Aggregate Confusion In Crypto Market Data

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

October 26, 2025†

Abstract

We present one of the first systematic audits of cryptocurrency market data quality across leading vendors. We document pervasive mislabeling, identifier instability, and large cross-provider discrepancies in prices, market caps, and volumes. To address these issues, we develop an aggregation method that yields asymptotically correct data by autonomously identifying and filtering unreliable observations. Using this framework, we construct an index to measure data quality over time and a grading system to benchmark providers. Our findings show that data inconsistencies can materially distort empirical research and investment analysis. They highlight the need for standardized reporting and oversight in crypto data markets.

Keywords: Cryptocurrencies, data quality, data aggregation, measurement errors, quality grading. JEL codes: C80, C10, C58, G19, G23.

*Schwenkler is at Santa Clara University. Shah and Yang are at Indicia Labs. Schwenkler is the corresponding author. Email: gschwenkler@scu.edu, website.

†This paper previously circulated under the title: "A Study Of The Reliability Of Crypto Data Provision." We thank Andrea Andolfatto (discussant), Agostino Capponi, Will Cong, Kay Giesecke, Cam Harvey, Jiasun Li, Dan Liebau, Markus Pelger, Donghwa Shin (discussant), and Geogii Zvonka (discussant); seminar participants at Stanford University; and conference participants at the 2024 CBER Annual Meeting, Economics of Financial Technology Conference, Wolfe Research Crypto Webcast, INFORMS Annual Meeting, 2025 SF Fed Fintech Conference, Future Finance Fest, and Knut Wicksell Conference on Crypto and Fintech for helpful suggestions. We also thank Romulo Farah, Zachary Fennie, Mark Filice, Sam Dines, Ismael IIboudo, William Morton, Chuanpu Yang, and Anthony Wang for outstanding research assistance at Santa Clara University.

1 Introduction

Crypto is once again experiencing extreme hype. After the crypto winter in 2022, Bitcoin has gained more than 550% since the start of 2023 and hit a new all-time high above $125,000 in 2025. A critical component for the broad adoption of crypto will be the degree to which crypto data is available in a reliable way. Academics and practitioners alike are demanding access to crypto data. Figure 1 displays the Google Trends time series for the term *"crypto data,"* which shows that interest in crypto data has been near its peak since 2023. Figure 1 also displays the number of new research publications that contain the words *"crypto"* and *"data"* in either abstract or title. It shows that more and more crypto data research is getting published.

Despite the growing interest, little research has been dedicated to studying the quality of the available crypto data resources. We fill this gap. We review some of the most common cryptocurrency market data vendors and show that quality issues are pervasive. We document inconsistencies in how cryptos are identified within and across providers, as well as substantial measurement errors. To resolve these problems, we propose an aggregation method that delivers asymptotically correct data when the number of providers grows large. We showcase the practical benefits of our approach and make our aggregated data available for public download. We use our aggregation method to develop an aggregate data quality index as well as a quality grading scheme for the various vendors in order to guide consumer choices. Our findings have a wide array of implications. We offer recommendations for academics, practitioners, and regulators in order to navigate the degree of aggregate confusion in crypto market data.

We run an in-depth analysis of eight crypto market data providers: CoinCap, CoinGecko, CoinMarketCap, Coinpaprika, CryptoCompare, Live Coin Watch, Messari, and Santiment. While this list is not a complete representation of all providers that exist in the market, it is representative because the providers are commonly used by practitioners and academics.[1] We analyze the data supplied by these providers for the time between November 2018 and October 2024. In doing so, we uncover pervasive quality issues. We find repeated issues with how different cryptocurrencies are labeled. We document that as much as 21% of all

[1]We complement our analysis by providing a survey of 20 market data vendors in our accompanying Online Appendix.

coins in a provider's sample may be subject to ID changes that are not disclosed. We find that providers may keep the same ID for a coin even after it goes through a fork or a swap that effectively renders a new coin. We also document instances in which one and the same ID is used for distinct cryptocurrencies within a provider. This happens in 16% of the sample on CoinGecko. The frequency with which labeling issues occur is somewhat unique to the crypto market given that traditional markets commonly use unique identifiers, such as CUSIP or ISIN. These inconsistencies make it difficult to ensure that data that are pulled from a provider actually belong to the targeted coins.

We document inconsistencies in the reported data across providers.[2] The daily close price reported by a provider can deviate from the median across all providers by as much as 10^6 on Coinpaprika or 10^{-9} on CoinGecko or CoinCap. Some inconsistencies are to be expected. The providers collect and aggregate data from different exchanges. Because they have discretion on how to do this, the supply of data naturally differs across providers. But the inconsistencies go well beyond this. Reported metrics commonly deviate by more than 5% across providers. This is especially pervasive for reported volumes. In almost 70% of the daily instances in our six-year sample did the daily aggregate volume for a coin reported by a provider deviate by more than 5% from the median volume reported across providers. We find that the distribution of trading volumes across exchanges is heavy-tailed and well-approximated by a power law with infinite variance. As a result, the choices that a provider makes in terms of which exchanges to pull volumes from in the process of aggregation have a major impact on the aggregate volumes that the provider ultimately reports to its customers. We find that this issue is particularly prominent for large coins that are listed on many exchanges, even though we would generally expect more reliable data for larger coins. The problem may be exacerbated by the practice of wash trading highlighted by Aloosh and Li (2024), Amiram et al. (2025), Cong et al. (2023), and Falk et al. (2025), whereby exchanges artificially inflate their reported trading volumes to appear more liquid.

The data discrepancies across providers permeate to the measurement of risks and returns for coins. We show that returns and volatilities computed using data from different providers are effectively uncorrelated if one does not use approaches to filter out large outliers. We also show that the discrepancies have a direct impact on portfolio construction and management. In analyzing the factor portfolios of Liu et al. (2022), we find that data from different

[2]Our approach follows Berg et al. (2022), who establish aggregate confusion across ESG ratings.

providers imply portfolio universes that share little overlap with each other. They also imply overstated portfolio returns that are difficult to realize based on more realistic data aggregated as the median across providers.

The large data discrepancies are puzzling. Despite the growing interest in crypto data, the providers offer little insights into how they construct their aggregated metrics. To understand what kind of behavior could give rise to the large data discrepancies we document, we develop a structural model that explains how providers compute aggregate daily close prices and 24-hour trading volumes using trade-level data from exchanges. Our model has two components. The first component specifies how trade-level data is generated on an exchange. The second component explains how providers collect and aggregate trade data from exchanges. Our model mirrors the aggregation approach followed by CryptoCompare, which we impose on all providers even though different vendors may use different methods.[3] We add a layer of errors and noise that is governed by five structural parameters. The first parameter captures how many exchanges a provider tracks when collecting trade data. The second governs the frequency with which a provider mislabels a coin in their internal ID system. The third captures how frequently a provider mismatches a coin on individual exchanges. The fourth parameter captures the volatility of measurement errors that a provider may add through fat-finger-like mistakes when collecting trade data. The final parameter captures the strength of wash trading controls that a provider may apply. We compute estimates of these parameters through a simulated method of moments (McFadden (1989), Pakes and Pollard (1989)). Our estimates are given by the parameters that minimize the sum of squared errors between moments of data discrepancies in the provider samples and moments of simulated data generated from our structural model.

Our estimates show that three key behaviors may be at play. First, our estimates suggest that providers may only collect data from small sets of exchanges that are not shared across providers. We find that the data discrepancies can be explained by our model if providers track no more than 50 exchanges on average across coins. This stands in stark contrast to the number of exchanges that providers claim to track, which can be as large as 1,500 for CoinGecko. Second, the estimates suggest that error rates among providers may have to be high in order to match the distribution of data discrepancies. Our model requires a mislabel-

[3]While this assumption is not realistic, it is not possible for use to implement individual methods because most providers offer little details on their aggregation approaches; see Online Appendix C.

ing rate of 0.7% for Live Coin Watch and a mismatch rate of 1% for CryptoCompare. These rates are higher than those that apply to traditional databases like CRSP or Compustat, which are expected to be below 0.1% (Bennin (1980)). Finally, the estimates suggest that measurement errors are pervasive. Our model requires measurement error volatilities above 1% for all providers to match the discrepancies in the data. Surprisingly, our estimates do not find differences in wash trading controls to be a significant driver of data discrepancies. Still, our findings raise questions about quality controls in the market for crypto data. While our estimates cannot be interpreted as proof of what the providers actually do because of their indirect nature, they offer evidence of potential shortcomings.

Our findings indicate that it is difficult to rely on a single provider for all crypto market data needs. We propose a methodology that aggregates data across providers and ensures consistency. Our methodology follows two key steps. In the first step, we cluster providers based how similar their data are. In the second step, we aggregate within the largest provider cluster using the median function to control for outliers. We repeat these steps twice, once for market caps and a second time for closing prices. Each time, we sieve out providers that report inconsistent data. We provide conditions under which our methodology delivers correct coin-level data in the asymptotic limit in which the number of data provider grows large. We also show that our methodology performs better than alternative approaches that rely on aggregating using the mean function rather than the median, or that do not employ any clustering whatsoever. Our methodology achieves a feat that at first seems hard to accomplish: to extract correct data values from a cross section of providers for which we do not ex-ante know which ones report correctly. In this sense, our methodology connects to truth inference algorithms that are common in the artificial intelligence literature; see Meng et al. (2015), Sheng and Zhang (2019), Zheng et al. (2017), and others. Our methodology is also robust to missing data issues that have recently been flagged by Bryzgalova et al. (2024) and Freyberger et al. (2024) for financial data.

We demonstrate the practical benefits of our methodology in several applications. We show that the coin returns implied by our aggregated data generally align with the median across providers without being susceptible to mislabeling issues. We also show that measures of volatilities and tail risk implied by our aggregated data are more conservative than those implied by the individual providers or alternative aggregation methods because our approach is robust to large outliers. Finally, we find that factor portfolios constructed using our

approach do not suffer under inflated portfolio returns that may be hard to replicate in realistic settings. To enable the adoption of our approach, we make our aggregate data available for public download in our Online Data Repository.

To facilitate consumer choices in the market for crypto data, we exploit our aggregation approach to develop a grading scheme that quantifies the quality of the market data supplied by a provider. We use a standard academic grading scheme based on the percentage of daily data instances of a provider that are discarded by our aggregation approach. We find that CoinMarketCap and Santiment fare best according to our grading scheme, achieving consistent monthly grades of A over our sample period. On the other hand, CoinGecko fares worst in our analysis, achieving an average grade of C. We also use our aggregation approach to construct an Aggregate Confusion Index that indicates when aggregate crypto market data may be less reliable. We find that the quality of aggregate crypto market data may be lower during periods of high market activity. In contrast, we find that the quality may be higher whenever demand for crypto market data is high. These results show that our aggregation approach offers benefits that go beyond the computation of coin-level and portfolio metrics. It also enables insights that can help consumers navigate the market for crypto data.

Our paper has important implications. For academics, our results raise questions about p-hacking (Harvey (2017)) and non-standard errors (Menkveld et al. (2023)) in cryptocurrency asset pricing research. Alexander and Dakos (2020) and Fieberg et al. (2024) were some of the first to document replicability issues in crypto studies. We complement these papers by providing a novel aggregation approach and offering quality gradings for the different providers. For practitioners, our results indicate that it is imperative to develop robust methods to clean market data obtained from commercial data vendors. We provide one approach that offers practical benefits over alternatives, especially when only a few data vendors are available. For market overseers, one recommendation that arises from our paper is the establishment of a unified identification system for cryptocurrencies that can be adopted by all providers, similar to CUSIP and ISIN in traditional markets. Unified crypto identifiers have recently been pushed by Bloomberg, Kaiko, and the Digital Token Identifier Foundation. Our results provide empirical support for these endeavors. Our findings also raise questions about a potential need for regulation from a consumer protection perspective. In the United States, the Securities and Exchange Commission (SEC) has issued Rule 603

of Regulation NMS known as the *Vendor Display Rule* that governs how data vendors have to report aggregate market data in any instance in which a customer may use the data for trading purposes. Our results suggests that some practices in the market for crypto data may be at odds with the Vendor Display Rule. However, this regulation does not currently apply to crypto, leaving the market exposed to data quality issues.

This paper is organized as follows. Section 2 provides an overview of the providers and their data. Section 3 outlines the quality issues we uncover. Section 4 shows how the data quality issues impact the measurement of coin-level risk and return metrics as well as portfolio construction. Section 5 introduces our structural model to back out the origins of the data discrepancies, together with our parameter estimates. Section 6 presents our aggregation approach while Section 7 demonstrates its practical benefits. Section 8 concludes and provides our recommendations. There is an Online Appendix as well as an Online Data Repository, both of which can be accessed here.

2 Providers & data

We evaluate the quality of the crypto market data supplied by the following providers: CoinCap, CoinGecko, CoinMarketCap, Coinpaprika, CryptoCompare, Live Coin Watch, Messari, and Santiment. There are many more providers that we do not consider in our analysis. We complement the analyses in this paper by providing a survey of other providers in our Online Appendix. While incomplete, our set of providers is representative of how crypto data is used. Figure 2 shows the number of research publications through the end of 2024 that mention the name of a providers together with the word *"crypto"* in the body of the paper. There are more than 6,500 distinct papers that mention the providers. The most commonly quoted provider is CoinMarketCap, followed by Nomics and CoinGecko. CoinMarketCap and CoinGecko are included in our list while Nomics has gone out of business. In addition, the providers in our list often serve as data sources for other providers, including some that we did not review. For example, Yahoo Finance pulls data from CoinMarketCap, Messari partially pulls data from CryptoCompare, and Santiment pulls data from both CoinMarketCap and CryptoCompare (see here). We therefore consider our list to be a comprehensive representation of the market for crypto data.

For each provider, we collect daily global open, high, low, and close prices, as well as

market caps and trading volumes at the individual coin-level. By "daily," we mean from the start to the end of a day in the Coordinated Universal Time Zone (UTC). Because cryptocurrencies trade around the globe across time zones, consensus has built to treat UTC as the reference time zone for daily crypto metrics. By "global," we mean market metrics that are aggregated from exchanges around the world into unique time series. Providers have discretion in determining which exchanges they collect data from. Different providers collect from different exchanges, giving rise to differences across providers. One of the goals of this paper is to study these differences.

We obtain our data from the providers' Application Programming Interfaces (APIs). We employ the following matching approach when comparing data for one coin across providers. At the beginning of each month, we load the list of all coins tracked by a provider from the API. We sort coins by market cap and track their tickers. Whenever there are repeated tickers in a provider, we keep the ticker associated with the largest market cap and discard data for the other coins. Then, we match data across providers based on the ticker and keep any ticker that is present in at least 3 providers. For simplicity, we track only the 250 largest coins each month for each provider. We do this every month between November 2018 and October 2024. This approach yields a sample of 553 distinct coins. We call this sample the sample of *primary coins*. In our analyses, we do not censor any data entries or control for outliers. Our goal is to assess the quality of the raw data supplied by the providers without imposing filters that may favor one provider over another.

3 Aggregate confusion

We document two key data issues we encountered in our analysis: erratic coin labels and large discrepancies across providers.

3.1 Inconsistent labels

There are no rules for how to label different cryptocurrencies. Because of this, the providers have come up with their own unique labels. Most data providers have three different types of labels for cryptocurrencies:

- The **name** of a coin, which is the text that is commonly used to refer to a coin.

- The **ticker** of a coin, which is the abbreviation that is commonly used to quote a coin on cryptocurrency exchanges. Not all providers refer to the ticker with the word "ticker." Some providers, like CoinCap, CoinGecko, Coinpaprika, and CryptoCompare, use the word "symbol."

- The **ID** of a coin, which is a provider-specific label assigned to identify coin. Not all providers refer to their internal coin label as "ID." Santiment uses "slug." Live Coin Watch uses "code."

When using APIs, most data providers require that one supplies the ID and not the ticker of a coin. While ID and tickers can be the same, this is not always the case. Table 1 shows that several providers have IDs that are different from the tickers. CoinCap, CoinGecko, CoinMarketCap, and Santiment use a variation of the name while Coinpaprika uses a combination of the ticker and name. Messari uses a unique alphanumeric combination that neither includes the ticker nor the name as ID.

Given the discrepancies between tickers and IDs, it is often necessary to collect all of a provider's IDs before any data can be accessed through an API. Still, caution is advised. Table 1 shows that 16% of the sample on CoinGecko is plagued with re-used IDs. For example, CoinGecko uses the same ID to refer to both the now defunct Terra as well as the new Terra Luna Classic. Table 1 also shows that there are instances in which the same ID is used to label a coin after a token swap or a fork. For example, SAFEMOON had a token swap on December 13, 2021, at a ratio of 1000-to-1 when it migrated to a new contract address (see here). Some providers, like Live Coin Watch, mixed the two tokens under the same ticker and therefore supplied mismatched data. Other providers, like CoinGecko and Coinpaprika, only reported data for the old Safemoon token. Santiment only reported data for the new token. Issues like this impact almost half a percent of primary coins on Messari and 0.7% on CoinGecko.

In other situations, the ID of a coin may change at some point during the lifetime of a coin. Table 1 shows that ID changes affect as much as 21% of primary coins on Coinpaprika and 4% on Messari. Below are some examples of instances in which the ID of a provider changed or got disabled even though the coins continued to exist.

- ARPA Chain (ticker: ARPA) had the CoinGecko ID "arpa-chain" through July 30, 2022. It then changed to "arpa." When we request data with the old ID, CoinGecko

returns a 404 error.

- BitDAO (ticker: BIT) had the Santiment ID "bitrewards" through May 31, 2023, and then changed to "bitdao." When we requested data with the old ID, Santiment returned an empty set.

- MCDEX (ticker: MCB) has the Coinpaprika ID "mcb-mcdex-token" through August 30, 2023, and then changed to "mcb-mcdex."

The extent to which identification issues are present is somewhat unique to the crypto market. Most securities in traditional markets are identified through a CUSIP or an ISIN. The ISIN of a security is unique and can never be reused according to the ISIN Uniform Guidelines.[4] A CUSIP may only be reused for certain asset classes that have rolling, large-volume issuances of securities with the same principal characteristics.[5] Indeed, the CUSIP system was established by the American Bankers Association in the 1960s exactly to deal with identification issues that plagued the US stock market; see Ritter and Wool (2021). We view the identification issues documented in this section as a challenge for the consumption of aggregated crypto market data.

3.2 Large data discrepancies

We document large discrepancies in daily market metrics reported by the different providers. Consider the daily global close price of a coin. When computing global close prices, the providers pull data from potentially different sets of exchanges. This introduces differences in global close prices. Our analyses shows that these differences can be significant.

For each coin in our sample of primary coins, Figure 3 plots what we call the *divergence score* of daily global close prices. We define the divergence score as a daily global metric for a coin divided by the median daily global metric across providers for the given coin. It is a measure of data discrepancies across providers. In the figure, we plot the close price divergence scores against the median daily global close price normalized by the average median daily global close price for the given coin over the sample period. If all providers

[4]Section 6 states: *"ISINs should never be re-used. This rule applies to all kinds of financial and referential instruments."*

[5]Only three asset classes may have re-used CUSIPs. They are discount commercial papers, Federal Agency discount notes, and to-be-announced mortgage pools; see CUSIP Global Services Process & Procedures.

reported the same global close price, all entries would land on a horizontal line centered at one. They would fall further on the right when the close price of a coin is high and further on the left when the close price is low. However, the figure shows that the daily global close prices can diverge by large amounts across providers. The divergences can be as large as a factor of 10^6 (Coinpaprika) or 10^{-9} (CoinGecko and CoinCap) relative to the median close price. The discrepancies are more prevalent when the close price fluctuates around its average value (close to one on the x-axis).

Figure 4 shows that cross-provider discrepancies are also present in daily global open, high, and low prices, as well as trading volumes and market capitalizations. The problem is especially pervasive for global trading volumes. Out of the 3,379,576 provider-crypto-day global volume entries in our sample of primary coins, only 30.4% fall in a $\pm 5\%$ band around the median global volume across providers. Table 2 shows the correlations of the different metrics across providers. Most providers report metrics that are generally strongly correlated. However, global volumes are the noisiest market metric. The correlation between global volumes can be as low as 27% between CryptoCompare and many of the providers.

Table 3 reports summary statistics of the divergence scores of Figures 3 and 4 for the different providers. These statistics reveal additional insights. First, we observe that some providers report data that deviates systematically from the other providers. For example, CryptoCompare appears to report excessively large high prices and excessively low trading volumes, as characterized by the average divergence for these market metrics and provider. On the other hand, Coinpaprika appears to report excessively large volumes, market caps, and close prices. Second, we observe that some coins are more likely to be impacted by the divergences of certain providers. For example, Metadium (ticker: META) appears to be frequently impacted by inconsistent pricing data from Messari, as can be seen by the large minimum and maximum pricing divergence scores. Similarly, Safemoon appears to be frequently impacted by price data inconsistencies at CryptoCompare. Finally, some providers appear to be specialized, meaning that they are more consistent for some metrics and more inconsistent for other metrics. For example, CoinGecko appears to have open and close price divergence scores that are highly concentrated around one, as can be seen from the low standard deviation of these scores. This suggests that global open and close prices from CoinGecko are fairly consistent with the median across providers. However, global market caps and trading volumes from CoinGecko appear to be inconsistent given the large standard

deviation of their divergence scores.

3.2.1 Noisy volumes

Crypto trading volumes are known to be intrinsically noisy. There is a natural statistical reason. Unlike for other market metrics, volumes are aggregated by summing up individual records rather than averaging them out through a robust method that controls for outliers. The distribution of trading volumes across exchanges is highly skewed. Figure 5 shows the distribution of global trading volumes across the largest 250 crypto exchanges in 2023. We see that the largest 10 exchanges recorded almost 55% of the total global trading volume. We estimate that the power law exponent for the distribution of exchange volumes is 2.99, suggesting an infinite variance. A bootstrapped Kolmogorov-Smirnov test cannot reject the null hypothesis that the empirical distribution in Figure 5 was generated from this power law. These observations suggest that aggregating trading volumes may be intrinsically difficult if a provider does not consider all of the largest exchanges in the world. Our observations imply that a provider that chooses only a subset of the largest exchanges to aggregate volumes from may end up with a vastly different global daily volumes than another provider that chooses a different set of large exchanges.

There is also an economic reason. Aloosh and Li (2024), Amiram et al. (2025), Cong et al. (2023), and Falk et al. (2025) document that exchanges may have incentives to engage in wash trading and therefore misreport their true trading volumes. Wash trading is the practice whereby exchanges move funds across internal wallets to inflate their trading volumes without actually seeing trades. They do so to appear more liquid and attract users. Wash trading is common among unregulated exchanges, with as much as two-thirds of all reported volumes being potentially fake according to the literature.

The providers recognize that wash trading is an issue. Many have developed their own methodologies to filter out true trading volumes from reported trading volumes. This adds another layer whereby data across providers may differ. Two providers may measure the same reported trading volume but end up quoting different *adjusted* trading volumes because they may use different methods to control for wash trading. To add to the confusion, many providers do not clarify whether they quote adjusted or reported trading volumes. Below, we list whether a provider indicates in their API documentation (when available) whether the supplied volume metrics are adjusted or reported:

- CoinCap, Coinpaprika, Live Coin Watch, and Santiment: these providers do not appear to adjust and only supply reported volumes.
- CoinGecko and CryptoCompare compute both adjusted and reported trading volumes. Their APIs only report single volume metrics and it is unclear which ones they are.
- Messari claims to transform "*raw trade data gathered from Kaiko, CCData, and internal trade ingestors*" in their API documentation. But it reports only one volume metric. It is unclear which metric it is.
- CoinMarketCap supplies both adjusted and reported trading volumes with clearly outlined variable names.

As these observations indicate, volumes across data providers may be vastly different because of the statistical distribution of volumes across exchanges and because providers have discretion on how to filter out fake volumes in their quoted metrics.

3.2.2 Interaction with size

We evaluate whether data discrepancies are more pervasive for smaller or larger coins. In Table 4, we report the percentage of instances in which a divergence score exceeds a $\pm 5\%$ band around one. We break down the measurements based on terciles of cross-provider median market caps measured each day.

We observe that inconsistencies in daily global prices (open, high, low, and close) are more pervasive for small coins than for large coins. Generally, high and low prices are noisier than open and close prices. We also observe that market caps are generally noisier for smaller coins, even though global market caps tend to be noisier than global prices. For volumes, we observe that discrepancies are more pervasive for large coins than for small coins. This is likely due to the fact that large coins are listed on more exchanges than small coins (see Schwenkler and Zheng (2025) and Shams (2022)). As a result, providers have more discretion to choose which exchanges to pull volume data from for large coins than for small coins. Combined with the fact that the distribution of volumes across exchanges follows a power law, this naturally yields noisier volume data for large coins than for small coins.

4 The cost of aggregate confusion

We demonstrate the impact of the data discrepancies highlighted in the previous section by testing how coin-level risk and return measurements deviate when evaluated with data from different providers. We also study their impact on portfolio construction and management.

4.1 Risk & return measurements

For each provider, we compute two measures of returns: an intraday return from the open to the close, and a full-day return from the previous close to the current close. We also evaluate two different measures of volatility. We compute an intraday volatility due to Garman and Klass (1980) using the open, high, low, and close prices as follows:

$$\sigma_{i,t}^2 = 0.5 \times (h_{i,t} - l_{i,t})^2 - (2\ln(2) - 1) \times (c_{i,t} - o_{i,t})^2$$

Here, $o_{i,t}$, $h_{i,t}$, $l_{i,t}$, and $c_{i,t}$ represent the log open, high, low, and close prices for coin i on day t. We also compute a historical volatility metric as the standard deviation of the last 30 full-day returns. We measure daily 95% value-at-risk at the coin-level as the fifth percentile of historical full-day return distributions over rolling 365-day windows, requiring a minimum of 180 non-missing daily data entries. Finally, we measure the daily ranking of a coin based on descending daily market caps, but restricted to the 100 largest coins for simplicity (i.e., any coin that would rank below the top 100 does not get ranked).

Figure 6 shows analogous divergence scores as in Figures 3 and 4, but for the different risk and return metrics. In addition, Table 5 shows the correlations for these metrics across providers. We observe that value-at-risk and size rankings are the most consistent metrics across providers. However, there is significant divergence when we compute returns and volatilities based on data from the different providers. This occurs even when two providers report pricing data that are highly correlated. For example, prices are highly correlated between CryptoCompare (CC) and CoinCap (CCP) in Table 2. However, the returns and volatilities implied by data from CryptoCompare and CoinCap are essentially uncorrelated according to Table 5.

To understand to which degree the outliers in the pricing data documented in Figure 3 drive the low correlations, we re-evaluate the correlations after removing the top and bottom

1% outliers from each provider's close price sample before computing full-day returns (rows labeled "Cleaned full-day return") and historical volatilities ("Cleaned historical volatility"). We observe that the correlations increase to 61% for returns and 50% for historical volatilities. While these values are higher than without removing the outliers, they are still relatively low in absolute terms. These results suggest that deviations in reported market data can lead to substantially different risk and return measurements for individual coins.

4.2 Portfolio construction

We evaluate the impact of the data discrepancies on portfolio construction by assessing the value-weighted market, size, and momentum factor portfolios of Liu et al. (2022). These results are directly relevant for institutions running factor strategies that are traded across multiple exchanges through over-the-counter brokers or smart order routers.[6] Because such trading mechanisms do not rely on single exchanges, portfolio construction decisions may be based on aggregated data rather than direct exchange-level data. By considering portfolios based both on traditional value weights as well as portfolio sorts, our findings may also be representative of the kinds of issues that portfolio managers may face more broadly beyond the factor setting of Liu et al. (2022).

We proceed as follows to construct the portfolios. We rebalance at the beginning of every Wednesday and restrict ourselves to the 100 largest coins from the set of primary coins of a provider based on the average market cap from the previous 7 days. For the market factor, we compute value weights at the start of every Wednesday based on the average daily market cap over the prior seven days. For the size factor, at the start of each Wednesday we determine quintile portfolios of coins with the largest and the smallest average market capitalizations over the past 7 days. We go short on the small size and long on the large size value-weighted portfolios. Similarly, for the momentum factor, we construct at the start of each Wednesday quintile portfolios of coins with the largest and lowest returns over the last 7 days. We then go long on the top and short on the bottom past-week-return vale-weighted portfolios. Note that data may be missing. When this is the case, we shrink the quintile portfolios to include only one-fifth of the coins with available data (which will be less than 20 coins when data are missing). Our portfolio construction takes place on

[6]Recently, Sparkline Capital introduced crypto factor portfolios; see here. MSCI's Barrera and Minovitsky (2021) suggest that factors can explain up to 45% of the cross-sectional variation of crypto returns.

a rolling basis only using contemporaneously available data, avoiding any look-ahead bias. Once the portfolio have been constructed, we measure their returns from the open on the Wednesday to the close the next Tuesday. We exclude CoinCap and CryptoCompare in this analysis because these providers do not report historical market caps, making it impossible to construct value-weighted portfolios.

Figure 7 shows cumulative weekly returns of the factor portfolios implied by the different providers. We observe that different providers imply different performance for the factor portfolios. The performance of the market portfolio is more or less consistent among all providers. However, the size and momentum factor portfolios show substantial inconsistencies. The differences between the different providers imply a cumulative return gap between the highest and the lowest value of 2.9x for the size factor and 7.9x for the momentum factor.

4.2.1 Portfolio differences

What drives the differences in the portfolio returns of Figure 7? We posit that there are three channels. First, portfolios constructed using data from different providers may contain different sets of assets because of the data discrepancies we documented in Section 3. Second, even among the set of assets that are common across portfolios, data from different providers may imply different value weights due to the market cap discrepancies across providers. Third, even if data from different providers implied the same portfolio allocations for the same set of assets, the cumulative returns of the portfolios may differ because of the return discrepancies across providers. We empirically evaluate these three channels.

In Table 6, we compare the weekly portfolio universes and allocations of the factor portfolios to what would be obtained if one always used the median daily market cap or price across providers when computing previous-week size rankings or returns while constructing the portfolios.[7] These tests consider the impact of the data issues on portfolio construction while ignoring portfolio returns, allowing us to assess how strongly the first two channels contribute to the portfolio performance discrepancies of Figure 7. In an additional test, we keep the portfolios fixed as implied by the different providers and we evaluate their weekly returns using different data: one time using the returns implied by the different provider data samples (as in Figure 7) and another time using the returns implied by the median open and

[7]While median-based aggregation may not necessarily be optimal (see Section 6.4), it offers a simple benchmark to compare the different portfolios.

close prices across providers in a week. Figure 8 plots the corresponding cumulative return differences. This test assesses how strongly the third channel contributes to the portfolio performance discrepancies of Figure 7.

We observe in Table 6 that, generally, low overlap between the portfolios implied by different providers is a critical issue. We find that the market portfolio implied by Coinpaprika can have as little as 14% overlap with the weekly median-based market portfolio. The long leg of the size factor portfolio (i.e., the small cap portfolio) implied by Live Coin Watch has on average only 11% overlap with the corresponding leg of the median-based portfolio. The long and short legs of the momentum size portfolio of Live Coin Watch only have around two-third overlap with the median-based legs. These measurements show that the data discrepancies of Section 3 have a direct impact on portfolio construction by implying different investment universes when using data from different providers. In contrast, we find that portfolio allocations among the set of common coins tend to be generally be fairly consistent. The average root mean squared error of the portfolio weights for common portfolio assets is relatively low, remaining below 3.5% on average. These results suggest that the main channel that drives the impact of the data quality issues of Section 3 on portfolio construction is by imposing differences in portfolio universes (first channel), not so much by introducing differences in portfolio allocations for common assets (second channel).

We extend our analysis in Figure 8 by considering the differences between the provider-implied and the median-implied portfolio returns. The cumulative return differences in these plots are all positive by the end of the sample period. This suggests that the factor portfolio returns implied by data from the different providers may be overstated, meaning that portfolio returns may not be fully realizable when considering the more realistic median weekly returns across providers. The overstatement of the portfolio returns can be as high as 2.8% per year for the market portfolio implied by Live Coin Watch, 10.8% per year for the size factor portfolio implied by CoinGecko, and 9.8% per year for the momentum factor portfolio implied by Coinpaprika. Such an overstatement is consistent with an interpretation that the factor portfolios are unable to capture the full mispricing attributed to them when they are constructed using data contaminated with quality issues as those of Section 3. These results have implications beyond the factor setting of Liu et al. (2022). They suggest that portfolios constructed using data from individual providers with the goal of exploiting mispricing or earning an alpha may have a hard time living up to their expected performance

when evaluated with realistic data. They show that the third channel by which cross-provider data discrepancies impact portfolio performance can be significant.

5 The origins of aggregate confusion

We evaluate the origins of the data discrepancies across providers. In an ideal world, we would collect exchange-level transactions and replicate the aggregation approaches of the different providers to back out where the discrepancies arise. However, such an approach is challenging for two reasons. We provide a survey of the methodology descriptions offered by the providers in Online Appendix C. There, we show that details on the aggregation methodologies are scant. The lack of detailed methodology information renders replication nearly impossible. Even if the providers were perfectly transparent, replicating their approaches requires an exorbitantly expensive amount of data for every exchange and every coin, in frequencies higher than daily. Given these frictions, we pursue an indirect approach to back out how the discrepancies arise by employing a simulated method-of-moment approach (McFadden (1989), Pakes and Pollard (1989)).

5.1 Structural model

We specify a structural model that explains how data discrepancies may arise. The model makes reduced-form assumptions on how exchange-level data is generated and how providers aggregate exchange data to compute aggregate prices and volumes. It has structural parameters that govern the aggregation methodology of a provider. We generate simulated samples from the model to compute divergence scores like those of Figures 3 and 4. Then, we then back out for each provider the structural parameters that minimize the distance between moments of the actual divergence scores and those that are simulated from our model. Our model has many components even though it makes simple assumptions. Figure 9 provides an overview of our model structure. We make our simulation algorithm as well as our simulated exchange data available in our data repository.

We first describe the model that governs how exchange-level data is generated. We consider a single coin (indexed with c) for which a provider (indexed with p) reports an aggregate close price $P^A_{c,p}$ and a 24-hour trading volume $V^A_{c,p}$. The coin has a true close global

price P_c^* and a true global daily volume V_c^* that apply exactly at midnight UTC. The true global price P_c^* is uniformly distributed in an interval ranging from $\$7 \times 10^{-5}$ (similar to Shiba Inu) to \$2,400 (similar to Ethereum). The true global daily trading volume V_c^* is uniformly distributed in an interval ranging from \$11,000 (extremely illiquid) to \$4.5 billion (liquid, on par with Ripple). We assume that there are $N_E = 1{,}500$ exchanges on which the coin may be traded, but the coin only experiences trades on $N_c \leq N_E$ exchanges. The number N_c of exchanges on which Coin c is traded is picked at random from a uniform distribution on $\{1, \ldots, N_E\}$. We assume that each exchange (indexed with e) reports its last trade before midnight including the price $P_{c,e}$ associated with the trade and the 24-hour volume $V_{c,e}$ at the time of the trade. Trades on an exchange occur with frequency $\xi_{c,e}$ which is uniformly distributed between 0.0001 (frequent trades) and 2 (infrequent trades). The trade frequency $\xi_{c,e}$ can be interpreted as the average number of minutes between consecutive trades. The last trade of the day occurs a certain number of minutes before midnight. We call this the trade delay $\Delta_{c,e}$ and assume that is exponentially distributed with rate $1/\xi_{c,e}$. This means that a coin can have different trade frequencies and trade delays on different exchanges, reflecting what we see in the real world.

The price $P_{c,e}$ associated with the trade is random and distributed according to a log-normal distribution. The mean of the price distribution is the true global price P_c^* at midnight. The log-volatility of the trade price is sampled from a uniform distribution ranging from $3\% \times \sqrt{\Delta_{c,e}}$ (comparable to Bitcoin over the time period $\Delta_{c,e}$) to $15.8\% \times \sqrt{\Delta_{c,e}}$ (comparable to a small cap coin like MANTRA over $\Delta_{c,e}$). This assumption implies that trades that occur with very small delays $\Delta_{c,e}$ are going to have prices that are very close to the true global price P_c^*, ensuring convergence as we approach midnight. The daily trading volume $V_{c,e}$ on an exchange is also log-normally distributed with mean given by the proportion $\pi_{c,e} V_c^*$ of the true global trading volume that is on average associated with the exchange, and log-volatility that is also uniformly distributed between 3% and 15.8%. Here, we do not apply the scaling factor $\sqrt{\Delta_{c,e}}$ for volatility because reported volumes are measured over 24 hours while reported prices are the last traded ones. We assume that the distribution $(\pi_{c,e} : \ e \in E_c)$ of trading volumes across exchanges follows the same power law as in Figure 5 but spanning the number N_c of exchanges on which the coin is listed. This model of exchange-level data is reduced-form, meaning that it does not reflect the economics of how and why trades arrive. Still, it makes realistic assumptions of possible price and

volume distributions while introducing substantial variability across coins.

We now describe how providers collect coin data from exchanges. For each Coin c, a provider only monitors a subset $E_{c,p} \subset E_c$ of the exchanges on which the coin is listed. The maximum number of exchanges $N_{E,p}$ that a provider tracks is a structural parameter that we back out from the divergence data. It holds that $|E_{c,p}| = \min\{N_c, N_{E,p}\}$, meaning that a provider cannot pull data from exchanges on which a coin is not listed. The constituents of $E_{c,p}$ are picked without replacement from a volume-weighted distribution so that exchanges that have higher volumes are more likely to be tracked. The structural parameter $N_{E,p}$ allows us to capture the empirical fact that different providers make different coverage choices (see Table 7).

A provider collects information on the last trade from each tracked exchange. However, a provider may be sloppy and introduce noise when collecting this information. There are two types of noise in our model. First, a provider may aim to collect data for Coin c but end up collecting data for another Coin $c' \neq c$ by mistake. This may happen if, for example, a coin just had a fork or a swap and the provider collects trade data for a stale coin. Such situations are possible given the findings of Section 3.1. We assume that the probability that a provider may mismatch a coin on an exchange is given by $p_{E,p} \in [0, 1]$. Our assumption implies that, for each tracked exchange $e \in E_{c,p}$, the provider flips a coin with success probability $1 - p_{E,p}$. If successful, the provider collects correct data for Coin c. If unsuccessful, the provider collects data for another Coin $c' \neq c$ and assumes that it is the data for Coin c. Regardless of whether the provider collects exchange data for the correct or wrong coin, we assume that the last traded price $\hat{P}_{c,e,p}$ and 24-hour volume $\hat{V}_{c,e,p}$ that a provider records are noisy, given by independent log-normally distributed random variables with mean equal to the true last reported price $P_{c,e}$ and 24-hour volume $V_{c,e}$, respectively, and log-volatilities given by $\sigma_p > 0$. Such noise me be due to fat finger mistakes when recording exchange-level data, which the findings of Section 3.2 suggest may be prevalent. The mismatch probability $p_{E,p}$ and the noise volatility σ_p are structural parameters that we back out from the divergence data. These structural parameters introduce measurement errors in the provider samples.

Finally, we describe how providers aggregate exchange-level data. We assume that all providers employ a slight variant of the methodology of CryptoCompare (see Online Appendix C). The methodology carries out a time-and-volume-weighted average of exchange data, subject to outlier removal and wash trading controls. The methodology proceeds as

follows. First, the provider tries to match the coin to an ID. The provider may get this wrong with probability $p_{C,p} \in [0,1]$. This can happen if, for example, there are multiple coins labeled with the same ID and the provider collects data for the right ID but matched to the wrong coin. Such instances are not uncommon based on Section 3.1. If the provider mislabels the coin, then the provider collects data for another Coin $c'' \neq c$ from all exchanges in $E_{c,p}$ and assumes that it belongs to Coin c. The mislabelling probability $p_{C,p}$ is a structural parameter that reflects the accuracy of a provider's ID system. We back it out from the divergence data.

Next, the provider checks how many exchanges have trade data for the targeted coin (which may the the wrong one in case of mislabeling). If there are fewer than three tracked exchanges with trade data ($|E_{c,p}| < 3$), then the provider does not return data for this coin. Otherwise, the provider moves on to detecting outliers. The provider first measures the median across all collected close prices and then discards any exchange for which the price is above a factor of four or below a factor of one-fourth relative to the median. A similar rule is followed by CryptoCompare. We collect the subset of tracked exchanges that survive these steps in $\hat{E}_{c,p}$. For all surviving exchanges, the provider imposes a control for wash trading of the following type. The provider scales down the collected volume $\hat{V}_{c,e,p}$ with a factor $\lambda_{c,e,p}$, where $\lambda_{c,e,p}$ is uniformly distributed in the interval $[1-\Lambda_p, 1]$ with $\Lambda_p \in [0,1]$. Here, Λ_p is final structural parameter that we back out from the divergence data. It governs how strongly a provider imposes wash trading controls, with higher values implying stronger wash trading controls and a value of $\Lambda_p = 0$ indicating that no controls are active.

In the last step, the provider aggregates data from the survived exchanges. The aggregate volume $V^A_{c,p}$ is given by the sum of all wash-trading-adjusted volumes:

$$V^A_{c,p} = \sum_{e \in \hat{E}_{c,p}} \hat{V}_{c,e,p} \times \lambda_{c,e,p}.$$

The aggregate price $P^A_{c,p}$ is a weighted average of the collected exchange prices given by:

$$P^A_{c,p} = \sum_{e \in \hat{E}_{c,p}} \hat{P}_{c,e,p} \times w_{c,e,p}.$$

The weights are based on relative volumes and time decay factors so that:

$$w_{c,e,p} \propto \hat{V}_{c,e,p} \times \lambda_{c,e,p} \times \tau_{c,e} \quad \text{and} \quad \sum_{e \in \hat{E}_{c,p}} w_{c,e,p} = 1.$$

The time decay factor $\tau_{c,e}$ is the same for all providers:

$$\tau_{c,e} = \begin{cases} 1, & \text{if } \Delta_{c,e} < 5h, \\ 0.8, & \text{if } 5h \leq \Delta_{c,e} < 10h, \\ 0.6, & \text{if } 10h \leq \Delta_{c,e} < 15h, \\ 0.4, & \text{if } 15h \leq \Delta_{c,e} < 20h, \\ 0.2, & \text{if } 20h \leq \Delta_{c,e} < 25h, \\ 0.001, & \text{otherwise} \end{cases}$$

Here, $\Delta_{c,e}$ is the trade delay of Coin c on Exchange e. This is the same formula also used by CryptoCompare. Finally, we compute divergence scores relative to the true global price and true global volume. The close price divergence for Provider p and Coin c is given by $P^A_{c,p}/P^*_c$ while the volume divergence is given by $V^A_{c,p}/V^*_c$.

5.2 Estimation approach

In total, our model has five structural parameters:

- $N_{E,p} \in \mathbb{N}$: Maximum number of exchanges tracked by Provider p
- $p_{C,p} \in [0,1]$: Probability that Provider p mislabels coins across all exchanges
- $p_{E,p} \in [0,1]$: Probability that Provider p mismatches a coin on an exchange
- $\sigma_p > 0$: Measurement error volatility of exchange-level trade records of Provider p
- $\Lambda_p \in [0,1]$: Strength of wash trading controls of Provider p

We back out estimates of these structural parameters by comparing moments of simulated divergence data generated from our model to actual divergence data from each provider as reported in Figures 3 and 4. To accomplish this, we generate 10,000 simulated samples from our model and evaluate the following ten moments:

- Fractions of close price divergences that land ±1, 2, 3, 4, and 5% outside of one.
- Fractions of volume divergences that land ±2, 4, 6, 8, and 10% outside of one.

We also compute the analogous moments of the actual close price and volume divergence scores of a provider. We then determine the structural parameters that minimize the sum of squared errors between the ten simulated and actual data moments. We chose these ten moments because they are evaluations of the respective cumulative distribution functions. With sufficient evaluations of the cumulative distribution function, we should be able to closely match the divergence distribution in the real data with that in the simulation. Note that the divergence scores in the data are recorded across coins and days in our sample. Because of this, we generate independent samples for different simulated coins, which can be interpreted as coin-day records.

Our structural model gives rise to substantial variation in the dispersion scores. This can be observed in Figure 10, which shows plots analogous to those in Figure 3 for the dispersion of close prices under different parametric assumptions. Even though providers in reality use different aggregation methodologies, we choose to impose the same methodology on all providers to back out structural parameters that can be compared across providers. A consequence of our simplifying assumptions is that the estimates of the structural parameters cannot be interpreted as being representative of what the providers actually do. Instead, they provide insights on whether the kinds of quality issues captured by our model can give rise to the discrepancies we observe in the data. In other words, our estimates are suggestive of what could be going wrong instead of being indicative of what is actually going wrong with the aggregation approaches of the different providers.

5.3 Estimates

Table 7 reports the estimates of the structural parameters for the different providers. There, we exclude CoinCap because it does not report volumes. The table also reports the moments implied by our simulated divergence samples and the actual provider divergence data, as well as the number of exchanges that each provider claims to track according to their websites.

We observe that our structural model can closely capture moments of the actual provider divergence scores. The root mean squared moment errors (RMSE) are lower than 9% for

all providers. Generally, the model is much better at capturing moments of close price divergences (RMSE lower than 1.1% across providers) than moments of volume divergences (average RMSE across providers is 7%). Our model is much better at capturing the divergences implied by the data from CryptoCompare (RMSE: 1.46%) and Messari (RMSE: 1.17%). This may be expected because our structural model closely follows the aggregation approach of CryptoCompare and Messari claims to partially pull data from CryptoCompare. These results validate that, while our model is a simplified version of reality, it captures the dynamics of data generation and data aggregation reasonably well.

We also observe that our model approximates the distribution of divergence scores by assuming that providers collect data only from a small amount of exchanges $N_{E,p}$, not exceeding 50 for any exchange. This happens even though the providers claim that they track upwards of 100 exchanges, with CoinGecko claiming they track more than 1,500 exchanges. This result should not be interpreted as meaning that the providers misrepresent the number of exchanges they track. Instead, our estimate $N_{E,p}$ is an average across all coins that a provider may report data for. Even though a provider may track a large number of exchanges overall, any single coin may only be listed on a small subset of exchanges. This phenomenon can result in a small estimate for $N_{E,p}$ even though the providers may overall track a large number of exchanges across all coins. Still, our estimates for $N_{E,p}$ suggest that part of the discrepancies we see in the data may be due to the fact that different providers collect coin data from small and differing subsets of exchanges.

We find that the wash trading intensity parameter Λ_p is estimated to be zero for all providers but CryptoCompare. The estimate of Λ_p for CryptoCompare is very small at less than 0.5%. These findings suggest that our model can approximate the distribution of divergence scores across providers without needing to assume that they impose strong wash trading controls. It implies that the discrepancies we observe in the data may not necessarily be due by differences in wash trading control approaches, as hypothesized in Section 3.2.1.

The mislabeling probability parameter $p_{C,p}$ is estimated to be positive for four out of seven providers, ranging between 0.1% for CoinGecko to 0.7% for Live Coin Watch. In contrast, the mismatch probability for a coin on an exchange is estimated to be positive for all providers. The estimates range from 0.3% for CoinGecko to 1% for CryptoCompare. While these numbers individually may appear small, they imply that our model requires significant error rates in order to match the distribution of divergence scores in the provider

data. As a reference, Bennin (1980) reports that the error rates in the CRSP and Compustat databases between 1962 and 1978 were 0.01% and 0.1%, respectively. These estimates applied before the advent of the advanced computational resources and data processing techniques that are available now, which has likely enabled lower error rates.[8] Our results suggest that mislabeling and mismatch errors may have to occur frequently in order to explain the discrepancies we observe in the aggregate data from the providers.

Finally, we estimate that that the measurement error volatility is positive across the board, with values larger than 1% for all providers except CoinMarketCap and Santiment. These estimates imply that our model requires significant measurement errors, on top of frequent mislabeling and mismatch errors, in order to match the distribution of divergence scores in the provider data. They suggest that fat-finger-like mistakes may drive part of the data discrepancies we observe across providers.

All in all, the estimates of our structural model suggest that there may be three phenomena that give rise to the data discrepancies we document. First, the providers may only collect coin-level data from a small number of exchanges that are not common across providers. Second, the providers may incur mislabeling and mismatch errors that are more frequent than in traditional databases. Third, the providers may introduce measurement errors when collecting trade data. Our estimates do not support the notion that differences in wash trading approaches may drive the discrepancies in the data. Still, the results of this section suggest that data quality issues, such as those we document in Section 3, may arise when providers do not exhaust all available trade data sources or when they do not impose sufficient quality controls to prevent mislabeling, mismatch, and measurement errors.

6 A solution: aggregation across providers

We propose a solution that aggregates data across providers and ensures quality.

Algorithm 6.1. *Fix the tolerance bounds* $\text{tol}_2, \text{tol}_3 \in [0, 1]$. *At the beginning of each aggregation period, we determine the set of unique tickers that exist across all providers. For each ticker, we do the following:*

(1) For each provider, we take the ID of the coin that matches the ticker. If there is more

[8] Newer estimates of error rates in commercial financial databases are scarce; see Liu (2020).

than one ID that matches the ticker for a provider, we take the ID of the coin with the largest market capitalization (that is, we neglect any smaller coins that may have duplicate tickers).

(2) Once we have one ID for each provider, we load historical market cap data and compute clusters based on the average market capitalizations over the past 7 days.

(a) We first use the function "Ckmeans.1d.dp" in R to determine the optimal number of clusters that minimizes the within-cluster sum of squares.

(b) Once the optimal number of clusters has been found, we merge any two clusters for which the respective centers are within tol_2 *percent of the average center across the two clusters.*

(c) We determine the cluster with the largest number of providers. We record the maximum number of providers in the cluster as N_{MC}*. If there is only one cluster with* N_{MC} *providers, we keep this cluster. If there are three ore more clusters with* N_{MC} *providers, we discard the ticker and iterate. If there are two clusters with* N_{MC} *providers, we keep the cluster with the lower average market capitalization. The providers in that cluster are more likely to quote circulating market capitalizations instead of fully diluted market capitalizations (see Online Appendix B.1). Else, we discard the ticker and iterate.*

(3) For the providers that survive, we load historical close prices and run a second clustering step based on the average close price over the previous week.

(a) We also use the function "Ckmeans.1d.dp" in R to determine the optimal number of clusters that minimizes the within-cluster sum of squares.

(b) Once the optimal number of clusters has been found, we merge any two clusters for which the respective centers are within tol_3 *percent of the average center across the two clusters.*

(c) We determine the cluster with the largest number of providers. We record the maximum number of providers in the cluster as N_{CP}*. If there is a single cluster with* N_{CP} *providers, we keep this cluster. Otherwise, we discard this ticker and iterate.*

(4) If two or fewer providers survive the previous steps, then we discard the ticker.

(5) If the ticker is not discarded, then the providers that survive are considered to supply data for the same coin. We only rely on these providers and aggregate their data using the median values for each metric.

Our aggregation approach repeats a set of two basic steps, once for market caps and then for closing prices. The first step is to build clusters of providers with similar data. The second step is to aggregate within the largest cluster using the median function. Both steps are important as they achieve different goals. By clustering providers with similar data, we seek to identify those providers that supply data for the same coin. This aims to resolve the identification issues highlighted in Section 3.1. By aggregating data in a cluster using the median function, we seek to remove outliers that may be due to the inconsistencies documented in Section 3.2. By repeating these two steps once for market caps and then for closing prices, we seek to filter out situations in which a provider kept an ID for a coin but the supplied data refers to a different coin. This can happen, for example, if an ID is reused for different coins by a provider or the underlying coin faced a token swap or fork. Such instances are not uncommon; see Table 1.

6.1 Validity

We face a unique problem because we are trying to aggregate data across providers without knowing the truth that we should benchmark against. By employing clustering algorithms and relying on the goodwill of the different providers, we are able to demonstrate that Algorithm 6.1 yields correct data on average in the asymptotic limit in which the number of data providers grows large.

Suppose that there are $J > 0$ providers that supply data for a given ticker. Provider $j \in \{1, \ldots, J\}$ supplies the past-week average market cap M_j and past week average closing price P_j. Provider j may provide correct data, wrong data, or no data for the given ticker. It is correct with probability $0 \leq p_{C,j} \leq 1$, in which case $P_j = \hat{P}$, where $\hat{P}$ is a random approximation that is drawn from a distribution with mean given by the true average closing price

P^* and variance σ_P^2. The reported market cap can take on one of two possible realizations:

$$M_j = \begin{cases} \hat{m} & \text{with probability } p_{m,j}, \\ \hat{M}, & \text{with probability } p_{M,j} = 1 - p_{m,j}. \end{cases}$$

Here, $\hat{m}$ ($\hat{M}$) is a random approximation of the average circulating (fully diluted) market cap, which is drawn from a distribution with mean given by the true average circulating (fully diluted) market cap m^* (M^*) with variance σ_m^2 (σ_M^2). It holds that $m^* < M^*$. The provider is wrong with probability $0 \leq p_{W,j} \leq 1$, in which case it reports an erroneous average market cap $\hat{X}$ and an erroneous average closing price $\hat{Y}$. We assume that the wrong data entries $\hat{X}$ and $\hat{Y}$ are drawn from independent distributions with means X^* and Y^* and variances σ_X^2 and σ_Y^2. Finally, a provider may not report any data at all with probability $1 - p_{C,j} - p_{W,j}$. The below theorem states conditions under which Algorithm 6.1 returns correct data in the asymptotic limit in which the number of data providers grows large ($J \to \infty$).

Theorem 6.2. *Take* tol_2, tol_3 *as fixed in Algorithm 6.1. Suppose that the following conditions are satisfied for a given coin:*

(A1) All means and variances are positive and finite: $0 < m^*$, M^*, X^*, P^*, Y^*, σ_m^2, σ_M^2, σ_X^2, σ_P^2, $\sigma_Y^2 < \infty$.

(A2) The probability that a provider reports correct data is asymptotically strictly positive: $\lim_{J\to\infty} \min\{p_{C,j} : j = 1, \ldots, J\} > 0$.

(A3) All providers are more likely to report circulating market cap over fully diluted or erroneous market caps: $p_{C,j} \times p_{m,j} > \max\{p_{C,j} \times (1 - p_{m,j}), p_{W,j}\}$ *for all* $j \geq 1$.

(A4) The erroneous market cap is not too close to the true circulating or fully diluted market caps: $|m^* - X^*| > \text{tol}_2 \times |m^* + X^*|$ *and* $|M^* - X^*| > \text{tol}_2 \times |M^* + X^*|$.

(A5) The erroneous closing price is not too close to the true closing price: $|P^* - Y^*| > \text{tol}_3 \times |P^* + Y^*|$.

(A6) The distribution of the circulating market cap is sufficiently far away from the erroneous market cap: if $m^* \leq X^*$, *then* $\liminf_{j=1,\ldots,J} \mathbb{P}\left(\hat{m} \leq \frac{|m^*+X^*|}{2}\right) \times p_{C,j} \times p_{m,j} > 0.5$. *Otherwise,* $\liminf_{j=1,\ldots,J} \mathbb{P}\left(\hat{m} \geq \frac{|m^*+X^*|}{2}\right) \times p_{C,j} \times p_{m,j} > 0.5$.

(A7) *The true and wrong close prices are separated by a sufficiently high margin in probability: if* $P^* \leq Y^*$, *then* $\mathbb{P}\left(\hat{P} \leq \frac{|P^*+Y^*|}{2}\right) > \mathbb{P}\left(\hat{Y} \leq \frac{|P^*+Y^*|}{2}\right)$. *Otherwise,* $\mathbb{P}\left(\hat{P} \geq \frac{|P^*+Y^*|}{2}\right) > \mathbb{P}\left(\hat{Y} \geq \frac{|P^*+Y^*|}{2}\right)$.

(A8) *There are no redundant data providers in the asymptotic limit: for all* $1 \leq i \neq j \leq J$, $\limsup_{J\to\infty} \mathbb{P}\left(M_i = M_j \text{ and } P_i = P_j\right) = 0$.

Then, our aggregation approach yields correct data for the requested ticker with increasing probability as the number of providers J *grows large:*

$$\lim_{J\to\infty} \mathbb{P}\left(\textit{Algorithm 6.1 yields correct data for the requested ticker}\right) = 1.$$

The proof of Thereom 6.2 can be found in Appendix A. The conditions required for convergence are fairly mild. Assumption (A1) is natural. It is the motivation for us to use variables like closing prices and market caps because they are more likely to have a well-defined distributions (unlike, e.g., volumes; see Section 3.2). Assumptions (A2) and (A3) impose goodwill on the providers. They imply that providers try to be truthful and report meaningful market metrics. Assumptions (A4) and (A5) are likely the strictest. They require that wrong data is centered away from correct data. These assumptions are hard to verify in practice when one does not know the correct data. The additional criteria in Steps 2 and 3 on when to merge or eliminate clusters are meant to address instances in which Assumptions (A4) and (A5) may not be valid in small samples ($J \approx 0$). Assumptions (A6) and (A7) are satisfied whenever the clusters are tightly concentrated around the correct data when the number of providers grows large. These assumptions ensure that Algorithm 6.1 can safely disregard information from the smaller clusters in the asymptotic limit $J \to \infty$ as those clusters are more likely to contain wrong data. Assumption (A8) implies that the probability that two different providers return exactly the same data converges to zero as the number of providers grows large. This assumption addresses the empirically observed phenomenon that providers reuse data from each other (Section 2) and suppresses it in the asymptotic limit in which infinite data vendors are available. We believe that this is a minor assumption because growing competition in the market for data as the number of vendors grows large would force redundant data vendors out of the market.

The methodological and theoretical results of this section connect to two strands of re-

search in the artificial intelligence literature. On the one hand, our approach aligns with truth inference algorithms that are commonly used to extract true values from crowdsourced data; see Sheng and Zhang (2019), Zheng et al. (2017), and others. Our algorithm relates to the truth discovery approach of Meng et al. (2015) that identifies correlated data entries through a decomposition of the variance-covariance matrix of the data providers. We, instead, rely on clustering to achieve this goal. On the other hand, our approach also relates to newer work on how to deal with missing data in financial applications. In particular, we show that Algorithm 6.1 is asymptotically robust to missing data. By relying on a cross-section of providers that report truthfully with sufficient probability according to Assumption (A2), we can complement missing data at one provider with data from the other providers. Because of this, we do not need to complete our dataset or impute missing data entries, as recently proposed by Bryzgalova et al. (2024) and Freyberger et al. (2024).

6.2 Simulation analysis

We evaluate the performance of our aggregation algorithm in a simulation case study. For this, we simulate market cap and close price data for a single ticker and a range of providers. We then assess the accuracy of Algorithm 6.1 as the number of providers grows large. We implement Algorithm 6.1 with tolerance bounds $\text{tol}_2 = 10\%$ for Step 2 and $\text{tol}_3 = 2\%$ for Step 3. We consider three different scenarios, parametrized as indicated in Table 8. The parameters are chosen as to pose different levels of challenges to Algorithm 6.1 while satisfying the conditions of Theorem 6.2. In Scenario 1, correct and wrong data are clearly differentiated from each other. Missing data is frequent (occurring in $1 - p_C - p_W = 20\%$ of instances), but the providers mostly report correct data when not missing. This is a scenario in which our aggregation methodology should have an easy time identifying the correct data providers. In Scenario 2, the gap between correct and wrong data is tight but the providers still report correct data frequently when not missing. This scenario should be more challenging for our aggregation algorithm. Scenario 3 is the most challenging for Algorithm 6.1. In this scenario, there is no missing data and the providers report correct and wrong data almost equally often. In all scenarios, we generate independent samples of market caps and close prices from log-normal distributions with mean and standard deviation parameters calibrated to match the coefficients of Table 8. We consider 1,000 simulated

samples covering up to 200 different providers.

Table 9 present the results of our simulation case study. Consistent with our goal of confronting Algorithm 6.1 with increasingly challenging tests, we see that the probability that the algorithm returns data from a provider that reports erroneously (Columns (2) & (5)) increases as we move from Scenario 1 to Scenario 3. In the most challenging Scenario 3, Algorithm 6.1 returns wrong data with up to 25% probability when less than 10 providers are available for aggregation. However, we observe that the probability that Algorithm 6.1 returns correct data (i.e., the target probability of Theorem 6.2, Columns (1) & (4) in Table 9) increases as the number J of providers grows large. This probability remains above 50% when there are more than six providers even in the most challenging Scenario 3. The simulation study confirms that Algorithm 6.1 is unaffected by missing data as convergence occurs even in Scenarios 1 and 2 that have 20% missing data on average. Table 9 indicates that Algorithm 6.1 detects and discards an increasing number of erroneous data providers as J grows large (Column (7)). It also shows that our approach often returns no data at all (Columns (3) & (6)), rather than reporting wrong data (Columns (2) & (5)), whenever it is unable to determine the correct set of providers in cases in which there is too much missing data or not enough providers survive the steps of Algorithm 6.1.

One issue we observe empirically is that some providers recycle data from others, resulting in repeated data entries in the provider cross section. Assumption (A8) requires that this does not occur asymptotically. But recycled data may hurt our aggregation approach in small samples ($J \approx 0$). To understand how our aggregation is impacted by recycled data, we repeat the simulation from Scenario 3 under the assumption that the first 10 providers in each simulation offer exactly the same data. Figure 11 reports the probabilities that Algorithm 6.1 returns correct, wrong, or no data as a function of the number J of providers in the presence of repeated data. We observe that the probability that Algorithm 6.1 delivers correct data increases to one as J grows large. This suggests that our convergence theorem still applies even when the first set of providers offer the same data. However, the rate of convergence is slower and our algorithm may deliver wrong data more frequently when some providers recycle data from each other. This is particularly the case when there are only a few providers available for aggregation. These results suggest that the practice of recycling data poses challenges for our algorithm when only a handful of providers are available, but does not hinder its convergence. All in all, our simulation study confirms the validity of

Theorem 6.2. It shows that our aggregation approach delivers asymptotically correct data as the number of data providers grows large, even in the presence of recycled data.

6.3 Empirical implementation

We implement Algorithm 6.1 with the same tolerance bounds as in our simulation study ($\text{tol}_2 = 10\%$ for Step 2 and $\text{tol}_3 = 2\%$ for Step 3). We apply it on a rolling daily basis and track the coins and providers that our algorithm identifies as having faulty data.

We begin with 2,852,943 provider-crypto-day instances of market capitalization and closing price combinations for the 553 primary coins in our crypto universe. Step 2 of our approach (market cap clustering) removes 231,584 (8.2%) data instances. Step 3 (close price clustering) removes additional 54,588 (1.9%) data entries. A final number of 72,830 (2.6%) data entries are removed due to Step 4 (minimum number of required providers). After our aggregation approach, we are left with 2,493,941 provider-crypto-day instances for 545 different cryptocurrencies. These statistics showcase an important aspect of our aggregation approach. It discards a substantial amount of provider data. We reject about 12.6% of all provider-crypto-day data instances. But we ultimately only discard few coins due to inconsistent data. We only lose 8 coins, or 1.4%, from our universe of 553 primary coins.

Figure 12(a) shows the coins that are most impacted by our aggregation approach. Several coins are impacted because the providers offer conflicting data. Some conflicts may be due to measurement errors. For example, on August 30, 2023, Live Coin Watch reported a close price for iExec (RLC) that was 25% lower than the median close price across all other providers. Our approach recognized this discrepancy and excluded Live Coin Watch as a provider for RLC starting on August 31 through September 10, 2023. Conflicts may also arise because coins are mismatched through their tickers. For example, the ticker SNT generally refers to the token Status on most providers. However, the largest token with the ticker SNT on Coinpaprika was Share NFT through May 2023. Our approach recognized this inconsistency and excluded Coinpaprika as a provider for SNT during this time.

We showcase the distribution of the instances in which our aggregation approach removes daily provider data in Figure 12(b). We see that CoinGecko is most frequently hit by our aggregation approach, while Coinpaprika and Santiment are least frequently hit. However, the frequencies in Figure 12(b) may misrepresent how reliable a data provider is because they

do not account for the amount of data that is supplied by the different providers. They also include instances in which a provider is discarded in Step 4 of our algorithm not because it has faulty data, but because there is not a large enough cross section of providers to draw reliable conclusions. We propose a grading scheme that more accurately reflects the reliability of a data provider in Section 7.3.

6.4 Alternative approaches

We compare our aggregation algorithm to two alternative approaches. One is a variant of our approach that uses the mean function instead of the median function in Step 5 of Algorithm 6.1. Such an approach may be more susceptible to the large outliers we document in Section 3 when the number of providers is small ($J \approx 0$). The second alternartive is a simpler variant that does not do Steps 1 through 4 and instead aggregates all providers using the median function. Such an approach may find it hard to discern between true and wrong data, and between circulating and fully-diluted market caps.

We show in Online Appendix D that our approach offers benefits over the two alternative approaches. In an extension of the simulation case study of Section 6.2 in which we also consider the two alternative approaches, we document that Algorithm 6.1 dominates the simpler median-based alternative that skips Steps 1-4 in small samples ($J \leq 10$). Our approach also has a higher rate of convergence as the number of providers grows large ($J \to \infty$). In addition, Algorithm 6.1 offers benefits over a variant that uses the mean in Step 5 for the aggregation of close prices when only a few providers are available ($J \leq 20$), especially in Scenarios 2 and 3 that pose the most challenging tests for aggregation. Applying the alternative approaches empirically as in Section 6.3, we show that our approach also offers practical benefits over the two alternatives. First, by using the median instead of the mean function, our approach becomes robust to outliers that may bias the aggregation in small samples when only a few providers are available. Second, by clustering providers prior to aggregating using the median, our approach removes wrongly reported data for coins that may be affected by ID issues (Section 3.2). These results show that, while our approach is not the only option to extract correct data from a cross section of crypto data vendors, it offers benefits over possible alternatives.

7 How our approach helps in practice

7.1 Sensible risk & return measurements

We use Algorithm 6.1 as implemented in Section 6.3 to aggregate data across providers and compute daily risk and return measurements for each coin in our sample. Table 10 reports the average metrics for the 10 largest coins over our sample, computed with data from the different providers as well as with our aggregated data. They also report the results of a two-sided paired difference test for the null hypothesis that the daily value implied by a provider (or our estimate) is equal to the median daily value across all providers (not including our estimate). These are quantitative tests for the graphical outcomes in Figure 6.

We observe that our aggregation approach yields estimates that are sensible. In Table 10, the average weekly returns computed using our approach are often in line with the median across providers, even though we show in Section 6.4 that our aggregates are often more accurate than simple median-based aggregates. Our aggregates are not impacted by some values that are deemed to be outliers by our paired difference tests: CoinMarketCap for ETH and XRP; Coinpaprika for BNB; and Santiment for SOL. We observe that full-day returns are more consistent across providers (as measured by the number of rejections in the paired difference test) than intraday returns. This is likely because open prices are noisier than close prices; see Section 3.2. Our volatility and value-at-risk estimates are on the lower side of the spectrum across providers. This is confirmed by our paired difference test, which mostly establishes that our estimates are statistically different than the median across providers. These results suggest that our approach is only weakly influenced by return outliers that may yield abnormally large volatility or value-at-risk estimates. Our aggregation approach yields a consistent number 1 ranking for Bitcoin (BTC), which is correct but would not be the case based on data from Coinpaprika.

The results of this section show that our aggregation approach enables the measurement of sensible risk and return metrics for individual coins. It achieves this by comparing data across providers and discarding any data pieces that appear abnormal. As a result, our approach yields return and risk measurements that are based on the highest quality data available across providers. We make our aggregated data openly available in our online repository. We include coin-level daily global open, high, low, and close prices, market caps, and volumes, as well as dispersion statistics of these metrics across providers.

7.2 Realistic portfolios

Our aggregation approach facilitates the construction of realistic portfolios. We show in Table 6 that our approach yields factor portfolios that are closely consistent with what would be obtained from a simple median-based aggregation approach. The overlap rates of the different legs of the factor portfolios of Liu et al. (2022) implied by our aggregation approach are higher than those implied by the different providers, often scoring above 99%. The portfolio weight root mean squared errors are also small. Figure 8 shows that the difference between the factor portfolio returns implied by our aggregated data and the factor portfolio returns implied by a simple median-based approach are negligible. This finding suggests that the factor portfolios implied by our aggregated data do not have overstated returns that could not be realistically captured out-of-sample, in contrast to the portfolios implied by the different provider datasets. We make our daily market, size, and momentum factor returns available in our online data repository.

Our results show that our aggregation approach yields portfolios that are robust to the data quality issues we document. We obtain portfolios that are similar to those implied by a simple median-based aggregation approach. Still, our approach offers one major benefit over a simple median-based aggregation alternative. As we show in Section 6.4, our approach discards coin data that may be impacted by mislabeling issues (such as those documented in Section 3.2) while a simple median-based aggregation approach does not. This difference may explain the small gaps in overlap and portfolio weights between our approach and the median-based approach in Table 6. It suggests that our approach may be a more reliable alternative when constructing portfolios over a simple median-based approach.

7.3 Quality ratings

Algorithm 6.1 requires a large cross section of providers as input. Obtaining access to multiple providers is costly and therefore not possible for many consumers of crypto data. To enable the consumption of high quality crypto data without the high costs, we propose a rating scheme that reflects the quality of a provider's data. We follow a standard academic grading scale based on the percentage of instances in the weekly data of a provider that our algorithm deems to be faulty; see Table 11 for our grading rules.

To fairly measure the quality of the data supplied by a provider, we compute the *discard*

rate which is given by the percentage of all daily data instances reported by a provider that are discarded by either Steps 2 or 3 of Algorithm 6.1. We intentionally do not include instances that are discarded by Step 4 of the algorithm to not penalize a provider that passes all of our quality checks but does not survive because there is not a large enough cross section of providers to aggregate from. We also intentionally normalize the discard rate with the number of daily instances reported by a provider, and not by the total number of discards as is done in Figure 12(b), to not penalize a provider that may naturally be more often discarded just because it provides more data. Table 11 reports the frequency of daily provider discard rates that fall within each grade bucket. We observe that the grading scheme is benevolent, with nearly 65% of all provider discard rates falling in the A range. These observations show that our grading scheme does not overly penalize providers for reporting data that may be slightly inconsistent. We also observe that low grades are rare, with only 5.1% of daily provider discard rates scoring D+ or below. This shows that it is generally hard to score poorly in our test, suggesting that low scores are particularly alarming.

The average discard rates of the individual providers over our sample period, together with the implied quality grades, are:

- CoinGecko: 26.8%. Grade: C.
- CoinMarketCap: 3.5%. Grade: A.
- Coinpaprika: 4.5%. Grade: A.
- Live Coin Watch: 7.7%. Grade: A-.
- Messari: 12.6%. Grade: B+.
- Santiment: 3.3%. Grade: A.

Figure 13 shows the time series of grades assigned to each provider based on the discard rates evaluated within a month. We observe that CoinMarketCap and Santiment generally score best, consistently achieving grades in the A range. That both of these providers would fare similarly well is unsurprising considering that Santiment recycles data from CoinMarketCap. CoinGecko performs worst overall, almost scoring an F for data from late 2019. In recent times, however, the quality of the data on CoinGecko has substantially improved. Coinpaprika and Live Coin Watch generally score well. However, the quality of the data supplied by these providers tends to be volatile, achieving scores as low at C in some months and A+ in other months. Finally, Messari performs average in our analysis. The quality of the data from Messari is better for early dates than for more recent dates. The low scores for Messari are surprising because Messari is the only enterprise-level provider we considered,

charging large annual fees to access their data (see Online Appendix A.1).

The results of this section provide guidance for consumers of crypto data as to which providers offer reliable quality market data. This may be helpful for consumers that only want to purchase access to single sources of data without needing to implement the aggregation approach of Algorithm 6.1.

7.4 Aggregate confusion index

Our approach enables the construction of an index that indicates how reliable aggregate market data from different providers is at any given point of time. Such an index may help consumers of crypto data identify periods of time in which data sold by the different providers may be of lower quality. We proceed as follows to construct the index. Every day, we count the number of instances in which Steps 2 or 3 of Algorithm 6.1 discard data from the different providers. We normalize by the number of available data points across providers. This rate represents our Aggregate Confusion Index. There is a negative relationship between data quality and the Aggregate Confusion Index: higher (lower) values of the Aggregate Confusion Index match periods of time in which data quality is lower (higher).

Figure 14 shows the time series of the index over our sample period. We observe that there is substantial time variation in the index. The minimum fraction of daily data discarded by Algorithm 6.1 is 4.8% while the maximum fraction is 22%. The index has generally been lower since 2023, suggesting an improvement in the quality of the data supplied by the different vendors in recent times.

We run regressions to understand the dynamics of the Aggregate Confusion Index. We conjecture that there may be several factors. For one, market activity may influence the quality of the data supplied by the providers. When the market runs hot and there are many transactions that the providers have to process and aggregate, it may be harder for them to deliver high quality data. Based on this intuition, we include in our regressions the following variables: market return, market volatility, total market cap, market volume, and total on-chain transactions (which we obtain from Token Terminal). We also conjecture that competition may push the providers to offer high quality data when the there is high demand for crypto data. To capture this mechanism, we also include the following variables in our regressions: Google Trends for the term "crypto data," total active addresses (also

from Token Terminal), as well as Total Value Locked (TVL) in decentralized applications that is a proxy for overall interest in the crypto landscape. We obtain TVL data from DefiLlama. We run these regressions at the weekly horizon using contemporaneous data. In the regressions, we measure market volatility as the standard deviation of the seven daily market portfolio returns (as implied by our aggregation approach) in the week. All other metrics are aggregated at the weekly frequency from daily data by either adding them up (volumes and transactions) or taking their averages (market cap, TVL, active addresses, and Google Trends interpolated with the most recent entry).

Columns (1)–(9) of Table 12 present our first set of results. We observe that certain market activity variables, such as total market cap and market volume, have strong positive relationship with the Aggregate Confusion Index. Total on-chain transactions has an ambivalent relationship while the market return and volatility have no statistically significant relationship. On the other hand, we find that proxies of interest for crypto data, such as Google Trends for the term "crypto data," total active addresses, and TVL, generally have a negative relationship with the Aggregate Confusion Index. One issue that may confound these estimates is that market activity and demand for crypto data may be high whenever the crypto market booms. To control for this effect, we normalize several variables with the concurrent market cap of the crypto market. Column (10) presents the estimates and further validates our findings. We observe that market activity variables, such as total market cap, market volume, and on-chain transactions, have a statistically significant positive relationship with the Aggregate Confusion Index. Market return and volatility continue to be statistically insignificant. These results suggest that the quality of crypto market data serviced by commercial data vendors may be lower during periods of time in which the crypto market runs hot, such as those when there are many transactions, large trading volumes, or high valuations. The estimates of Column (10) also show that the quality of aggregated crypto data may be higher whenever demand for crypto data is high, as proxied by high Google Trends for the term "crypto data," a large number of active addresses on different blockchains, or large value locked in different decentralized applications (TVL). These results suggest that the quality of commercial crypto data may be higher when there is high demand for it.

8 Conclusion & recommendations

We document pervasive quality issues in the supply of crypto data across many commonly used providers. We study the origins of these quality issues as well as their impact on coin and portfolio performance measurement. We propose an aggregation approach that addresses the quality issues and demonstrate its benefits both theoretically and in practical applications. We make our aggregated data available for public download in our online repository.

Our findings have important implications. They first imply that relying on a single provider for all crypto data needs is challenging due to quality issues that impact all providers. We provide an aggregation approach that yields sensible coin-level risk and return measurements, as well as portfolios, for those users that have the financial bandwidth to subscribe to multiple providers. For users that can only subscribe to a single provider, we offer an intuitive grading scale that reflects the quality of the data provided by the different vendors. In our analyses, CoinMarketCap and Santiment shine as providers offering market data with consistently high quality. We also provide insights on when the quality of aggregate crypto market data may be compromised.

From an academic point of view, our findings suggest that cryptocurrency studies may be particularly prone to p-hacking (Harvey (2017)) and non-standard errors (Menkveld et al. (2023)). This is because small choices in how to collect and aggregate crypto data can yield vastly different coin-level risk and return measurements. They also suggest that the replication of empirical findings of cryptocurrency studies may be challenging unless the exact input data is provided. Similar points have also been highlighted by Alexander and Dakos (2020) and Fieberg et al. (2024). We complement these papers by developing an aggregation approach to resolve some of the quality issues we document and delivering a quality grading scheme that can guide consumer choices in the market for crypto data.

For practitioners, our findings highlight the need to implement robust data collection and cleaning processes. We recommend that practitioners ensure that:

(1) Coins and IDs are thoroughly matched for each data provider to ensure that the data collected indeed belongs to the targeted cryptocurrency,

(2) Coins and IDs are thoroughly matched across data providers to ensure that data from different providers corresponds to the same cryptocurrency, and

(3) Robust methods that control for outliers are used when aggregating across data providers to control for measurement errors.

While even the most robust methods may still be affected by the data issues we highlight in our paper, the above recommendations can mitigate their effects. We offer one possible aggregation approach and demonstrate its validity both theoretically and in simulations.

For market overseers, our findings highlight the need for unified data approaches in the cryptocurrency market. At the minimum, a recommendation that arises from our paper is the establishment of a unified identification system for cryptocurrencies, similar to the CUSIP in North America or ISIN globally. Such systems can be developed and pushed by trade associations, as was done with CUSIP by the American Bankers Association (Ritter and Wool (2021)). Currently, there are several initiatives to establish global identifiers for crypto assets, including the Digital Token Identifier that follows an ISO standard similar to ISIN, and the Financial Instrument Global Identifier issued by Bloomberg and Kaiko. Our research provides empirical support for the establishment of these systems.

More broadly, our findings suggest that there may be a need to oversee crypto data vendors from a consumer protection point of view. In traditional markets, the Vendor Display Rule of the SEC governs how aggregate market data has to be reported in any setting in which a customer may use that data to make trading decisions. The regulation requires that vendors have to aggregate data from all national exchanges and securities associations. In a consequential 2015 letter denying a no-action request by BATS Global Markets, the SEC emphasized that data vendors cannot selectively choose a subset of exchanges to aggregate from. Failing to comply with the Vendor Display Rule can trigger penalties. In 2012, NYSE Euronext paid a civil penalty of $5 million, and faced censure and cease-and-desist orders, for providing unfair access to better data to some clients. Our results suggest that the market for crypto data may not fully align with the Vendor Display Rule. However, the Rule does not currently apply to the crypto market because many cryptocurrencies are not classified as securities that would be regulated by the SEC. As a result, the supply of aggregated crypto market data remains prone to quality issues such as those we document in our paper.

A Proof of Theorem 6.2

We proceed step by step. We can neglect Step 1 of the algorithm because the uniqueness of a ticker within a provider only affects the probability $p_{C,j}$ that the provider reports correct data. That is, whenever a provider uses the same ticker for different coins, the probability $p_{C,j}$ that a provider is correct will be lower and that impacts the other conditions of the Theorem. So we can assume for the sake of simplicity that the tickers and IDs are uniquely matched for each provider.

Because we employ a clustering method that minimizes the within-cluster sum-of-squares, Assumptions (A1) and (A2) implies that Step 2a of Algorithm 6.1 identifies 3 unique market cap clusters in the asymptotic limit in which we have infinite providers ($J \to \infty$). These clusters are centered around the true circulating market cap m^*, the true fully diluted market cap M^*, and the erroneous market cap X^*. Assumption (A4) implies that the center of the erroneous cluster is more than tol_2 percent away from either the true circulating market cap or the true fully diluted market cap cluster. As a result, the cluster containing primarily wrong providers is not merged with any of the clusters containing mostly correct providers in Step 2b as $J \to \infty$. The Law of Large Numbers together with Assumptions (A1) through (A3) imply that the cluster centered around the true circulating market cap m^* will be the most populated one in the asymptotic limit $J \to \infty$. As a result, the probability that this cluster will be the one picked by Step 2c of our aggregation approach converges to 1 as $J \to \infty$. The criteria for which cluster to pick in Step 2c only apply in small samples ($J \approx 0$) and are not binding in the asymptotic ($J \to \infty$).

For Step 3, because the reported price is either correct or wrong based on Assumption (A2), the fact that we again pick clusters as to minimize the within-cluster sum-of-squares in Step 3a means that we will end up with two distinct cluster in the limit $J \to \infty$: one centered around the true price P^* and one centered around the wrong price Y^*. Assumption (A5) implies that these two clusters are not merged in Step 3b when $J \to \infty$. To determine which cluster is the most populated one and ultimately picked by Step 3c, note that the probability that a correct provider is among the providers picked by Step 2 is bounded below by $\mathbb{P}\left(\hat{m} \leq \frac{|m^*+X^*|}{2}\right) \times p_{C,j} \times p_{m,j}$ if $m^* \leq X^*$ and $\mathbb{P}\left(\hat{m} \geq \frac{|m^*+X^*|}{2}\right) \times p_{C,j} \times p_{m,j}$ if $m^* > X^*$. Assumption (A6) implies that this probability is larger than 0.5. The Law of Large Numbers therefore implies that the cluster picked by Step 2 is primarily populated by

correct providers that report close prices that land close to P^*. This means that the cluster centered around P^* will be the most populated cluster in Step 3c as $J \to \infty$.

Finally, Assumption (A7) implies that the cluster picked by Step 3 contains more correct providers than wrong providers. Assumption (A8) implies that the probability that this cluster contains repeated entries vanishes when $J \to \infty$. These assumptions imply that the probability that the provider that reports the median value in the cluster is correct converges to one as $J \to \infty$. Step 4 does not bind in the asymptotic limit $J \to \infty$ because we proved that the cluster centered around P^* will be the most populated one.

References

Alexander, C. and M. Dakos (2020), 'A critical investigation of cryptocurrency data and analysis', *Quantitative Finance* **20**(2), 173–188.

Aloosh, Arash and Jiasun Li (2024), 'Direct evidence of bitcoin wash trading', *Management Science* **70**(12), 8875–8921.

Amiram, Dan, Evgeny Lyandres and Daniel Rabetti (2025), 'Trading volume manipulation and competition among centralized crypto exchanges', *Management Science* **71**(10), 8604–8622.

Barrera, Daniel R. and Simon Minovitsky (2021), 'Could Factors Have Explained Cryptocurrency Risk?', https://www.msci.com/research-and-insights/blog-post/could-factors-have-explained-cryptocurrency-risk. MSCI. Accessed: 2025-10-02.

Bennin, Robert (1980), 'Error Rates in CRSP and COMPUSTAT: A Second Look', *The Journal of Finance* **35**(5).

Berg, Florian, Julian F Kölbel and Roberto Rigobon (2022), 'Aggregate confusion: The divergence of esg ratings', *Review of Finance* **26**(6), 1315–1344.

Bryzgalova, Svetlana, Sven Lerner, Martin Lettau and Markus Pelger (2024), 'Missing financial data', *The Review of Financial Studies* **38**(3), 803–882.

Cong, Lin William, Xi Li, Ke Tang and Yang Yang (2023), 'Crypto wash trading', *Management Science* **69**(11), 6427–6454.

Falk, Brett Hemenway, Gerry Tsoukalas and Niuniu Zhang (2025), Can ai detect wash trading? evidence from nfts. Working Paper.

Fieberg, Christian, Steffen Günther, Thorsten Poddig and Adam Zaremba (2024), 'Non-standard errors in the cryptocurrency world', *International Review of Financial Analysis* **92**, 103106.

Freyberger, Joachim, Bjoern Hoeppner, Andreas Neuhierl and Michael Weber (2024), 'Missing data in asset pricing panels', *The Review of Financial Studies* **38**(3), 760–802.

Garman, Mark B. and Michael J. Klass (1980), 'On the estimation of security price volatilities from historical data', *The Journal of Business* **53**(1), 67–78.

Harvey, Campbell R. (2017), 'Presidential address: The scientific outlook in financial economics', *The Journal of Finance* **72**(4), 1399–1440.

Liu, Grace (2020), 'Data quality problems troubling business and financial researchers: A literature review and synthetic analysis', *Journal of Business & Finance Librarianship* **25**(3-4), 315–371.

Liu, Yukun, Aleh Tsyvinski and Xi Wu (2022), 'Common risk factors in cryptocurrency', *The Journal of Finance* **77**(2), 1133–1177.

McFadden, Daniel (1989), 'A method of simulated moments for estimation of discrete response models without numerical integration', *Econometrica* **57**(5), 995–1026.

Meng, Chuishi, Wenjun Jiang, Yaliang Li, Jing Gao, Lu Su, Hu Ding and Yun Cheng (2015), Truth discovery on crowd sensing of correlated entities, *in* 'Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems', SenSys '15, p. 169–182.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, ... and Zhen-Xing Wu (2023), 'Non-standard errors', *The Journal of Finance* **79**(3), 2339–2390.

Pakes, Ariel and David Pollard (1989), 'Simulation and the asymptotics of optimization estimators', *Econometrica* **57**, 1027–1057.

Ritter, Jay R. and Phillip Wool (2021), Anatomy of a World-Class Standards Body: The Origins and Future of the CUSIP System. Working Paper.

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

Shams, Amin (2022), Cryptocurrency exchanges and comovements of cryptocurrency returns. Working Paper.

Sheng, Victor S. and Jing Zhang (2019), 'Machine learning with crowdsourcing: A brief summary of the past research and future directions', *Proceedings of the AAAI Conference on Artificial Intelligence* **33**(1), 9837–9843.

Zheng, Yudian, Guoliang Li, Yuanbing Li, Caihua Shan and Reynold Cheng (2017), 'Truth inference in crowdsourcing: is the problem solved?', *Proc. VLDB Endow.* **10**(5), 541–552.

	CC	CCP	CG	CMC	CP	LCW	M	S
Ticker = ID (% of all coins)	100.00	13.1	16.2	15.00	0.01	100.00	0.00	13.59
Ticker contained in ID (% of all coins)	100.00	44.1	57.48	57.06	99.78	100.00	0.14	49.03
Name = ID (% of all coins)	17.14	61.95	39.94	42.58	0.01	12.65	0.00	48.78
Name contained in ID (% of all coins)	20.68	90.7	92.87	90.59	99.03	20.4	0.02	88.27
Id refers to more than one coin (% of all coins)	0.00	0.00	16.01	0.00	0.00	0.00	0.00	0.00
Id changes over time (% of primary coins)	0.00	0.00	0.91	0.58	21.3	0.61	4.17	2.90
Id unchanged after fork or swap (% of primary coins)			0.74	0	0.21	0.21	0.43	0.18

Table 1: *Coin identification.* This table reports summary statistics of the sample of labels for a coin, including the name, the ticker, and the provider ID. We consider only the primary data providers. We abbreviate the provider names as follows: CC: CryptoCompare; CCP: CoinCap; CG: CoinGecko; CMC: CoinMarketCap; CP: Coinpaprika; LCW: Live Coin Watch; M: Messari; and S: Santiment. Whenever we refer to "all coins," we mean the sample of all coins tracked by a provider as reported through their asset list API on February 14, 2025. Whenever we refer to "primary coins," we mean the union of the 250 largest coins each month sampled from November 2018 through October 2024. The top coin sample covers 553 distinct coins. In order to identify a token fork or swap, we assume that the return on the day in which such an event occurs will be close to the swap multiple without having any impact on the market capitalization of that token. So we identify token forks or swaps as days in which the relative error between either the open-to-close or close-to-close return of a coin and the swap multiple rounds to zero but the relative error between the market cap growth rate and the swap multiple does not. We consider swap multiples of 5, 10, 20, 25, 50, 100, 1,000, 10,000, and 1 million. We exclude CryptoCompare and Coincap from this analysis due to their lack of market cap data.

Providers	CC CCP	CC CG	CC CMC	CC CP	CC LCW	CC M	CC S
Open	0.9998	0.9999	0.9999	1.0000	0.9980	0.9999	1.0000
High			-0.0001	-0.0001		0.9710	-0.0001
Low			0.9999	1.0000		1.0000	0.9999
Close	0.9998	0.9999	0.9999	0.9865	0.9980	0.9999	1.0000
MC							
Volume		0.2967	0.2752	0.3025	0.2962	0.2693	0.2936

Providers	CCP CG	CCP CMC	CCP CP	CCP LCW	CCP M	CCP S	CG CMC
Open	0.9998	0.9998	0.9999	0.9954	0.9998	0.9999	1.0000
High							
Low							
Close	0.9998	0.9998	0.9999	0.9954	0.9998	0.9999	1.0000
MC							1.0000
Volume							0.9048

Providers	CG CP	CG LCW	CG M	CG S	CMC CP	CMC LCW	CMC M
Open	1.0000	0.9979	1.0000	1.0000	1.0000	0.9979	1.0000
High					0.9870		0.9999
Low					1.0000		1.0000
Close	0.9866	0.9979	1.0000	1.0000	0.9864	0.9979	1.0000
MC	0.2451	1.0000	1.0000	1.0000	0.2450	1.0000	1.0000
Volume	0.9534	0.9173	0.7994	0.9654	0.8965	0.9011	0.9360

Providers	CMC S	CP LCW	CP M	CP S	LCW M	LCW S	M S
Open	1.0000	1.0000	0.9999	1.0000	0.9979	0.9985	0.9985
High	0.9999		0.9999	0.9864			
Low	1.0000		0.9999	1.0000			
Close	1.0000	0.9859	0.9999	0.9858	0.9979	0.9979	0.9979
MC	1.0000	0.2445	1.0000	0.2450	1.0000	1.0000	1.0000
Volume	0.9360	0.9224	0.8430	0.9570	0.7926	0.9134	0.9134

Table 2: *Correlations across providers.* This table shows the pairwise correlations across the global metrics reported by the different providers. The providers are abbreviated as in Table 1. The samples are the same as in Figures 3 and 4. An empty value indicates that not enough data were available to compute the correlation.

		CC	CCP	CG	CMC	CP	LCW	M	S
Open	Mean	1.0682	1.0011	0.9999	1.0006	1.0013	1.0069	12.3528	1.0008
	St. dev.	28.29	0.05	0.02	0.02	0.02	1.71	1.1e+03	0.03
	Median	0.9999	1.0000	1.0000	1.0000	1.0004	1.0005	0.9998	1.0000
	Min.	3.39e-04	1.72e-09	7.02e-10	4.24e-01	4.62e-04	6.07e-05	2.99e-09	1.67e-03
		(AVAX)	(META)	(DGD)	(VTC)	(GTC)	(LEO)	(META)	(ALI)
	Max.	1.37e+04	1.29e+01	2.88e+00	8.95e+00	3.69e+00	7.49e+02	1.37e+05	8.95e+00
		(SAFEMOON)	(YOYOW)	(XBY)	(CQT)	(RDD)	(BTT)	(META)	(CQT)
High	Mean	1.39e+05			1.0001	1.0047		12.7755	1.0017
	St. dev.	9.74e+07			0.03	1.17		1.10e+03	0.38
	Median	1.0009			1.0000	1.0000		1.0010	0.9996
	Min.	3.42e-04			1.64e-07	4.53e-04		2.95e-09	1.63e-07
		(AVAX)			(IDEX)	(GTC)		(META)	(IDEX)
	Max.	6.79e+10			1.43e+01	4.97e+02		1.40e+05	9.00e+01
		(LYXE)			(NPXS)	(BTT)		(META)	(FLOW)
Low	Mean	1.0755			1.0006	1.0037		11.8633	1.0013
	St. dev.	2.90e+01			0.01	0.78		1.02e+03	0.02
	Median	0.9986			1.0000	1.0005		0.9973	1.0003
	Min.	5.78e-07			1.86e-02	4.54e-04		2.96e-09	1.76e-03
		(RDD)			(LPT)	(GTC)		(META)	(ALI)
	Max.	1.37e+04			2.00e+00	4.73e+02		1.35e+05	4.28e+00
		(SAFEMOON)			(BOND)	(BTT)		(META)	(HYN)
Close	Mean	1.0703	1.0012	0.9999	1.0006	30.4238	1.0077	12.3909	1.0007
	St. dev.	2.83e+01	0.05	0.02	0.05	1.98e+04	1.95	1.07e+03	0.03
	Median	0.9999	1.0000	1.0000	1.0000	1.0004	1.0005	0.9998	1.0000
	Min.	3.49e-04	1.69e-09	7.03e-10	2.95e-01	4.58e-04	6.07e-05	2.95e-09	1.73e-03
		(AVAX)	(META)	(DGD)	(REQ)	(GTC)	(LEO)	(META)	(ALI)
	Max.	1.37e+04	1.29e+01	3.54e+00	3.17e+01	1.33e+07	7.48e+02	1.39e+05	8.97e+00
		(SAFEMOON)	(YOYOW)	(XCP)	(HXRO)	(IDEX)	(BTT)	(META)	(CQT)
MC	Mean			1.6819	1.0193	32.3647	1.0032	1.0398	1.0014
	St. dev.			4.78e+01	7.07	2.02e+04	0.42	0.90	7.85e-02
	Median			1.0000	1.0000	1.0001	1.0004	0.9998	1.0000
	Min.			7.03e-10	7.56e-02	4.70e-03	1.48e-09	1.47e-10	2.50e-03
				(DGD)	(STRK)	(BTT)	(CHZ)	(SAFEMOON)	(ALI)
	Max.			3.60e+03	2.94e+03	1.33e+07	8.83e+01	5.47e+01	1.05e+01
				(OKT)	(DFI)	(IDEX)	(CDT)	(ONG)	(OXT)
Volume	Mean	0.6451		18.8909	1.5583	29.9538	1.5413	1.1948	1.4783
	St. dev.	1.04		9.10e+03	4.73e+01	1.91e+04	2.26e+02	2.54e+02	1.42e+01
	Median	0.5976		1.0977	1.0856	0.9975	0.9924	0.6690	1.0828
	Min.	5.31e-13		7.77e-12	9.43e-07	1.58e-08	2.61e-08	7.27e-08	2.39e-05
		(HOT)		(XBY)	(HXRO)	(STMX)	(FTT)	(NFT)	(BTT)
	Max.	1.95e+02		5.92e+06	2.82e+04	1.28e+07	1.32e+05	1.77e+05	8.13e+03
		(RFR)		(XCP)	(TBTC)	(IDEX)	(XBY)	(TBTC)	(NPXS)

Table 3: *Summary statistics of divergence scores.* This table shows summary statistics of the divergence scores in Figures 3 and 4. For the minimum and maximum divergence scores of each provider, we indicate the affected coin in parentheses. The providers are abbreviated as in Table 1.

	Size	Close	Open	High	Low	MC	Volume
CC	Big	1.08%	1.08%	5.21%	3.81%		95.18%
	Medium	2.39%	2.43%	9.49%	6.93%		88.00%
	Small	4.28%	4.31%	11.94%	9.26%		84.61%
CCP	Big	7.48%	7.42%				
	Medium	10.09%	10.16%				
	Small	12.34%	12.41%				
CG	Big	0.34%	0.34%			23.72%	73.27%
	Medium	0.90%	0.91%			39.52%	68.24%
	Small	1.72%	1.75%			43.07%	58.58%
CMC	Big	0.21%	0.19%	0.26%	0.16%	2.23%	62.13%
	Medium	0.63%	0.62%	0.67%	0.50%	4.48%	60.39%
	Small	0.92%	0.93%	0.82%	0.56%	8.95%	53.96%
CP	Big	0.43%	0.40%	0.60%	0.51%	5.90%	58.04%
	Medium	1.11%	1.06%	1.48%	1.22%	6.54%	64.05%
	Small	1.89%	1.81%	2.40%	2.00%	7.92%	57.82%
LCW	Big	1.15%	1.13%			9.37%	68.48%
	Medium	1.81%	1.73%			16.73%	59.03%
	Small	2.25%	2.27%			17.36%	55.88%
M	Big	0.59%	0.61%	1.78%	1.36%	7.68%	92.07%
	Medium	2.10%	2.29%	4.45%	4.45%	17.00%	90.13%
	Small	3.34%	3.50%	6.47%	6.45%	21.74%	88.61%
S	Big	0.24%	0.28%	0.39%	0.30%	2.60%	62.36%
	Medium	0.66%	0.79%	0.82%	0.66%	4.91%	60.45%
	Small	1.03%	1.14%	1.05%	0.68%	8.83%	53.52%

Table 4: *Large divergence instances.* We report the percentage of instances in which a divergence score deviates by more than ±5% from one. We split our sample based on whether a coin's market cap (as measured by the median market cap across all providers) on a day falls in the top ("Big"), middle ("Medium"), or bottom ("Small") tercile of market caps across coins on that day. Empty values indicate that the provider does not report the requested data metric. The providers are abbreviated as in Table 1. The samples are the same as in Figures 3 and 4.

Providers	CC CCP	CC CG	CC CMC	CC CP	CC LCW	CC M	CC S
Full-day return	0.0009	-0.0000	0.0098	-0.0000	0.0045	-0.0000	0.0027
Cleaned full-day return	0.6133	0.9820	0.9849	0.9814	0.9821	0.9848	0.9834
Intraday return	0.0157	-0.0000	0.0291	-0.0000	0.0296	0.0011	0.0285
Historical volatility	-0.0034	-0.0001	0.0037	-0.0001	0.0050	-0.0001	0.0177
Cleaned historical volatility	0.5033	0.8357	0.8565	0.8993	0.9188	0.8595	0.8340
Intraday volatility			0.3778	0.3927		0.2876	0.3308
95% value-at-risk	0.8076	0.8513	0.8569	0.8447	0.8216	0.7401	0.8513
Ranking							

Providers	CCP CG	CCP CMC	CCP CP	CCP LCW	CCP M	CCP S	CG CMC
Full-day return	-0.0019	0.4513	0.0750	0.3818	0.0001	0.4845	-0.0000
Cleaned full-day return	0.6334	0.6342	0.6305	0.6301	0.6167	0.6340	0.9936
Intraday return	-0.0019	0.4503	0.0749	0.3826	0.0012	0.4816	-0.0000
Historical volatility	-0.0016	0.7793	0.0762	0.6115	0.0030	0.5145	-0.0003
Cleaned historical volatility	0.5829	0.5501	0.5625	0.4469	0.6768	0.6563	0.8018
Intraday volatility							
95% value-at-risk	0.8660	0.8712	0.8660	0.8600	0.6288	0.8735	0.9613
Ranking							0.9778

Providers	CG CP	CG LCW	CG M	CG S	CMC CP	CMC LCW	CMC M
Full-day return	-0.0005	0.0000	-0.0000	-0.0000	-0.0007	0.9265	0.4044
Cleaned full-day return	0.9922	0.9892	0.9873	0.9936	0.9937	0.9915	0.9996
Intraday return	-0.0005	0.0000	-0.0000	-0.0000	-0.0007	0.9283	0.4026
Historical volatility	0.0005	-0.0007	-0.0001	-0.0013	0.0018	0.0657	0.1920
Cleaned historical volatility	0.8175	0.8082	0.8537	0.8742	0.9050	0.8287	0.9893
Intraday volatility					0.8090		0.8771
95% value-at-risk	0.9526	0.9439	0.7838	0.9554	0.9338	0.9289	0.9965
Ranking	0.9581	0.9659	0.9815	0.9733	0.9776	0.9827	0.9974

Providers	CMC S	CP LCW	CP M	CP S	LCW M	LCW S	M S
Full-day return	0.4044	-0.0009	-0.0005	-0.0007	-0.0004	0.9172	0.9172
Cleaned full-day return	0.9996	0.9898	0.9857	0.9933	0.9857	0.9907	0.9907
Intraday return	0.4026	-0.0009	0.0037	-0.0007	0.0016	0.9154	0.9154
Historical volatility	0.1920	0.0009	0.0255	0.0018	0.0131	0.7866	0.7866
Cleaned historical volatility	0.9893	0.8791	0.8599	0.8899	0.7870	0.8052	0.8052
Intraday volatility	0.8771		0.4715	0.7213			
95% value-at-risk	0.9965	0.9318	0.7490	0.9393	0.7309	0.9296	0.9296
Ranking	0.9974	0.9656	0.9682	0.9771	0.9735	0.9845	0.9845

Table 5: *Correlations across providers for risk and return metrics.* This table shows the pairwise correlations across metrics computed with data from the different providers. See Figure 6 for descriptions of how these metrics are computed and the samples used for their computation. The providers are abbreviated as in Table 1. An empty value indicates that not enough data was available to compute the correlation.

		CG	CMC	CP	LCW	M	S	Us
(a) Market portfolio:								
Overlap (l)	Mean	0.9389	0.9603	0.8694	0.7599	0.9285	0.9528	0.9993
	St. dev.	0.0249	0.0231	0.1141	0.0656	0.0364	0.0271	0.0026
	Median	0.9400	0.9700	0.8900	0.7500	0.9400	0.9500	1.0000
	Min.	0.8400	0.8600	0.1429	0.6300	0.8300	0.8600	0.9900
	Max.	0.9900	1.0000	0.9900	0.9500	0.9900	1.0000	1.0000
RMSE	Mean	0.0006	0.0002	0.0036	0.0016	0.0002	0.0002	0.0000
	St. dev.	0.0003	0.0002	0.0174	0.0011	0.0002	0.0002	0.0000
	Median	0.0006	0.0001	0.0008	0.0014	0.0002	0.0002	0.0000
	Min.	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
	Max.	0.0021	0.0016	0.1839	0.0079	0.0015	0.0016	0.0001
(b) Size factor portfolio:								
Overlap (l)	Mean	0.6793	0.8426	0.5151	0.1100	0.7240	0.7875	0.9934
	St. dev.	0.1247	0.1048	0.2931	0.2130	0.1759	0.1354	0.0169
	Median	0.6500	0.8500	0.5000	0.0000	0.8000	0.8000	1.0000
	Min.	0.2500	0.3500	0.0000	0.0000	0.1500	0.3000	0.9500
	Max.	0.9500	1.0000	1.0000	0.9500	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.9606	0.9912	0.9000	0.8926	0.9799	0.9907	0.9998
	St. dev.	0.0284	0.0199	0.1298	0.0572	0.0307	0.0203	0.0028
	Median	0.9500	1.0000	0.9000	0.9000	1.0000	1.0000	1.0000
	Min.	0.8500	0.9000	0.0000	0.7500	0.9000	0.9000	0.9500
	Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RMSE	Mean	0.0122	0.0074	0.0183	0.0197	0.0095	0.0095	0.0006
	St. dev.	0.0034	0.0036	0.0331	0.0032	0.0041	0.0045	0.0021
	Median	0.0125	0.0077	0.0163	0.0205	0.0092	0.0103	0.0000
	Min.	0.0005	0.0001	0.0002	0.0005	0.0003	0.0001	0.0000
	Max.	0.0214	0.0185	0.4089	0.0253	0.0203	0.0195	0.0102
(c) Momentum factor portfolio:								
Overlap (l)	Mean	0.8911	0.9305	0.8131	0.6852	0.8795	0.9225	0.9955
	St. dev.	0.0618	0.0482	0.1380	0.1045	0.0745	0.0526	0.0149
	Median	0.9000	0.9500	0.8500	0.7000	0.9000	0.9474	1.0000
	Min.	0.6842	0.8000	0.0000	0.3500	0.5789	0.7500	0.9000
	Max.	1.0000	1.0000	1.0000	0.9500	1.0000	1.0000	1.0000
Overlap (s)	Mean	0.8700	0.9144	0.7959	0.6647	0.8667	0.9072	0.9936
	St. dev.	0.0685	0.0589	0.1434	0.1058	0.0887	0.0593	0.0168
	Median	0.8947	0.9000	0.8421	0.6500	0.9000	0.9000	1.0000
	Min.	0.6000	0.6000	0.0000	0.3125	0.4737	0.6500	0.9500
	Max.	1.0000	1.0000	1.0000	0.9500	1.0000	1.0000	1.0000
RMSE	Mean	0.0203	0.0165	0.0279	0.0353	0.0193	0.0148	0.0023
	St. dev.	0.0237	0.0260	0.0458	0.0338	0.0263	0.0215	0.0105
	Median	0.0108	0.0056	0.0133	0.0202	0.0076	0.0064	0.0000
	Min.	0.0003	0.0001	0.0006	0.0007	0.0003	0.0001	0.0000
	Max.	0.1165	0.1127	0.6084	0.1336	0.1194	0.1105	0.1025

Table 6: *Summary statistics of portfolio differences.* This table shows summary statistics of the weekly differences of the market, size, and 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. The market portfolio is long-only so it does not have an "Overlap (s)" entry. "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.

		CC	CG	CMC	CP	LCW	M	S
Claimed number of exchanges tracked		160	1,500	790	500		210	
Structural parameter estimates:								
$N_{E,p}$		15	40	47	48	46	14	47
$p_{C,p}$		0.0062	0.0010	0.0000	0.0037	0.0072	0.0000	0.0000
$p_{E,p}$		0.0104	0.0027	0.0046	0.0060	0.0056	0.0098	0.0045
σ_p		0.0135	0.0162	0.0032	0.0105	0.0178	0.0126	0.0032
Λ_p		0.0046	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Overall moment RMSE		0.0146	0.0570	0.0841	0.0774	0.0716	0.0117	0.0852
Close price divergence moments:								
Outside ±1% from 1	Actual	0.1392	0.1043	0.0467	0.0920	0.1479	0.1179	0.0491
	Simulated	0.1400	0.1115	0.0428	0.0803	0.1539	0.1161	0.0416
Outside ±2% from 1	Actual	0.0715	0.0379	0.0214	0.0393	0.0640	0.0556	0.0225
	Simulated	0.0595	0.0296	0.0300	0.0438	0.0580	0.0469	0.0290
Outside ±3% from 1	Actual	0.0463	0.0207	0.0126	0.0228	0.0363	0.0351	0.0134
	Simulated	0.0460	0.0154	0.0230	0.0331	0.0379	0.0354	0.0221
Outside ±4% from 1	Actual	0.0329	0.0138	0.0083	0.0154	0.0241	0.0254	0.0090
	Simulated	0.0396	0.0119	0.0186	0.0272	0.0309	0.0310	0.0178
Outside ±5% from 1	Actual	0.0250	0.0098	0.0058	0.0112	0.0171	0.0198	0.0064
	Simulated	0.0355	0.0101	0.0154	0.0241	0.0266	0.0263	0.0149
RMSE of these five moments		0.0077	0.0055	0.0089	0.0107	0.0065	0.0055	0.0081
Volume divergence moments:								
Outside ±2% from 1	Actual	0.9325	0.7605	0.6698	0.6919	0.7099	0.9356	0.6670
	Simulated	0.9646	0.8917	0.8705	0.8686	0.8753	0.9653	0.8705
Outside ±4% from 1	Actual	0.9070	0.6943	0.6127	0.6258	0.6417	0.9143	0.6081
	Simulated	0.9265	0.7856	0.7430	0.7417	0.7528	0.9295	0.7433
Outside ±6% from 1	Actual	0.8836	0.6426	0.5677	0.5758	0.5902	0.8914	0.5627
	Simulated	0.8908	0.6782	0.6147	0.6107	0.6327	0.8962	0.6148
Outside ±8% from 1	Actual	0.8627	0.5981	0.5286	0.5339	0.5474	0.8672	0.5241
	Simulated	0.8566	0.5768	0.5016	0.4974	0.5206	0.8635	0.5017
Outside ±10% from 1	Actual	0.8422	0.5584	0.4942	0.4968	0.5100	0.8417	0.4901
	Simulated	0.8237	0.4871	0.3931	0.3870	0.4155	0.8329	0.3931
RMSE of these five moments		0.0192	0.0804	0.1187	0.1089	0.1011	0.0157	0.1203

Table 7: *Estimates of structural parameters for different providers.* This table lists the estimates of the structural parameters of the model of Section 5.1 and Figure 9 for the different providers (excluding CoinCap because it does not report volumes). The providers are abbreviated as in Table 1. We compute these estimates via a simulated method of moments. We generate 10,000 samples of coins from our structural models and compute the following 10 moments of the simulated divergence scores: fractions of close price divergences that land ±1, 2, 3, 4, and 5% outside of one, and fractions of volume divergences that land ±2, 4, 6, 8, and 10% outside of one. We then determine the set of structural parameters that minimize the sum of squares between the simulated moments and the actual moments of the divergence scores of a provider in Figures 3 and 4. Because the number of tracked exchanges $N_{E,p}$ is an integer, we carry out a stepwise search across values of $N_{E,p} \in \{1, \ldots, 100\}$. For each value of $N_{E,p}$, we determine the parameter vector $(p_{C,p}, p_{E,p}, \sigma_p, \Lambda_p)$ that minimizes the root mean squared error across simulated and actual moments. We use the "L-BFGS-B" optimization routine in R subject to boundary constraints on the parameters. We then pick the value of $N_{E,p}$, together with its minimal parameter vector $(p_{C,p}, p_{E,p}, \sigma_p, \Lambda_p)$, that minimizes the overall root mean squared error. The minimal root mean squared error is reported in row "Moment RMSE." We also report the root mean squared error for the moments of close price and volume divergences separately. Finally, the table also lists the number of exchanges that a provider claims to track based on the information on their websites. An empty value means that no information is provided.

Parameter	Interpretation	Scenarios 1	2	3
p_C	Probability that provider reports correctly	0.7	0.7	0.6
p_W	Probability that provider reports wrongly	0.1	0.1	0.4
p_m	Probability that correct provider reports circulating market cap	0.8	0.8	0.84
m^*	Correct circulating market cap mean	1 billion	1 billion	1 billion
σ_m	Correct circulating market cap standard deviation	1 million	35 million	35 million
M^*	Correct fully-diluted market cap mean	100 billion	100 billion	100 billion
σ_M	Correct fully-diluted market cap standard deviation	10 million	100 million	100 million
X^*	Wrong market cap mean	50 million	800 million	800 million
σ_X	Wrong market cap standard deviation	10 million	100 million	100 million
P^*	Correct close price mean	1,000	1,000	1,000
σ_P	Correct close price standard deviation	100	100	100
Y^*	Wrong close price mean	100	900	900
σ_Y	Wrong close price standard deviation	20	1,000	1,000

Table 8: *Simulation parametrizations.* This table reports the parameters assumed in the different simulation scenarios for the test of Algorithm 6.1. The parameter choices ensure that the conditions of Theorem 6.2 are satisfied.

	J	(1) Probability that algorithm returns market cap from a correct provider	(2) Probability that algorithm returns market cap from a wrong provider	(3) Probability that algorithm is unable to return market cap	(4) Probability that algorithm returns close price from a correct provider	(5) Probability that algorithm returns close price from wrong provider	(6) Probability that algorithm is unable to return close price	(7) Probability that provider with wrong data is removed
Scenario 1	3	0.146	0.000	0.854	0.146	0.000	0.854	0.358
	4	0.312	0.000	0.688	0.312	0.000	0.688	0.615
	5	0.456	0.002	0.542	0.456	0.002	0.542	0.752
	6	0.623	0.010	0.367	0.623	0.010	0.367	0.833
	7	0.726	0.015	0.259	0.726	0.015	0.259	0.920
	8	0.791	0.013	0.196	0.791	0.013	0.196	0.959
	9	0.838	0.012	0.150	0.838	0.012	0.150	0.968
	10	0.848	0.012	0.140	0.848	0.012	0.140	0.977
	15	0.908	0.004	0.088	0.908	0.004	0.088	0.995
	20	0.936	0.002	0.062	0.936	0.002	0.062	0.998
	30	0.956	0.001	0.043	0.956	0.001	0.043	0.999
	40	0.960	0.000	0.040	0.960	0.000	0.040	1.000
	50	0.966	0.000	0.034	0.966	0.000	0.034	1.000
	100	0.984	0.000	0.016	0.984	0.000	0.016	1.000
	200	0.991	0.000	0.009	0.991	0.000	0.009	1.000
Scenario 2	3	0.170	0.008	0.822	0.177	0.001	0.822	0.267
	4	0.346	0.012	0.642	0.351	0.007	0.642	0.455
	5	0.503	0.022	0.475	0.513	0.012	0.475	0.563
	6	0.654	0.024	0.322	0.660	0.018	0.322	0.629
	7	0.750	0.021	0.229	0.753	0.018	0.229	0.682
	8	0.805	0.016	0.179	0.806	0.015	0.179	0.718
	9	0.858	0.011	0.131	0.853	0.016	0.131	0.746
	10	0.877	0.013	0.110	0.877	0.013	0.110	0.760
	15	0.922	0.009	0.069	0.920	0.011	0.069	0.791
	20	0.953	0.004	0.043	0.949	0.008	0.043	0.828
	30	0.977	0.003	0.020	0.975	0.005	0.020	0.869
	40	0.980	0.007	0.013	0.982	0.005	0.013	0.889
	50	0.988	0.005	0.007	0.990	0.003	0.007	0.909
	100	0.989	0.007	0.004	0.990	0.006	0.004	0.906
	200	0.997	0.003	0.000	0.994	0.006	0.000	0.916
Scenario 3	3	0.199	0.148	0.653	0.243	0.104	0.653	0.321
	4	0.313	0.203	0.484	0.370	0.146	0.484	0.403
	5	0.467	0.204	0.329	0.510	0.161	0.329	0.528
	6	0.518	0.250	0.232	0.574	0.194	0.232	0.602
	7	0.544	0.240	0.216	0.597	0.187	0.216	0.629
	8	0.578	0.244	0.178	0.623	0.199	0.178	0.638
	9	0.609	0.217	0.174	0.667	0.159	0.174	0.680
	10	0.639	0.222	0.139	0.715	0.146	0.139	0.672
	15	0.708	0.197	0.095	0.779	0.126	0.095	0.676
	20	0.767	0.162	0.071	0.817	0.112	0.071	0.711
	30	0.824	0.113	0.063	0.857	0.080	0.063	0.792
	40	0.893	0.064	0.043	0.900	0.057	0.043	0.867
	50	0.914	0.047	0.039	0.922	0.039	0.039	0.897
	100	0.948	0.017	0.035	0.951	0.014	0.035	0.936
	200	0.952	0.014	0.034	0.950	0.016	0.034	0.942

Table 9: *Results of simulation case study.* This table reports outcome probabilities in the simulation study for the convergence of Algorithm 6.1 in the three different parametric scenarios described in Table 8. We consider 1,000 simulated samples and evaluate probabilities via the frequency of occurrence in the simulated samples.

		BTC	ETH	BNB	SOL	XRP	ADA	DOT	DOGE	SHIB	AVAX
Full-day return	CC	0.0017	0.0021	0.0035	0.0057	0.0014	0.0138	0.0013	0.0047	0.0025	1.9934
	CCP	0.0020	0.0010	0.0013	0.0040	0.0009	0.0004	-0.0004	0.0018	0.0015	0.0013
	CG	0.0017	0.0021	0.0030	0.0057	0.0013	0.0019	0.0013	0.0045	0.0025	0.0032
	CMC	0.0017	0.0021	0.0030	0.0043	0.0013	0.0020	0.0013	0.0045	0.0024	0.0032
	CP	0.0017	0.0021	0.0030	0.0053	0.0015	0.0021	0.0020	0.0045	0.0020	0.0022
	LCW	0.0017	0.0020	0.0029	0.0055	0.0017	0.0019	0.0014	0.0037	0.0004	0.0032
	M	0.0017	0.0022	0.0030	0.0057	0.0014	0.0020	0.0013	0.0046	0.0025	0.0032
	S	0.0017	0.0021	0.0030	0.0043	0.0013	0.0020	0.0014	0.0045	0.0025	0.0032
	Us	0.0017	0.0021	0.0030	0.0034	0.0013	0.0020	0.0013	0.0046	0.0002	0.0033
Intraday return	CC	0.0017	0.0021	0.0028	0.0056	0.0014	0.0022	0.0013	0.0047	0.0024	0.0032
	CCP	0.0020	0.0010	0.0013	0.0040	0.0009	0.0004	-0.0004	0.0018	0.0014	0.0013
	CG	0.0017	0.0021	0.0030	0.0056	0.0013	0.0019	0.0013	0.0045	0.0024	0.0032
	CMC	0.0017	* 0.0020	0.0029	0.0041	* 0.0013	0.0019	0.0013	0.0045	0.0021	0.0032
	CP	0.0017	0.0022	** 0.0028	0.0053	0.0015	0.0021	0.0022	0.0046	0.0020	0.0022
	LCW	0.0017	0.0020	0.0029	0.0055	0.0016	0.0019	0.0014	0.0038	0.0004	0.0031
	M	0.0017	0.0021	0.0030	0.0056	0.0014	0.0020	0.0013	0.0046	0.0024	0.0032
	S	0.0018	0.0021	0.0030	** 0.0044	0.0016	0.0019	0.0015	0.0047	0.0025	0.0034
	Us	0.0017	0.0021	0.0029	0.0034	0.0014	0.0020	0.0013	0.0046	0.0002	0.0032
Historical volatility	CC	*** 0.0327	*** 0.0415	*** 0.0466	*** 0.0626	*** 0.0465	*** 0.1133	*** 0.0478	*** 0.0600	** 0.0533	*** 12.8178
	CCP	*** 0.0191	*** 0.0216	*** 0.0203	*** 0.0357	*** 0.0266	*** 0.0281	*** 0.0264	*** 0.0308	*** 0.0338	*** 0.0345
	CG	*** 0.0320	*** 0.0408	*** 0.0410	0.0624	*** 0.0453	*** 0.0465	*** 0.0476	0.0583	*** 0.0549	*** 0.0579
	CMC	*** 0.0320	*** 0.0406	*** 0.0407	*** 0.0598	*** 0.0456	*** 0.0472	*** 0.0475	*** 0.0583	*** 0.0539	** 0.0581
	CP	0.0323	*** 0.0409	*** 0.0409	** 0.0608	0.0459	*** 0.0472	0.0482	0.0583	0.0428	*** 0.0505
	LCW	*** 0.0320	*** 0.0403	*** 0.0411	*** 0.0615	*** 0.0437	*** 0.0468	*** 0.0472	*** 0.0570	*** 0.0457	*** 0.0573
	M	*** 0.0327	*** 0.0417	*** 0.0415	*** 0.0627	*** 0.0466	*** 0.0480	*** 0.0479	*** 0.0596	*** 0.0546	*** 0.0581
	S	*** 0.0321	*** 0.0406	*** 0.0407	*** 0.0597	*** 0.0456	*** 0.0472	** 0.0476	0.0585	*** 0.0546	*** 0.0581
	Us	*** 0.0322	*** 0.0408	*** 0.0409	* 0.0611	*** 0.0457	*** 0.0472	*** 0.0474	*** 0.0581	*** 0.0486	*** 0.0579
Intraday volatility	CC	*** 0.0295	** 0.0332	0.0341	*** 0.0612	*** 0.0431	*** 0.0479	*** 0.0467	*** 0.0543	*** 0.0512	*** 0.0569
	CCP										
	CG										
	CMC	*** 0.0259	* 0.0329	0.0342	*** 0.0529	*** 0.0371	*** 0.0411	*** 0.0436	*** 0.0428	*** 0.0457	*** 0.0525
	CP	*** 0.0249	*** 0.0320	*** 0.0334	*** 0.0524	*** 0.0358	*** 0.0404	*** 0.0446	*** 0.0414	*** 0.0371	*** 0.0458
	LCW										
	M	*** 0.0298	*** 0.0373	*** 0.0383	*** 0.0589	*** 0.0426	*** 0.0459	*** 0.0461	*** 0.0526	*** 0.0495	*** 0.0554
	S	*** 0.0253	*** 0.0320	0.0344	*** 0.0520	*** 0.0362	0.0411	0.0432	*** 0.0414	0.0462	0.0519
	Us	*** 0.0259	*** 0.0328	0.0343	*** 0.0531	*** 0.0371	*** 0.0413	*** 0.0436	*** 0.0433	*** 0.0414	*** 0.0525
95% value-at-risk	CC	*** -0.0504	*** -0.0643	*** -0.0636	** -0.0898	*** -0.0688	-0.0719	-0.0751	*** -0.0740	*** -0.0748	*** -0.0889
	CCP	*** -0.0265	*** -0.0297	*** -0.0289	*** -0.0499	*** -0.0363	*** -0.0388	*** -0.0407	*** -0.0424	*** -0.0406	*** -0.0483
	CG	*** -0.0497	*** -0.0631	*** -0.0599	*** -0.0887	*** -0.0669	*** -0.0711	*** -0.0745	* -0.0714	*** -0.0762	-0.0885
	CMC	*** -0.0491	*** -0.0630	*** -0.0602	*** -0.0905	*** -0.0677	*** -0.0717	*** -0.0753	*** -0.0713	*** -0.0751	*** -0.0886
	CP	*** -0.0498	-0.0634	*** -0.0603	*** -0.0870	*** -0.0673	*** -0.0709	*** -0.0835	*** -0.0702	*** -0.0611	*** -0.0816
	LCW	*** -0.0491	*** -0.0630	-0.0607	-0.0890	*** -0.0639	*** -0.0707	*** -0.0747	*** -0.0708	*** -0.0671	-0.0884
	M	*** -0.0505	*** -0.0646	*** -0.0621	*** -0.0902	*** -0.0691	*** -0.0731	*** -0.0753	-0.0715	*** -0.0758	*** -0.0886
	S	*** -0.0491	*** -0.0627	*** -0.0602	*** -0.0906	*** -0.0682	*** -0.0722	-0.0752	*** -0.0718	*** -0.0757	*** -0.0887
	Us	*** -0.0494	*** -0.0631	*** -0.0604	*** -0.0915	*** -0.0679	*** -0.0719	*** -0.0749	*** -0.0705	-0.0710	*** -0.0884
Size rankings	CC										
	CCP										
	CG	1.0000	2.0366	*** 4.9177	*** 14.7807	3.9973	*** 6.6903	*** 8.7433	*** 15.0645	*** 12.7878	*** 18.6924
	CMC	1.0000	2.0366	*** 4.8664	13.5858	4.0037	7.0229	*** 8.8200	14.7017	*** 13.3187	18.2406
	CP	1.0005	2.0380	*** 4.8280	*** 15.6409	*** 3.9147	*** 7.0075	*** 9.3680	*** 14.4610	*** 10.2662	*** 14.3098
	LCW	1.0000	2.0228	* 4.8563	*** 12.3904	*** 3.9220	*** 6.9423	*** 8.7112	*** 12.9675	*** 10.5612	*** 14.4841
	M	1.0000	2.0366	*** 4.8728	*** 14.5068	** 4.0073	*** 7.0567	*** 8.6823	*** 14.8225	*** 12.6061	*** 18.4068
	S	1.0000	2.0366	*** 4.8669	** 13.5969	4.0037	** 7.0137	8.8015	14.7045	*** 13.2882	18.2480
	Us	1.0000	2.0366	4.8614	13.7872	4.0041	** 7.0329	8.8016	*** 14.7198	*** 10.8241	*** 18.1247

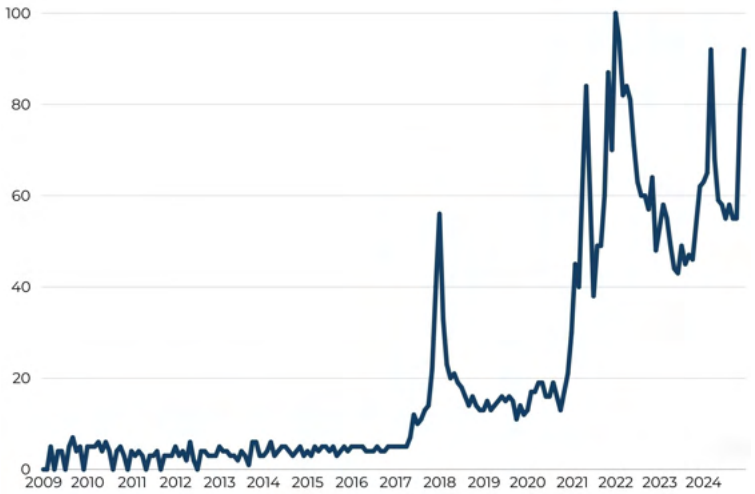
Table 10: *Daily coin-level risk & return measurements from providers.* This table reports the time series average of a metric reported in Table 5 for the 10 largest coins based on average market capitalization over our sample, where we compute market caps using the aggregation approach of Section 6. For each provider, we compute a metric using the same approach as in Section 4.1. The providers are abbreviated as in Table 1. "Us" refers to estimates based on data aggregated using Algorithm 6.1. Stars indicate significance for a two-sided paired difference test of the null hypothesis that the daily value is equal to the median daily value across providers (not including our estimate). ***, **, and * denote significance at the 99.9%, 99%, and 95% confidence levels, respectively. Empty values indicate that no data was available for aggregation.

Grade	Discard rate needs to be		Empirical
	More or equal than	Less than	frequency
A+	0%	3%	28.0%
A	3%	7%	27.9%
A-	7%	10%	9.0%
B+	10%	13%	7.6%
B	13%	17%	9.5%
B-	17%	20%	2.9%
C+	20%	23%	3.1%
C	23%	27%	4.1%
C-	27%	30%	2.9%
D+	30%	33%	1.3%
D	33%	37%	2.2%
D-	37%	40%	1.3%
F	40%	100%	0.3%

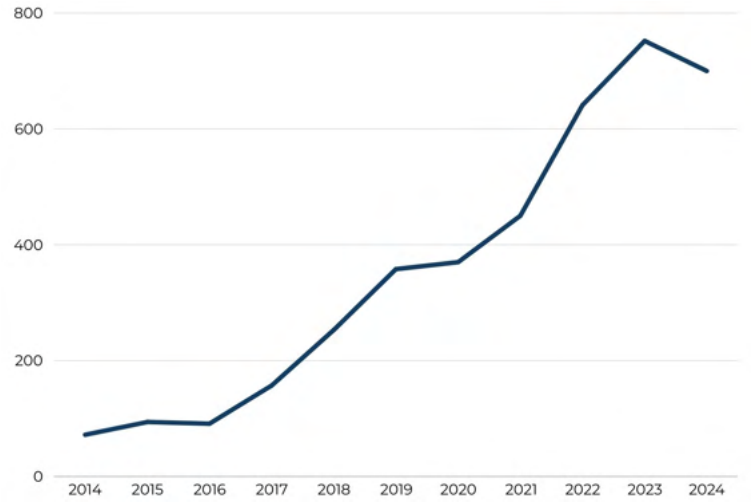
Table 11: *Provider quality grading.* This table reports the rules we use to grade the quality of the data supplied by a provider. We measure the discard rate as the percentage of all daily data instanced reported by a provider that are discarded by Steps 2 or 3 of Algorithm 6.1. Column "Empirical frequency" reports the fraction of daily provider discard rates that fall within each one of the grade buckets. There are a total of 13,116 daily provider discard rate instances over our sample period.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	*** 0.1005	*** 0.0928	*** 0.2640	*** -0.2345	*** 0.1377	*** 0.1630	*** 0.1660	*** 0.1125	*** -0.3760	-0.1221
	(76.5800)	(36.3610)	(7.1232)	(-4.5059)	(10.9261)	(16.4044)	(19.9892)	(43.9364)	(-5.1772)	(-1.2312)
Market return	0.0201								0.0037	0.0022
	(1.4860)								(0.4053)	(0.2362)
Market volatility		*** 0.2510							-0.0099	-0.0574
		(3.5810)							(-0.1715)	(-0.9834)
Market cap (log)			*** -0.0060						* 0.0089	* 0.0090
			(-4.4085)						(2.0659)	(2.4801)
Market volume (log)				*** 0.0126					** 0.0085	
				(6.4424)					(3.2382)	
TVL (log)					** -0.0016				0.0008	
					(-2.9492)				(0.4973)	
On-chain trans. (log)						*** -0.0039			*** 0.0439	
						(-6.3220)			(7.8917)	
Active addresses (log)							*** -0.0049		*** -0.0532	
							(-7.9495)		(-9.5776)	
Google Trends								*** -0.0266	*** -0.0306	
								(-5.2943)	(-4.3063)	
Market volume (norm.)										*** 0.0949
										(5.3668)
TVL (norm.)										* -0.1904
										(-2.1378)
Trans. (norm., $\times 10^3$)										*** 1.2934
										(6.8833)
Addr. (norm., $\times 10^4$)										*** -3.1593
										(-10.3914)
Google Trends (norm.)										** -0.2121
										(-2.8304)

Table 12: *Regressions of Aggregate Confusion Index.* We regress the Aggregate Confusion Index of Figure 14, aggregated at weekly horizons, on several weekly metrics that capture either market activity or demand for crypto data. Weeks are measured from Wednesday to Tuesday. "Market return" is the weekly return of the market portfolio of Section 7.1. "Market volatility" is measured as the standard deviation of daily market returns, which we compute from coin-level data based on market cap weights computed at the start of a week. "Market cap" is the sum of the weekly average market caps of the 100 coins included in the construction of the market portfolio each week as outlined in Section 7.1. "Market volume" is the weekly sum of the daily coin volumes for the 100 coins included in the market portfolio. "TVL" is the average of daily aggregate TVL data obtained from DefiLlama. "Transactions" and "Active addresses" represents the sum of all on-chain transactions and average of all daily active addresses recorded on Token Terminal during a week. "Google Trends" is the most recent monthly entry in the Google Trends database for the term "crypto data" between 2017 and 2024. We apply certain transformations for some variables: either logarithmic (log), or a normalization (norm.) with the concurrent total weekly market cap of the market. When required, we abbreviate active addresses as "Addr." and on-chain transactions as "trans."



(a) Google Trends.



(b) Crypto data publications.

Figure 1: *Crypto data trends.* The top half of the figure shows the Google Trends time series for the term "*crypto data.*" It is given by the number of worldwide Google web searches for a given month, divided by the maximum value in the time series. The sample interval is January 1, 2009, through December 31, 2024. The bottom half shows the number of new publications that were published in a year for which either the title or the abstract contained both of the words *"crypto"* and *"data."* We obtain these data from Dimensions.ai.

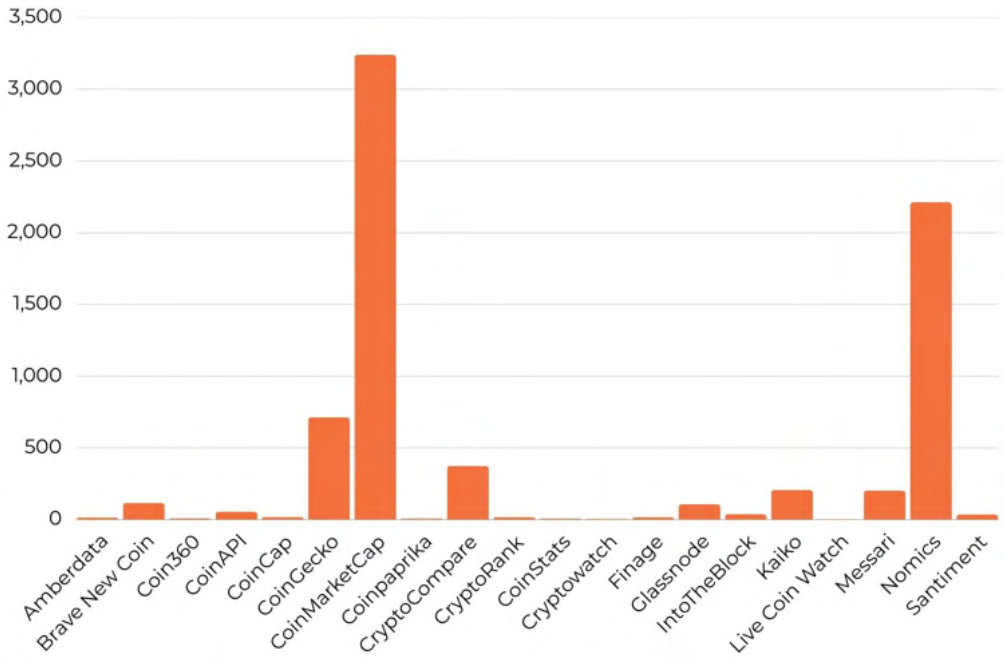


Figure 2: *Research papers that mention the data providers.* This image shows the total number of publications that mention the names of the providers together with the word *"crypto"* in the body of the paper. There are a total of 6,570 distinct research papers that mention one of the provider names and crypto in their bodies. We obtain these data from Dimensions.ai. The sampled time frame runs through the end of 2024.

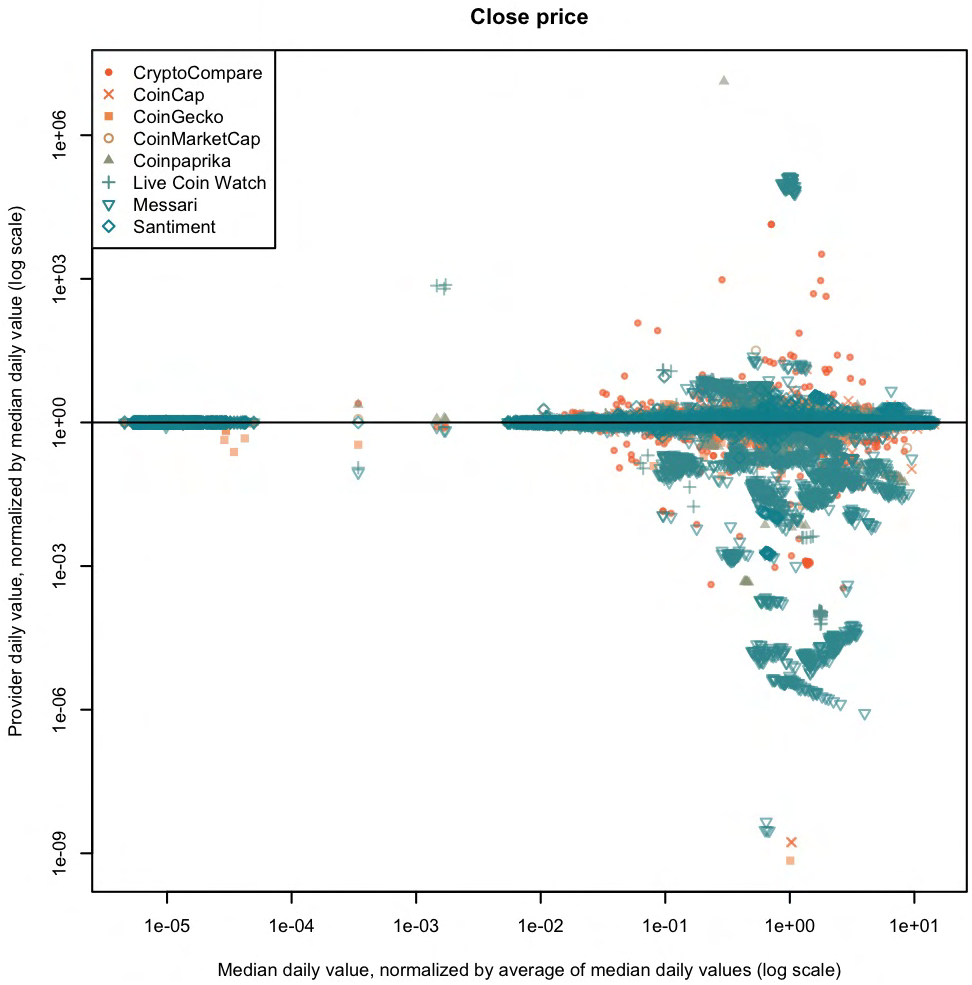
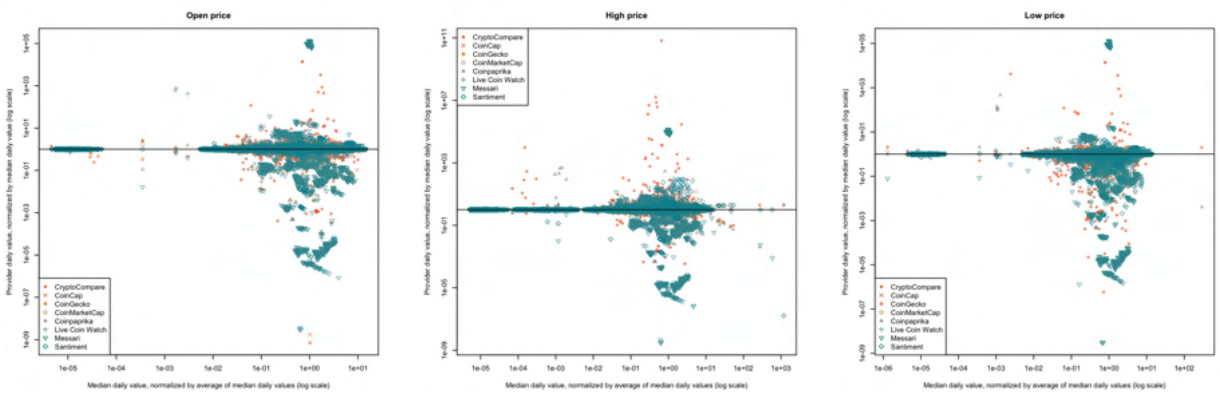


Figure 3: *Divergence scores of daily global close prices.* This plot shows the divergence scores of daily global close prices, where we define the divergence score for a metric as the metric for a coin on a day divided by the median metric across providers for that coin of the given day. The figure plots close price divergence scores against the corresponding daily cross-provider median close price, normalized by the average daily cross-provider median close price over the lifespan of the sampled coin that intersects with our sample period. The sample period is November 1, 2018, through October 31, 2024, and covers 553 unique cryptocurrencies. The plot includes 3,620,955 provider-crypto-day close price observations.



(a) Open prices. (b) High prices. (c) Low prices.

(d) Market caps. (e) Trading volumes.

Figure 4: *Divergence scores for additional market data.* These graphs show divergence scores analogous to those of Figure 3 for the daily global open, high, and low prices, as well as market capitalizations and trading volumes reported the different providers. The sample period is November 1, 2018, through October 31, 2024, and covers 553 unique cryptocurrencies. The number of provider-crypto-day entries for each plot is as follows. Open prices: 3,620,768. High prices: 2,475,103. Low prices: 2,475,058. Market caps: 2,808,467. Trading volumes: 3,379,576.

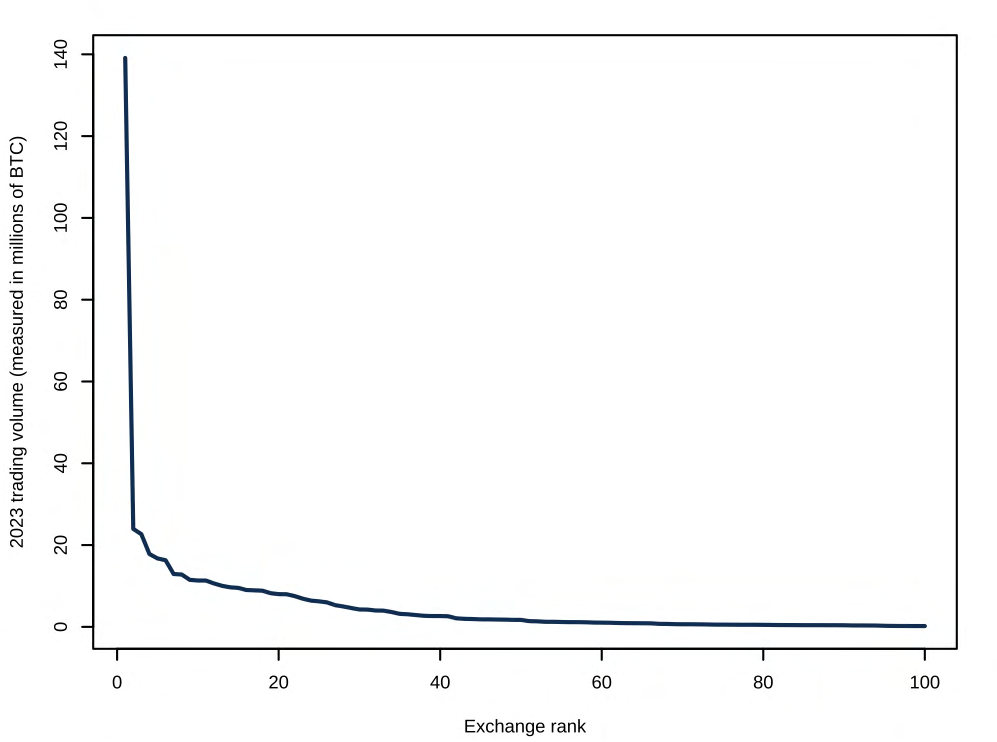
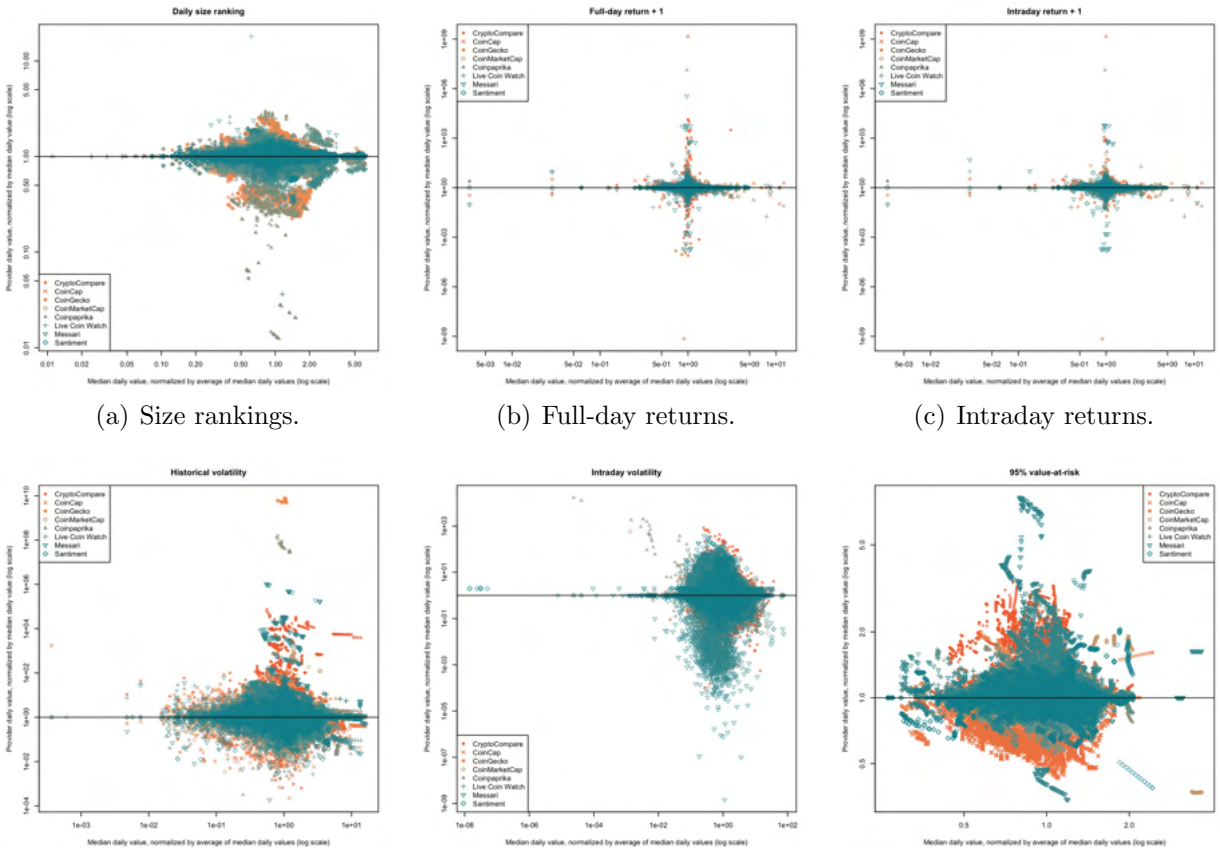


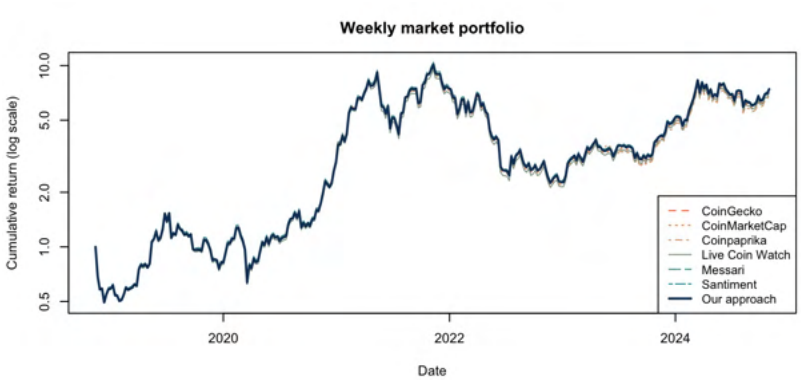
Figure 5: *Distribution of annual volume in 2023 across exchanges.* We restrict ourselves to the largest 100 exchanges. We consider the total reported volume of an exchange across all crypto pairs it offers for trading. We obtain these data from CoinGecko.



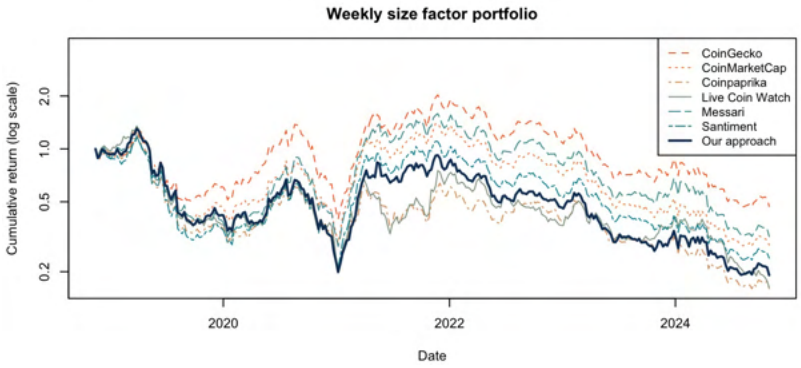
(a) Size rankings. (b) Full-day returns. (c) Intraday returns.

(d) Historical volatilities. (e) Intraday volatilities. (f) 95% value-at-risk.

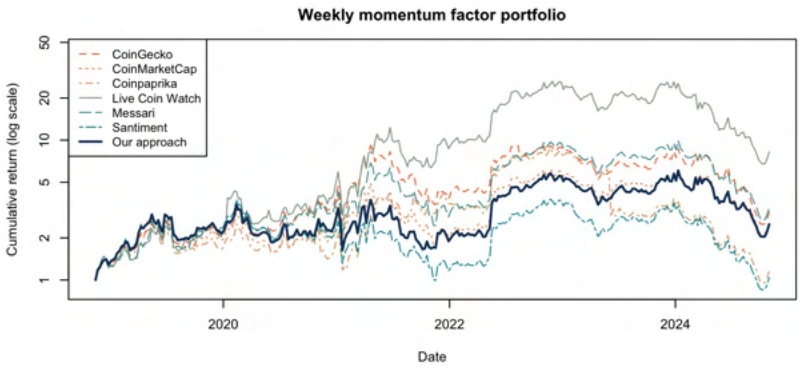
Figure 6: *Divergence scores for daily risk and return measurements.* These figures plot divergence scores analogous to those of Figures 3 and 4 for different daily risk and return metrics. The metrics include the open-to-close intraday return, the close-to-close full-day return, the intraday volatility as in Garman and Klass (1980), the historical volatility over 30-day rolling windows, the 95% value-at-risk computed over rolling 365-day windows with a minimum of 180 non-missing entries, and the daily size ranking of a coin. The sample period is November 1, 2018, through October 31, 2024, and covers 553 unique cryptocurrencies. The number of provider-crypto-day entries for each plot is as follows. Size rankings: 1,311,794. Full-day returns: 3,610,369. Intraday returns: 3,620,664. Historical volatilities: 3,753,251. Intraday volatilities: 2,450,286. 95% value-at-risk: 3,012,797.



(a) Market portfolio.

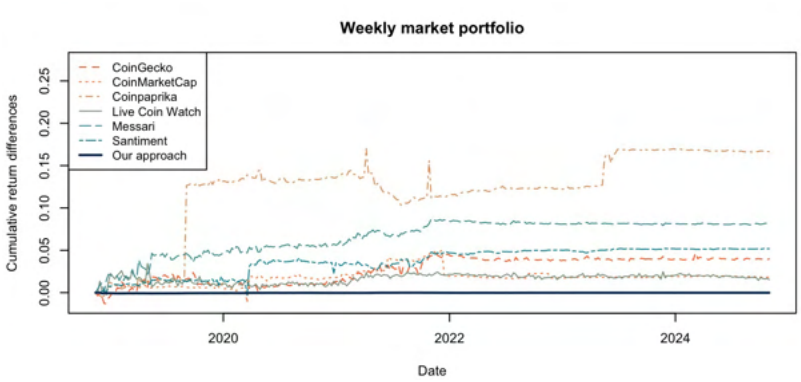


(b) Size factor portfolio.

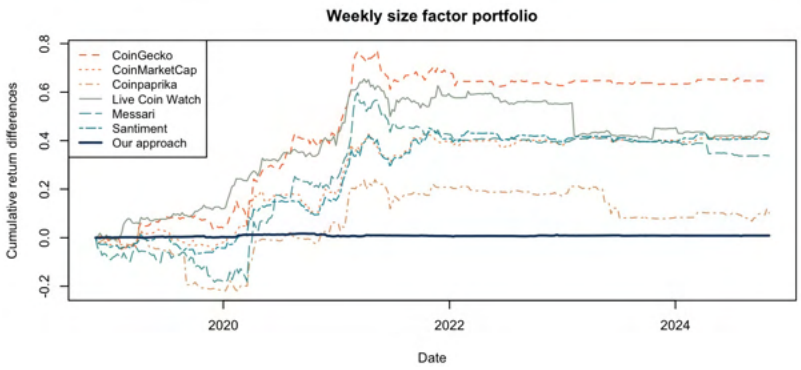


(c) Momentum factor portfolio.

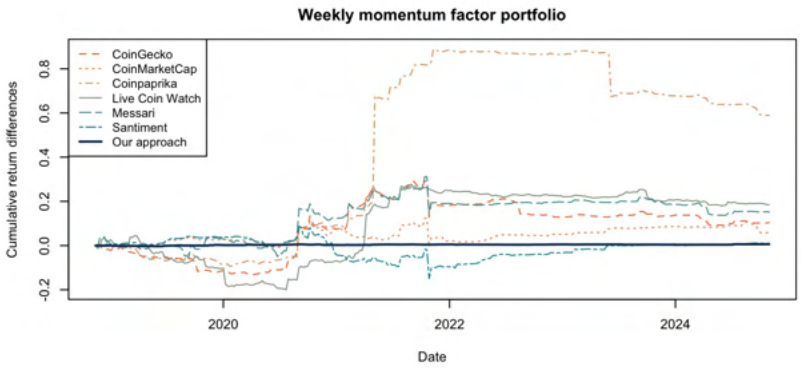
Figure 7: *Cumulative returns of factor portfolios.* This figure shows the cumulative returns of the market portfolio, as well as the size and momentum factor portfolios, based on data from different providers. It also shows the cumulative return of the factor portfolios implied our aggregation approach from Section 6.



(a) Market portfolio.



(b) Size factor portfolio.



(c) Momentum factor portfolio.

Figure 8: *Cumulative return differences of factor portfolios.* Each week, we measure the return of a factor portfolio in two different ways while keeping the portfolio composition fixed. Once based on the return data of the provider (matching Figure 7) and a second time by taking the median weekly open and close prices across providers to compute weekly coin returns. We plot the cumulative differences between these two alternative ways of measuring portfolio returns for each provider as well as our aggregation approach from Section 6.

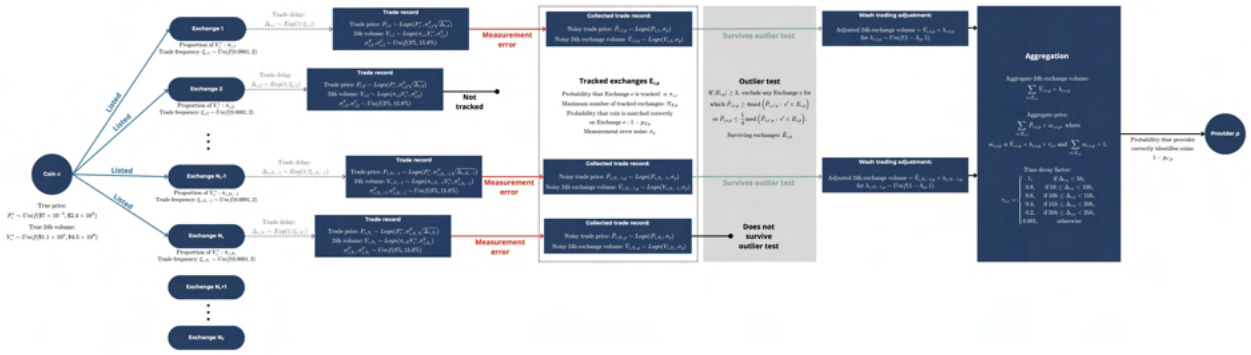
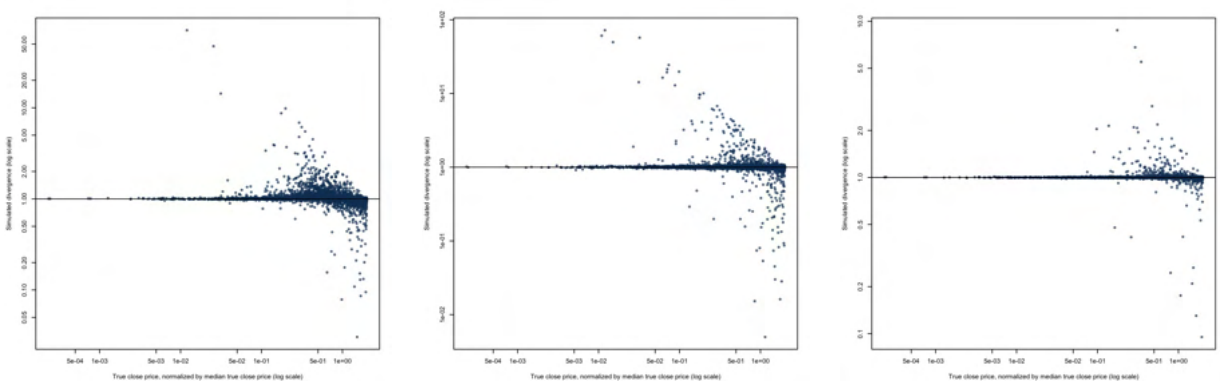


Figure 9: *Structural model for data discrepancy origins.* This graph summarizes our structural model to study the origins of the data discrepancies documented in Section 3. We provide details of the model structure in Section 5.1.



(a) Case 1. (b) Case 2. (c) Case 3.

Figure 10: *Simulated close price divergence scores.* These figures show divergence scores for the close price simulated from our structural model under different parametric assumptions. Each figure includes 10,000 independently simulated divergence scores. We use the same model as described in Figure 9 and Section 5.1. In Case 1, the structural parameters are $N_E = 50$, $p_{C,p} = 0.01$, $p_{E,p} = 0.05$, $\sigma_p = 0.03$, and $\Lambda_p = 0$. Case 2: $N_E = 100$, $p_{C,p} = 0.025$, $p_{E,p} = 0.01$, $\sigma_p = 0.05$, and $\Lambda_p = 0.3$. Case 3: $N_E = 30$, $p_{C,p} = 0.001$, $p_{E,p} = 0.005$, $\sigma_p = 0.01$, and $\Lambda_p = 0.8$.

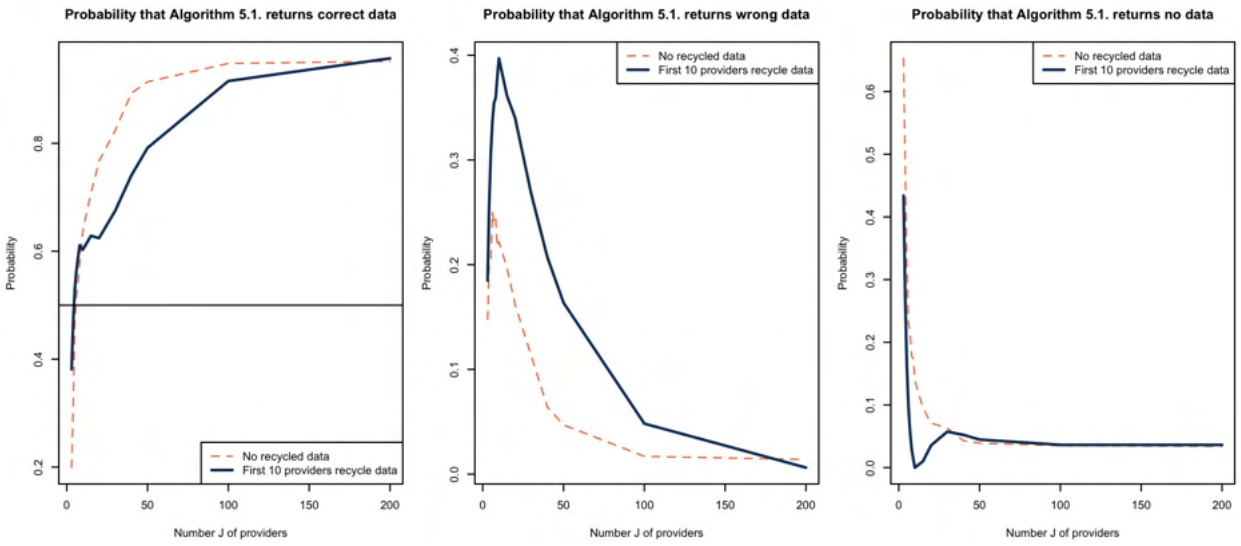
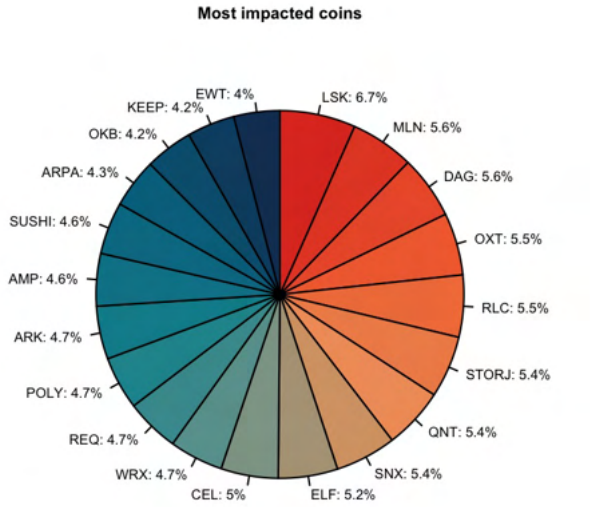
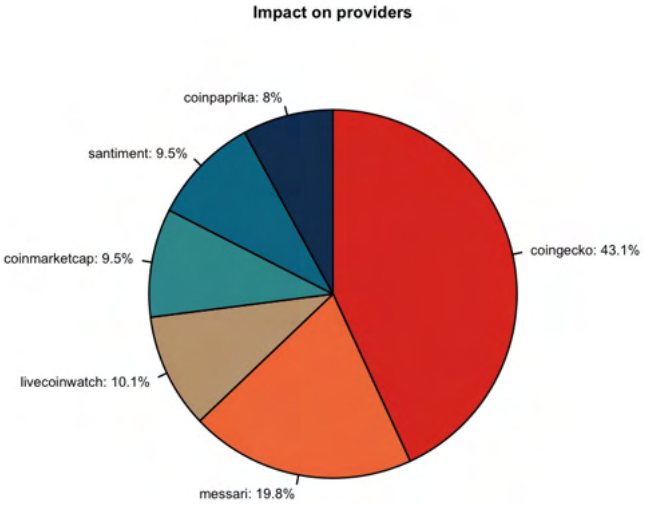


Figure 11: *Simulation case study in the presence of recycled data.* These plots show outcome probabilities in the simulation study for the convergence of Algorithm 6.1 under the assumption that the first 10 providers all return the same data. We consider only the parametric Scenario 3 of Table 8.



(a) Most impacted coins.



(b) Impacted providers.

Figure 12: *Impact of our aggregation methodology.* We measure impact as the proportion of daily data entries for a coin or provider that are removed by a step of our aggregation approach relative to the total number of data instances removed. In Panel (a), we restrict ourselves to the 20 most impacted coins by our algorithm.

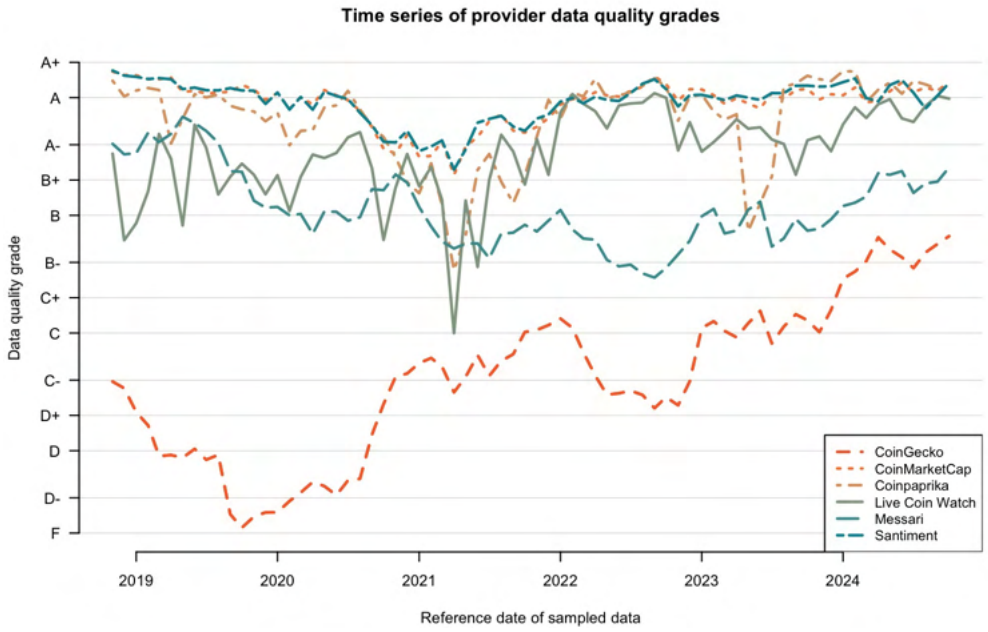


Figure 13: *Time series of data quality grades.* This graph shows the monthly time series of the quality grades assigned to the different providers based on the data provided in each month. To compute these grades, we first measure the monthly discard rate as the proportion of a provider's monthly data that is discarded by either Steps 2 or 3 of Algorithm 6.1. Then, we use the grading rules of Table 11 to determine the monthly grade.

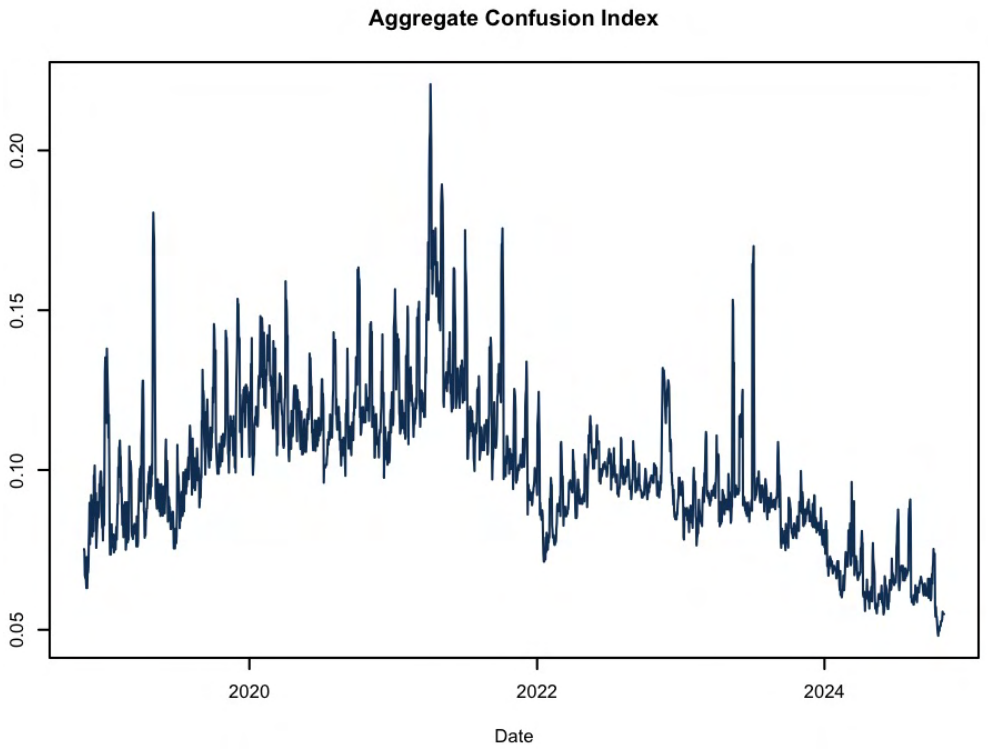


Figure 14: *Aggregate Confusion Index.* We compute the Aggregate Confusion Index at the daily frequency as the ratio of the number of daily data points that Algorithm 6.1 discards in Steps 2 or 3 across providers over the total number of data points available across providers on that day.