SANJAY GOEL AND STEPHEN F. BUSH

biological models of security for virus propagation in computer networks

Dr. Goel is an assistant professor in the School of Business and director of research at the Center for Information Forensics and Assurance at SUNY Albany. His research interests include distributed computing, computer security, risk analysis, biological modeling, and optimization algorithms.

goel@albany.edu
http://www.albany.edu/~goel

Dr. Bush is a researcher at GE Global Research. He continues to explore novel concepts in complexity and algorithmic information theory with a spectrum of applications ranging from network security and low-energy wireless ad hoc sensor networking to DNA sequence analysis for bioinformatics.

bushsf@research.ge.com
http://www.research.ge.com/~bushsf

THIS ARTICLE DISCUSSES THE SIMILARITY between the propagation of pathogens (viruses and worms) on computer networks and the proliferation of pathogens in cellular organisms (organisms with genetic material contained within a membrane-encased nucleus). It introduces several biological mechanisms which are used in these organisms to protect against such pathogens and presents security models for networked computers inspired by several biological paradigms, including genomics (RNA interference), proteomics (pathway mapping), and physiology (immune system). In addition, the study of epidemiological models for disease control can inspire methods for controlling the spread of pathogens across multiple nodes of a network. It also presents results based on the authors' research in immune system modeling.

The analogy between computers and communication networks and living organisms is an enticing paradigm that researchers have been exploring for some time. In 1984 Fred Cohen, in his Ph.D. dissertation, first put the term "computer virus" into print, although there he credits Len Adleman with coining the term used to describe the malicious pieces of code that can proliferate on a network and infect multiple computers. Since then, advances in bioinformatics (that is, the modeling of biological processes as well as storage, retrieval, and analysis of biological data through the use of information technology) have helped to define these analogies more precisely, to the point where results in bioinformatics can often be leveraged for use in computer networking and security. The challenges faced in bioinformatics are quite similar to those in computer network security. Several mechanisms have been devised in biological organisms to protect against pathogen invasion. It is important to learn from these biological phenomena and devise innovative solutions to protect computer systems from software pathogens.

Virus detection systems prevalent today are based on data analysis which looks for the presence of specific patterns. The data may be composed of header information in incoming packets at a firewall, data resident on a node, or behavioral patterns of programs resident on a computer. In most cases, the patterns of behavior (signatures) are defined a priori based on knowledge of existing pathogens. The signatures are usually gleaned from virus code by teams of virus experts who dissect

the code and identify strings that uniquely identify the virus. The signature database in virus detection programs becomes obsolete rapidly, as new virus strains are released, and is updated as these new viruses are discovered. However, with the speed of virus propagation increasing—as is evident from the spread of the Slammer worm, which infected more than 90% of vulnerable hosts in 10 minutes—this mechanism is proving inadequate to control the spread of viruses, with its consequent loss of data and services. It is imperative to develop new virus detection software that does not rely solely on external intervention but can detect new strains of viruses by organically generating "antibodies" within a node. The physiology of cellular organisms contains several paradigms that can be used as inspiration for developing such autonomous security systems in computer networks. Several streams of research on automatic detection of virus (and worm) signatures are in progress (Kim and Karp, 2004), but this research is still preliminary and not mature enough for commercial deployment.

One of the initial areas explored in the realm of biological models of computer security involves the work of Forrest et al. (1994) with regard to virus detection. Here the similarities are strikingly clear regarding the need to quickly and efficiently identify viruses, generate "antibodies," and remove them from the system before they cause damage and propagate throughout the system. Prior to this, Kauffman (1969) had been focused on understanding and modeling the mechanics of gene transcription and translation within the cell. The concept of a complex network of interactions describing gene regulation had been born in the form of the Boolean network model. Now that the human genome has been fully sequenced, the task of determining gene function is a significant focus. However, specific genes identified in the sequence can interact with other genes in complex ways. Some portions of the genome can turn off the expression of other genes. These portions are called the structural and regulatory genes. Their behavior is thought to be a defense against foreign sequences, perhaps passed on from ancient viruses, from being expressed and potentially harming the organism (Hood, 2004). In fact, in this mechanism one can draw upon concepts that apply directly to network security, namely, the idea of defensive code that can be inherently activated to turn off dangerous code or viruses within the network. One of the problems in virus protection systems is the result of false positives, when portions of the code that provide legitimate functionality may be turned off accidentally. The authors propose use of surrogate code that can replicate the functionality of the pieces of code that are shut off, maintaining continuity in the operations of the node. Specifically, fault-tolerant networks are capable of surviving attacks and dynamically reconstituting services. Bush (2003) explores the ability of a communication network to genetically constitute a service. The network service evolves in real time using whatever building blocks are available within the network. Thus, a service damaged by a virus attack may be genetically reconstituted in real time. The general concept was illustrated using a specific example of a genetic jitter-control algorithm which evolved a 100-fold decrease in jitter in real time.

Another biological paradigm which lends itself well to adaptation as a computer security paradigm is protein pathway mapping. Living organisms have complex metabolic pathways consisting of interactions between proteins and enzymes, which may themselves have multiple subunits, alternate forms, and alternate specificities. Molecular biologists have spent decades investigating these biochemical pathways in organisms. These pathways usually relate to a known physiological process or phenotype and together constitute protein networks. These networks are very complex, with several alternate pathways through the same start and end point. The partitioning of networks into pathways is, however, often arbitrary, with the start and finish points chosen based on "important" or easily understood compounds. The models for biochemical pathways that have been developed thus far primarily demonstrate the working of the cel-

lular machinery for specific tasks, such as metabolic flux and signaling. Several different modeling techniques are used: (1) classical biochemical pathways (e.g., glycolysis, TCA cycle); (2) stoichiometric modeling (e.g., flux balance analysis); and (3) kinetic modeling (e.g., CyberCell, E-Cell). More recently, cell metabolism is being studied using cellular networks that are defined from large-scale protein interaction and gene expression measurements.

Similar to the cellular networks in organisms, computer networks are complex in nature and collectively exhibit complex behavior. In these networks, start and end points can be arbitrarily chosen, and multiple paths may exist between the same nodes. Protein networks are predetermined and stay fairly static, whereas computer networks are constantly evolving with the addition of new nodes and network links. In protein networks, interactions among proteins, enzymes, and catalysts culminate in specific events. Analogously to protein networks, interactions among nodes of computer networks result in specific events or conditions in the network. The events may include propagation of viruses, denial-of-service attacks, and congestion on the network. Investigation of the network pathways along which the events propagate will enable us in forensic analysis to determine the root cause of the failures, as well as helping in developing intelligence for prediction of network events.

One biological paradigm that is not directly related to the physiology of living organisms is epidemiology that involves statistical analysis of disease propagation. Three basic models of disease propagation have been used extensively in epidemiological studies. Kephart and White (1991) first used these epidemiological models to study the spread of viruses on computer networks. Williamson and Léveillé (2003) have also developed virus spread models in computer networks using the epidemiological metaphor. Since then, several researchers have used variations of these basic models for studying the spread of computer viruses on computer networks.

The authors (Goel and Bush, 2003) have used the biological paradigm of the immune system, coupled with information theory, to create security models for network security. Information theory allows generic metrics and signatures to be created which transcend the specific details of a system or an individual piece of code. They compare information-theoretic approaches with traditional string-matching techniques. They also provide an analytic model that uses the epidemiological paradigm to study the behavior of the nodes. This article discusses several different biological paradigms which inspire defense against pathogens that invade computer networks, but it focuses on in-depth analysis of the immune system model. Some of the other innovative biological models that are currently being researched will be discussed in depth in a series of future articles.

Immune System Models

The role of the human immune system is to protect our body from pathogens such as viruses, bacteria, and microbes. The immune system consists of various kinds of cells, which operate autonomously and through interaction with each other to create complex chains of events leading to the destruction of pathogens. At a high level, cells can be categorized into two groups: detectors and effectors. Detectors identify pathogens, and effectors neutralize them. There are two kinds of immune responses evoked by the immune system: innate response and adaptive response. The innate immune response is the natural resistance of the body to foreign antigens and is non-specific toward invaders in the body. During this response, a specialized class of cells called phagocytes (macrophages and neutrophils) is used. These specialized cells, which have surface receptors that match many common bacteria, have remained unchanged throughout evolu-

tion. This system reacts nearly instantaneously to detect pathogens in the body. However, it is incapable of recognizing viruses and bacteria that mutate and evolve.

The innate immune response is complemented by the adaptive immune response, in which antibodies are generated to specific pathogens that are not recognized by the phagocytes. The adaptive response system uses lymphocytes, which have receptors for a specific strain instead of having receptors for multiple strains as phagocytes do. Lymphocytes are produced in the bone marrow, which generates variants of genes that encode the receptor molecules and mature in the thymus. When an antigen is encountered, it is presented to the lymphocytes in the lymphatic system. The lymphocytes that match proliferate by cloning and subsequently differentiate into B-cells, which generate antibodies, and T-cells, which destroy infected cells and activate other cells in the immune system. Most effectors that proliferate to fight pathogens die; only 5–10% are converted into memory cells which retain the signature of the pathogen that was matched. These memory cells permit a rapid response the next time a similar pathogen is encountered, which is the principle used in vaccinations and inoculations. The number of memory cells produced is directly related to the number of effector cells in the initial response to a disease. While the total number of memory cells can become quite large, still, as an organism is exposed to new pathogens, newer memory cells may take the place of older memory cells, due to competition for space (Ahmed, 1998). This decrease in memory cells leads to weakened immunity over time. Another reason for weakened immunity is an immune response rate that is not sufficiently rapid to counteract the spread of a powerful exotoxin, such as that produced by tetanus (Harcourt et al., 2004). Lymphocytes have a fixed lifetime, and if during this period they do not match a pathogen, they automatically die.

The key to the functioning of the immune system is detection. Recognition is based on pattern matching between complementary protein structures of the antigen and the detector. The primary purpose of the genetic mechanism in the thymus and bone marrow is to generate proteins with different physical structures. The immune system recognizes pathogens by matching the protein structure of the pathogen with that of the receptor. If the receptor of the antigen and the detector fit together like a three-dimensional jigsaw puzzle, a match is found. A fundamental problem with the detection mechanism of the immune system is its computational complexity. For example, if there are 50 different attributes with four different values, over six million different detectors are required to cover the search space. The number of virus structures that can arise due to different protein configurations is virtually infinite. In spite of high efficiency in creating detectors and pattern matching at the molecular level, maintaining a detector for each possible pathogen protein structure is not feasible. The human immune mechanism solves this problem by using generalizations in matching—that is, some features of the structure are ignored during matching. This is called specificity of match; the more features are ignored, the lower the specificity. The lower the specificity, the fewer the number of detectors required for matching a population of pathogens and the more nonspecific is the response. An explanation of specificity is elegantly described in J.H. Holland's description of classifier systems (1985). To cover the space of all possible nonself proteins, the immune system uses detectors with low specificity. This enables the immune system to detect most pathogens with only a few detectors; however, it results in poor discrimination ability and a weak response to pathogen intrusion. The immune system counters this problem by employing a process called affinity maturation (Bradley and Tyrrell, 2000). Several methods have been proposed for analytic representation of matching pathogen signatures in the immune system, such as bit-strings (Farmer, Packard, and Perelson, 1986; De Boer, Segel, and Perelson, 1992), Euclidean parameter spaces (Segel and

Perelson, 1988), polyhedron models (Weinand, 1991), and, more recently, Kolmogorov Complexity (Bush, 2002; Goel and Bush, 2003).

Several applications based on immune systems outside the area of biology have recently emerged, the most notable of these being computer security. Kephart (1995) was perhaps the first to introduce the idea of using biologically inspired defenses against computer viruses and immune systems for computer security. Forrest et al. (1994) also proposed the use of immune system concepts for design of computer security systems and provided an elaborate description of some immune system principles applicable to security. They presented three alternate matching schemes—Hamming distance, edit distance, and r-contiguous bits—arguing that the primary premise behind a computer immune system should be the ability to distinguish between self and non-self. They presented a signature scheme where a data tuple consisting of source IP address, destination IP address, and a destination port number were used to distinguish self-packets from non-self packets. Hofmeyr (1999) presented a detailed architecture of a computer immune system. He analytically compared different schemes for detection of pathogens, such as Hamming distance and specificity. There are several other works in the literature on the use of immune systems for network security, including Murray (1998), Kim and Bentley (1999), and Skormin et al. (2001). Kephart and White (1991, 1993) present an architecture for an immune system and the issues involved in its commercialization. They incorporate a virus analysis center to which viruses are presented for analysis through an active network. The Kolmogorov Complexity approach (Goel and Bush, 2003) demonstrated a 32% decrease in the time required to detect a signature over two common Hamming distance–based matching techniques, i.e., a sliding window and the number of contiguous bit matches. The Kolmogorov Complexity–based technique estimates the information distance of entire code sequences, not just specific segments or bits. Using the entire code sequence makes it more difficult to modify the virus so that it can hide in another portion of a legitimate code segment.

Artificial immune systems consist of detectors and effectors that are able to recognize specific pathogen signatures and neutralize the pathogens. To detect pathogens, the signature of incoming traffic packets is matched against signatures of potential viruses stored in an immune system database. An immune system that is capable of recognizing most pathogens requires a large number of detectors. Low-specificity detectors that identify and match several viruses are often used to reduce the number of detectors at the cost of increased false positives. The computational complexity of a computer immune system remains fairly high, and individual nodes are incapable of garnering enough resources to match against a large signature set. The computational complexity gets worse as network traffic grows due to use of broadband networks, and it is straining the capacities of conventional security tools such as packet-filtering firewalls. Massive parallelism and molecular-level pattern matching allow the biological immune system to maintain a large number of detectors and efficiently match pathogens. However, artificial immune systems have not achieved these levels of efficiency. To reduce the computational burden on any individual node in the network, all nodes need to pool their resources, share information, and collectively defend the network. In addition, such inspection should be done within the network itself, to improve efficiency and reduce the time required for reacting to an event in the network. This concept of collective defense enabled by a unified framework is the primary premise of the authors' research. To enable this concept of collective network defense, they have proposed an approach based on information theory principles using Kolmogorov Complexity measures.

To study the parameters and different schemes of detection and sampling in the immune system, Goel et al. (working paper) have developed a simulation model using RePast (Schaeffer et al., 2004), a simulation tool typically used for model-

ing self-organizing systems. The simulation models a classical immune system, where new signatures are created by mutation of existing signatures which then go through a maturation phase. The simulation also models a cooperative immune system, where multiple nodes on the network share virus detection information prevalent in the network to improve the efficiency of each immune system. The research will investigate the trade-off between the additional burden of sharing information across nodes and the benefit of improving scanning efficiency by obtaining intelligence information on active or new pathogens. Figures 1a and 1b show the impact of the match threshold and sampling rate, respectively, on the performance of the immune system. Figure 1a shows a high gradient between a threshold match of 0.2 and 0.4, which is the practical operating region for the immune system. Figure 1b shows an improved performance with the sampling rate, which asymptotes around 70%.

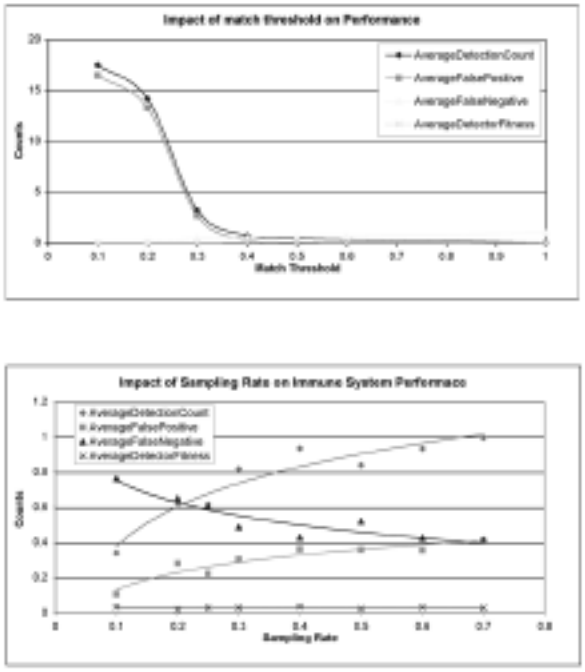

Figure 1. Plots showing impact of match threshold and sampling rate on immune system-metrics

Goel and Bush (2003) have also compared different signature metrics and have demonstrated that Kolmogorov Complexity is a feasible metric for the signature of pathogens.

Conclusion

The security models for detection and elimination of pathogens that invade computer networks have been based on perimeter defense. Such defenses are proving inept against fast-spreading viruses and worms. The current tools are unable to guarantee adequate protection of data and unfettered access to services. It is imperative to complement these existing security models with reactive systems that are able to detect new strains of pathogens reliably and are able to destroy them before they can cause damage and propagate further. Several biological paradigms provide a rich substrate to conceptualize and build computer security models that are reactive in nature. Three specific mechanisms in mammalian organisms present the most potential: (1) the RNAi mechanism, (2) protein pathway mapping, and (3) the immune mechanism. In addition, the models

of disease control that study the spread and control of viruses suggest ways to throttle the spread of viruses. Current work has mainly focused on the use of immune and epidemiological models. It is time to move beyond these existing models to other innovative models, such as those based on genomics and proteomics. Such reactive models provide a scalable, resilient, and cost-effective mechanism that may keep pace with constantly evolving security needs.

Acknowledgments

The authors are grateful to Damira Pon, from the School of Information Science at the State University of New York at Albany, for a thorough review of this article and her useful suggestions.

REFERENCES

Ahmed, R., February 5, 1998. "Long-term Immune Memory Holds Clues to Vaccine Development," Emory Health Sciences Press Release.

Bradley, D.W., and Tyrrell, A.M., April 2000. "Immunotronics: Hardware Fault Tolerance Inspired by the Immune System," *ICES 2000* (Springer-Verlag, 2000), pp. 11–20.

Bush, S.F., 2002. "Active Virtual Network Management Prediction: Complexity as a Framework for Prediction, Optimization, and Assurance," *Proceedings of the 2002 DARPA Active Networks Conference and Exposition (DANCE 2002)* (Los Alamos, CA: IEEE Computer Society Press), pp. 534–553: http://www.research.ge.com/~bushsf/ftn/005-FINAL.pdf.

Bush, S.F., 2003. "Genetically Induced Communication Network Fault Tolerance," *Complexity Journal*, vol. 9, no. 2, Special Issue: "Resilient & Adaptive Defense of Computing Networks": http://www.research.ge.com/~bushsf/pdfpapers/ComplexityJournal.pdf.

Cohen, F., 1987. "Computer Viruses Theory and Experiments," *Computers and Security*, vol. 6, pp. 22–35.

De Boer, R.J., Segel, L.A., and Perelson, A.S., 1992. "Pattern Formation in One- and Two-Dimensional Shape-Space Models of the Immune System," *J. Theor. Biol.*, pp. 155, 295–333.

Farmer, J.D., Packard, N.H., and Perelson, A.S., 1986. "The Immune System, Adaptation, and Machine Learning," *Physica D*, vol. 22, pp. 187–204.

Forrest, S., Perelson, A.S., Allen, L., and Cherukuri, R., 1994. "Self–Nonself Discrimination in a Computer," *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy* (Los Alamos, CA: IEEE Computer Society Press).

Goel, S., and Bush, S.F., 2003. "Kolmogorov Complexity Estimates for Detection of Viruses in Biologically Inspired Security Systems: A Comparison with Traditional Approaches," *Complexity*, vol. 9, no. 2: http://www.research.ge.com/~bushsf/pdfpapers/ImmunoComplexity.pdf.

Goel, S., Rangan, P., Lessner, L., and Bush, S.F., [working paper]. "Collective Network Defense: A Network Security Paradigm Using Immunological Models."

Harcourt, G.C., Lucas, M., Sheridan, I., Barnes, E., Phillips, R., Klenerman, P., July 2004. "Longitudinal Mapping of Protective CD4 T Cell Responses Against HCV: Analysis of Fluctuating Dominant and Subdominant HLA-DR11 Restricted Epitopes," *Journal of Viral Hepatitis*, vol. 11, no. 4, p. 324.

Hofmeyr, S.A., May 1999. "An Immunological Model of Distributed Detection and Its Application to Computer Security," Ph.D. thesis, University of New Mexico.

Holland, J.H., 1985. "Properties of the Bucket Brigade Algorithm," *Proceedings of the 1st international Conference on Genetic Algorithms and Their Applications*, ed. Grefenstette, J.J., L.E. Associates, pp. 1–7.

Hood, E., 2004. "RNAi: What's All the Noise About Gene Silencing?" *Environmental Health Perspectives*, vol. 112, no. 4.

Kauffman, S.A., 1969. "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets," *J. Theor. Biol.*, vol. 22, pp. 437–467.

Kephart, J. O., 1995. "Biologically Inspired Defenses Against Computer Viruses," *Proceedings of IJCA '95*, pp. 985–996.

Kephart, J.O., and White, S.R., 1991. "Directed Graph Epidemiological Models of Computer Viruses," *Proceedings of the 1991 IEEE Computer Security Symposium on Research in Security and Privacy*, pp. 343–359.

Kephart, J.O., and White, S.R., May 1993. "Measuring and Modeling Computer Virus Prevalence," *Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, pp. 2–15.

Kim, H.-A., and Karp, B. 2004. "Autograph: Toward Automated, Distributed Worm Signature Detection," *Proceedings of the 13th USENIX Security Symposium*, pp. 271–286.

Kim, J., and Bentley, P., 1999. "Negative Selection and Niching by an Artificial Immune System for Network Intrusion Detection," *Late-Breaking Papers at the 1999 Genetic and Evolutionary Computation Conference (GECCO '99),*, pp.149-158.

Murray, W.H., 1998. "The Application of Epidemiology to Computer Viruses," *Computer Security*, vol. 7, pp. 139–150.

Schaeffer, S.E., Clemens, J.P., and Hamilton, P., 2004. "Decision Making in a Distributed Sensor Network," *Proceedings of the Santa Fe Institute Complex Systems Summer School:* http://www.tcs.hut.fi/~satu/online-papers/sensor.pdf.

Segel, L.A., and Perelson, A.S., 1988. "Computations in Shape Space: A New Approach to Immune Network Theory," in *Theoretical Immunology Part 2*, ed. Perelson, A.S. (Redwood City: Addison-Wesle)y, pp. 377–401.

Skormin, V.A., Delgado-Frias, J.G., McGee, D.L., Giordano, J.V., Popyack, L.J., Gorodetski, V.I., and Tarakanov, A.O., 2001. "BASIS: A Biological Approach to System Information Security," *Mathematical Methods, Models, and Architectures for Network Security Systems (MMM-ACNS) 2001*, pp. 127–142.

Weinand, R.G., 1991. "Somatic Mutation and the Antibody Repertoire: A Computational Model of Shape-Space," in *Molecular Evolution on Rugged Landscapes*, ed. Perelson, A.S. and Kaufman, S.A., *SFT Studies in the Science of Complexity*, vol. 9 (Redwood City: Addison-Wesley), pp. 215–236.

Williamson, M.M., and Léveillé, J., 2003. "An Epidemiological Model of Virus Spread and Cleanup": http://www.hpl.hp.com/techreports/2003/HPL-2003-39.html.