



# Worm Evolution Tracking via Timing Analysis

**Moheeb Abu Rajab** Fabian Monrose Andreas Terzis

Computer Science Department

Johns Hopkins University

# Outline

- Motivation and Background
- Problem Statement and Goals
- Analysis
- Evaluation

# Post-mortem analysis

- *Xie et al.* proposed a random walk algorithm on the hosts contact graph
  - Provides who infected whom tree
  - Requires extensive logging throughout the network
- Reverse Engineering Approaches [*Kumar et al.*]
  - Can reveal several interesting pertinent characteristics of the worm
  - Can be tedious and not easily generalizable



# Background:

## Network Monitors (Telescopes)

- Passive Monitors listening on portions of routable unused IP space.
- Used to make inferences about different global scale security events that exhibit a random behavior:
  - Random scanning worms
  - DDoS attacks employing random spoofed addresses

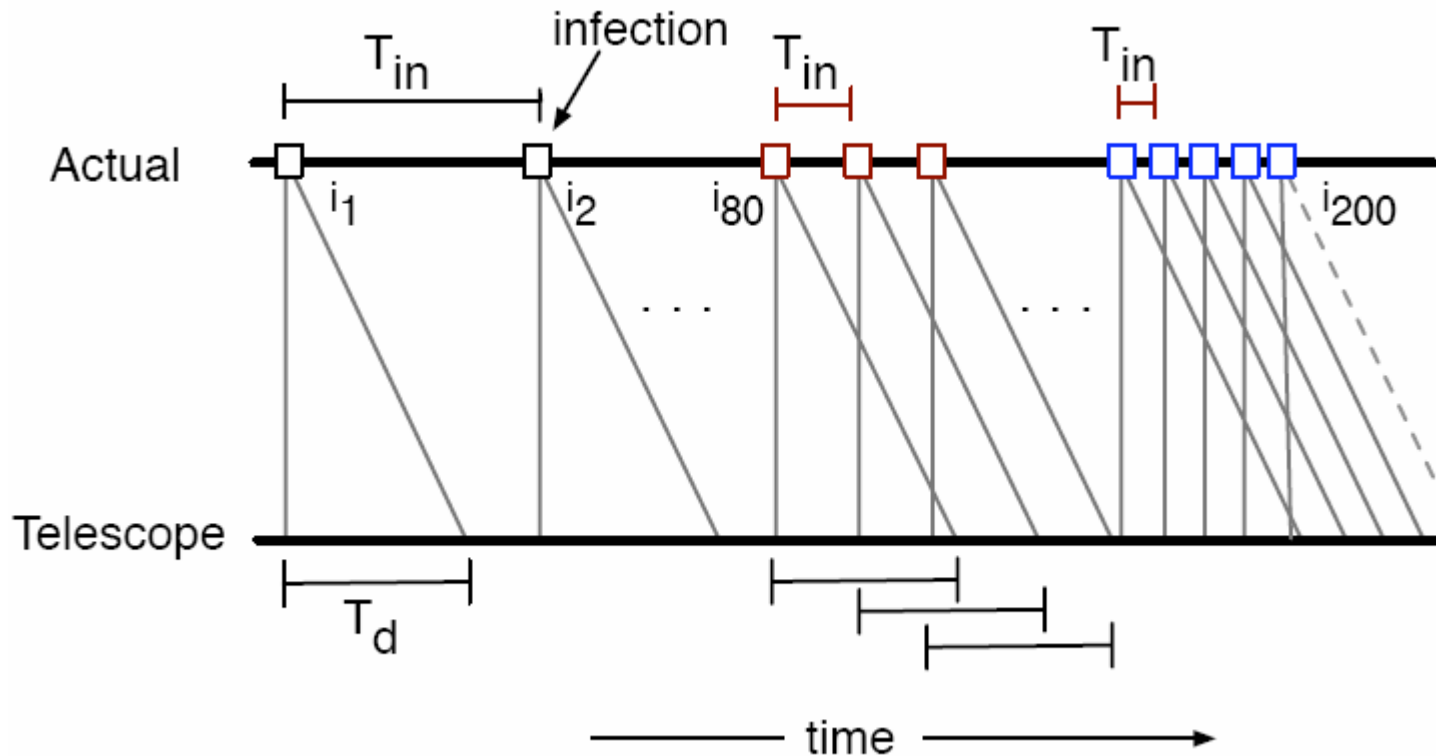
# Problem Statement and Goals

- Consider a uniform scanning worm with scanning rate  $s$  and vulnerable population size  $V$  and a monitor with effective size  $M$ .
  - To what extent can a network monitor trace the infection sequence back to patient zero through looking at the order of unique source contacts?
  - For worms that start with a hitlist, can we use network monitors to detect the existence of the hitlist and determine its size?

# Evolution Sequence and “Patient Zero”

- We distinguish between two processes:
  - Time to Infect  $T_{in}$ 
    - Time elapsed before the worm infects an additional host
  - Time to Detect  $T_d$ 
    - The time interval within which a monitor can reliably detect at least one scan from a *single* newly infected host

# Time to Infect and Time to Detect



# Time to Infect and Time to Detect

- From the AAWP model [Chen et al]

$$n_{i+1} = n_i + (V - n_i) \left[ 1 - \left( 1 - \frac{1}{2^{32}} \right)^{sn_i} \right]$$

- Time to infect a new host  $T_{in}$

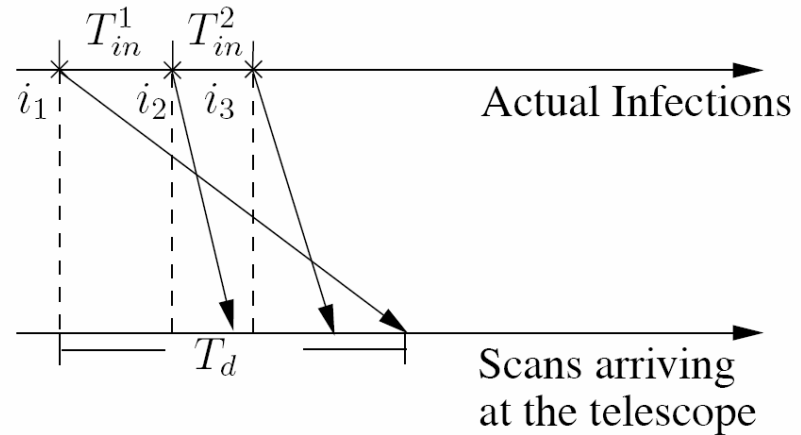
$$T_{in} = \frac{\log \left( 1 - \frac{1}{V - n_i} \right)}{sn_i \log \left( 1 - \frac{1}{2^{32}} \right)}$$

# Single host detection

- For a Monitor of size  $M$
- Time to detect a certain host with a certain confidence  $\alpha$

$$T_d = \frac{\log(1 - \alpha)}{s \log\left(1 - \frac{M}{2^{32}}\right)}$$

# Monitor Accuracy



## ■ Probability of error

$$P_e = 1 - \prod_{i=1}^n \left( 1 - \frac{M}{2^{32}} \right)^{\left( T_d - \sum_{j=1}^i T_{in}^j \right) s}$$

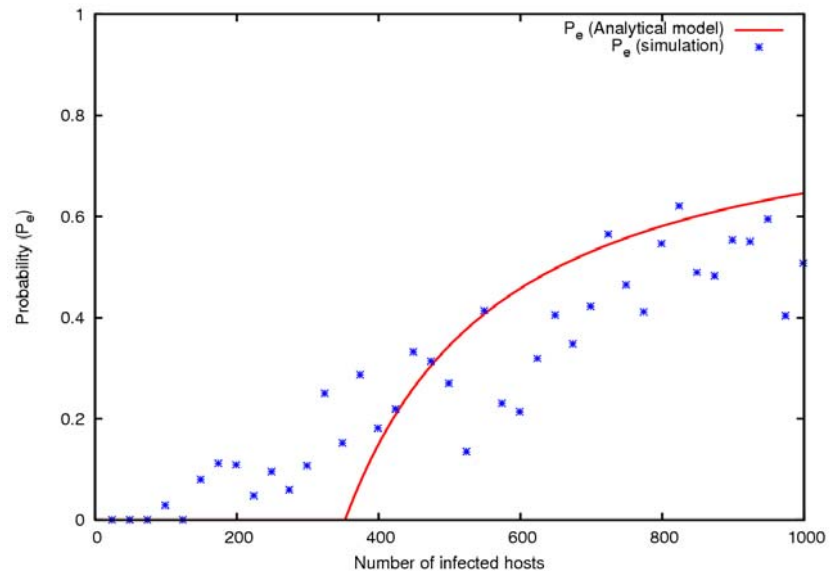
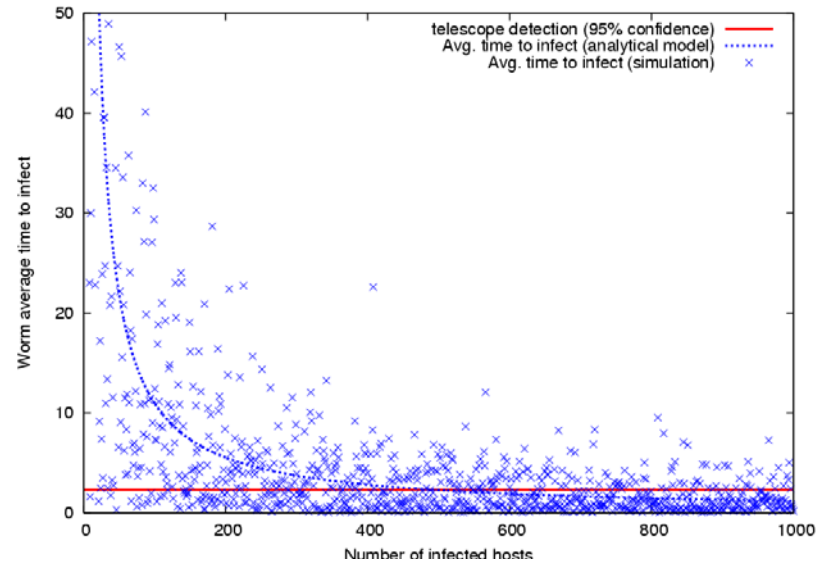
$T_{in}$  and  $T_d$

*Uniform scanning worm:*

$s = 350$  scans/sec,

$V = 12,000$

Monitor size = /8



# Similarity

## ■ Sequence Similarity

Actual Sequence (A)

1	2	3	4	5	6	7	8	9				m-1	m
---	---	---	---	---	---	---	---	---	--	--	--	-----	---

Monitor Sequence (B)

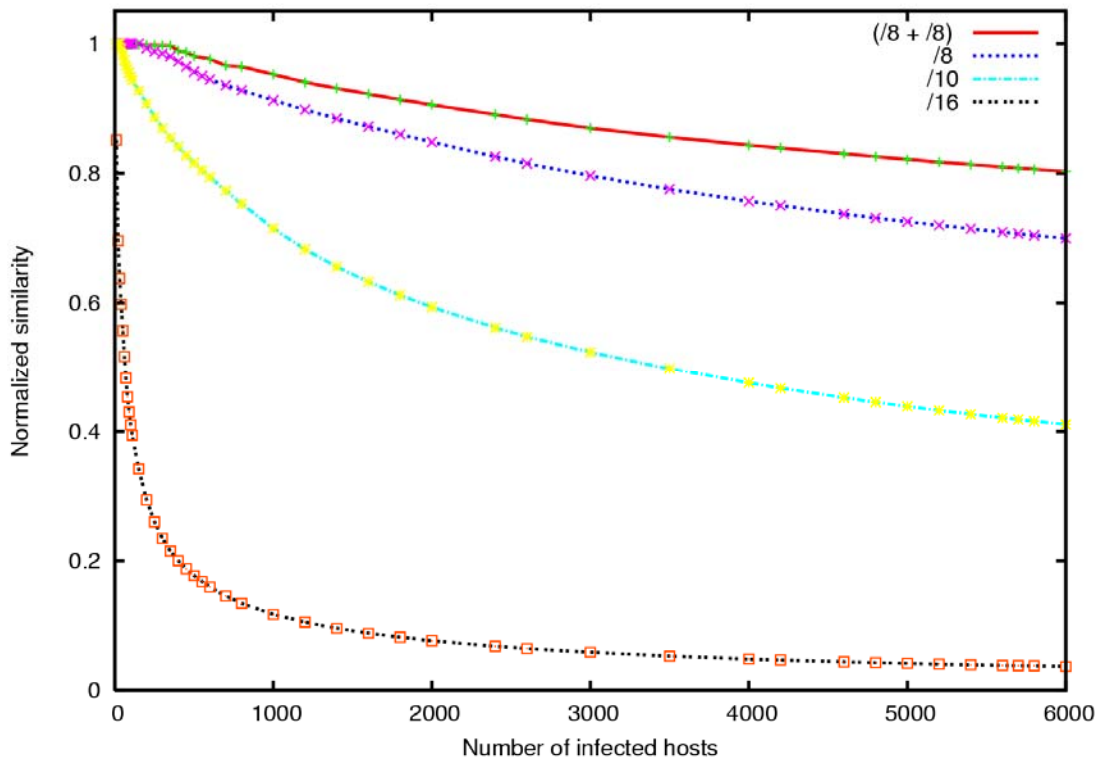
1	2	3	4	9	6	7	8	5				m-1	m
---	---	---	---	---	---	---	---	---	--	--	--	-----	---

$$Y_{B \rightarrow A} = \sum_{i=0}^m \frac{(m - r_{(e_i, A)})}{1 + \underbrace{|r_{(e_i, B)} - r_{(e_i, A)}|}$$

# Evaluation of Monitor Accuracy

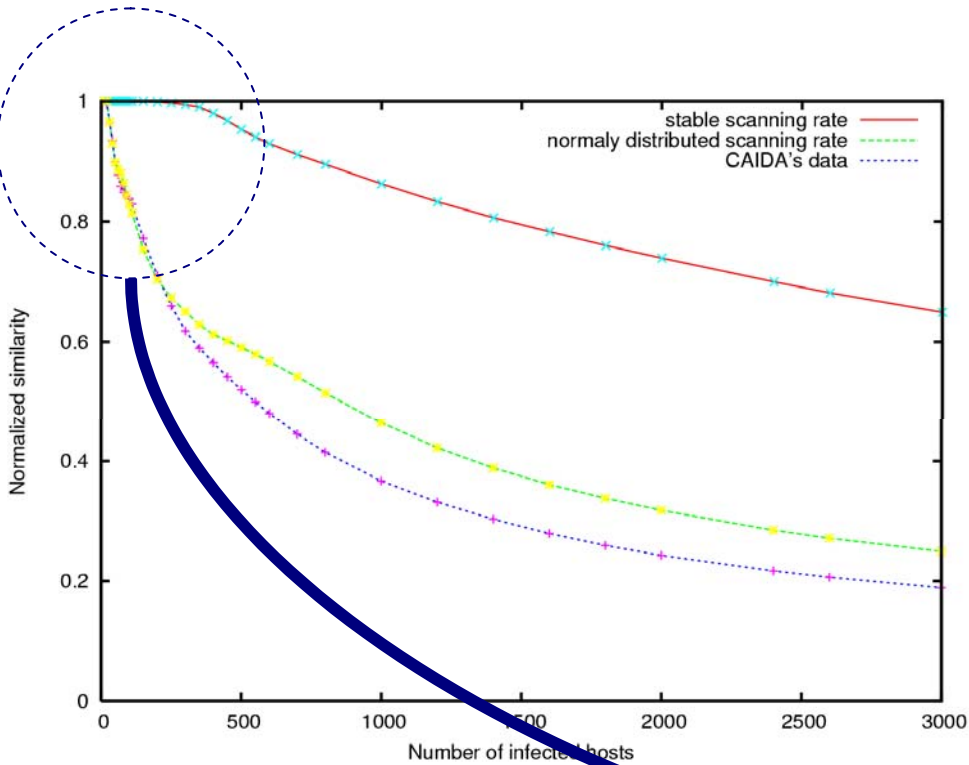
- Worm Evolution and Patient Zero:
  - Telescopes size
  - Vulnerable population distribution
  - Worm scanning rate in-homogeneity

# Worm Evolution Similarity (Impact of monitor size)

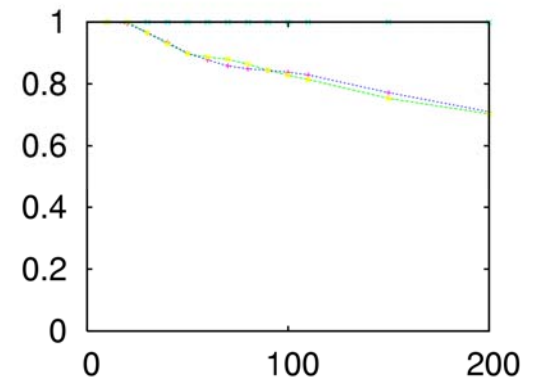


Large Telescopes give a highly similar view to the actual worm evolution

# Scanning Rate non-homogeneity



Scanning rate distribution from CAIDA's dataset

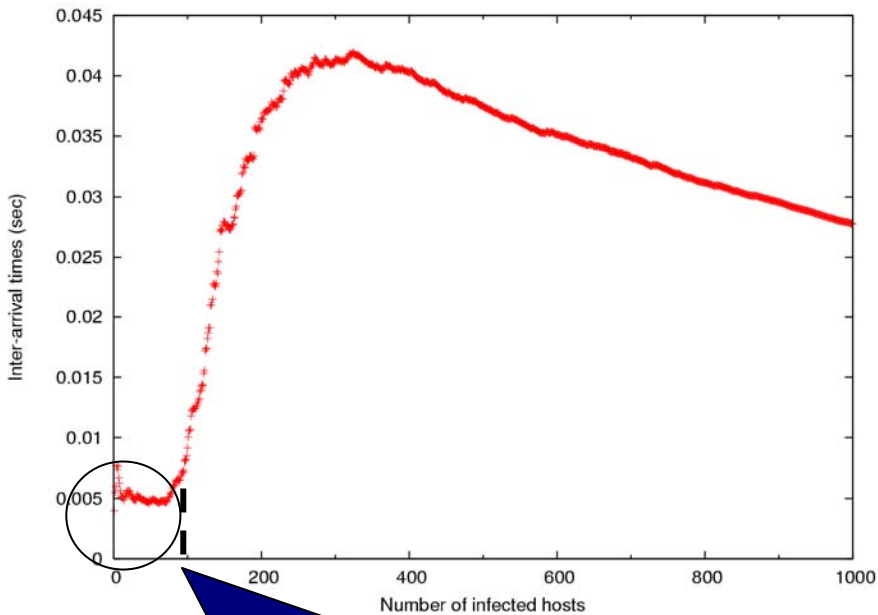


# Hitlist detection

- Hitlist is a set of sources targeted by an out of band mechanism (eg: Botnet, pre-established list of vulnerable machines)
- Exploit the *pattern of inter-arrival times* of unique sources contacts at the monitor to infer the existence and estimated size of the hitlist

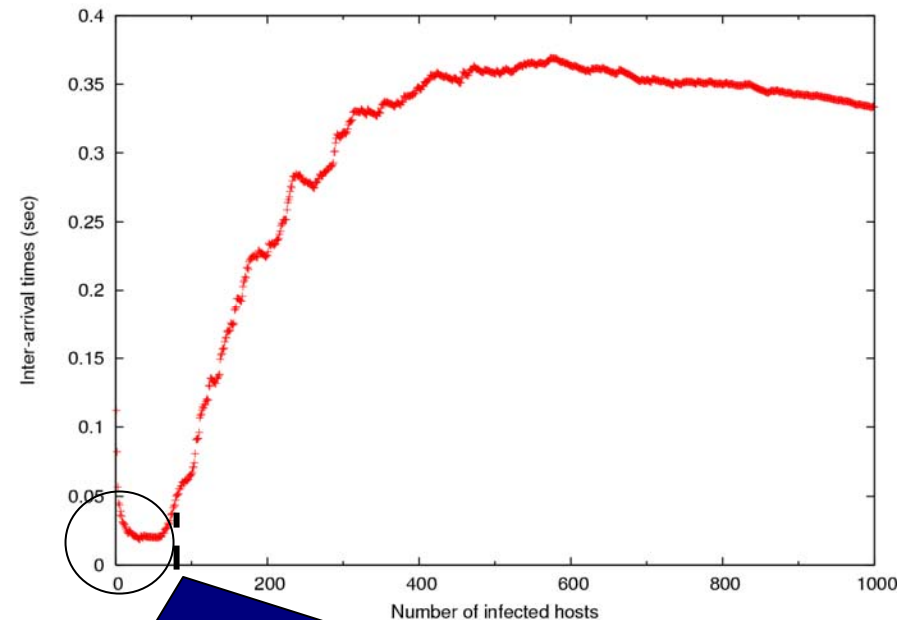
# Hit-list detection and size estimation

## Simulation



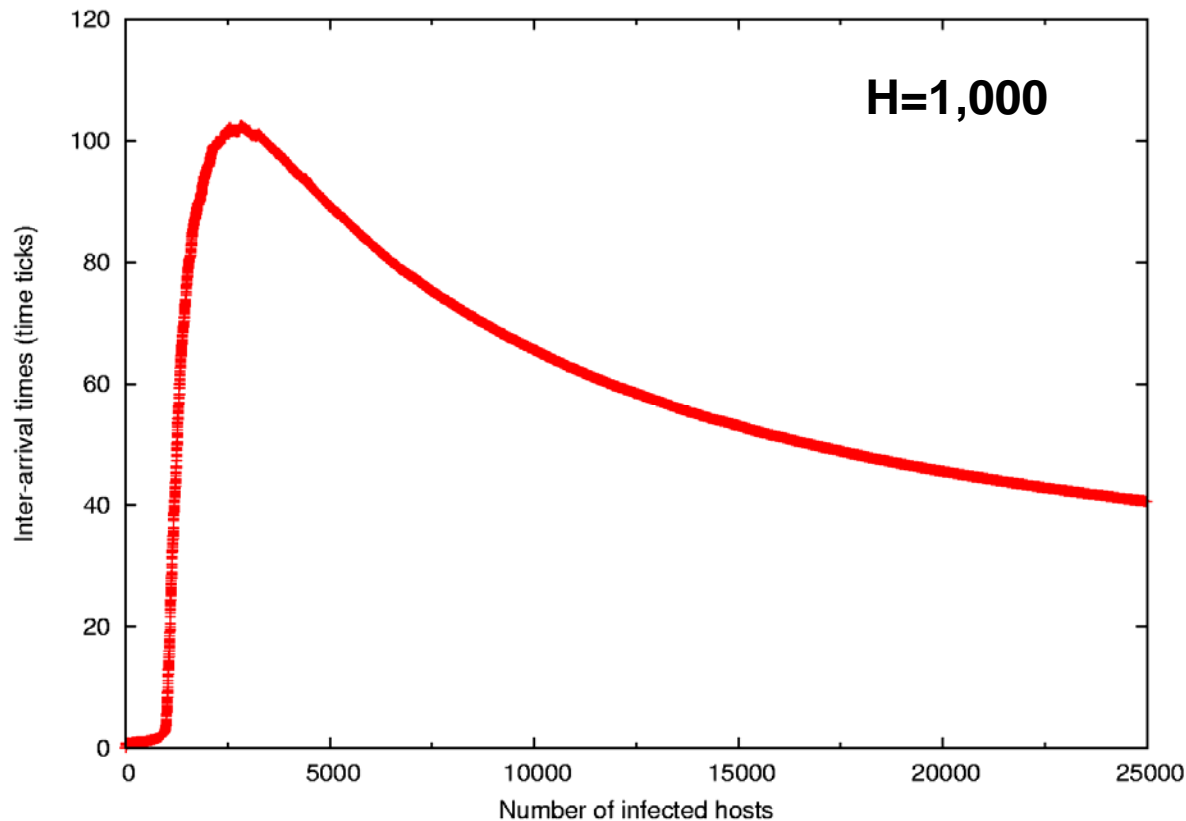
**Pattern Change  
around the hit-list  
boundaries  
 $H = 100$**

## Witty Worm



**Estimated hit-list  
 $H = 80$   
80% in the same /16  
88% belong to the same institution**

# Varying the hit-list size



- Same pattern was noticed when varying the scanning rate and for inhomogeneous scanning rates.

# Hit-list identification

- With a hit-list of size  $h_0$  the average worm infection time  $T_{in}$  should be less than  $T_d / h_0$

$$\log \left( 1 - \frac{1}{(V - h_0)} \right) \leq \frac{\log(1 - \alpha) \log \left( 1 - \frac{1}{2^{32}} \right)}{\log \left( 1 - \frac{M}{2^{32}} \right)}$$

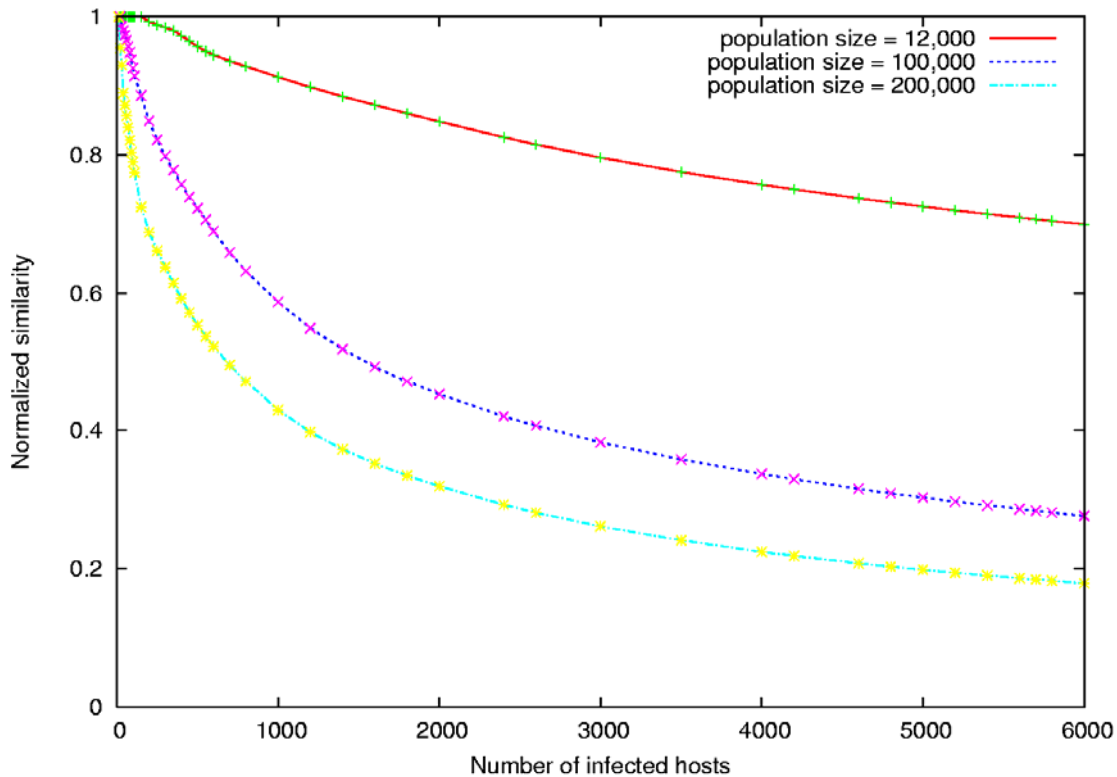
# Summary

- Network monitors give a reliable view of uniform scanning worm infection sequence back to patient zero.
- The accuracy of the monitor deteriorates further throughout the infection progression for higher population sizes and for highly inhomogeneous scanning rates
- Hit-list existence can be detected at the monitor by detecting the pattern change of unique infections' inter-arrivals at the monitors.



# Questions !

# Effect of Population size



Monitor accuracy degrades with the increase of the vulnerability density.