Analysis and Forecasting of the Main Pocket Network Chains Traffic Using Statistical Tools

Pablo Frigeiro[1] and Ramiro Rodríguez Colmeiro[1]

POKTscan, Data Science Team, {pablo,ramiro}@poktscan.com
https://www.poktscan.com/

Abstract. The Pocket Network is an infrastructure layer which provides Web3 applications with access to cryptographic data structures (blockchains). As such, the health of the network should be analyzed in terms of the served blockchains. Analyzing the statistics behavior of blockchains served by Pocket provides models that show trends and the stability of the network.
By means of statistical time series analysis we fit blockchain history to a series of models which explain their evolution and characterize their behavior. Our study shows that, under the current conditions, the main network chains are actively growing and in good shape in terms of number of relays.
Finally, we intend to show the Pocket community that with a rigorous mathematical foundation we can show that the network is thriving.

Keywords: POKT network · relay traffic · statistical analysis · forecast.

1 Introduction

The Web 3.0 (or Web3) is the concept of a decentralized world wide web [2]. One of the main goals entailed in the Web3 concept is the elimination of the dominating powers which exist in current world wide web (also known as Web2), creating a blockchain-oriented structure of the internet. As of its current state, the Web3 relies on tamper resistant digital ledgers (blockchains) technology [19] to achieve its goal of maintaining a decentralized ledger of the web. However blockchain technologies are struggling with decentralization in three fonts, the *Governance* layer, the *Network* layer and the *Storage* layer [3,6,7]. One of these, the *Storage* layer, refers to the cost of storing the bockchain data. Some solutions were already proposed to reduce the size of the blockchain data and hence enable an easier access to the blockchain data [11,16]. Nevertheless running such blockchain nodes (no mater the size of the blockchain) signifies a cost for which the largest blockchains [1] provide no incentives [4].

In this scope the Pocket Network [13] seeks to incentive the servicing of different blockchains nodes and create a truly decentralized access to blockchain data without a single source of failure and reduced costs for the application

[1] Such as Bitcoin, Ethereum or Solana, to name a few.

ecosystem [4]. In order to achieve such objectives the Pocket Network should also provide high quality service for all the served blockchains to gain trust in the applications ecosystem and increase its acceptance in the blockchain community.

This work pretends to measure the acceptance and health of the Pocket Network. We try to accomplish this by measuring analyzing the currently served blockchains in terms of the evolution of network relays and the number of active nodes. We also provide some insights on the state of the industry from the perspective of a node runner participant, showing what is to be expected for current node runners and future participants. Both of these objectives are developed in a strict mathematical, impartial and scientific approach.

This paper is organized as follows: In section 1 we introduce the subject under study and the generalities of the used techniques. Then, in section 2 the data origin is presented along the analysis tools, followed by section 3 where the results of the analysis are presented. Finally in section 4 the main conclusions are drawn.

1.1 Nature of the Pocket Network relays

The relays in the Pocket Network are fundamentally network relays and as such their nature is stochastic [12]. In addition to this, the Pocket Network is a young protocol and is affected by several non-deterministic factors, including but not restraining to: Addition of new blockchains ; Changes in the protocol ; Dependence on few applications ; Cryptocurrency markets fluctuations. All this results in a large amount of variables, some endogenous and some exogenous, that cannot be measured in a systematic way. This restrings the analysis of the relay evolution by means of a regression tool over other variables. The proposed approach to model the relay data is to use the same relay data, meaning its historical values. Such model should only observe the history of the relay data including its stochastic shocks and noise and predict the futures values of the series, with a given confidence interval.

As stated before, the present work is done from a node runner perspective and for this reason the analysis of the relays data is done by observing the mean number of relays served by a given node in intervals corresponding to a Pocket Network session [5] [2]. This way we analyze not only the growth of the relays in the Pocket Network but also the response of the node runners (the growth of the available nodes) and the utilization of such nodes [3].

1.2 Stochastic Variable Forecasting

Provided that the relays in the Pocket Network have a stochastic nature, they can be described as a stochastic time series [14]. There is a large pool of methods from where to choose in order to analyze a stochastic time series, ranging from

[2] More information on the construction of the observed time series in section 2.
[3] See section 4 for a larger discussion on this subject.

simple regressions [10] to large deep neural models [17]. In this work we focus on the ARIMA models [14], due to their acceptance and relative explicability.

The ARIMA models are a composition of three elements: The Auto Regressive (AR) element ; The Integral (I) element ; The Moving Average (AM) element. The AR and MA elements can be by themselves models, i.e. an ARIMA model without the I nor MA elements will be a purely AR model. We will lightly introduce the basic notions of the ARIMA models, we refer the reader to [14,8,18,9] for more details in this subject.

Auto Regressive element: This element is by itself a model which is only based on historical data of the time series. The order "p" of this model represents the number of historical values that are used to produce the forecast. The number of historical data points to be used is defined by the correlation between the current value of the series and the historical data. For example, an AR model of order 1 can be expressed as:

$$y_t = \Delta + \Phi_1 y_{t-1} + \epsilon_t, \tag{1}$$

where y_t and y_{t-1} are the current and previous data point, Δ is the tendency of the process, Φ_1 is a real valued coefficient that weights the variable at time $t-1$ and $\epsilon_t \sim \mathcal{N}(0, \sigma)$ a random shock affecting the process at time t. The same expression can be extended to an AR model of order p. Such model will have the following expression:

$$y_t = \Delta + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + ... + \Phi_p y_{t-p} + \epsilon_t, \tag{2}$$

normally expressed in its compact notation:

$$y_t = \Delta + \Phi(L) y_t + \epsilon_t, \tag{3}$$

where $\Phi(L)$ is a lag polynomial and L is the lag operator. Ideally, the more distant the term (in time) the lesser its influence in the current value.

Moving Average element: This element is also a model by itself, it uses only the information of random shocks (white noise) affecting the series. In this case the order "q" of the model is defined by the amount of random variables used to predict the current value of the series. These variables are defined by the auto-correlation function of the current value against its historical values. An MA model of order 1 is expressed as:

$$y_t = \Delta + \epsilon_t + \Theta_1 \epsilon_{t-1}, \tag{4}$$

here Θ_1 is a real valued coefficient that weights the shocks at time $t-1$. This expression, extended for q historical points, takes the form:

$$y_t = \Delta + \epsilon_t + \Theta_1 \epsilon_{t-1} + \Theta_2 \epsilon_{t-2} + \Theta_3 \epsilon_{t-3} + ... + \Theta_q \epsilon_{t-q}, \tag{5}$$

which also has a compact notation:

$$y_t = \Delta + \Theta(L) \epsilon_t, \tag{6}$$

where $\Theta(L)$ is a lag polynomial.

Integration element: So far we have supposed that the polynomials $\Phi(L)$ and $\Theta(L)$, of the AR and MA elements/models respectively, are good predictors of the current value of the series y_t. They both describe the ARMA model, which can be expressed as:

$$y_t = \Delta + \Phi(L)y_t + \epsilon_t + \Theta(L)\epsilon_t, \tag{7}$$

or similarly as:

$$[1 - \Phi(L)]y_t = \Delta + \epsilon_t[1 + \Theta(L)]. \tag{8}$$

However this can only be true if the following holds:

- $\epsilon_t \sim \mathcal{N}(0, \sigma)$, with σ being finite.
- The process Y_t is stationary, meaning that its probability distribution does not change across time. Their variances and autocovariances are finite and do not depend on the time t, only on the lag k.

When this conditions are met the roots of the polynomial $\Phi(L)$ are outside the unit circle (roots are often complex numbers). If these conditions are violated the roots can all be inside the unit circle (explosive process) or have at least one unitary root (the process is summing, integrating, over time). The last case is also non stationary, however we can transform the series to eliminate such root, this transformation is the differentiation of the series. The transformation is defined as:

$$\Delta y_t = y_t - y_{t-1} = (1 - L)y_t. \tag{9}$$

Similarly the differentiated series Δy_t could be further differentiated obtaining $\Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1}$. The general notation is that an order d of differentiation is notated as $\Delta^d y_t = (1-L)^d y_t$. Introducing the differentiated series in equation 8 we obtain the expression for the ARIMA model with an integration of order d:

$$[1 - \Phi(L)](1 - L)^d y_t = \Delta + \epsilon_t[1 + \Theta(L)], \tag{10}$$

Now the polynomial $\Phi(L)$ is formed by the rest of the roots outside the unit circle.

Finding the order and elements of the ARIMA model: The ARIMA(p, d, q) model is defined by three parameters:

- p: The degree of the lag polynomial $\Phi(L)$. This value is determined by the Partial Auto-Correlation Function (PACF) which measures the association between elements of the series at different lags k: (y_t, y_{t-k}). This function cancels at a given lag k depending on the function, then $p < k$.
- d: The number of integrations required to obtain an stationary series. This value is obtained by using the Augmented Dickey–Fuller (ADF) test to check the series for stationarity.
- q: The degree of the lag polynomial $\Theta(L)$, obtained by means of the Auto-Correlation Function (ACF). The ACF cancels at a lag j, then $q < j$.

While elements of the $\Phi(L)$ and $\Theta(L)$ are obtained by means of a maximum likelihood estimation, not a single needs to be fitted. It can be seen that the mentioned criteria for the p, q and d parameters describe a search space rather than a single model, i.e. there are many possible ARIMA models that fulfill these constraints. To select the best model for a given series two metrics can be applied: The Akaike Information Criterion (AIC):

$$\text{AIC} = \frac{-2\mathcal{L}}{n} + \frac{2m}{n}, \tag{11}$$

where $\mathcal{L}$ is the Likelihood of the obtained model against the observed time series, m is the amount of parameters and n is the number of observations in the time series. The secon criterion is the Bayesian Information Criterion (BIC):

$$\text{BIC} = \frac{-2\mathcal{L}}{n} + \frac{m\log(n)}{n}. \tag{12}$$

The best model will then be the one with the lowest AIC and BIC.

2 Materials and Methods

In this section the data origin and shape is presented followed by a the model fitting and test procedure.

2.1 Data Origin

The data used in this study was obtained from the data *blocks* of the Pocket Network main-net ledger. The data is acquired using the POKTscan synchronization tools and stored in internal databases. The data gathered for this work ranges from $2022-04-08$ to $2022-05-09$, composing a total of $D = 30$ days, the number of days was chosen empirically. An smaller data set is gathered for testing purposes. The test data-set is composed of a total of $D' = 2$ days, ranging from $2022-05-09$ to $2022-05-11$.

Using the database the top six chains, ordered by total number of nodes is selected. These chains make up for more than the 90% of the network traffic. The selected networks are, in order of relevance [4]

1. Polygon Mainnet (Code: 0009)
2. Gnosis - xDai (Code: 0027)
3. Harmony Shard 0 (Code: 0040)
4. Fantom (Code: 0049)
5. FUSE Mainnet (Code: 0005)
6. Ethereum (Code: 0021)

Also using these databases, a series of queries are processed to obtain the following data for a given chain at a given block height:

[4] See section 3.

– Number of nodes serving the selected chain.
– Number of relays performed in the selected chain (counting only confirmed and claimed relays).

The data is summarized in chunks of $S = 4$ blocks aligned with the genesis block, according to the current session tumbling procedure (defined by the governance and found in the block data). This summation ensure that for each data point the total number of nodes was able to report the relays done. It is worth noting that nodes are able to claim relays up to 12 blocks later, however we accept this error in our calculation since we observe that most claims are done as soon as possible by the nodes.

2.2 Time Series Construction

Each point in the analyzed time series is the quotient of the number of claimed relays and the number of staked nodes:

$$y_t^c = \frac{R_t^c}{V_t^c}, \tag{13}$$

where y_t^c is the value of the time series, R_t^c is the number of relays and V_t^c the number of nodes (or validators), all evaluated at the moment t for chain c. The time series has an element for each session in the D observed days. The expected number of elements in a given series is given by:

$$\mathbb{E}(N) = \frac{D \times 24 \times 60}{T_{\text{block}} \times S} = 720, \tag{14}$$

where $T_{\text{block}} = 15$ min is the target block time. This number may change if the chain was not served at any moment in the observed period or if the block time is not met.

2.3 ARIMA Model Fitting

For each of the constructed time series an ARIMA model is fitted using the R software [15]. Each time series is tested for stationarity using the ADF test prior fitting. The best model is chosen using the AIC and BIC metrics. The resulting model is inspected by means of the residuals analysis, the error distribution histograms and the ACF values.

3 Results

This section summarizes the results of the present work. The Pocket Network observed relays by session, for the last 48 sessions, are shown in figure 1, where the top six chains are colored. These chains are the ones being analyzed in the rest of the work.

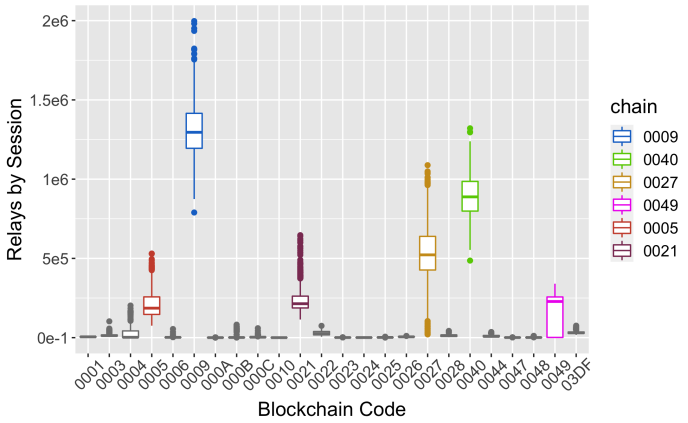

Fig. 1. BoxPlot of all the served chains of the Pocket Network. The top six chains are shown in color.

The blockchain data at the last observed session is presented in table 1.

Table 1. Data of selected blockchains for the last 48 observed sessions (48 Hs). The relays data is the summation of all the relays in the given period. The number of staked nodes and staked apps is the mean of the observed staked nodes and apps.

chain	staked nodes	staked apps	relays	Prop. of total relays [%]
Polygon Mainnet (0009)	44829	193	1.54×10^8	30.60
Gnosis - xDai (0027)	45693	234	1.25×10^8	25.00
Harmony Shard 0 (0040)	46178	223	1.05×10^8	20.80
Fantom (0049)	23322	16	3.71×10^7	7.40
FUSE Mainnet (0005)	45712	69	3.19×10^7	6.36
Ethereum (0021)	46354	673	2.66×10^7	5.33

Each of the following sub-sections contains the particular results for each of the studied blockchains.

3.1 Polygon Mainnet (0009)

The Polygon Mainnet comprises over the 30% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 2.

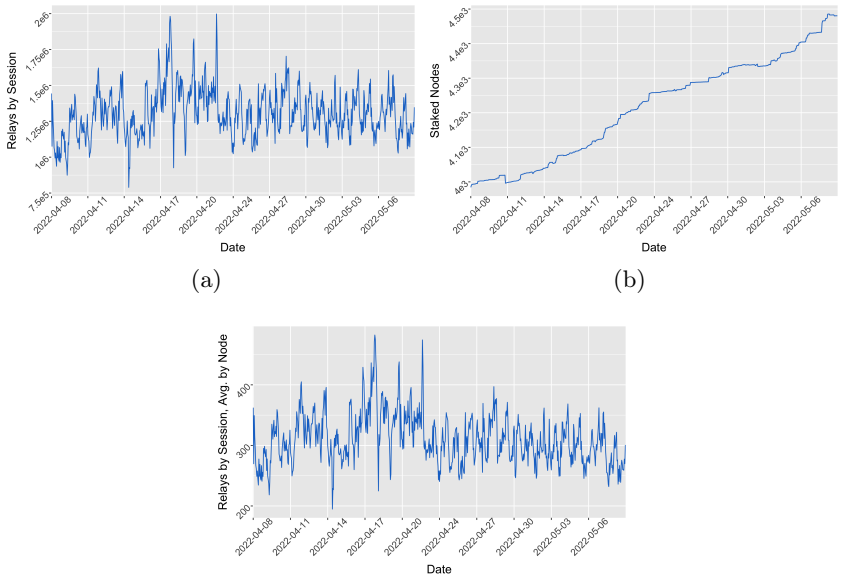

Fig. 2. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the Polygon Mainnet (0009) time series.

The best model parameters are: $p = 2$; $d = 1$; $q = 3$, with $AIC = 6649.71$ and $BIC = 6677.18$. The model coefficients are summarized in table 2.

Table 2. ARIMA model coefficients for the Polygon Mainnet (0009) time series.

ar1	ar2	ma1	ma2	ma3
1.473	−0.603	−1.778	0.999	−0.204

The residual analysis, the error distribution histograms, the PACF/ACF values and the Q-Q plots of the resulting model can be seen in figure 3. Finally the forecasting and the 95% confidence bands of the best ARIMA model is observed in figure 4.

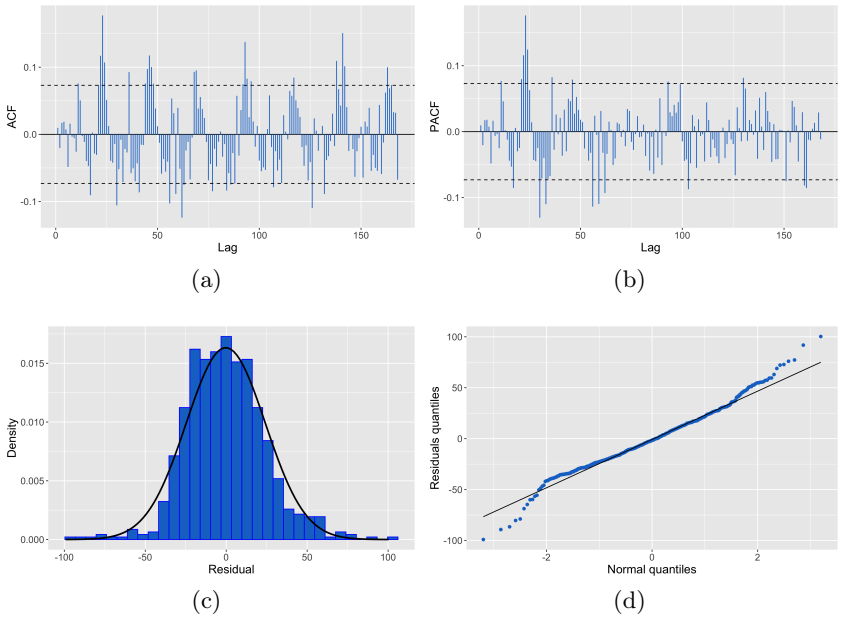

Fig. 3. Auto-Correlation Factors (a), Partial Auto-Correlation Factors (b), error histogram (c) and Q-Q plot (d) of the best ARIMA model for the Polygon Mainnet (0009) time series.

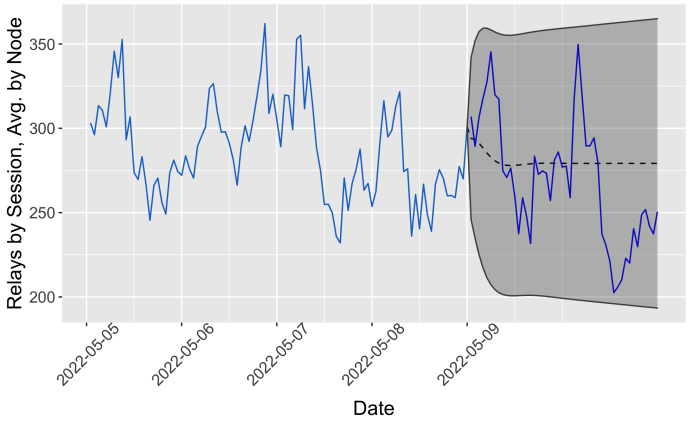

Fig. 4. Historical data of the time series (blue line) the ARIMA model forecast (black dashed line), the 95% confidence interval (grey shade) and test set data (bright blue line) for the Polygon Mainnet (0009) time series.

3.2 Gnosis - xDai (0027)

The Gnosis - xDai (0027) comprises over the 25% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 5. The best model parameters are: $p = 4$; $d = 1$

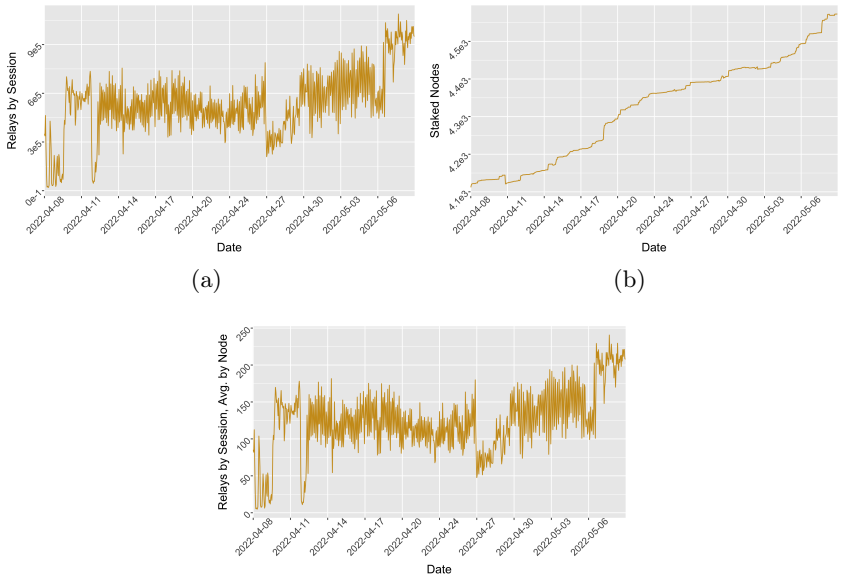

Fig. 5. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the Gnosis - xDai (0027) time series.

; $q = 1$ (with drift), with $AIC = 6437.85$ and $BIC = 6469.9$. The model coefficients are summarized in table 3. The residual analysis, the error distribution

Table 3. ARIMA model coefficients for the Gnosis - xDai (0027) time series.

ar1	ar2	ar3	ar4	ma1	drift
0.579	−0.306	0.726	−0.171	−0.981	0.168

histograms, the PACF/ACF values and the Q-Q plots of the resulting model can be seen in figure 6. Finally the forecasting and the 95% confidence bands of the best ARIMA model is observed in figure 7.

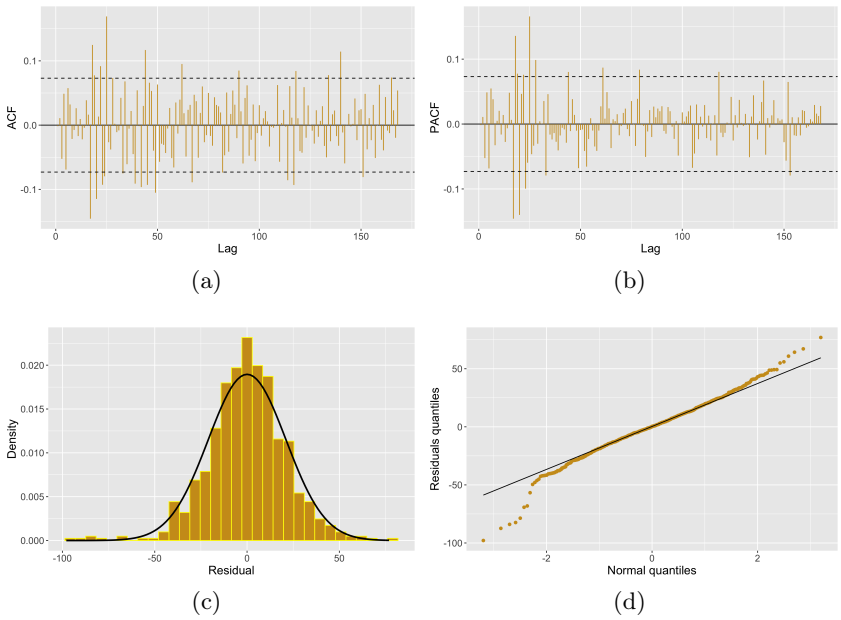

Fig. 6. Auto-Correlation Factors (a), Partial Auto-Correlation Factors (b), error histogram (c) and Q-Q plot (d) of the best ARIMA model for the Gnosis - xDai (0027) time series.

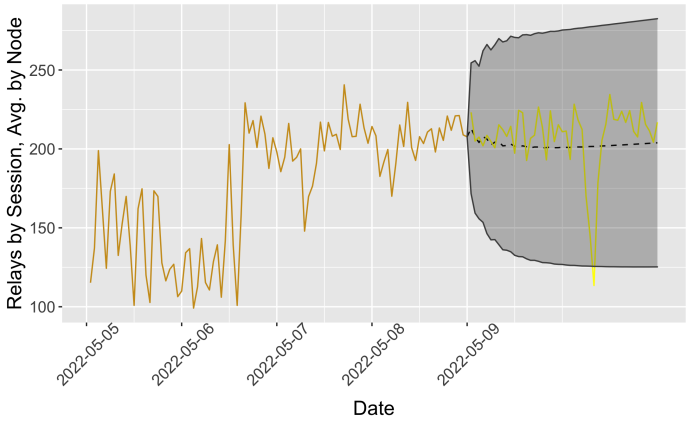

Fig. 7. Historical data of the time series (dark gold line) the ARIMA model forecast (black dashed line), the 95% confidence interval (grey shade) and test set data (yellow line) for the Gnosis - xDai (0027) time series.

3.3 Harmony Shard 0 (0040)

The Harmony Shard 0 (0040) comprises over the 20% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 8.

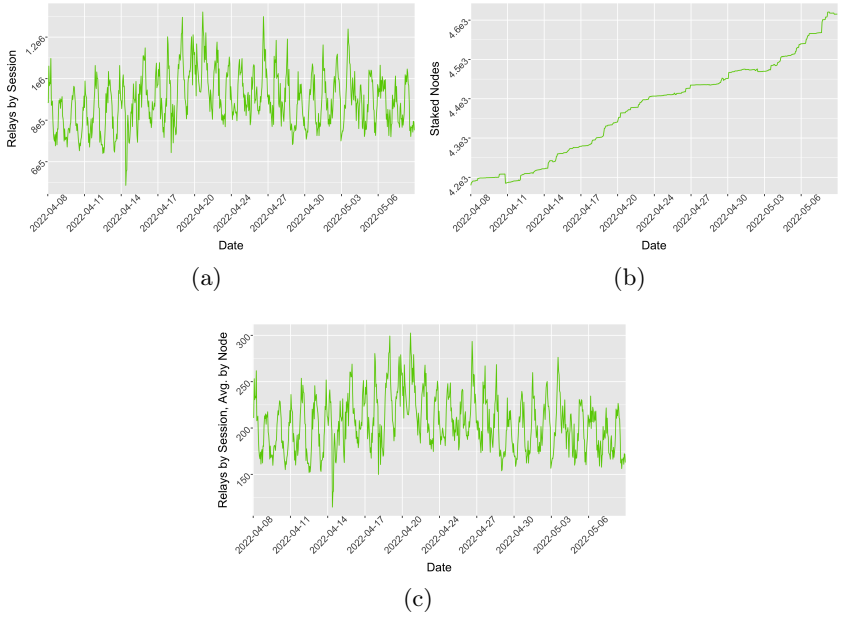

Fig. 8. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the Harmony Shard 0 (0040) time series.

The best model parameters are: $p = 1$; $d = 1$; $q = 0$, with $AIC = 6088.53$ and $BIC = 6097.68$. The model has one coefficient, $ar1 = -0.164$.

The residual analysis, the error distribution histograms, the PACF/ACF values and the Q-Q plots of the resulting model can be seen in figure 9. Finally the forecasting and the 95% confidence bands of the best ARIMA model is observed in figure 10.

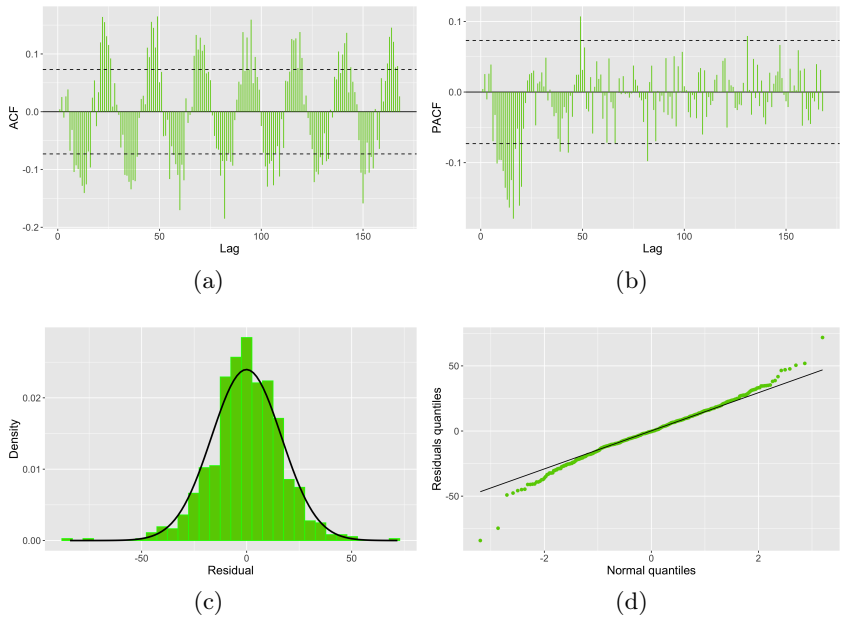

Fig. 9. Auto-Correlation Factors (a), Partial Auto-Correlation Factors (b), error histogram (c) and Q-Q plot (d) of the best ARIMA model for the Harmony Shard 0 (0040) time series.

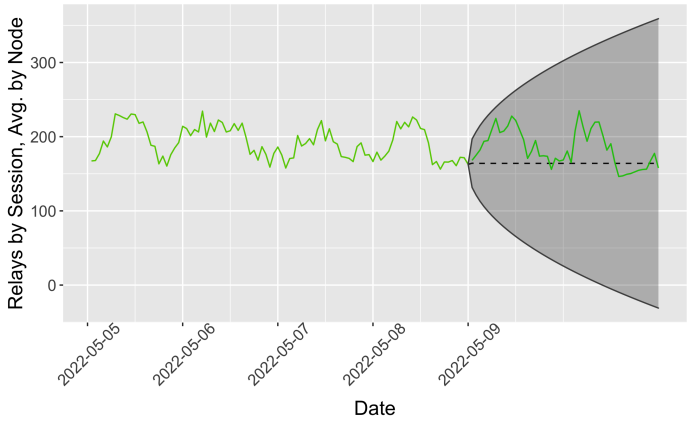

Fig. 10. Historical data of the time series green line) the ARIMA model forecast (black dashed line), the 95% confidence interval (grey shade) and test set data (bright green line) for the Harmony Shard 0 (0040) time series.

3.4 Fantom (0049)

This new supported blockchain comprises over the 7% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 11.

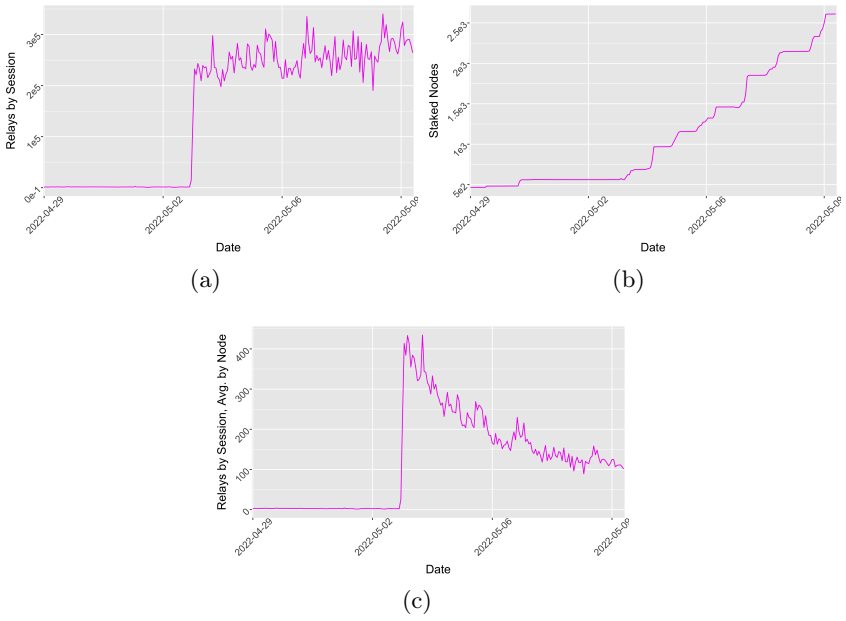

Fig. 11. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the Fantom (0049) time series.

The network is not stationary. Dickey-Fuller test value : -1.8105 ; p-value $= 0.6555$. No further analysis is done.

3.5 FUSE Mainnet (0005)

The FUSE Mainnet (0005) comprises over the 6% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 12.

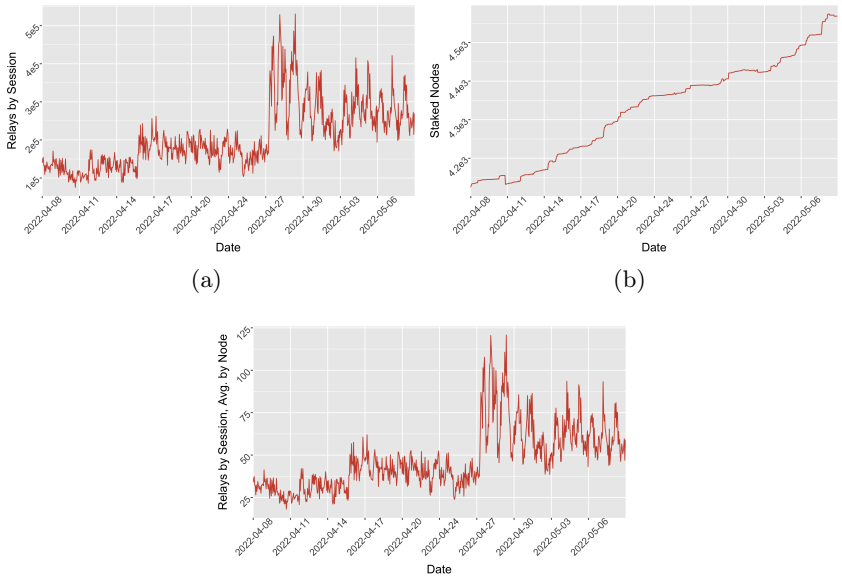

Fig. 12. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the FUSE Mainnet (0005) time series.

The best model parameters are: $p = 2$; $d = 1$; $q = 3$, with $AIC = 4959.61$ and $BIC = 4987.08$. The model coefficients are summarized in table 4. The

Table 4. ARIMA model coefficients for the FUSE Mainnet (0005) time series.

ar1	ar2	ma1	ma2	ma3
1.416	−0.572	−1.904	1.353	−0.413

residual analysis, the error distribution histograms, the PACF/ACF values and the Q-Q plots of the resulting model can be seen in figure 13.

Finally the forecasting and the 95% confidence bands of the best ARIMA model is observed in figure 14.

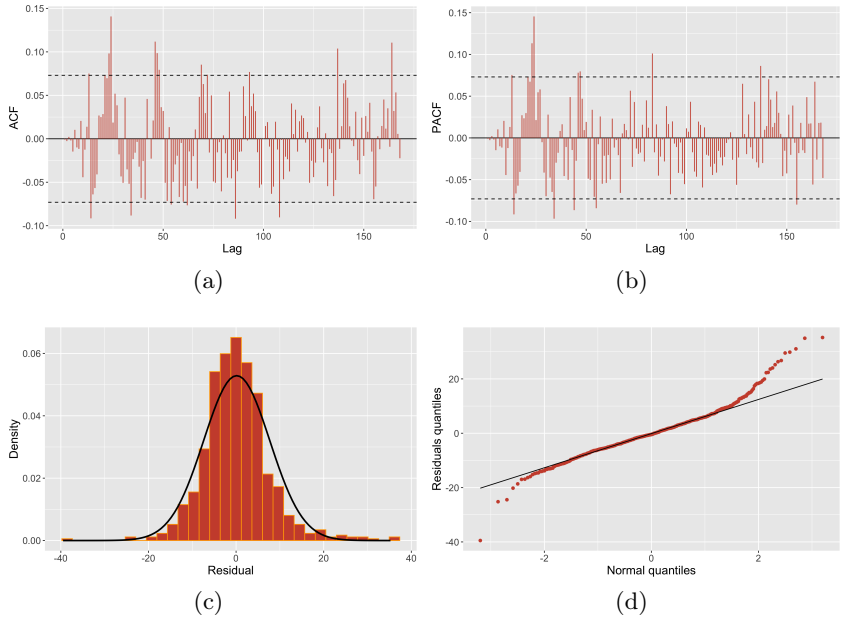

Fig. 13. Auto-Correlation Factors (a), Partial Auto-Correlation Factors (b), error histogram (c) and Q-Q plot (d) of the best ARIMA model for the FUSE Mainnet (0005) time series.

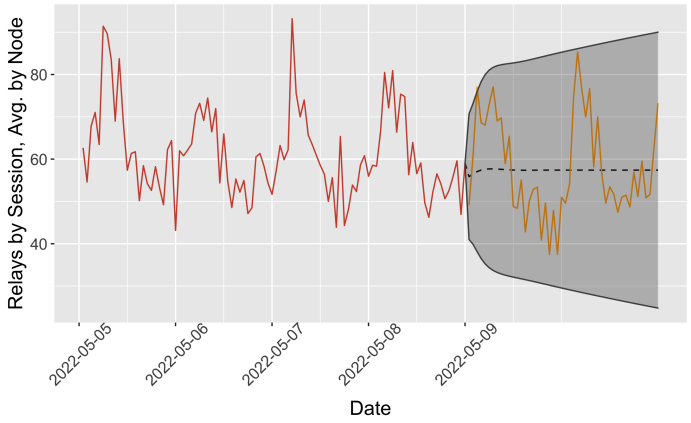

Fig. 14. Historical data of the time series (red line) the ARIMA model forecast (black dashed line), the 95% confidence interval (grey shade) and test set data (orange line) for the FUSE Mainnet (0005) time series.

3.6 Ethereum (0021)

The Ethereum (0021) comprises over the 5% of the relays. The evolution of the network relays, the staked nodes and the avg. relays by node, in the observed period, are presented in figure 15.

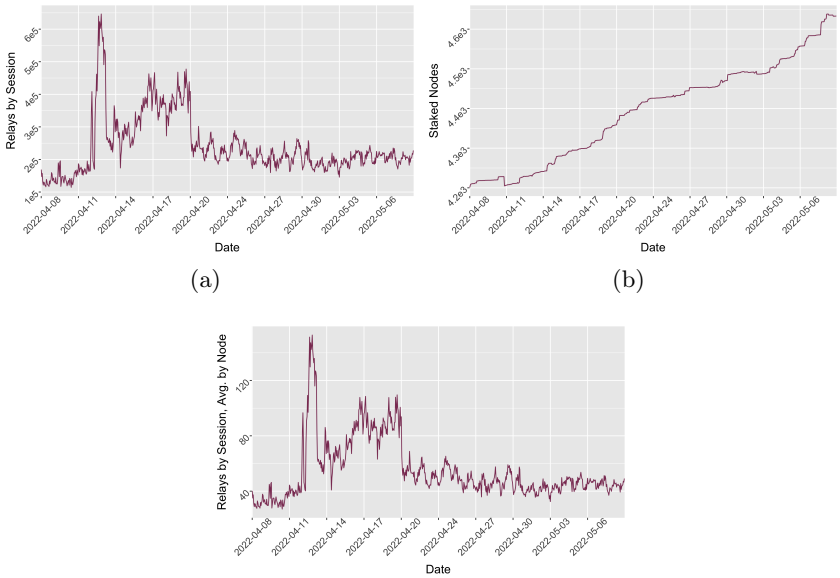

Fig. 15. Evolution of the number of relays (a), the number of staked nodes (b) and the average number of relays by node (c) in the observed time frame for the Ethereum (0021) time series.

The best model parameters are: $p = 1$; $d = 1$; $q = 1$, with $AIC = 4573.22$ and $BIC = 4586.95$. The model coefficients are summarized in table 5. The

Table 5. ARIMA model coefficients for the Ethereum (0021) time series.

ar1	ma1
0.934	−0.980

residual analysis, the error distribution histograms, the PACF/ACF values and the Q-Q plots of the resulting model can be seen in figure 16. Finally the forecasting and the 95% confidence bands of the best ARIMA model is observed in figure 17.

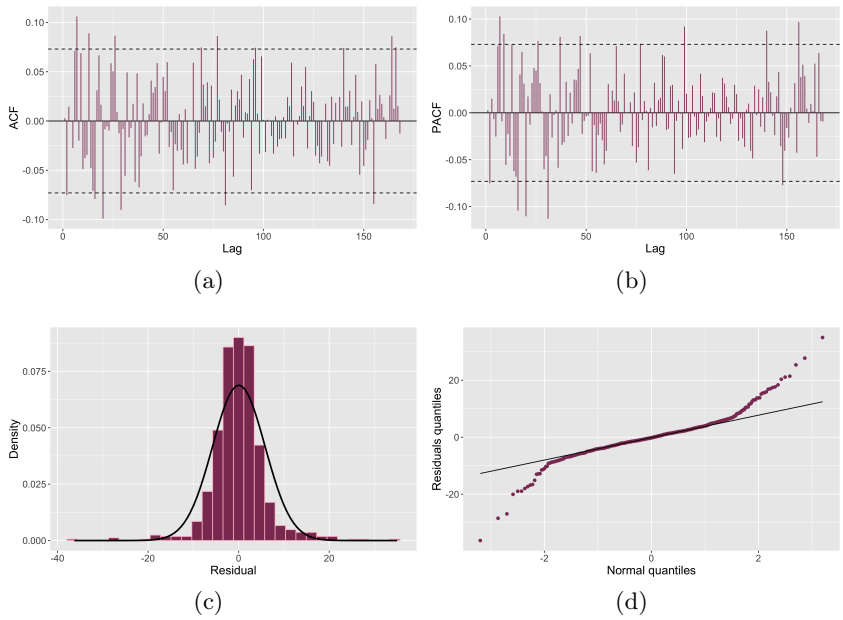

Fig. 16. Auto-Correlation Factors (a), Partial Auto-Correlation Factors (b), error histogram (c) and Q-Q plot (d) of the best ARIMA model for the Ethereum (0021) time series.

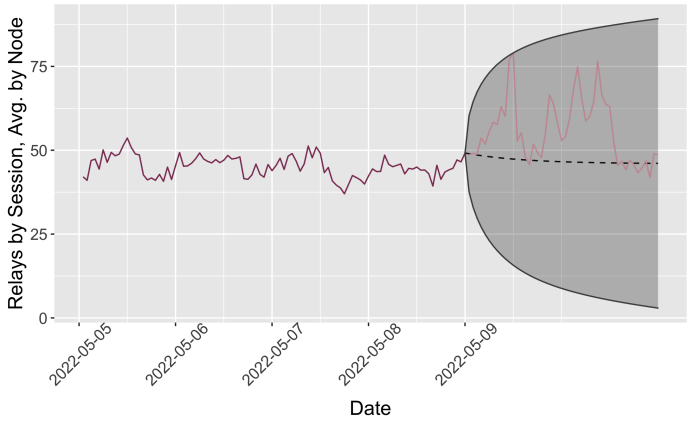

Fig. 17. Historical data of the time series (violet line) the ARIMA model forecast (black dashed line), the 95% confidence interval (grey shade) and test set data (pink line) for the Ethereum (0021) time series.

4 Discussion and Conclusions

The analysis of the average nodes by relay for each chain is successful for most of the analyzed chains, given the quality of the historical data. By quality we mean that the historical data should be:

- Long enough. Valid for all the chains except 0049, since it has less than 30 days of data.
- Stable in the sense that all external factors should remain the same (ceteris paribus). Valid for all chains except 0049, 0005 and 0021 which presented abrupt changes in the analyzed dates.

The particular findings for each chain is reported below.

4.1 Polygon Mainnet (0009)

The model of this chain presents a residuals distribution with some deviation from the expected normal distribution (figures 3(c) and 3(d)), more visible in the left side of the distribution. It has several ACFs values outside the $p - value = 0.015$ significance bands (figure 3(a)). On the other hand, the PACF values are more consistently inside the significance bands (figure 3(b)), as expected in a model with more MA elements. The overall fit is not perfect but it shows stability. The forecast mean and 95% region is correct when overlaid with the test data (figure 4). While the number of relays in the chains appears to be growing (figure 2(a)), so is the number of nodes (figure 2(b)). This is a good indicator of the application confidence and node runner confidence, respectively. Nevertheless, the downward trend of the number of relays by node by session (figure 2(c)) indicates that the number of nodes is growing faster than the number of relays. If this trend continues, the expected rewards of a node runner serving this chain will start to decrease.

4.2 Gnosis - xDai (0027)

This chain is presented an explosive growth around the $2022 - 04 - 08$, (by the start of this analysis) and still show some abrupt changes in the analyzed period (figure 5(c)). The model is not a perfect fit, it has 4 auto regressive parameters and drift (table 3). The model presents a near normal distribution but with large deviations to the left (figures 6(c)) and 6(d)). Several ACF and PACF values are outside the significance bars but present an attenuation trend in both cases (see figure 6(a) and 6(b)). The predicted values (figure 7) are inline with the test data in all but a single data sample. A larger history and stabilization of the network traffic can improve the obtained results. Nevertheless the network presents a growth of relays (figure 5(a)) higher than the growth of available nodes (figure 5(b)), which indicates that new nodes serving this blockchain will receive the same (or more) amount of traffic than current nodes [5]. The network

[5] In average terms, this really depends on the node service quality, a subject out of the scope of this work.

appears to be stabilizing in an upward trend for both app relays and serving nodes.

4.3 Harmony Shard 0 (0040)

The model presents residual distribution near the normal distribution (figures 9(c)) and 9(d)). The ACF plot (figure 9(a)) is oscillating, which expected in a purely AR model. The PACF show some values outside the significance bands but disappear quickly (figure 9(b)). Despite of having a single auto-regressive parameter, the model is a good fit. The forecast mean is correct but the model seem to be pessimistic, displaying larger possible ranges than what is observed in the test data (see figure 10). This network traffic seems to be stable and the number of nodes is meeting the demands of the blockchain, the growth of the relays is almost in balance with the growth of the nodes (see figures 8(a) and 8(b)). New nodes are expected to observe an amount of traffic similar to the historical values in this chain.

4.4 Fantom (0049)

The Fantom blockchain was added to the Pocket Network in around the $2022-05-03$ (figure 11(a)) and shows a high volume of relays despite of having only 16 staked apps (see table 1). This network does not present enough history to enable a proper analysis using the proposed tools and criteria. The time series of this blockchain is not stable and hence no model was fitted. This is a good example of the limitations of the presented method. A time series analysis will be made when the history of the blockchain data reaches the required 30 days. An interesting detail about this network is that while the number of relays is growing since $2022-05-03$ (figure 11(a)), the growth of the number of nodes is much faster (figure 11(b)), resulting in a decreasing number of relays by node by session (figure 11(c)). This shows that the blockchain is saturated with nodes, and the new nodes are seeing less traffic than the expected. This behavior is normal in young chains like this, where node runners rush to include chains and those able to set the service earlier obtain larger gains for a short period of time.

4.5 FUSE Mainnet (0005)

The resulting model presents a near-normal residual distribution with some excess kurtosis (leptokurtic) and extreme high errors to the right of figure 13(c) and 13(d), nevertheless the model shows a good fit with almost all ACF and PACF values within the limits, and those outside the significance level present an oscillating behavior (see figure 13(a) and 13(b)). The forecast of the data using this model shows a good fit, with all test data point within the expected range and a mean value in the center of the test data (figure 14). The quality of the model and the fact that the number of relays and nodes is still growing (see figures 12(a) and 12(b)) indicates that the blockchain is in good health with an stable growth of relays trend that is slightly faster than the node growth trend.

4.6 Ethereum (0021)

This blockchain presents abrupt changes along the observed period (see figure 15(a)). These changes are presumably due to external factors and affect the quality of the obtained model. The error distribution (figures 16(c) and 16(d)) is highly leptokurtic and the ACF/PCFS graphs show values outside the significance bands (see figure 16(a) and 16(b)). The forecast shows a negative bias in the mean value and a low model confidence when compared to the test data (figure 17). The model stability is not observed. Being one of the oldest supported chain in the Pocket Network the relays by node is not expected to stabilize in the near future. Deeper analysis of the behabiour of this chain is needed.

4.7 Closing Remarks

The interpretation is important when dealing with ARIMA models. Common sense should be always prime when using these models, i.e. models that predict a negative amount of relays as a possible outcome should not be used, neither models whose predicted values conditioned to the known data do not differ from those predicted without conditioning, etc. This means that the forecast ability of the created models is constrained to the near future and remote predictions should not be pretended from these models. The observed forecast ability of the models is presumably restricted around 24 to 48 sessions (with 1 session $\approx$ 1 hour). The forecast figures (4, 7, 10, 14 and 17) the forecast is shown for 48 sessions to illustrate this limit.

Overall the health of the network is good. Since we analyzed the average number of relays by node by session, the resulting models take into account the balance of the relays growth and the nodes growth. Thus we can say that the largest networks seem to be stable and growing in terms of relays by node, meaning that the relay growth is slightly higher than the node growth. However this does not mean that relays are going to saturate the network, it only shows that there is room for more nodes and that the expected number of relays by node is not decreasing. This growing trend in the number of relays by node also means that nodes are expected to work more (which in turn is an expectation of earning more at a given relays-to-token rate [6]). The analysis of the saturation of the network in terms of relays to nodes is left for future work as the information provided in this work is not enough to address the matter.

Even though the presented models are not suitable for long term predictions [7] they are good to evaluate the current state in an strict mathematical way. These models are also good predictors of the expected number of relays by node in the short term, an useful tool for a node runner wishing to know if their nodes are running correctly or not.

[6] This rate is actually subject to change due to the WAGMI proposal [1], the POKT token earn by node analysis requires to take this into account.

[7] In fact, the authors of this work believe that long term forecasting at this stage of the network is not possible.

All the data associated with this work is being released to be used by the community.

Acknowledgements We would like to thank the Pocket Network core team for their support and helpful discussions and the all the POKTscan team for their readiness and providing the necessary data.

Conflict of interest This work was founded by POKTscan, a company operating on the Pocket Network.

References

1. adam: Pup-11: Wagmi inflation (2021), https://forum.pokt.network/t/pup-11-wagmi-inflation/1369
2. Chohan, U.W.: Web 3.0: The future architecture of the internet? Available at SSRN (2022)
3. Chu, S., Wang, S.: The curses of blockchain decentralization. arXiv preprint arXiv:1810.02937 (2018)
4. Comunity, P.: Economics of pocket network (2021), https://docs.pokt.network/home/v0/economics/
5. Comunity, P.: Protocol of pocket network (2021), https://docs.pokt.network/home/v0/protocol
6. Gencer, A.E., Basu, S., Eyal, I., Renesse, R.v., Sirer, E.G.: Decentralization in bitcoin and ethereum networks. In: International Conference on Financial Cryptography and Data Security. pp. 439–457. Springer (2018)
7. Gochhayat, S.P., Shetty, S., Mukkamala, R., Foytik, P., Kamhoua, G.A., Njilla, L.: Measuring decentrality in blockchain based systems. IEEE Access **8**, 178372–178390 (2020)
8. González Casimiro, M.P.: Análisis de series temporales: Modelos arima (2009)
9. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2018)
10. Kedem, B., Fokianos, K.: Regression models for time series analysis. John Wiley & Sons (2005)
11. Kim, Y., Raman, R.K., Kim, Y.S., Varshney, L.R., Shanbhag, N.R.: Efficient local secret sharing for distributed blockchain systems. IEEE Communications Letters **23**(2), 282–285 (2018)
12. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Transactions on networking **2**(1), 1–15 (1994)
13. Correa de Leon, L.R., O'Rourke, M.P.: Pocket network (2018), https://www.pokt.network
14. Mauricio, J.A.: Análisis de series temporales. Universidad Complutence de Madrid (2007)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018), https://www.R-project.org/

16. Raman, R.K., Varshney, L.R.: Distributed storage meets secret sharing on the blockchain. In: 2018 Information Theory and Applications Workshop (ITA). pp. 1–6. IEEE (2018)
17. Torres, J.F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., Troncoso, A.: Deep learning for time series forecasting: a survey. Big Data **9**(1), 3–21 (2021)
18. Viñals, M.P.: Series temporales, vol. 64. Univ. Politèc. de Catalunya (2009)
19. Yaga, D., Mell, P., Roby, N., Scarfone, K.: Blockchain technology overview. arXiv preprint arXiv:1906.11078 (2019)