INFRACON

<Pocket Network Community Conference 2022>

Statistical Analysis of the Pocket Network

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

Pablo Frigerio

Advanced Mathematics & Statistics,
PoktScan Data Science Team

- DonPablo#1563

Ramiro Rodríguez Colmeiro

Data Analysis and Machine Learning Methods,
PoktScan Data Science Team

- RawthiL#6804
- @Rama_stdout

Motivation

- Check the state of the Pocket Network chains using an impartial and scientific criteria.
- Analyze the health and acceptance of the different supported networks.
- Measure what is to be expected for a node runner in terms of relays by node.
- Answer the question: Are my nodes running properly?
- Give some insights on the impact of new node runners in the nodes workload (that results in node gains)

INFRACON

Analysis of Network Relays

A short review of the characteristics of the Pocket Network relays

The Pocket Network relays are...

- **Heterogeneous**: The network is composed of different blockchains and each has its own traffic profile.

- **Stochastic:** It is internet traffic!

- **Multi-Causal:** The network is affected by multiple sources, most of them not mensurable, i.e. Addition of new chains, Staking/Un-staking of large applications, changes in the cryptocurrency markets, etc.

How do we analyze such traffic?

Regression model?

How do we analyze such traffic?

Regression model? 

Many independent variables

that cannot be measured..

How do we analyze such traffic?

Regression model?

Many independent variables that cannot be measured..

Modeling of the whole network?

How do we analyze such traffic?

Regression model? → Many independent variables that cannot be measured..

Modeling of the whole network? → Interesting but we lack some data and is also dependent on non-observable phenomena.

How do we analyze such traffic?

Regression model? → Many independent variables that cannot be measured..

Modeling of the whole network? → Interesting but we lack some data and is also dependent on non-observable phenomena.

Deep Models!

How do we analyze such traffic?

Regression model? → Many independent variables that cannot be measured..

Modeling of the whole network? → Interesting but we lack some data and is also dependent on non-observable phenomena.

Deep Models!

500 Billon Parameters!!

How do we analyze such traffic?

Regression model? → Many independent variables that cannot be measured..

Modeling of the whole network? → Interesting but we lack some data and is also dependent on non-observable phenomena.

Deep Models!

500 Billon Parameters!! → Few data points... explicability... stop reading cheap tech blogs...

Our approach:

ARIMA(p,d,q) Models

Our approach:

ARIMA(p,d,q) Models → Build upon historical data of the same analyzed variable.

Our approach:

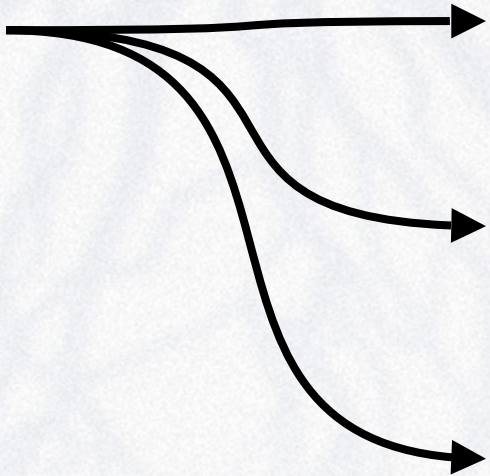
ARIMA(p,d,q) Models

Build upon historical data of the same analyzed variable.

A good fit for stochastic time series. They can model random shocks.

Our approach:

ARIMA(p,d,q) Models



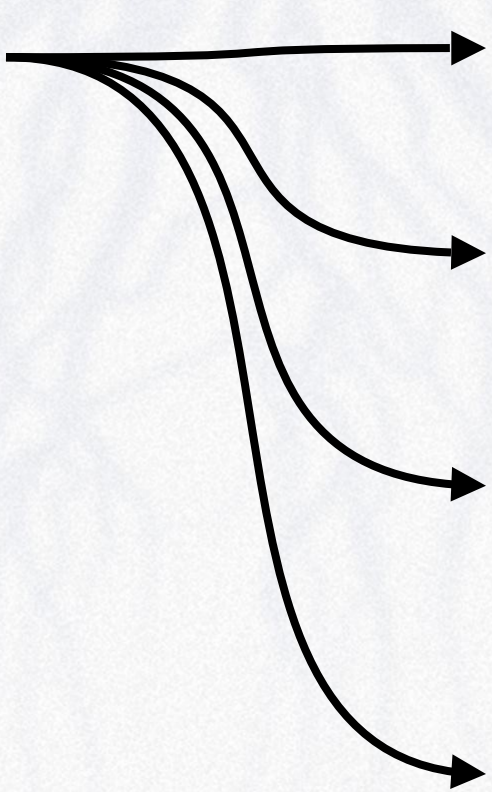
Build upon historical data of the same analyzed variable.

A good fit for stochastic time series. They can model random shocks.

Offers some degree of explicability.

Our approach:

ARIMA(p,d,q) Models



Build upon historical data of the same analyzed variable.

A good fit for stochastic time series. They can model random shocks.

Offers some degree of explicability.

Allow forecasting (with some limitations)

Our approach:

ARIMA(p,d,q) Models

$$[1-\Phi(L)](1-L)^d y_t = \delta + \epsilon_t[1+\Theta(L)]$$

Main parameters:

- **Φ(L)**: Polynomial of order “p”. Reflects the amount of steps in the past that the model observes.
- **Θ(L)**: Polynomial of order “q”. Reflects how many “random shocks” the model is including.
- **(1-L)ᵈ**: Integration element, the signal is differentiated “d” times to achieve an stationary form.

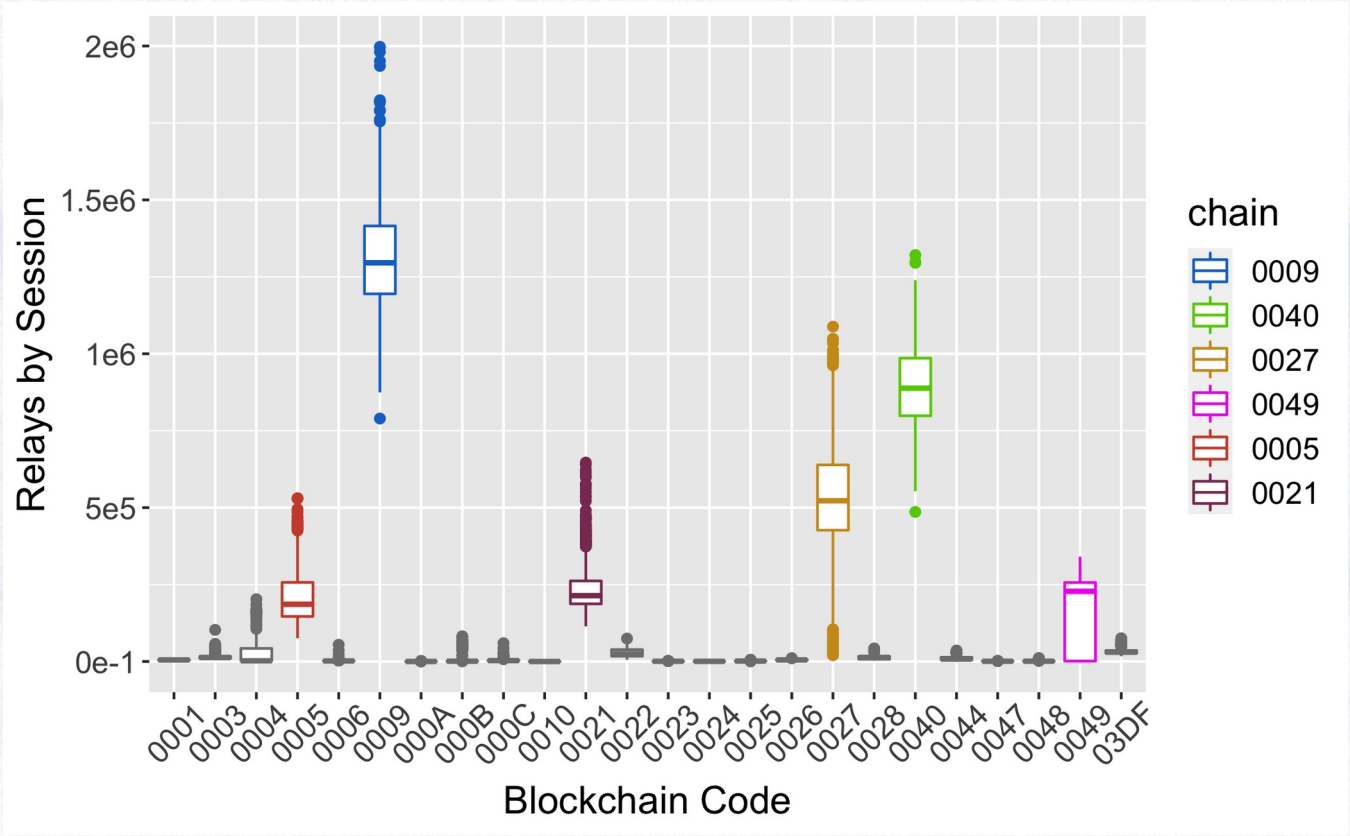
Build upon historical data of the same analyzed variable.

A good fit for stochastic time series. They can model random shocks.

Offers some degree of explicability.

Allow forecasting (with some limitations)

Our approach, also: Divide the data



(48 Hs of data, from 2022-05-07 to 2022-05-09)

Our approach, also: Divide the data

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

Results by Blockchain

Results and analysis of each network relay data

Some generalities on the analysis...

- The analyzed period is from 2022-04-08 to 2022-05-09.
- The data was obtained from the POKTScan database, whose source is the Pocket Network blocks.
- The “R” software was used to perform the computation of the presented models.
- The used data for each blockchain was obtained in the same period and analyzed using the same processing steps to allow inter-comparability.
- All the data is being released, including older/newer datasets and an extended paper.

Some generalities on the analysis...

The elements of the time series to be analyzed are:

The average number of relays served by a node during a session.

$$y_t^c = \frac{R_t^c}{V_t^c}$$

For a given session “t” and a given blockchain “c”.

Some generalities on the analysis...

The elements of the time series to be analyzed are:

The average number of relays served by a node during a session.

Element (sample)
of the time series

$$y_t^c = \frac{R_t^c}{V_t^c}$$

For a given session “t” and a given blockchain “c”.

Some generalities on the analysis...

The elements of the time series to be analyzed are:

The average number of relays served by a node during a session.

Total number of relays observed in a session

Element (sample) of the time series

$$y_t^c = \frac{R_t^c}{V_t^c}$$

For a given session “t” and a given blockchain “c”.

Some generalities on the analysis...

The elements of the time series to be analyzed are:

The average number of relays served by a node during a session.

Element (sample) of the time series

$$y_t^c = \frac{R_t^c}{V_t^c}$$

Total number of relays observed in a session

Total number of nodes staked in the blockchain

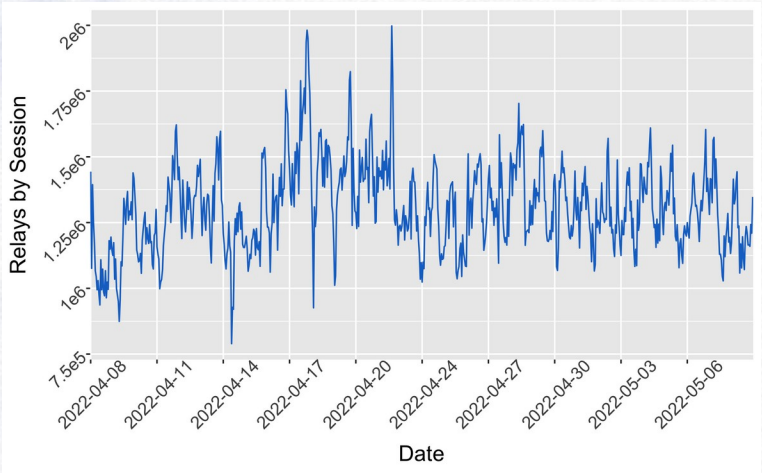
For a given session “t” and a given blockchain “c”.

Polygon Mainnet (0009)

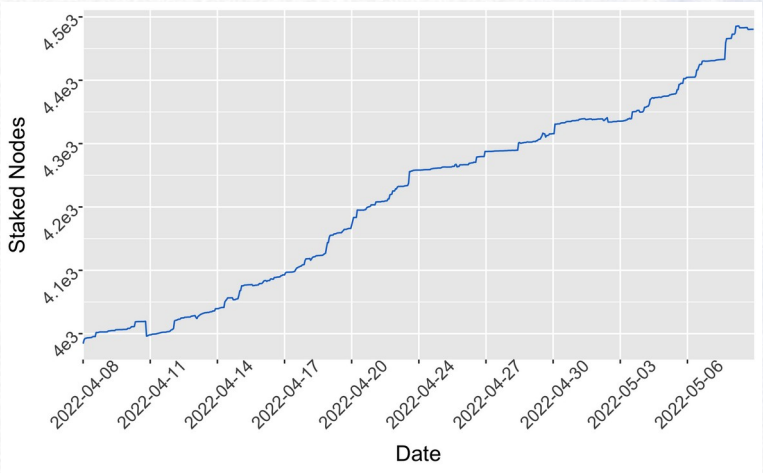
30% of the total Pocket Network relays. Stable and getting crowded.

Polygon Mainnet (0009)

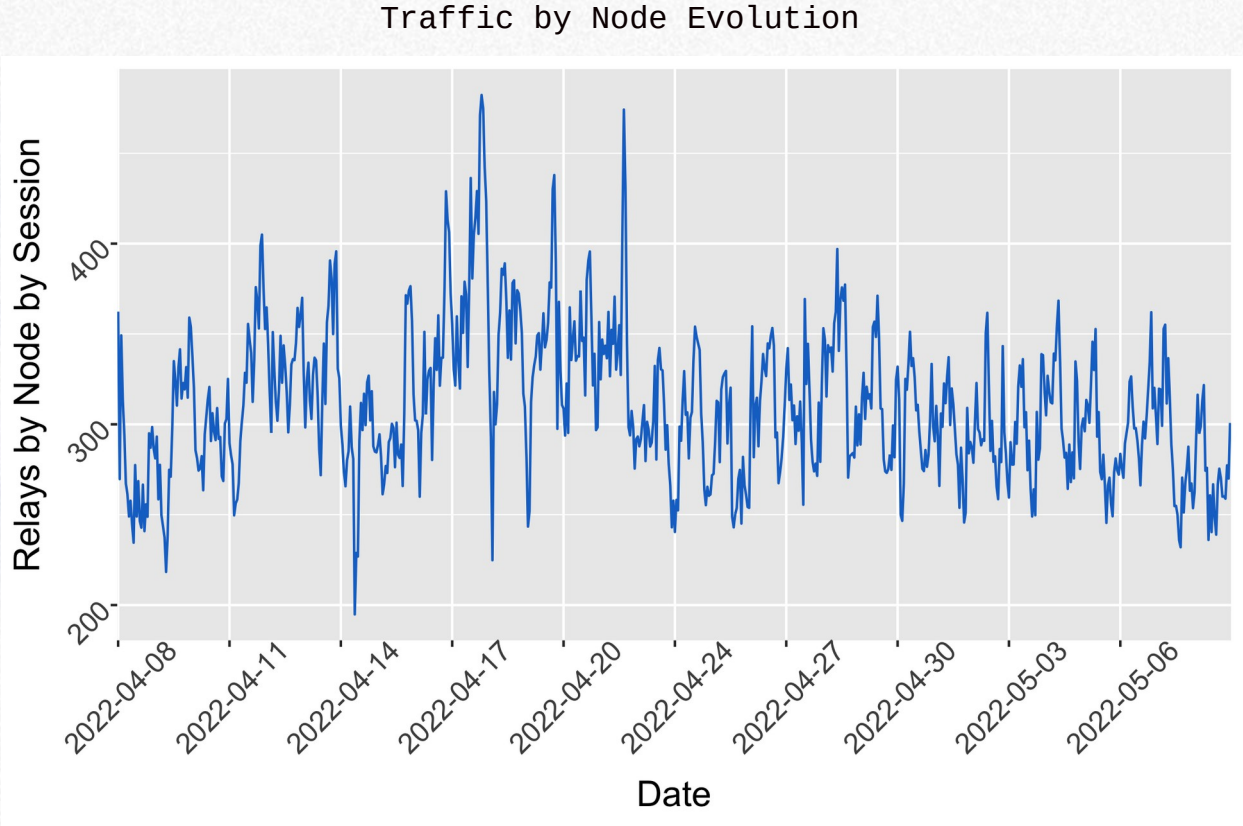
Relays Evolution



Staked Nodes Evolution



Polygon Mainnet (0009)

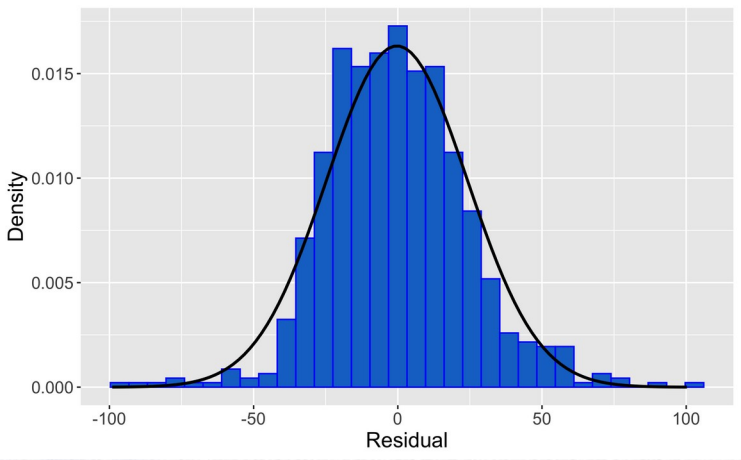


Polygon Mainnet (0009)

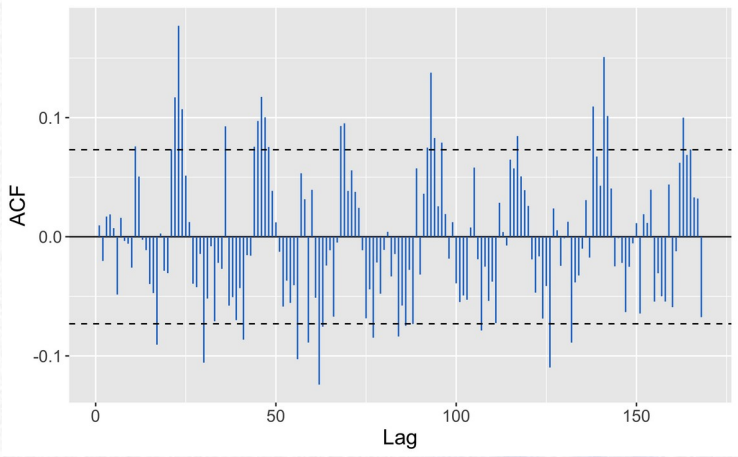
Resulting model: ARIMA(2,1,3)

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

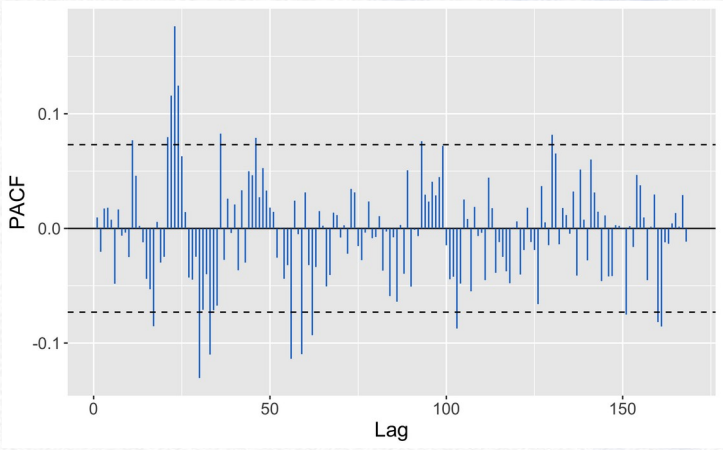
Residuals distribution



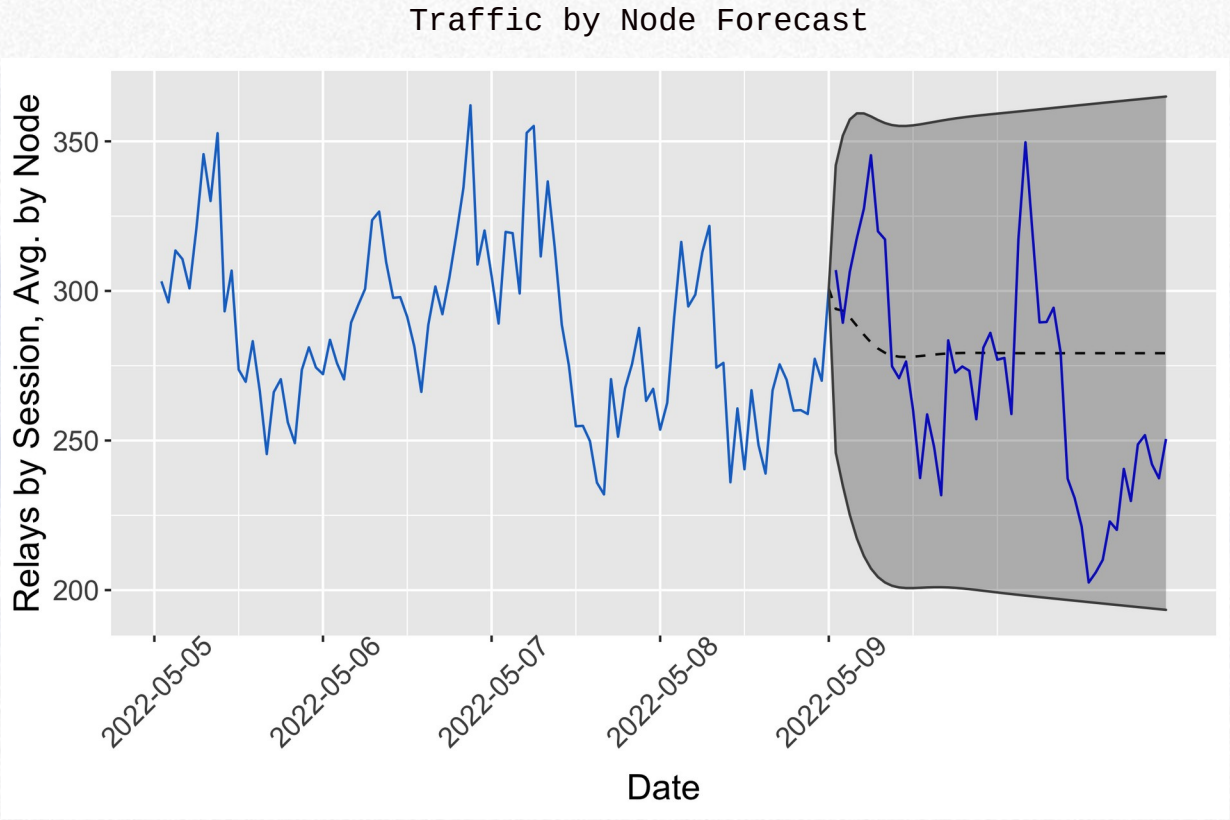
Auto-Correlation Factors



Partial Auto-Correlation Factors



Polygon Mainnet (0009)



Polygon Mainnet (0009)

Conclusions:

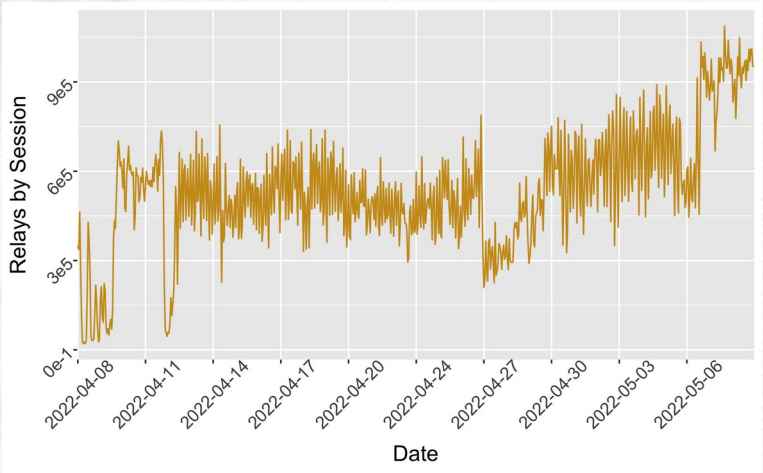
- Not a perfect model. The residual distribution deviates a little from the normal distribution and some ACFs/PACFs are outside the significance bands.
- The model forecast is correct in mean and expected ranges.
- The model shows stability, indicating a healthy (stable) network.
- Both the number of relays and number of staked nodes is growing.
- The number of nodes is growing slightly faster than the number of relays. This creates a downward trend in the number of relays by node. Less work by node is to be expected if this trend continues.

Gnosis - xDai (0027)

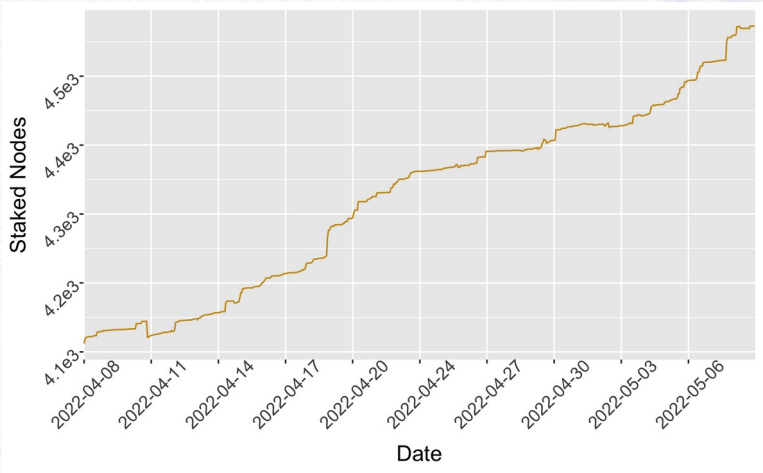
25% of the total Pocket Network relays. Recent and great growth.

Gnosis - xDai (0027)

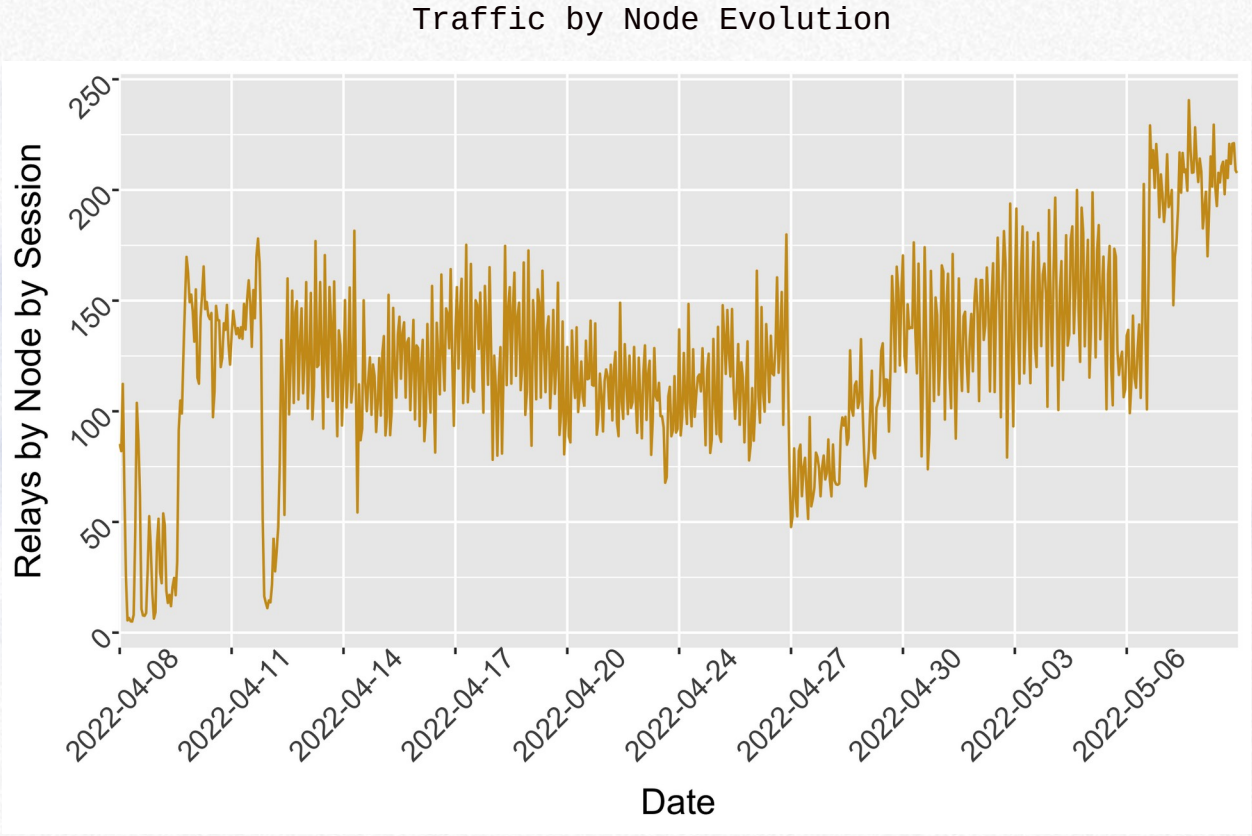
Relays Evolution



Staked Nodes Evolution



Gnosis - xDai (0027)

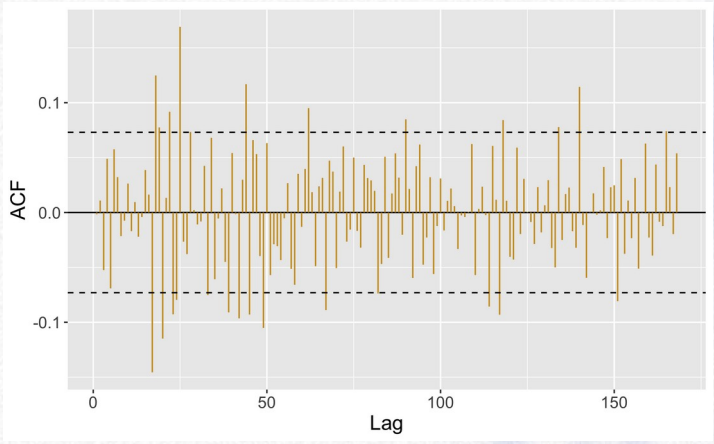


Gnosis - xDai (0027)

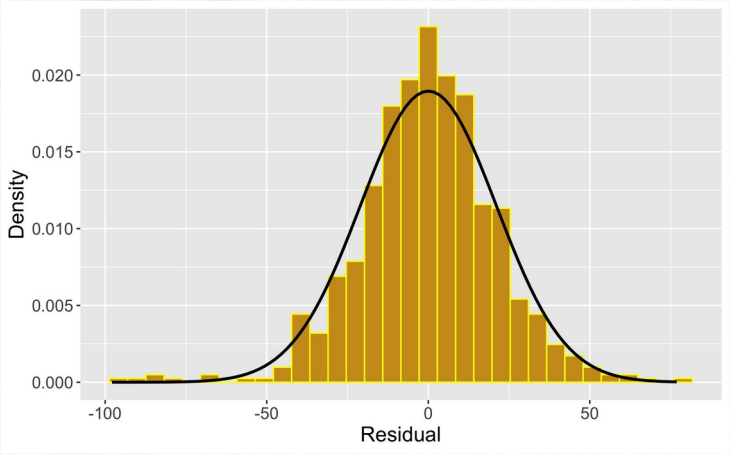
Resulting model: ARIMA(4,1,1)

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

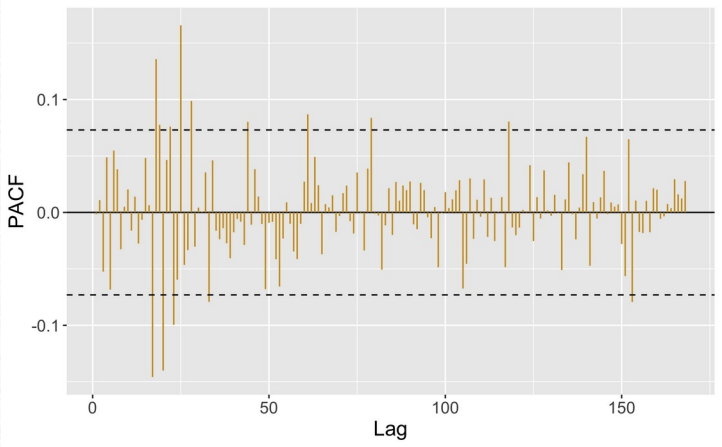
Auto-Correlation Factors



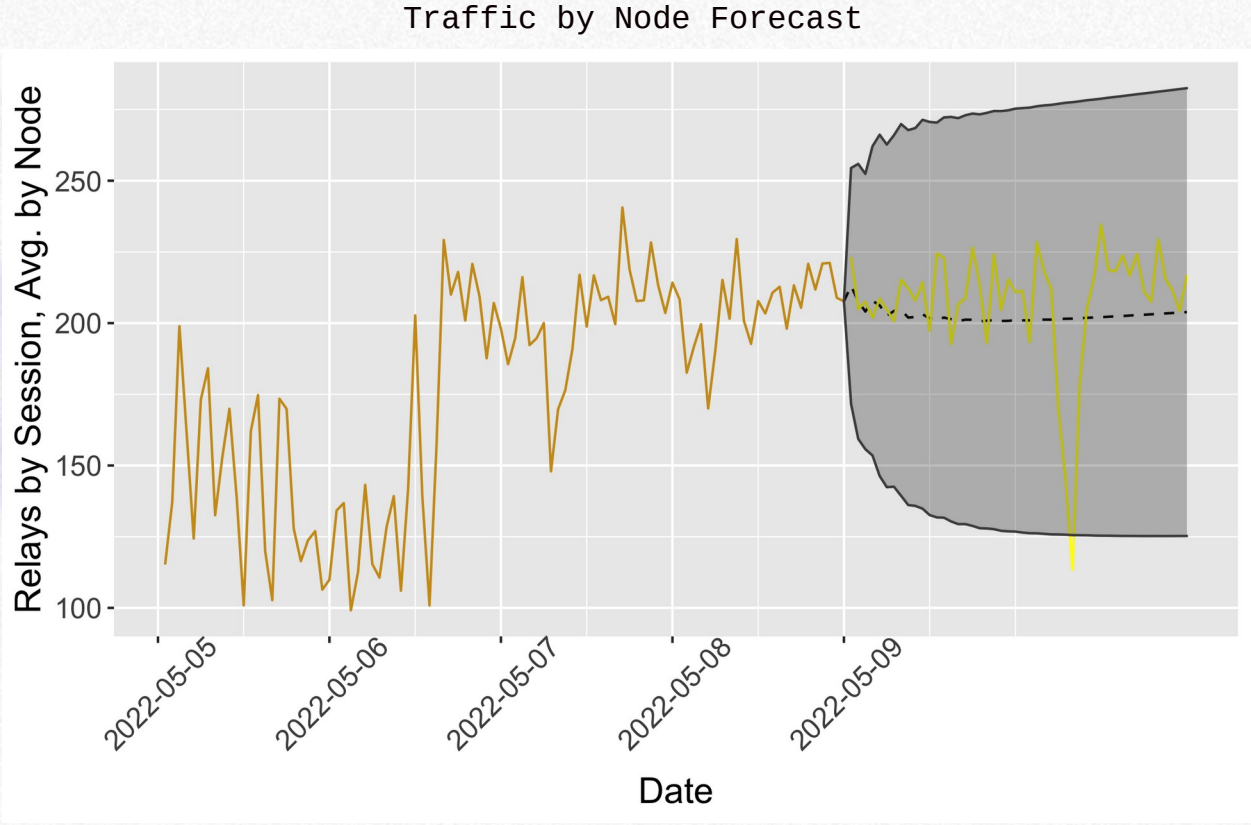
Residuals distribution



Partial Auto-Correlation Factors



Gnosis - xDai (0027)



Gnosis - xDai (0027)

Conclusions:

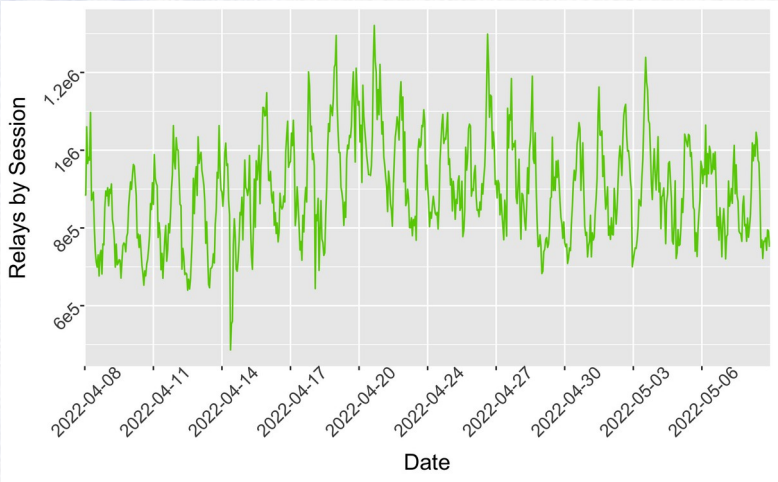
- Features several explosive growths, including one around 2022-04-08 which makes the analysis not so easy. Also the model is not a perfect fit, it has 4 auto-regressive parameters ans some ACFs/PACFs outside the significance values.
- The predicted zone is correct. Only a single sample is outside the expected range.
- Larger history is required to improve to model, but so far the network looks good.
- Both the number of relays and number of staked nodes is growing.
- The number of relays is growing faster than the number of nodes, which means that new nodes can enter the blockchains and similar or better work loads are to be expected.

Harmony Shard 0 (0040)

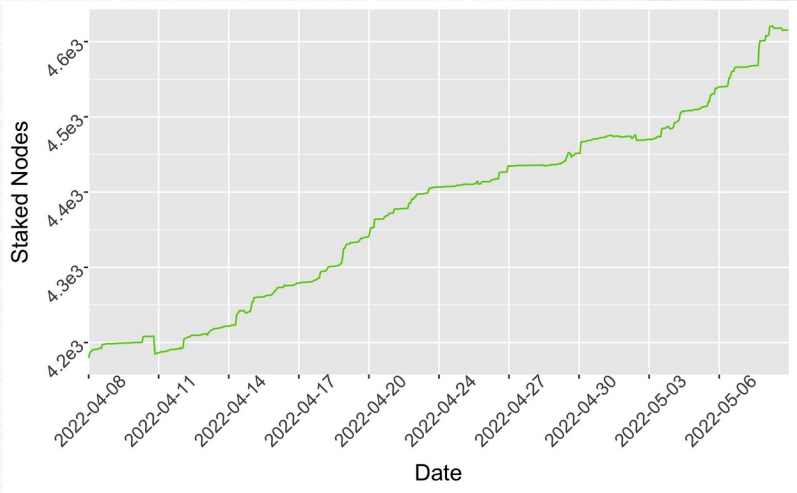
20% of the total Pocket Network relays. Old, stable and strong.

Harmony Shard 0 (0040)

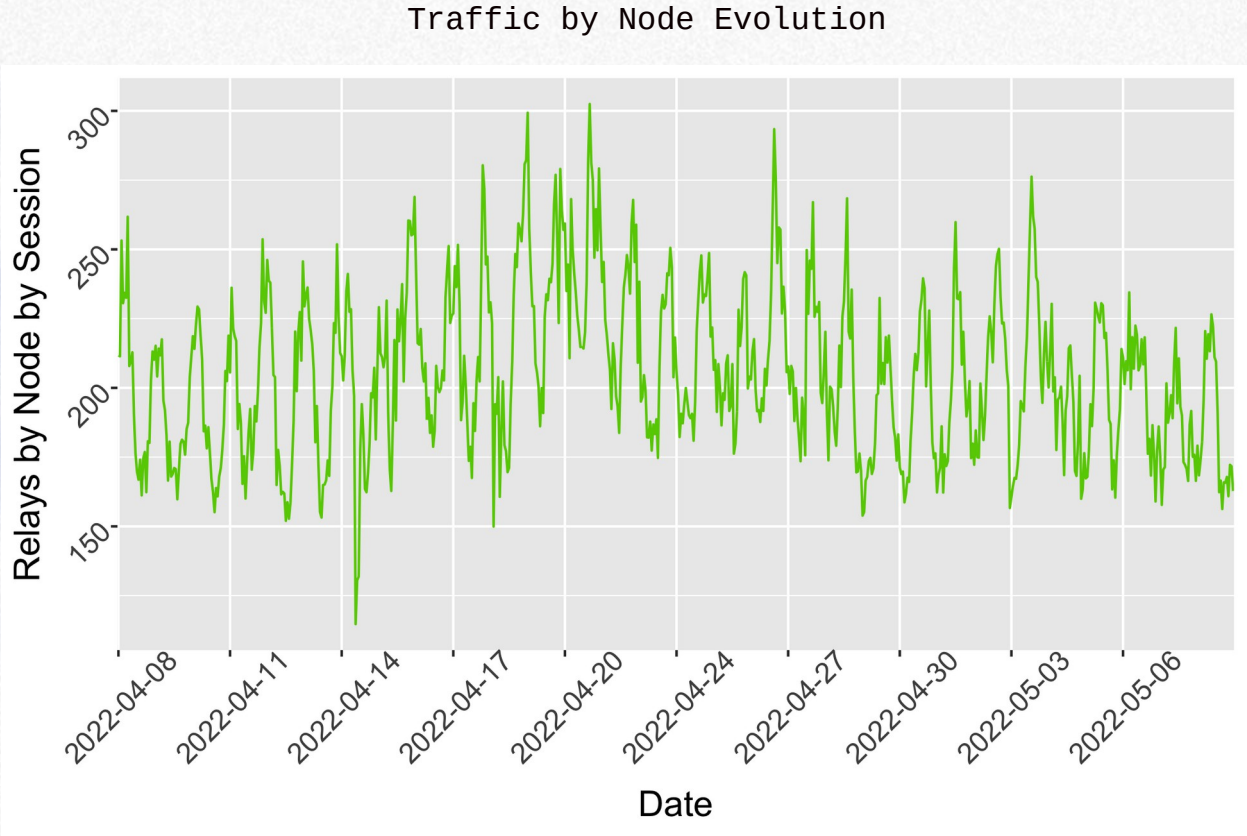
Relays Evolution



Staked Nodes Evolution



Harmony Shard 0 (0040)

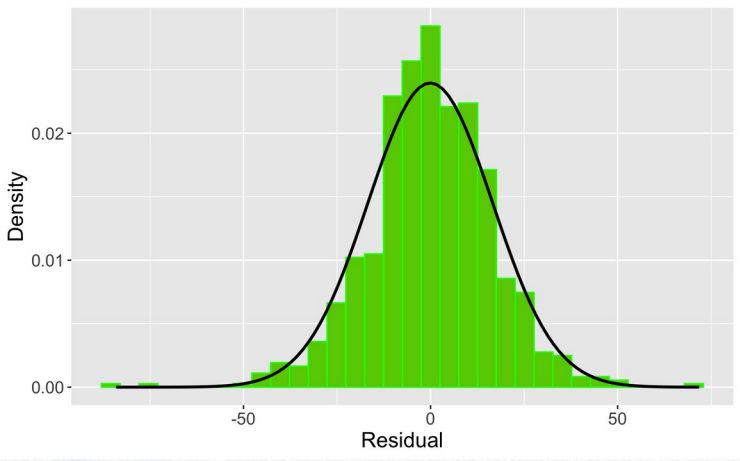


Harmony Shard 0 (0040)

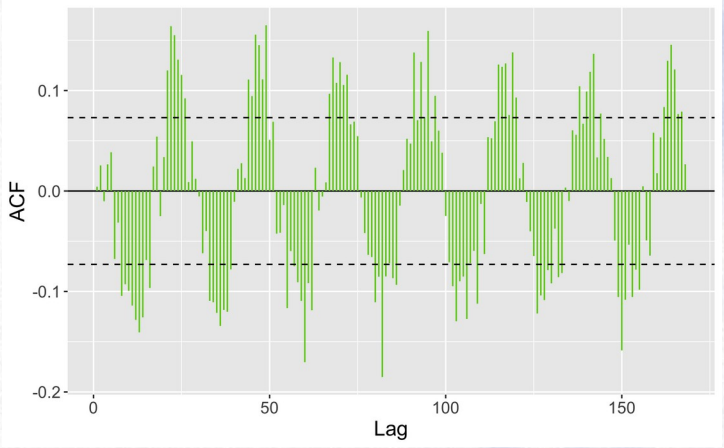
Resulting model: ARIMA(1,1,0)

ar1
−0.164

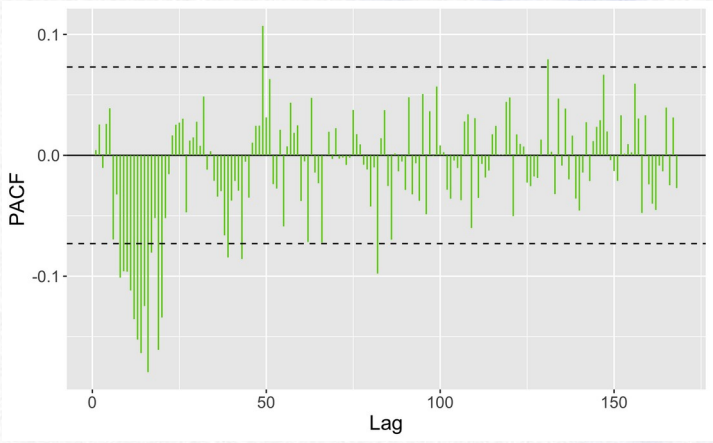
Residuals distribution



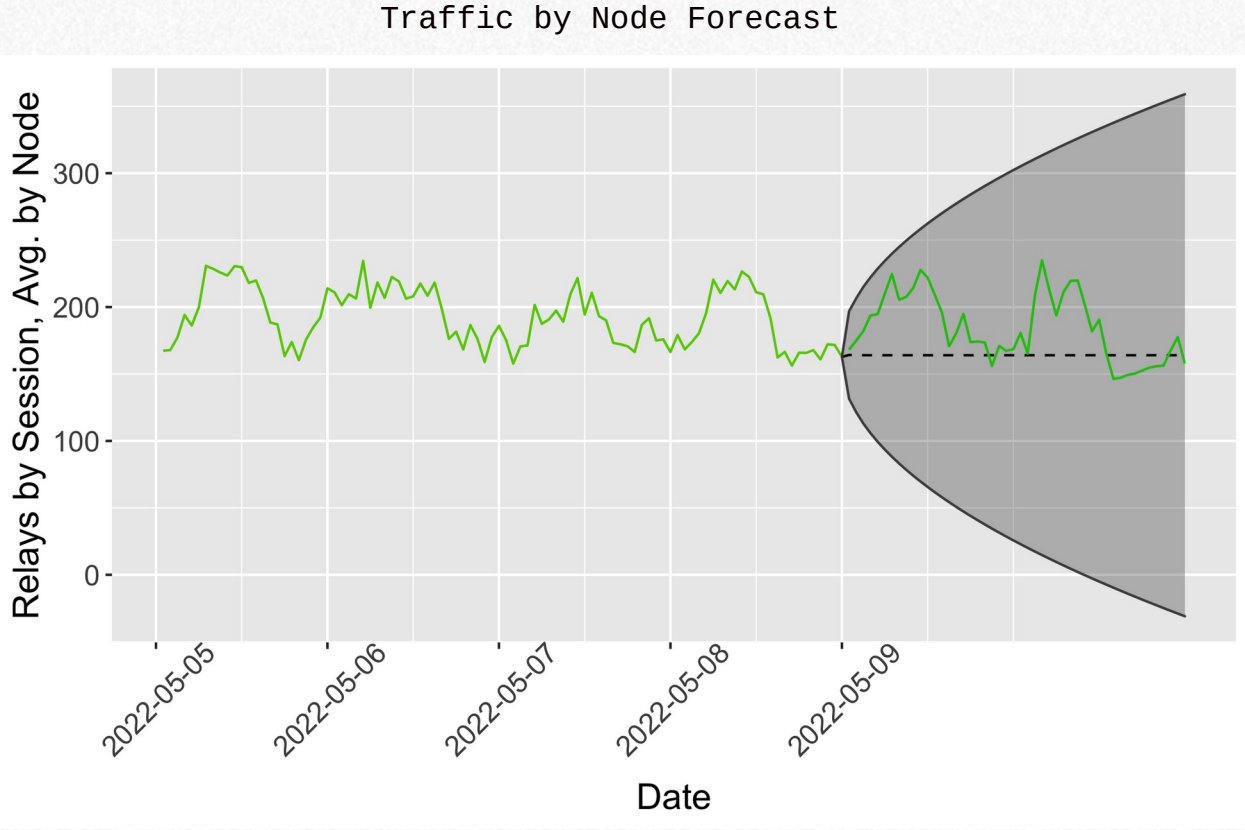
Auto-Correlation Factors



Partial Auto-Correlation Factors



Harmony Shard 0 (0040)



Harmony Shard 0 (0040)

Conclusions:

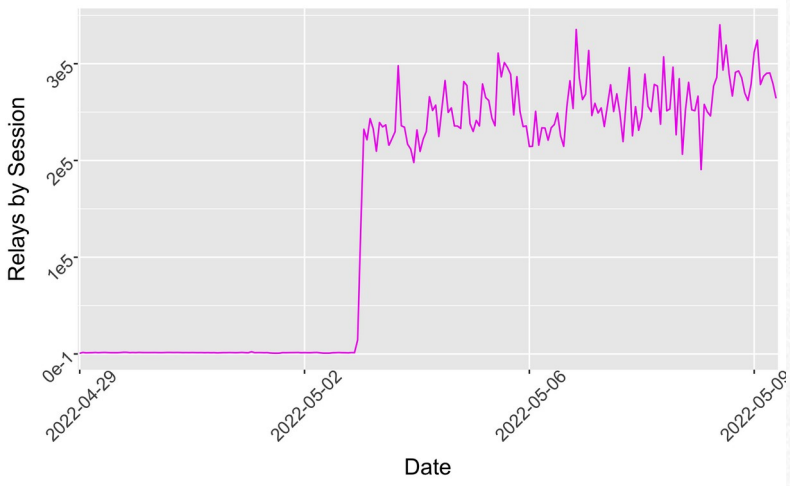
- The model presents a near-normal residual distribution with an oscillating ACFs plot but a damping PACFs plot, resulting in a good fit despite having a single auto-regressive parameter.
- The forecast mean is accurate but a little pessimistic in its ranges.
- The network is healthy in terms of relays by node stability.
- The growth of relays is in balance with the growth of staked nodes.

Fantom (0049)

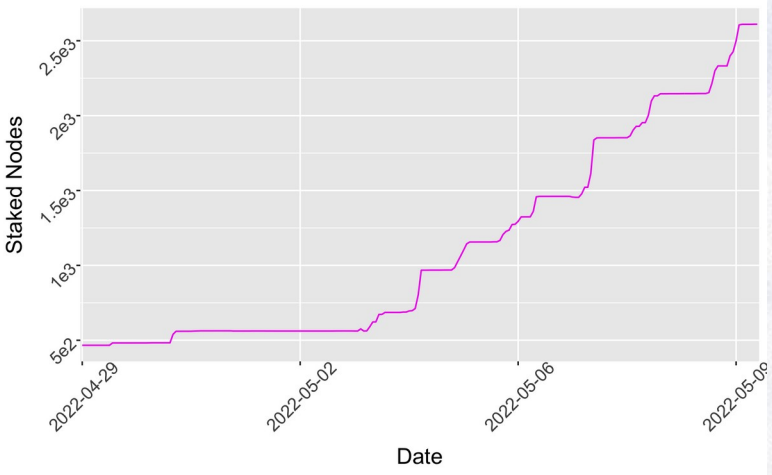
7% of the total Pocket Network relays. Nice to meet you!

Fantom (0049)

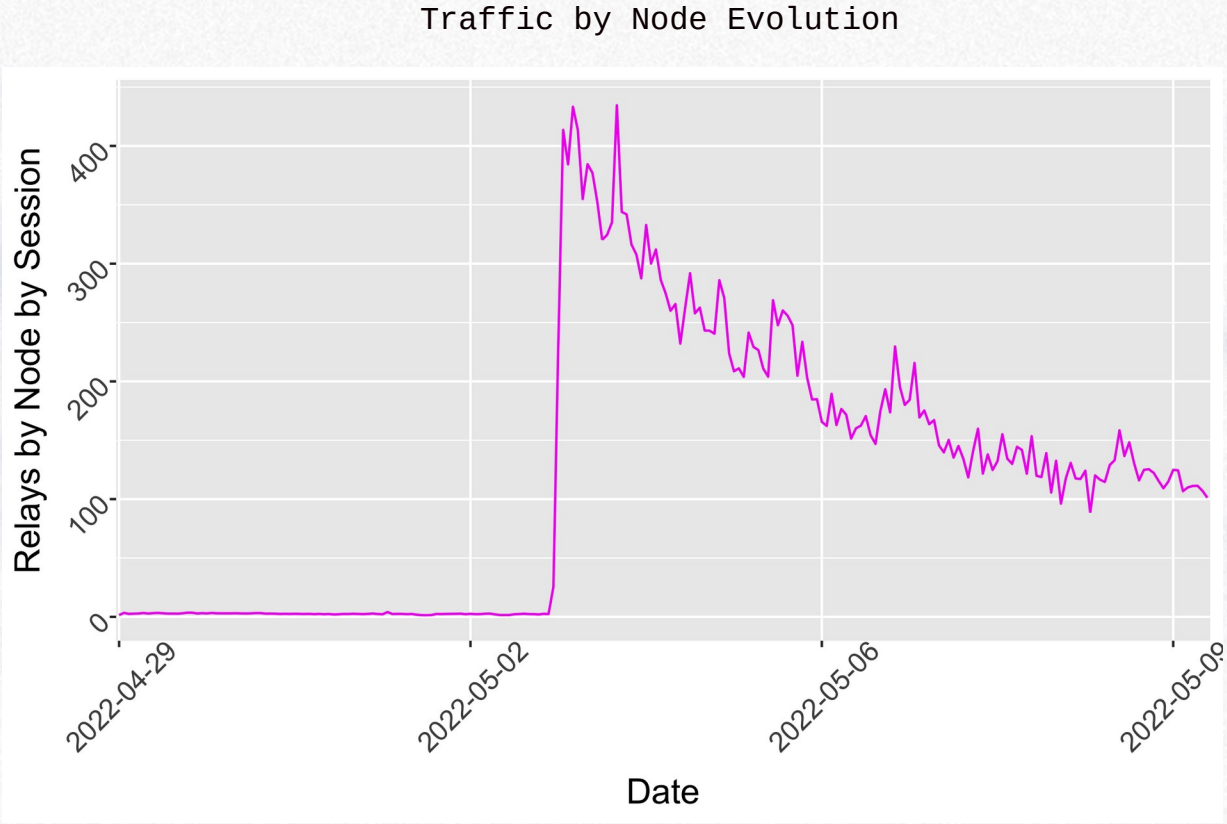
Relays Evolution



Staked Nodes Evolution



Fantom (0049)


Traffic by Node Evolution
Relays by Node by Session
400
300
200
100
0
2022-04-29
2022-05-02
2022-05-06
Date

Fantom (0049)

Resulting model: oh… man…

Dickey-Fuller test value : −1.8105 ; p-value = 0.6555

Fantom (0049)

Resulting model: oh… man…

Dickey-Fuller test value : −1.8105 ; p-value = 0.6555

The time series is

not stationary

Fantom (0049)

Resulting model: oh… man…

Dickey-Fuller test value : −1.8105 ; p-value = 0.6555

The time series is not stationary

We cannot fit a model here.

Fantom (0049)

Conclusions:

- The time series is a good example of the limitations of the ARIMA models. It cannot be fitted right now, may be we can try latter when it stabilizes.

- The network illustrates clearly the behavior of new blockchains. It presents a spike of relays by nodes (a large gain for the early adopters) and an exponential decrease due to the addition of new serving nodes.

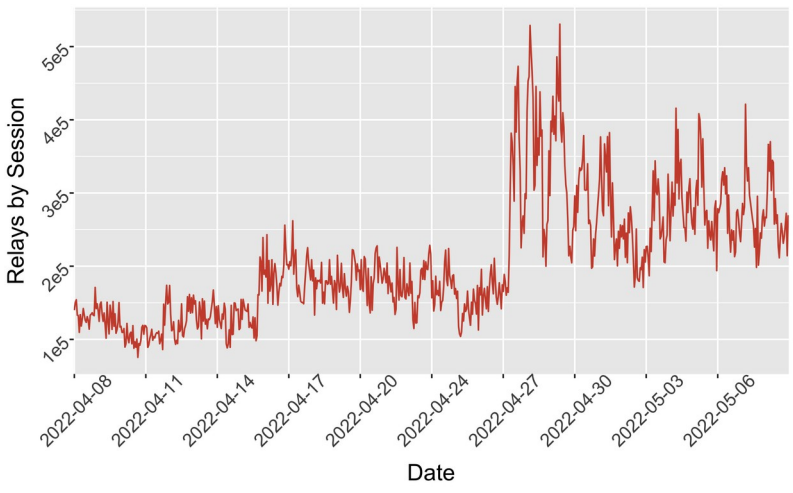
- The number of relays is growing but the node growth is much faster, meaning that new nodes will receive less and less workload.

FUSE Mainnet (0005)

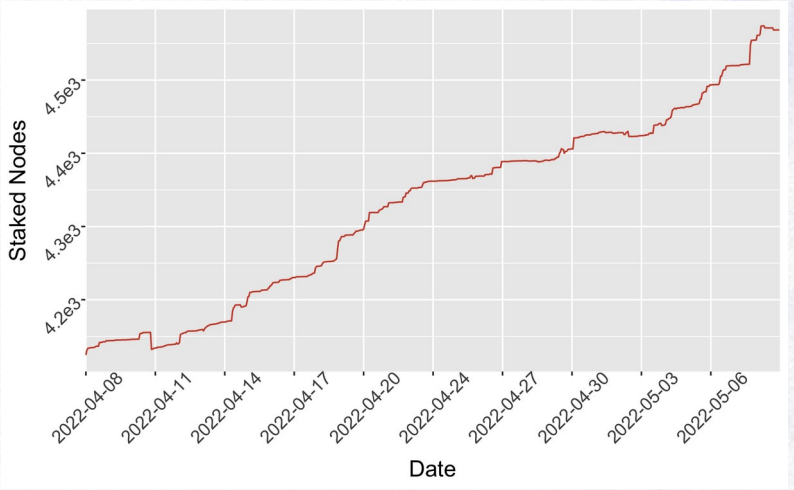
6% of the total Pocket Network relays. Eclectic and recent chain.

FUSE Mainnet (0005)

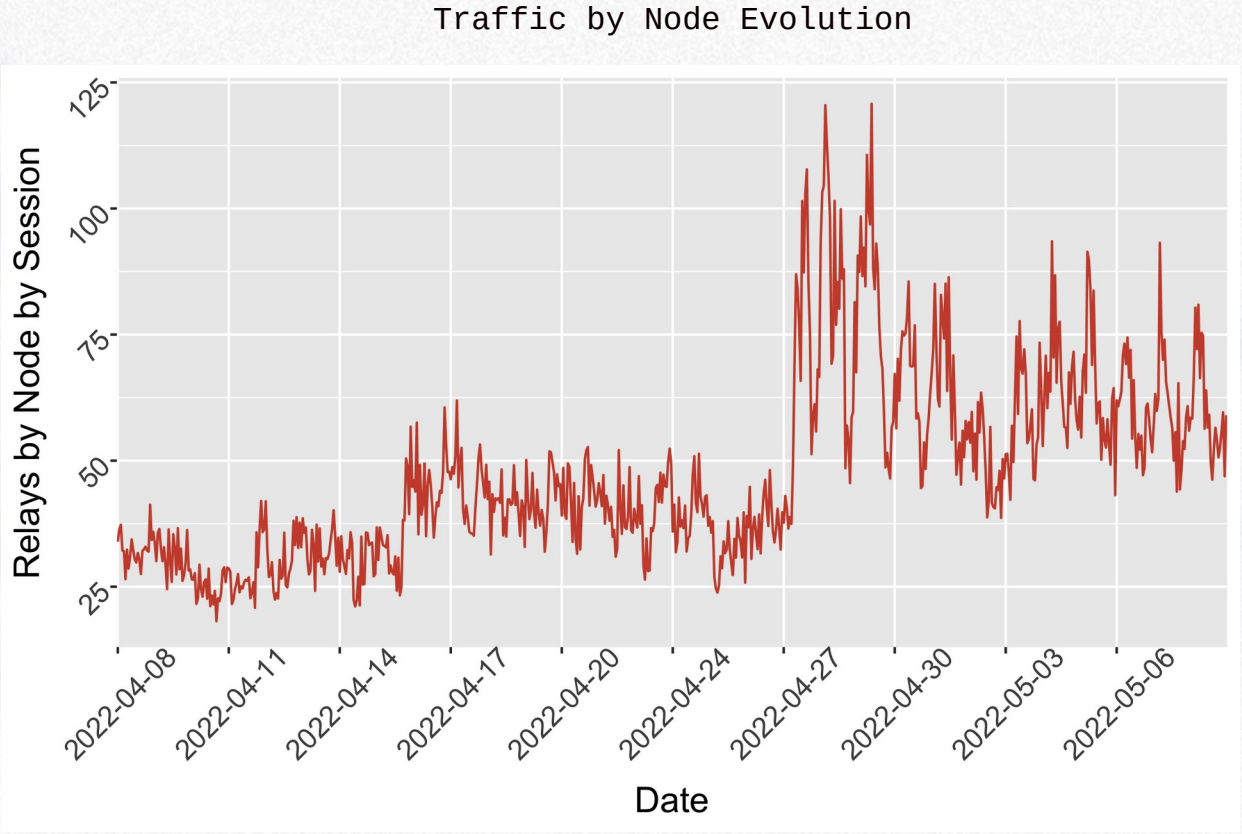
Relays Evolution



Staked Nodes Evolution



FUSE Mainnet (0005)

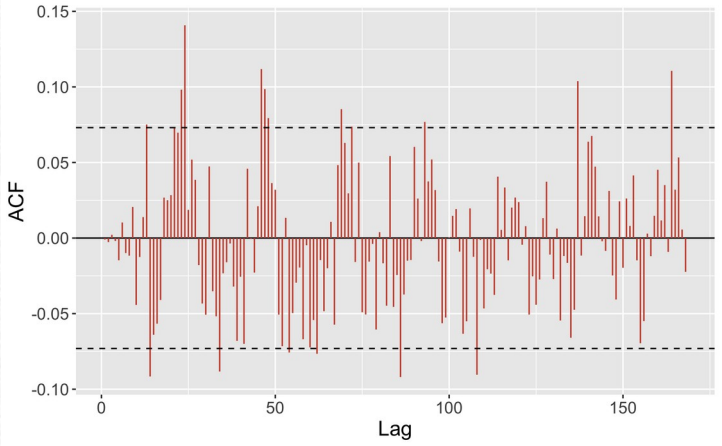


FUSE Mainnet (0005)

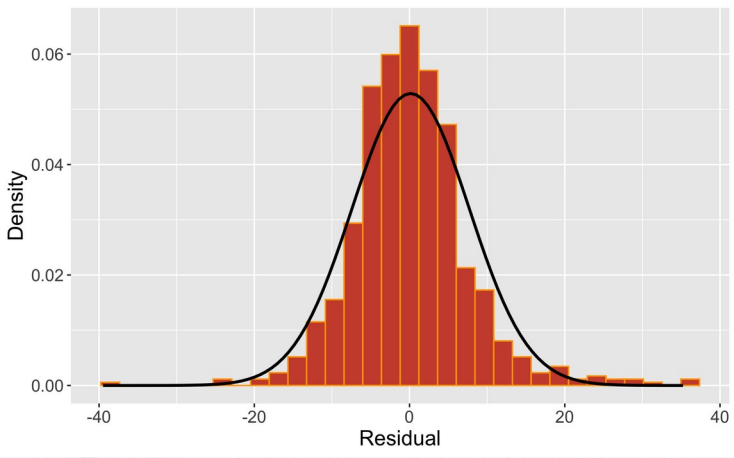
Resulting model: ARIMA(2,1,3)

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

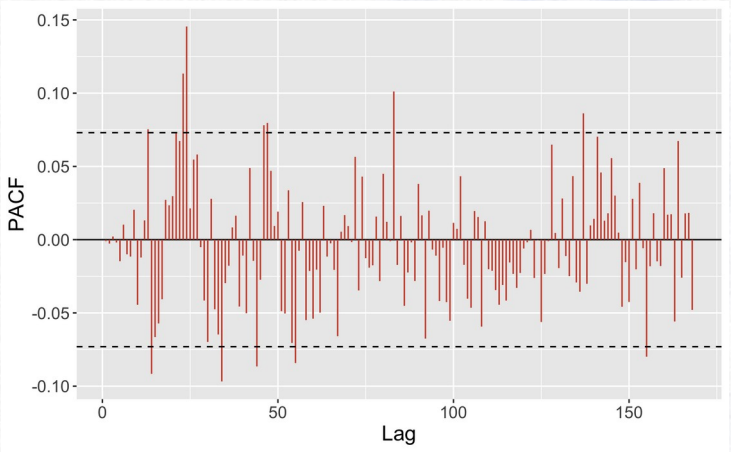
Auto-Correlation Factors



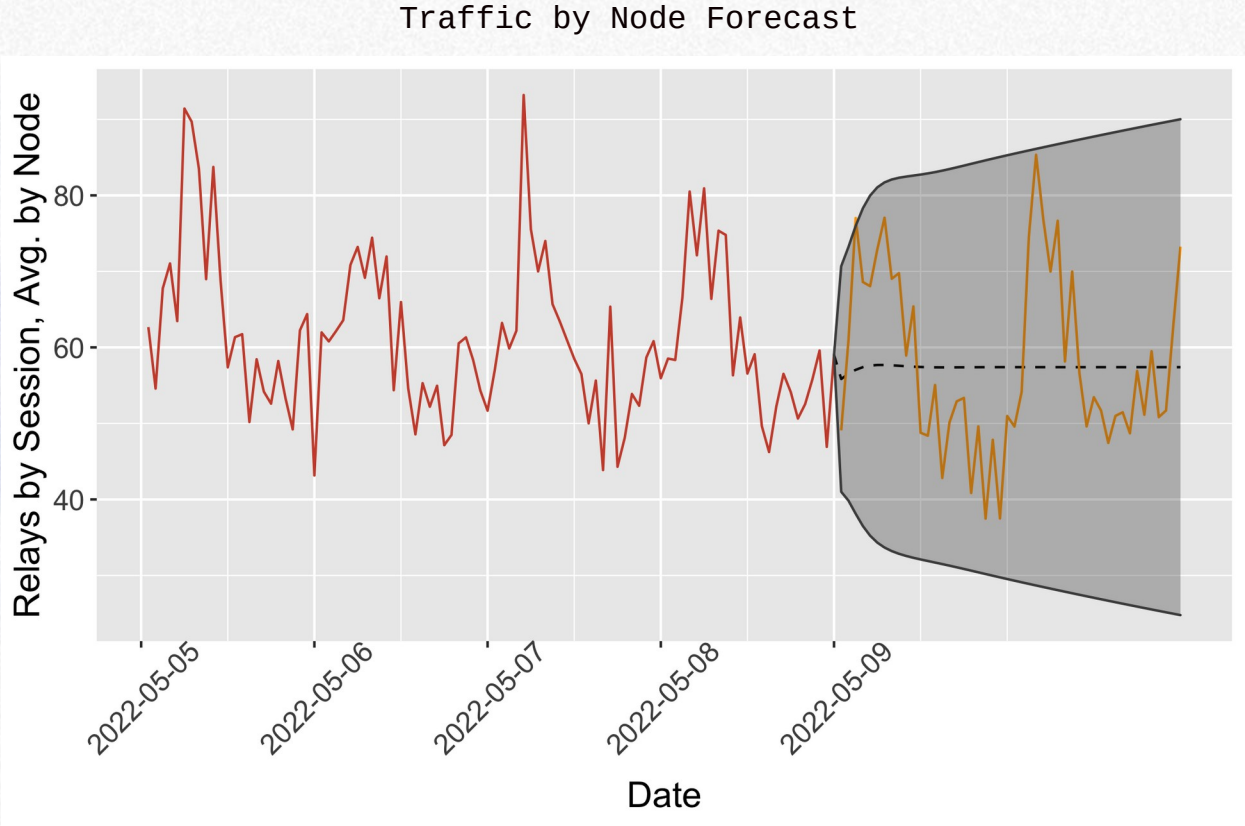
Residuals distribution



Partial Auto-Correlation Factors



FUSE Mainnet (0005)



FUSE Mainnet (0005)

Conclusions:

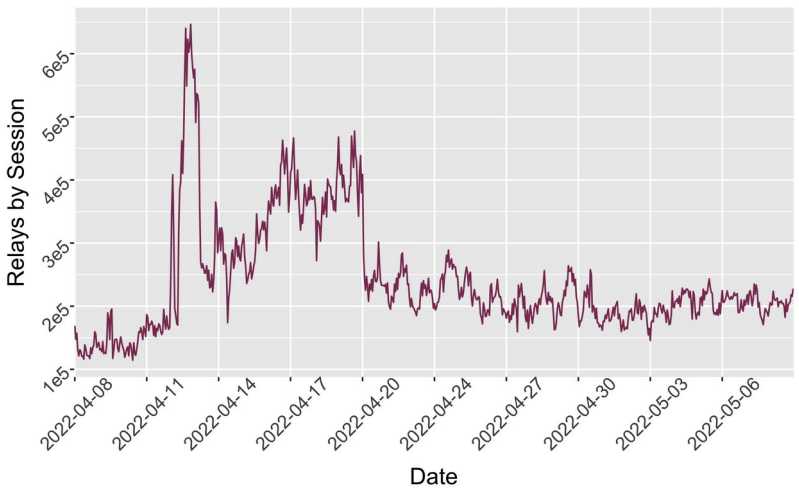
- The model presents a near-normal residual distribution with some excess kurtosis (clustered in the center), only a few ACFs/PACFs outside the significance bands.
- The model forecast is good in terms of means and expected range.
- Despite the unstable behaviour, the network is healthy in terms of relays by node stability.
- The growth of relays is faster than the growth of servicing nodes, resulting in an upward trend in the node workload.

Ethereum (0021)

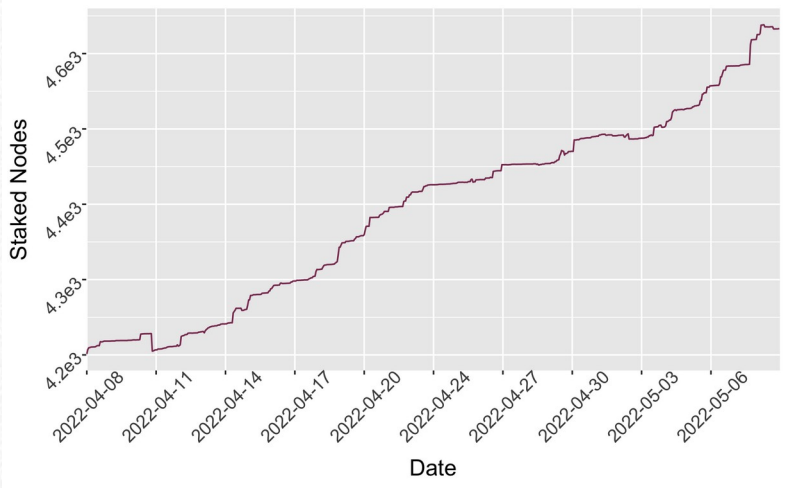
5% of the total Pocket Network relays. Oldest (and craziest) chain.

Ethereum (0021)

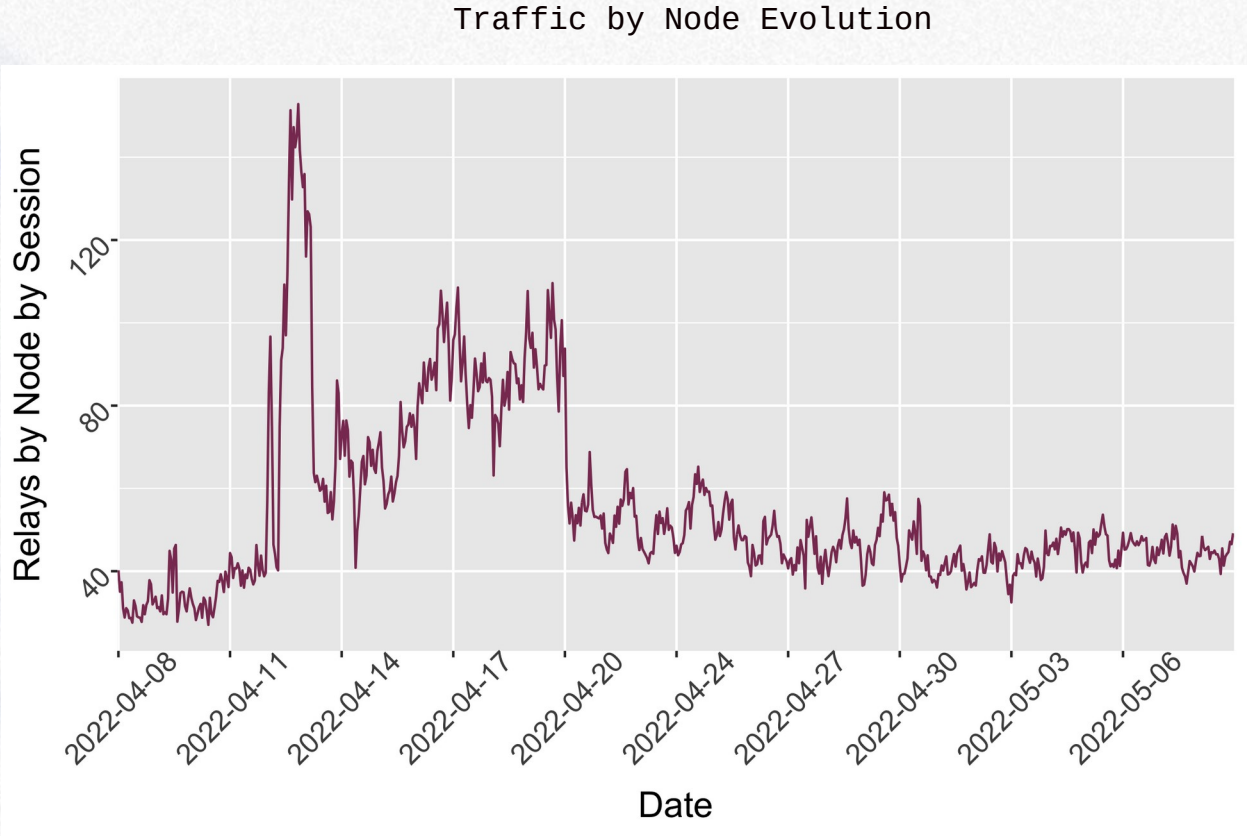
Relays Evolution



Staked Nodes Evolution



Ethereum (0021)

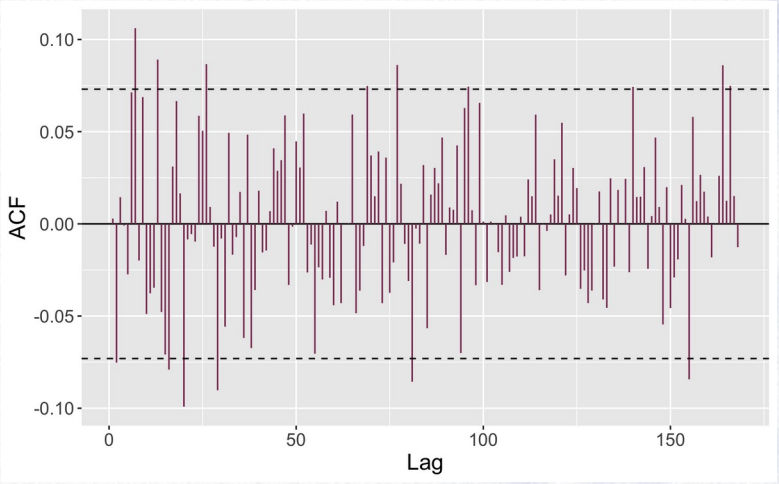


Ethereum (0021)

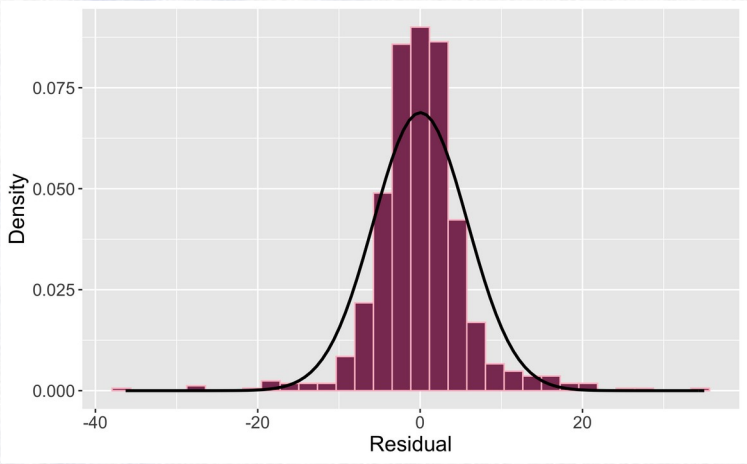
Resulting model: ARIMA(1,1,1)

ar1	ma1
0.934	−0.980

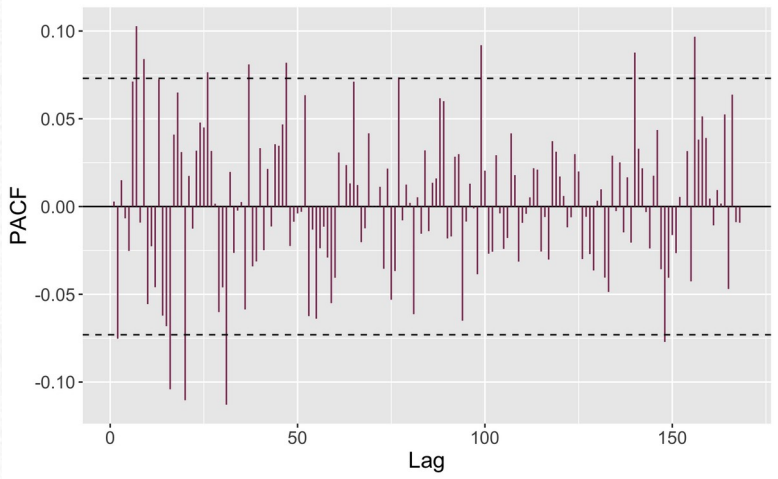
Auto-Correlation Factors



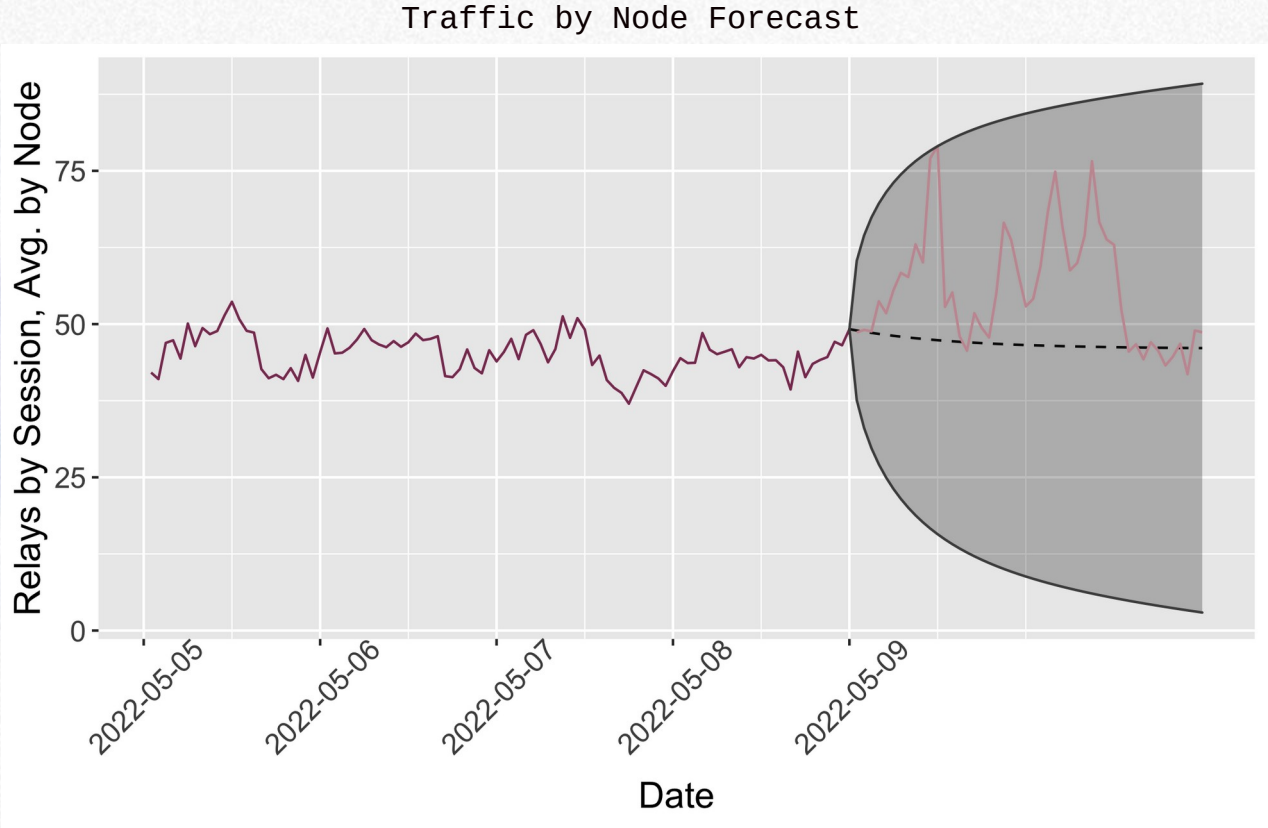
Residuals distribution



Partial Auto-Correlation Factors



Ethereum (0021)



Ethereum (0021)

Conclusions:

- The model presents a highly clustered distribution (excess kurtosis) and many ACFs/PACFs outside the significance bands. The resulting model is not good.
- The forecast mean has a large negative bias and large expected ranges.
- Even though the blockchain is one of the oldest in the Pocket Network its behavior is not stable and makes the analysis difficult using ARIMA models.
- Deeper analysis of this blockchains is needed to draw more conclusions, probably resourcing to other methods.

Conclusions

Method capabilty:

- The ARIMA models work when the blockchain presents a *good* history: Not too short and stable.

Method capabilty:

- The ARIMA models work when the blockchain presents a *good* history: Not too short and stable.

- The forecasting capability of these models must be used with caution and common sense. They are limited to a short time forecast (approx. 48 Hs in our tests).

Network Status:

- This study reveals that most of the top chains are performing good. They are stable and growing.

Chain	Prop. of total relays [%]	Stable? (30 days)	Relays by node trend
Polygon Mainnet (0009)	30.6	yes	decreasing
Gnosis - xDai (0027)	25	yes*	increasing
Harmony Shard 0 (0040)	20.8	yes	stable to decreasing
Fantom (0049)	7.4	no	—
FUSE Mainnet (0005)	6.36	yes*	increasing
Ethereum (0021)	5.33	no	—

*Not a very good fit.

Network Status:

- This study reveals that most of the top chains are performing good. They are stable and growing.

Chain	Prop. of total relays [%]	Stable? (30 days)	Relays by node trend
Polygon Mainnet (0009)	30.6	yes	decreasing
Gnosis - xDai (0027)	25	yes*	increasing
Harmony Shard 0 (0040)	20.8	yes	stable to decreasing
Fantom (0049)	7.4	no	—
FUSE Mainnet (0005)	6.36	yes*	increasing
Ethereum (0021)	5.33	no	—

*Not a very good fit.

- Additional work is required to analyze the actual workload on the nodes and state the real capacity of the network in terms of relays. (future work!)

Node Runner:

- The work expectation of new nodes is affected by many variables, this work shows only one of them: The interaction of relay and node growth.

Node Runner:

- The work expectation of new nodes is affected by many variables, this work shows only one of them: The interaction of relay and node growth.

- The models presented in this work can be used for short term analysis of node performance. A healthy node should not be outside the expected 95% confidence interval of the number of relays expected in a given chain.

Bonus Track!

How to use this in the day-to-day of a node runner

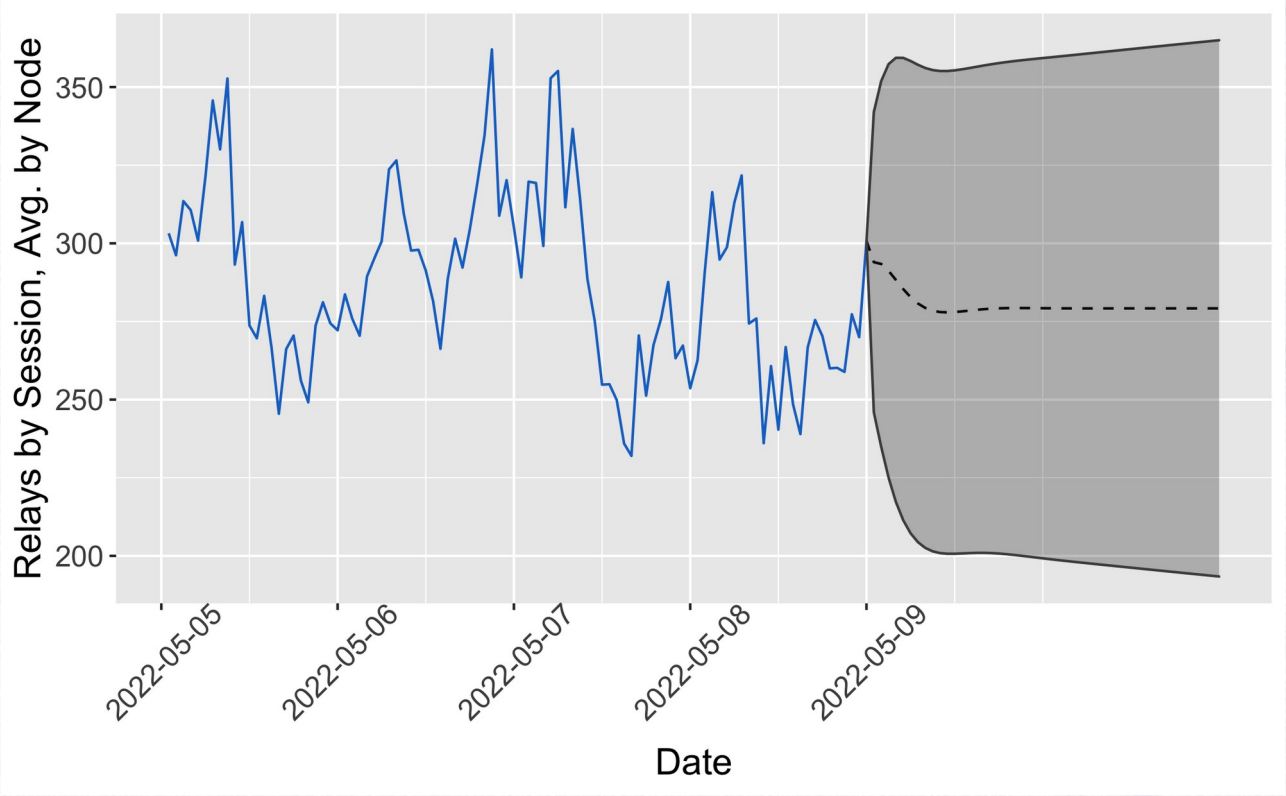
Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

1 – Fit a model and forecast a given period.

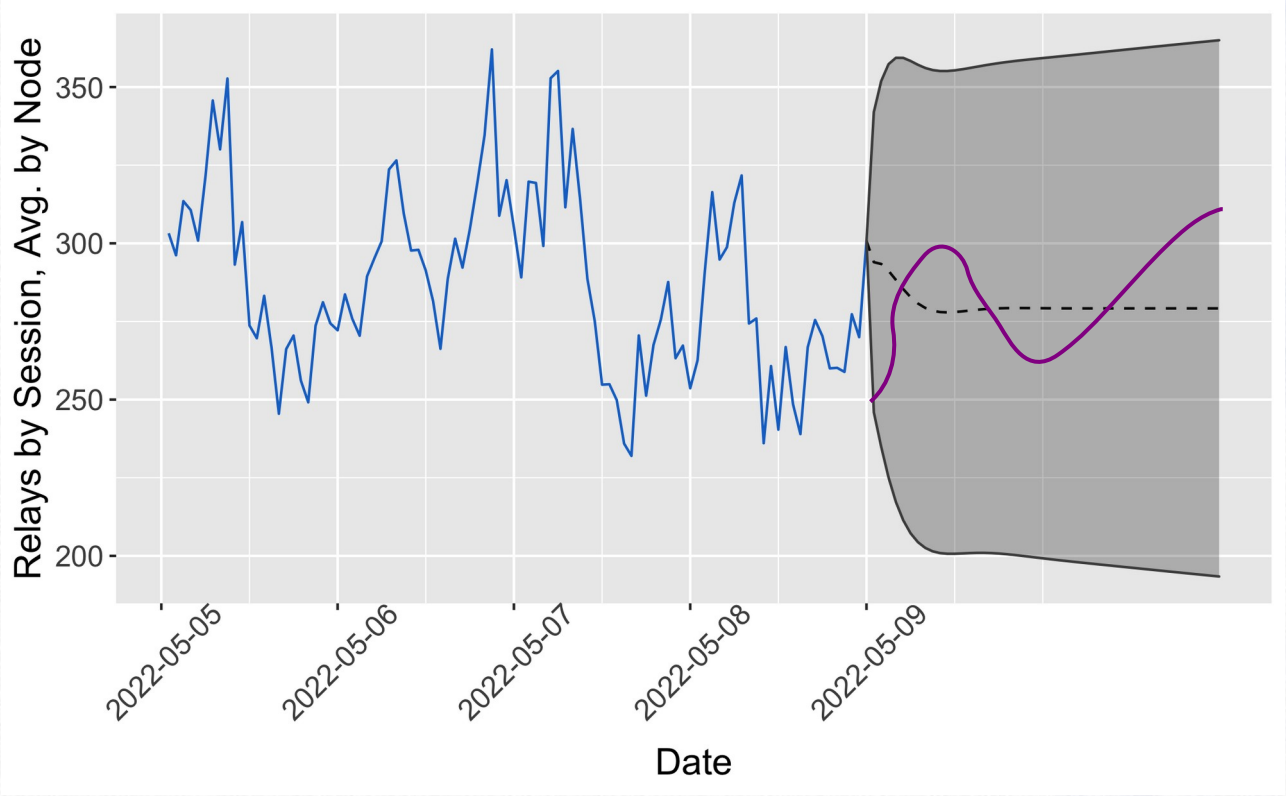


Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

1 – Fit a model and forecast a given period.

2 – Measure the Relays by Session average of your node group.



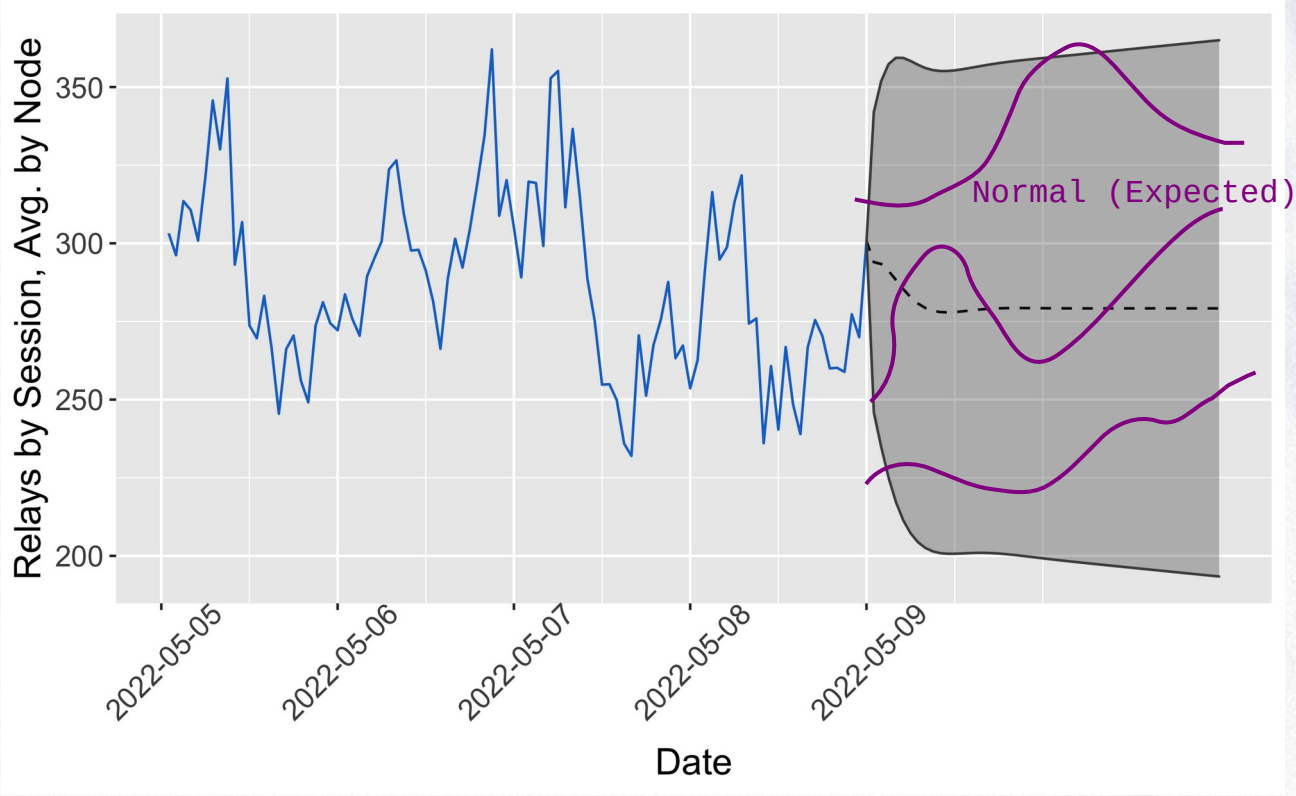
Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

1 – Fit a model and forecast a given period.

2 – Measure the Relays by Session average of your node group.

3 – Check if your values are consistently inside the grey area.



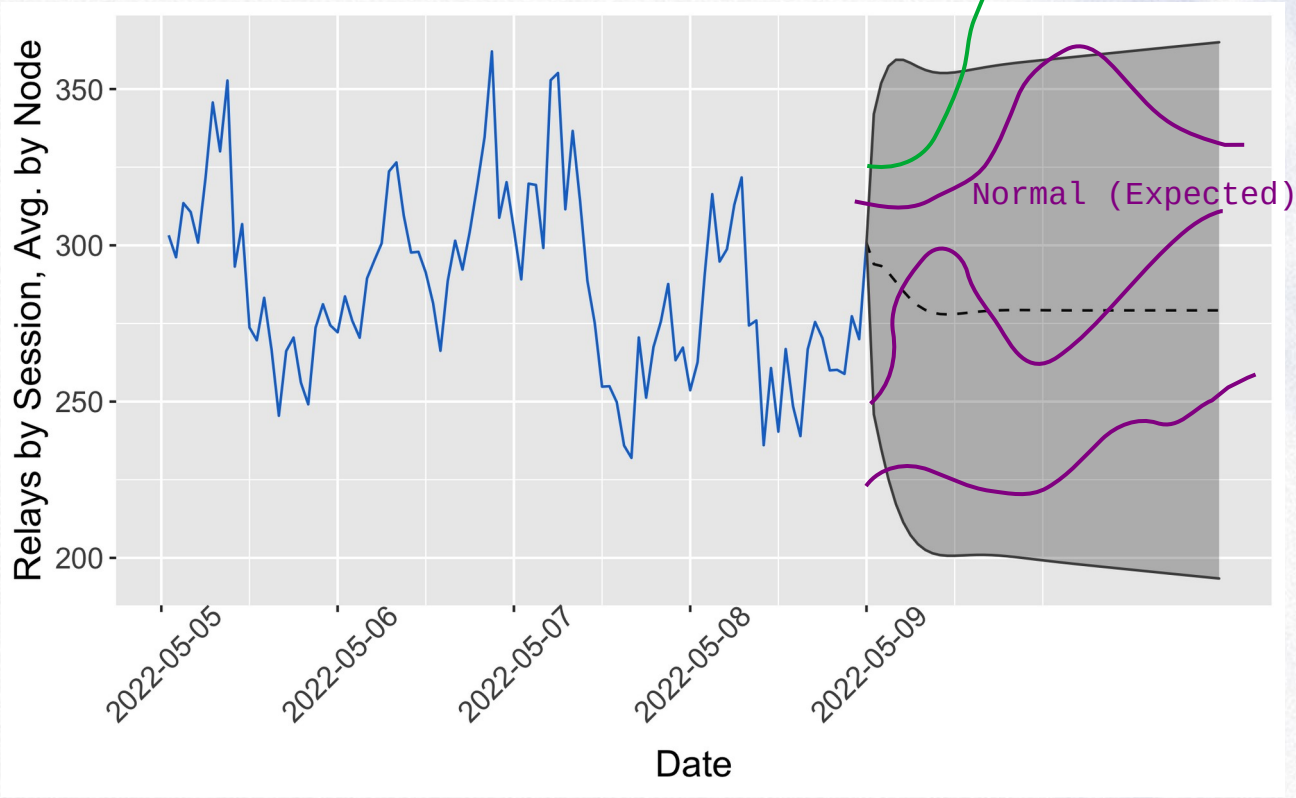
Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

1 – Fit a model and forecast a given period.

2 – Measure the Relays by Session average of your node group.

3 – Check if your values are consistently inside the grey area.



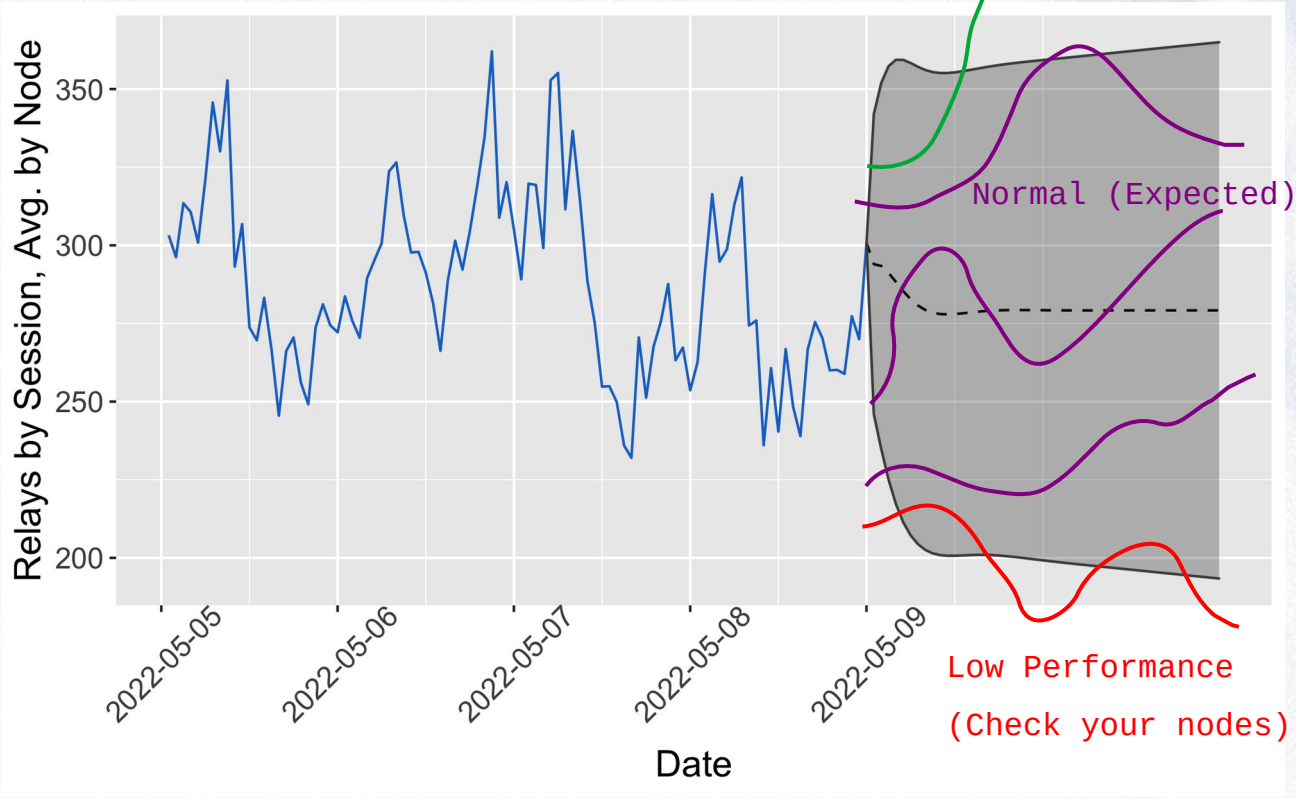
Using this model as a node provider:

(or someone with more than 1% of the network, ~500 nodes)

1 – Fit a model and forecast a given period.

2 – Measure the Relays by Session average of your node group.

3 – Check if your values are consistently inside the grey area.



And if I only have a few nodes?

This model will not work, because your sample won’t reflect the network behavior.

And if I only have a few nodes?

This model will not work, because your sample won’t reflect the network behavior.

Workaround:

1 – Construct a new time series using the average of relays of nodes **with sessions** (not all the staked nodes as we do here)

And if I only have a few nodes?

This model will not work, because your sample won’t reflect the network behavior.

Workaround:

1 – Construct a new time series using the average of relays of nodes **with sessions** (not all the staked nodes as we do here)

2 – Fit the model to the obtained time series and check if it is a good fit.

And if I only have a few nodes?

This model will not work, because your sample won’t reflect the network behavior.

Workaround:

1 – Construct a new time series using the average of relays of nodes **with sessions** (not all the staked nodes as we do here)

2 – Fit the model to the obtained time series and check if it is a good fit.

3 – Forecast a given time and wait until any of your nodes enters a session.

And if I only have a few nodes?

This model will not work, because your sample won’t reflect the network behavior.

Workaround:

1 – Construct a new time series using the average of relays of nodes **with sessions** (not all the staked nodes as we do here)

2 – Fit the model to the obtained time series and check if it is a good fit.

3 – Forecast a given time and wait until any of your nodes enters a session.

4 – Take the average of relays of your nodes **with sessions** and proceed as before. (check if it falls in the greyed area)

Thanks!

The POKTScan Team

INFRACON

<Pocket Network Community Conference 2022>