Online Appendix to: Aggregate Confusion In Crypto Market Data

G. Schwenkler, A. Shah, and D. Yang

April 20, 2026

This appendix provides supplementary data to the paper "Aggregate Confusion In Crypto Market Data" by G. Schwenkler, A. Shah and D. Yang. All figures and tables contained in this Online Appendix are labeled "OA." All other tables and figure are contained in the main body of the paper.

A Provider survey

We provide a survey of the data supplied by different data providers and the associated costs. We tracked the following data providers:

(1) Amberdata
(2) Brave New Coin
(3) Coin360
(4) CoinAPI
(5) CoinCap
(6) CoinGecko
(7) CoinMarketCap
(8) Coinpaprika
(9) CryptoCompare (now CCData)
(10) CryptoRank
(11) CoinStats
(12) Cryptowatch
(13) Finage
(14) Glassnode
(15) IntoTheBlock
(16) Kaiko
(17) Live Coin Watch
(18) Messari
(19) Nomics
(20) Santiment

It may seem surprising that there are so many different crypto data vendors. However, the providers serve a real market need. One of their main jobs is to aggregate data across the different platforms and exchanges in which different cryptocurrencies are traded simultaneously around the globe, and provide clean and unique time series that can be used for different analyses. This work is arduous and costly. We show in Online Appendix A.2 that the providers often raise around the median deal size or above, suggesting that the providers operate businesses that are highly capital intensive. These observations indicate that it is likely infeasible for consumers of aggregated crypto market data to carry out the aggregation on their own.

A.1 Overview and costs

In general, the data supply can be categorized as follows:

- **Broad:** The supplied data includes both market and blockchain data. There are few providers that offer broad data services. They include Amberdata, CoinMarketCap, CryptoCompare, Messari, and Santiment.

- **Narrow:** The supplied data only covers market or blockchain data, but not both. The vast majority of the providers focus on offering market data. That list includes well-known providers such as CoinGecko and Kaiko. Some providers, like Glassnode and IntoTheBlock, mostly offer blockchain data.

We break down providers based on their data supply categorization in Figure OA.1. Online Appendix B contains tables that summarize the data supplied by the different providers and provides additional details, including information about how the data series are constructed and the amount of data available at each provider.

The providers charge a varying range of fees. We summarize the fee structures in Figure OA.1. Few providers offer their data fully for free. CoinCap and Live Coin Watch are two free providers that are included in our primary list. Most providers operate under freemium models. They offer limited access for free while more extensive access can be purchased through recurring subscriptions with increasing costs. Subscriptions are mostly offered at the monthly or annual frequencies. Some providers charge expensive recurring fees in their freemium models. CoinGecko and Coinpaprika are two expensive freemium providers that are included in our primary list. However, most freemium providers strike a balance in their fee structure: the average monthly fee, across providers and subscription plans, is $613.35. There are some providers that cater exclusively to institutions without offering obvious free options. Messari is an enterprise-level provider that is also included in our primary list.

A.2 Provider fundraising

Figure OA.2 displays a normalized measure of the amount of capital raised by the providers in different fundraising rounds. Below, we summarize publicly available information about fundraising rounds closed by the different providers:

- IntoTheBlock raised $1.85 million in its seed round in 2020 according to Pitchbook data.
- Nomics raised $3 million in its Series A in 2018 from Arthur Ventures, Coinbase Ventures, and others according to Crunchbase data.
- CoinStats announced that it raised $3.2 million in its seed round in 2022.
- Amberdata raised $2 million in its seed round in 2019, $15 million in its Series A in 2021, and $30 million in its Series B in 2022 according to Crunchbase and TechCrunch.
- Messari raised $21 million in its Series A in 2021 and $35 million in its Series B in 2022 according to CoinDesk.
- Kaiko announced that it had raised $24 million in its Series A round in 2021 and $53 million in its Series B round in 2022.
- Binance purchased CoinMarketCap for $400 million in 2020, according to Decrypt.
- Santiment completed a $12 million initial coin offering (ICO) in 2017.

The amounts raised are not necessarily indicative of the book value of the operational costs of the different providers. But they suggest that the data providers in our list operate in investment-intensive business segments.

B The supplied data

We provide an overview of the available crypto data within the different providers.

B.1 Market data

Tables OA.1–OA.2 summarize the market data supplied by the different providers. Under market data, we understand any realizations of global coin-level market variables, including prices, volumes, and market capitalizations. The variables are aggregated from exchange-level trades.

Prices and market capitalizations are aggregated differently than volumes. Volumes are simply summed up across exchanges over certain time horizons. The data providers often report 24-hour rolling trading volumes both on their websites and their APIs, though the aggregation period can be adjusted through personalized API calls. But while the aggregation of volumes appears to be straight-forward, it poses serious challenges because the distribution of trading volumes across exchanges is heavy-tailed. We document in Section 3.2 that we cannot reject the null hypothesis that the distribution of cross-exchange volumes follows a power law with finite mean but infinite variance. This implies that choices that a data provider has in the aggregation process could lead to vastly different measurements of trading volumes across providers.

Prices are generally aggregated using a value-weighted average price (VWAP) approach (see Paine and Knottenbelt (2016)). This means that a data provider reports as the price of a cryptocurrency the average of the latest prices observed on exchanges, weighted by the exchange-specific volumes associated with the trades at those latest prices. Most data providers allow a user to request the latest price right before a given time.

Several providers offer historical open, high, low, close, and volume metrics (OHLCV *candles*). Unless otherwise indicated, these values are measured over the course of a day in the Coordinated Universal Time (UTC) zone. Generally, historical OHLCV data are restricted to paid customers. Many data providers require larger monthly fees for longer historical OHLCV time series. Tables OA.1 through OA.2 document how much data each provider makes available in their free plans and in their most expensive plan. We see that the providers generally limit the amount of data and the frequency with which data can be called in the free plan. Data access is generally unlimited in premium plans. Only few providers offer data that goes back earlier than 2014.

Other metrics of interest are the amount of coins that are in circulating supply as well as the implied market cap of a coin. These data are often available in real time but rarely in the history. Most enterprise-focused providers supply historical data. But few providers offer these with freemium models. The providers that offer market cap or circulating supply data in freemium models include CoinGecko, CoinMarketCap, Coinpaprika, CryptoRank, Glassnode, and Santiment.

Historical market cap data are important for portfolio applications as it allows the creation of historical ranks of cryptocurrencies by size and the computation of value-weighted portfolios. Historical market caps can naturally be constructed if the time series of prices and circulating supplies are available. However, one peculiarity of cryptocurrencies is that data providers are often inconsistent in terms of what kind of circulating supply or market cap they report.

Several cryptocurrencies, like Bitcoin, have a limited amount of tokens that will ever be issued to be in circulation. This amount of tokens is often referred to as the *maximum supply*. In contrast, the amount of tokens that has already been issued by a cryptocurrency provider is called the *circulating supply*. The circulating supply is naturally smaller than the maximum supply, often by orders of magnitude. All cryptocurrencies have a finite circulating supply of coins. But only some coins have a finite maximum supply. For example, Bitcoin has a maximum supply of 21 million coins but Ethereum has an unlimited maximum supply.

The discrepancy between circulating and maximum supply is often inherited by the reported market caps of the data providers. Some providers report the *fully diluted* market cap implied by the maximum supply of tokens. Other providers report the current market cap implied by the circulating supply of coins. And then there are some providers that mix up these two metrics and sometimes report one or the other metric. This can lead to inconsistent market cap data being supplied by the data providers. For example, CoinGecko, Coinpaprika, and Santiment appear to report circulating market cap for OKB Coin while Live Coin Watch appears to report fully diluted market cap. Furthermore, Coinpaprika appears to have reported fully diluted market caps for OKB Coin prior to July 2023 and CoinGecko appears to have stopped reporting market cap for OKB Coin starting in October 2023. Users should be careful when collecting market cap data and ensure that the actual required metric is supplied by the data provider. Alternatively, users can construct their own market caps by collecting either circulating or maximum coin supplies and multiply these time series with the corresponding coin prices.

Tables OA.1 through OA.2 indicate that few providers also supply market metrics specific to decentralized finance (DeFi) applications, such as total value locked (TVL). Out of our

20 providers, only 5 providers supply DeFi data: Amberdata, Glassnode, IntoTheBlock, Messari, and Santiment. Even fewer providers supply transactions data that may include large movements of crypto across wallets that may be recorded on-chain (so-called *whale* transactions). They are CryptoCompare, Glassnode, IntoTheBlock, and Santiment.

All in one, we observe that the supply of market data is spotty across providers. The granularity and history of market data strongly differs across providers. Most of the enterprise-focused providers offer comprehensive and granular market data over long histories. Only some of the providers that offer accessible freemium subscription models also supply comprehensive market data with similar degrees of granularity and histories. One provider with accessible premium plans that offers all types of market data is Santiment. Another provider is Glassnode. But Glassnode focuses on collecting on-chain data so that the prices recorded by this provider do not include off-chain transactions observed in most centralized exchanges. Furthermore, Glassnode restricts itself to a limited number of coins only. As a result, the market data provided by Glassnode may not be as comprehensive as that of Santiment.

B.2 Blockchain data

Some statistical applications may require historical characteristics of the blockchains underpinning the individual cryptocurrencies. These data may include, for example, the type of algorithms used to reach consensus when verifying transactions, the number of transactions or users in a network, the fees paid for on-chain transactions on a network, and how a cryptocurrency may be classified based on its usage. These kinds of metrics have been used in academic studies of cryptocurrency markets by Biais et al. (2023), Liu and Tsyvinski (2020), Cong et al. (2025), Schwenkler and Zheng (2025), and others.

Table OA.3 provides an overview of the availability of blockchain data across providers. We see that blockchain data are generally sparse and not often supplied by the providers in our list. The providers that specifically focus on blockchain data (Glassnode and IntoTheBlock) offer comprehensive histories. However, as indicated in Tables OA.1 through OA.2, these narrow data providers do not provide comprehensive market data. This limits their usage in broad statistical applications that require both market and blockchain data.

The broad data providers of Figure OA.1 offer both market and blockchain daat minute frequencyta. Figure OA.1 shows that most of these providers cater to enterprise customers without any accessible freemium plans. A notable exception here is Santiment again, which offers broad market and blockchain data with accessible freemium models.

C Categorical factors

We repeat the exercises of Table 6 and Figure 7 of Section 4.2 by considering categorical factors. These are factors are constructed as in Section 4.2 with the difference being that we restrict the universe of the 100 largest coins of a provider to those that fall within one of the following categories: payment, product, platform, and security. We obtain categorical classifications directly from Cong et al. (2025). Tables OA.4–OA.8 and Figures OA.3–OA.5 show our results. Section 4.2.2 of the main body provides a summary.

D Provider methodologies

We provide a survey of the methodologies used by the primary providers of Section 2 to aggregate market data from exchange-level transactions.

- CryptoCompare provides an extensive summary of their methodology through a white paper. This white paper, however, has not been updated since 2022 even though CryptoCompare claims that *"the methodology is reviewed at least every quarter by the Technical Committee"*. From every tracked exchange, CryptoCompare collects the last traded price and 24-hour volumes. These transactions are then aggregated at the global level using a volume-and-time-weighted average. The aggregation weight of a transaction is proportional to the exchange-level 24-hour volume times a time decay factor for transactions older than 5 hours, reducing to 0.001 after 25 hours. There is an outlier control, removing prices that deviate by more than a factor of 4x above or below the previous aggregated price. This control only applies whenever a coin is traded in more than two exchanges, however. There are wash trading and friction controls, but details are not provided. This makes it hard to exactly replicate the methodology of CryptoCompare.

- CoinCap has a methodology webpage that describes the principles of their aggregation approach. It mentions that CoinCap considers every possible transaction, long or short, in which a coin is either the base or the quote. It also mentions that transactions from different exchanges are volume weighted. Exchange volumes are not adjusted for wash trading. CoinCap applies a specific outlier detection approach, claiming: *"If the price is more than 10 percent outside (above or below) the current global price, the data is flagged as an outlier. [...] These outlier prices will appear on the website with notation, but will not be used in the global price calculation on the digital asset."* CoinCap also applies age controls, discarding transactions that are older than 4 hours. There also appears to be human inspection of newer data. Beyond these principles, the exact aggregation approach is not provided. For example, it is not clear whether a time window is used to determine closing prices at midnight UTC, or whether it considers only the last transaction prior to the close. Without this information, it is not possible to replicate the aggregation approach.

- CoinGecko provides a comprehensive summary of their methodology through a white paper that is regularly updated, version-tracked, and controlled by a Governance Forum. The first step of their aggregation approach is to compute a universal Bitcoin Price Index (BPI) that is then used to aggregate data for different coins. The BPI is computed as a volume-weighted average of Bitcoin transactions within the previous 5 minutes. Any BTC pair is considered, and exchange rates are obtained from Open Exchange Rates. Any price that deviates by more than 4% from the median is discarded. Once the BPI has been computed, CoinGecko computes aggregate coin prices as follows. It considers the 600 most liquid trading pairs for a coin, filtering out any pair that has not been updated in 8 hours, that has zero volume or price, for which the price cannot be converted to BTC, and for which the volume is larger than 5-times the previous aggregate coin volume (meaning that some wash trading controls are in place). There are outlier controls depending of how common the coin is, including removing transactions for which the price is 100-times larger than the last aggregate price, or the price deviates substantially from the volume-weighted median. After this, transaction pairs are first converted to a BTC price using the BPI and then BTC-converted prices are aggregated using a volume-weighted approach over 5 minute intervals. The extensive documentation makes it possible to replicate the price aggregation approach of CoinGecko. For volumes, CoinGecko applies some wash trading controls but no details are provided.

- CoinMarketCap provides superficial details of their aggregation methodology on a unified website. For prices, CoinMarketCap claims to aggregate all transactions available at minute frequency. Prices are aggregated using a proprietary USD reference price. There are controls for outliers, wash trading, and market frictions, but details are scarce. This makes it impossible to replicate CoinMarketCap's price aggregation methodology. For volumes, it indicates that the only controls are excluding exchanges with no fees or transaction mining. Still, CoinMarketCap reports both adjusted and unadjusted volumes.

- Coinpaprika does not have a methodology page hosted on their website. However, we found an older Medium page from 2018 that describes principles of the approach they may still be following. From that page, we interpret that Coinpaprika follows a 24-hour-volume-weighted aggregation approach, where prices are first converted to USD or BTC using previous prices from Coinpaprika. The page does not indicate whether outlier, wash trading, or market frictions are adopted. We are therefore unable to replicate their methodology.

- Live Coin Watch only provides superficial information about their aggregation methodology in a FAQ page. There, it states that it employs a volume-weighted average to aggregate prices. There is no information about outlier, wash trading, or market friction controls. This makes it impossible to replicate their methodology.

- Messari provides basic details of their aggregation methodology in their Glossary. It states that prices are volume-weighted averages based on what they call the *"Real Volume"*. We take this to mean that Messari applies some wash trading and market friction controls. There is no indication of whether outlier controls are in place. This makes it hard to replicate Messari's price and volume aggregation approaches.

- Santiment has an extensive Academy page where they define all the different metrics they track. However, they do not provide details of their aggregation methodology, making it hard to replicate. Santiment mentions that they collect data from other providers, like CryptoCompare and CoinMarketCap (see here).

E Our aggregation approach vs alternatives

It may seem counterintuitive that our aggregation approach requires both clustering and median aggregation. A much simpler approach would be to aggregate using only the median function without needing to do any clustering beforehand (i.e., neglecting Steps 1 through 4 of Algorithm 6.1). An additional alternative approach would be to cluster as we do but then aggregate using the mean function rather than the median. In this section, we show that our aggregation approach offers benefits over these alternative approaches, especially when only few providers are available to aggregate.

We expect that the variant of our approach that aggregates using the mean instead of the median within a cluster would perform similarly. Indeed, under the conditions of Theorem 6.2, we can show that both our approach and the mean-based alternative have root mean squared errors that converge to zero as the number of providers grows large ($J \to \infty$). However, in small samples when only few providers are available, our median-based aggregation approach is likely to perform better than the mean-based aggregation approach. That is because the median function is robust to outliers while the mean function is not. We study the small sample performance of the mean-based alternative in the simulation case study of Section 6.2. We evaluate the relative root mean squared error of the aggregated market cap and close price, relative to the true circulating market cap m^* and the true close price P^*, across the 1,000 simulations in the three different scenarios. Figure OA.6 displays our findings. We find that using the mean instead of the median generally delivers equally accurate market cap estimates compared to our approach. However, estimates of the close price are less accurate using the mean-based approximation when there are few data vendors to aggregate from ($J \leq 20$). The problem is exacerbated in Scenarios 2 and 3 that pose the most challenging test for aggregation. These findings suggest that our approach may offer benefits for the aggregation of close prices over a mean-based alternative when only few data vendors are available.

Our approach also has an advantage over a simple median-only aggregation approach. By first clustering providers based on how similar their data are, our approach favors providers that are more likely to report about the same coin. This is likely to result in more accurate aggregation by excluding providers with inconsistent labels. Table 1 indicates that this occurs sufficiently often to be of concern. We measure how often our aggregation approach removes data for a coin if a provider either leaves the ID unchanged after a major fork or split, or changes the coin's ID over time. Table OA.9 summarizes our findings. We find that our aggregation approach corrects over almost 100% of the instances in which a coin has a faulty ID from a provider. By doing this, our approach ensures that we only aggregate data that belongs to the targeted coin. In contrast, a naïve median-based approach would include data from providers with faulty IDs in its aggregation, resulting in contaminated data.

We extend the analysis of the median-based alternative by considering its performance in the simulation case study of Section 6.2. Figure OA.6 shows the relative root mean squared error of the market caps and close prices aggregated using the median-based alternative in the three different scenarios. We observe that the median-based alternative generally has higher error rates when only few data vendors are available for aggregation ($J \leq 10$). We also observe that the median-based alternative generally has a lower rate of convergence as $J \to \infty$. These results indicate that Algorithm 6.1 offers benefits both in small and large samples over a simpler median-based alternatives that skips Steps 1 through 4.

References

Biais, Bruno, Christophe Bisière, Matthieu Bouvard, Catherine Casamatta and Albert J. Menkveld (2023), 'Equilibrium bitcoin pricing', *The Journal of Finance* **78**(2), 967–1014.

Cong, Lin William, George Andrew Karolyi, Ke Tang and Weiyi Zhao (2025), Crypto value, factor pricing, and market segmentation. Working Paper.

Liu, Yukun and Aleh Tsyvinski (2020), 'Risks and Returns of Cryptocurrency', *The Review of Financial Studies* **34**(6), 2689–2727.

Paine, Andrew and William J. Knottenbelt (2016), Analysis of the cme cf bitcoin reference rate and real time index, Technical report, Imperial College Centre for Cryptocurrency Research and Engineering.

Schwenkler, Gustavo and Hannan Zheng (2025), 'News-driven peer co-movement in crypto markets', *Journal of Corporate Finance* **93**, 102772.

Provider	Free plan	Premium plan
Amberdata	N / A	Includes historical OHLCV, market cap, circulating supply, and DeFi data but no transactions data. 12 months of daily data, 30 days of hourly data, and 24 hours of minutely data. Both centralized & decentralized exchanges. Going back to 2014 for Coinbase.
Brave New Coin	N / A	Includes historical OHLCV & circulating supply data, but no historical market cap, DeFi, or transactions data. Maximum 200-day trailing aggregation windows through pagination. Going back to April 2014.
CoinAPI	Same as premium, just with a lower API call limit.	Includes OHLCV but no market cap, circulating supply, DeFi, or transactions data. Maximum 100,000 days at a time going back to 2010.
CoinCap	Includes historical OHLCV but no market cap, circulating supply, DeFi, or transactions data. Minutely, hourly, daily, and weekly aggregation windows. Claims to have data going back to the "*inception of an exchange.*"	N / A
CoinGecko	Includes historical OHLCV and market cap data but no circulating supply DeFi, or transactions data. Intervals of 30 mins when requesting 1-2 days, 4 hours when requesting 3-30 days, and 4 days when requesting more than 31 days. Data go back to 2013.	Same as free plan but with higher granularity.
CoinMarketCap	Not available in the free plan	Includes historical OHLCV, market cap, and circulating supply but no DeFi or transactions data. Only daily and hourly intervals for volumes going back to September 22, 2020. Otherwise, can only call 6-years worth historical data.
Coinpaprika	Same as premium but restricted to 1-day's worth of data.	Includes historical OHLCV and market cap but no circulating supply, DeFi, or transactions data. Can only request daily data over a maximum 1-year interval. (Unlimited in the unpriced enterprise plan).
CryptoCompare	Same as premium plan but restricted to a 3-month history.	Includes historical OHLCV and transactions data but no market cap, circulating supply, or DeFi data. Minutely (max 7 days), hourly, and daily aggregation intervals. Can only request 2,000 entries at a time. Limited to a 1-year history (unlimited in the unpriced enterprise plan).
CryptoRank	No historical data.	Includes historical OHLCV and market cap but no circulating supply, DeFi, or transactions data. Limited to a 90-day history. No aggregation, OHLCV are limited to individual exchanges.

Table OA.1: *Availability of market data data across providers.* This table only contains the top third of providers in alphabetical order. There is no API for Coin360 and CryptoWatch.

Provider	Free plan	Premium plan
CoinStats	Same as premium with stricter limits.	No historical high or low prices, market cap, circulating supply, DeFi, or transactions data. 10 years of historical data otherwise. Limited numbers of requests and call rates.
Finage	N / A	Includes historical OHLCV, market cap, and circulating supply data but no DeFi or transactions data. Minutely, hourly, daily, weekly, and monthly aggregation intervals. Maximum 30,000 requests at a time. 8 years worth of historical data.
Glassnode	Same as premium but restricted to daily aggregation intervals.	Includes historical OHLCV, market cap, circulating supply, DeFi, and transactions data. Hourly and daily aggregation intervals. Supports on-chain data for BTC, ETH, LTC, and ERC-20 tokens. History goes back partially to 2009.
IntoTheBlock	No API available in the free plan.	Includes DeFi and transactions data, as well as some prices, but no OHLCV, market cap, or circulating supply. Data goes back partially to 2008.
Kaiko	N / A	Includes historical OHLCV and circulating supply but no market cap, DeFi, or transactions data. Aggregation intervals of seconds, minutes, hours, or days. Maximum 100,000 requests at a time. Data goes back to 2010.
Live Coin Watch	No historical high prices, low prices, circulating supply, DeFi, or transactions data. Otherwise, 8 years worth of history.	N / A
Messari	N / A	Includes historical OHLCV, market cap, and DeFi but no circulating supply or transactions data. Minutely, hourly, daily, and weekly aggregation intervals. Rate limits apply in all but the unpriced enterprise plan. History goes back to 2011.
Nomics	No longer operating.	
Santiment	Same as premium plan but limited to a 2-year history and with limited amount of API calls.	Includes OHLCV, market cap, DeFi, and transactions data but no circulating supply. Highest aggregation frequency is 5 minutes. Unlimited historical access, partially through 2010.

Table OA.2: *Availability of market data across providers.* This table only contains the bottom third of providers in alphabetical order. There is no API for Coin360 and CryptoWatch.

Provider	Blockchain data availability
Amberdata	Includes algorithm and network but no category data. Data only go back to 2018 for a restricted number of coins.
Brave New Coin	No blockchain data.
CoinAPI	No blockchain data.
CoinCap	No blockchain data.
CoinGecko	Only offers category data.
CoinMarketCap	Includes historical algorithm, network, and category data.
Coinpaprika	Includes algorithm and category but no network data.
CryptoCompare	Includes historical algorithm, network, and category data.
CryptoRank	Includes category data but no algorithm or network data.
CoinStats	No blockchain data available.
Finage	Includes algorithm and category data but no network data.
Glassnode	Includes historical network data but no algorithm or category data.
IntoTheBlock	Includes all blockchain data.
Kaiko	No blockchain data available.
Live Coin Watch	Only includes category data.
Messari	Includes algorithm & category data but no network data.
Nomics	No longer operating.
Santiment	Includes network data and categories but no algorithm data.

Table OA.3: *Availability of blockchain data across providers.* The same rate limits apply as in Tables OA.1–OA.2. There is no API for Coin360 and Cryptowatch.

		CG	CMC	CP	LCW	M	S	Us
(a) Payment tokens:								
Overlap	Mean	0.9983	0.9983	0.9866	0.9697	0.9972	0.9983	1.0000
	St. dev.	0.0175	0.0175	0.0539	0.0724	0.0219	0.0175	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.8000	0.8000	0.6667	0.6667	0.8000	0.8000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0003	0.0002	0.0077	0.0009	0.0003	0.0002	0.0000
	St. dev.	0.0012	0.0012	0.0610	0.0025	0.0012	0.0012	0.0000
	Median	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.0184	0.0184	0.5532	0.0201	0.0184	0.0184	0.0000
(b) Platform tokens:								
Overlap	Mean	0.9283	0.9606	0.8354	0.7037	0.9095	0.9550	0.9991
	St. dev.	0.0376	0.0300	0.1392	0.0862	0.0586	0.0353	0.0046
	Median	0.9286	0.9722	0.8780	0.7045	0.9231	0.9545	1.0000
	Min.	0.8205	0.8605	0.1429	0.4545	0.7436	0.8409	0.9730
	Max.	1.0000	1.0000	1.0000	0.9512	1.0000	1.0000	1.0000
RMSE	Mean	0.0033	0.0006	0.0099	0.0047	0.0008	0.0008	0.0000
	St. dev.	0.0025	0.0008	0.0361	0.0055	0.0007	0.0008	0.0000
	Median	0.0026	0.0004	0.0030	0.0036	0.0006	0.0005	0.0000
	Min.	0.0001	0.0000	0.0001	0.0001	0.0001	0.0000	0.0000
	Max.	0.0102	0.0062	0.3880	0.0527	0.0059	0.0062	0.0003
(c) Product tokens:								
Overlap	Mean	0.9313	0.9580	0.9086	0.7929	0.9170	0.9504	0.9992
	St. dev.	0.0794	0.0723	0.1222	0.1270	0.0772	0.0722	0.0106
	Median	0.9167	1.0000	0.9091	0.8000	0.9091	1.0000	1.0000
	Min.	0.5714	0.5714	0.0000	0.2500	0.5714	0.5714	0.8333
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0162	0.0076	0.0284	0.1260	0.0126	0.0107	0.0007
	St. dev.	0.0179	0.0183	0.0670	0.0809	0.0166	0.0201	0.0019
	Median	0.0099	0.0041	0.0086	0.1056	0.0067	0.0057	0.0001
	Min.	0.0028	0.0001	0.0002	0.0005	0.0002	0.0002	0.0000
	Max.	0.2171	0.2252	0.6346	0.5900	0.2172	0.2261	0.0197
(d) Security tokens:								
Overlap	Mean	0.8733	0.9429	0.8596	0.6776	0.8818	0.9390	1.0000
	St. dev.	0.1505	0.0994	0.1928	0.2499	0.1419	0.1052	0.0000
	Median	1.0000	1.0000	1.0000	0.6667	1.0000	1.0000	1.0000
	Min.	0.3750	0.6000	0.3333	0.1667	0.4286	0.5000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0437	0.0093	0.0742	0.0874	0.0335	0.0106	0.0009
	St. dev.	0.0387	0.0178	0.0894	0.1148	0.0376	0.0198	0.0027
	Median	0.0348	0.0017	0.0274	0.0358	0.0214	0.0020	0.0001
	Min.	0.0002	0.0001	0.0001	0.0004	0.0003	0.0002	0.0000
	Max.	0.1687	0.1272	0.3433	0.4482	0.1684	0.1257	0.0266

Table OA.4: *Summary statistics of differences for the market portfolio across categories.* This table repeats Exercise (a) of Table 6 by restricting to different token categories. It shows summary statistics of the weekly differences of the categorical market factor portfolios implied by the different providers relative to a portfolio that is constructed using daily market caps and close prices given by the median across all providers on a given day; see Section 6.4. "Overlap" measures the fraction of a provider's portfolio in a week that is also contained in the median-based portfolio during the same week. "RMSE" measures the root mean squared error of a provider portfolio weights in a week, relative to the same-week median-based portfolio weights, for those coins that are common across the two portfolios in the week. The summary statistics are computed across all weeks in the sample. For each provider, we have 311 weekly observations. The providers are abbreviated as in Table 1. Column "Us" considers the portfolios computed using data aggregated using our approach described in Section 6.

		CG	CMC	CP	LCW	M	S	Us
(a) Payment tokens:								
Overlap (l)	Mean	0.9775	0.9807	0.9091	0.7685	0.9678	0.9678	1.0000
	St. dev.	0.1486	0.1378	0.2879	0.4225	0.1767	0.1767	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	1.0000	1.0000	0.9870	1.0000	1.0000	1.0000	1.0000
	St. dev.	0.0000	0.0000	0.1134	0.0000	0.0000	0.0000	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0113	0.0095	0.0555	0.1196	0.0157	0.0154	0.0000
	St. dev.	0.0743	0.0677	0.1796	0.2218	0.0866	0.0847	0.0000
	Median	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.5000	0.5000	1.0000	0.7071	0.5000	0.5000	0.0000
(b) Platform tokens:								
Overlap (l)	Mean	0.5907	0.8197	0.4411	0.0744	0.6492	0.7695	0.9872
	St. dev.	0.1811	0.1447	0.3188	0.1923	0.2547	0.1809	0.0380
	Median	0.6250	0.8750	0.4444	0.0000	0.7500	0.8571	1.0000
	Min.	0.0000	0.2500	0.0000	0.0000	0.0000	0.2222	0.8571
	Max.	1.0000	1.0000	1.0000	0.8889	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.9172	0.9802	0.8540	0.8355	0.9589	0.9799	0.9984
	St. dev.	0.0757	0.0459	0.1829	0.1137	0.0623	0.0451	0.0137
	Median	0.8750	1.0000	0.8750	0.8750	1.0000	1.0000	1.0000
	Min.	0.7143	0.8571	0.0000	0.5556	0.7500	0.8571	0.8750
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0357	0.0214	0.0484	0.0523	0.0263	0.0251	0.0025
	St. dev.	0.0133	0.0128	0.0815	0.0111	0.0142	0.0135	0.0077
	Median	0.0378	0.0239	0.0415	0.0524	0.0261	0.0266	0.0001
	Min.	0.0001	0.0001	0.0004	0.0057	0.0004	0.0002	0.0000
	Max.	0.0671	0.0484	1.0000	0.0959	0.0559	0.0594	0.0409

Table OA.5: *Summary statistics of differences for the size portfolio across categories (Part I).* This table repeats Exercise (b) of Table 6 by restricting to different token categories. It shows summary statistics of the weekly differences of the categorical size factor portfolios implied by the different providers relative to a portfolio that is constructed using daily market caps and close prices given by the median across all providers on a given day; see Section 6.4. “Overlap” measures the fraction of a provider’s portfolio in a week that is also contained in the median-based portfolio during the same week. “RMSE” measures the root mean squared error of a provider portfolio weights in a week, relative to the same-week median-based portfolio weights, for those coins that are common across the two portfolios in the week. The summary statistics are computed across all weeks in the sample. For each provider, we have 311 weekly observations. The providers are abbreviated as in Table 1. Column “Us” considers the portfolios computed using data aggregated using our approach described in Section 6.

		CG	CMC	CP	LCW	M	S	Us
(c) Product tokens:								
Overlap (l)	Mean	0.5531	0.8264	0.6254	0.1736	0.5236	0.7658	0.9871
	St. dev.	0.3805	0.3039	0.4104	0.3461	0.3856	0.3361	0.1055
	Median	0.5000	1.0000	0.5000	0.0000	0.5000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.9352	0.9598	0.8269	0.5675	0.9421	0.9443	0.9952
	St. dev.	0.1672	0.1438	0.2801	0.4215	0.1568	0.1637	0.0489
	Median	1.0000	1.0000	1.0000	0.5000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.1785	0.0780	0.1404	0.4047	0.1714	0.0965	0.0082
	St. dev.	0.1714	0.1381	0.1500	0.2192	0.1613	0.1386	0.0548
	Median	0.1620	0.0016	0.1371	0.3793	0.1620	0.0034	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.7071	0.6325	0.6993	1.0000	0.7071	0.7071	0.6325
(d) Security tokens:								
Overlap (l)	Mean	0.5129	0.7894	0.5469	0.2540	0.6013	0.7540	0.9936
	St. dev.	0.4801	0.3808	0.4986	0.4360	0.4754	0.4073	0.0801
	Median	0.5000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.8939	0.9566	0.7120	0.6238	0.9405	0.9598	1.0000
	St. dev.	0.2603	0.1466	0.4482	0.4852	0.2020	0.1420	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.2648	0.1115	0.3194	0.5312	0.2032	0.1311	0.0037
	St. dev.	0.2693	0.2075	0.3282	0.4219	0.2553	0.2215	0.0462
	Median	0.2105	0.0000	0.2306	0.5774	0.0000	0.0000	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.8660	0.7071	1.0000	2.0000	0.8660	0.7071	0.5774

Table OA.6: *Summary statistics of differences for the size portfolio across categories (Part II).* This table repeats Exercise (b) of Table 6 by restricting to different token categories. It shows summary statistics of the weekly differences of the categorical size factor portfolios implied by the different providers relative to a portfolio that is constructed using daily market caps and close prices given by the median across all providers on a given day; see Section 6.4. "Overlap (l)" and "Overlap (s)" measure the fraction of provider portfolio's long or short legs in a week that is also contained in the corresponding leg of the median-based portfolio during the same week. "RMSE" measures the root mean squared error of a provider portfolio weights in a week, relative to the same-week median-based portfolio weights, for those coins that are common across the two portfolios in the week. The summary statistics are computed across all weeks in the sample. For each provider, we have 311 weekly observations. The providers are abbreviated as in Table 1. Column "Us" considers the portfolios computed using data aggregated using our approach described in Section 6.

		CG	CMC	CP	LCW	M	S	Us
(a) Payment tokens:								
Overlap (l)	Mean	0.9582	0.9807	0.9448	0.8585	0.9614	0.9711	1.0000
	St. dev.	0.2005	0.1378	0.2287	0.3491	0.1929	0.1679	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.9678	0.9614	0.9188	0.8489	0.9453	0.9614	1.0000
	St. dev.	0.1767	0.1929	0.2735	0.3587	0.2277	0.1929	0.0000
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0438	0.0334	0.0754	0.1573	0.0569	0.0380	0.0000
	St. dev.	0.1566	0.1434	0.2043	0.2677	0.1795	0.1552	0.0000
	Median	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.7071	1.0954	0.8660	1.0000	0.7071	1.0954	0.0000
(b) Platform tokens:								
Overlap (l)	Mean	0.8612	0.9191	0.7601	0.6236	0.8492	0.9165	0.9936
	St. dev.	0.1055	0.0854	0.1946	0.1558	0.1248	0.0867	0.0274
	Median	0.8750	0.8889	0.8571	0.6250	0.8750	0.8889	1.0000
	Min.	0.4286	0.6250	0.0000	0.1250	0.2857	0.6250	0.8571
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.8404	0.9115	0.7357	0.5982	0.8315	0.9050	0.9924
	St. dev.	0.1135	0.0917	0.1941	0.1574	0.1373	0.0935	0.0299
	Median	0.8750	0.8750	0.7500	0.6000	0.8750	0.8750	1.0000
	Min.	0.4286	0.5714	0.0000	0.2222	0.2857	0.5714	0.8571
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0570	0.0386	0.0682	0.0765	0.0494	0.0403	0.0040
	St. dev.	0.0515	0.0479	0.0614	0.0609	0.0514	0.0482	0.0175
	Median	0.0396	0.0169	0.0505	0.0528	0.0294	0.0199	0.0001
	Min.	0.0006	0.0002	0.0004	0.0018	0.0003	0.0001	0.0000
	Max.	0.2322	0.2039	0.3160	0.2606	0.2258	0.2010	0.1908

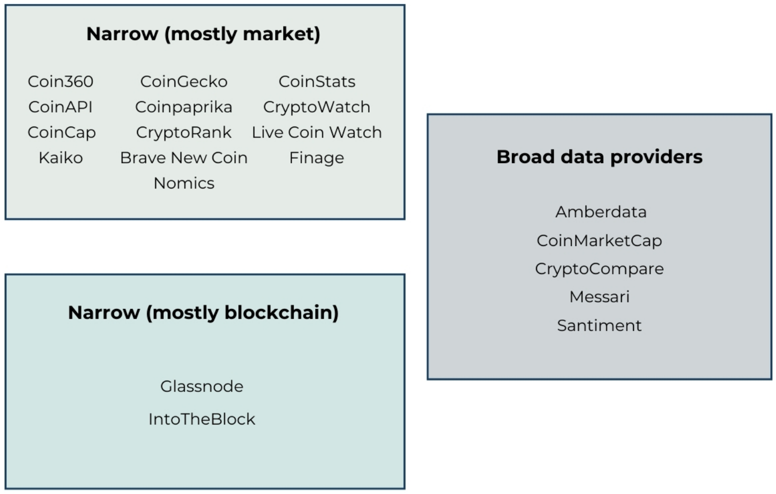
Table OA.7: *Summary statistics of differences for the momentum portfolio across categories (Part I).* This table repeats Exercise (c) of Table 6 by restricting to different token categories. It shows summary statistics of the weekly differences of the categorical momentum factor portfolios implied by the different providers relative to a portfolio that is constructed using daily market caps and close prices given by the median across all providers on a given day; see Section 6.4. "Overlap" measures the fraction of a provider's portfolio in a week that is also contained in the median-based portfolio during the same week. "RMSE" measures the root mean squared error of a provider portfolio weights in a week, relative to the same-week median-based portfolio weights, for those coins that are common across the two portfolios in the week. The summary statistics are computed across all weeks in the sample. For each provider, we have 311 weekly observations. The providers are abbreviated as in Table 1. Column "Us" considers the portfolios computed using data aggregated using our approach described in Section 6.

		CG	CMC	CP	LCW	M	S	Us
(c) Product tokens:								
Overlap (l)	Mean	0.8505	0.9218	0.8124	0.5756	0.8328	0.8826	1.0000
	St. dev.	0.2964	0.1998	0.3108	0.4435	0.3078	0.2370	0.0000
	Median	1.0000	1.0000	1.0000	0.5000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.8189	0.8810	0.7846	0.5852	0.7910	0.8167	0.9818
	St. dev.	0.3244	0.2482	0.3328	0.4399	0.3443	0.2902	0.1221
	Median	1.0000	1.0000	1.0000	0.5000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.1395	0.0785	0.1551	0.3245	0.1545	0.1148	0.0098
	St. dev.	0.1994	0.1415	0.1905	0.2406	0.2145	0.1584	0.0660
	Median	0.0225	0.0015	0.0679	0.3377	0.0300	0.0049	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	0.5878	1.0000	2.0000	1.0000	0.6542	0.5774
(d) Security tokens:								
Overlap (l)	Mean	0.8103	0.9100	0.7104	0.5145	0.7556	0.9100	0.9968
	St. dev.	0.3561	0.2474	0.4516	0.5006	0.4151	0.2507	0.0567
	Median	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.7910	0.8891	0.6990	0.4791	0.6945	0.8617	0.9968
	St. dev.	0.3764	0.2835	0.4559	0.5004	0.4489	0.3216	0.0567
	Median	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.1798	0.0990	0.2453	0.4141	0.2664	0.1141	0.0036
	St. dev.	0.2535	0.2076	0.3158	0.3902	0.3185	0.2267	0.0445
	Median	0.0000	0.0000	0.0000	0.5451	0.0000	0.0000	0.0000
	Min.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Max.	0.8165	0.8165	1.0000	2.0000	1.4142	0.8660	0.5774

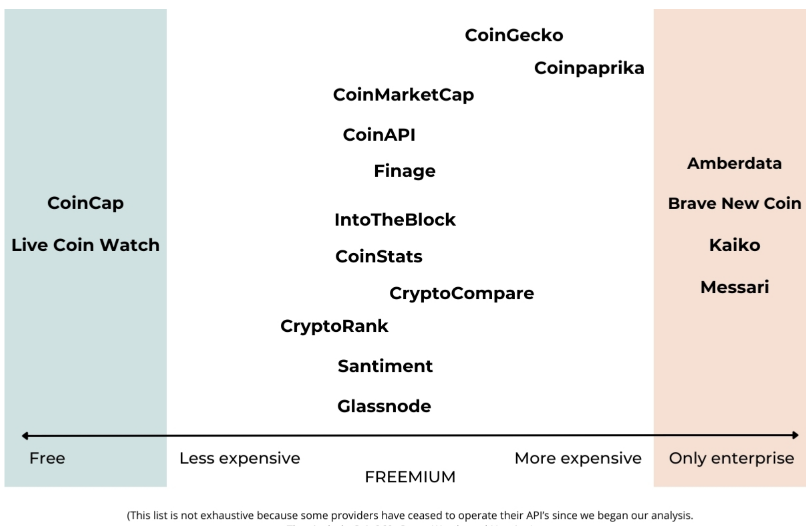
Table OA.8: *Summary statistics of differences for the momentum portfolio across categories (Part II).* This table repeats Exercise (c) of Table 6 by restricting to different token categories. It shows summary statistics of the weekly differences of the categorical momentum factor portfolios implied by the different providers relative to a portfolio that is constructed using daily market caps and close prices given by the median across all providers on a given day; see Section 6.4. "Overlap (l)" and "Overlap (s)" measure the fraction of provider portfolio's long or short legs in a week that is also contained in the corresponding leg of the median-based portfolio during the same week. "RMSE" measures the root mean squared error of a provider portfolio weights in a week, relative to the same-week median-based portfolio weights, for those coins that are common across the two portfolios in the week. The summary statistics are computed across all weeks in the sample. For each provider, we have 311 weekly observations. The providers are abbreviated as in Table 1. Column "Us" considers the portfolios computed using data aggregated using our approach described in Section 6.

	Provider					
	CG	CMC	CP	LCW	M	S
Number of primary coins affected by ID changes	5	3	108	3	23	16
% of these that are resolved by Algorithm 6.1	100.00%	100.00%	96.30%	100.00%	100.00%	93.75%
Number of primary coins affected by token forks or swaps	13	48	78	84	89	12
% of these instances that are resolved by Algorithm 6.1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table OA.9: *Frequency of ID issue cleanup.* This table shows the frequency with which the clustering step of our aggregation approach remove data for a coin that has ID issues from a provider. We identify ID issues as instances in which a provider leaves the ID unchanged after a major fork or split, or changes the coin's ID over time (see Table 1). We abbreviate the provider names as follows: CG: CoinGecko; CMC: CoinMarketCap; CP: Coinpaprika; LCW: Live Coin Watch; M: Messari; and S: Santiment.



(a) Types of data.



(b) Costs.

Figure OA.1: *Data providers broken up by the types of data they offer and the fees they charge.* For providers that operate under freemium models, we compute an average monthly fee as follows. We take all the monthly fees listed on the provider's website. We remove the fee for the free plan. When a provider offers an unpriced enterprise plan, we assume that its monthly fee is the maximum between $1,000 and twice the most expensive subscription price listed on the website. For each provider, we take the average of all the monthly fees. We then sort providers by their average monthly fee.

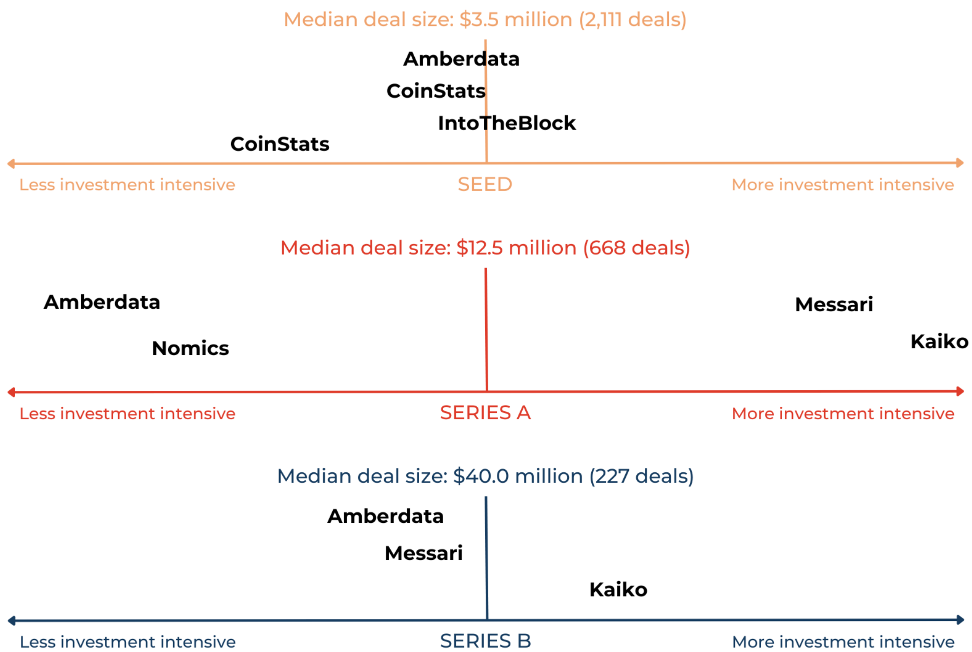
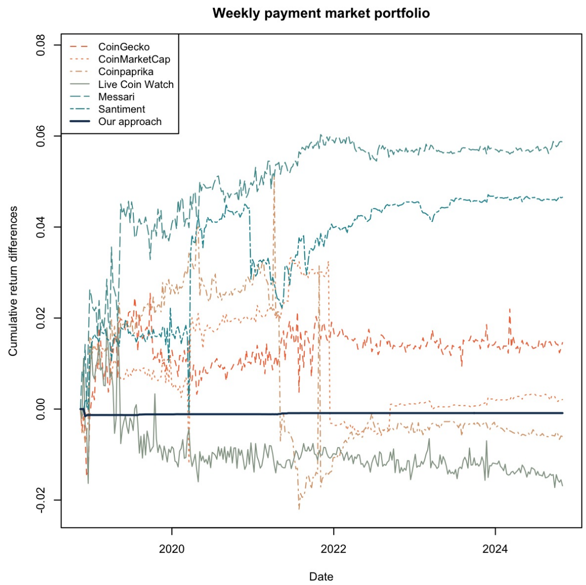
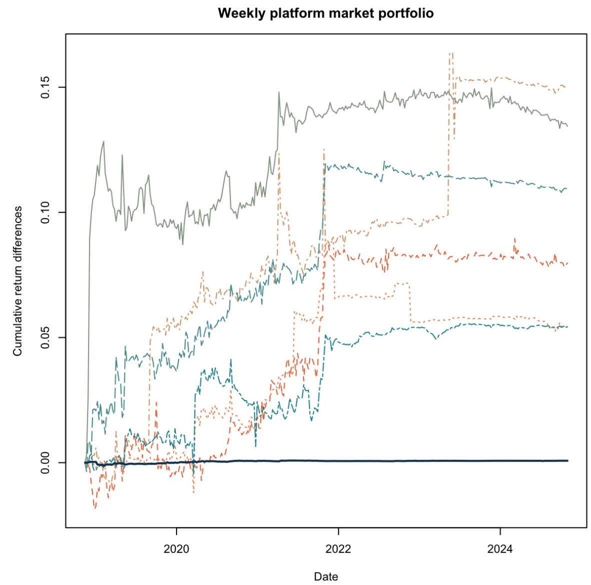


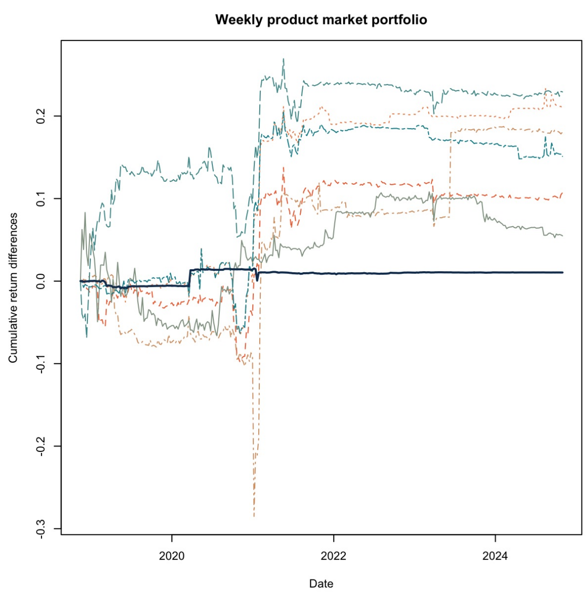
Figure OA.2: *Fundraising by data providers.* This chart shows a normalized measure of the amounts raised by the different data providers in the different funding rounds. For each provider and round, we take the amount raised and divide by the median deal size in the year in which the provider raised for the respective round. For details, see Online Appendix A.2.



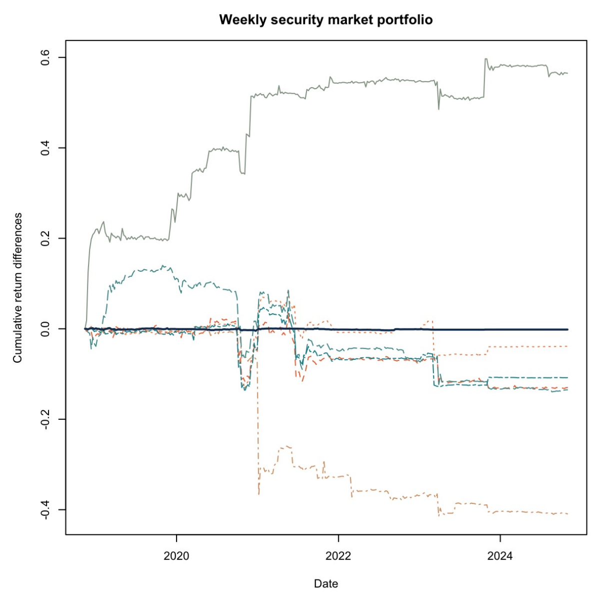
(a) Payment tokens.



(b) Platform tokens.

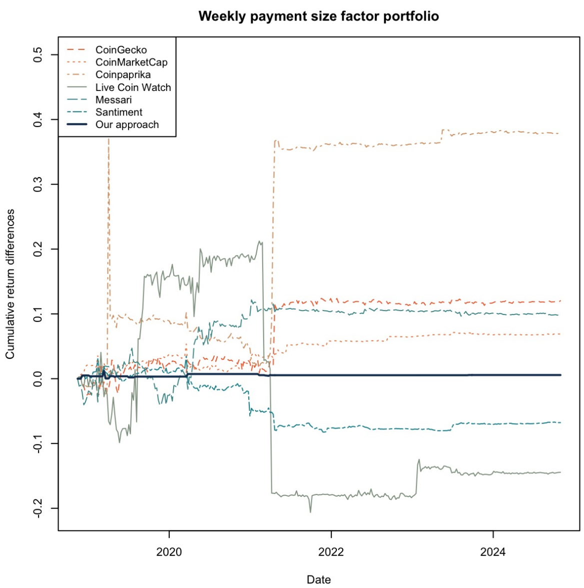


(c) Product tokens.

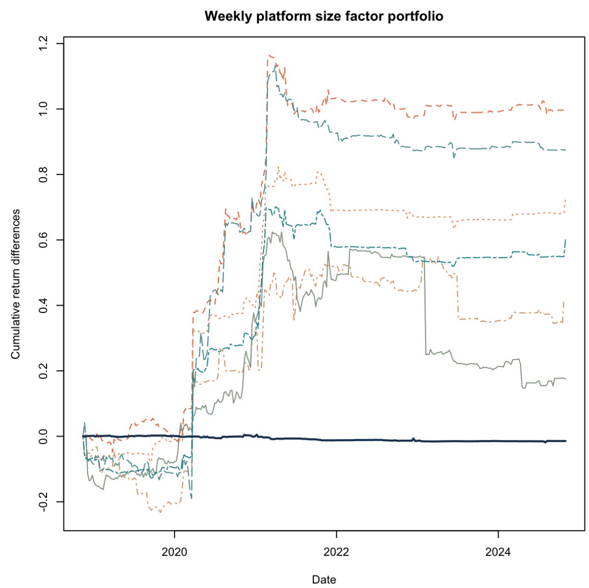


(d) Security tokens.

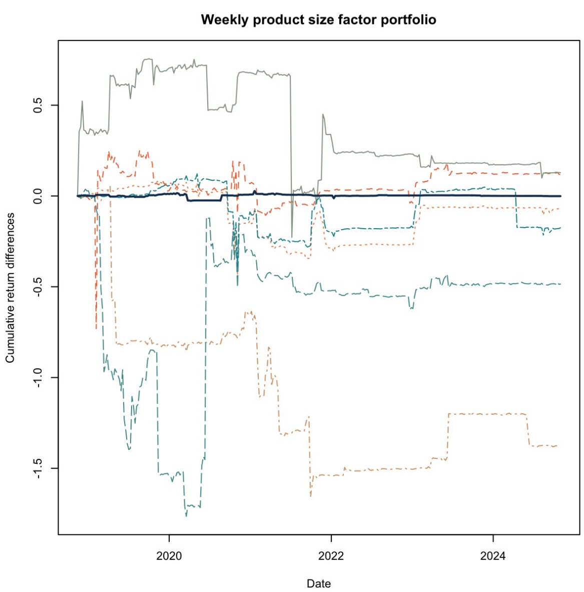
Figure OA.3: *Cumulative return differences of market factor portfolios for different categories.* We follow the same approach as in Figure 7 but restricted to coins that fall in each of the categories.



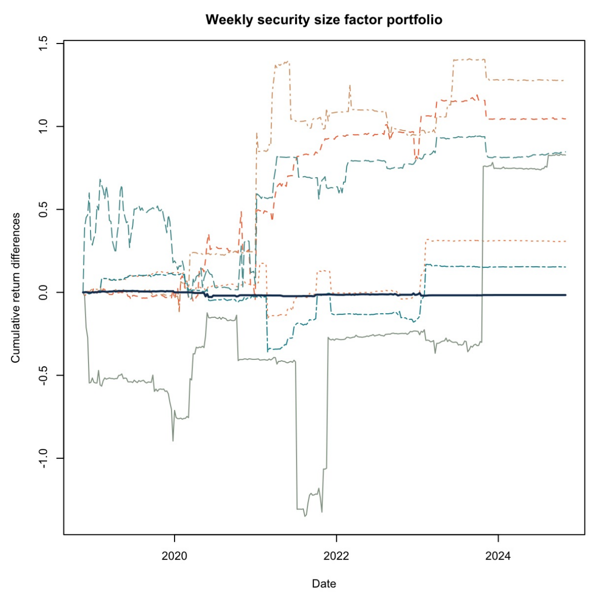
(a) Payment tokens.



(b) Platform tokens.

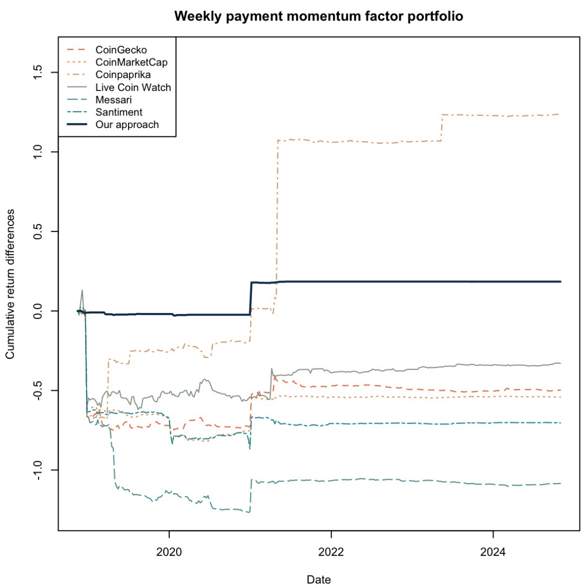


(c) Product tokens.

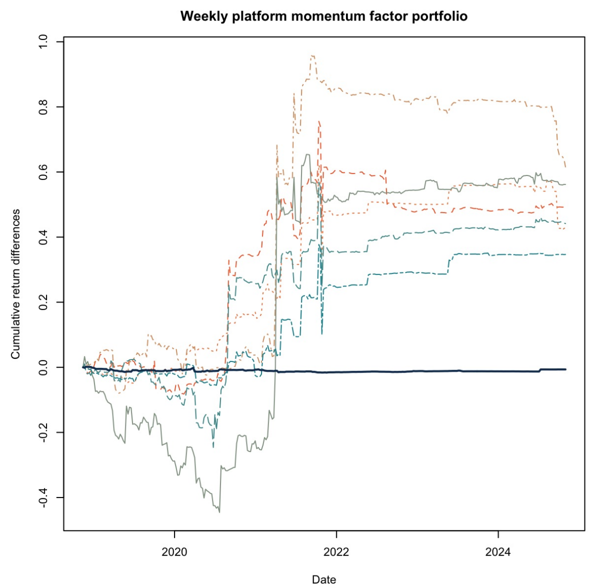


(d) Security tokens.

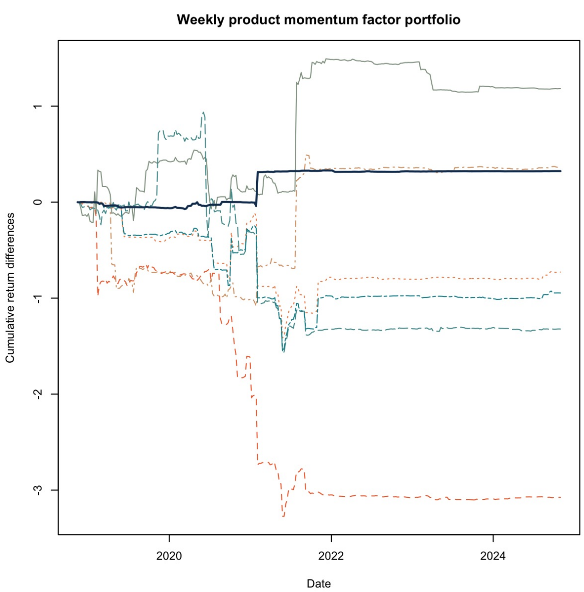
Figure OA.4: *Cumulative return differences of size factor portfolios for different categories.* We follow the same approach as in Figure 7 but restricted to coins that fall in each of the categories.



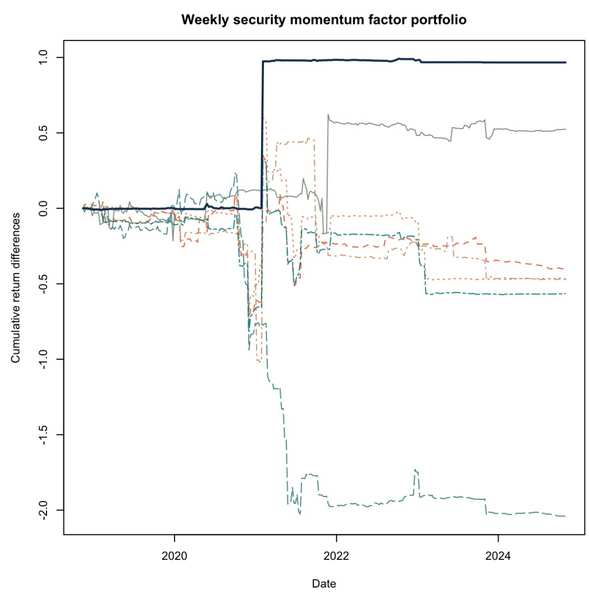
(a) Payment tokens.



(b) Platform tokens.

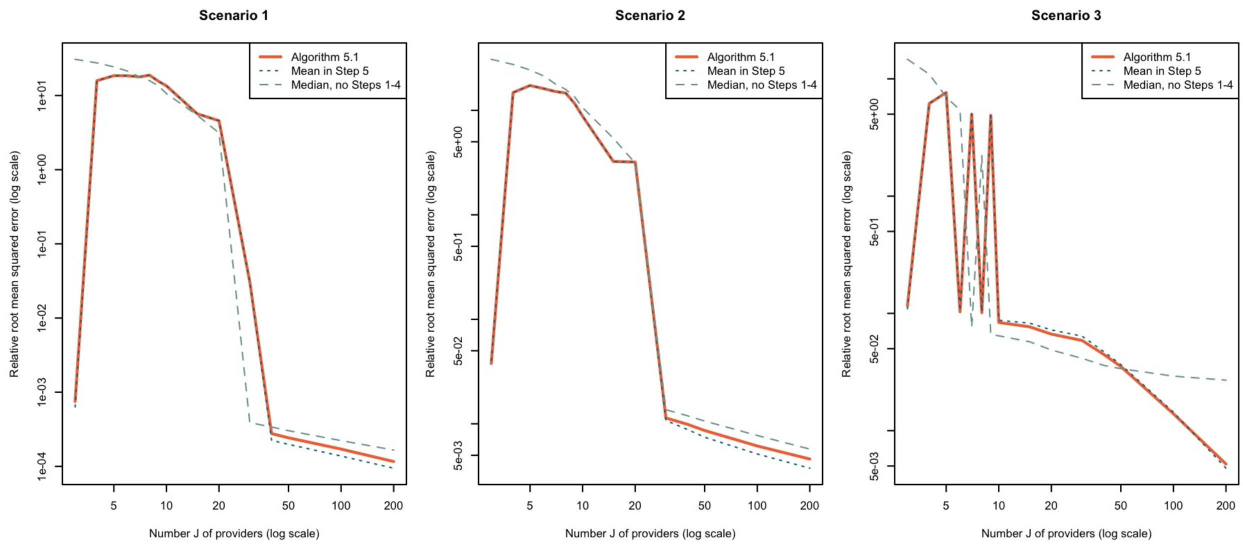


(c) Product tokens.

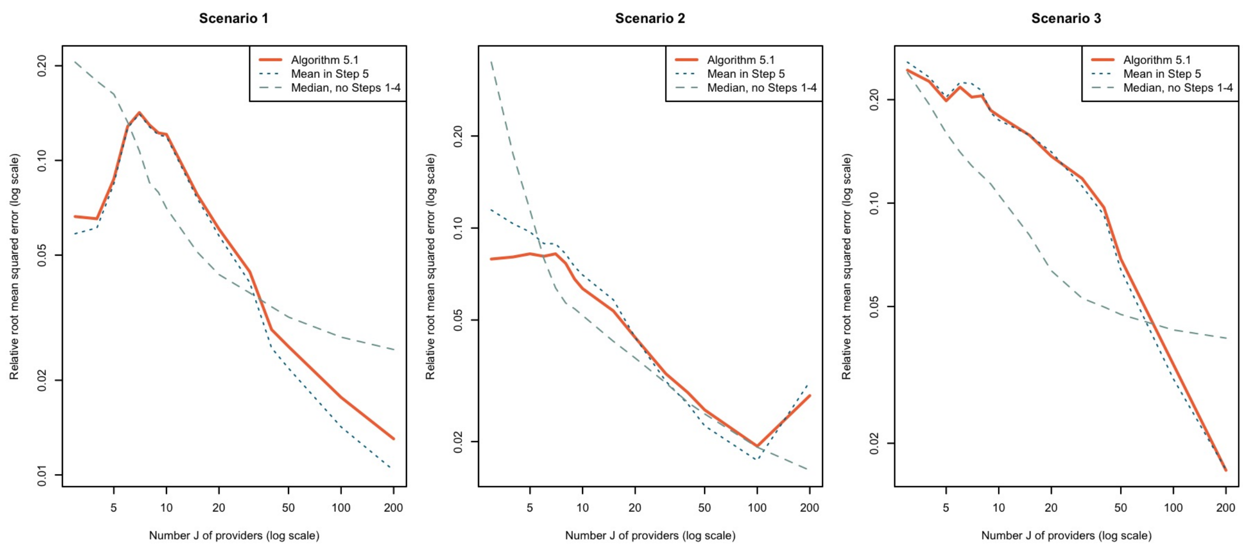


(d) Security tokens.

Figure OA.5: *Cumulative return differences of momentum factor portfolios for different categories.* We follow the same approach as in Figure 7 but restricted to coins that fall in each of the categories.



(a) Market cap.



(b) Close price.

Figure OA.6: *Relative root mean squared error of aggregation alternatives.* We evaluate the relative error of the aggregated market caps and close prices obtained from different aggregation approaches. Suppose we are aiming to estimate metric z and z^* is the true metric mean as in Theorem 6.2. Let $\hat{z}_k^i$ be the metric aggregated with method i for simulated data sample k. Then we estimate the relative root mean squared error as $\frac{1}{z^*}\sqrt{\sum_{k=1}^{K}(\hat{z}_k^i - z^*)^2/K}$, where $K = 1{,}000$ is the number of simulated samples.