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)
Alexel Vazquez (SISSA-ICTP)

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 $\lambda$ if connected to one or more infected nodes.

Infected nodes are recovered (cured) with rate $\delta$ without loss of generality $\delta = 1$ (sets the time scale).

Definition of an effective spreading rate $\lambda = \delta / \delta$

- re-infection is possible.

Dynamical Mean-Field equation for the order parameter $\rho$ - density of infected nodes

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

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 $\rho$ = 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.

- CAIDA
- INLAN
- Mercker project
- IIPM
- Bell Labs
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

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

Scale-free properties

Diverging fluctuations

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

Data carried viruses
- ftp, file exchange, etc.

Internet topology

E-mail network

Computer worms
- e-mail diffusing
- self-replicating

The web of Human sexual contacts
- [L. Bijeos 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: ROOT viruses

Real data from viruses in the wild

- Prevalence data from large monitored samples

- Just a few viruses are lucky enough to prevail (sub-critical or very close to criticality?)
- In the endemic case prevalence is always very small (p<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 problems in computer virus epidemiology.

Survival probability

\[ P(t) = \text{fraction of viruses still in the wild at time } t \text{ after their birth} \]

\[ P_0 \sim \exp(-t/\tau) \]

\[ \tau = \text{average lifetime (characteristic time) of the virus strain} \]

Average lifetime extremely long compared to the virus rates time scale

- Anti-virus software is delivered in a few hours after the first detection
- I Love You 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

- All virus strains share the same characteristics

- The MACRO strain (particularily) 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 $<k^2> \to \infty$

- Connectivity fluctuations must be included

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

$$\frac{\partial \rho_k(t)}{\partial t} = -\rho_k(t) + \lambda k [1 - \rho_k(t)] \Theta(\rho(t)),$$

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

- $\Theta$ is function of the average density of infected nodes
- Links point with higher probability to highly connected nodes

---

**Stationary state**

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

Mean field 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
The probability of pointing to an infected node is

\[ \Theta(\lambda) = \sum_k \frac{k P(k) \rho_k}{\sum s P(s)} \]

\( \rho_k \) 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

\[ \Theta(\lambda) = m\lambda \Theta(\lambda) \int \frac{1}{\sum k^3 1 + k\lambda \Theta(\lambda)} \]

which yields the solution

\[ \Theta(\lambda) = \frac{e^{-1/m\lambda}}{\lambda m^2} (1 - e^{-1/m\lambda})^{-1} \]

The order parameter equation is

\[ \rho = 2m^2 \lambda \Theta(\lambda) \int \frac{1}{\sum k^3 1 + k\lambda \Theta(\lambda)} \]

Obtaining

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

Epidemic threshold in scale-free networks

\[ \lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \]

\[ \langle k^2 \rangle \to \infty \]

\[ \lambda_c \to 0 \]

Order parameter behavior in an infinite system

\[ \rho = 2e^{-1/m\lambda} \]
\[ \rho = 2e^{-\frac{1}{m\lambda}} \]

**Numerical simulations in a BA network**

**Zoom in lin-log scale**

**Network sizes**
\[ N=10^4 \text{ to } N=10^7 \]

**Spreading of a virus starting from a localized seed**

**Surviving probability with the same definition used to analyze data.**

**Exponential behavior with characteristic time increasing with the network size**

**Time behavior of the prevalence (algebraic)**

\[ \lambda = 0.06 \]

**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

\[ \lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \]

\[ \langle k^2 \rangle \to \infty \]

\[ \lambda_c \to 0 \]

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

Order parameter behavior in an infinite systems

Finite size scale-free networks

\[ P(k) \sim k^{-\gamma} \exp\left(-\frac{k}{k_c}\right) \]

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 \( k_c N = 10^5 \)

\[ \lambda_c / \lambda_c^H < 10^{-1} \]

Immunization

Random immunization:

\( g = \) density of immune nodes

\[ \lambda \to \lambda (1 - g) \]

Epidemic dies if

\[ \lambda (1 - g) \leq \lambda_c \]

Regular or random networks

\[ \rho_g = \rho_c (g_c - g) / g_c \]

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**
**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, ERJ (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, PRE 86, 026117 (2001)
- R. Pastor-Satorras and A. Vespignani, PRE 63, 066117 (2001)
- Y. Moreno, R. Pastor-Satorras and A. Vespignani, cond-mat/0107267 (2001)
**Long-term projects**


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

(CAIDA, NLANR, INFN)

**How the internet looks like**

**Graph representation**

- $P(k)$ - 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:

\[
\mathcal{P}(k_i) = \frac{k_i}{\sum_j k_j}
\]

The generated connectivity distribution is

\[ P(k) \sim k^{-3} \]

Connectivity distribution

BA network

**Computer virus 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 "Virlen" 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 "1890" virus (real danger)
- **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 “Kanot” is posted on the alt.binaries.pictures.erotic news group. The file is called Sexotica and downloaded by a large number of users in few hours Small but very fast epidemics

- **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
Winter School on Complex Systems – Alessandro VESPIGNANI

**Antivirus Research Center**

**W95.MTX**

**Description:**
W95.MTX has a virus component and a worm component. It propagates using email. Also it often uses Word macros in specific documents. The virus can be triggered to block email to certain addresses. This can prevent network access.

**CBR:**
How to download and remove W95.MTX

**Also known as:**
W95.Gater, W95.MTXa, W95.MTXd, W95.MTXc, W95.MTXa, W95.MTXd, W95.MTXc

**Category:** Virus, Worm

**Infection length:** 5495 (not calculated)

**Threat assessment:**
- Threat level: High
- Damage: Medium
- Distribution: High

**Features:**
- Number of infections: 50-999
- Number of infected computers: 10
- Geographic distribution: High
- Removal: Difficult

**Definition:**
August 14, 2000

---

**Description:**
- **Subject:** A malware file is sent to the user.
- **Malicious file:** The file is sent in an email attachment.
- **Target:** Windows executables
- **Behavior:**,

**Technical description:**
The malware component sends a copy of W95.MTX and removes it.

**Worm component:**
The worm component sends a copy of itself to other infected files and then modifies the process to point to itself. The worm attaches the worm to each copy of the attachment and runs with it.

Technical description:
- **Subject:** A file is sent in an email attachment.
- **Malicious file:** The file is sent in an email attachment.
- **Target:** Windows executables
- **Behavior:** The worm attaches a copy of itself to each attachment and runs with it.

**Technical description:**
The worm component sends a copy of itself to other infected files and then modifies the process to point to itself. The worm attaches the worm to each copy of the attachment and runs with it.

---

**A Fresh look at the data**

**WildList Index**

- [http://www.virusbfa.com](http://www.virusbfa.com)
<table>
<thead>
<tr>
<th>Virus Name</th>
<th>Type</th>
<th>Number of Infecteds</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wau-235Paarty</td>
<td>Iles</td>
<td>500</td>
<td>19.21%</td>
</tr>
<tr>
<td>Wau-235Paarty</td>
<td>Iles</td>
<td>121</td>
<td>12.15%</td>
</tr>
<tr>
<td>E718</td>
<td>Iles</td>
<td>118</td>
<td>11.72%</td>
</tr>
<tr>
<td>Marcus</td>
<td>Mature</td>
<td>105</td>
<td>4.00%</td>
</tr>
<tr>
<td>E874</td>
<td>Mature</td>
<td>93</td>
<td>7.10%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>80</td>
<td>2.50%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>26</td>
<td>2.72%</td>
</tr>
<tr>
<td>Dian</td>
<td>Mature</td>
<td>74</td>
<td>2.51%</td>
</tr>
<tr>
<td>E24</td>
<td>Mature</td>
<td>22</td>
<td>0.70%</td>
</tr>
<tr>
<td>Dian</td>
<td>Mature</td>
<td>34</td>
<td>2.60%</td>
</tr>
<tr>
<td>Bana</td>
<td>Mature</td>
<td>11</td>
<td>1.09%</td>
</tr>
<tr>
<td>Bana</td>
<td>Mature</td>
<td>17</td>
<td>1.83%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>17</td>
<td>1.07%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>5</td>
<td>0.52%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>3</td>
<td>0.17%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>7</td>
<td>0.05%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>6</td>
<td>0.05%</td>
</tr>
<tr>
<td>Wau-201Paarty</td>
<td>Iles</td>
<td>9</td>
<td>0.09%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>5</td>
<td>0.17%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>4</td>
<td>0.12%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>5</td>
<td>0.16%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>3</td>
<td>0.10%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>2</td>
<td>0.09%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>2</td>
<td>0.09%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>2</td>
<td>0.09%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>5</td>
<td>0.14%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>3</td>
<td>0.11%</td>
</tr>
<tr>
<td>Ghan</td>
<td>Mature</td>
<td>2</td>
<td>0.09%</td>
</tr>
</tbody>
</table>