Theorem-Based, Data-Driven, Cyber Event Detection

Lee M. Hively
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831-6418
865-574-7188
hivelylm@ornl.gov

J. Todd McDonald
University of South Alabama
150 Jaguar Drive
Mobile, AL 36688
251-460-7555
jtmcdonald@usouthal.edu

ABSTRACT
Nonlinear dynamics and graph theory may provide a theorem-based path to improve design security and aid detection of anomalous events in cyber applications. Using side-channel information such as power taken from underlying computer components and analyzing noisy data such as timing, we ask the question of whether such data can reveal anomalous activity or verify the changing dynamics of an underlying computer system. Takens' theorem in nonlinear dynamics allows reconstruction of topologically invariant, time-delay-embedding states from the computer dynamics in a sufficiently high-dimensional space. The resultant dynamical states are vertices, and the state-to-state transitions are edges in a graph. Graph theorems guarantee topologically invariant measures to quantify the dynamical changes, based on the applications that are executing. This paper highlights recent applications of the phase-space analysis technique in the non-cyber realm (forewarning of biomedical events and equipment failures), and proposes new applications that would bolster cyber event detection.

Categories and Subject Descriptors
G.2.2 [**Mathematics of Computing**]: Discrete mathematics – *graph theory* J.2 [**Computer Applications**]: Physical Sciences and Engineering – *engineering, physics, mathematics.* K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *invasive software, unauthorized access.*

General Terms
Measurement, Performance, Design, Reliability, Experimentation, Security.

Keywords
Cyber anomaly detection, phasespace analysis, nonlinear dynamics, graph theory, power measurement.

1. INTRODUCTION
Recently, computers have been modeled using nonlinear dynamics-based measurement frameworks [1-2]. Appealing to a physics-based view of the system, results indicate that the dynamics of a computer can be described by an iterated map representing the software and hardware. Based upon time-series data from simple programs running on common computers, researchers have used delay-coordinate embedding [3] to study the associated dynamics and found strong indications of a low-dimensional attractor in the dynamics of simple programs, as well as showing the first experimental evidence of chaos in real computer hardware [1].

Side-channel information (particularly differential power analysis) has been used extensively for adversarial compromise of algorithms and has allowed exploitation of cryptographic operations implemented in hardware [4-6]. As a novel approach for detecting cyber anomalies, we consider whether side-channel power information sampled from various computer components (external aggregate AC power, internal aggregate DC power, motherboard, CPU, disk drive, memory, network interface cards, and graphics cards) can be used to characterize normal operational behavior in cyber systems. Given historical success for using theorem-based, data-driven phase-space analysis techniques in biomedical and industrial applications [7-8], we postulate that side-channel characterization from non-invasive sensors may provide indicators for predicting failures in physical devices and detecting execution of anomalous software.

In this paper, Section 2 describes the technical approach for phase-space analysis of time serial, process-indicative data. Section 3 presents typical results for diverse, non-cyber applications. Section 4 discusses the proposed application to cyber events, for which the necessary dynamical data are presently unavailable (e.g., timing data, side-channel power). Section 5 discusses the conclusions for event detection via dynamical variability.

2. PHASE-SPACE ANALYSIS APPROACH
Dynamical systems are mathematical models describing real-world phenomena such as weather systems, fluid flow, population growth, and mechanical motion. Dynamical systems are defined by a fixed rule that describes time-dependent behavior at a point in some geometrical shape. At any given time, the dynamical system is described by some real-valued vector state that is a point in a state space that characterizes transitions from one state to another. Our analysis technique combines dynamical systems theory with standard time series analysis approaches. We start with a process-indicative signal, e, that is sampled at equal time intervals, τ, starting at an initial time, t_0, yielding a time-serial set of N points (cutset), $e_i = e(t_0 + i\tau)$. The garbage-in-garbage-out syndrome is avoided by rejecting data that fails any of the following tests: proper number of data points; intervals with unchanged amplitude; saturation at high or low limits; consistent amplitude across datasets; adequate sampling rate; excessive periodic content; and excessive noise.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CSIIRW '12, October 30 - November 2, Oak Ridge, Tennessee, USA
Copyright 2012 ACM 978-1-4503-1687-3 ... $15.00

Artifacts are removed with a zero-phase quadratic filter that performs better than conventional filters. This filter fits a parabola in the least-squares sense over a moving window of $2w+1$ data points. The central point of the fit estimates the low-frequency artifact, f_i. The residual (artifact-filtered) signal is then, $g_i = e_i - f_i$, and has essentially no low-frequency artifact activity. The g_i-data are symbolized into S discrete values, s_i, namely $0 \le s_i \le S - 1$. Equiprobable symbols are formed by ordering all of the baseline data from the smallest to largest value. The first N/S of these ordered values correspond to the first symbol, 0; data values $(N/S) + 1$ through $2N/S$ correspond to the second symbol, 1, and so on. Uniform symbols, $s_i = \mathrm{INT}[S\,(g_i - g_n)/(g_x - g_n)]$, use g_x and g_n, which are the maximum and minimum in the g_i-data, respectively. One can choose either uniform or equiprobable symbols; experience to date favors the use of uniform symbols.

A time-delay vector, $y(i) = [s_i, s_{i+L}, \ldots, s_{i+(d-1)L}]$, uniquely defines a dynamical state by base-S arithmetic. Thus, the dynamical domain is partitioned into S^d bins. Several channels of data add more information for: $y(i) = [s_i(1), s_{i+L}(1), \ldots, s_{i+(d-1)L}(1), \ldots, s_i(C), s_{i+L}(C), \ldots, s_{i+(d-1)L}(C)]$. Here, $s(k)$ denotes symbols from the k-th channel up to C channels with a total of S^{Cd} dynamical states. Here, L is the time delay, and d is the dimension of the time-delay space on the basis of Takens' theorem [3]. Figure 1 summarizes the major points of the approach.

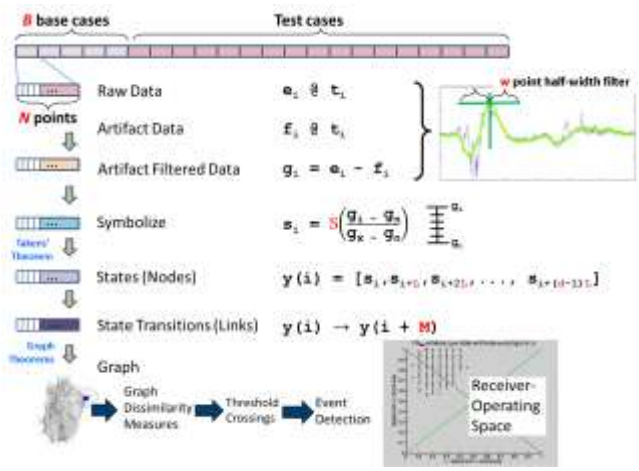

Figure 1. Phase-Space Analysis from Time Series Data

Takens' theorem [3] provides a diffeomorphism that guarantees topology (connectivity) and directivity, but not a density of states. Consequently, the analysis tabulates the unique, time-delay states as nodes, $y(i)$. The process flow, $y(i) \rightarrow y(i + M)$, is also extracted as directed, state-to state links. The nodes and links form what mathematicians call a "graph." Graph theorems guarantee topologically-invariant measures. The present work uses four dissimilarity measures between graphs from different cutsets: (1) nodes in A but not in B; (2) nodes in B but not in A; (3) links in A but not in B; and (4) links in B but not in A. These dissimilarity measures sum the absolute value of differences, providing better discrimination then traditional nonlinear measures, which are based on a difference of averages. For each A-B comparison, the dissimilarity measure is normalized to the total number of nodes (links) in A (for A not in B) or in B (for B not in A).

Normalized measures, $U_i(V) = |V_i - \overline{V}|/\sigma$, account for the disparate range and variability of the dissimilarities. The mean dissimilarity measure, $\overline{V}$, is obtained by comparison among the $B(B-1)/2$ unique combinations of the B base case segments, with a corresponding sample standard deviation, σ. Each contiguous, non-overlapping test case is subsequently compared to each of the B base case intervals to obtain the corresponding average dissimilarity, V_i, of the i-th analysis window for each dissimilarity measure. U_i is then the number of standard deviations that the i-th test case (unknown dynamics) deviates from the base case (nominal-state). Classification of dynamical changes uses several successive occurrences of U_i above a threshold to provide indicate a significant change in the dynamics of the process. Present applications of this approach include event (anomaly) detection and forewarning, which we discuss further in section 3.

Table 1 summarizes the trainable parameters, along with the corresponding trade-offs between small and large values of those parameters. For example, the number of data points in a cutset is a trade-off between inadequate sampling of the dynamics at small N (thus giving scarce statistics) against excessive blurring of dynamical change at large N. We implement our analysis in research-class FORTRAN for computational speed. Typical training of the classifier analysis is very compute intensive, typically involving CPU weeks to months on a modern desktop computer. The measures of success are the number of true positives (TP) from known-event datasets (Ev), and the number of true negatives (TN) from known-nonevent datasets (NEv). Best TP and TN rates are obtained by minimizing the prediction distance:

$$D = \{[1 - (TP/Ev)]^2 + [1 - (TN/NEv)]^2\}^{1/2}. \qquad (1)$$

Table 1. Summary of the Trainable Parameters

Algorithm Step	Specific Parameter	Small value	Large value	Present work
Digitize data	data points in cutset (N)	scarce statistics	blurred change	$5000 \le N < 100000$
Remove artifact	half-window width (w)	faster artifacts	slower artifacts	$2 \le w < 100$
Symbolize data	number of symbols (S)	noise rejection	excess precision	$2 \le S < 100$
Phase space	dimensions (d)	under-fitting	over-fitting	$1 \le d \le 26$
	time delay lag (λ)	small unfolding	large unfolding	$1 \le \lambda < 100$
Connected phase Space	inter-symbol lag (μ)	short correlation	long correlation	$1 \le \mu < 100$
Phase space dissimilarity	base-case cutsets (B)	short baseline	long baseline	$5 \le B \le 20$

The forewarning analysis proceeds as follows: (a) choose specific values for each of the training parameters in the set, $\{d, S, M, L, w, B, N\}$; (b) search exhaustively over N_{OCC} (the number of successive cutsets where dissimilarity between test and basecases is above a threshold) and U_C (the threshold for a normalized dissimilarity measure) for each of the dissimilarity measures to find the smallest prediction distance, D, or for smallest forewarning time if no improvement in D occurs; (c) search randomly over the parameter space in (a)-(b) until no further improvement is found; and (d) search exhaustively over the pruned domain from (c) to find the largest region of smallest D. The search strategies under (c) and (d) use the falsifiability theorem to eliminate (prune) statistical models that do not match the data [9]. Random and exhaustive searches are needed (rather than steepest-descent methods, for example), because the objective function over the search space has very irregular

features, as shown in figure 2. Our current analysis is >200-fold faster than real time on a desktop or hand-held device. The random and exhaustive searches over the parameter space allow partitioning of the computation into parallel instantiations, which scale linearly in the number of processors.

3. REPRESENTIVE RESULTS

The phase-space analysis technique has shown historical success for forewarning of events in diverse application areas. Part of our future work involves acquisition of cyber data of various kinds in order to determine applicability of the theorem-based approach for forewarning and detection. In lieu of cyber data, the success of this approach in other domains establishes its relevance for cyber applications as a statistical analysis method for noisy, time-serial data. Two specific examples are discussed next: epileptic seizure prediction and forewarning of helicopter rotor gearbox failure.

3.1 Biomedical Application

A promising application of phase-space analysis involves forewarning of epileptic seizures. Historically, we have used 32-channel, human scalp brain waves (electroencephalography or EEG) sampled at 250 Hz for algorithmic training. Currently we use 940,104 seconds (261 hrs and 8 min) of physician-provided "representative" EEG data. The data represents 60 human EEG observation datasets (27 GB). Forty (40) datasets have seizure events; twenty (20) datasets have no event. Due to lack of additional characterized data, we only use the data for training the parameter space (discussed in section 2). Typical results from EEG analysis fill a receiver-operating space, arising from dependence on the large training-parameter space, $\{d, S, \mu, \lambda, w, B, N\}$, summarized in table 1. To date, the best point has a true positive rate of 39/40 and a true negative rate of 19/20. Figure 2 shows the prediction distance for an exhaustive search through a slice of the search space.

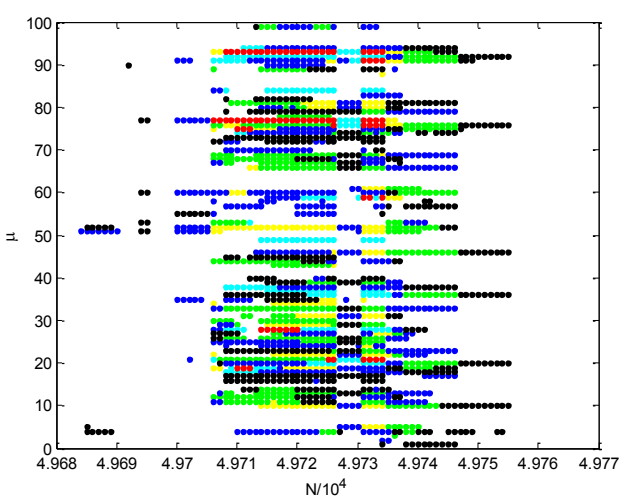

Figure 2. Minimum prediction distance (D_N) versus N and μ, for values of the other parameters fixed at d=7, S=3, λ=56, w=29, B=12. The colors correspond to the following values of D_N: red (0.0559); yellow (0.0707); green (0.0901); cyan (0.1031); blue (0.1118); black (0.1250); and white (≤0.1346).

The fractal features in figure 2 show that the objective function is not smooth, and has isolated maxima (and minima) that cannot be predicted from nearby points. Only a locally exhaustive search strategy can find the best values. The best points are an improvement over the previous results [8] for three reasons, namely: (1) D_N = 0.055902 (versus D_N = 0.1581 previously [8]); (2) TN = 19/20, and TP = 39/40 (versus 17/20 and 39/40 previously); (3) the best solutions occur in regions, versus an isolated point previously.

3.2 Mechanical Application

Another successful application of phase-space analysis involves forewarning of failure in the main rotor gearbox for a helicopter [10]. In this experiment, a seeded fault (notch) was formed in one gear tooth, with accelerated failure testing at 1.5-times the maximum design torque. Accelerometer data were sampled during contiguous, non-overlapping 15-second intervals. Forewarning corresponds to several successive values above a threshold of a composite measure C_i as the sum of the four dissimilarities. Figure 3 shows C_i versus time for one accelerometer channel.

Figure 3 illustrates the forewarning threshold value, U_{FW}=2.701, as a green horizontal line. Many C_i values fall below this threshold before 2.43 hours, corresponding to nominal gearbox operation. Failure forewarning spans 464 successive occurrences of $C_i > U_{FW}$ after 2.43 hours. Before this time (between 2.3 and 2.425), the largest number of successive occurrences of $C_i > U_{FW}$ is 31, denoted in Figure 3. This false indication is excluded by requiring at least 32 successive occurrences above the threshold, to yield a forewarning criterion of ≥ 32 successive occurrences for $C_i > U_{FW}$ = 2.7015 with the start of forewarning at 2.56 hours (32 time windows after 2.43 hours at 15 seconds per time window). Figure 3 also displays failure onset as 40 successive occurrences of the composite dissimilarity above the red horizontal line, $C_i \geq U_{\mathrm{FAIL}}$=21.68. The study [10] also found analogous forewarning or failure onset prediction for other accelerometer channels. Overall, the phase-space dissimilarity analysis showed promise in main rotor gearbox failure prediction.

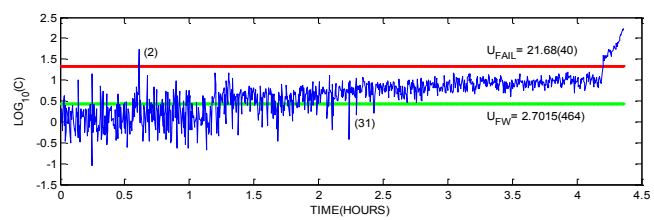

Figure 3. Failure forewarning and onset via the composite dissimilarity for accelerometer A1.

4. CYBER APPLICATION

Cyber observables (e.g., from a PC's hardware performance monitoring facility) form a sequence of emergent, dynamical states [1]. Examples include the instructions executed per cycle, the total number of references to the data cache, and component level power measurements. As with biomedical, industrial, and mechanical applications, Takens' theorem [3] can guarantee topologically-invariant construction of these dynamics in a sufficiently high-dimensional state-space, assuming a real, twice-differentiable observable without special symmetries. Forewarning and detection of anomalous events in the cyber domain has massive implications for improved operational readiness and ensuring security of mission-critical systems. Unfortunately, forewarning and detection are only useful if they are reliably actionable. Although high true positive rates can be achieved for various methods of anomaly detection using a wide variety of data correlation, high false positive rates cause a loss of

confidence in applied tools and reduced efficiency for operational communities in both the government and commercial sector. Two specific areas in phase-space analysis may be fruitful.

4.1 Component Failure

Failures in mechanical and electrical systems involve dynamical changes (e.g., whisker growth to create an electrical short, or crack growth to create a connectivity or structural failure). Modern high-performance computers require several reboots per day to recover from failures. Exascale computers will experience thousand-fold more faults, for which failure prediction is then essential. Present prediction techniques use empirical analysis, and achieve a true positive rate of 94% and a false positive rate of 55-59% [11]. The high false-positive rate does not allow reliable predictive maintenance or other pro-active steps, such as computational migration away from the failing component(s). In addition, studies in large disk drive populations [12] indicate that several parameters from the self-monitoring facility (SMART) correlate highly with failures, but models based on SMART analysis alone are not likely to be useful for predicting failures in individual drives. Phase-space analysis promises an alternative approach with higher accuracy for such failures.

4.2 Host-Based Anomalies

A key question of basic research is whether malicious software operates outside of normal operational constraints. Phase-space analysis can utilize electrical power consumed by each active component (e.g., CPU, memory, disk-drive, network access card) to determine anomalous patterns, if they are exhibited by certain malware. The specific measures are electrical current (I) and electrical voltage (V) to yield power ($P = I*V$). Typical voltages are 3-12 VDC, which vary with instantaneous load. Typical currents are < 1a and are highly load dependent. Typical data-sampling rates are 50 kHz (20 microsecond intervals) to resolve processor dynamics. As figure 4 illustrates, we envision appropriate sensors on key components to collect various types of power data. Hardware-based counters in modern computers also expose process and host-level information such as timing, processor load, instructions per cycle, and memory operations: all these data are adaptable to phase-space methodologies because they can have temporal relationships.

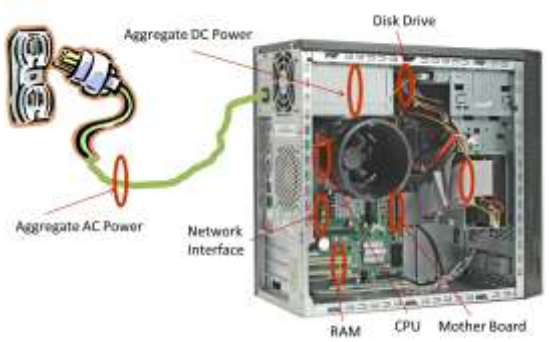

Figure 4. Conceptual Component-Level Power Monitoring.

Our research entails three phases of future work: 1) configuration of data acquisition equipment and monitoring sensors in controlled laboratory environments; 2) data collection experiments to identify and define baseline behavior for different cyber environments under different operating regimes; and 3) detection of variations from baseline behavior using phase-space dissimilarity analysis during malicious program execution or hard drive failure scenarios. The results will guide creation of real-time sensor based applications for field testing and deployment.

5. CONCLUSIONS

In this paper we propose a novel approach for prediction and detection of anomalies in cyber applications. We show the technical strength of the approach is a theorem-based, data-driven phase-space analysis. Prior work demonstrates the success of the technique for forewarning in diverse environments with disparate forms of time serial data. We illustrate how several forms of cyber data are adapted for this analysis method and make the argument for future research using this framework.

6. ACKNOWLEDGMENTS

This research was supported in part by an appointment to the U.S. Department of Energy (DOE) Higher Education Research Experiences (HERE) for Faculty at the Oak Ridge National Laboratory (ORNL) administered by the Oak Ridge Institute for Science and Education.

7. REFERENCES

[1] Mytkowicz, T., Diwan, A., and Bradley E. Computer systems are dynamical systems. *Chaos* 19: 033124 (2009).

[2] Alexander, Z., Mytkowicz, T., Diwan, A., and Bradley, E. Measurement and dynamical analysis of computer performance data. *IDA* 2010: 18-29 (2010).

[3] Takens, F. Detecting strange attractors in turbulence. In Rand, D.A. and Young L.S., *Dynamical Systems and Turbulence*, *Lecture Notes in Mathematics* 898: 366–381, Springer-Verlag (1981).

[4] Kocher, P., Jaffe, J., and Jun, B. Introduction to differential power analysis and related attacks. *Technical Report*, Cryptography Research Inc. (1998).

[5] Kocher, P., Jaffe, J., and Jun, B. Differential power analysis. In *Proc. 19th Annual Int'l Cryptology Conference on Advances in Cryptology (CRYPTO '99)*, pp. 388–397, Springer-Verlag, (1999).

[6] Mangard, S., Oswald, E., and Popp, T. Power analysis attacks: Revealing the secrets of smart cards. In *Advances in Information Security,* Springer-Verlag (2007).

[7] Protopopescu, V. and Hively, L.M. Phase-space dissimilarity measures of nonlinear dynamics: Industrial and biomedical applications. *Recent Res. Devel. Physics*, **6**, 649-688 (2005).

[8] Hively, L.M., Protopopescu, V.A., and Munro, N.B. Epilepsy forewarning via phase-space dissimilarity. *J. Clin. Physiol.* **22,** 402-409 (2005).

[9] Tarantola, A. Popper, Bayes and the inverse problem. *Nature Phys*. **2**, 492-494 (2006).

[10] Hively, L.M. Prognostication of helicopter failure. *ORNL/TM-2009-244*, Oak Ridge National Laboratory, Oak Ridge, TN (2009).

[11] Cappello, F. Advanced fault resilience techniques for exascale numerical simulations. Presentation at Oak Ridge National Laboratory, Oak Ridge, TN (16 May 2012).

[12] Pinheiro, E., Weber, W.D., Barraso, L.A. Failure trends in a large disk drive population," *Proc. 5th USENIX Conf. on File and Storage Technologies* (February 2007).