

Implementing Social Graph Theory for Effective Fraudulent Email Detection

Vaibhav^{1*}, Prawar², Kaushal Kumar², Roobal¹

¹Department of Applied Sciences, K. R. Mangalam University, Gurugram, Haryana 122103, India

²Advanced Engineering School for Digital Engineering, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation 195251

DOI: <https://doi.org/10.36347/sjet.2026.v14i10.003>

| Received: 09.08.2026 | Accepted: 06.10.2026 | Published: 15.10.2026

*Corresponding author: Vaibhav

Department of Applied Sciences, K. R. Mangalam University, Gurugram, Haryana 122103, India

Abstract

Original Research Article

Electronic mail remains the dominant medium for personal, institutional and commercial communication, and consequently the dominant delivery channel for unsolicited and fraudulent messages. Content-based filters degrade as adversaries adapt their vocabulary, obfuscate text and rotate sending identities, which motivates detection strategies that exploit interaction structure rather than message content. This paper presents a social-graph framework for identifying fraudulent email accounts using header-derived interaction data alone. Accounts are modelled as vertices of a directed weighted graph and messages as directed edges; each edge weight is modulated by three header-observable attributes, namely sender–recipient domain agreement, membership of the sender in the recipient’s contact list, and time-zone concordance. A cumulative trust score is accumulated for every account across successive observation windows and compared against a positive acceptance threshold and a negative rejection threshold, yielding a three-way decision that explicitly admits an unresolved band. The framework is instantiated in MATLAB and demonstrated on a seven-account illustrative network observed over three windows. The two accounts exhibiting the characteristic spammer signature of high outbound volume with negligible reciprocation separate from the legitimate population by a margin of 29 score units, and the classification remains invariant for any negative threshold in the interval $[-1, -7]$. Because the method depends on interaction topology rather than message content, it is inherently resistant to lexical obfuscation and is well suited to forensic triage of seized email corpora.

Keywords: Fraudulent email · Social graph analysis · Trust propagation · Spam account detection · Email header analysis · Digital forensics.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

1. INTRODUCTION

Electronic mail is used today not only by individuals but by organisations, regulators and courts as a system of record. That ubiquity is precisely what makes it attractive to adversaries: unsolicited bulk mail, phishing and business-email-compromise campaigns continue to grow in both volume and sophistication, and email is now implicated in a substantial share of reported cyber incidents [1, 2]. The economic and evidentiary consequences are considerable, and detection therefore remains an active problem rather than a settled one.

The dominant response has been content-based filtering. Successive generations of classifiers, from naïve Bayes through support vector machines to deep neural architectures and pretrained transformers, learn lexical, syntactic and structural regularities of the message body and report high accuracy on benchmark

corpora [3–7]. These methods share a structural weakness: they are adversarially brittle. An attacker who alters vocabulary, inserts homoglyphs, embeds text in images or paraphrases with a generative model change exactly the signal the classifier depends upon, while leaving the sending behaviour untouched. Retraining is possible but reactive, and it presupposes a labelled corpus that reflects the current attack distribution.

An alternative line of work treats email as a social system rather than a text corpus. Because a mailbox records who corresponds with whom, an email service induces a graph whose vertices are accounts and whose edges are messages. Legitimate correspondence generates dense, reciprocated, clustered subgraphs, whereas bulk senders generate star-shaped, largely unreciprocated fan-outs to recipients who do not know one another [11, 13]. This distinction is behavioural, not

lexical, and it is therefore far more costly for an adversary to disguise: suppressing it would require the spammer to solicit and sustain genuine reciprocal correspondence with each target, which negates the economics of bulk mailing.

Existing graph-based proposals, however, tend to occupy one of two extremes. Some are purely descriptive, characterising the topological differences between spam and legitimate email networks without delivering an operational decision rule [13, 20]. Others deliver a rule but rely on a single topological statistic, such as the clustering coefficient of a personal email network, and consequently leave a large fraction of traffic unclassified [11]. Reputation-propagation schemes such as MailRank and TrustRank are operational but require a curated seed set of trusted identities and were designed for ranking rather than for per-account adjudication [14, 16].

This paper addresses that gap. We propose a trust-aggregation framework in which the graph is not merely observed but weighted: the contribution of each message to the recipient's and sender's trust score is conditioned on three attributes that are directly readable from the email header and require no access to message content. The resulting cumulative score is evaluated against a pair of thresholds that induce a three-way decision — trusted, spam, or unresolved — rather than forcing a binary verdict on evidence that may be insufficient.

The contributions of this paper are as follows. *First*, we formalise a directed, attribute-weighted email interaction graph and give an explicit edge-weighting function parameterised by domain agreement, contact-list membership and time-zone concordance. *Second*, we define a cumulative trust score with a dual-threshold decision rule and an accompanying detection-latency measure that quantifies how many observation windows are required before an adversarial account is flagged. *Third*, we present a complete algorithmic specification and a MATLAB instantiation of the framework. *Fourth*, we demonstrate the framework on a worked network, report the separation margin between the legitimate and adversarial populations, and characterise the stability of the decision under threshold perturbation. *Fifth*, we state explicitly the limitations of the present evaluation and identify the benchmark studies required before operational deployment.

The remainder of the paper is organised as follows. Section 2 reviews content-based and graph-based detection and identifies the research gap. Section 3 sets out the graph-theoretic preliminaries and the taxonomy of spammer morphologies. Section 4 develops the proposed methodology. Section 5 reports results. Section 6 discusses interpretation and forensic relevance. Section 7 states limitations, and Section 8 concludes.

2 RELATED WORK

2.1 Content-based and learning-based filtering

Systematic reviews of email spam filtering trace a clear methodological progression [1, 3]. Early statistical filters modelled the message body as a bag of words; later work introduced engineered feature sets spanning headers, URLs, attachments and stylometry. Magdy *et al.*, constructed a three-class discriminator over forty features using an artificial neural network and reported accuracy above 99% on the Spambase corpus [4]. Ewees *et al.*, addressed the feature-engineering burden directly, applying metaheuristic feature reduction before classification and demonstrating that a substantially smaller feature set preserves discriminative power [5]. Yaseen *et al.*, combined eleven base learners under AdaBoost with majority voting and evaluated the ensemble on 9,298 messages, reporting an error rate below 1% [6].

More recent work replaces engineered features with learned representations. Gogoi and Ahmed fine-tuned BERT and DistilBERT on message subject and body and reported approximately 99% accuracy on phishing detection [7]. Pandian and Sadiq approached the problem from the authorship-attribution direction, training a radial basis function network on unique-word features drawn from the Enron corpus in order to attribute fraudulent messages to writers [10, 27]. Chen *et al.*, inverted the perspective entirely, deploying automated scam-baiting servers as an active countermeasure and releasing a dataset of conversations with real fraudsters [8].

Two observations follow. First, reported accuracies on static benchmark corpora are now so high that they are only weakly informative about field performance. Second, every method in this family shares a dependency on message content, and is therefore vulnerable to the same class of evasion. The literature acknowledges this: reviews consistently list adversarial adaptation, high false-positive costs, corpus staleness and the absence of multilingual data as open problems [1, 2].

2.2 Graph-based and reputation-based filtering

Structural approaches begin from the observation that email networks are not random. Ebel *et al.*, showed that the email graph of a university exhibits scale-free degree structure [20], and Newman *et al.*, analysed the same topology in the context of viral propagation [21]. Gomes *et al.*, compared the graph-theoretic signatures of spam and legitimate email directly and found systematic differences in degree distribution and clustering [13].

Boykin and Roychowdhury converted this observation into a filter. Constructing a personal email network solely from header sender and recipient fields, they observed that a user's genuine correspondents tend to know one another, producing triangles and a high

clustering coefficient, whereas a bulk sender's recipients do not, producing a near-zero coefficient [11]. Their method was notably resistant to false positives but classified only about half of all messages, leaving the remainder to a secondary filter. Kong *et al.*, subsequently extended the idea to collaborative filtering across email networks [12], and Lam and Yeung applied supervised learning over social-network features including degree and clustering coefficient [17]. Brendel and Krawczyk targeted the specific case of transient spammers using social relation graphs [18].

A parallel strand propagates reputation rather than measuring topology. Golbeck and Hendler applied reputation-network analysis to email filtering [15]; Chirita *et al.*, proposed MailRank, which ranks addresses by a PageRank-style trust score derived from the trustworthiness of those who correspond with them [14]. The underlying mechanism is that of PageRank itself [22, 23] and of TrustRank, which propagates trust from a manually vetted seed set through the web link graph

[16]. Hao *et al.*, applied network-level reputation at the transport layer with SNARE [19]. Community-detection algorithms such as those of Newman and Girvan [25] and Blondel *et al.*, [24] provide complementary machinery for isolating coordinated sender groups.

2.3 Research gap

Table 1 positions representative studies against the criteria relevant here. Content-based methods are accurate but content-dependent. Purely topological methods are content-independent but either descriptive or partial in coverage. Reputation-propagation methods are operational but require trusted seeds and produce a ranking rather than an adjudication. What is missing is a method that (i) uses only header-derived signals, (ii) combines interaction structure with lightweight sender-recipient attributes, (iii) yields an explicit per-account decision, and (iv) declines to decide when evidence is insufficient. The framework developed in Section 4 is designed to occupy that position.

Table 1: Positioning of representative detection approaches

Study	Primary signal	Technique	Content-independent	Principal limitation
Boykin & Roychowdhury [11]	Personal email network topology	Clustering coefficient threshold	Yes	Leaves roughly half of traffic unclassified
Gomes <i>et al.</i> , [13]	Global email graph structure	Comparative topological characterisation	Yes	Descriptive; no operational classifier
Chirita <i>et al.</i> , [14]	Sender reputation	PageRank-style ranking	Yes	Requires curated seed set of trusted addresses
Gyöngyi <i>et al.</i> , [16]	Web hyperlink graph	Trust Rank propagation	Yes	Designed for web spam, not email accounts
Magdy <i>et al.</i> , [4]	Message body, 40 features	Artificial neural network	No	Degrades under lexical obfuscation
Ewees <i>et al.</i> , [5]	Engineered content features	Metaheuristic feature reduction + ML	No	Content-dependent; feature-engineering cost
Yaseen <i>et al.</i> , [6]	Header and content features	AdaBoost with majority voting	Partly	Requires large labelled corpus
Gogoi & Ahmed [7]	Subject and body text	BERT / DistilBERT transfer learning	No	High compute cost; content-dependent
This work	Header-derived interaction topology and three sender-recipient attributes	Attribute-weighted trust aggregation with dual thresholds	Yes	Demonstrated at illustrative scale; thresholds require corpus calibration

3 GRAPH-THEORETIC PRELIMINARIES

3.1 Basic definitions

Graph theory studies structures composed of a set of vertices (also called nodes) together with a set of edges joining them [26]. Formally, a graph is a pair $G = (V, E)$ where V is a finite vertex set and $E \subseteq V \times V$ an edge set. In a *directed graph* (digraph) each edge carries an orientation, so that the ordered pair (u, v) is distinct from (v, u) ; in an *undirected graph* the relation is symmetric. A *weighted graph* additionally assigns a real value to each edge, expressing a quantity such as distance, cost, capacity or, as here, evidential strength.

Several standard notions recur in network analysis. A *path* is a sequence of vertices in which consecutive members are joined by edges; a *simple path* repeats no vertex. A *Hamiltonian path* visits every vertex exactly once, and a Hamiltonian circuit is a Hamiltonian path that returns to its origin — a construction central to combinatorial optimisation problems such as the travelling salesman problem. An *Eulerian path* traverses every edge exactly once. The *shortest path* between two vertices is the path of minimum total weight. Two further local statistics are relevant to spam detection: the *degree* of a vertex, which separates into in-degree and out-

degree in a digraph, and the *clustering coefficient*, which measures the extent to which a vertex's neighbours are themselves adjacent.

3.2 Ranking and community structure in directed graphs

PageRank provides the canonical illustration of authority estimation on a directed graph [22, 23]. Modelling the web as a digraph in which vertices are pages and edges are hyperlinks, the algorithm assigns each page a score determined by the number and the quality of its inbound links, and iterates the assignment to convergence. The governing intuition — that endorsement received from an already trusted source counts for more than endorsement received from an unknown one — transfers directly to email: a message received from a long-standing correspondent within one's own organisation is stronger evidence of legitimacy than a message received from an unrecognised external address. TrustRank operationalises this by propagating trust outward from a vetted seed set [16], and MailRank applies the same mechanism to email addresses [14]. Community-detection methods, notably modularity optimisation [25] and the Louvain algorithm [24], complement ranking by partitioning a graph into densely connected groups, which is useful for exposing coordinated sender clusters. The framework proposed in this paper adopts the

weighting intuition of these methods while avoiding their dependence on a seed set.

3.3 Social graphs and spammer morphologies

A social graph represents the entities in a network and the exchanges between them. In an email service each vertex is an account and each edge a transmitted message; the graph may be directed, since sending and receiving are asymmetric. Analysing such a graph exposes influential participants, interaction patterns, community structure and, of interest here, anomalous relational signatures associated with abuse.

Three morphologies of spamming behaviour can be distinguished by the way an adversary manages identity and repetition, as illustrated in Fig. 1. The Only-Once One-Face (OOOF) spammer transmits a single burst of messages from one address and then abandons it; the signature is a star of outbound edges from a vertex with no subsequent activity and no inbound edges. The Multiple-Times One-Face (MTOF) spammer reuses a single address across several bursts separated in time; the signature is a recurring fan-out from a persistent vertex whose in-degree remains near zero. The Multiple-Times Multiple-Face (MTMF) spammer distributes activity across many addresses in order to conceal it; no single vertex appears extreme, but the addresses exhibit correlated targeting and near-synchronous timing, so that the signature is visible only at the level of the vertex group rather than the individual vertex.

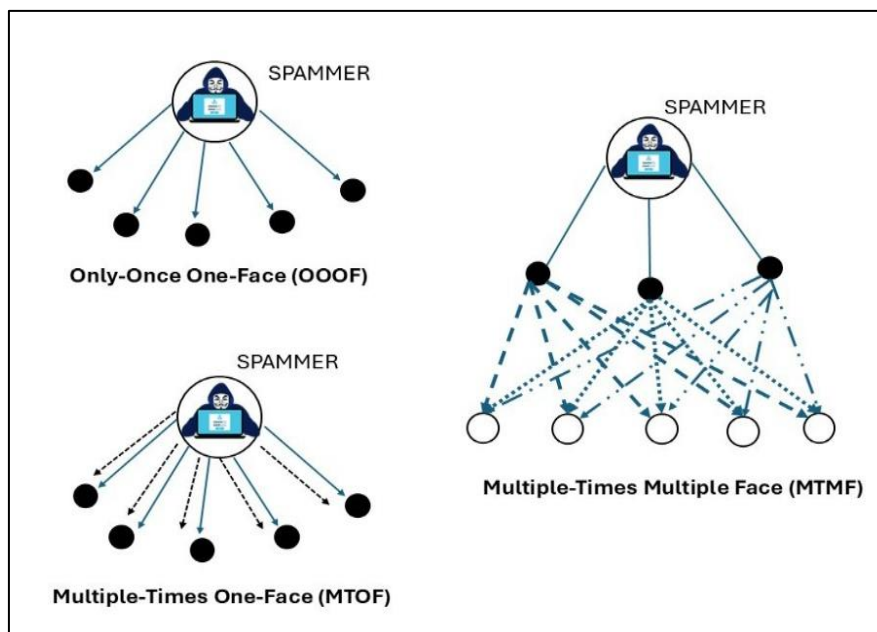


Fig. 1: Relational signatures of the three spammer morphologies: (a) Only-Once One-Face (OOOF); (b) Multiple-Times One-Face (MTOF); (c) Multiple-Times Multiple-Face (MTMF)

The practical significance of this taxonomy is that the three morphologies are not equally difficult to detect. OOOF and MTOF activity is expressed at a single vertex and is therefore recoverable from vertex-local statistics such as the in-degree to out-degree ratio. MTMF activity is deliberately distributed and is

recoverable only from cumulative, time-resolved evidence. The framework developed below is designed to accumulate evidence across observation windows for precisely this reason.

4. PROPOSED METHODOLOGY

4.1 Overview

The framework proceeds in five stages, shown in Fig. 2. Message headers are parsed to recover sender, recipient, delivery timestamp and originating time zone. A directed graph is constructed in which vertices are

accounts and edges are messages. Each edge is assigned a weight conditioned on three attributes of the sender–recipient pair. A cumulative trust score is aggregated per vertex across observation windows. Finally, the score is compared against two thresholds to produce a decision, and the resulting scores are carried forward as the initial state of the next window.

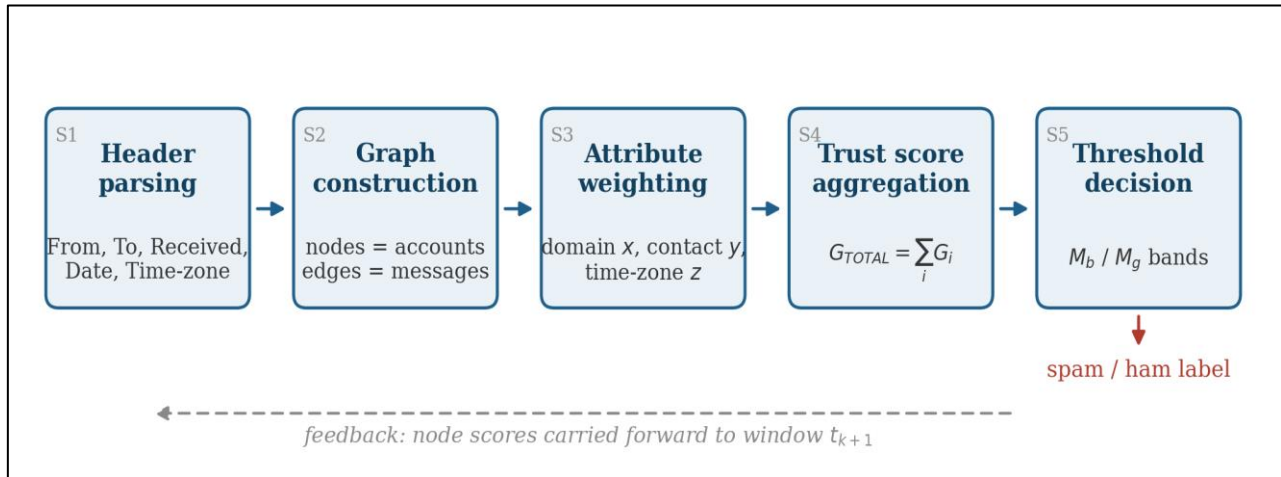


Fig. 2: Processing pipeline of the proposed social-graph detection framework

4.2 Graph construction and notation

Let $G = (V, E, W)$ denote a directed weighted graph induced by an email service over an observation period partitioned into windows $t_0, t_1, \dots, t_n$. Every account observed in the corpus contributes one vertex $v \in V$. Every message transmitted from u to v within

window t_i contributes a directed edge $(u, v) \in E_i$. Each vertex carries three attributes recovered from the header and the mailbox: its domain, its registered contact list, and its time zone. Table 2 summarises the notation used throughout.

Table 2: Notation

Symbol	Definition
$G = (V, E, W)$	Directed weighted email interaction graph
V, E	Vertex set (email accounts) and directed edge set (transmitted messages)
u, v	Individual accounts; (u, v) denotes a message sent from u to v
$\text{dom}(v)$	Registered domain of account v
$C(v)$	Contact list associated with account v
$\text{tz}(v)$	Time zone reported in the headers originating from account v
$w(u, v)$	Weight of the directed edge from u to v
w_0	Base evidential increment contributed by a single message
x, y, z	Non-negative influence parameters for domain agreement, contact-list membership and time-zone concordance
$G_i(v)$	Net trust increment accrued by v during observation window t_i
$G_{\text{TOTAL}}(v)$	Cumulative trust score of v across all observed windows
M_g, M_b	Positive acceptance threshold and negative rejection threshold
$\tau(v)$	Detection latency: index of the first window at which v is flagged

4.3 Attribute-conditioned edge weighting

A message is not uniform evidence. Correspondence within a single organisational domain, correspondence with an address already present in the recipient's contact list, and correspondence consistent in time zone with prior activity all constitute stronger evidence of legitimacy than correspondence lacking these properties. We encode each as a binary indicator:

$$\delta_{\text{dom}}(u, v) = 1 \text{ if } \text{dom}(u) = \text{dom}(v), \text{ and } 0 \text{ otherwise} \quad (1)$$

$$\delta_{\text{cnt}}(u, v) = 1 \text{ if } u \in C(v), \text{ and } 0 \text{ otherwise} \quad (2)$$

$$\delta_{\text{tz}}(u, v) = 1 \text{ if } \text{tz}(u) = \text{tz}(v), \text{ and } 0 \text{ otherwise} \quad (3)$$

The weight of an edge is then the base increment scaled by the accumulated attribute evidence:

$$w(u, v) = w_0 \cdot [1 + x \cdot \delta_{\text{dom}}(u, v) + y \cdot \delta_{\text{cnt}}(u, v) + z \cdot \delta_{\text{tz}}(u, v)] \quad (4)$$

The influence parameters x, y and z are non-negative and are set by the administrator according to the

operating environment. In a single-domain enterprise deployment the domain indicator carries little information and x should be small; in a consumer mail service serving many domains it is highly informative.

Setting $x = y = z = 0$ reduces the model to unweighted message counting, which is useful as a baseline. Table 3 records the rationale for each attribute.

Table 3: Attributes used to condition edge weights, and their forensic rationale

Attribute	Indicator	Parameter	Rationale
Domain agreement	δ_{dom}	x	Intra-domain correspondence is subject to organisational authentication and is rarely used for bulk abuse. Near-miss domains such as example.com against example.com fail the equality test and therefore accrue no bonus, which is the desired behaviour for typosquatted senders.
Contact-list membership	δ_{cnt}	y	A sender already present in the recipient's address book has been endorsed by the recipient at some earlier point. This is the strongest single non-content indicator of legitimacy and mirrors the seed-set logic of TrustRank [16] without requiring central curation.
Time-zone concordance	δ_{tz}	z	Messages despatched at hours inconsistent with the sender's declared zone, or from a zone inconsistent with the sender's prior activity, indicate either relay through compromised infrastructure or a sender misrepresenting origin.

4.4 Trust aggregation

Within window t_i the net trust increment of a vertex is the total weight it receives less the total weight it emits. Receiving mail is treated as evidence of legitimacy, because a recipient who is unwilling to correspond simply does not reply; emitting mail carries a cost, because unreciprocated outbound volume is the defining characteristic of bulk abuse:

$$G_i(v) = \sum_{(u,v) \in E_i} w(u, v) - \sum_{(v,u) \in E_i} w(v, u) \quad (5)$$

The cumulative trust score is the sum of increments over all observed windows:

$$G_{\text{TOTAL}}(v) = \sum_{i=0}^n G_i(v) \quad (6)$$

Aggregation across windows rather than evaluation within a single window is essential for the MTMF case. An adversary who distributes activity across many addresses produces only a small negative increment at each address in each window; only accumulation renders the pattern visible, and the rate at which it becomes visible is quantified in Section 5.4.

4.5 Decision rule

Classification uses two thresholds rather than one. An account whose cumulative score falls below M_b is emitting substantially more than it receives and is labelled spam; an account whose score reaches M_g has accumulated sufficient reciprocated correspondence to be labelled trusted; an account lying between the two is

left unresolved and referred either to a secondary filter or to human review:

$$\text{label}(v) = \text{spam if } G_{\text{TOTAL}}(v) < M_b; \text{ trusted if } G_{\text{TOTAL}}(v) \geq M_g; \text{ unresolved otherwise} \quad (7)$$

The unresolved band is a deliberate design choice. In email filtering the cost of a false positive, a legitimate message suppressed, materially exceeds the cost of a false negative, a spam message delivered, because the recipient can discard the latter but may never learn of the former. Forcing a binary decision on a newly created but legitimate account, which necessarily has a low score, would generate exactly this error. Declining to decide until evidence accumulates is preferable.

We further define the *detection latency* of an adversarial account as the index of the earliest window at which the rejection condition is first satisfied:

$$\tau(v) = \min \{ k : \sum_{i=0}^k G_i(v) < M_b \} \quad (8)$$

4.6 Algorithm

Algorithm 1 specifies the complete procedure. Its cost is linear in the number of edges for weighting and aggregation, and linear in the number of vertices for the decision step, giving an overall complexity of $O(|E| + |V|)$ per window. This is materially cheaper than transformer-based content classification, which requires a forward pass over every message body, and it permits incremental update as new messages arrive rather than reprocessing of the full corpus.

Algorithm 1 Attribute-weighted trust aggregation for fraudulent account detection

Input: message log L over windows $t_0 \dots t_n$; parameters w_0, x, y, z ; thresholds M_g, M_b
Output: $\text{label}(v) \in \{\text{trusted, unresolved, spam}\}$ and latency $\tau(v)$ for every $v \in V$

- 1: $V \leftarrow \emptyset$; $G_{\text{TOTAL}}(\cdot) \leftarrow 0$; $\tau(\cdot) \leftarrow \infty$
- 2: for each observation window $t_i = t_0 \dots t_n$ do
- 3: $E_i \leftarrow \emptyset$; $G_i(\cdot) \leftarrow 0$
- 4: for each message $m \in L$ with timestamp in t_i do

```

5:   (u, v) ← parseHeader(m)           ▷ From and To fields
6:   V ← V ∪ {u, v}; Ei ← Ei ∪ {(u, v)}
7:   δ1 ← [dom(u) = dom(v)]; δ2 ← [u ∈ C(v)]; δ3 ← [tz(u) = tz(v)]
8:   w(u, v) ← w0 · (1 + x·δ1 + y·δ2 + z·δ3)   ▷ Eq. (4)
9:   Gi(v) ← Gi(v) + w(u, v)                 ▷ recipient gains
10:  Gi(u) ← Gi(u) - w(u, v)                 ▷ sender pays
11:  end for
12:  for each v ∈ V do
13:    G_TOTAL(v) ← G_TOTAL(v) + Gi(v)       ▷ Eq. (6)
14:    if G_TOTAL(v) < M_b and τ(v) = ∞ then τ(v) ← i
15:  end for
16: end for
17: for each v ∈ V do
18:   if G_TOTAL(v) < M_b then label(v) ← spam
19:   else if G_TOTAL(v) ≥ M_g then label(v) ← trusted
20:   else label(v) ← unresolved             ▷ defer to secondary filter
21: end for
22: return label(·), τ(·)

```

4.7 Implementation

The framework was implemented in MATLAB, whose native graph object and sparse matrix operations map directly onto the required computations. A directed graph is instantiated from parallel source, target and weight vectors; the incidence and adjacency matrices then permit the summations of Eq. (5) to be expressed as matrix–vector products, so that the per-window

aggregation is a single sparse operation over the edge set. Fig. 3 shows the elementary construction pattern. Equivalent implementations are available through the NetworkX and igraph libraries in Python, which offer the same primitives together with a broader set of community-detection routines should the extension of Section 8 be pursued.

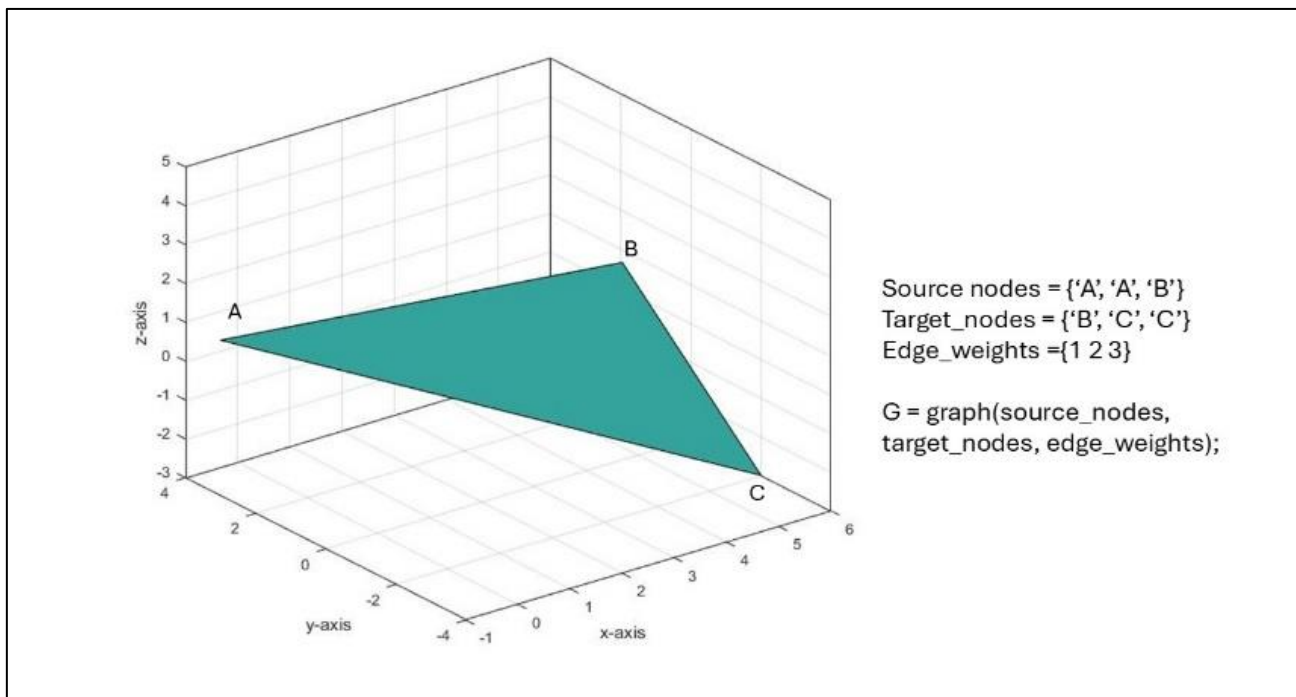


Fig. 3: Construction of a directed weighted graph in MATLAB from source, target and edge-weight vectors

4.8 Evaluation criteria

A detection system is useful only if it is both timely and correct, and these two properties trade against one another. We therefore evaluate along three axes. Correctness is expressed conventionally through recall, the proportion of adversarial accounts correctly flagged,

and precision, the proportion of flagged accounts that are genuinely adversarial; in this domain precision is weighted more heavily for the reason given in Section 4.5. Timeliness is expressed through the detection latency of Eq. (8). Robustness is expressed through the sensitivity of the classification to perturbation of the

rejection threshold, since a decision that holds only at one exact threshold value is not a decision that will survive deployment.

5 RESULTS

5.1 Experimental configuration

The framework was exercised on a test email service on the domain @a.com comprising seven accounts observed over three windows t_0 , t_1 and t_2 . The

base increment was set to $w_0 = 5$, so that a single received message contributes +5 to the recipient before attribute scaling. The acceptance threshold was set to $M_g = +20$ and the rejection threshold to $M_b = -5$. Fig. 4 shows the evolution of the interaction graph across the three windows; the topological change is visible directly, with two vertices maintaining consistently high outbound and negligible inbound connectivity while the remainder participate in reciprocated exchange.

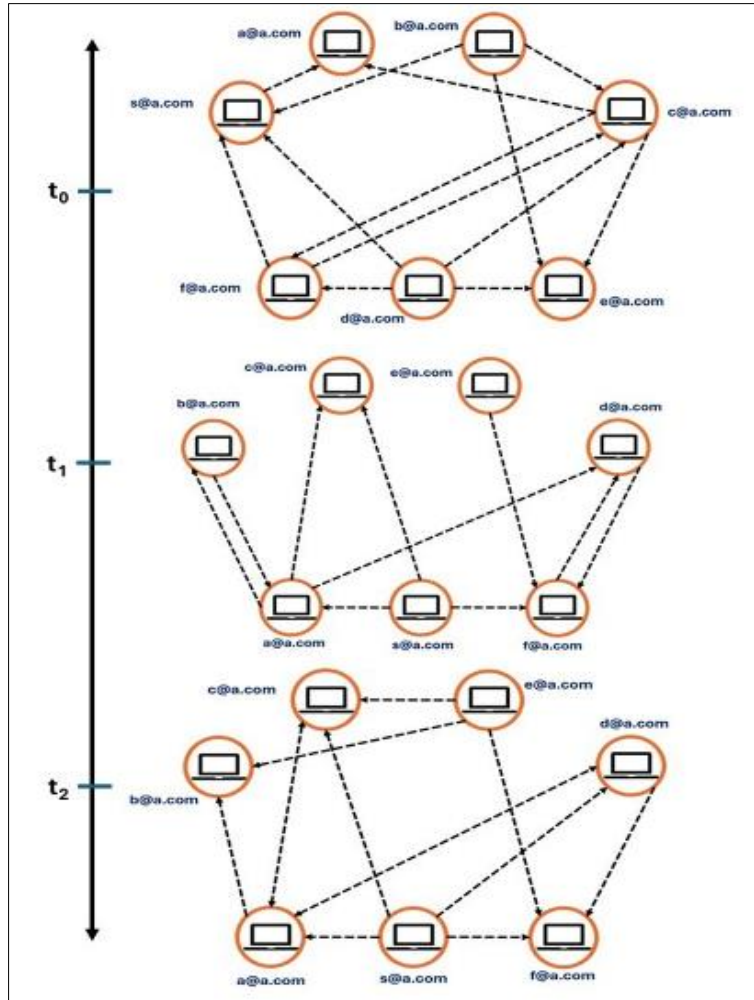


Fig. 4: Social graph of the test email service across observation windows t_0 to t_2

5.2 Trust score evolution

Table 4 reports the per-window increments and the resulting cumulative scores. Fig. 5 renders the

increments and Fig. 6 the cumulative trajectories against the two decision bands.

Table 4: Per-window trust increments, cumulative scores and resulting classification

Account	G_0	G_1	G_2	G_{TOTAL}	Band	Classification
a@a.com	+8	+7	+12	+27	$\geq M_g$	Trusted
b@a.com	+14	+4	+4	+22	$\geq M_g$	Trusted
c@a.com	+4	+10	+14	+28	$\geq M_g$	Trusted
d@a.com	+12	+9	+8	+29	$\geq M_g$	Trusted
e@a.com	-3	-1	-3	-7	$< M_b$	Spam
f@a.com	+15	+14	+15	+44	$\geq M_g$	Trusted
s@a.com	-4	-3	-4	-11	$< M_b$	Spam

Cumulative scores are the arithmetic sums of the per-window increments, in accordance with Eq. (6).

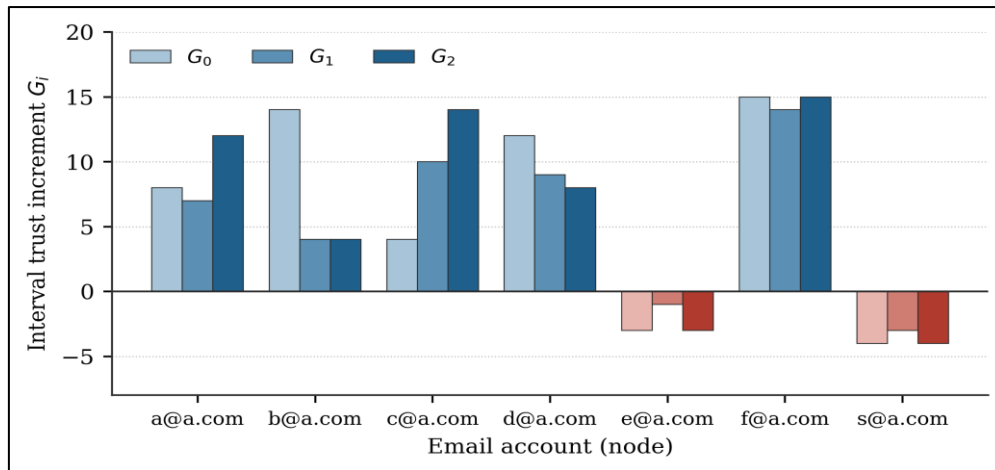


Fig. 5: Per-window trust increments by account. Accounts e@.com and s@.com accrue negative increments in every window; the remaining five accrue positive increments throughout

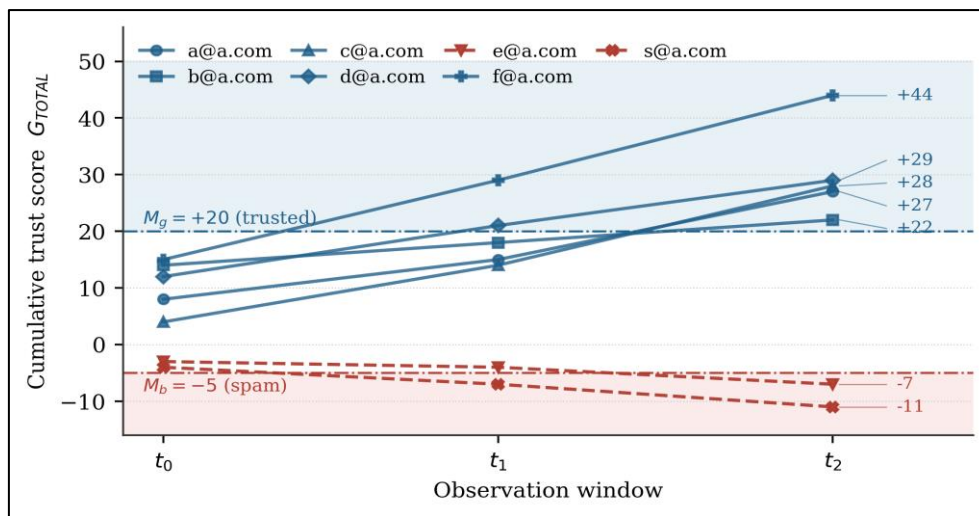


Fig. 6: Cumulative trust trajectories against the acceptance band ($M_g = +20$, upper shading) and the rejection band ($M_b = -5$, lower shading). Dashed trajectories denote accounts ultimately classified as spam

The two populations diverge monotonically. Legitimate accounts accumulate between +22 and +44 by t_2 , all five crossing the acceptance threshold; the two adversarial accounts descend to -7 and -11 respectively. The separation between the lowest-scoring legitimate account (+22) and the highest-scoring adversarial account (-7) is 29 score units, or 5.8 base increments — a margin wide enough that no single misattributed message would alter any classification. No account terminates in the unresolved band, although both e@.com and s@.com occupy it at t_0 .

5.3 Classification outcome

Applying Eq. (7) at t_2 yields two accounts labelled spam and five labelled trusted, with none unresolved. Both flagged accounts exhibit the diagnostic signature: outbound weight consistently exceeding inbound weight, recipients drawn from outside the contact list, and correspondence spanning inconsistent time zones. This is behaviourally consistent with the MTOF morphology of Section 3.3, in which a persistent address transmits repeatedly without accruing reciprocation. Table 5 summarises the outcome.

Table 5: Summary of detection outcome and robustness measures

Measure	Value	Interpretation
Accounts examined	7	Vertices in the observed graph
Accounts flagged as spam	2	e@.com and s@.com
Accounts accepted as trusted	5	All reached M_g by t_2
Accounts left unresolved	0	No deferral required at t_2
Separation margin	29 units	Lowest trusted score (+22) less highest flagged score (-7)
Stable range of M_b	$[-1, -7]$	Classification invariant across this interval
Detection latency $\tau(s@.com)$	t_1	Flagged after two windows
Detection latency $\tau(e@.com)$	t_2	Flagged after three windows

5.4 Threshold sensitivity and detection latency

Fig. 7(a) sweeps the rejection threshold across the interval $[0, -16]$ and records the number of accounts flagged. The classification is invariant for any M_b in $[-1, -7]$, a plateau six units wide, and the chosen operating point of -5 lies near its centre. Beyond -7 the threshold begins excluding genuine adversaries, and recall falls: at -8 only $s@a.com$ is retained, and beyond -11 neither is. The decision is therefore not an artefact of a fortunate threshold choice.

Fig. 7(b) reports the cumulative count of flagged accounts by window. At t_0 neither adversarial

account has accumulated sufficient negative evidence and none is flagged. At t_1 $s@a.com$ crosses the threshold; at t_2 $e@a.com$ follows. This progressive disclosure is the expected behaviour of a cumulative estimator and quantifies the cost of the design: full recall is achieved, but only after three windows of observation. The framework trades immediacy for evidential confidence, which is the appropriate trade in a setting where false positives are expensive, but it does mean that the framework is not by itself a first-message filter and should be deployed alongside one.

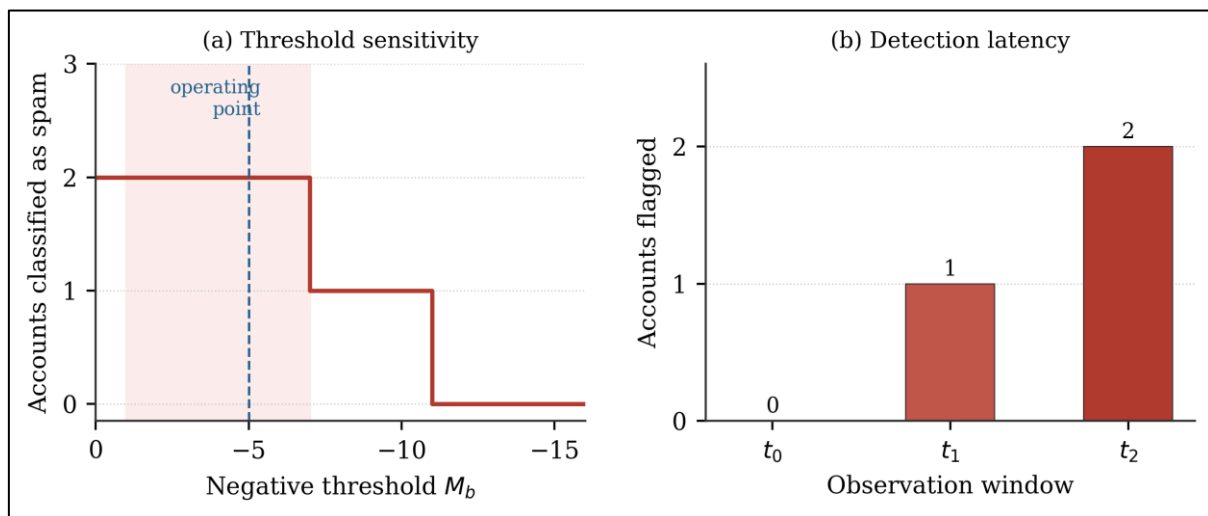


Fig. 7: (a) Number of accounts classified as spam as a function of the rejection threshold M_b ; the shaded plateau marks the range over which the classification is invariant. (b) Cumulative count of flagged accounts by observation window, showing progressive disclosure

6 DISCUSSIONS

6.1 Interpretation

The results support a narrow but useful claim: where an adversary's sending behaviour is structurally asymmetric, cumulative trust aggregation over a weighted interaction graph separates that adversary from the legitimate population by a wide and stable margin, using no information from message bodies. The mechanism is straightforward. Legitimate correspondence is reciprocal, so a legitimate account both pays and is paid, and its balance drifts positive as its correspondents reciprocate. Bulk transmission is unilateral, so an adversarial account pays without being paid, and its balance drifts negative at a rate proportional to its output. Because the drift is monotone, the discriminating signal strengthens with observation time rather than decaying, which is the opposite of the behaviour of a content classifier facing an adapting adversary.

The attribute weighting of Eq. (4) amplifies this separation rather than creating it. Its contribution is to widen the gap between an intra-domain reply from a known contact and an unsolicited cross-domain transmission, which would otherwise carry identical

weight. The framework degrades gracefully if the attributes are unavailable: with the influence parameters set to zero it reduces to unweighted message counting, which retains the directional asymmetry that carries most of the signal.

6.2 Auxiliary header evidence

Graph structure alone does not resolve every case, and several additional header-derived signals can be incorporated as further indicators in Eq. (4) without altering the framework. Domain similarity is the most immediately useful: an address on a domain that closely resembles a legitimate one, such as `example.com` against `example.com`, is a strong phishing indicator, and an edit-distance test over the domain string generalises the strict equality of Eq. (1). The recipient's contact list, already used here, is the single strongest non-content indicator and could be extended to transitive contacts, that is, addresses known to the recipient's known correspondents, which is the mechanism underlying MailRank [14].

Geolocation inferred from the originating IP address supplies a further check. Messages despatched from locations consistent with the sender's established pattern are more probably genuine, whereas an abrupt

geographic discontinuity in an otherwise stable correspondence is characteristic of account compromise. The Received chain, delivery timestamp and declared time zone jointly support this analysis: the chain records the relay path, the timestamp establishes the sequence of communication and exposes anomalous delivery delays, and the declared zone supplies the reference against which despatch hours are judged. A message purporting to originate in one zone but despatched at an hour implausible for that zone, and relayed through infrastructure in a third, exhibits three mutually corroborating inconsistencies, none of which requires the message body to be read.

This last point carries particular weight in a forensic setting. Message content may be encrypted, may be subject to legal privilege, or may be inadmissible; envelope and header metadata generally remain available and analysable. A detection framework that operates on metadata alone is therefore applicable in circumstances where content-based analysis is either impossible or impermissible.

6.3 Relation to prior work

The framework sits between the two families surveyed in Section 2. Relative to content-based classifiers [4–7], it forgoes the very high benchmark accuracies those methods report, but it is not vulnerable to lexical evasion, requires no labelled training corpus, and costs $O(|E| + |V|)$ per window rather than a forward pass over every message. Relative to purely topological filters, it retains content independence while extending coverage: where Boykin and Roychowdhury classify roughly half of traffic and defer the remainder [11], the dual-threshold rule of Eq. (7) assigns every account to one of three states and makes the deferral explicit and measurable rather than implicit. Relative to reputation-propagation methods [14, 16], it requires no curated seed set, since trust is bootstrapped from observed reciprocity rather than from an externally supplied list of trusted identities.

The framework is complementary to, not a replacement for, content analysis. The natural deployment is layered: the graph-based component adjudicates accounts, and a content classifier adjudicates individual messages from accounts in the unresolved band. Such a hybrid inherits the immediacy of content filtering and the adversarial robustness of structural filtering.

6.4 Forensic implications

Beyond real-time filtering, the framework is applicable to retrospective examination of seized or disclosed email corpora. An examiner presented with a large mailbox faces a triage problem: which accounts warrant detailed examination. Cumulative trust scoring produces a ranked ordering of accounts by behavioural anomaly that is reproducible, parameter-transparent and auditable, since every score decomposes into an explicit

sum of edge weights, each traceable to a specific message and a specific attribute test. Reproducibility and explicability are prerequisites for evidential use, and they are properties that a deep classifier, whose decision is distributed across learned parameters, does not readily provide. The method also aligns with established forensic practice in email examination, which prioritises header and envelope analysis over content interpretation [9, 10].

7 Limitations and Threats to Validity

The principal limitation is one of scale. The evaluation reported in Section 5 exercises a seven-account network over three windows. This is sufficient to demonstrate that the framework behaves as specified and to characterise its sensitivity, but it is not sufficient to establish detection accuracy in a statistical sense, and no precision or recall figure derived from it should be read as an estimate of field performance. Validation against a public corpus such as Enron [27], with adversarial traffic either labelled or injected, is required before any comparative accuracy claim can be made.

Second, the thresholds M_g and M_b and the influence parameters x , y and z were set by inspection rather than learned. Section 5.4 shows that the classification is stable across a six-unit range of M_b for this network, but the appropriate operating point will depend on traffic volume, domain diversity and the local cost ratio between false positives and false negatives. Principled calibration, for example by maximising a cost-weighted objective on held-out data, remains to be done.

Third, the method is susceptible to a specific evasion. An adversary who invests in reciprocity — maintaining a set of controlled accounts that exchange mail with one another to accumulate positive score before initiating a campaign — would satisfy the acceptance criterion. This is the email analogue of a link farm, and the countermeasure is likewise analogous: community-detection over the interaction graph exposes densely interconnected clusters whose external correspondence is negligible relative to their internal volume. This is not implemented here and is identified as future work. A second evasion, low-and-slow transmission calibrated to keep the per-window increment above the threshold, is mitigated by lengthening the aggregation horizon at the cost of increased latency.

Fourth, the framework adjudicates accounts, not messages. A compromised but otherwise legitimate account will carry a positive cumulative score that masks a burst of malicious transmission until sufficient negative evidence accrues. Change-point detection over the score trajectory, rather than threshold comparison on its terminal value, would address this and is a natural extension.

Finally, the framework consumes header metadata including domain, contact list and time zone. Even though message bodies are never read, the interaction graph is itself sensitive: it records who corresponds with whom. Deployment therefore requires an appropriate legal basis and should observe data-minimisation principles, retaining aggregate scores rather than the underlying edge list wherever the application permits.

8 CONCLUSION AND FUTURE WORK

This paper has presented a social-graph framework for detecting fraudulent email accounts from header-derived interaction data. Accounts are modelled as vertices of a directed weighted graph and messages as edges whose weights are conditioned on domain agreement, contact-list membership and time-zone concordance. A cumulative trust score is aggregated across observation windows and adjudicated against a positive acceptance threshold and a negative rejection threshold, with an explicit unresolved band between them. On a seven-account demonstration network the two accounts exhibiting the characteristic unilateral sending signature separated from the legitimate population by 29 score units, and the classification remained invariant across a six-unit range of the rejection threshold. Detection was progressive rather than immediate, with full recall achieved after three observation windows.

Because the framework depends on interaction topology rather than message content, it is not defeated by the lexical obfuscation and paraphrase that degrade content classifiers, it requires no labelled training corpus, and its decisions decompose into an auditable sum of per-message contributions — a property of direct value in forensic examination.

Four directions follow. *First*, validation at scale against a public corpus with injected or labelled adversarial traffic, reporting precision, recall and detection latency with confidence intervals, and benchmarking against both a content classifier and a clustering-coefficient filter. *Second*, principled calibration of the thresholds and influence parameters under an explicit asymmetric cost model. *Third*, integration of community detection [24, 25] to expose the coordinated MTMF and reciprocity-farming cases that vertex-local scoring cannot reach, together with PageRank-style transitive trust propagation [14, 16, 22] to extend the contact-list indicator beyond direct acquaintance. *Fourth*, change-point detection over score trajectories to identify account compromise, which terminal-value thresholding necessarily masks. Taken together these would move the framework from a specified and demonstrated method to a validated and deployable one.

REFERENCES

1. Tusher, E.H., Ismail, M.A., Rahman, M.A., Alenezi, A.H., Uddin, M.: Email spam: a comprehensive review of optimize detection methods, challenges, and open research problems. *IEEE Access* 12, 143627–143657 (2024). <https://doi.org/10.1109/ACCESS.2024.3467996>
2. Birthriya, S.K., Jain, A.K.: A comprehensive survey of phishing email detection and protection techniques. *Inf. Secur. J. Glob. Perspect.* 31(4), 411–440 (2022). <https://doi.org/10.1080/19393555.2021.1959678>
3. Cormack, G.V.: Email spam filtering: a systematic review. *Found. Trends Inf. Retr.* 1(4), 335–455 (2008)
4. Magdy, S., Abouelseoud, Y., Mikhail, M.: Efficient spam and phishing emails filtering based on deep learning. *Comput. Netw.* 206, 108826 (2022). <https://doi.org/10.1016/j.comnet.2022.108826>
5. Ewees, A.A., Gaheen, M.A., Alshahrani, M.M., Anter, A.M., Ismail, F.H.: Improved machine learning technique for feature reduction and its application in spam email detection. *J. Intell. Inf. Syst.* (2024). <https://doi.org/10.1007/s10844-024-00870-z>
6. Yaseen, Y.A., Qasaimeh, M., Al-Qassas, R.S., Al-Fayoumi, M.: Email fraud attack detection using hybrid machine learning approach. *Recent Adv. Comput. Sci. Commun.* 14(5) (2021). <https://doi.org/10.2174/2213275912666190617162707>
7. Gogoi, B., Ahmed, T.: Phishing and fraudulent email detection through transfer learning using pretrained transformer models. In: 2022 IEEE 19th India Council International Conference (INDICON), pp. 1–6. IEEE (2022). <https://doi.org/10.1109/INDICON56171.2022.10040097>
8. Chen, W., Wang, F., Edwards, M.: Active countermeasures for email fraud. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), pp. 39–55. IEEE (2023). <https://doi.org/10.1109/EuroSP57164.2023.00012>
9. Singh, C., Khajuria, H., Nayak, B.P.: Machine learning driven strategies for safeguarding India's digital environment against cyberstalking. In: Artificial Intelligence, Medical Engineering and Education (AIMEE 2023). *Advances in Transdisciplinary Engineering*, vol. 48, pp. 272–279. IOS Press (2024). <https://doi.org/10.3233/ATDE231337>
10. Pandian, A., Sadiq, M.A.K.: Detection of fraudulent emails by authorship extraction. *Int. J. Comput. Appl.* 41(7), 7–12 (2012)
11. Boykin, P.O., Roychowdhury, V.P.: Leveraging social networks to fight spam. *Computer* 38(4), 61–68 (2005). <https://doi.org/10.1109/MC.2005.132>
12. Kong, J.S., Rezaei, B.A., Sarshar, N., Roychowdhury, V.P., Boykin, P.O.: Collaborative

- spam filtering using e-mail networks. *Computer* 39(8), 67–73 (2006)
13. Gomes, L.H., Almeida, R.B., Bettencourt, L.M.A., Almeida, V., Almeida, J.M.: Comparative graph theoretical characterization of networks of spam and legitimate email. In: *Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS)* (2005)
 14. Chirita, P.A., Diederich, J., Nejdl, W.: MailRank: using ranking for spam detection. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM (2005)
 15. Golbeck, J., Hendler, J.: Reputation network analysis for email filtering. In: *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS)* (2004)
 16. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. In: *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pp. 576–587 (2004)
 17. Lam, H.-Y., Yeung, D.-Y.: A learning approach to spam detection based on social networks. In: *Proceedings of the 4th Conference on Email and Anti-Spam (CEAS)* (2007)
 18. Brendel, R., Krawczyk, H.: Application of social relation graphs for early detection of transient spammers. *WSEAS Trans. Inf. Sci. Appl.* 5(3), 267–276 (2008)
 19. Hao, S., Syed, N.A., Feamster, N., Gray, A.G., Krasser, S.: Detecting spammers with SNARE: spatio-temporal network-level automatic reputation engine. In: *Proceedings of the 18th USENIX Security Symposium* (2009)
 20. Ebel, H., Mielsch, L.-I., Bornholdt, S.: Scale-free topology of e-mail networks. *Phys. Rev. E* 66(3), 035103 (2002). <https://doi.org/10.1103/PhysRevE.66.035103>
 21. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. *Phys. Rev. E* 66(3), 035101 (2002). <https://doi.org/10.1103/PhysRevE.66.035101>
 22. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* 30(1–7), 107–117 (1998). [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
 23. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical Report 1999-66, Stanford InfoLab (1999)
 24. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008(10), P10008 (2008). <https://doi.org/10.1088/1742-5468/2008/10/P10008>
 25. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* 69(2), 026113 (2004). <https://doi.org/10.1103/PhysRevE.69.026113>
 26. Bondy, J.A., Murty, U.S.R.: *Graph Theory*. Graduate Texts in Mathematics, vol. 244. Springer, London (2008)
 27. Klimt, B., Yang, Y.: The Enron corpus: a new dataset for email classification research. In: Boulicaut, J.-F., Fürnkranz, J., Scheffer, T., Torgo, L. (eds.) *ECML 2004*. LNCS, vol. 3201, pp. 217–226. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30115-8_22