
# Computer Viruses: A View from the Trenches

Matthew Williamson

HP Labs Bristol

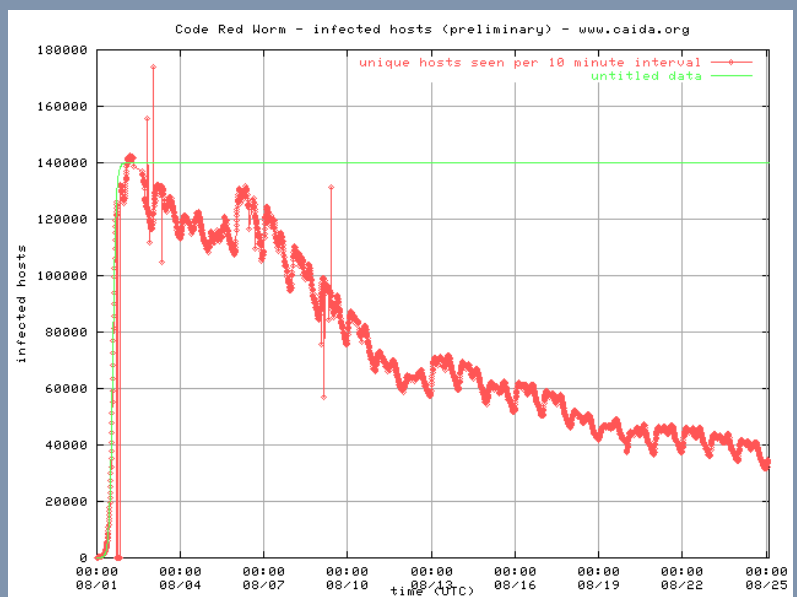
June 5, 2002

# Summary

- Epidemiology concentrates on spreading
- Computer viruses are not just about spreading
  - Remediation is where the $$s go
- Topology of real virus-threatened networks
  - Email, http, IM, etc.
  - Shape of address space
  - Virus behaviour

# Not just spreading

## Timescale of a virus attack




- Very quick rise
- Slow fall
- Why?
  - Reaction is late
  - Reaction is slow

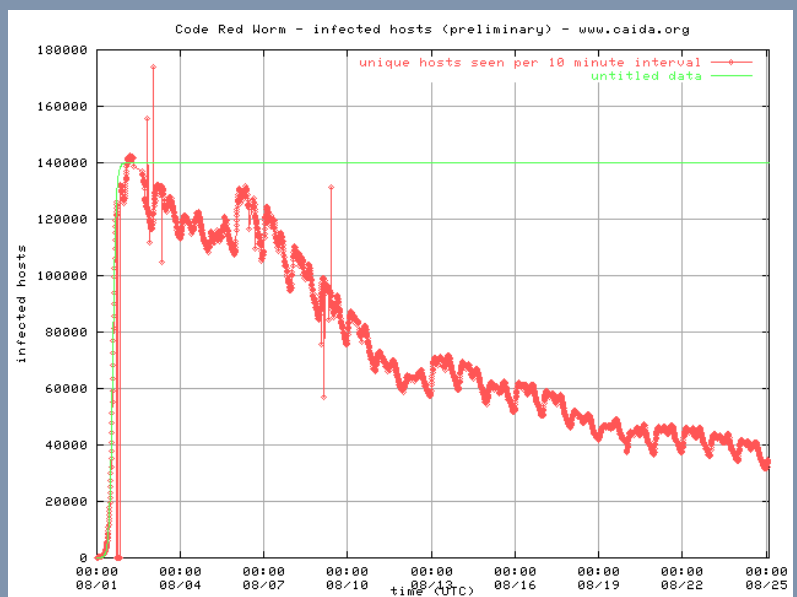
# Why late?

- Predominant technology used is signature based virus scanner
- Needs human to examine virus and generate signature and antidote
  - Takes time
- This information then needs to be distributed to all the machines
  - Takes time
- May need to reengineer the scanner itself
  - E.g. polymorphic viruses

# Why slow?

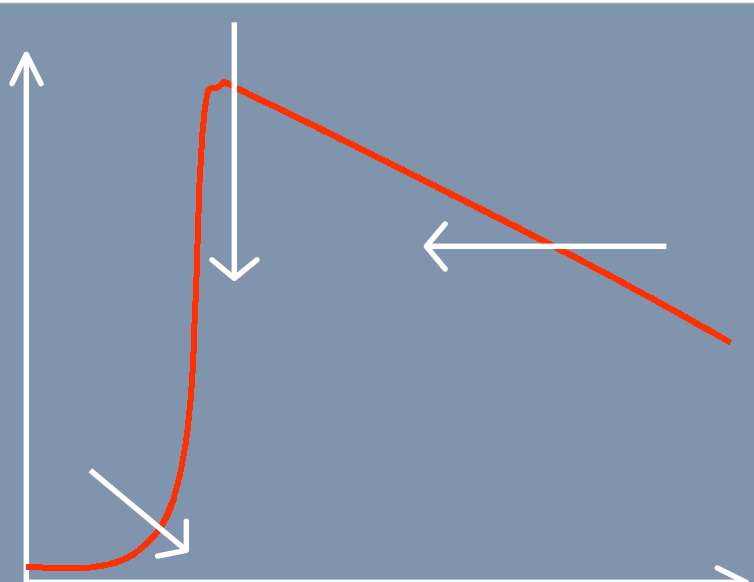
- Once got the signatures and distributed them, need to fix machines
- Enough machines not remotely manageable to be awkward
- Fundamentally manual process
  - It is slow and expensive
- Added difficulty of mapping IP address to physical location
  - Slow and expensive

# Costs are in the remediation



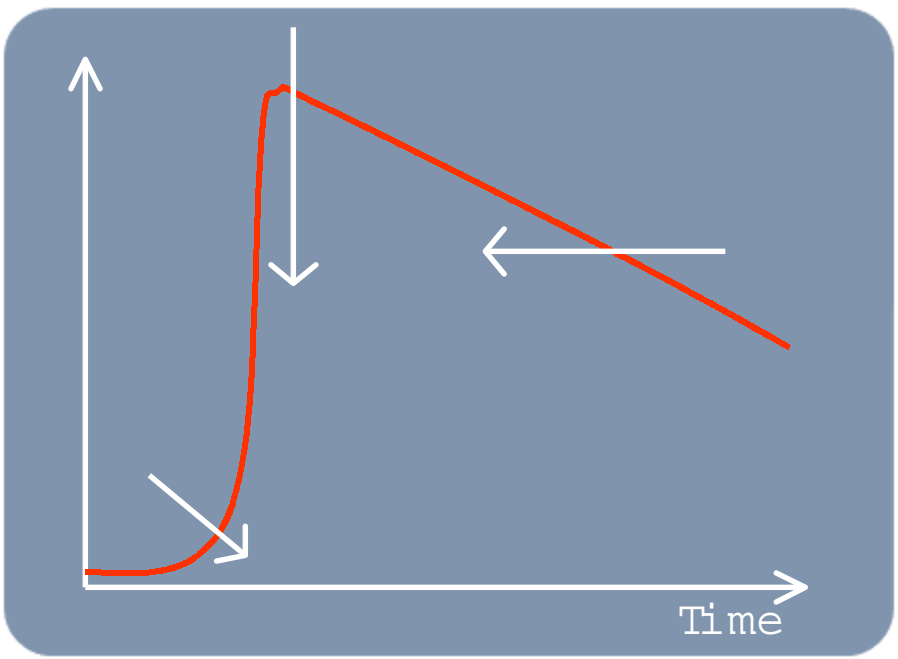

- Fixing it
  - IT staff
  - Employee's time
- Lost productivity
  - Lost connectivity
  - Lost services e.g. printing
  - Lost business
    - Business processes
    - Order fulfilment
    - Comms with vendors
  - Lost brand
- Longer it goes on, the costlier it is
- Worse the infection, the costlier it is

# Three things to do




- Stopping spreading is important
- Reducing number of susceptible machines is important
- Speeding up remediation is important
- How can epidemiology help?

# Ideas




- SIR models not SIS
- Death rate varies with time
  - Zero until get a virus signature
  - I.e. cannot assume =1

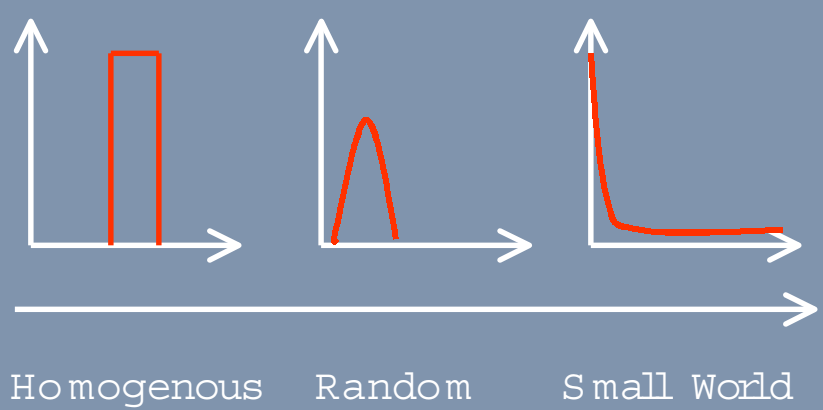
# Summary

- Reactions to real viruses late and slow
- Long time to clean up
- $$ lost
  - Fixing it
  - Lost productivity
- If could clean up faster, reduce this loss
- SIR/SIS
- Death rate as function of time

# Topology of real networks

- Fundamental assumption:
  - Modelling spread of viruses over networks will help us control them
- But – only if the models are realistic
- Tension between complexity of network and usefulness of result???

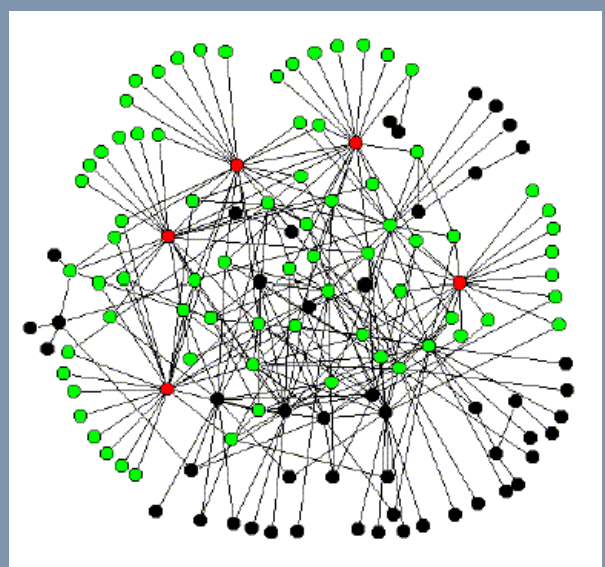
# Topology of real networks




- What is the topology of an email, http, mobile phone, instant messaging, shared drive etc. virus?
- Is it homogeneous, random, small world/scale free?
- Depends on shape of address space
  - Not the machines
- Depends on virus behaviour

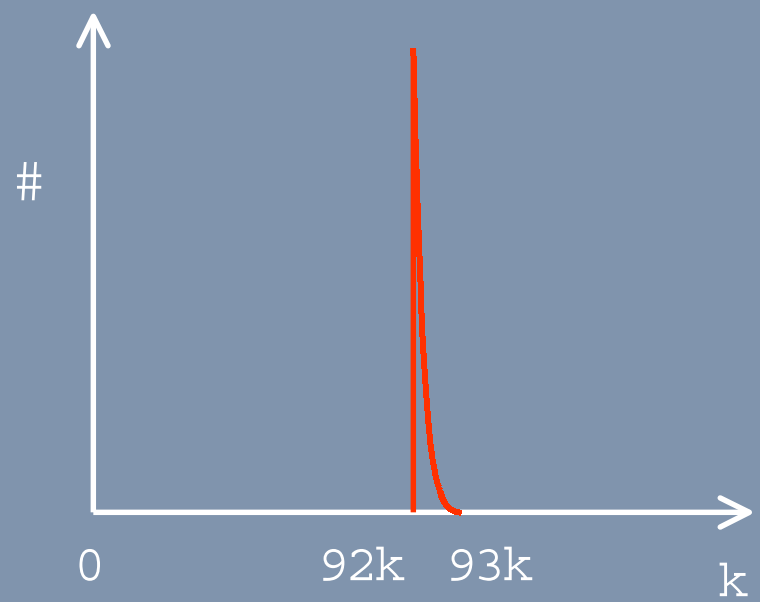
# Email viruses

## What is the shape of the address space?



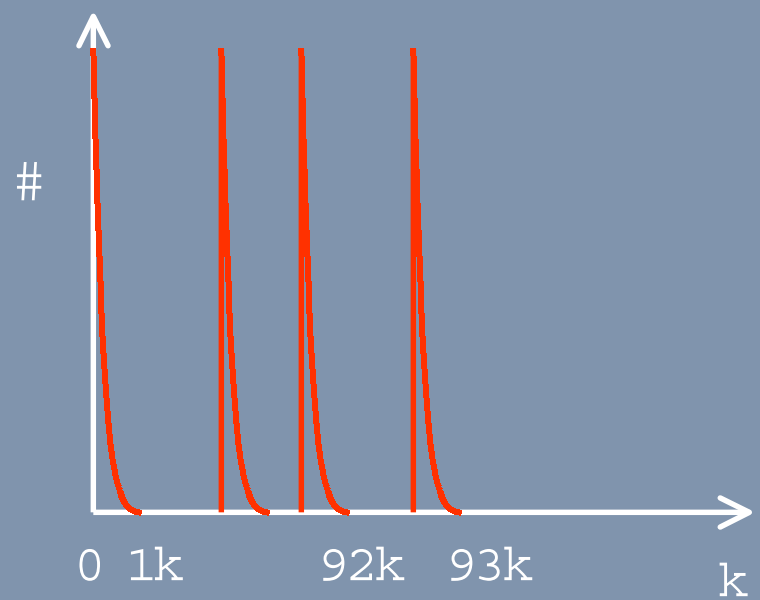
- Normally spread by virus sending itself to every entry in the email address book
- Topology?
  - Some researchers think small world
    - Ebel, Mielsch & Bornholdt (2002)
  - ISPs Hotmail etc. probably true
  - Corporate world is different.

# Corporate email e.g. HP



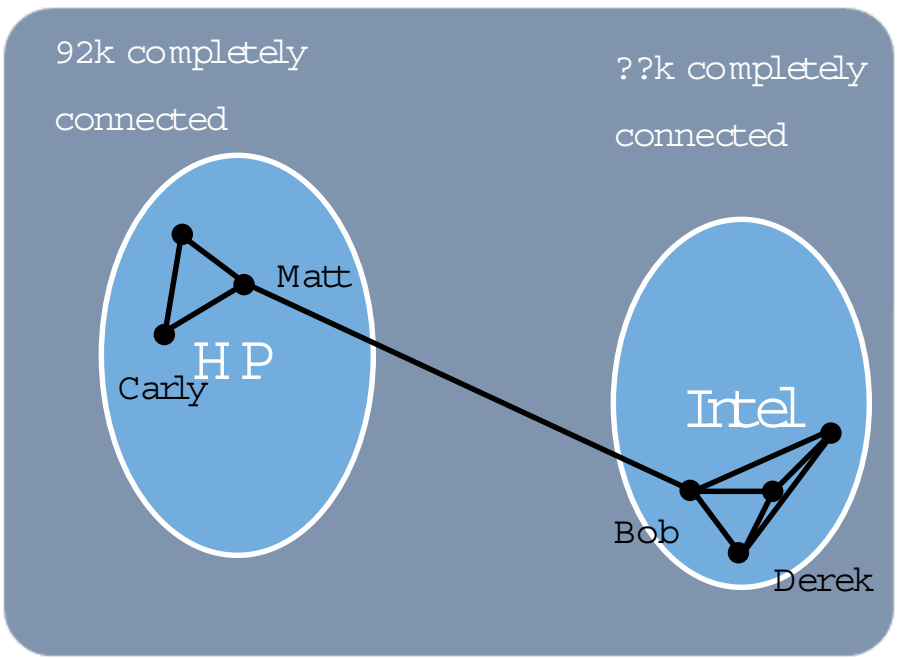

- Pre-merger HP had 92k employees
- Each with an address book with 92k entries
- This is more like a homogenous network?
- Other corporate sites likely to follow same pattern.

# Overall address space




- Some parts small world
- Some parts homogeneous
- Lots of spikes of different sizes
- Small world with large homogeneous hot-spots?
- Graph misleading – fully connected areas

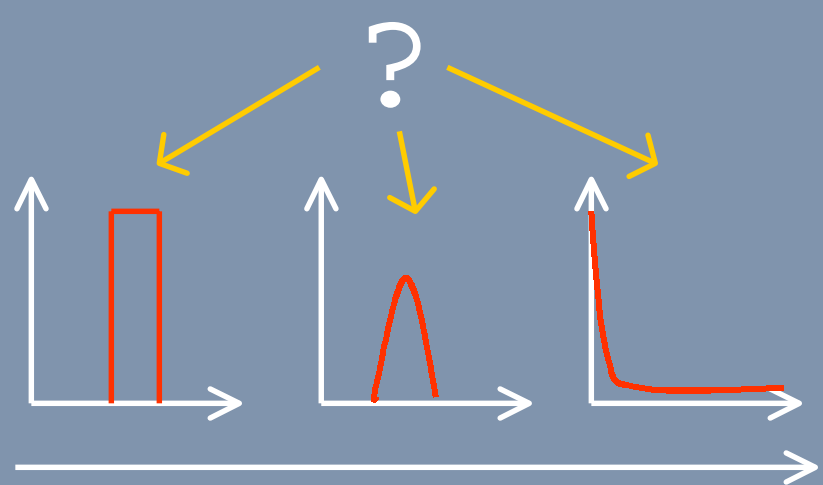
# What does it look like?




- Large completely connected blobs
- Sparser connections between blobs
- Length between 2 people quite small
- Takes time to find link though

# Virus behaviour

## Large effect on topology




- Some viruses only used top 100 entries in address book
- What if virus took a random sample
- What if virus looked at addresses of email received?
- Etc. etc.

# Http virus topology

e.g. Code Red, Code Red II, Nimda

- Recent highly damaging viruses
- Automatic spread – no human in the loop (as with email)
- Affected IIS web server, virus spreading using http requests
- Spread by "guessing" IP addresses and attacking machines at those addresses.

# Internet IP topology
## Background

Class A

15.*.*.*

Class B

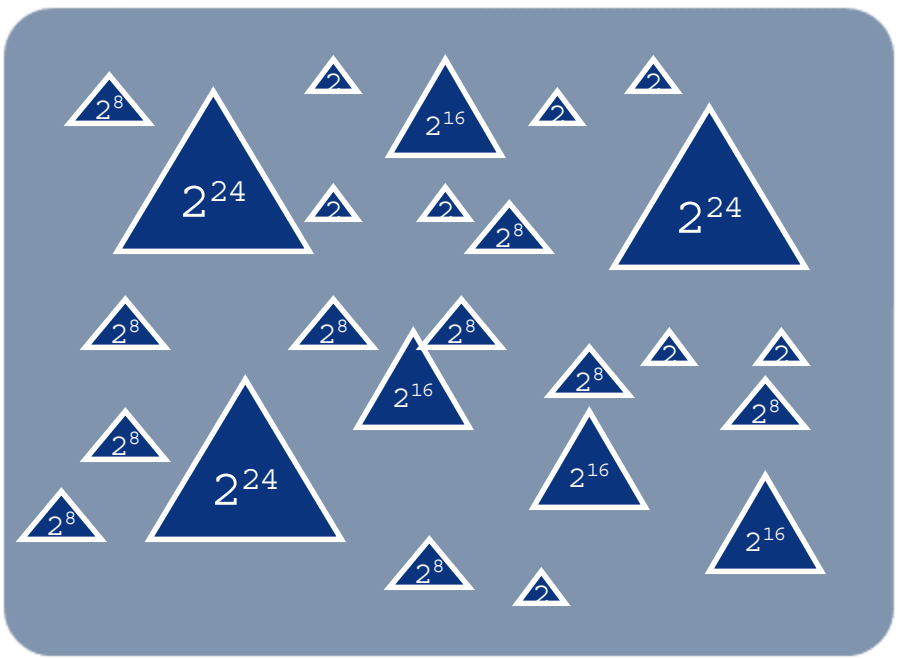
15.193.*.*

Class C

15.193.45.*

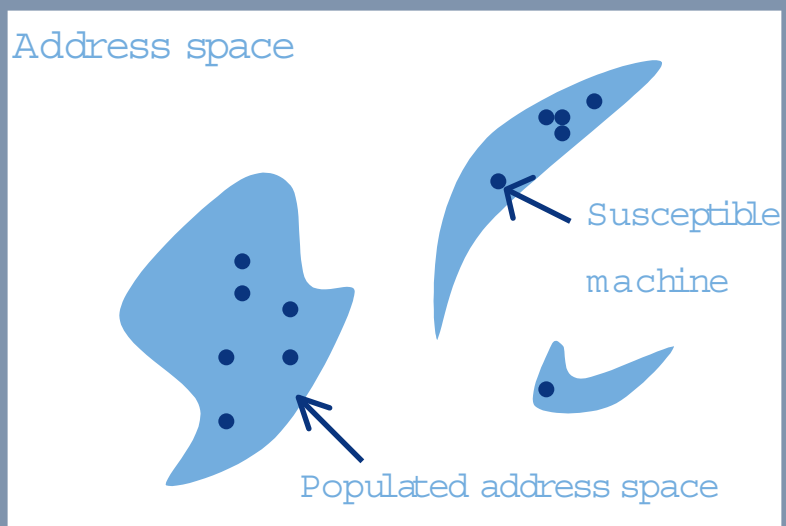
- Address space is defined by 32 bit number – IP address.
- IP address split into 4 8-bit chunks.
  - E.g. 15.193.45.76
  - Hierarchical
- Address space "owned" by companies/institutions
  - E.g. HP has 15.*.*.* and (now) 16.*.*.*
- Theoretically completely connected

# Hard outside, chewy inside



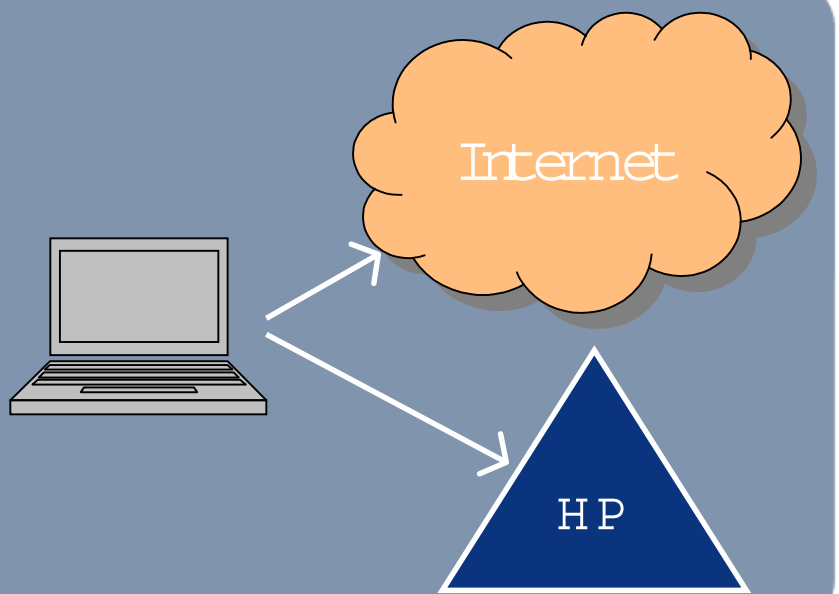

- Links are directional
- Firewalls restrict traffic in and out of chunks of the address space
- Generally no restrictions within an intranet
- Network consists of various sized chunks with
  - Outsides completely connected
  - Insides completely connected
  - Very rare outside-to-inside connections

# Variable density



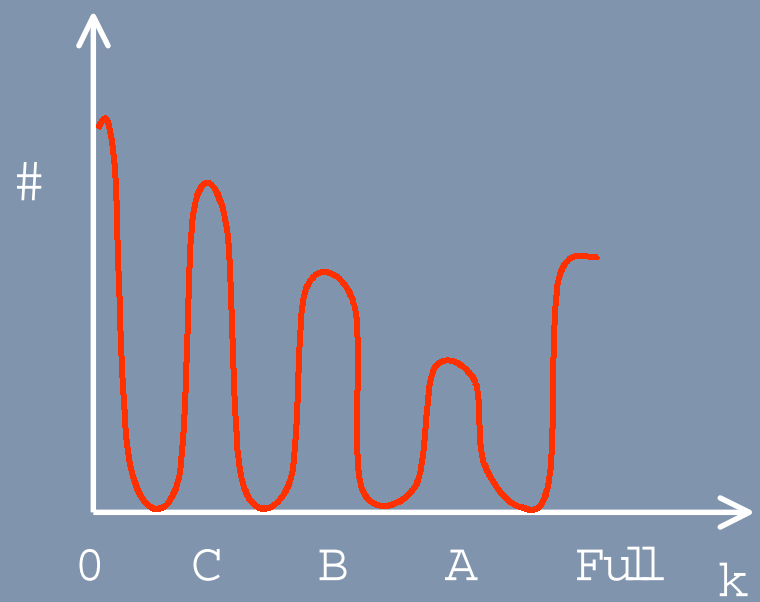

- Not all addresses are populated with machines
  - Some areas sparse e.g. HP with $2^{24}$ addresses and 600k machines
  - Some areas dense e.g. AOL and NAT
- Not all machines are susceptible
  - What % have the vulnerability?
- As you connect machines of the same OS, more vulnerable machines
  - E.g. mobile phones

# Further complications...




- Machines change address and move
  - DHCP
  - Laptops
  - Virtual Private Networks (VPN)
- Effect of bandwidth
  - On intranets, links between sites can saturate giving a loss in connectivity
  - Routers max out

# So what does the graph look like?




- Roughly????
- C, B, A are size of networks
- Not at all sure about the magnitudes
- It is a directed graph, with fully connected chunks

And then there is the effect of virus behavior!!

- Code Red generated IP addresses at random
- Code Red II generated IP addresses biased towards the local subnet
- Nimda used multiple infection routes – multiple topologies?
- If Code Red had looked at any html pages on the compromised server, it could have spread along hyperlinks
  - Small world topology?

# Theoretical viruses

(Staniford, Paxson and Weaver 2002)

- Warhol worm
  - Effective scanning
  - Starts with list of vulnerable hosts
  - Deals with complexity of IP network
- Flash worm
  - Pre-compute all vulnerable hosts
  - Go infect, do not scan
  - Fully connected network

# What about other viruses?

Hopefully simpler!

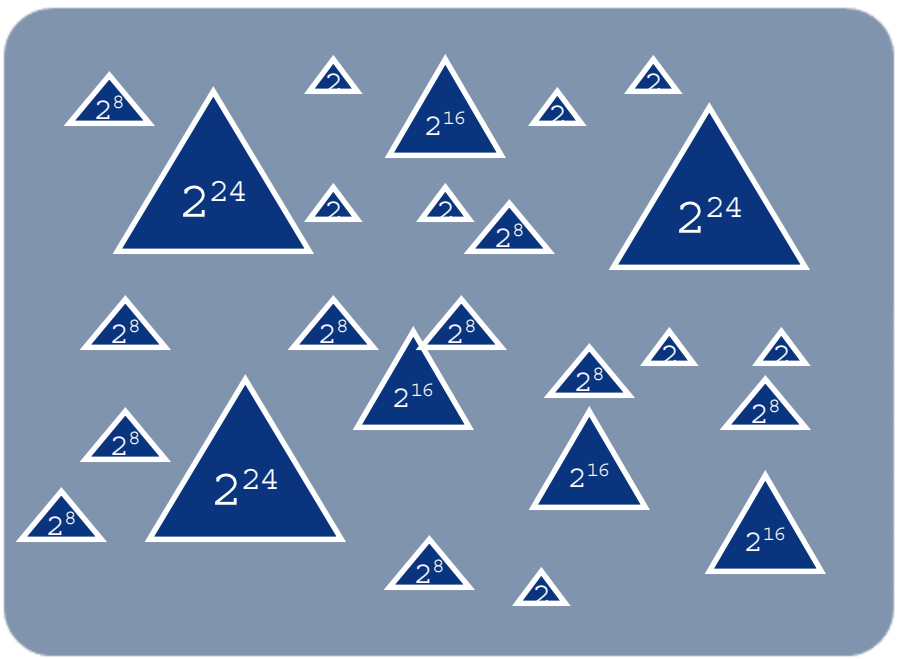
- Instant Messaging
  - Probably small world if use "address lists"
  - Corporate IM??
  - Goner, (ICQ); CoolNow (MSN Messenger)
- Mobile phones?
  - Small world if use addresses
  - Could scan for vulnerable phones
  - Aiello, Chung & Lu (2000)
- Shared drives
  - Could be small world??
  - Worth looking at
  - E.g. Nimda, FunLove

# Summary of network topologies

| Protocol | Type of network |
|---|---|
| Email | Small world + areas of homogeneous + directed |
| http | Directed + dynamic + fully connected in places |
| Mobile Phone, shared drive, Instant messaging | Probably small world |

- Topology depends on
  - Shape of address space
  - Virus behavior
  - Distribution of vulnerable hosts in the address space
- Real networks appear to be
  - Partly small world/social
  - Partly homogeneous
  - Directed
  - Complex

# Ideas




- Networks that are completely connected with tightly defined chunks
  - Get slopping?
- Dynamic (mobile computers)
- Asynchronous
- Directed (firewalls/NAT)
- Not all addresses map to machines
- Not all machines are vulnerable

# Summary

- Not just about spreading
  - Slowing
  - Vaccination
  - Clean up
- Clean up is where the $$'s are spent and lost
- Topology of real virus-threatened networks
  - Address space + virus behaviour
  - Most have some sort of small world structure
  - Completely connected chunks
  - Real email/web networks messy

# Making models more realistic

- SIR not SIS
- Death rate changing with time, and zero at first.
- Directed graphs
- Graphs with local areas of complete connectivity
- Not all machines susceptible
- Asynchronous updates?


hp
invent