Invited paper

Probability References to Apply in the Detection of Anomalous Public Key Infrastructures

Antonio Castro Lechtaler[1,2], Marcelo Cipriano[1], Eduardo Malvacio[1]

[1]*CriptoLab. Escuela Superior Técnica - Universidad Nacional de la Defensa, Cabildo 15. C1426AAA*
Ciudad Autónoma de Buenos Aires, Argentina
{antonio.castrolechtaler; cipriano1.618; edumalvacio}@gmail.com
[2]*CISTIC, Facultad de Ciencias Económicas, Universidad de Buenos Aires, Córdoba 2122, C1120AAQ*
Ciudad Autónoma de Buenos Aires, Argentina

Abstract

This article calculates the theoretical probability of finding repeated primes in a given sample of unbiased issued digital certificates. These values can be used as reference for developing a statistical procedure to audit and control the behavioral pattern of a Public Key Infrastructure (PKI), thus allowing the detection of operational anomalies and the prevention of vulnerabilities of this nature.

Keywords: PKI, RSA, Digital Certificates.

1. Introduction

The certificates issued by a public key infrastructure (PKI) are widely used in military environments and systems, as well as in civil, public or private networks, LANs or WANs, and even in the Internet. Among other applications, they can be used for the login and authentication of users, equipments and systems, encryption and digital signatures, non-repudiation, determination of session keys, etc.

The certificates issued by a PKI include, among others, a module *m* and a number *e* (usually 65537), known as "public key", and a number *d*, known as "private key". The *m* value, which has a *t* size (measured in bits), is obtained by multiplying 2 prime values. This trio *(m, e, d)* is calculated by the PKI when the pertinent digital certificate is requested and is delivered to a user who will own it.

A vulnerability occurs[1] if the PKI reveals an anomaly when calculating the *m* values or upon the issuance of certificates in which two or more users share any prime factor of their respective modules.

This information allows to circumvent the security provided by RSA and to easily obtain the private key, thus enabling access to the information intended to be protected.

Current systems are highly complex and it is not easy to detect certain types of errors [7]. The traditional method consists in the reading and control of the code lines that make up the PKI. The detection of errors, on the other hand, is a reality with many precedents that can be reviewed [1,9].

An interesting discussion may take place regarding the nature of such errors: innocent "bugs" that beat the tests and leaked through only to be detected years after their creation or that were "planted" with the intention of debilitating the security system.

Other researchers [8] evaluated more than a million public key certificates and discovered that about *5%* of them shared prime factors. Is this a value to be expected, considering the magnitude of the analyzed sample, or is it beyond possibilities, considering the size of the analyzed modules and the number of possible primes?

This work and its precedents [2-6] determine the Probability Function of finding or not finding collisions of prime factors in a variable sample space composed by certificates. Such probabilities (unbiased and free of anomalies because they were calculated theoretically) could be used as reference values to audit and control the behaviour of a specific PKI.

Sections 2 and 3 present a probabilistic model to perform experiment *E*, its possible results, the methodology to determine the aforementioned results and, finally, the resulting formulas that will allow to estimate the probabilities of finding repeated primes size *t* in samples size *mu*, whether repeated primes size t are present or not.

Section 4 presents the Probability Function that prevails in a theoretical PKI that is unbiased and free of anomalies.

Section 5 presents different formulas to calculate large factorials, as required by the formulas obtained in the previous sections.

Section 6 presents conclusions and possible follow-ups to this research: determination of the probabilistic model hidden in a PKI to be analyzed, using tools derived from statistical inference. And, lastly, the comparison between theoretical values and those obtained empirically that will allow to

[1] The security of the RSA system is based on the difficulty of factorizing the modules *m* (i.e., *t*=1024, 2048 or 4096 bits as currently used) within an acceptable time frame in order to preserve the secret *d* key. Knowing one of the prime factors of a given module makes it possible to calculate the other factor and the *d* key with no difficulty.

determine the presence of biases or anomalies. The final goal would be the development of a software to make the auditing of the PKI possible.

2. Number of certificates issuable by a PKI

Given hypothesis[2] *H1*: the size of prime values generated by the PKI, shown here as b, is half the value t of the modules. For instance, if t=1024 the prime values shall then be size b=512 bits.
Being P_1 the set of prime numbers size b.

$$P_1=\{p \,/\, p\ primo;\ 2^{b-1}< p <2^b\}. \qquad \text{(Eq.1)}$$

The cardinal or number of elements of P_1 – shown here as p_1 – can be calculated with an equation associated to the Theorem of Prime Numbers[3]:

$$p_1 = Card(P_1) \approx \pi(2^b) - \pi(2^{b-1}). \qquad \text{(Eq.2)}$$

$$p_1 \approx \frac{2^b}{ln\,2^b} - \frac{2^{b-1}}{ln\,2^{b-1}} = \frac{2^{b-1}}{ln\,2}\left(\frac{2}{b} - \frac{1}{b-1}\right). \qquad \text{(Eq.3)}$$

Being M_1 the set of all the public modules that can be determined based on the number of P_1 elements:

$$M_1 = \{m/m = pq;\ p \neq q;\ p,q \in P_1\}. \qquad \text{(Eq.4)}$$

We assume here a different work hypothesis *H2:* the PKI shall not issue certificates resulting from the same prime factor. This means that the public module shall not be a square number.

The cardinal of M_1 (shown as m_1) is the number of subsets of P_1 consisting of 2 elements, since each public module is the product of 2 prime values and because the order of multiplication is irrelevant due to the commutativity of the product.

$$m_1 = Card(M_1) = \binom{p_1}{2} = \frac{p_1(p_1-1)}{2}. \qquad \text{(Eq.5)}$$

3. Experiment E, Probabilistic Model and its Probability Function

3.1 Definition of Experiment E

Experiment *E* is laid out: request the PKI a number mu of certificates and assemble with them the set *MU*, called "sample".

$$MU=\{m \,/\, m\ is\ a\ public\ module\ size\ t\} \qquad \text{(Eq.6)}$$

$$Card(MU)=mu. \qquad \text{(Eq.7)}$$

We assume here hypothesis *H3*: the resulting modules m are unbiased. This determines a probabilistic model in which the probability of obtaining any module is equiprobable.

3.2 Experiment Outcomes

This experiment can have 2 *outcomes* or *events:*

$$R = \{r1\ ;\ r2\}. \qquad \text{(Eq.8)}$$

- r_1: where in the set *MU* there are no modules *m* sharing any prime factor. In this case it will be said that *there are no collisions of prime factors.*
- r_2: where in the set *MU* there are 2 or more modules *m* repeating prime factors. In this case it will be said that *there are collisions of prime factors.*

To determine the experiment outcome[4] we will use the Highest Common Denominator of all modules, taken 2 at a time:

$$\forall m_i,m_j \in MU(i \neq j); mcd(m_i,m_j) \in \{1;p\}, p \in P_1. \qquad \text{(Eq.9)}$$

Therefore, if all the values obtained from *mcd* are 1, the experiment had an outcome r_1 because it was not verified that there were repeated primes in those modules. Otherwise, the experiment had an outcome r_2.

The process of applying experiment *E* to all possible *MU* sets, determines 2 sets: R_1 and R_2:

$$R_1 = \{\ MU \,/\, R(MU) = r_1\}. \qquad \text{(Eq.10)}$$

$$R_2 = \{\ MU \,/\, R(MU) = r_2\}. \qquad \text{(Eq.11)}$$

These sets have the following properties:
Disjoints:

[2] Different hypotheses regarding context and environment will be presented throughout this work. They will be described and numbered in order of appearance.

[3] Conjectured by German mathematician *Carl Gauss* (1777-1855) and confirmed independently by Belgian mathematician *Charles-Jean de la Vallée Poussin* (1866-1962) and French mathematician *Jacques Hadamard* (1865-1963).

[4] It could be requested from the PKI to provide the prime factors of the module together with the rest of the certificate information. Then, it will suffice to just check whether there are repeated primes or not in the sample. Both tests, the mcd estimation and the latter, have a computational cost that will not be analyzed here. It is recommended to choose the least complex of the two.

$$R_1 \cap R_2 = \varnothing. \qquad \text{(Eq.12)}$$

a) Complementary:

$$R_1 \cup R_2 = EM(E). \qquad \text{(Eq.13)}$$

Where *EM(E)* is the *Sample Space of Experiment E*, meaning all *MU* sets size *mu*.

$$EM(E)=m_1^{mu} \qquad \text{(Eq.14)}$$

Assuming hypothesis *H4*: the PKI does not "remember" the issued certificates. Therefore, it could repeat modules in its sample space[5].

We define the Probability Function in accordance with the properties of R_1 and R_2

$$p(R_1)+p(R_2)=1\,. \qquad \text{(Eq.15)}$$

$$p(R_2)=1-p(R_1)\,. \qquad \text{(Eq.16)}$$

We will use the classic theory[6] to determine the value of these probabilities. To this purpose, the cardinal of each set and the total of the sample space must be calculated.

3.3 Cardinality of R1 and R2

The cardinal of R_1 is the number of sets *MU*, size *mu*, consisting of modules size *t*, in which it is verified that there are *no* collisions of primes.

Given the sets P_1 y M_1 shown in Eq(1) and Eq(3), respectively, the first element of *MU* could be any of the m_1 elements of M_1.

The second element should be a co-prime module of the first element in the sample. To this purpose, we determine set P_2 as the set of numbers that results from removing numbers *p* and *q* that make up the first element of set P_1.

$$P_2 = P_1 - \{p;q\}. \qquad \text{(Eq.17)}$$

$$p_2 = Card(P_2) = Card(P_1) - 2. \qquad \text{(Eq.18)}$$

Being M_2 is the set of all the modules that can be generated with P_2, whose cardinal m_2 is calculated as the combinatorial of all the elements of P_2, taken 2 at a time.

$$M_2 = \{m/m = pq;\, p \neq q;\, p,q \in P_2\} \qquad \text{(Eq.19)}$$

$$m_2 = Card(M_2) = \binom{p_2}{2} = \frac{p_2(p_2-1)}{2}. \qquad \text{(Eq.20)}$$

$$m_2 = \binom{p_1-2}{2} = \frac{(p_1-2)(p_1-3)}{2} \qquad \text{(Eq.21)}$$

Following the same reasoning, the third element of the sample will be a co-prime module with the first and the second element, taken 2 at a time. To this purpose, we determine the set of prime numbers P_3 that results from removing from P_2 factors *p* y *q* that make up its second element.

$$P_3 = P_2 - \{p;q\}. \qquad \text{(Eq.22)}$$

$$p_3 = Card(P_3) = Card(P_2) - 2 = Card(P_1.) - 4. \qquad \text{(Eq.23)}$$

Being M_3 the set all the modules that can be generated with P_3.

$$M_3 = \{m/m = pq;\, p \neq q;\, p,q \in P_3\} \qquad \text{(Eq.24)}$$

$$m_3 = Card(M_3) = \binom{p_3}{2} = \frac{p_3(p_3-1)}{2} \qquad \text{(Eq.25)}$$

$$m_3 = \binom{p_1-4}{2} = \frac{(p_1-4)(p_1-5)}{2} \qquad \text{(Eq.26)}$$

In general, for any value *i* between *1* and *mu*, we have:

$$Pi = P_{i-1} - \{p;q\}. \qquad \text{(Eq.27)}$$

$$p_i = Card(P_i) = Card(P_{i-1}) - 2 = Card(P_1.) - 2(i-1). \qquad \text{(Eq.28)}$$

Being M_i the set of all the modules resulting as a product of the elements of P_i:

$$M_i = \{m/m = pq;\, p \neq q;\, p,q \in P_i\} \qquad \text{(Eq.29)}$$

$$m_i = Card(M_i) = \binom{p_i}{2} = \frac{p_i(p_i-1)}{2}\cdot \qquad \text{(Eq.30)}$$

$$m_i = \binom{p_i - 2(i-1)}{2} = \frac{\lfloor p_1 - 2(i-1) \rfloor \lfloor p_1 - 2(i-1) - 1 \rfloor}{2} \qquad \text{(Eq.31)}$$

It can be observed that each cardinal of the M_i sets is expressed in function of the cardinal of M_1.

This procedure is followed until the last module of the sample is reached: The module number *mu* is the quantity expected in *experiment E*.

[5] The assumption of another hypothesis regarding the PKI would imply a change in the size of the sample space.

[6] The Theory of Probabilities was initiated by *Pierre de Fermat (1601-1665)* and *Blaise Pascal (1623-1662)*. However, the first axiomatic definition of probability is attributed to *Pierre-Simon Laplace (1749-1827)*: the probability of an event being the ratio between the number of favorable outcomes and the total number of possible outcomes.

$$P_{mu} = P_{mu-1} - \{p;q\}. \quad \text{(Eq.32)}$$

$$p_{mu} = Card(P_{mu}) = Card(P_{mu-1}) - 2 = Card(P_1) - 2(mu-1). \quad \text{(Eq.33)}$$

Given M_{mu} the set of all the modules that can be generated with P_{mu}

$$M_{mu} = \{m/m = pq; p \neq q; p,q \in P_{mu}\}. \quad \text{(Eq.34)}$$

$$m_{mu} = Card(M_{mu}) = \binom{p_{mu}}{2} = \frac{p_{mu}(p_{mu}-1)}{2} \quad \text{(Eq.35)}$$

$$m_{mu} = \binom{p_1 - 2(mu-1)}{2} = \frac{[p_1 - 2(mu-1)][p_1 - 2(mu-1)-1]}{2} \quad \text{(Eq.36)}$$

All the modules of the *MU* set are co-primes taken 2 at a time.

Finally, the cardinal of the set of all the samples of modules *mu* in which there are no collisions of primes is determined by:

$$Card(R_1) = \prod_{i=1}^{mu} m_i = \prod_{i=0}^{mu-1} \binom{p_1 - 2i}{2}. \quad \text{(Eq.37)}$$

$$Card(R_1) = \frac{\prod_{i=0}^{2(mu-1)} (p_1 - i).}{2^{mu}}. \quad \text{(Eq.38)}$$

Then,

$$Card(R_1) = \frac{p_1!}{2^{mu}(p_1 - 2(mu-2))!}. \quad \text{(Eq.39)}$$

3.4 Cardinality of R2

Since R_1 y R_2 are disjoints and complementary, as shown in Eq(12) to Eq(14), then:

$$Card(R_2) = EM(E) - Card(R_1). \quad \text{(Eq.40)}$$

$$Card(R_2) = m_1^{mu} - \frac{p_1!}{2^{mu}(p_1 - 2(mu-2))!} \quad \text{(Eq.41)}$$

Being m_1 the number of modules that can be calculated with the prime factors of the set P_1, whose cardinal is the value p_1 and *mu* being the number of modules of each sample.

3.5 Probability Function

As indicated in (16-17), then:

$$p(R_1) = \frac{card(R_1)}{m_1^{mu}}. \quad \text{(Eq.42)}$$

$$p(R_1) = \frac{\dfrac{p_1!}{2^{mu}[p_1 - 2(mu-2)]!}}{m_1^{mu}}. \quad \text{(Eq.43)}$$

Then, as per (Eq.5):

$$p(R_1) = \frac{\dfrac{p_1!}{2^{mu}[p_1 - 2(mu-2)]!}}{\dfrac{[p_1(p1-1)]^{mu}}{2^{mu}}}. \quad \text{(Eq.44)}$$

$$p(R_1) = \frac{p_1!}{[p_1 - 2(mu-2)]![p_1(p1-1)]^{mu}}. \quad \text{(Eq.45)}$$

$$p(R_2) = 1 - \frac{p_1!}{[p_1 - 2(mu-2)]![p_1(p1-1)]^{mu}} \quad \text{(Eq.46)}$$

4. Calculating large factorials

These equations require solving quite large factorials whose computational complexity hinder the calculation. As an example, we will present some equations to estimate the factorial value:

$$n! = e^{\ln n!} \approx e^{n(\ln n - 1)}. \quad \text{(Eq.47)}$$

$$n! \approx n^n e^{-n} \sqrt{2\pi n}. \quad \text{(Eq.48)}$$

$$n! \approx n^n e^{-n} \sqrt{\pi} \sqrt[6]{8n^3 + 4n^2 + n + \frac{1}{30}}. \quad \text{(Eq.49)}$$

Eq(47) and Eq(48) are known as *Stirling*[7]'s and Eq(49) as *Ramanujan*[8]'s equations:

$$n! \approx \sqrt{2\pi} \left(\frac{n + \frac{1}{2}}{e} \right)^{n+\frac{1}{2}}. \quad \text{(Eq.50)}$$

$$n! \approx n^n e^{-n} \sqrt{\pi} \sqrt{2n + \frac{1}{3}}. \quad \text{(Eq.51)}$$

$$n! \approx n^n e^{-n} \sqrt{2\pi\left(n + \frac{1}{6} + \frac{1}{72n} - \frac{31}{6480n^2} - \frac{139}{155520n^3} + \frac{9871}{6531840n^4}\right)} \quad \text{(Eq.52)}$$

Eq(50), Eq(51) and Eq(52) are known as Burnside's[9], Gosper's[10] and Batir's[11] equations, respectively.

5. Conclusions and future work

We have presented equations to calculate the mathematical probabilities of finding collisions of primes in a sample based on a source that is unbiased and free of anomalies.

The existence of statistical permanence can be

[7] *James Stirling* (1692-1770). Scottish mathematician.

[8] *Srinivasa Ramanujan* (1887-1920). Indian mathematician. He did not leave behind a demonstration of his equation. It was demonstrated by the Russian mathematician *Ekatherina Karatsuba* in the year 2000.

[9] *William Burnside* (1852-1927). English mathematician.

[10] *Ralph Gosper, Jr.* (1943-). American mathematician and computer scientist.

[11] *Necdet Batir* (1959 -). Turkish mathematician.

assumed as a hypothesis, which means that throughout the performance of experiment E applied to a specific PKI, its unknown probabilistic model can be discovered with the use of statistical tools.

Should discrepancies between the reference values proposed by this work and those obtained by “direct experience” be verified, it would indicate the existence of an anomaly in the behavior of the PKI.

In sum, this work sets out issues to continue researching and enable the development of a software to audit and control PKI anomalies.

References

[1] Bello L, Bertacchini M. *“Generador de Números Pseudo-Aleatorios Predecible en Debian”*. III International Cyber Security Conference. Manizales, Colombia. Octubre 2009.

[2] Benaben, A; Castro Lechtaler, A; Cipriano, M; Foti, A. *“Development, testing and performance evaluation of factoring algorithms whit additional information*” XXVIII International Conference of the Chilean Society of Computer Sciences. Santiago de Chile. 2009.

[3] Castro Lechtaler, C; Cipriano, M; Benaben A; Quiroga, P. *“Study on the effectiveness and efficiency of an algorithm to factorize N given e and d*”. IX Latin American Seminar on Information Technology Security, La Habana, CUBA. 2009.

[4] Castro Lechtaler, A; Cipriano, M. “*Detection of anomalies in Oracles such as OpenSS through the analysis of probabilities*”. XVII Argentine Convention of Computer Sciences CACIC 2011. La Plata, Buenos Aires, October 2011.

[5] Castro Lechtaler, Antonio, Cipriano Marcelo; Malvacio Eduardo; Cañón, Sebastián; *Procedure for the Detection of Anomalies in Public Key Infrastruture (RSA Systems)*. XIII Argentine Technological Symposium, 41 Argentine Meetings on Information Technology and Operational Research JAIIO – SADIO. La Plata, Buenos Aires, August 2012.

[6] Castro Lechtaler, Antonio; Cipriano, Marcelo; Malvacio, Eduardo. *Experi-mental detection of anomalies in public key infrastructure*. XVIII Argentine Convention on Computer Sciences CACIC 2012. Bahía Blanca, Buenos Aires, October 2011.

[7] Glass, Robert *“Facts and Fallacies of Software Engineering*”. Addison-Wesley Professional, 2003.

[8] Lenstra, A; Hughes, J; Augier, M and others. Ron was wrong, Whit is right. e-print International Association for Cryptologic Research. 15 Feb 2012. http://eprint.iacr.org/2012/064.

[9] Young A and Yung M. *An Elliptic Curve Asymmetric Backdoor in Open-SSL RSA Key Generation.* Chapter 10. Cryptovirology. 2006.