Decisions in Economics and Finance (2021) 44:141–160
https://doi.org/10.1007/s10203-020-00301-y

On the determinants of data breaches: A cointegration analysis

Domenico De Giovanni[1] · **Arturo Leccadito**[1] · **Marco Pirra**[1]

Received: 24 January 2020 / Accepted: 7 August 2020 / Published online: 5 September 2020

Abstract
Cyber risks and particularly data breaches constitute one of the new frontiers of risk modeling for insurers across the world. We use the cointegration methodology to uncover the relation between data breaches and Bitcoin-related variables. We perform our analyses on two different datasets of data breaches. In both cases, we provide statistical evidence of a bidirectional lead–lag relation in the short run between the variables under investigation. Moreover, the existence of a cointegrating vector suggests that this relation is likely to persist in the long run. To evaluate the quantitative implications of the relations found, we complement the study with Granger causality tests, impulse response analyses and variance decompositions of the forecasting errors.

Keywords Emerging risks · Cyber risk · Data breaches

1 Introduction

Concerns on cybersecurity threats are growing across all sectors of the global economy, as cyber risks have increased and cyber criminals have become progressively more dangerous.

In the last years, many stealthy and sophisticated cyber attacks targeted public and private sector organizations. The annual cost of cybercrime study conducted by

✉ Marco Pirra
marco.pirra@unical.it

Domenico De Giovanni
ddegiovanni@unical.it

Arturo Leccadito
arturo.leccadito@unical.it

[1] Università della Calabria, Rende (CS), Italy

Ponemon Institute[1] confirms that, combined with the expanding threat landscape, organizations are noticing a steady rise in the number of security breaches: The average number has moved from 130 in 2017 to 145 in 2018 (+11% last year,+67% last 5 years). The impact of these cyber attacks on organizations, industries and society is relevant, as the total cost of cybercrime for each company has increased from $11.7 million in 2017 to a new high of $13.0 million in 2018 (+12% last year, +72% last 5 years). The 2018 study reports that the global average cost of a data breach is up 6.4% over the previous year to $3.86 million. The average cost for each lost or stolen record containing sensitive and confidential information has also increased by 4.8% year over year to $148. According to the Online Trust Alliance,[2] the number of cyber attacks worldwide doubled in 2017 to 160,000, although endemic underreporting means that the true figure could be as high as 350,000.

Despite the improvements in security countermeasures and practices, the statistics presented above highlight how cyber insurance represents an important tool for risk managers to mitigate the economic impact of cyber attacks. The demand for cyber insurance is expected to experience a huge growth, as people and companies become aware of the economic risk behind cyber attacks. However, the market for cyber insurance is undersized, mainly because insurance and reinsurance companies are still unprepared to offer coverage for such kind of risks. As KPMG highlights in one of its report[3] on cyber insurance, insurers still need to improve their modeling capabilities with respect to these specific types of risk.

With the availability of new databases, academic research has started offering its contribution to understanding a particular class of cyber risk, namely data breaches. The literature on cyber risk and information security is plenty of papers in the area of information technology, while less work has been proposed in economics, finance and insurance. A comprehensive reference for an overview on the latter is Xu et al. (2018). In the study, the authors discuss a statistical analysis of a breach incident dataset obtained from the Privacy Rights Clearinghouse[4] and use stochastic processes to fit and predict inter-arrival times and breach sizes. The work includes a detailed review of prior contributions on the topic: Among others, it is worth mentioning Eling and Loperfido (2017) that analyzes the PRC Database with some actuarial insights, and the related studies on data breach statistical evaluations such as Maillart and Sornette (2010); Edwards et al. (2016); Wheatley et al. (2016, 2019, 2020). The PRC Database is also used in Farkas et al. (2019): The authors investigate the heterogeneity of the reported cyber claims using regression trees. The economical value of cyber risk is discussed in Eling and Wirfs (2019) where the authors focus the attention on cyber losses from an operational risk database and analyze the dataset with methods from

[1] The Ponemon Institute is dedicated to independent research and education that advances responsible information and privacy management practices within business and government. 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. More details can be found on the official Web site https://www.ponemon.org/.

[2] Allianz Barometer 2018, https://www.internetsociety.org/ota/. Last accessed December 2019.

[3] https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/07/cyber-insurance-report.pdf, pag 10.

[4] P. R. Clearinghouse. Privacy Rights Clearinghouse's Chronology of Data Breaches. Accessed: Dec. 2019. [Online]. Available: https://www.privacyrights.org/data-breaches.

statistics and actuarial science. As far as cyber insurance is concerned, a review of the available scientific approaches for the analysis of the cyber insurance market is Marotta et al. (2017), where the authors offer insights from both market and scientific perspectives.

In this paper, we go a step ahead in the understanding of data breaches by providing a dynamic analysis in which we find a causal relation between the intensity of data stolen and some metrics of the cryptocurrency market. Our original conjecture is as follows: If hackers perform data attacks to make a profit, they must somehow cash the attack. In this case, some cryptocurrencies offer the quickest and most anonymous way to monetize the attack. To test our postulate, we perform, on two distinct datasets of data breaches, a rigorous cointegration analysis between the daily number of stolen data, the daily Bitcoin's price and the daily number of transactions in Bitcoin. In addition, we run Granger causality tests between the three variables. In both datasets, we find strong empirical evidence of the existence of a causal relationship between the number of data breaches and the Bitcoin-related variables both in the short run and in the long run.

To the best of our knowledge, we provide for the first time a set of easily measurable variables that explain data breaches. Thus, our findings offer new insights into the statistical estimation and forecasts of data breaches. This might guide insurers and reinsurers in the process of building new products that offer protection against such kinds of risk.

In the remaining of the paper, we proceed as follows: In Sect. 2, we summarize the methodology used. In Sect. 3, we describe the datasets used and present the results of the cointegration analysis and Granger causality tests. To quantify the impact of Bitcoin-related variables on data breaches, we perform an impulse response analysis and a variance decomposition of the forecasting errors. In Sect. 4, we conclude highlighting our results and suggesting new directions of research.

2 Methodology: cointegration analysis

Cointegration analysis has been widely used in finance and economics. Among the other applications, it has been employed to investigate the lead–lag relationship between spot and futures prices (see for instance Tse 1995) and the integration and efficiency of international bond market (Mills and Mills 1991; Ciner 2007). As for the applications involving economic data, many authors have relied on cointegration analysis to test the purchasing power parity (Pippenger 1993; Chen 1995) or to examine the expectations theory of the term structure of interest rates (see Campbell and Shiller 1987; Shea 1992, among others).

A d-dimensional time series $\mathbf{Y}_t$ is said to be cointegrated of order (a, b) if each series is integrated of order a,[5] i.e., each series becomes stationary after taking first differences a times, and there exists a linear combination of the d variables, $\tilde{\mathbf{Y}}_t = \boldsymbol{\beta}'\mathbf{Y}_t$ with $\boldsymbol{\beta}$ nonzero $d \times 1$ vector, such that $\tilde{\mathbf{Y}}_t$ is integrated of order $a - b$. As with several economic time series, we are interested in the case in which $a = b = 1$, meaning that

[5] We use the notation $I(a)$ for a time series that is integrated of order a.

each of the one-dimensional components of $\boldsymbol{Y}_t$ has a unit root (i.e., it is integrated of order one), and their linear combination $\tilde{\mathbf{Y}}_t$ is instead $I(0)$. Hence, the starting point of cointegration analysis consists in establishing the order of integration of all the series of interest.

2.1 Unit root tests

The first step needed in a cointegration analysis involves testing if all the time series in $\mathbf{Y}_t$ are integrated of order one. To this end, besides the usual augmented Dickey–Fuller (ADF) tests, we consider the ADF-GLS test of Elliott et al. (1996). The authors proposed a variant of the ADF test which involves an alternative method of handling the parameters pertaining to the deterministic term: These are estimated first via generalized least squares, and in a second stage an ADF regression is performed using the GLS residuals. The usual ADF tests are based on the t-statistic on ϕ in the following regression:

$$\Delta y_{i,t} = \mu_t + \phi y_{i,t-1} + \sum_{j=1}^{p} \gamma_i \Delta y_{i,t-j} + \epsilon_{i,t}, \tag{1}$$

where $y_{i,t}$ is the ith component of $\mathbf{Y}_t$. The null hypothesis of a unit root is $\phi = 0$, tested against the alternative $\phi < 0$. Therefore, large negative values of the test statistic lead to the rejection of the null. If all the components of $\mathbf{Y}_t$ are found to be $I(1)$, then the econometrician can move to the next step, i.e., the Johansen procedure. Its aim is to establish whether $\mathbf{Y}_t$ is cointegrated and, if this is the case, how many cointegrating relations exist.

2.2 The Johansen procedure

A general vector autoregression (VAR) model with deterministic part $\boldsymbol{\mu}_t$ of the form:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Pi}_1 \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Pi}_k \mathbf{Y}_{t-k} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{2}$$

can be rewritten using the following vector error correction (VECM) specification[6]:

$$\Delta \mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Pi} \mathbf{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{k-1} \Delta \mathbf{Y}_{t-k+1} + \boldsymbol{\varepsilon}_t \tag{3}$$

where

$$\begin{aligned} \boldsymbol{\Gamma}_i &= -(\boldsymbol{\Pi}_{i+1} + \cdots + \boldsymbol{\Pi}_k), \quad i = 1, \ldots, k-1, \\ \boldsymbol{\Pi} &= -(\mathbf{I} - \boldsymbol{\Pi}_1 - \cdots - \boldsymbol{\Pi}_k), \end{aligned}$$

and Δ is the first difference operator, i.e., $\Delta \mathbf{Y}_t = \mathbf{Y}_t - \mathbf{Y}_{t-1}$.

[6] To derive the VECM specification, it suffices to use, for $i = 2, \ldots, k$, the identity $\mathbf{Y}_{t-i} = \mathbf{Y}_{t-1} - \sum_{h=1}^{i-1} \Delta \mathbf{Y}_{t-h}$ in the VAR equation.

We implement the tests developed by Johansen (1991) to test the hypothesis that $\mathbf{Y}_t$ is cointegrated of order (1, 1). Such hypothesis involves r, the rank of $\boldsymbol{\Pi}$. If $r \leq d-1$, one can write $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $d \times r$ matrices. The matrix $\boldsymbol{\beta}$ contains r linear cointegration parameter vectors, whereas $\boldsymbol{\alpha}$ is a matrix consisting of d error-correction parameter vectors (the so-called loadings). The maximum likelihood estimate of $\boldsymbol{\alpha}$ is obtained using the OLS regression of $\Delta\mathbf{Y}_t$ on $\Delta\mathbf{Y}_{t-1}, \ldots, \Delta\mathbf{Y}_{t-k+1}$ and a constant. Denote by $\hat{\boldsymbol{\epsilon}}_{0t}$ the residuals. Similarly, the maximum likelihood estimate of $\boldsymbol{\beta}$ can be obtained from the OLS regression of $\mathbf{Y}_t$ on $\Delta\mathbf{Y}_{t-1}, \ldots, \Delta\mathbf{Y}_{t-k+1}$ and a constant. In this case, denote by $\hat{\boldsymbol{\epsilon}}_{1t}$ the residuals. Given the residuals, it is possible to calculate for $j = 0, 1$ the matrices $\boldsymbol{S}_{ij} = T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\epsilon}}_{it}\hat{\boldsymbol{\epsilon}}'_{jt}$. Let $\hat{\lambda}_1 > \cdots > \hat{\lambda}_d$ be the eigenvalues obtained from solving the eigenvalue system $\left|\lambda\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}\right| = 0$, and $(\hat{\boldsymbol{\psi}}_1, \ldots, \hat{\boldsymbol{\psi}}_d)$ the corresponding eigenvectors. The estimate for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is obtained as the juxtaposition of the eigenvectors associated with the r largest eigenvalues, and the one for $\boldsymbol{\alpha}$ is obtained as $\hat{\boldsymbol{\alpha}} = \boldsymbol{S}_{01}\hat{\boldsymbol{\beta}}$. Two Johansen's maximum likelihood tests, the maximal eigenvalue test and the trace test, can then be used to determine the number of cointegration vectors. The statistic from the maximal eigenvalue test for the null hypothesis of r cointegration vectors against the alternative of $r+1$ cointegration vector is $\hat{\lambda}_{\max} = -T\log(1-\hat{\lambda}_{r+1})$. The trace test statistic for the null hypothesis of at most r cointegration vectors is $\hat{\lambda}_{\text{trace}} = -T\sum_{i=r+1}^{d}\log(1-\hat{\lambda}_i)$. If the results are consistent with the hypothesis of at least one cointegration vector, one then uses the maximum likelihood method to test the hypotheses regarding the restriction on $\boldsymbol{\beta}$.

3 The relation between data breaches and Bitcoin metrics

3.1 Data

In this paper, we look for short-term and/or long-term relations between data breaches and Bitcoin-related variables. More specifically, we perform two distinct analyses based on two publicly available databases of data breaches.

The first database is taken from the Chronology of Data Breaches provided by the Privacy Rights Clearinghouse[7] (PRC). The PRC dataset is publicly available and constantly updated on the PRC Web site and has been used in other recent investigations (see for instance, Eling and Loperfido 2017; Maillart and Sornette 2010; Edwards et al. 2016; Wheatley et al. 2016; Farkas et al. 2019; Wheatley et al. 2019, 2020).

The second dataset is obtained from the Breach Level Index (BLI) Data Breach Database, a centralized, global database of data breaches with calculations of their severity based on multiple factors. The BLI tracks publicly disclosed breaches and also allows organizations to do their own risk assessment since, on the basis of a few simple inputs, it calculates their risk score, overall breach severity level, and summarizes possible actions to reduce the risk score. The dataset has been downloaded

[7] The Privacy Rights Clearinghouse is a US non-profit organization founded in 1992 whose aim is the privacy protection for US citizens by empowering individuals and advocating for positive change. The dataset is available at https://privacyrights.org/data-breaches.

from the Web site of Gemalto, part of the Thales Group, one of the world leaders in digital security.[8]

We mention that the databases do not necessarily contain all of the hacking breach events because there may be unreported ones. The exposure to this type of risk is not easy to be tracked, since the population of potential victims that would report to registers is not stable through time or, at least not known in opposition to the type of information that comes from an insurer that might have a clearer view on the exposure, for example. Many organizations are not aware they have been breached or they are not required to report it according to the reporting laws. PRCs Chronology is limited to those reported in the USA. If a data breach affects individuals in other countries, it is included only if individuals in the USA are also affected. The data contain only the number of records affected by data breaches and do not include financial losses.

The PRC database is organized in industries and type of attack. The BLI database also includes the country interested by the breach.

As for the Bitcoin-related variables, we are mainly interested in the daily price and the daily number of transactions of Bitcoins. Our data source is taken from DataHub, [9] a project by Datopian and Open Knowledge International that provides publicly available high-quality datasets. As for the cryptocurrency, we focus on the historical prices (USD), on the number of transactions happening on the public blockchain during a given day.

The period of time we refer to, considering the dimensions of the datasets and in order to make them comparable, goes from January 1, 2013, to December 31, 2018. The work aims at investigating the idea that some breaches are treated by criminal organizations to make money: For this reason, the analysis focuses on malicious breaches related to hackers while negligent breaches and other sub-categories of malicious breaches included in the databases (i.e., insider, payment card fraud, ransomware, accidental unknown) have been ignored.

When researchers apply cointegration, both the sample size and the time span are relevant. Hakkio and Rush (1991) argue that cointegration is a long-run concept and hence long spans of data are needed for cointegration tests to have power and that gains from using more frequently sampled observation while keeping the same time span are "*more apparent than real*." The Monte Carlo study of Zhou (2001), while confirming the importance of the time span, reveals that increasing the data frequency may yield substantial power gains. Since the considered time series on data breaches goes back only to 2013 and is obviously available only at a daily frequency, in the present paper, we use the longest span at the highest possible frequency. We believe that the time span is long enough to have reliable results regarding the short-term and long-term relations between the variables of interest.

The dynamic behavior of data breaches is represented by integer-valued time series displaying an unusual pattern which resembles a point process. Figure 1 plots the two time series generated with the datasets used. The figure clearly shows the impulsive nature intrinsic in time series of this kind. In such cases, standard cointegration analysis

[8] We downloaded the database from https://breachlevelindex.com in April 2019. However, to the best of our knowledge, the accessibility policy of Gemalto seems to be changed.

[9] https://datahub.io/.

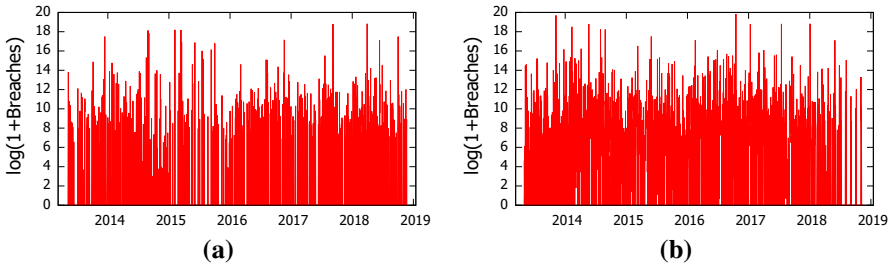

Fig. 1 Plot of the observed time series. Left panel: PRC dataset. Right panel: BLI dataset. All plots are in logarithmic scale for visualization purposes

cannot be applied directly to the time series.[10] Our empirical strategy to overcome this problem is to extrapolate from the original dataset a new latent time series that generated the observed pattern of data breach. Then, we perform the cointegration analysis using the extrapolated latent time series instead of the observed data. This approach is quite common in cointegration analysis. We refer to Niu and Melenberg (2014) for cointegration analysis which uses latent factors.

We report additional details in the subsequent subsection.

3.2 Time series of counts and their intensities

Integer-valued GARCH models (henceforth INGARCH) constitute a popular class of models for time series of counts. Although the name might suggest some sort of affinity with the well-known GARCH models, INGARCHs are auto-regressive moving average processes constructed to model the dynamics of phenomena that are discrete. An INGARCH model allows the conditional expected value of a discrete random variable (or some transformation) that models a countable phenomenon to depend on its previous values and on previous observations of the phenomenon itself. As a general discussion of such processes is far beyond the scope of the present paper, we restrict ourselves to the presentation of model used to extract the latent time series and refer to Weiß (2018) for an outstanding introduction on INGARCH models.

Let B_t be the observed breach size at time t. We assume that the conditional distribution of $B_t|(B_0, \ldots, B_{t-1})$ of the observed breach size given the previous realizations follows a Negative Binomial distribution $NB\left(s,\, p_t = \frac{\mu_t}{s+\mu_t}\right)$ with probability mass

$$P(B_t = k) = \frac{\Gamma(s+k)}{\Gamma(k+1)\Gamma(s)} \left(\frac{\mu_t}{s+\mu_t}\right)^k \left(\frac{s}{s+\mu_t}\right)^s,$$

so that we have $E\left(B_t|(B_0, \ldots, B_{t-1})\right) = \mu_t$ and $Var\left(B_t|(B_0, \ldots, B_{t-1})\right) = \mu_t + \frac{\mu_t^2}{s}$. In addition, we allow $\lambda_t = \log(\mu_t)$ to follow an ARMA process of order (p, q) of the following type:

[10] We are extremely grateful to an anonymous reviewer for pointing out this issue.

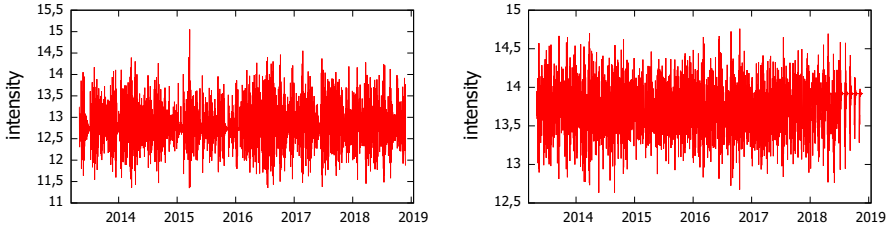

Fig. 2 Estimated logarithmic conditional expectations of the INGARCH model. Left panel: PRC dataset. Right panel: BLI dataset

$$\lambda_t = a_0 + \sum_{i=1}^{p} a_i \log\left(B_{t-i} + 1\right) + \sum_{j=1}^{q} b_j \lambda_{t-j}.$$

This model appeared the first time in Fokianos and Tjøstheim (2011). It is a generalization of the basic INGARCH model that allows for both positive and negative serial correlation. The choice of a logarithmic scale for the observed time series is needed to ensure the positivity of the conditional expectation μ_t. Fokianos and Tjøstheim (2011) also show that adding a constant to the logarithmic transformation of the time series does not alter the estimation process. Although originally proposed in association with a Poisson distribution for the observed time series, the strong over-dispersion present in the breach data motivates our choice of a negative binomial distribution.

We use maximum likelihood to fit the INGARCH model with the main goal to extrapolate the time series of (logarithmic) conditional expected values λ_t. We plot the resulting extrapolated time series in Fig. 2 and observe that the dynamic patterns of the latent time series are well suited for standard cointegration analysis. Thus, in what follows we will use the extrapolated time series to question cointegration by using λ_t instead of the original dataset. Accordingly, we specify the vector in (1) and (2) as $\mathbf{Y}_t = \left(C_t, P_t, \lambda_t^h\right)'$, where C_t and P_t are, respectively, the logarithm of the daily number of transactions in Bitcoin and the daily Bitcoin's price and λ_t^h as the logarithmic conditional expectation $h = \text{PRC, BLI}$.

3.3 Empirical evidence

The standard cookbook of cointegration analysis requires first a preliminary test to check the order of integration of the time series under investigation. According to the methodology explained in Sect. 2, to check whether or not each time series is integrated of order one, we perform the ADF-GLS test based on regressions (1) for each variable and each first-order difference and report the results in Table 1. From Table 1, we conclude that the time series under investigation are integrated with order of integration one. Indeed, the unit root tests performed on the levels of each variable under consideration lead to not rejecting the null hypothesis of $\phi = 0$, suggesting that the time series has a unit root, while the first-order difference leads to the rejection of the null hypothesis, meaning that stationarity is achieved after applying the first difference

Table 1 Unit root tests by means of ADF-GLS

Variable	Lags	t-value	p-value
C_t	79	0.834	0.891
ΔC_t	0	−35.656	4.96E-21
P_t	99	−0.922	0.317
ΔP_t	21	−7.479	1.07E-12
λ_t^{PRC}	142	−0.391	0.544
$\Delta\lambda_t^{\text{PRC}}$	0	−50.1001	0.0001
λ_t^{BLI}	132	0.549398	0.8349
$\Delta\lambda_t^{\text{BLI}}$	21	−84.8524	0.0001

Table 2 Johansen's cointegration tests for the PRC dataset

Rank	Eigenvalue	$\hat{\lambda}_{\text{trace}}$		$\hat{\lambda}_{\text{max}}$	
		Statistics	p value	Statistics	p-value
0	0.128	507.110	0.000	277.010	0.000
1	0.106	230.100	0.000	227.630	0.000
2	0.001	24.695	0.116	24.695	0.116

operator. In what follows, we discuss the results from vector error-correction model for each of the two datasets.

PRC dataset

Having established that all the series involved in the analysis are $I(1)$, here we determine whether there exists a cointegration relation between the variables. The Johansen's tests are based on the rank of the matrix $\boldsymbol{\Pi}$ of equation, r. The null hypothesis $r = 0$ implies no cointegration, while $r > 0$ ($r = 1, \ldots, d-1$) means that there are r cointegrating relations. In the latter case, r distinct linear combinations of the variables—the cointegrating vectors—represent the long-run relation between the components of the multivariate time series. Table 2 presents the Johansen's tests on the PRC dataset, where we follow the Box–Jenkins' model selection technique to select the optimal order in the VAR model (3), identified to be 8 according to the Bayesian information criterion (BIC). Both the trace test and the maximal eigenvalue test agree to accept the null hypothesis $r = 2$.

Having identified $r = 2$ the rank of $\boldsymbol{\Pi}$, we proceed by estimating the vector error-correction model with one cointegrating vector. Table 3 reports the resulting vector error-correction model for data breaches of PRC dataset, the daily number of transactions of Bitcoin and the Bitcoin's price, from which we identify both a short-run and a long-run relation between the lagged variables ΔC_t and $\Delta\lambda_t^{\text{PRC}}$ and a long-run relation between ΔP_t and $\Delta\lambda_t^{\text{PRC}}$.

In Table 3, the coefficients of the vector auto-regression marked as underlined highlight the short-run relation between the lagged logarithm of conditional expectations of data breaches and the lagged logarithm of the number of transactions in Bitcoin.

Table 3 Vector error-correction estimation. PRC dataset[a,b]

	ΔP_t		ΔC_t		$\Delta\lambda_t^{\text{PRC}}$	
	Statistics	p value	Statistics	p-value	Statistics	p value
Intercept	0.055	0.749	−1.018	0.009	201.072	0.000
ΔP_{t-1}	−0.092	0.120	−0.440	0.001	0.236	0.564
ΔP_{t-2}	−0.107	0.052	−0.302	0.015	0.263	0.491
ΔP_{t-3}	−0.112	0.027	−0.190	0.095	0.218	0.535
ΔP_{t-4}	−0.084	0.065	−0.127	0.210	0.114	0.716
ΔP_{t-5}	−0.028	0.472	0.002	0.979	−0.069	0.799
ΔP_{t-6}	0.032	0.305	0.062	0.378	0.053	0.807
ΔP_{t-7}	0.011	0.613	0.041	0.406	0.079	0.608
ΔC_{t-1}	0.027	0.007	−0.420	0.000	−0.158	0.021
ΔC_{t-2}	0.017	0.100	−0.444	0.000	0.195	0.008
ΔC_{t-3}	0.005	0.613	−0.429	0.000	0.263	0.000
ΔC_{t-4}	0.024	0.030	−0.359	0.000	0.135	0.076
ΔC_{t-5}	0.014	0.180	−0.438	0.000	0.198	0.007
ΔC_{t-6}	0.016	0.134	−0.188	0.000	0.063	0.395
ΔC_{t-7}	0.014	0.148	0.163	0.000	−0.326	0.000
$\Delta\lambda_{t-1}^{\text{PRC}}$	0.005	0.689	−0.078	0.007	**0.626**	**0.000**
$\Delta\lambda_{t-2}^{\text{PRC}}$	0.005	0.671	−0.077	0.004	**−0.180**	**0.029**
$\Delta\lambda_{t-3}^{\text{PRC}}$	0.005	0.654	−0.073	0.002	−0.135	0.066
$\Delta\lambda_{t-4}^{\text{PRC}}$	0.006	0.525	−0.067	0.001	−0.063	0.320
$\Delta\lambda_{t-5}^{\text{PRC}}$	0.006	0.438	−0.029	0.074	−0.087	0.085
$\Delta\lambda_{t-6}^{\text{PRC}}$	0.004	0.347	−0.025	0.014	−0.013	0.668
$\Delta\lambda_{t-7}^{\text{PRC}}$	0.003	0.390	−0.005	0.496	−0.009	0.694
$\Psi^{PRC,1}$	−0.915	0.000	0.617	0.000	−0.194	0.655
$\Psi^{PRC,2}$	0.001	0.000	−0.002	0.000	0.029	0.000

[a] Additional details about the estimation are provided in Appendix 1.

[b] Cointegrating vectors: $\beta_1 = (1, 0, -0.08233)$, $\beta_2 = (0, 1, -55.434)$; adjustment vectors: $\alpha_1 = (-0.91548, 0.61663, -0.19402)$, $\alpha_2 = (0.00146, -0.002388, 0.02903)$

More precisely, almost all the lagged variables ΔC have a strong negative impact on the lagged conditional expectations of data breaches today, as one may observe from the value and the highly significance of the regression coefficients. This suggests that the number of transactions in Bitcoin might be a good predictor for data breaches. The intuition behind this result is that hackers prepare themselves to monetize the data attack (either by selling the data or by extorting money to the legitimate data proprietor) some days before, by operating on the Bitcoin market.

Furthermore, when looking at the equation for the Bitcoin transaction in the short-run component of the estimated VECM model, we find that almost all lagged variables $\Delta\lambda^{\text{PRC}}$ are statistically significant when regressing ΔC_t. This result, coupled with the one regarding the equation for $\Delta\lambda_t^{\text{PRC}}$, implies that there is a lead–lag relation between data breaches and transactions in Bitcoins and the relation is bidirectional. Hence, our

results confirm that some of the movements in cryptocurrency markets depend on illegalities (in our case cyber attacks). Also looking at Table 3, we find no short-run link between lagged conditional expectations of data breaches and Bitcoin price. Although the number of transactions has a strong impact on data breaches, this link is not necessarily reflected in a short-run impact on Bitcoin price.

The long-run relation between the variables under investigation is described by the existence of two cointegrating variables $\Psi_t^{\mathrm{BRC},1} = P_t - 0.08233\lambda_t^{\mathrm{PRC}}$ and $\Psi_t^{\mathrm{BRC},2} = C_t - 55.434\lambda_t^{\mathrm{PRC}}$, obtained by the estimation procedure detailed in Sect. 2.2. The double-underlined coefficient in Table 3 highlights that changes of $\Delta\lambda_t$ are affected with very high statistical significance by the second cointegrating variable. Since both the logarithm of the number of transactions and the logarithm of conditional expectations of data breaches enter the second cointegrating vector, we highlight that the short-run impact found is likely to persist in the long time. Our conjecture about the intuition behind this long-term link between transactions in Bitcoin and data breaches is as follows. On the one hand, once the hackers have monetized the breach, they possess a bunch of Bitcoins that will later be used in some different contexts. On the other hand, a *remunerative* data breach creates incentives to prepare more cyber attacks, which in turn create the needs of more transactions in Bitcoin. We also find high statistical significance of both cointegrating variables in the equation for the change in Bitcoin prices. This implies that the effects of Bitcoin metrics and conditional expectations of data breach will impact also on Bitcoin returns in the long run. For instance, the negative and significant coefficient associated with Ψ^1 indicates that if the difference between the linear combination of Bitcoin price and the logarithm of the conditional expectations of data breaches is positive in one period, the price will fall during the next period to restore equilibrium, and vice versa.

Summarizing the empirical evidence discussed in this section, the change of expected number of data breaches recorded in the PRC dataset is statistically (and negatively) influenced in the short run by its lagged variables $\Delta\lambda_{t_i}^{\mathrm{PRC}}$, $i = 1, \ldots, 3$,[11] and by the lagged levels of the number of transactions in Bitcoin some days before the attack. In the long run, deviations from the cointegration link, whose components are data breaches and transactions of Bitcoin, cause changes in the data breach intensities.

BLI dataset

The cointegration analysis of the BLI dataset fully confirms the existence of a strong, statistically significant, link between data breaches and number of transactions in Bitcoin, both in the short run and in the long run. Table 4 reports Johansen's cointegration tests and highlights the existence of two cointegrating vectors. The optimal order for the VAR model is once again 8 and has been selected according to Box–Jenkins' technique. The strong significance of underlined coefficients in Table 5 indicates the short-term relation. More specifically, the lagged variables ΔC_{t-1}, ΔC_{t-2} and ΔC_{t-6} impact heavily on the lagged value of the logarithmic conditional expectations of data breaches. We also find statistical significance for the reverse relation, especially in

[11] See the statistical significance of coefficients highlighted in bold in Table 3. This is a statistical evidence that hackers usually perform their attacks to different organizations in a small period of time.

Table 4 Johansen's Cointegration Tests for BLI dataset

Rank	Eigenvalue	$\hat{\lambda}_{\text{trace}}$ Statistics	p-value	$\hat{\lambda}_{\max}$ Statistics	p value
0	0.110	392.760	0.000	235.540	0.000
1	0.074	157.220	0.000	154.930	0.000
2	0.001	22.898	0.130	22.898	0.130

the variables $\Delta\lambda_{t-6}^{\text{BLI}}$ and $\Delta\lambda_{t-7}^{\text{BLI}}$. This confirms the intuition provided in the previous section for which data breaches have a statistical effect in the number of transactions in Bitcoin. The analysis also confirms the lack of a significant short-term relation between data breaches and price of Bitcoin.

The two cointegrating variables $\Psi_t^{\text{BLI},1} = P_t - 0.524\lambda_t^{\text{BLI}}$ and $\Psi_t^{\text{BLI},2} = C_t - 354.95\lambda_t^{\text{BLI}}$, obtained by inserting the estimated cointegrating of Table 5 into the VECM equation (3), describes the long-term relation among the three variables under investigation. We see that changes in logarithmic expectations of data breaches are due to changes in lagged logarithms of the number of transactions, the cointegrating variable itself (double-underlined coefficient of Table 5) and lagged logarithm of data breaches (bold coefficients in Table 5). This highlights once again the autoregressive structure of data breaches. Once again, the short-term relation between Bitcoin metrics and data breach is reflected in a long-run impact in the Bitcoin rate of return.

3.4 Granger causality tests

In this section, we perform a series of Granger causality tests (Granger 1969, 1988; Sims 1972) to provide further evidence on the lead–lag relationships between data breaches and Bitcoin metrics. Granger causality test is a standard tool for uncovering lead–lag relationships among economic variables. With reference to the most successful applications, we mention Chan (1992); Abhyankar (1998) among others. However, it is worth mentioning that novel methodologies for determining time-dependent lead–lag relations based on optimal thermal paths appeared recently, as in Meng et al. (2017); Xu et al. (2017).

We use first differences of the variables $\left(C_t, P_t, \lambda_t^h\right)'$ and perform the tests on both bivariate and trivariate VAR models. In the case of bivariate models, testing for instance the null that ΔC_t does not Granger-cause $\Delta\lambda_t$, amounts to estimating the VAR comprising the two variables, and testing the null that the coefficients associated with the first variable are all zero in the equation for $\Delta\lambda_t$, against the alternative that at least one is different than zero. In the case of the trivariate model (like the one in Eq. (3)), we perform the same test, but the VAR model includes also the remaining variable. We choose the order of autoregression according to the best Bayesian information criterion. We report the results in Tables 6 and 7. Underlined p values indicate rejection of the null hypothesis and thus Granger causality between the variables under investigation. We observe that both tables agree in all the cases under consideration but in one. Moreover, the results are fully consistent with the VECM models previously considered.

Table 5 Vector error-correction estimation. BLI dataset[a,b]

	ΔP_t		ΔC_t		$\Delta\lambda_t^{\text{BLI}}$	
	Statistics	p-value	Statistics	p-value	Statistics	p-value
Intercept	0.056	0.725	−0.291	0.422	129.025	0.000
ΔP_{t-1}	−0.097	0.101	−0.475	0.000	−0.048	0.898
ΔP_{t-2}	−0.110	0.047	−0.338	0.007	0.100	0.777
ΔP_{t-3}	−0.115	0.024	−0.226	0.049	0.242	0.453
ΔP_{t-4}	−0.084	0.063	−0.164	0.110	0.140	0.627
ΔP_{t-5}	−0.027	0.496	−0.017	0.847	0.106	0.670
ΔP_{t-6}	0.034	0.279	0.056	0.430	−0.041	0.838
ΔP_{t-7}	0.012	0.588	0.042	0.404	−0.010	0.943
ΔC_{t-1}	0.027	0.005	−0.417	0.000	−0.224	0.000
ΔC_{t-2}	0.016	0.120	−0.450	0.000	−0.162	0.016
ΔC_{t-3}	0.002	0.886	−0.431	0.000	−0.063	0.344
ΔC_{t-4}	0.021	0.056	−0.355	0.000	−0.068	0.326
ΔC_{t-5}	0.012	0.246	−0.455	0.000	0.016	0.812
ΔC_{t-6}	0.013	0.203	−0.212	0.000	0.177	0.008
ΔC_{t-7}	0.013	0.181	0.160	0.000	0.075	0.234
$\Delta\lambda_{t-1}^{\text{BLI}}$	0.003	0.820	−0.013	0.609	**−0.649**	**0.000**
$\Delta\lambda_{t-2}^{\text{BLI}}$	−0.004	0.693	−0.007	0.781	**−0.506**	**0.000**
$\Delta\lambda_{t-3}^{\text{BLI}}$	−0.008	0.415	−0.007	0.751	**−0.355**	**0.000**
$\Delta\lambda_{t-4}^{\text{BLI}}$	−0.008	0.393	−0.009	0.677	**−0.241**	**0.000**
$\Delta\lambda_{t-5}^{\text{BLI}}$	−0.010	0.238	−0.021	0.261	**−0.181**	**0.001**
$\Delta\lambda_{t-6}^{\text{BLI}}$	−0.011	0.102	*−0.033*	0.026	**−0.123**	**0.003**
$\Delta\lambda_{t-7}^{\text{BLI}}$	−0.004	0.215	*−0.024*	0.003	−0.043	0.056
$\Psi^{\text{BLI},1}$	−0.910	0.000	0.646	0.000	0.173	0.665
$\Psi^{\text{BLI},2}$	0.001	0.000	−0.001	0.000	0.002	0.000

[a] Additional details about the estimation are provided in the Appendix
[b] Cointegrating vectors: $\beta_1 = (1, 0, -0.524)$, $\beta_2 = (0, 1, -354.95)$; adjustment vectors: $\alpha_1 = (-0.91, 0.646, 0.173)$, $\alpha_2 = (0.00136, -0.001016, 0.0024)$

3.5 Impulse response analysis

Having discovered a causal link between Bitcoin metrics and conditional expectations of data breaches recorded in two different databases, we analyze the impulse response function (IRF) of the estimated VECM to evaluate the response of conditional expectations of data breaches with respect to unexpected shocks of the Bitcoin-related variables. The possibility of studying impulse responses in the context of cointegration analysis and the feasibility of combining the two approaches has been demonstrated by Lütkepohl and Reimers (1992).

In what follows, the IRF of variable i to shock j is defined as the sequence of the elements in the ith row and jth column of the matrices $\{\boldsymbol{\Phi}_k\}_{k=0,1,\ldots}$. Assuming

Table 6 Granger causality tests for bivariate VARs

Null			Test stat		p-value
PRC dataset					
$\Delta\lambda_t^{\text{PRC}}$	Does not Granger cause	ΔC_t	F(15, 1990)	4.7275	0
ΔP_t	Does not Granger cause	ΔC_t	F(21, 1972)	2.445	0.0003
ΔC_t	Does not Granger cause	ΔP_t	F(21, 1972)	1.5969	0.0418
$\Delta\lambda_t^{\text{PRC}}$	Does not Granger cause	ΔP_t	F(16, 1987)	0.58303	0.8988
ΔC_t	Does not Granger cause	$\Delta\lambda_t^{\text{PRC}}$	F(15, 1990)	8.5168	0
ΔP_t	Does not Granger cause	$\Delta\lambda_t^{\text{PRC}}$	F(16, 1987)	1.3138	0.1791
		BLI dataset			
Null			Test stat		p-value
$\Delta\lambda_t^{\text{BLI}}$	Does not Granger cause	ΔC_t	F(14, 1992)	1.7334	0.0434
ΔP_t	Does not Granger cause	ΔC_t	F(21, 1971)	2.4406	0.0003
ΔC_t	Does not Granger cause	ΔP_t	F(21, 1971)	1.5963	0.042
$\Delta\lambda_t^{\text{BLI}}$	Does not Granger cause	ΔP_t	F(10, 2004)	1.1557	0.3165
ΔC_t	Does not Granger cause	$\Delta\lambda_t^{\text{BLI}}$	F(14, 1992)	4.6984	0
ΔP_t	Does not Granger cause	$\Delta\lambda_t^{\text{BLI}}$	F(10, 2004)	0.42712	0.9341

Table 7 Granger causality tests for trivariate VARs

Null			Test Stat		p-value
PRC dataset					
$\Delta\lambda_t^{\text{PRC}}$	Does not Granger cause	ΔC_t	F(13, 1983) =	3.4644	0
ΔP_t	Does not Granger cause	ΔC_t	F(13, 1983) =	5.058	0
ΔC_t	Does not Granger cause	ΔP_t	F(13, 1983) =	1.8495	0.0314
$\Delta\lambda_t^{\text{PRC}}$	Does not Granger cause	ΔP_t	F(13, 1983) =	0.95824	0.4909
ΔC_t	Does not Granger cause	$\Delta\lambda_t^{\text{PRC}}$	F(13, 1983) =	10.5	0
ΔP_t	Does not Granger cause	$\Delta\lambda_t^{\text{PRC}}$	F(13, 1983) =	0.9807	0.4678
		BLI dataset			
Null			Test Stat		p-value
$\Delta\lambda_t^{\text{BLI}}$	Does not Granger cause	ΔC_t	F(14, 1978)	1.7534	0.0402
ΔP_t	Does not Granger cause	ΔC_t	F(14, 1978)	3.2021	0
ΔC_t	Does not Granger cause	ΔP_t	F(14, 1978)	1.4278	0.1317
$\Delta\lambda_t^{\text{BLI}}$	Does not Granger cause	ΔP_t	F(14, 1978)	1.1912	0.2749
ΔC_t	Does not Granger cause	$\Delta\lambda_t^{\text{BLI}}$	F(14, 1978)	4.5617	0
ΔP_t	Does not Granger cause	$\Delta\lambda_t^{\text{BLI}}$	F(14, 1978)	0.58771	0.8767

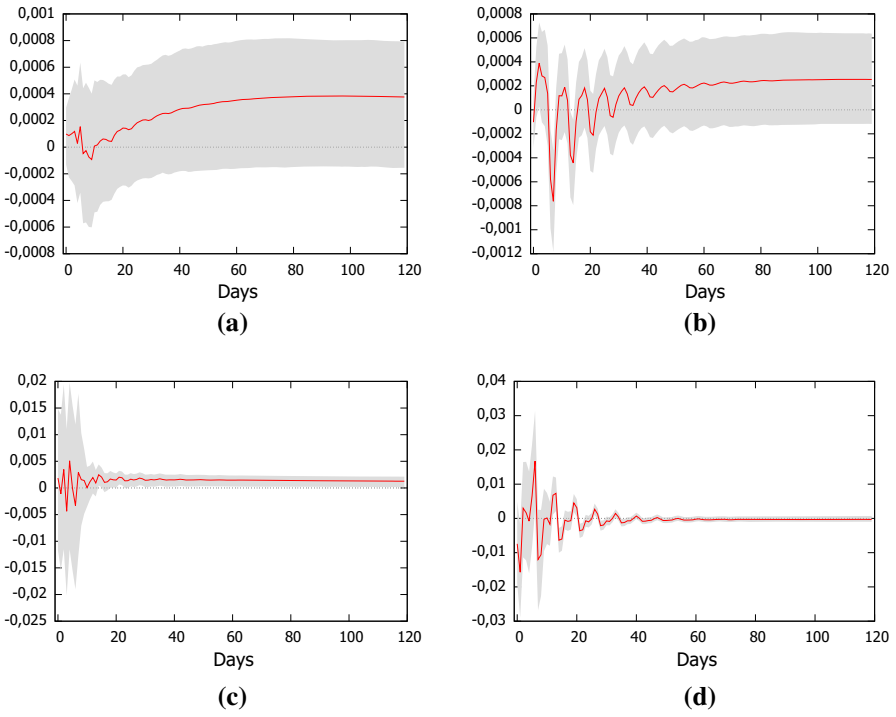

Fig. 3 Impulse response point estimates and 95% confidence bands. **a**: Shock variable P_t, response variable λ_t^{PRC}. **b**: Shock variable C_t, response variable λ_t^{PRC}. **c**: Shock variable P_t, response variable λ_t^{BLI}. **d**: Shock variable C_t, response variable λ_t^{BLI}

that the error term in (2)–(3) can be written as a linear combination of mutually uncorrelated shocks with unit variance, i.e., $\boldsymbol{\varepsilon}_t = \boldsymbol{H}\boldsymbol{u}_t$, these matrices are obtained as $\boldsymbol{\Phi}_k = \frac{\partial \boldsymbol{Y}_t}{\partial \boldsymbol{u}_{t-k}} = \frac{\partial \boldsymbol{Y}_t}{\partial \boldsymbol{\varepsilon}_{t-k}}\boldsymbol{H}$. The matrix $\boldsymbol{H}$ is assumed to be lower triangular, and its estimate is obtained as the Cholesky decomposition of the estimated variance covariance matrix of $\boldsymbol{\varepsilon}_t$ (see Lütkepohl 2006, Chapter 9). Choosing $\boldsymbol{H}$ to be lower triangular implies that, in general, the ordering of variables in the vector $\boldsymbol{Y_t}$ is important. Since we are interested in the effects of Bitcoin-related variables on data breaches, we put the latter as the last variable in the ordering, so that in this setting the variable Delta λ responds instantaneously to shocks associated with the remaining two variables. However, we have verified that, in our case, a different order does not change much the estimated impulse response function.

Figure 3 reports point estimates and 95% confidence intervals of the response of λ^h, $h = PRC, BLI$, with respect to exogenous shocks of Bitcoin's price and number of transactions of one standard deviation, for a period of 120 days, although the impact of the C_t variable seems stronger in the short run (up to about two weeks), especially for the BLI dataset. For both datasets, the impulse response estimates associated with the C_t variable cross the zero axis more often than the ones associated with the Bitcoin's price do. More specifically, Panels (a) and (b) refer to changes of PRC data breaches, while panels (c) and (d) refer to changes of BLI data breaches due to shocks of Bitcoin-related variables. We note that the logarithm of the number of transactions in Bitcoins

Table 8 Variance decomposition of forecast errors of data breaches intensities

	PRC				BLI			
Days	Variance of forecast error	C (%)	P (%)	λ^{PRC} (%)	Variance of forecast error	C (%)	P (%)	λ^{BLI} (%)
1	0.0048	0.05	0.04	99.91	0.274	0.03	0.01	99.96
5	0.0101	0.36	0.04	99.59	0.357	0.36	0.04	99.60
10	0.0133	0.76	0.05	99.20	0.361	0.96	0.08	98.97
15	0.0153	0.74	0.04	99.22	0.362	1.06	0.09	98.86
30	0.0181	0.60	0.13	99.27	0.362	1.10	0.12	98.78
60	0.0195	0.73	0.81	98.45	0.362	1.11	0.17	98.72
90	0.0198	1.12	1.84	97.04	0.362	1.11	0.22	98.66
120	0.02	1.58	2.90	95.53	0.362	1.11	0.27	98.62
180	0.0203	2.42	4.65	92.93	0.362	1.12	0.34	98.54
240	0.0205	3.08	5.89	91.02	0.362	1.12	0.39	98.48
360	0.0208	4.02	7.31	88.67	0.363	1.13	0.47	98.40

has a relevant impact on the future data breaches in both datasets. The size of the response is significantly different from zero, and the phenomenon continues to persist in the long run.

The confidence intervals associated with the points estimates are very tight in the BLI dataset, indicating low variability in the estimates, less tight in the PRC dataset. Less relevant is the response with respect to unexpected shocks of the Bitcoin's price, at least until 15 days. Since after about 15 days the confidence interval is above the zero line, the impact of a shock of the Bitcoin's price on data breaches becomes positive and significant in the case of the PRC dataset. On the other hand, a shock of the Bitcoin's price has a negative and significant impact after about 15 days on the data breaches from the BLI dataset. The effect of a shock of Bitcoin's price starts declining after 10 days and almost vanishes by the 30th day in both datasets.

3.6 Variance decomposition of forecasting errors

In this section, we offer further details on the contribution of each Bitcoin-related variable to the forecast power of the estimated VAR model.[12]

The variance of forecast error after h steps for variable i, $\omega^2_{h;i}$, is defined as the element in position i in the main diagonal of the matrix $\sum_{k=0}^{h} \boldsymbol{\Phi}_k \boldsymbol{\Phi}'_k$. The contribution of variable j to the variance of forecast error after h steps for variable i is calculated as $\text{VDFE}_{h;i,j} = \frac{\sum_{k=0}^{h}(\phi_{k;i,j})^2}{\omega^2_{h;i}}$, where $\phi_{k;i,j}$ is the element in the ith row and jth column of matrix $\boldsymbol{\Phi}_k$.

Table 8 presents the variance decomposition of forecasting errors (VDFE) associated with the VECM estimations presented in Tables 3 and 5. VDFE is a classical tool

[12] We have performed a complete variance decomposition analysis of the forecasts errors, showing also the contribution of data breaches to the forecast power of the VAR model. However, to keep the paper in focus, we omit the presentation of such results. We make this additional material available upon request.

used by econometricians to understand the impact of exogenous shocks of a given (independent) variable on the forecasting errors of a different (dependent) variable. In other words, VDFE helps us understand the contribution of Bitcoin-related variables in the forecasting power of our model for data breaches.

Looking at the results presented in Table 8, we see that, in the PRC dataset, the contribution of number of transactions in Bitcoin in explaining the variance of forecasting errors of data breaches is irrelevant when the forecasting horizon is short (up to 5 days). However, for forecasting horizon greater than 5 days an exogenous shock in the independent variable is able to explain up to 4% of the variance of the forecasting vector. The contribution of Bitcoin's price seems to be to be even more important, as it can explain up to 7.3% of the total variance. In total, for the PRC dataset, Bitcoin metrics are able to explain more than 11.3% of the total variance of the forecasting errors. The VDFE of the BLI dataset displays a reduced relevance of Bitcoin metrics. The number of transactions in Bitcoins is able to explain at most 1.1% of the entire variance and Bitcoin's prices only up to 0.47%.

4 Concluding remarks

In this paper, we uncover the strong, bidirectional, relation between data breaches and Bitcoin-related variables. Our analysis suggests that in the short run the lagged values of the number of transactions in Bitcoin have a strong negative impact on data breaches today. In the long run, the existence of a cointegrating vector including all variables under investigation implies that the short-run relation will persist in the long run. Moreover, we find almost identical results on two different datasets, confirming the robustness of our result. The impulse response analyses highlight the relevant quantitative impact of both the number of transactions and the Bitcoin price on future data breaches, while the variance decomposition of the forecasting errors suggests that the same variable can explain up to 5% of the variability.

Our results might open up new research directions. First, on the econometric side, a deeper understanding of the relation between cyber risk and cryptocurrencies might be helpful. Indeed, one might wonder whether the relations found in this paper extend to other class of cyber risk and to different cryptocurrencies. With the availability of new datasets of cyber attacks, such analyses are going to become feasible in the near future. Second, on the actuarial side, the next step is to create a risk model for data breaches which includes exogenous factors as cryptocurrency-related variables. This might improve the forecasting procedures and give insurers a better understanding of cyber risk. An analysis of the impact of our findings on classical actuarial measures might also be of great interest. On the other hand, the relevance of cryptocurrencies in an international financial context is increasing. In this scenarios, understanding the connections among cryptocurrencies, macroeconomic variables and other factors such as data breaches is surely an interesting point to further explore.

Acknowledgements We wish to thank the Lead Guest Editor, Marcello Galeotti, and two anonymous referees for very useful comments which have helped to improve the paper. The usual disclaimer applies.

Funding Open access funding provided by Università della Calabria within the CRUI-CARE Agreement.

Additional details of VECM estimation

See Tables 9, 10, 11 and 12.

Table 9 Detailed statistics for VECM: BLI dataset

	ΔC_t	ΔP_t	$\Delta B_t^{\mathrm{BLI}}$
Mean dependent var	0.000	0.001	0.000
S.D. dependent var	0.061	0.126	0.653
Sum of squared residuals	3.770	19.250	153.115
S.E. of regression	0.043	0.098	0.276
R^2	0.507	0.402	0.823
Adjusted R^2	0.501	0.395	0.821
$\hat{\rho}$	−0.001	−0.024	0.001
Durbin–Watson statistics	2.000	2.048	1.995

Log-likelihood = 5,094; AIC = −4.95; BIC = −4.742; HQC = −4.874

Table 10 Cross-section covariance matrix: BLI dataset

	ΔP_t	ΔC_t	$\Delta \lambda_t^{\mathrm{BLI}}$
ΔP_t	0.0018581	0.000338	0.000333
ΔC_t	0.0003377	0.009493	−0.000852
$\Delta \lambda_t^{\mathrm{BLI}}$	0.0003331	−0.00085	0.075501

Table 11 Detailed statistics for VECM: PRC dataset

	ΔP_t	ΔC_t	$\Delta \lambda_t^{\mathrm{PRC}}$
Mean dependent var	−0.000069	0.001	−0.000283
S.D. dependent var	0.061	0.126	0.774
Sum of squared residuals	3.778	19.043	181.849
S.E. of regression	0.043	0.097	0.301
R^2	0.506	0.408	0.850
Adjusted R^2	0.500	0.401	0.849
$\hat{\rho}$	−0.001111	−0.018177	−0.011623
Durbin–Watson statistics	1.999	2.036	2.022

Log-likelihood = 4940.226; AIC = −4.796; BIC = −4.588; HQC = −4.720

Table 12 Cross-section covariance matrix: PRC dataset

	ΔP_t	ΔC_t	$\Delta\lambda_t^{\text{PRC}}$
ΔP_t	0.00186	0.00034	0.00027
ΔC_t	0.00034	0.00939	−0.00271
$\Delta\lambda_t^{\text{PRC}}$	0.00027	−0.00271	0.08963

References

Abhyankar, A.: Linear and nonlinear granger causality: Evidence from the UK stock index futures market. J. Futur. Mark. (1986-1998) **18**(5), 519 (1998)

Campbell, J.Y., Shiller, R.J.: Cointegration and tests of present value models. J. Polit. Econ. **95**(5), 1062–1088 (1987)

Chan, K.: A further analysis of the lead-lag relationship between the cash market and stock index futures market. Rev. Financ. Stud. **5**(1), 123–152 (1992)

Chen, B.: Long-run purchasing power parity: evidence from some European monetary system countries. Appl. Econ. **27**(4), 377–383 (1995)

Ciner, C.: Dynamic linkages between international bond markets. J. Multinatl. Financ. Manag. **17**(4), 290–303 (2007)

Edwards, B., Hofmeyr, S., Forrest, S.: Hype and heavy tails: A closer look at data breaches. J. Cybersecur. **2**(1), 3–14 (2016)

Eling, M., Loperfido, N.: Data breaches: Goodness of fit, pricing, and risk measurement. Insur. Math. Econ. **75**, 126–136 (2017)

Eling, M., Wirfs, J.: What are the actual costs of cyber risk events? European J. Operat. Res. **272**(3), 1109–1119 (2019)

Elliott, G., Rothenberg, T.J., Stock, J.H.: Efficient tests for an autoregressive unit root. Econometrica **64**(4), 813–836 (1996)

Farkas, S., Lopez, O., Thomas M.: Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance pricing and reserving (2019). https://hal.archives-ouvertes.fr/hal-02118080

Fokianos, K., Tjøstheim, D.: Log-linear poisson autoregression. J. Multivar. Anal. **102**(3), 563–578 (2011)

Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969)

Granger, C.W.: Some recent development in a concept of causality. J. Econom. **39**(1–2), 199–211 (1988)

Hakkio, C.S., Rush, M.: Cointegration: how short is the long run? J. Int. Money and Finance **10**(4), 571–581 (1991)

Johansen, S.: Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. Econometrica **59**(6), 1551–1580 (1991)

Lütkepohl, H.: New Introduction To Multiple Time Series Analysis. Springer, Berlin (2006)

Lütkepohl, H., Reimers, H.-E.: Impulse response analysis of cointegrated systems. J. Econ. Dyn. Control **16**(1), 53–78 (1992)

Maillart, T., Sornette, D.: Heavy-tailed distribution of cyber-risks. European Phys. J. B **75**(3), 357–364 (2010)

Marotta, A., Martinelli, F., Nanni, S., Orlando, A., Yautsiukhin, A.: Cyber-insurance survey. Comput. Sci. Rev. **24**, 35–61 (2017)

Meng, H., Xu, H.-C., Zhou, W.-X., Sornette, D.: Symmetric thermal optimal path and time-dependent lead-lag relationship: novel statistical tests and application to uk and us real-estate and monetary policies. Quant. Finance **17**(6), 959–977 (2017)

Mills, T.C., Mills, A.G.: The international transmission of bond market movements. Bull. Econ. Res. **43**(3), 273–281 (1991)

Niu, G., Melenberg, B.: Trends in mortality decrease and economic growth. Demography **51**(5), 1755–1773 (2014)

Pippenger, M.K.: Cointegration tests of purchasing power parity: the case of Swiss exchange rates. J. Int. Money and Finance **12**(1), 46–61 (1993)

Shea, G.S.: Benchmarking the expectations hypothesis of the interest-rate term structure: An analysis of cointegration vectors. J. Bus. Econ. Stat. **10**(3), 347–366 (1992)

Sims, C.A.: Money, income, and causality. Am. Econ. Rev. **62**(4), 540–552 (1972)
Tse, Y.K.: Lead-lag relationship between spot index and futures price of the Nikkei stock average. J. Forecast. **14**(7), 553–563 (1995)
Weiß, C.H.: An Introduction to Discrete-valued Time Series. Wiley, New York (2018)
Wheatley, S., Hofmann, A., Sornette, D.: Data breaches in the catastrophe framework & beyond (2019) arXiv:1901.00699
Wheatley, S., Hofmann, A., Sornette D.: Addressing insurance of data breach cyber risks in the catastrophe framework. The Geneva Papers on Risk and Insurance-Issues and Practice (2020). https://doi.org/10.1057/s41288-020-00163-w
Wheatley, S., Maillart, T., Sornette, D.: The extreme risk of personal data breaches and the erosion of privacy. European Phys. J. B **89**(1), 7 (2016)
Xu, H.-C., Zhou, W.-X., Sornette, D.: Time-dependent lead-lag relationship between the onshore and offshore renminbi exchange rates. J. Int. Financ. Markets Inst. Money **49**, 173–183 (2017)
Xu, M., Schweitzer, K.M., Bateman, R.M., Xu, S.: Modeling and predicting cyber hacking breaches. IEEE Trans. Inf. Forensics Secur. **13**(11), 2856–2871 (2018)
Zhou, S.: The power of cointegration tests versus data frequency and time spans. South. Econ. J. **67**(4), 906–921 (2001)