BDEX WHITE PAPER

Invalid by Default

The State of Mobile Identity in 2026 November 2026 www.bdex.com

Executive Summary

Nearly 95% of the mobile advertising IDs (MAIDs) in the programmatic bid stream cannot be trusted to reach a real person. In October 2026, BDEX observed about 52 billion MAIDs, and only 2.8 billion cleared its quality filters. BDEX's original study found about 1 in 4 identifiers invalid. Today, roughly 18 in 19 bid-stream MAIDs fail.

- **Most of the bid stream is unaddressable.** More than 76% of MAIDs observed in October 2026 were temporary. They resolve to no persistent device or person.
- **Most of what remains is invalid.** Only 23.3% of the non-temporary MAIDs passed validation.
- **Malformed IDs still arrive by the tens of millions.** Just 0.155% of MAIDs were malformed, but that was about 79 million IDs in a single month.
- **Over-linking is now the norm.** More than 60% of non-temporary MAIDs were tied to too many emails, or shared an email with too many other MAIDs. Both are signatures of bot manipulation.
- **AI buying amplifies the damage.** Automated platforms learn from the signals they are fed. Invalid IDs now waste budget *and* train algorithms toward audiences that do not exist.
- **Bad IDs turn profitable campaigns into losses.** An average campaign returns 2.87x on a validated audience but only about 2.0x when 30% of its IDs are bad. At about 65% bad IDs, it returns less than it spends.
- **Validation, not matching, is the fix.** Every MAID must clear fraud screening, household validation and continuous re-checking before it enters the BDEX graph. In October 2026, about 1 in 19 did.

Introduction

Bad identity data costs marketers twice. Money is spent reaching people who do not exist, and the results of that spend are fed back into planning as if they were real. Every downstream decision, from audience selection to budget allocation, inherits the error.

The scale of the problem keeps growing. Statista put global ad fraud losses at $88 billion in 2023 and projects $172 billion by 2028 (NEXD, citing Statista). Estimates vary widely by method, but none of them are shrinking.

Three shifts since BDEX's original study have changed the shape of the problem:

1. **Opt-in tracking became the norm.** Apple's App Tracking Transparency (ATT) left only about 35% of iOS users who see the prompt opting in as of Q2 2025 (Adjust). The rest generate identifiers that are temporary, zeroed or rotated. These IDs pass format checks and still point

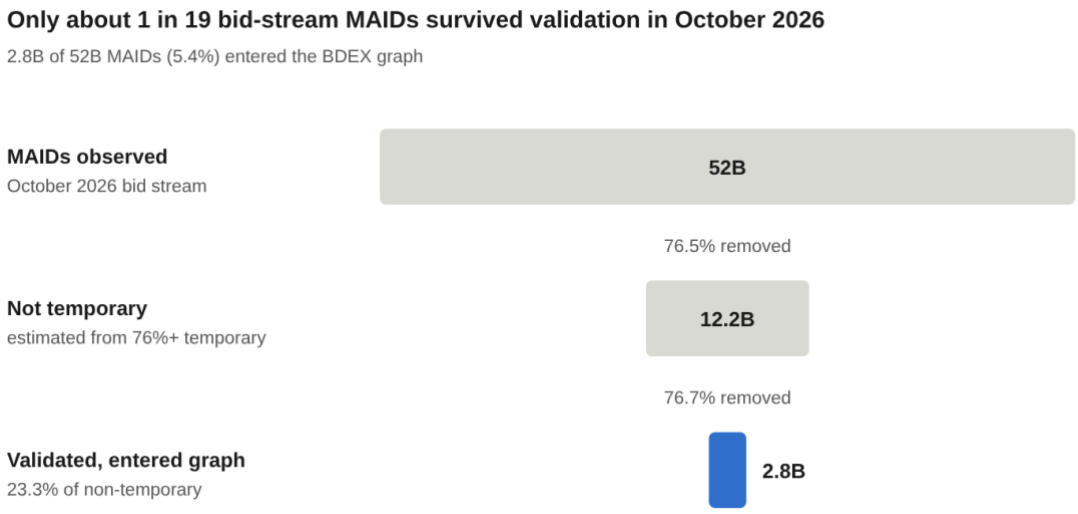
to no persistent consumer. By October 2026, more than 76% of MAIDs in the bid stream were temporary.

2. **Buying became automated.** Spider Labs estimates advertisers lost more than $32.6 billion to ad fraud in 2025. It found fraud risk in AI-optimized campaigns up to twice the average (Spider Labs via ACCESS Newswire). When an algorithm optimizes against fake signals, it learns to find more of them.
3. **Click farms moved into new formats.** The same report measured a 12.79% fraud rate in short-form video, mostly tied to organized click farms. Those farms run real devices, so their identifiers survive filters built to catch software bots.

The answer is the same as it was, only more urgent: buy identity data from providers who validate it continuously, and treat any identifier that cannot prove it belongs to a real person as noise.

October 2026: What We Found in the Bid Stream

Only 2.8 billion of the 52 billion MAIDs BDEX observed in October 2026 were good enough to enter its graph. BDEX watched more than 13 billion MAIDs pass through the programmatic bid stream each week and ran every one through the validation pipeline described later in this paper.



Source: BDEX bid-stream test, October 2026. Non-temporary count estimated from the 76% temporary share.

Two filters did most of the work. More than 76% of MAIDs were temporary, so no advertiser could ever reach the same device twice. Of the rest, only 23.3% passed validation.

Two more findings show how the failures happen:

- **Malformed IDs.** 0.155% of MAIDs were simply broken: wrong length, wrong characters, or hashed when they should not be. Small as a share, that was about 79 million IDs in one month.

- **Over-linking.** More than 60% of non-temporary MAIDs were over-linked: one MAID tied to too many emails, or one email tied to too many MAIDs. Real people do not look like this. Bots that recycle and share identifiers do.

Then and now

Measure	Original study (2017)	October 2026
What was examined	1B+ identifiers sold by top U.S. data providers	52B MAIDs in the programmatic bid stream
MAIDs that failed	20% invalid	About 95% temporary or invalid
Over-linking	21% of hashed emails linked to more than 10 MAIDs	More than 60% of non-temporary MAIDs over-linked

The two tests measured different pools with different methods, so the numbers are not a like-for-like trend. The direction is not in doubt: the share of mobile identifiers that cannot be trusted has grown from a minority to the overwhelming majority.

The Problem: Where False Identifiers Come From

False identifiers enter the ecosystem through error, through fraud and through the privacy layer itself. Once in, resellers spread them.

What bad identity data costs

- **Lower ROAS.** Impressions served to invalid IDs return nothing, and frequency caps fire on people who are not there.
- **Wasted analyst time.** Analytics teams already spend 46% of their time preparing data. Bad identifiers add cleanup that should never be needed.
- **Corrupted decisions and models.** Lookalikes, attribution and automated bidding all train on the data they receive. Fraud in the seed becomes fraud in the output.

Sources of false identifiers

Source	How it shows up	Why filters miss it
User error	Typos, placeholder emails (example@example.com), duplicates, inconsistent formatting	Valid format, invalid person
Spoofing	One device reports itself as many unique devices	Each ID looks new
Bots and emulators	Software creates thousands of fake devices with fabricated IDs	Behavior mimics humans
Click farms	People operate dozens of real devices for pay	The traffic is human, just not a customer
Geomasking	A device's location is altered to look like a premium market	Geo data is self-reported

Source	How it shows up	Why filters miss it
Temporary and rotating MAIDs	Opted-out or reset devices emit short-lived IDs; fraud schemes rotate IDs on purpose	Passes MAID format checks
Invalid hashed emails	Inactive domains, no MX record, impossible prefixes	A hash hides the bad address
Unactionable IDs	Wrong country, commercial or government IPs, data-center traffic	Valid ID, wrong audience
False linkages	Real IDs tied to the wrong person, or to other invalid IDs	Each ID passes alone; the link is the error

Why it compounds: resellers and co-ops

Most identity data passes through several hands before a marketer uses it. A secondary reseller buys from a tertiary one, and one bad source corrupts the whole pool.

Data co-ops face the same trap. When several members contribute the same bad linkage, the co-op counts it as confirmed. Agreement among sources is not validation if the sources share the same polluted inputs.

Why Deterministic and Probabilistic Aren't Enough

A link can be deterministic and still be fake. That is why BDEX describes its graph as validated rather than deterministic or probabilistic.

Approach	How links are made	Where it breaks
Probabilistic	Infers that devices belong together from shared IPs, behavior and device traits	Over-clusters shared networks: apartments, offices, libraries, campuses
Deterministic	Links IDs through a login, hashed email, account or loyalty ID	A bot can log in. A click farm can register an email. The link is certain; the person may not be real
Co-op	Treats agreement among contributing members as confirmation	Members often share the same polluted upstream sources
Validated (BDEX)	**Every ID and link must clear fraud screening, household validation and ongoing re-checks before release**	**Built to remove what the other three let through**

The distinction matters most where false positives are expensive: fraud prevention, suppression, measurement and high-value audience activation.

The Baseline: BDEX's Original Study

BDEX's original study, conducted in 2017, set the baseline: on average, 25% of device identifiers sold in the U.S. market may be invalid. BDEX reached this figure by analyzing more than 1 billion identifiers sourced from a broad set of top U.S. data providers. The analysis looked at identifier match pairs and at markers of invalidity and fraud.

Invalid identifiers by type

Identifier	Share flagged	Reason
Mobile advertising IDs (MAIDs)	20%	Invalid format, unintended hashing, inactive, or tied to fraud
Email MD5s	21%	Linked to more than 10 MAIDs
Email MD5s	2%	Invalid email (dead domain, no MX record, bad prefix)
Consumer IP addresses	2%	Invalid or non-consumer

What normal looks like

Real people have a small, stable footprint. Fraud shows up as identifiers that sit far outside it.

- **Email to IP:** 95% of emails were associated with 4 or fewer IP addresses, and 95% of IPs with 6 or fewer emails.
- **MD5 to MAID:** only 40% of MD5s linked to a single MAID; nearly 80% linked to 6 or fewer; 95% linked to 20 or fewer.
- **MAID to MD5:** the same pattern. More than 40% of MAIDs linked to one MD5, and 80% to 7 or fewer.
- **Extreme outliers:** some single MD5s were linked to 1,000 MAIDs. Causes include fraud and placeholder addresses such as example@example.com.

The effect of cleaning

Removing invalid MD5s tightened the distribution sharply. Before cleaning, 95% of MD5s linked to 20 or fewer MAIDs. After, 95% linked to 10 or fewer. A small share of bad emails had been distorting the shape of the entire dataset.

How BDEX Validates Identifiers

BDEX screens every identifier for fraud before it is linked to anyone, then keeps re-proving it for as long as it stays in the graph. The graph is built entirely in-house, not licensed or resold from third parties. That gives BDEX control over every step where bad data could enter.

1. Ingest raw signals, not pre-linked pairs

BDEX ingests billions of raw records per day from bid-stream data, email service providers, apps and SDKs. It deliberately takes unlinked signals rather than pre-built ID pairs. Pre-built pairs are where third-party fraud hides, so BDEX makes every link itself.

2. Screen for fraud before linking

Incoming MAIDs, IPs and hashed emails pass through proprietary fraud filters before any linking happens. Only IDs that clear screening move forward. The filters look for:

- **Bot and emulator patterns:** device traits and activity that no real handset produces.

- **Click-farm clustering:** many devices sharing infrastructure, timing or behavior in ways households do not.
- **Abnormal fan-out:** one email tied to dozens of MAIDs, or one IP tied to an implausible number of emails. The study thresholds above define normal.
- **Temporary and rotating MAIDs:** IDs with short lifespans, frequent resets or no behavioral history, including those produced by ATT and Android tracking opt-outs.
- **Non-consumer traffic:** data-center, proxy, commercial and government IPs, and IDs whose origin contradicts their stated country.

3. Link to a real household

Surviving IDs are linked to a consumer and household: name, postal address, email, MAIDs, CTV IDs and household IP. BDEX treats the household as the meaningful unit. A stray individual IP says little; a stable home IP says a lot.

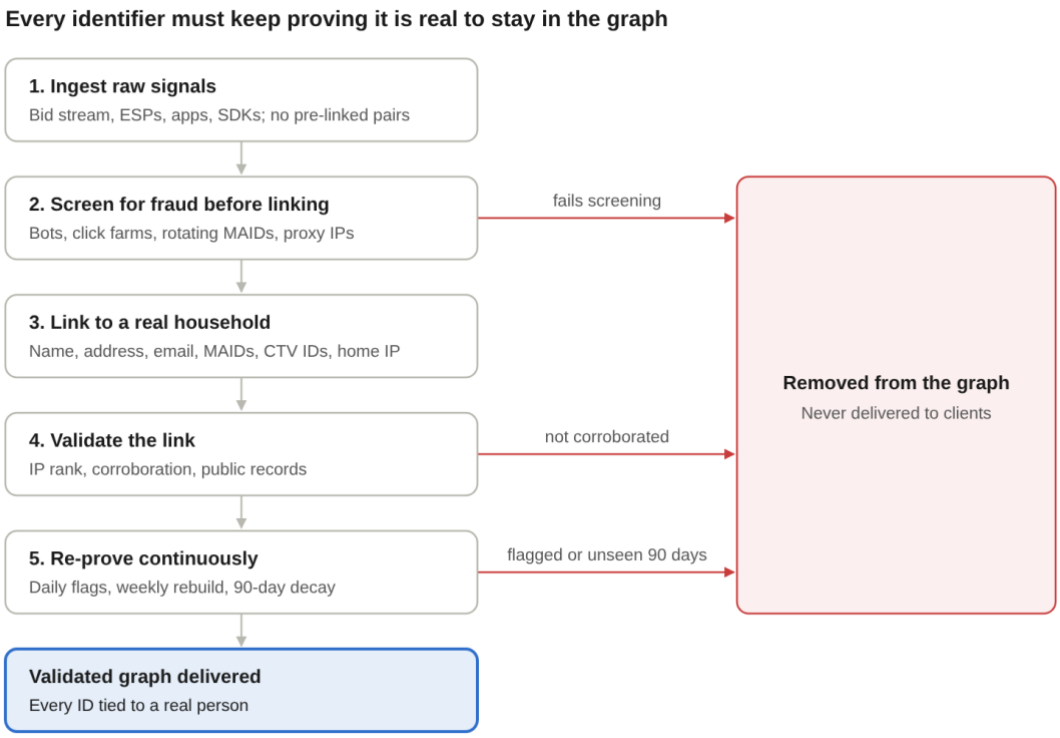
4. Validate the link

- **Household IP rank.** BDEX scores how confident it is that an IP is a consumer's real home. It compares where the IP is observed against the consumer's known address.
- **Move versus visit.** Repeated sightings at a different IP trigger re-evaluation, not an instant change. A weekend at a friend's house is not a move.
- **Cross-source corroboration.** A link must be supported by independent sources, not echoes of the same upstream feed.
- **Public-record checks.** PII and email data are cross-checked against public databases.
- **Temporal consistency.** Attributes that change faster than real life allows are flagged.

5. Keep re-proving it

- **Daily:** incoming feeds are analyzed and fraud flags and signals are updated.
- **Weekly:** the full graph is rebuilt from the week's signals, adding new validations and stripping IDs now judged fraudulent or unresolved.
- **90-day decay:** any ID not observed in 90 days is dropped.

This is what lets BDEX stand behind every ID in its graph as linked to a real person. The graph clients use is clean as of this week, not as of the last one-time cleanse. At U.S. scale, that covers 200M+ consumers and roughly 100M households, more than 80M of them mapped to a household IP, informed by 2 to 3 trillion signals a month. In October 2026, that standard admitted 2.8 billion of the 52 billion MAIDs BDEX observed.



An identifier can be removed at three points: before linking, at validation, or on any later re-check.

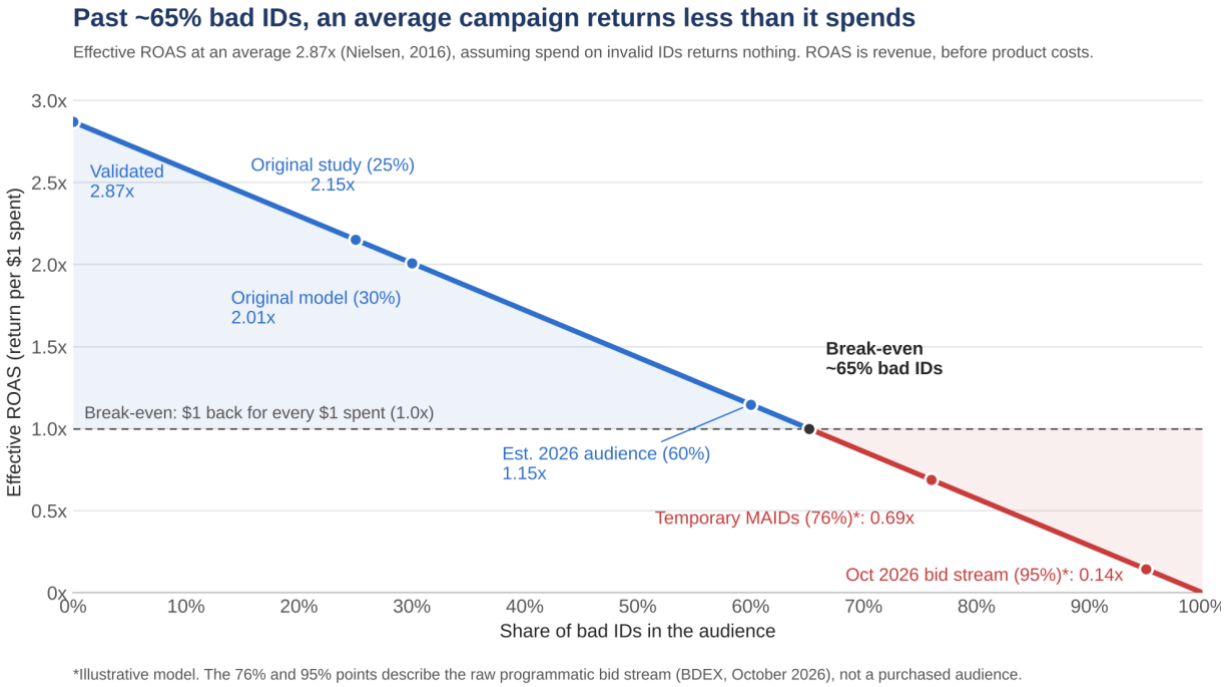
The ROAS Impact

At about 65% bad IDs, an average campaign returns less than it spends. In October 2026, 95% of bid-stream MAIDs failed validation. Even at a modest 30% bad-ID rate, validating the audience raises return by 43%.

The model applies the average ROAS of 2.87x found by Nielsen (2016) to a $100,000 campaign bought at a $5 CPM. That buys 20 million impressions, enough to reach about 4 million people five times each. Spend that lands on an invalid ID returns nothing. Spend that lands on a real person returns the average.

Share of bad IDs	What it represents	Effective ROAS	Return on $100K	Lift from validating
0%	**BDEX-validated audience**	**2.87x**	**$287,000**	**Baseline**
10%	A cleaner-than-average file	2.58x	$258,300	+11%
25%	BDEX original study (2017)	2.15x	$215,250	+33%
30%	Rate used in the original paper's model	2.01x	$200,900	+43%
60%	Estimated 2026 purchased audience	1.15x	$114,800	+150%
~65%	**Break-even**	**1.00x**	**$100,000**	**+187%**
76%*	Temporary MAIDs, Oct 2026 bid stream	0.69x	$68,880	+317%
95%*	MAIDs failing validation, Oct 2026 bid stream	0.14x	$14,350	+1,900%

The 76% and 95% rows describe the raw programmatic bid stream, not a purchased audience. Audiences are filtered before they are sold, so these rows show the upper bound of the problem rather than a typical campaign.



Where 30% comes from. The study itself averaged 25%. The original paper's ROAS model used 30%, and this paper keeps it as a conservative anchor. The 60% row estimates a purchased audience in 2026: well above the original rate, well below the raw bid stream.

The lift comes from the bad share, not the benchmark. Nielsen's 2.87x sets the dollar amounts, but the lift from validating depends only on how much spend is wasted. It is the same at any average ROAS:

$$\text{Lift} = 1 \div (1 - \text{share of bad IDs}) - 1$$

Revenue is not profit. ROAS counts revenue, so once product costs come out, a campaign is underwater well before 65% bad IDs.

The model counts only wasted impressions. It does not count the second cost: models, lookalikes and automated bidders trained on fake signals, which then find more fake signals.

Recommendations

Audit your own identity data against the norms real people produce, then hold data providers to a validation standard rather than a match-rate standard.

Red flags to check in your data

- **IP fan-out:** about 95% of IPs have 6 or fewer emails; about 95% of emails have 4 or fewer IPs. Treat anything beyond that as suspect.
- **MAID fan-out:** an MD5 linked to more than 10 MAIDs is suspect; hundreds or thousands is almost certainly fraud or a placeholder address.
- **Short-lived MAIDs:** IDs with no history, frequent resets or a lifespan of days rather than months.

- **Proxy and data-center IPs:** not proof of fraud on their own, but a reason to weigh the identifier lower.
- **Implausible behavior:** thousands of page likes, unnaturally fast page traffic, or activity sustained around the clock.
- **Stale IDs:** identifiers not seen in 90 days or more.

Questions to ask any identity data provider

1. Is your graph built in-house, or licensed and resold from other sources?
2. Do you ingest raw signals, or buy pre-linked ID pairs?
3. Is fraud screening applied before linking, or only to finished output?
4. How do you detect and remove temporary and rotating MAIDs?
5. How do you validate that an IP is a household, not an office, campus or proxy?
6. How often is the full graph rebuilt, and when are unseen IDs dropped?
7. Will you guarantee that each ID you deliver links to a real person?

See What Is in Your Data

BDEX will score a sample of your audience file for invalid IDs at no charge. The BDEX Fraud Score flags temporary, malformed and over-linked identifiers and shows how much of your file would clear BDEX validation. To request a scan, contact BDEX at www.bdex.com.

About BDEX

BDEX is an identity resolution and data infrastructure company built around one question: is this a real person? Founded in 2014, BDEX has focused on validated identity data since 2017.

Its identity graph is built entirely in-house and covers more than 200 million U.S. consumers and roughly 100 million households, more than 80 million of them mapped to a household IP. It links MAIDs, hashed emails, household IPs, CTV IDs and postal data, and is delivered through real-time APIs, batch files and an MCP server for AI agents. Enterprise data, identity, adtech/martech platforms and measurement companies rely on BDEX data. Learn more at bdex.com.

References

- Adjust, ATT opt-in rates: The latest benchmarks by app category and country, July 2025.
- Spider Labs, As AI Ad Buying Expands, Global Fraud Losses Hit $32.6 Billion, ACCESS Newswire, April 2026.
- NEXD, Ad Fraud Statistics: Current Costs, Rates, and Industry Data, citing Statista, updated November 2025.
- Nielsen, Benchmarking Return on Ad Spend: Media Type and Brand Size Matter, 2016.
- Gartner, Marketing Data and Analytics Survey 2018.
- ANA, 2018–2019 Bot Baseline: Fraud in Digital Advertising.