FITTING A CODE-RED VIRUS SPREAD MODEL

A. Kolesnichenko, B.R. Haverkort, P.-T. de Boer
(Univ. of Twente, The Netherlands)
A. Remke (Univ. Münster, Germany)

Motivation

- Increased interest in virus spread models
- Need for realistic parameters

- Mean-field models
- Large number of interacting similar objects
- No assumption w.r.t. topology
- Model the spreading phase of a computer virus

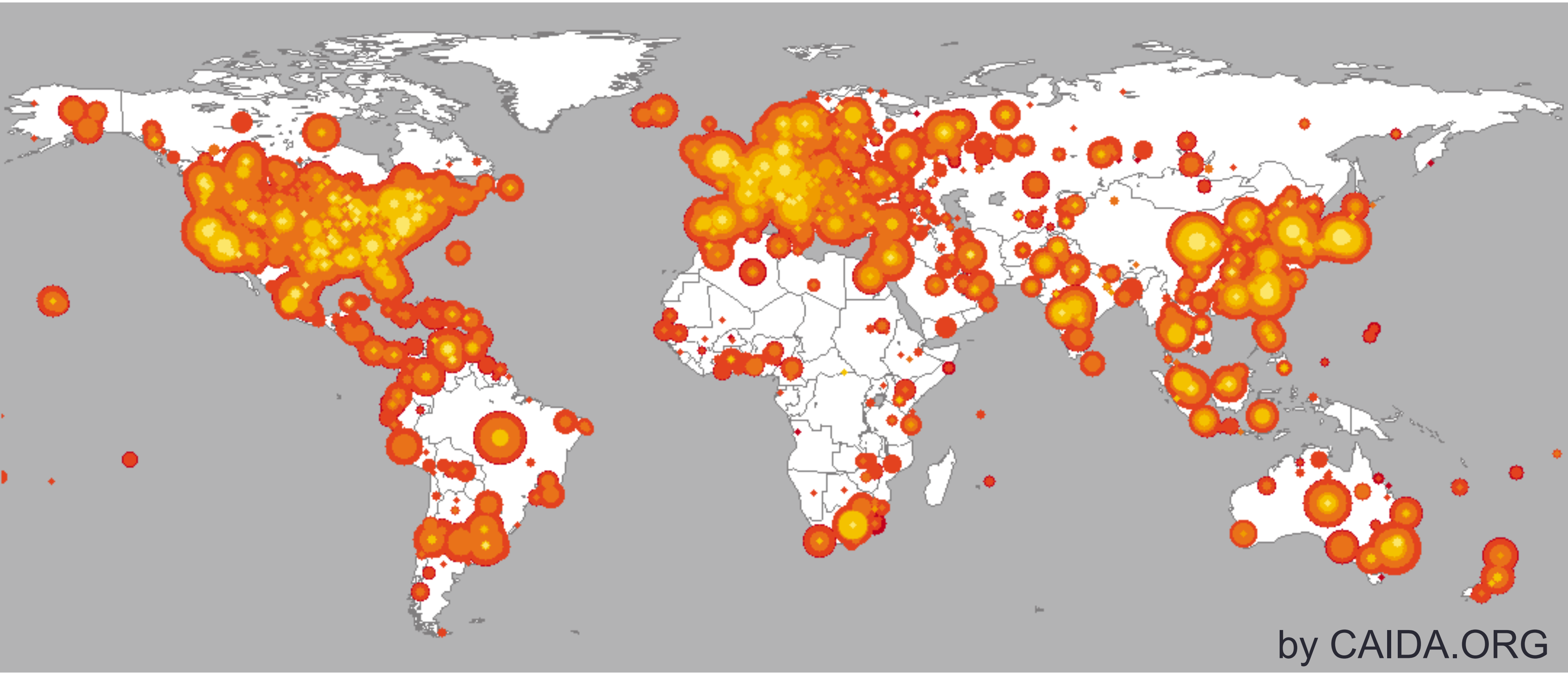
- Illustrate the fitting procedure on the case of Code Red
- An account of putting theory into practice

Code Red

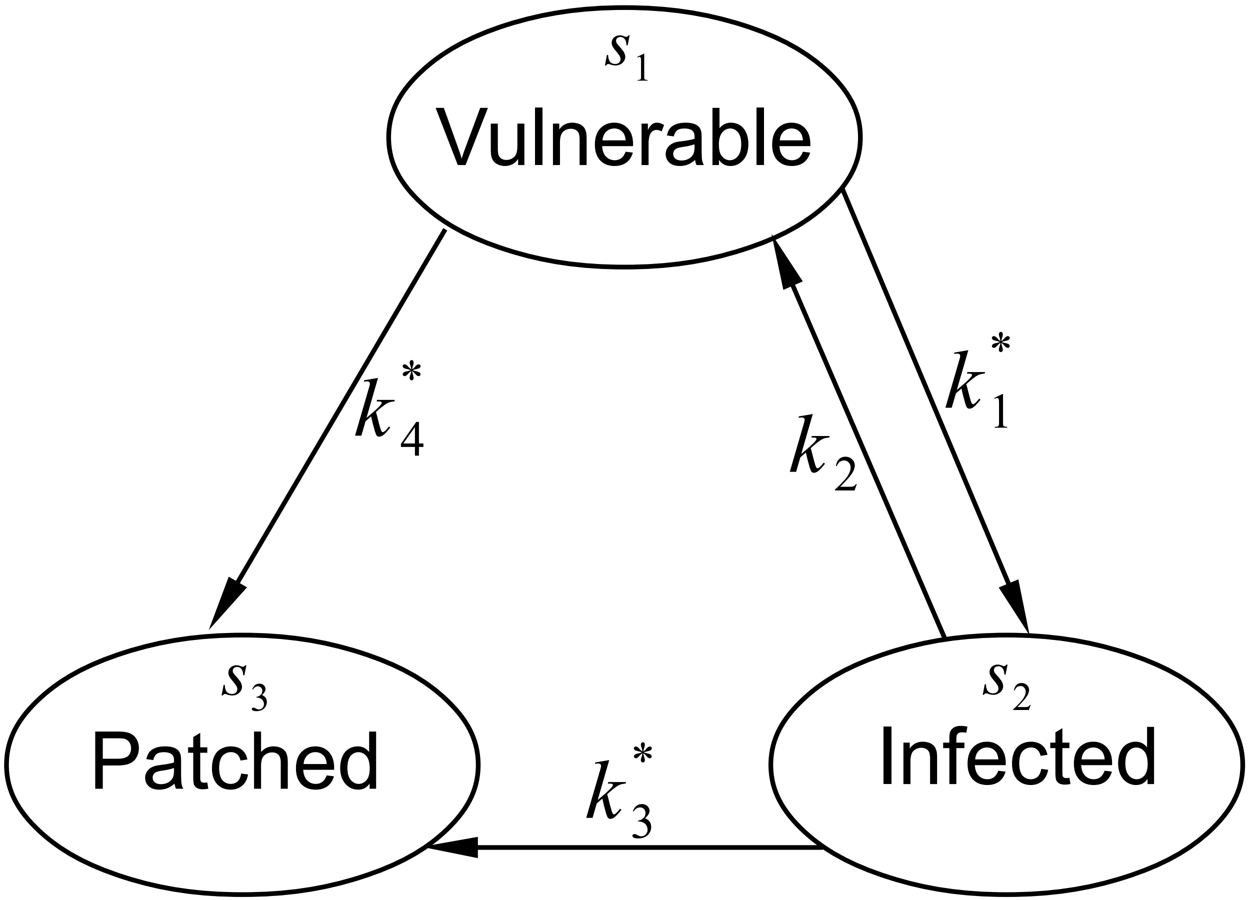
- buffer-overflow vulnerability in Microsoft's IIS web servers
  - Information on vulnerability released June 18, 2001
  - Patch released June 26, 2001
  - July 12, 2001 Code Red version 1 (CRv1) started spreading
  - July 19, 2001 Code Red version 2 (CRv2) spreading 10:00 UTC
  - August 1, 2001 CRv2 started spreading again

- Spreading phase (between 1st and 19th of each month)
  - Generates random list of IP addresses trying to connect to TCP port 80
- Attacking phase (between 20th and 28th of each month)
  - Starts DoS attack to www.whitehouse.gov through fixed IP address

Spread of the Code Red worm on July 19



A first spreading model



Mean-field model

- For a network of N nodes
- State space of fractions $\overline{m} = (m_1, m_2, m_3)$
- Transition probabilities

$$k_1^*(t) = k_1 \cdot m_2(t), \quad k_3^*(t) = k_3 \cdot m_2(t), \quad k_4^*(t) = k_4 \cdot m_2(t),$$

- Transient behaviour given by ODEs

$$\begin{cases} \dot{m}_1(t) &= k_2 m_2(t) - k_1 m_2(t) m_1(t) - k_4 m_1(t) m_2(t), \\ \dot{m}_2(t) &= k_1 m_2(t) m_1(t) - k_2 m_2(t) - k_3 m_2(t) m_2(t), \\ \dot{m}_3(t) &= k_4 m_1(t) m_2(t) + k_3 m_2(t) m_2(t), \end{cases}$$

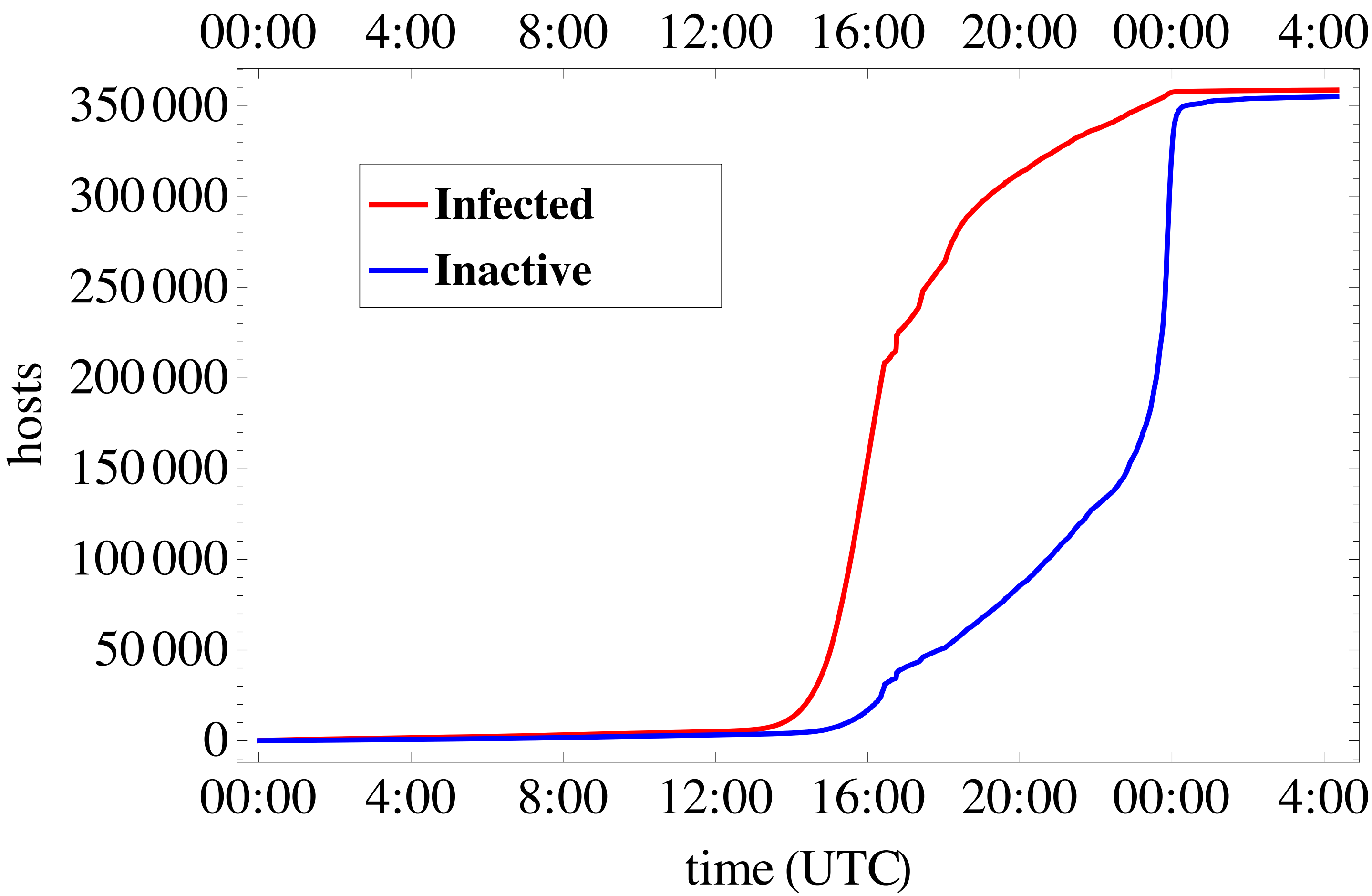
- Actual numbers $M_1(t)$, $M_2(t)$, $M_3(t)$ result from multiplying with N

How to obtain parameter values?

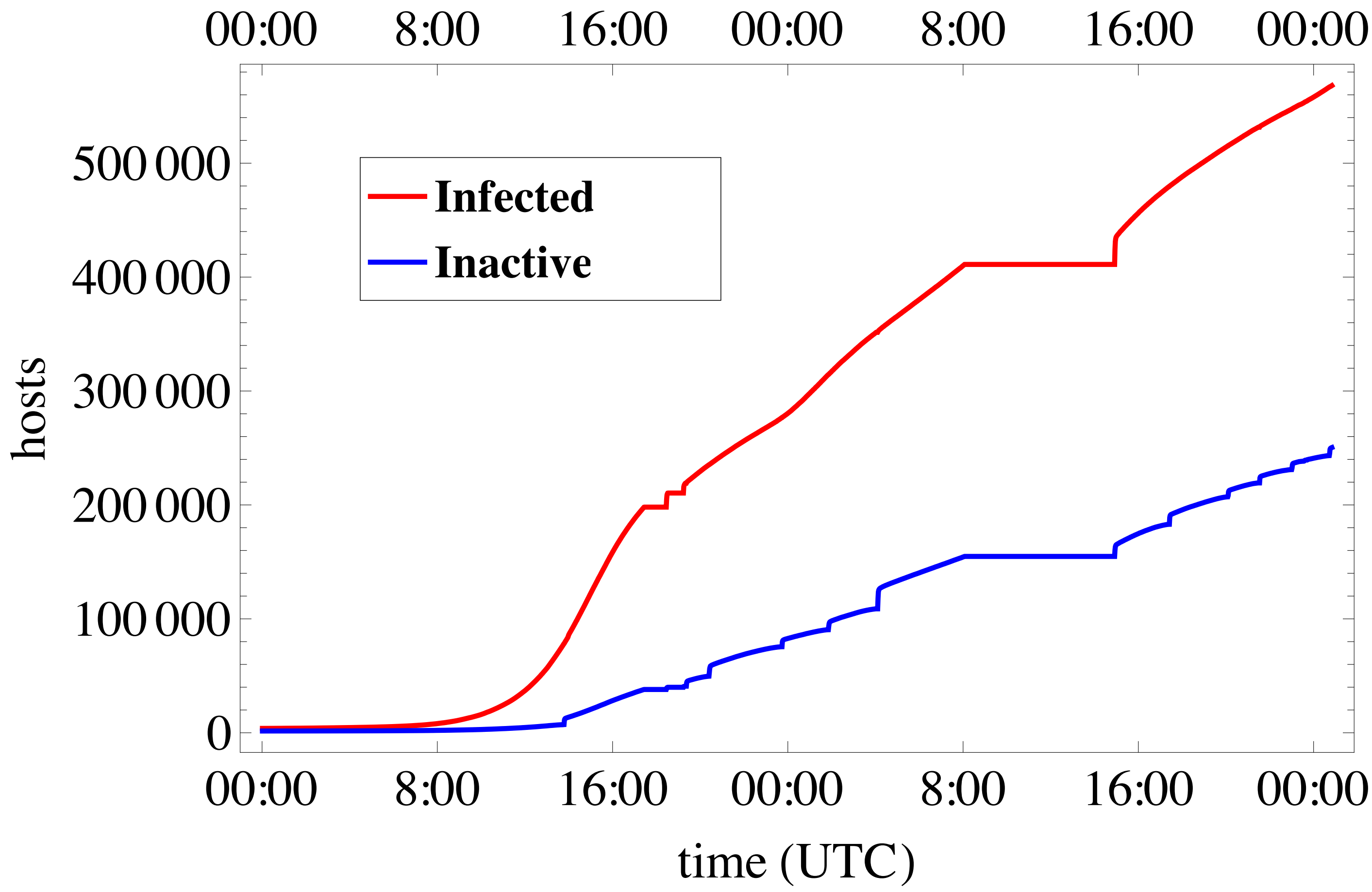
Data set by CAIDA

- Data is based on combined measurements
  - /8 Telescope network at UCSD until 16:30
  - Sampled netflow data from a router upstream after 16:30
  - Data from two /16 networks at Lawrence Berkeley Labartory
- Two traces from this data have been used
  - Number of new unique infected hosts
  - Number of hosts that have stopped being infected

Measurement Data July (total number of infected hosts)



Measurement Data August (total number of infected hosts)

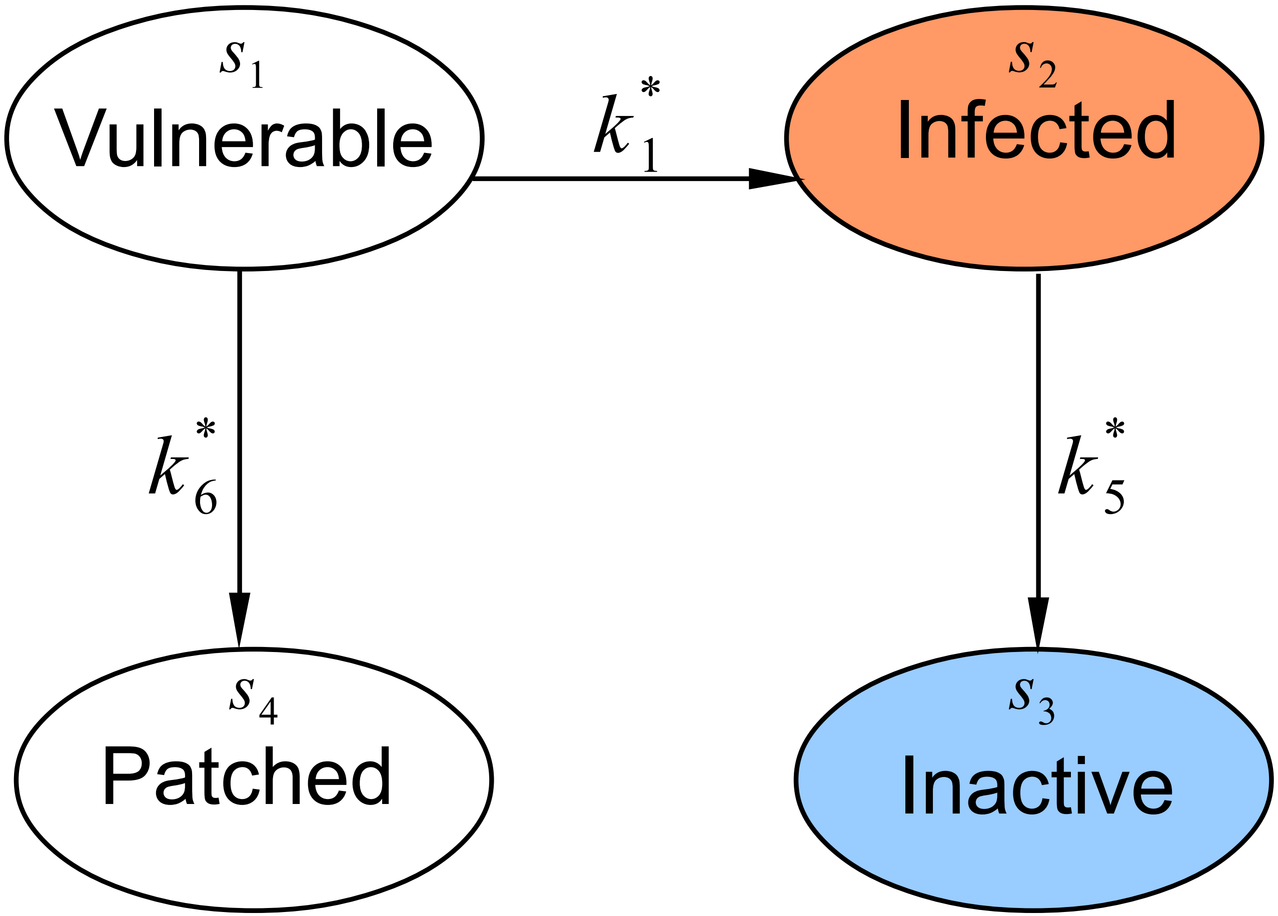


Challenges when working with old data

- Why does spreading slow down before midnight?
- Why does rate of increase decline?
  - Overloaded networks due to worm
  - Unavailability of vulnerable hosts
  - Many infected machines were office desktops

- Need to adapt the model to match the available data
  - Rebooting of infected hosts not measured in dataset
    - Distinguish between vulnerable and inactive hosts
  - Split patched hosts into two groups
    - Hosts which became inactive after being infected
    - Hosts which were never infected before getting patched

Model reconsideration



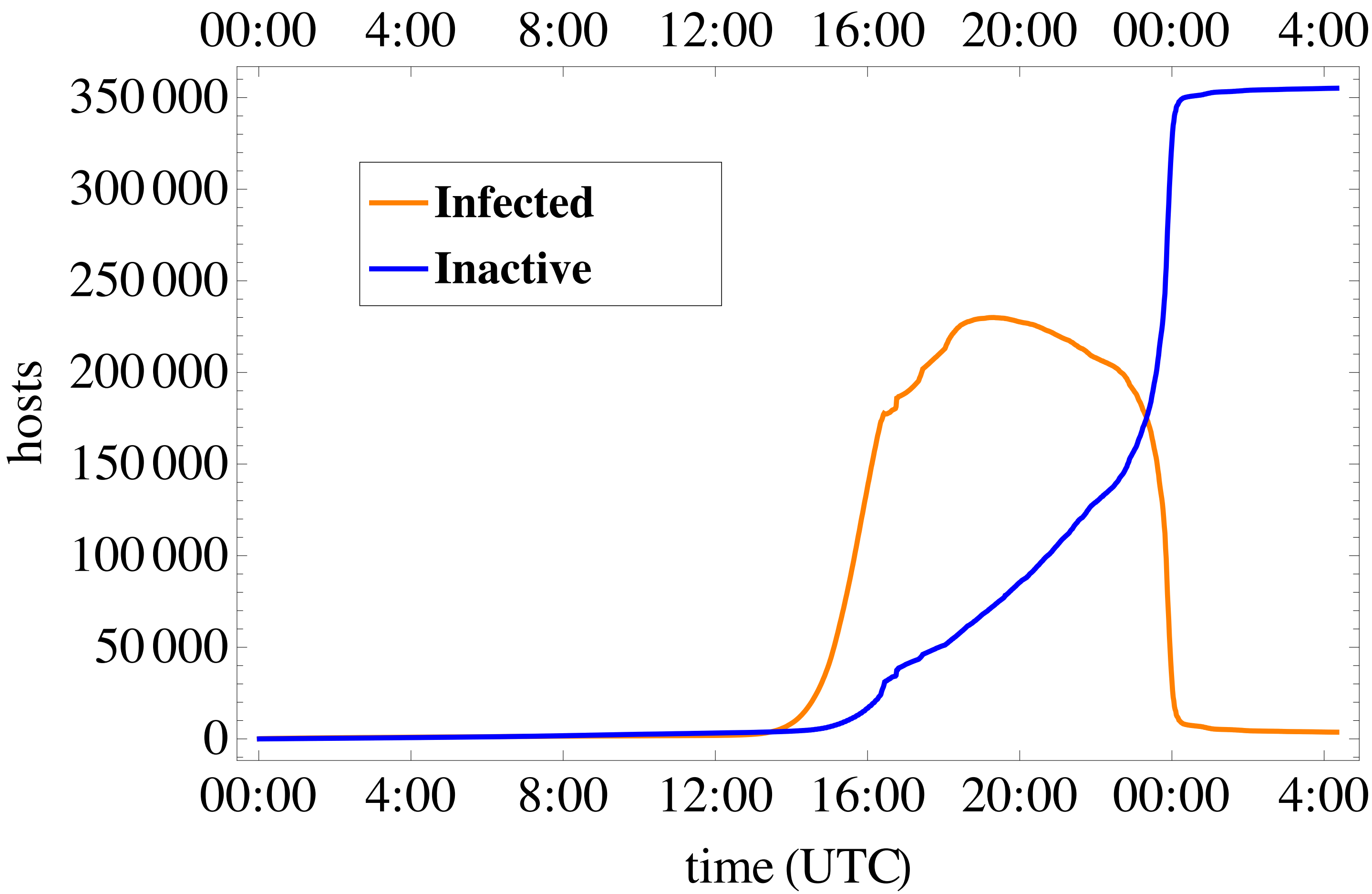
Model details

- A *vulnerable* machine becomes infected $k_1^*(t) = k_1 \cdot m_2(t)$
- *Infected* machines are patched $k_5^*(t) = k_5 \cdot m_2(t)$
- *Vulnerable* machines are patched $k_6^*(t) = k_6 \cdot m_2(t)$

- Dynamics are given by

$$\begin{cases} \dot{m}_1(t) &= -k_1 \cdot m_2(t) \cdot m_1(t) - k_6 \cdot m_1(t) \cdot m_2(t), \\ \dot{m}_2(t) &= k_1 \cdot m_2(t) \cdot m_1(t) - k_5 \cdot m_2(t) \cdot m_2(t), \\ \dot{m}_3(t) &= k_5 \cdot m_2(t) \cdot m_2(t), \\ \dot{m}_4(t) &= k_6 \cdot m_1(t) \cdot m_2(t), \end{cases}$$

Number of hosts still infected (July)



Parameter Fitting

- Minimize the relative squared error

$$\mathcal{E}_{rel} = \frac{\sum_{r=1}^{R} ||\mathcal{O}(t_r) - m(t_r)||^2}{\sum_{r=1}^{R} ||\mathcal{O}(t_r) - \overline{\mathcal{O}}||^2},$$

- Which is in our case equivalent to least squared error and the maximum liklelihood methods

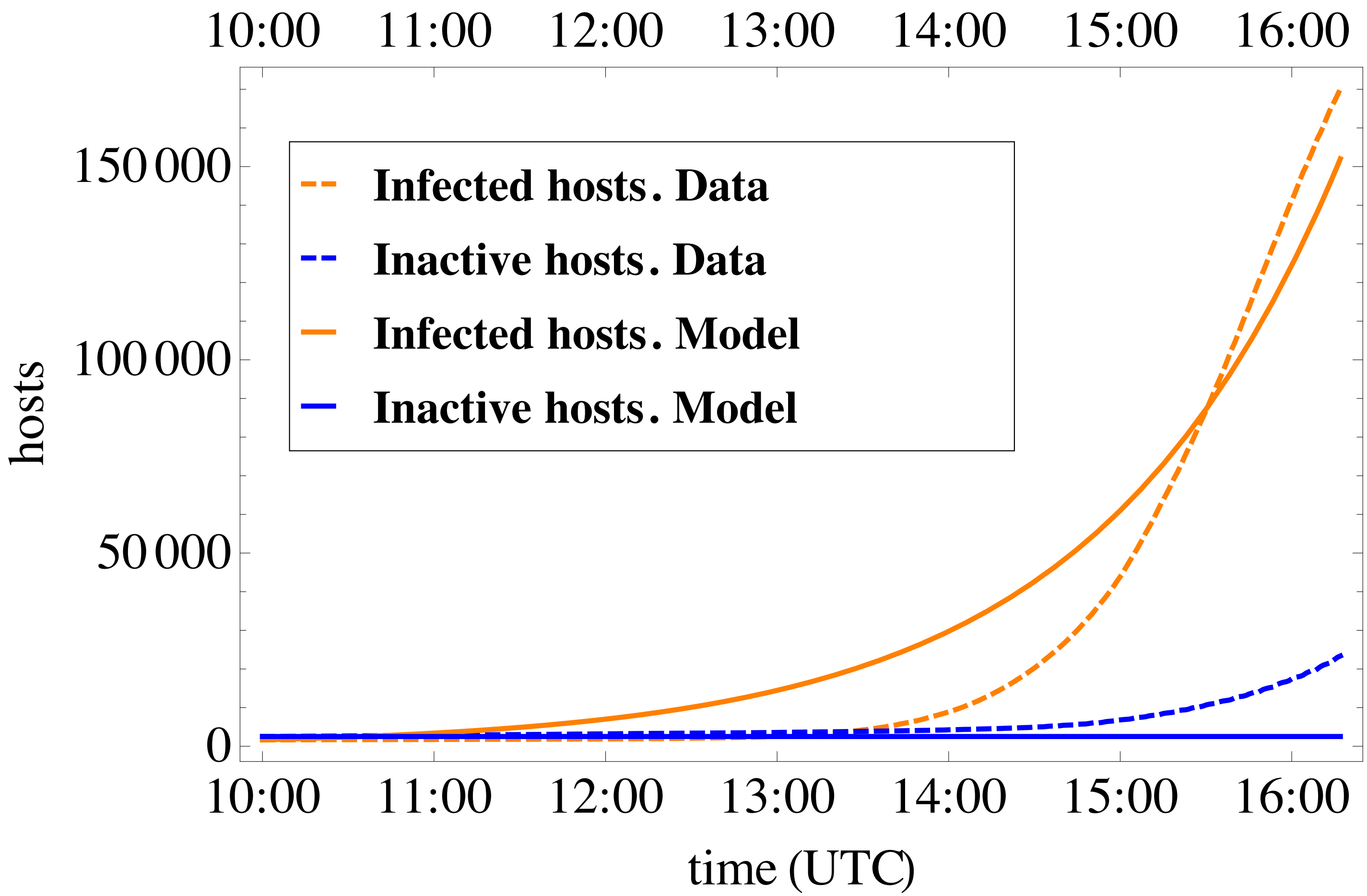
Set the inital conditions

- According to literature: CRv2 infected between 1 and 2 million out of a potential 6 million hosts

- $M_1(0)=(6H)$ (vulnerable hosts)
- $M_4(0)=(0)$ (patched nodes)
- No data available to fit against

- Number of infected and inactive hosts obtained from measurement data at 10:00 UTC
- $M_2(0) = 4181$
- $M_3(0) = 2528$

Fit for CRV 2 Outbreak in July 2011



Improving the inital conditions

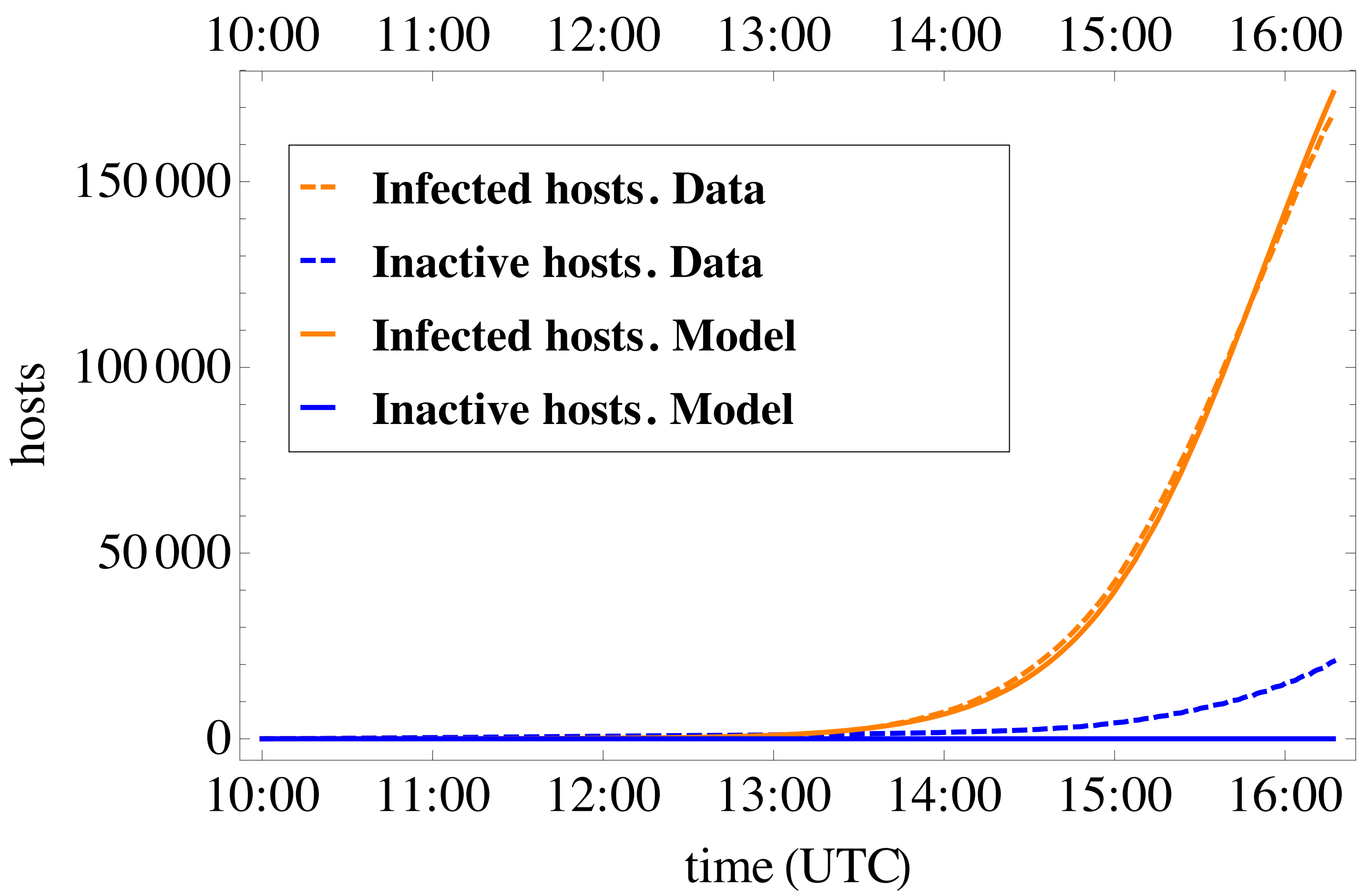
- Relative squared error of approx. 10%

- Spead of virus propagation is overestimated
- Number of initially infected hosts is too big

- Activity of CRv1 and other background unsolicited SYN probes already registerd before CRv2 started to spread.
- Substract all infections that took place before 10:00 UTC

- New initial conditions $M(0) = (6H - 3; 3; 0; 0)$

Improved fit of July data

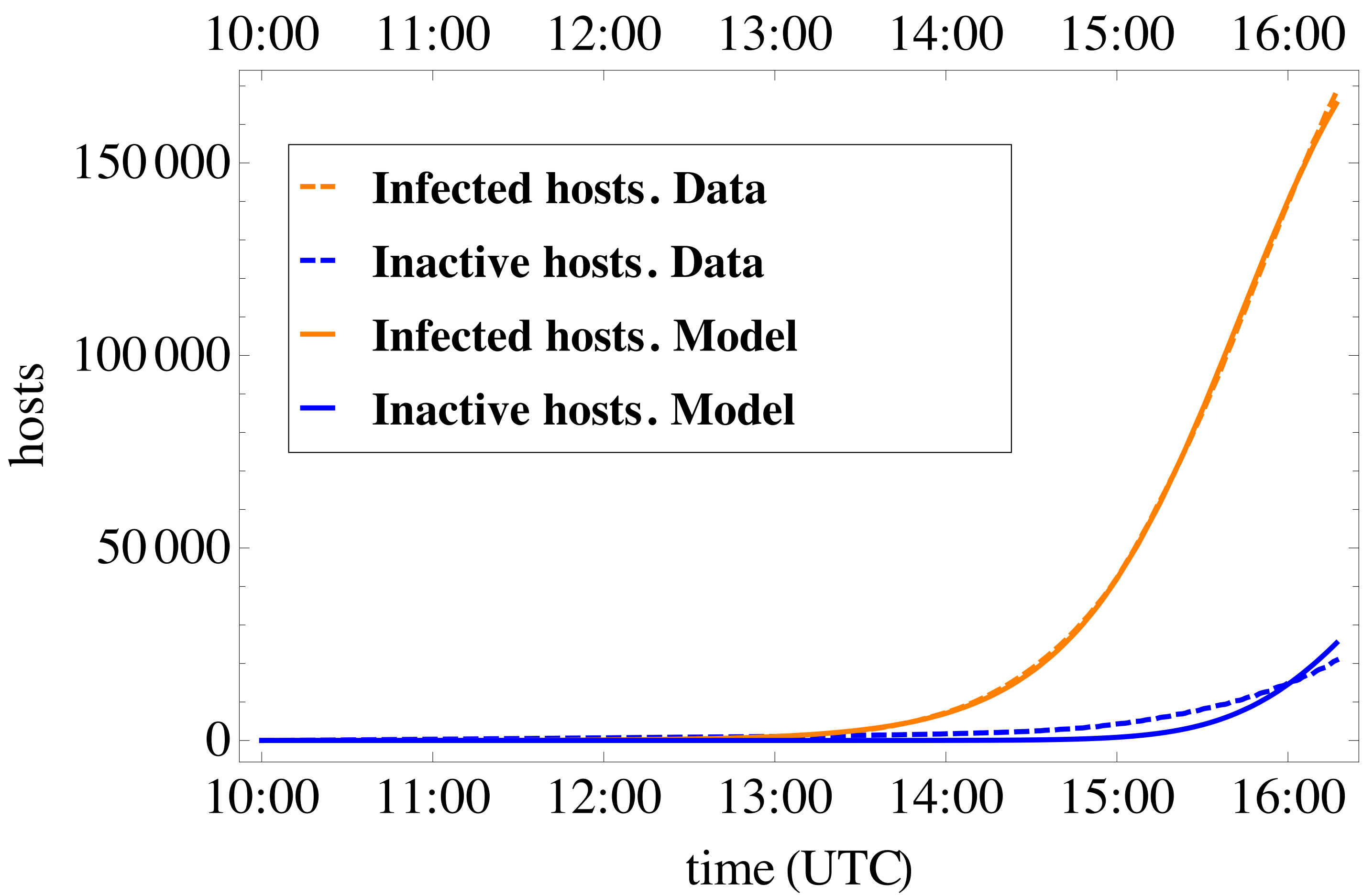


Initial conditions ctd.

- Relative squared error of 1.6%
- Mostly due to number of inactive hosts
- Difficult to model since it includes human behaviour

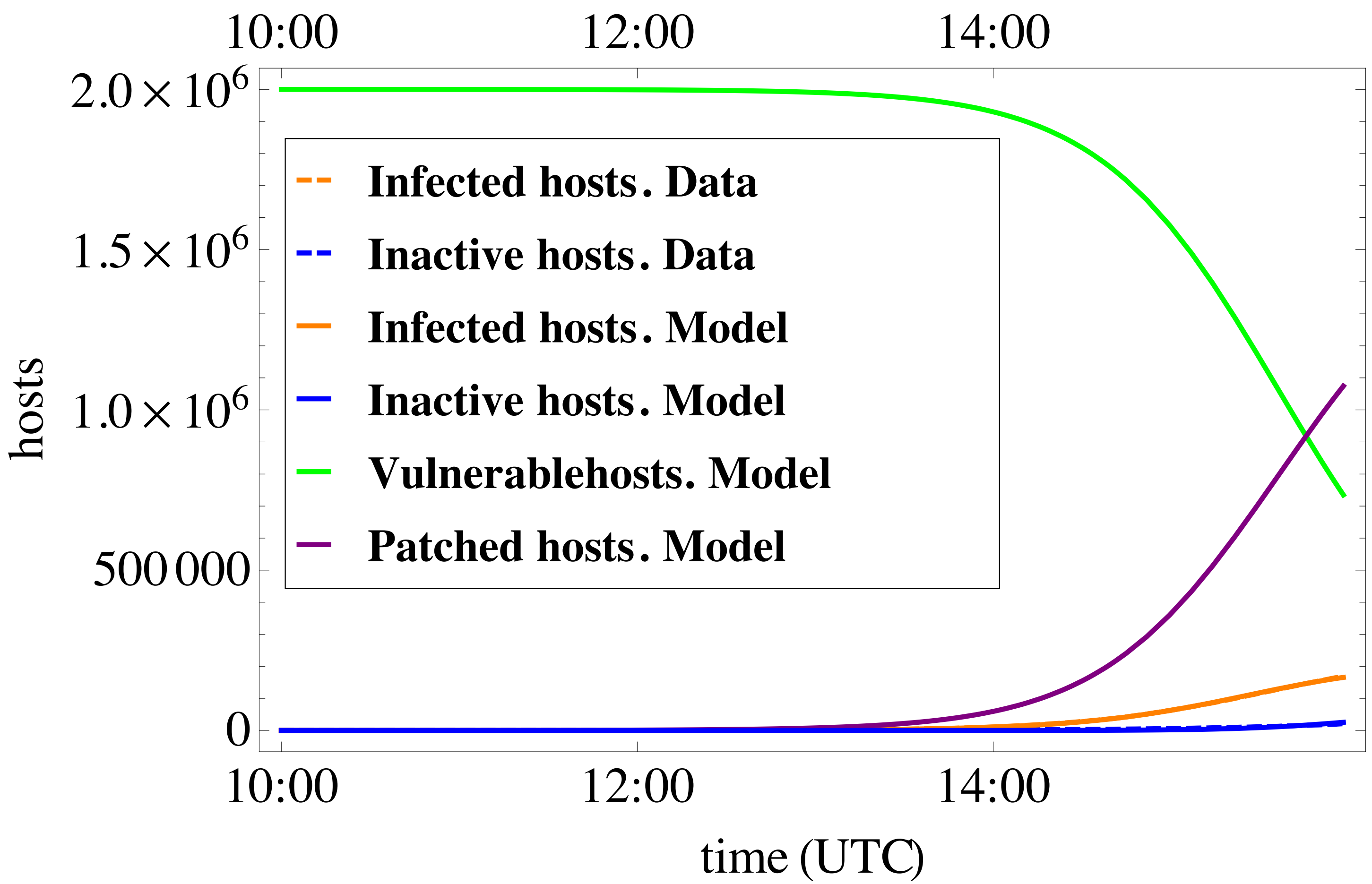
- Another uncertainty: number of initially vulnerable hosts
- 60 experiments, with populations from 500.000 to 6H
- Results in a smallest relative error of 0.2 for $M_1(0) \leq 2H$

M(0) = (2H - 3; 3; 0; 0)



10:00
11:00
12:00
13:00
14:00
15:00
16:00
150 000
100 000
50 000
0
hosts
Infected hosts. Data
Inactive hosts. Data
Infected hosts. Model
Inactive hosts. Model
time (UTC)

M(0) = (2H - 3; 3; 0; 0)



10:00
12:00
14:00
2.0 × 10^6
1.5 × 10^6
1.0 × 10^6
500 000
0
hosts
Infected hosts. Data
Inactive hosts. Data
Infected hosts. Model
Inactive hosts. Model
Vulnerablehosts. Model
Patched hosts. Model
10:00
12:00
14:00
time (UTC)

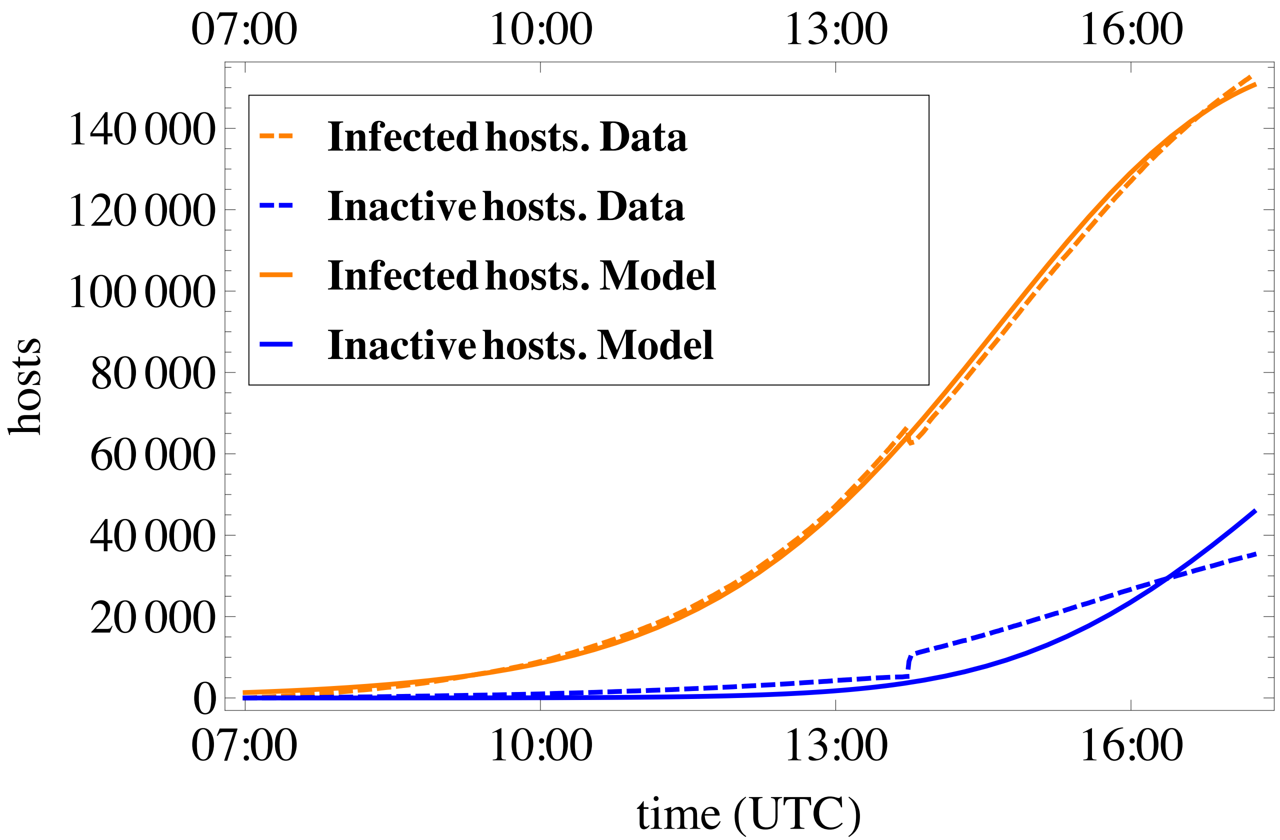
How to fit August data?

- Again difficult to find good initial values
  - All CRv2 activity before 00:00 UTC has to be taken into account
  - Other background activity should be substracted

- Use $M_1(0) = 1.5\ H – M_2(0)$
  - Only a limited number of hosts was patched during July outbreak
- Add $M_2(0)$ as extra parameters to fitting procedure
  - Extra degrees of freedom can lead to a worse result
- Take $M_3(0) = 0$ and $M_4(0) = 0$
  - As any patching before midnight is not related to CRv2 spreading

- Minimizing relative squared error leads to 0.7% error

Fit August outbreak



Related work on Code Red

- Staniford presented epidemiological model for infected hosts and a manually made fit to data
- Zou et al. presented a two-factor worm model including
  - Human counter-measures
  - Slowing down due to impact on internet traffic

- We do not take into account data after 16:20 UTC
- Based our model on insight in actual operation
- Use well-known parameter estimation techniques

Conclusions

- Parametrizing a large-scale distributed system

- Need to change model to match data available for fitting
- Handle measurement data very carefully
  - Missing or incomplete measurement intervals
  - Available data only reflects part of the system

- Possible to find a model and a set of parameters that closely captures the first part of virus spreading
  - Do not know whether these are *ultimate correct parameters*

- Resulting squared error of 0.2% and 0.7% for July and August outbreaks, respectively