Towards a Science of Trust

Dusko Pavlovic
University of Hawaii
Honolulu, USA
dusko@hawaii.edu

It is unclear how to think about trust or to model its ebb and flow. Is there some sort of Second Law of Thermodynamics of trust, where trust starts high and is dissipated over time? Or is it the contrary, that trust starts low and can grow through a series of good experiences? Is it more complex, and how can the waxing and waning be thought about?

JASON Report on Science of Cyber-Security [17]

ABSTRACT

The diverse views of science of security have opened up several alleys towards applying the methods of science to security. We pursue a different kind of connection between science and security. This paper explores the idea that security is not just a suitable *subject* for science, but that the process of security is also *similar* to the process of science. This similarity arises from the fact that both science and security depend on the methods of *inductive inference*. Because of this dependency, a scientific theory can never be definitely proved, but can only be disproved by new evidence, and improved into a better theory. Because of the same dependency, every security claim and method has a lifetime, and always eventually needs to be improved.

In this general framework of *security-as-science*, we explore the ways to apply the methods of scientific induction in the process of trust. The process of trust building and updating is viewed as hypothesis testing. We propose to formulate the trust hypotheses by the methods of algorithmic learning, and to build more robust trust testing and vetting methodologies on the solid foundations of statistical inference.

Categories and Subject Descriptors

K.4.4 [**Computers and society**]: Electronic commerce— *Security*; K.6.5 [**Management of computing and information systems**]: Security and protection; D.4.6 [**Operating Systems**]: Security and Protection

General Terms

Theory, Security, Experimentation

Keywords

trust, hypothesis testing, statistical inference, science of security, intrusion

1. INTRODUCTION

The effort towards science of security was born from the need for a more systematic approach to security [17, 26, 22, 38, 34]. It resulted in new empiric and experimental approaches to cyber security [3, 29, 30]. The fact that science of security still means many things to many people should perhaps be seen as a feature and not a bug, since already security on its own means many things to many people, and it is natural that they study it from many directions [34]. On the other hand, it seems that each step of scientific progress requires a unifying idea, each of them showing that a certain group of trees is actually a forest [20]. What is then the unifying idea of science of security?

1.1 Science is something else

1.1.1 *What science is not*

Every known civilization seems to have developed technology, art, and religion. But only the Western Civilization has developed science. Science emerged in Europe during the Renaissance, and caused the Industrial Revolution. This unique stream of events is analyzed in some detail in [20].[1]

There are, of course, many definitions of science. Some of them are shaped to include the teachings of Ron L. Hubbard; some to include marxism, or even the daily thoughts of the current leader of North Korea. Most definitions, however, point to some of the features of the methodological movement that led to understanding the natural processes

---

[1] It has been objected that this view can be construed as eurocentric. While the word *"science"* can, of course, be used to denote many things, as explained in the next paragraph, theory of science *defines* science as the movement that led to the Industrial Revolution. Since the fact that the Industrial Revolution emerged in Europe is historically uncontestable, the fact that science emerged in Europe follows from this definition. Moreover, as incisive critics of the Industrial Revolution even before its current destructive consequences became clear, the theorists of science can hardly be accused of *praising* Europe for being the cradle of science [11].

like heat, electricity, magnetism, radiation, or networking. Although the notion of science can be extended to include astrology, scientology, theology, mathematics, or engineering, it does not seem useful to stretch it too much. Assigning the status of science, say, to the engineering principles and processes (whether those that enabled the public works of Ancient Egypt, or those that emerged in medieval alchemy, or in Renaissance architecture, or in modern software engineering) might conceal something essential about science. Can science be reduced to its technological thrust [22]? Or does it boil down to the view that the world is governed by a system of laws [38]? Or is there more to it?

Many ancient civilizations developed the *quantitative methods* that enabled them to plan and execute extensive engineering projects, and change the landscapes of their environment. Many of them also explained the world around them through sophisticated *theoretical edifices* and that included the *'Laws of Nature'*, formalized as mythologies, or gathered in sacred texts, often equipped with extensive symbolic systems. But no one until the Age of Science came anywhere near to understanding and reproducing, e.g., the thermo-nuclear processes of Sun; or the space-time curvature, without which our GPS systems could not surf on the geodesics, and would keep sending us wrong coordinates. No one before science came anywhere near to understanding genomics and to engineering the basic processes of life; and nowhere near to connecting our world into a network of networks, and spanning a distance-free space, where every two nodes are neighbors, and where our joint problem solving, and problem creating capabilities seem to be reaching a completely new level. This network of networks is what we call cyber space. Inhabited by the processes that we programmed, but whose interactions we cannot control, cyber space hosts a new nature in need of a new science. Is this new level of our civilization just another new level of yet another civilization, or is it something else? Is the science that brought it all about just another way that we found to generate new technologies, or just another religion that tells us the laws of the world, or is it something essentially new?

There is a qualitative difference between the science-generated technologies, and the spontaneously evolved technologies. There is also a qualitative difference between the symbolic systems of theologies and mythologies that emerge from religions, and the symbolic systems of mathematics and computation that underlie science. There is a qualitative difference between the religious rituals on one hand and the scientific protocols on the other. The essence of these differences is not in the levels of complexity or effectiveness. There are complex religious systems, and there are simple scientific theories. Many religions and even superstitions postulate their *'Laws of Nature'* that are structurally indistinguishable from those postulated by science. Astrology and phrenology have in their time been tested as scientific theories, by scientific methods, and rejected not for structural reasons, not as unscientific, but as wrong. And there are also effective religious systems, and there are ineffective sciences. E.g., although the processes of photosynthesis are everywhere around us, at the bottom of all of our food chains, science has remained unable to understand what do the plants really do when they bind photons into sugars. There is a quantum effect, but science has been ineffective in explaining it. It has also been less effective than most religions in addressing people's emotional and social needs.

So what really distinguishes scientific theories, if it is not complexity, and not effectiveness?

1.1.2 What science is

I propose to consider the logical pattern of inductive inference as the essence of science: *While religion claims to provide the truth, science only seeks to disprove false hypotheses.*[2]

In a formal sense, science is the quest for *disproving* theories. This formal sense was fully implemented for the first time in Ronald Fisher's practical methods of scientific inference [13, 14], and then analyzed theoretically in Karl Popper's extensive and influential work [36]. The historic support for this view of science was provided by Thomas Kuhn [20], while the scientists themselves provided some of the most compelling examples from their current practices [12]. Other leading templates of scientific inference (e.g. the Neyman-Pearson testing [28], or Bayesian inference [1, 5]) may appear to offer ways beyond this negative logic of science, as the quest for merely *improving* scientific theories through *disproving* false hypotheses. But a closer look shows that they only formalize the task of hypothesis selection, and thus support formation of new theories, not proving. They do not provide a method to definitely prove anything. Richard Feynman announced this with compelling simplicity in his lectures on *'The Character of Physical Law'* [12]:

> If we have a definite theory, from which we can compute the consequences which can be compared with experiment, then in principle we can prove that theory wrong. But notice that we can never prove it right. Suppose that you invent a theory, calculate the consequences, and discover every time that the consequences agree with the experiment. The theory is then right? No, it is simply not proved wrong! In the future you could compute a wider range of consequences, there could be a wider range of experiments, and you might then discover that the thing is wrong. [...] — *We never are definitely right; we can only be sure when we are wrong.*

This is perhaps the best kept secret of science: *Science does not provide persistent theories; it only provides methods to disprove and improve our hypotheses.*

1.2 Security is like science

The fact that the process of security is of the same type like the process of science can be illustrated by translating Feynman's statement from the language of science to the language of security:

> If we have a precisely defined security claim about a system, from which we can derive the consequences which can be tested, then in principle we can prove that the system is insecure. But we can never prove that it is secure. Suppose that you design a system, calculate some security claims, and discover every time that the system remains

secure under all tests. The system is then secure? No, it is simply not proved insecure! In the future you could refine the security model, there could be a wider range of tests and attacks, and you might then discover that the thing is insecure. — *We never are definitely secure; we can only be sure when we are insecure.*

A scientific approach to security must therefore begin with the realization that there is no persistent security. Cryptographers have known for a long time that every key has a lifetime. It is time that we recognize that every security claim has a lifetime. The designers of protocols and systems have, of course, accumulated a lot of empiric evidence about this phenomenon [32]. The point is to understand it as a *logical* phenomenon.

Upon the admission that theories cannot be definitely proved, but only disproved and improved, science has gained its current unparalleled power to harness nature. Upon the realization that security guarantees cannot be definitely assured, but only broken and strengthened, science of security will gain the ability to tap its power to protect from the same methodological source.

1.3 Zoom on trust

In this paper we focus on the scientific approaches to a special family of security claims: the statements of *trust*. While a general security claim says that a key K is uncompromised, or that a protocol P guarantees an authentic channel, a statement of trust says that Alice trusts the key K for use in a particular cipher, or that Bob trusts the protocol P to establish an authentic channel with Alice. A statement of trust is thus a security statement *bound* to two subjects and an object: *who trusts what to whom*. The parallel between the security processes of trust building and the scientific methods of hypothesis testing seems like a particularly good illustration of the general logical link of security and science, so we pursue it in the rest of this paper.

Outline of the paper

In Sec. 2 we briefly explain the concept of trust used in the paper, and why is it interesting to model the process of trust as hypothesis testing. In Sec. 3 we show on toy examples how to apply the three standard methods of statistical inference in trust testing. In Sec. 4 we show how to formulate the best trust hypotheses *a priori*, since it is notoriously difficult to extract the normal behavioral profiles from empiric data. In Sec. 5 we comment about the relations of the presented ideas with the other views of trust, and with the application of statistics in intrusion detection.

2. TRUST AS HYPOTHESIS TESTING

2.1 What is trust?

Security analyses often begin with the assumptions that some of the subjects are *honest*, i.e. that they behave according to some prescribed protocol rules, whereas the others are dishonest, and launch attacks. Trust internalizes the honesty assumptions into beliefs of subjects about each other. E.g., we say that Alice trusts Bob if she believes that he will behave honestly, according to some protocol agreed implicitly or explicitly. In such a trust statement, Alice is the

trustor, and Bob is the *trustee*. In social and electronic networks, and on the web, trust is implemented in a variety of ways: as feedback services in web commerce, as the web of trust or certificate authorities in the various versions of Public Key Infrastructure, etc. The underlying trust models often include *trust ratings*, which quantify trust, and the *entrusted concepts*, which qualify trust. A survey of the models of trust used in computer security research can be found in [18]. Dynamics of the trust processes in network computation were analyzed in [15, 31, 33], and the problem of trust was introduced in the framework of science of security in [16].

2.2 Inductive inference of trust

Just like science can never settle but has to keep testing its theories and refining its hypotheses, trust can also never settle and needs to keep testing its hypotheses. Just like a scientific theory can always turn out to be wrong, trust can always be broken. The reasoning about such ongoing processes goes under the name of inductive logics, which is quite different, and much less familiar than deductive logics. The central problem of the inductive inference of trust is expressed by the central principle of the modern court of law, i.e. the principle of *due process*: that the accused must be *presumed innocent until proven guilty* [35]. But this is just the legal form of a more general social principle of trust: that people should be *trusted until proven untrustworthy.* The burden of proof is here on the prosecution, or on the accusers. The dual principle of *ordeal*, typical of medieval trials, places the burden of proof on the defense, and requires that the accused be *presumed guilty until proven innocent*. The corresponding social maxim is the principle of distrust (or suspicion), namely that people should be *trusted only if they are proven trustworthy*. These two views of trust, the optimistic and the pessimistic one, correspond to the two social functions of trust:

- to support stable social links based on *cumulative trust*: "I trust you because I know you"

- to enable new social links through a *leap of trust*: "I trust you although I don't know you"

Note that both the trust principle and the suspicion principle are asserted in a logical process akin to science: they are hypotheses that need to be tested. The logical parallel described in the Introduction emerges again: just like a scientific theory can always be disproved by a new experiment, but can never be definitely proven, trust can always be broken, and can never be settled. We just follow this parallel.

2.3 How to trust methodically?

The scientific method is the method for hypothesis testing through empiric validation. This means that a scientific theory can only be validated on a finite number of samples or instances, since the empiric data are always finite. Hence the asymmetry of inductive inference: while a counterexample can definitely disprove a theory, no amount of experience can definitely prove it. This is where the *problem of induction* emerges [21].

Statistical methods have been developed as tools for deciding when to reject a hypothesis [13, 14], and also which

alternative hypothesis to endorse [27, 28]. In the experimental setting, statistical methods moreover allow testing multiple hypotheses and quantifying their likelihood [9].

2.4 How many trust values?

Up to the point where the trust *decisions* need to be made, trust can be quantified in many ways, reflected to some extent by the trust ratings, as mentioned in Sec. 2.1. There may be many colors, shades, and nuances of trust, in-between trusting and not trusting. At the end of the day, though, a trust decision must be extracted: *Will the trustor trust the trustee enough to enter into the entrusted transaction?* At the moment of decision, all previous considerations are reduced to one of the two answers: *yes* or *no*. This simple outcome is not only the *process requirement* of trust, akin to the process requirement of justice, where the verdict of *guilty* or *not guilty* must be extracted from whatever mixture of subtle and dubious concerns may precede it. More importantly, the final trust decision is in principle also the *only* observable manifestation of trust. The rich models of trust are our theories, attempting to explain the unobservable causes of the trust decisions. With such theories, science always does the same thing: it tests them as hypotheses, and decides whether they should be rejected or not yet. The good news is that the trust process seems similar. The bad news is that the *yes-no* decisions are not simple.

In Sec. 3, we sketch how the basic statistical methodologies apply to trust decisions, i.e. how the trust hypotheses can be tested scientifically. In the subsequent Sec. 4, we discuss a harder problem of trust science, that does not yield to the standard methodologies: *how to formulate the trust hypotheses for testing.*

3. TESTING TRUST HYPOTHESES

Suppose that you are interacting with a system $\mathcal{S}$ presented by a set of observable behaviors $\mathcal{B}$. Depending on the ongoing observations of the system behaviors, you must make decisions whether to entrust the system with some critical or security sensitive operations. For instance, if $\mathcal{S}$ is a computational device, then $\mathcal{B}$ can consist of the various computational behaviors: it may run fast or slowly, it may crash or spontaneously restart, it may show high or low CPU load, frequent or intermittent network accesses, various power usage behaviors, etc. If $\mathcal{S}$ is a closed network or a large organization, then the observable behaviors $\mathcal{B}$ may consist of the various network phenomena, such as local load imbalances, clustering and community formations, network chatter or its absence, and so on. If $\mathcal{S}$ is a market segment or a network of contractors, then $\mathcal{B}$ consists of the various market behaviors: clear or unclear market positions and strategies, pricing drift, shifts in supply or demand, overt or covert information propagation. In all cases, it is interesting to assume that the observable behaviors conceal some ultimately unobservable causes: the computational device may have a firmware virus or a hidden hardware component; the organization may be penetrated by undetectable moles, or bubbling with defectors; the market may be manipulated by a colluding cluster, or swayed by hidden incentives. — *Science offers methods to detect the unobservable causes of some observable phenomena.*

The observations of the observable behaviors $\mathcal{B}$ are modeled by a real function $f : \mathcal{B} \to \mathbb{R}$, which is often called a *statistic*. A statistic may list the raw measurements of a sample, but it more often displays some property, e.g. the mean, the deviation, a higher-order moment, or some other combination of data.

One thing that a statistic does not display is a *distribution* of the behaviors in $\mathcal{B}$. The distribution of the behaviors, i.e. how often does a behavior $b \in \mathcal{B}$ come about in a system $\mathcal{S}$, is what a scientific analysis attempts to induce from the observations. More precisely, a scientific analysis proceeds by

(1) setting a hypothesis θ, presented by a probability distribution $\mathrm{Pr}_\theta : \mathcal{B} \to [0, 1]$, and then

(2) testing whether the statistic $f : \mathcal{B} \to \mathbb{R}$ supports or disproves the hypothesis θ.

In the context of trust, the probability distribution $\mathrm{Pr}_\theta : \mathcal{B} \to [0, 1]$ is intended to capture the trust profile of the system $\mathcal{S}$: e.g., how often does it manifest the undesirable behaviors, how reliable is its track record, etc. Testing the trust hypothesis θ should tell us whether to stick with it, or replace it with another trust statement.

In this section, we assume that the trust hypothesis Pr_θ is given: e.g. from the records of past behaviors. The statistic f presents a new record, capturing recent behaviors. The task is to align the two. The problem of formulating Pr_θ will be discussed in the next section.

3.1 Significance testing of trust

For simplicity, assume that the system $\mathcal{S}$ has just 4 observable behaviors, collected in the set $\mathcal{B} = \{a, b, c, d\}$. To be trustworthy, the system should manifest the acceptable behavior a at least 98% of time. It may block b, or crash c for .5% of the time, and it may delay d its functioning for 1% of the time. So we postulate the *null hypothesis* that the system $\mathcal{S}$ behaves according to the probability distribution $\mathrm{Pr}_0 : \{a, b, c, d\} \to [0, 1]$ displayed on Table 3.1. For

$\mathcal{B}$	a	b	c	d
Pr_0	.98	.005	.005	.01

Table 1: Trustworthy behavior

even more simplicity, assume that we observe just one of the events from the set $\{a, b, c, d\}$. This means that the statistic $f : \{a, b, c, d\} \to \mathbb{R}$ will have the value 1 for one event, and 0 for the rest. Should we continue to trust the system $\mathcal{S}$?

In statistics, the answer to this question is reduced to determining whether the sample represented by the statistic f is *significant* enough to reject the null hypothesis (which was in our case that the system $\mathcal{S}$ was trustworthy). The idea of statistical *significance testing* is that the observation f is significant enough to reject the null hypothesis just when the observation f is extremely unlikely according to the null hypothesis. So we could fix a very small number $\alpha > 0$ and say that the null hypothesis should be rejected if x is observed such that

$$\mathrm{Pr}_0\big(f(x) = 1\big) \quad < \quad \alpha \qquad (1)$$

Since the times before computers, the scientists got in the habit of tabulating and using $\alpha = 5\%$ and $\alpha = 1\%$. So if we

use $\alpha = 1\%$ and observe b or c, we would have to reject the null hypothesis, and stop trusting the system $\mathcal{S}$; and if we observe a or d we could continue to trust it.

But to not oversimplify things, we should mention that already the founder of statistics, Ronald Fisher, argued in [13, 14] that a test should be considered significant and the null hypothesis rejected only when

$$\sum_{\mathrm{Pr}_0(y) \leq P} \mathrm{Pr}_0(y) \quad < \quad \alpha \qquad (2)$$

where $P = \mathrm{Pr}_0\big(f(x) = 1\big)$ for the observed event x. In words, the total probability of all events y that are at least as unlikely as the observed event x should be less than α. The left-hand side of (2) is the *p-value* of the observation f under the hypothesis Pr_0. The *p-value* of both b and c is now .1, and the null hypothesis is never rejected. The *p*-values for a and d are 1 and .2 respectively.

Remark. It should be noted here that significance testing is a typical embodiment of the *negative* logics of scientific induction: a test is only significant if it *disproves* the null hypothesis. This aspect of inductive logic is similar to the proof by contradiction in deductive logic; but it is different from deductive logic in that this is the *only* inductive proof schema, while deductive logic also has the positive schemas. This logical constraint is *just* what makes inductive logic and the scientific methodologies built upon it, suitable for the reasoning about security and trust, as it echoes the fact that they can always be broken, and cannot be assured by logics.

3.2 Powerful testing of trust

While the significance testing allows rejecting the null hypothesis when significant tests are found, it does not allow drawing any conclusions about the null hypothesis when it is not rejected, and no conclusions about the other hypotheses when the null hypothesis is rejected. The testing method devised by Neyman and Pearson [27, 28] considers two competing hypotheses $\mathrm{Pr}_\theta : \mathcal{B} \to [0,1]$, for $\theta \in \{0,1\}$, and maximizes the probability that the null hypothesis $\theta = 0$ is rejected when the alternate hypothesis $\theta = 1$ happens to be true. This probability is called the *power* of a test.

It is assumed that the null hypothesis $\theta = 0$, claiming that the observed sample will be distributed according to $\mathrm{Pr}_0 : \mathcal{B} \to [0,1]$, is the one that is currently accepted, whereas the alternate hypothesis $\theta = 1$, claiming that the observations will be distributed according to $\mathrm{Pr}_1 : \mathcal{B} \to [0,1]$, will gain validity if the test turns out to be significant and rejects the null hypothesis. For instance, when a scientist hypothesizes that a phenomenon A is the cause of the phenomenon B, then the null hypothesis is usually taken to be the claim that the phenomenon B is not correlated to A, whereas the alternate hypothesis is the claim A and B are correlated. When a judge needs to decide whether the accused A has committed a crime B, then the null hypothesis is that A is innocent with respect to B, whereas the alternate hypothesis is that A is guilty of B.

To continue with the example from Sec. 3.1, now consider the two hypothetic distributions of the behaviors in the system $\mathcal{S}$ displayed in Table 3.2. In the last line of the table is the *likelihood ratio* $\frac{\mathrm{Pr}_1(x)}{\mathrm{Pr}_0(x)}$. Neyman and Pearson [27] use the likelihood ratio to decide when to reject the null hypothesis $\theta = 0$ in favor of the alternative hypothesis $\theta = 1$. For this

$\mathcal{B}$	a	b	c	d
Pr_0	.98	.005	.005	.01
Pr_1	.098	.001	.001	.9
$\frac{\mathrm{Pr}_1(x)}{\mathrm{Pr}_0(x)}$	.1	.2	.2	90

Table 2: Trustworthy vs untrustworthy behavior

purpose, they introduce the decision thresholds α and β, displayed in Table 3.2, which define the error probabilities as follows

- α is the probability that the null hypothesis is rejected when it is true, whereas

- β is the probability that the null hypothesis is not rejected when it is false.

		reality	
		$\theta = 0$	$\theta = 1$
decision	$\theta = 0$	true $1-\alpha$ confidence	false negative $\beta = \mathrm{Pr}(0\vert 1)$
	$\theta = 1$	false positive $\alpha = \mathrm{Pr}(1\vert 0)$	true $1 - \beta$ strength

Table 3: Decision thresholds α and β

Since the rejection of the null hypothesis is conventionally viewed as the positive outcome a statistical test, the first type of error is called a *false positive* decision, whereas the second type of error is called a *false negative*. E.g. in the court of law, sentencing an innocent person is a false positive, and letting a guilty person free is a false negative, since the null hypothesis is that the accused is innocent, and the burden of proof towards rejecting this hypothesis is on the prosecution. In a fire alarm system, the null hypothesis is that there is no fire, and the false positive is when the alarm rings without fire, whereas a false negative is when the alarm does not ring when there is fire. It is generally accepted as worse to have false positives, since they lead to switching off the fire alarms, rejecting the entire testing frameworks, and thus impelling the negatives as the only outcomes. Neyman and Pearson therefore design the testing frameworks where the upper bound α of the false positive decisions is freely chosen by the tester, and then the upper bound for the false negative decisions is minimized. The *power* of a test is defined to be the probability $1 - \beta$ that the null hypothesis is rejected when it is really false. The Neyman-Pearson Lemma [27] characterizes the decision rules of maximally powerful tests.

In the case of the trust test $f : \{a, b, c, d\} \to \mathbb{R}$ from Sec. 3.1, which boils down to an observation of a single system event, the Neyman-Pearson powerful testing rejects the null hypothesis $\theta = 0$ in favor of the alternative hypothesis $\theta = 1$ at the level $\alpha = 1\%$ only if the event d is observed, and otherwise fails to obtain a significant result. This means that we should only reject the trust hypothesis $\theta = 0$ and endorse the hypothesis $\theta = 1$ that the system $\mathcal{S}$ is not trustworthy if the observed delays d amount to more than 1% of the sampled performance time. Crashing or blocking .5% of the time should not trigger our distrust.

Note that the threshold $\alpha = 1\%$, imposed in the powerful testing as the upper bound of the false positives, has eliminated the significance of the observations b and c, which were significant enough to cause the rejection of the null hypothesis at the same threshold level $\alpha = 1\%$ in Sec. 3.1. On the other hand, the minimization of the false negatives in the powerful testing has now introduced the observation d as significant, which it was not the significance testing. The two testing approaches thus implement two incomparable views of trust. It seems worth while to further explore which one might be more suitable for which application domains.

Although the powerful testing allows comparing pairs of hypotheses (albeit in essentially asymmetric roles of the null hypothesis and its alternative!), it actually provides little help in selecting between *multiple* alternative hypothesis. The best we can do with powerful testing in such situations is to test the null hypothesis against each of the candidate alternatives. However, such approaches lead to pathological situations, where the hypothesis 0 is rejected against 1, 1 against 2, and 2 against 1. Similar phenomena arise when the same significance test is applied to several hypotheses, in the hope that some will be rejected and some not. Overcoming such difficulties requires randomized sampling, Bayesian reasoning, and controlled experiments.

3.3 Experimental testing of trust

If I know an overall probability $\Pr(0)$ that a system similar to $\mathcal{S}$ might be trustworthy, and $\Pr(1) = 1 - \Pr(0)$ that it might not be trustworthy, then I could derive the probability $\Pr(0|x)$ that the system $\mathcal{S}$ is trustworthy after the observed behavior $x \in \mathcal{B}$ using the Bayes' law:

$$\Pr(0|x) \quad = \quad \frac{\Pr_0(x)\Pr(0)}{\Pr_0(x)\Pr(0) + \Pr_1(x)\Pr(1)} \qquad (3)$$

If there are several hypotheses $\theta \in \Theta = \{0, 1, 2, \ldots, n\}$ about the behavioral profiles of the systems, then I can calculate the probability of each of them after the observation $x \in \mathcal{B}$ by the general formula

$$\Pr(\theta|x) \quad = \quad \frac{\Pr_\theta(x)\Pr(\theta)}{\sum_{\psi \in \Theta} \Pr_\psi(x)\Pr(\psi)} \qquad (4)$$

However, the only way to control the distribution $\Pr : \Theta \to [0, 1]$ of the trust profiles of a population of systems to which $\mathcal{S}$ belongs is to model this population in the experimental environment of a laboratory, where I could control that the sample is distributed according to $\Pr : \Theta \to [0, 1]$. Sampling the behaviors of the system $\mathcal{S}$ in this controlled environment would then allow me to calculate $\Pr(\theta|x)$ according to (4) for all profiles $\theta \in \Theta$, and to select the most likely profile $\theta = 0 \in \Theta$ as my current trust hypothesis about $\mathcal{S}$.

But even this experimental environment, where I can impose the prior probability $\Pr : \Theta \to [0, 1]$ by controlling the sample, does not give me the prior probabilities $\Pr_\theta : \mathcal{B} \to [0, 1]$, which express the trust hypotheses to be tested. Where do they come from?

4. FORMULATING TRUST HYPOTHESES

How exactly should I find the trust hypotheses suitable for testing? How should I select the most important ones?

4.1 Scientific presumption of innocence

Both the scientific methodology and the sound legal practices suggest that the null hypothesis should be that the system is trustworthy, i.e. "innocent until proven guilty" [35]. The alternate hypotheses should describe the various forms of undesired behavior, which the tested sample might uncover if the null hypothesis is rejected.

If I know the statistical profile of the desired normal behavior of a system, then I should take that profile as the null hypothesis $\Pr_0 : \mathcal{B} \to [0, 1]$. But it is usually difficult to specify the desired normal behavior as a single profile. It is much easier to characterize each of the abnormal behaviors, which we learn from the anomalies experienced in the past. That is why the statistical intrusion detection systems [10, 24, 7] and forensics mostly work with the statistical profiles of intruders and criminals, and test these profiles as the null hypotheses.

The problem with this "guilty until proven innocent" approach is not just that it is unfair in court. A greater problem arises from the logical limitation of inductive inference: that *the null hypothesis can never be proved by a finite number of tests, but can only be disproved.* By testing the profiles of guilt on the given samples of behaviors, we can never demonstrate anyone's guilt; we can only fail to disprove it. The consequence in the realm of security is that the trust based on testing and rejecting every known form of undesirable behavior is not only impractical, but also the weakest possible form of trust. All that you know is that no guilt has been proven yet. The complexity and the ineffectiveness of this method is illustrated time and again by the complexity and the ineffectiveness of the vetting procedures, which often admit untrustworthy subjects, while regularly rejecting trustworthy subjects. Scientifically based trust, based on testing the null hypothesis that the subject is trustworthy, would obviously be simpler and more effective, both because it allows sound statistical controls of the false positives and the false negatives, and also because it eliminates not only the known anomalies, but all anomalies that are inconsistent with the normal behavior profile described by the null hypothesis.

But where can I find the statistical profile $\Pr_0 : \mathcal{B} \to [0, 1]$ characterizing the trustworthy behavior of the system $\mathcal{S}$? I could log the normal functioning of the system for a long time; but which observable system events $\mathcal{B}$ yield the relevant observations?

The *first limitation* of scientific induction, discussed so far, is that it never proves, but only disproves its hypotheses. Here we confront its *second limitation*: the null hypotheses cannot be extracted from the empiric data, but always have to be formulated *a priori*.

4.2 Compressing trust

The problem of formulating *a priori* hypotheses was discussed in philosophy of science several centuries ago, but remained unsolved. The path towards the modern solutions was opened by Ray Solomonoff [39], and cleared by Andrei Kolmogorov [19] and his school. The versions suitable for practical applications in machine learning and in statistics were developed by Jorma Rissanen [37], Chris Wallace [40], and many others. Very roughly, the idea is as follows.

Continuing with the notation from Sec. 3.3, we still denote the set of hypotheses by Θ. The problem is that we do

not know the probabilities $\Pr_\theta : \mathcal{B} \to [0,1]$. We are, however, given a sufficiently large data sample, from which we extract the frequency distribution $\Pr : \mathcal{B} \to [0,1]$ of each observation.

The task is now to find a hypothesis $\theta = \theta_0 \in \Theta$ such that $\Pr_0 : \mathcal{B} \to [0,1]$ maximizes the conditional probability $\Pr(\theta_0|x)$ in (4) when the behavior $x \in \mathcal{B}$ is observed. Since

$$\Pr(x) \quad = \quad \sum_{\psi \in \Theta} \Pr_\psi(x) \Pr(\psi) \qquad (5)$$

the Bayes' formula (4) now becomes

$$\Pr(\theta|x) \quad = \quad \frac{\Pr_\theta(x) \Pr(\theta)}{\Pr(x)} \qquad (6)$$

The null hypothesis θ_0 gives the probability distribution $\Pr_0 : \mathcal{B} \to [0,1]$ such that for the observed x holds $\Pr(\theta_0|x) \geq \Pr(\theta|x)$ for all $\theta \in \Theta$. Since the probability $\Pr(x)$ is given by the observed data, the task only depends on the unknown hypotheses $\theta \in \Theta$.

The idea used by Solomonoff, Kolmogorov and others is to apply Occam's razor here, and to postulate that *the simplest hypotheses have the highest* a priori *probability*. The idea is implemented by taking into account the *lengths of the descriptions* of the probabilities in (6). Using the optimal Shannon-Fano encodings [8], we can write a number $p \in [0,1]$ using $-\log p$ bits. The task of maximizing (6) now becomes the task of minimizing

$$-\log \Pr(\theta|x) \quad = \quad -\log \Pr_\theta(x) - \log \Pr(\theta) + \log \Pr(x)$$

Since $\Pr(x)$ is fixed, this means that

$$\theta_0 \quad = \quad \operatorname*{argmin}_{\theta \in \Theta} \{-\log \Pr_\theta(x) - \log \Pr(\theta)\} \qquad (7)$$

This is equivalent to $\Pr_0(x) \cdot \Pr(\theta_0) \geq \Pr_\theta(x) \cdot \Pr(\theta)$ for all $\theta \in \Theta$, which picks θ_0 to maximize the chance that x is observed. This is what makes θ_0 the best *a priori* null hypothesis. The minimality of the description length $-\log \Pr(\theta_0)$ means that θ_0 is the simplest. The minimality of $-\log \Pr_0(x)$, or equivalently the maximality $\Pr_0(x)$, means that x is the most likely prediction of θ_0. The minimality of $-\log \Pr_0(x) - \log \Pr(\theta_0)$ means that θ_0 is the simplest hypothesis among those that predict x.

Instantiated to the realm of trust, (7) thus says that the best trust hypothesis is the one that provides the shortest description of my notion of trust, which fits the observations that I have made.

The rapidly expanding research area of algorithmic learning and statistical inference is concerned not only with the effective computations of the *a priori* hypotheses, but also with the situations where the succinct descriptions of the data and the hypotheses need to be combined with empiric data. The right-hand side of (7) is roughly Rissanen's Minimum Description Length (MDL) [37] of the distribution of the observed data x. Wallace's Minimum Message Length (MML) [37] differs in the compression methods used. Kolmogorov's minimal sufficient statistic [19] uses the optimal computable encodings as the compression method. The standard compression algorithms, e.g. based on the very efficient Lempel-Ziv algorithms [41, 42] are also often used, and give reasonable results. In any case, the best null hypothesis is the one which best compresses the observed data x, within some given family of compression algorithms. The underlying idea is that the better we understand the data, the better we compress them.

Although these methods give somewhat degenerate results when applied to our toy examples from Sec. 3, just slightly larger trust hypotheses show the intuitive meaning of (7) in the realm of trust. My trust hypothesis should be the simplest description of the desired behaviors which best approximates the observed behaviors of the tested system.

5. BACKGROUND AND FUTURE WORK

The main claim of this paper is that the methods of statistical inference, on which modern science has been built, can be used to analyze and secure trust. We close the paper relating this idea with the general context of trust research, and in particular with the existing application of statistical methods to trust testing in the framework of intrusion detection.

The literature about trust is very extensive, as it is studied in psychology, social sciences, economics, game theory [4, 6, 23]. Even within the closely related security research communities, the word 'trust' is used in several different meanings [18]. The notion of trust used in this paper is based on [33].

A quantitative analysis of the process of trust building was initiated in [31]. The question of trust decisions was, however, avoided by reducing them to the preferences extracted from the trust ratings. The question of trust measurements was avoided by reducing them to user ratings and feedback, which are usually available in web commerce, but not in general. In system security, the task of quantifying security in general and trust in particular becomes a problem [2, 25]. In the present paper, we did not consider the problem of quantifying trust *a posteriori*, i.e. using the measurements of the past performance, but focused on the harder problem of formulating the trust hypotheses *a priori*, i.e. before any empiric data are available. This problem arises even if the satisfactory methods for quantifying trust and security *a posteriori* are available, because the data are not always available. On the other hand, understanding how to express the *a priori* trust beliefs may also help in devising and validating the methods to quantify them *a posteriori*.

The idea of *statistical intrusion detection*, going back to Dorothy Denning [10] and her work with Peter Neumann at SRI in the 80s, can be viewed as an application of statistics to detect the subjects or the components that are not trustworthy. An early survey is [24]. The practices of intrusion detection have evolved a lot since those early days, and the rule based methods seem to have found broader applications than the statistical methods. One of the reasons often mentioned is the difficulty to control the false positives that arise when statistical tests are used to detect the intruders. We explained in Sec. 4.1 why the statistical methodologies suggest that trust testing should be based on taking a trustworthy behavior as the null hypothesis, and why testing for anomalies and the untrustworthy behaviors leads to the false positives that are harder to control, and to less reliable results overall. In statistics, proving that someone is not trustworthy is not equivalent to disproving that they are trustworthy. The general method for controlling the false positives when disproving trust is outlined in Sec. 3.2. The false positives thus emerge as a hard problem in statistical intrusion detection because it tests for the in-

trusions, and not for trust. The reason is, of course, that the intruder profiles are much easier to come by than the trustworthy profiles. In the Sec. 4, we discussed the way to solve this problem using the methods of algorithmic learning. Whether that brief discussion explained or obscured the idea, there is very little doubt that at least a theoretical solution lies in this direction. But the practical work towards implementing such computation-based scientific methodologies on the concrete problems of trust lies ahead.

Acknowledgements.

6. REFERENCES

[1] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Soceity. of London*, 53:370–418, 1763.

[2] Steven M. Bellovin. On the brittleness of software and the infeasibility of security metrics. *IEEE Security & Privacy*, 4(4):96–96, 2006.

[3] Terry Benzel. The science of cyber security experimentation: The DETER Project. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 137–148, New York, NY, USA, 2011. ACM.

[4] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, July 1995.

[5] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009.

[6] V. Buskens. *Social Networks and Trust*. Theory and Decision Library C. Springer US, 2002.

[7] Phuong Cao, Key-whan Chung, Zbigniew Kalbarczyk, Ravishankar Iyer, and Adam J. Slagell. Preemptive intrusion detection. In *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*, HotSoS '14, pages 21:1–21:2, New York, NY, USA, 2014. ACM.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.

[9] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1990.

[10] Dorothy E. Denning. An intrusion-detection model. *IEEE Trans. Softw. Eng.*, 13(2):222–232, February 1987.

[11] P. Feyerabend. *Against Method*. Verso, 1993.

[12] Richard P. Feynman. *The Character of Physical Law*. Penguin Books, 1992.

[13] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

[14] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, UK, second edition, 1959.

[15] Jingwei Huang and David Nicol. A formal-semantics-based calculus of trust. *IEEE Internet Computing*, 14(5):38–46, September 2010.

[16] Jingwei Huang and David M. Nicol. Evidence-based trust reasoning. In *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*, HotSoS '14, pages 17:1–17:2, New York, NY, USA, 2014. ACM.

[17] JASON Defense Advisory Panel. *Science of Cyber-security*. The MITRE Corporation, 2010.

[18] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43:618–644, March 2007.

[19] Andrei Kolmogorov. On the logical foundations of probability theory. In A.N. Shiryayev, editor, *Selected Works of A. N. Kolmogorov*, volume 26 of *Mathematics and Its Applications (Soviet Series)*, pages 515–519. Springer Netherlands, 1992.

[20] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 2012.

[21] Imre Lakatos. *The Problem of Inductive Logic*. Proceedings of the International Colloquium in Philosophy of Science (London, 1965). North Holland Publishing Company, 1968.

[22] Carl E. Landwehr. Cybersecurity: From engineering to science. *The Next Wave*, 19(2):2–5, 2012.

[23] N. Luhmann. *Trust ; And, Power: Two Works*. Number pts. 1-2 in UMI Books on Demand. Wiley, 1979.

[24] Teresa F Lunt. A survey of intrusion detection techniques. *Computers & Security*, 12(4):405–418, 1993.

[25] Andrew Meneely, Ben Smith, and Laurie Williams. Validating software metrics: A Spectrum of Philosophies. *ACM Trans. Softw. Eng. Methodol.*, 21(4):24:1–24:28, February 2013.

[26] Robert Meushaw. NSA initiatives in cybersecurity science. *The Next Wave*, 19(4):8–13, 2012.

[27] Jerzy Neyman and Egon S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I, Part II. *Biometrika*, 20A(1/2, 3/4):175–240, 263–294, 1928.

[28] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933.

[29] D.M. Nicol, W.H. Sanders, W.L. Scherlis, and L.A. Williams. Science of security hard problems: A Lablet Perspective. cps-vo.org/file/6394/download/47034, retrieved on 2015/1/10.

[30] D.M. Nicol and M.P. Singh, editors. *HotSoS '14: Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*, New York, NY, USA, 2014. ACM.

[31] Dusko Pavlovic. Dynamics, robustness and fragility of trust. In Pierpaolo Degano, Joshua Guttman, and Fabio Martinelli, editors, *Proceedings of FAST 2008*, volume 5491 of *Lecture Notes in Computer Science*, pages 97–113. Springer Verlag, 2008. arxiv.org:0808.0732.

[32] Dusko Pavlovic. The unreasonable ineffectiveness of security engineering: An overview. In José Luiz Fiadeiro and Stefania Gnesi, editors, *Proceedings of IEEE Conference on Software Engineering and Formal Methods, Pisa, Italy, 2010*, pages 12–18. IEEE, 2010.

[33] Dusko Pavlovic. Quantifying and qualifying trust:

Spectral decomposition of trust networks. In Pierpaolo Degano, Sandro Etalle, and Joshua Guttman, editors, *Proceedings of FAST 2010*, volume 6561 of *Lecture Notes in Computer Science*, pages 1–17. Springer Verlag, 2011. arxiv.org:1011.5696.

[34] Dusko Pavlovic. On bugs and elephants: Mining for science of security. *The Next Wave*, 19(2):23–29, 2012.

[35] Kenneth Pennington. Innocent until proven guilty: The origins of a legal maxim. *Jurist*, 63:106–124, 2003.

[36] Karl Popper. *The Logic of Scientific Discovery*. Routledge Classics. Taylor & Francis, 2002.

[37] Jorma Rissanen. *Information and Complexity in Statistical Modeling*. Information science and statistics. Springer, New York, 2007.

[38] Fred B. Schneider. Blueprint for a science of cybersecurity. *The Next Wave*, 19(2):47–57, 2012.

[39] Ray J. Solomonoff. A formal theory of inductive inference. Part I., Part II. *Information and Control*, 7:1–22, 224–254, 1964.

[40] C.S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer, 2005.

[41] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[42] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.