

Analysis of the predictability power of Ethereum blockchain data

1st Giro Tomas
EPFL MFE student
Lausanne, Switzerland
tomas.girolarraz@epfl.ch

2nd Meylan Valentin
EPFL MFE student
Lausanne, Switzerland
valentin.meylan@epfl.ch

Abstract—We investigate if we can find correlations between deposits resulting from transactions to exchanges' wallets, and price returns of **ERC-20** tokens. We find no evidence of predictability of returns with a linear regression model. Furthermore, we find non-negligible predictability power for daily volatility.

Index Terms—ERC-20, cryptocurrency, Ethereum, blockchain, prediction, volatility, returns.

INTRODUCTION

Tether issuance is linked with Bitcoin price movements, according to M. Griffin & Shams [1]. They show evidence of unbacked digital money (Tether) inflating cryptocurrency prices. From these findings, we wondered whether on-chain transactions for ERC-20 tokens had predictive power on off-chain movements, with respect to price returns and volume. We analyzed if deposits directed towards exchanges wallets had any predictive power on ERC-20 token returns and volatility on centralized exchanges. We start our analysis by analysing the returns of the token POLY. Then, we compare our results with two other tokens THETA and ENJ.

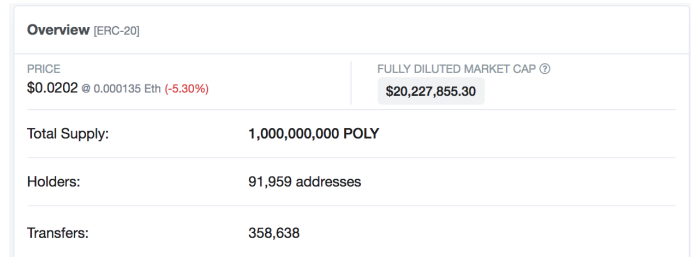
We decided to focus on these three tokens because they have a relatively small market cap in the order of magnitude of 50 million dollars. From Bogousslavsky & Collin-Dufresne [2], we know that "microstructure models based on Kyle, 1985, [3] suggest that higher volume should be associated with lower trading costs, as volume is mainly driven by uninformed trading which reduces adverse selection risk". Our naive assumption is that relatively big on-chain transactions could constitute a predictor of future off-chain transactions' volume, namely trades on centralized exchanges.

I. ANALYSIS WITH POLY

A. Presentation of Polymath network

The following paragraph explains the motivation behind the token POLY: create a platform to simplify selling security tokens. Securities represent an ownership position in a publicly-traded corporation, a creditor relationship with a governmental body or corporation, or rights to ownership. A security token is a tokenized, digital form of these traditional securities.

"Polymath (POLY) is creating a global platform for issuing and investing in securities tokens. Polymath's standard for blockchain security tokens aims to integrate the necessary regulatory requirements into smart contracts and comply with



Overview [ERC-20]	
PRICE \$0.0202 @ 0.000135 Eth (-5.30%)	FULLY DILUTED MARKET CAP ⓘ \$20,227,855.30
Total Supply:	1,000,000,000 POLY
Holders:	91,959 addresses
Transfers:	358,638

Figure 1: Main Poly characteristics as presented in etherscan.io

regulations. The project simplifies the legal process of creating and selling security tokens. It establishes a new token standard (ST20) and enforces compliance by whitelisting authorized investors and their Ethereum wallet addresses. The POLY token is used for payments on the platform, which facilitates exchanges between issuers, investors, service providers, and developers." ¹

B. Data acquisition process

1) *Transaction data*: We found a detailed dataset aggregating the whole Ethereum blockchain: all transactions for all ERC-20 tokens (see Google BigQuery [Ethereum blockchain data](#)). The dataset size is around 850 GB and contains the set of all Ethereum blocks and their attributes. The data is exported regularly using this [GitHub](#) open source code. The dataset contains a token_transfers table with the following columns:

- token_address (ERC-20 token address);
- from_address (address of the sender);
- to_address (address of the receiver);
- value (amount of tokens transferred);
- transaction_hash;
- log_index (log index in the transaction receipt);
- block_timestamp (timestamp of the block where this transfer was in);
- block_number (block number where this transfer was in);
- block_hash (hash of the block where this transfer was in);

The dataset is hosted on Google Cloud by Google Big Query.

¹source: [coinmarketcap](#)

Using [this query](#), we obtained all POLY input transactions delivered to Binance addresses, since the token's quoting. We found 15,000 transactions, the first occurring on February 2018, while the last had been registered on October 2019. This correlates with the timeline of its Initial Coin Offering (ICO) which occurred on January 13th, 2018. For illustration purposes we included the SQL query in figure 2.

```
SELECT
  *
FROM
  `bigquery-public-data.crypto_ethereum\
  token_transfers` AS transfers
WHERE to_address IN
  ('0x3f5ce5fbfe3e9af3971dd833d26ba9b5c936f0be',
  '0xd551234ae421e3bcb99a0da6d736074f22192ff',
  '0x564286362092d8e7936f0549571a803b203aaced',
  '0x0681d8db095565fe8a346fa0277bffd9c0edbbf',
  '0xfe9e8709d3215310075d67e3ed32a380ccf451c8',
  '0x4e9ce36e442e55ecd9025b9a6e0d88485d628a67',
  '0xbe0eb53f46cd790cd13851d5eff43d12404d33e8',
  '0xf977814e90da44bfa03b6295a0616a897441acec')
AND token_address=\
  '0x9992ec3cf6a55b00978cddf2b27bc6882d88d1ec'
```

Figure 2: SQL query to get all transfers of Polymath to Binance. Binance has 8 Ethereum wallets. token_address refers to Polymath's smart contract.

2) *Price data*: We queried hourly data using the [API](#) of Crypto Compare. By iterating requests of 2000 candles, we downloaded the whole hourly price data available. The code of the requests can be found [here](#). We downloaded the price data in Ethereum amounts because it was the most direct pricing method. To get the price in USD the API would have to convert into the traded currency (usually BTC / ETH), then to USD.

C. Data Visualization

We started using data after September 2018, as the transaction volume on Binance prior to that date was null. We also removed two data points that were 50-sigma event on the return of POLY. It was due to a sudden variation of POLY price for a single datapoint. We aggregated on-chain transactions into hourly bins. We obtained the hourly volume of deposits to Binance wallets. Additionally, we computed the daily volumes of POLY deposit-transactions to Binance wallets. We plotted the three time-series in Figure 3.

We can observe the Autocorrelation and Partial Autocorrelation of POLY returns in figure 4. We see that there is a significative (-25%) negative correlation between a return in one hour and the next. We can also observe that there is a small negative autocorrelation between one day and the next (-2%).

D. Regression

In order to find and test meaningful relationships between the transaction data and the price data, we carried out several

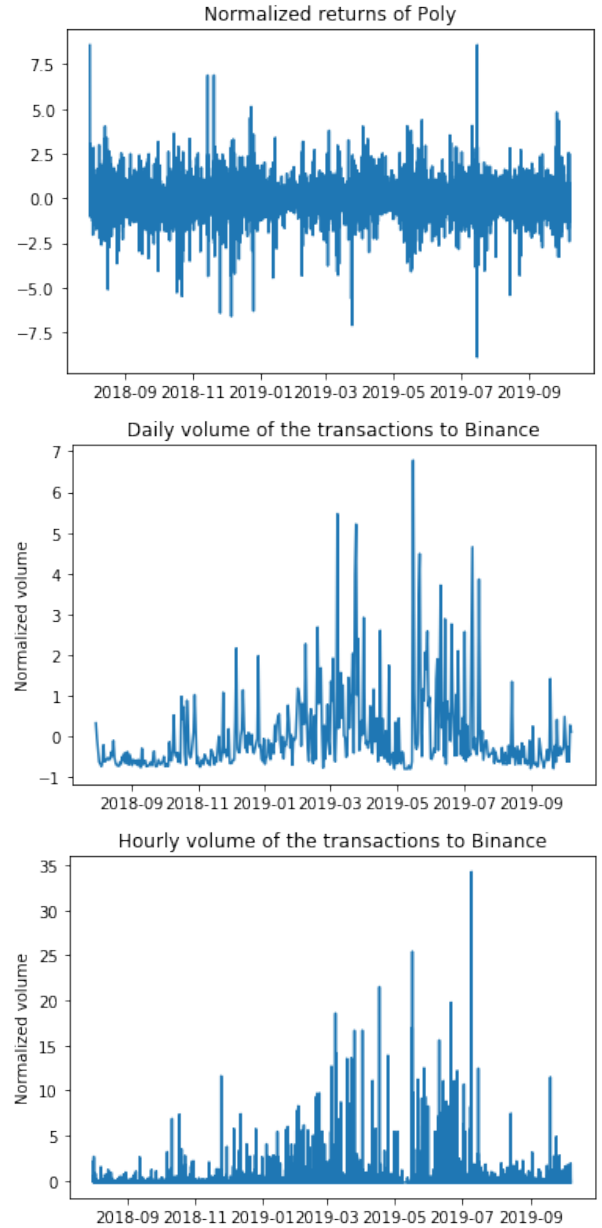


Figure 3: Normalized returns of POLY, and daily and hourly on-chain transaction volume from unknown accounts to Binance accounts

ordinary linear regressions.

1) *Returns vs. Transaction volume* : We regressed the normalized hourly returns with the normalized on-chain transaction volume as described in this formula (that is the volume of the blockchain transactions to Binance addresses supposedly to load Binance accounts):

$$R_{\text{hourly}} = \alpha + \beta V_{\text{hourly}}^{\text{on chain}}$$

The regression results can be found in figure 5. They show no significant relationship. We tried in the same way the same regression for daily returns and transaction volume of the same day. We found nothing significant.

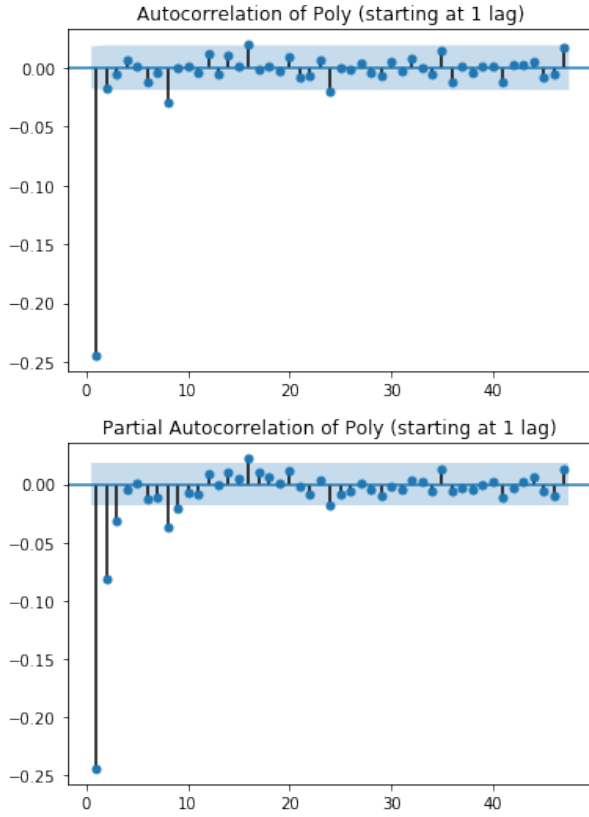


Figure 4: ACF and PACF of POLY returns

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.000			
Model:	OLS	Adj. R-squared:	-0.000			
Method:	Least Squares	F-statistic:	0.1004			
Date:	Sat, 12 Oct 2019	Prob (F-statistic):	0.751			
Time:	19:45:03	Log-Likelihood:	-12668.			
No. Observations:	10395	AIC:	2.534e+04			
Df Residuals:	10393	BIC:	2.535e+04			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.0011	0.008	-0.142	0.887	-0.017	0.015
x1	0.0001	0.000	0.317	0.751	-0.001	0.001
Omnibus:		1952.597	Durbin-Watson:		2.275	
Prob(Omnibus):		0.000	Jarque-Bera (JB):		50790.074	
Skew:		-0.181	Prob(JB):		0.00	
Kurtosis:		13.823	Cond. No.		17.4	

Figure 5: Normalized returns and POLY transactions' volume to Binance wallets. The regressors are not significant. The data and regression can be found on this [GitHub](#)

2) *Volatility vs. Transaction volume*: We regressed also the daily volatility of the POLY price quoted in ETH and the daily transaction volume. To do so, we tested the following linear regression model where σ_{daily} stands for the volatility and $V_{\text{daily}}^{\text{on chain}}$ for the transaction volume.

$$\sigma_{\text{daily}} = \alpha + \beta V_{\text{daily}}^{\text{on chain}}$$

The regression results show an extremely significant relationship. With a t-statistic of 6.3 for the proportionality coefficient.

The results of the regression can be found on figure 6.

OLS Regression Results						
Dep. Variable:	std	R-squared:	0.085			
Model:	OLS	Adj. R-squared:	0.083			
Method:	Least Squares	F-statistic:	40.15			
Date:	Sat, 12 Oct 2019	Prob (F-statistic):	5.91e-10			
Time:	19:27:53	Log-Likelihood:	-161.93			
No. Observations:	434	AIC:	327.9			
Df Residuals:	432	BIC:	336.0			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.7362	0.017	43.361	0.000	0.703	0.770
x1	0.3794	0.060	6.337	0.000	0.262	0.497
Omnibus:		145.021	Durbin-Watson:		1.149	
Prob(Omnibus):		0.000	Jarque-Bera (JB):		448.765	
Skew:		1.559	Prob(JB):		3.56e-98	
Kurtosis:		6.885	Cond. No.		3.54	

Figure 6: Volatility of returns vs. POLY transaction volume to Binance wallets. The findings are extremely significant and indicate a relationship. The data and regression can be found on this [GitHub](#).

II. ANALYSIS OF THETA

In this section we compare our previous results with the same analysis for the token THETA.

A. Presentation of THETA

The following extract describes the purpose of this token: to finance a P2P network for a modern video sharing platform.

Overview [ERC-20]	
PRICE \$0.0719 @ 0.000482 Eth (-2.60%)	FULLY DILUTED MARKET CAP ⓘ \$71,856,187.80
Total Supply:	1,000,000,000 THETA
Holders:	26,663 addresses
Transfers:	191,115

Figure 7: Main THETA characteristics as presented in etherscan.io

Theta aims to tackle the costly, centralized and inefficient video streaming infrastructures. Their blockchain technology solution provides a decentralized, peer-to-peer content delivery network to bypass costly infrastructure and rewards users allocating their bandwidth and resources to the network, ensuring its continuous high performance. ²

B. Comparison of the autocorrelation

We can observe the Autocorrelation and Partial Autocorrelation of THETA returns in figure 8. On first sight, the figures look quite different than the ACF and PACF from POLY. The negative correlation between one return and the next for THETA is the half of the one for POLY (-11%). This negative correlation is followed by a positive correlation of 7%, and no significant effect from one day to the next (at the 24th lag).

²source: [thetatoken](#)

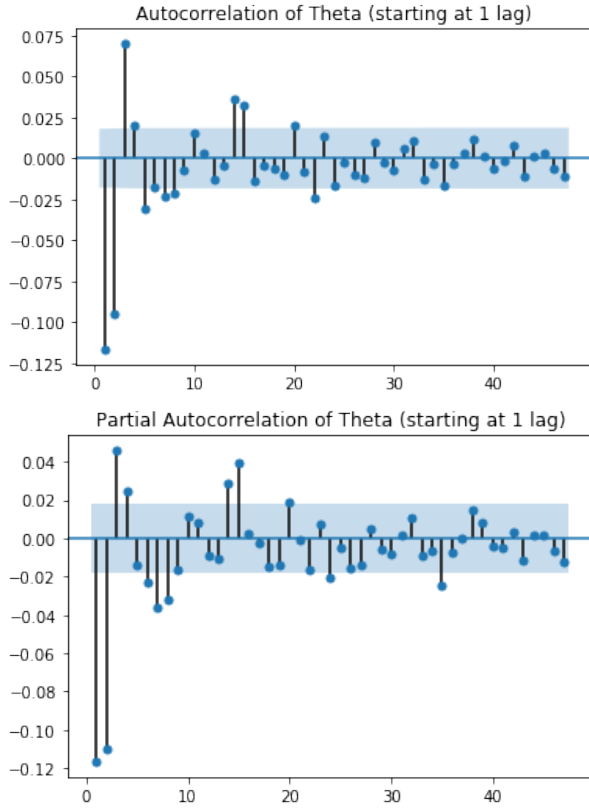


Figure 8: ACF and PACF of THETA returns

C. Comparison of the results of the return predictability

In the first regression, we compare, as before, hourly returns of THETA and hourly volume of on-chain transactions of THETA towards exchanges. We also do not find any predictability with an R^2 of 0.002. The results of the regression can be seen in figure 10.

D. Comparison of the results of the volatility predictability

In the second regression, we compare, as before, daily variance of hourly returns of THETA and daily volume of on-chain transactions of THETA towards exchanges. We find a lot of correlation an R^2 of 0.62 and a p-value of 0.000. The results of the regression can be seen in figure 11.

III. ANALYSIS OF ENJ

Finally we compare our results of the two other tokens with the token ENJ.

A. Presentation of the token ENJ

The following paragraphs present the purpose of the Enjin token: a token for gamers to purchase in-game assets, and for gaming companies to grow revenue with blockchain in-game transactions. The main characteristics of the Enjin token are presented in figure 12.

Enjin token (ENJ) is a blockchain platform focusing on the video game industry. Enjin allow in-game assets backing via

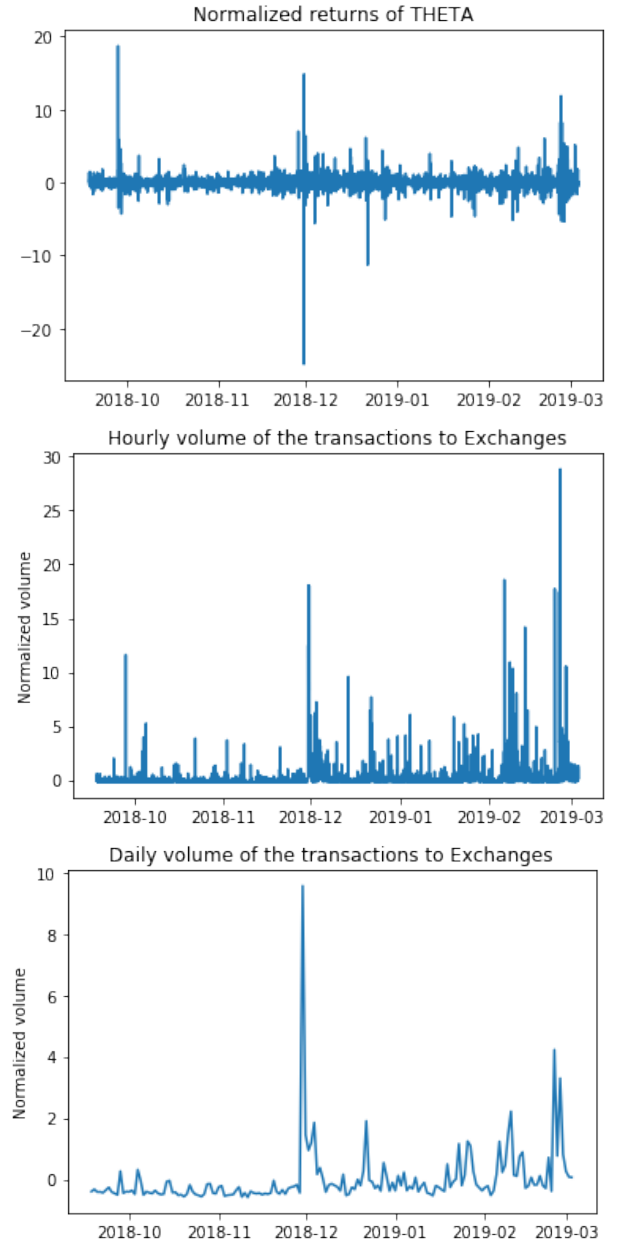


Figure 9: Normalized hourly returns of THETA, and daily and hourly on-chain transaction volume from unknown accounts to Exchange accounts

their ENJ tangible tokens, increasing the exchange gateways while maintaining consumer trust and item authenticity.³

B. Comparison of the autocorrelation

We can observe the Autocorrelation and Partial Autocorrelation of ENJ returns in figure 13. The pattern is similar to the previous two tokens: There is a small negative correlation from one return to the next in the next hour (-7%). There is a positive correlation from one return to the one 7 hours later

³enjin.io

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.002			
Model:	OLS	Adj. R-squared:	0.002			
Method:	Least Squares	F-statistic:	7.915			
Date:	Mon, 25 Nov 2019	Prob (F-statistic):	0.00493			
Time:	12:23:01	Log-Likelihood:	-5875.2			
No. Observations:	4000	AIC:	1.175e+04			
Df Residuals:	3998	BIC:	1.177e+04			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	0.0063	0.017	0.379	0.705	-0.026	0.039
x1	0.0023	0.001	2.813	0.005	0.001	0.004
=====						
Omnibus:	2008.351	Durbin-Watson:	2.184			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2624404.291			
Skew:	-0.822	Prob(JB):	0.000			
Kurtosis:	128.474	Cond. No.	20.1			
=====						

Figure 10: Normalized hourly returns vs. THETA transaction volume to Exchange wallets. The regressors are not significant.

OLS Regression Results						
=====						
Dep. Variable:	std		R-squared:	0.622		
Model:	OLS		Adj. R-squared:	0.620		
Method:	Least Squares		F-statistic:	270.2		
Date:	Mon, 25 Nov 2019		Prob (F-statistic):	1.68e-36		
Time:	12:23:02		Log-Likelihood:	-99.292		
No. Observations:	166		AIC:	202.6		
Df Residuals:	164		BIC:	208.8		
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
const	0.6116	0.036	17.068	0.000	0.541	0.682
x1	0.8819	0.054	16.437	0.000	0.776	0.988
=====						
Omnibus:	165.692		Durbin-Watson:	1.735		
Prob(Omnibus):	0.000		Jarque-Bera (JB):	4098.798		
Skew:	3.627		Prob(JB):	0.000		
Kurtosis:	26.238		Cond. No.	1.65		

Figure 11: Volatility of hourly returns vs. THETA transaction volume to Exchange wallets. The findings are extremely significant and indicate a relationship.

(around 4%). And finally there is a small negative correlation from one return to the one one day later (-5%).

All in all, there is no big significant correlation between returns.

C. Comparison of the results of the return predictability

In the first regression, we compare, hourly returns of ENJ and hourly volume of on-chain transactions of ENJ towards exchanges. We also do not find any predictability with an R^2 of 0.001. The results of the regression can be seen in figure 15.

D. Comparison of the results of the volatility predictability

In the second regression where we compare, daily variance of hourly returns of ENJ and daily volume of on-chain transactions of ENJ towards exchanges. We find a lot of correlation an R^2 of 0.39 and a p-value of 0.000. The results of the regression can be seen in figure 16.

CONCLUSION

We have seen that the linear regression model is not able to predict the returns of these ERC-20 tokens. On the other

Overview [ERC-20]	
PRICE	FULLY DILUTED MARKET CAP ⓘ
\$0.0528 @ 0.000353 Eth (-7.13%)	\$52,845,876.40
Total Supply:	1,000,000,000 ENJ
Holders:	30,856 addresses
Transfers:	624,481

Figure 12: Main ENJ characteristics as presented in etherscan.io

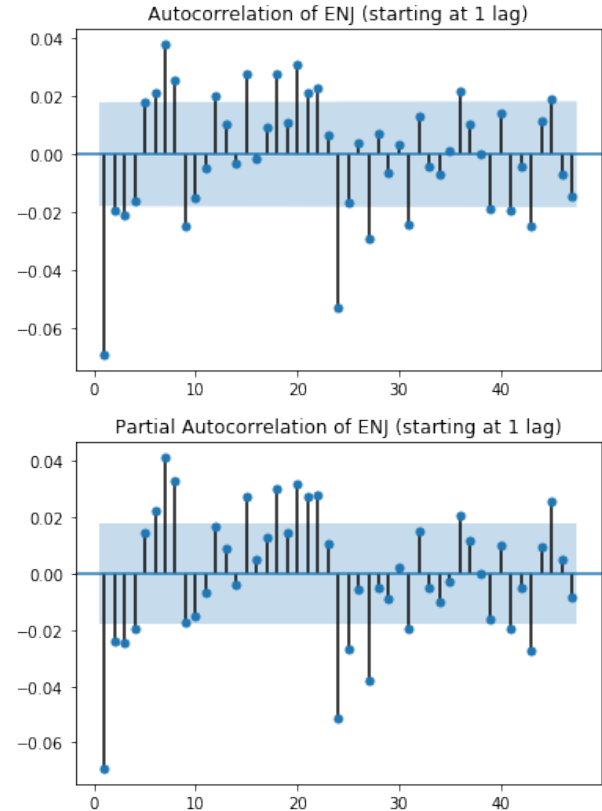


Figure 13: ACF and PACF of ENJ returns

hand, volatility is correlated with the on-chain volume directed towards centralized exchanges.

We believe that these results were expected. In fact, the predictive power of on-chain data for crypto-asset returns might be negligible compared to other variables, e.g. market microstructure or news, especially for smaller market capitalizations. However, as the share of monetary supply credited to (liquid and trustworthy) exchanges is high, it does not seem implausible that return volatility can be explained.

Furthermore, on-chain transactions are arguably costly and bear execution risks, see congestion related problems for [Bitcoin](#), and the case of Ethereum's [cryptokitties](#). Additionally, usually both deposits towards- and withdrawals from exchanges have their set of constraints, e.g. maximum deposit

or withdrawal amount on a daily basis. These characteristics are common to all three concerned tokens. Therefore, on-chain transactions towards centralized exchanges seem to constitute a valuable attribute with respect to prediction of returns' volatility.

While all three analyzed tokens are relatively small in terms of market capitalization ($< \$100M$), they do not share the exact same monetary characteristics. Each token's total supply is one billion token, as seen in, 1, 7, 12. However, their circulating supply varies greatly: in the order of 443M⁴ for POLY, 870M⁵ for THETA and 784M⁶ tokens in circulation for ENJ. Analyzing the distribution of tokens among its holders, as well as the distribution of the number of tokens traded at each on-chain transaction, could highlight the transactions which are most likely to explain the volatility of hourly returns.

Other directions to explore would be to correlate the atomicity of holders, their first on-chain activity prior to exchange deposits, and how that is susceptible to influence transaction volume both on-chain and off-chain. One might be able to derive additional attributes to explain both our current results and data observations.

Studying the release schedule of the currently non-circulating supply could be an additional dimension to look at to have a more thorough explanation of our regression results. In fact, depending on the reason for which this supply is currently non-circulating, outlining an exact explanation of the regression results once we incorporate on-chain data seems less straightforward. We could have for instance, bulk monetary supply emission following the end of a stakeholder' token lock-up period. The existing holders would technically be diluted. It is unclear how this would impact returns and volatility both long and short term. At last, an aggregate analysis of relevant ERC-20 tokens might also yield different results depending on the implementation of [token burning](#) mechanisms to slow down or reduce the monetary supply, or incentivization mechanisms resulting in supply release or monetary creation to selected economic agents.

REFERENCES

- [1] J. M. Griffin and A. Shams, "Is Bitcoin Really Un-Tethered?", en, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3195066, Oct. 2019. [Online]. Available: <https://papers.ssrn.com/abstract=3195066>.
- [2] V. Bogousslavsky and P. Collin-Dufresne, "Liquidity, Volume, and Volatility", en, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3336171, Feb. 2019. [Online]. Available: <https://papers.ssrn.com/abstract=3336171>.
- [3] A. S. Kyle, "Continuous Auctions and Insider Trading", *Econometrica*, vol. 53, no. 6, p. 1315, Nov. 1985, ISSN: 00129682. DOI: [10.2307/1913210](https://doi.org/10.2307/1913210). [Online]. Available: <https://www.jstor.org/stable/1913210?origin=crossref>.

⁴POLY circulating supply

⁵THETA circulating supply

⁶ENJ circulating supply

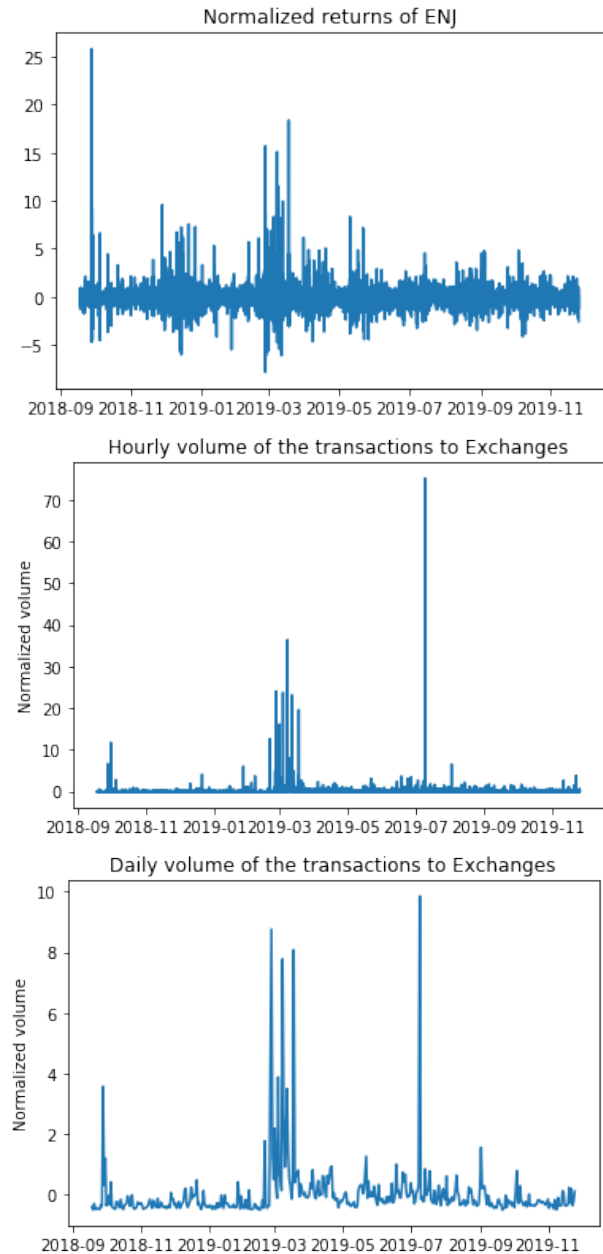


Figure 14: Normalized hourly returns of ENJ, and daily and hourly on-chain transaction volume from unknown accounts to Exchange accounts

```

=====
                        OLS Regression Results
=====
Dep. Variable:          y      R-squared:          0.001
Model:                OLS    Adj. R-squared:       0.001
Method:              Least Squares  F-statistic:      6.722
Date:                Mon, 25 Nov 2019  Prob (F-statistic): 0.00954
Time:                12:10:36  Log-Likelihood:   -14941.
No. Observations:    10397    AIC:              2.989e+04
Df Residuals:        10395    BIC:              2.990e+04
Df Model:            1
Covariance Type:      nonrobust
=====
                        coef    std err          t      P>|t|      [0.025    0.975]
-----
const          -0.0018      0.010     -0.179     0.858     -0.021     0.018
x1              0.0004      0.000      2.593     0.010      0.000     0.001
=====
Omnibus:            10342.465  Durbin-Watson:      2.100
Prob(Omnibus):      0.000    Jarque-Bera (JB):    2577883.398
Skew:               4.319    Prob(JB):            0.00
Kurtosis:           79.655    Cond. No.            58.1
=====

```

Figure 15: Normalized hourly returns vs. ENJ transaction volume to Exchange wallets. The attributes are not significant.

```

=====
                        OLS Regression Results
=====
Dep. Variable:          std  R-squared:          0.393
Model:                OLS  Adj. R-squared:      0.392
Method:              Least Squares  F-statistic:    279.3
Date:                Mon, 25 Nov 2019  Prob (F-statistic): 1.06e-48
Time:                12:10:36  Log-Likelihood:   -292.21
No. Observations:    433    AIC:              588.4
Df Residuals:        431    BIC:              596.6
Df Model:            1
Covariance Type:      nonrobust
=====
                        coef    std err          t      P>|t|      [0.025    0.975]
-----
const           0.7895      0.023     34.452     0.000      0.744     0.835
x1              1.2010      0.072     16.713     0.000      1.060     1.342
=====
Omnibus:            160.750  Durbin-Watson:      1.565
Prob(Omnibus):      0.000    Jarque-Bera (JB):    10585.579
Skew:               0.666    Prob(JB):            0.00
Kurtosis:           27.186    Cond. No.            3.14
=====

```

Figure 16: Volatility of hourly returns vs. ENJ transaction volume to Exchange wallets. The findings are extremely significant and indicate a relationship.