The impact of ad fraud and false identifiers on marketing strategy and return on advertising spend (ROAS)

BDEX

New York City | Miami

TABLE OF CONTENTS

SUMMARY

- How widespread is the problem of bad data? Utilizing data from a broad sample of top providers, BDEX analyzed over 1 billion device identifiers and found 25% of them to be invalid.

- Errors were found in 20% of MAIDs and 2% of email MD5s sold in the U.S. data market. Additionally, findings showed 21% of email MD5s were linked to more than 10 separate MAIDs, indicating invalid and/or fraudulent identifiers.

- This erroneous consumer data is caused by factors such as intentional fraud and user error, and exacerbated by the churning of bad data through resellers.

- Bad data is usually mixed in with other data used by marketers and advertisers, clouding their ability to make sound strategic and tactical decisions for campaigns.

- Because data sold and used in the U.S. market is often fraudulent, in 2019 the Association of National Advertisers estimated losses on return on ad spend (ROAS) to be $5.4 billion globally.

- Utilizing industry rates of useable consumer data and average ROAS, BDEX found marketers can increase ROAS by as much as 43% by taking steps to eliminate bad data.

- Companies should take steps to evaluate identifier data for signs of fraud such as excessive MD5s on one IP address, use of proxy services and tell-tale spoofed device characteristics.

INTRODUCTION

When bad data enters the digital ecosystem, the impact to marketers is two-fold: not only is money wasted targeting non-existent consumers, but also, forward-looking data-based strategy decisions are muddied. And in a landscape where 76% of marketing leaders use data to drive key decisions and allocate an average of 25% of budgets to digital ad spends, now more than ever marketers need to ensure their consumer data reflects real, genuine targets.

Fraud is a big part of the problem. Throughout the years, the industry has somewhat thwarted passive fraudsters by developing fraud detection filters; however, as digital marketing has become more valuable, ad-fraud simultaneously has become more lucrative, spurring an evolution of creative techniques. Some of the most incessant forms of ad-fraud, such as bot traffic, domain spoofing, and click-farms, have forced industry leaders to throw significant resources into solving these problems.

To fight ad fraud, companies should align with consumer data providers who prioritize data quality. These firms should have robust big data and automation tools to scour identifiers for bad data characteristics and strive to acquire impartial, quality-scored third-party data at all times. Only by putting quality first can marketers make informed decisions and salvage ROAS — imperative in this increasingly competitive digital world.

PROBLEM STATEMENT

Challenges Created by Bad Identifier Data

In an economy driven increasingly by customer data, bad identifier data is a major concern. Here are ways this erroneous data impacts business:

- **Low Return on Ad Spend (ROAS):** Ads purchased and wasted on targeting bad identifiers reduces ROAS and, subsequentially, brand awareness.
- **Wasted Employee Time:** Analytics teams spend 46% of their time preparing data for analysis. Additionally, when ads are used on identifiers not attached to potential and target audiences, marketers' time is also wasted.
- **Poor Marketing Decisions:** Not having the proper data leads to poor planning and strategies as well as wasting crucial time and significant resources.

Where does erroneous device identifier data come from?

Erroneous device identifier data stems from multiple sources, but most frequently from:

User Error

Typos, incorrectly inputted information, duplicate entries and inconsistent formatting all contribute to erroneous information in the data ecosystem. While most of these errors are unintentional, they create large quantities of unusable data that is being sold through the data ecosystem.

Fraud

There are several motives behind ad fraud:

- To improve rankings for social media accounts by artificially increasing engagement
- To inflate website traffic in order to attract advertisers or potential buyers
- To deplete a competitor's ad budget and weaken their position in the market
- To generate ad revenue by boosting clicks or views of the ads on their own sites

Fraud that contributes to false identifier data includes (among others):

- **Spoofing:** A single device is spoofed to resemble different unique devices. Clicks or other actions performed by one device are counted as multiple devices.
- **Bots:** Using emulation software, fraudsters create thousands of fake devices with spoofed identifier information.

CONTINUED ON NEXT PAGE »»

PROBLEM STATEMENT

- **Geomasking:** The geographic location of an identifier is changed to disguise a less desirable location as a premier one. This has significant impact on ROAS for marketing campaigns that target an audience from a specific geographic location.
- **Click-Farms:** Operations where people are recruited to click on multiple ads in exchange for money, usually operating dozens of devices at a time. It is difficult for automated filters to recognize this traffic as erroneous because the traffic is itself not fake, but not provided by genuine, potential consumers.

Invalid Hashed Email (MD5)

Invalid hashed emails (MD5) can be the result of invalid email domains, such as:

- Domains that are not active or do not exist
- No mail exchanger record (MX record), so the data cannot receive email
- Invalid email prefixes, making email delivery impossible because there is no actual recipient

Invalid Mobile Advertising IDs (MAIDs)

- MAIDs that contain invalid formatting, are unintentionally hashed, are no longer active, or are linked to fraudulent activity

Unactionable Identifiers

Some data have identifiers that make them unactionable, such as:

- The data originated from a country different than the indicated country
- The ID traces back to a bot or click farm
- It has been associated with commercial or government IPs

Invalid Identity Linkages

Some identifiers are verified as valid but have been linked to other data incorrectly, or they are linked to other invalid identifiers, which makes them unreliable.

Compounding the Problem: Data Resellers

Companies often partner with secondary data resellers, who may partner with tertiary data resellers. If one of these downstream resellers has bad data, the entire data pool becomes corrupted.

STUDY RESULTS

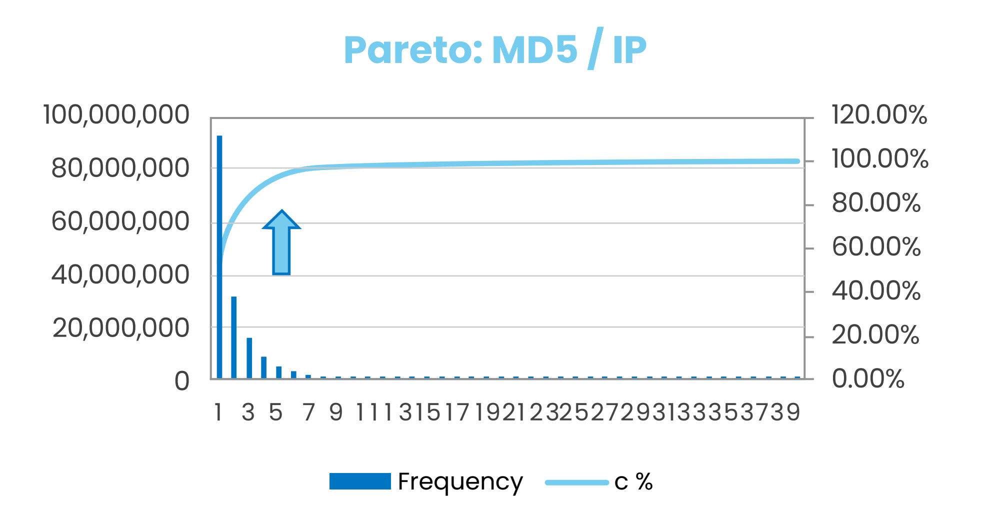
Study Methodology:

Utilizing in-house data from a broad swath of top providers, BDEX analyzed over 1 billion device identifiers sold in the U.S. market for identifier match pairs, and indicators of invalidity and fraud. All result findings are concluded using this sample, except for the analysis on impact of removal of bad consumer data on ROAS, which utilized industry rates of useable consumer data and average ROAS by Nielsen (2016).

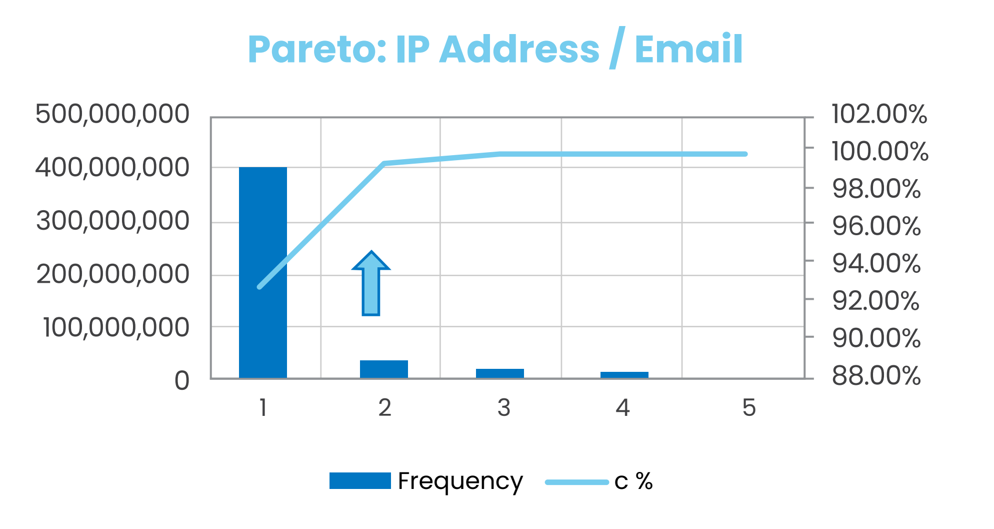
Overview:

BDEX analysts found on average **25% of device identifiers sold in the U.S. market may be invalid,** including:

- 20% of all MAIDs
- 2% of consumer IP addresses
- 2% of email MD5s, due to invalidity
- 21% of Email MD5s, due to linkage to more than 10 MAIDs



This Pareto chart signifies the frequency of IP address to Email. The arrow indicates 95% of Emails are associated to less than 2 IP Addresses

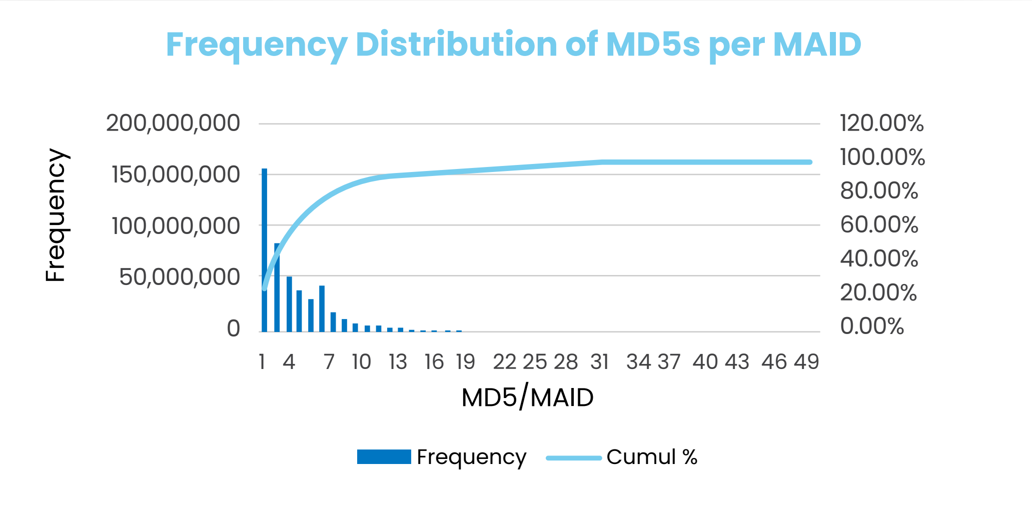


STUDY RESULTS

Frequency Distribution Tests

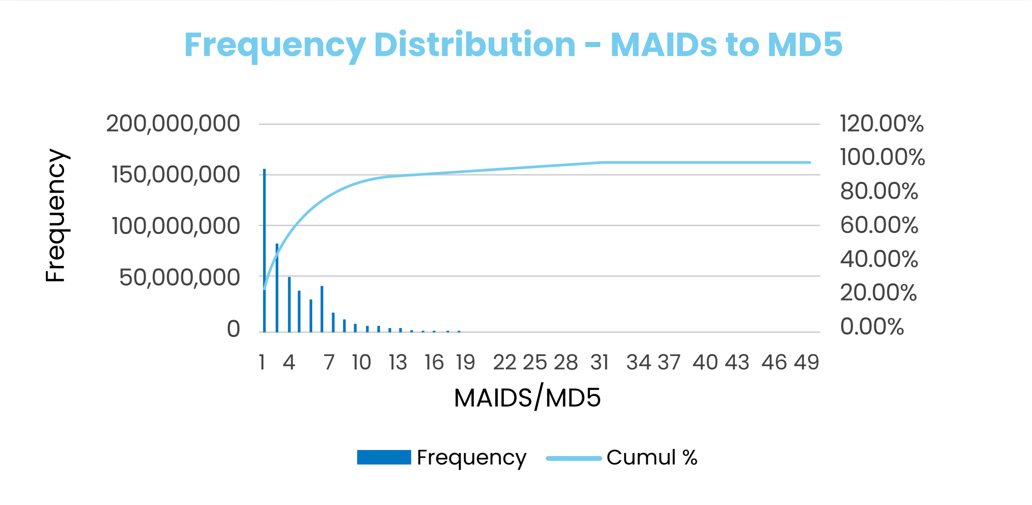
BDEX also performed a frequency distribution analysis of MAIDs per MD5s and MD5s per MAIDs. For both distribution tests, 95% of data points fell within a frequency of 20. In addition, BDEX found very extreme outliers, such as cases of 1000 MAIDs per one MD5. Such outliers can be attributed to fraudulent practices, but also to human error, with offenders such as example@example.com.

Common Frequency Distribution of MD5/MAID



BDEX analysts found that only 40% of MD5s in the study were linked to one MAID. The distribution of the sample fell significantly after a frequency of 6 MD5/MAIDs, with nearly 80% of MD5s attributed to ≤ 6 MAIDs. 95% of MD5s were attributed to ≤ 20 MAIDs.

Common Frequency Distribution of MAID/MD5

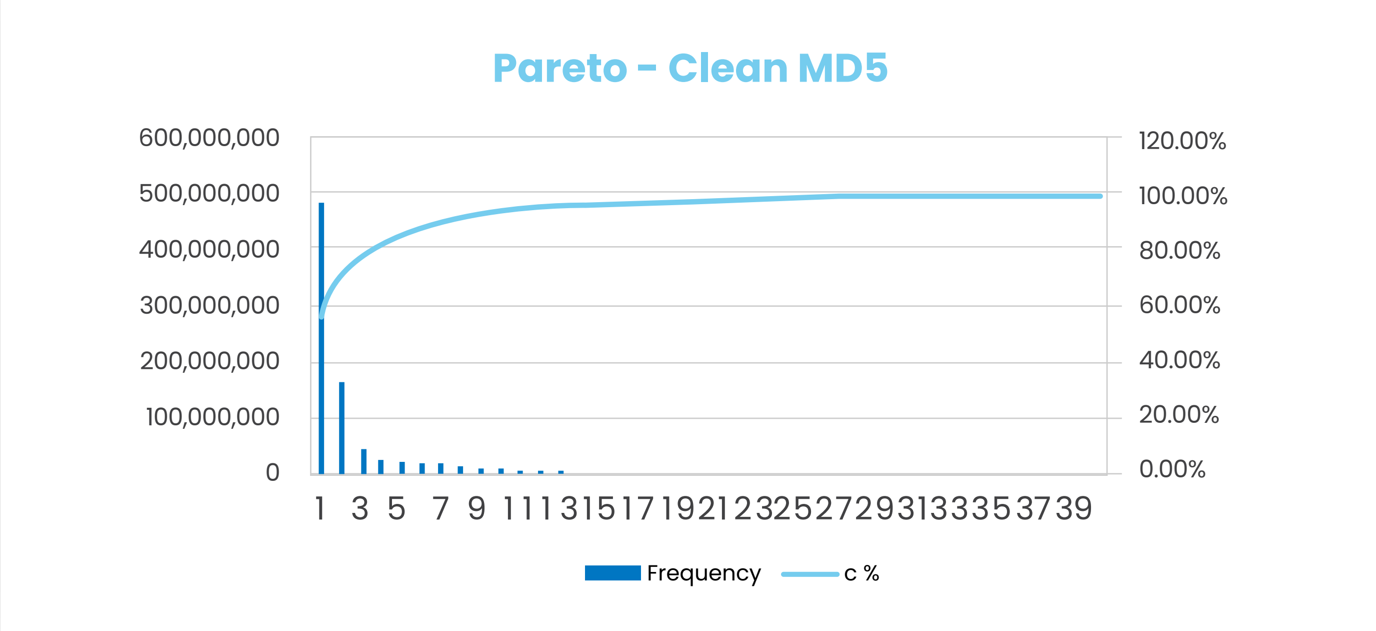


Distribution of MAID per MD5 frequencies followed the same pattern as MD5/MAID findings: more than 40% of MAIDs were attributed to one MD5 and 80% of MAIDs were attributed to ≤ 7 MD5s.

RECOMMENDATION

BDEX analysts found that by removing invalid MD5s, the frequency distribution now shows that 95% of MD5s are linked to 10 or less MAIDs.

Frequency Dist. of MD5/MAID – after removing invalid MD5s



In comparison to pre-removal frequency distribution of 95% of MD5/MAID having a distribution of f=20, 95% of MD5s had less than 10 MAIDs connected to the identifier (f=10), demonstrating a strong effect of invalid MD5s on the distribution and the positive impact of removing such erroneous data from sets.

Further, BDEX found, through utilizing industry rates of ROAS and usable consumer data by Nielsen (2016), **companies can increase ROAS by as much as 43%** by eliminating bad data. Therefore, companies should be vigilant at removing these invalid identifiers from their data to not only save their ROAS but also to maintain the accurate data they need to make informed decisions.

Eliminating Bad Data on ROAS Percent Increase

Audience	Bad Data	Good Data	Spend	Cost	Avg. ROAS %	ROAS ($)
1,000,000	30%	70%	$5	$3,500.00	287%	$ 10,045.00
1,000,000	0%	100%	$5	$5,000.00	287%	$ 14,350.00
					% Increase	**43%**

CONCLUSION

Companies cannot make proper marketing, sales, and strategic business decisions without consumer data that is reliable and accurate.

While this issue is challenging for firms to tackle, there are actions companies can take to decrease bad consumer data and avoid ad fraud, such as evaluating the following characteristics in their data sets:

- **IP addresses**: About 95% of all IP addresses have six or fewer email addresses associated with them. 95% of all email addresses have four or fewer IP addresses associated with them. Companies should be wary of IP addresses associated more than these numbers.
- **Proxy server information:** While not a direct indicator of fraud, the use of proxy servers can be a red flag for fraudulent activity, and companies should consider this factor when evaluating the validity of an identifier.
- **Phone characteristics:** Fraudulent identifiers may have obvious tells, such as accounts liking thousands of pages, unnaturally quick traffic on web pages or excessive traffic for extended periods of time.

In order to get their data as close to accurate as possible, companies should also consider aligning with data providers who have been assessed across these metrics:

- **Verified Across Multiple Channels:** Data has been cross-referenced by multiple sources and linked accurately.
- **Filtered for Erroneous Data:** Fraudulent, outdated, and inaccurate data has been removed from the data set, assuring identifiers can be used for decision making and limiting loss of ROAS due to identifier data.
- **Real-Time Metric:** Data sources use SDKs, APIs, App Publishers, and Aggregators to provide the latest, up-to-the-minute customer data available.
- **Allows for Profile Building:** Connecting device identifier information to other forms of consumer data such as visits to various websites, social media platforms, geo-locations, and more provides comprehensive profiles that allow you to know when a specific customer is right for your goods and services.

In a world where consumer data has become our most valuable commodity, ensuring its accuracy must become our top priority.

ABOUT BDEX

Established in 2014, BDEX is the first ever Data Exchange Platform (DXP). Combining the functionality, data, and reach of a traditional data management platform in a true marketplace environment with the most powerful identity graph in the US, BDEX empowers B2C companies to use the power of data to understand consumer behaviors and intents helping them reach the right people at the right time.

Find out more at **www.bdex.com**.

REFERENCES

"2018-2019 Bot Baseline: Fraud in Digital Advertising." ANA. Accessed August 21, 2020. https://www.ana.net/miccontent/show/id/rr-2019-bot-baseline.

"56% Of Digital Ads Served Are Never Seen, Says Google." Ad Age, December 3, 2014. https://adage.com/article/digital/56-digital-ads-served-google/296062?_ptid=%7Bjcx%7DH4sIAAAAAAAAAI2RyW7CMBRF_8VrLHmIY-MdFAQFUagEoe3OJC_BUiYSh6FV_71JVCqx68aSfe55T7r-QsZGSCOvOq3eb5_T2wQNUGkSCCxcnjvCCCOYSEw9TBkWAguKqZCYEPVypXPrC7jQdLzBUvJIKB6DzwBUZIBRZoSMhwcZM8XidjBcS6gs5CH0o6dvezmWw8VcBv4DnV4hbJwt8j5GFRGpSYgN26WEsPoEcHZJ5iWFaLjJY8KL8MEfhX9yfSwuW8jK1DjYe8sFC9YfwWbHJW2No6nvDGIXNTBA7vfey-stGcl1MJ7TXRe_s8BU1uSui-RNmg5QaLLS2CSv7w9nW9ueozN-LNBTmHJMqcKz7ewUrCavOQ-X5VMR_KdAW3Y_4hOiqWwPTnyiWduPlor5OvaHQvsy1BHQQ5tuaqhGCeSulaKsq8i5FGkqhp7iQgr2_QMdxnVe_wEAAA.

"Poor Data Quality: Marketers Waste 21 Cents of Every Dollar Spent on Media." 09/10/2019. Accessed August 21, 2020. https://www.mediapost.com/publications/article/340443/poor-data-quality-marketers-waste-21-cents-of-eve.html.

"The New Ad Fraud Scheme That Is Costing Advertisers an Estimated $130 Million." Business Insider. Business Insider, July 27, 2020. https://www.businessinsider.com/business-insiders-biggest-advertising-and-media-stories-for-july-27-2020-7.

"4 Key Findings in the Annual Gartner CMO Spend Survey 2019-2020." Gartner. https://www.gartner.com/en/marketing/insights/articles/4-key-findings-in-the-annual-gartner-cmo-spend-survey-2019-2020.

"Key Findings from Gartner Marketing Analytics Survey 2018." Gartner. https://www.gartner.com/en/marketing/insights/articles/key-findings-from-gartner-marketing-analytics-survey-2018.

Gartner Inc. "Marketing Data and Analytics Survey 2018: Messy Data and Mismatched Resources Undermine Marketing Teams." Gartner. https://www.gartner.com/en/documents/3883171.

"Benchmarking Return on Ad Spend: Media Type and Brand Size Matter." Nielsen. Accessed August 24, 2020. www.nielsen.com/us/en/insights/article/2016/benchmarking-return-on-ad-spend-media-type-brand-size-matter/.

Duczeminski, Matt. Using ROAS to Measure the Effectiveness of Your Ad Campaigns. Accessed August 24, 2020. https://blog.salesandorders.com/roas.