BLOODHOUND: Multi-Layer Behavioral Detection of Coordinated Inauthentic Behavior Without Content Analysis

Working Paper

Jack Roelofs
Watchfire AI
jack@jackroelofs.com

March 2026

Abstract

This paper presents BLOODHOUND, a multi-layer detection system for coordinated inauthentic behavior (CIB) on social media that operates primarily on behavioral metadata rather than content analysis. The system implements a layered detection cascade: Layer 1 (L1) applies sample entropy analysis to individual account posting timelines to identify mechanical regularity; Layer 2 (L2) computes pairwise Jensen-Shannon divergence across behavioral distributions and applies Leiden community detection to identify coordinated clusters; and Layer 3 (L3) uses content hash co-sharing and transfer entropy to recover operationally isolated accounts. Applied to the Internet Research Agency (IRA) dataset of 9,041,308 tweets from 3,836 accounts, the behavioral layers (L1 and L2) achieve a 93.7% detection rate among the 2,634 accounts with sufficient activity history for analysis, rising to 99.9% when content fingerprint matching is added. Critically, the system reveals a coverage gap: 1,202 accounts (31.3% of the known population) fall below the minimum activity threshold for behavioral profiling, receiving no analytical coverage despite constituting known members of the operation. This paper reports the methodology, results, negative findings, and all identified limitations, and proposes a two-phase synthesis architecture that explicitly separates behavioral assessment from content-dependent analysis. The behavioral-only capability is motivated by operational scenarios involving encrypted communications, where content may be unavailable but behavioral metadata remains observable.

Keywords: coordinated inauthentic behavior, information operations, sample entropy, Jensen-Shannon divergence, community detection, behavioral analysis, influence operations, bot detection

1. Introduction

The detection of coordinated inauthentic behavior (CIB) on social media platforms has become a critical challenge for platform integrity, national security, and democratic governance. State-sponsored influence operations, exemplified by the Russian Internet Research Agency's campaign targeting the 2016 United States presidential election, demonstrate that sophisticated actors can operate at scale while maintaining plausible deniability at the individual account level.

Existing detection approaches broadly fall into two categories: content-based methods that analyze what accounts say (narrative themes, linguistic patterns, sentiment coordination), and behavior-based methods that analyze how accounts act (posting frequency, temporal patterns, interaction networks). Content-based methods require access to plaintext message content, which may not always be available. Encrypted messaging platforms, deleted content, and privacy-restricted datasets all present scenarios where behavioral metadata—timestamps, action types, interaction targets—may be the only observable signal.

This observation motivates the central hypothesis of the BLOODHOUND project: *behavioral metadata alone, without access to message content, can detect a substantial proportion of coordinated inauthentic behavior, and the limits of this capability can be precisely characterized.* The corollary hypothesis is that content analysis, when available, provides incremental detection gains that can be quantified and separated from behavioral findings.

BLOODHOUND implements this hypothesis through a layered detection cascade in which each layer operates on increasingly complex behavioral signals. Layer 1 examines individual account temporal regularity via entropy measures. Layer 2 examines pairwise behavioral similarity and group coordination via distributional divergence and community detection. Layer 3 examines influence propagation through content fingerprint co-sharing and directed information transfer. A proposed Layer 4 would add semantic content analysis for narrative classification, explicitly separated from the behavioral layers.

The system is evaluated on the Internet Research Agency (IRA) dataset released by Twitter, Inc. in October 2018, comprising 9,041,308 tweets from 3,836 accounts active between May 2009 and June 2018. This dataset provides ground truth: every account is a confirmed participant in a state-sponsored influence operation, enabling measurement of detection rates (true positive rates) at each layer. However, the absence of a matched organic control dataset means that false positive rates cannot be estimated from this analysis. This limitation is explicitly acknowledged throughout.

This paper makes four contributions. First, it demonstrates that entropy-based behavioral analysis alone detects 79.5% of IRA accounts with sufficient activity history. Second, it shows that adding distributional coordination detection raises the behavioral ceiling to 93.7%. Third, it identifies and quantifies a coverage gap: 31.3% of known IRA accounts fall below the minimum activity threshold for behavioral profiling and receive no analytical coverage. Fourth, it proposes a two-phase synthesis architecture that separates behavioral assessment from content-dependent analysis, motivated by operational scenarios where content is unavailable.

2. Related Work

2.1 Entropy-Based Bot Detection

Gilmary et al. (2023) applied approximate entropy and sample entropy to Twitter account timelines for automated behavior detection, reporting an F1 score of 0.95 on their evaluation dataset. Their approach computes entropy over inter-tweet interval sequences, with lower entropy indicating more regular (and therefore more suspicious) posting patterns. The same research group extended this work using DNA-inspired profiling techniques (Gilmary et al., 2022), reporting an F1 of 0.95 on a binary classification task. The SampEn < 0.2 threshold used in BLOODHOUND's Layer 1 is adopted from this work, though its applicability to the IRA corpus specifically has not been independently validated.

Mazza et al. (2019) demonstrated that retweet temporal patterns alone, without content analysis, can achieve F1 = 0.87 for botnet detection (RTbust). This result is particularly relevant to BLOODHOUND's behavioral-only premise, as it establishes a precedent for content-free detection using temporal metadata.

2.2 Sample Entropy

Sample entropy (SampEn) was introduced by Richman and Moorman (2000) as a bias-corrected alternative to approximate entropy (ApEn) for physiological time series analysis. Unlike ApEn, SampEn excludes self-matches in template comparison, producing more consistent estimates for short time series. Richman and Moorman recommend a minimum of 200 data points for reliable SampEn computation with embedding dimension m = 2. BLOODHOUND adopts this threshold, which has consequences for population coverage discussed in Section 6.

2.3 Coordination Detection

Jensen-Shannon divergence (JSD), introduced by Lin (1991), provides a symmetric, bounded measure of distributional similarity suitable for comparing behavioral profiles. Unlike Kullback-Leibler divergence, JSD is always finite and symmetric, making it appropriate for pairwise similarity matrices.

The Leiden algorithm (Traag et al., 2019) improves upon Louvain community detection by guaranteeing well-connected communities. BLOODHOUND uses Leiden for decomposing the behavioral similarity graph into operationally meaningful clusters.

Normalized compression distance (NCD), based on the theoretical framework of Kolmogorov complexity (Cilibrasi & Vitanyi, 2005), measures sequence-level similarity by comparing compression ratios. While effective for detecting copy-paste bot behavior, NCD produced a negative result on the IRA dataset, as discussed in Section 5.3.

2.4 Transfer Entropy

Transfer entropy, introduced by Schreiber (2000), quantifies the directed information flow between two time series. It measures the reduction in uncertainty about one process's future state given knowledge of another process's past, beyond what the first process's own history provides. BLOODHOUND applies transfer entropy within detected communities to characterize directional influence patterns and identify account roles (broadcaster, amplifier, bridge).

3. Dataset

The evaluation dataset is the Internet Research Agency (IRA) hashed dataset released by Twitter, Inc. in October 2018.[1]

The dataset comprises two files. The tweets file contains 9,041,308 rows across 31 columns, totaling approximately 5.4 GB. Each row represents a single tweet with metadata including timestamp, user identifier, tweet text, action indicators (is_retweet, in_reply_to_tweetid), engagement counts, URLs, hashtags, and language. The users file contains 3,836 rows across 10 columns, listing all known IRA account identifiers with profile metadata.

All user identifiers in the dataset are SHA-256 hashed, preventing re-identification.[2] Timestamps are truncated to minute-level granularity (YYYY-MM-DD HH:MM), which introduces a known limitation for entropy analysis: events occurring within the same minute produce inter-tweet intervals of zero, creating artificial regularity in the time series. This limitation is mitigated by the choice of 15-minute binning for Shannon entropy (bin width exceeds granularity) and is documented as a factor affecting SampEn discriminative power.

3.1 Population Structure

Of the 3,836 accounts in the users file, 3,667 (95.6%) have at least one event in the tweets file. The remaining 169 accounts appear in the users file but have no associated tweets. The activity distribution is highly skewed:

Activity Level	Accounts	% of Total	Events	% of Events
≥200 events	2,634	68.7%	8,991,038	99.4%
50–199 events	397	10.3%	≈45,000	≈0.5%
1–49 events	636	16.6%	≈5,200	≈0.1%
0 events	169	4.4%	0	0%

[1] The dataset was released by Twitter, Inc. in October 2018 and is archived at https://archive.org/details/twitter-ira

[2] The hashed dataset uses SHA-256 for all user identifiers. Timestamps are truncated to minute-level granularity.

This distribution has a direct consequence for detection coverage: the 2,634 accounts with ≥200 events account for 99.4% of all observed activity but only 68.7% of the known population. The remaining 1,202 accounts, while collectively producing only 0.6% of events, are confirmed members of the influence operation and represent a coverage gap in any analysis that requires minimum activity thresholds.

3.2 Canonical Event Schema

All tweets are transformed into a platform-agnostic canonical event schema prior to analysis. Each event contains: a unique event identifier, an account identifier, a Unix epoch timestamp (milliseconds), an action type (one of six primitives: post_original, amplify, reply, quote, react, link_share), a SHA-256 content hash of normalized text, an optional target identifier, extracted URLs, and platform-specific raw metadata preserved for audit. The schema is implemented as an immutable Python dataclass with validation constraints. All analytical layers consume canonical events exclusively; no layer accesses raw platform data directly.

3.3 Action Type Distribution

The full corpus exhibits the following action type distribution across 9,041,308 events:

Action Type	Count	Percentage
Amplify (retweet)	3,333,184	36.9%
Post original	2,823,439	31.2%
Link share	2,567,851	28.4%
Reply	266,208	2.9%
Quote	50,626	0.6%

4. Methodology

BLOODHOUND implements a layered detection cascade in which each layer operates on a distinct class of behavioral signal. The layers are designed to be applied sequentially, with each subsequent layer targeting accounts that evade detection by preceding layers. The first two layers (L1 and L2) are purely behavioral, requiring only temporal metadata and action type classifications. Layer 3 (L3) introduces content fingerprints (hash-based) and temporal influence analysis. A proposed Layer 4 (L4) would add natural language content analysis, explicitly separated from the behavioral layers.

4.1 Layer 1: Entropy Anomaly Detection

Layer 1 examines individual account timelines for statistical regularity inconsistent with organic human behavior. Three entropy measures are computed for each account.

4.1.1 Shannon Entropy

For each account, the time-of-day distribution of events is computed by binning timestamps into 96 intervals of 15 minutes each across a 24-hour cycle. Shannon entropy H is computed over this distribution:

$$H = -\Sigma\, p_i \log_2(p_i)$$

where p_i is the proportion of events in bin i. The theoretical maximum for 96 bins is 6.58 bits (uniform distribution). Lower values indicate temporal concentration of activity. Shannon entropy captures daily rhythm patterns but does not capture sequential regularity.

4.1.2 Sample Entropy

Sample entropy (SampEn) is computed over the inter-tweet interval (ITI) sequence, following Richman and Moorman (2000). Given a sequence of ITI values, SampEn measures the conditional probability that sequences that match for m consecutive points will also match for m + 1 points, within a tolerance r. BLOODHOUND uses m = 2 and r = 0.2 × std(ITI). Lower SampEn indicates more predictable (more regular) posting patterns. A minimum of 200 events is required for reliable computation.[3]

For accounts with more than 5,000 events, the ITI sequence is subsampled to 5,000 points using deterministic stride sampling (every k-th interval) to maintain computational tractability while preserving temporal structure.

4.1.3 Approximate Entropy

Approximate entropy (ApEn) is computed with identical parameters for comparison with published baselines that report ApEn rather than SampEn. ApEn includes self-matches, producing an upward bias. It is retained for completeness but SampEn is the primary discriminant.

4.1.4 L1 Flagging

Accounts are flagged by L1 when SampEn falls below 0.2, following the threshold established by Gilmary et al. (2022).[4] Shannon entropy was found to be insufficiently discriminant as a standalone measure for this dataset: the IRA operation was sophisticated enough to distribute activity across the day, producing high Shannon entropy even in automated accounts. SampEn, which captures sequential regularity rather than distributional uniformity, proved to be the stronger signal.

[3] Richman and Moorman (2000) recommend a minimum of 200 data points for SampEn with embedding dimension m = 2.

[4] The Gilmary et al. (2022) threshold of SampEn < 0.2 was derived from a binary classification task on a different Twitter dataset. Its applicability to the IRA corpus is assumed but not independently validated.

4.2 Layer 2: Coordination Detection

Layer 2 examines pairwise behavioral similarity between accounts to identify coordinated clusters. Two measures were implemented; one produced a positive result and one a negative result.

4.2.1 Jensen-Shannon Divergence

For each account, a 480-dimensional behavioral distribution is constructed: 5 action types × 96 time-of-day bins. Each dimension represents the proportion of an account's activity of a given action type in a given 15-minute window. Laplace smoothing (1e-10) is applied to ensure well-defined divergence computation.

Pairwise JSD is computed using the entropy decomposition $JSD(P,Q) = H(M) - 0.5H(P) - 0.5H(Q)$, where $M = 0.5(P + Q)$. This vectorized approach processed 3,467,661 unique pairs across 2,634 accounts in 6.0 seconds.

4.2.2 Leiden Community Detection

A weighted similarity graph is constructed where edges connect account pairs with JSD below a configurable threshold, weighted by $(1 - JSD)$. The Leiden algorithm (Traag et al., 2019) is applied with modularity optimization to detect communities. At the selected threshold of $JSD < 0.15$, this produces 58 communities with 3 or more members and a modularity of 0.3379. An account is flagged by L2 if it belongs to any detected community.

4.2.3 Normalized Compression Distance (Negative Result)

NCD was implemented as a second independent coordination measure, following Cilibrasi and Vitanyi (2005). Account activity sequences were encoded as character strings (action type characters interleaved with time-gap delimiters) and compared using LZMA compression ratios.[5]

Result: NCD did not produce a useful coordination signal on the IRA dataset. All within-community NCD values exceeded 0.5, with large communities averaging approximately 0.92. The random baseline NCD was 0.93, making within-community pairs barely distinguishable from random. The planned NCD threshold of 0.3 was unreachable. This negative result is consistent with a sophisticated operation that shares distributional parameters (detectable by JSD) but uses independent posting schedules (invisible to NCD). NCD remains in the codebase for potential applicability to less sophisticated campaigns but is excluded from the active detection pipeline.

4.3 Layer 3: Influence Mapping

[5] NCD threshold of 0.3 originates from copy-paste bot detection literature (Cilibrasi & Vitanyi, 2005). The IRA operation is more sophisticated than the campaigns studied in that work.

Layer 3 targets the accounts that evade L1 and L2 detection. It employs two complementary methods, one behavioral and one content-based. The content-based component (content hash co-sharing) represents a departure from the purely behavioral approach of L1 and L2.

4.3.1 Content Hash Co-Sharing

Content propagation analysis identifies accounts that share identical content (by SHA-256 hash) with accounts already detected by L1 and L2. For each undetected account, the number of distinct detected accounts sharing at least one content hash is computed. An account is flagged by L3 when this count exceeds a configurable threshold (default: ≥3 distinct detected accounts).

This is explicitly a content-based signal, not a behavioral one. The content hash identifies *what* was posted, not *when* or *how*. The distinction matters for the two-phase architecture proposed in Section 7: content hash analysis belongs in the content phase, not the behavioral phase, even though it does not require access to plaintext content (only to content fingerprints).

4.3.2 Transfer Entropy

Transfer entropy TE(X→Y) is computed between account pairs within detected Leiden communities. Activity timelines are discretized into binary time series (active/inactive per hour-long bin), and TE is estimated from the joint probability distributions of current and lagged activity states:

$$TE(X \rightarrow Y) = \Sigma\, p(y_{t+1}, y_t, x_t) \cdot \log_2[p(y_{t+1}|y_t, x_t) \,/\, p(y_{t+1}|y_t)]$$

TE is computed per-community using community-specific time windows (the span of activity within each community), not a global window spanning the full 9-year dataset. This correction is essential: a global window would produce 76,000+ time bins, diluting TE to the noise floor. Community-specific windows concentrate the analysis on periods of actual coordinated activity.

Transfer entropy is a purely behavioral measure (it examines temporal activity patterns, not content). It provides characterization of influence roles within communities—broadcaster, amplifier, bridge, isolated—but did not contribute additional detection recovery beyond content hash co-sharing in this evaluation.

4.4 Behavioral Tiering

To facilitate analysis across the behavioral spectrum, the 2,634 L1-eligible accounts are stratified into five tiers based on SampEn percentile:

Tier	SampEn Range	Accounts	Description
T1	< 0.0180 (P10)	264	Blatant bot
T2	0.0180–0.0445 (P10–P25)	395	Likely automated

T3	0.0445–0.1690 (P25–P75)	1,316	Ambiguous
T4	0.1690–0.3047 (P75–P90)	395	Likely human-operated
T5	≥0.3047 (P90)	264	Near-organic

Tier boundaries are set by percentile (P10, P25, P75, P90) rather than fixed thresholds, allowing the tiering to adapt to the dataset's own distribution. T5 accounts, in particular, are individually indistinguishable from organic accounts on entropy measures alone: 247 of 264 exhibit both high Shannon entropy ($H > 5.0$) and high SampEn (≥0.2), indicating human-like daily activity patterns and unpredictable posting cadences.

5. Results

5.1 Layer 1: Entropy Detection

L1 analysis was performed on all 2,634 accounts meeting the ≥200 event threshold. At the SampEn < 0.2 threshold, L1 flags 2,095 accounts (79.5%). Detection is complete for T1 through T3 (100% flagged), partial for T4 (30.4%), and zero for T5 (0%). The T5 result is expected: these accounts were designed to appear human, and their entropy profiles are consistent with that design.

Shannon entropy alone was found to be insufficiently discriminant. The IRA operation distributed activity across the day, producing Shannon entropy values (median 5.55 bits, max 6.58) that overlap heavily with expected organic behavior. SampEn, which captures sequential regularity in the posting timeline rather than daily uniformity, proved far more effective.

5.2 Layer 2: Coordination Detection

L2 analysis was performed on the same 2,634 accounts. At JSD < 0.15, Leiden community detection identifies 58 communities encompassing 1,665 accounts. Combined with L1, the L1 ∪ L2 detection rate reaches 93.7% (2,468 of 2,634 accounts).

The top four communities exhibit distinct operational signatures:

Community	Size	Mean JSD	Mean SampEn	Character
0	480	0.122	0.132	Mixed operational
1	316	0.255	0.126	Mixed, dispersed
2	258	0.244	0.218	Human-operated
3	212	0.095	0.122	Tight automation

Community 3, characterized as "tight automation," has the lowest mean JSD (0.095), indicating highly similar behavioral distributions among its members. Community 2, with the highest mean SampEn (0.218), contains predominantly human-operated accounts that coordinate on distributional patterns rather than mechanical posting schedules. The remaining 54 smaller communities (3–47 members each) include several pure-T3 bot swarms.

L2's primary value is in recovering accounts that L1 misses. Of the 539 accounts not flagged by L1, L2 recovers 373 (69.2%) through community membership. The combined L1 ∪ L2 per-tier detection is:

Tier	L1	L2	L1 ∪ L2	Detection
T1 (264)	264	236	264	100.0%
T2 (395)	395	191	395	100.0%
T3 (1,316)	1,316	576	1,316	100.0%
T4 (395)	120	134	280	70.9%
T5 (264)	0	213	213	80.7%

Note: The L1 ∪ L2 values reported here use the Leiden community detection at JSD < 0.15, which produces the 93.7% combined rate (2,468/2,634). The 166 accounts missed by both L1 and L2 are predominantly low-volume T4 and T5 accounts: 115 T4 and 51 T5, with median event counts of 588 and 440 respectively and high mean inter-tweet intervals (2,006 minutes for missed T4). These accounts are behaviorally isolated—insufficient activity for strong distributional fingerprinting and not tightly coordinated with any detected cluster.

5.3 Layer 3: Influence Mapping

5.3.1 Content Hash Co-Sharing

Content propagation analysis was run on the 166 accounts missed by L1 ∪ L2. The primary finding: 165 of 166 missed accounts share at least one content hash with a detected account. At the ≥3 detected links threshold, 163 of 166 are recovered, bringing the cumulative L1 ∪ L2 ∪ L3 detection to 99.9% (2,631 of 2,634).

Threshold (≥N detected links)	Newly Flagged	Cumulative	Detection
≥1	165	2,633/2,634	100.0%
≥3	163	2,631/2,634	99.9%
≥5	161	2,629/2,634	99.8%
≥10	155	2,623/2,634	99.6%
≥50	146	2,614/2,634	99.2%
≥100	145	2,613/2,634	99.2%

The majority of recovered accounts share content with over 2,000 detected accounts, indicating deep integration with the IRA content ecosystem despite temporal and behavioral isolation. Three accounts remain undetected at all layers: one with zero content overlap (201 events, predominantly link_share), and two with only 2 detected content links (below the ≥3 threshold). All three are T5 accounts.

5.3.2 Transfer Entropy Within Communities

Transfer entropy was computed for the top four Leiden communities using 30-member subsamples (full pairwise computation on the 480-member community 0 would require 230,000+ pairs):

Community	Size	TE Mean (bits)	TE Max	Significant Pairs	Character
0	480	0.0076	0.0515	31.5%	Mixed operational
1	316	0.0017	0.0450	2.8%	Weakest, dispersed
2	258	0.0051	0.1062	14.4%	Human-operated, directional
3	212	0.0178	0.0415	78.3%	Tight automation

Community 3 (tight automation) exhibits the highest TE density: 78.3% of sampled account pairs show significant directional information transfer, consistent with centrally orchestrated posting schedules. Community 2 (human-operated) has the highest single-pair TE (0.106 bits), suggesting targeted influence chains between specific account pairs. The three remaining undetected accounts show no significant bridge TE with any community members, confirming their genuine operational isolation from the temporal coordination structure.

5.4 Detection Cascade Summary

The cumulative detection cascade across layers, measured against the 2,634 accounts with ≥200 events:

Layer	Method	Standalone	Cumulative	Rate
L1	SampEn < 0.2	2,095/2,634	2,095/2,634	79.5%
+ L2	Leiden JSD < 0.15	1,665/2,634	2,468/2,634	93.7%
+ L3	Content ≥3 links	163/166 missed	2,631/2,634	99.9%

The behavioral-only ceiling (L1 + L2) is 93.7%. The jump from 93.7% to 99.9% is achieved through content hash co-sharing, which is a content-based signal. This distinction is central to the interpretation of results: behavioral analysis alone, applied to the eligible population, detects approximately 94% of accounts. The remaining 6% requires content-level signals for recovery.

6. Coverage Gap Analysis

The detection rates reported in Section 5 are computed against a denominator of 2,634 accounts—those meeting the ≥200 event threshold for SampEn eligibility. This denominator excludes 1,202 accounts (31.3% of the 3,836 known IRA accounts) that fall below this threshold.[6]

Measured against the full population of accounts with any activity (3,667), the actual detection rates are substantially lower:

Metric	vs. Eligible (2,634)	vs. Active (3,667)	vs. Total (3,836)
L1 detected	79.5%	57.1%	54.6%
L1 + L2 detected	93.7%	67.3%	64.3%
L1 + L2 + L3 detected	99.9%	71.7%	68.6%
Never analyzed	0	1,033 (28.2%)	1,202 (31.3%)

The coverage gap arises because every analytical layer inherits L1's 200-event threshold. This inheritance was not a deliberate analytical decision. L2's JSD computation does not require 200 events—a behavioral distribution can be constructed from 50 or fewer events, though with increased noise. Content hash co-sharing has no minimum event threshold: an account with a single tweet sharing a content hash with detected accounts produces signal. The 200-event gate was set for SampEn reliability and was applied uniformly to all layers without independent evaluation of each layer's minimum requirements.

The practical consequence is modest in event volume (the 1,202 excluded accounts produced only 50,270 events, 0.6% of the corpus) but significant in population coverage. In a real-world deployment, low-activity accounts may represent sleeper accounts, seed accounts, or accounts performing specific single-use functions within a broader operation. Excluding them from analysis by default creates a blind spot that sophisticated operators could exploit.

6.1 Proposed Coverage Expansion

The proposed remediation is to decouple layer eligibility thresholds:

Layer	Current Threshold	Proposed Threshold	Population
L1 (SampEn)	≥200 events	≥200 events (unchanged)	2,634
L2 (JSD + Leiden)	≥200 events	≥50 events	3,031 (+397)
L3 (Content prop.)	≥200 events	≥1 event	3,667 (+1,033)
TE	Within Leiden only	Within expanded Leiden	Expands with L2

[6] This 200-event threshold was set for SampEn statistical reliability and was not independently evaluated for JSD or content propagation methods.

This expansion has not been implemented in the current evaluation. The detection rates reported in Section 5 reflect the original 2,634-account population only. The expanded analysis is proposed as future work to be validated in a production deployment.

7. Proposed Architecture: Two-Phase Synthesis

The results motivate a two-phase analysis and reporting architecture that explicitly separates behavioral findings from content-dependent analysis. This separation is driven by an operational premise: in scenarios involving encrypted communications, end-to-end encrypted platforms, or datasets where content has been redacted, behavioral metadata (timestamps, action types, interaction graphs) may be the only available signal. A system designed to produce useful output from behavioral metadata alone has broader applicability than one that assumes content availability.

7.1 Behavioral Analysis Phase

The behavioral phase encompasses L1 (entropy anomaly detection), L2 (JSD coordination and Leiden community detection), and transfer entropy analysis within detected communities. No content is read or analyzed. The phase produces:

Per-account assessments translating raw scores into structured findings (e.g., “This account exhibits mechanical posting regularity with SampEn in the 5th percentile, consistent with scheduling automation. It belongs to a 212-member coordination cluster characterized by tight temporal synchronization”). Per-community characterizations describing operational signatures, tier composition, and TE-derived influence structure. An operation-level summary with the population coverage funnel, detection rates against honest denominators, and explicit statements about what cannot be determined from behavior alone.

The behavioral phase ceiling on the IRA dataset is 93.7% of the eligible population (2,468 of 2,634 accounts with ≥200 events). This ceiling is a finding, not a failure. It means that approximately 6% of accounts in a sophisticated operation are designed to be individually indistinguishable from organic users and not closely coordinated with detectable clusters. A behavioral report should communicate this clearly.

7.2 Content Analysis Phase

The content phase is optional and is only offered when content data (at minimum, content fingerprints; ideally, plaintext) is available. It encompasses two sub-components:

Content fingerprint analysis (content hash co-sharing, URL co-sharing) identifies accounts that distribute the same content as detected accounts. This requires content hashes but not plaintext. On the IRA dataset, this recovers 163 of the 166 accounts missed by behavioral analysis, raising the detection rate from 93.7% to 99.9% within the eligible population.

Semantic content analysis, powered by large language model classification, categorizes account content into narrative themes, identifies thematic coordination within detected communities, and provides human-interpretable descriptions of what a campaign is about (as opposed to the behavioral phase, which describes that a campaign exists). This component requires an API integration with a language model and is specified separately for cloud deployment.

7.3 User-Facing Flow

The intended user experience is: (1) the user provides an input file or data stream; (2) BLOODHOUND runs the behavioral analysis phase; (3) the system presents behavioral synthesis results with explicit coverage and methodology disclosures; (4) the system offers to proceed with content analysis if content data is available; (5) if the user opts in, the content phase runs and presents incremental findings over the behavioral baseline. At each synthesis point, the user receives a structured report that identifies what was analyzed, what was found, what the confidence level is, and what the method cannot determine.

8. Assumptions and Limitations

The following assumptions underlie this work. Each represents a potential point of failure if violated.

Assumption 1: Ground truth validity. All 3,836 accounts in the IRA dataset are treated as confirmed influence operation participants. This assumes Twitter's attribution is correct and complete. If the dataset contains false positives (legitimate accounts misattributed to the IRA) or false negatives (IRA accounts not included), the detection rates reported here are affected.

Assumption 2: SampEn < 0.2 threshold transferability. The flagging threshold is adopted from Gilmary et al. (2022), who derived it on a different Twitter dataset. Its optimality for the IRA corpus is assumed, not validated. A proper ROC analysis requires organic control data, which is not available in this evaluation.

Assumption 3: Minimum event thresholds. The 200-event minimum for SampEn follows Richman and Moorman (2000). The extension of this threshold to L2 and L3 was an implementation convenience, not an analytically justified decision. The proposed threshold reduction for L2 (≥50 events) has not been empirically validated.

Assumption 4: Behavioral stationarity. Entropy measures are computed over the full account lifetime (up to 9 years of activity). This assumes behavioral patterns are sufficiently stationary to produce meaningful aggregate statistics. Accounts that changed operational modes (e.g., from automated to human-operated) may exhibit entropy values that average across distinct behavioral regimes.

Assumption 5: Platform specificity. BLOODHOUND was developed and evaluated exclusively on Twitter data. The canonical event schema is designed for cross-platform applicability, but the entropy thresholds, JSD features, and detection rates are specific to Twitter's interaction model and the IRA's operational patterns. Transferability to other platforms (Telegram, Reddit, Bluesky) is not established.

The following limitations are known:

No false positive estimation. All detection rates are true positive rates computed against the known IO population. Without matched organic control data, precision, recall, F1, and false positive rates cannot be computed. A method that flags 93.7% of IO accounts is useful only if it does not also flag a comparable proportion of organic accounts.

Timestamp granularity. Minute-level timestamps introduce artificial regularity: events occurring within the same minute produce zero-second ITI values. This affects SampEn computation for high-frequency accounts. The magnitude of this effect has not been precisely quantified.

Subsampling effects. Transfer entropy computation within large communities uses 30-member subsamples due to computational constraints. TE statistics for communities with 200+ members are estimates based on these subsamples, not exact values for the full community.

Single-operation evaluation. Results are derived from a single influence operation (Russian IRA). The behavioral signatures, detection rates, and layer-specific contributions may differ substantially for operations by other state actors, commercial influence firms, or grassroots coordination campaigns.

Coverage gap. As detailed in Section 6, 31.3% of known IO accounts receive no analytical coverage due to the 200-event minimum threshold. The proposed coverage expansion has not been implemented or validated.

9. Conclusion

BLOODHOUND demonstrates that behavioral metadata analysis, without content, detects a substantial majority of accounts in a sophisticated state-sponsored influence operation. The behavioral ceiling of 93.7% against the eligible population establishes an empirical bound on what temporal and distributional analysis can achieve against a well-resourced adversary. The remaining accounts—designed to be operationally indistinguishable from organic users—require content-level signals for detection, a finding that is itself analytically valuable for understanding the limits of metadata-based surveillance.

The identification of a 31.3% coverage gap, arising from an inherited activity threshold that was never independently evaluated for downstream layers, illustrates the importance of explicitly auditing analytical assumptions at each stage of a detection pipeline. Reporting detection rates against an artificially favorable denominator—while technically accurate—obscures the system's actual coverage of the target population.

The proposed two-phase synthesis architecture, separating behavioral assessment from content analysis, is motivated by the practical reality that content is not always available. Encrypted platforms, privacy regulations, and data retention limitations all create scenarios where behavioral metadata is the primary or sole signal. A detection system that produces structured, honest, and useful output under these constraints has broader operational applicability than one that assumes full content access.

References

Cilibrasi, R., & Vitanyi, P. M. B. (2005). Clustering by compression. IEEE Transactions on Information Theory, 51(4), 1523–1545. https://doi.org/10.1109/TIT.2005.844059

Gilmary, R., Venkatesan, A., & Vaiyapuri, G. (2022). DNA-influenced automated behavior detection on Twitter through relative entropy. Scientific Reports, 12, 8022. https://doi.org/10.1038/s41598-022-11854-w

Gilmary, R., Venkatesan, A., & Vaiyapuri, G. (2023). Detection of automated behavior on Twitter through approximate entropy and sample entropy. Personal and Ubiquitous Computing, 27, 91–105. https://doi.org/10.1007/s00779-021-01647-9

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151. https://doi.org/10.1109/18.61115

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting temporal patterns for botnet detection on Twitter. In Proceedings of the 11th ACM Conference on Web Science (pp. 183–192). ACM. https://doi.org/10.1145/3292522.3326015

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology–Heart and Circulatory Physiology, 278(6), H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039

Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, 85(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. Scientific Reports, 9, 5233. https://doi.org/10.1038/s41598-019-41695-z

Twitter, Inc. (2018). Information operations dataset: Internet Research Agency [Data set]. Internet Archive. https://archive.org/details/twitter-ira