Comment

Network security and data integrity in academia: an assessment and a proposal for large-scale archiving

Andrew Smith*†, Dov Greenbaum‡, Shawn M Douglas*, Morrow Long§ and Mark Gerstein*¶†

Addresses: *Department of Molecular Biophysics and Biochemistry, †Department of Computer Science, §Information Technology Services and ¶Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. ‡University of California, Berkeley, CA 94720, USA.

Correspondence: Mark Gerstein. E-mail: mark.gerstein@yale.edu

Published: X Month 2005

Genome ***Biology*** 2005, **6:**119 (doi:10.1186/gb-2005-6-9-119)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/9/119

Academic scientific research, particularly in genomics, is becoming increasingly dependent on computers, networks, and online databases. The future will see a continuing increase in the number and importance of new discoveries and insights gained via computational analyses of large datasets rather than through direct experimentation in a laboratory. Moreover, integrated analysis of multiple distributed databases will become increasingly important as the number of online scientific resources continues to rise exponentially (Greenbaum *et al., Nat Biotechnol* 2004, **22**(6):771-772): the whole is definitely greater than the sum of its parts.

A direct impediment to the optimal use of online databases and their interoperation is the increasing prevalence, severity, and toll of computer and network security incidents. The security problem is more common and invasive than commonly thought. A recent experiment [http://www.usatoday.com/money/industries/technology/2004-11-29-honeypot_x.htm] was conducted using 'honeypots' - computers and networks specifically set up to be attacked in order to collect information on the frequency and types of attack. In addition to validating the protection that firewalls and regular updates of operating systems can provide (options which are regrettably often neglected), the most interesting finding of this study is the prevalence and frequency of attacks. A computer is almost guaranteed to be the target of incessant and recurrent attacks within minutes of being connected to the internet. The various honeypots in the study were attacked anywhere from 2 to 341 times per hour. Other previous studies by the HoneyNet project [http://www.honeynet.org] have reached similar conclusions [http://www.schneier.com/crypto-gram-0106.html#1].

To highlight the security problem particularly for academic genomics research, consider actual intrusion-detection data provided by the network intrusion detection system SNORT [http://www.snort.org] that cover the first 198 days of 2002 for a server that hosts a number of commonly used, publicly available genomics databases and that we feel is typical of academic genomics server setups. SNORT works by checking incoming packets of network data against a large database of likely attack patterns; packets that match patterns in the database are flagged for the system administrator and written into logfiles for subsequent analysis. Figure 1a shows graphs of daily event counts over the first 198 days of 2002. These data echo the honeypot results: attempted attacks are a daily occurrence, with usually around 6-10 per day. In addition, there is wide variability with some days showing concerted attempted attacks resulting in hundreds or even thousands of events. Figure 1b,c shows a breakdown of specific event types. The frequency and nature of attacks are unpredictable: on two days that showed a massive spike in events, a single event type accounted for over 90% of events, while for another sequence of days with fewer, but more consistent, attacks no single event dominated the SNORT data. The details of these event types can be found at the SNORT website [http://www.snort.org]. While intrusion-detection systems such as SNORT can trigger false positive 'events', it seems likely that at least the most common "SHELLCODE x86 inc ebx NOOP" events are real attempted attacks, given that they are used for buffer overflow attacks in attempts to gain control over machines. (With such buffer overflow attacks attempts are made to write past the legal boundaries of allocated computer memory; these are exceptional events whose consequences might be exploitable by an attacker.)

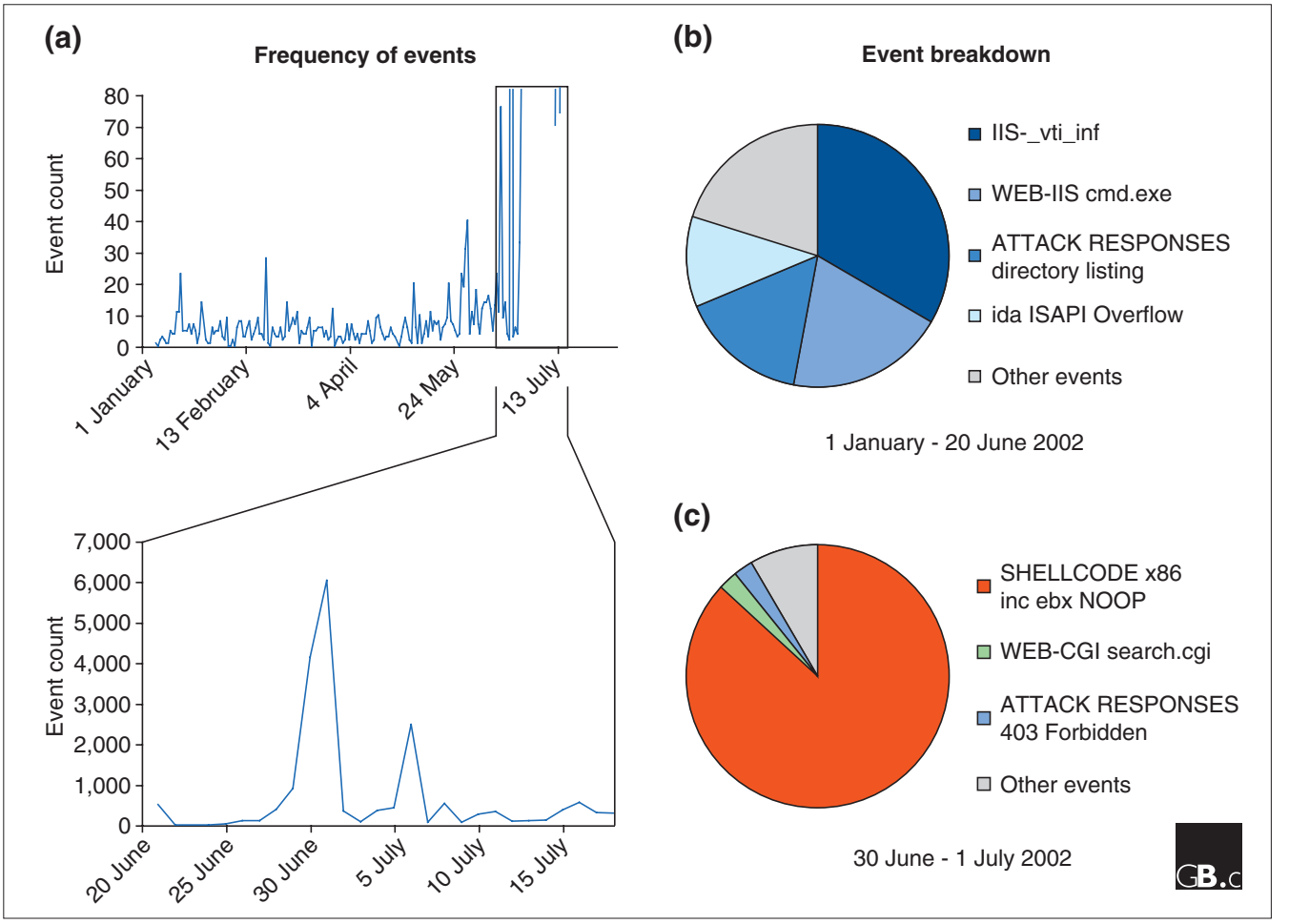

Figure 1
The frequency of security events on a typical genomics server. **(a)** A plot of daily security-event counts for the first 198 days of 2002; the expanded region had a large increase in daily counts. Attack attempts are an everyday occurrence and there can be large spikes in attack activity. **(b,c)** Aggregate breakdown and relative proportions of the most common security events for, (b) days with small, regular event counts or (c) two days showing a massive spike in events as evident on the graph in (a). For the two days with the massive spike a single event type "SHELLCODE x86 inc ebx NOOP", which is used in buffer overflow attacks (attacks that attempt to write past the legal boundaries of allocated computer memory) and thus is likely to represent real and serious attack attempts, accounts for over 90% of events. For the more regular days there is no single dominating event, and it is not clear whether these events are genuine attack attempts.

If attempted attacks are incessant, successful attacks are also relatively common. The clearest demonstration of this is the often-publicized and seemingly endless stream of widely propagated email virus and worm attacks. While the majority of these viruses and worms do not cause any data loss, with many simply written for the virus- or worm-writer's amusement or to enable the writer to use other people's computers for 'spam relays', it is worth noting that any successful attack gives access to the compromised computer, potentially allowing all files to be compromised or erased. The potential for real data loss is great; we should not be lulled into a false sense of security simply because hackers do not often take advantage of their attacks. There is a large but stratified hacker society and culture with just a few very technically knowledgeable and skilled hackers dedicated to ferreting out new exploits; these leaders then package up their exploits into easily executable cracking programs and make them available to legions of novice but eager 'script kiddies' who help enact large and widespread attacks. And there have been real and significant cases of data loss. *The New York Times* recently reported on a successful attack involving hundreds of computers in government and academic research labs ('Internet Attack Called Broad and Long Lasting by Investigators', *New York Times Online*, 5 October 2005 [http://nytimes.com]) that is a perfect example of why we need something like the recommendations made below. While the extent of the attack and any data loss is still being investigated, a geophysics graduate student at University of California, Berkeley had all her files and many emails erased by this hacker. In another incident at UC Berkeley a bioinformatics lab was successfully hacked. Key data on several machines were erased and permanently lost; the only backup

was weeks old and progress was significantly impeded. Finally, on the other side of the USA, the Dana Farber Cancer Institute in Boston had a high-throughput sequencing machine successfully hacked, and data files and programs were deleted; these were, fortunately, recovered from backup (for details see [http://research.dfci.harvard.edu/ news.html#hack]).

There are common and effective lines of defense, such as firewalls and antivirus software, but "security is a process, not a product" (Schneier B: *Secrets and Lies: Digital Security in a Networked World*. New York: John Wiley; 2000): the most important parts of a solution are vigilance, good policy and planning, and attention to detail in a three-pronged strategy of prevention, detection, and response. Fortunately, academia has it somewhat easier than the military, government, and business, where security is generally a very serious business - because of the free and open nature of academia, the key requirement is not to prevent unauthorized access at all costs, but rather to maintain the integrity and robustness of data and scientific results and to ensure this for posterity. We feel that the open nature of academic genomics research, analogous to the open-source software movement, makes it possible to make use of cooperative economies of scale to deal effectively and efficiently with security.

Academia needs to explore the specifics and scale of how best to aggregate security expertise, personnel, and resources for the use and benefit of all. We believe there is great potential in aggregation and we offer the following to demonstrate the possibilities of what might be called 'Open Genomics'. Funding agencies such as the National Institutes of Health (NIH) should set up working groups dedicated to computer/network security issues, and should provide aid in this area to government-grant-funded members of the academic community. We believe there are many positive things such a working group could do, such as: provide security guidelines, help documentation, and possibly even Linux distributions, tailored specifically to the genomics community; provide custom and third-party security scripts/programs, such as hardening scripts from the Bastille Linux project [http://www.bastille-linux.org]; setup and monitor intrusion-detection systems such as SNORT or via honeypots/honeynets and/or perform security scans using programs such as Nessus [http://www.nessus.org] and SARA [http://www-arc.com/sara/] on community members' machines, allowing community-wide attack patterns to be detected; provide central hosting; and provide central authentication, enabling distributed collaborations. Finally, and most importantly, one can never fully prevent successful attacks and it must be assumed that in the worst case everything will be lost. Ultimately a security solution must be to do the best you can to secure your computing infrastructure, but, more importantly, to perform regular and redundant backups that can be quickly and efficiently restored. Thus universal backup, archival storage, and mirroring of community resources are the most essential services such a working group can provide, consistent with the key goal of security in academia: to preserve data and results for posterity.

While it might seem a daunting task regularly to backup all online genomics resources, in fact this is realizable today without excessive difficulty or cost. Pointing the way are sites such as Google [http://www.google.com], which maintains a cache of the most recent crawling of most pages it indexes, and the Internet Archive [http://www.archive.org], which goes further and maintains an archive of the web's pages at different time points, thus allowing one to view the history of particular sites and how they have evolved over time. Since most or all genomics resources are web-accessible, a similar webcrawler-based solution could provide a simple, 'rough-and-ready' way to backup genomics resources. This would require a lot of storage space, but space is relatively cheap today: for example, Google offers free email accounts with 2 gigabytes of storage to anyone. One problem, however, is that most scientific webpages are not static but are generated dynamically by programs accessing databases for user-submitted forms. It is estimated this so-called 'hidden web' is 500 times the size of the static web, and it is challenging to crawl it, but there is some research and products addressing this that could be leveraged (Mostafa J: *Seeking better Web searches*. *Sci Am* 2005, **292:**67-73.)

The ideal solution would be to backup the databases and programs used to generate content, or even better, the entire 'virtual machine' if virtualization software such as Vmware [http://www.vmware.com] or Xen [http://www.cl.cam.ac.uk/Research/SRG/netos/xen/] were used; from these any site's full functionality can be reproduced. Given that these programs and files in which data are stored are generally not directly web-accessible, this would involve more user intervention than simply having your site crawled. But, the proposed working group could create custom-configurable scripts that users could install in their web server's executable content area; these scripts would authenticate incoming connections, only executing if the request comes from the working group's crawlers, and then run code to dump the database's content to a text file, and send it, as well as the site's programs, to the working group's computer for backup. Finally, any such backup webcrawlers would need to know which sites to crawl, and there are various ways this could be determined, such as crawling any sites associated with PubMed [http://www.ncbi.nlm.nih.gov/Entrez] records or having some kind of registration system whereby researchers could, after authentication, register their site to be crawled and backed up. Possibly systems such as DSpace [http://www.dspace.org], which has been used to create a digital archive of research documents at Massachusetts Institute of Technology (MIT) [http://dspace.mit.edu], and the various 'web services' technologies, such as Universal Description, Discovery and Integration (UDDI) [http://www.uddi.org/] and Simple Object Access Protocol

(SOAP) [http://www.w3.org/TR/soap/], could be used to help enable this.

In this short article, we have shown that attempted and successful network attacks are fairly common in academic scientific research. We suggested that the open nature of academia allows it to address security issues cooperatively and that funding agencies should set up working groups to provide security services for the community. Finally, we have suggested that a large-scale backup system be set up to archive academia's digital data and to ensure its integrity for posterity. It is important to note that most of the data and information about scientific research is now primarily stored digitally; gone are the days when archiving of scientific research amounted solely to physically storing lab notebooks and copies of old journals. Digital data are more ephemeral and, because modern technology allows information to be generated at a much faster rate, the scale of digital information is vastly greater than physically printed information ever was – a fact that is causing headaches for the US government's National Archives and Records Administration (NARA), responsible for maintaining archives of all government correspondence and records (Talbot D: *The Fading Memory of the State*. *Technology Review*, July 2005 [http://www.technologyreview.com/articles/05/07/issue/feature_memory]). Nevertheless, we feel it is more important than ever to archive the record of scientific progress in order to avoid the curse that "those who cannot remember the past are condemned to repeat it" (Santayana G: *Life of Reason*. Scribner's; 1905:284). We sincerely hope that governments and funding agencies give serious attention to the issues raised in this article and implement systems for addressing them along the lines suggested herein.

Acknowledgements

This work was supported, in part, by NIH/NHGRI Centers of Excellence in Genomic Science grant P50 HG02357-01.