Coping with Instant Messaging Worms - Statistical Modeling and Analysis

Zhijun Liu and David Lee

Department of Computer Science and Engineering,

The Ohio State University

THE OHIO STATE UNIVERSITY

Outline

- What is IM worm?
- Statistical IM Worm Modeling
- IM Worm Modeling Analysis
- Discussion of IM Worm Defense Approaches
- Conclusion

What is IM worm?

- Instant Messaging (IM)
  - MSN, Yahoo!, AIM, Jabber, Skype…
- IM Worm = A self-replicating malicious code that propagates over IM networks
  - “Bropia@MSN”, “Kelvir@MSN”, “Velkbot@ AIM/MSN/ YAHOO”, “MyDoom.AL@AIM”, …
  - Two weeks ago, a new worm called “SendPhoto.a @MSN”….
- By IMlogic Threat Center, IM and P2P exploits exploded in 2005, and have grown 50% each month

IM Worm vs. Internet/Email Worm

- Internet Worm
  - Target discovery: Scanning/Pre-generated target list
- Email Worm
  - Target discovery: Email Address Book
  - Email message delivery: Connectionless-oriented, non-real time
  - User presence: not aware of user presence
- IM Worm
  - Target discovery: Buddy List
  - IM message delivery: Connection-oriented, real time, instance user response
  - User presence: rich presence information (online or offline, etc.)

Existing Worm Modeling Techniques

- Deterministic epidemic disease propagation modeling
  - Biological epidemiology based worm model [KW93]
  - SIS/SIR, Kermack-Mckendrick [ZGT02]
  - P2P virus modeling [TC06]
- Discrete time worm model
  - Analytical active worm propagation [CGK03]
  - E-mail worm model [ZTG04]
- Stochastic process based modeling
  - Worm modeling based on interactive Markov chains in small-world topology [GGT03]
  - Modeling and Automated Containment of Worms (Branching process worm modeling) [SSB05]

IM Worm Modeling - Assumptions & Symbols

- Model Assumptions
  - Worm only infect online users
  - Worm propagates immediately after infection
  - Worm re-infection

- Model Symbols
  - $G = <V,E>$: Directed graph representing IM network
  - User response time $1/\lambda_i$
  - User Presence probability Pst_i
  - Open probability Po_i
  - Re-infection factor d_f

IM Modeling based on Branching Processes with General Variable Lifetime - 1

Branching Process with General Variable Lifetime	IM Worm Propagation Process
A population of particles	A collection of infected nodes
Particle splits into a random number of new particles	Worm multicasts from infected nodes and infects a random number of victims
Particle lifetime	User response time

- Model Notations
  - *T: random variable* for user response time
  - *N(t):* number of infected nodes existing at time *t*
  - $P_k(t) = P\{N(t)=k\}$ for k=0,1,2…..

    $P_0(t)=0$ for all $t \leq 0$ and $P_1(t) = P\{N(t)=1\} = P\{T>t\}=1-F(t)$
  - *G(s,t):* probability generating function of *N(t)*

$$G(s,t) = \sum_{k=0}^{\infty} P_k(t)s^k = \sum_{k=1}^{\infty} P_k(t)s^k$$

IM Modeling based on Branching Processes with General Variable Lifetime - 2

- IM worm process with two online responsive buddies

$$P_k(t) = \int_0^t f(\tau) d\tau \sum_{l=1}^{k-1} p_l(t-\tau) p_{k-1-l}(t-\tau), k = 2,3...$$

$$G(s,t) = s\int_0^t [G(s,t-\tau)]^2 f(\tau) d\tau + [1-F(t)]s$$

- IM worm process with *r* online responsive buddies

$$G(s,t) = s\int_0^t [G(s,t-\tau)]^r f(\tau) d\tau + [1-F(t)]s$$

- IM worm process with random number of online responsive buddies

$$G(s,t) = s\int_0^t h(G(s,t-\tau)) f(\tau) d\tau + [1-F(t)]s \qquad h(s) = \sum_{l=0}^{\infty} q_l s^l$$

IM Modeling Analysis - 1

- Assume
  - Each user has same use response time distribution
  - Same open probability

- $P_k(t)$ and $E[N(t)]$ analysis

$$G(s,t) = s\int_0^t h(G(s,t-\tau))f(\tau)d\tau + [1-F(t)]s \qquad h(s) = \sum_{l=0}^{\infty} q_l s^l$$

$$m(t) = E(N(t)) = \frac{dG(s,t)}{ds}\Big|_{s=1} = \sum_{k=1}^{\infty} kP_k(t)$$

$$m(t) = \frac{dG(s,t)}{ds}\Big|_{s=1} = 1 + E[l]\int_0^t m(t-\tau)f(\tau)d\tau$$

IM Modeling Analysis - 2

$$m(t) = 1 + E[l]\int_0^t m(t-\tau) f(\tau) d\tau$$

- Average user response time conforms exponential distribution $t \sim Exp(\lambda)$.

$$m(t) = \frac{1}{1-d} + \frac{e^{(d-1)\lambda t}}{1-1/d}$$

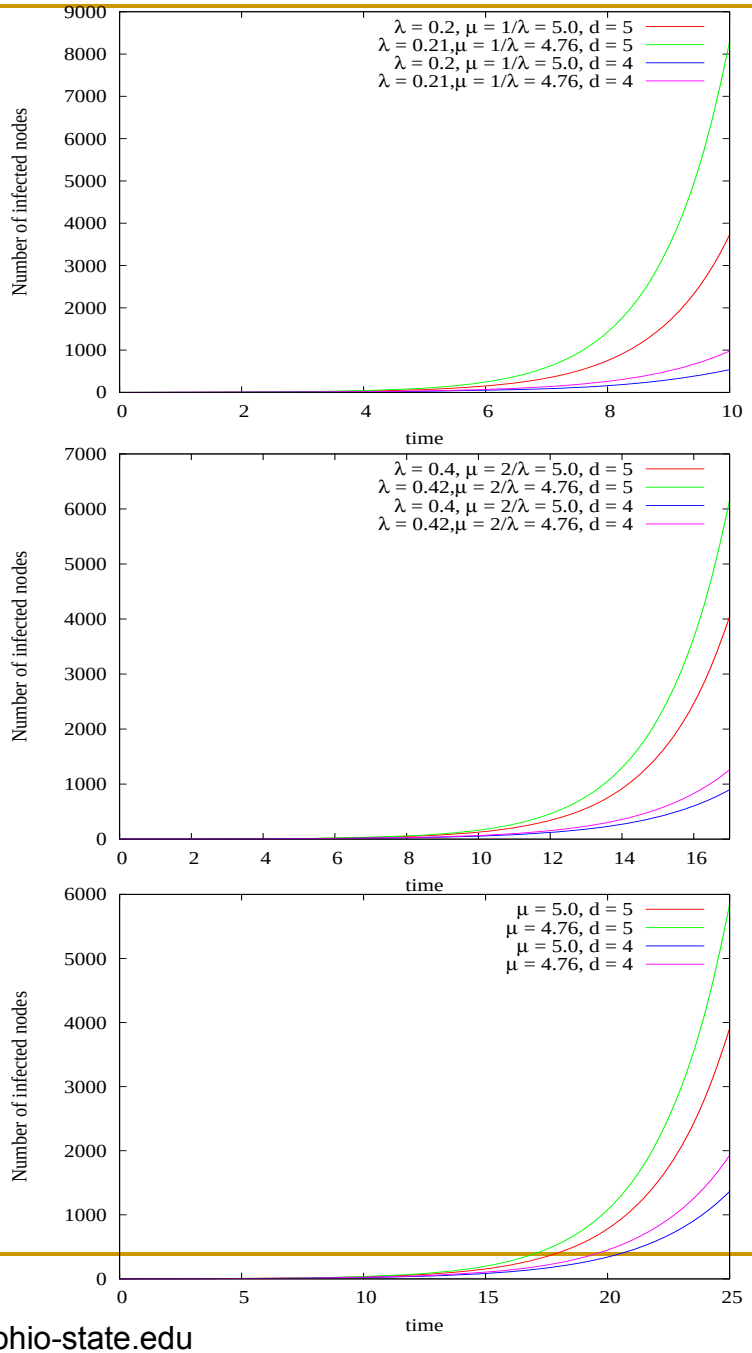
$$P_k(t) = e^{-\lambda t}(1-e^{-\lambda t})^{\frac{k-1}{2}} \qquad (d=2)$$

- Average user response time conforms Erlang distribution $t \sim Erlang(\lambda, 2)$.

$$m(t) = \frac{1}{1-d} + \frac{\sqrt{d}\, e^{(-\lambda+\lambda\sqrt{d})t}}{2\sqrt{d}-2} + \frac{\sqrt{d}\, e^{(-\lambda-\lambda\sqrt{d})t}}{2\sqrt{d}+2}$$

- User periodically checks IM message (average check period μ)

$$m(t) = \frac{d^{t/\mu+1}-1}{d-1}$$



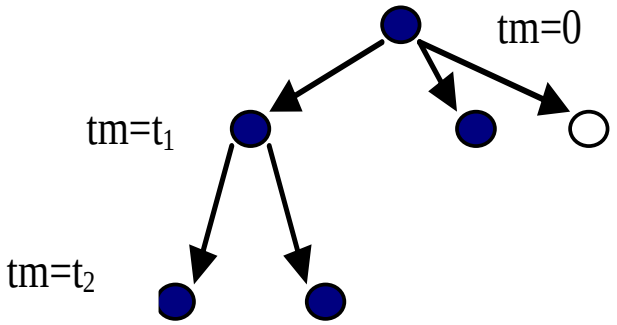
IM Modeling Analysis with $t \sim Exp(\lambda)$ - asymptotically constant

- τ_n: the time of the n^{th} worm multicast in the population
- T_i : the time between the $(i-1)^{th}$ and i^{th} multicasts (splits)
- S_i : the number of infected nodes in the worm population at time τ_i
- $Y_n = \sum_{i=1}^{n}(T_i - \frac{1}{\lambda S_{i-1}})$ is a Martingale $(E[Y_{n+1}|Y_0, Y_1, \ldots, Y_n] = Y_n)$
- $\tau_{2n} - \tau_n \cong \frac{\log 2}{(d-1)\lambda}$ is asymptotically constant

IM Modeling Analysis with $t \sim Exp(\lambda)$ - worm multicast event tree analysis

- Worm multicast event tree with leveled Gamma distribution
  - $t_1 \sim Exp(\lambda)$ and $t_i \sim Gamma(i, \lambda)$, i.e. $t_i \sim Erlang(i, \lambda)$
  - Given a window of size W, CDF is $P(t_i \le W)$, denoted by P_i
  - Traffic analysis:

$$M_w = 1 + \sum_{i=1}^{\infty} N_i = 1 + \sum_{i=1}^{\infty} (P_i \times d^i)$$



- Worm multicast event tree distribution
  - Worm propagation as a Poisson process with branches at each event

$$P(n=r) \approx \sum_{\sum_{j=1}^{k} l_j = r \wedge l_0 = 1 \wedge l_{k+1} = 0 \wedge l_j \le l_{j-1} \times d} \left(\sum_{i=1}^{k+1} C_{l_{i-1} \times d}^{l_i} \times \left(\frac{P_i}{P_{i-1}}\right)^{l_i} \times \left(1 - \frac{P_i}{P_{i-1}}\right)^{l_{i-1} \times d - l_i} \right)$$

Discussion on IM Worm Defending Approaches

- Existing IM worm defending techniques
  - Anti-virus protections
  - Temporary Server Shutdown [HC03]
  - Temporarily Disabling the Most Connected Users [S02]
  - Virus throttling [WP04, MV05]
  - Traditional count-based or trend-based worm detection approaches […]
- Proposed worm defense approaches
  - Worm multicast tree monitoring based detection and containment
  - Clustering based worm throttling

Conclusion

- Defending against IM worm poses new challenges
- A general IM worm statistical modeling based on branching process with variable lifetime
- Model analysis:
  - Study the impact of user response behavior on IM worm propagation of speed
  - Martingale analysis
  - Worm multicast event tree analysis.
- Discussion of IM worm defense approaches

Future Work

- Smart Worm Modeling and Defense
- Investigation IM Worm under P2P based IM Systems

Thank you.

liuzh@cse.ohio-state.edu