

Distributive immunization of networks against viruses using the 'honey pot' architecture

Jacob Goldenberg¹, Yuval Shavitt², Eran Shir², Sorin Solomon¹

¹*Hebrew University, Jerusalem, Israel*

²*Tel-Aviv University, Tel-Aviv, Israel*

Although computer viruses cause tremendous economic loss, defense mechanisms fail to adapt to their rapid evolution. Previous immunization strategies have been characterized as being static and centralized, which has made virus containment difficult or even impossible. We suggest, instead, to propagate the immunization agent as an epidemic.

The main problem with epidemic vaccine propagation is that it is bound to lag behind the virus. We suggest to give the vaccine an advantage over the virus by allowing it to leapfrog through a separate, overlapping, partially correlated network. This enables the anti-virus to contain the epidemic efficiently. We systemize this concept with a “honey pots” architecture which achieves both early virus discovery and rapid immune dissemination. We present analytic, as well as simulation, results for a set of realistic topologies that illustrate the effectiveness of this approach.

Epidemic Modeling

The realization that network models possess non-trivial properties¹⁻³ such as a diameter which grows logarithmically with network size⁴ and a non-existent percolation threshold⁵, implies that for predominant epidemic models the epidemic will not stop by immunizing any finite sub-set of nodes.

However, current immunization strategies⁶⁻¹¹ focus on removing nodes from the network a-priori by immunizing them before the epidemic outburst. In the absence of complete knowledge of the network topology, these strategies are confined to a random character. Thus, these strategies require, in most cases, the removal of almost all of the nodes, and in all cases⁷ the removal of at least a quarter of the nodes.

In contrast, we introduce a dynamic distributed immunization strategy, where the vaccine development and immunization processes depend on, and interact with, the virus dissemination process itself, thus creating a co-dependency between virus dissemination and immunization.

In the context of traditional biological epidemiology, there was little sense in considering dynamic, distributed immunization strategies. This is mainly due to the fact that the time-scale gap between epidemic outburst and vaccine creation is very large, and that there is no 'infectious' delivery mechanism available for biological vaccines.

The world of computer viruses has diametrically different characteristics. First, new viruses emerge at an increasing pace. Second, computer viruses are much less complex than their biological counterparts, and are much easier to analyze and to characterize^{12,13}. Thus, vaccine development can be achieved in a time-scale comparable to that of the infection process. On the negative side, however, the viral process possesses an inherent lead-time advantage: It appears before the vaccine, since a new vaccine can be created only after the new virus has started percolating the network. This fact, in itself, imposes strong constraints¹⁴ on the usability of dynamic approaches. However, as we will demonstrate below, one can devise design principles that compensate for the virus' lead-time advantage and that support the deployment of efficient dynamic immunization systems.

We discuss the concrete example of the e-mail network. In this network, an e-mail account constitutes a node while the directed edges of the network are the entries in the account's address book. The virus spreads through the account address book with a timescale that ranges from several hours to a number of days. This timescale is an upper-bound of the vaccine generation timescale for any effective epidemic containment solution. Studies^{9,15} show that this network's degree distribution (i.e. the distribution, $P(k)$, which governs the probability that a node will have degree k , that is have k edges attached to it) is very broad and can be modeled by a scale-free network, which is a network with a power law degree distribution. We verified this through a survey of 502 individuals, which we also used to calibrate our simulations' parameters.

In the last several years, virus spreading on such networks has been studied intensively using the static percolation framework¹¹. The dynamics that we introduce stem from this framework and allow for a richer set of effects.

Distributed Immunization

In the present communication we define a framework for the study of immunization strategies that react in real-time to the emergence and propagation pattern of a virus. The objective is to find the strategies that minimize the size of the infected cluster. The size of the virus cluster is the portion of infected nodes after a time period that we take to infinity. The size of the immunized clusters is defined consistently as the aggregate number of immunized nodes. The underlying assumptions of our model are:

1. Like the SIR epidemiologic model^{16,17}, a node can be in one of three modes with respect to a specific virus: susceptible, infected, or immunized (removed). However, unlike the SIR model, a node cannot change its mode once it is either infected or immunized during the relevant time

scale. This model is in close agreement with the behavior observed on the Internet today, as an increasing number of viruses shut down security related software upon infecting a new node.

2. An infected node releases the virus to all of its neighbors with a delay time that is either deterministic or stochastic. The virus is transmitted to all neighbors, and not a stochastic subset, which makes fighting the virus harder.

3. Some nodes, in accordance with a given probability function, may recognize their own infection, identify its characteristics, and create an immunization agent, as the infection process progresses^{13,18}. The agent then spreads to all neighbors and immunize the susceptible ones.

In addition, we define:

- Average infection delay, also known as the disease generation time, t_{inf} = the average time required by a virus in a given node to infect a neighboring node.
- Average immunization delay, t_{imm} = the average time required by an immunization agent in a given node to immunize a neighboring node.
- Average development delay, t_{dev} = the average time required by an infected node to develop an immunization agent.

In essence, the described dynamics involve a competition between two types of branching processes on a network¹⁹, where the first type creates a connected virus cluster, and the second creates a collection of immunized clusters. Unlike centralized approaches, this one nullify the need for a global knowledge of the topology. We consider the deterministic case, where the various delays associated with the infection, agent creation, and immunization, are all constant, and where all neighboring nodes become infected or immunized simultaneously. The resulting dynamics exhibit a sharp transition at the point where $t_{inf} = t_{imm} + t_{dev}$. In the deterministic

case, when t_{inf} is below this threshold, the virus infects the entire network, whereas when above it, the dynamics are governed by the agent development pace and the network topology characteristics. This sharp transition around the critical value remains true also when the delays are stochastic variables. In the discrete time simulations we present below, all of the time parameters equals one time step, which in turn gives the virus a head start of one time step. As presented in Figure 1, this difference is enough to let the virus infect the entire network when the virus and immunization agent spread on the same network.

Partially Correlated Networks

To unleash the potential of the immunization system, we offer a slight modification to the problem by introducing a relatively small number of edges to the network topology. These immunization edges, which are used exclusively by the immunization agents, have a dramatic effect on the ability of a dynamic scheme to contain the virus by offering access to a parallel network with identical nodes and almost identical edges as the original network. These edges connect the node that produced the immunization agent to nodes that are beyond its immediate neighborhood as defined by the the initial network. In our example, the parallel network is the phone book network, which is strongly correlated with the e-mail network.

The study of networks that connect the same nodes but have different sets of links is only in its infancy^{20,21} but even now it is clear that such networks are qualitatively different from each of their components taken separately. This is due to the complete change in the topology and the metrics that are induced in each of the networks through their interaction. In practical terms, this means that the immunization agents are effectively deployed "behind enemy lines", unconstrained by the boundaries of the surrounding virus cluster. Once in this position, they

can alter the topology of the space remaining at the virus' disposal by immunizing nodes that otherwise would have belonged to the infected cluster. In Figure 1 we illustrate the difference between a network with no extra immunization edges and a network that does possess a number of these edges. The difference in the dynamics is further illustrated in Supplementary Video 1.

The effect of introducing additional immunization edges, along with the original network amounts to the generation of a pair of *Partially Correlated Networks*^{22,23}, which we define as follows:

Two given networks $G_1 = (V, E_1), G_2 = (V, E_2)$ are partially correlated with overlap p if $p = \frac{|E_1 \cap E_2|}{\max(|E_1|, |E_2|)}$ is greater than zero.

Starting with our initial network G_1 , we created a new network G_2 for the immunizing agent by adding to G_1 a set of edges e_1 that do not belong to G_1 . Using the relative edge addition, $q = |e_1|/|E_1|$, the overlap will be:

$$p = \frac{|E_1|}{|E_2|} = \frac{|E_1|}{|E_1 \cup e_1|} = \frac{1}{1 + q}. \quad (1)$$

Next, we alter the Distributed Immunization Dynamics in the following way: The virus spreads through the original network G_1 , while the immunizing agent is deployed through the partially correlated network G_2 , which is obtained by randomly adding $q|E_1|$ edges to G_1 . By doing so, we enable the immunizing agent to break through the virus cluster and to immunize the network.

In the methods section we show analytically that for the discrete-time deterministic model the relative size of the infected cluster (i.e., the ratio of infected to immunized clusters), as a function of the relative edge addition q , has a power law upper-bound with a -1 power exponent.

In addition, we have studied the problem through network simulations. In Figure 2 we present simulation results that show a power law ratio dependence with an exponent that approaches $-4/3$.

Thus, we can conclude that dynamic immunization, which is employed over partially-correlated networks can reduce the size of a virus cluster considerably with negligible costs.

Honey Pots

To systemize and improve our scheme we present the *Honey Pots Architecture*¹³ (originally termed due to their function as traps). This architecture has two main benefits over the random solution. First, it is much more realistic and technically feasible. Second, it is considerably more efficient than a random deployment of immunizing edges, and given the same immunization edge budget, it minimizes the virus cluster to sizes that can be as small as a fourth of the respective cluster in the random-edges case. These features make this architecture a most attractive alternative to current immune systems.

The aim of this architecture is to introduce a virtual super-hub, which transforms the shortcomings of a scale-free network – that is considerably impaired when its largest hubs are removed³ – into an advantage.

The Honey Pot Architecture is constructed in the following manner: We exclusively implant the ability to develop an immunizing agent to a set of randomly-distributed nodes in our network (“the Honey Pots”). The Honey Pots are embedded randomly within the network, such that any virus that spreads through the network will be likely to promptly reach them. Finally, all Honey Pots are connected in a complete graph topology using special edges that only allow

the immunizing agents to traverse.

Initially, the virus spreads freely, until it infects the first honey pot and thus triggers an immunization agent development process. By this time, the expected size of the virus cluster equals the size of the network divided by the number of honey pots. As the virus continues to spread, all honey pots are informed of the new virus, and each honey pot then begins to function as the root of a separate infectious immunization process. The honey pots have the effect of a super-hub, with a degree which is the sum of the degrees of the separate honey pots.

In the methods section we calculate an upper bound for the relative virus cluster using the honey pot architecture. We show that if the amount of honey pots, as a function of the network size N , $f(N)$, grows faster than \sqrt{N} , the size of the virus cluster will approach a zero portion of the network, as network size approaches infinity. In the case where $f(N) = \beta N$, we get

$$\lim_{t \rightarrow \infty} \frac{V_t(N)}{A_t(N)} = \frac{1}{\beta^2 N} \cdot (\alpha - 1) \quad (2)$$

where α is a characteristic constant of the topology, $V_t(N)$ is the size of the virus cluster and $A_t(N)$ is the aggregate size of the immunized clusters after time step t , as a function of network size N . When $f(N) = \sqrt{N}$, the relative special edge addition due to the honey pots is kept constant in the infinite size limit. This analytic estimation is validated through simulations and is presented in Figures 3 and 4 and illustrated in Supplementary Video 2. Comparing the random architecture to the honey pot architecture, we observe in Figure 5 a significant improvement due to the honey pots that grows with network size.

In Figure 6 we address the question of robustness of these approaches to different topology characteristics, presenting an analysis of the effect of varying the degree distribution power exponent on the virus cluster, and show that it is minor compared with the dependence on the

immune edges density.

Deployment and Feasibility

Faced with the systematic defeats in the war against computer viruses, a paradigm-shift may be required. We propose such a shift from the current, static, centralized immunization strategies to a dynamic-distributed-immune-system approach. We demonstrate the effectiveness of such an architecture in protecting large networks, both when built randomly or when designed artificially. Although the presentation of a practical system design is outside the scope of this paper, such a system is certainly deployable. In the recent past, it has been shown that distributed systems that monitor the Internet in realtime^{24,25} are not only feasible, but are also very effective. Shifting the focus of an anti-virus system from cleaning a single machine to containing the epidemic, enables the introduction of accurate automatic triggering within a timescale of less than a minute, which allows such a system to surpass the $t_{inf} = t_{imm} + t_{dev}$ barrier. This enables the system to compete with and defeat the spread of the epidemic. The architectures we presented constitute a starting point that can be further improved, e.g. by designing algorithms for the placement of the honey pots⁷.

Methods

Analysis of the random edges effect Given a network with degree distribution $P(k)$, let us calculate an upper bound on the rate of growth of the virus cluster, $V_t(N)$ where N is the network's size and t is the time index: Let us examine the portion of the $t + 1$ time layer, l_{t+1} , with degree k .

$$l_{t+1}(k) = \frac{kP(k)}{\sum_{k'} k' P(k')} \cdot \sum_{k'} l_t(k') \cdot C \cdot (k' - 1) = \frac{kP(k)}{m} \cdot \sum_{k'} l_t(k') \cdot C \cdot (k' - 1) \quad (3)$$

where m is the average node degree and C holds the topological clustering properties of the network, which reduce the number of effective neighbors. While in general C may be a complicated expression, and may also depend on k , in our mean field approximation we refer to it as a constant of the topology. Since the sum does not depend on k , we can calculate it independently and call it a_{t+1} . Then, $l_{t+1}(k) = a_{t+1}kP(k)$. Substituting in (3), gives us

$$l_{t+1}(k) = l_t(k) \cdot \sum_{k'} \frac{k'P(k')}{m} \cdot C \cdot (k' - 1) \quad (4)$$

We call the outcome of the new sum α . Since it does not depend on k , we get that $l_{t+1} = l_t \cdot \alpha$. If α is larger than 1 we get an exponential growth. If N is large enough so that finite size effects are irrelevant we get:

$$V_t(N) = (1 + \alpha + \alpha^2 + \dots + \alpha^t) = \frac{\alpha^{t+1} - 1}{\alpha - 1}. \quad (5)$$

Let us turn to the immunized cluster(s). Define $A_t(N)$ to be the aggregate size of the immunized clusters at time step t as a function of N . Given a relative edge addition q , and an average degree m , the expected number of immune specific edges is qm , each may initiate an immunized cluster. Once started, the immunized clusters, also grow with ratio α . Thus, the aggregate immunized clusters' size when N is large enough, A_t is:

$$A_t(N) = qm \left[(t-1) \cdot \alpha^{t-2} + (t-2) \cdot \alpha^{t-3} + \dots + 1 \right]. \quad (6)$$

which can be compacted:

$$A_t(N) = qm \left[\frac{t\alpha^{t-1}}{\alpha - 1} - \frac{\alpha^t - 1}{(\alpha - 1)^2} \right] \quad (7)$$

The ratio between the size of the virus cluster and immunized clusters is:

$$\frac{V_t(N)}{A_t(N)} = \frac{1}{qm} \left[\frac{(\alpha^{t+1} - 1)(\alpha - 1)}{(t-1)\alpha^t - t\alpha^{t-1} + 1} \right] \quad (8)$$

from which we get an upper-bound (since all our assumptions were in favor of the virus cluster) power law dependence with exponent -1 on q for the discrete time, deterministic model.

Analysis of the Honey Pot architecture effect We would like to calculate an upper-bound for the ratio $\frac{V_t(N)}{A_t(N)}$ for very large N s and when t approach infinity.

Let us assume that there are $f(N)$ honey pots distributed randomly in the network, all connected in a complete graph using immunization edges. Then, clearly, the expected virus cluster size when a honey pot is infected with a virus for the first time is: $\frac{N}{f(N)}$. At that time, the boundary outside the virus cluster will have $\frac{N}{f(N)} \cdot (\alpha - 1)$ nodes. At the next time step, $f(N)$ nodes will be ‘infected’ with the immunization agent. From this point forward, we assume that (in the deterministic case) the virus cluster and the immunized clusters grow as an uninterrupted geometric series. Then, with increasing t , their ratio approaches:

$$\frac{V_t(N)}{A_t(N)} = \frac{\frac{N}{f(N)} \cdot (\alpha^t - 1)}{f(N) \cdot \frac{\alpha^t - 1}{(\alpha - 1)}} = \frac{N}{f(N)^2} \cdot (\alpha - 1) \quad (9)$$

From this equation we can see that whenever $f(N)$ grows faster than \sqrt{N} the size of the virus cluster will approach a zero portion of the network, as the network size approaches infinity. In the case where $f(N) = \beta N$, we get

$$\frac{V_t(N)}{A_t(N)} = \frac{1}{\beta^2 N} \cdot (\alpha - 1) \quad (10)$$

which means that we have a power law relation between this ratio and the relative amount of honey pot nodes, β , with an exponent equal to -2 . This result is not surprising, since $f(N) = \sqrt{N}$ is the function for which the relative special edge addition due to the honey pots is kept constant in the infinite size limit.

1. Watts, D. J. and Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
2. Albert, R., Jeong, H., and Barabasi, A.-L. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

3. Albert, R., Jeong, H., and Barabasi, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
4. Chung, F. and Lu, L. The average distances in random graphs with given expected degrees. *PNAS* **99**, 15879–15882 (2002).
5. Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
6. Pastor-Satorras, R. and Vespignani, A. Immunization of complex networks. *Physical Review E* **65**, 036104 (2002).
7. Havlin, S., Cohen, R., and Ben-Avraham, D. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.* **91**, 247901 (2003).
8. Dezsó, Z. and Barabasi, A.-L. Halting viruses in scale-free networks. *Physical Review E* **65**, 055103 (2002).
9. Newman, M., Forrest, S., and Balthorp, J. Email networks and the spread of computer viruses. *Physical Review E* **66**, 035101 (2002).
10. Zou, C. C., Gong, W., and Towsley, D. Code red worm propagation modeling and analysis. In *The 9th ACM conference on Computer and communications security*, 138–147, (2002).
11. Pastor-Satorras, R. and Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
12. Kephart, J., Sorkin, G., Swimmer, M., and White, S. *Blueprint for a Computer Immune System*, chapter 13, 242–261. Springer-Verlag (1999).

13. Kreibich, C. and Crowcroft, J. Honeycomb-creating intrusion detection signatures using honeypots. *Computer Communication Review* **34(1)**, 51–56 (2004).
14. Moore, D., Shannon, C., Voelker, G., and Savage, S. Internet quarantine: Requirements for containing self-propagating code. In *IEEE Infocom 2003*, (2003).
15. Ebel, H., Mielsch, L.-I., and Bornholdt, S. Scale-free topology of e-mail networks. *Physical Review E* **66**, 035103 (2002).
16. Newman, M. Spread of epidemic disease on networks. *Physical Review E* **66**, 016128 (2002).
17. May, R. M. and Lloyd, A. L. Infection dynamics on scale-free networks. *Physical Review E* **64**, 066112 (2001).
18. Kephart, J. and Arnold, W. C. Automatic extraction of computer virus signatures. In *The 4th Virus Bulletin International Conference 1994*, 179–194, (1994).
19. Huang, Z.-F. Self-organized model of information spread in financial markets. *Eur. Phys. J. B* **16**, 379 (2000).
20. Erez, T., Hohnisch, M., and Solomon, S. Statistical economics on multi-variable layered networks. In *Economics: Complex Windows*, Salzano, M. and Kirman, A., editors, 201–216. Springer (2005).
21. Palla, G., Derényi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818, 9 June (2005).
22. Malarz, K. Social phase transition in Solomon network. *Int. Journal of Mod. Phys. C* **14**, 561–565 (2003).

23. Chen, L.-C. and Carley, K. M. The impact of countermeasure propagation on the prevalence of computer viruses. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics* **34(2)**, 823–833 (2004).
24. Buchanan, M. Data-bots chart the internet. *Science* **308**, 813 (2005).
25. Shavitt, Y. and Shir, E. DIMES: Let the Internet measure itself. *ACM Computer Communication Review* **35(3)**, 71–74 (2005).

Acknowledgements This work was supported by a grant from Israel Science Foundation. E.S. was partially supported by the “Yeshaya Horowitz Association through the Center for Complexity Science”.

Author Information The authors declare that they have no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (email: shire@eng.tau.ac.il).

Figure 1 Comparing infection process evolution with (bottom) and without (top) immunization edges. On the top the network is being infected fully by the virus. On the bottom the virus cluster is reduced by more than half by introducing immunization edges. The blue (dark green) edges represent the original network (additional immunization) edges. During the spread, an edge is colored in red (light green) if it was used to infect (immunize) a node. In both cases we present four snapshots of each network in different times. In addition we present at the center the time varying graphs for the clusters development over time. The blue, red and green are used to present the size of the susceptible, infected, and immunized clusters respectively. Note that at the bottom set, initially in snapshots 1 and 2, the virus cluster develops uninterrupted until the immunization agent manages to escape the border of the virus cluster, in snapshot 3, and start immunizing the network and manages to immunize most of the network eventhough the virus had a head start.

Figure 2 Relative virus cluster size as a function of immunization links density (log-log scale). The dependence of the relative infected cluster size on the relative edge addition q , as resulting from simulations over uncorrelated, scale free networks with power exponent -3 , mean degree equal to 4, and network size ranging from 50000 to 200000 nodes. Ratio dependence exhibits a power law form, with an exponent close to $-4/3$. In this Figure and the ones that follow we use error bars to present the 95% confidence interval.

Figure 3 Relative virus cluster size as a function of system size for different honey pots densities. The simulations were ran over uncorrelated, scale free networks with power exponent -3 , mean degree equal to 4, and network size ranging from 25000 to

200000 nodes. The virus cluster is multiplied by the square of honey pots density to cancel its effect in accordance with equation 2. The curves show a power law dependence with exponent equals to -1.08 , compared to the expected -1 . Error bars are used as in Figure 2.

Figure 4 Relative virus cluster size as a function of honey pots density for different system sizes. The simulations were ran over uncorrelated, scale free networks with power exponent -3 , mean degree equal to 4, and network size ranging from 25000 to 200000 nodes. The virus cluster is multiplied by the network's size to cancel its effect in accordance with equation 2. The curves show a power law dependence with exponent equals to -1.8 , compared to the expected -2 . Error bars are used as in Figure 2.

Figure 5 Comparison of virus cluster sizes for the random edges and the honey pots architectures. The different sets present different immunization edge densities, q . The clusters in the random case are always larger than the clusters in the honey pots case; as the network size grows, so grows the gap between the two architectures. The reason being that while in the random case the cluster size remains fairly constant as we vary the network size, in the honey pots case, as the network grows, so grows the effectiveness of the honey pots. This effect is mostly noted in the middle-range of the density values where the immunization has an effect but the virus cluster is not extremely small. Error bars are used as in Figure 2.

Figure 6 Dependence of the virus cluster on the degree distribution power exponent. We have ran a sensitivity analysis where we varied the power exponent of the Pareto degree distribution characterizing the underlying topology between 1.8 and 3, which includes all degree distributions found in real scale free networks. As can be seen,

the effectiveness of the immunization process grows with the power exponent, due to the fact that lower exponents entail higher density of edges, which allows the virus to advance faster. However, this variation is still minor compared to variations in the immunization edges density, q , and in the architecture type, which are illustrated by the different data sets presented. Error bars are used as in Figure 2.