Technical Disclosure Commons

Defensive Publications Series

August 13, 2018

MANAGING UNKNOWN-UNKNOWNS IN CYBER-SECURITY

David A. Maluf

MANAGING UNKNOWN-UNKNOWNS IN CYBER-SECURITY

AUTHORS:
David A. Maluf

ABSTRACT

Techniques are described herein for managing unknown-unknowns in cyber-security. Trust degradation is a precursor index to failure. The use cases of scoring the trust degradation in a system span to almost every aspect in networking, edge and cloud included. A well devised Trust Evaluation Function (TEF) will cover many use cases: for example (1) better and adaptive private key management (e.g., re-keying); (2) better and adaptive end user experience password management and its fine grain monitoring in a data center; (3) better and adaptive digital asset certifications; (4) troubleshooting; and (5) real-time scalability and risk assessment for extremely large network, for example in federated cloud environment. The features of a digital trust scoring will start to reflect the likelihood of erosion of trust created on day 0. Platform independency is achieved when the score is a degradation of the trust and not the trust value alone. A trust value may start erroneously, but the rate of change may lead to continuous evaluation. Therefore, the originating trust is set as a prior. Erosion will thus work with time against the assumed original trust. In the example of an expiration date or a combinatorial complexity erosion of a private key, the realization of a trust erosion is not a Boolean fail pass type, but a relative factor number. On a comprehensive integrated analytical dashboard, the trust factor produces the percent life left of given a digital secret.

DETAILED DESCRIPTION

Trust under certainty does not recognize the element of unknown | unknown. Simply, the foundational models of certainty lie in the notion that the definition of prior is what has been known or measured. No provisions are made otherwise for undeclared variables. Trust under uncertainty intends to model the unknown.

But the argument remains that trust under uncertainty will use certainty approaches once certain criteria are met. For general discussion purposes, trust under uncertainty obeys the law of information preservation as defined in Information Theory principles. The clear

difference is that trust under uncertainty will optimally extrapolate, in terms of new a posteriori hypothesis, trust outside the scope of the known models and priors.

With lack of ideal models, the concept of discarding random variables for density functions is of interest. For example, the individual outcome of a specific atom in a Brownian motion is not of interest. On the other hand, the density function of few atoms is useful. Thus the kinetic energy of the collective makes sense based on the sampling of density functions.

Similarly, trust under uncertainty relies on density functions and is measured in terms of entropy uncertainty. It reflects the uncertainty of a whole system not dependent on the assumed models. There are three basic definitions that trust may fall under.

First, outcomes are known. In this case, the range of outcomes is known and the outcome is also known. This is the easiest way to make risk management decisions. This reflects a deterministic event process. This is also known as "you know what you know" paradigm or the kk paradigm.

Trust can be summarized as making decisions based on what is "known to be known" or $\{k \mid k\}$. The model parameters are defined and deterministic. Observations in $\{k \mid k\}$ have known means but zero variances: $y(t) = f(\mu)$.

Second, outcomes are unknown, but probabilities are known. In this case, the range of outcomes are known but the individual outcome is unknown. This reflects a deterministic process with known density functions. All of the outcomes are known as are the probabilities of each. This is also known "you know what you don't know paradigm" or "k ¬k paradigm."

Trust expands to this category as what is "known to be unknown" or $\{k \mid \neg k\}$. The model parameters are defined with density mass functions. Observations in $\{k \mid \neg k\}$ have known means but positive variances: $y(t) = f(\mu, \sigma)$.

Third, outcomes are unknown and probabilities are unknown. In this case, the distribution of outcomes are unknown and the individual outcomes must be unknown. This is also known as the "you don't know what you don't know" paradigm or the "¬k ¬k paradigm."

Trust extends to the empty knowledge (or lack thereof). It deals with decision when nothing is known as it is unknown what is unknown $\{\neg k \mid \neg k\}$. The consideration of

external priors such as boundary conditions and reasoning assume sampling for data around asymptotic directions in the direction of these boundaries.

A clarification on the meaning of "data" is provided as follows. Computer sciences conceptualize data to be defined to meet known types. The generalization of these definitions presents a diverse, but limited, representation that extends from basic binary representations (i.e. [0,1]) to templated sequences to ordered complex data structures. An example would be the Internet Protocol (IP) network header stack and yet another a database record.

Data analysis for cyber-security can therefore be processed in two different approaches. The first approach is by the introspection and analysis of the data in these structures. The second approach is the process of segmenting the data into types and structures that brings forth the corpus of analysis about the segmentations. The formulation of an objective analysis about the occurrence of data becomes critical as the new information about the data is made as an objective measure. Each process of formulating the new information about the data is referred to as an experiment.

Most solutions in machine learning and artificial intelligence assume that deterministic models have strong notions of priors and models. Conversely, trust under uncertainty investigates how to rationalize data with incomplete priors and weak models.

Trust under uncertainty is a form of risk assessment for complex systems and can be illustrated by the complex aspects of cyber-security. Complex systems are hard to model without making assumptions and assumptions are detrimental in cyber-security given their impact. In accordance with models taking assumptions, uncertainty is treated herein as the unknown growth beyond the made assumptions. In experimentations, the boundary estimate for the gap between the models and the measurements can be estimated in relative terms and is proportional to the models estimates. However, when reaching beyond the boundary, the process of extrapolation is necessary. Optimal extrapolation (and interpolation) is thus established in probabilistic terms beyond the model's boundaries. For these multidimensional complex problems, uncertainty is computed by experimental sampling with maximum entropy estimations.

Suppose an extrapolation outcome function optimizes a certain objective. It becomes pivotal to sample in the newly extrapolated neighborhood. Given an iterative

process, stochastic interpolation and extrapolation is an equivalent process of an event based optimization search. With the notion of time, the process defines a time series optimization search as in system theory. Independency is defined when the data is calculated on Information Theory principles. Maximizing the information is an entropy minimization search. The final outcome of an iterative process is a continuous search with optimal sampling objectives minimizing criteria.

Optimal sampling leads to better prior knowledge. Optimal sampling turns the objective towards optimizing the covariance density instead of optimizing the individual variances or the diagonal elements. Maximum Entropy Sampling (MES) does that. Sampling might have to touch deeper in the underlying law of physics to fill in the correlating elements of the prior among the sparse but measured density functions do not.

An example of maintaining an analytical continuity is to assume some given data points over a grid represent a random surface obeying the law of continuous surface. Regardless of the data density, interpolation and extrapolation obey the continuity constraints. It is expressed mathematically as a smoothing function in the inverse covariance of the grid points. This operation is critical when the data points are sparse and when extrapolation can be evaluated over the range $[-\infty; +\infty]$.

The case above distinguishes key aspects of the heart rate monitor in two-fold. First, it separates the data as part of an experiment, say ξ, being designed independently of the models' g(t) parameters. Second, the mapping of the experiment data to the model parameters has the system realizations that the data complexity may subceed or exceed the model's characteristics.

In the counterintelligence case, it would be naive to assume that "more data" is the only answer, as the latter is not often feasible when time is limited. In the example of measuring hidden intent, data sparsity limits any reasoning into a comprehensive approximation. Inversely, models will also perform poorly given the spectrum spread of the deviations. These examples emphasize cyber-security frameworks, as well as human behavior in Human Intelligence (HUMINT) cases, both of which are very complex systems.

To further illustrate a hidden intent, uncovering deviant intent in the constituent of HUMINT evaluations (e.g., stress detectors) can be used. Suppose the only data available is the highly dependent heart rate measurements and a minimal model g(t) representing the

expected prior in terms of means and distributions. In HUMINT, the deviations from the prior constitutes the "intelligence" and not the expected values themselves. The outcome is therefore to compute for the deviation deduced (subtracted) from the experiment and model.

In the cases where the model characteristic is less than ideal, the model realization Degree of Freedom (DOF) is less to the data degree of freedom available from the data, the deviations constitute as loss of information when not accounted. The magnitude of the loss is in the order of number of degree of freedom away from what is expected. For example, if g(t) does not model data modulation, assume min, max and average and distributions, the information in first and second moment of the data variability (e.g., heart rate variability) are lost.

While multi-objective optimizations fall under constraint optimization, the focus is on the byproduct of convoluting the models (e.g., g(t)) with measurements from a given experiment ξ with parameters Θ. The assumption is that the experiment parameters vector Θ is not dependent on the models. In perspective, Θ represents the total DOF among the data. This relationship can also be defined as a coherence analysis between the models and Θ. With a model less than ideal, it obeys the Cauchy-Schwarz inequality and the lost information the model sustains is the difference in the entropy measured as defined by the Shannon theorem.

With an experiment ξ not dependent on the model g(t), but both systems satisfy the Cauchy-Schwarz inequality the result is a positive number. This brings forward the result than while Θ is time dependent and g(t) is assumed a fixed model, the Cauchy-Schwarz inequality is not guaranteed to be constant. It is imperative that g(t) is designed or the selection of the experiment ratio at the time of origin is a positive number. The Variability is an index of the degradation. A decreasing index is a degradation of trust in the static model g(t). However, ideally a convergence towards unity is sought. In the cyber-security example, priors are considered in issuing a secret and the model expresses in characteristics such as encryption bits, algorithms and expiration dates. The degradation of the expiration date in that case is computed on constructed experiment data sampled from actual data.

Finding adequate models g(t) for experimental data is not necessary conclusive or achievable. The approach of trust erosion is computed by marginalizing the prior g(t)

asymptotically along the exponential growth function at the time of origin t_0. For an upper bound measure of a degradation, the difference between the distributions for the experiment $\xi|_{t0}$ and the distributions at $\xi|_t$ are evaluated. In Bayesian sense, the trust degradation is therefore the maximum likelihood (upper bound) of the ratio of the posterior over the prior. For the upper boundary conditions, the maximum likelihood is therefore evaluated. Both the prior and the posterior are experimentally derived at t_0 and time t.

The outcome is consequently two folds: 1) determine a generalized measure for the impact of the deviation and 2) determine a trust degradation system as a dynamic time-varying process.

The interaction between the measurement and models can be thought of as the resolution of the information uncertainty. The mismatch is in the imparity of the minimal system realizations underlying the measurements and models. An outcome of the disproportionality in dimensions is categorized by the $\{k \mid \neg k\}$ and $\{\neg k \mid \neg k\}$ sets depending on the direction of the imparity.

The special case from above is that for an ideal model (system), the conservation of information exchange with the measurement is lossless.

The use case of the HUMINT is a segue to cyber-security and therefore extends to Cyber Intelligence (CYBINT) and Digital Network Intelligence (DNINT). The foundational aspects follow two distinct vectors: First, to illustrate the notion of trust under uncertainty, second, to illustrate the notion of degradation of assumptions being proportional to time and to the unaccounted (lost) information. The foundation of cyber-security is based on series of non-ideal statistics crafted to remedy current challenges. These statistics have specific models and require specific configurations to function. The current cyber-security paradigm will fall short as data holds are information centric and the security holds as models. Degradation will also be referred to as the erosion of the expected trust in the models given the loss of the divergence between the data and the models.

Addressed herein is cyber-security in regards to its digital assets where it would make sense to think that digital assets have inherent non-ideal models by design. The objective is to solve for the expected models and deviations as a dynamic time-varying process. Inversely, this demonstrates the aggregate risk factor when the degradation is left unchecked. The degradation is computed as the difference of the normalized uncertainty

measures between the expected and actual measurements. A degradation index is then constructed in the range [0,1].

Model independent system realization is discussed as the underlying computational framework for MES to provide optimal methods.

Examining cyber-security history, one would conclude it has been an evolutionary path. In fact, Internet security had been non-existent in its early days until the consequences were substantial and could not be marginalized. There was a time when base 64 encoding was the recommended approach. It is arguable that today's cyber-security surpasses the encoded secrets approach of base 64. However, the principle of security challenges remains the same but at a different level; still evolutionary, but they are more complex algorithms to decode. Either way the approach models have been probabilistic when the system's risk is measured as:

$$\text{Risk} = \text{Likelihood} * \text{Impact} \qquad (1)$$

There are many standards that govern the computation of the risk assessment. The likelihood is probabilistic measured through simulations and white hat assessments (e.g., third party certifications).

In general, this example describes a clear outcome with some probability distribution and as described, a known threat with known or unknown occurrences. In other words, the threat is stated as "the probability to decode a known secret."

It is clear that after-the-fact analysis has driven security to model assumptions about the threats. Thus, the modalities of a secure system have been to meet the conditions set by the probabilities analyzed from known (and hypothetical) threats.

The observations point out that cyber-security challenges still exist, and the question remains if what has been computed for and set forth in execution is not adequate. There should be no surprises since the Risk = Likelihood * Impact is still accepted as a positive number; in other words, the Impact is normalized. For example, in the economic sense worldwide (e.g. industry-wide) a few targets are sacrificed as long as the risk is still below a certain threshold. This is the true status-quo of Internet security.

This shortfall will be addressed herein on how to avoid marginalizing systems in cyber-security.

Currently, cyber-security creates solutions that are constantly devised to counter threats with complexity (e.g. algorithmic) but are built around existing threat models. These threat models do take additional measures into consideration for futuristic enhancements but in truth, these assumptions (considerations) are still projections with assumed variances. Meanwhile, as more complex algorithms are devised new threats are discovered outside the assumed models. After decades of cyber-security failures over the same principles of security, one would wonder if the stated evolutionary strategies are still the right ones.

The unknown-unknown type of problems are addressed herein as they relate to cyber-security shortfalls. Though, one could argue that the previous shortfalls are a subclass of the unknown-unknown type problem.

Loosely, the existence of a failure is addressed from an existentialism view. To express this further a new analogy will be expressed. Fundamentally, a threat only exists to a system if the system exists in the first place. This notion is known among economics; existentialism takes the form of economic terms and existence is some dollar value. Simply put, the maximum threat to a $1-dollar investment is $1 whereas the maximum threat to $1B is $1B. It can also be understood that the risks are proportional to their underlying complexities. Internet security is not that different and has grown very complex on its own, but only to the extent of the impact (profit and loss).

Defining the economics case as a trust under uncertainty, one can realize that the market is constantly challenged with its own unknown-unknown. In fact, economic recession has occurred around ⅘ of the time for centuries and still no one can predict the next one. If a hundred percent secure solution was devised there would be a better payoff applying it instead of the methodology of market prediction.

In perspective, one can argue that the goal is to maximize the ratio of trust degradation over its design closer to unity. A distinction between digital assets (e.g., digital secret) and physical assets (e.g., rover on Mars) is the capitalization and cost of opportunity. While a rover would cost $1B in capitalization and many years to make, digital assets will far exceed their value proposition, risks and liabilities included, because of the premise in the economics of digitization. Cyber-security in general and CYBINT differ at this level as

CYBINT has a substantial upfront capitalization. The premise of this maximization is solely dependent on the interpretation of "impact." Obviously, the meaning of impact differs for both use cases. The volatility in the digital asset context dictates the purpose of this discussion, which is to model such a ratio in the digital world that is less volatile and more physically defined in the analog world.

The goal is not to devise a new secure system, but rather to assert the foundational science of the risk factors for a system attributed to its own growth (e.g., usage). The assumption is that the set {¬k | ¬k} grows proportionally with complexity. Assuming the underlying complexity growth follows the Solow economic model, the risk R(t) will follow a similarly an exponential growth.

The assumption that the underlying risks are proportional to underlying complexities takes the form of an assumed known risk R_{t0} at time t_0 and unknown deviation risks $R_d(t)$. An exposure factor $E_F(t)$ as an integral of all materialized and unmaterialized risks (i.e. deviation risks), or

$$E_F(t) = \int \big(R_{t0} + R_d(t)\big)dt = \int R_d(t)dt \qquad (2)$$

Given the complexity growth model, the unknown risk when left unchecked, will grow exponentially from t_0, or

$$R(t) = R_{t0} + R_d * e^{xt} \qquad (3)$$

Where x is the growth factor. The exposure risk factor follows as

$$E_F(t) = R_{t0}t + R_d \int e^{xt} \qquad (4)$$

It is clear that the unaccounted exposure is a monotonically increasing function for $x > 0$. It is notable that most systems rely on (R_{t0} t) only for a limited period $t_0 < t < t_{max}$ where t_{max} is the lifespan. One can further the assertion as the exposure factor, when factored back into the system model, reflects a degradation factor to the known assumptions.

As illustrated in Figure 1 below, the total area grows exponentially on the premise of {¬k | ¬k}. Both {k | k} and {k | ¬k} shrink at the same rate of the lack of new knowledge, {¬k | ¬k} grows. In layman terms, the cost of ignorance over time is not forgiving.

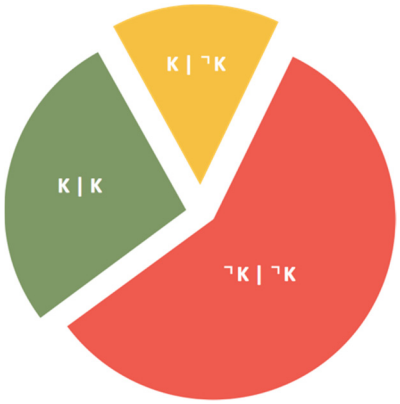

Figure 1: { K | K }, { K | ¬K } and { ¬K | ¬K }. One can envisage the boundary problem for the three concepts is limited the underlying physical concept. One would need to identify the boundary of the physical context. Reasoning under uncertainty is feasible once a model is conceived.

The growth of {¬k | ¬k} is non-linear, but the assumption made of its growth being exponential is fair choice. The choice of the growth factor x is determined from the underlying system model and can be independently modeled and evaluated as the underpinning economics and infrastructure complexity. Complexity is a foundational aspect of measure to many systems including cyber security. There are enough correlations to point out a causality at play. The latter assumption makes a big difference on what is being optimized. The applications (among others) are numerous. Though, in cyber security cases, the focus is on when to upgrade digital assets like secrets (e.g. secret keys or when to rekey). This is because the nature of security exposure is no longer a fixed assumption but a non-linear degradation as a process of complexity and time.

The concept of {k | k}, {k | ¬k} and {¬k | ¬k} is a very challenging concept to represent mathematically as one can model the { k | k} framework in set theory and model {k | ¬k} in the sense of probability theory. On the other hand, {¬k | ¬k} is boundless and unknown by definition. This problem relates to the modality of an inverse problem type constrained by its boundary problem.

At this level the systems' components {k | k} and {k | ¬k} have been exposed as a deterministic and stationary Gaussian process. The input of the system is not deterministic

nor a stationary Gaussian process. The inverse problem is formulated such that to identify the impact of {¬k | ¬k} on {k | k} and {k | ¬k} at least as a ratio. As {¬k | ¬k} expands exponentially, {k | k} and {k | ¬k} decline in the same order of scale. The ratio is thus a degradation index. One goal is to solve for the magnitude of the uncertainty (deviations) and not to search for predictability (building models). The magnitude of the uncertainty is entropic and will be addressed with these mathematical measures.

Sampling in the {¬k | ¬k} for deviations is a stochastic process for the points on a multidimensional function whose integral exhibits exponential behavior. But more precisely, the delta growth of the uncertainty is taken at an origin with a reference point in time t_0 of the system with prior as R_{t0} (i.e., {k | k} {k | ¬k} sets).

Optimally sampling in the {¬k | ¬k} space coincides with the MES theory thus maximizing the information and minimizing the entropy for non-stationary Gaussian processes. This discovery leads to two possible inversion problem perspectives as crossing the same boundary conditions either from left to right or from right to left and in terms of use cases. As illustrated in Figure 2 below, for the use cases where a deterministic system (i.e., stationary process) intersects with a non-stationary process, interpolation occurs for the area between the two intersecting boundaries, rendered deterministic and augmented to {k |¬k} models. Moreover, for the use cases where a non-stationary process intersects a stationary process system, extrapolation occurs based on the intersecting non-stationary process boundary curvature. Extrapolation for the rest of the non-stationary process is determined from the origin which is defined by the intersecting boundary, curvatures and an assumed growth rate of the non-stationary process.

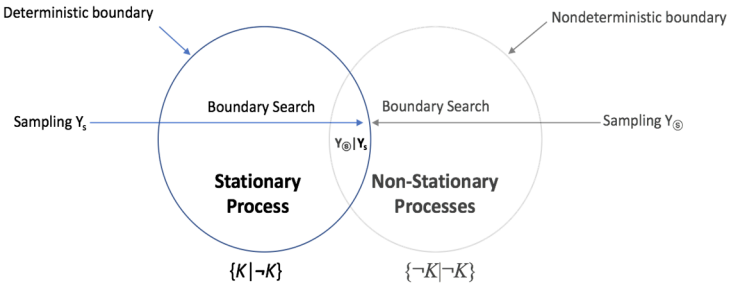

Figure 2: The non-stationary process is illustrated by the circle. The extrapolation attempts to compute the area/circumference dimensions based on the intersecting arc. The inverse problem is a boundary condition search for the intersection between stationary and non-stationary processes. A boundary can be crossed either from the left or from the right between {K |¬K} and {¬K |¬K}. An identification minimization search (convergence) can begin once the boundary conditions are exposed. Maximum entropy solves for the stationary process boundary.

Gaming theory makes sense in this context. For a conceptual "Red" and "Blue" teams division, cyber-security becomes a subset use case. In terms of intersection, "Blue" is the deterministic system at hand, and "Red" is the unknown unknown group.

Most gaming takes root from warfare. Warfare is the oldest school ever known to the human civilization and precedes literacy. The premise has been touched upon as Sun Tzu states in The Art of War, "To know your Enemy, you must become your Enemy." Fundamentally, cyber warfare is not much different in principle. One example is when digital secret keys are changed every now and then, which mimics an old strategy of moving the residency of a king every now and then.

The value proposition herein is analogous to the analysis of the ingress and egress behaviors made at the entrance of a city. Historically, travelers were denied pass-through when the traveler ingress exceeded its egress. This made the generals evaluate the risk ratio at the first travelers count, and later, at the maximum volume and capacity of their transports (reflecting disguised materials), against an existing defense system. The formula was simple: assess the worst-case scenario given the ingress minus egress capacity. Optimal reasoning mimicked the maximum entropy assessment of a threat if all travelers and all their transports constitute a threat. Inversely, reasoning with confidence intervals, or less than the maximum, is a limited system. An example in the prior model for the threat above is limiting the threat to men for an age range. Cyber-security models today mimics the stated assumptions as sets limits as well as the remedy to deter threats. However, they fall short as digital concealments has been a successful work around.

(As a result, cities were built as hub and worked to maximize the control over the transport characteristics.)

The maximum likelihood (paranoid) approach has both advantages and disadvantages. The maximum likelihood approach penalizes the optimization at some cost (impact). With the latter example, the impact may be represented as a trade-off analysis between trading economic loss against safety.

The objective in the analogy of the ingress and egress analysis over known physical quantities, such as volume and weight over time, serves the purpose of distinguishing the separation of the quantitative analysis (the data) from its qualitative counterpart (the

models). The qualitative analyses from the analogy are numerous such as safety threats, illegal trafficking, etc. The physical measurements do not change.

Not surprisingly, this methodology is still an optimal approach used in modern warfare, social and economic behaviors. Military checkpoints will report the ingress and egress quantitative measures regardless of the qualitative assessment to command and control. In unperfected warfare methods decisions are qualitatively made at checkpoints.

A risk assessment is not limited to the mean, variances, or the maximum confidence region of previous conflicts but a differential sensitivity analysis. Sensitivity analysis over quantitative data is described to measure an information index.

Using the former equations, the risk R_{t0} is a quantitative measure at time t_0. The exposure risk factor E_{rf} at time t_N is an integral risk carried forward because it is the maximum exposure. The exposure risk factor considers the risk at a future moment t_N and factors in the former risks as persistent intelligence (e.g., informers in the context of warfare).

The digital world is not different from the physical one, and digital Input/Output (I/O) equates to the ingress and egress in physical dimensions.

The techniques described herein build formally on this principle. The clear distinction, as in many analytical threat systems, is the premise that there is no knowledge about the threat, the enemy, or the "Red" team. This applies to every aspect underlying the $\{\neg k \mid \neg k\}$ assumptions. However, the assumption that the system boundary is measurable (the input and output interfaces are known) is a fair one.

There are two main topics concerning boundary discovery and how sampling experiments can be used to track dynamic system boundaries. The first is how to establish formal representations in experimental sampling and selecting the sampling points. The second is how to examine them, conduct the actual experiments, and collect the data. Formal representations such as D-optimality designs are particularly useful when classical designs do not apply in nonlinear systems. The optimality criterion used in generating D-optimal designs maximizes the determinant of the information matrix.

In complex cases, in both non-linear and non-normal estimation designs, when the posterior distribution cannot be modeled in a design criterion the estimation is not possible.

The solution described herein relies on asymptotic approximations thus the sense of direction becomes the basis of the posterior estimation of the design.

Sebastiani and Wynn have introduced MES for estimating similar design problems, see [5]. Using this approach, if the entropy of the sampling distribution is not functionally dependent on the experiment, then an experiment minimizing the expected posterior entropy can be found. This experiment is found by maximizing the marginal entropy of the data.

The techniques described herein may use MES as the evaluation criterion where the most informative design with respect to the parameters is of interest. The design builds on the D-optimality design for a continuous analysis. The end result is its combinatorial (regression) optimization over the continuous process.

In the context of building the experiment, different criteria exist that match different goals. Most supervised learning algorithms focus on A-optimality designs where a strict sense of models exist *a priori* and therefore the goal is to minimize the variances. In unsupervised learning, models are not known *a priori* and one is left to rely on said asymptotic approximations.

The entropy criterion is used for the purpose of choosing enough experiments $\xi_{1..k}$ that collectively maximize the amount of information about Θ, the model parameters, when intersecting $\{k \mid \neg k\}$ or $\{k \mid \neg k\}$ with $\{\neg k \mid \neg k\}$. It will be shown later that the choice of Θ is a frequency spectrum. The experiments $\xi_{1..k}$ will be chosen as quantitative measures that parallel underlying physical quantities such as I/O in digital systems.

Suppose one experiment ξ enables measuring the intersecting boundary. The outcome of the experiment is a random N vector y, and the subset $\Theta = \theta$ is assumed to be the intersecting boundary parameters where y has a known distribution with density $p(y \mid \theta, \xi)$. The parameter vector Θ has a prior density $p(\theta)$ that is not dependent on the experiment ξ and therefore θ is quantitative. In system theory, the experiment takes the role of an observer. The parameter Θ models the data and its dimensionality is in the order of the complexity of the data to preserve the maximum information.

The problem is to obtain the maximum amount of information about Θ as the negative Shannon entropy. For a sample measurement vector Γ, entropy is defined as the negative measure of information,

$$Ent(\Theta, y) = E_y(\Theta, y)\{-log\, E_y\,(\Theta, y_s)\} \qquad (5)$$

Working with {¬k |¬k}, it is assumed that an unknown vector r can be decomposed as (r_s : $r_{Ⓢ}$), where s represents the sampled set for N points on a boundary. The criterion identity decomposition for a sampled and unsampled random vector is

$$Ent(r) = Ent(r_s) + E(r_s) * Ent(r_{Ⓢ}|r_s) \qquad (6)$$

Sebastiani and Wynn rewrote the entropy decomposition for r = (Θ, y), see [5], such that

$$\begin{aligned} Ent(\Theta, y) &= Ent(y) + E_y\{Ent(\Theta|y)\} \qquad (7)\\ &= Ent(\Theta) + E_\Theta\{Ent(y|\Theta)\} \end{aligned}$$

These results may appear in several forms, as seen in [10].

For a given Ent(Θ, y), minimizing E_y {Ent(Θ| y, ξ} is equivalent to maximizing Ent(y), or the marginal entropy of the data.

The theoretical Bayesian information optimization consists of the experiment that minimizes the expected posterior (Shannon) entropy of Θ. This method applies the MES principle because of how the experiment optimally minimizes the expected posterior entropy of Θ. Therefore, the experiment which maximizes the entropy of the marginal distribution of y will be most informative for Θ.

The decomposition of the entropy with a minimization term and maximization term is in fact a divide and conquer strategy. In this strategy information is learned, extracted from the measurements and isolated from the ongoing remaining information. This is critical for time series regression experiments where deviation in the entropy is a measured quantity.

To apply the MES principle, a regression type experiment is sought for time series in order to sample models with additive deviations x and error variances. The additive

deviations and error variances are not dependent on the experiment ξ. The regression may be rewritten as

$$y|(\xi,\theta) = f(\xi,\theta) + x + \sigma\varepsilon \qquad (8)$$

f is a function of ξ and θ, conditional on Θ = θ, and x + σε are the vectors of independently distributed deviations and errors. The vectors x + σε are also independent of ξ. The vectors σε will be considered noise and are distinct from the deviation x. N is a control variable for the number of data points for the regression type and for a given period of time τ. Given θ and y(ξ, θ) are known (fixed), the following equation can be written.

$$Ent(y|\xi,\theta) = Ent(x|\xi,\theta) \qquad (9)$$

A key aspect is that f(ξ, θ) can be any function of ξ and θ and applies for nonlinear regression frameworks. Ent(x|ξ, θ) would be the most informative information.

For the assumption that the system is inherently nonlinear but follows an asymptotic approximation, Θ is assumed to have a continuous prior and that the sampling and prior density of θ is differentiable near the MES of θ. Examining the curvature of the posterior dispersion inverse leads to the Fisher information matrix

$$I(\theta,y|\xi) = \left[\frac{\delta^2 \ell(\theta,y)}{\delta\theta_i \delta\theta_j}\right]$$

$$= D.D^T \qquad (10)$$

ℓ(θ, y) is the negative log-likelihood function and D is its derivative matrix. The posterior can then be approximated as $\theta,y|\xi \sim N(\theta, I^{-1})$. Chaloner and Verdinelli prove, see [12], that the expectation may be switched to the prior distribution of Θ and the expected information gain from the experiment becomes, when minimized, D-optimal (asymptotic) as follows.

$$Min(E_u\{\Theta \mid y, \xi\}) = E_\Theta\{-\log \det I\} \qquad (11)$$

For the maximum information and approximate for the marginal density, a smoothing local constraint may be enforced on variations of the parameters. $\Sigma^{-1}(\theta_i, \theta_j)$ is designed as an isotropic Gaussian matrix of an arbitrary dispersion control size (e.g., typical size 5x5 Gaussian distribution) where σ is a dispersion control set. Many dispersion templates can be used for σ. As the range of Θ values becomes large, the smoothing function will propagate and increase the upper bound of the MES design.

The previous Fisher matrix may be replaced with an updated inverse covariance as shown below. A is the inverse covariance matrix, taking the form of a Hessian matrix, and b is the gradient of the likelihood function and is therefore the quantum measure of the information learned.

$$A = \Sigma^{-1} + D.D^T \qquad (12)$$
$$b = xD$$

For a regression, Θ is updated with A as the inverse covariance and b as the update vector. The information gain becomes E_y { - log det A^{-1} } and b is the model update change. The relationship of the derivatives into the estimated linearized regression becomes

$$\hat{y}_{s+1}|(\xi, \theta) = f(\theta_s) + D(\theta_s - \theta_{s-1}) \qquad (13)$$

The outcome of the experiment may be chosen for the random n-vector y to be a transformed vector Y, a Discrete Fourier Transform. Given $\sum_k | Y |^2 = 1$, then θ is considered a density function. Let $\theta = | Y |^2$ be the spectral density outcome. The entropic uncertainty expressed in terms of the deviation of the Shannon entropy presents the analogous form for y as

$$Ent(y|\xi, \theta) = -\sum_k D(\theta_k - \theta_{k-1}) \log\{D(\theta_k - \theta_{k-1})\} \qquad (14)$$

Θ may be used as the parameter vector of all θ for {y, t → ∞}. Nonetheless, the equation above needs to be optimized for regression purposes.

For complex communication systems g(t), the sense of time becomes artificial when it crosses the digital boundary. This means that time series analyses need to be analyzed for their maximum information content on the true physical boundaries before any loss of information occurs. That is, the digital sampling of the information needs to be measured in the absolute sense. In particular, non-real-time systems suffer losses because of the lack of absolute time and because their ability to reliably model the information at their interfaces becomes uncertain.

Neglect occurs when digital systems interact with the real word at the boundary of the underlying physical system. The deficiency occurs when the used information data is less than the maximum information index. A good example of deficiency occurs when there is variability in the rate of data arrivals. Timeouts are used as workarounds in systems but do not account for the variability and information is therefore lost to what is north-bound to the interface.

For illustrative purposes, the Open Systems Interconnection (OSI) model segmentation is discussed herein. The OSI model segmentation will show the derivation of the analytical transformation of time into an artificial one (other boundary types can also be found in other I/O driven systems). In general, I/O is subject to the communication theory where systems can be decomposed at their boundaries through analog to digital conversion or vice versa.

Suppose an experiment ξ is constructed and data collected. First, a received signal $y_a(t)$ on an interface may be considered. It is demonstrated, see [7], that the analog signal $y_a(t)$ may be written as

$$y_a(t) = f_a(\theta, t) + x_a(t) + \sigma_a \varepsilon_a(t) \qquad (15)$$

where y(t) are the observed sequences, $f_a(\theta, t)$ is the expected trend function, and x(t) and $\sigma\varepsilon(t)$ are the independent random deviations from the expected function of f(θ, t). The decomposition may be extended into a digital counterpart which may be written as

$$y_d(t) = f_d(\theta, t) + x_d(t) \qquad (16)$$

This equation assumes that $\sigma_a \varepsilon_a(t)$ is limited to analog signals reflecting hardware noise and vanishes in digital observations. The signal y(t) is defined as a stochastic process with a mean function $f_d(\theta, t)$ and a stochastic deviation x(t). The behavior of x(t) is analyzed herein and x is not discarded as noise.

Optimal parametric representation may be used for the estimation of the relevant statistical components and the estimates of the coefficients may be made using parametric methods. Furthermore, to maximize the analysis over a flexible time range, the Fourier function may be used as a basis function. Comparatively, Gebraeel and al. analyze their signals non-parametrically, but they similarly assume a localized smooth curvature model to achieve the surface continuity [7]. The curvature is considered to be smooth and continuous even though the nature of data contains sharp changes for neighboring characteristics. For that purpose, Fourier is the method to use for modeling. Therefore, taking the Fourier transform over a period τ for a total of N samples yields

$$Y(f) = \theta_d(f) + X(f) \qquad (17)$$

θ is in the frequency domain and Y and X are the corresponding discrete Fourier transforms for size N.

The choice of using the OSI model is intentional as it points out that the main innovation in the Internet is artificial time. Artificial time occurs with the reordering of data arrival frames on a new sequence such as demultiplexing. This novelty renders the communication channel asynchronous and as such, a manifestation of a frame (i.e., packet) can arrive any time and out of order.

In an OSI model, which includes layer 1 and 2, layer 1 is analytical in the spectrum. This means that layer 2 should also be treated analytically.

Comparatively, serial communication, such as in a Controller Area Network (CAN bus), is synchronous and deterministic. The time shuffling in OSI renders time unreliable in a deterministic sense. Indeed, the nature of packet reordering renders the element of time

artificial and thus the data set cannot be integrated over ∞ if the time reordering is not accounted for.

A major outcome of the derived arbitrary sense of time is the need to analyze the newly created data-time and thus their frequency spectrums. With demultiplexing, it is often expected that the derived spectrums will be different from the true input to accommodate for the allowed data reordering among the data sets. The newly created spectrums are artifacts of demultiplexing. Demultiplexing observations are linear in their spectrum mappings. This spectrum assumption also holds inversely on a multiplexing process.

The focus is on the constant that is pointed out when applying Shannon's theorem on the information content carried by the packet reordering. It can be deduced that any transformation scheme cannot increase the total information exchange but can only decrease or equate it. Measuring this constant and its variability over time is essential in determining the maximum information that occurs at the interface.

To access the analytical features of OSI layer 1 across the boundary the linearity aspect of the Fourier transform must be recalled. The aspect of linearity is the equivalence key of the Fourier of layer 1 with the Fourier Transforms rendered over synthetic channels created at layer 2. Synthetic channels are the logical experiments ξ and are independent of the data. Demultiplexing analysis, such as accounting for Domain Name System (DNS) traffic independently from the rest of a packet flow, is an example of an experiment as one channel.

For the proposed digital data, the linearly reversible nature of discrete Fourier transforms enables the desired analytical function to occur over an arbitrary time index. For completeness in the introspection of the data, the signal phases must be considered. Digitization fails to account for the known impact signal phases have and they are therefore considered the lossy aspect of the digitization, not the analytical functions.

To assert a sense of continuity over the data signals, the analytical spectral function S_d is done in terms of the discrete Fourier transform underlying the frequencies of the data points in the sequence y(t).

In order to follow numerical solution spectral methods for the partial differential equations required in the Fisher Information Matrix (inverse covariance), the

transformations must use cosines. The ipso facto uses of cosines in the transformations lead to an analytical spectral density function. Furthermore, for differential equations the cosines express a particular choice of boundary conditions. Discrete Cosine Transform (DCT) is a lossless transformation that renders the parameters not dependent. Digital data points are analytically computed over

$$Y(k) = \sum_{n=0}^{N-1} y_n cos[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k] \quad for\ k = 0,1, \ldots . N-1 \qquad (18)$$

The data set in this equation represents the arrival of layer 2 over an IP network. The quasi-physical quantity parameters modeled in the table below are the differential times between real time and their corresponding offsets, or the delays. Lengths are also valid dimensions. In general, one can devise a model for the data sets for I/O such as physical entities derived from the frames over serialization transmissions or other transmission/reception. This analogy extends to numerous observation models.

The table below summarizes common data sets calculated for y_d and their reflection of physical world measurements:

Data Sets and Their Physical Measurements

$\mathbf{Y_d} \mid \xi$	**Variables**	**Original Spectrum**	**Channels**
Dimension(s)	e.g Bandwidth	Size	Size
Absolute(s)	e.g Time	Differential	
Relativity(ies)	e.g. Offsets		OSI model suggests the option to reshuffle the measurements ordering for layer 3.
Type(s)	E.g. Channels and mux/demux		OSI model suggests new virtual channels based on unique identifications (IP, protocol, ports).

At this stage, it is assumed that the spectrum demultiplexing for digital systems into their corresponding channels has been addressed. The consideration made here is that the N samples were collected in the relative time period τ. The notation of $y(t-\tau)$ may be expanded to $y_i(t-\tau)$ because the representation of data sets from an arbitrary channel i, representing the i^{th} experiment ξ_i for N data sets, matches the artificial time period τ. $Y_i(f)$ is the DCT and S_Y is the spectral density function. For simplicity S_i is referred to as S(i, k), a vector of k elements for a channel i.

Suppose the system in question is a non-linear multiple-input multiple-output (MIMO), a bounded-input bounded-output (BIBO) and a non-continuous discrete system. Meaning that if $|y(t)| \leq N$ for all values of t when the input is a finite-variance random process, the system is being deterministic by design and is thus a Wide-Sense Stationary (WSS). The input is a WSS process as the expected input for all t is

$$E\left[y_{0(t)}\right] = f_d(\theta_0, t)$$
$$E[y_i(t)] = f_d(\theta_0, t) \int \xi_i(s)ds$$
$$= f_d(\theta_i, t) \qquad (19)$$

where $\xi_i(t)$ is the response of the interface as an experiment for channel i. In fact, $\{ y_i(t) \}$ is also a WSS process like $\{ y_0(t) \}$. The corresponding relative deviations can also be written:

$$y_i(t) = f_d(\theta_i, t) + x_i(t) \qquad (20)$$

and through the linearity aspect of the Fourier transform, the following equation may be written:

$$Y_i(f) = \theta_i(f) + X_i(f) \qquad (21)$$

Given equation 20, the deviation of $x_i(t)$ appears to be the most informative component in the isolated stochastic process. For completeness, accounting for any

discrepancy in the de-sequencing between $y_0(t)$ and $y_i(t)$, the vector $y_0(t)$ may be asserted as

$$y_0(t) = \sum_i y_i(t) \qquad (22)$$

0 is the zero[th] channel or the original channel in a demultiplexed system for a total of (i+1) channels. Let $\bar{S}(f) = \| S(f) \|$ be the normalized outcome, one can write the partial normalization as

$$\bar{S}_i(f) = S_i(f) * [|S_0(f)|]^{-1} \qquad (23)$$

The magnitude of the spectral density is denoted as | S |. For practical purposes, S(f) has been quantized to a smaller size M, being as positive number ≤ N.

Since each channel may represent an independent experiment or SDOF, the modal analysis on the data is a possible response of multi-output data. In such case, the modal analysis will generate the realization of a nonlinear system F being sampled in the period τ. Using the frequency response with multi-output data a linearized transfer function may be written as

$$F_{partial} = \left[\frac{\delta S_i}{\delta S_0}\right] \qquad (24)$$

Where $F_{partial}$ is the partial spectral power distribution matrix.

When extending the integral of the auto-spectral and cross-spectral density function of a signal to the covariance notation of the random variables, the following relationships are established: the variance is $\sigma^2 = \frac{1}{N} \sum_I S_i$, the covariance is $\sigma^2_{ij} = \frac{1}{N} \sum_{ij} S_{ij}$ and N is the number of frequencies.

The way to consider the variability in the transfer of the power spectrum into the multi-output system, in relation to the variance and covariance, is to reflect the relationship through the cross correlations between channels.

The partial spectral power distribution matrix is one modality underlining the power transfer that is not descriptive. On the other hand, the variability analysis in spectral signatures leads to the measure of coherence. This is essential to measuring the transferability of the energy and therefore measuring the effectiveness of the experiments.

The spectral coherence is used to examine the relation between data sets. This is essential to assess if the designed experiments have drawn enough data as it estimates the power transfer between input and output of the demultiplexing process. The system is lossless as the signals are ergodic and the power transfer is linear. The causality between the input and output is therefore calculated. The coherence, being the magnitude-squared coherence between the deviations $x_0(t)$ and $x_i(t)$, is a real-valued function and is defined as:

$$C_Y = \sum_i |S_{0i}(f)|^2 * [S_0(f) * \sum_i S_i(f)]^{-1} \qquad (25)$$

and for a corresponding transfer matrix

$$F_Y = \left[\frac{\delta S_{ij}{}^2}{\delta S_i \, \delta S_j}\right] \qquad (26)$$

where $S_0(f)$ and $S_i(f)$ are the auto-spectral densities of y_0 and y_i, respectively, and $S_{ij}(f)$ is the cross-spectral density between y_i and y_j. Given the restrictions noted above concerning ergodicity and linearity, the coherence function estimates the extent to which $y_i(t)$ is composed from $y_j(t)$. F_{ij} follows the same Fisher Information Matrix as described above.

The coherence will thus always satisfy $0 \leq C_y \leq 1$, or an Identity Matrix. For an ideal system with input $y_0(t)$ and multiple outputs $y_i(t)$, the coherence will be equal to 1 or the identity. In cases where the ideal linearized system assumptions are insufficient, the Cauchy-Schwarz inequality guarantees a value of $C_y \leq 1$.

The deviation of the realization of the system in the physical world from its ideal counterpart culminates to the unaccounted spectrum and loss of information when integrated over time. While these deviations are an inherent part of the system life cycle, few known approaches are sufficient to model the deviations of the complete system dynamics.

Tracking the coherence is an element of understanding that is lost for complex systems. The information is lost because it is tracking the derivative functions of the change from the expected values for non-stationary and non-linear systems.

If the coherence C_y is not one but is greater than zero it is an indication that several different things could be happening. It could be that strong deviations are entering the data, that the assumed function relating $y_i(t)$ and $y_j(t)$ is highly nonlinear at that moment, or that boundary discontinuity is being produced by $y_i(t)$ due to input $y_j(t)$ as well as other inputs from different time segments. If the coherence is equal to zero, it is an indication that $y_i(t)$ and $y_j(t)$ are completely unrelated given the constraints mentioned above.

The coherence therefore represents the fractional part of the output signal power that is produced by the input at that frequency. The quantity (1−C) may be viewed as an estimate of the output that is not contributed by the input at a particular frequency. The quantity (I-F_Y) is used in a matrix notation where I is the identity matrix. This leads to the definition of the coherence output spectrum C_Y. $\sum_i S_i(f)$, which provides a spectral quantification of the output power that is uncorrelated with other inputs and is therefore a useful quantity.

With the parameters being differentiable, the optimization of the upper bound is reached with lower curvature dispersion in the Fisher Information matrix. In reality, achieving a system independent realization will be different from the ideal world. For $\tau > 0$, AA^T is a design matrix that is asymptotically approximate as described by Brincker et al [8]. A way to further smooth the dispersion in the Fisher Information is by using the quantization effect on the parameters instead of putting them into lesser dimensions. This also results in smaller computational complexity.

For a recursive decomposition, the variance associated with the deviation $S_{X,s}$ of an unknown true measure $S_{Y,s}$ is given by

$$Q_{s|s-1=}E[(S_{X,s}) * (S_{X,s})^T] \qquad (27)$$

and evaluated recursively with $S_{Y,s|s-1}$

$$Q_{s|s-1} = E[(S_{Y,s} - \theta_{s|s-1})(S_{Y,s} - \theta_{s|s-1})^T] \qquad (28)$$

For a system transfer function F evaluated at sample s, the influence of the deviation on the expected value is solved for, minimizing the variances, and the following equation is obtained:

$$w_{s|s-1} = P_{s|s-1}F_s^{T}(A_s P_{s|s-1} A_s^{T} + Q_{s|s-1})^{-1} \qquad (29)$$

$P_{s|s-1}$ is the prior distribution and w_s is the weighted gain from the deviation. The posterior θ_s estimates and covariance are thus correspondingly updated to

$$\theta_s(f) = \theta_{s-1} + w_s(\theta_{s-1} - S_Y)$$
$$P_{s|s} = P_{s|s-1} - w_s F_s Q_{s|s-1} \qquad (30)$$

The impact of the deviation gain w_s and the parameters θ_s are outcomes and come from tracking the covariance P_s. The entropy of the parameters and the Fisher matrix are also two key measures.

One principle discussed herein is tied to Shannon's Information Theory in that {0, 1} cannot be created without an analog counterpart. In that case, cyber-security would need to consider unexpected data a threat. In the very least, unmodeled data should be assessed for what it is worth in information value. Comparatively in the digital world, the trend has been that unexpected and unprocessed data is discarded in favor of the event that arises. Digital forensic tools have been a proposed remedy but are post process in most cases.

Considering when data is collected in a continuous manner, the variability in the outcome may induce deviations from and expected trend and yet but both, deviation and expected trends, are new measures being context-independent from the actual data. Deviations are exposed with differential methods. For example, if a measured result is correctly predicted, the *a priori* entropy of the result is small relative to that expected. The only information at hand, is that it is predictable at that moment. On the other hand, when the deviations become substantially evident, the amount of new information is carried to the extent of the deviations.

To analyze a diverse context-independent data set, one may seek to measure the information index rather than attempting to understand the data. This approach takes a

direct measure of the maximum information per bit. Roughly speaking, when applying Shannon's theorem to the information content carried by the deviation, one can deduce that any transformation scheme cannot, on average, have more than one bit of information per bit of deviation. This means that any value less than one bit of information per bit of deviation can be attained. The entropy of a deviation per bit multiplied by the length of that deviation is a measure of how much total information the deviation contains. Therefore, the maximum amount of information is calculated as the entropy of all the deviations.

Thus, the entropy becomes a measure of unpredictability, or equivalently, of its average information content. Considering the normalized spectrum $\bar{S}_i$, when the deviations are normalized and constrained the following equation is obtained

$$\sum_j \left|\bar{S}_{ij}\right|^2 = 1 \qquad (31)$$

Summing all j deviations in channel i, the following equation can be written

$$P_i = |\bar{S}_i|^2 \qquad (32)$$

to be considered as a discrete probability mass function, with an associated probability mass function constructed from the transformed variable x_i. Shannon's entropy H for the spectrum $\bar{S}_i$ and probability mass function $P(\bar{S}_{Xi})$ can be written as

$$H(\bar{S}_i) = -\sum_j P(\bar{S}_{ij}) \log P(\bar{S}_{ij}) \qquad (33)$$

In cryptography, a "key" is information that determines the output of a cryptographic algorithm. These algorithms are typically difficult to protect and as such are often assumed to be known. A key is byproduct of a cryptographic algorithm, is easier to change if compromised. As such, the security of a system usually relies on the security of the key. The risk is therefore computed based on the prior of being able to compromise the key with certain confidence.

Maintenance of the key carries the process of changing the keys is deterministically defined where the maintenance follows a fixed time period. In parallel, security products have followed an evolutionary strategy of patching and therefore the system overall

weakens over time. Failure modes become the opportune times when the weaknesses of the security systems put in place align. For example, double factor authentication may appear to increase the overall authentication but it may be hiding the weakness of the keys. In other words, the keys may have already been compromised but because the complete authentication process is still effective the compromise is unknown.

These secrets, used in special or generic purpose connected networks, can be expected to degrade and fail as a function of time and usage. The convention of secrets being replaced or upgraded periodically, in a fairly arbitrary fashion, is a backdoor to the new layers of security solutions. Furthermore, digital asset protection in security revolves around designing very complex secrets and it therefore stops short in deterring failure modality beyond the threat of deciphering the secrets.

Secrets are originally designed with a difficulty level that reflects an index of randomness and complexity. With the diverse types of configurations in security setups, it is challenging to quantify an index strength for original setups to start with. Typical usage and conditions on the difficulties in many cyber-security instantiations are reflected in the secrets expiration date. Expiration dates are derived analytically from the secrets known states in certain usage conditions.

Suppose a secret is replaced every x days but is used enough in those x days that the secret has been compromised for x - T days. The compromise period for the secret is therefore T days. If a compromise does occur it does not necessarily mean that the secret was deciphered; in some compromised security cases the secrets are hijacked. If the key was a private key, it could be used to create others keys (i.e. symmetric keys) and perform certain illicit transactions without the need to decipher the secret.

The information entropy refers to a dynamic value that is independent of the underlying keys or digital assets. It reflects new information certainties about the assets as erosion factors. In MES, these information certainties are density functions computed from the marginal distributions on secrets usage. For the compromise mentioned above with the compromise period of T days, the registered trust score of the key will have degraded differently than expected. Decisions are therefore made based on the tolerance of the degradation.

A review on how experiments and channels are constructed from the materials and methods section for the OSI framework is necessary. Channels are constructed with parameter vectors about the network protocol, packet payload size, time delays among network packets, DNS answer vector, and another vector for errors, etc.

If one needs to place digital watchdog experiments on devices because observers are constantly updating models on ingress and egress traffic, it would be a trivial problem if the sentries are only knowledgeable about the traffic *a priori*. Suppose these digital watchdogs are placed as observers on an Ethernet layer 2 network. These watchdogs are designed to process layer 2 packets for IP sources, destinations, protocols, packets size, time delays, errors, drop packets, repeat packets, etc. Furthermore, suppose these watchdogs can produce the indices about the deviations as explained in previous sections. The watchdog entropy indices H(t) therefore indicate the quality of the coherence measure, the system entropy of the models Θ degradation over time, the deviation entropy of the normalized covariance w update, and the update entropy of b. This means that the watchdogs are bounded to the scope of the sampling vector.

Whether the end example is a device such a switch, router or any electronic device; the calculation to generate the differential entropy H(t) reflects the degradation index, use, and span of the device.

When the scope of the watchdog experiment is coherent with the data a threat could build its own independent analysis. The final degradation, or the Degradation Factor (DF), may be computed as a function of the entropy degradation against the entropy of the likelihood distribution of the device behavior.

Knowledge about the device manufacturing would become imperative if measured in universal units, such as entropy, and would create a generalization set of dimensions, such as expiration dates. For example, if the device strength is about a secret, the DF is used against the *a priori* entropy of the strength. For clarification, entropy is also used in key cryptography as the measure of the randomness. A degradation can therefore be done as a direct multiplication. In general, other parameters that govern the keys are also considered in quantitative fashion.

Cyber-security enterprises will immediately benefit from these techniques and can extend to compliance or audit datasets, cyber datasets, and network management or

assurance dataset use cases. These watchdogs may be distributed to run efficiently on endpoints such as APs, edge routers, IoT devices, etc. These endpoints may observe communications in the network or the application layer. Either way, a DF factor is reported on the device telemetry to a data center.

These techniques may be applied to all layers of the OSI stack that utilize keys for session or application authentication of communication protections (e.g., encryption). The extensibility of these methods in the Digital world reaches beyond the network. For example, if a secret protects an application, as opposed to a network, messages such as database access and password systems pertaining to that application may be observed.

A user password is one type of private key where a typical user is familiar with the strength during the creation process. While degradation in digital assets covers a diverse range of assets, login passwords substantially degrade with little oversight. The impact of these secret degradations has grown substantially in the consumer world. Systematically, the applicability from personal computers, Internet of Things (IoT) devices, mobile devices, to online accounts in the cloud follow suit. In these use cases, the physical or logical model vectorizations are evaluated as density functions. These functions extend to other forms of I/O such as keyboard I/O and password database access I/O (e.g., keychain databases).

In the paradigm of MES, information entropy may be aggregated across similar systems. In the case of a host and multiple physical interfaces, it is assumed that each localized entropy is calculated for each device and channel in network switches, routers or simply IoT devices. Comparative analysis is constructed on each individual component and therefore differentiates the accentuating usage. The advantage of using Shannon's entropy for the information index is that because it is linear, basic linear algebra over subsystems and systems on a large network is allowable. One can therefore devise a sharing mechanism (i.e., over the network) and cause an individual device to potentially have its entropy aggregate at a global scale.

For large scale system distributions, individual and class degradation factors are analyzed with basic theoretical graph frameworks for decisions. An example is a Naïve Bayes graph measuring the propagation of the degradation over the network. This phenomenon is extremely useful when certain behaviors need to be isolated and investigated, also known as real time forensics. There are numerous forensic approaches

that allude to reinforcement with the appropriate policies in the control plane. The end result is an isolation of the systems and devices until the scores are restored to the expected level.

Productization of the proposed approach using MES in an environment can be achieved either locally or remotely on the device or system in question depending on telemetry data. The weakness of the analysis from telemetry is that it fails the Turing test since there is always a question on the authenticity of the telemetry itself. In fact, a computed system score has the authenticity and complexity of a ledger when computed in situ.

Many software-based techniques develop probabilistic models of the validity or security of a private or public network. They may employ machine learning over network telemetry and big data in order to indicate whether the actual data can compromise a target system. These techniques for detecting threats may involve using supervised learning to technique and pattern match data for certain known goals. Unsupervised learning has also been used for classification and identification of new dimensions.

This is close in principle to unsupervised learning only when the new classified dimensions are assessed in their statistical meaningfulness and computed by the unknown.

Not many embedded solution and runtime approaches proactively track DFs. Typical architectural environments in enterprises and service provider environments would be similar to the Fair Isaac Score (FICO score) tailored for security assessment on device level IoT switches and gateways. FICO scores have been instrumental in financial security.

Historically, the principles in security have been to exclusively rely on inputs. These techniques have evolved from supervised learning to unsupervised learning. This leads to classification and pattern encoding and/or inverse covariance models. In many situations, these methods require searching through multiple terabytes of data.

The originating problem in security is whether or not the calculation of the risk for a secret is compromised. This problem is a multivariate in which the solution space is not known *a priori*. That is the nature, type, and approach of possible attacks are not known *a priori*. Unsupervised learning intends to address this kind of problem. Nonetheless, modern reinforcement learning has shown its capability in handling unsupervised auxiliary input

without biasing the base models learnt by the engine while allowing for identifying outlier situations.

The entropy computation leverages the statistical distribution and shaping from the smoothed discrete Fourier transform in order to increase the accuracy and intelligence of scoring the risk and/or threat associated with secrets. This mathematical approach of computational scores being the output of the logic running on the platform may improve the accuracy of the learning supervised system running, for example, a theoretical aggregate decision in the cloud.

Accordingly, the proposed system can also be embedded into unsupervised learning for digital asset usage and degradation in a system or a network. Secret analytic models are applied to control plane Software Defined Networking (SDN) for analytic model exchanges, known as orchestration.

This may apply to various IoT edge use cases such as connected vehicle energy or utilities, oil and gas, and manufacturing. Embedded binaries may be built for edge devices, such as connected grid routers and integrated services routers, that perform assurance monitoring and local anomaly detection.

Tampering of far edge communication systems and compute devices is therefore preventable. Furthermore, avoiding being tampered by the far edge becomes also feasible. This may apply in various use cases, such as in surveillance, transportation and asset tracking for military and civil use.

The techniques described herein provide embedded endpoint self-protection, which leads to sustainable approaches for long running periods over multiple years. The software may run in the embedded environment, container, Operating System (OS), etc. to avoid the heavy investment of big data analytics in the cloud. The runtime may relate to conducting the learning and security threat as a measured part of the exposure risk at the edge.

The analytics are based on Information Theory. In general, the more certainty or data, the lower the entropy. In the real world, behavior is rarely random and patterns are evaluated for what they are. Patterns lead to certainty and certainty reflects a lower entropy. The amount of information collected by simply accessing a secret is measured. Interfaces may be used to measure such information about access as reflected in an entropy index.

Trust erosion is nothing more than the ratio of the impact, both perceived and hypothetical, of the unknown over what is known. While there is no clue whatsoever about the unknown, the impact is estimated to follow an exponential growth over time. Therefore, the known will decline in the same order of scale. The break-even is reached when learning occurs fast enough to catch up with the rate of the growth of the unknown. In truth, cloud systems are far from being as adaptive and dynamic as they ought to be given today's total dependency on a digital world.

When these concepts are formalized into a Bayesian framework, they take the form of a time series likelihood optimization. The framework is a continuous differential analysis comparing measurements to a posterior of a likelihood estimation. The resulting error or deviation is the most informative gain. The deviations (unknown) are marginalized (unmodeled), and it becomes imperative to solve for their magnitude, or $||\{\neg k| \neg k\}||$, as an information measure. It is not necessary to have models of the deviations to make decisions. In many cases, high entropy data - as it may get closer to random - are mathematically complex as it may be highly non-linear and close to noise in dimensionality. Measuring the magnitude (information content) is a simpler task and can quantify uncertainty. The measure of uncertainty is computed using Shannon Information Theory.

Naturally, this lends itself to finding the maximum bit encoding measures as information. This is readily available from the data without the need to decode the data at all time. Independency is achieved across platforms, and the information index measure spans all data, whether encrypted or not. The magnitude of the uncertainty is entropic and will be addressed with corresponding information theory mathematical measures.

Optimally sampling in the uncertainty space coincides with the MES theory. It lends itself to maximizing the information and minimizing the entropy for non-stationary processes, which coincides with the nature of sampling the unknown. Sampling for deviations is the same problem as sampling points on a multidimensional function and is a stochastic process. The usage of maximum likelihood estimation and MES approaches to maximize information follows the typical implementation of these theories.

In practical example, measuring the weight of a bucket is informative to a user who has developed a sensitivity sense to weights without having to sort out what is in the bucket. When bucket weights comparison sampled in different location will become informative

about the underlying topology changes, or the deviations in a cyber-security case. A more popular use case than bucket comparison analysis is the science behind Searching for Extraterrestrial Intelligence (SETI) computational algorithms, even though in SETI "Intelligence" is not even well-defined.

The science behind building a Trust Evaluation Function (TEF) score, roughly speaking, is when applying Shannon's theorem to the information content carried by the deviation, one can deduce that any transformation scheme cannot, on average, have more than one bit of information per bit of deviation. This means that any value less than one bit of information per bit of deviation can be attained. The entropy of a deviation per bit multiplied by the length of that deviation is a measure of how much total information the deviation contains. Therefore, the maximum amount of information is calculated as the entropy of all the deviations.

In summary, techniques are described herein for managing unknown-unknowns in cyber-security. Trust degradation is a precursor index to failure. The use cases of scoring the trust degradation in a system span to almost every aspect in networking, edge and cloud included. A well devised TEF will cover many use cases: for example (1) better and adaptive private key management (e.g., re-keying); (2) better and adaptive end user experience password management and its fine grain monitoring in a data center; (3) better and adaptive digital asset certifications; (4) troubleshooting; and (5) real-time scalability and risk assessment for extremely large network, for example in federated cloud environment. The features of a digital trust scoring will start to reflect the likelihood of erosion of trust created on day 0. Platform independency is achieved when the score is a degradation of the trust and not the trust value alone. A trust value may start erroneously, but the rate of change lead to continuous evaluation. Therefore, the originating trust is set as a prior. Erosion will thus work with time against the assumed original trust. In the example of an expiration date or a combinatorial complexity erosion of a private key, the realization of a trust erosion is not a Boolean fail pass type, but a relative factor number. On a comprehensive integrated analytical dashboard, the trust factor produces the percent life left of given a digital secret.

References

[1] H. P. Wynn, H. Läuter, and A. Bucchianico, *Maximum Entropy Sampling and General Equivalence Theory*. (eds) mODa 7 — Advances in Model-Oriented Design and Analysis. Contributions to Statistics. Physica, Heidelberg (2004).

[2] J. Sacks, W. J. Welch, T. J. Mitchell, & H. P. Wynn, "Design and Analysis of Computer Experiments." *Statist. Sci*., 4(4):409–435, 1989.

[3] N. Youssef, "Optimal Experimental Design for Computer Experiments," Ph.D. dissertation, London School of Economics, 2011.

[4] C. J. Paciorek & M. J. Schervish, "Spatial Modelling Using a New Class of Nonstationary Covariance Functions." *Environmetrics,* 17(5):483–506, 2006.

[5] P. Sebastiani & H. P. Wynn, "Maximum Entropy Sampling and Optimal Bayesian Experimental Design." J. R. Stat. Soc. Ser. B Stat. Methodol., 62(1):145–157, 2000.

[6] M. C. Shewry & H. P. Wynn, "Maximum Entropy Sampling." *Journal of Applied Statistics*, 14:165–170, 1987.

[7] R. R. Zhou, N. Serban, and N. Gebraeel, "Degradation Modeling Applied to Residual Lifetime Prediction Using Functional Data Analysis," *Annals of Applied Statistics* 2011, Vol. 5, No. 2B, 1586-1610.

[8] R. Brincker, L. Zhang, and P. Andersen, *Modal Identification from Ambient Responses using Frequency Domain Decomposition*. Proc. of the 18th International Modal Analysis Conference (IMAC), San Antonio, Texas, 2000.

[9] R. M. Solow, "A Contribution to the Theory of Economic Growth," *The Quarterly Journal of Economics*, Vol 70, No. 1, pp. 65-94, 1956.

[10] T. M. Cover, J. A. Thomas, "Elements of Information Theory," *Wiley Series in Telecommunications,* Donald L. Schilling, Editor, 1991.

[11] P. Sebastiani, H. P. Wynn, *Bayesian Experimental Design and Shannon Information*. Proceedings of the Section on Bayesian Statistical Science, 1997.

[12] K. Chaloner and I. Verdinelli, "Bayesian Experimental Design: A Review," *Statist. Sci*., Volume 10, No. 3, 273-304, 1995.

[13] G. J. Klir, "Uncertainty and Information: Foundations of Generalized Information Theory," *Wiley-IEEE Press.*