Temporal Semantic Proximity Analysis: A Method for Detecting Coordinated Information Operations in Multilingual Social Media Data

Cristian-Teodor Băbălău
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—Coordinated information operations (IO) on social media are most commonly detected through platform-dependent behavioral traces such as shared retweets, URLs, or hashtags. These signals, however, vary across platforms and across the differing playbooks of state actors, and they overlook the core objective of an operation: propagating a narrative. We argue that greater emphasis should be placed on *what* is communicated and *when*, rather than on *how* users interact. To this end, we introduce Temporal Semantic Proximity Analysis (TSPA), a language-independent, platform-agnostic method that links accounts solely through the semantic and temporal proximity of their posts, using multilingual sentence embeddings, sliding-window cosine similarity, and Leiden community detection. We evaluate TSPA on 14 campaigns, drawn from a public labeled dataset, spanning diverse languages, durations, and IO concentration. Our method consistently separates IO-IO pairs, yielding high-modularity communities with IO purities of up to 100%, and, on several campaigns, matches or exceeds every individual behavioral trace. Our results show that temporal semantic proximity is an effective, cross-platform backbone for surfacing coordinated communities, complementing rather than replacing other approaches.**

I. Introduction

In the recent decades, social media has become the primary channel for people looking to keep abreast of global and local events, making it a potentially powerful tool for actors seeking to influence public beliefs or propagate narratives.

The problem is structural: unlike traditional media, social platforms allow content to be published without any prior verification by an impartial side. Furthermore, a considerable share of false and unverified stories travels faster and wider than true ones. A study tracking roughly 126,000 stories shared on Twitter between 2006 and 2017 found that false news was 70% more likely to be retweeted than true stories [1]. A 2023 UNESCO-Ipsos survey across 16 countries found that 87% of respondents believe disinformation has already had a significant impact on their country's political life [2].

A growing range of state and non-state-backed actors exploit this environment, with state-aligned operators being the most studied. By the end of 2022, Meta reported having disrupted more than 200 global influence networks originating in 68 countries and operating in at least 42 languages, with Russia, Iran, and Mexico appearing as the most active sources of disrupted operations [3]. The Oxford Internet Institute's 2019 Global Inventory of Organised Social Media Manipulation found evidence of coordinated manipulation campaigns in 70 countries [4].

A 2025 Meta takedown surfaced a network targeting Romania's electoral process that used fake personas to manipulate local political discourse across Facebook, YouTube, X, and TikTok [5]. Social media has also proven to be a highly effective space for spreading conspiracy theories.

The QAnon movement grew to a network of Facebook groups whose membership exceeded three million users [6]. Twitter had decided to act against more than 150,000 accounts it identified as malicious in a single enforcement wave [7]. Networks of accounts have also been used to push health misinformation related to COVID-19 [8].

All of the above suggests that staying aware of who initiates a narrative's discussions is an important task in our current age. Tools that can identify and flag coordinated inauthentic behavior must be developed and deployed to keep social media a transparent and safe space.

A. Research Questions

In this thesis we address the following research questions:

1) How does a sliding-window semantic proximity analysis perform as a coordination detection signal across IO campaigns originating from different countries?
2) What are the results of a clustering method applied on a user-to-user network built of temporal textual similarity proximity?
3) How effectively does temporal semantic proximity analysis detect coordinated information operations when combined with behavioral interactions?

II. Related Work

State-sponsored information operations on social media have been extensively documented [9], [10], with studies increasingly focusing on the detection of *coordinated* behavior

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Computer Science, Maastricht University. Supervisor(s): Dr. Adriana Iamnitchi

as a key signal [11]–[13]. A growing body of work try to identify malicious coordinated networks or users using platform-specific behavioral traces: co-retweets, co-URLs, hashtag sequences, and fast-retweets [12]–[14].

Nizzoli et al. [11] proposed a non-binary coordination framework using multiscale network backbone extraction and Louvain community detection, highlighting continuous coordination signals on the 2019 UK General Election dataset. Luceri et al. [12] showed that eigenvector-centrality based node pruning, combined with a fused similarity network, achieves precision up to 0.95 across six verified IO campaigns. Alizadeh et al. [15] demonstrated that platform-agnostic content features alone leave a temporally stable, discriminative signal across campaigns from multiple countries, motivating text-first approaches such as ours.

For the semantic representation of posts, Reimers and Gurevych [16] introduced Sentence-BERT, encoding sentences into a 768-dimensional space suited to cosine-similarity comparison. A multilingual extension of this embedding tool [17] aligns this space across 50+ languages via knowledge distillation.

For community detection, Traag et al. [18] identified a structural defect in the widely used Louvain algorithm [19], whose instability on Twitter retweet networks has itself been characterised by Evkoski et al. [20], and proposed Leiden, which guarantees well-connected partitions. A direct comparison on Twitter data from 2022 related to the war in Ukraine [21] reports more cohesive communities under Leiden than under Louvain.

III. Data

In this work we rely on the public dataset introduced by Seckin et al. [22], composed of *tweets* from the platform X (formerly Twitter) that are associated with information operations across multiple countries. The dataset spans 26 campaigns attributed by the platform to 16 distinct state actors, ranging from state governments (e.g., Russia, China, Iran) to sub-national political movements (e.g., Catalonia).

Alongside the platform-labelled IO participants of each campaign, the authors of the dataset present a topically and temporally matched set of *control* tweets: posts by accounts that discussed comparable topics in the same time periods but were not flagged as part of any operation. The IO timelines are released in full, while control timelines are capped at 100 posts per account and truncated at the date the account first met the inclusion criteria. The presence of this matched control set enables an objective evaluation of the methods introduced and reproduced in this work.

IV. Temporal Semantic Proximity Analysis

A. Motivation and Rationale

This study proposes a new approach to IO analysis, identification, and clustering, emphasizing semantic features of textual data rather than relying on platform-dependent behavioral traces (e.g., Retweet, Share, Mention). A motive backing up our proposed approach is covered by the idea that the objective

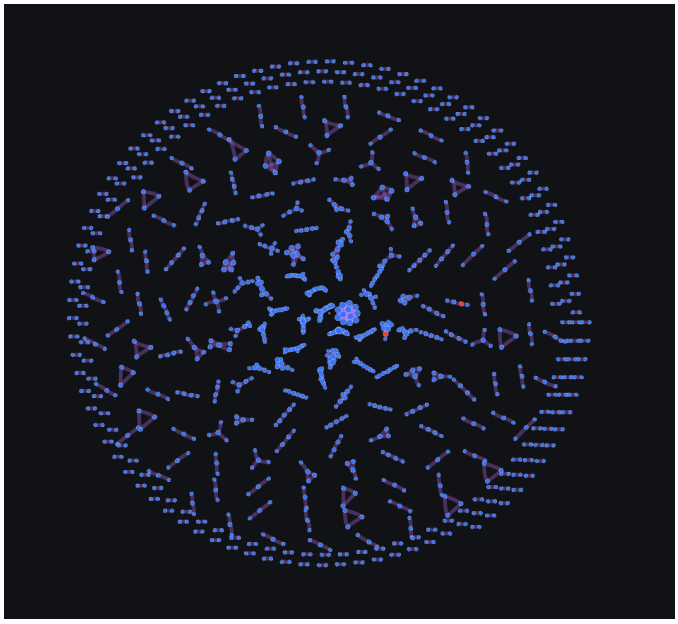

Fig. 1. Co-retweet network for the *Venezuela_1* campaign, showing connections between accounts that repeatedly retweet the same content. The count of IO accounts in this graph is 2 among 2,054 vertices.

of an information operation is to propagate a narrative or idea, without necessarily leaving distinguishable patterns of platform specific interaction. As a consequence, we consider that greater emphasis should be placed on the questions of **"What"** is communicated and **"When,"** rather than **"How"** users post.

Multiple studies and investigations, suggest that campaigns originating in different countries, have different playbooks with distinct operational signatures. [23].

Our empirical observations further indicate that there is no universal trace or signature that can be used to highlight IO members across campaigns originating in different geographical spaces. As an example, the *Venezuelan_1 Co-Retweet* Network (**Figure 1**) campaign contains only 2 IO nodes out of 2,054 total vertices, whereas *Iran_1 Co-Retweet* graph has a share of 421 IO accounts out of 716 nodes.

It is worth mentioning that the application of the textual information as an indicative coordination factor was inspired by previous work on textual similarity analysis [12], [15]. Nevertheless, the proposed method of **Temporal Semantic Proximity Analysis (TSPA)**, comes as an extension to previous work, aiming to research the capacity and efficacy of using textual and temporal similarity to identify coordinated activity.

B. Methodology

As a first step of the **TSPA** pipeline (High-level overview of the pipeline visualized in **Figure 2**), we are applying a set of regular expressions on all of the tweets, in order to strip the mentions, URLs, retweet prefixes, and hashtags. We further filter out any rows that are labeled as retweets and ones with less than 3 words post-filtering. Doing so, minimizes

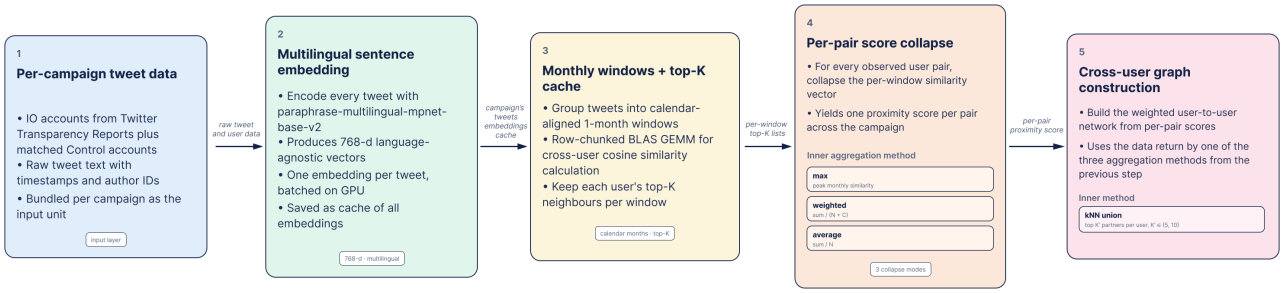

Fig. 2. High-level overview of the Temporal Semantic Proximity Analysis **(TSPA)** pipeline, from tweet preprocessing and multilingual sentence embeddings to within-window similarity computation, aggregation, and graph construction.

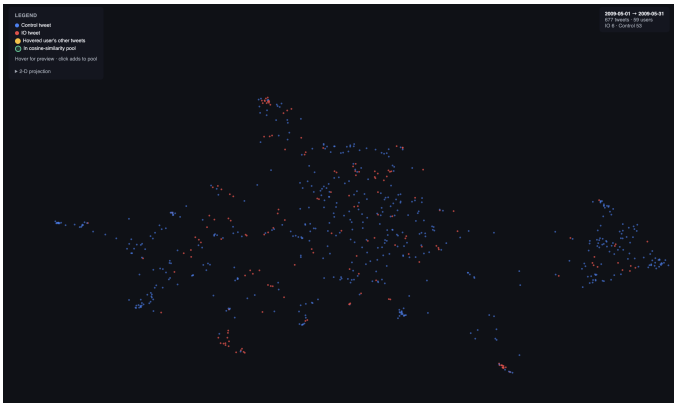

Fig. 3. 2D UMAP projection of tweet embeddings for a single monthly window in the *China_1* campaign. Each point represents one tweet, coloured by ground-truth label as indicated in the legend (top left).

the potential presence of noise further down the pipeline, along with shaping this method as a more platform agnostic approach. A flexible limit of tweets per campaign is set, with consideration to the original proportion of IO and Control posts.

Next, we employ the Sentence Transformer model *paraphrase-multilingual-mpnet-base-v2* [17] from *Hugging Face* in order to encode each tweet into a 768-D float32 vector. This model is selected as it is efficient at preserving the semantic meaning rather than the syntactic structure of tweets. It shows to be an important property in our pipeline, since the same narrative can be paraphrased, and presented as independent, fresh opinion [24], [25]. Encoding tweets in a paraphrase-aware embedding space lets two posts that share an underlying narrative remain close to one another even when no individual token is shared.

For each calendar-aligned window t_i containing N tweets, each represented by its 768-dimensional embedding $\mathbf{e}_k \in \mathbb{R}^{768}$ (already ℓ_2-normalised by the encoder, so cosine similarity coincides with the dot product), we construct the within-window similarity matrix

$$\mathbf{M} = \mathbf{E}\mathbf{E}^\top, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{N\times 768}$ is the vector of all $\mathbf{e} \in t_i$. The full $N \times N$ similarity matrix $\mathbf{E}\mathbf{E}^\top$ for large windows t_i, with the number of tweets $N = 100{,}000$, can be using up to 40GB of RAM memory. To keep the computation efficient and feasible, we are slicing $\mathbf{E}$ into chunks of a size $c \leq N$ as follows :

$$\mathbf{E} = \begin{bmatrix} \mathbf{E}_{[0:c]} \\ \mathbf{E}_{[c:2c]} \\ \vdots \\ \mathbf{E}_{[(K-1)c:N]} \end{bmatrix}$$

Each entry

$$M_{k,\ell} = \mathbf{e}_k \cdot \mathbf{e}_\ell,$$

thus represents the cosine similarity between tweets k and ℓ.

Two classes of entries are masked in $\mathbf{M}$ by assigning the value of -2, which lies outside the cosine similarity range $[-1, 1]$: (i) diagonal elements ($k = \ell$), and (ii) off-diagonal pairs where both tweets originate from the same author, as the analysis focuses on interactions between users.

For each row, the top-$K_{\max}$ nearest neighbours are extracted using `numpy.argpartition`, which runs in $O(N \log K_{\max})$ time per row, compared to $O(N \log N)$ for a full sort. We have selected $K_{\max} = 50$, covering the experimental range $K \in \{1, 3, 5, 10, 20, 50\}$, so that smaller K values can be obtained via simple slicing.

For each window t_i, the resulting neighbour indices, cosine similarities, and author identifiers are stored in a compressed NumPy `.npz` archive, together with a JSON file containing metadata such as the number of windows, $K_{\max}$, user labels (IO/control), and total tweet count. The resulting file sizes range from 1.8 MB (7,057 tweets) to 121 MB (500,000 tweets), which allows efficient storage and reproducible downstream analysis. An example of a single monthly window for the *China_1* campaign, projected to 2D via UMAP can be seen in **Figure 3**.

Three aggregation methods are then used to collapse the top-K similarity hits between a user pair (A,B) into a single score for each pair.

- **Maximum** - takes the single highest value of cosine similarity observed between any tweet of A and any tweet

of B, over the span of all time windows. This approach presents some potential limitations, since it is extremely favourable towards a single strong signal, but fails to describe any sort of long lasting cooperation.

- **Average** - returns the mean cosine over all N accumulated hits,

$$\text{Average}(A,B) = \frac{1}{N}\sum_{i=1}^{\text{count}_A}\sum_{j=1}^{\text{count}_B} s_{ij},$$

where s_{ij} is the cosine similarity between tweet i of A and tweet j of B. This method comes as a more equilibrated tradeoff, aimed to reach a more reasonable estimation of similarity between users.

- **Weighted** - applies a Bayesian-style shrinkage with a manual prior count C,

$$\text{Weighted}(A,B) = \frac{\sum_{i=1}^{\text{count}_A}\sum_{j=1}^{\text{count}_B} s_{ij}}{N+C}, \qquad (2)$$

with $C = 10$ by default, which suppresses spontaneous high scores from pairs with a low count of similar tweets, while still letting heavily-supported pairs converge to the empirical mean.

C. Graph construction

We summarise each campaign as an undirected weighted graph $G_{K'} = (V, E_{K'})$. For every user we select its K' strongest similarities to other users, and an edge $\{A, B\} \in E_{K'}$ is kept whenever B is among A's K' strongest matches or vice versa. Each edge is weighted, and comprises the precomputed pairwise similarity score based on one of the two aggregation methods: *maximum, weighted*. We build $G_{K'}$ for $K' \in \{5, 10\}$, and pass the resulting graphs to the community-detection step described further.

V. Experiments

A. Experimental Setup

1) Datasets Selection: Our experiments and analysis of the **TSPA** method begin with the selection of 14 campaigns out of the full dataset (**Section III**). The campaigns were selected with consideration to potential biases arising from the structure or other characteristics of the data. We have chosen campaigns that differ in the proportion of IO and Control users, the dominant tweet language, duration of campaigns, and per-window density of tweets, in order to obtain more comprehensive and critical result with our methodology. Additionally, we set a flexible limit of cleaned tweets per campaign, preserving the original proportion of IO/Control tweets. **Table I** presents a quantitative description of the filtered datasets, used in the experiments section.

*2) **Parameters**:* We establish the following set of variables: (i) the number K of neighbouring tweets that each tweet is compared against (default $K = 10$), (ii) the temporal length of each window (Granularity = 1 month). (iii) the parameter C for the weighted aggregation mode (default $C = 10$). (iv) the parameter K' that is the maximal number of edges per vertex within graph $G_{K'}$ (K' = 10). (v) maximal number of tweets per window. We limit the count of tweets in a window at $N_{\max} = 100{,}000$.

TABLE I
Campaign-level statistics for the selected post-filtering datasets, including remaining tweet and user counts, number of IO accounts, and the IO prevalence p_I

campaign	n_{tweets}	n_{users}	n_{IO}	p_I	p_I^2
Armenia	29,714	1,356	31	0.0229	0.001
China_1	500,000	23,332	426	0.0183	0.000
Cuba	500,000	19,636	458	0.0233	0.001
Egypt_UAE	20,561	513	236	0.4600	0.212
Iran_2	120,000	5,492	273	0.0497	0.002
Iran_6	120,000	11,067	201	0.0182	0.000
Russia_1	120,000	22,039	2,795	0.1268	0.016
Russia_2	71,654	2,209	344	0.1557	0.024
Russia_3	7,057	544	54	0.099	0.009
Russia_4	120,000	20,337	22	0.0011	0.000
Russia_5	120,000	9,219	45	0.0049	0.000
Spain	26,106	1,221	209	0.1712	0.029
Venezuela_1	300,000	4,307	585	0.1296	0.017
Venezuela_2	300,000	2,632	33	0.0125	0.000

B. Pair-class score distributions

We analyse two types of similarity connections: IO–Control (IC) pairs and IO–IO (II) pairs. Our goal is to identify if any of the proposed aggregation methods can capture a difference between the cumulative weights of IC and II connections. **Figure 4** presents pair-class CDFs for three campaigns *(Venezuela_1, Iran_2, Russia_1)*, selectively chosen in order to show how considerably different results look. The three graphs indicate that the *Venezuela_1* campaign shows a significantly larger gap between the II and IC connections than the other two. *Russia_1* is the campaign with the least distinguishable distribution of II, IC scores out of the selected scope.

The numerical differences between these two types of connections can be found in **Table IV (Appendix A)**

C. Aggregation Method Comparison (max / weighted / avg)

To compare the three aggregation methods, we compute the heatmaps for five evaluation metrics that we consider informative and suggestive.

Below we describe each metric, with the corresponding heatmaps present in **Appendix B, Figure 10**:

- **PR-AUC$_{II}$ (IO–IO ranking quality)** - Measures how well the score ranks IO–IO pairs above all other pairs. Values are interpreted relative to the campaign's random baseline p_I^2; only values roughly $p_I^2 + 0.10$ or higher indicate a genuinely informative signal. Campaign's baseline p_I^2 are included in the **Table I**.
- **Cohen's** d - Standardized difference between mean scores for the IO-IO and IO–Control pairs. Positive and large d means IO–IO pairs get substantially higher scores, compared to IO-Control pairs' score.
- **mean** H_I **(normalised IO homophily at** K**)** - Average tendency of IO users to have IO neighbours among their top-K neighbours, corrected for the IO base rate p_I.

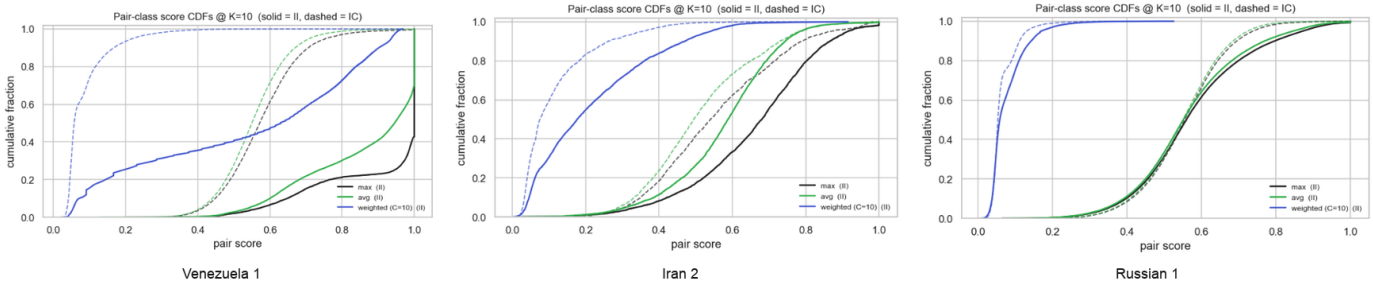

Fig. 4. Cumulative distribution functions (CDFs) of pairwise similarity scores for IO–IO (II) and IO–Control (IC) connections in three campaigns (*Venezuela_1*, *Iran_2*, *Russia_1*), illustrating how strongly **TSPA** separates coordinated IO pairs from mixed pairs in each case.

TABLE II
NUMBER OF CAMPAIGNS (OUT OF 14) IN WHICH EACH AGGREGATION VARIANT (*max*, *avg*, *weighted*) REACH THE BEST SCORE FOR A GIVEN METRIC.

Metric	max	avg	weighted	best method
PR-AUC (II)	**7**	2	5	max
Cohen's d (II vs IC)	5	3	6	weighted
mean $H_I(K=10)$	5	0	**9**	weighted
$L_I(K=10)$	5	0	**9**	weighted
Assortativity r	4	**4**	6	weighted

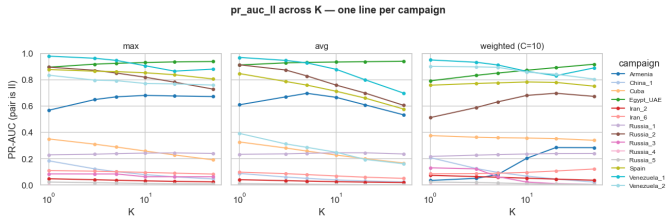

Fig. 5. PR-AUC$_{\text{II}}$ as a function of neighbourhood size K for all campaigns and aggregation modes.

A maximal $H_I(K) = 1$ means every IO user's top-K neighbours are all IO.

- L_I **(IO to Control leakage at** K**)** - The average fraction of Control neighbours in IO users' top-K lists. The random reference is $p_C = 1 - p_I$. Values clearly below p_C indicate that the method effectively positions IO users to IO dense neighbourhoods.
- **Newman assortativity** r**, (top-5% edges)** - Assortativity of the IO/Control labels on the strongest 5% of edges. $r = 0$ indicates chance mixing, $r > 0$ means same class edges are dominating, and $r < 0$ to an excess of cross-class links.

Table II summarises how often do particular aggregation variants *(max, average, weighted)* achieve the best performance across the 14 campaigns.

D. Sensitivity of Metrics to Neighbourhood Size

We examine the parameter $K \in \{1, 3, 5, 10, 20, 50\}$ to assess how neighbourhood size affects the discriminative power of our method.

The curves for PR-AUC$_{\text{II}}$ (**Figure 5**) indicate that performance tends to saturate for intermediate values $K \in \{5, \ldots, 20\}$.

The figures in Appendix A, namely **Figures 8 and 9**, provide a more detailed view of how K affects other metrics,$H_I(K)$ and $L_I(K)$, across all campaigns. Our findings indicate that higher values of K lead to an increasing of $L_I(K)$ values for the majority of campaigns, since we select a broader space of analysis for each tweet.

E. Application of Leiden community clustering algorithm

Following that, we have decided to use the computed network graphs (**Section IV-C**), to receive a clustering classification of the data. Motivated by prior work on clustering social media users [11], [13], [20], we apply the Leiden algorithm [18] for community detection.

The optimization objective maximised by Leiden is the weighted modularity Q (Formula (3)).

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{3}$$

Where A_{ij} is the pair-score of the edge, $k_i = \sum_j A_{ij}$ is the node's weighted degree, $m = \frac{1}{2} \sum_{ij} A_{ij}$ is the total edge weight. The δ is an indicator for whether nodes i and j are in the same community.

Figure 7 depicts the *Largest Connected Component* of the *Egypt_UAE*'s network graph, constructed using the *weighted* aggregation method with $C = 10$ (**Section IV-B**) and $K' = 10$ (**Section IV-C**). The resulting graph exhibits a clear structure of communities, with coherent organization of users into clusters. Clustering effort reaches a high modularity score of $Q \approx 0.863$ *(slightly fluctuating because Leiden is a non-deterministic algorithm)* and **purity = 0.929** compared to a **chance level of 0.563**.

We have also achieved suggestive results for other datasets, such as *Russia_1*. With **51** out of **59** communities identified by Leiden algorithm, being part of the *Largest Connected Component* (**Figure 6**). Clustering efforts grouped **1115** users into community *number three*, with **94%** of users **(1048)** being IO members. On top of that, cluster *number six* contained **982** IO-members with a **100%** intra-community purity.

A more detailed insight on the application of the Leiden algorithm against all 14 campaigns can be found in **Table III**.

Columns $\mathbf{IO\%}_{\max/w}$ present the share of **IO** members within the purest community of the networks built based on the *Maximum* and *Weighted* aggregation method, respectively. $\mathbf{IO}_{\max/w}$ show the count of IO members within that community.

VI. Combining Behavioral Traces and TSPA

We have also looked into how well our method **(TSPA)** complements an analysis based on the behavioral actions inspired from previous work [11]–[13], [26]. The following traces were chosen for an individual, per-campaign graph construction and analysis:

- **Co-Retweet** - a TF-IDF cosine similarity computed over each user's vector of retweeted content.
- **Co-URL** - a TF-IDF cosine similarity over each user's vector of shared URLs, connecting accounts that share the same links with overlapping frequency.
- **Hashtag Sequence** - a TF-IDF cosine similarity over ordered hashtag sequences of length at least five, capturing accounts that produce the same chains of hashtags.
- **Co-Mention** - a TF-IDF cosine similarity over each user's vector of mentioned accounts, linking users who direct their mentions at the same set of targets.
- **Fast Retweet** - a Jaccard-style overlap on the set of accounts each user retweeted within sixty seconds of the original posting.

Following that, we combine all of the traces into *fused networks*, in the same way as described by Luceri et al. [12]. We compute it by taking the union of nodes and edges, such that two accounts become connected whenever *any* trace links them. If the same pair is identified by more than one trace, the fused edge is assigned the maximum weight observed across those traces. We build two such networks: one over the five behavioral traces alone, and other as a union of all traces with the TSPA graphs.

Inspired by the work of Luceri et al. [12], we have decided to use the eigenvector centrality as a metric to evaluate our method's discriminatory power. In order to identify coordinated accounts we take every network's *Largest Connected Component* and compute the *weighted eigenvector centrality*, leaving every excluded account with a score of zero. The AUC-ROC scores are presented in **Table V Appendix B**.

VII. Evaluation of Results

A. Assessing IO detection via TSPA (RQ1)

In the first experiment **(Section V-B)**, we were determined to find if our method (TSPA) is able to capture differences between the similarity scores of *IO-IO* and *IO-Control* pairs. **Figure 4** and **Table IV (Appendix A)** show that there is consistently a considerable gap between the two types of connections, attesting the delimitation power of our analysis. Column **KS** D holds the largest vertical gap between the two CDFs.

An evaluation of the five selected metrics (**Experiment V-C**), against three proposed aggregation methods, has validated the discriminatory power and usefulness of our method. The heatmap found at **Appendix B, Figure 10**, displays the consistent effectiveness of our method across the majority of campaigs, thus answering the first *RQ*.

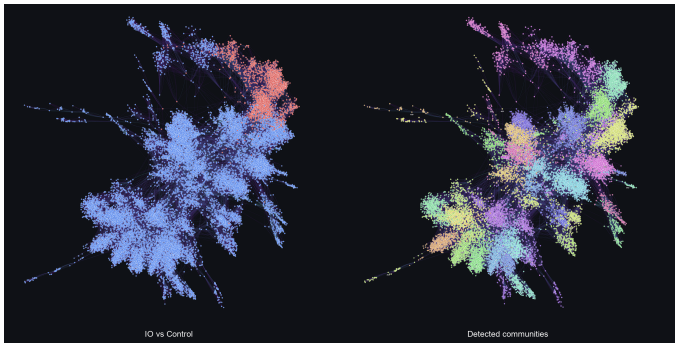

Fig. 6. *Russia_1* TSPA graph (largest connected component). Left: ground-truth IO labels, with IO accounts in red and Control accounts in blue. Right: communities detected by the Leiden algorithm on the same graph.

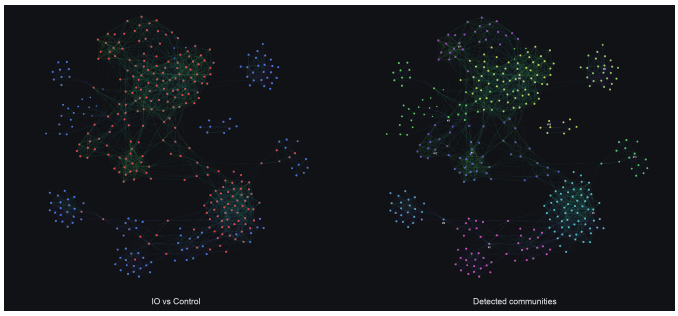

Fig. 7. *Egypt_UAE* TSPA graph (largest connected component). Left: ground-truth IO labels, with IO accounts in red and Control accounts in blue. Right: communities detected by the Leiden algorithm on the same subgraph.

B. Clustering applied to TSPA (RQ2)

By applying the weighted Leiden algorithm on each campaign, we have received some indicative results, that suggest the powerful grouping capacity for most of the campaigns.

Strong **IO-dominant** communities were formed within such campaigns as: *Cuba, Venezuela_1, Russia_1, Russia_2, Iran_2* (**Table III**).

A qualitative examination of the campaigns that formed less evident communities (*Russia_3, Russia_4, Russia_5, Venezuela_2)*, showed that these datasets contained very few tweets posted by IO accounts, uniformly spread across the time-windows. The nonexistence of any dense discussions of a similar topic facilitated by IO members affects the efficiency of TSPA in these cases. Our method classifies users into appropriate clusters, as long they took part in the discussion of a topic.

Altogether, TSPA serves as a great backbone for community-formation, making a further individual qualitative analysis of each community feasible.

TABLE III
MODULARITY AND IO PURITY OF THE MOST IO-DENSE COMMUNITY FOR THE **MAX** AND **WEIGHTED** TSPA AGGREGATION GRAPHS ($K' = 10$). FOR EACH CAMPAIGN AND AGGREGATION METHOD WE REPORT MODULARITY, THE PROPORTION OF IO ACCOUNTS IN THE PUREST COMMUNITY, AND THE ABSOLUTE IO COUNTS IN THAT COMMUNITY.

campaign	$Q_{\max}$	$Q_{\mathbf{w}}$	**IO%**$_{\max}$	**IO%**$_{\mathbf{w}}$	**IO**$_{\max}$	**IO**$_{\mathbf{w}}$
Armenia	0.847	0.852	15.7	15.7	19	19
China_1	0.849	0.792	7.1	16.3	98	119
Cuba	0.733	0.712	85.7	89.6	397	397
Egypt_UAE	0.773	0.863	100	100	118	70
Iran_2	0.897	0.900	95.9	96.7	236	237
Iran_6	0.928	0.927	60.9	61.5	185	187
Russia_1	0.900	0.899	100	100	967	982
Russia_2	0.905	0.878	97.6	100	160	147
Russia_3	0.676	0.676	2.0	2.4	2	2
Russia_4	0.910	0.910	2.2	1.8	8	9
Russia_5	0.806	0.805	2.1	2.4	5	5
Spain	0.516	0.501	28.2	47.9	138	126
Venezuela_1	0.671	0.663	41.1	58.2	165	166
Venezuela_2	0.594	0.557	1.9	2.5	14	9

C. Evaluating the combination of behavioral traces and TSPA (RQ3)

Our third research question was formulated to evaluate what are the gains from combining the behavioral traces with our temporal similarity proximity analysis. We decided to look into the AUC-ROC of each signal separately, as well as combined (described in **Section VI**, with the results presented in **Table V Appendix B**).

Our method scores above the random baseline *(0.50)* in 13 out of 14 campaigns, suggesting that IO-associated nodes have a higher centrality when compared to Control vertexes.

It is worth mentioning, that other selected behavioral traces, also lead to strong signals, depending on the playbook of methods used in a specific campaign. However, nitpicking a specific trace for each Informational Operation is unrealistic in an actual scenario. This fact shapes both Fused Networks as a great tradeoff between granularity and accuracy.

Throughout the experimental work, we have came to a conclusion that the separate study of Behavioral Traces and TSPA, is the best approach one can take. In a real life scenario, analysts can first cluster the observed users using the TSPA method, find any controversial topics, and back their observation based on anomalous behavioral activity.

VIII. DISCUSSIONS AND CONCLUSION

A. Limitations and the FAISS Trade-off

The principal limitation of TSPA is concentrated in a single stage: the per-window cosine similarity that produces each tweet's top-K neighbour matrix. For a window of N tweets we form the full similarity matrix $M = EE^\top$ from the $d = 768$-dimensional `mpnet` embeddings at a cost of $\mathcal{O}(N^2 d)$. Across a campaign the work scales with the sum of squared window sizes, $\mathcal{O}\big(d \sum_w N_w^2\big)$. To keep this quadratic term manageable we currently cap each window at $N_{\max} = 100{,}000$ tweets by uniform subsampling, trading statistical completeness for runtime on the busiest windows.

A considerable improvement would be to replace the exact $M = EE^\top$ matrix with an *approximate nearest-neighbour* index such as FAISS [27] or HNSW [28]. Because we only ever consume each tweet's top-K neighbours, materialising the entire $N \times N$ matrix is wasteful. Any of these methods would return those neighbours in roughly $\mathcal{O}(N \log N)$ and never store M in memory, which would collapse the computing cost considerably and remove the need for the subsampling altogether.

The trade-off is exactness. An ANN index is approximate, so it may miss a small fraction of true top-K neighbours. However, TSPA would still be tolerant of this, the cosine scores feed a *thresholded* aggregation rather than being consumed at full precision, so small perturbations in the neighbour list are very unlikely to alter the downstream community structure.

B. Conclusion

In this thesis we introduced **Temporal Semantic Proximity Analysis (TSPA)**, a platform-agnostic method for detecting coordinated information operations that relies solely on the semantic and temporal proximity of posts, rather than platform-dependent behavioral traces. Across 14 campaigns, we addressed three research questions.

For *RQ1*, our pair-class analysis revealed a pronounced separation between IO-IO and IO-Control similarity scores, confirmed across five complementary metrics and three aggregation schemes, establishing temporal semantic proximity as an informative coordination signal.

For *RQ2*, applying the Leiden algorithm to the TSPA graphs produced high-modularity partitions with strongly IO-dominant communities in campaigns such as *Cuba*, *Russia_1*, *Russia_2*, *Iran_2*, and *Venezuela_1*, reaching purities of up to 100%. Where campaigns contained few, temporally dispersed IO tweets, performance degraded, indicating that TSPA depends on the presence active topical discussion.

For *RQ3*, TSPA consistently exceeded the random baseline and, in several campaigns, surpassed every individual behavioral trace. Fused behavioral networks offered a robust, playbook-independent reference, with an open-ended question about the benefit of combining behavioral and textual signals. Overall, TSPA serves as an effective, language-independent backbone for highlighting coordinated communities and enabling targeted qualitative inspection. We conclude that analysts are best served by treating TSPA and behavioral analysis as complementary lenses, pointing toward scalable, cross-platform IO detection.

While the developed framework offers a high degree of predictive accuracy, it is not intended to serve as a standalone arbiter for final classification. Instead, it should function as a tool that must be complemented by rigorous qualitative analysis, in order to conclude any verdicts.

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] UNESCO and Ipsos, "Survey on the impact of online disinformation and hate speech," UNESCO, Tech. Rep., 2023. [Online]. Available: https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf

[3] B. Nimmo and D. Agranovich, "Recapping our 2022 coordinated inauthentic behavior enforcements," Meta Newsroom, 2022. [Online]. Available: https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements/

[4] S. Bradshaw and P. N. Howard, "The global disinformation order: 2019 global inventory of organised social media manipulation," Oxford Internet Institute, University of Oxford, Tech. Rep., 2019. [Online]. Available: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2019/09/CyberTroop-Report19.pdf

[5] Meta, "Adversarial threat report, first quarter 2025," Meta Platforms, Tech. Rep., 2025. [Online]. Available: https://transparency.meta.com/sr/Q1-2025-Adversarial-threat-report/

[6] A. Sen and B. Zadrozny, "QAnon groups have millions of members on Facebook, documents show," NBC News, 2020. [Online]. Available: https://www.nbcnews.com/tech/tech-news/qanon-groups-have-millions-members-facebook-documents-show-n1236317

[7] K. Conger, "Twitter takedown targets QAnon accounts," NPR / The New York Times reporting on Twitter's July 2020 enforcement action, 2020. [Online]. Available: https://www.npr.org/2020/07/21/894014810/twitter-removes-thousands-of-qanon-accounts-promises-sweeping-ban-on-the-conspir

[8] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, "How coordinated link sharing behavior and partisans' narrative framing fan the spread of COVID-19 misinformation and conspiracy theories," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 118, 2022.

[9] S. Tardelli, L. Nizzoli, M. Avvenuti, S. Cresci, and M. Tesconi, "Multifaceted online coordinated behavior in the 2020 US presidential election," *EPJ Data Science*, vol. 13, no. 1, p. 33, 2024.

[10] F. Cinus, M. Minici, L. Luceri, and E. Ferrara, "Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 U.S. election," in *Proc. ACM Web Conference (WWW)*, 2025, pp. 541–559.

[11] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated behavior on social media in 2019 UK general election," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 15, 2021, pp. 443–454.

[12] L. Luceri, V. Pantè, K. Burghardt, and E. Ferrara, "Unmasking the web of deceit: Uncovering coordinated activity to expose information operations on Twitter," in *Proc. ACM Web Conference (WWW)*, 2024, pp. 2530–2541.

[13] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: Methods and case studies," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 15, 2021, pp. 455–466.

[14] L. H. X. Ng and A. Iamnitchi, "Coordinated information campaigns on social media: A multifaceted framework for detection and analysis," in *Proc. Int. Conf. Multidisciplinary Int. Symp. Disinformation in Open Online Media (MISDOOM)*, ser. LNCS, vol. 14397. Springer, 2023, pp. 123–139.

[15] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, "Content-based features predict social media influence operations," *Science Advances*, vol. 6, no. 30, p. eabb5824, 2020.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[17] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525.

[18] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," *Scientific Reports*, vol. 9, p. 5233, 2019.

[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[20] B. Evkoski, I. Mozetič, N. Ljubešič, and P. Kralj Novak, "Community evolution in retweet networks," *PLOS ONE*, vol. 16, no. 9, p. e0256175, 2021.

[21] K. Sliwa, E. Kušen, and M. Strembeck, "A case study comparing Twitter communities detected by the Louvain and Leiden algorithms during the 2022 war in Ukraine," in *Companion Proc. ACM Web Conference (WWW)*, 2024, pp. 1376–1381.

[22] O. C. Seckin, M. Pote, A. Nwala, L. Yin, L. Luceri, A. Flammini, and F. Menczer, "Labeled datasets for research on information operations," *arXiv preprint arXiv:2411.10609*, 2024. [Online]. Available: https://arxiv.org/abs/2411.10609

[23] D. A. Martin, J. N. Shapiro, and J. G. Ilhardt, "Introducing the Online Political Influence Efforts dataset," *Journal of Peace Research*, vol. 60, no. 5, pp. 868–876, 2023.

[24] M. Richard, L. Giordani, C. Brokate, and J. Liénard, "Unmasking information manipulation: A quantitative approach to detecting copy-pasta, rewording, and translation on social media," *arXiv preprint arXiv:2312.17338*, 2023. [Online]. Available: https://arxiv.org/abs/2312.17338

[25] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," Georgetown CSET / OpenAI / Stanford Internet Observatory, Tech. Rep. arXiv:2301.04246, 2023. [Online]. Available: https://arxiv.org/abs/2301.04246

[26] T. Magelinski, L. H. X. Ng, and K. M. Carley, "A synchronized action framework for detection of coordination on social media," *Journal of Online Trust and Safety*, vol. 1, no. 2, 2022.

[27] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[28] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020.

APPENDIX A
ADDITIONAL EXPERIMENTAL FIGURES AND PAIR-LEVEL STATISTICS

This appendix provides supplementary visualisations and statistics that support the quantitative evaluation of TSPA in the main text. The figures summarise how key metrics behave as the neighbourhood size K varies, and the table reports detailed pair-level score distributions for IO-IO and IO-Control pairs.

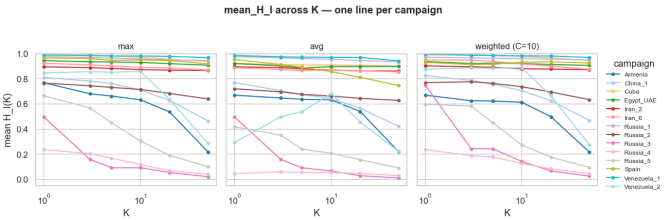

Fig. 8. Normalised IO homophily $H_I(K)$ across neighbourhood sizes K for all campaigns, illustrating how strongly IO users remain surrounded by other IO users as the top-K list widens under different aggregation modes.

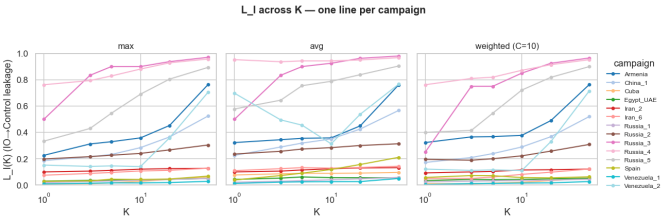

Fig. 9. IO-to-Control leakage $L_I(K)$ across neighbourhood sizes K for all campaigns, showing how often IO users' top-K neighbours include Control accounts as K increases.

APPENDIX B
OTHER FIGURES AND TABLES

TABLE IV
PAIR-LEVEL STATISTICS FOR IO–IO (II) AND IO–CONTROL (IC) SIMILARITY SCORES UNDER THE WEIGHTED AGGREGATION AT $K = 10$. FOR EACH CAMPAIGN, WE REPORT THE NUMBER OF II AND IC PAIRS, THE MEDIAN SIMILARITY FOR EACH CLASS, AND THE KOLMOGOROV–SMIRNOV DISTANCE D BETWEEN THEIR EMPIRICAL CDFS, QUANTIFYING HOW STRONGLY THE TWO SCORE DISTRIBUTIONS DIFFER.

Campaign	n_{II}	n_{IC}	**median II**	**median IC**	**KS** D
Armenia	172	1,274	0.500	0.096	0.614
China_1	5,834	93,103	0.113	0.084	0.157
Cuba	42,685	92,867	0.241	0.105	0.320
Egypt_UAE	10,793	479	0.254	0.134	0.268
Iran_2	4,699	1,917	0.177	0.075	0.310
Iran_6	6,969	4,245	0.170	0.079	0.361
Russia_1	137,630	20,808	0.058	0.053	0.156
Russia_2	11,608	8,930	0.463	0.084	0.665
Russia_3	4	449	0.405	0.108	0.688
Russia_4	27	2,416	0.186	0.058	0.545
Russia_5	174	8,337	0.180	0.078	0.387
Spain	8,274	21,544	0.389	0.070	0.653
Venezuela_1	35,159	64,812	0.630	0.058	0.683
Venezuela_2	258	19,526	0.760	0.143	0.896

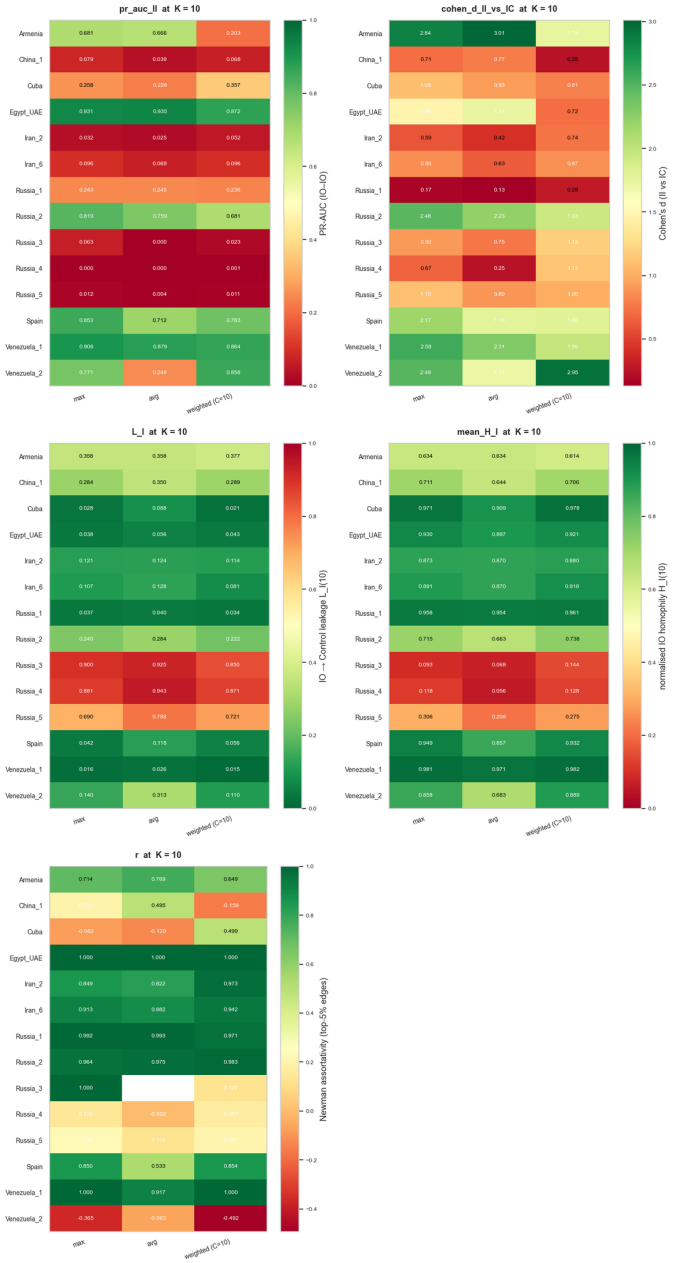

Fig. 10. Heatmaps of key metrics ($\mathrm{PR\text{-}AUC}_{\mathrm{II}}$, Cohen's d, $H_I(K{=}10)$, $L_I(K{=}10)$, and assortativity r) across campaigns and aggregation methods (max, average, weighted).

TABLE V

AUC-ROC SCORES OBTAINED BY EIGENVECTOR-CENTRALITY RANKING FOR EACH BEHAVIORAL TRACE, FOR TSPA ALONE, AND FOR TWO FUSED NETWORKS ACROSS ALL CAMPAIGNS. HIGHER VALUES INDICATE BETTER SEPARATION BETWEEN IO AND CONTROL ACCOUNTS.

campaign	Co-Retweet	Co-URL	Hashtag Seq.	Fast RT	Co-Mention	TSPA	Fused (Behav.)	Fused (Behav.+TSPA)
Armenia	0.471	0.786	0.804	0.500	0.672	**0.935**	0.892	0.890
Egypt_UAE	0.898	**0.972**	0.530	0.509	0.798	0.964	0.883	0.902
Spain	0.783	0.875	0.499	0.601	0.978	0.767	0.990	**0.991**
Iran_2	0.804	**0.850**	0.492	0.519	0.439	0.172	0.550	0.501
Iran_6	0.686	0.679	0.499	0.535	0.795	**0.970**	0.793	0.794
Russia_1	0.736	**0.861**	0.583	0.505	0.590	0.662	0.680	0.647
Russia_2	0.718	**0.838**	0.492	0.505	0.326	0.802	0.413	0.305
Russia_3	**0.776**	0.445	0.498	0.500	0.489	0.633	0.643	0.609
Russia_4	0.610	0.757	0.499	0.500	0.797	0.709	0.812	**0.827**
Russia_5	0.609	0.664	0.632	0.529	0.574	**0.826**	0.589	0.723
Venezuela_1	0.321	0.144	0.565	0.499	0.450	**0.939**	0.366	0.679
Venezuela_2	0.445	0.363	0.499	0.532	0.710	**0.882**	0.705	0.858
China_1	0.608	0.729	0.499	0.500	0.839	0.554	**0.856**	0.850
Cuba	0.957	**0.967**	0.669	0.688	0.891	0.843	0.910	0.908