Salvatore J. Stolfo

Salvatore J. Stolfo is Professor of Computer Science at Columbia University. He received his Ph.D. from NYU Courant Institute in 1979 and has been on the faculty of Columbia ever since. He won an IBM Faculty Development Award early in his academic career in 1983. He has published several books and well over 200 scientific papers since then, several winning best paper awards, in the areas of parallel computing, AI knowledge-based systems, data mining and most recently computer security and intrusion detection systems (see www.cs.columbia.edu/ids). He has been granted 26 patents in the areas of parallel computing and database inference and computer security; most have been licensed or sold. His research has been supported by DARPA, NSF, ONR, NSA, CIA, IARPA, AFOSR, ARO, NIST, DHS and numerous companies and state agencies over the years while at Columbia. Professor Stolfo has mentored over 30 PhD students (26 have graduated to date) and many dozens of Master's students. His most recent research is devoted to payload anomaly detection for zero-day exploits, secure private querying, private and anonymous network trace synthesis for Predict.org, Symbiotic embedded machines, automatic bait generation for trap-based defense to mitigate the insider threat and he recently conducted a study in the area of multi-core parallel computing.

The following describe a number of systems and algorithms he invented and some of his professional activities over the years as a professor or as a consultant to industry or the US government. Stolfo's activities conducted in confidential settings is not reported here.

ACE Expert System: the First Deductive Database System and Application

Among his earliest work, Stolfo along with colleague Greg Vesonder of Bell Labs, developed a large-scale expert data analysis system, called ACE (Automated Cable Expertise) for the nation's phone system. ATT Bell Labs distributed ACE to a number of telephone wire centers to improve the management and scheduling of repairs in the local loop. ACE is likely to have been the first system to combine rule-based inference (an AI expert system) with a relational database management system, the ATT CRAS system, and serves as a model for deductive data base systems that were the subject matter of research for many years in the data base community. ACE was the first expert system of its kind that was commercialized and widely distributed.

Merge/Purge, De-duplication of large datasets

In other work related to the "merge/purge" problem, an algorithm developed by him and student Mauricio Hernandez has been used in large-scale commercial systems for data cleansing. Identifying and purging duplicates from large data sets is a very important part of large-scale data analysis systems, especially in commercial data analytics. The algorithms invented provided a means of scaling to very large data sets while balancing the requirement to produce accurate results in the presence of arbitrary noise and error in the data base. The patented technology was licensed by Informix, a company that was later acquired by IBM.

Improved Credit Card Fraud Detection

Stolfo consulted to the CTO of Citibank for several years and conducted research on machine learning algorithms applied to the credit card fraud problem. Much of that work with students Phil Chan and Andreas Prodromidis published as "meta-learning"-based strategies, demonstrated how to improve the accuracy of fraud detectors and substantially reduce loss due to fraud. This work is among the first examples of privacy-preserving, distributed data mining.

April 2011

DADO Parallel Computer

Stolfo and students Dan Miranker, Mike van Biema, Alexander Pasik and Steve Taylor, designed the architecture and software systems for the DADO parallel computer, an example "fifth generation computer" sponsored by DARPA's high performance parallel computing initiative in the mid-1980's. The DADO research group designed and built in a lab at Columbia University a fully functional a 1023-processor version of the machine that was the first parallel machine providing large-scale commercial speech recognition services. The DADO occupied about 2 cubic feet of cabinet space. The DADO was tested at sea in a Navy research vessel to test its capabilities for related acoustic analyses and detection capabilities. A parallel broadcast and resolve/report function introduced by the DADO machine apparently influenced part of the design of the IBM Blue Gene parallel computer.

The DADO technology was the first invention claimed by Columbia University for ownership of a faculty member's intellectual property under the 1980 Bayh-Dole Act. A company called Fifth Generation Computer was formed by Columbia and outside investors to commercialize the DADO machine. The company subsequently developed a commercially deployed speech recognition system operated by Qwest.

Sometimes important inventions create adversarial relationships. Unfortunately, a dispute between the small company and a large telecommunications provider and Columbia University caused a six year detour into the US court system where ultimately Stolfo prevailed and "the parties resolved the dispute that was the subject of the litigation". (Stolfo often advises young faculty to avoid complex business relationships that are not well managed by all the parties. Six years of litigation is not a productive use of a young faculty member's time and effort, especially when little to no support, financial or otherwise, was provided by any of the parties who created the confrontation and originated the disputes. When he retires from Columbia, Stolfo plans to complete a book that will explain the details of the original landmines, explain how to avoid irrational confrontation between management of companies and universities, and hopefully show how a pathway from chaos to success can be achieved.)

PARULLEL: Parallel Rule Processing

Early work in deductive databases, as represented by early systems such as ACE, required scaling inference processes to large distributed data sets, and maintaining the logical consistency of the distributed inference process. Stolfo and student Hasan Dewan and colleague Ori Wolfson of UI Chicago, invented a number of algorithms to execute parallel rule evaluation against distributed data sets. In this work, the copy-and-constrain algorithm studied by student Alexander Pasik provided a formal means of efficiently partitioning data and allocating rule processing to balance the load across distributed sites.

Data mining-based Intrusion Detection Systems

Stolfo's Intrusion Detection System (IDS) lab, established in 1996 and sponsored by DARPA's Cyber Panel program, pioneered the use of data analysis and machine learning techniques for the adaptive generation of novel sensors and anomaly detectors for a variety of tasks in computer security. One of Stolfo's papers co-authored with student Wenke Lee was identified as one of the most influential papers of the IEEE Security and Privacy Symposium over the last 30 years. Another paper won runner-up best paper award at the SIG KDD conference. To date this line of work has been cited and referenced by many hundreds of researchers and papers.

KDD CUP Data set

The DARPA IDS evaluation datasets were constructed by Lincoln Labs in 1998 and 1999 for the DARPA Cyber Panel program. These network trace data sets were used to evaluate the performance of different intrusion detection systems; they were the only network trace data with ground truth available to the open research community. The data, however, were difficult to use directly by a wider community of data mining researchers. Stolfo and his associates in the IDS lab including Wenke Lee created the KDD Cup dataset derived from the DARPA IDS datasets. The DARPA network trace data were converted to "connection records" making the data more suitable for data mining researchers to test various machine learning algorithms. This data created as a community service is extensively used in IDS research, even today.

Email Mining Toolkit (EMT)

The EMT system sponsored by DARPA contracts was among the first machine learning system to incorporate social network analyses in important security problems, including spam detection and virus propagation. The extensive set of analyses in EMT, developed by Stolfo and student Shlomo Herskhop and others, allowed analysts, forensics experts, students and researchers the opportunity to explore large corpora of email messages and discover a wide range of important derivative knowledge about the communication dynamics of a user or an organization. Among its applications, EMT models user behavior to identify uncharacteristic email flows indicative of spam bots and viral propagations. The toolkit has been downloaded by well over a 100 users and elements of the analyses introduced by EMT serve as a model for other analytical systems. The entire body of analyses demonstrated a general description of all IDS network and communication analysis systems conveniently described by the acronym, CV^5.

CV^5

Stolfo coined the term CV^5, meaning **C**orrelation of **V**iolations of **V**olume, **V**elocity, **V**alues and **V**ertices to serve as a general model of how IDS and network analysis systems operate. Features are extracted from temporal streams of network, host, user, or communication data and that are modeled to identify anomalies or suspicious events. The statistical features that are modeled are volume (amounts of data, or number of events) and velocity (rates or speed), values (content of messages, or network packets datagrams), or vertices (connectivity of endpoints when participants in communication are considered nodes in a graph). The term "violations" refers to an unexpected value of the computed model of the system. When a sufficient number of violations are correlated (across sensors, across sites, or across layers of a system) an IDS will typically alert. Stolfo's course on Intrusion and Anomaly Detection Systems taught at Columbia University explores the many different audit sources, features, and IDS systems that are each described as a specific instance of the **CV^5** framework.

Anomaly Detection: Algorithms, Sensors and Systems, Payl, Anagram, Spectrogram

A number of anomaly detection algorithms have been invented in Stolfo's lab with students Eleazar Eskin, Ke Wang, Janak Parehk, Yingbo Song and Gabriela Cretu, and have been deployed in commercial products licensed by Columbia University. Some of the mature content-based anomaly detectors, Payl and Anagram, have been deployed by the government in critical systems. Payl has been directly licensed and is in use in certain products. A great deal of work in the IDS lab focused on making AD systems practical and easy to deploy and use, and to essentially debunk the "folk-theorem" that AD systems generate too many false positives to be useful. The bulk of this work is reported in Cretu's

thesis describing a system called STAND. Recent and ongoing work with student Nathaniel Boggs, and joint with colleague Angelos Stavrou at GMU, have demonstrated the utility of a system, called AutoSense. The algorithms and distributed systems operate across two sites on the internet. Many of the AV technologies in wide use employ elements of anomaly detection technologies at their core.

Worminator

Stolfo was an early proponent of collaborative security and distributed IDS technology and systems. Stolfo and students Ke Wang and Janek Parehk developed a fully functional IDS alert exchange system that introduced a new means of sharing sensitive data in a privacy-preserving manner. The technique involved communicating network packet content found to be anomalous or verified as an attack after converting the raw packet content into a statistical representation allowing accurate correlation of common attacks across sites. The method invented by Stolfo and students to share and correlate content across administrative domains without disclosing sensitive information introduced the use of Bloom filters storing n-gram content of network packet datagrams. The method was extensively studied and continues to be used in several ongoing experiments. The method also formed the basis of a recent project with colleagues Steve Bellovin and Tal Malkin for the secure querying of encrypted document databases without requiring the insecure decryption of any document when searching for relevant content.

SPARSE – Statistical Parsing of Document Content

Just a few years ago most users felt entirely safe in exchanging Adobe PDF documents in emails and file shares. Other common file formats are often considered dangerous and are typically filtered by email service providers since they may harbor malware. Today, PDF's are as dangerous as Word documents, and zip files. Malware-laden documents are pervasive on the internet today. Stolfo and student Wei-Jen Li created a system that parses the binary format of Word documents and extracts and models statistical features based on n-gram analysis of the large number of object types embedded in Word documents. The resultant system was tested by a red team and demonstrated a number of techniques to identify suspicious documents, and prevent malware exploitation.

Application Communities

Stolfo invented with Angelos Keromytis and colleagues at Columbia the concept of "Application Communities" and assisted DARPA in running a workshop on that topic that ultimately led to the DARPA research program of the same name. The patent pending concept involves profitably using a monoculture of application as a "security sensor grid" to improve the security of all community members. The method and system is designed to rapidly generate a patch and update all hosts when an attacks is identified and validated by the initial victims, a small set of members of the application community. (Oddly enough, the Director of DARPA at that time did not select Stolfo's team proposal for funding giving other groups and organizations contracts to develop the technology Stolfo and Keromytis invented. Such was the possible outcomes of DARPA proposals from academic institutions in the "Tether era.")

The Insider Threat: RUU?

The insider threat remains the most vexing of all security problems. In 2005 Stolfo received funding from ARO to conduct a workshop to bring together a group of researchers to help identify a research program to focus on this important topic. Since then the IDS group at Columbia working with other researchers at the I3P developed several demonstration systems integrating a number of host and network sensors to provide

evidence of insider malfeasance. The work includes user profiling techniques (especially for masquerader detection. "RUU" is a spoken acronym for Are You You?) studied by Stolfo and student Malek Ben Salem, and a number of decoy generation facilities studied jointly with co-PI Angelos Keromytis and student Brian Bowen. A recent paper on the RUU project won best paper award. The effort has developed a number of interesting publicly available data sets for user studies, and novel "advanced behavioral sensors", including a decoy generator system, accessible through a public website, that produces decoy documents with embedded beacons and a relatively well developed theory of describing and measuring the properties of a decoy to guide the generation system.

Symbiotic Embedded Machines (SEM) and Insecure Embedded Systems

Student Ang Cui working with Stolfo in the IDS lab invented a concept to embed arbitrary code into legacy embedded devices. The symbiotic embedded machine technology has been demonstrated to provide a direct means to inject security features into operational CISCO IOS routers *in situ* without any significant performance degradation and without any negative impact on the routers primary function. The Symbiote technology is being explored for use in a number of different platforms and devices (ARM, X86, MIPS) and several interesting applications, especially for a large set of existing insecure embedded devices found on the internet. This line of work is supported by the DARPA CRASH program that has brought together a very large number of computer science researchers focused on clean slate design for a new generation of safe and secure computer systems. Preliminary work performed by Cui and Stolfo in the IDS lab performed a wide area scan of the internet counting the number of vulnerable devices. To date over 1.1 million have been found. The results were recently published and won best paper award.

Professional Activities

FSTC: Stolfo, as a consultant to Citicorp at that time, assisted Dan Schutzer, an executive at Citibank, in the formation of the Financial Services Technology Consortium (FSTC). The FSTC recently merged with the FS Roundtable organization, the nation's leading representative organization of the US banking and financial services industry.

Visa 3D Secure: Stolfo also served as a member of a committee created by VISA to recommend new online mutual authentication measures for more secure transactions.

Digital Government: With Herb Schorr of USC/ISI and with support from NSF he led the development of the Digital Government research community that ultimately led to the dg.o yearly conference. ISI and Columbia CS formed a collaboration that lasted several years.

Expert Witness: Stolfo was an expert witness in the DOJ versus Microsoft "browser wars" case having filed an *amicus brief* to the court. Stolfo assists law firms in various litigation and intellectual property disputes.

Government contracting: Stolfo as a consultant for several large defense contractors has played a leading scientific role in a number of projects in the area of cyber security and parallel computing.

Government advisory positions

DARPA: Stolfo served on the Futures Panel as a consultant to the Director of DARPA's IPTO Office (now called I20). He also participated in various DARPA ISAT study groups, and other DARPA invitational workshops.

National Academy: Stolfo served on the National Research Council Naval Studies Board subcommittee on Information Assurance for Network-Centric Naval Forces.

NCDI: Stolfo also served as a member of an informal group called the National Cyber Defense Initiative that provided advice to various agencies of the US Government on strategies for securing cyberspace.

Academic administrative roles

Professor Stolfo served as the Acting Chairman of Computer Science at Columbia for one year at a time when the existing chair left to pursue other challenges, and the CS department's finances were in a critical position. He also served as the Director of the New York State Center for Advanced Technology of Columbia University and the NSF sponsored Digital Government Research Center at Columbia University. He is not enthusiastic about holding administrative roles in academia.

Boards and Committees

Professor Stolfo is a member of the editorial boards of IEEE Security and Privacy Magazine and Data Mining and Knowledge Discovery Journal. He has served as special issue editor for the IEEE S&P several times.

Over 30 years he has chaired, co-chaired and served on the program committees of numerous workshops and conferences in the areas of parallel processing, data mining, computer security, intrusion detection and digital government.

Start Ups and Entrepreneurial Activities

Professor Stolfo was founder or co-founder of and advisor to several startups. One of these, a company called StackSafe, is not described here since the core of the company's product was based upon an invention of another colleague.

Fifth Generation Computer Corp.

The first start up Stolfo was involved in developed a commercial product based upon his DARPA-supported DADO Parallel computer, described above in some detail. The most recent incarnation of this company shows little interest in mentioning its Columbia roots.

iPrivacy: Online Private Shopping and Shipping

A few friends and co-founders including Jonathan Smith and Yechiam Yemini, developed with Stolfo an internet privacy company called iPrivacy, that was way ahead of its time. (iPrivacy developed in the 1990' should not to be confused with a more recent company that now owns that name and deals with privacy products to secure personal information). The iPrivacy company, started in the late 1990's with angel investors, developed and deployed a complete private browsing, private shopping and private shipping system that was completely deployed and ready for test in 2001. iPrivacy was the first to solve the problem of shipping physical objects bought on the internet! The company had a contract with the US Post Office and a large commercial bank who signed up over 30,000 customers in Silicon Valley, and a fully fielded system ready for use. The tragic, cowardly events of 9/11 at Ground Zero of the World Trade Center in lower Manhattan, nearby iPrivacy's offices, caused a rapid economic downturn that forced iPrivacy to shut down in the wake of the nation-wide crisis. (The Past Research page on Stolfo's webpage points to a number of patents and applications describing iPrivacy's technology.)

System Detection Inc.,

System Detection was one of the companies founded by Prof. Stolfo to commercialize the Anomaly Detection technology developed in the IDS lab. The company

ultimately reorganized and was rebranded as CounterStorm, and later was acquired by Trusted Computing Systems. That company was recently acquired by Raytheon.

Allure Security Technology

The Columbia IDS lab has produced well over a dozen patent applications filed by Columbia University for security and privacy technologies some of which have been very recently licensed to commercial enterprises. His most recent spinout developed with colleague Angelos Keromytis is a company called Allure Security Technology, Inc. devoted to developing products and services that are based upon the decoy technologies invented in the IDS lab, especially the new area Stolfo calls "Fog Computing". Fog computing provides a means of securing sensitive information one may store "in the cloud" by embedding a great deal of "cover" bogus data making it difficult for an adversary to know what is real and what is not. Stolfo introduced the concept at the RSA 2011 conference.

Personal philosophy

Stolfo is not done yet! His personal philosophy, especially after having experienced an awakening after a challenging physical ailment in 2007, is this: key both eyes wide open. Life is too short!

April 2011