From Slang to Standards: Consensus-Driven Airdrop Hunter Definition as a Baseline for Cryptocurrency Ecosystem Security and Governance

Chunyang Li
University of Washington
Tacoma, Washington, USA
chunyang0409@outlook.com

Hongzhou Chen
CKB Eco Fund
Shenzhen, China
hongzhouchen1@link.cuhk.edu.cn

Wei Cai*
Tacoma School of Engineering and Technology
University of Washington
Tacoma, Washington, USA
weicaics@uw.edu

Abstract

Cryptocurrency airdrops power the growth and governance of the cryptocurrency ecosystem, yet attract airdrop hunters, who coordinate wallets, script interactions, and cash out quickly, distorting metrics and fairness. Prior detection strands (heuristics/clustering, light-supervised community partitioning, and graph learning) face three fundamentals: inconsistent definitions, weak explainability, and poor cross-context generalization. We distill expert knowledge into a computable, interpretable baseline: open/axial coding of expert narratives followed by two Delphi rounds to (1) formalize a consensus, operational definition with six contrasts to regular users; (2) derive 15 measurable indicators spanning operations and fund-flow, tempered by human-ness counter-evidence; and (3) report thresholds as reference distributions (medians, quartiles). The baseline supplies shared semantics and computation for labeling/evaluation, yields inspectable why-flagged rationales for audit and governance, and offers context-aware guidance across chains, campaign designs, and market phases, thereby strengthening on-chain security while informing the design of socio-technical systems perceived as fair, trustworthy, and resistant to strategic misuse.

CCS Concepts

• **Security and privacy → Security services**; • **Human-centered computing → Empirical studies in HCI**; • **Social and professional topics → Computing occupations**.

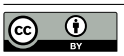
Keywords

Airdrop Hunters, Cryptocurrency, Decentralized Governance, On-Chain Behavior, Explainable Detection, Fairness, Community Trust

ACM Reference Format:

Chunyang Li, Hongzhou Chen, and Wei Cai. 2026. From Slang to Standards: Consensus-Driven Airdrop Hunter Definition as a Baseline for Cryptocurrency Ecosystem Security and Governance. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3790777

*Wei Cai is the corresponding author (weicaics@uw.edu).



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790777

1 Introduction

Decentralized Finance, built on blockchains and smart contracts, is reshaping trading, lending, derivatives, and governance with properties of programmability, permissionlessness, and composability [36, 44, 47]. Within this landscape, cryptocurrency airdrops have become a core mechanism for cold-start growth and the distribution of governance rights: projects allocate tokens to early or prospective users to catalyze adoption and embed participation incentives [2, 3]. Strong incentives, however, create strategic externalities: once eligibility rules and scoring recipes can be learned and replicated, profit-seeking participants may rely on multi-address coordination, scripted interactions, and rapid liquidation—diluting genuine contributions and inducing sell pressure [16, 29]. Following Fan et al. [12], we refer to these strategically profit-seeking participants as **airdrop hunters**. Yet despite its frequent use in industry discourse, the term "airdrop hunter" has remained a community slang rather than a scientific concept, with no standardized or reproducible definition. As a result, labeling in prior studies has been fundamentally ad hoc: what one project flags as a hunter may be considered legitimate elsewhere, leaving ground truth uncertain.

Empirically, and regardless of the definitional ambiguities noted above, airdrop hunters have evolved from sporadic actors into a structural force shaping program outcomes. Their behaviors include batch account creation, short holding with quick sell-offs, fund consolidation, and address-loop arbitrage; hunters can inflate superficial activity while increasing Sybil risk, raising distribution costs, and eroding governance fairness [2, 12, 46]. Accurate identification and management of such behavior is therefore closely tied to on-chain security, market integrity, and transparent governance. For ordinary users, these strategic externalities manifest not only as diluted rewards but also as frustration, loss of motivation, and erosion of trust in the fairness of the process. Taken together, airdrop hunting is not only a technical or economic security concern, but a socio-technical problem that emerges at the intersection of system rules, incentive design, and user experience.

Prior detection efforts follow three main technical strands. (i) Heuristics and unsupervised clustering extract patterns from interaction sequences and transfer graphs [26]; (ii) community partitioning with light supervision leverages graph structure and handcrafted features [34]; and (iii) graph learning and multimodal fusion perform supervised detection on transaction graphs and asset metadata [25, 38, 48]. Despite empirical promise, the field faces three foundational gaps: **(a) definitional inconsistency**. "airdrop

hunter" lacks a unified, operational definition. Existing studies therefore rely on project-specific or heuristic labeling, making accuracy claims difficult to interpret and undermining comparability across datasets, methods, and campaigns; **(b) missing context-aware thresholding**. There is no shared, reproducible basis for what constitutes "highly suspicious" values under different chains, phases, or campaign designs; and **(c) limited explainability**. Black-box models provide weak rationales for governance and compliance stakeholders [2, 29]. More broadly, the hunter phenomenon exemplifies a recurring challenge in socio-technical systems: how to distinguish authentic, organic participation from strategic, profit-driven manipulation, which also resonates with prior CHI work on fairness, abuse prevention, and trust in online communities [8, 22, 27].

To address these challenges, we aim to distill front-line expert knowledge into a computable and interpretable framework that can serve as a reproducible baseline for detection and audit. In particular, we ask **How can we transform the community slang of "airdrop hunter" into a consensus-driven, computable, and explainable framework that spans definition, behavioral indicators, and thresholds, and strengthens on-chain security and governance fairness?** We organize this overarching question into three sub-questions: **RQ1 (Definition),** How do domain experts define "airdrop hunter", and along which systematic dimensions do hunters differ from non-hunter users? **RQ2 (Behavior),** Which on-chain behaviors typify hunters, and how can they be operationalized into *computable* items? **RQ3 (Thresholds),** How should thresholds for these items be *reported* so that future work can replicate and align without relying on a one-size-fits-all cutoff?

We follow a "qualitative generation → expert consensus → formalization" pipeline. First, we elicit expert narratives via open-ended questionnaires and conduct open/axial coding to shape an initial structure spanning definition, behaviors, and contrasts [5, 40]. Second, we translate the codes into *item-level* indicators and computation recipes (15 items in total: 11 hunter-behavior indicators and 4 human-ness counter-evidence indicators), then run two Delphi rounds for importance and threshold elicitation, reporting medians, quartiles, and dispersion statistics [6, 18, 20, 31].

We provide a unified, operational definition of "airdrop hunter," and articulate six systematic contrasts with non-hunter users: motivation, temporal cadence, organization/automation, breadth/depth of behaviors, fund-flow structure, and human-ness cues. At the item level, two *diagnostic axes* emerge from importance consensus across two Delphi rounds: on the *operations* side, *Automated Scripted Trading* (BT), *Batch New-Wallet Activation* (BW), and *High-Frequency Trading arbitrage* (HF); on the *fund-flow* side, *Rapid Consolidation of Airdrop Funds* (RF) and *Multi-Address circular Arbitrage* (MA). A human-ness counter-evidence layer—*Long-Term Natural Activity* (LA), *High-Value Asset Diversity* (AD), *Clear DID Identity Binding* (DID), and *Natural Trading Behavior* (NT)—is jointly recognized as important for reducing false positives. Remaining items either lack consensus or are context-sensitive; we therefore position them as auxiliary evidence or review triggers rather than isolated signals.

This work contributes on three levels. First, **we resolve the field's most fundamental obstacle: the absence of a standardized definition.** By formalizing the community slang of "airdrop hunter" into a reproducible, expert-consensus definition, we provide the first stable ground truth on which labeling, evaluation, and detection can be consistently built. Second, **we deliver an engineering contribution: a computable joint-evidence framework of on-chain indicators (operations × fund-flow tempered by human-ness)**, with Delphi-validated importance and threshold reference distributions, forming a deployable baseline for labeling, detection, and review. Third, at a broader level, **we compress dispersed expert knowledge into a public baseline that improves comparability, strengthens auditability, and supplies context-aware guidance for governance and security.**

Taken together, these contributions connect methodological rigor with the practical need for airdrop detection that is explainable, fair, and adaptable across contexts, while also speaking to HCI's interest in socio-technical systems that are experienced by users as transparent and trustworthy. More concretely, they improve procedural fairness for legitimate users by reducing misclassification, give the "airdrop hunter" category clearer and more publicly understandable boundaries, and support community-level participation by making the underlying rules more legible and open to scrutiny. In doing so, we show how a once-contested community slang term can be formalized into a reproducible, consensus-based construct, linking technical detection with human-centered concerns for fairness, trust, and meaningful participation.

2 Related Work

The rise of decentralized finance has spurred the widespread adoption of airdrop mechanisms to rapidly expand community user bases and enhance project influence. However, while airdrops incentivize community participation, they have also triggered a series of user arbitrage and ecosystem fairness issues, most notably the emergence of "airdrop hunters," an arbitrage-oriented group whose activities can significantly undermine fairness and community health. Accordingly, this section first reviews scholarly work on airdrop mechanisms and user behavioral characteristics, then examines the state-of-the-art in detecting airdrop hunters, highlights the limitations of existing approaches, and lays the groundwork for the subsequent introduction of expert-consensus methodology and on-chain user-profiling techniques.

2.1 Airdrop Design and User Behavior Analysis

Airdrops have become a key incentive instrument for the cold-start phase of cryptocurrency and DeFi projects. Allen et al. [3] argue that distributing tokens for free can rapidly aggregate early users, spur community growth and network effects, and lay the groundwork for decentralized governance and ecosystem vitality, while a study also finds that airdrops significantly expand the user base of decentralized exchanges and boost early-stage activity [28].

Yet severe arbitrage and airdrop-farming behaviors have followed. Based on large-scale on-chain empirical analysis, a study shows that in several well-known airdrops, over 60% of the distributed tokens were liquidated within 24 hours by profit-seeking users, diverting incentives originally intended for community building and user retention toward short-term arbitrageurs [29]. Using on-chain data, Fan et al. [12] further reveal a clear bifurcation between *altruistic users* (focused on long-term ecosystem value) and *profit-oriented users* (focused on arbitrage and quick cash-out), with

the latter often employing multiple accounts and automation scripts to batch-claim and dump airdropped assets. These behaviors dilute the long-term effectiveness of airdrop incentives and raise broad concerns about fairness and project sustainability.

The large-scale impact of arbitrage has forced projects to continually refine and upgrade airdrop designs. From an evolutionary-economics perspective, Allen [2] notes that projects have adopted more complex rules, including multi-round distributions, task gating, tiered rewards, and contract lockups, to raise the cost for profit-driven participants and protect loyal users. Yaish and Livshits [46] go further by explicitly incorporating "airdrop farmers (a concept which is similar to airdrop hunter)" into the incentive framework via a tiered-reward mechanism, aiming both to stimulate genuine user activity and to harness the short-term growth effects generated by arbitrage-oriented participants. Extending this line, Halaburda et al. [16] argue from a platform-design standpoint that properly aligned incentives can leverage airdrop farmers to attract real users and increase platform revenue.

In sum, while airdrops greatly facilitate cold starts and network expansion, the arbitrage activities of profit-driven users (i.e., *airdrop hunters*) impose substantial financial and community costs. How to effectively identify and manage airdrop hunters and to steer incentive mechanisms toward healthier evolution has become a core challenge for both academia and industry. The next section examines current detection methodologies and their limitations.

2.2 Airdrop Hunter Detection: Current Research Status and Challenges

Existing detection approaches for airdrop hunters mainly include heuristic rules, graph clustering and community detection, and machine learning. Yet a core obstacle underlying all these efforts is that "airdrop hunter" has never had a standardized definition, remaining an industry slang rather than a scientific construct.

Representative of the heuristic line, Liu and Zhu [26] combines on-chain behavioral sequence clustering with handcrafted rules to identify suspected Sybil groups in the Hop Protocol airdrop. While simple and effective, such approaches are constrained by specific heuristic assumptions, struggle with evolving strategies, and lack unified data-labeling standards, which limits their generalization.

To overcome the limitations of heuristics, researchers have developed more generalizable graph-based methods. *ARTEMIX* by Qin et al. [34] integrates Louvain community detection with feature engineering, using heuristic categorizations of transactional and temporal patterns to flag airdrop-hunter accounts. Trusta Labs' *TrustScan* framework (2023) adopts a two-stage graph clustering pipeline with manual threshold tuning and has been applied to several real-world airdrops. However, these methods still depend on particular labeling protocols and heuristic assumptions (e.g., star- or chain-shaped fund flows); once attackers adopt more covert or novel strategies, performance can degrade substantially.

In parallel, supervised machine-learning approaches have advanced. Zhou et al. [48] propose *ARTEMIS*, which leverages graph neural networks together with multimodal NFT metadata (images and descriptions) to achieve strong performance in detecting airdrop hunters in NFT markets. Moreover, Liu et al. [25] present a subgraph-based feature propagation method coupled with a LightGBM classifier, demonstrating efficiency and accuracy on larger-scale datasets. Nevertheless, these supervised methods commonly face a "black-box" challenge, making decisions hard to interpret for communities and project teams, and their strong results often rely on project-specific labels and feature engineering, hindering cross-context generalization. An even more fundamental issue is that such evaluations rest on project-specific or ad-hoc definitions of what counts as an "airdrop hunter." Without a shared operational definition, each dataset effectively encodes its own labeling logic, so results cannot be meaningfully compared across studies. Different projects, and even different studies, may treat similar behaviors in inconsistent ways—an address flagged as a hunter in one context might be considered legitimate in another. As a result, claims of accuracy or effectiveness are difficult to interpret, since there is no standardized baseline against which results can be meaningfully compared. This lack of a shared, operational definition is precisely what motivates our RQ1: to establish, through expert consensus, a reproducible definition of "airdrop hunter" that can serve as common ground for subsequent detection and evaluation.

Overall, current detection work faces several challenges:

- Inconsistent labeling and lack of consensus: the very term "airdrop hunter" remains undefined beyond industry slang, so annotation standards vary widely and are often project-specific or heuristic, complicating fair and unified evaluation.
- Limited generalization: many methods are tailored to specific projects and airdrop rules, making them brittle to new campaign designs or unknown attack strategies.
- Interpretability gap: especially for ML and deep-learning models, the decision process is not readily understandable, reducing trust and hindering deployment.

Although existing methods have made progress in identifying airdrop-hunter behavior, these issues limit their practical utility and governance effectiveness. There is an urgent need for a more systematic and transparent expert-consensus methodology to address label inconsistency and generalization difficulties, combined with on-chain user profiling to enable deeper analysis and precise recognition of airdrop-hunter behavior patterns.

3 Methodology

Prior detection studies of airdrop hunters have been largely data driven and bottom-up, focusing on clustering interaction traces, transfer graphs, or training supervised classifiers on labeled datasets [25, 26, 34, 48]. These methods surface statistical regularities in large-scale traces, but they leave unresolved a definitional layer: "airdrop hunter" is a socially constructed category shaped by community discourse, project teams' intentions, and normative judgments about fairness. Bottom-up clustering or supervision can group similar behaviors, yet cannot, on their own, determine which behaviors ought to count as diagnostic or where to draw context dependent suspicion thresholds across chains and campaign designs, thereby necessitating a complementary top-down, expert-consensus view.

To capture this missing layer, we adopt an expert-consensus methodology. We first deployed open-ended questionnaires to elicit expert narratives, then applied open and axial coding to distill recurring behavioral dimensions. Building on these qualitative results,

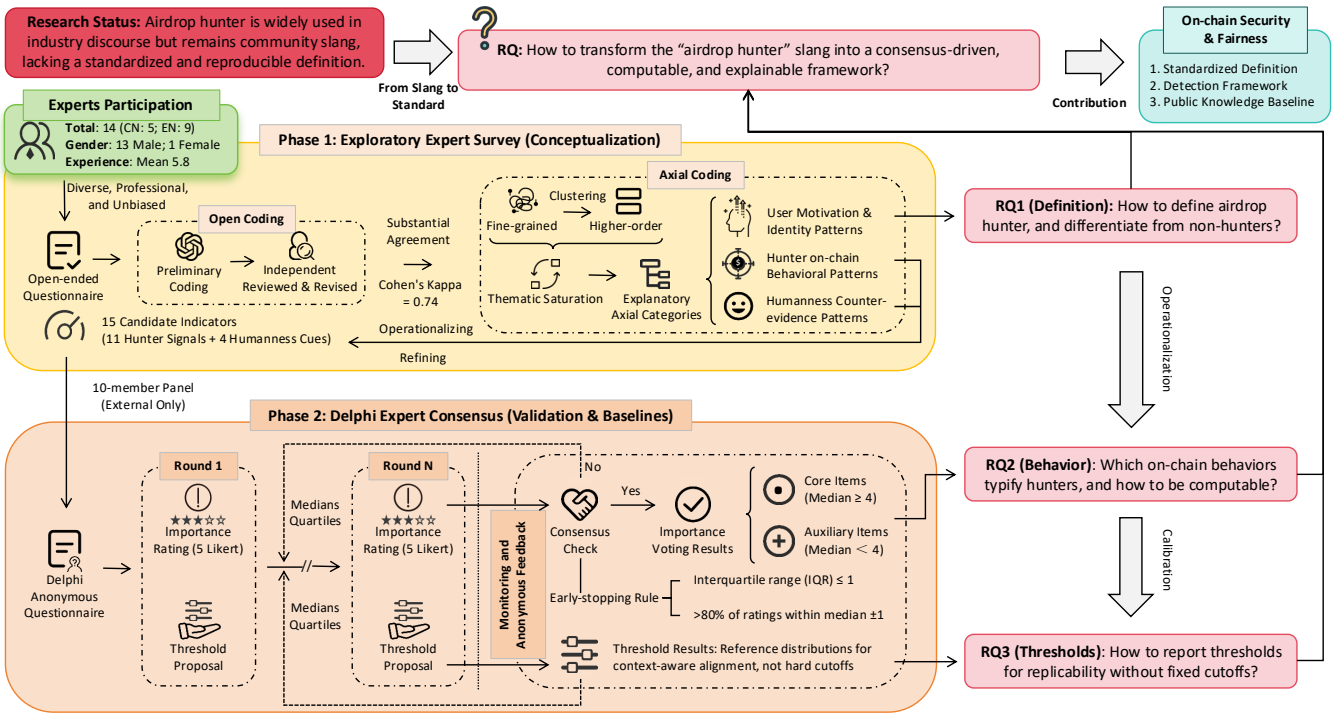


Figure 1: Study overview: open-ended expert survey, qualitative coding, behavior-pattern induction, indicator operationalization, Delphi rounds, and consensus convergence.

we employed the Delphi technique to iteratively converge on the importance and threshold reference ranges of candidate indicators. Delphi is particularly suited for domains with fuzzy definitions and high uncertainty, as it provides structured, anonymous, multi-round feedback to aggregate expert judgment into reproducible consensus [18, 31]. This combination allows us to translate dispersed expert intuitions into *computable, consensus-based indicators and reference thresholds*, yielding both interpretability and methodological rigor.

3.1 Positionality Statement

This study was conducted by three core researchers, all male, currently based in Canada, China, and the United States, respectively. The authors have years of interdisciplinary experience in blockchain security, the cryptocurrency ecosystem, and artificial intelligence, with substantial practical and scholarly expertise in the Decentralized Finance ecosystem, on-chain user behavior analysis, and AI-assisted data analysis methods.

Given the ambiguous definitional boundaries of airdrop hunter detection and the risk of subjective judgment, we paid particular attention to biases stemming from our industry backgrounds. To mitigate these effects, we held regular reflexive discussions throughout data collection, coding, and indicator design, critically examining potential biases in study design, coding, and interpretation to promote objectivity, fairness, and transparency in both the research process and reporting [13, 21, 35].

3.2 Study Design Overview

We adopt a rigorous multi-stage mixed-methods design organized into two sequential phases (see Figure 1).

- **Phase 1: Exploratory expert survey and behavior-pattern induction.** We invited *12 external experts and 2 author-participants* to complete an open-ended questionnaire that elicited free-text descriptions of the definition of "airdrop hunters," typical on-chain behavioral patterns, identification procedures, and related on-chain indicators. After collecting responses, we conducted a qualitative analysis pipeline consisting of AI-assisted pre-coding [45], dual human adjudication and thematic saturation analysis [5], and axial coding to synthesize a core framework of hunter behavioral patterns [40]. This yielded a unified list of patterns and a definitional schema.
- **Phase 2: Delphi consensus on computable indicators.** Building on Phase 1, we operationalized the qualitative patterns into a set of computable on-chain indicators (definitions and computational recipes). For the anonymous Delphi rounds, the two author-participants were *excluded* from rating to avoid authority effects, and two of the 12 external experts *withdrew*; the panel therefore comprised 10 external experts. Experts rated indicator *importance* (Likert 1–5) and proposed numeric *thresholds* over two rounds. We used an *early-stopping* rule: *only* indicators that *did not* reach importance consensus in Round 1 (defined as *both* interquartile

Table 1: Participant summary for Phase 1 (n = 14; 12 external experts + 2 author-participants).

Section	Item	Value	Item	Value
Summary	Total respondents	14 (CN: 5; EN: 9)	Sampling	Via authors' networks
	Gender	13 male; 1 female	Ethics	IRB; voluntary; anonymized
	Years in industry	Mean 5.8; range < 1 to > 10	Data retention	3 years

Section	Item	Value
Region	Canada	3
	United States	1
	Germany	1
	Singapore	1
	United Arab Emirates	1
	China	1
	North America (unspecified)	2
	Asia (unspecified)	1
	Korea	1
	Europe (unspecified)	1
	United Kingdom	1

Section	Item	Value
Domain	Academia / Research	3 (21.4%)
	Engineering / Development	3 (21.4%)
	Industry application / Management	4 (28.6%)
	Community / Operations / Growth	3 (21.4%)
	Multi-role	1 (7.1%)
Experience	< 1 year	1 (7.1%)
	1–2 years	2 (14.3%)
	3–5 years	3 (21.4%)
	6–9 years	4 (28.6%)
	≥ 10 years	4 (28.6%)

range, IQR ≤ 1 *and* consensus proportion $\geq 80\%$ within the median band $\tilde{r} \pm 1$) were re-rated for importance in Round 2; *all* indicators provided threshold proposals in *both* rounds.

3.3 Participants & Recruitment

3.3.1 Recruitment Strategy and Rationale. We employed purposive sampling to select informative and representative industry experts, ensuring data quality and analytic depth [33]. To build a unified and transparent expert-consensus indicator system for identifying airdrop hunters, it was essential that participants possess substantial experience in blockchain security or airdrop analysis and a deep understanding of arbitrage behaviors in airdrop campaigns.

Experts were recruited according to four criteria: (i) research or hands-on experience in blockchain security or airdrop behavior analysis; (ii) current or recent activity in the cryptocurrency industry, enabling up-to-date practitioner perspectives on on-chain user behavior; (iii) clear knowledge of, and extensive practical experience with, airdrop-arbitrage behaviors, especially airdrop hunters; and (iv) willingness to participate without monetary compensation to avoid bias introduced by financial incentives. Rather than maintaining a separate expert registry, we issued targeted invitations through the authors' long-term cryptocurrency industry networks and ecosystem partners to ensure high-quality expertise and credibility, obtaining rich information for the research questions [11].

3.3.2 Recruitment Procedure. **Step 1: Shortlisting candidates (Phase 1).** Drawing on cryptocurrency ecosystem collaborations, the authors compiled a shortlist based on inclusion criteria and invited *12 external experts*. In addition, *2 authors* completed the same open-ended questionnaire as participants for Phase 1 only, contributing domain knowledge to the inductive coding stage.
Step 2: Formal invitations and informed consent. All 14 Phase 1 respondents (12 external experts + 2 authors) participated under IRB oversight. Experts were invited individually via email that included the study purpose, the estimated time required for each round, and the informed-consent statement. Participation was explicitly described as voluntary and uncompensated. Data were anonymized and scheduled for deletion three years after study completion.
Step 3: Confirming the Delphi panel (Phase 2). For the anonymous Delphi rounds, *the two authors were excluded from all scoring*. Two external experts *withdrew*, leaving a 10-member Delphi panel, consistent with common recommendations for panel size [31].
Step 4: Mitigating researcher authority bias. Author exclusion from Phase 2 ratings and full anonymity across respondents were used to reduce authority and conformity biases.

3.3.3 Sample Description. Phase 1 included *14 respondents* (*12 external experts and 2 author-participants*) with summary statistics in Table 1. The diversified sample was obtained via purposive sampling [11, 33], solidifying the empirical basis for subsequent analyses.

3.4 Phase 1: Exploratory Expert Survey

3.4.1 Questionnaire Design. The core objective of this phase was to collect experts' understandings, definitions, and judgment procedures regarding "airdrop hunters" through an open-ended questionnaire, laying the groundwork for the subsequent Delphi phase. The instrument comprised four sections:
Section 1: Background (approx. 1 minute). Field of work, years of experience, and optional gender and region.
Section 2: Definition and behavioral characteristics (approx. 10 minutes). A detailed definition of airdrop hunters, typical on-chain indicators, and cross-scenario differences in arbitrage behaviors; experts were encouraged to share concrete cases.
Section 3: Methods and tools (approx. 3 minutes). Practical steps, commonly used tools, and key decision indicators.
Section 4: Additional suggestions and follow-up (approx. 1 minute). Open-ended suggestions and willingness to participate in the Delphi consensus round.
The complete questionnaire is provided in Appendix A.1.

3.4.2 Data Collection. Data collection proceeded in three steps: (1) the research team distributed a bilingual (Chinese and English) Word-format questionnaire to 14 respondents; (2) completed forms were returned via email; and (3) responses were consolidated into a restricted-access Google Docs spreadsheet accessible only to core research members to ensure data security and privacy.

We followed established practices for cross-language instrument translation and quality assurance [1, 7, 17, 37]. Phase 1 data collection proceeded asynchronously over approximately two weeks, allowing experts to complete the questionnaire at their own pace. Each Delphi round in Phase 2 lasted about one week, providing sufficient time to review aggregated feedback and submit ratings. We used an open-ended questionnaire rather than semi-structured interviews for two reasons. First, given experts' geographic dispersion across time zones, asynchronous self-completion maximized scheduling flexibility. Second, written questionnaires ensured identical prompts for all respondents, reduced interviewer effects, and aligned with the anonymity principles of the Delphi method to maintain methodological consistency [10, 31].

Box 1. LLM Pre-coding Configuration

Model: GPT-4.5 (OpenAI web interface, default parameters)

Input unit: One expert response per session

Task prompt: *"Conduct open coding on the following text: {response}."*

Output: Descriptive labels paired with source text spans

3.4.3 Qualitative Analysis: Open Coding, Axial Coding, AI Assistance. We employed a grounded-theory qualitative analysis pipeline comprising open coding and axial coding, augmented with AI assistance to improve consistency and efficiency.

(1) Open coding with AI-assisted pre-coding. We used GPT-4.5 to generate preliminary open codes over all expert responses. As detailed in Box 1, the prompt was designed to elicit descriptive labels close to participants' own language while avoiding theoretical imposition, consistent with grounded-theory principles of open coding [40]. The model output served solely as a lightweight reference for subsequent human coding and all analytical decisions regarding label refinement, merging, and conceptual abstraction were made by human researchers [45].

(2) Dual human adjudication and agreement checking. Two researchers independently reviewed and revised the AI-generated preliminary codes and reached consensus through item-by-item discussion. Inter-rater agreement yielded Cohen's $\kappa = 0.74$, which is commonly interpreted as substantial agreement [23], supporting the objectivity and reliability of the coding process.

(3) Axial coding and core-theme induction. Building on the open codes, two researchers conducted axial coding to cluster fine-grained labels into higher-order dimensions [40]. Guided by our research aims, we focused on three explanatory axial categories: ***User motivation & identity patterns*** (addressing RQ1 by producing a qualitative definition of airdrop hunters in contrast with regular users), and two sets of behavioral modes—***Hunter on-chain behavioral patterns*** and ***Human-ness counter-evidence patterns***—which together yielded 15 candidate indicators (11 hunter-behavior signals and 4 human-ness counter-evidence signals). These categories formed both the conceptual foundation for defining "airdrop hunter" and the empirical basis for subsequent Delphi rounds.

(4) Thematic saturation and narrative richness. We monitored the emergence of new themes throughout open and axial coding and observed a marked decline after the 10th response; by the 12th response, no additional themes appeared, indicating theoretical saturation [5]. In addition to saturation as evidence of conceptual sufficiency, the responses themselves were paragraph-level and multi-sentence, and many experts used the optional remark fields to provide further clarifications, examples, or methodological reflections. Together, this combination of saturation and narrative richness ensured sufficient depth for reliable open and axial coding and underpins the validity of the behavioral pattern taxonomy used in the subsequent Delphi phase.

3.5 Phase 2: Delphi Expert Consensus

3.5.1 Operationalizing Behaviors into On-chain Indicators. The behavioral patterns distilled in Phase 1 were refined and operationalized into a clear, actionable set of on-chain quantitative indicators, ensuring objectivity and data-driven evaluation in the subsequent Delphi phase. To support portability across different campaigns and chains, we intentionally do not prescribe a universal time window Δt; instead, each indicator leaves Δt as a tunable parameter, and the selected value should be reported for reproducibility. As a practical reference, major airdrop programs have employed varied time windows depending on the behavioral pattern of interest. Arbitrum's official anti-sybil rules use 48 hours to penalize addresses whose

Table 2: Delphi questionnaire examples (excerpt). Blanks are completed by experts during rating.

Category	Code	Indicator definition	Computation	Importance	Threshold
Operational pattern	BT	Multiple addr. executing highly similar transaction sequences with nearly identical amounts.	Num. of times multiple addr. execute completely or highly similar transaction sequences with nearly identical amounts within a short period.	□ 1 □ 2 □ 3 □ 4 □ 5	Hunter: identical sequences ≥ {*expert fill-in*} times
Operational pattern	BW	A single funding source activating multiple new wallets in a short period.	Num. of new addr. activated by a single source address within a short period, previously without transaction history.	□ 1 □ 2 □ 3 □ 4 □ 5	Hunter: number of new wallets ≥ {*expert fill-in*}
Positive reference	LA	Stable long-term trading activities by an addr.	Num. of months of continuous activity by an addr.	□ 1 □ 2 □ 3 □ 4 □ 5	Non-Hunter: continuous activity ≥ {*expert fill-in*} months

Note. Numeric thresholds were intentionally left unspecified *a priori*; experts filled them in based on practice and observation.

entire transaction history falls within a narrow window [4]. Empirical studies report that over 60% of airdropped tokens are liquidated within 24 hours by profit-seeking users [29]. Optimism defines "distinct weeks" (7-day periods) to measure repeated usage, requiring activity across four or more such weeks to qualify as a repeat user [32]. ZKsync uses 30 days as a threshold for contract activity maturity evaluation [49]. For cross-chain or long-term behavioral indicators, windows of 90 days or longer are common [32, 49]. These ranges are illustrative rather than normative; practitioners should calibrate Δt based on chain throughput, gas economics, and campaign-specific eligibility rules. Following standard Delphi procedures [18, 31], we designed a structured expert questionnaire in which experts rated the identification importance of each indicator (Likert 1–5) and proposed threshold ranges that would indicate either highly suspicious (Hunter) or clearly non-Hunter cases. An excerpted template is shown in Table 2; the full Delphi questionnaire (rating prompts and threshold fields) is included in Appendix A.2.

3.5.2 Anonymous Rounds and Feedback. A 10-member Delphi panel (external experts only) completed all ratings anonymously. The procedure followed two rounds:

Round 1: Initial importance ratings and threshold proposals. Experts independently rated the importance of each indicator (1–5) and proposed numeric thresholds.

Between-round feedback. For each indicator, we reported the Round-1 *median and quartiles (Q1, Q3) for importance ratings*, as well as the *median and quartiles (Q1, Q3) for threshold proposals*, accompanied by brief distribution notes.

Round 2: Targeted re-rating for non-consensus items only. *Only* indicators that had *not* reached consensus on importance in Round 1 were carried forward for a second importance rating. Here, "consensus on importance" is defined as *both* (i) interquartile range (IQR) ≤ 1 and [42](ii) a consensus proportion $\geq 80\%$, computed as the share of experts whose ratings fell within the median band $\tilde{r} \pm 1$. Regardless of Round-1 status, *all* indicators provided threshold proposals again in Round 2.

Stopping rule. We pre-specified a two-round Delphi design to balance convergence opportunities with participant burden. After two rounds, the vast majority of importance indicators had reached consensus; we therefore terminated the process, while transparently reporting the few items that remained non-consensual.

3.5.3 Consensus Criteria and Finalization. **Importance (consensus rule).** An indicator reaches *consensus on importance* when its ratings satisfy *both*: (i) IQR ≤ 1 and (ii) consensus proportion $\geq 80\%$ within the median band $\tilde{r} \pm 1$. Indicators meeting this rule after Round1 were *not* re-rated for importance in Round 2.

Core indicators (inclusion rule). After 2 Delphi rounds, indicators meeting *(a)* the above importance-consensus rule and *(b)* a final-round median importance ≥ 4 were designated *core indicators*.

Thresholds (reference-only). Threshold proposals were treated as *reference* rather than an inclusion criterion. For *all* indicators, we conducted *two* rounds of threshold elicitation. In each round, we computed a *reference consensus proportion* as the share of experts whose proposed values fell within $\tilde{t} \pm 10\%$ of the round median $\tilde{t}$. We report final medians and dispersion for interpretability. Thresholds were *not* used to determine core-set membership.

3.6 Reliability and Bias Mitigation

To ensure the reliability and credibility of both process and results, we implemented four mitigation measures:

(1) Reducing coding subjectivity. As reported in Section 3.4.3, inter-rater agreement yielded Cohen's $\kappa = 0.74$, commonly interpreted as substantial agreement [23], supporting the objectivity and reliability of our analysis.

(2) Minimizing translation error. We followed a rigorous translation protocol with machine-assisted back-translation [7] and line-by-line bilingual checks by a native Chinese author, following established guidance on translation, adaptation, and instrument design [17]. These steps maximized semantic equivalence between the Chinese and English versions and reduced ambiguity [1, 37].

(3) Preventing authority and conformity bias in Delphi. All Phase 2 ratings were anonymous; panelists were unaware of one another's identities or scores. The two authors who completed the Phase 1 open-ended questionnaire were *explicitly excluded* from all Delphi ratings, and two external experts withdrew before/during Delphi; the final panel size was 10 [18, 31].

(4) Ensuring analytic sufficiency and saturation. We tracked new themes throughout coding. Novel themes declined after the 10th response, and none emerged by the 12th, indicating theoretical saturation and analytic completeness [5].

Together, these measures reinforce the study's methodological rigor and bolster confidence in the validity of our findings.

3.7 Ethics and Open Science

This study followed relevant ethical research guidelines and was approved by our institutional review board (IRB protocol). All experts provided electronic informed consent after being briefed on the study purpose, procedures, and anonymity principles, protecting their rights to informed participation and privacy.

All data were anonymized and de-identified, stored under expert IDs in a restricted-access Google Docs folder accessible only to core research members. We will permanently delete these data within three years of study completion to protect privacy and security.

Where feasible, and subject to de-identification constraints, we plan to release a subset of materials, including selected data excerpts, the coding handbook, and analysis scripts, on the Open Science Framework (OSF) to support transparency and reproducibility.

4 Results

4.1 RQ1: Definition and Conceptual Dimensions

In Phase 1, we distributed an open-ended questionnaire to 14 experts and analyzed their narratives using open and axial coding. We clustered the narratives into six higher-order categories that consistently distinguish airdrop hunters from regular users, and present each category with representative verbatim excerpts.

Motivation & Value Orientation. Experts agreed that motivation is the primary divide. Hunters were portrayed as participants whose "only goal is to obtain rewards and eventually liquidate, with no long-term significance beyond inflating surface-level metrics." One expert noted that "their core motivation is not to use product functions, participate in governance, or recognize the project's value, but to extract rewards strategically." By contrast, regular users were

described as joining for genuine use: "they may use a DeFi protocol for lending or participate in NFTs or games out of real interest, with airdrops seen as an incidental bonus." Several experts stressed that this divergence harms fairness, as hunters "dilute the share of genuine contributors and may harm token value through short-term sell-offs."

Temporal Rhythm. Hunters were described as concentrating activity around key moments: "they often enter just before an airdrop is expected, and quickly exit once rewards are distributed." Activity was said to "spike at snapshots or claim points, then go silent afterward." By contrast, regular users were seen as more continuous and dispersed: "often continue engaging across cycles, while hunters go dormant after claiming." This temporal contrast highlights opportunistic bursts versus sustained participation.

Organization & Automation. Experts highlighted hunter activity as industrialized: "more like studios, using bots and systematic playbooks to maximize efficiency." Another noted that hunters "register many wallets at low cost and operate them in batches through group control." Some added that hunters "hire people or use coordinated studios to simulate usage." By contrast, regular users were portrayed as acting individually, at a human pace: "their interactions are organic and not batch-style or automated."

Behavioral Scope & Depth. Hunters were described as doing only the bare minimum: "they only complete the tasks necessary for eligibility," often through "volume boosting, fake transactions, or repeated small interactions to fabricate activity." Others noted that hunters "precisely hit the minimum task requirement and then stop." While regular users were seen as engaging more broadly and deeply: "they naturally use swaps, staking, governance, or games, with behavior that is more diverse and sustained." This dimension contrasts shallow compliance with authentic, multifaceted activity.

Fund & Address Topology. Funding flows were another consistent marker. Hunters' wallets were "often funded by the same address" and quickly consolidated: "airdropped funds from multiple wallets are rapidly gathered into one." Experts also observed cyclical and multi-hop transfers: "funds circulate repeatedly among controlled addresses." By contrast, regular users' flows were "more heterogeneous, with no systematic consolidation or loops." This highlights structured, sybil-like topologies versus organic funding.

Identity & Human-ness Cues. Finally, experts highlighted cues of authenticity. Hunters were described as "new wallets with no prior history," "low-cost accounts," and "addresses lacking DID bindings, POAPs, or diverse assets." Their traces looked mechanical: "regular intervals, low diversity, little evidence of genuine community involvement." While regular users were said to "bind to ENS or other DIDs, have older addresses, richer portfolios of NFTs and tokens, and more natural rhythms of interaction." These cues distinguish automated traces from human-like participation.

Synthesis and Definition. Through iterative open and axial coding, we synthesized six core dimensions: motivation, temporal rhythm, organization/automation, behavioral scope and depth, fund

Table 3: Conceptual Contrast Matrix (RQ1, Six Dimensions)

Dimension	Airdrop Hunters (Expert Consensus)	Regular Users (Contrast)	Representative Excerpts
Motivation & Value Orientation	Reward-centric; maximize expected token returns; eligibility-first; not driven by product/mission.	Utility/community value; the airdrop is an "extra reward," not the primary reason for participation.	"The primary goal is to maximize airdrop rewards." (transl.)
Temporal Rhythm	Bursting around snapshot/claim points; quickly goes quiet after claiming; rapid cash-out.	Behavior is more temporally dispersed; continues outside campaign windows.	"They stop being active soon after the airdrop." (transl.)
Organization & Automation	Studio-like operation; multiple wallets/scripts; reuse of playbooks/SOPs across projects; low-cost, high-efficiency.	Individualized, non-batch, human pace; lack of orchestration.	"By using bots and studio-style operations." / "Very familiar with the standard procedures." (transl.)
Behavioral Scope & Depth	Minimal compliance under task/points/ranking systems; simulated use via volume-boosting/faked transactions/check-ins.	Diverse dApp use (swaps, staking, governance, NFTs/games, etc.); deeper and sustained.	"They only do the bare minimum for the airdrop." (transl.)
Fund & Address Topology	Multi-wallet shared/common-origin funds; consolidation after claiming; cyclical/multi-hop transfers.	More mixed funding sources; lack of systematic consolidation or cycles.	"Funded by the same address." (transl.)
Identity & "Humanness" Cues	Avoid DID/identity binding; many new/low-history wallets; sparse POAP/NFT and asset diversity; mechanical temporal distributions.	Bind DID/ENS; older addresses; richer POAP/NFT and asset diversity; natural temporal distributions.	"Not genuinely interested in the project." (transl.)

Language note: Chinese originals translated (marked as transl.); English originals retained.
Traceability: The six dimensions derive from our open coding and axial aggregation of expert materials.

and address topology, and identity/human-ness cues, which define a stable conceptual contrast between hunters and regular users.

Definition. We define *airdrop hunters* as actors—individuals, teams, or automated systems—whose primary objective is to extract value from airdrop campaigns. They employ structured processes such as multiple wallets, bots, or scripted interactions to achieve minimal eligibility across projects, often through simulated use or fabricated traces. They rapidly consolidate or liquidate rewards after distribution, with participation driven by reward extraction rather than product utility or community identification.

For clarity, Table 3 summarizes these six dimensions in contrast form, aggregating open codes and representative excerpts.

Further analysis shows that this definition can be elucidated more clearly by contrasting with regular users along six core dimensions (see Table 3). The table follows our axial–open coding scheme for RQ1: each row is an *axial category* (dimension); the two middle columns summarize the *aggregated open codes* for hunters vs. regular users; and the rightmost column provides a *representative verbatim excerpt* (translated where noted). Experts conceptualize airdrop hunters as playbook-based, profit-maximizing batch operators: they reuse the same standard operating procedure (SOP) across multiple projects, use multiple wallets or scripts to achieve minimum compliance and volume-boosting *and* faked check-ins (e.g., repeated, very-small-amount interactions to fabricate activity or task-completion traces), and rapidly consolidate or liquidate post-distribution. By contrast, regular users are driven by utility/community, exhibit natural and sustained engagement, and treat the airdrop as an incidental reward rather than the core motivation for participation. This conceptual difference is stable across six dimensions (motivation, temporal rhythm, organization/automation, behavioral scope and depth, fund *and* address topology, and identity *and* human-ness cues), laying the groundwork for the diagnostic taxonomy in RQ2 and threshold setting in RQ3.

4.2 RQ2: Indicators and Consensus Importance

4.2.1 RQ2-A: From Open/Axial Coding to Operationalized Indicators. Tables 4 and 5 summarize the computable universe distilled from open and axial coding. For traceability, each table includes an *Open-code anchors* column listing 2-4 English tags per row. Each row provides a short code and name, a concise behavioral definition, and the exact *computational specification* to reproduce the signal on chain. Read these as reusable measurement recipes rather than labels: BT/BW/RF/TL/MA/FC/HF/MC/RC/SP/CA are detection-oriented operational patterns; LA/AD/DID/NT are human-ness counter-evidence intended to mitigate false positives. These definitions result in an "annotation handbook" and underpin the importance ratings (RQ2-B/C) and threshold descriptions (RQ3).

4.2.2 RQ2-B: Delphi Round 1 (R1) Importance. Tables 6 and 7 report Round 1 (R1) importance statistics (Q1/Median/Q3, IQR, and the share of ratings within Median ±1), and Figure 2 visualizes the distribution of expert ratings as 100% stacked shares (5–1).

Per the pre-specified rule detailed in Methods, an indicator reaches *importance consensus* when IQR ≤ 1 and the consensus ratio $\geq 80\%$; *core* additionally requires Median ≥ 4. Read row-wise: R1 consensus items skip R2 re-rating; others proceed to R2. Results show a stable two-axis structure: template/batch operations (BT, BW, HF) and consolidation/cycling (RF, MA) are both high-importance and high-agreement and thus core; all human-ness indicators (LA, AD, DID, NT) are likewise core; TL is consensus but non-core (Median = 3); RC, SP, FC, CA, MC lack R1 importance consensus.

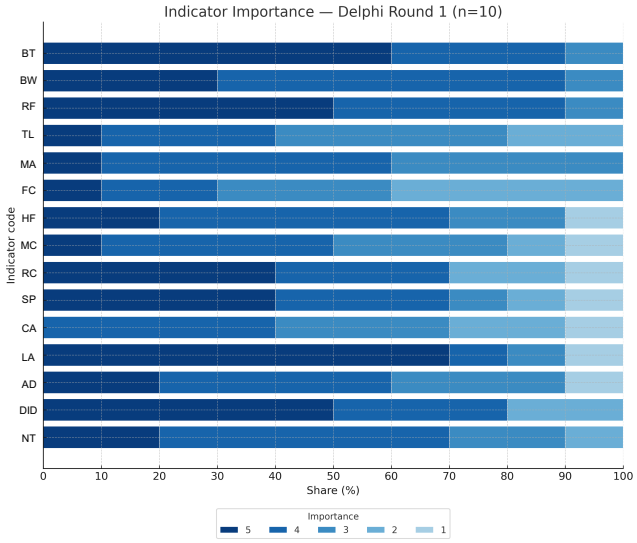


Figure 2: Round 1 Delphi importance (n=10). The figure visualizes the distribution of expert ratings as 100% stacked shares (5–1; darker shade = higher importance).

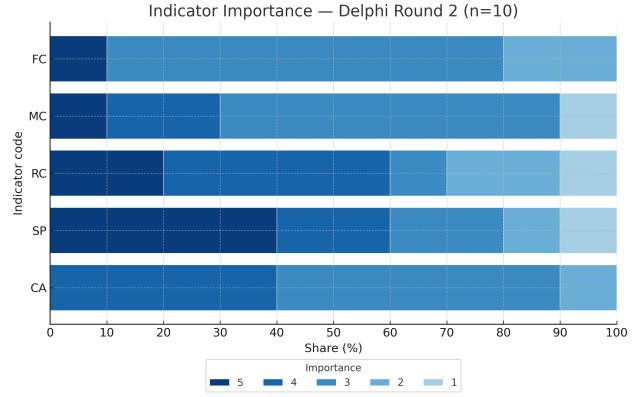


Figure 3: Round 2 Delphi importance (n=10). The figure visualizes the distribution of expert ratings as 100% stacked shares (5–1; darker shade = higher importance).

4.2.3 RQ2-C: Delphi Round 2 (R2) Importance and R1⟶R2 Changes. Figure 3 and Table 8 report R2 importance statistics for indicators that lacked R1 importance consensus (per Methods, only non-consensus are re-rated). By same columns and decision rule, Table 9 compares R1 and R2 medians and dispersion. No additional items become core in R2: FC/CA/MC reach importance consensus but remain non-core (Median < 4), while RC and SP retain high dispersion and fail to reach importance consensus despite Medians of 4, suggesting context sensitivity rather than standalone diagnostic.

Table 4: Operational indicators: definitions and computational specifications

Code	Name	Definition	Computational specification	Open-code anchors
BT	Automated Scripted Trading	Multiple addresses executing highly similar transaction sequences with nearly identical amounts.	Number of times multiple addresses execute completely or highly similar transaction sequences with nearly identical amounts within a short period.	automation/bots/scripts; parameter homogeneity; mechanical timing; repeated small-amount ops
BW	Batch New-Wallet Activation	A single funding source activating multiple new wallets in a short period.	Number of new addresses activated by a single source address within a short period, previously without transaction history.	batch wallet operations; new/low-history wallets; organized group control
RF	Rapid Consolidation of Airdrop Funds	Multiple addresses quickly consolidating received airdrops to a single address.	Proportion of airdropped funds quickly consolidated to a single address from multiple addresses within a short period.	reward consolidation; common-origin funding; short-hop paths; abnormal inflows
TL	Two-Address Cyclical Arbitrage	Frequent reciprocal transfers between two addresses within a short period.	Number of reciprocal fund transfers between two addresses within a short period.	repeated dyadic interactions; two-address cycles; strategy-driven timing
MA	Multi-Address Circular Arbitrage	Funds forming closed-loop circulation among multiple addresses within a short period.	Number of closed-loop token interactions among multiple addresses.	multi-address cycles; multi-hop transfers; group arbitrage
FC	Highly Concentrated Funding Sources	Funds of a single address predominantly sourced from a few addresses.	Proportion of funds from the top two addresses contributing most significantly to a single address.	common-origin funding; source concentration; cost-control pattern
HF	High-Frequency Trading Arbitrage	Exceptionally frequent transactions by a single address to meet airdrop task thresholds.	Proportion of short-term on-chain transactions to the total historical transactions of the address.	high-frequency bursts; volume boosting; task/points campaigns
MC	Precise Task Threshold Arbitrage	Address stops activity immediately after reaching minimum airdrop task requirement.	Whether the address's on-chain activity precisely meets the minimum task requirement with no subsequent additional transactions.	minimum-compliance camouflage; minimum task threshold; rule exploitation
RC	Rapid Asset Liquidation Arbitrage	Quickly liquidating assets shortly after receiving an airdrop.	Proportion of assets liquidated shortly after receiving an airdrop.	post-claim cash-out; short-term participation; volatility avoidance
SP	Snapshot-focused Arbitrage	Address activity heavily concentrated around the airdrop snapshot date.	Proportion of transactions within a specific short window around the snapshot date to total transactions of the address.	snapshot-proximate activity; concentrated entry; bursty rhythm
CA	Cross-Chain Arbitrage Participation	Frequent cross-chain participation in multiple ecosystems within a short period.	Number of different chain ecosystems activated within a short period, each meeting minimum airdrop eligibility.	cross-chain farming; chain-type diversity; gas/threshold adaptation

Table 5: Human-ness counter-evidence indicators: definitions and computational specifications

Code	Name	Definition	Computational specification	Open-code anchors
LA	Long-term Natural Activity	Stable long-term trading activities by an address.	Number of months of continuous activity by an address.	long-run engagement; utility/community identification; extra-reward framing
AD	High-value Asset Diversity	Long-term holding of multiple blockchain assets with clear long-term value.	Number of asset types with clear long-term value held by the address (e.g., blue-chip NFTs, mainstream tokens BTC/ETH, established ecosystem POAPs).	asset diversity; POAP/NFT holdings; portfolio richness
DID	Clear DID Identity Binding	Address explicitly bound to decentralized identity.	Whether the address is explicitly bound to a DID (e.g., ENS, Lens).	DID binding; identity signals; on/off-chain linkage
NT	Natural Trading Behavior	Transaction intervals not exhibiting obvious scripting patterns.	Maximum number of consecutive transactions within a short period having identical or highly similar intervals (error $< 5\%$)	natural timing distribution; reduced regularity; human-ness cues

Table 6: R1 importance results (11 operational indicators)

Code	Name	Q1	Median	Q3	IQR	Consensus ratio (Mdn. ± 1) and status
BT	Automated Scripted Trading	4.00	5.0	5.00	1.00	90% **Core**
BW	Batch New-Wallet Activation	4.00	4.0	4.75	0.75	100% **Core**
HF	High-Frequency Trading Arbitrage	3.25	4.0	4.00	0.75	90% **Core**
RF	Rapid Consolidation of Airdrop Funds	4.00	4.5	5.00	1.00	90% **Core**
MA	Multi-Address Circular Arbitrage	3.00	4.0	4.00	1.00	100% **Core**
TL	Two-Address Cyclical Arbitrage	3.00	3.0	4.00	1.00	90% **Consensus, non-core**
MC	Precise Task Threshold Arbitrage	3.00	3.5	4.00	1.00	70% **No consensus**
RC	Rapid Asset Liquidation Arbitrage	2.50	4.0	5.00	2.50	70% **No consensus**
SP	Snapshot-focused Arbitrage	3.25	4.0	5.00	1.75	80% **No consensus**
FC	Highly Concentrated Funding Sources	2.00	3.0	3.75	1.75	90% **No consensus**
CA	Cross-Chain Arbitrage Participation	2.25	3.0	4.00	1.75	90% **No consensus**

Table 7: R1 importance results (4 human-ness indicators)

Code	Name	Q1	Median	Q3	IQR	Consensus ratio (Mdn. ± 1) and status
LA	Long-term Natural Activity	4.25	5.0	5.00	0.75	80% **Core**
AD	High-value Asset Diversity	3.00	4.0	4.00	1.00	90% **Core**
DID	Clear DID Identity Binding	4.00	4.5	5.00	1.00	80% **Core**
NT	Natural Trading Behavior	3.25	4.0	4.00	0.75	90% **Core**

Table 8: R2 importance results for R1 non-consensus indicators ("±1" ratio is descriptive)

Code	Name	R2 Median	R2 IQR	R2 Q1	R2 Q3	R2 consensus ratio (Median ± 1) and status
CA	Cross-Chain Arbitrage Participation	3.0	1.00	3.00	4.00	100% **Consensus, non-core**
FC	Highly Concentrated Funding Sources	3.0	0.00	3.00	3.00	90% **Consensus, non-core**
MC	Precise Task Threshold Arbitrage	3.0	0.75	3.00	3.75	80% **Consensus, non-core**
SP	Snapshot-focused Arbitrage	4.0	*2.00*	3.00	5.00	80% **No consensus**
RC	Rapid Asset Liquidation Arbitrage	4.0	*1.75*	2.25	4.00	70% **No consensus**

Table 9: R1→R2 contrast in importance (medians and IQRs)

Code	Name	R1 Median	R2 Median	ΔMedian	R1 IQR	R2 IQR	ΔIQR
CA	Cross-Chain Arbitrage Participation	3.0	3.0	0.0	1.75	1.00	−0.75
FC	Highly Concentrated Funding Sources	3.0	3.0	0.0	1.75	0.00	−1.75
MC	Precise Task Threshold Arbitrage	3.5	3.0	−0.5	1.00	0.75	−0.25
SP	Snapshot-focused Arbitrage	4.0	4.0	0.0	1.75	2.00	+0.25
RC	Rapid Asset Liquidation Arbitrage	4.0	4.0	0.0	2.50	1.75	−0.75

4.2.4 RQ2-D: Summary. Across two rounds, the *core* set is fixed by importance consensus *and* Median ≥ 4: five operational indicators (BT, BW, HF, RF, MA) and four human-ness indicators (LA, AD, DID, NT). TL (R1) and FC/CA/MC (R2) are consensus but non-core; RC and SP remain non-consensus, reinforcing their auxiliary, context-sensitive roles rather than standalone determinants.

4.3 RQ3: Thresholds and Reference Ranges

4.3.1 RQ3-A: Measurement Notes and Procedural Context. Thresholds were elicited for all indicators in both rounds according to each computational specification. For every indicator, Round 1 medians and quartiles (Q1, Q3) were summarized and fed back to experts for Round 2 reflection. Thresholds serve as descriptive references for interpretability and deployment tuning only.

4.3.2 RQ3-B: Round 1 Threshold Distributions. Table 10 aggregates each indicator's Round 1 (R1) threshold distribution-Q1/Median/Q3, IQR, sample size n, and the proportion within *Median* ±10% (a concentration cue, not a decision rule; units follow each specification). Read each row by taking the Median as a representative suspicious level and IQR as the degree of agreement/context sensitivity, with

Table 10: R3-1: R1 threshold distributions (units follow each indicator's specification)

Code	Behavior (aligned with RQ2)	Q1	Median	Q3	IQR	Prop. (±10% Median)	n
BT	Identical-sequence occurrences (times)	5.00	5.0	10.00	5.00	0.40	10
BW	New wallets per time unit (count)	10.00	10.0	10.00	0.00	0.70	10
RF	consolidation ratio (%)	50.00	50.0	80.00	30.00	0.44	9
TL	Two-address cycles (times)	5.00	7.5	10.00	5.00	0.00	10
MA	Multi-address closed loops (times)	3.00	5.0	5.00	2.00	0.40	10
FC	Top-2 funding share (%)	55.00	80.0	90.00	35.00	0.00	10
HF	Short-term share (%)	65.00	80.0	87.50	22.50	0.40	10
MC	—	—	—	—	—	—	—
RC	liquidation ratio (%)	78.75	90.0	90.00	11.25	0.60	10
SP	Snapshot-window share (%)	60.00	77.5	83.75	23.75	0.40	10
CA	Chains meeting eligibility (count)	3.00	3.0	3.75	0.75	0.50	10
LA	Months of sustained activity	3.75	6.0	6.75	3.00	0.40	10
AD	Asset categories (types)	3.00	3.0	4.00	1.00	0.56	9
DID	—	—	—	—	—	—	—
NT	Max consecutive identical-interval runs	2.00	3.0	4.75	2.75	0.20	10

the ±10% proportion indicating clustering around the Median. Results show a layered structure: BW (≈ 10) and MA (≈ 5) act as low-dispersion integer anchors; RC is high (Median ≈ 90%) yet relatively tight; SP/RF/FC/TL exhibit heavier tails or multimodality that likely reflect chain/campaign idiosyncrasies and motivate joint adjudication or manual review; LA (≈ 6 months), AD (≈ 3 types), and NT (≈ 3 runs) provide natural-user baselines, while NT's wider IQR indicates differing tolerance for mechanical rhythms.

4.3.3 RQ3-C: Round 2 Threshold Distributions. Table 11 reports Round 2 (R2) threshold statistics in the same format; experts adjusted values after reviewing R1 medians and quartiles. R2 medians serve as engineering defaults, IQR reflects remaining dispersion, and the ±10% proportion indicates concentration. Medians are largely stable (RC: 90% → 85%, SP: 77.5% → 82.5%), while several indicators converge (IQR ↓), yielding a practical palette with narrow anchors (BW/MA), a moderate midlayer (HF/RC/CA/AD/NT), and a wider, context-sensitive set (SP/RF/FC/TL).

4.3.4 RQ3-D: R1→R2 Threshold Contrast. Table 12 compares R1 and R2 medians and IQRs and reports their differences (Δ = R2 − R1), separating threshold re-positioning (Median changes) from agreement tightening (IQR changes). The dominant pattern is median stability with systematic convergence for several indicators (RF/FC/HF/MA/NT); context-sensitive ones retain wider spreads, warranting joint-signal logic and, where needed, manual review.

4.3.5 RQ3-E: Summary. Across indicators, thresholds were treated as reference distributions rather than decisive inclusion rules. For each item we reported the final-round median, IQR, and the proportion of experts whose values fell within a relative $\tilde{t}$ ± 10% band. This allowed us to characterize both central tendency and dispersion. The results reveal a practical gradient: some indicators converged to extremely narrow anchors (e.g., BW ≈ 10 new wallets, MA ≈ 5 closed-loop), others showed medium-width bands (HF/RC/CA/AD/NT), while a third group remained wide and context-sensitive (SP/RF/FC/TL). This gradient highlights which thresholds can serve as stable "hard triggers" across settings, and which are better interpreted as flexible, environment-dependent reference points. Thus, instead of a one-size-fits-all global constant, our findings provide situationally grounded ranges that projects and analysts can adapt when operationalizing detection systems.

Table 11: R3-2: R2 threshold statistics

Code (unit)	Mdn.	IQR	Q1	Q3	R2 prop.
BT (times)	5.0	5.00	5.00	10.00	0.60
BW (count)	10.0	0.00	10.00	10.00	0.60
RF (%)	50.0	15.00	50.00	65.00	0.60
TL (times)	7.0	4.75	5.25	10.00	0.20
MA (times)	5.0	0.00	5.00	5.00	0.60
FC (%)	80.0	15.00	72.50	87.50	0.40
HF (%)	80.0	7.50	80.00	87.50	0.50
RC (%)	85.0	10.00	80.00	90.00	0.70
SP (%)	82.5	17.00	72.50	89.50	0.70
CA (chains)	3.0	0.75	3.00	3.75	0.50
LA (months)	6.0	4.00	4.25	8.25	0.30
AD (types)	3.0	0.75	3.00	3.75	0.60
NT (times)	3.0	0.75	2.25	3.00	0.50

Note: "R2 prop.", here denotes "prop." within ±10% of the median.

Table 12: R3-3: Median and IQR changes (Δ = R2 − R1)

Code	R1 Mdn.	R2 Mdn.	ΔMdn.	R1 IQR	R2 IQR	ΔIQR	R2 prop.
BT	5.0	5.0	0.0	5.00	5.00	0.00	0.60
BW	10.0	10.0	0.0	0.00	0.00	0.00	0.60
RF	50.0	50.0	0.0	30.00	15.00	−15.00	0.60
TL	7.5	7.0	−0.5	5.00	4.75	−0.25	0.20
MA	5.0	5.0	0.0	2.00	0.00	−2.00	0.60
FC	80.0	80.0	0.0	35.00	15.00	−20.00	0.40
HF	80.0	80.0	0.0	22.50	7.50	−15.00	0.50
RC	90.0	85.0	−5.0	11.25	10.00	−1.25	0.70
SP	77.5	82.5	+5.0	23.75	17.00	−6.75	0.70
CA	3.0	3.0	0.0	0.75	0.75	0.00	0.50
LA	6.0	6.0	0.0	3.00	4.00	+1.00	0.30
AD	3.0	3.0	0.0	1.00	0.75	−0.25	0.60
NT	3.0	3.0	0.0	2.75	0.75	−2.00	0.50

Note: "R2 prop.", here denotes "prop." within ±10% of the median.

5 Discussion

5.1 Revisiting the Questions and Core Findings

RQ1: Is there a clear and consistent definition of "airdrop hunter"? The term has long lacked a unified, operational definition that distinguishes hunters from ordinary users. Drawing on expert narratives elicited via open-ended questionnaires and analyzed through open and axial coding, we derive a workable definition: an airdrop hunter is an actor (individual, team, or automated) primarily motivated by extracting value from airdrops, often meeting eligibility across multiple projects with scripted, multi-wallet, and batch-style procedures, followed by rapid consolidation and/or liquidation, rather than sustained, utility- or community-driven use. Relative to ordinary users, hunters differ along six dimensions (motivation, temporal cadence, organization/automation, behavioral breadth/depth, fund/address topology, and human-ness cues), providing a baseline for subsequent quantification and validation.

RQ2: How do we operationalize expert narratives into computable indicators and validate their importance? Building on RQ1, we created a joint signal inventory: 11 hunter-behavior indicators and 4 human-ness counter-evidence indicators, each with a definition and a computation recipe, followed by two Delphi rounds of importance assessment. Five behavioral indicators reached importance consensus—*Automated Scripted Trading* (BT), *Batch New-Wallet Activation* (BW), *High-Frequency Trading arbitrage* (HF), *Rapid Consolidation of Airdrop Funds* (RF), and *Multi-Address circular Arbitrage* (MA). In parallel, the four human-ness indicators—*Long-term Natural Activity* (LA), *High-value asset diversity* (AD), *Clear DID identity binding:* (DID), and *Natural trading behavior* (NT)—also reached importance consensus, forming a counter-evidence layer to reduce false positives. Other behavioral indicators (e.g., TL/FC/RC/SP/CA/MC) did not reach importance consensus or are context-sensitive; we therefore position them as auxiliary evidence or review triggers rather than stand-alone determinants. Importantly, this divergence does not indicate analytical weakness but rather reflects genuine heterogeneity across chains and campaign designs. Notably, SP shows a slight increase in importance IQR from Round 1 to Round 2; based on the cross-round feedback pattern, we infer that the modest widening of the distribution reflects genuine context-dependent heterogeneity rather than instability in individual judgments.

RQ3: How should thresholds be quantified and presented? Across two Delphi rounds, we elicited threshold distributions (medians and IQRs) for all indicators and tracked changes. Most items converge in Round 2, with stable medians and narrower IQRs, suggesting shared expert intuition about suspicious ranges. A few indicators (e.g., FC and TL) retain wider dispersion, reflecting different but reasonable operational contexts rather than conceptual ambiguity, so a single cross-context "global threshold" is unlikely. We therefore report thresholds as *reference distributions*: the median captures central tendency and the IQR bounds plausible variation, enabling practitioners to align with the distribution while adapting thresholds to local conditions rather than applying rigid cutoffs.

How these findings address the field's core challenges. Taken together, these findings address the core challenges in airdrop-hunter detection, starting with the absence of a standardized definition and labeling baseline. First, to **inconsistent labels and lack of consensus**, we provide a unified operational definition and item-level computation recipes, establishing a shared baseline for fair comparison and reproducibility across methods and projects. Second, for **limited generalization**, we avoid one-size-fits-all cutoffs and report *reference distributions*, treating some signals as auxiliary evidence or review triggers, which supports adaptation across chains, designs, and market conditions. Third, for **limited explainability**, the joint-evidence structure, *operational* (BT/BW/HF) crossed with *fund-flow* (RF/MA) and tempered by *human-ness* counter-evidence (LA/AD/DID/NT), provides clear, inspectable reasons for flags, improving interpretability, trust, and adoption in audit and governance workflows. Overall, our approach unifies definition, computation, and context-aware thresholding, establishing a reproducible baseline that supports subsequent detection methods while strengthening ecosystem security and governance.

5.2 Design & Engineering Implications

Grounded in the importance consensus from RQ2 and the threshold distributions and convergence patterns from RQ3, we distill seven transferable implications as a reusable reasoning framework for engineering and governance. To minimize reader burden, we give each indicator as "full name + code" at first mention, then use the code thereafter. Our core decision logic is stated explicitly as **adjudicate using the co-occurrence of template/batch operations and structured fund consolidation/cycling, together with a human-ness counter-evidence layer (LA/AD/DID/NT)**.

I1: Prioritize co-occurrence over single-point extremes. Expert consensus on BT/BW/HF (operational side) and RF/MA (fund-flow side) indicates that strong evidence arises when *template/batch* signals and *structured consolidation/cycling co-occur*, rather than from any single metric at an extreme. In practice, surface candidates where operational intensity and fund-flow structure rise together before adding more complex models. This improves interpretability and aligns with the strong axes observed in our results [34, 48].

I2: Treat human-ness as an evidence layer to systematically reduce false positives. LA/AD/DID/NT are as important as behavioral items; they are not mere filters (e.g., "bound DID" is not a sufficient compliance criterion). When multiple human-ness cues *co-occur* (e.g., sustained long-run activity together with natural timing), reduce hunter confidence or route to manual review. This preserves detection power while protecting genuine users, aligning with community expectations of fairness [2, 29].

I3: Use thresholds as reference distributions*, not universal constants.* Across two Delphi rounds, most medians are stable and several IQRs converge, yet a single cross-context threshold does not emerge. Report medians and quartiles as reference distributions and align locally by chain environment, campaign design, and market state—use within-context quantiles rather than transplanting absolute values [2, 46].

I4: Explain first, automate second. Item-level, evidence-based outputs (e.g., "BT and RF are both high, while LA is also high") help stakeholders understand *why* a case is flagged. Provide traceable rationales and evidence snapshots (hit items, relative position within

Table 13: Delphi-validated high-consensus indicators (5 operational + 4 human-ness counter-evidence), including definitions, computation, and Round-2 threshold reference ranges. Thresholds are reference distributions rather than universal cutoffs.

Code	Name	Definition & Computation	R2 Threshold (Median / IQR / Q1–Q3; ±10%)
BT	Automated Scripted Trading	Multiple addresses executing highly similar transaction sequences with nearly identical amounts. *Computation:* Number of times multiple addresses execute completely or highly similar transaction sequences with nearly identical amounts within a short period.	5 / 5.0 / 5–10 / 0.60
BW	Batch New-Wallet Activation	A single funding source activating multiple new wallets in a short period. *Computation:* Number of new addresses activated by a single source address within a short period, previously without transaction history.	10 / 0.0 / 10–10 / 0.60
HF	High-Frequency Trading Arbitrage	Exceptionally frequent transactions by a single address to meet airdrop task thresholds. *Computation:* Proportion of short-term on-chain transactions to the total historical transactions of the address.	80% / 7.5 / 80–87.5 / 0.50
RF	Rapid Consolidation of Airdrop Funds	Multiple addresses quickly consolidating received airdrops to a single address. *Computation:* Proportion of airdropped funds quickly consolidated to a single address from multiple addresses within a short period.	50% / 15.0 / 50–65 / 0.60
MA	Multi-Address Circular Arbitrage	Funds forming closed-loop circulation among multiple addresses within a short period. *Computation:* Number of closed-loop token interactions among multiple addresses.	5 / 0.0 / 5–5 / 0.60
LA	Long-term Natural Activity	Stable long term trading activities by an addr. *Computation:* Number of months of continuous activity by an addr.	6 / 4.0 / 4.25–8.25 / 0.30
AD	High-Value Asset Diversity	Long-term holding of multiple blockchain assets with clear long-term value. *Computation:* Num. of asset types with clear long-term value held by the addr. (e.g., blue-chip NFTs, mainstream tokens BTC/ETH, established ecosystem POAPs).	3 / 0.75 / 3–3.75 / 0.60
DID	Clear DID Identity Binding	Addr. explicitly bound to decentralized identity. *Computation:* Whether the addr. is explicitly bound to a DID (e.g., ENS, Lens).	N/A
NT	Natural Trading Behavior	Transaction intervals not exhibiting obvious scripting patterns. *Computation:* Maximum num. of consecutive transactions within a short period having identical or highly similar intervals (error <5%)	3 / 0.75 / 2.25–3.0 / 0.50

the distribution, temporal window) before automated actions; this reduces contention costs and supports appeals and review.

I5: Engineer for contested cases instead of chasing zero false positives. Several items (TL/FC/RC/SP/CA/MC) did not reach importance consensus or are context-sensitive. Build explicit channels for contestation: use them as *auxiliary evidence* for ranking or as *review triggers* in conjunction with the human-ness layer, rather than simply raising thresholds at the expense of recall.

I6: Assume adversarial adaptation; favor simple, explainable combinations. As thresholds become targets, they can be gamed, and conditions evolve. Maintain a *compact, explainable* signal set with periodic recalibration (e.g., audits of alerts and false positives). When needed, run a small *mini-Delphi* to update definitions/weights, formalizing adaptation while preserving interpretability.

I7: Design for rehabilitation, not just punishment. Beyond exclusionary measures, consider graduated responses that educate and redirect suspicious participants before permanent bans. For example, platforms could provide transparency dashboards or "second-chance" participation schemes that explain why certain behaviors appear hunter-like and how to avoid them. Such restorative mechanisms align with human-centered and justice-oriented design principles, potentially converting hunters into genuine contributors rather than treating them solely as adversaries.

Integrating the Framework into Detection and Audit Workflows. The indicator set and threshold references produced by this study can be integrated into practical systems at multiple points. First, as a *rule-based pre-filter*: projects may implement the five core behavioral indicators (BT/BW/HF/RF/MA) as initial screening criteria, flagging addresses that exceed reference thresholds for manual review. Second, as a *labeling source for supervised learning*: the operational definitions and consensus thresholds provide a principled basis for generating training labels, reducing reliance on ad-hoc heuristics. Third, as an *audit checklist*: the human-ness counter-evidence layer (LA/AD/DID/NT) offers structured criteria for reviewing edge cases and supporting appeals, improving procedural fairness. In each scenario, the framework supplies interpretable rationales: specific indicators, threshold positions, and co-occurrence patterns, that can be logged and communicated to stakeholders, addressing the explainability gap identified in prior detection approaches. To support direct adoption, Table 13 consolidates the nine high-consensus indicators with their definitions, computation specifications, and Round-2 threshold reference ranges.

5.3 Generalisability & Contextualisation

Detection is not about fixing a single threshold on one dataset. Differences in chain environments (fees/throughput), campaign designs (snapshots vs. tasks/points), and market states (liquidity/cycles) alter costs and incentives, reshaping hunter behavior distributions. Consistent with this, our RQ2 and RQ3 results show that some indicators reach importance consensus while their thresholds display heavy tails, multimodality, or context-dependent concentration. Thresholds should therefore be treated as *interpretable*

reference distributions, not cross-context constants; in deployment, **adjudicate via the co-occurrence of template/batch operations and structured fund consolidation/cycling, tempered by a human-ness counter-evidence layer (LA/AD/DID/NT)**, to balance detection power and fairness.

5.3.1 Chain/Fees (L1 vs. L2). **Mechanism.** Low-fee/high-throughput environments such as rollups lower the marginal cost of micro/batch interactions via batching and data-availability cost sharing, expanding feasible behavior sets [15, 41].
Affected signals. *Operational:* Baselines for HF/BT/BW rise naturally on low-fee chains, making L1-based absolute thresholds overly tight and error-prone. *Funds:* RF/MA cadence accelerates; fixed 24h/7d windows may overstate abnormal speed.
Practice. Prefer rolling-window *quantiles* on the target chain; shorten RF/MA windows modestly and increase BT/BW evidential weight. Always trigger on the co-occurrence of BT/BW/HF with RF/MA, then use the human-ness layer [24].

5.3.2 Campaign Design (Snapshot vs. Tasks/Points). **Mechanism.** Design changes the shortest path to arbitrage: snapshot-style campaigns magnify SP peaks and, around claim events, elevate RC and RF; tasks/points-style campaigns encourage MC/TL with higher HF/BT/BW [2, 3, 46].
Affected signals. SP/RC/MC/TL shift with campaign cadence, increasing the risk of penalizing genuine "deadline-pushers."
Practice. Model campaign windows explicitly: use *relative* within-window quantiles for SP/RC; handle MC/TL with project-internal thresholds plus review priority. Jointly consider LA and NT to avoid misclassifying normal sprints/returns as hunters [34, 48].

5.3.3 Market State (Liquidity / Cycle). **Mechanism.** Market conditions affect the opportunity cost of liquidation and fund movement: in hotter/high-liquidity periods, RC and RF shift rightward and thicken; in colder/low-liquidity periods, RC abnormality may occur at lower levels, while TL/MC/BT (low-cost probing) are comparatively stable [2, 3, 29].
Affected signals. RC and RF are most sensitive to market state, so fixed cutoffs across cycles bias results.
Practice. Re-estimate baselines by market regime and explain context; treat RC/RF as auxiliary ranking factors, decided jointly with the core co-occurrence axes and the human-ness layer.

Summary. Context determines cost structures, and cost structures shape distributions. We therefore advocate treating RQ3 thresholds as **interpretable reference distributions**: in any new scenario, first characterize the local distributions, then **adjudicate using the co-occurrence of template/batch operations and structured fund consolidation/cycling, together with a human-ness counter-evidence layer (LA/AD/DID/NT)**. This extends the importance consensus from RQ2 and translates RQ3 heterogeneity into robust, transferable engineering and governance practices.

5.4 Method & Validity

We frame Delphi rigor and adversarial considerations together: the evidence chain from qualitative data to computable items, the split between importance and thresholds with bias controls, and Goodhart risks with robustness measures.

5.4.1 From Qualitative to Computable: Evidence Chain and Explainability. We begin with open/axial coding to standardize expert narratives into items (definition + computation), then run Delphi scoring. This makes the mapping from *concepts* to *indicators* to *thresholds* traceable and reusable: concept generation and consistency control via open coding with dual independent review (Cohen's $\kappa = 0.74$, "substantial agreement"), then axial aggregation [23, 40, 45]; item standardization and anonymous consensus through itemized definitions and computation with anonymous scoring and cross-round feedback [6, 18, 31]. This "qualitative-to-computable" route compresses dispersed expert tacit knowledge into a shared baseline, enhancing explainability and transferability.

5.4.2 Bifurcating "Importance" and "Threshold". Importance (inclusion/priority) follows a threefold rule—median ≥ 4, IQR ≤ 1, and proportion within median±1 $\geq 80\%$—to decide whether an item enters the core signal set [19]. *Thresholds (contextual reference)* are presented only as interpretable distributions (median and quartiles, with the proportion within median±10% as a descriptive concentration index), not as hard cutoffs—directly reflecting the heavy tails, multi-modality, and context sensitivity observed in RQ3.

5.4.3 Bias Controls and Residual Threats. Multiple anonymous rounds reduce authority and conformity biases; IQR and "±1 band" proportions describe opinion shape and capture convergence and divergence [6, 18]. Delphi remains an "expert snapshot," not first-order on-chain truth; hence we call for ROC/PR evaluation on real on-chain datasets, cross-project transfer tests, and temporal robustness backtests to externally validate threshold reference distributions [31]. For translation and coding risks, we used back-translation and dual review, yet residual model/researcher bias may remain [1, 7, 9, 17, 37, 45].

5.4.4 Goodhart Risks and Robustness: Why Co-judgment, Quantiles, and Counter-evidence. Once thresholds become targets, they invite gaming [14, 39]. Studios can reduce frequency, split flows, or delay movements to evade any single-point rule. Our three-part response is: (1) **co-judgment via structural co-occurrence**—make the joint appearance of template/batch (BT/BW/HF) and structured consolidation (RF/MA) the primary trigger, forcing adversaries to hide on both the operational and fund-flow sides simultaneously; (2) **contextual quantiles**—use within-context quantiles rather than global constants, hindering stable alignment to any fixed cutoff; and (3) **human-ness counter-evidence**—use LA/AD/DID/NT to systematically reduce false positives (DID alone is not sufficient). This elevates single-parameter checks into structural evidence with relative positioning and counter-evidence, aligning with our RQ2–RQ3 findings and practical needs [24].

5.4.5 Synthesis: Unifying Credibility and Robustness. Our framework's credibility rests on a qualitative-to-computable evidence chain (open/axial coding → items → Delphi consensus), importance consensus for inclusion/priority, and thresholds reported as reference distributions for contextual interpretation, with robustness from co-judgment, quantiles, and counter-evidence to mitigate Goodhart risks and reduce false positives. Next, we close the loop between expert evidence and behavioral data via on-chain backtests and cross-context transfer evaluations [19, 31].

5.5 Limitations & Future Work

Despite our efforts toward rigor and transparency, several limitations remain and motivate future work:

(1) **Gender imbalance.** The gender distribution of our expert panel is highly uneven (approximately 92% male). This pattern is also reflected in recent industry statistics—for example, a 2025 market report estimates that women represent about 12% of the global crypto senior workforce [30]. While such external figures contextualize our sample composition, limited gender diversity may constrain the breadth of perspectives represented in the consensus process and influence how certain behavioral patterns are interpreted or weighted. Future work should include more gender-diverse experts to enhance representativeness and external validity.
(2) **Regional bias.** Our panel is concentrated in a limited set of regions (e.g., North America, Europe, East Asia, and the Middle East), where participation costs, infrastructure, and interaction patterns may be more similar. This may shape intuitions about normal rhythms, cost structures, and abnormality thresholds, limiting the cross-ecosystem applicability of some threshold distributions or behavioral interpretations. Future work should recruit experts from a wider range of regions and chain ecosystems to improve robustness across heterogeneous contexts.
(3) **Language equivalence.** We followed a cross-lingual translation workflow but did not run cognitive interviews with native English speakers. Future work should add such interviews to strengthen equivalence [1, 7, 17, 37].
(4) **Threshold validation.** Delphi-derived thresholds are not yet validated on real on-chain datasets. Future work should test them via ROC/PR analysis against labeled events, cross-project transfer to assess generalization, and temporal robustness checks across market regimes to detect drift.
(5) **AI-assisted coding risks.** While LLM assistance improved efficiency, it may introduce misclassification risks. We used dual independent review and κ checks to mitigate this [23, 45]; ongoing audits and calibration remain necessary [9, 43].

These limitations clarify current boundaries and outline concrete avenues for future research.

Summary. Our findings directly respond to the long-standing gaps in airdrop-hunter detection: definitional inconsistency, weak generalization, and poor explainability. Yet their significance goes beyond methodological rigor. At stake is not only whether blockchain projects can deter manipulative behaviors, but whether participants perceive these socio-technical systems as fair, trustworthy, and legitimate. By grounding detection in expert consensus, our framework provides interpretable rationales that governance communities can inspect and contest, rather than opaque algorithmic labels. This shifts the emphasis from "catching bad actors" toward designing infrastructures of *explainability*, *contestability*, and *fairness*, principles at the heart of human–computer interaction research. In doing so, we show how a contested industry slang ("airdrop hunter") can be transformed into a reproducible construct that supports both technical security and fairness in decentralized governance.

6 Conclusion

We introduced the first Delphi-based, computable, and explainable framework for defining and detecting airdrop hunters, transforming what has long remained community slang into a reproducible research construct. By combining open and axial coding with expert consensus, we derived a standardized definition, a taxonomy of behavioral indicators, and reference threshold distributions that together provide a transparent and transferable baseline.

Our findings address three long-standing gaps in the field: definitional inconsistency, limited generalization, and weak explainability. The consensus-based definition supplies a stable ground truth for consistent labeling and evaluation. The joint-evidence framework—spanning operational and fund-flow signals tempered by human-ness counter-evidence—offers interpretable rationales that support both audit and governance. Reporting thresholds as context-aware reference distributions, rather than universal cutoffs, further enables adaptation across heterogeneous chains, campaign designs, and market conditions.

Beyond methodological contribution, this work speaks directly to the health of the broader cryptocurrency ecosystem. By providing shared semantics and computable evidence, our framework strengthens the foundations of cryptocurrency ecosystem security and governance. Future work will validate and extend these baselines on large-scale datasets, enabling robust, fair, and transparent socio-technical infrastructures in the face of strategic manipulation.

References

[1] Dagmar Abfalter, Julia Mueller-Seeger, and Margit Raich. 2021. Translation decisions in qualitative research: a systematic framework. *International Journal of Social Research Methodology* 24, 4 (2021), 469–486.
[2] Darcy WE Allen. 2024. Crypto airdrops: An evolutionary approach. *Journal of Evolutionary Economics* 34, 4 (2024), 849–872.
[3] Darcy WE Allen, Chris Berg, and Aaron M Lane. 2023. Why airdrop cryptocurrency tokens? *Journal of Business Research* 163 (2023), 113945.
[4] Arbitrum Foundation. 2023. $ARB Airdrop Eligibility and Distribution Specifications. Arbitrum DAO Governance Docs. https://docs.arbitrum.foundation/airdrop-eligibility-distribution
[5] Sarah Elsie Baker and Rosalind Edwards. 2012. *How many qualitative interviews is enough.* Technical Report. National Centre for Research Methods, Southampton.
[6] Daniel Beiderbeck, Nicolas Frevel, Heiko A von der Gracht, Sascha L Schmidt, and Vera M Schweitzer. 2021. Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX* 8 (2021), 101401.
[7] Richard W Brislin. 1970. Back-translation for cross-cultural research. *Journal of cross-cultural psychology* 1, 3 (1970), 185–216.
[8] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, USA, 3175–3187. doi:10.1145/3025453.3026018
[9] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. arXiv:2310.15100 [cs.CL]
[10] D.A. Dillman, J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Wiley, United States.
[11] Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, et al. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5, 1 (2016), 1–4.
[12] Sizheng Fan, Tian Min, Xiao Wu, and Wei Cai. 2023. Altruistic and Profit-oriented: Making Sense of Roles in Web3 Community from Airdrop Perspective. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany). ACM, USA, Article 551, 16 pages.
[13] Linda Finlay. 2002. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. *Qualitative research* 2, 2 (2002), 209–230.
[14] C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience.* Macmillan Education UK, London, 91–121.

[15] Lewis Gudgeon, Pedro Moreno-Sanchez, Stefanie Roos, Patrick McCorry, and Arthur Gervais. 2020. SoK: Layer-Two Blockchain Protocols. In *Financial Cryptography and Data Security: 24th International Conference, FC 2020 , Kota Kinabalu, Malaysia, February 10–14, 2020 Revised Selected Papers* (Kota Kinabalu, Malaysia). Springer-Verlag, Berlin, Heidelberg, 201–226. doi:10.1007/978-3-030-51280-4_12

[16] Hanna Halaburda, Benjamin Livshits, and Aviv Yaish. 2025. *Platform Building With Fake Consumers: On Double Dippers and Airdrop Farmers.* Research Paper 5364583. NYU Stern School of Business, New York, NY.

[17] Janet A. Harkness, Ana Villar, and Brad Edwards. 2010. *Translation, Adaptation, and Design.* John Wiley & Sons, Ltd, US, Chapter 7, 115–140. doi:10.1002/9780470609927.ch7

[18] Chia-Chien Hsu and Brian A Sandford. 2007. The Delphi technique: making sense of consensus. *Practical assessment, research, and evaluation* 12, 1 (2007), 10.

[19] Susan Humphrey-Murto, Lara Varpio, Carol Gonsalves, and Timothy J Wood. 2017. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Medical teacher* 39, 1 (2017), 14–19.

[20] Susan Humphrey-Murto, Lara Varpio, Timothy J Wood, Carol Gonsalves, Lee-Anne Ufholz, Kelly Mascioli, Carol Wang, and Thomas Foth. 2017. The use of the Delphi and other consensus group methods in medical education research: a review. *Academic Medicine* 92, 10 (2017), 1491–1498.

[21] Rose Johnson, Yvonne Rogers, Janet van der Linden, and Nadia Bianchi-Berthouze. 2012. Being in the thick of in-the-wild studies: the challenges and insights of researcher participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, US, 1135–1144.

[22] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, USA, 543–550.

[23] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* 33, 1 (1977), 159–174.

[24] Zhong Li, Yuxuan Zhu, and Matthijs Van Leeuwen. 2023. A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data* 18, 1 (2023), 1–54.

[25] Qiangqiang Liu, Qian Huang, Frank Fan, Haishan Wu, and Xueyan Tang. 2025. Detecting Sybil Addresses in Blockchain Airdrops: A Subgraph-based Feature Propagation and Fusion Approach. arXiv:2505.09313

[26] Zheng Liu and Hongyang Zhu. 2022. Fighting Sybils in Airdrops. arXiv:2209.04603

[27] Xiao Ma, Justin Cheng, Shankar Iyer, and Mor Naaman. 2019. When do people trust their social groups?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, US, 1–12.

[28] Christos A Makridis, Michael Fröwis, Kiran Sridhar, and Rainer Böhme. 2023. The rise of decentralized cryptocurrency exchanges: Evaluating the role of airdrops and governance tokens. *Journal of Corporate Finance* 79 (2023), 102358.

[29] Johnnatan Messias, Aviv Yaish, and Benjamin Livshits. 2025. Airdrops: Giving Money Away Is Harder Than It Seems. arXiv:2312.02752

[30] Sarah Mitchell. 2025. Diversity, Equity and Inclusion in the Crypto Industry Statistics. https://gitnux.org/diversity-equity-and-inclusion-in-the-crypto-industry-statistics/. Accessed: 2026-01-20.

[31] Chitu Okoli and Suzanne D Pawlowski. 2004. The Delphi method as a research tool: an example, design considerations and applications. *Information & management* 42, 1 (2004), 15–29.

[32] Optimism Foundation. 2022. Airdrop 1. Optimism Docs. https://github.com/ethereum-optimism/community-hub/blob/main/pages/op-token/airdrops/airdrop-1.mdx

[33] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42, 5 (2015), 533–544.

[34] Yuyang Qin, Tengfei Ma, Hongzhou Chen, and Haihan Duan. 2024. ARTEMIX: A community-boosting-based framework for air-drop hunter detection in the Web3 community. *Blockchain* 2, 2 (2024), 1–24.

[35] Jennifer A Rode. 2011. Reflexivity in digital anthropology. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, US, 123–132.

[36] Fabian Schär. 2021. Decentralized finance: On blockchain-and smart contract-based financial markets.

[37] Mandy Sha and Stephen Immerwahr. 2018. Survey Translation: Why and How Should Researchers and Managers be Engaged? *Survey Practice* 11, 2 (2018), 1.

[38] Jiajie Shi, Yuyang Qin, Hengming Dai, Xiaoyi Fan, and Haihan Duan. 2025. Airdrop Hunter Detection via PageRank-Augmented Multimodal Graph Neural Networks. In *2025 IEEE International Conference on Cloud Computing Technology and Science (CloudCom).* IEEE, US, 1–8.

[39] Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European review* 5, 3 (1997), 305–321.

[40] Anselm Strauss and Juliet Corbin. 1998. *Basics of qualitative research techniques.* Thousand oaks, CA: Sage publications, US.

[41] Domenico Tortola, Andrea Lisi, Paolo Mori, and Laura Ricci. 2024. Tethering Layer 2 solutions to the blockchain: A survey on proving schemes. *Computer Communications* 225 (2024), 289–310.

[42] Heiko A Von Der Gracht. 2012. Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technological forecasting and social change* 79, 8 (2012), 1525–1536.

[43] Qile Wang, Moath Erqsous, Kenneth E Barner, and Matthew Louis Mauriello. 2025. LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–28.

[44] Sam Werner, Daniel Perez, Lewis Gudgeon, Ariah Klages-Mundt, Dominik Harz, and William Knottenbelt. 2022. Sok: Decentralized finance (defi). In *Proceedings of the 4th ACM Conference on Advances in Financial Technologies.* ACM, US, 30–46.

[45] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces.* ACM, USA, 75–78.

[46] Aviv Yaish and Benjamin Livshits. 2024. Tierdrop: Harnessing airdrop farmers for user growth.

[47] Dirk A Zetzsche, Douglas W Arner, and Ross P Buckley. 2020. Decentralized finance. *Journal of Financial Regulation* 6, 2 (2020), 172–203.

[48] Chenyu Zhou, Hongzhou Chen, Hao Wu, Junyu Zhang, and Wei Cai. 2024. Artemis: Detecting airdrop hunters in nft markets with a graph learning system. In *Proceedings of the ACM Web Conference 2024.* ACM, USA, 1824–1834.

[49] ZK Nation. 2024. ZK Airdrop. ZK Nation Docs. https://docs.zknation.io/zk-token/zk-airdrop

A Survey Instruments

A.1 Phase 1: Exploratory Expert Questionnaire

Part1: Expert Background Information

(1) What is your current primary area of work?

(2) How many years of experience do you have in blockchain security or airdrop analysis?

(3) (Optional) Please indicate your gender.

(4) (Optional) In which region are you currently based?

Part2: Airdrop Hunter Definition and Typical On-chain Behavior

(5) In your own words, how would you define an "Airdrop Hunter," and what is the primary difference between an Airdrop Hunter and a regular user?

(6) Based on your experience, which typical on-chain indicators do you usually rely on to identify an address as a potential Airdrop Hunter? Please list freely (no limit on quantity; specify on-chain behavior characteristics in detail).

(7) Considering different types of airdrops or blockchain environments, what notable differences in on-chain behaviors of Airdrop Hunters have you observed? Please provide examples or observed differences.

(8) (Optional) Please briefly share an impressive case of an Airdrop Hunter. You may provide a specific address or transaction hash and briefly describe the typical on-chain behaviors at the time and why it raised your suspicion.

Part3: Methods and Tools for Identifying Airdrop Hunters

(9) What steps or tools do you typically use when determining whether an address is an "Airdrop Hunter"?

Part4: Additional Comments and Future Participation

(10) (Optional) Apart from the topics discussed above, are there any other critical clues or recommendations you believe are essential for identifying Airdrop Hunters, which were not mentioned?

(11) Would you be willing to participate in the second-round Delphi consensus confirmation? (Approximately 5–10 minutes)

A.2 Phase 2: Delphi Expert Questionnaire

Table 14: Delphi Questionnaire — Airdrop Hunter Behavior Patterns

Code	Indicator Definition	Measurement	Importance	Suspicious Threshold
BT	Automated script trading: Multiple addresses executing highly similar transaction sequences with nearly identical amounts	Number of times multiple addresses execute completely or highly similar transaction sequences with nearly identical amounts within a short period	□ 1 □ 2 □ 3 □ 4 □ 5	Identical sequences ≥ {*expert fill-in*} times
BW	Batch new wallet activation: A single funding source activating multiple new wallets in a short period	Number of new addresses activated by a single source address within a short period, previously without transaction history	□ 1 □ 2 □ 3 □ 4 □ 5	Number of new wallets ≥ {*expert fill-in*}
RF	Rapid consolidation of airdropped funds: Multiple addresses quickly consolidating received airdrops to a single address	Proportion of airdropped funds quickly consolidated to a single address from multiple addresses within a short period	□ 1 □ 2 □ 3 □ 4 □ 5	Consolidation ratio ≥ {*expert fill-in*} %
TL	Two-address cyclical arbitrage: Frequent reciprocal transfers between two addresses within a short period	Number of reciprocal fund transfers between two addresses within a short period	□ 1 □ 2 □ 3 □ 4 □ 5	Reciprocal transfers ≥ {*expert fill-in*} times
MA	Multi-address circular arbitrage: Funds forming closed-loop circulation among multiple addresses within a short period	Number of closed-loop token interactions among multiple addresses	□ 1 □ 2 □ 3 □ 4 □ 5	Closed-loop transactions ≥ {*expert fill-in*} times
FC	Highly concentrated funding sources: Funds of a single address predominantly sourced from a few addresses	Proportion of funds from the top two addresses contributing most significantly to a single address	□ 1 □ 2 □ 3 □ 4 □ 5	Top two addresses fund ratio ≥ {*expert fill-in*} %
HF	High-frequency trading arbitrage: Exceptionally frequent transactions by a single address to meet airdrop task thresholds	Proportion of short-term on-chain transactions to the total historical transactions of the address	□ 1 □ 2 □ 3 □ 4 □ 5	Short-term transaction proportion ≥ {*expert fill-in*} %
MC	Precise task threshold arbitrage: Address stops activity immediately after reaching minimum airdrop task requirement	Whether the address's on-chain activity precisely meets the minimum task requirement with no subsequent additional transactions	□ 1 □ 2 □ 3 □ 4 □ 5	Evaluate importance only
RC	Rapid asset liquidation arbitrage: Quickly liquidating assets shortly after receiving an airdrop	Proportion of assets liquidated shortly after receiving an airdrop	□ 1 □ 2 □ 3 □ 4 □ 5	Liquidation ratio ≥ {*expert fill-in*} %
SP	Snapshot-focused arbitrage: Address activity heavily concentrated around the airdrop snapshot date	Proportion of transactions within a specific short window around the snapshot date to total transactions of the address	□ 1 □ 2 □ 3 □ 4 □ 5	Transaction concentration ratio ≥ {*expert fill-in*} %
CA	Cross-chain arbitrage participation: Frequent cross-chain participation in multiple ecosystems within a short period	Number of different chain ecosystems activated within a short period, each meeting minimum airdrop eligibility	□ 1 □ 2 □ 3 □ 4 □ 5	Number of chains meeting minimum eligibility ≥ {*expert fill-in*}

Note. Experts can add remarks on each code.

Table 15: Delphi Questionnaire — Genuine Human User Positive Control Patterns

Code	Indicator Definition	Measurement	Importance	Non-Hunter Standard
LA	Long-term natural activity: Stable long-term trading activities by an addr.	Number of months of continuous activity by an addr.	□ 1 □ 2 □ 3 □ 4 □ 5	Continuous activity ≥ {*expert fill-in*} months
AD	High-value asset diversity: Long-term holding of multiple blockchain assets with clear long-term value	Num. of asset types with clear long-term value held by the addr. (e.g., blue-chip NFTs, mainstream tokens BTC/ETH, established ecosystem POAPs)	□ 1 □ 2 □ 3 □ 4 □ 5	Long-term holding of high-value assets ≥ {*expert fill-in*} types
DID	Clear DID identity binding: Addr. explicitly bound to decentralized identity	Whether the addr. is explicitly bound to a DID (e.g., ENS, Lens)	□ 1 □ 2 □ 3 □ 4 □ 5	Explicit identity binding: □ Yes □ No Example: {*expert fill-in*}
NT	Natural trading behavior: Transaction intervals not exhibiting obvious scripting patterns	Maximum num. of consecutive transactions within a short period having identical or highly similar intervals (error <5%)	□ 1 □ 2 □ 3 □ 4 □ 5	Consecutive highly similar transactions ≤ {*expert fill-in*}

Note. Experts can add remarks on each code.