PENNSTATE

Worm Spread in Scale-Free Networks

A Model Using Random Graph Theory

PRESENTED TO:
CSIIR Workshop
Oak Ridge National Lab
PRESENTED BY*:
Christopher Griffin
Penn State Applied Research

*Richard R. Brooks of Clemson University contributed to this study.

1

PENNSTATE

Goals of Presentation

- Summarize the epidemiological models of worm spread in the Internet
- Introduce *Random Graphs* as models of the Internet
- Propose a natural model of worm spread using *Random Graphs*
- Demonstrate quantitative results showing this model may be appropriate

2

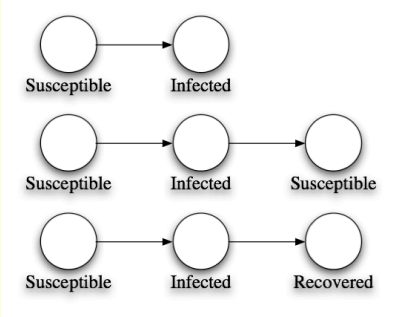
Computer Worms

- Computing a self-replicating program able to propagate itself across a network, typically having a detrimental effect.
- The name 'worm' comes from The Shockwave Rider, a science fiction novel published in 1975 by John Brunner.
- Researchers John F Shoch and John A Hupp of Xerox PARC chose the name in a paper published in 1982; The Worm Programs, Comm ACM, 25(3):172-180, 1982), and it has since been widely adopted.

3

Epidemic Models

- **Epidemiology:** The study of the spread of disease in populations.
- Diseases may spread quickly and then die out (Ebola) or remain endemic within a population (Chicken Pox)
- Populations can be modeled in a number of ways:
- “SI”, “SIS” or “SIR” models are most common.



4

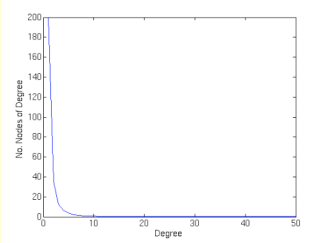
Mathematical Epidemiology

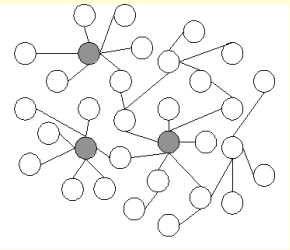
- Classical Mathematical Epidemiology Uses 3 Key Parameters in SIS/SIR Models:
  - R_0: The number of secondary infections that occur when **one** infective agent is introduced into a population.
  - $\sigma(t)$: The average number of effective contacts an individual has during his/her infected period.
  - R(t) : The average number of secondary infections produced by an individual during his/her infected period.
- In general, epidemic is only possible if $R_0 > 1$.

5

Scale-Free Networks

- A graph G=(V,E) is scale free if the number of vertices with degree *d* follows an inverse power law. That is:
- $n(d) = k/d^{\alpha}$
  - $n(d)$ is the number of vertices with degree d.
  - k is a constant of proportionality, and
  - α is the scaling parameter.
- Scale-free graphs have gained popularity in recent years.
- Examples: The World Wide Web, Human Sexual Contacts, Protein-Protein Interaction Networks.

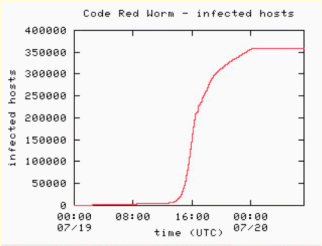


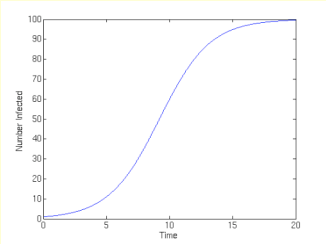


6

Worm Models with Epidemiology

- R. Pastor-Satorras and A. Vespigani studied the spread of worms in Internet-like networks using classical mathematical epidemiology.
  - Differential Equation Model of Infection Spread
  - Mean-Field Theory Approximations
- They show that for certain scale-free networks with scaling parameter < 3, epidemics will occur for all diseases with $R_0 > 1$.





7

OK, this model "looks" good. Why not use it?

- Three reasons to search for a different model:
  - These models assume a completely mixed population.
  - Classical mathematical epidemiology assumes a fluid-like behavior of individuals.
  - R. Pastor-Satorras and A. Vespigani were studying general scale-free networks, not computer networks specifically.
- Two dangers to note:
  - In the absence of *ad hoc* mesh networks, computers do not mix.
  - The effective R_0 is highly dependent on the initial infection position.

8

Graphs and Random Graphs

- A graph G=(V,E) is said to have a *giant component* H if H is a subgraph and contains a majority of the vertices of G.
- A random graph is a misnomer. A random graph is a tuple (Γ,*p*), where Γ is a set of graphs and *p* is an appropriately defined probability measure on a sigma algebra of Γ.
- The *most widely studied* random graph family is $\Gamma(n,p)$, where each graph in Γ has *n* vertices when any graph G is chosen from Γ the probability that there is an edge between two arbitrarily chosen vertices is *p*.
- These are the *Erdös-Renyi* Random Graphs.

9

Random Graph Model of SF Graphs

- Aiello et al. have formulated a random graph model of SF graphs.
- Let $\Gamma(\alpha,\gamma)$ be the collection of graphs whose degree distribution follows the curve $n(d)=\lfloor \exp(\alpha)/d^{\gamma} \rfloor$.
  - Here $\lfloor x \rfloor$ denotes the greatest integer lower bound for *x*.
  - Aiello et al. have shown that this definition is mathematically sufficient and that a reasonable probability measure can be defined.
- In this model, α (roughly) controls the size of the graph while γ controls the scaling of the graph.

10

Relation to Epidemic Models

- **Lemma [Griffin & Brooks 2006]:** If G is an element of $\Gamma(\alpha,\gamma)$, and vertices of G are uniformly randomly *kept* with probability $0<p\leq 1$ to produce G', then a.s. G' has the same properties as G.
- **Theorem [Griffin & Brooks 2006]:** For any infection in graph $G\in\Gamma(\alpha,\gamma)$ with $\gamma>2$, and with nodes having susceptibility probability *p*, then for all time

$$R_0 = \sigma(t) = p\frac{\zeta(\gamma-1)}{\zeta(\gamma)}$$

11

Infection Potential

- **Theorem [Griffin & Brooks 2006]:** If $2<\gamma<\gamma_0$, and for any infectious agent with infection probability *p*, a.s. $\lim_{t\to\infty} i(t) = p$. Where $i(t)$ is the proportion of infected nodes.
- This result is particularly interesting:
  - Often the affects of Internet worms have been blamed on the monoculture of Microsoft products.
  - This theorem suggests that even in the absence of a network monoculture, for appropriate Internet structures, 100% infection would occur among the susceptible nodes.

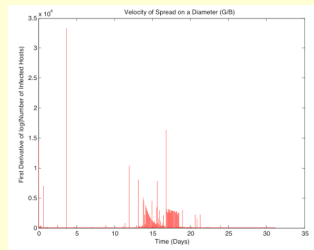
12

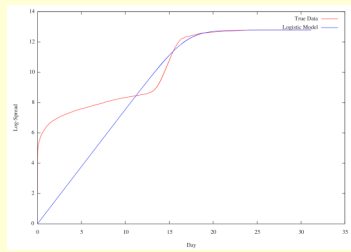
Rate of Infection

- **Theorem [Griffin & Brooks]:** Suppose that the rate of infection is constant, then the time required to achieve total infection is a.s. O(log|G|).
- Suppose that the infection rate is $r(t)$, then:

$$I(t) = \exp\left(\int_0^t r(t)dt\right)$$

- *For certain* $r(t)$ we can obtain an "S" curve matching the data.
- This gives a *natural* model of infection rate that matches the given data and does not appeal to continuous mixing models.



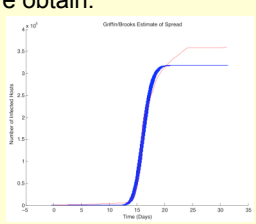


Comparison of Approaches

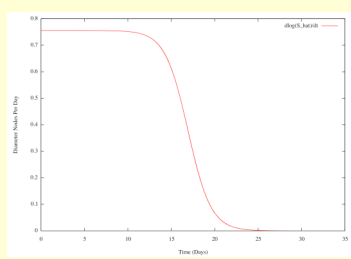
- When we try a model:

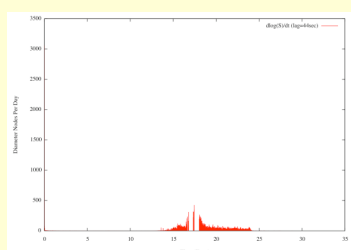
$$r(t) = \beta_1 \exp\left(\beta_2 (t-\mu)^2\right)$$

- We obtain:



- The model is seen to be imperfect because the true "logarithmic rate" does have hump, but it is probably not Gaussian in nature.
- G/B puts the diameter of this monitored network at ~13--this is a bit smaller than most estimates of the diameter of the Internet.





Infection Countermeasures

- **Theorem [Griffin & Brooks 2006]:** Centralized patch distribution runs in O(|G|), while decentralized "white worms" can inoculate machines in time O(log|G|) assuming a constant rate of transmission.
- This theorem was "experimentally verified" by Chen and Carley (2005).
- What does this mean?
  - Centralized patch distribution is inefficient but…
  - Centralized patch distribution is safe.
  - Inefficiency is the cost of safety.
- Here is a real tradeoff: either we distribute patches quickly and prevent global infection at the risk of creating patch-based errors or we live with our current security model.

15

Conclusions

- Infectious agents in computer networks can be modeled using natural "random graph" models.
- These models are more appropriate than continuous mixing models.
- For scale-free random graph models, total infection is a.s. whenever $\gamma<\gamma_0$, hence infections are a function of network structure as much as pathogen.
- Infection rates can be well described using the random graph model.
- There is a natural trade-off between security countermeasures efficiency and safety. This confirms experimental results presented by Chen and Carley.

16