EPIDEMIC SPREADING IN SCALE-FREE NETWORKS

or

Epidemic modeling of computer viruses

Alessandro Vespignani (ICTP-Trieste)

Romualdo Pastor-Satorras (Tech. Univ. Barcelona)

Recently :

Yamir Moreno Vega (ICTP)

Alexei Vazquez (SISSA-ICTP)

APS March Meeting

Two levels

•Microscopic level
Researchers who disassemble and try to kill off new viruses.

Corresponds to the quest for new vaccines and medicines

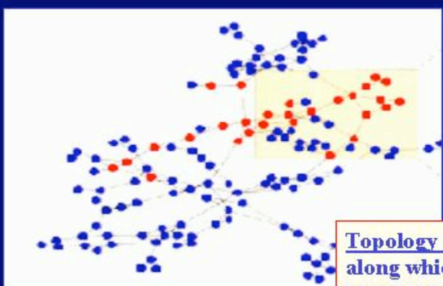
•Macroscopic level
Statistical analysis and modeling of epidemiological data in order to find informations and policies aimed at lowering epidemic outbreaks

Macroscopic prophylaxis , Vaccination campaigns

Mathematical models of epidemics

Coarse grained description of individuals and their state

- •Individuals exist only in few states:
- •Healthy or Susceptible * Infected * Immune * Dead
- •Particulars on the infection mechanism on each individual are neglected.



Topology of the system: the pattern of contacts along which infections spread in population is identified by a network

- •Each node represents an individual
- •Each link is a connection along which the virus can spread

The Susceptible-Infected-Susceptible (SIS) model

- Each node is infected with rate ν if connected to one or more infected nodes
- Infected nodes are recovered (cured) with rate δ without loss of generality $\delta = 1$ (sets the time scale)
- Definition of an effective spreading rate $\lambda = \nu/\delta$

re-infection is possible.

Dynamical Mean-Field equation for the order parameter ρ = **density of infected nodes**

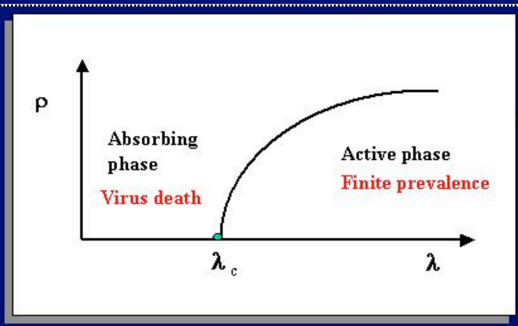
$$\partial_t \rho(t) = -\rho(t) + \lambda \langle k \rangle \rho(t) [1 - \rho(t)] + h.o. \; .$$

In the stationary state $\partial_t \rho = 0$, we have

$$\rho[-1 + \lambda \langle k \rangle (1 - \rho)] = 0$$

Definition of the epidemic threshold $\lambda_c = \langle k \rangle^{-1}$

$$\rho = 0 \quad \text{if } \lambda < \lambda_c,$$
$$\rho \sim \lambda - \lambda_c \quad \text{if } \lambda > \lambda_c,$$

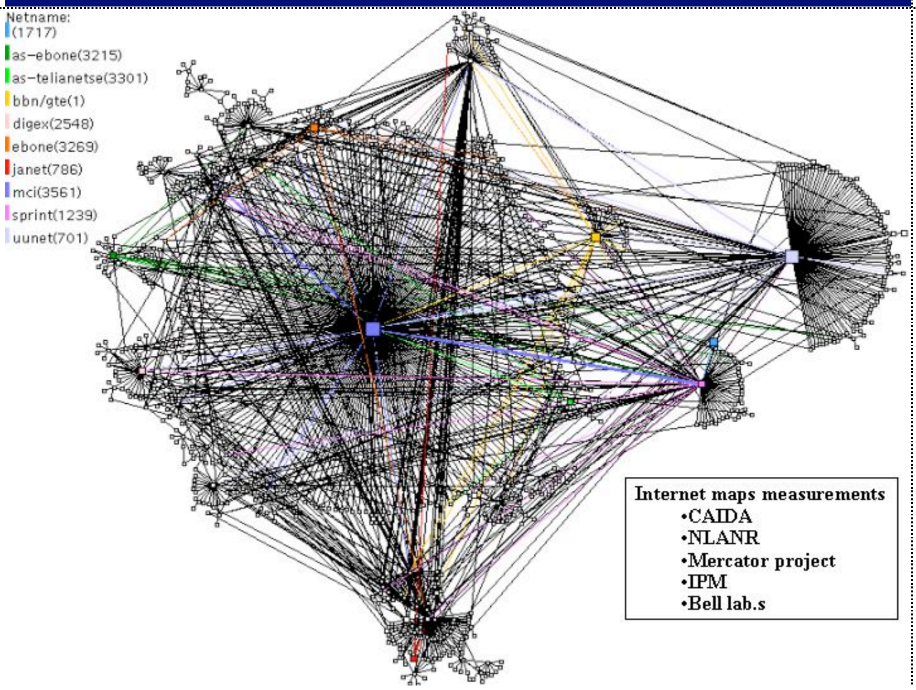


- Non-equilibrium phase transition
- SIS model is a variation of the contact process
- epidemic threshold = critical point
- prevalence ρ = order parameter

Similar models with immunity and death (removal) can be defined (SIR etc)

The epidemic threshold is a general result

The question of thresholds in epidemics is central



Main properties
- **complex network**
- **preferential attachment**
- **local clustering**

Modeling of scale-free networks by Barabasi et al. (1999)

- **The Internet and the World-Wide-Web**
- **Protein networks**
- **Metabolic networks**
- **Social networks**
- **Food-webs and ecological networks**

• $<k^2> \to \infty$

Scale-free properties ⟷ Diverging fluctuations

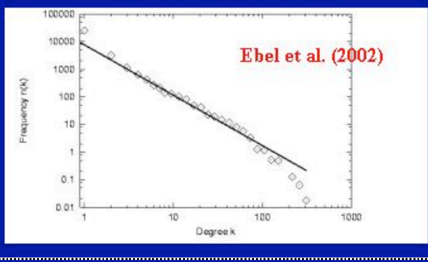
Natural computer virus
- DNS-cache computer viruses
- Routing tables corruption

Data carried viruses
- ftp, file exchange, etc.

→ Internet topology

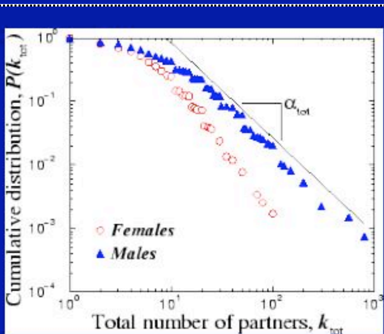
Computer worms
- e-mail diffusing
- self-replicating

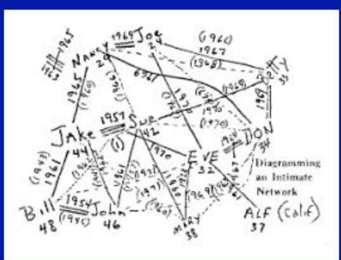
→ E-mail network

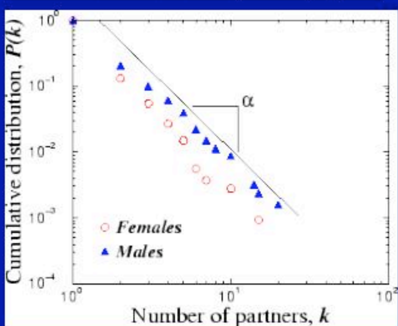


The web of Human sexual contacts

[Lilijeros et al., Nature (2001)]



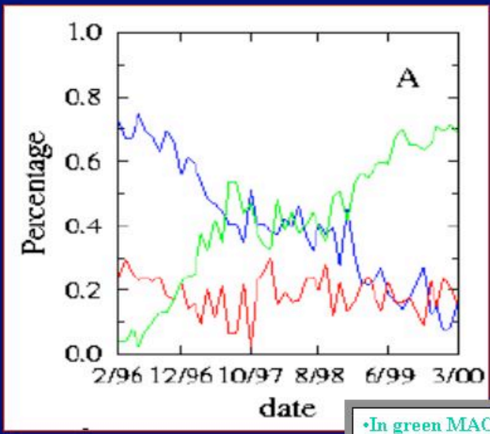




Strain data analysis

We analyzed homogeneous groups of viruses

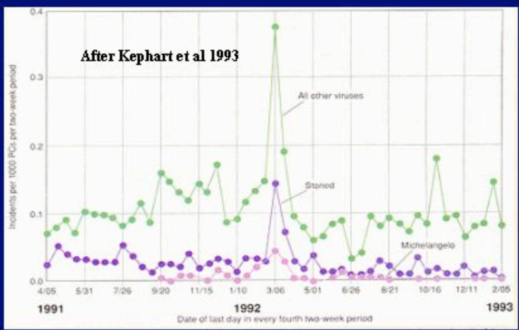
- effective parameters
- similar properties within the strain



- In green MACRO viruses
- In red FILE viruses
- In blue BOOT viruses

Real data from viruses in the wild

- **Prevalence data from large monitored samples**



- Just a few viruses are lucky enough to pervade (sub-critical or very close to criticality ??)
- In the endemic case prevalence is always very small (ρ<0.01) but stationary for long period.

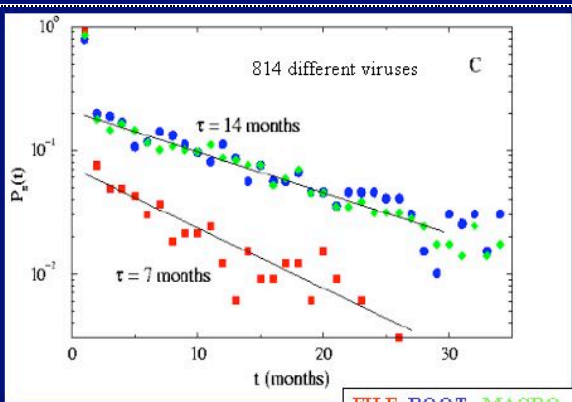
Why computer viruses are always tuned infinitesimally close to the epidemic threshold??

According to S.White this is one of the most relevant open problem in computer virus epidemiology.

- Survival probability $P_s(t)$ = fraction of viruses still in the wild at time t after their birth

$$P_s(t) \sim \exp(-t/\tau)$$

τ = average lifetime (characteristic time) of the virus strain



FILE BOOT MACRO

Average lifetime extremely long compared to the virus rates time scale

- Anti-virus software is delivered in a few hours after the first detection
- ILoveYou virus is still present in the wild list after two years

What we do have learned

•Strain data analysis is reasonably consistent (definition of effective parameters)

•Long lifetime of viruses is not compatible with anti-virus software delivery time-scale.

•Data strengthen the question of why according to standard models all viruses seems very close to the epidemic threshold

Various kind of topology have been attempted (Random graph, local lattice etc.) (Kephart et al)

•All virus strains share the same characteristics

•The MACRO strain (particularly) is platform independent and travel essentially on the internet

•The Internet topology should be included in the virus spreading

Epidemic spreading on Scale-Free networks

• Highly connected nodes are statistically significant $\langle k^2 \rangle = \infty$

• Connectivity fluctuations must be included

Relative density $\rho_k(t)$ of infected nodes with given connectivity k

$$\partial_t \rho_k(t) = -\rho_k(t) + \lambda k [1 - \rho_k(t)] \Theta(\rho(t)),$$

$\Theta(\rho(t))$= Prob. that any given link points to an infected node.

Annihilation term

Creation term

• Θ is function of the average density of infected nodes

• Links point with higher probability to highly connected nodes

[Pastor Satorras &Vespignani, PRL 86, 320(2001)]

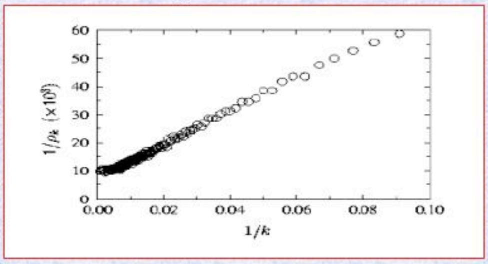
Stationary state

In the stationary state $\partial_t \rho_k(t) = 0$ we have that $\Theta(\rho) \Rightarrow \Theta(\lambda)$.

MF equations yield

$$\rho_k = \frac{k\lambda\Theta(\lambda)}{1 + k\lambda\Theta(\lambda)}.$$

Simulations in a BA network



•Higher is the node connectivity and higher is the probability to be in an infected state

•Strong inhomogeneity

•A link is more likely connected to a node with high connectivity

•The probability that a link points to a node with *s* links is proportional to *sP(s)*.

The probability of pointing to an infected node is

$$\Theta(\lambda) = \sum_k \frac{kP(k)\rho_k}{\sum_s sP(s)}.$$

ρ_k themselves are a function of $\Theta(\lambda)$

Self-consistent equation

Finally the equation for the order parameter is

$$\rho = \sum_k P(k)\rho_k$$

In the case of the BA-model ($P(k) = 2mk^{-3}$), we consider k as a continuous variable and $\langle k \rangle = 2m$.

The first self-consistent equation is

$$\Theta(\lambda) = m\lambda\Theta(\lambda) \int_m^\infty \frac{1}{k^3} \frac{k^2}{1 + k\lambda\Theta(\lambda)}$$

which yields the solution

$$\Theta(\lambda) = \frac{e^{-1/m\lambda}}{\lambda m}(1 - e^{-1/m\lambda})^{-1}$$

The order parameter equation is

$$\rho = 2m^2\lambda\Theta(\lambda) \int_m^\infty \frac{1}{k^3} \frac{k}{1 + k\lambda\Theta(\lambda)},$$

Obtaining

$$\rho = 2e^{-1/m\lambda} + h.o.$$

•Absence of any epidemic threshold (critical point)

•Active state for any value of λ

•The infection pervades the system whatever spreading rate

•In infinite systems the infection is infinitely persistent (indefinite stationary state)

Epidemic threshold in scale-free networks

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle}$$

$$\langle k^2 \rangle \rightarrow \infty$$

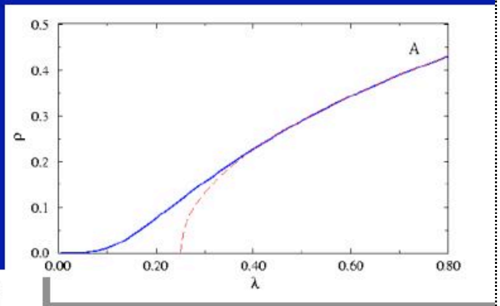
$$\lambda_c \rightarrow 0$$

Order parameter behavior in an infinite systems

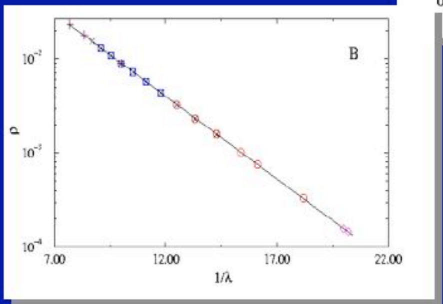
$$\rho = 2e^{-1/m\lambda}$$

$$\rho = 2e^{-1/m\lambda}$$

Numerical simulations in a BA network

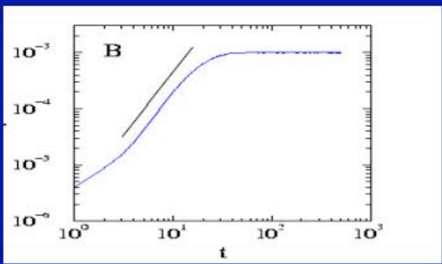


Zoom in lin-log scale



Network sizes $N=10^3$ to $N=10^7$

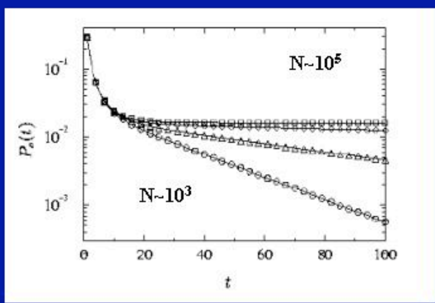
Spreading of a virus starting from a localized seed



Time behavior of the prevalence (algebraic)

$\lambda=0.06$

Surviving probability with the same definition used to analyze data.



Exponential behavior with characteristic time increasing with the network size

Results can be generalized to generic connectivity distributions $P(k) \sim k^{-\gamma}$

- **If $2 < \gamma \leq 3$ we have absence of an epidemic threshold and no critical behavior.**
- **If $3 < \gamma \leq 4$ an epidemic threshold appears, but it is approached with vanishing slope (no criticality).**
- **If $\gamma > 4$ the usual MF behavior is recovered. SF networks are equal to random graph.**

Epidemic threshold in scale-free networks

$$\lambda_c = \frac{<k>}{<k^2>}$$

$$<k^2> \rightarrow \infty$$

$$\lambda_c \rightarrow 0$$

Order parameter behavior in an infinite systems

$$\rho = 2e^{-1/m\lambda}$$

Finite size scale-free networks

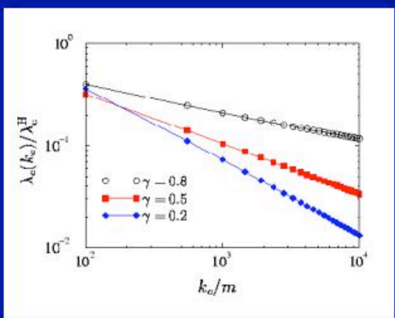
$P(k) \sim k^{-\gamma} \exp(-k/k_c)$

Exponentially bounded

$P(k) \sim k^{-\gamma}\, \theta(k - k_c)$

Hard cut-off

$$\lambda_c \sim k_c^{\gamma-3}$$

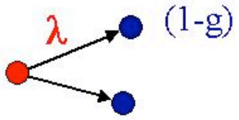


Ratio between SF and homogeneous Epidemic threshold for k_c, $N=10^4$

$$\Lambda_c / \lambda_c^H < 10^{-1}$$

Immunization

- Random immunization:
  g = density of immune nodes



$$\lambda \rightarrow \lambda\,(1-g)$$

Epidemic dies if $\lambda\,(1-g) \leq \lambda_c$

Regular or random networks

$$\rho_g = \rho_0 (g_c - g)/g_c$$

Immunization threshold

$$g_c = (\lambda - \lambda_c)/\lambda$$

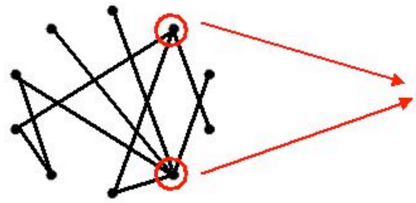
Scale-free networks → $\rho_g \sim \exp(-C/(1-g))$

Immunization threshold $g_c = 1$

•Random immunization is totally ineffective

•Different immunization specifically devised for highly heterogeneous systems

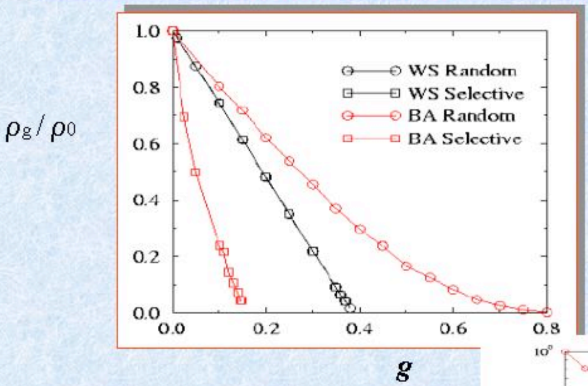
Targeted immunization strategies

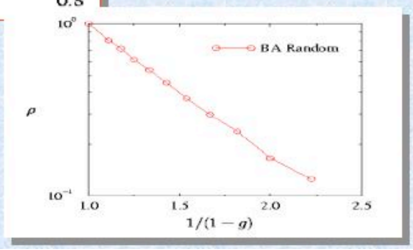


Progressive immunization of crucial nodes

Epidemic threshold is reintroduced → $g_c = \exp(-2/m\lambda)$

Numerical Simulations on Scale-free and Random Networks





[Pastor Satorras &Vespignani, PRE 65, 036104 (2002)]
[Dezso & Barabasi cond-mat/0107420]

MAIN RESULTS FOR S-F NETWORKS

- Absence of an epidemic/immunization threshold
- The network is prone to infections (endemic state always possible)
- Small prevalence for a wide range of spreading rates
- Progressive random immunization is totally ineffective
- Lifetime is related to the network size

Rationalization of computer virus data

NEXT STEPS

Short and mid-term projects

- SIR epidemic outbreaks (non-closed population)
  May & Lloyd PRE 2001;
  Moreno, Pastor-Satorras &Vespignani EPJB (2002);
  Newman cond-mat (2002)
- Latency effects and population heterogeneity
- Modeling and simulations on real Internet maps
- Optimal immunization schemes for real maps
- Finite network effects
- Pollution of food-webs

References :

- R.Pastor Satorras and A.Vespignani, PRL 86, 3200 (2001)
- R.Pastor Satorras and A.Vespignani, PRE 63, 066117 (2001)
- R.Pastor Satorras, A. Vazquez and A.Vespignani, PRL 87, 258701 (2001)
- R.Pastor Satorras and A.Vespignani, PRE 65, 036104 (2002)
- Y. Moreno, R.Pastor Satorras and A.Vespignani, cond-mat/0107267 (2001)
- A. Vazquez, R.Pastor Satorras and A.Vespignani, cond-mat/0112400 (2001)

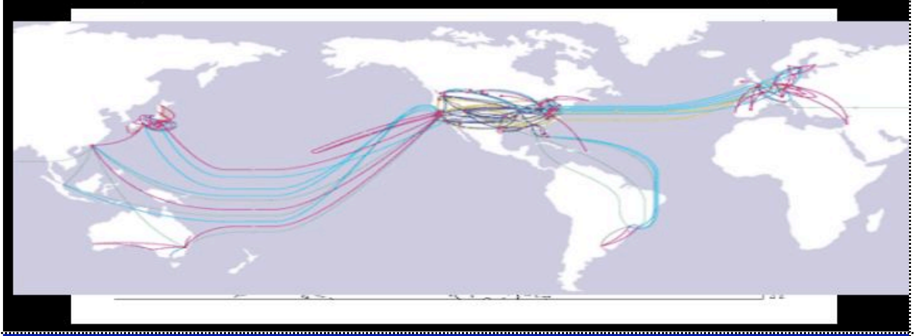
Long-term projects

•Relevance of the new epidemic framework in human epidemiology.
In particular sexually transmitted human diseases.

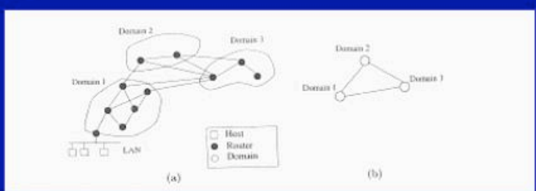
•Characterization of real Internet maps
- •Topology
- •Connectivity
- •Correlation properties
- •Hierarchical structure

(CAIDA, NLANR, INFN)

How the internet looks like



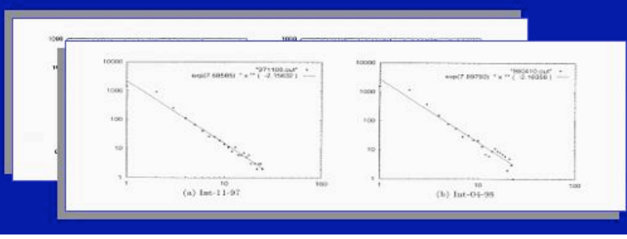
Graph representation



<u>This happens at both domain and router server</u>

•$P(k)$ = probability that a node has k links

Faloutsos et al. (1999)



How to generate scale-free graph

Growth : at each time step a new node is added with *m* links to be connected with previous nodes

Preferential attachment: The probability that a new link is connected to a given node is proportional to the number of node's links.

(by Barabasi & Albert (1999))

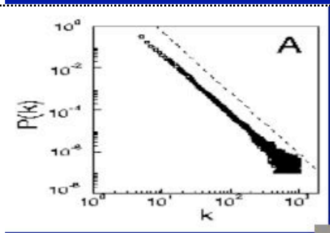
The BA model

The preferential attachment is following the probability distribution :

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j},$$

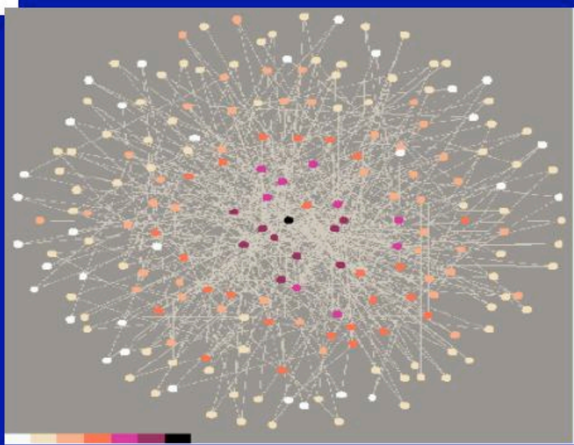
The generated connectivity distribution is

$$P(k) \sim k^{-3}$$



Connectivity distribution

BA network



Computer viruses timeline

- **1986** JUST A CURIOSITY
- MS-DOS 3.2 (top-line processor 386)

First virus created in Pakistan ("Brain").

BOOT sector virus = spreads via infected applications but copies itself in the boot sector/ immune to reboot

Lab experiment creates "Virden" in Germany

File virus = it infects the computer running a specific application

- **1987** IN THE WILD
- Windows 2.0 is released

"Brain" is discovered in the wild in Delaware. "Jerusalem" makes its appearance. First outbreak. "Stoned" and "Vienna", viruses written by high school students, appear. A book with a disassembly of "Vienna" is published becoming a source code for many other viruses

- **1988/1989** THEME VARIATIONS
- (Top-line processor 486)

"Cascade" virus (encrypted).
"Ping-Pong" virus (large outbreak in Italy)
Starts research on antivirus products

Dark Avenger in Sophia delivers the "1800" virus (real danger)

Computer viruses timeline

•1990 IT'S WAR and MONEY

•Windows 3.0

•Viruses get stronger (stealth, armoring, multipartite)

•IBM starts the "High Integrity Laboratory".
Anti-virus software houses.

•1991/93 MEDIA PANIC and PROLIFERATION

• Windows 3.1 and notebook

•Dark Avenger announces the release of a mutant virus (Mutation engine) "polymorphic".
•Virus construction sets appear.
•"Michelangelo" appears in the wild and hits the news!!
•First official wild-list with 100 viruses

•1994 INTERNET OUTBREAK

•A virus called "Kaos4" is posted on the alt.binaries.pictures.erotica news group.The file is called Sexotica and downloaded by a large number of users in few hours Small but very fast epidemics

Computer viruses timeline

•1995 NEW CONCEPTS

•Windows 95 is released

"Concept" the first virus written in WordBASIC.
Macro virus =infects data files and wordprocessors
They are platform independent!!!
All virus scanners fail the detection of "Concept"

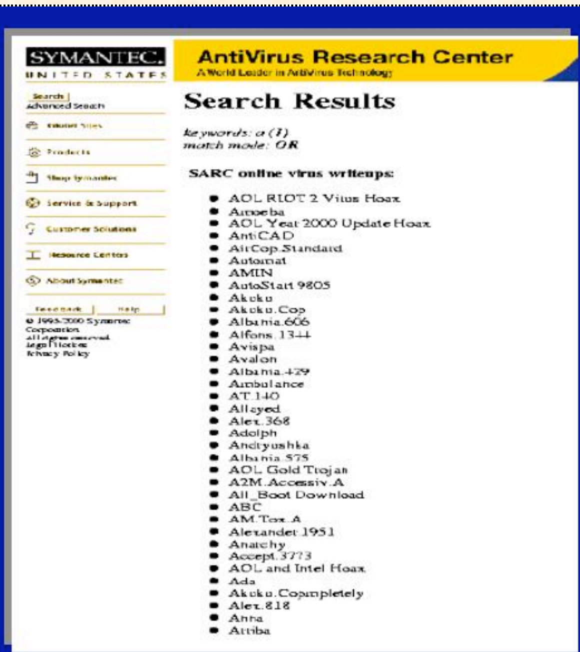
•1997 MACRO STRAIN TAKES OVER

•Pentium II

• 1000 Macro viruses identified

•2000 NOWADAYS

•Virus List with 48000 different viruses
•Wild-List with more than 1000 viruses
•I-LOVE-YOU causes $8 Billion in damage



SYMANTEC.
UNITED STATES

AntiVirus Research Center
A World Leader in AntiVirus Technology

Search | Advanced Search
Products
Shop Symantec
Service & Support
Customer Solutions
Resource Centers
About Symantec
Feedback | Help

W95.MTX

Discovered on: August 17, 2000
Last Updated on: October 16, 2000 04:25:59 PM PST

W95.MTX has a virus component and a worm component. It propagates using email. Also it infects some Win32 executables in specific directories. The virus also has the capability to block access to certain web sites. This may prevent users from downloading new virus definitions.

Click here to download tool to repair W95.MTX damage

Also known as: W95.Cisdbo, W95.MTX.dr, W95.MTX (.dll)

Category: Worm, Virus

Infection length: 9250 (variable)

Virus definitions: August 28, 2000

Threat assessment:

Wild

- Number of infections: 50-999
- Number of sites: More than 10
- Geographical distribution: High
- Threat containment: Moderate
- Removal: Difficult

Damage

- Payload:
  - Modifies files: *Some* infected files are corrupted beyond repair.

Distribution

- Subject of e-mail: None
- Name of attachment: Variable (see below)
- Size of attachment: Variable
- Target of infection: Windows executables
- Time stamp of attachment: Immediately after a new email message is sent, a second message is sent with no subject and the worm attached.

Technical description:

Worm component
The worm component makes a copy of Wsock32.dll and names it Wsock32.mtx. The *Send* export function of this .mtx file is then modified to point to its own code. This allows the virus to mail a copy of the worm infected with this virus to the same person to whom the user sends an email (using the same program).

Here are a list of file names that this virus might use when it sends the infected worm to other people. For those files with .pif extensions, the .pif extension might not be visible in your mail program.

I_wanna_see_you.txt.pif
Matrix_screen_saver.scr
Love_letter_for_you.txt.pif
New_playboy_screen_saver.scr
Bill_gates_piece.jpg.pif
Tiazinha.jpg.pif
Feiticeira_nua.jpg.pif
Geocities_free_sites.txt.pif
New_napster_site.txt.pif
Metallica_song.mp3.pif
Anti_cih.exe
Internet_security_forum.doc.pif
Alanis_screen_saver.scr
Reader_digest_letter.txt.pif
Win_$100_now.doc.pif
Is_linux_good_enough!.txt.pif
Qi_test.exe
Avp_updates.exe
Seicho_no_ie.exe
You_are_fat!.txt.pif
Free_xxx_sites.txt.pif
I_am_sorry.doc.pif
Me_nude.avi.pif

A Fresh look at the data

Up | Home page

WildList Index

Home
VB 100% Award
Subscriptions
Conference
Virus Hoaxes
Project VGrep
Prevalence Tables
WildLists
AV Links
Contact Us

Joe Wells has, for a few years now, been collecting virus reports from anti-virus experts around the world. He combines these to produce the WildList, a list of those viruses currently in the wild.

In recent times, the list has started to be used by Virus Bulletin and other anti-virus product testers as the definitive guide to the the viruses found in the real world.

An anti-virus product is expected to score 100% detection against this group of viruses.

In addition to viewing the lists here, you may download individual WildLists (in ZIP or text format) from links at the top of each list's page.

http://www.virusbtn.com

VB Prevalence Table, March 2000

Virus Name	Type	Number of Incidents	Percentage
Win32/Pretty	File	200	19.01%
Win32/Ska	File	131	12.45%
Kak	Script	118	11.22%
Marker	Macro	102	9.70%
Laroux	Macro	94	8.94%
Freelinks	Script	80	7.60%
Ethan	Macro	56	5.32%
Class	Macro	37	3.52%
Pri	Macro	22	2.09%
Thus	Macro	21	2.00%
Story	Macro	17	1.62%
Win32/ExploreZip	File	17	1.62%
Win95/CIH	File	16	1.52%
Myna	Macro	14	1.33%
Cap	Macro	12	1.14%
Melissa	Macro	11	1.05%
Win32/Fix	File	9	0.86%
Titch	Macro	8	0.76%
ColdApe	Macro	7	0.67%
Proverb	Macro	6	0.57%
Chack	Macro	5	0.48%
Form	Boot	5	0.48%
Locale	Macro	4	0.38%
Tristate	Macro	4	0.38%
AntiCMOS	Boot	3	0.29%
Niceday	Macro	3	0.29%
Smack	Macro	3	0.29%
Broken	Macro	2	0.19%
Divi	Macro	2	0.19%
Empire Monkey	Boot	2	0.19%