A Statistical Data Analysis on GitHub Dynamics and Market Synchronization for fifteen Cryptocurrencies

Honeylyn Garde
1169386
ghoneylyn@student.unimelb.edu.au

Xueying LI
1129769
xueying4@student.unimelb.edu.au

1 INTRODUCTION

1.1 Background

The value and the adoption of cryptocurrencies often depends on the demand based on how buyers perceive it (García-Monleón et al., 2021). It can also depend on the current market conditions and because of its decentralized nature, each cryptocurrency token has different characteristics to compete and to strive in the ever-changing competitive market to help retain its value. Once listed on an exchange, its price can further go up or skyrocket as determined by variable market forces, including demand and supply (Giudici et al., 2019). Previous studies have suggested that popularity of a cryptocurrency is one of the main drivers of its value (Kristoufek, 2014). The more widely known it is, the higher its value rises, despite any potential questions on the validity and credibility of the project. Often, intelligent market participants and developers would further inspect the whitepapers and documentation repositories such as published in Github (Github, 2022) to inspect reliability and project quality of a cryptocurrency (Dabbish et al., 2012; Trockman et al., 2019).

1.2 Research Questions

These studies indicate that drawing insights using repository mining techniques may potentially enable financial profitability. Inspired by them, the current paper aims to investigate the the features present in repository activities and how that is synchronized with outside events, especially the market behavior. In particular, we are interested to learn (i) how do developers of each cryptocurrency project share the editing workload (ii) how does the development records change over time (iii) what are driven factors for these changes (iv) how can statistical results further prove our findings.

2 METHODOLOGY

2.1 Datasets

GitHub Dataset

Weighing on the diversity in blockchain technologies, we selected the following 15 cryptocurrency projects that are also popular in the market. *Proof-of-Work* projects: **Bitcoin**, and its derivatives **Litecoin** and **Zcash**; **Dogecoin**, the meme coin derived from Litecoin, initially featuring a randomized reward but having moved to a static return. *Proof-of-Stake* based projects: **Cardano**, **Avalanche**, **TRON**, top POS/Delegated POS projects; **Polkadot**, using nominated POS, featuring multi-chain and forkless upgrade; **Polygon**, the Ethereum compatible project; **Cosmos**, incorporated with *Byzantine Fault Tolerance*; **Algorand**, pure POS. **Ethereum**, which is moving from POW to POS. **Solana**, known for its own *Proof-of-History* protocol that facilitates fast transaction processing. **Stellar**, known for the *Stellar Consensus Protocol* based on trustworthy nodes and aimed at supporting cross-border transactions. Powered by *Google BigQuery* and GitHub's open database *githubarchive* (GitHub, 2022), we retrieved the daily statistics (push, pull request, watch, fork events and the corresponding developer) for all the 15 projects from January 1, 2021 to May 20, 2022.

Market Dataset

The market data we collected from the web sources: CoinCodex (CoinCodex, 2022) and CoinMarketCap (CoinMarketCap, 2022) [1] is comprised of the daily close price, traded volume and market capitalization of each project. The close price, as the price the cryptocurrency was lastly traded, and the market cap, as the product of the price and the

[1] Due to the limitations in downloading historical data from CoinMarketCap, the site was used solely for validating the downloaded data from CoinCodex

number of coins in circulation, were processed and USD was set to be the base currency reference. The trading volume pertains to the total trading volume from USD to one cryptocurrency (Lucchini et al., 2020). All the data are according to all the market activities across exchange markets.

2.2 Theories and Models

Detection of linked project pairs

To investigate the impacts of co-developing cryptocurrency projects, we identified *linked pairs* as, according to (Lucchini et al., 2020), two cryptocurrencies that share at least one developers contributing to a Github repository. On the other hand, *random pairs* would be two cryptocurrencies that do not have any common developers contributing to the Github repositories. In this project, we analyse the characteristics and relationships of Github and market activities of the linked pairs mentioned in table 2 and assess how these differ to the random pairs.

Models for correlation analysis

We also explored the pairwise Pearson correlations between the market data such as average volume, average market cap and the Github metrics such as push and pull events, watched and forked events, averages over the entire time period in scope. We then fit linear models of average market capitalization as a function of the Github activity metrics to assess the predictive relationship between the Github metrics on the average market cap.

Time series analysis

To have a pairwise comparison of cryptocurrency characteristics, we analysed different time series aspects of financial key metrics such as change in price over time t and Pearson correlation within the linked pairs and within the random pairs of cryptocurrencies (Lucchini et al., 2020). The cryptocurrency change in price at time t is defined as $\frac{(Price_t - Price_{t-1})}{Price_{t-1}}$. The normalized Pearson correlation between two cryptocurrencies was calculated over a rolling period of 14-day window as follows:

$$Corr_k = \frac{C_k(t) - M_k(t)}{\sigma(t)}$$

where $C_k(t)$ is the correlation time series from the linked pair of cryptocurrency A and B at time t, and $M_k(t)$ and $\sigma(t)$ are the average correlation and corresponding standard deviation across pairs (Lucchini et al., 2020).

	Github_Name	Market_Cap	Mean_Daily_Edit	Median_Daily_Edit	Total_Edit	#Developer
Cardano	input-output-hk	8	149.275100	176.0	74339	279
Polkadot	paritytech	11	111.016097	124.0	55175	165
Ethereum	ethereum	2	69.957490	72.0	34559	156
Solana	solana-labs	9	57.975659	59.0	28582	84
Cosmos	cosmos	28	50.146939	51.0	24572	135
Stellar	stellar	27	20.401345	18.0	9099	55
Avalanche	ava-labs	12	18.220000	14.0	8199	52
Algorand	algorand	29	16.285425	13.0	8045	50
Zcash	zcash	42	11.389755	10.0	5114	23
Terra	terra-money	211	15.429448	13.0	5030	49
Polygon	maticnetwork	17	11.126521	9.0	4573	66
TRON	tronprotocol	14	6.698113	5.0	2130	39
Bitcoin	bitcoin	1	4.912037	4.0	2122	9
Dogecoin	dogecoin	10	2.335260	2.0	404	7
Litecoin	litecoin-project	18	2.829787	2.0	133	4

Table 1. The editing statistics of the 15 projects: number of developers, total edits, mean and median number of daily edits

	Common_Devloper	First_Connection_Time
Ethereum ++ Cardano	KonradStaniec	2022-01-12 11:37:13 UTC
Cardano ++ Cosmos	Mr-Leshiy	2021-10-13 06:53:40 UTC
Solana ++ Terra	JoshTDLester	2022-04-14 13:48:59 UTC
Polkadot ++ Terra	jarcodallo	2022-02-16 13:12:57 UTC
Polygon ++ Cosmos	anilCSE	2021-01-07 09:26:24 UTC

Table 2. The Linkage between projects on GitHub: common developer and the first connection time

3 RESULTS

3.1 Developer Analysis

Overall statistics

We identified a total of 262,566 push events from 1,190 individual developers as the edits in total. As shown in table 1, the rankings of projects in mean (or median) daily edits, are close to the rankings in sums. The Cardano project, which belongs to the "Input-Output HK" organization, is the most developed project on GitHub, holding the largest developer group and soaring above others in the number of edits. Other big projects (>10k edits) are Polkadot, Ethereum, Solana, and Cosmos; Dogecoin and Litecoin are the least in push records. For larger projects, the number of developers strongly correlates to the records of edits; however, this correlation appears weaker in their smaller counterparts, leading us to further look at the roles of the developers.

Project contribution share

As the distribution of each project suggests (figure 1), we have determined a substantial centralization in submitting code. 10 out of the 15 projects (Bitcoin, Solana, Dogecoin, Avalanche, Tron, Litecoin, Stellar, Algorand, Zcash, and Terra) have at least

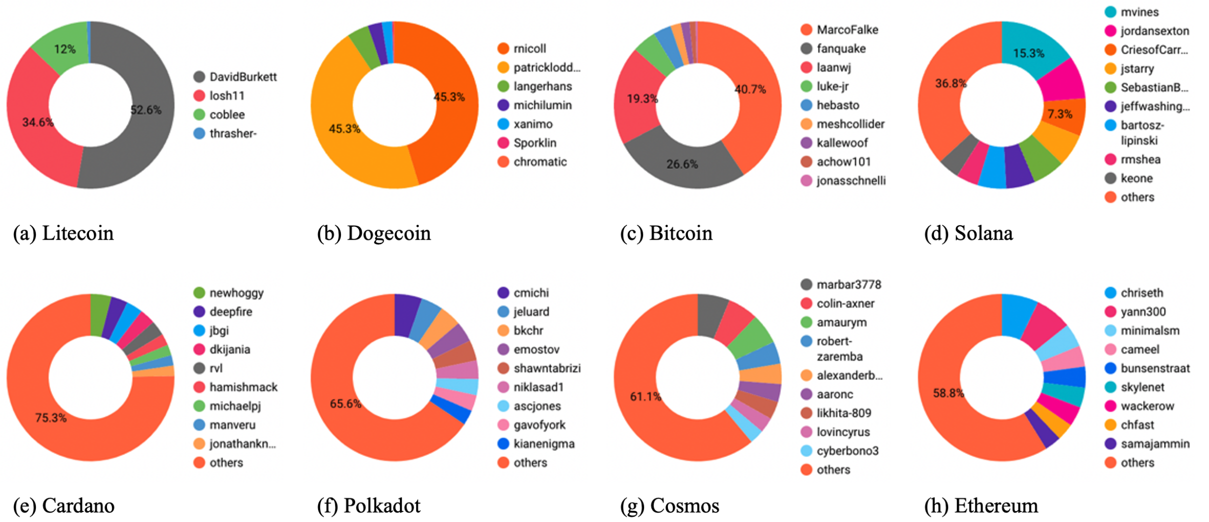


Figure 1: The contribution share of 8 example projects. Key developer (edits>10%) are common in projects (a)-(d), while projects (e)-(h) present a more decentralized distribution. See Appendix A for more projects.

one developer contributing more than 10% of the total edits. We also found that the concentration grows against the project size. For smaller projects such as Litecoin or Dogecoin, their key developers have contributed nearly half of all the edits. For bigger projects such as Cardano or Polkadot, the distribution of work tends to be more decentralized as more developers have joined in. However, key developers can exist in larger projects; 15% of Solana edits come from a single user, which counts as over 4k push events.

Linked projects

Similar to the paper by Lucchini et al., we also identified projects that are "linked" in Github by sharing a common developer. Linked pairs and the first time they connected (when the common developer edited the second project) are shown in table 2. Cardano, Cosmos, and Terra have at least one common developer from other projects.

3.2 Temporal Analysis

Overall trends

Here we provided the stacked area chart to demonstrate how the editing records evolved (figure 2). From the entire perspective, a remarkable weekly pattern can be captured, showing that the volume of edits downgrades on the weekends but recovers when a new week starts. The total volume increased from January 1st to late March 2021, possibly due to the re-emerging wave of COVID-19. At this point, multiple countries re-introduced the lockdown policy, leading to a new wave of working from home; developers hence had more time to take care of their GitHub projects. Following that the overall trend keeps declining until mid-September when the second peak emerges (a point when the major projects were experiencing an upward trend). Several valley points are also remarkable in figure 2; we can assume that the steeper decline at April 2021 and 2022 is due to the Easter break, and the crucial decrease from late December 2021 to early January 2022 lasting for weeks is related to the Christmas-New Year Holiday. Given so, the strict cut-off around October 26, 2021, cannot be simply explained as a holiday or any major event. We found this is hugely possible due to a database error; our query [2] result shows that in the meantime for any GitHub project, none of their push events had been recorded in the githubarchive.

Project and market dynamics

To better understand the GitHub dynamics, we extended our scope to multiple GitHub events, including push, pull request, watch, and fork. We determined that the number of watches or forks captures the popularity of the project, whilst the push and pull requests are significant indicators of activity. Therefore we were able to explore the relationship between these attributes and two significant market indicators, namely the market cap and transaction volume, for each project. We found that the correlation between the market and GitHub metrics varies among cryptocurrencies (See Appendix B).

[2] We queried the database if any push events of any project took place in October 26-27, 2021; the result was none.

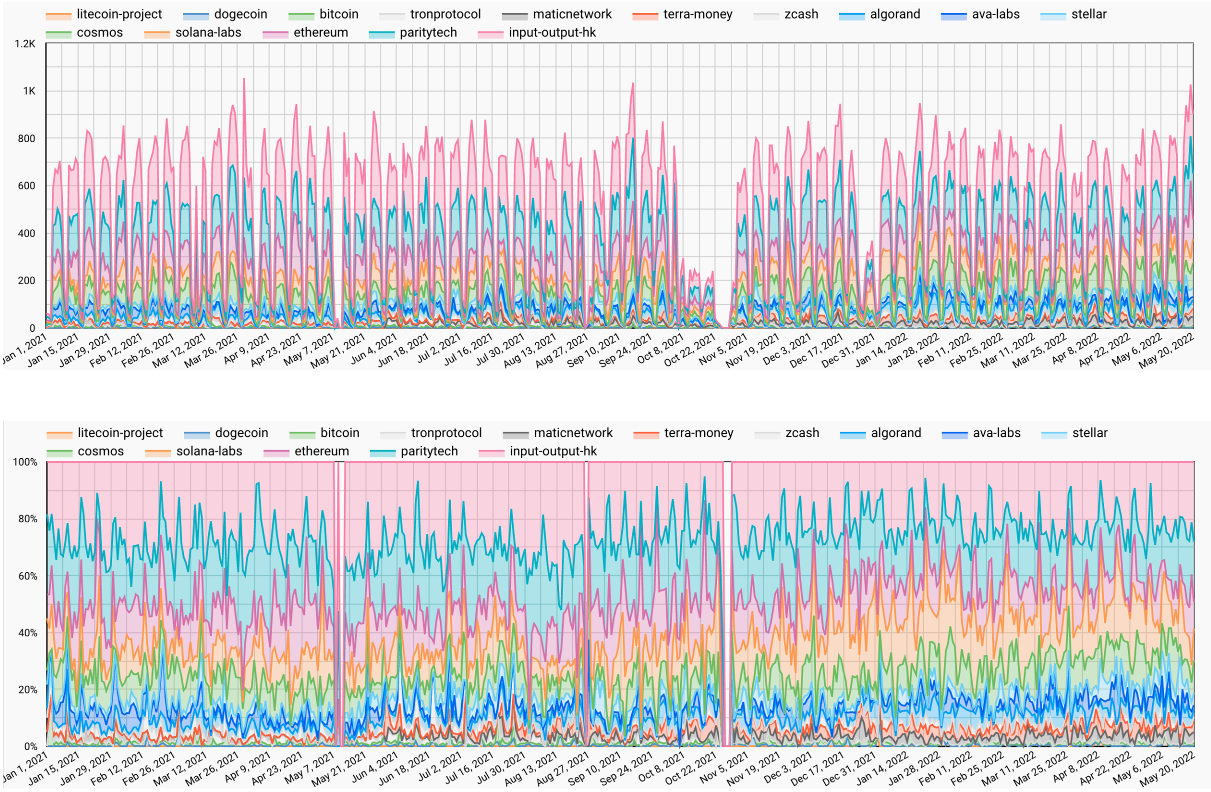


Figure 2: (a) Stacked project edits over time (b) Percentage stacked project edits over time. The cut-off in October may caused by an error in the githubarchive.

(a) Ethereum

(b) Dogecoin

(c) Solana

Figure 3: GitHub and market dynamics of Ethereum, Dogecoin, and Solana

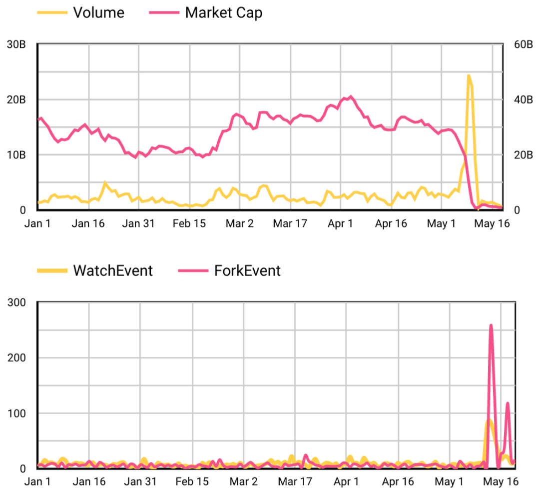


Figure 4: Market and popularity dynamics of Terra

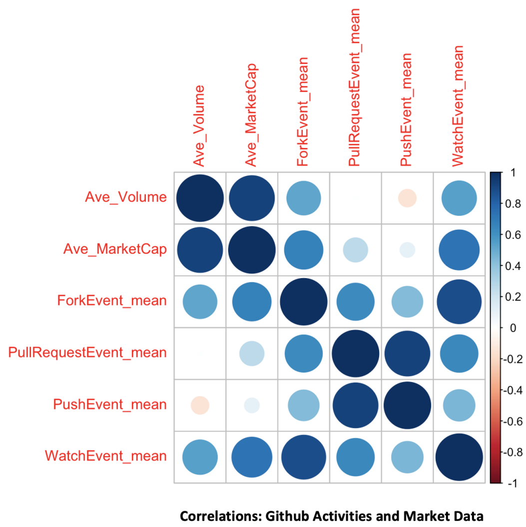


Figure 5: Correlations between Github Metrics and Market Data for all projects in scope

For the majority of these projects, exemplified by Ethereum, the popularity metrics (watch, fork) instead of activity metrics (push, pull) have shown correlations to either one or both of the market indicators. As illustrated in figure 3, the Ethereum market cap reached its peak in November 2021, when a stinging rise also appeared in both watches and forks. In May 2021, the increase in Ethereum watches also synchronized with its growth in the market. Although generally, the activity metrics are not strongly tied with the market, the second-highest volume of daily edits occurred in April, which can be attributed to the surge in the transaction volume. It is worth mentioning that the edits of Ethererum reached their maximum recently (May 2022); we consider this should be associated with the "Merge" that Ethereum has been shifting to proof of stake. Technology update also boosted the market capitalization and the GitHub forks for Polkadot (Appendix B) in November 2021, when the organization was about to launch the parachain.

For Dogecoin and Solana, we observed a synchronization between the market metrics and both GitHub metrics. First, the watches Dogecoin receives strongly correlate to its market growth, which can be associated with major investing events. Second, the push or pull requests also peaked at the end of January 2021 (May 2021 again), tuning with their market growth. A similar trend appeared in the Solana series in September, when the project featuring the proof of history protocol actually became popular in the community. For the recent trending project Terra, when the market collapsed in May 2022, massive selling transactions took place, and the surge in watches depicted the trend in volume. In general, as mentioned earlier, the project activity is hugely affected by holidays or weekends, but it also changes at the time when new features are released; the popularity metrics are found more subjective to the trend in the market.

3.3 Quantitative Correlation Analysis

To support the findings from the overall trend analysis and the project market dynamics, we perform quantitative correlation analysis.

Pearson Correlation. The market data and the Github data were aggregated by computing the average over the entire time period. The Pearson correlation among the market and Github metrics were calculated and we found some significant correlations: watch events 75%, fork events 68%, pull requests 27% and push events 10%. In practice, the highly correlated variables must be further removed before any model fitting as these metrics may potentially skew the results (for instance, projects with more watchers are more likely to suggest more forks events to occur). See figure 5.

Linear Models. We then fit linear models of the average market cap as a function of Github data. The aim is to attempt to model the relationship between the Github metrics and average market price by fitting a linear equation to observed data (Yan and Su, 2009). Before fitting, we first transformed all continuous variables into their logarithmic form to keep all the scale and ranges of values standard-

```
lm(formula = log(Ave_MarketCap) ~ log(ForkEvent_mean) + log(PushEvent_mean) +
    log(WatchEvent_mean), data = coin)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4581 -0.5688 -0.1493  0.2736  2.1762

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           21.9130     0.6335  34.588 2.17e-13 ***
log(ForkEvent_mean)    0.2276     0.5486   0.415   0.6856
log(PushEvent_mean)   -0.2805     0.1874  -1.497   0.1603
log(WatchEvent_mean)   0.9442     0.5100   1.851   0.0889 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.089 on 12 degrees of freedom
Multiple R-squared:  0.6148,    Adjusted R-squared:  0.5185
F-statistic: 6.385 on 3 and 12 DF,  p-value: 0.007835
```

Figure 6: Linear Regression Model 1 ("Full Model") of Market Cap.

```
lm(formula = log(Ave_MarketCap) ~ log(PushEvent_mean) + log(WatchEvent_mean),
    data = coin)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6574 -0.5685 -0.1972  0.3311  2.1607

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           21.8547     0.5978  36.561 1.71e-14 ***
log(PushEvent_mean)   -0.2745     0.1808  -1.518 0.152871
log(WatchEvent_mean)   1.1260     0.2522   4.465 0.000637 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.054 on 13 degrees of freedom
Multiple R-squared:  0.6093,    Adjusted R-squared:  0.5492
F-statistic: 10.14 on 2 and 13 DF,  p-value: 0.002224
```

Figure 7: Linear Regression Model 2 ("Final Model") of Market Cap.

ized. The resulting fitted linear model is in figure 6. We applied stepwise Aikake Information Criterion (Akaike, 1976) to investigate the optimal explanatory variables and remove those that don't have significant contribution to the predictive model. As a result we have a full model based on all the watch events, fork events, push events and pull events. After performing the stepwise Aikake Information Criterion, the resulting model found push events and watch events as significant to the model (see figure 7).

To further investigate, we tested joint hypothesis using F-tests (Trockman et al., 2019) to assess the significance of pull events and fork events (i.e. testing the null hypothesis H_0: forks = pull = 0). After running this test, we found that, at p-value = 0.05, we fail to reject the null hypothesis stated above, which means that the fork events and pull events are not significantly contributing to the model, hence, we exclude in the final model. In this final model (i.e. no fork events and pull events), the watch events were found to be highly significant to the model: a 1% increase in the watch events is associated with 1.13% increase in the average market cap. On the other hand, in the full model, the watch events were also found to be significant:

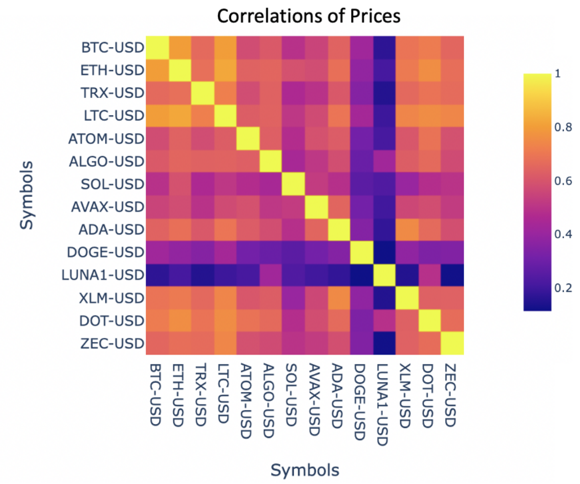


Figure 8: Correlation of Prices between the projects in scope

a 1% increase is associated with 0.94% increase in the average market cap. In terms of the adjusted R^2 (i.e. percentage of variation explained by the model) (Miles, 2014), the full model explains 52% of the variance, while the final model explains 55%. Thus, we conclude that the Github metrics of popularity (i.e. the watch events) (Github, 2022), as supported by the significance of push events, explains its significant relationship with the average market cap.

3.4 Time Series Analysis

Now that we have investigated the role of watch events and pull events in explaining the cryptocurrencies and market cap using the regression analysis, here we discuss the temporal aspect of market data that will support our analysis in the Github activities mentioned in previous sections. We will also discuss in what ways does the linked pairs and the random pairs of cryptocurrencies differ.

Correlation of Closing Prices. To have a holistic view, we have inspected the correlation of the closing prices between all of the cryptocurrencies in scope in this project. In figure 8, we can infer that majority of the prices are moderately strongly and positively correlated with another cryptocurrency in this pool (correlation coefficient around > 0.5). This means the the market prices of these cryptocurrencies tend to go along in the same direction (Kirch, 2008). We will now further investigate this relationship in time series representations.

Market Properties of Github-linked cryptocurrencies We have investigated the time series of prices for each of the linked pairs and found an interesting observation. When a linked pair comprises of cryptocurrencies that belong to top list in terms of market capitalization (CoinMarketCap, 2022), there are some evident patterns of market synchronization after the time that a developer or group of developers from a cryptocurrency A in the linked pair contributes codes into the Github repository of the paired cryptocurrency B. This supports the observation we can see in figure 9. Take a look at the upper left chart of prices for Cardano and Ethereum. From the time a developer started contributing into Cardano, the cryptocurrency prices of this paired projects tend to have similar trends, tighter than it was before the co-development. Similar trends can be seen with linked pair Cardano and Cosmos, as well as Solana and Luna. Same observations can also be draw on the cryptocurrency asset return (calculated as the rate of change in price of the cryptocurrency (Lucchini et al., 2020)) as well as the normalized correlations of prices among lined pairs. Refer to figures 11 and 13, respectively. However, these findings may need to be backed by further analysis as purely relying on reading trends may need a great amount of carefulness when used in interpreting results. Meanwhile, this observation cannot be extended to the case of random pairs. In figure 10, the trend charts seem to lack the observed patterns described. Same observations can also be draw on the cryptocurrency asset return as well as the normalized correlations of prices among lined pairs. Refer to figures 12 and 14, respectively.

4 Conclusion and Future Work

The current paper presents a statistical analysis on the GitHub dynamics of 15 cryptocurrencies. By investigating the editing events, we detected the shared developers and the centralization in developing cryptocurrency projects. We also found that holidays (weekends) and technology updates are the primary factors that affect the daily rate of push. Based on which, we associated the trends in development with the market behavior, and performed a quantitative correlation analysis. We then further explored the market relation between projects that are co-developed by the common developers, and compared such correlation with non-linked or randomly paired projects.

We recognize that our study has some limitations. First, we could expand our datasets to include a wider range of cryptocurrencies (not just the popular ones) to back our correlation analysis and reduce the variation. Second, we could combine more human effort in detecting what GitHub repositories are "in scope"; mature cryptocurrencies like Bitcoin can have multiple relevant projects that are maintained by the community. The project namely "bitcoin" may not fully reflect the activeness of this cryptocurrency. Third, project management are highly variable among organizations; the amount of edits does not imply the merit of the code, and so does not give enough information about the value of the project.

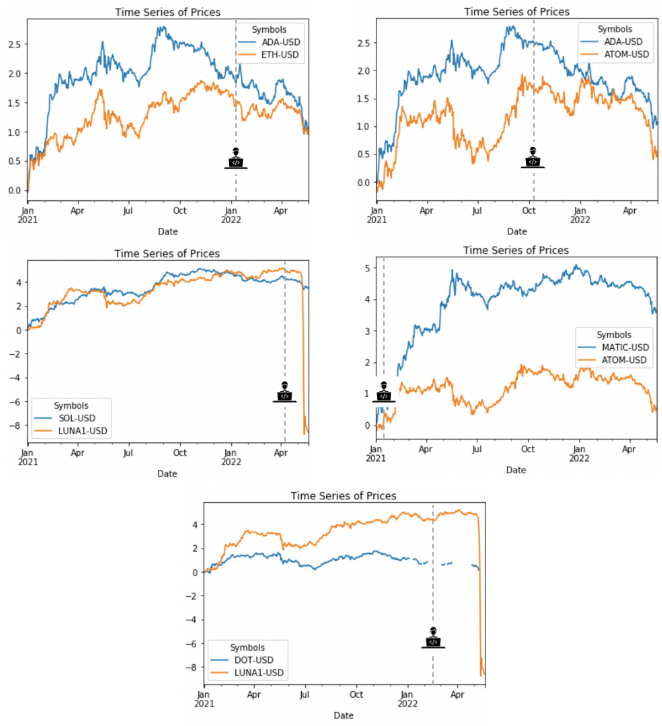


Figure 9: GitHub co-development and cryptocurrency market impacts. some patterns reveal some impacts to prices when a developer of "cryptocurrency A" co-develops another project "cryptocurrency 2".

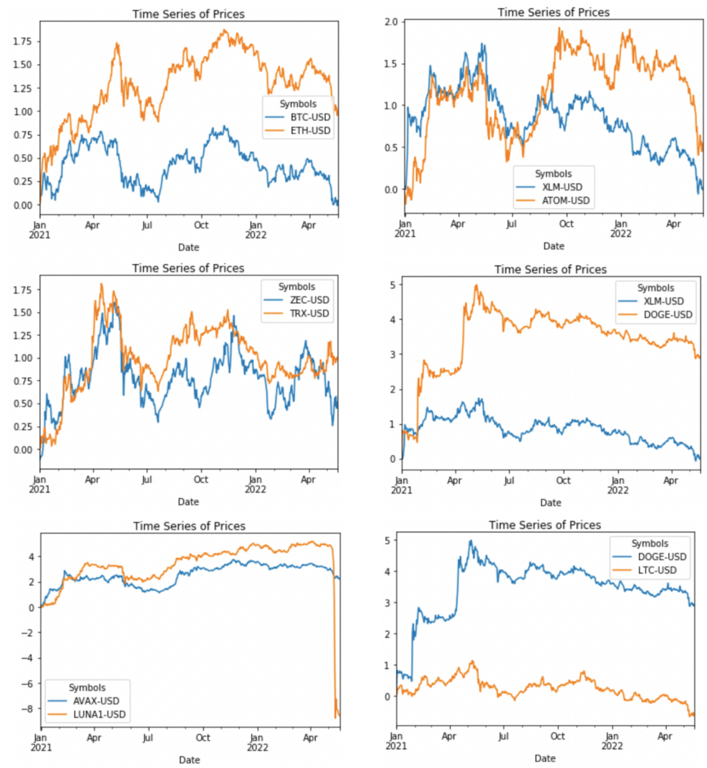


Figure 10: GitHub and market dynamics of random pairs of cryptocurrencies

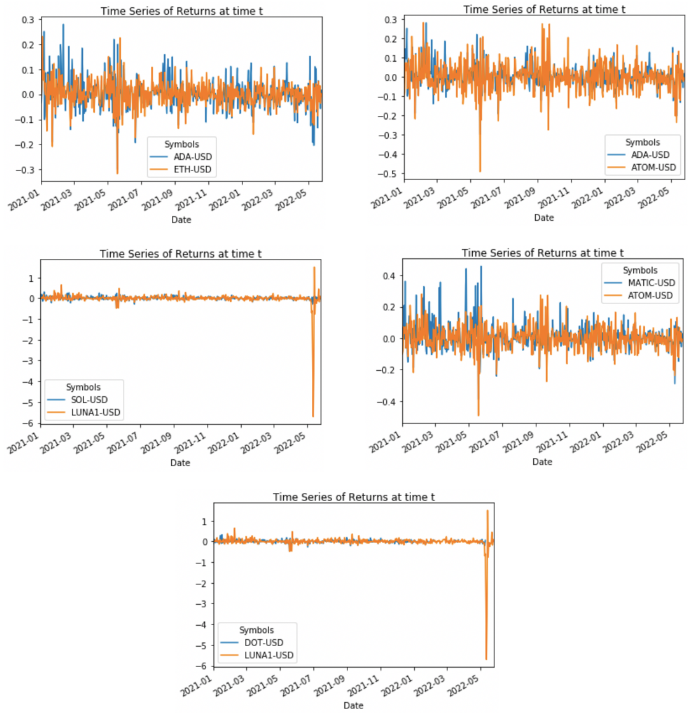


Figure 11: Cryptocurrency asset return at time t: Linked Pairs

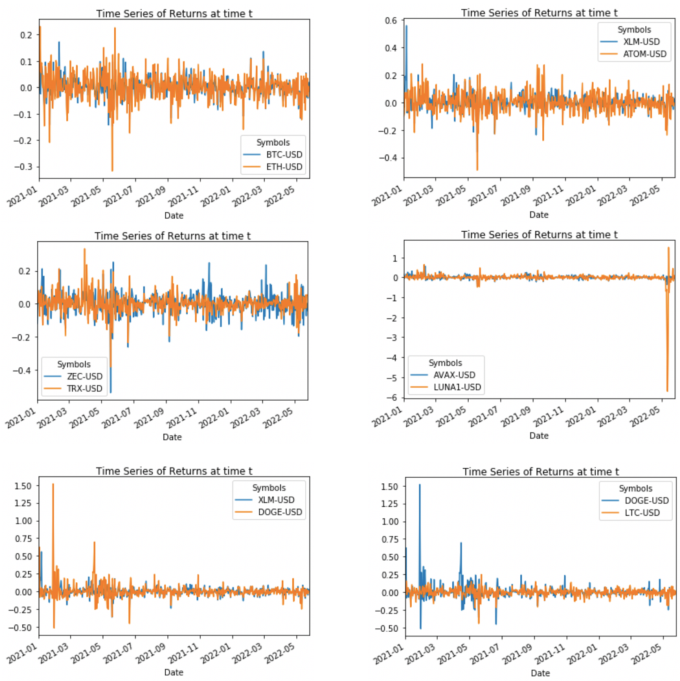


Figure 12: Cryptocurrency asset return at time t: Random Pairs

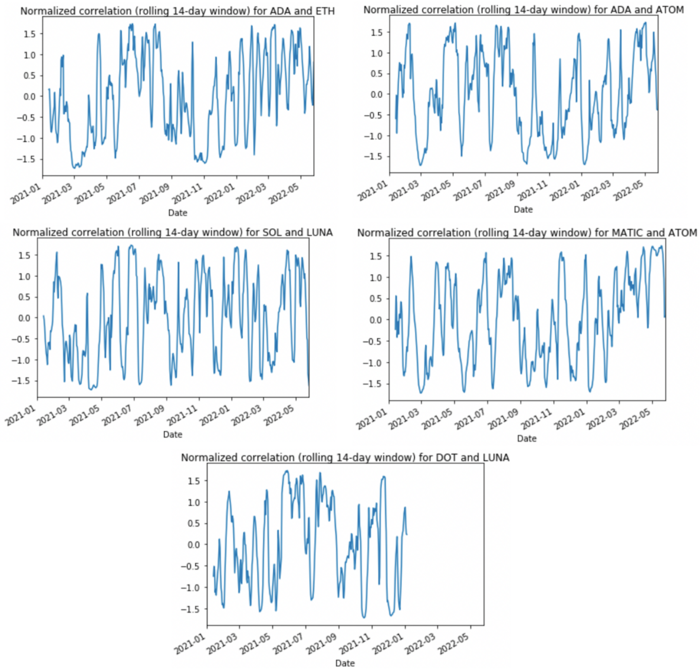


Figure 13: Normalized Correlations of Cryptocurrency Linked Pairs (rolling window of 14 days)

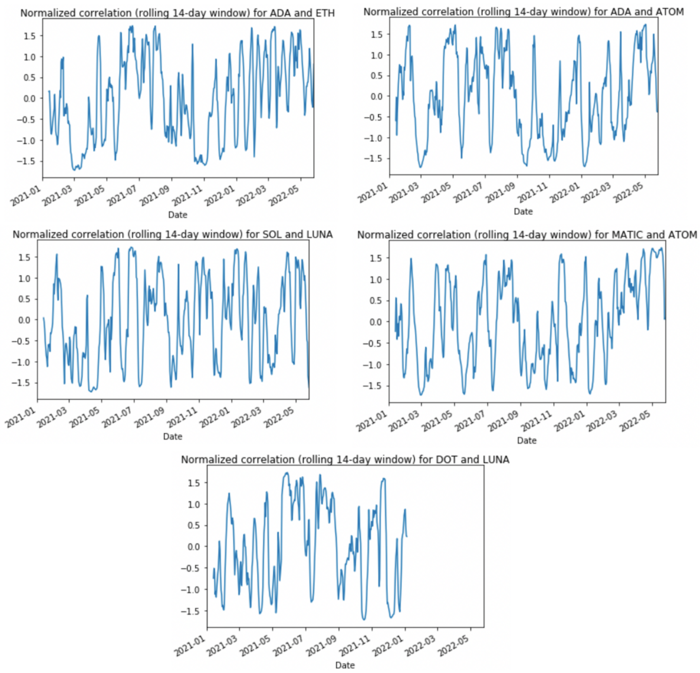


Figure 14: Normalized Correlations of Cryptocurrency Random Pairs (rolling window of 14 days)

References

Hirotugu Akaike. 1976. Canonical correlation analysis of time series and the use of an information criterion. In Raman K. Mehra and Dimitri G. Lainiotis, editors, *System Identification Advances and Case Studies*, volume 126 of *Mathematics in Science and Engineering*, pages 27–96. Elsevier.

CoinCodex. 2022. Live crypto prices and cryptocurrency market cap. https://coincodex.com/. Accessed: 2022-05-20.

CoinMarketCap. 2022. ryptocurrency prices by market cap. https://coinmarketcap.com/. Accessed: 2022-05-20.

Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in github: Transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 1277–1286, New York, NY, USA. Association for Computing Machinery.

Fernando García-Monleón, Ignacio Danvila del Valle, and Francisco Lara. 2021. Intrinsic value in crypto currencies. *Technological Forecasting and Social Change*, 162:120393.

GitHub. 2022. Gh archive. https://www.gharchive.org/. Accessed: 2022-05-18.

Github. 2022. Github docs. https://docs.github.com/en/developers. Accessed: 2022-05-18.

Giancarlo Giudici, Alistair Milne, and Dmitri Vinogradov. 2019. Cryptocurrencies: market analysis and perspectives. *Journal of Industrial and Business Economics*, 47.

Wilhelm Kirch, editor. 2008. *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht.

Ladislav Kristoufek. 2014. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PLoS ONE*.

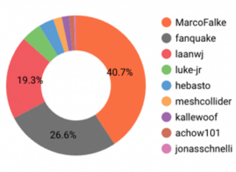
Lorenzo Lucchini, Laura Alessandretti, Bruno Lepri, Angela Gallo, and Andrea Baronchelli. 2020. From code to market: Network of developers and correlated returns of cryptocurrencies. *Science Advances*, 6(51):eabd2204.

Jeremy Miles. 2014. *R Squared, Adjusted R Squared*. John Wiley Sons, Ltd.

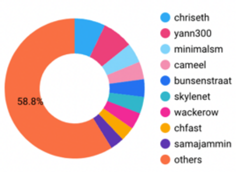
Asher Trockman, Rijnard van Tonder, and Bogdan Vasilescu. 2019. Striking gold in software repositories? an econometric study of cryptocurrencies on github. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 181–185.

Xin Yan and Xiao Gang Su. 2009. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., USA.

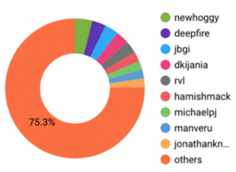
Appendix A. Contribution share of 15 cryptocurrency projects



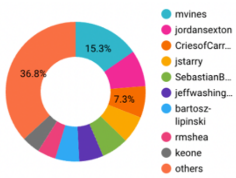
(a) Bitcoin



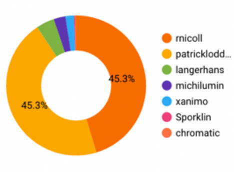
(b) Ethereum



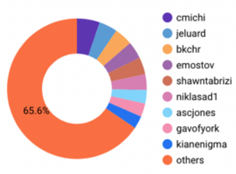
(c) Cardano



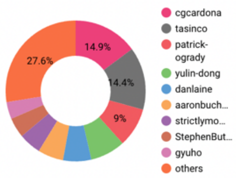
(d) Solana



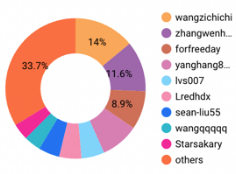
(e) Dogecoin



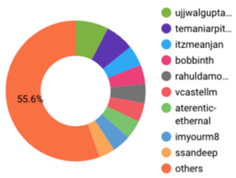
(f) Polkadot



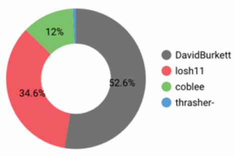
(g) Avalanche



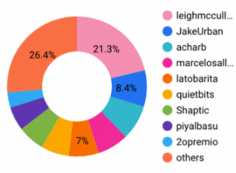
(h) TRON



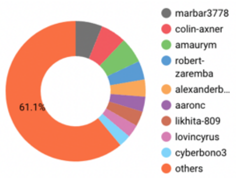
(i) Polygon



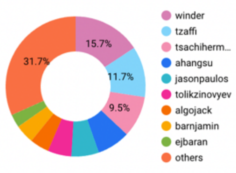
(j) Litecoin



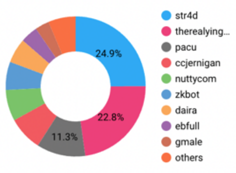
(k) Stellar



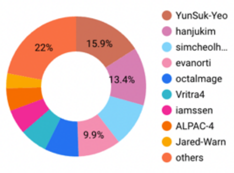
(l) Cosmos



(m) Algorand

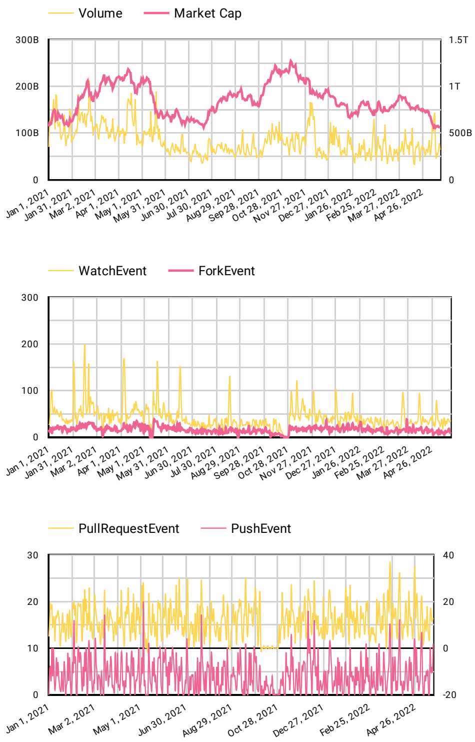


(n) Zcash

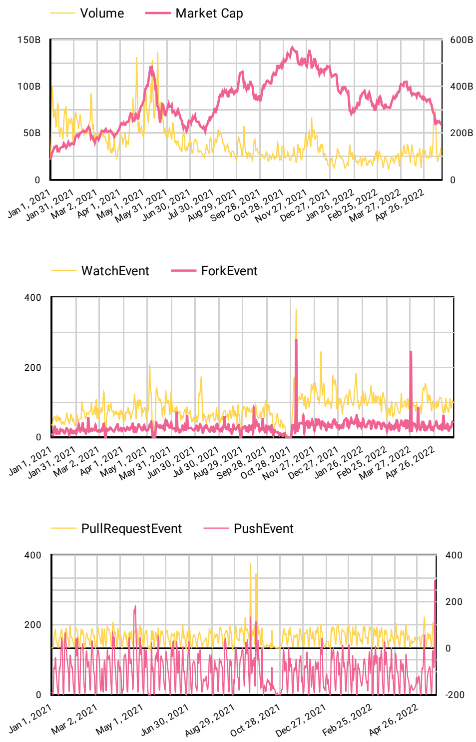


(o) Terra

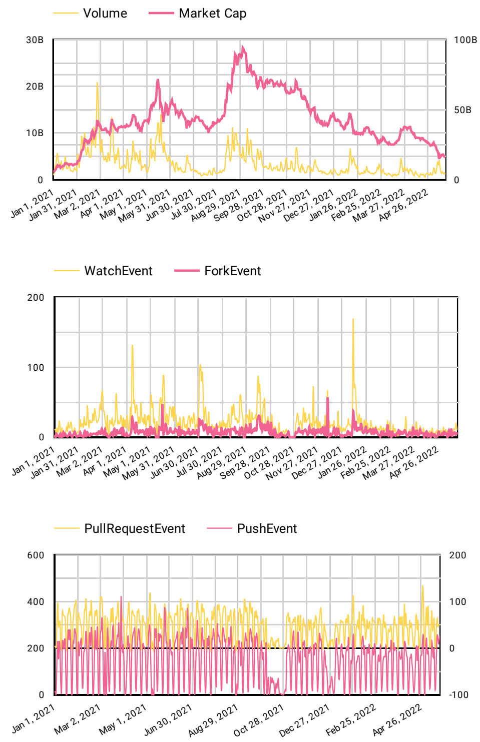
Appendix B. GitHub and market dynamics of 15 cryptocurrency projects



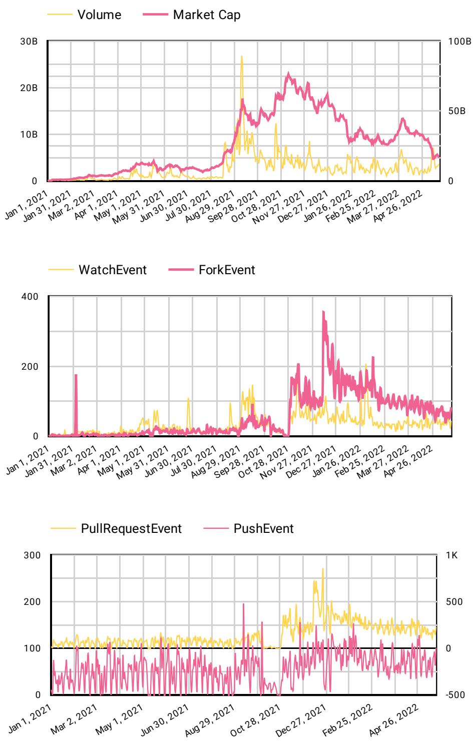
(a) Bitcoin



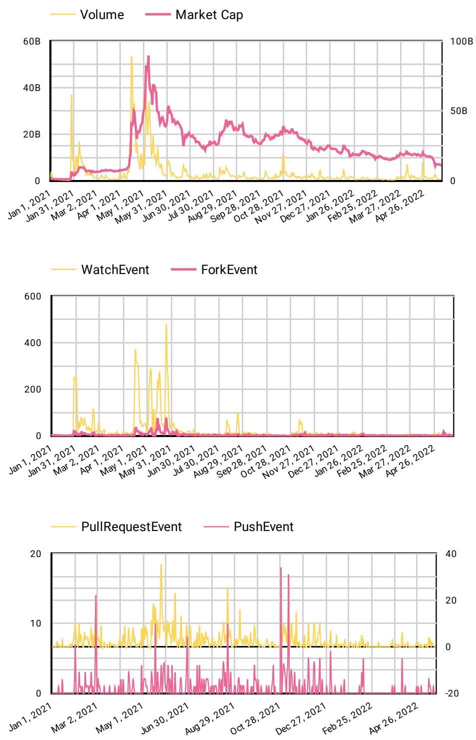
(b) Ethereum



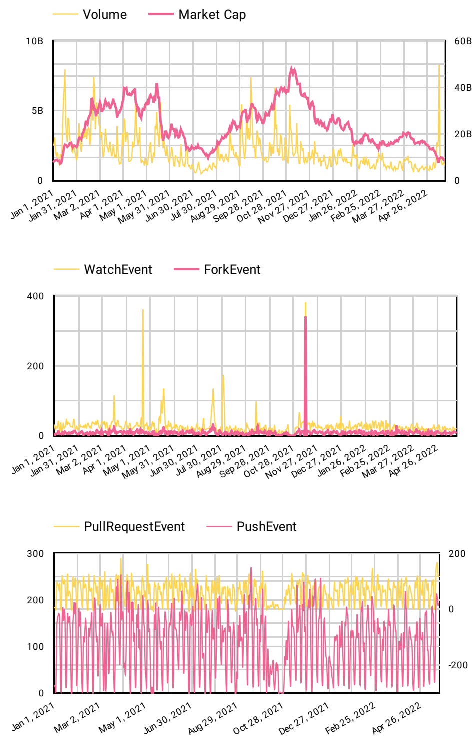
(b) Cardano



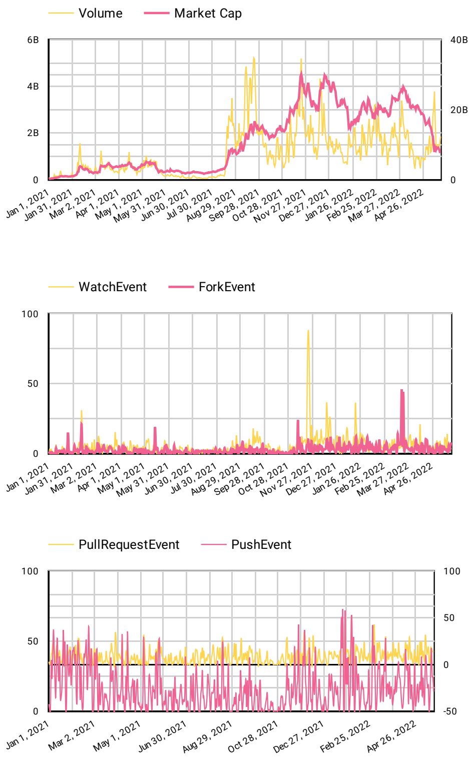
(d) Solana



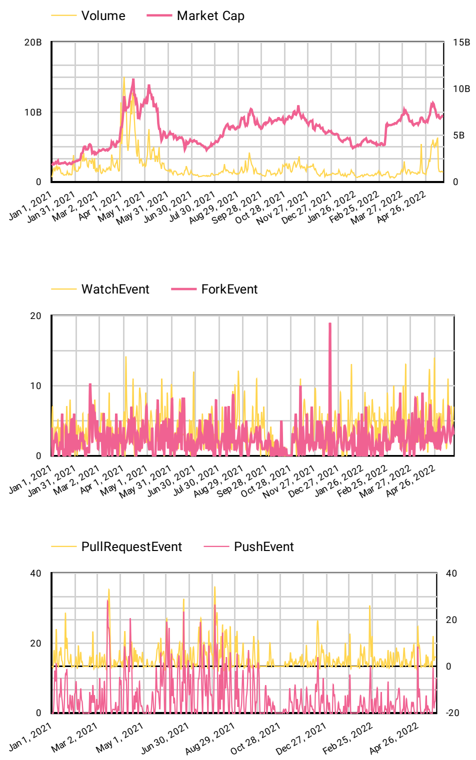
(e) Dogecoin



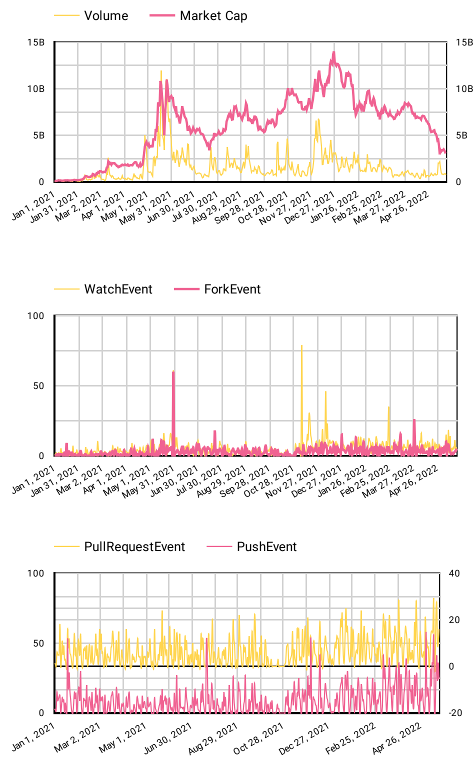
(f) Polkadot



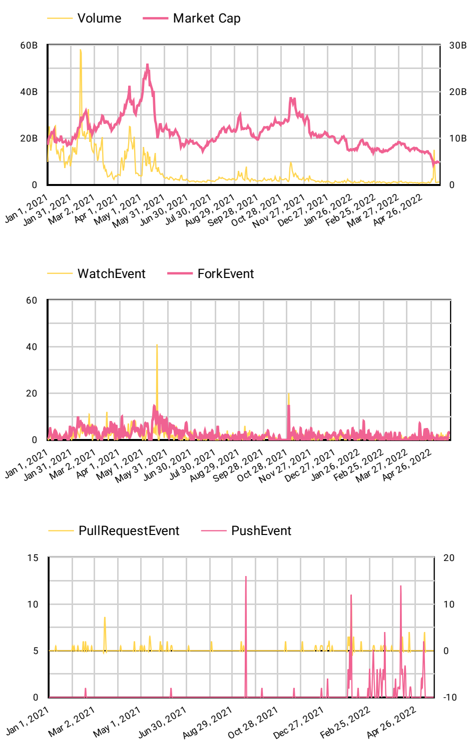
(g) Avalanche



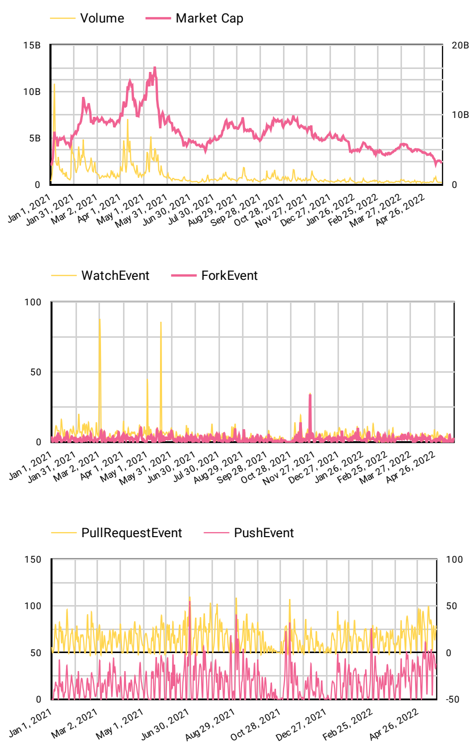
(h) TRON



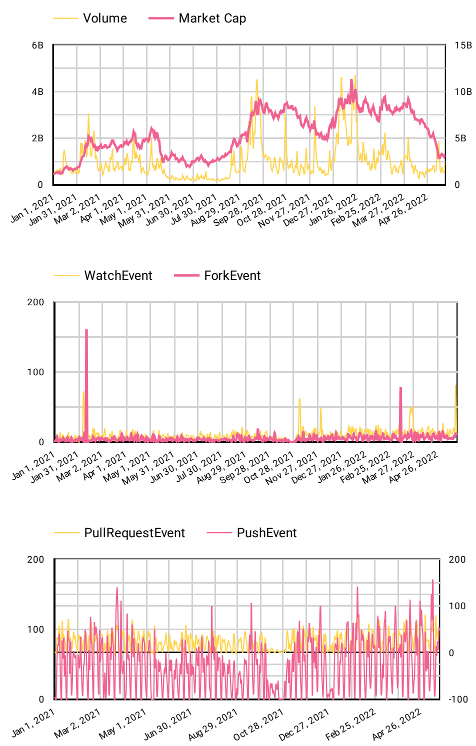
(i) Polygon



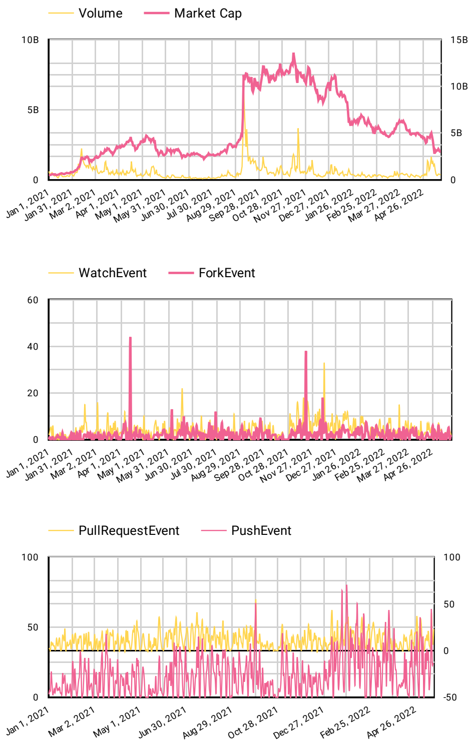
(j) Litecoin



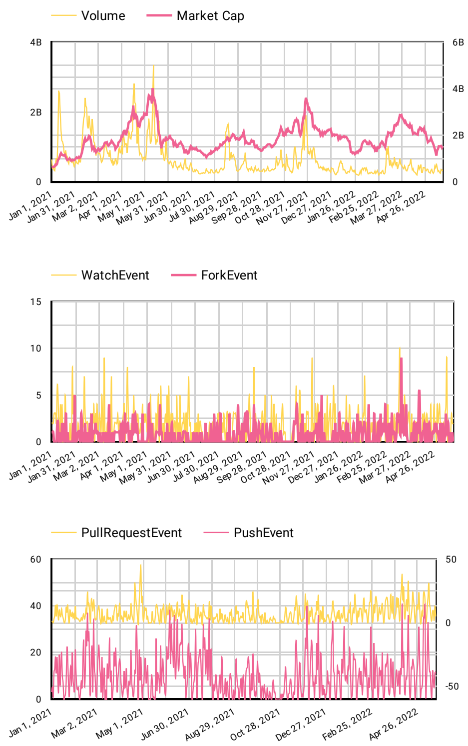
(k) Stellar



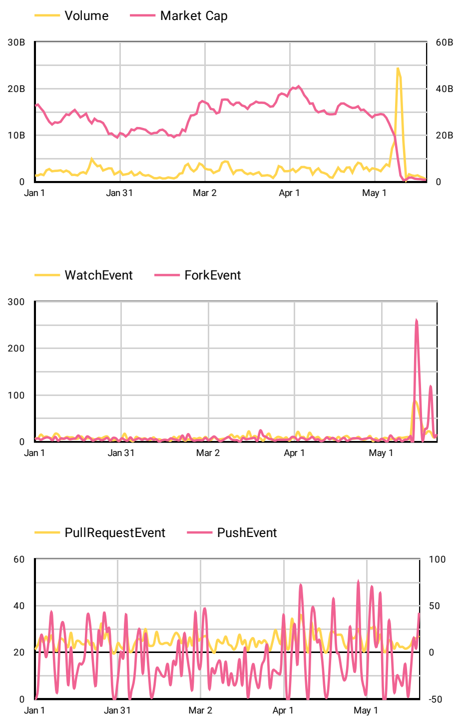
(l) Cosmos



(m) Algorand



(n) Zcash



(o) Terra