The dynamic structure of data breaches

Domenico De Giovanni, Arturo Leccadito, Marco Pirra
Department of Economics, Statistics and Finance "Giovanni Anania"
Università della Calabria

Marco Pirra
marco.pirra@unical.it

22 May 2019

AFIR ERM 19 Florence

IAA AAI AFIR-ERM Finance, Investment & ERM

ISOA

Agenda

- Introduction
- Methodology
- Results
- Conclusions

Cyber risk

There is **no standardised definition** of the term "cyber risk."

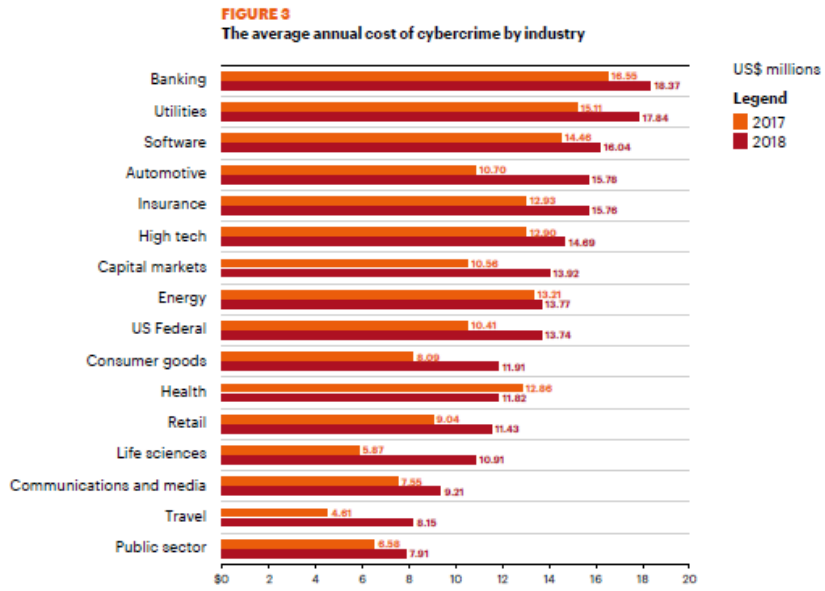
The CRO Forum has broadly described "cyber risk" to mean: "*Any risks that emanate from the use of electronic data and its transmission, including technology tools such as the internet and telecommunications networks*.

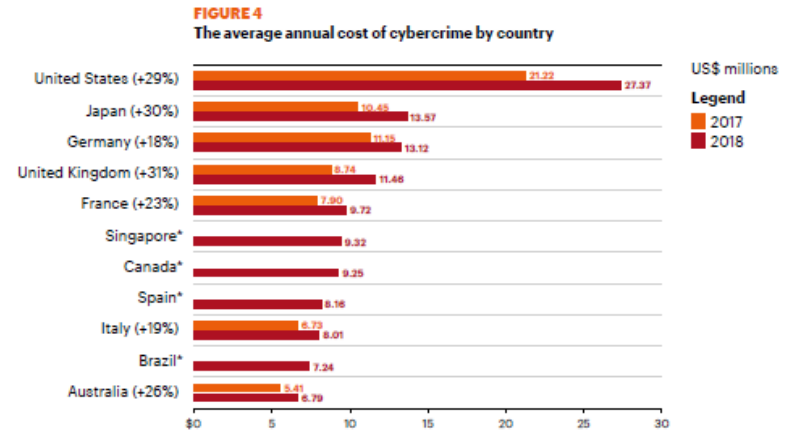
It also encompasses physical damage that can be caused by cybersecurity incidents, fraud committed by misuse of data, any liability arising from data storage, and the availability, integrity and confidentiality of electronic information – be it related to individuals, companies or governments."

The cost of Cybercrime [Ponemon Institute LLC]

In the last year, many stealthy and sophisticated cyberattacks targeted public and private sector organizations.
Combined with the expanding threat landscape, organizations are seeing a steady rise in the number of security breaches—from 130 in 2017 to 145 this year **(+11% last year, +67% last 5 years)**.

The impact of these cyberattacks to organizations, industries and society is substantial.
Alongside the growing number of security breaches, the total cost of cybercrime for each company increased from US$11.7 million in 2017 to a new high of US$13.0million (**+12% last year, +72% last 5 years**).





Annual Cost of a Data Breach Study 2018

2018 Cost of a Data Breach Study: Global Overview

Benchmark research sponsored by IBM Security

Independently conducted by Ponemon Institute LLC

This year's study reports the global average cost of a data breach is **up 6.4 percent over the previous year** to $3.86 million.

The average cost for each lost or stolen record containing sensitive and confidential information also **increased by 4.8 percent** year over year to $148

For the past 13 years, the Ponemon Institute has conducted an annual Cost of a Data Breach Study in order to measure exactly how much lost and stolen records could cost companies around the world.

Economical Impact [Allianz Risk Barometer]

The eighth Allianz Risk Barometer incorporates the views of a record 2,415 respondents from 86 countries.

For the first time, cyber incidents is neck-and-neck with business interruption at the top of the Risk Barometer – with the two risks increasingly interlinked, reflecting the magnitude of the threat now posed by a growing dependence on technology and the malicious actions of nation states and criminals.

Increasing concern over cyber incidents follows a watershed year of activity. Cyber crime costs an estimated **$600bn a year** up from $445bn in 2014. This compares with a 10-year average economic loss from natural catastrophes of around **$200bn** – three times as much.

The number of **cyber-attacks** worldwide **doubled** in 2017 to 160,000, although **endemic underreporting** means the true figure could be as high as 350,000, according to the Online Trust Alliance

IAIS - International Association of Insurance Supervisors

ISSUES PAPER ON CYBER RISK
TO THE INSURANCE SECTOR (2016)

IAIS
INTERNATIONAL ASSOCIATION OF INSURANCE SUPERVISORS

Concern over **cybersecurity** **is growing across all sectors of the global economy**, as cyber risks have grown and cyber criminals have become increasingly sophisticated. For **insurers, cybersecurity incidents can harm the ability to conduct business, compromise the protection of commercial and personal data, and undermine confidence in the sector**. The IAIS has noted that the level of awareness of cyber threats and cybersecurity within the insurance sector, as well as supervisory approaches to combat the risks, appear to vary across jurisdictions.

These factors prompted the IAIS to consider the area of cybersecurity in the insurance sector, **including the involvement of insurance supervisors in assessing and promoting the mitigation of cyber risk**. While many of the most widely publicised cybersecurity incidents involving consumer data have affected retailers, companies in the financial services sector, including insurers, have been victimised as well.

AFIR ERM 19 Florence

IAA AAI AFIR-ERM Finance, Investment & ERM

ISOA

IAIS - International Association of Insurance Supervisors

All insurers, regardless of size, complexity, or lines of business, **collect, store, and share with various third-parties** (e.g., service providers, reinsurers) **substantial amounts of private and confidential policyholder information**, including in some instances sensitive health-related information.

Information obtained from insurers through cyber crime may be used for financial gain through extortion, identity theft, misappropriation of intellectual property, or other criminal activities. Exposure of private data can potentially result in severe and lingering harm for the affected policyholders, as well as reputational damage to insurer sector participants.

IAIS - International Association of Insurance Supervisors

The objectives of the Issues Paper are to **raise awareness** for **insurers and supervisors of the challenges presented by cyber risk**, including current and contemplated supervisory approaches for addressing these risks. As an Issues Paper, it provides background, describes current practices, identifies examples, and explores related regulatory and supervisory issues and challenges.

The Issues Paper focuses on **cyber risk to the insurance sector and the mitigation of such risks**, but does not cover IT security risks more broadly. It also does not specifically address insurers' underwriting of cyber risk (i.e., cyber insurance) or risks arising from cybersecurity incidents involving supervisors.

Data breaches

A **data breach** is an incident where information is stolen or taken from a system without the knowledge or authorization of the system's owner.

A small company or large organization may suffer a data breach. Stolen data may involve sensitive, proprietary, or confidential information such as credit card numbers, customer data, trade secrets or matters of national security.

The effects brought on by a data breach can come in the form of damage to the target company's reputation due to a perceived 'betrayal of trust.' Victims and their customers may also suffer ***financial losses*** should related records be part of the information stolen.

Literature Overview

Modeling and Predicting Cyber Hacking Breaches (2018) Maochao Xu, Kristin M. Schweitzer, Raymond M. Bateman, and Shouhuai Xu IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 13, NO. 11

T. Maillart and D. Sornette, "*Heavy-tailed distribution of cyber-risks*," Eur. Phys. J. B, vol. 75, no. 3, pp. 357–364, 2010.
B. Edwards, S. Hofmeyr, and S. Forrest, "*Hype and heavy tails: A closer look at data breaches*," J. Cybersecur., vol. 2, no. 1, pp. 3–14, 2016.
S. Wheatley, T. Maillart, and D. Sornette, "*The extreme risk of personal data breaches and the erosion of privacy*," Eur. Phys. J. B, vol. 89, no. 1, p. 7, 2016.

R. Böhme and G. Kataria, "*Models and measures for correlation in cyber-insurance*," in Proc. Workshop Econ. Inf. Secur. (WEIS), 2006, pp. 1–26.
H. Herath and T. Herath, "*Copula-based actuarial model for pricing cyber-insurance policies*," Insurance Markets Companies: Anal. Actuarial Comput., vol. 2, no. 1, pp. 7–20, 2011.
A. Mukhopadhyay, S. Chatterjee, D. Saha, A. Mahanti, and S. K. Sadhukhan, "*Cyber-risk decision models: To insure it or not?*" Decision Support Syst., vol. 56, pp. 11–26, Dec. 2013.
M. Xu and L. Hua. (2017). *Cybersecurity Insurance: Modeling and Pricing*. [Online]. Available: https://www.soa.org/research-reports/
M. Xu, L. Hua, and S. Xu, "*A vine copula model for predicting the effectiveness of cyber defense early-warning*," Technometrics, vol. 59, no. 4, pp. 508–520, 2017.
C. Peng, M. Xu, S. Xu, and T. Hu, "*Modeling multivariate cybersecurity risks*," J. Appl. Stat., pp. 1–23, 2018.

M. Eling and N. Loperfido, "*Data breaches: Goodness of fit, pricing, and risk measurement*," Insurance, Math. Econ., vol. 75, pp. 126–136, Jul. 2017.
K. K. Bagchi and G. Udo, "*An analysis of the growth of computer and Internet security breaches*," Commun. Assoc. Inf. Syst., vol. 12, no. 1, p. 46, 2003

Contribution of the paper

The **contribution of this paper** to the recent and fast-growing literature on Cyber risk modelling can be summarized as follows:

I. We build the class of **Zero-Inflated INGARCH models** to accommodate for the possibility of unreported data breaches, thus providing a methodological contribution.
II. We **uncover the dynamics** present in two different datasets, thus producing empirical evidence that data breaches possess an autoregressive structure.
III. We find statistical evidence of **explicative variables** explaining data breaches
IV. We apply the methodology developed to the problem of **insurability** of Cyber risk.

Privacy Rights Clearinghouse [*www.privacyrights.org*]

Records Breached: 11,575,804,706
from 8,804 DATA BREACHES
made public since 2005

The first dataset we analyze was obtained from the **Privacy Rights Clearinghouse (PRC)** which is one of the largest and most extensive datasets that is also publicly available.

PRC maintains the Chronology of Data Breaches as a source of information to assist in research involving reported data breaches from 2005 to present.

Many organizations are not aware they've been breached or are not required to report it based on reporting laws. PRC's Chronology is limited to data breaches reported in the U.S. If a data breach affects individuals in other countries, it is included only if individuals in the U.S. are also affected.

Privacy Rights Clearinghouse [*www.privacyrights.org*]

Year	Events	Records
2005	136	55,101,241
2006	482	68,580,749
2007	456	149,957,921
2008	355	130,896,900
2009	270	251,575,814
2010	801	140,937,393
2011	793	447,901,379
2012	886	298,766,833
2013	890	158,789,584
2014	869	1,313,623,927
2015	547	318,837,458
2016	826	4,815,012,420
2017	863	2,051,896,420
2018	828	1,371,001,705
2019	16	321,922

Types of data breach

CARD	Payment Card Fraud – fraud involving debit and credit cards that is not accomplished via hacking (e.g., skimming devices at point-of-service terminals)
DISC	Unintended disclosure – sensitive information either posted publicly on a website, mishandled, or sent to the wrong party via email, fax, or mail
HACK	Hacking or malware – electronic entry by an outside party, malware, and spyware
INSD	Insider – someone with legitimate access, such as an employee or contractor, intentionally breaches information
PHYS	Physical loss – lost, discarded, or stolen non-electronic records, such as paper documents
PORT	Portable device – lost, discarded, or stolen laptop, PDA, smartphone, portable memory device, CD, hard drive, data tape, etc.
STAT	Stationary device – lost, discarded, or stolen stationary electronic device, such as a computer or server not designed for mobility
UNKN	Unknown or other

Entity types

BSF	BSF Businesses – Financial and insurance services
BSO	BSO Businesses – Other
BSR	BSR Businesses – Retail/Merchant
EDU	EDU Educational institution
GOV	GOV Government and military
MED	MED Healthcare – Medical providers
NGO	NGO Nonprofit organizations

Type	Events	%	Records	%
CARD	68	0.75%	9,203,036	0.08%
DISC	1802	19.98%	2,815,845,013	24.33%
HACK	2584	28.65%	8,207,451,875	70.92%
INSD	608	6.74%	83,580,453	0.72%
PHYS	1735	19.24%	40,769,571	0.35%
PORT	1172	13.00%	185,650,895	1.60%
STAT	249	2.76%	16,235,932	0.14%
UNKN	800	8.87%	214,464,891	1.85%

Entity	Events	%	Records	%
BSF	788	8.74%	643,820,265	5.56%
BSO	1047	11.61%	8,990,170,575	77.68%
BSR	623	6.91%	1,383,161,417	11.95%
EDU	862	9.56%	66,376,099	0.57%
GOV	781	8.66%	227,483,420	1.97%
MED	4321	47.92%	242,968,015	2.10%
NGO	119	1.32%	8,444,531	0.07%
UNKN	477	5.29%	10,777,344	0.09%

Breach Level Index [*breachlevelindex.com*]

Data Breach Statistics
Data Records Lost or Stolen Since 2013
14,717,618,286 records
ONLY 4% of breaches were “Secure Breaches” where encryption was used and the stolen data was rendered useless.

The second dataset we analyze was obtained from the **Breach Level Index Data Breach Database** a centralized, global database of data breaches with calculations of their severity based on multiple factors.

The Breach Level Index not only tracks publicly disclosed breaches, but also allows organizations to do their own risk assessment based on a few simple inputs that will calculate their risk score, overall breach severity level, and summarize actions IT can take to reduce the risk score.

Gemalto is the world leader in digital security, helping the largest and most respected brands protect their data, identities, and intellectual property.

Breach Level Index [*breachlevelindex.com*]

YEAR	Events	Records
2013	1217	2,107,666,417
2014	1746	2,888,466,820
2015	1887	743,462,574
2016	1993	1,388,190,640
2017	1958	2,962,190,464
2018	1505	4,876,541,349

#	Source	Events	%	Records	%
1	**Accidental Loss**	2428	24%	4,532,637,539	30.3%
2	**Hacktivist**	164	2%	65,343,200	0.4%
3	**Lost Device**	5	0%	115,007	0.0%
4	**Malicious Insider**	1194	12%	306,945,069	2.1%
5	**Malicious Outsider**	6298	61%	9,430,616,718	63.0%
6	**Ransomware**	5	0%	-	0.0%
7	**State Sponsored**	130	1%	628,967,833	4.2%
8	**Stolen Device**	15	0%	59,069	0.0%
9	**Unknown**	67	1%	1,833,829	0.0%

#	Industry	Events	%	Records	%
1	**Education**	879	8.5%	126,843,836	0.8%
2	**Entertainment**	104	1.0%	502,594,229	3.4%
3	**Financial**	1301	12.6%	552,524,623	3.7%
4	**Government**	1418	13.8%	1,298,531,178	8.7%
5	**Healthcare**	2714	26.3%	291,675,274	1.9%
6	**Hospitality**	106	1.0%	527,606,802	3.5%
7	**Industrial**	138	1.3%	21,119,009	0.1%
8	**Insurance**	83	0.8%	12,700,290	0.1%
9	**Non-profit**	74	0.7%	410,488	0.0%
10	**Other**	1324	12.8%	3,110,303,702	20.8%
11	**Professional Services**	202	2.0%	147,140,489	1.0%
12	**Retail**	1131	11.0%	1,228,013,093	8.2%
13	**Social Media**	34	0.3%	2,758,853,076	18.4%
14	**Technology**	798	7.7%	4,388,202,175	29.3%

Framework

Count time series $\{Y_t : t \in N\}$. Y_t models the number of records stolen at time t.

Time-varying regressors $X_t = \left(X_{t,1}, \ldots, X_{t,r}\right)^T$

Conditional mean $E[Y_t | F_{t-1}] = \lambda_t$,
where F_t is the history generated by the joint process $\{Y_t, \lambda_t, X_t : t \in N\}$

General form:

$$\log(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \log(Y_{t-k} + 1) + \sum_{j=1}^{q} \alpha_j \log(\lambda_{t-j}) + \eta^T X_{t-1}$$

Specific form with p=q=1

$$\log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}) + \eta^T X_{t-1}$$

Distributions

Distributional assumption **Negative Binomial**

$$Y_t|F_{t-1} \sim NB(\lambda_t, \phi)$$

$$\text{with } P(Y_t|F_{t-1} = n) = p_n^Y = \frac{\Gamma(\phi+n)}{\Gamma(n+1)\Gamma(\phi)} \left(\frac{\phi}{\phi+\lambda_t}\right)^{\phi} \left(\frac{\lambda_t}{\phi+\lambda_t}\right)^{n}, n = 0,1, \ldots$$

Distributional Assumption **Poisson**

$$Y_t|F_{t-1} \sim Poiss(\lambda_t)$$

Zero-Inflated INGARCH models

Distributional Assumption **0-I Negative binomial** (*our own specification*)

$$Y_t|F_{t-1} \sim 0I - NB(\lambda_t, \phi, r)$$

$$\text{with } P(Y_t|F_{t-1} = n) = \tilde{p}_n^Y = \begin{cases} (1-r) + r\left(\frac{\phi}{\phi+\lambda_t}\right)^{\phi} & \text{if } n = 0 \\ r\, p_n^Y & \text{if } n > 0 \end{cases}$$

$$\tilde{Y}_t \sim NB(\lambda_t, \phi)$$

Y_t observed data breaches
$\tilde{Y}_t$ occurred data breaches

$$Y_t = I_t \tilde{Y}_t$$

$$I_t \sim Bern\,(r) \begin{cases} I_t = 1 & data\ breaches\ detected\ and\ reported \\ I_t = 0 & data\ breaches\ not\ detected\ or\ not\ reported \end{cases}$$

Explicative Variables

A data breach occurs when a cybercriminal successfully infiltrates a data source and extracts sensitive information.

Hackers search for these **data** because they can be used to **make money**

As part of their strategy, the attackers hold the information for ransom and demand a payment in order to have the data removed from the host website.
The motive of a cybercriminal defines what company he/she will attack. Different sources yield different information.

Criminal organizations now are treating this like a **business** “They’re going to plan, they’re going to make sure they understand how they’re going to execute and then they’re going to set out and see where they can execute.”

Bitcoins: Why do we care? What is the relationship with data breaches?

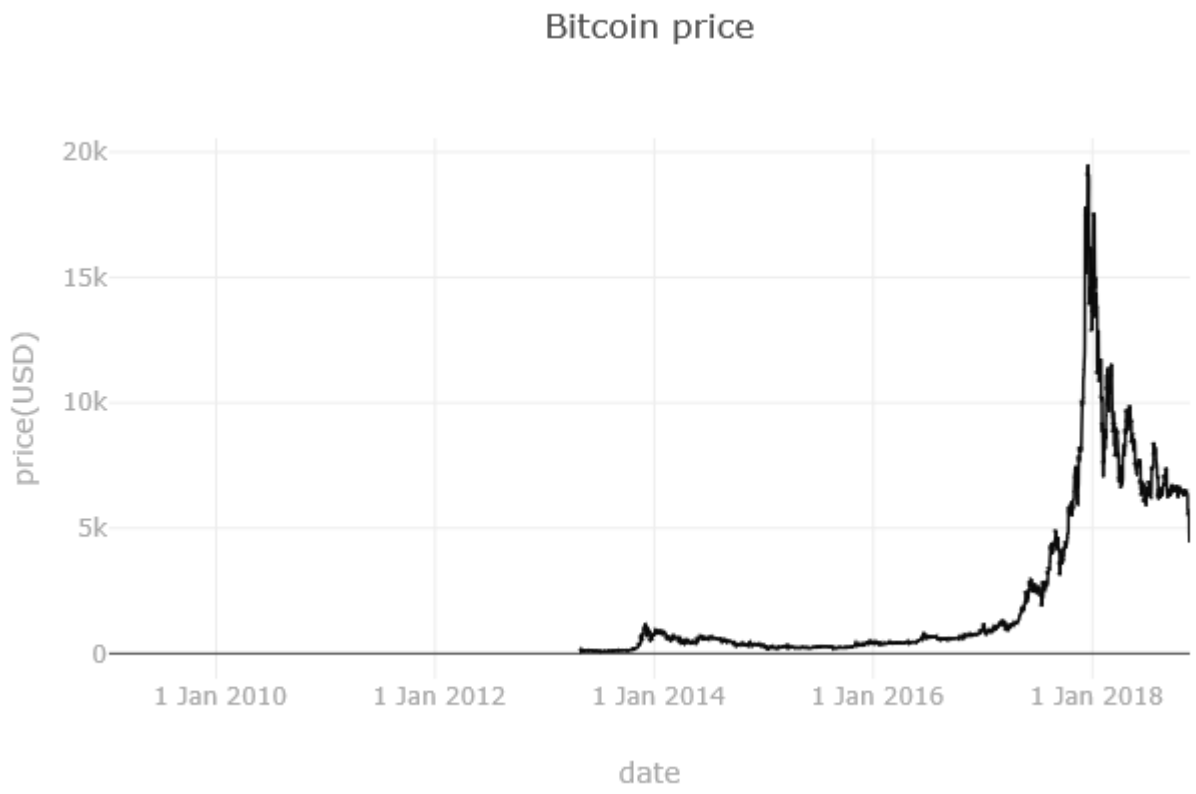
BitCoins

Bitcoin is a **digital payment currency** that utilizes cryptocurrency (a digital medium of exchange) and peer-to-peer (P2P) technology to create and manage monetary transactions as opposed to a central authority. The open source Bitcoin P2P network creates the bitcoins and manages all the bitcoin transactions.
Often referred to as "cash for the Internet," Bitcoin is one of several popular digital payment currencies along with Litecoin, Peercoin and Namecoin.

Bitcoin is considered the **biggest cryptocurrency**. It was first introduced in 2009 and is the most widely-traded cryptocurrency.

Bitcoin as an implementation of the cryptocurrency concept was described by Wei Dai in 1998 on the cypherpunks mailing list. Dai suggested a new form of money that uses cryptography to control its creation and transactions, rather than a central authority. In 2009, the Bitcoin specification and proof of concept was published in a cryptography mailing list by Satoshi Nakamoto. As noted in the Official Bitcoin FAQ, Satoshi Nakamoto left the project in late 2010 without revealing much about himself.

BitCoins [source datahub.io]



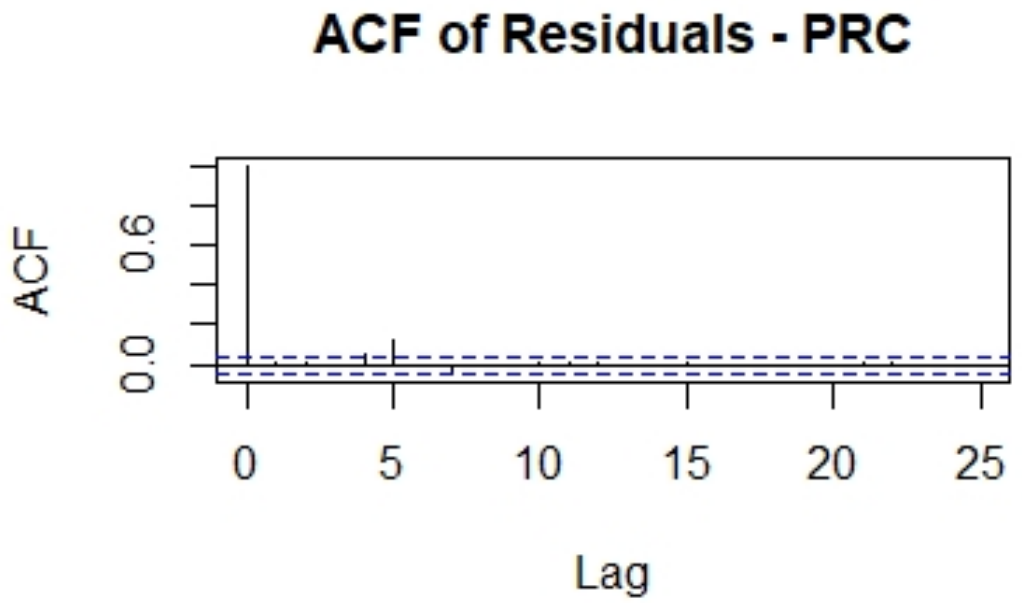
Field Name
date
txVolume(USD)
adjustedTxVolume(USD)
txCount
marketcap(USD)
price(USD)
exchangeVolume(USD)
generatedCoins
fees
activeAddresses
averageDifficulty
paymentCount
medianTxValue(USD)
medianFee
blockSize
blockCount

txCount - refers to the number of transactions happening on the public blockchain a day. Be aware that for low-fee blockchains, it's really easy to fabricate a whole bunch of transactions.

generatedCoins - refers to the number of new coins that have been brought into existence on that day. Actual number of newly-minted coins.

Results Database PRC

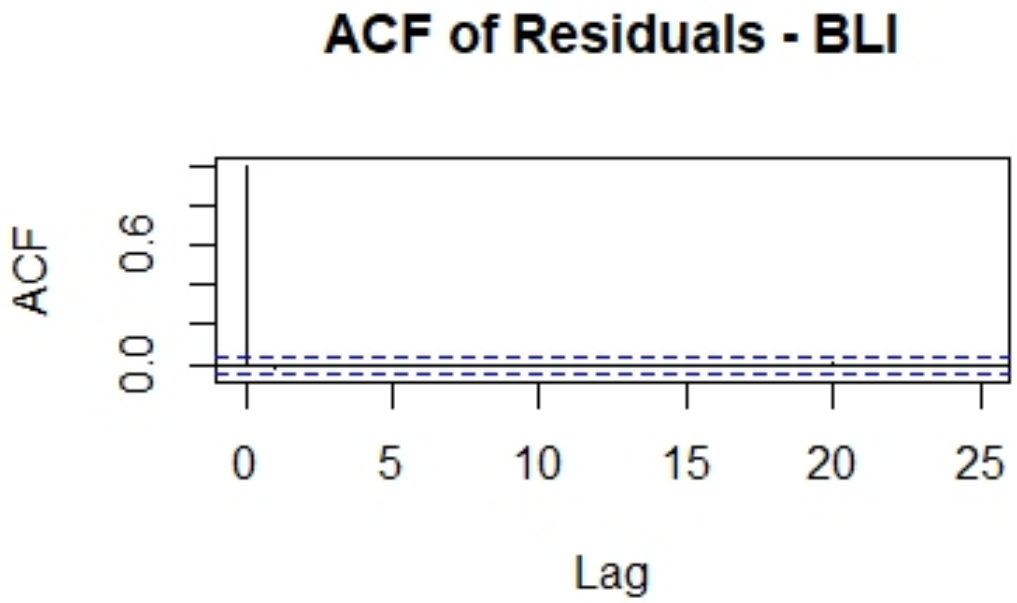
	Estimate	Std. Error	t value	Pr(>|t|)
gamma	- 0.15069	0.08290	-1.81777	0.06910
(Intercept)	4.16140	2.73801	1.51986	0.12855
beta_1	0.02239	0.02336	0.95840	0.33786
alpha_1	0.16208	0.14225	1.13938	0.25454
logGenerated	0.85439	0.31113	2.74611	0.00603
Return	-13.14767	3.84581	-3.41870	0.00063
interv_1	7.11124	5.49732	1.29358	0.19581
interv_2	6.23616	3.57748	1.74317	0.08130
interv_3	5.21903	5.58300	0.93481	0.34989
phi	0.07590	0.01001	7.58468	0.00000



In the PRC estimation the dynamics is NOT YET well captured. This is signaled by the non-significance of both beta_1 and alpha_1 and by the autocorrelation plot, where some lag is out of bounds.

Results Database BLI

	Estimate	Std. Error	t value	Pr(>|t|)
gamma	5.36868	658.19279	0.00816	0.99349
(Intercept)	0.02436	0.66901	0.03641	0.97096
beta_1	-0.00995	0.01186	- 0.83877	0.40160
alpha_1	0.72794	0.05380	13.53153	0.00000
logGenerated	0.45363	0.10871	4.17280	0.00003
Return	-4.35294	1.63529	- 2.66188	0.00777
interv_1	6.45844	2.34383	2.75551	0.00586
interv_2	5.48563	2.81723	1.94717	0.05151
interv_3	5.22553	2.55030	2.04899	0.04046
phi	0.05818	0.00174	33.47602	0.00000



In the BLI, the dynamics is fully captured. The value alpha_1 and its strong significance reflects a strong impact of the TODAY breach intensity on the TOMORROW intensity. The value of beta_1 is very close to zero. This means that the TODAY Size of the breach does not affect the TOMORROW intensity.

Comments

In both databases the value of the coefficients associated to *logGenerated* and *return* are **significant** and show the same **qualitative effect**.
An increase in log-generated rises the Tomorrow intensity.
An increase in the return reduces the Tomorrow intensity.

The value of phi is small; for this reason we have not considered the Poisson distribution (no good fit)

“Interventions” spike variables for outliers

Results Insurability

DB PRC	Records	Cost A ($ mln)	Cost B ($ mln)
E[1year]	44,014,919	6,514	1,384
DevStd[1year]	12,889,934	1,908	308
Var99.5%[1Year]	81,000,458	11,988	2,217

DB BLI	Records	Cost A ($ mln)	Cost B ($ mln)
E[1year]	122,690,591	18,158	3,026
DevStd[1year]	27,419,349	4,058	513
Var99.5%[1Year]	209,567,936	31,016	4,567

Cost A: Ponemon Cost per stolen record 2018

Cost B: Jacobs, J., 2014. Analyzing Ponemon Cost of Data Breach.

Conclusions

- Further steps
- Feedbacks appreciated, thank you for the attention

marco.pirra@unical.it