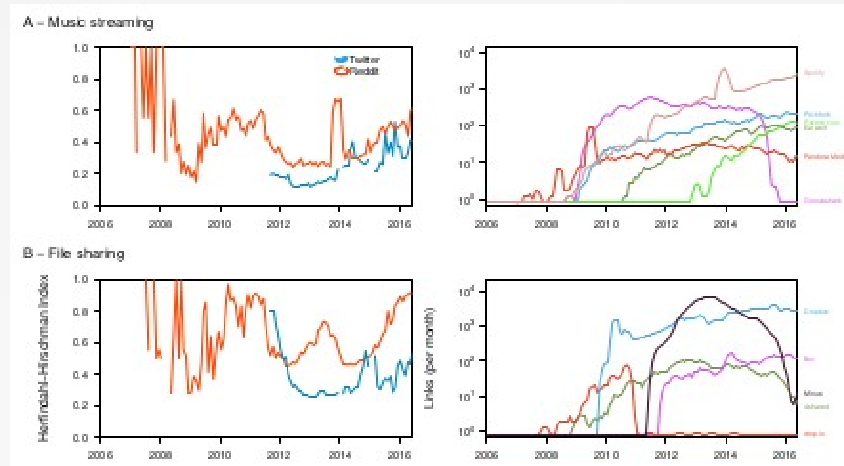
3.

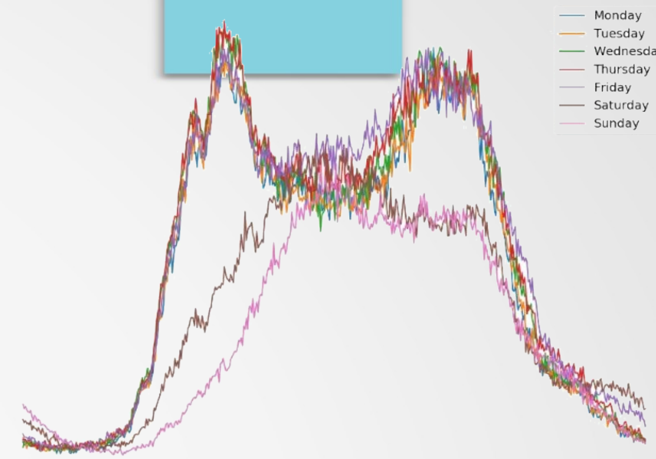
2.



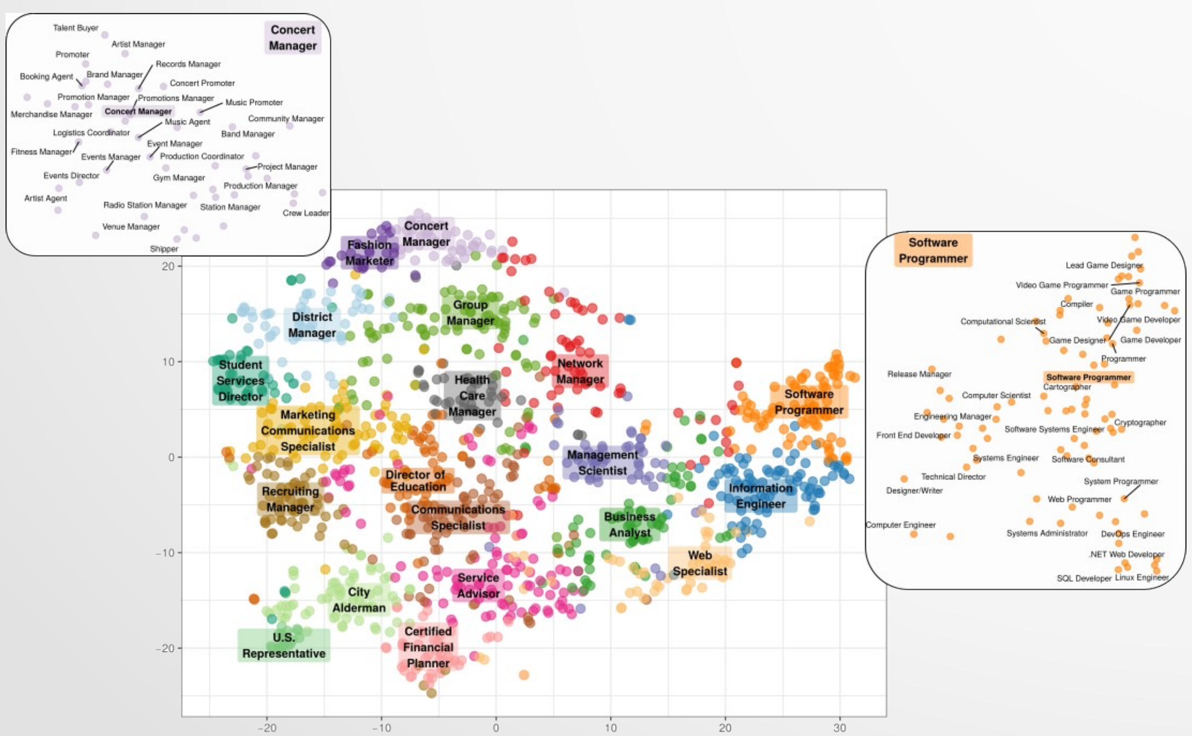
Other projects

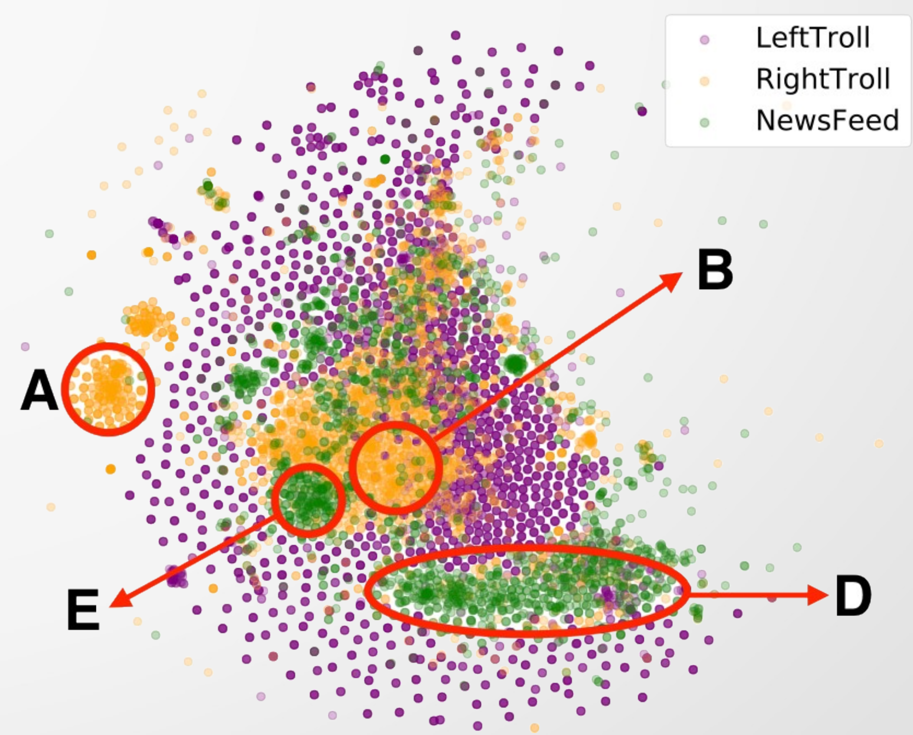
Behavioral Data Science





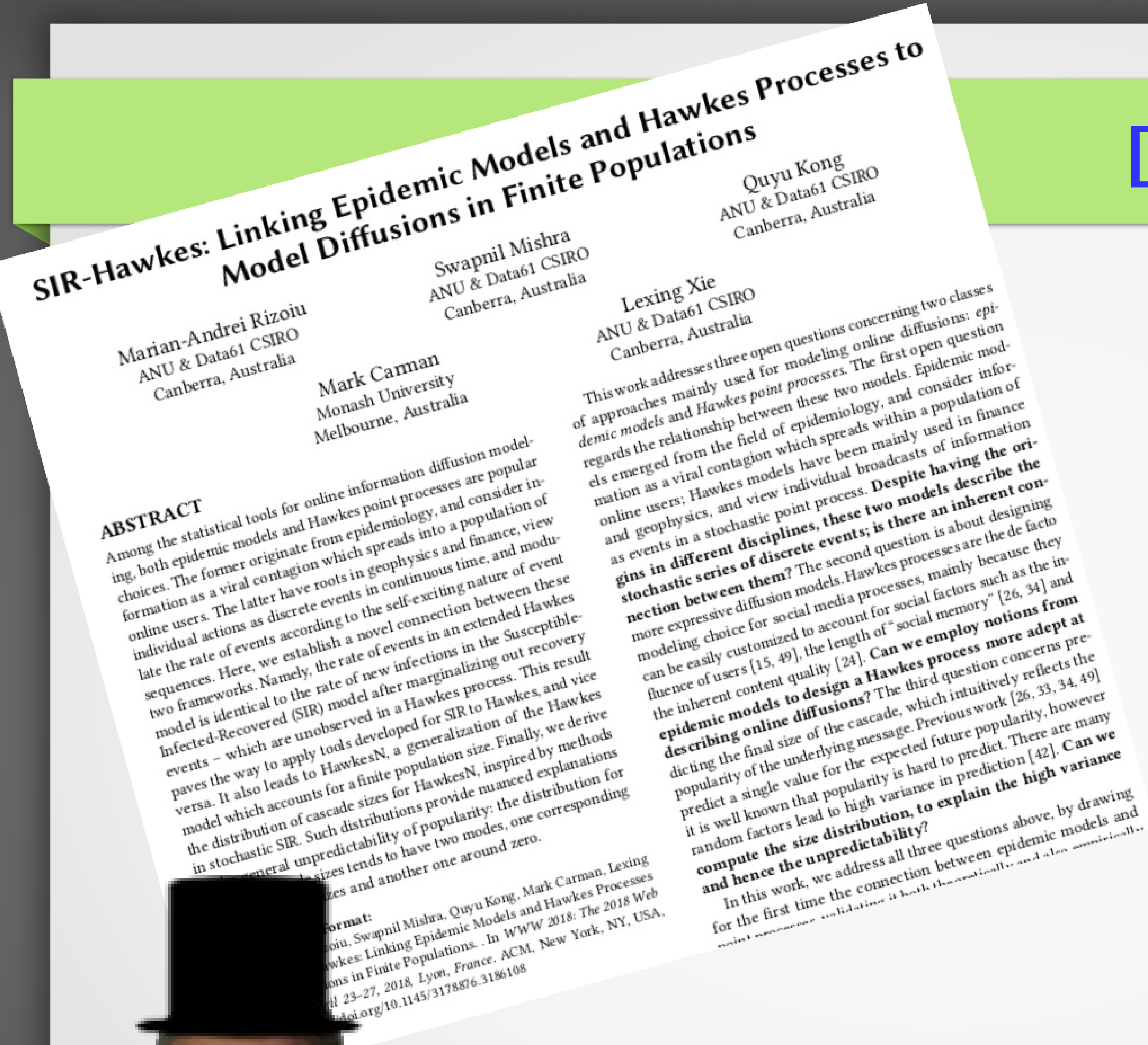
Wikipedia privacy **Online Diversity** **Smart traffic**





Vocation compass **Busting Russian Trolls**

[Rizoiu, WWW'18]



SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations

Marian-Andrei Rizoiu
ANU & Data61 CSIRO
Canberra, Australia

Swapnil Mishra
ANU & Data61 CSIRO
Canberra, Australia

Quyu Kong
ANU & Data61 CSIRO
Canberra, Australia

Mark Carman
Monash University
Melbourne, Australia

Lexing Xie
ANU & Data61 CSIRO
Canberra, Australia

ABSTRACT

Among the statistical tools for online information diffusion modeling, both epidemic models and Hawkes point processes are popular choices. The former originate from epidemiology, and consider information as a viral contagion which spreads into a population of online users. The latter have roots in geophysics and finance, view individual actions as discrete events in continuous time, and modulate the rate of events according to the self-exciting nature of event sequences. Here, we establish a novel connection between these two frameworks. Namely, the rate of events in an extended Hawkes model is identical to the rate of new infections in the Susceptible-Infected-Recovered (SIR) model after marginalizing out recovery events – which are unobserved in a Hawkes process. This result paves the way to apply tools developed for SIR to Hawkes, and vice versa. It also leads to HawkesN, a generalization of the Hawkes model which accounts for a finite population size. Finally, we derive the distribution of cascade sizes for HawkesN, inspired by methods in stochastic SIR. Such distributions provide nuanced explanations for the [illegible] unpredictability of popularity: the distribution for [illegible] sizes tends to have two modes, one corresponding [illegible] sizes and another one around zero.

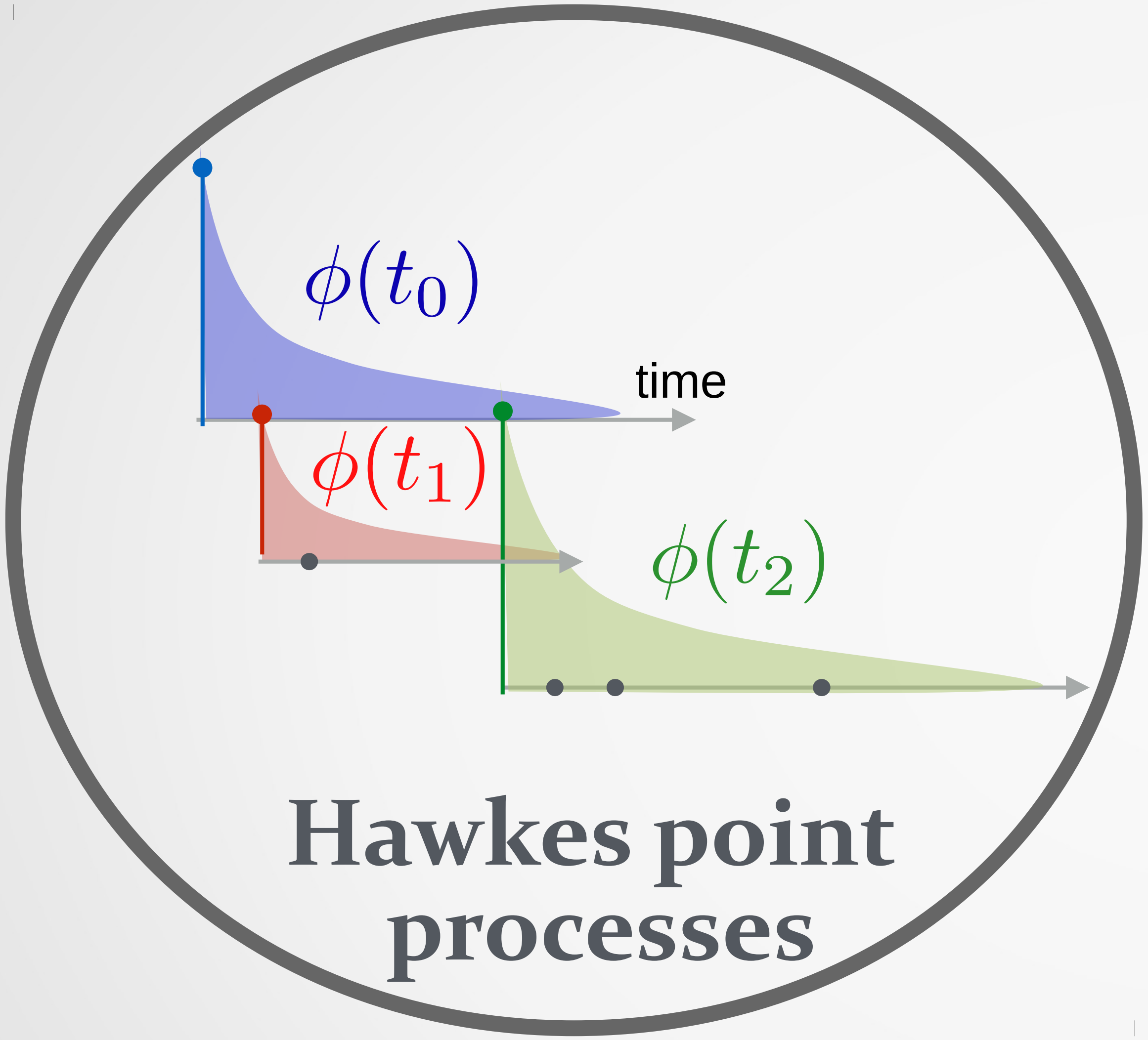
[illegible] Format:
[illegible] oiu, Swapnil Mishra, Quyu Kong, Mark Carman, Lexing [illegible] wkes: Linking Epidemic Models and Hawkes Processes [illegible] ons in Finite Populations. . In *WWW 2018: The 2018 Web* [illegible] il 23–27, 2018, *Lyon, France*. ACM, New York, NY, USA, [illegible] doi.org/10.1145/3178876.3186108

This work addresses three open questions concerning two classes of approaches mainly used for modeling online diffusions: *epidemic models* and *Hawkes point processes*. The first open question regards the relationship between these two models. Epidemic models emerged from the field of epidemiology, and consider information as a viral contagion which spreads within a population of online users; Hawkes models have been mainly used in finance and geophysics, and view individual broadcasts of information as events in a stochastic point process. **Despite having the origins in different disciplines, these two models describe the stochastic series of discrete events; is there an inherent connection between them?** The second question is about designing more expressive diffusion models. Hawkes processes are the de facto modeling choice for social media processes, mainly because they can be easily customized to account for social factors such as the influence of users [15, 49], the length of "social memory" [26, 34] and the inherent content quality [24]. **Can we employ notions from epidemic models to design a Hawkes process more adept at describing online diffusions?** The third question concerns predicting the final size of the cascade, which intuitively reflects the popularity of the underlying message. Previous work [26, 33, 34, 49] predict a single value for the expected future popularity, however it is well known that popularity is hard to predict. There are many random factors lead to high variance in prediction [42]. **Can we compute the size distribution, to explain the high variance and hence the unpredictability?**

In this work, we address all three questions above, by drawing for the first time the connection between epidemic models and [illegible]

SIR-Hawkes: Linking Epidemic Models and Hawkes Processes

Divided we model



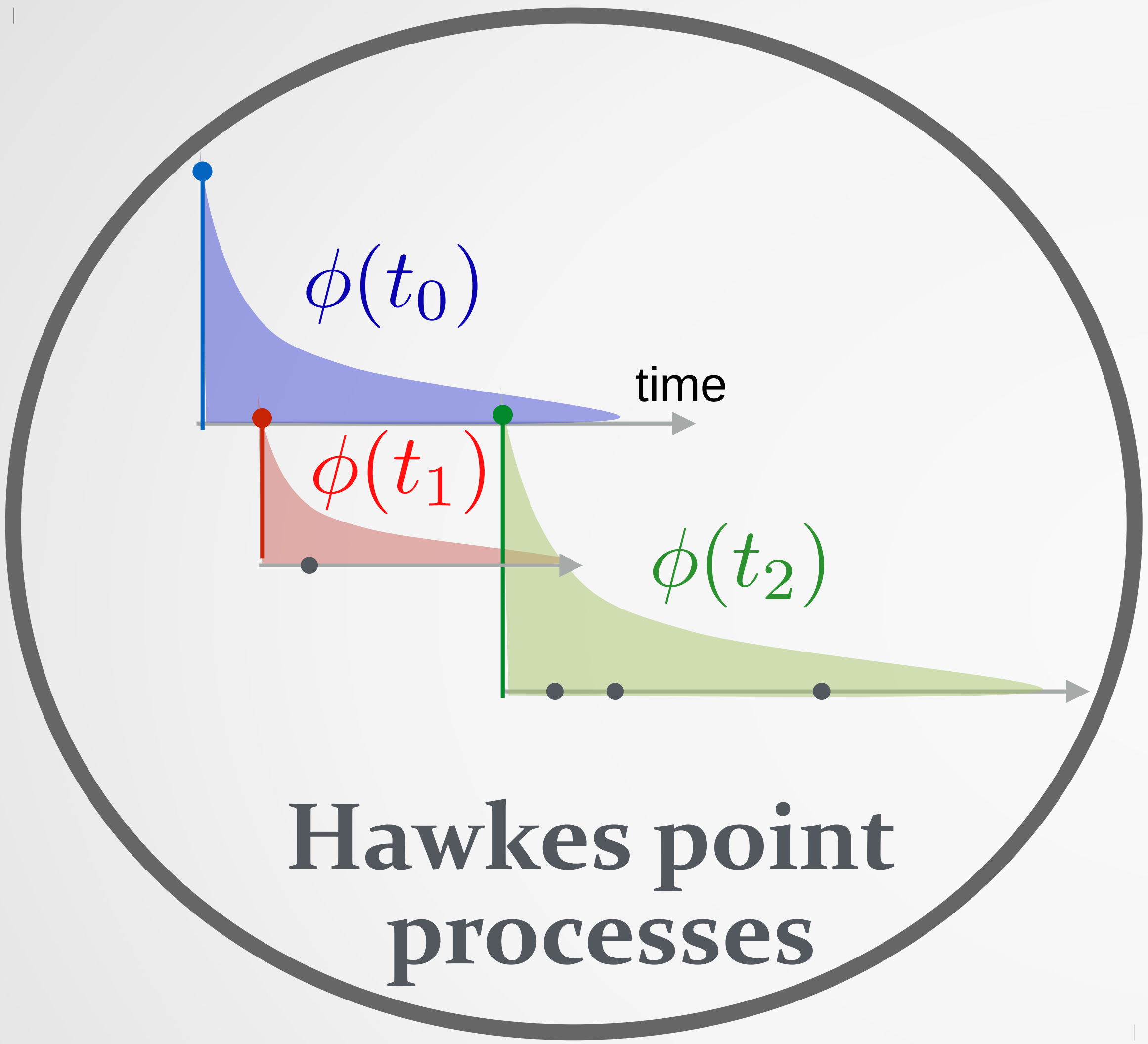
[Zhao et al KDD'15]

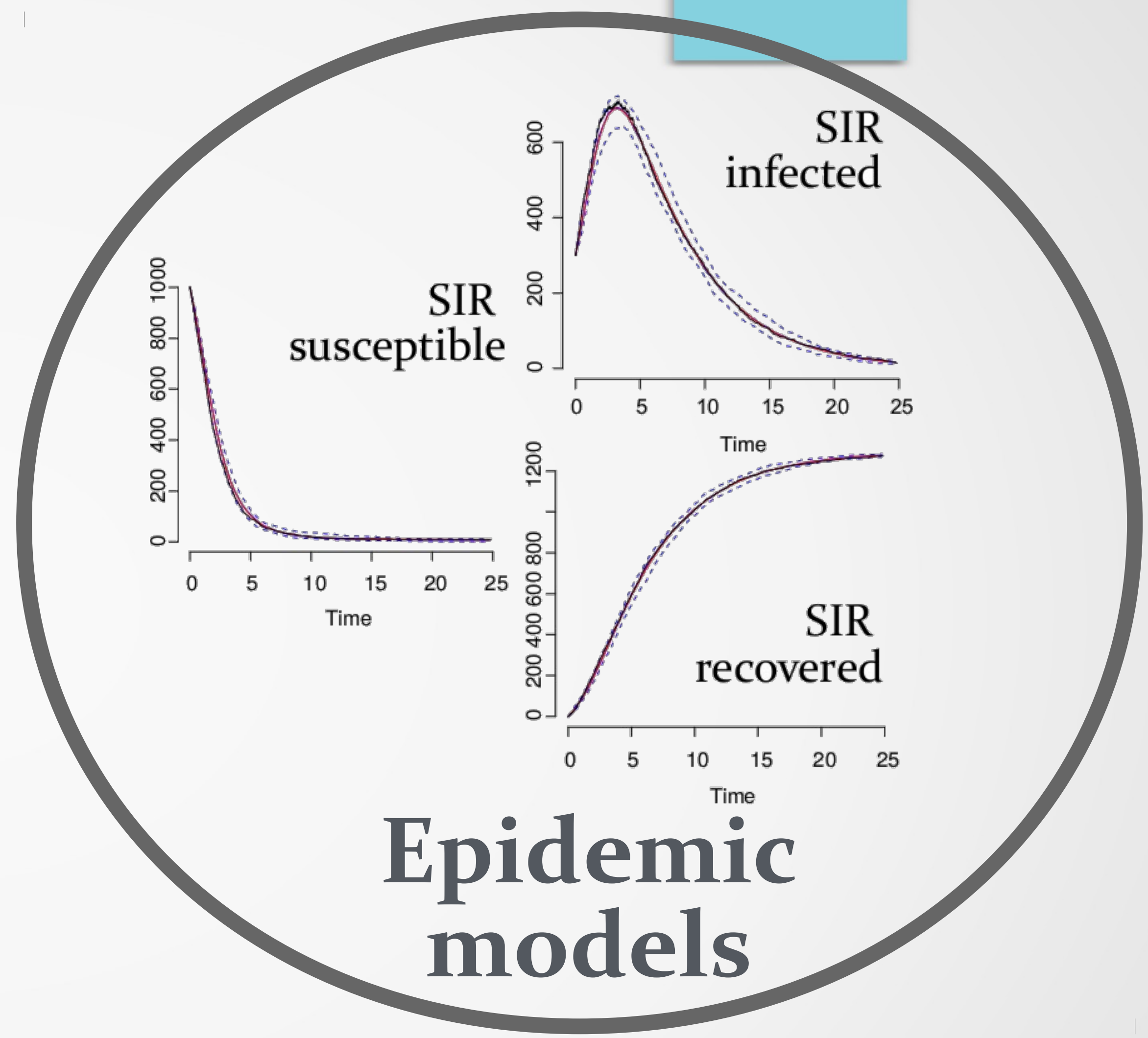
[Mishra et al CIKM'16]

[Farajtabar et al NIPS'15]

[Shen et al AAAI'14]

Divided we model





[Zhao et al KDD‘15]

[Mishra et al CIKM‘16]

[Farajtabar et al NIPS‘15]

[Shen et al AAAI‘14]

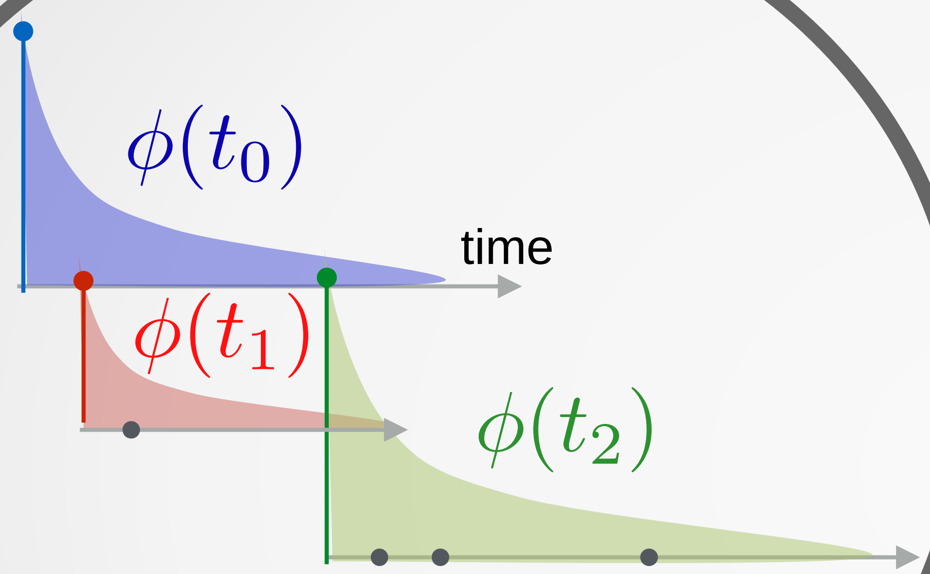
[Martin et al WWW‘16]

[Wu and Chen Springer+‘16]

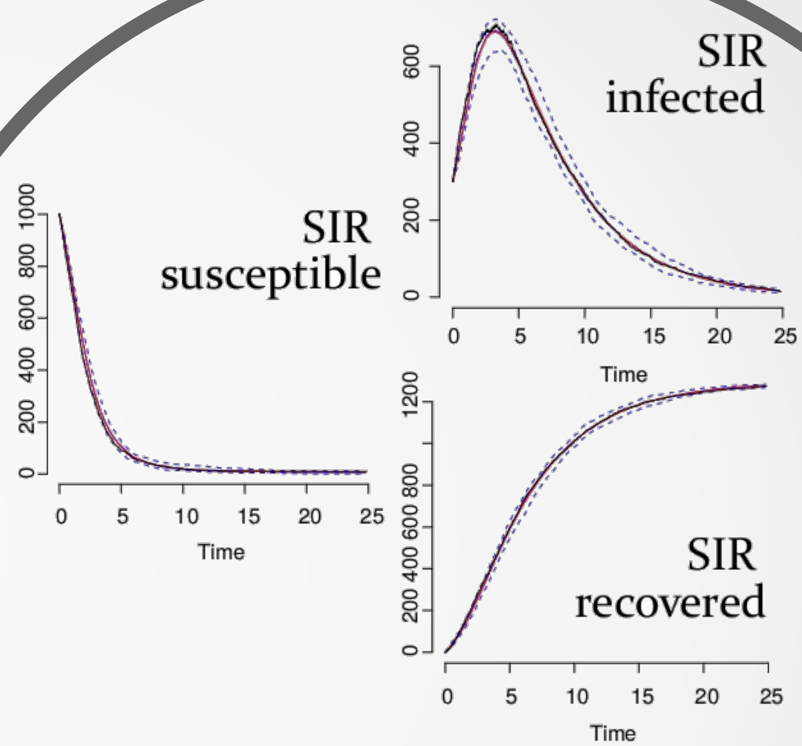
[Bauckhage et al ICWSM‘15]

[Goel et al Manag.Sci.‘15]

Divided we model



Hawkes point processes

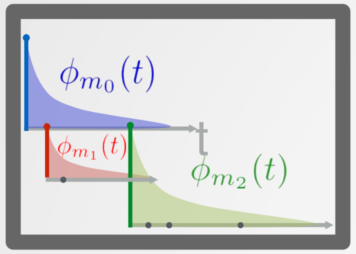


Epidemic models

[Zhao et al KDD'15]
[Mishra et al CIKM'16]
[Farajtabar et al NIPS'15]
[Shen et al AAAI'14]

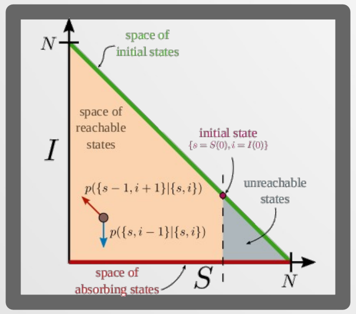
[Martin et al WWW'16]
[Wu and Chen Springer+'16]
[Bauckhage et al ICWSM'15]
[Goel et al Manag.Sci.'15]

Presentation outline



Prerequisites: Hawkes point processes and SIR infectious models

Linking SIR and the Hawkes processes



Computing the distribution of diffusion size

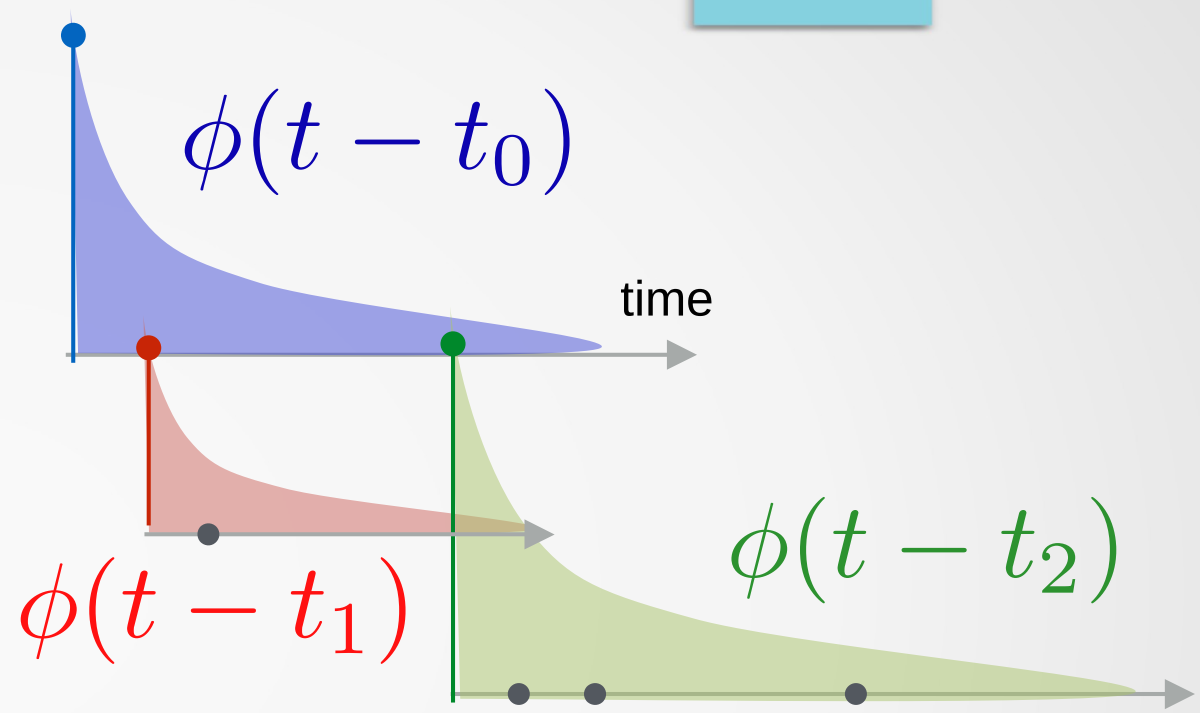
The Hawkes Process [Hawkes '71]

background event rate

$$\lambda(t) = \mu + \sum_{t_j < t} \phi(t - t_j)$$

event intensity

self-excitation



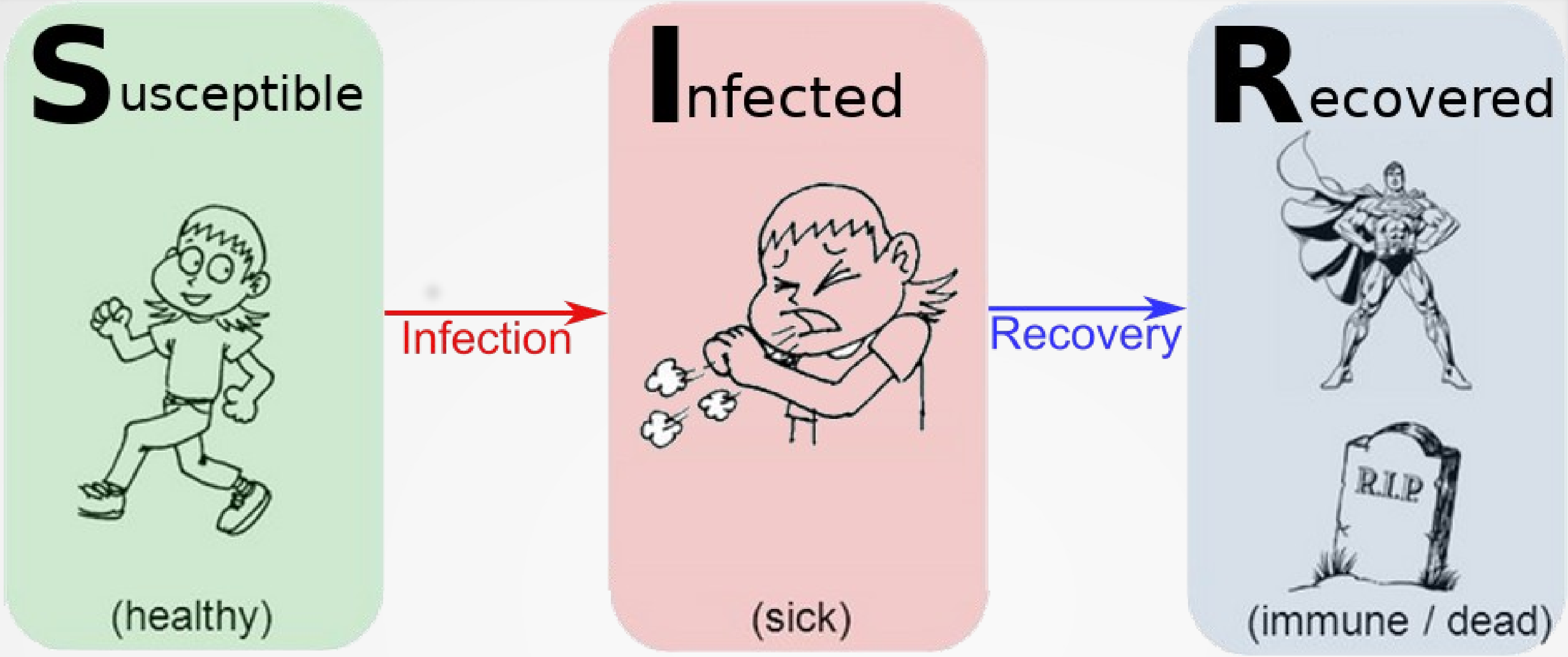
the rate of 'daughter' events

content virality

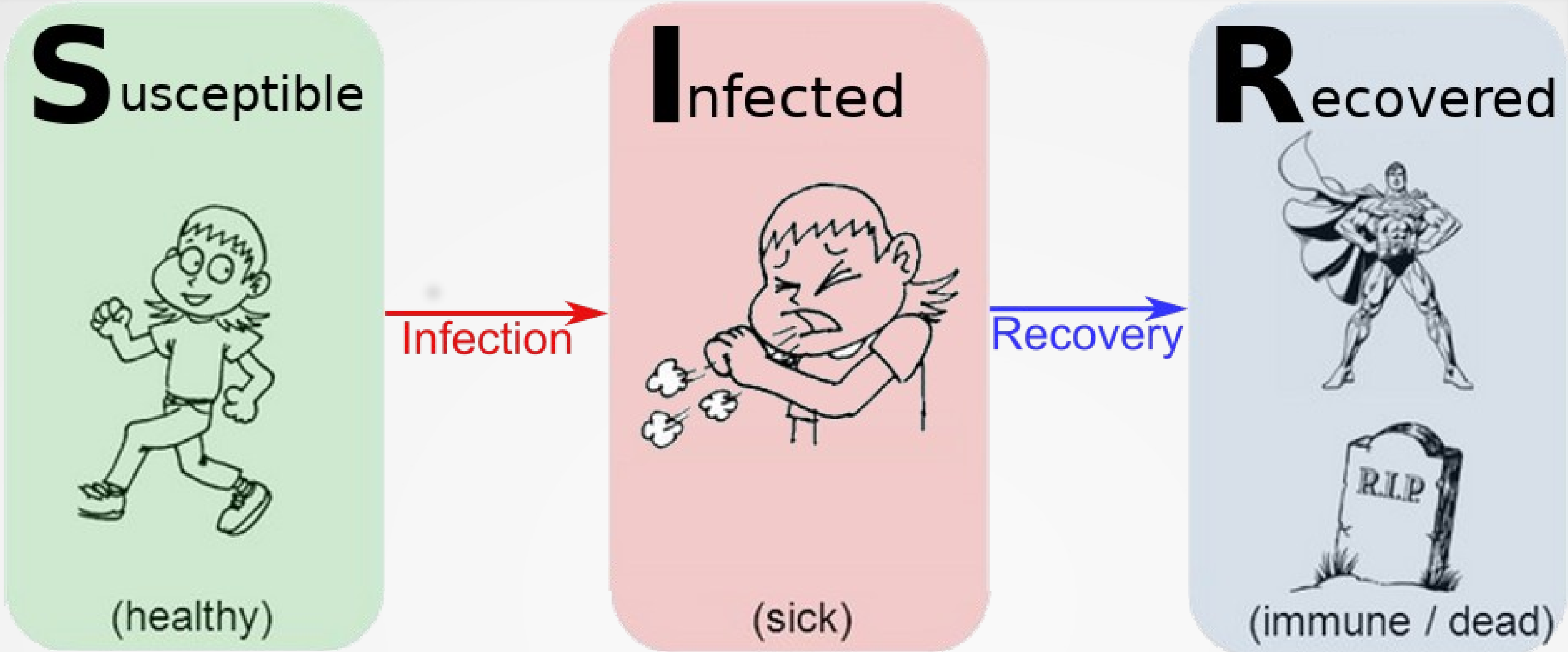
memory decay

$$\phi(\tau) = \kappa\, \theta e^{-\theta\tau}$$

The SIR epidemic model



The SIR epidemic model



infection rate

$$\frac{dS(t)}{dt} = -\beta\frac{S(t)}{N}I(t)$$

$$\frac{dI(t)}{dt} = \beta\frac{S(t)}{N}I(t) - \gamma I(t)$$

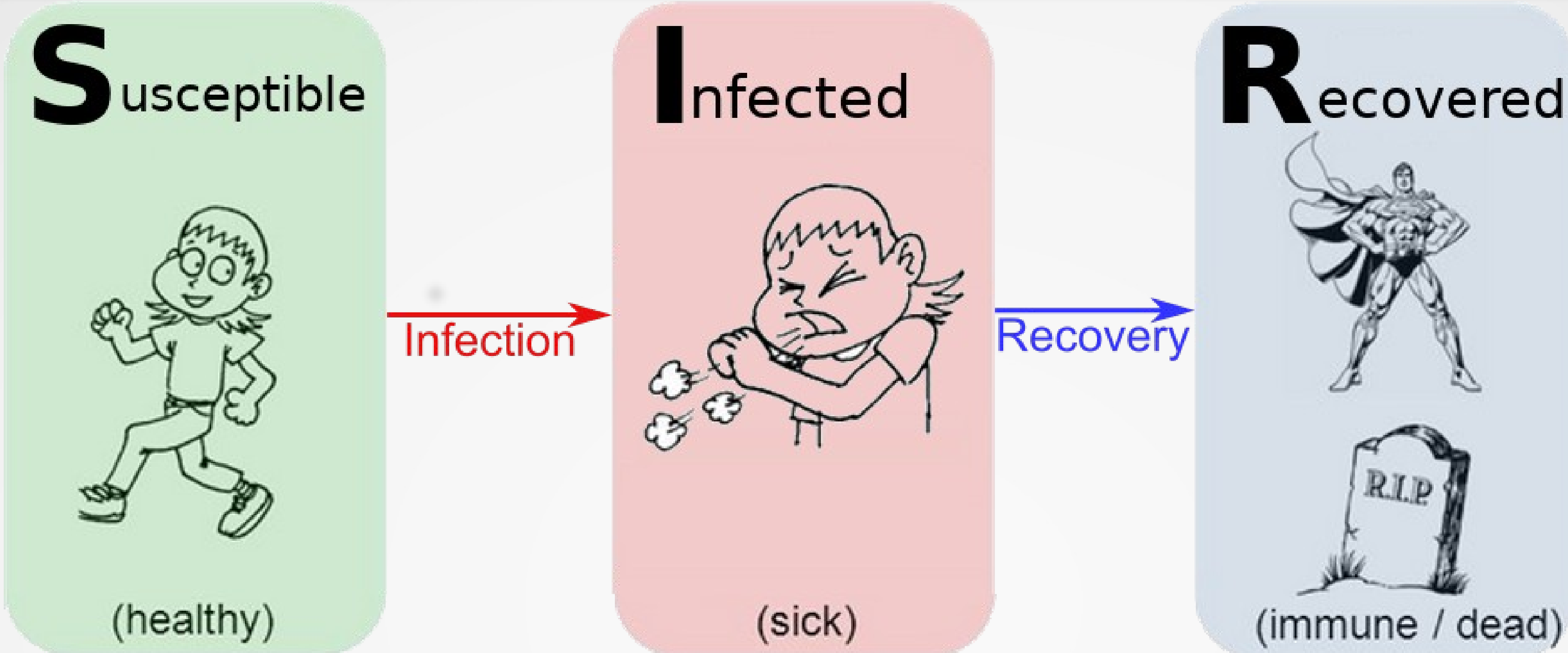
$$\frac{dR(t)}{dt} = \gamma I(t)$$

Population size (known and fixed)

recovery rate

Deterministic SIR

The SIR epidemic model



infection rate

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)}{N} I(t)$$

$$\frac{dI(t)}{dt} = \beta \frac{S(t)}{N} I(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

Population size (known and fixed)

recovery rate

Deterministic SIR

$$\lambda^I(t) = \beta \frac{S_t}{N} I_t$$

$$\lambda^R(t) = \gamma I_t$$

Stochastic SIR

SIR as a bivariate point process

Infection process C_t

Recovery process R_t

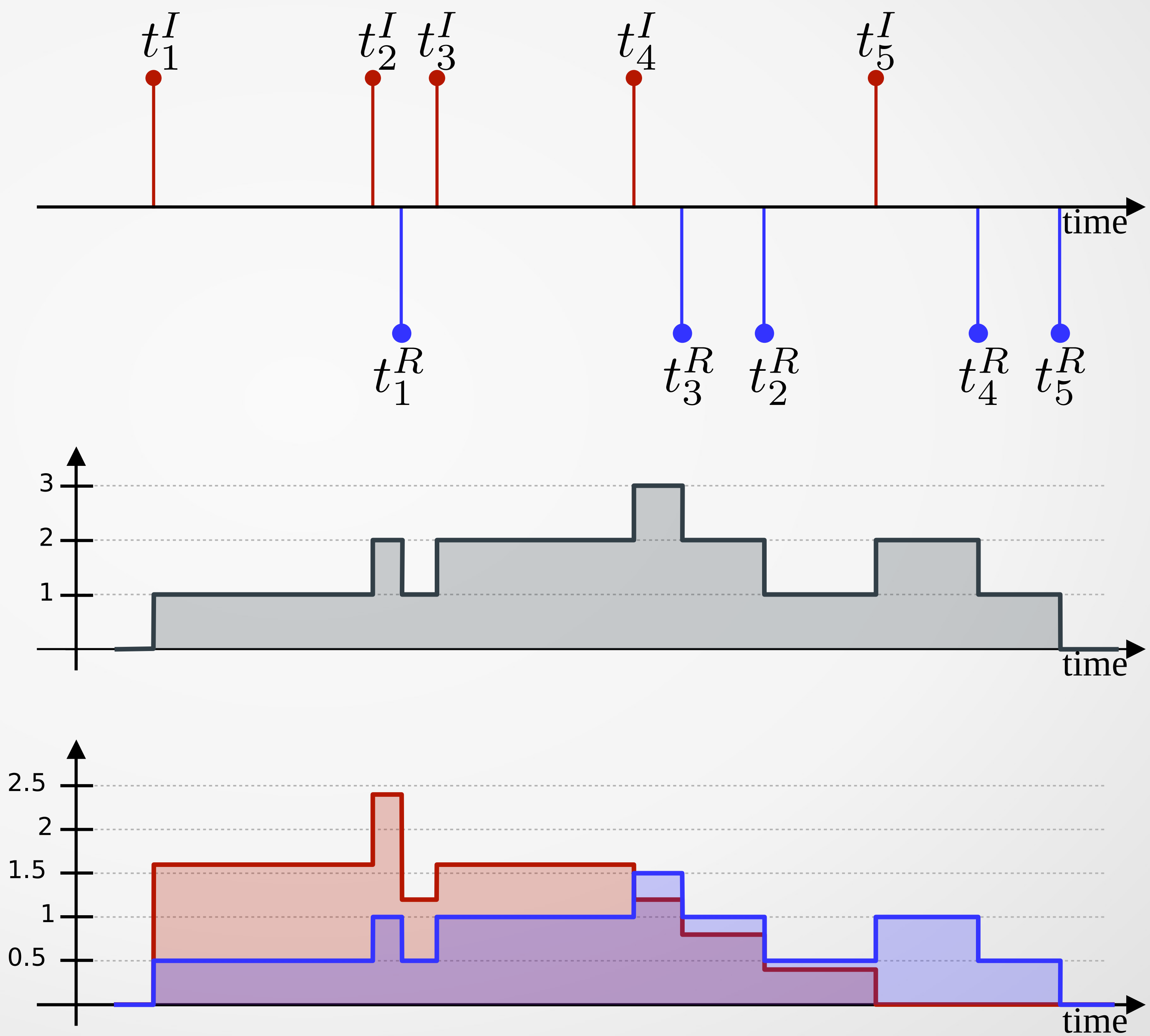
Number of infected I_t

New infection rate

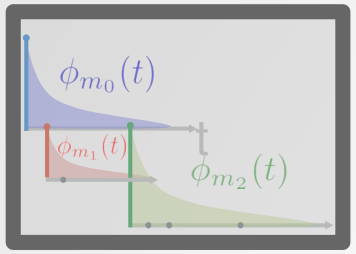
$$\lambda^I(t) = \beta \frac{S_t}{N} I_t$$

New recovery rate

$$\lambda^R(t) = \gamma I_t$$

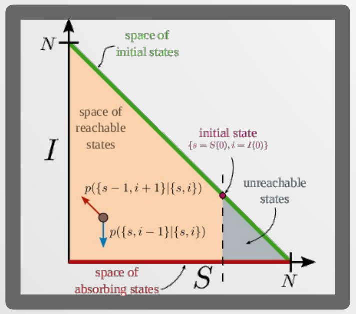


Presentation outline



Prerequisites: Hawkes point processes and SIR infectious models

Linking SIR and the Hawkes processes



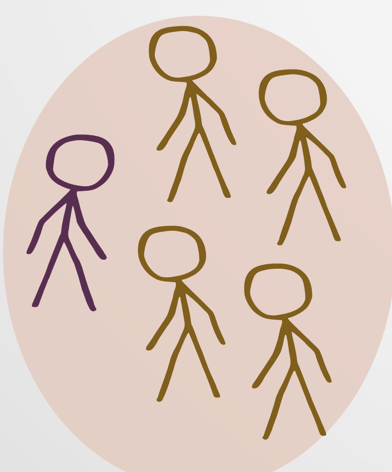
Computing the distribution of diffusion size

A finite population Hawkes model

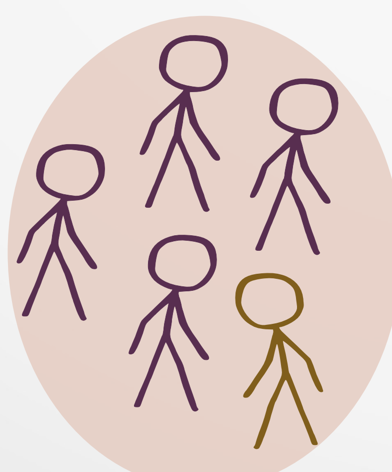
Goal: Introduce population size in Hawkes

HawkesN: modulate the event intensity by the size of the available population:

$$\lambda^H(t) = \left(1 - \frac{N_t}{N}\right) \underbrace{\left[\mu + \sum_{t_j < t} \phi(t - t_j)\right]}_{\textbf{Hawkes}\text{ intensity}}$$



80% susceptible

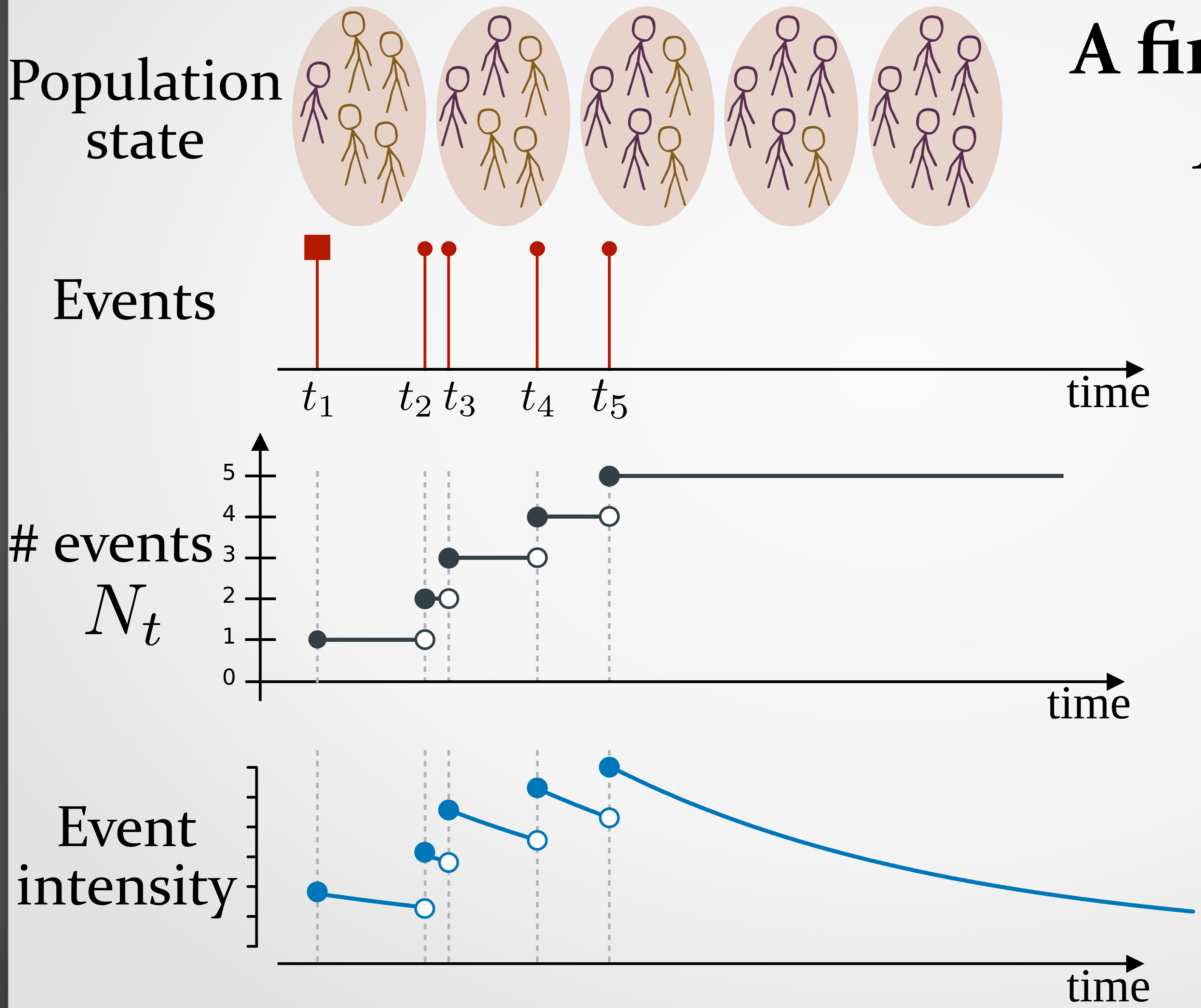


20% susceptible

Example: a HawkesN diffusion

A finite population

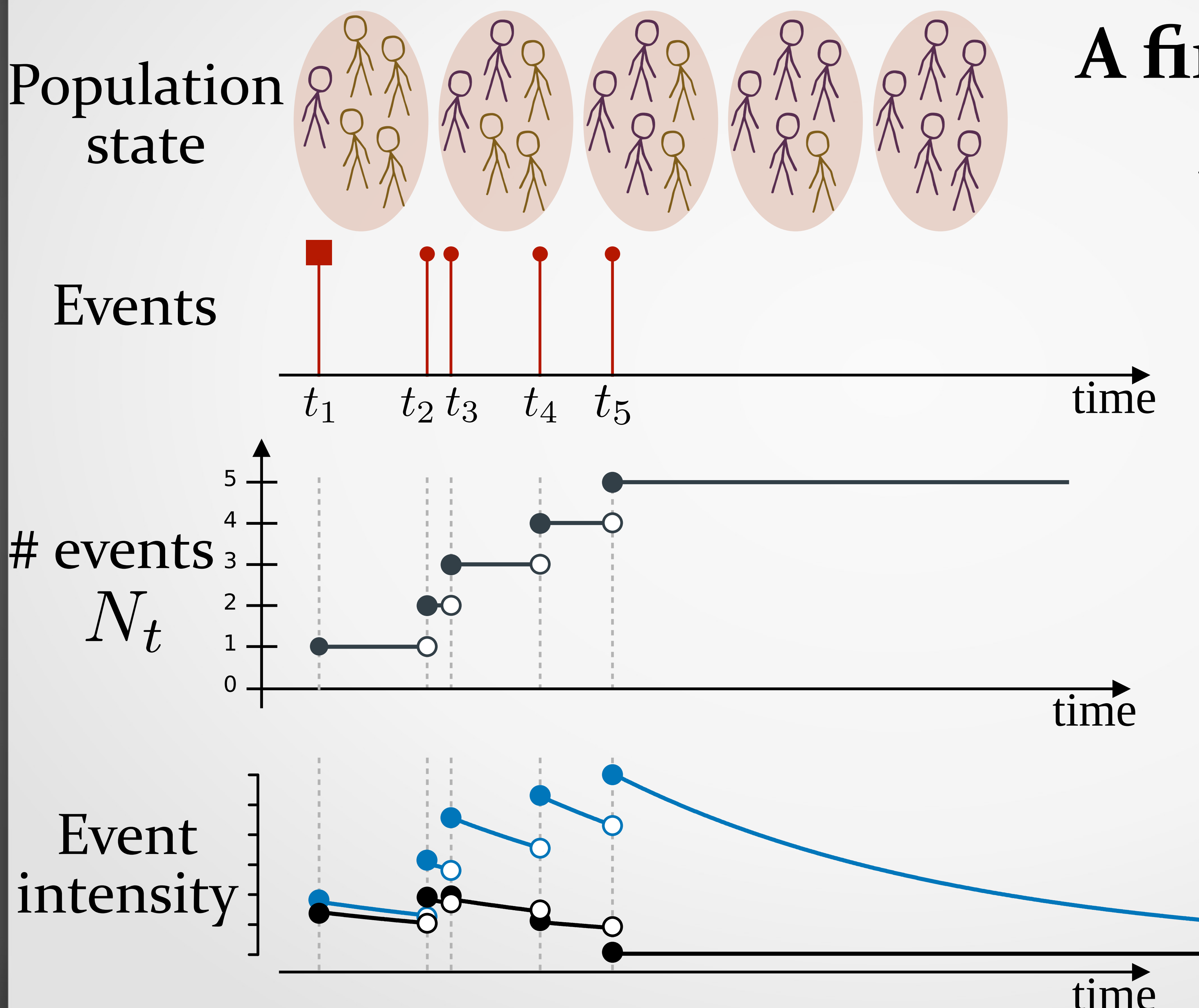
$$N = 5$$



Hawkes

$$\lambda(t) = \mu + \sum_{t_j < t} \phi(t - t_j)$$

Example: a HawkesN diffusion



A finite population

$$N = 5$$

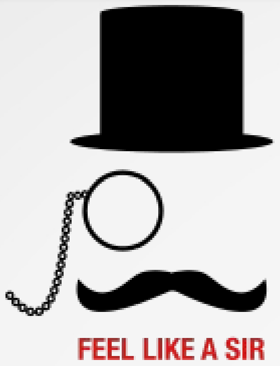
Hawkes

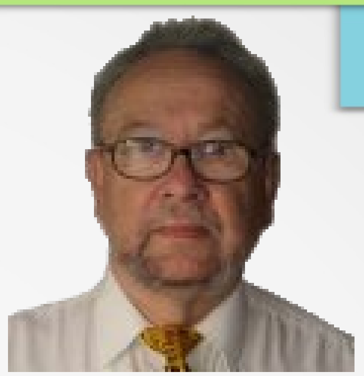
$$\lambda(t) = \mu + \sum_{t_j < t} \phi(t - t_j)$$

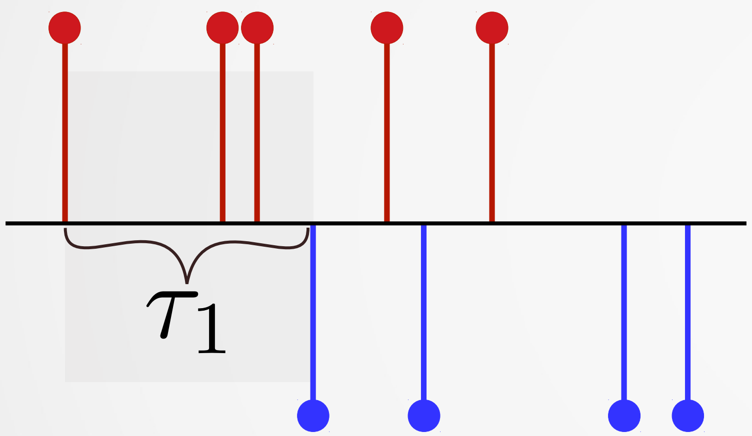
HawkesN

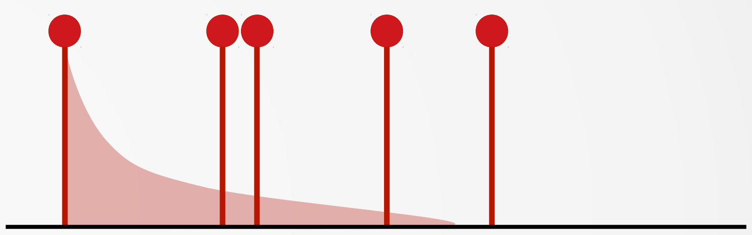
$$\lambda^H(t) = \left(1 - \frac{N_t}{N}\right) \lambda(t)$$

Linking SIR and Hawkes





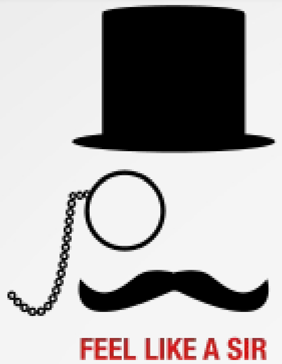


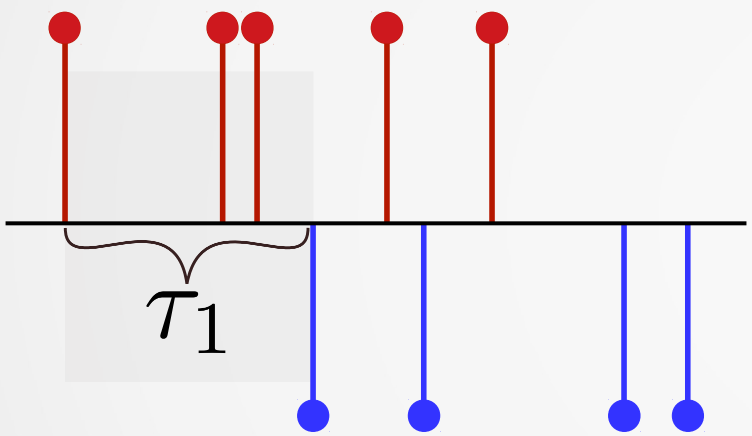


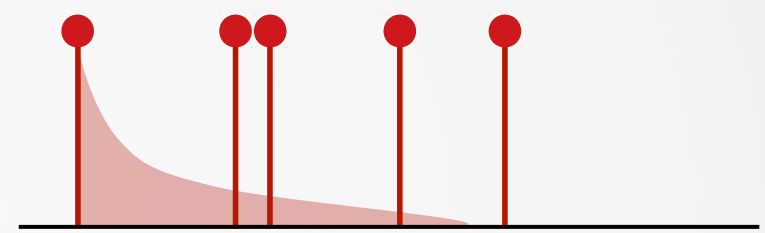
$SIR(\beta, \gamma)$

$HawkesN(\mu, \kappa, \theta)$

Linking SIR and Hawkes







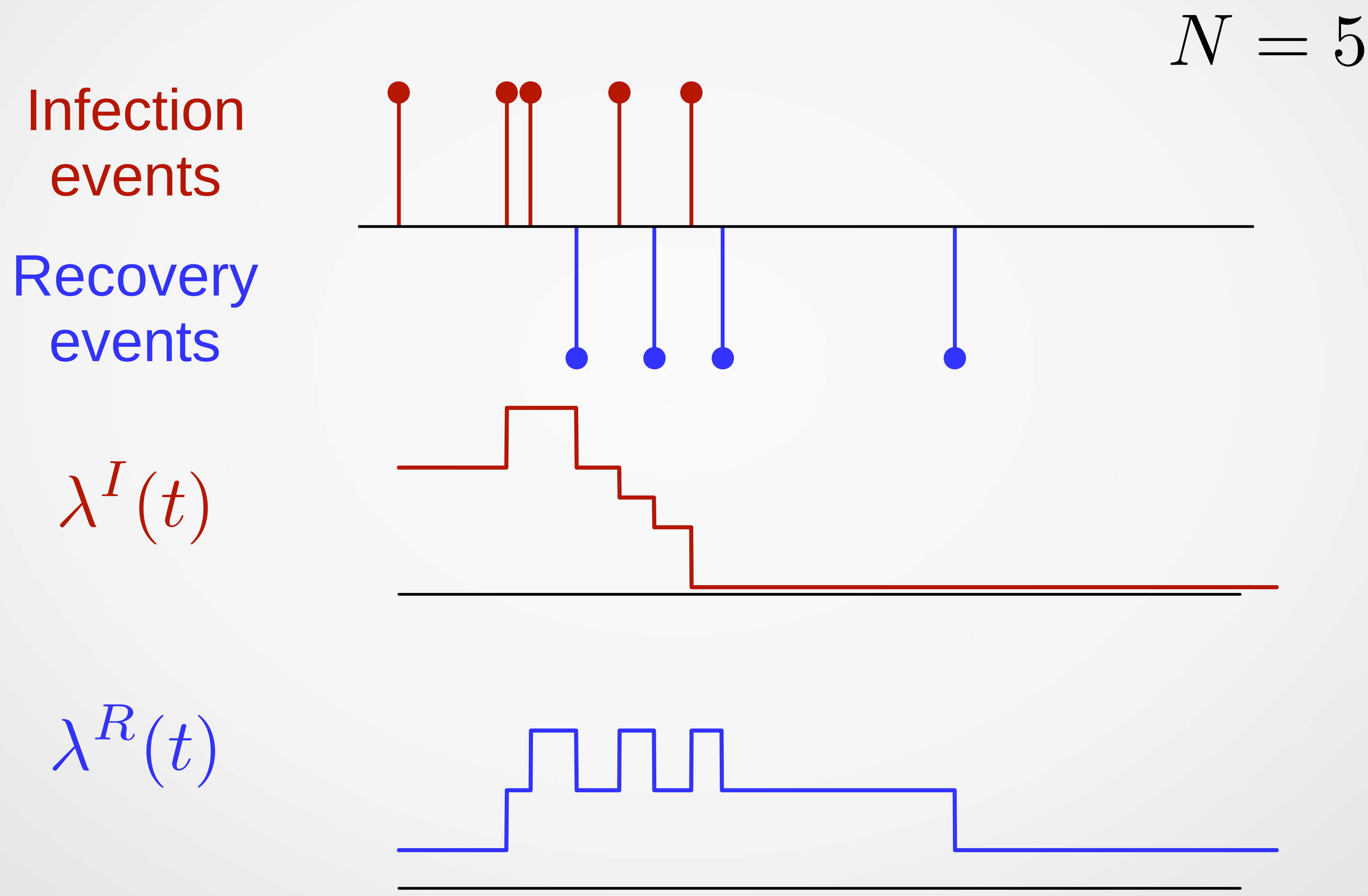
$$SIR(\beta, \gamma)$$

$$HawkesN(\mu, \kappa, \theta)$$

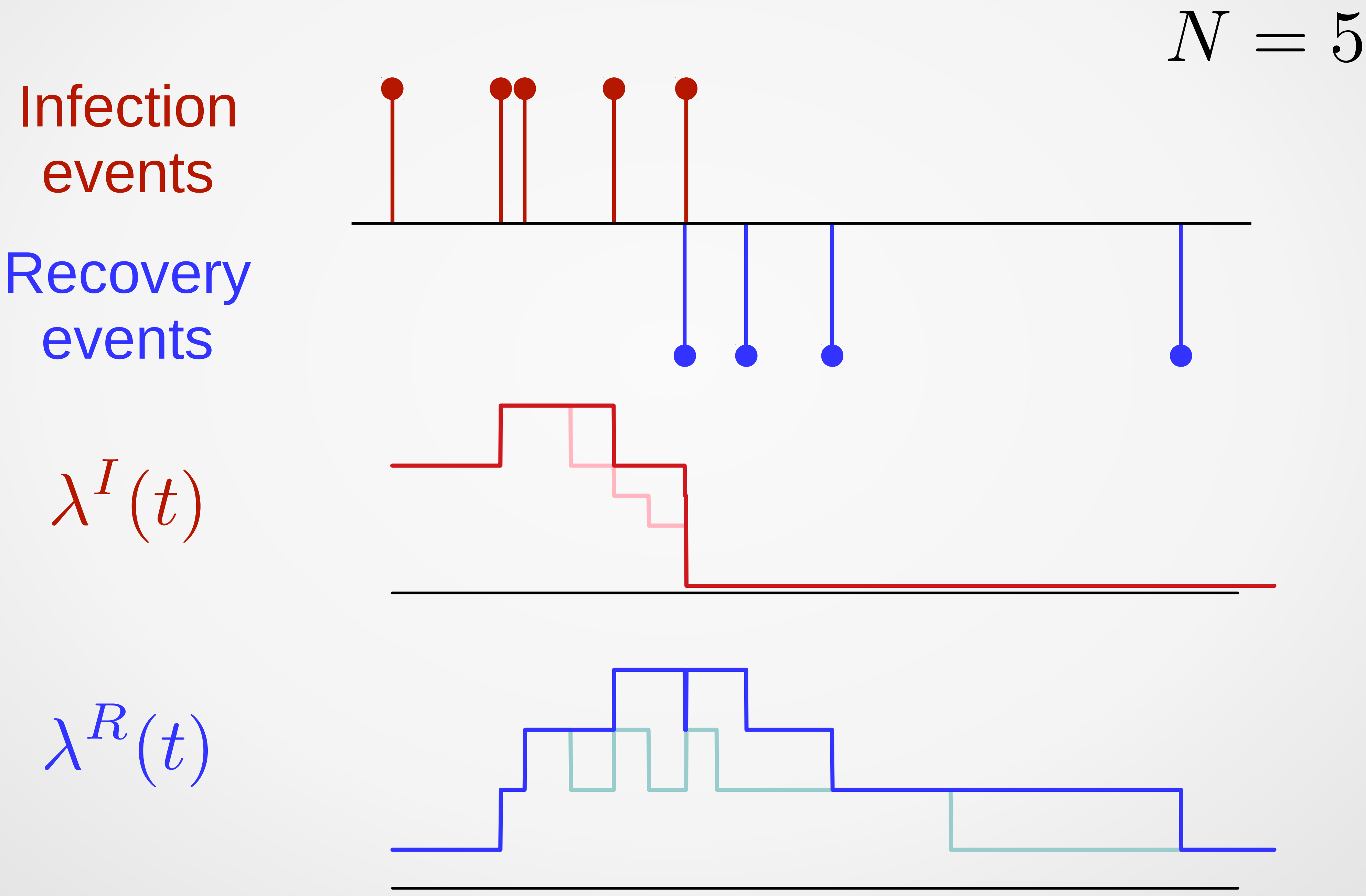
$$\mathbb{E}_{t^R}[\lambda^I(t)] = \lambda^H(t)$$

where $\mu = 0,\ \beta = \kappa\theta,\ \gamma = \theta$

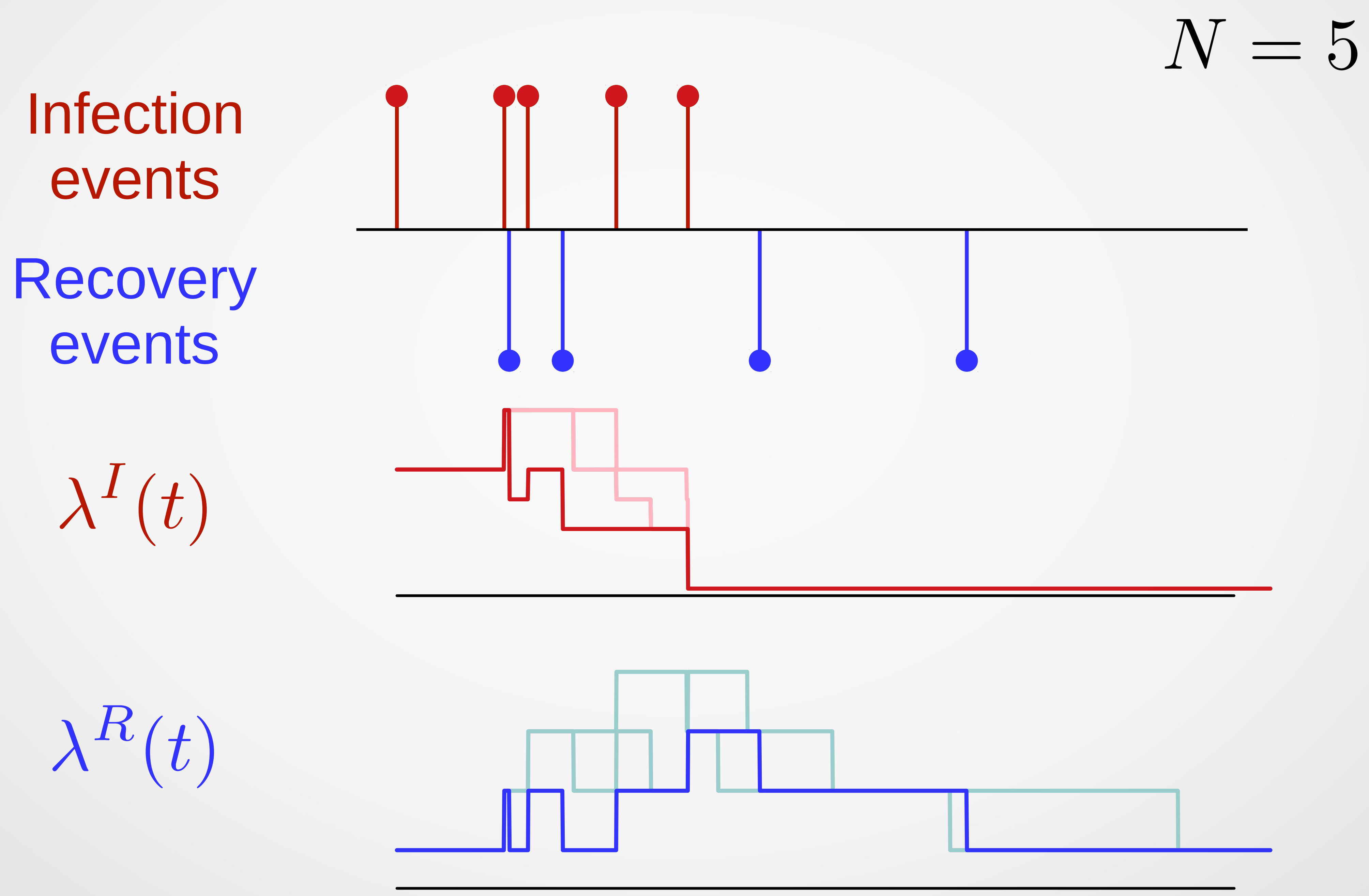
From SIR to HawkesN



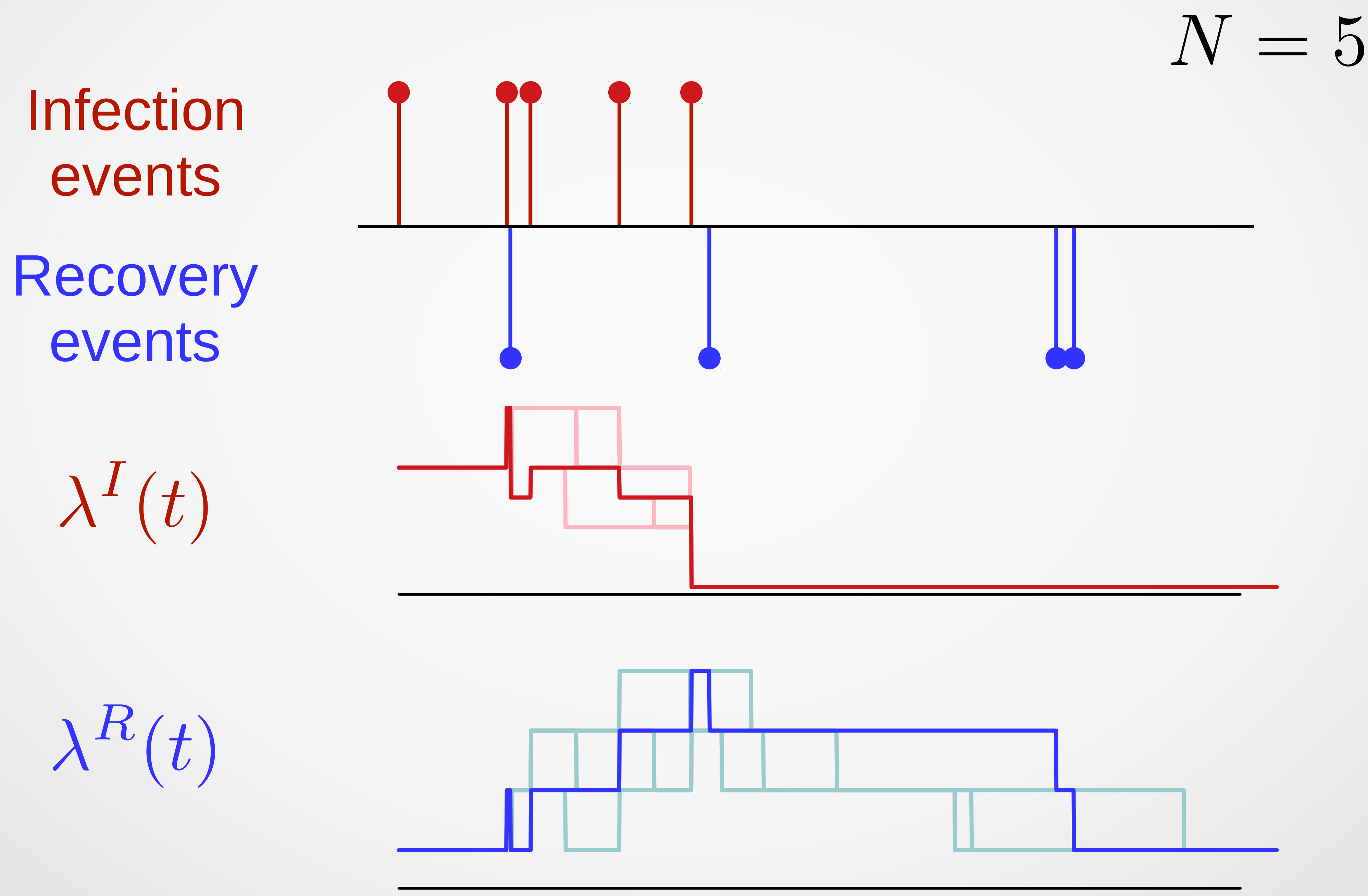
From SIR to HawkesN



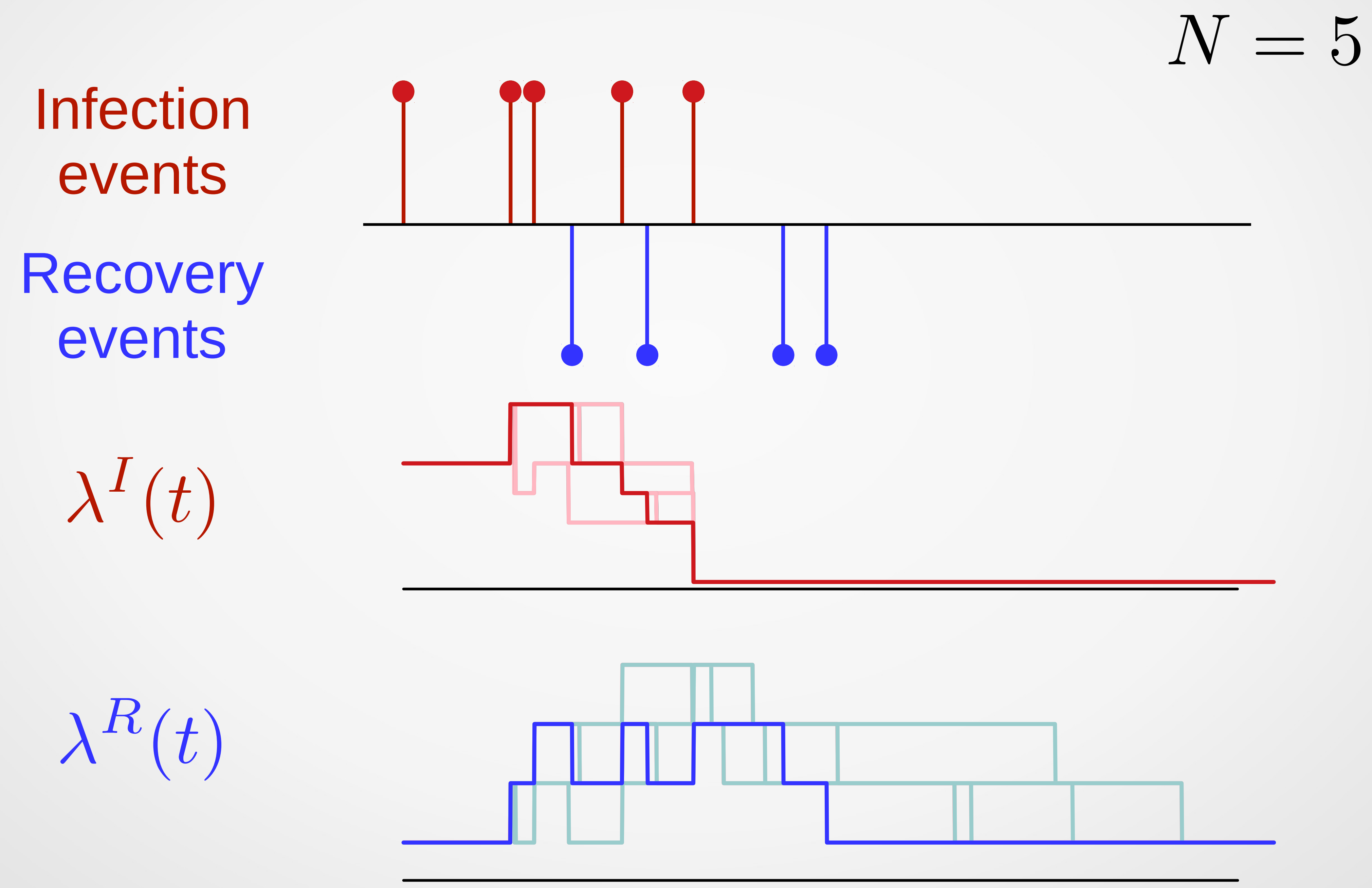
From SIR to HawkesN



From SIR to HawkesN

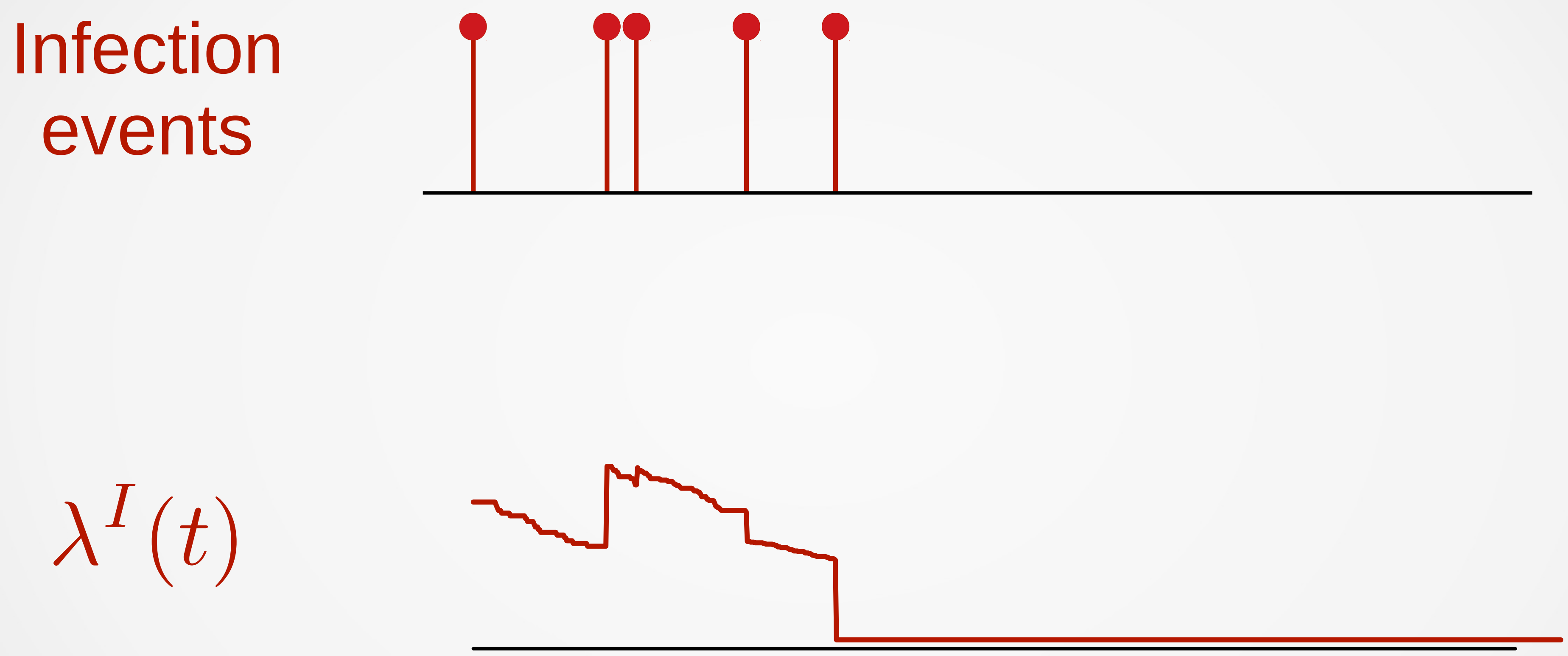


From SIR to HawkesN



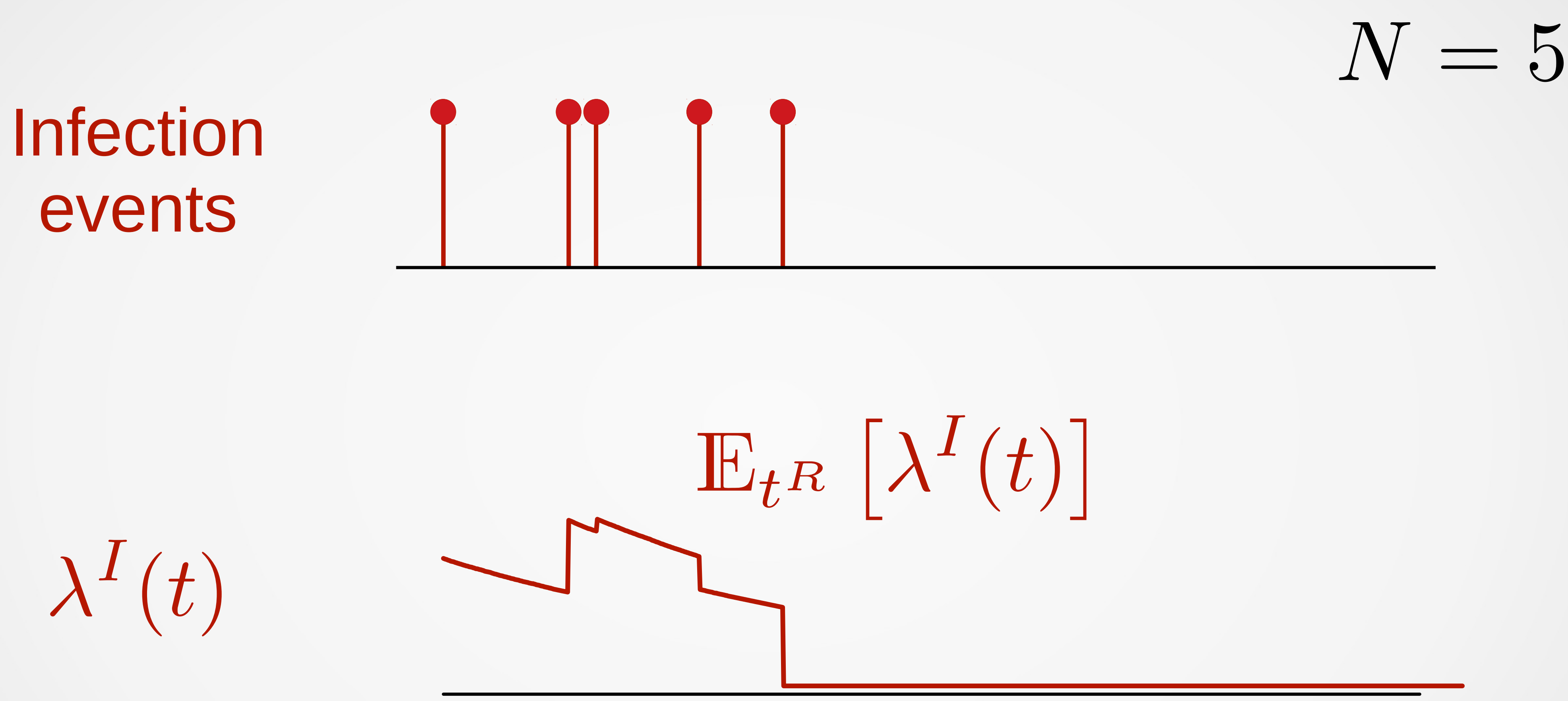
From SIR to HawkesN

$N = 5$



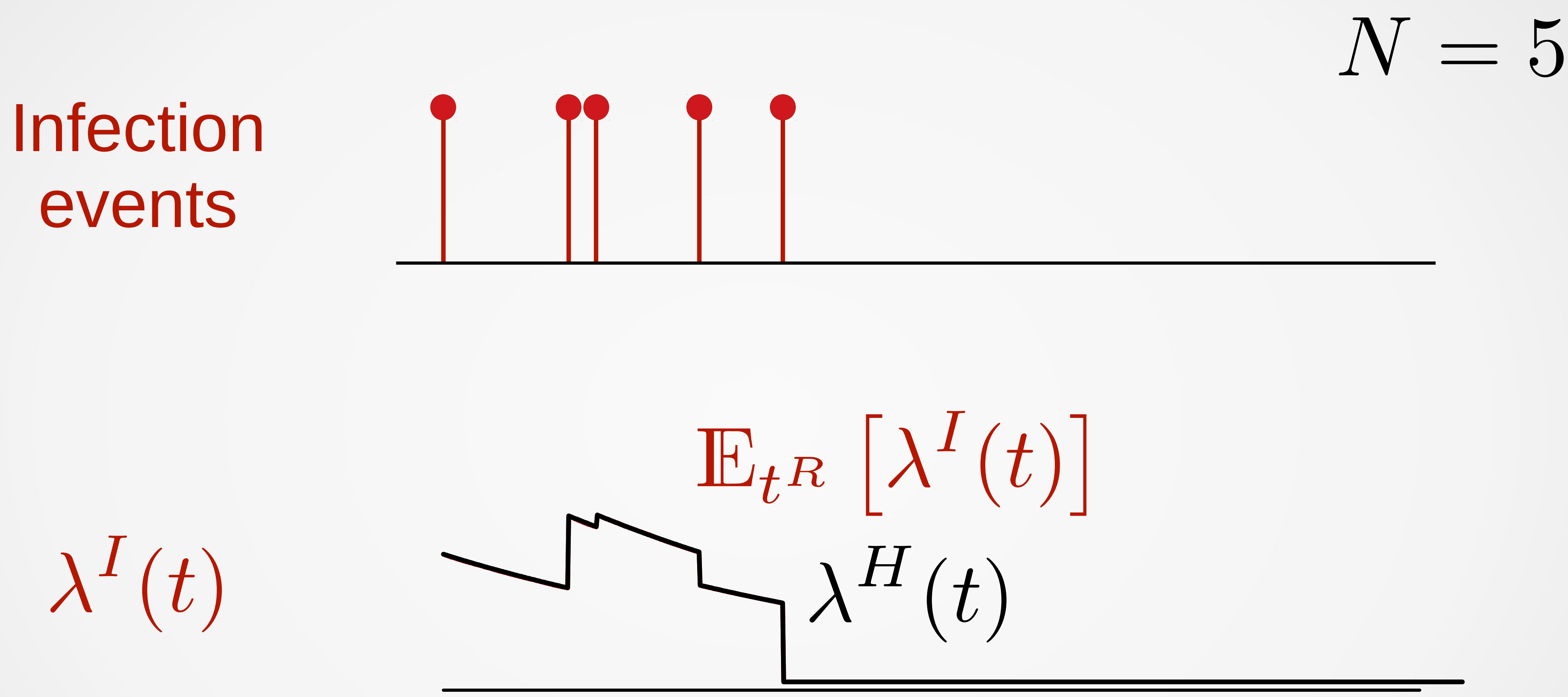
Aggregated over 50 recovery realizations

From SIR to HawkesN



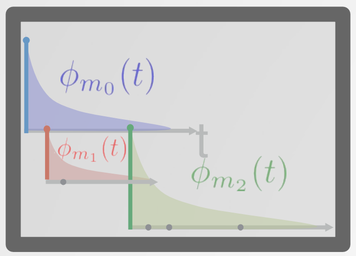
Aggregated over 10,000 recovery realizations

From SIR to HawkesN



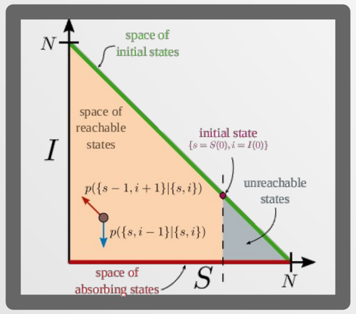
The event intensity of the equivalent HawkesN is the expected new infections intensity

Presentation outline



Prerequisites: Hawkes point processes and SIR infectious models

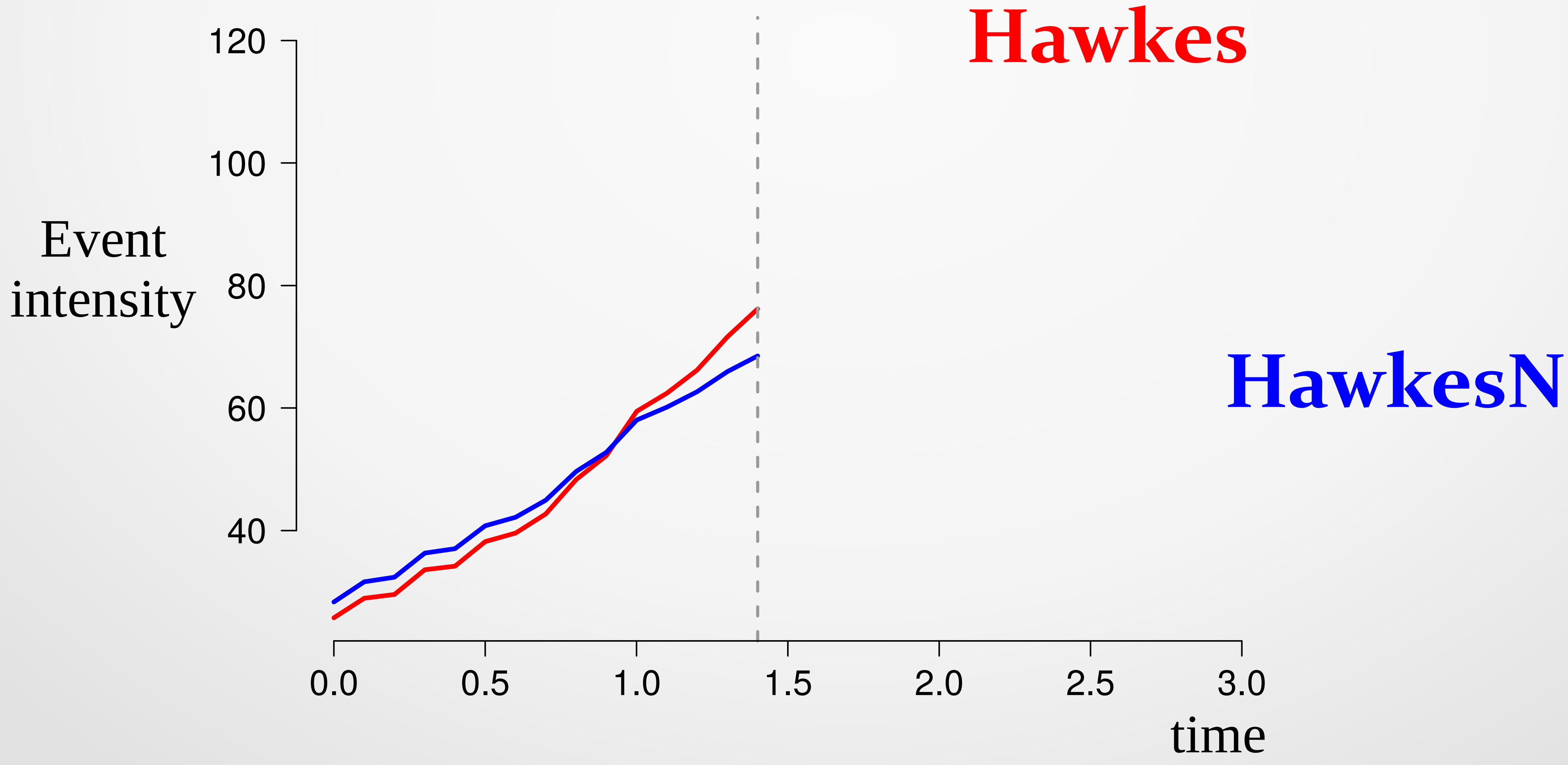
Linking SIR and the Hawkes processes



Computing the distribution of diffusion size

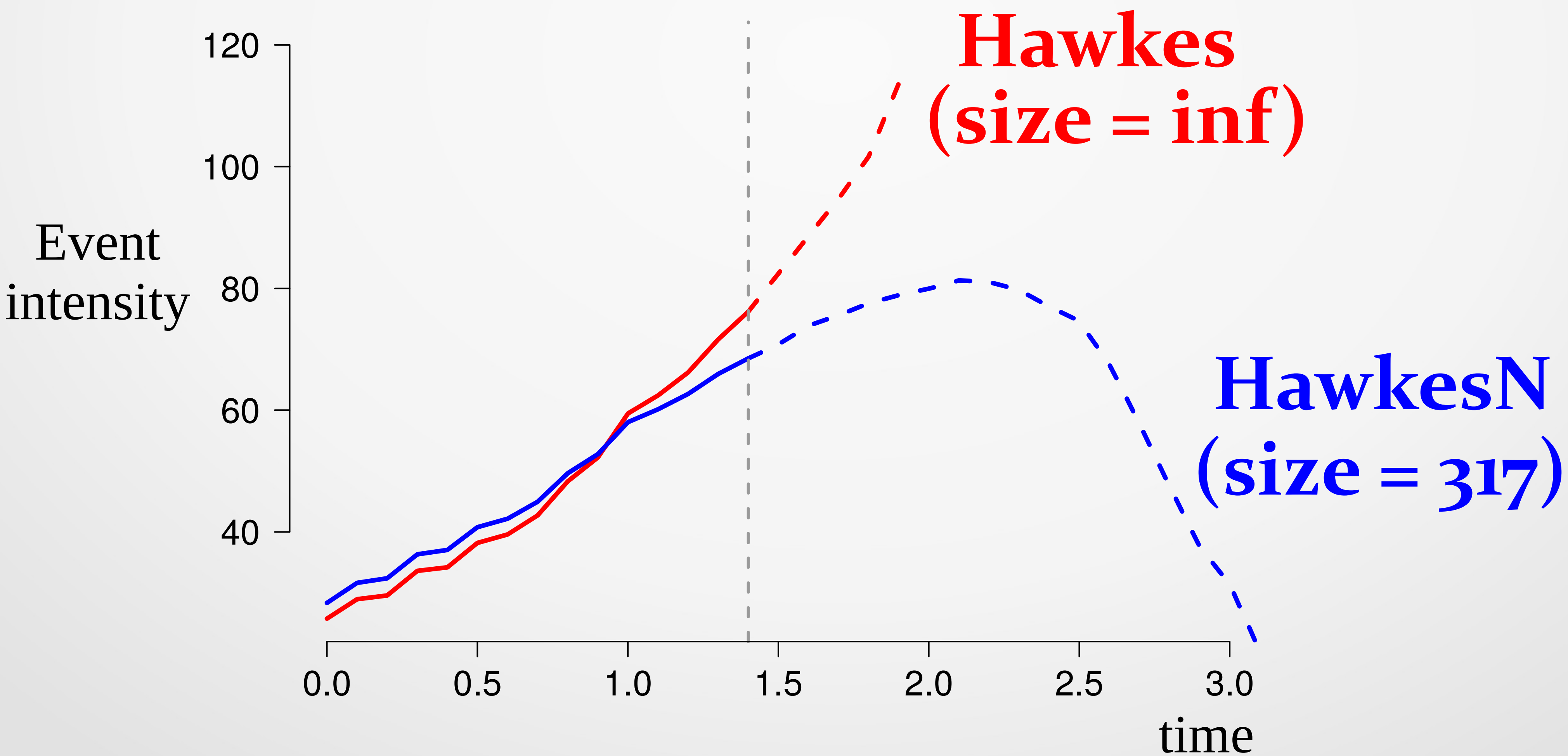
Hawkes and HawkesN in prediction

- 100 observed events;
- predict the final size of the cascade.



Hawkes and HawkesN in prediction

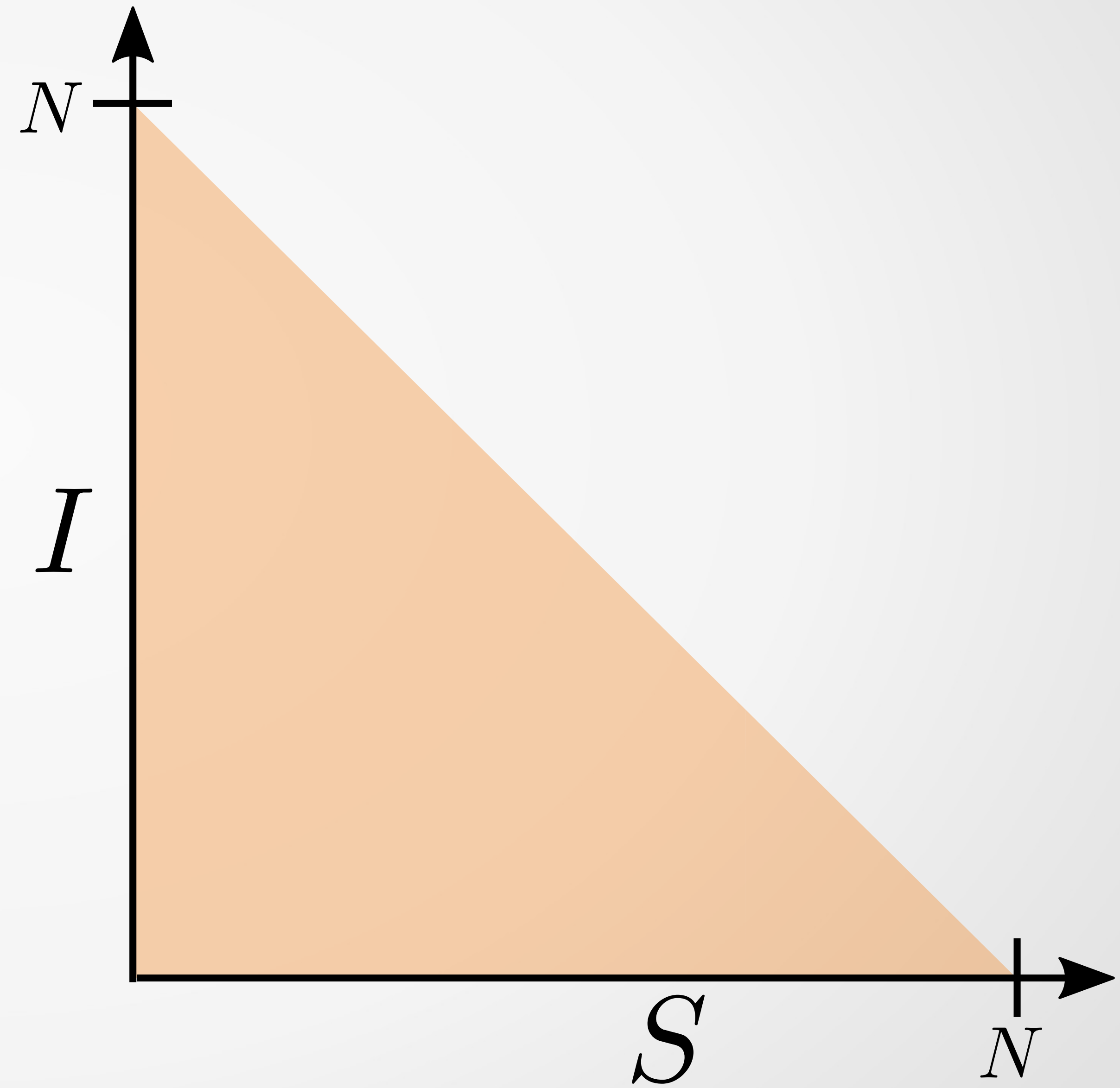
- 100 observed events;
- predict the final size of the cascade.



Distribution of total size

using an SIR Markov chain technique

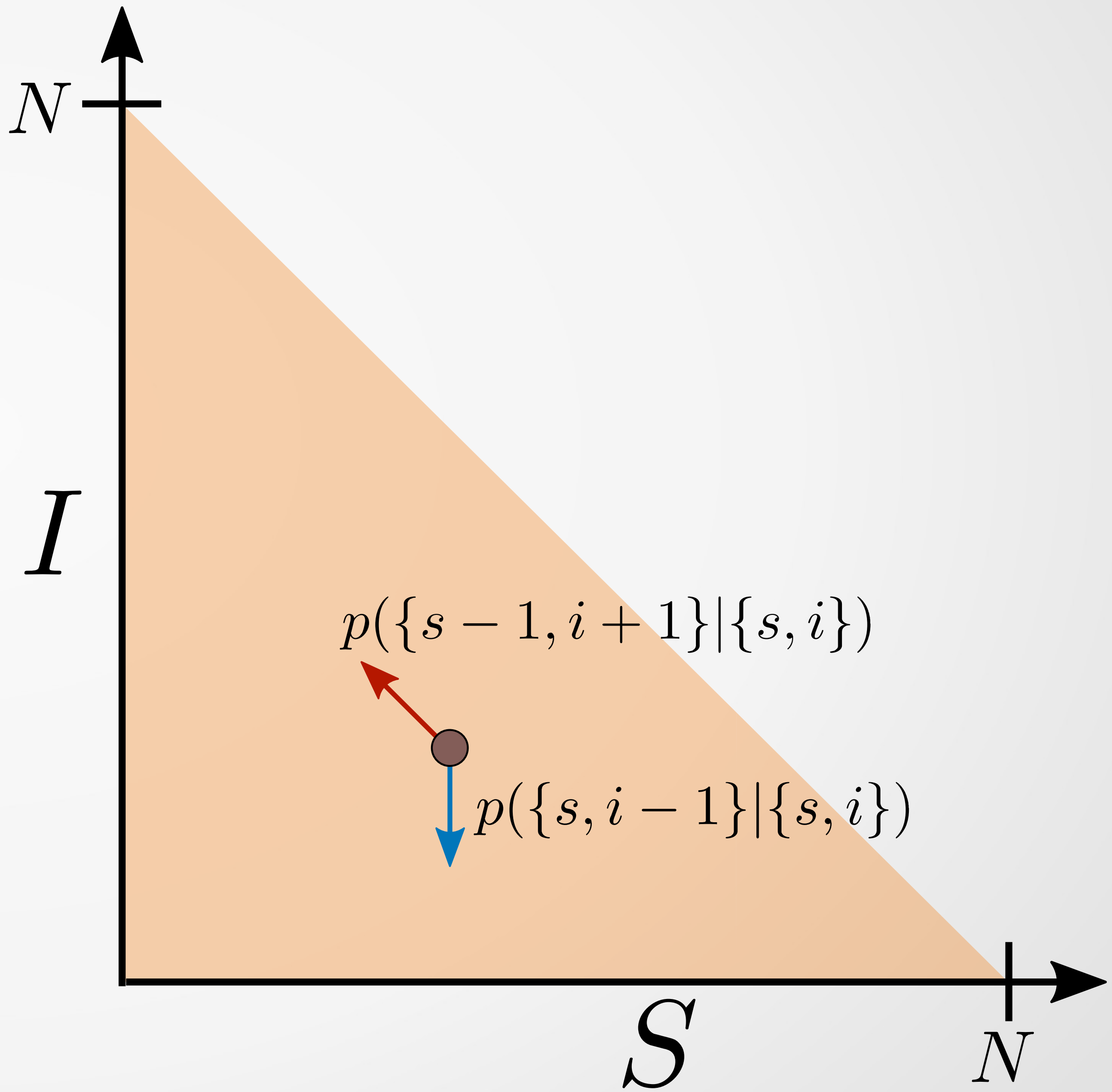
- 2-D space of (S, I)



Distribution of total size

using an SIR Markov chain technique

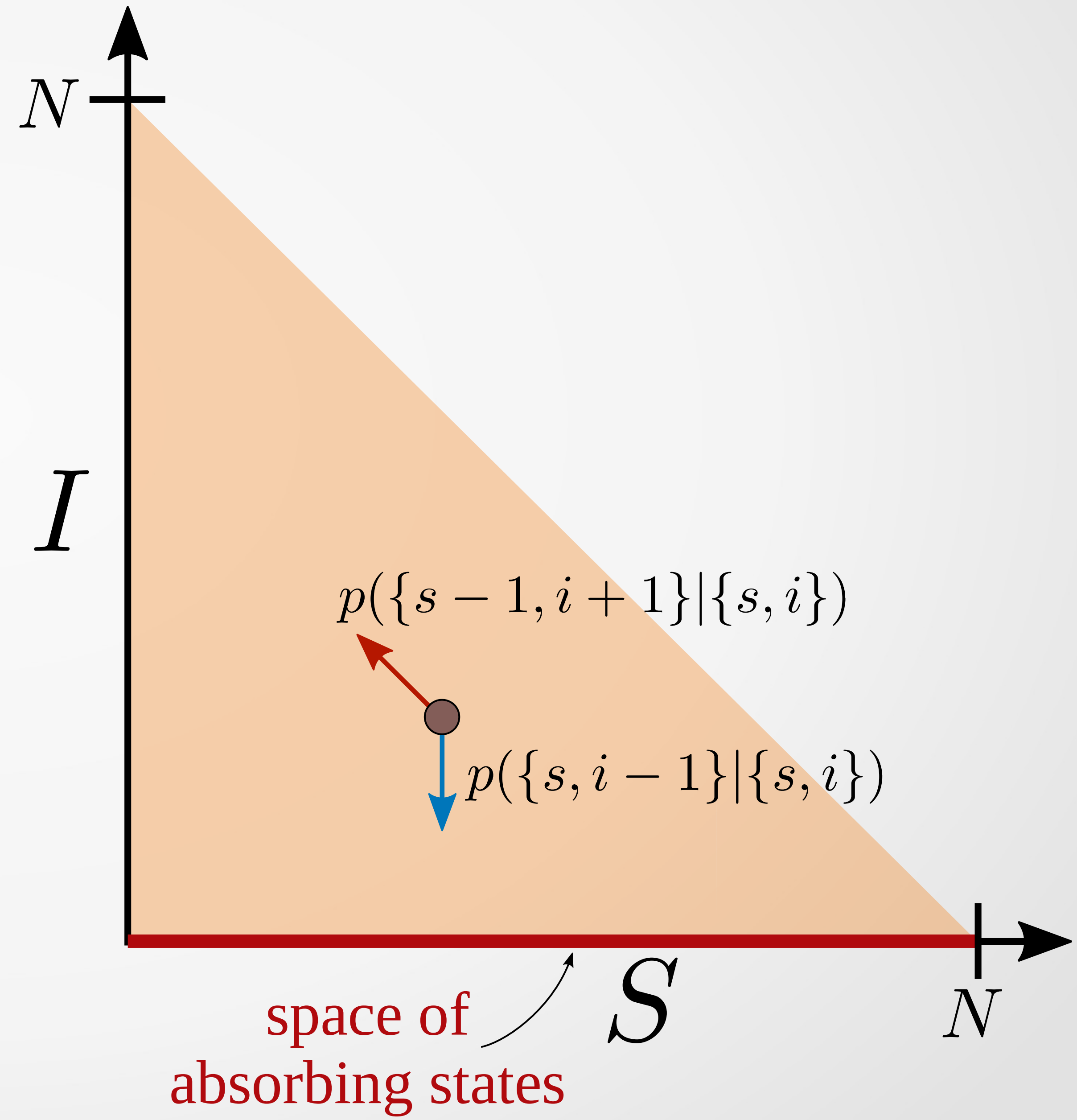
- 2-D space of (S, I)
- From $(S(t) = s, I(t) = i)$:
  - New infection → $(s\text{-}1, i\text{+}1)$
  - New recovery → $(s, i\text{-}1)$



Distribution of total size

using an SIR Markov chain technique

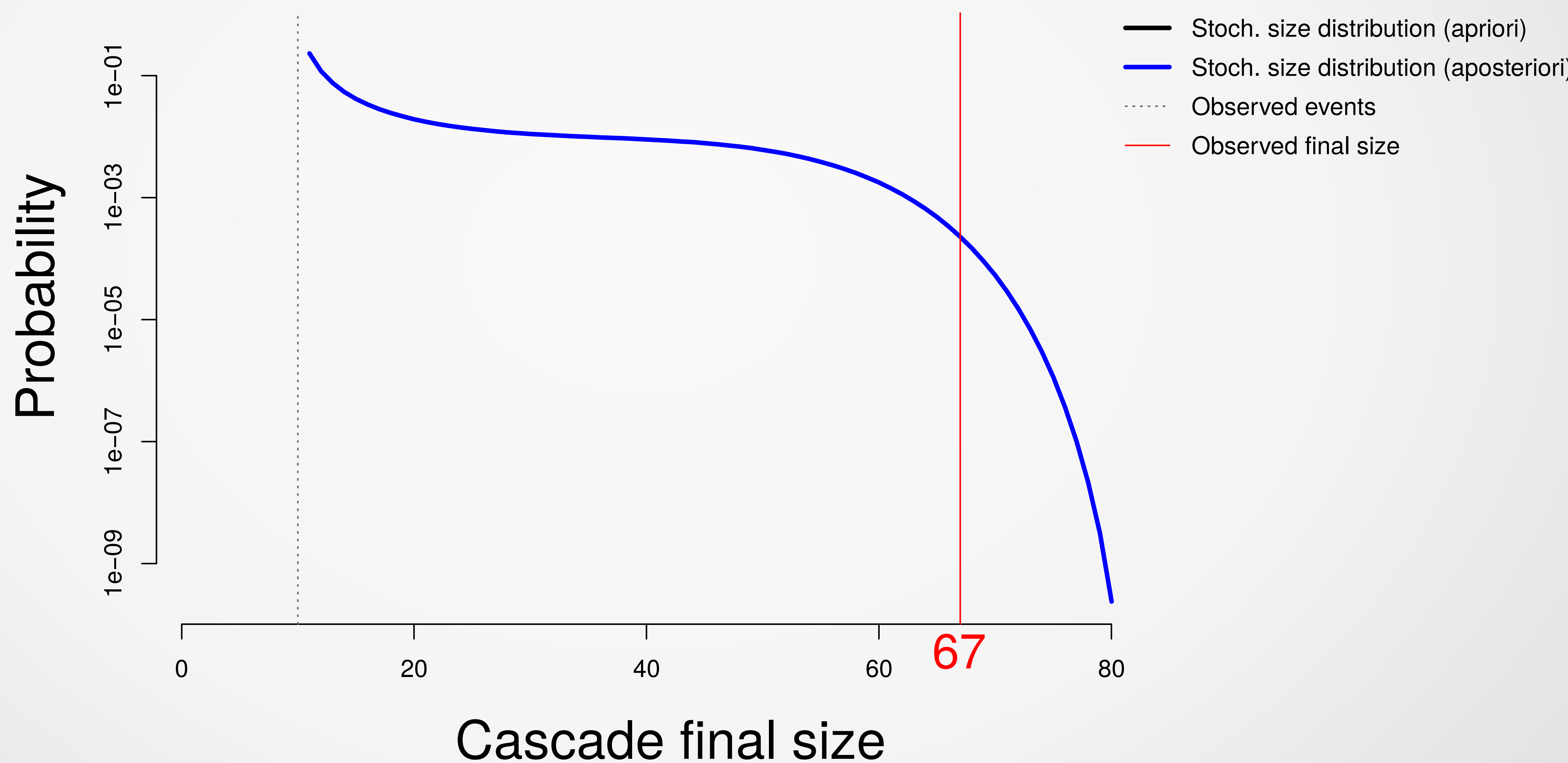
- 2-D space of (S, I)
- From ($S(t) = s$, $I(t) = i$):
  - New infection → (s-1, i+1)
  - New recovery→ (s, i-1)
- States (s, 0) are absorbing
- Probability of total size is the probability of N-s



Example: a tweet cascade

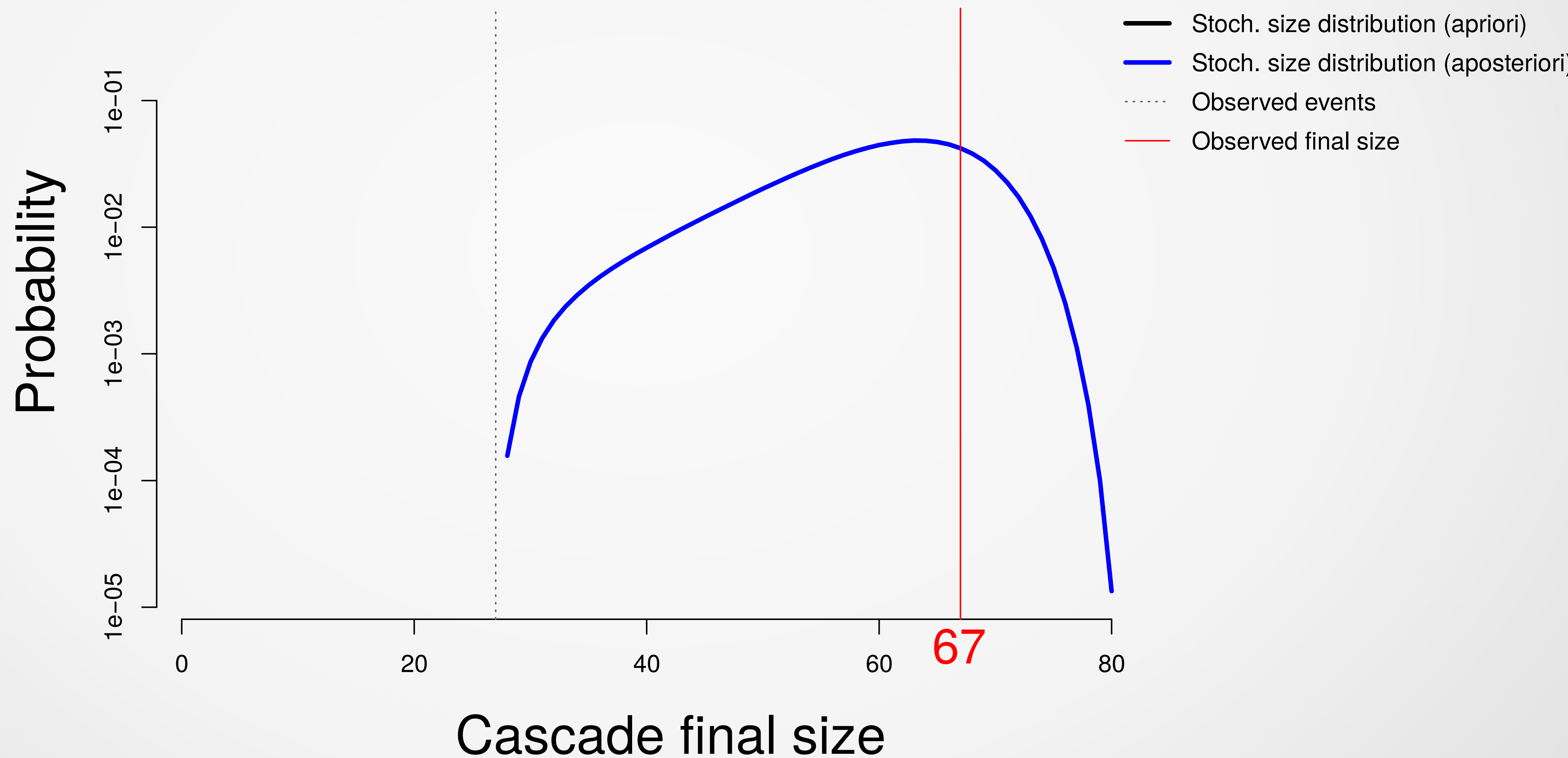
ScreenCrush @screencrushnews

The New York Times reports Leonard Nimoy, 'Star Trek''s beloved Mr. Spock, has died. nytimes.com/2015/02/27/art …



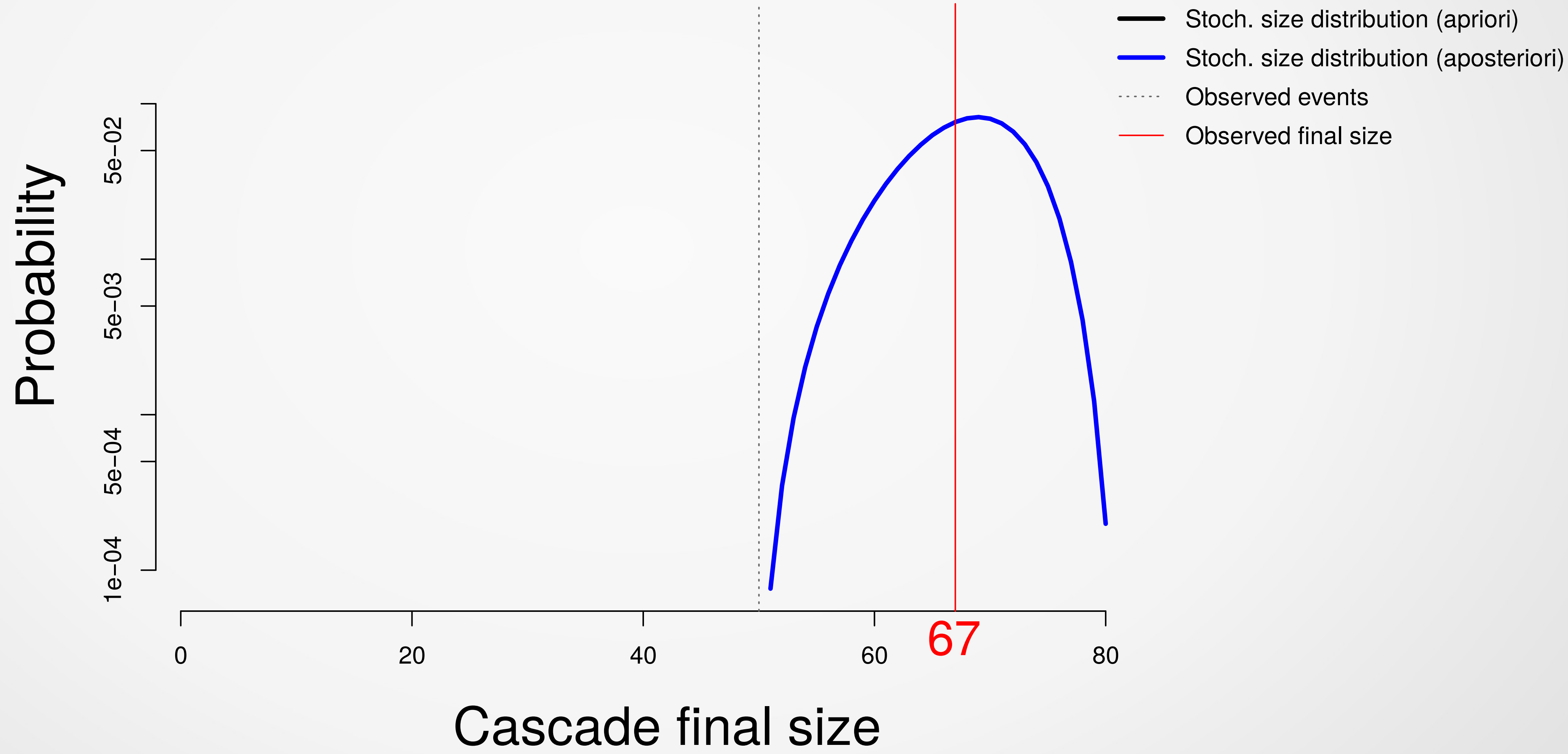
10 seen events

Example: a tweet cascade



27 seen events

Example: a tweet cascade

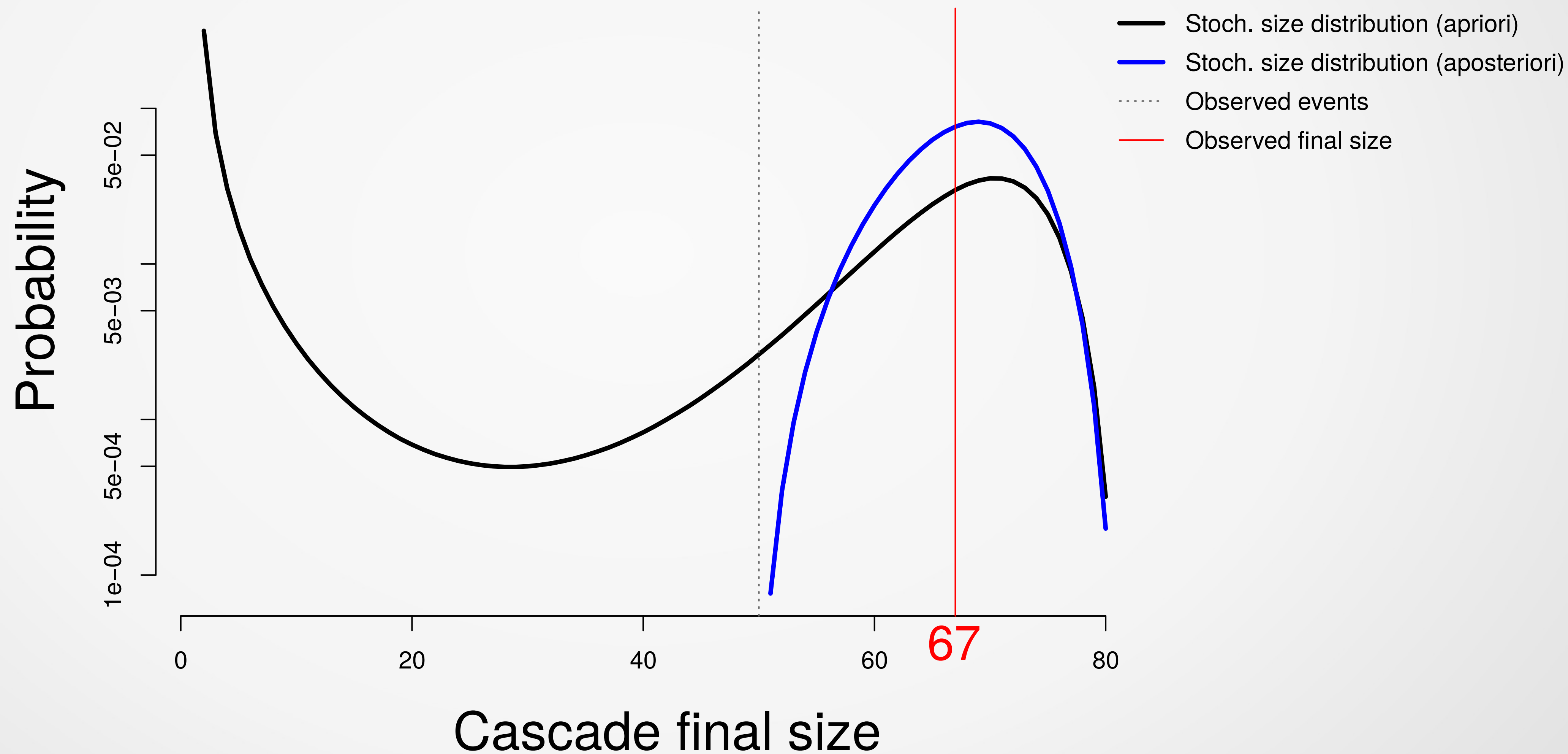


50 seen events

Example: a tweet cascade

ScreenCrush
@screencrushnews

The New York Times reports Leonard Nimoy, 'Star Trek''s beloved Mr. Spock, has died. nytimes.com/2015/02/27/art …

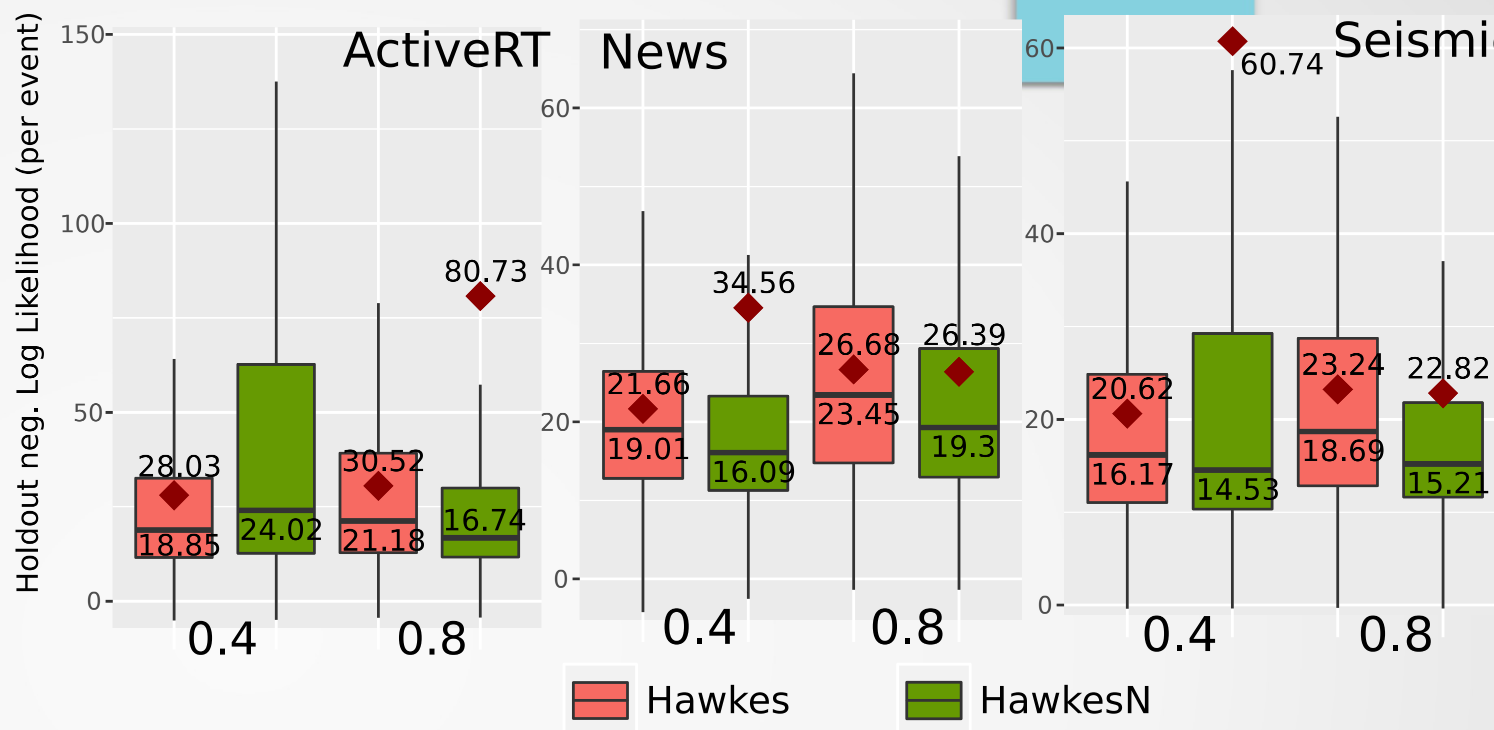


50 seen events

Explanation for the unpredictability of online popularity

HawkesN generalization

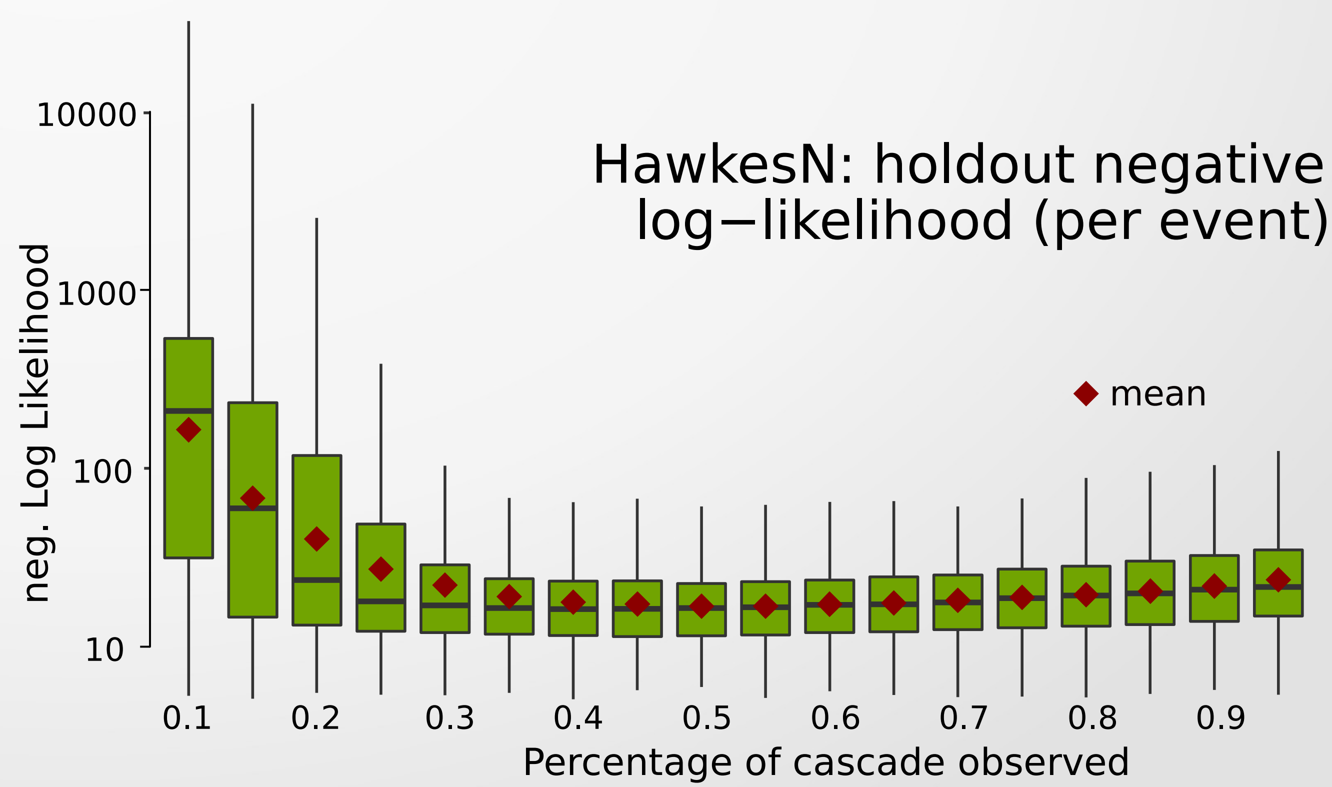
HawkesN generalizes better than Hawkes on real-life cascades



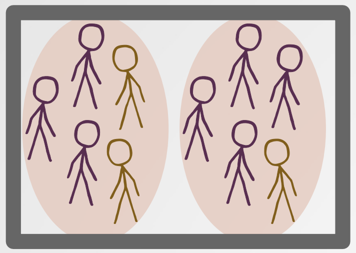
Caveat:

Estimating N from data is unreliable.

New statistic for diagnostic.

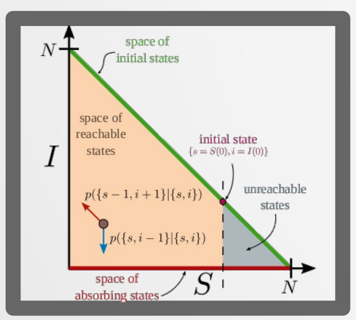


Summary



HawkesN: an extension of Hawkes accounting for a finite population

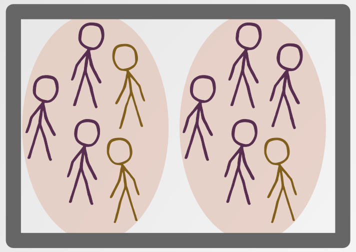
Connecting SIR epidemic models and HawkesN through the expected new infection intensity



A Markov Chain tool for computing the distribution of final size adapted to HawkesN

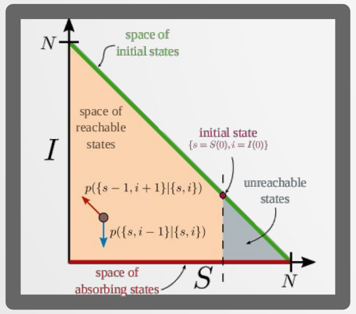
Limitations & future work: Fixed population, N estimated from each cascade, other kernels in HawkesN.

Thank you!



HawkesN: an extension of Hawkes accounting for a finite population

Connecting SIR epidemic models and HawkesN through the expected new infection intensity



A Markov Chain tool for computing the distribution of final size adapted to HawkesN

Limitations & future work: Fixed population, N estimated from each cascade, other kernels in HawkesN.

Data & code: https://github.com/computationalmedia/sir-hawkes

Supp: Estimating *I(0)* in HawkesN

Issue:

Recovery events are unobserved in HawkesN → the number of infected is unknown.

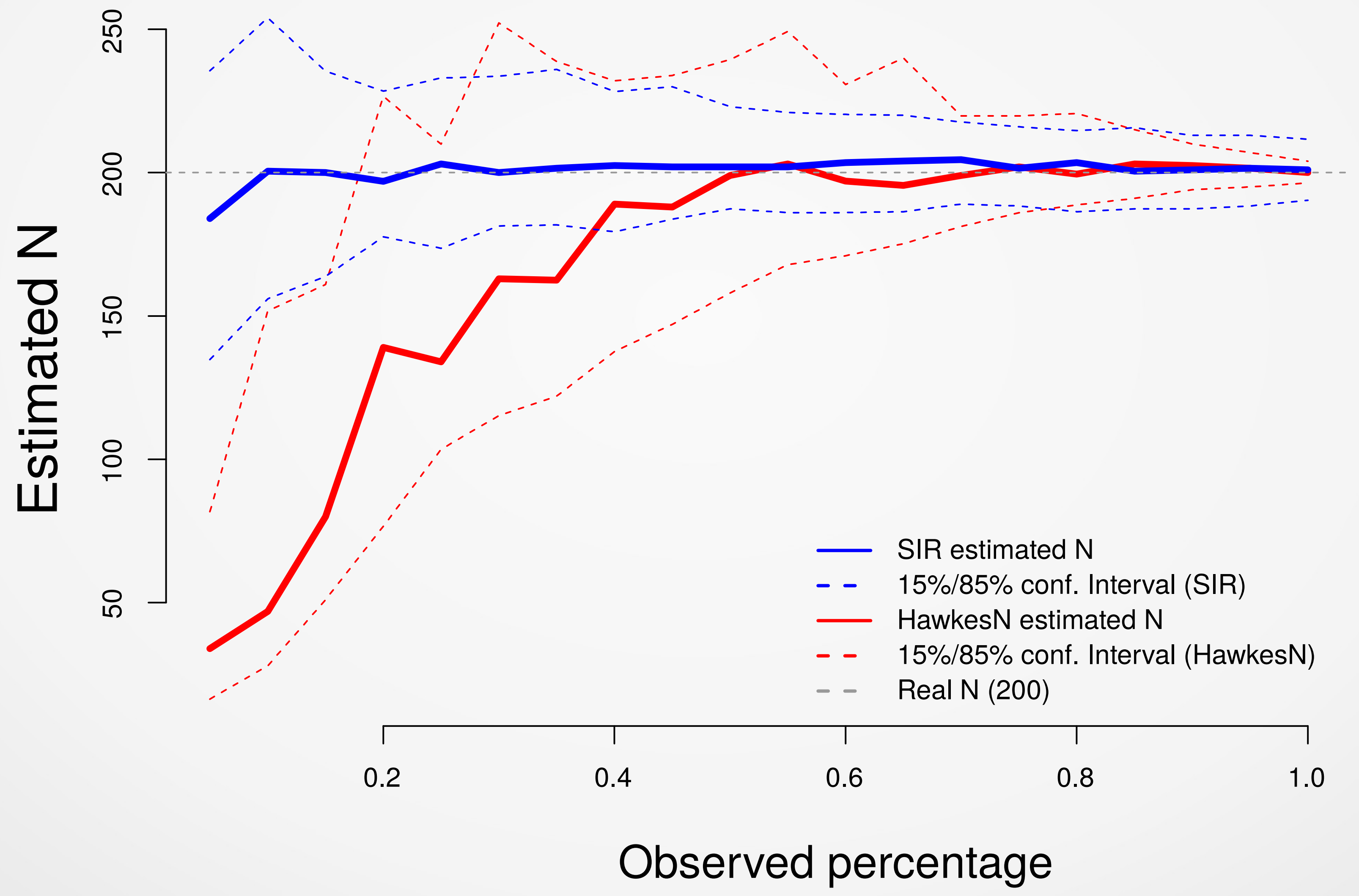
Solution:

Estimate its expected value

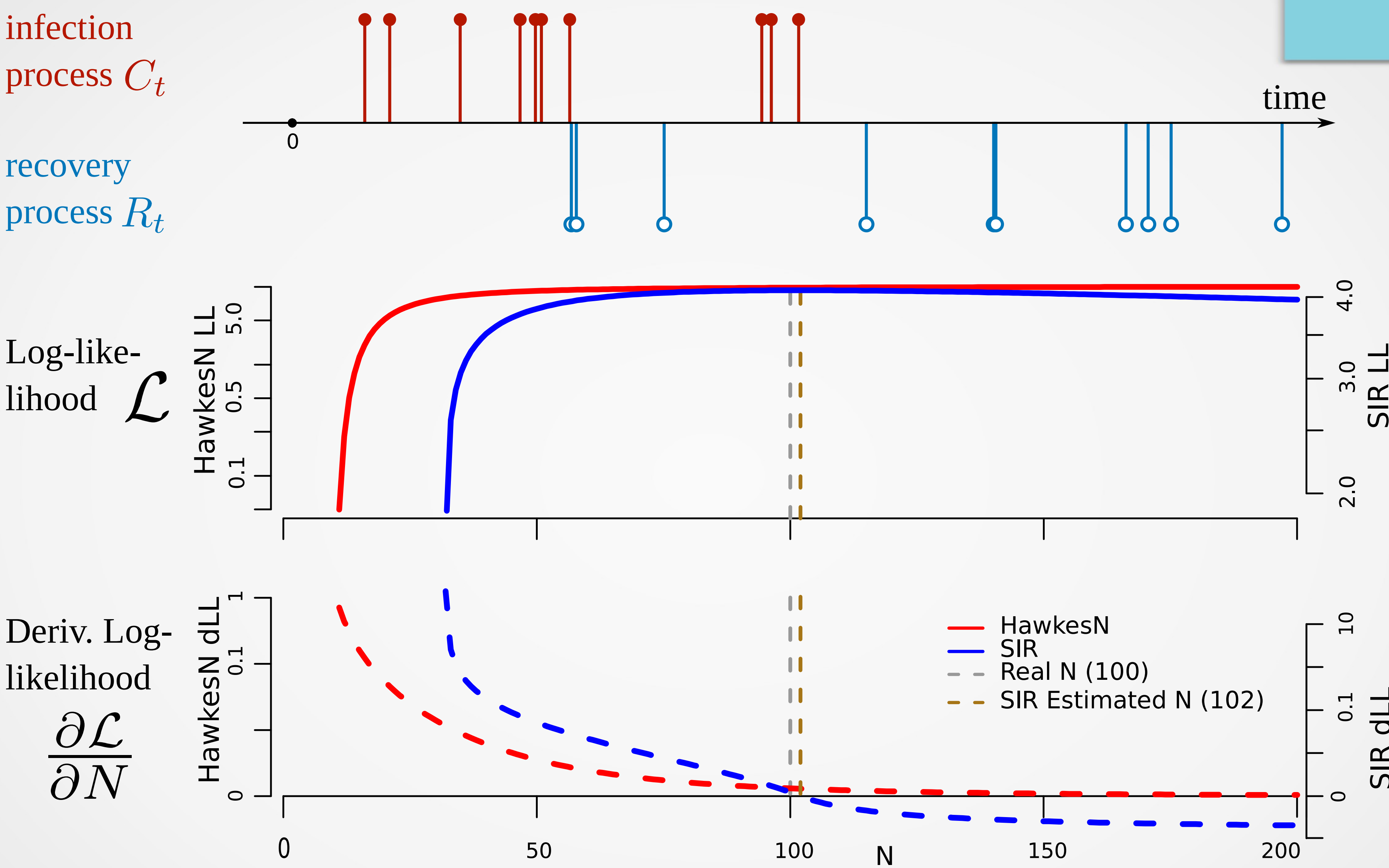
$$\mathbb{E}_{t^R}[I(0)] = \mathbb{E}_{t^R}\left[\sum_{j=1}^{l} \mathbb{1}(t_j^R > t_l)\right] = \sum_{j=1}^{l} e^{-\gamma(t_l - t_j^I)}$$

when $t_1, t_2, \ldots, t_l$ are the *l* observed events.

Supp: (under) Estimating N from data



Supp: Estimating N from data



$$\frac{\partial \mathcal{L}}{\partial N} \geq \frac{1}{N^2} \underbrace{\left(\frac{(n-1)n}{2} - \sum_{j=0}^{n-1} \sum_{l=j}^{n-1} l\kappa \left[e^{-\theta(t_l - t_j)} - e^{-\theta(t_{l+1} - t_j)} \right] \right)}_{\mathcal{S}(\kappa, \theta, \{t_1, t_2, \ldots, t_n\})}$$