An Epidemiological Model of Internet Worms with Hierarchical Dispersal and Spatial Clustering of Hosts

David E. Hiebeler^{a,*}, Andrew Audibert^{a,b}, Emma Strubell^{c,d}, Isaac J. Michaud^{a,e}

^a*Department of Mathematics and Statistics, University of Maine, Orono, ME 04469*

^b*Bangor High School, Bangor, ME 04401*

^c*School of Computing and Information Science, University of Maine, Orono, ME 04469*

^d*College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003*

^e*Department of Statistics, North Carolina State University, Raleigh, NC 27695*

Abstract

Beginning in 2001, many instances of malicious software known as Internet worms have been using biological strategies such as hierarchical dispersal to seek out and spread to new susceptible hosts more efficiently. We measured the distribution of potentially susceptible hosts in the space of Internet addresses to determine their clustering. We have used the results to construct a full-size simulated Internet with 2^{32} hosts with mean and variance of susceptible hosts chosen to match our measurements at multiple spatial scales. Epidemiological simulations of outbreaks **among the roughly 2.8×10^6 susceptible hosts** on this full-sized network show that local preference scanning greatly increases the chances for an infected host to locate and infect other susceptible hosts by a factor of as much as several hundred. However, once deploying this strategy, the overall success of a worm is relatively insensitive to the details of its dispersal strategy over a wide range of parameters. In addition, although using localized interactions may allow malicious software to spread more rapidly or to more hosts on average, it can also lead to increased variability in infection levels among replicate simulations. Using such dispersal strategies may therefore be

*Corresponding author

Email address: hiebeler@math.umaine.edu (David E. Hiebeler)

URL: <http://www.math.umaine.edu/~hiebeler> (David E. Hiebeler)

a high risk, high reward strategy for the authors of such software.

Keywords: network models, malicious software

1. Introduction

Malicious software such as worms spreading through computer networks is a routine problem faced by all computer or network users and administrators. Attacks by such malware can cost billions of dollars per incident and spread around the world in a matter of hours or even minutes (Moore et al., 2002, 2003a; Albanese et al., 2004).

Since the emergence of Internet worms in the late 1980s, many attempts have been made to characterize their proliferation. Just a few years after Morris, the first computer worm, infected thousands of university and corporate machines (Eichin and Rochlis, 1989), IBM sponsored a comprehensive study on the dynamics and dangers of these Internet infections. One of the earliest applications of models from biological epidemiology to the spread of Internet worms (Kephart et al., 1993) found that the simplifying assumption of homogeneous contact used in many traditional epidemiological models greatly affected the dynamics of simulated outbreaks. Network topology is a major factor in the spread of worms, whether through peer-to-peer file sharing networks, email contacts, or the IP addressing system. Just as the spatial distribution of a population changes the dynamics of an epidemic (Duryea et al., 1999), so too can the topological distribution of computer hosts on a network have a major impact on the success of an Internet worm. Random graphs were used to simulate the small software sharing networks that served as vectors for early worm infections (Kephart and White, 1991, 1993), and small-world social networks have been used to model the spread of email worms (Moore and Newman, 2000; Zou et al., 2004). Connectivity of some social networks as well as the Internet as a whole has also been described using scale-free networks. Consequently, a great deal of work has been done regarding the spread of disease, both biological and computer-based, on scale-free networks (Pastor-Satorras and Vespignani,

28 2001; Albert and Barabási, 2002; Newman, 2002; Balthrop et al., 2004; Pastor-
29 Satorras and Vespignani, 2004; Zou et al., 2004; Barthélemy et al., 2005; Doyle
30 et al., 2005; Hwang et al., 2005). Others have looked at hierarchical (Wang
31 et al., 2000; Grabowski and Kosiński, 2004) and random (Keeling, 2005) net-
32 work models with clustering, but not in conjunction with the topology of the
33 **Internet Protocol version 4 (IPv4)** Internet.

34 In the past 15 years, Internet worms have begun using biologically-inspired
35 strategies for seeking out hosts to invade (Chen and Ji, 2005; Avlonitis et al.,
36 2007). These tactics exploit inherent vulnerabilities in the structure of the
37 Internet to spread more swiftly and efficiently than their predecessors (Zou
38 et al., 2006).

39 In particular, a technique known as *local preference scanning* takes advantage
40 of the tendency for machines that are nearby in IPv4 address space to make up
41 a homogeneous network maintained by the same administrators and thus share
42 the same software and vulnerabilities. To quickly infect clusters of susceptible
43 hosts, local preference scanning worms choose new targets within the same
44 local network (“subnet”) as their infected host machine more often than they
45 choose a machine completely at random. That is, they use localized dispersal
46 when choosing new hosts to contact for attempted infection. This strategy
47 for propagation typically results in persistent and widespread infections while
48 entrenching extensive invasion within protected networks once a single host there
49 has been compromised (Nazario, 2004; Pastor-Satorras and Vespignani, 2004).
50 For example, one version of the Code Red worm infected more than 350,000
51 computers within 14 hours, for a time infecting more than 2,000 new hosts per
52 minute (Moore and Shannon, 2001).

53 Our study is based on the IPv4 addressing scheme, in which each IP address
54 is a 32-bit number commonly expressed in the form $w.x.y.z$, where each of the
55 four values w , x , y , and z is between 0 and 255 inclusive. (These values are
56 often referred to as *octets*, as they are 8-bit values.) Although there are $256^4 \approx$
57 4.3×10^9 possible IP addresses, many of them are reserved for various special
58 purposes. The space of IP addresses can be thought of as a $256 \times 256 \times 256 \times 256$

59 four-dimensional lattice. Note that there is not a one-to-one correspondence
60 between active IP addresses and computer hosts; a single computer host may
61 have multiple IP addresses, and it is very common for a single public IP address
62 to have many computer hosts “behind” it (such as multiple devices in a home
63 sharing a network router with a single publicly-visible IP address).

64 In summer 2001, Code Red II, hereafter simply Code Red, was the first
65 worm to successfully employ a local preference scanning strategy. Although
66 this paper focuses on the specific strategy used by this worm, many subsequent
67 worms have used similar methods to spread. Code Red’s strategy was as follows
68 (Moore et al., 2002). When a computer with IP address $w.x.y.z$ is infected and
69 attempts to contact a new host to attempt infection:

- 70 1. With probability 0.375, it performs short-range dispersal, generating ran-
71 dom IP addresses of the form $w.x.r_1.r_2$ where r_1 and r_2 are randomly
72 chosen between 0 and 255. These addresses are referred to as being “in
73 the same /16 subnet” as that of the infected host (they share the same
74 first 16 bits with the source address).
- 75 2. With probability 0.5, the worm performs medium-range dispersal, gen-
76 erating random IP addresses of the form $w.r_1.r_2.r_3$, with r_1 , r_2 , and r_3
77 randomly chosen between 0 and 255. These addresses are referred to as
78 being in the same /8 subnet as that of the infected host (they share the
79 same first 8 bits with the source address).
- 80 3. With probability 0.125, it performs long-range dispersal, generating an IP
81 address $r_1.r_2.r_3.r_4$ consisting of four randomly generated values, avoiding
82 addresses of the form $127.x.y.z$ and $224.x.y.z$, corresponding to subnets
83 reserved for loopback and multicast routing, respectively.

84 In April and May 2004, the Sasser worm utilized a similar dispersal strategy with
85 a slightly different probability distribution. **In August 2003, the Blaster**
86 **worm used a combination of dispersal strategies. It would first use**
87 **a variation of medium-range dispersal, and then after 20 connection**
88 **attempts, it would switch to long-range dispersal. Since that time,**

89 **worms generally use various types of hierarchical dispersal strategies**
90 **such as the above.**

91 To improve our understanding of how local preference scanning worms spread,
92 we have performed scans of web servers on the Internet to determine the extent
93 to which machines running certain server software and versions are clustered in
94 IPv4 address space. We use that information to partition a simulated population
95 of 2^{32} computer hosts into a hierarchy reflecting the observed structure imposed
96 by IPv4 addressing, and apply state-based epidemiological models to simulate
97 the dynamics of infected hosts **among the roughly 2.8×10^6 susceptible**
98 **hosts in the population.**

99 Different types of worms have been classified based on the way they identify
100 new hosts to infect (Staniford et al., 2002; Weaver, 2002) and their effectiveness
101 assessed (Vogt, 2004; Zou et al., 2006). Worms using simple random scanning
102 have been modeled a great deal due to their simplicity (Chen et al., 2003; Moore
103 et al., 2003b), while less has been done to study other methods of dispersion
104 such as preference scanning. Some have assumed a homogeneous distribution
105 of susceptible hosts throughout the Internet to model local preference scanning
106 worms (Chen et al., 2003), while others developed a model on a network made
107 of smaller networks of varying sizes (Avlonitis et al., 2007), but with no cor-
108 relation based on IP addresses alone. Because of their lack of an underlying
109 network topology, plant epidemiological models that incorporate multiple dis-
110 persal distances (Filipe and Gibson, 1998; Filipe and Maule, 2004) are difficult
111 to apply to preference scanning worms. Here we employ a hierarchical house-
112 hold/community model (Barbour, 1978; Dye and Hasibeder, 1986; Kotliar and
113 Wiens, 1990; Daley and Gani, 1994; Hiebeler et al., 2011) with clustering of
114 susceptible hosts as measured from scans of the Internet.

115 **In this model, the population is completely connected, in the sense**
116 **that any host can contact any other host. However, local preference**
117 **scanning leads to variability in the rates of contact among different**
118 **hosts depending on their location within the hierarchically structured**
119 **population.**

120 2. Model and Measurements

121 2.1. The Model

122 To apply the household epidemiological model to Internet worms, we parti-
123 tion a population of 2^{32} hosts into hierarchical groups reflecting the structure of
124 IPv4 addressing. Various other topologies, such as physical network connectiv-
125 ity, links between web sites, or social connections among users' e-mail address
126 books are also sometimes imposed by people or by worms and viruses as they
127 spread (Balthrop et al., 2004; Alderson et al., 2005). In our model, we simulate
128 the spread of worms over the IPv4 system, since it was the addressing scheme
129 exploited by Code Red and many subsequent Internet worms, and is still the
130 most prevalent addressing system used.

131 Although the space of IPv4 addresses may be thought of as a 4-D $256 \times 256 \times$
132 256×256 integer lattice, Code Red and other worms using similar strategies
133 do not separately utilize the final dimension of the lattice. The population may
134 therefore be thought of as a collection of hosts in a 3-D $256 \times 256 \times 65536$
135 lattice, each characterized by address of the form $w.x.y$, where w and x are
136 between 0 and 255, while y is between 0 and $65535 = 256^2 - 1$. We describe the
137 population as containing 256 *neighborhoods* (representing /8 networks), each
138 of which contains 256 *households* (representing /16 networks), each of which
139 contains 65536 individual *hosts*. Here we make the simplifying assumption that
140 a given IP address corresponds to a single potential computer host.

141 To explore the dynamics of worm proliferation, we use an SIR-V epidemio-
142 logical model (Hiebeler et al., 2011) in which each host is in one of four states:
143 Susceptible, Infected, Recovered/Removed, or Vaccinated. A susceptible host is
144 one running software that can be exploited by the Internet worm in question; in
145 the case of Code Red, this would be a vulnerable version of Microsoft's Internet
146 Information Services (IIS). Infected hosts are those containing the worm of in-
147 terest actively trying to infect others. Recovery corresponds to the worm being
148 removed and the security hole patched. Here, once a host is recovered/resistant,
149 it remains so. We also explored an SIRS-V model in which resistant machines

150 slowly transition back to the susceptible state, representing turnover of ma-
 151 chines on the Internet as new (vulnerable) machines come on-line; qualitatively
 152 similar results were obtained from that model, and are not reported here. A
 153 vaccinated host corresponds to an IP address with a machine running software
 154 which cannot be infected by the worm in question (e.g., a different operating
 155 system), or which does not contain a machine. Both resistant and vaccinated
 156 hosts are immune to infection; the only difference is that a resistant host is one
 157 which was previously infected at some time.

158 In our model, each infected host contacts others and attempts infection at
 159 rate ϕ per unit time and recovers at rate μ , according to a Poisson process: the
 160 time until the next contact follows an exponential probability distribution with
 161 mean $1/\phi$, and similarly for the time until recovery. When contact is made,
 162 the target host is chosen at random from the entire Internet with probability
 163 α_0 , from the same neighborhood (/8 network) as the infected attacker with
 164 probability α_1 , and from the same household (/16 network) with probability α_2 ;
 165 these are referred to as long-, medium-, and short-distance dispersal respectively.
 166 The subscript i on α_i indicates how many leading octets of the originating host's
 167 IP address are preserved when generating a new address to contact. When
 168 a target is contacted, if it is susceptible, it immediately becomes infectious;
 169 otherwise, the contact has no effect.

170 Ordinary differential equations can be written to describe the proportions of
 171 hosts infected within given households, neighborhoods, or the entire population.
 172 Here we use deterministic differential equations, implicitly taking the mean over
 173 realizations of the stochastic process. Let I_{jk} be the proportion of hosts in
 174 household k , neighborhood j (i.e., hosts whose IP addresses begin with the
 175 octets $j.k$) that are infected, for $j, k \in \{0, 1, 2, \dots, 255\}$. Similarly, let S_{jk} , R_{jk} ,
 176 and V_{jk} be the proportions of susceptible, recovered, and vaccinated hosts. Let

$$I_{j\cdot} = \sum_{k=0}^{255} I_{jk} / 256 \quad (1)$$

177 be the proportion of hosts within neighborhood j that are infected, computed

178 by averaging among the 256 households within that neighborhood, and

$$I_{..} = \sum_{j=0}^{255} \sum_{k=0}^{255} I_{jk} / 65536 \quad (2)$$

179 be the proportion of hosts infected within the entire population, as averaged
 180 among all 65536 households. The quantities $S_{j.}$, $S_{..}$, $R_{j.}$, $R_{..}$, $V_{j.}$, and $V_{..}$ are
 181 similarly defined. For convenience, $U = S + I + R = 1 - V$ will refer to all
 182 unvaccinated hosts combined.

183 New infections may arise via long-, medium-, or short-distance dispersal.
 184 A differential equation describing the rates of these three types of newly arising
 185 infections along with the rate of recovery of infected hosts for a particular
 186 household is

$$\frac{dI_{jk}}{dt} = \phi \{ \alpha_0 I_{..} S_{jk} + \alpha_1 I_{j.} S_{jk} + \alpha_2 I_{jk} S_{jk} \} - \mu I_{jk}. \quad (3)$$

187 This differential equation may be averaged across all households to obtain a
 188 differential equation describing the dynamics of the infection level among the
 189 entire population

190

$$\begin{aligned} \frac{dI_{..}}{dt} = & \phi \left\{ \alpha_0 I_{..} S_{..} + \alpha_1 \sum_{j=0}^{255} I_{j.} S_{j.} / 256 \right. \\ & \left. + \alpha_2 \sum_{j=0}^{255} \sum_{k=0}^{255} I_{jk} S_{jk} / 65536 \right\} - \mu I_{..} \end{aligned} \quad (4)$$

191 The quantities I_{jk} , $I_{j.}$, and $I_{..}$ indicate the prevalence of infection at various
 192 spatial scales within the network. We also quantify the clustering of infected
 193 (or unvaccinated) hosts at two spatial scales, as follows. Following earlier
 194 work (Hiebeler, 2006), let $Q_{g,a|b}$ be the probability that if one randomly selects
 195 a host in state b from the population, then a second host selected at random
 196 within the same group g is in state a , where a and b are one of S , I , R , U ,
 197 V . The “group” may refer to a neighborhood or household. For example, the

neighborhood-level clustering of unvaccinated hosts is given by

$$Q_{n,U|U} = \frac{\sum_{j=0}^{255} (U_{j\cdot})^2}{256U_{\cdot\cdot}},$$

while the household-level clustering of unvaccinated hosts is

$$Q_{h,U|U} = \frac{\sum_{j=0}^{255} \sum_{k=0}^{255} (U_{jk})^2}{65536U_{\cdot\cdot}}.$$

The variance of the level of unvaccinated hosts among groups g (neighborhoods or households) may be expressed in terms of the global level and the corresponding clustering parameter, $\text{Var}_g[U] = U_{\cdot\cdot}(Q_{g,U|U} - U_{\cdot\cdot})$. The minimum variance of 0 occurs when the local clustering is the same as the global density, $Q_{g,U|U} = U_{\cdot\cdot}$ (all groups have the same proportion $U_{\cdot\cdot}$ of unvaccinated hosts), while the maximal variance of $U_{\cdot\cdot}(1 - U_{\cdot\cdot})$ occurs when $Q_{g,U|U} = 1$ (fraction $U_{\cdot\cdot}$ of groups are completely unvaccinated, while the rest are completely vaccinated).

The first-generation reproduction ratio $\mathcal{R}_0$ can be calculated for the above model as a function of the dispersal parameters,

$$\mathcal{R}_0 = (\phi/\mu) (\alpha_0 U_{\cdot\cdot} + \alpha_1 Q_{n,U|U} + \alpha_2 Q_{h,U|U}). \quad (5)$$

This quantity gives the expected number of secondary infections caused by a single randomly-chosen unvaccinated host (before it recovers) which has been made infectious in an otherwise-healthy population, taking into account the prevalence and clustering of vaccination. Note that the number of tertiary infections per secondary infection and so on will tend to increase over subsequent generations; details also depend on where the initial infection is released (Hiebeler et al., 2011).

2.2. Measurement

Over the past decade, we have scanned random samples of hosts on the Internet to estimate the prevalence ($U_{\cdot\cdot}$) and clustering ($Q_{n,U|U}$ and $Q_{h,U|U}$) of potentially susceptible hosts. For simplicity in our study, we assume that all servers running the currently-dominant version of IIS are susceptible, although

221 in practice, worms such as Code Red were only able to infect hosts running
222 particular unpatched versions of IIS.

223 Measurements were taken during January 14–16, 2002, January 15–19, 2005,
224 and January 7–11, 2010. Each time, we generated **approximately** 10^6 unique
225 random IP addresses, excluding the reserved addresses **127.0.0.0/8, 224.0.0.0/8,**
226 **172.16.0.0/12, 10.0.0.0/8, 192.168.0.0/16,** and addresses **for which the final**
227 **(fourth) octet is 0 or 255. Note that Code Red only avoided ad-**
228 **dresses of the first two forms. To keep their code simple and small**
229 **(which allows them to spread to a new host more quickly and with a**
230 **smaller chance of their network traffic being detected), many worms**
231 **use fairly simple strategies for excluding reserved addresses and sim-**
232 **ply live with the slightly increased failure rate when seeking new hosts**
233 **to infect.** For each address, an attempt was made to contact a web server at
234 that address on HTTP port 80. If contact was established, a GET request was
235 made for the “robots.txt” file, any response was recorded, and the connection
236 closed. We allotted a timeout period of two minutes for the initial contact,
237 and thirty seconds for a response after sending the GET request. When a web
238 server responds to a request, it also provides information about what software
239 and version the server is running; this information was recorded for each web
240 server we found. Although easily spoofed, we assume the number of servers
241 reporting incorrect or no information is relatively small. The most common
242 version of IIS was 4.0 in 2002, 5.0 in 2005, and 6.0 in 2010.

243 To measure clustering, during each scan, a set of 249 IP addresses of comput-
244 ers running the then-dominant version of IIS was chosen from the computers
245 which responded during the measurement of the random sample above (this
246 number was chosen because in the initial scan in 2002, we found 249 machines
247 running the dominant version of IIS). **The version of IIS running on the**
248 **hosts at these 249 IP addresses is referred to as the “Source” version**
249 **in Table 1. For each address within the set of 249 initial IP addresses,**
250 **1000 new distinct addresses were generated using short-range (same**
251 **/16 network) dispersal, generating a new sample of 2.49×10^5 addresses**

252 **which were measured in the same manner as above. Another sample**
 253 **of 2.49×10^5 addresses were then generated using medium-range (same**
 254 **/8) dispersal.** In 2002, only short dispersal was used for the second phase of
 255 measurements; medium jumps were not performed. In 2005, we additionally
 256 scanned 15 addresses in the same /24 network for each of the 249 machines to
 257 look at “very short” range dispersal; in 2010, we scanned 20 such additional
 258 addresses per host.

259 Table 1 shows the results of the scans, indicating all hosts that we found
 260 running web servers of any kind. Table 2 shows results indicating only hosts
 261 running the dominant version of IIS. In 2010, of the 10^6 randomly selected IP
 262 addresses scanned, we found 1.196% running web servers, 7.564% of which were
 263 some version of IIS. Out of all the machines we scanned, 0.0655% were running
 264 IIS version 6.0. From these measurements, we projected that out of the 2^{32}
 265 IPv4 addresses, approximately 5.138×10^7 belonged to machines running web
 266 servers, 3.887×10^6 of which were running IIS, and 2.813×10^6 IIS 6.0, the
 267 most popular version of IIS at the time. As predicted, hosts with more leading
 268 octets in common with a host known to have a web server were more likely
 269 to also be running a web server with similar software. We used the results
 270 from the 2010 scan within our epidemiological simulations, assuming the worst-
 271 case scenario that all hosts running IIS 6.0 are unvaccinated (and therefore
 272 initially susceptible), namely $U_{..} = 6.55 \times 10^{-4}$, $Q_{n,U|U} = 2.892 \times 10^{-3}$, and
 273 $Q_{h,U|U} = 4.539 \times 10^{-2}$.

274 **3. Simulation**

275 *3.1. Generating Simulated Internets*

276 We first generate a 3-D *simulated Internet*, a population of $256^4 \approx 4.29 \times 10^9$
 277 hosts distributed as a set of 256 neighborhoods, each containing 256 households,
 278 each with 65536 hosts. This is done by adapting a similar method for generating
 279 2-D spatially structured heterogeneous lattices (Thomson and Ellner, 2003).
 280 **The previous work on lattices generated distributions of suitable and**

Year	“Source” version	Servers found out of approx. 10^6 scanned $U..$	Medium jumps (same /8)		Short jumps (same /16)		Very-short jumps (same /24)	
			Servers found out of 2.49×10^5 $Q_{n,U U}$	Times increase $Q_{n,U U}/U..$	Servers found out of 2.49×10^5 $Q_{h,U U}$	Times increase $Q_{h,U U}/U..$	Servers found increase	Times increase
2002	IIS 4.0	3629= 0.3658 %	–	–	21769=8.743% 23.90 ×		–	–
2005	IIS 5.0	5218= 0.5259 %	9587=3.850% 7.32 ×		28522=11.45% 21.77 ×		1420/3735=38.02% 72.29 ×	
2010	IIS 6.0	11964= 1.206 %	8231=3.306% 2.74 ×		33463=13.44% 11.14 ×		1568/4980=31.49% 26.11 ×	

Table 1: Results of scans to determine prevalence of all web servers and their clustering at various scales. **The total numbers of IP addresses scanned were: 992,116 (2002); 992,214 (2005); 992047 (2010).** “Times increase” indicates how much more likely one is to find a web server of **any kind** when performing a medium-, short-, or very short-distance jump (preserving 1, 2, or 3 leading octets of the IP address respectively) when starting from a web server running the then-dominant version of Microsoft IIS. **(The version of IIS on the hosts that served as the starting points for these subsequent localized scans is referred to as the “Source” version.)**

Year	Server version	Servers found out of approx. 10^6 scanned $U..$	Medium jumps (same /8)		Short jumps (same /16)		Very-short jumps (same /24)	
			Servers found out of 2.49×10^5 $Q_{n,U U}$	Times increase $Q_{n,U U}/U..$	Servers found out of 2.49×10^5 $Q_{h,U U}$	Times increase $Q_{h,U U}/U..$	Servers found increase	Times increase
2002	IIS 4.0	249= 0.0251 %	–	–	8747=3.513% 140.0 ×		–	–
2005	IIS 5.0	558= 0.0562 %	1691=0.6791% 12.08 ×		12920=5.189% 92.33 ×		1122/3735=30.04% 534.5 ×	
2010	IIS 6.0	655= 0.0660 %	720=0.2892% 4.382 ×		11301=4.539% 68.77 ×		1062/4980=21.33% 323.2 ×	

Table 2: Results of scans to determine prevalence of web servers running the dominant version of IIS and their clustering at various scales. Results are as in Table 1, but only servers running the then-dominant version of Microsoft IIS, and therefore potentially vulnerable to a worm targeting that server software, are counted.

281 unsuitable habitat, as follows. First, the desired proportion of sites
282 in the lattice are marked as suitable. Then, a pair of randomly-
283 chosen sites (one suitable, one unsuitable) is chosen. If swapping the

states of the two sites will move the measured clustering in the lattice closer to the desired clustering, the swap is performed. The process is then repeated until the desired clustering is reached, or a maximum number of iterations has been exceeded. This method is modified for our hierachically-structured population, with “suitable habitat” now corresponding to “unvaccinated hosts,” as described below.

First, the population is initialized so that each household has the same proportion $U.. = 6.55 \times 10^{-4}$ of unvaccinated hosts as estimated from our scans. We then cluster the population distribution, first among neighborhoods as follows, and subsequently among households:

1. Choose two neighborhoods at random, where the probability of choosing a neighborhood is proportional to the number of unvaccinated hosts within that neighborhood. Neighborhoods that are either completely unvaccinated or vaccinated are excluded from this choice.
2. Choose a random household within each of those two neighborhoods, where the probability of choosing a household is proportional to the number of unvaccinated hosts within that household.
3. Convert one vaccinated host to an unvaccinated host in the household whose neighborhood was less vaccinated, and convert one unvaccinated host to a vaccinated one in the other household. This increases the neighborhood-level clustering $Q_{n,U|U}$ while maintaining the overall prevalence $U..$ of unvaccinated hosts.
4. Repeat until $Q_{n,U|U}$ has reached or exceeded the desired level as measured from our scans, 2.892×10^{-3} .

This is then similarly performed at the household level, where hosts from two households within the same neighborhood are randomly chosen and adjusted, which maintains the neighborhood clustering $Q_{n,U|U}$ and prevalence $U..$ but increases the household clustering $Q_{h,U|U}$. This is repeated until $Q_{h,U|U}$ reaches or exceeds the desired value from our measurements, 4.539×10^{-2} .

The above procedure generates a simulated Internet with 2^{32} hosts, with the

314 desired mean and variance of susceptible hosts at the household and neighbor-
315 hood scales; note however, that third and higher moments of the distribution of
316 susceptible hosts are not specified (and were not measured in our scans). Our
317 simulated Internet has a total of 2,813,203 unvaccinated hosts **that are eligible**
318 **to be actively involved in the epidemiological simulations.**

319 3.2. Worm simulation

320 When simulating spatial epidemiological models, it is fairly typical to sim-
321 ulate all attempted infection events, i.e., contacts between infected hosts and
322 others, including those events where attempts to infect non-susceptible hosts fail.
323 Because the proportion of unvaccinated hosts is extremely low in our model ($\approx$
324 0.06%) and the **total** population size so large (2^{32}), this straightforward simu-
325 lation methodology is too inefficient for practical use here. We therefore used
326 the differential equations (3)–(4) to construct an improved simulation algorithm
327 which only simulates successful infections; failed infection attempts are filtered
328 out of the simulation.

329 Population-wide rates of successful infections arising from long-, medium-,
330 and short-distance contacts are denoted by r_0 , r_1 , and r_2 , respectively, while
331 the total rate of recoveries is given by r_μ . These rates, described in Table 3,
332 neglect the rates of failed infection attempts, where an infected host contacts
333 another host which is not susceptible. Let $r_T = r_0 + r_1 + r_2 + r_\mu$ be the total
334 rate of all successful events in the SIR-V simulation. The simulation proceeds
335 as follows:

- 336 1. The time until the next event is chosen from an exponential distribution
337 with mean $1/r_T$.
- 338 2. The type of the next event is chosen at random; the probability that the
339 next event will be of a given type is the ratio of the rate of that event-type
340 to r_T . For example, the probability that the next event is a successful
341 short-distance infection is r_2/r_T .
- 342 3. The location of the event is chosen.

Event type	Attempted rate	Successful rate
Long infection	$256^4 \phi \alpha_0 I_{..}$	$r_0 := 256^4 \phi \alpha_0 I_{..} S_{..}$
Medium infection	$256^4 \phi \alpha_1 I_{..}$	$r_1 := 256^3 \phi \alpha_1 \sum_{j=0}^{255} I_j \cdot S_j$
Short infection	$256^4 \phi \alpha_2 I_{..}$	$r_2 := 256^2 \phi \alpha_2 \sum_{j=0}^{255} \sum_{k=0}^{255} I_{jk} S_{jk}$
Recovery	$256^4 I_{..} \mu$	$r_\mu := 256^4 I_{..} \mu$

Table 3: The four event types within the epidemiological model and their population-wide rates. A contact between an infectious host and a non-susceptible host produces no change in the state of the system and is therefore a “wasted” or failed event. By using the probability of infections succeeding with contacts at various scales, the rates of only successful infections may be computed.

- (a) A recovery event occurs in neighborhood j , household k with probability $\mu I_{jk}/r_\mu$.
- (b) A long-distance infection occurs in neighborhood j , household k with probability $S_{jk}/(256^2 S_{..})$.
- (c) A medium-distance infection occurs in neighborhood j with probability $I_j \cdot S_j / \sum_{m=0}^{255} I_m \cdot S_m$; once the neighborhood has been chosen, the infection occurs specifically within household k in neighborhood j with probability $S_{jk}/(256 S_j)$.
- (d) A short-distance infection occurs in neighborhood j with probability $\sum_{\ell=0}^{255} I_{j\ell} S_{j\ell} / \sum_{m=0}^{255} \sum_{k=0}^{255} I_{mk} S_{mk}$; once the neighborhood has been chosen, the infection occurs specifically within household k with probability $I_{jk} S_{jk} / \sum_{\ell=0}^{255} I_{j\ell} S_{j\ell}$.

The simulation begins with a single infectious host. However, only infection events occur (i.e., no recoveries) until a specified number of initially infected hosts, N_{I_0} , is achieved. In a real worm outbreak, the individual responsible for first releasing the worm would likely similarly ensure (perhaps via repeated introductions) that the worm had infiltrated some minimum number of hosts before leaving it to proliferate unsupervised. It is also likely that fewer recovery events occur at such an early stage in the epidemic, before anti-malware software

has been updated to identify and remove the new threat. The simulation runs until there are no more infectious hosts remaining.

4. Results

Simulations were performed for a variety of dispersal parameters α_0 , α_1 , and α_2 , by letting each parameter vary between 0 and 1 in steps of $1/32$, subject to the constraint $\alpha_0 + \alpha_1 + \alpha_2 = 1$. There are 561 such parameter combinations; we ran 100 replicated simulations for each. We used an initial infection size of $N_{I_0} = 100$ infectious hosts. Other parameters used were contact rates of $\phi = 100$ and 1850, and recovery rate $\mu = 1$ which can be assumed without loss of generality by rescaling the units of time in the model, as only the ratio ϕ/μ affects qualitative dynamics. The reproduction ratio $\mathcal{R}_0$ from Eqn. (5) is minimized when $\alpha_0 = 1$, i.e. pure long-distance dispersal (random scanning) is used. In that case, the critical value $\mathcal{R}_0 = 1$ needed for the infection to avoid immediately decreasing to extinction occurs when $\phi = 1/U_{..} \approx 1527$. We chose our two values of ϕ to be slightly above and well below that value. In practice, the recovery rate μ is likely time-dependent, most likely increasing as both public awareness of the malware and the number of anti-malware software tools targeting it grow; this would be equivalent to decreasing ϕ over time in our model. Our study showed that results **varied gradually as ϕ varied over a fairly wide range of values (including several more values explored but not reported here), with no abrupt changes in results.**

Three quantities were recorded for each simulation: initial growth rate, final proportion $R_{..}/U_{..}$ of unvaccinated hosts affected by the infection, and household-scale relative clustering $Q_{h,I|I}/I_{..}$ of infected hosts at the time of peak infection, i.e. when $I_{..}$ achieves its maximum. The initial growth rate was measured as the slope of an interpolated least-squares fitted line for the logarithm of the infection level over an initial **period of 10 time units, which closer exploration and inspection of simulation results indicated was fairly indicative of initial behavior before density-dependent effects**

391 **tend to become significant.**

392 Fig. 1 displays the initial growth rate as a function of the dispersal param-
 393 eters. The location of Code Red’s dispersal parameters are highlighted within
 394 the figure. For comparison, the instantaneous initial growth rate may be cal-
 395 culated theoretically as follows. Using (1), (2), and (4), note that when the
 396 infection is at low density then $U_{jk} \approx S_{jk}$. Given that an infectious host which
 397 is randomly located from among all unvaccinated hosts will be in neighborhood
 398 j and household k with probability $U_{jk}/(65536U_{..})$, the instantaneous initial
 399 growth rate is

$$\phi(\alpha_0 U_{..} + \alpha_1 Q_{n,U|U} + \alpha_2 Q_{h,U|U}) - \mu. \quad (6)$$

400 Because $Q_{h,U|U}$ is much larger than $Q_{n,U|U}$ and $U_{..}$ as shown in Table 2, this
 401 quantity increases as one approaches $\alpha_2 = 1$, i.e. entirely short-distance disper-
 402 sal. When measured over an initial time interval rather than instantaneously at
 403 $t = 0$, saturation of local networks can begin to play a role; this happens more
 404 quickly with larger values of ϕ and when dispersal is more localized.

405 The final outbreak size $R_{..}/U_{..}$ is shown in Fig. 2; this indicates what propor-
 406 tion of unvaccinated hosts are infected over the entire course of the outbreak.
 407 This quantity quickly grows as α_0 and α_2 move away from zero, showing that a
 408 little bit of localized contacts and random scanning working together go a long
 409 way. Slightly higher infection levels can be reached by including more localized
 410 dispersal than Code Red used, but the results are not very sensitive to the exact
 411 values of α_0 , α_1 , and α_2 over much of their range.

412 Fig. 3 displays the household-scale relative clustering of infectious hosts,
 413 which indicates how much more likely it is that a second host chosen at random
 414 from within the same household as a randomly-chosen infected host will also
 415 be infected, as compared with a second host chosen from the entire population.
 416 With only short dispersal ($\alpha_2 = 1$), household-scale relative clustering $Q_{h,I|I}/I_{..}$
 417 is approximately 65536 because the infection is contained within one household.
 418 With only long-distance dispersal ($\alpha_0 = 1$), we obtain $Q_{h,I|I}/I_{..} \approx Q_{h,U|U}/U_{..} \approx$

Figure 1: The initial growth rate for the SIR-V model with $\phi = 100$ is shown as a function of the proportions of contacts which are short, medium, and long distance. Those proportions each varied from 0 to 1 in steps of $1/32$, subject to the constraint that they sum to one. Results here are displayed using barycentric coordinates. The location of parameters for Code Red ($\alpha_0 = 0.125, \alpha_1 = 0.5, \alpha_2 = 0.375$) is highlighted. Top: measurements from stochastic simulations; each hexagon within the triangle displays the mean of 100 replicated simulations. Bottom: theoretical instantaneous growth rate from Eqn. (6).

Figure 2: The rescaled final resistance level (i.e. R_{∞}/U_{∞} , the proportion of unvaccinated hosts that were affected during the entire course of the outbreak) is shown. Top: $\phi = 100$; Bottom: $\phi = 1850$ (note logarithmic scale for color values in bottom image).

Figure 3: The relative clustering (logarithmically scaled) of infectious hosts at the household scale $Q_{h,I|I}/I..$ when the system is at its peak infection level, with $\phi = 100$.

69 (from Table 2) for sufficiently large values of ϕ . For smaller values of ϕ , stochastic fluctuations in the model cause the clustering to deviate substantially from this (≈ 697 in Fig. 3).

Figs. 1–3 display sample means among 100 replicate simulations for each combination of parameter values; Fig. 4 displays the standard errors for the sample means of the final outbreak size with $\phi = 100$ from Fig. 2. **If there is no long-distance dispersal, the infection never spreads very far; if there is too much, the infection spreads relatively equally among replicated simulations. Variability tends to be highest when there is a small amount of long-distance dispersal. Along the strip of parameters slightly inside the top-left edge of the triangle, variability tends to increase as one moves up toward a higher proportion of short-range interactions with less medium-range contact.** Localized dispersal is therefore a relatively higher-risk strategy with potentially higher rewards for a worm seeking to maximize its number of hosts affected. Similar results (not

Figure 4: The standard errors of the final resistance level with $\phi = 100$ whose means are shown in the top of Fig. 2.

shown) were seen in the variability of the initial growth rate.

5. Discussion

Population models on heterogeneous landscapes in both lattices and community-structured models have shown that spatial clustering of suitable habitat or susceptible hosts strongly impacts the spread of populations or epidemics (Hiebeler, 2004; Hiebeler and Criner, 2007; Hiebeler et al., 2011; Liao et al., 2013a,b). One extreme case of clustered immunity in biological systems is with childhood vaccines, where some communities may have 18 times as many unvaccinated individuals as the population at large (McNeil Jr., 2002). Our measurements show very strong clustering of software platforms among web servers in IPv4 address space, far greater than the clustering of habitat typically observed in biological systems.

Because of the clustering of unvaccinated hosts, localized movement is much more likely to find new hosts to infect (several hundred times more likely, using

448 very-short jumps within the same /24 network). However, this strategy also has
 449 several disadvantages. Localized spreading can lead to greater local saturation
 450 of the infection, eventually causing infected hosts to waste their time contacting
 451 already-infected hosts nearby; this is a likely cause of the increased variability
 452 seen among replicated simulations with localized dispersal as seen in Fig. 4.
 453 To avoid this problem, some worms, such as Zotob, have begun using adaptive
 454 strategies whereby the values of α_0 , α_1 , and α_2 are modified based on the success
 455 or failure of infection attempts so far. Rapid spreading within a local area is
 456 also more likely to cause detection. A user walking into a lab with a single
 457 malfunctioning machine is likely to simply use another machine; if the entire
 458 lab is exhibiting problems, it's more likely to be reported. A local cluster of
 459 infected machines will also produce network traffic which is more likely to be
 460 detected by administrators than a single infected machine.

461 To be most successful, some proportion of infection attempts must take place
 462 over larger spatial scales in order to colonize new regions of Internet address
 463 space, along with local dispersal to exploit those newly colonized regions. Fig. 2
 464 shows that once α_2 increases somewhat modestly, the proportion of the popula-
 465 tion affected by an outbreak can be quite large relative to its maximum possible
 466 value. For example, with $\phi = 1850$, **the overall maximum final propor-**
 467 **tion of unvaccinated hosts affected (R_{∞}/U_{∞}) was 0.94 when $\alpha_0 = 3/8$,**
 468 **$\alpha_1 = 7/32$, and $\alpha_2 = 13/32$. With a relatively modest value of $\alpha_2 = 1/16$**
 469 **(with $\alpha_0 > 0$ to avoid the worm being completely trapped within the**
 470 **region it starts in), the *minimum* final R_{∞}/U_{∞} was approximately 0.86.**
 471 **This is in comparison with a value of only 0.33 with pure long-distance**
 472 **contact.**

473 IPv6 addressing will increase the size of the Internet address space to to 2^{128}
 474 possible addresses, 2^{96} times the size of IPv4. **One often-used example to**
 475 **illustrate the magnitude of this larger address space is the following:**
 476 **assume that the world population is 8×10^9 people, and that there**
 477 **are roughly 10^{14} cells per human body. There would still be more**
 478 **than 10^{14} IPv6 addresses available per cell in every human body. Put**

479 another way, suppose each of the 4.3×10^9 possible IPv4 addresses has
 480 a single computer host; if we were to place these hosts within the
 481 IPv6 address space and release a new worm capable of scanning 10^9
 482 hosts per second (roughly 300×10^6 times faster than Code Red), it
 483 would still take roughly 3×10^{12} years to find a host. The distribution
 484 of hosts in the address space will therefore be much too sparse for long-distance
 485 scanning to have any reasonable hope of success; even preferential random scan-
 486 ning will only be viable if hosts are extremely clustered in IPv6 address space.
 487 **Worms spreading in IPv6 address space will need to carry information**
 488 **about where they should seek new susceptible hosts, at the expense**
 489 **of making such worms larger and slower to copy.** It is likely that IPv4
 490 will remain in widespread use for quite some time, and consequently worm cre-
 491 ators will continue to favor the simple and effective scanning methods there over
 492 the more elaborate and ambitious techniques that would be necessary to infect
 493 machines over IPv6 (Albanese et al., 2004; Bellovin et al., 2006; Chen and Ji,
 494 2007). **However, the continued rapid growth of mobile devices such as**
 495 **smartphones with IPv6 addresses may eventually become a lucrative**
 496 **population for malicious software to target.**

497 Our measurements show that the proportion of IP addresses with web servers
 498 increased more than threefold over eight years as the Internet grew. Interest-
 499 ingly, the clustering of servers at the household level $Q_{h,U|U}$ did not follow a
 500 similar trend, while clustering at the neighborhood level $Q_{n,U|U}$ actually de-
 501 creased from 2005 to 2010. The fractions of all web servers running the dom-
 502 inant version of IIS were 6.86%, 10.7%, and 5.48% respectively in 2002, 2005,
 503 and 2010. The clustering of servers around these hosts running IIS reflects the
 504 pattern of deployment of new versions of IIS rolling out, which may vary among
 505 different versions. **Additional scans have been performed more recently**
 506 **than 2010. However, the methodology used for those scans was quite**
 507 **different, and detailed results of those scans are not included here.**
 508 **But to give one point of comparison with our historical scans, in**
 509 **September 2015, 1.988% of approximately 5.65×10^6 measured IP ad-**

510 dresses were running a web server of any kind, with $1.629\times$ and $10.05\times$
511 increases when using medium and short jumps. Version 7.5 was the
512 most prevalent version of IIS at that time; it was found on 0.0610%
513 of scanned hosts, with $2.796\times$ and $150.5\times$ increases at the medium and
514 short scales.

515 Although many worms propagate by exploiting vulnerabilities in other soft-
516 ware, measuring web servers is a relatively unobtrusive way to obtain a sample
517 which is likely to be qualitatively representative of the distribution of hosts in the
518 Internet. Performing more detailed scans of a given host is more likely
519 to set off intrusion detection systems, causing additional worry and
520 work for the administrators of those machines; we have attempted to
521 make our scanning process minimally disruptive. Also, even if hosts
522 which were actually susceptible to a particular worm were randomly distributed
523 among all hosts running, say, IIS 6.0, such machines would still be highly clustered
524 in the IP address space because of the strong clustering of IIS 6.0 hosts
525 themselves. In fact, unpatched hosts are likely to be even more clustered within
526 IIS 6.0 servers, because machines with similar IP addresses are more likely to
527 be maintained by the same group of people exercising a common set of security
528 practices. The spread of worms could be slowed by randomizing the
529 locations of hosts in IP address space, i.e., by randomly assigning IP
530 addresses to hosts regardless of their location or the organization they
531 belong to. This would make the distribution of susceptible hosts less
532 clustered, i.e., more uniform, among different regions in IP address
533 space. Previous epidemiological models have shown that varying the
534 clustering of unvaccinated hosts among communities within a pop-
535 ulation can change the proportion of hosts affected by an order of
536 magnitude or more (Hiebeler et al., 2011). Unfortunately, randomiz-
537 ing the assignment of IP addresses would also make it more difficult
538 to administer their allocations among different organizations.

539 Because the spread of a worm is not uniform across the space of
540 IP addresses, strategies for placing monitors to detect the spread of

malicious software may also be designed to take advantage of this heterogeneity. In particular, it would be advantageous to focus monitoring and detection efforts in those regions of the Internet that contain more susceptible hosts, as worms are both more likely to appear in those regions, and also able to spread more quickly (and do more damage) once there.

Future work with these models will explore the effects of worms such as Zotob using adaptive strategies to adjust dispersal parameters over time, and allowing for a time-varying recovery rate μ . Initial measurements taken during the spread of Code Red I showed a contact rate orders of magnitude larger than those studied here, but the rate surely decreased as awareness of the worm spread. The recovery rate may in fact be tied to prevalence of infections at local and global spatial scales, as media coverage of an outbreak will affect people's tendency to employ protective measures.

In conclusion, we have measured the prevalence and clustering of web servers in IPv4 address space over the past decade, showing a growing number of web servers and quantifying their clustering. Assuming a worst-case scenario that all web servers running IIS 6.0 are vulnerable, we constructed an artificial network with the same mean and variance of susceptible machines as seen in our measurements. Simulations of a full-size IPv4 address space (2^{32} hosts) **with approximately 2.8×10^6 susceptible hosts** showed that the local preference scanning strategies used by Code Red II and later worms greatly increases the ability of infected hosts to successfully locate additional susceptible hosts to infect. Highly localized scanning can however lead to greater variability in the success of a worm, and may increase chances of detection. We hope this paper serves as an introduction to quantify some of the well-known clustering of hosts in IPv4 address space, and that improved measurements and simulations may guide placement of monitoring systems to increase chances of early detection of intrusions.

570 6. Acknowledgments

571 This material is based upon work supported by the National Science Foun-
572 dation under Grant No. DMS-0746603 to D.E.H.

573 References

- 574 Albanese, D. J., Wiacek, M. J., Salter, C. M., Six, J. A., 2004. The
575 case for using layered defenses to stop worms. Network Architecture
576 and Applications Division of the Systems and Network Attack Center
577 (SNAC), Information Assurance Directorate, U.S. National Security Agency;
578 <http://www.nsa.gov/snac/support/WORMPAPER.pdf>.
- 579 Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks.
580 *Reviews of Modern Physics* 74 (1), 47–97.
- 581 Alderson, D., Li, L., Willinger, W., Doyle, J. C., 2005. Understanding Inter-
582 net topology: Principles, models, and validation. *IEEE/ACM Transactions on*
583 *Networking* 13 (6), 1205–1218.
- 584 Avlonitis, M., Magkos, E., Stefanidakis, M., Chrissikopoulos, V., Jun. 2007.
585 A spatial stochastic model for worm propagation: Scale effects. *Journal in*
586 *Computer Virology* 3 (2), 87–92.
- 587 Balthrop, J., Forrest, S., Newman, M., Williamson, M. M., Apr. 2004. Techno-
588 logical networks and the spread of computer viruses. *Science* 304, 527–529.
- 589 Barbour, A. D., 1978. Macdonald’s model and the transmission of bilharzia.
590 *Transactions of the Royal Society of Tropical Medicine and Hygiene* 72 (1),
591 6–15.
- 592 Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A., 2005. Dy-
593 namical patterns of epidemic outbreaks in complex heterogeneous networks.
594 *Journal of Theoretical Biology* 235, 275–288.

595 Bellovin, S., Cheswick, B., Keromytis, A., 2006. Worm propagation strategies
596 in an IPv6 Internet. *Login* 31 (1), 70–76.

597 Chen, Z., Gao, L., Kwiat, K., 2003. Modeling the spread of active worms. *Proc.*
598 *of 22nd Annual Joint Conference of the IEEE Computer and Communications*
599 *Societies (INFOCOM'03)*, San Francisco, California, USA, March 2003.
600 URL http://www.ieee-infocom.org/2003/papers/46_03.PDF

601 Chen, Z., Ji, C., Sep. 2005. Spatial-temporal modeling of malware propagation
602 in networks. *IEEE Transactions on Neural Networks* 16 (5), 1291–1303.

603 Chen, Z., Ji, C., May 2007. Measuring Network-Aware Worm Spreading Abil-
604 ity. Vol. 26 of *IEEE International Conference on Computer Communications*
605 *(INFOCOM 2007)*. pp. 116–124.

606 Daley, D., Gani, J., 1994. A deterministic general epidemic model in a stratified
607 population. In: Kelly, F. (Ed.), *Probability, Statistics, and Optimisation*.
608 John Wiley & Sons, Ltd., New York, NY, Ch. 8, pp. 117–132.

609 Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S.,
610 Tanaka, R., Willinger, W., Oct. 2005. The “robust yet fragile” nature of
611 the Internet. *Proceedings of the National Academy of Sciences of the United*
612 *States of America* 102 (41), 14497–14502.

613 Duryea, M., Caraco, T., Gardner, G., Maniatty, W., Szymanski, B. K., 1999.
614 Population dispersion and equilibrium infection frequency in a spatial epi-
615 demic. *Physica D* 132, 511–519.

616 Dye, C., Hasibeder, G., 1986. Population dynamics of mosquito-borne disease:
617 Effects of flies which bite some people more frequently than others. *Transac-*
618 *tions of the Royal Society of Tropical Medicine and Hygiene* 80, 69–77.

619 Eichin, M. W., Rochlis, J. A., May 1989. With Microscope and Tweezers: An
620 Analysis of the Internet Virus of November 1988. In: *Proceedings of the 1989*
621 *IEEE Symposium of Research in Security and Privacy*. Oakland, CA, pp.
622 326–343.

623 Filipe, J., Gibson, G., 1998. Studying and approximating spatio-temporal mod-
624 els for epidemic spread and control. *Phil. Trans. R. Soc. Lond. B* 353, 2153–
625 2162.

626 Filipe, J., Maule, M., 2004. Effects of dispersal mechanisms on spatio-temporal
627 development of epidemics. *Journal of Theoretical Biology* 226, 125–141.

628 Grabowski, A., Kosiński, R., 2004. Epidemic spreading in a hierarchical social
629 network. *Physical Review E* 70, 031908.

630 Hiebeler, D. E., 2004. Competition between near and far dispersers in spatially
631 structured habitats. *Theoretical Population Biology* 66 (3), 205–218.

632 Hiebeler, D. E., 2006. Moment equations and dynamics of a household SIS
633 epidemiological model. *Bulletin of Mathematical Biology* 68 (6), 1315–1333.

634 Hiebeler, D. E., Criner, A. K., 2007. Partially mixed household epidemiological
635 model with clustered resistant individuals. *Physical Review E* 75, 022901.

636 Hiebeler, D. E., Michaud, I. J., Ackerman, H. H., Iosevich, S. R., Robinson, A.,
637 2011. Multigeneration reproduction ratios and the effects of clustered unvac-
638 cinated individuals on epidemic outbreak. *Bulletin of Mathematical Biology*
639 73 (12), 3047–3070.

640 Hwang, D.-U., Boccaletti, S., Moreno, Y., López-Ruiz, R., April 2005. Thresh-
641 olds for epidemic outbreaks in finite scale-free networks. *Mathematical Bio-*
642 *sciences and Engineering* 2 (2), 317–327.

643 Keeling, M., 2005. The implications of network structure for epidemic dynamics.
644 *Theoretical Population Biology* 67, 1–8.

645 Kephart, J. O., Chess, D. M., White, S. R., May 1993. Computers and epidemi-
646 ology. *IEEE Spectrum*, 20–26.

647 Kephart, J. O., White, S. R., 1991. Directed-graph epidemiological models of
648 computer viruses. In: *Proceedings of the IEEE Computer Society Symposium*
649 *on Research in Security and Privacy*. Oakland, CA, pp. 343–359.

650 Kephart, J. O., White, S. R., May 1993. Measuring and modeling computer virus
651 prevalence. IEEE Symposium on Research in Security and Privacy. Oakland,
652 CA, pp. 2–15.

653 Kotliar, N. B., Wiens, J. A., 1990. Multiple scales of patchiness and patch
654 structure: A hierarchical framework for the study of heterogeneity. *Oikos* 59,
655 253–260.

656 Liao, J., Li, Z., Hiebeler, D. E., El-Bana, M., Deckmyn, G., Nijs, I., 2013a.
657 Modelling plant population size and extinction thresholds from habitat loss
658 and habitat fragmentation: Effects of neighbouring competition and dispersal
659 strategy. *Ecological Modelling* 268, 9–17.

660 Liao, J., Li, Z., Hiebeler, D. E., Iwasa, Y., Bogaert, J., Nijs, I., 2013b. Species
661 persistence in landscapes with spatial variation in habitat quality: A pair
662 approximation model. *Journal of Theoretical Biology* 335, 22–30.

663 McNeil Jr., D. G., Nov. 2002. When parents say no to child vaccinations. *The*
664 *New York Times*.
665 URL [http://www.nytimes.com/2002/11/30/us/
666 when-parents-say-no-to-child-vaccinations.html](http://www.nytimes.com/2002/11/30/us/when-parents-say-no-to-child-vaccinations.html)

667 Moore, C., Newman, M., 2000. Epidemics and percolation in small-world net-
668 works. *Physical Review E* 61 (5), 5678–5682.

669 Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., Weaver, N., Jul.
670 2003a. Inside the Slammer worm. *IEEE Security and Privacy* 1 (4), 33–39.

671 Moore, D., Shannon, C., 2001. The spread of the Code-Red worm (CRv2). *The*
672 *Cooperative Association for Internet Data Analysis*.
673 URL [http://www.caida.org/research/security/code-red/coderedv2_
674 analysis.xml](http://www.caida.org/research/security/code-red/coderedv2_analysis.xml)

675 Moore, D., Shannon, C., Brown, J., 2002. Code-Red: A case study on the
676 spread and victims of an Internet worm. In: *Proceedings of the Second ACM*
677 *SIGCOMM Workshop on Internet Measurement*. New York, NY, pp. 273–284.

678 Moore, D., Shannon, C., Voelker, G., Savage, S., 2003b. Internet Quarantine:
679 Requirements for Containing Self-Propagating Code. In: Proceedings of the
680 22nd Annual Joint Conference of the IEEE Computer and Communications
681 Societies (INFOCOM '03). San Francisco, CA, pp. 1901–1910.

682 Nazario, J., 2004. Defense and Detection Strategies Against Internet Worms.
683 Artech House, Boston.

684 Newman, M., 2002. The spread of epidemic disease on networks. *Phys. Rev. E*
685 66, 016128.

686 Pastor-Satorras, R., Vespignani, A., 2001. Epidemic dynamics and endemic
687 states in complex networks. *Physical Review E* 63, 066117.

688 Pastor-Satorras, R., Vespignani, A., 2004. Evolution and Structure of the Inter-
689 net. Cambridge University Press, Cambridge, United Kingdom.

690 Staniford, S., Paxson, V., Weaver, N., 2002. How to Own the Internet in Your
691 Spare Time. In: Proceedings of the 11th USENIX Security Symposium.
692 Berkeley, CA, pp. 149–167.

693 Thomson, N. A., Ellner, S. P., 2003. Pair-edge approximation for heterogeneous
694 lattice population models. *Theoretical Population Biology* 64, 271–280.

695 Vogt, T., February 2004. Simulating and Optimising Worm Propagation Algo-
696 rithms.
697 URL <http://web.lemuria.org/security/WormPropagation.pdf>

698 Wang, C., Knight, J., Elder, M., 2000. On computer viral infection and the effect
699 of immunization. Vol. 16 of IEEE Computer Security Applications Conference
700 (ACSAC '00). New Orleans, LA, pp. 246–256.

701 Weaver, N., March 2002. Potential Strategies for High Speed Active Worms: A
702 Worst Case Analysis.
703 URL www.cgisecurity.com/lib/worms.pdf

- 704 Zou, C. C., Towsley, D., Gong, W., October 2004. Email Worm Modeling and
705 Defense. In: 13th International Conference on Computer Communications
706 and Networks (ICCCN'04). Chicago, IL, pp. 409–414.
- 707 Zou, C. C., Towsley, D., Gong, W., Jul. 2006. On the performance of Internet
708 worm scanning strategies. Performance Evaluation 63 (7), 700–723.