Measuring the Scapegoat Mechanism: Evidence in Real Networks and in Parsimonious ABMs Gini, centralization, assortativity, and distance to the victim.

We present a mixed-methods study that operationalizes René Girard's mimetic theory in digital social networks, combining empirical data from X/Twitter with an agent-based model (ABM) on a small-world topology. We built time series and graphs for four cancellation episodes in Brazil and generated comparable synthetic data. We evaluated concentration and power structure using engagement Gini (per tweet and per user), centralization (in-degree) and assortativity by stance, in addition to peak indicators (peak/median) and top-k user participation. The ABM incorporates collective tension, contingent emergence of "leaders," selection of "victims" and "substitute victims," and a rite that temporarily relieves tension and reconfigures ties. The results show (i) "star" profiles in empirical events with high centralization and strong Gini, (ii) multi-community patterns in the synthetic before the rite, followed by post-rite recentralization, and (iii) a damping effect of skeptic agents on spread and on assortativity. The correspondence between empirical and simulated temporal metrics suggests that the scapegoat mechanism is detectable through signatures of concentration and (re)linking of ties. We contribute: (a) a bridge-metric between mimetic theory and network science; (b) an ABM-data validation protocol; and (c) implications for predictive dashboards and early detection of reputational crises.

1. Introduction

"Cancellation" on social networks is usually described as moral outrage amplified by algorithms. (1) (VAN BAVEL et al., 2024; BRADY et al., 2021; HUSZÁR et al., 2022). This description is useful, but insufficient to explain why certain episodes evolve up to a peak of symbolic violence — and why, after the climax, groups often "relieve" tension and reconstitute their ties. We propose that many of these episodes are better understood as mimetic crises: dynamics in which desires, fears, and hostilities spread by imitation, converging on a sacrificial target that concentrates accusations and releases, at least temporarily, the collective pressure.

This reading, inspired by René Girard's mimetic theory (2), offers a mechanistic key for recurring patterns on digital platforms: the sudden centralization of attention around a victim, the opportunistic emergence of accusing "leaders," the formation of substitute victims, and the subsequent reconfiguration of ties.

The literature on virtual lynchings, coordinated harassment and "outrage cascades" has already documented unequal distribution of engagement, the roles of influencers, and the effects of information bubbles (3) (McLOUGHLIN et al., 2024). However, two gaps remain. First, we lack a processual model that explicitly connects Girardian mechanisms (tension, imitation, accusation, sacrifice) to measurable signatures in real graphs. Second, there is a lack of cross-validation protocols that align, over time, empirical network metrics and synthetic data generated by a model that represents such mechanisms. Without these bridges, the theory remains suggestive, and the metrics unmoored from social causality.

This article aims to fill these gaps in three integrated steps. First, we formalize the mimetic crisis in an agent-based model (ABM) with a small-world topology (4), in which (i) collective tension grows endogenously; (ii) agents transfer tension through mutual accusations; (iii) upon

surpassing a certain tension threshold, leaders emerge who accuse victims; and (iv) a rite occurs that reduces tension and induces rewiring of ties. Second, we derive bridge-metrics between theory and data: engagement Gini (per tweet and per user) to capture concentration, centralization (in-degree and Freeman centralization (5)) to capture star-shape and accusatory power, and assortativity by stance to capture accuser/skeptic segregation. We complement these with peak indicators (peak/median and peak/p90) and top-k shares (top-1, top-5, top-10) to quantify the capture of attention by a few actors. Third, we apply the protocol to real episodes of cancellation on X/Twitter (Brazil), building time series and empirical graphs comparable to the synthetic ones.

Studying cancellation as a mimetic crisis brings theoretical and practical advantages. From an explanatory standpoint, it shifts the focus from individual traits (e.g., "toxicity") to relational mechanisms: how the imitation of accusing spreads; why targets become focal points; how opportunistic leaderships consolidate; and how punitive rites reorder the social fabric. From a methodological standpoint, it enables testable predictions: for example, we expect an abrupt increase in centralization near the climax, a rise in Gini, and a reduction (or reversal) of assortativity when accusatory bridges cross bubbles. From an applied standpoint, these signatures can support early-warning dashboards and reputational risk management, with value for journalists, moderators, and organizations.

Our framing also dialogues critically with current approaches to polarization and "echo chambers." While many analyses emphasize stable communities, the mimetic crisis highlights transversal moments in which lines of cleavage give way before the convergence against a target — a movement that tends to produce star-shaped graphs and peaks of concentration. After the event, however, rearrangements of ties and symbolic memories can reinstate cleavages, explaining temporal trajectories in which centralization declines and the network returns to a more modular regime.

Empirically, we deal with four Brazilian cases of high repercussion, allowing us to compare common signatures (e.g., elevated top-1 share, maximum centralization near the peak) and idiosyncrasies (e.g., the role of skeptics as dampers, variation in the queue depth of secondary victims). On the synthetic side, we explore scenarios with different proportions of skeptic and friendly agents, assessing their effects on the metrics. We find evidence that skeptics slow accusatory imitation and decrease assortativity by stance, while greater tie density facilitates peaks of centralization but can accelerate post-rite relief.

The central contribution is twofold. Conceptually, we translate elements of mimetic theory into implementable state variables and decision rules, proposing a common vocabulary between the humanities and network science. Empirically, we show that metrics of concentration and centralization function as markers of the scapegoat mechanism, and that their temporal evolution aligns with the phases of our ABM (accumulation of tension → accusation → rite → reconfiguration). Thus, the study illuminates how and when the transformation from "event" to "symbol" occurs — that is, when the case ceases to be only a controversy and comes to organize collective memory and future trajectories of accusation.

Finally, we offer an ABM–data validation protocol that can be replicated in other contexts, in addition to discussing ethical implications: detecting early signatures of sacrificial convergence can reduce harm, but also raises questions about freedom of expression, moderation, and algorithmic responsibility.

The remainder of the article describes the model and the data (Section 2), presents the metrics and the protocol (Section 3), reports empirical and synthetic results (Section 4), discusses implications and limitations (Section 5), and concludes with perspectives for predictive dashboards and comparative research (Section 6).

2. Model and Data

2.1 Empirical cases and collection

The empirical collection focused on four episodes of digital cancellation crises that occurred in Brazil, selected for their wide repercussion and diversity of profiles of the victims. In all cases, the data were collected on Twitter, using hashtags and keywords associated with the episodes, covering time windows that captured the peak and the decline of interactions. The cases function as natural experiments to observe the emergence of tensions, leaders, and scapegoats in real networks.

Case 1 – Monark (2022). Bruno "Monark" Aiub, presenter of the Flow podcast, was widely criticized after defending the right to the existence of a Nazi party during a live broadcast. The episode provoked a strong reaction on social networks, loss of sponsorships and his removal from the program. The case is exemplary for evidencing the rapid formation of negative consensus and the digital isolation of the central agent.

Case 2 – Karol Conká (2021). The singer and TV presenter Karol Conká became the target of massive rejection during her participation in the reality show Big Brother Brasil. Her behavior inside the house generated a collective mobilization on the networks, which went beyond the show and resulted in the highest rejection rate in the history of the show. The case illustrates the ritualized formation of a televised scapegoat amplified by social media.

Case 3 – Wagner Schwartz (2017). The performance artist was accused of "pedophilia" after a presentation at the MAM of São Paulo, in which he interacted nude with the public. The video of the performance was edited and viralized out of context, generating attacks and threats. The case highlights how moral contagion and mimetic panic spread through disinformation and polarized interpretations.

Case 4 – Eduardo Bueno (2025). The writer and historian was "canceled" after publishing comments and ironies about the assassination of the conservative commentator Charlie Kirk in the USA. The online reaction led to the cancellation of events and invitations, followed by public statements by the author (between partial retraction of the "form" and maintenance of the "content"). The case is relevant for showing transnational coupling (trigger event external to Brazil) and cycles of institutional sanction connected to network dynamics.

The four cases were processed in Python and converted into GEXF graphs for analysis of centralities, modularity, assortativity, and inequality of attention; then, we compared these empirical patterns to the synthetic data generated by our ABM.

Pre-processing. We applied: (i) deduplication by tweet_id; (ii) language filtering (pt-BR priority), (iii) normalization of users (stable mapping user_id → handle), (iv) rate-limiting aware batching and collection logs, (v) expansion of reply threads when available. Reposts (retweets/quotes) were preserved as signals of diffusion; obvious spam was removed by minimal rules (repeated URLs + low lexical diversity).

Stance labeling (stance). We followed a triadic scheme: accuser, skeptic (opposed to lynching/defends guarantees) and neutral/undefined. Labeling occurs preferably on the edge (which avoids "essentializing" users): each interaction receives a label by supervised/semi-

supervised language model, with human samples for spot-check and precision calculation (we report the confusion matrix in the Appendix). For aggregated analyses, a user can receive the modal stance of their edges.

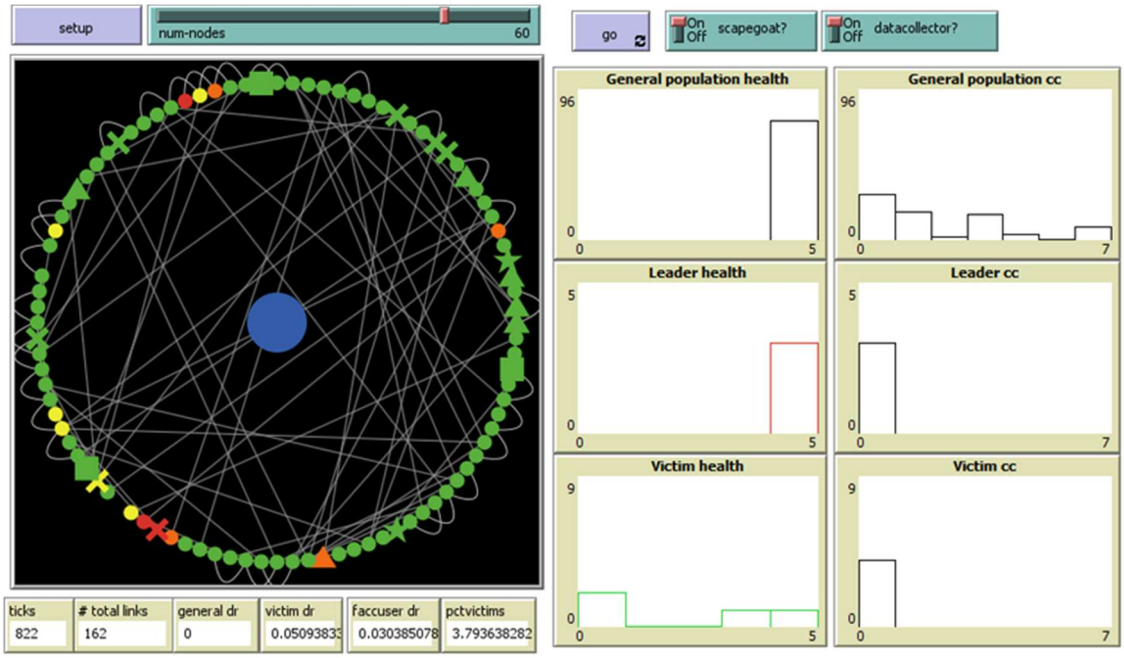
2.2 Construction of the empirical graphs

We defined directed graphs ($G_t = (V_t, E_t)$) by sliding windows (hour/day). A node ($u \in V_t$) represents a user active in the window; an edge ($u \rightarrow v \in E_t$) represents retweet, reply, quote or mention.

We generated series of metrics per window: engagement Gini (per tweet and per user), in-degree centralization (Freeman), assortativity by stance, peak ratios (peak/median), and top-k shares (top-1/5/10 by engagement). The empirical peak is the maximum of the volume/engagement series; we anchor all the series at (t=0) at the peak for case-aligned comparisons.

2.3 ABM (Small-World Scapegoat): ODD-lite

Purpose. Reproduce signatures of mimetic crises (accusatory convergence, emergence of leaders and victims, rite and reconfiguration of ties) and compare them with empirical series.



Entities and states.
• Agents (i=1..N) with: initial stance ($s_i \in \{\text{friendly}, \text{skeptic}, \text{neutral}\}$), tension ($T_i \in [0,1]$), health ($H_i \in [0,1]$).
• Initial network (G_0): small-world (Watts–Strogatz) with parameters and dynamic rewire.
• Dynamic roles: leaders, victims, failed accusers and failed victims.

Agenda at each tick.

1) Tension update (well as stressor).

The 'well' (global pollution variable at levels 0–3) acts as a systemic stressor that diffuses tension through the network. At each tick, live agents update their tension in a discrete way and limited to the interval [0,3]. Higher pollution values increase the probability of tension increment; low values reduce or stabilize tension. The color state (green→yellow→orange→red) is a visualization of the agent's tension band.

2) Interaction and imitation (transfer of tension).
When an agent becomes tense, they tend to transfer that tension to some neighbor with whom they have a link. Upon "passing" the tension, the emitting agent is relieved and returns to a state of lower tension, while the receiver increases their tension. This local relief/contagion mechanism implements the circulation of tension without the need for symbolic memory or edge weights.

2) Emergence of leaders.
Agents who reach high tension in a propitious context (neighborhood also tense and local responses) can assume the role of leader (marked in the code by change of shape/color). Once leaders, they have greater propensity to start accusations and to attract links (rewires) from other agents.

4) Selection of victim.
When a leader accuses, the victim emerges with greater probability among the structurally most vulnerable agents — notably those with few ties (low connectivity). In the code, several rules choose targets with small 'count link-neighbors' or with vulnerability markings; after the accusation, the victim is visually highlighted and has their 'health' reduced.

5) Mimetic crisis and rite.

Once the central accusation is triggered, the climax (rite) occurs with two main consequences: (i) tension relief — leader/victim agents return to states of lower tension; (ii) reconfiguration of ties — part of the links incident to the victim is broken and new links tend to form toward the leaders. There is no symbolic memory nor edge weights in the current mechanism; the post-rite effect appears as structural rearrangement and momentary reduction of collective tension.

Initialization and parameters. (N,k,p,) fractions ((f_{\text{skeptics}}, f_{\text{friendly}})), (\lambda,\beta_1,\beta_2,\tau_L,\tau_R,\alpha). No leader at (t{=}0). Initial health/tension may be homogeneous and default.

2.4 Experimental scenarios and calibration
The calibration of the model was conducted in an exploratory and simple way, based on manual adjustments of the main parameters — such as the total number of agents, the initial fraction of skeptics and friendly, and the intensity of the global stressor — until the simulated trajectories presented patterns qualitatively similar to those observed in the empirical data. The objective was not to achieve a statistically optimal fit, but to ensure that the general behavior of the model (growth of tension, emergence of leaders, choice of victims and post-rite relief) was plausible and coherent with the mechanisms expected by mimetic theory. This preliminary calibration allowed us to generate comparable synthetic data and to evaluate the conceptual validity of the model, serving as a proof of concept of the ABM–data protocol. However, we recognize that the procedure still lacks quantitative refinement: no systematic parameter sweep was carried out nor formal optimization by error metrics.
In future versions, we plan to implement a more precise and extensive calibration,

incorporating automatic sweeps (grid search or Bayesian optimization), multiple replicas for control of stochastic noise, and multi-metric fitting criteria based on empirical–synthetic discrepancies (for example, RMSE between normalized series of Gini, centralization and assortativity). This improvement will allow us to quantify parameter sensitivity and strengthen the reproducibility of the model.

2.5 Ethical considerations and reproducibility

The data were anonymous in the quantitative results (only aggregates and metrics). The focus is mechanical/structural, not reputational. We disclose: (i) collection/cleaning scripts; (ii) ABM generators; (iii) metric and calibration notebooks; (iv) empirical/synthetic graphs in open formats (.csv, .gexf). (6) github.com/carlospaes120/scapegoat. CC-BY-NC licenses (data/artifacts) and OSI-compatible (code) ensure citation and non-commercial academic use.

2.6 Design limitations

The collection guided by target terms tends to enhance star-shape (attention bias). Stance labeling, although by edge, inherits semantic uncertainty and irony. The ABM abstracts platforms and does not explicitly model recommendation algorithms; still, the rewiring and salience mechanism captures analogous effects. We opted for parsimony (without RL/ML in the agents) to isolate accusatory mimesis; future extensions can couple learning, degree heterogeneity and exogenous shocks.

3. Metrics and Validation Protocol

Our methodological protocol integrates metrics of different natures to capture the complete mimetic cycle: the convergence of attention, the structural reconfiguration of the network, the expulsion of the victim and the persistence of a symbolic memory.

3.1 Bridge metrics between theory and data

(a) Concentration and Convergence (Gini and Centralization). To measure the capture of attention by a few actors, we use:
• Engagement Gini: Calculated both per tweet (distribution of attention among contents) and per user (distribution of influence among actors). An increase of Gini close to the rite indicates the formation of a collective focus.
• In-degree centralization (Freeman): Measures the emergence of a star-shaped topology, where the victim becomes the center of a large volume of received interactions, signaling the convergence of accusatory power.
• Top-k shares: Proportion of total engagement concentrated in the top-1, top-5 and top-10 users/tweets, estimating the oligopoly of attention.

(b) Transversality and Community Reconfiguration. To capture how the crisis crosses and reorganizes the network, we use:
• Assortativity by stance: Measures the tendency of nodes with the same stance (accuser, skeptic) to connect. A drop in assortativity near the climax suggests that the crisis creates accusatory bridges that cross previously segregated bubbles.
• Modularity: We calculate the modularity of detected communities (via Leiden/Louvain algorithm) as an indicator of Girardian indifferentiation. A significant reduction in modularity indicates that boundaries between groups lose sharpness during sacrificial convergence,

reflecting the temporary collapse of social distinctions that characterizes the moment of the mimetic crisis.

(c) Structural Isolation of the Victim. To quantify the expulsion of the target, we introduce three isolation markers:
• Reciprocal degree of the victim: The number of mutual ties (A interacts with B and B interacts with A). Reaching reciprocal degree = 0 means that the victim lost all its bidirectional connections, a clear sign of ostracism.
• Size of the Strongly Connected Component (SCC_size): The size of the largest group in which all members can reach each other. An SCC_size = 1 for the victim indicates that they no longer belong to any conversation cycle, remaining structurally alone.
• Density of the ego-network: The internal connectivity of the immediate neighborhood of the victim. A collapse in this density shows that the social mesh that supported the victim has disintegrated.

(d) Peak Indicators. To identify the triggering of the crisis, we use Peak/Median, which capture the discontinuity and intensity of the climax.

(e) Sacrificial Residue and Symbolic Memory. To measure the lasting consequences of the rite, we analyze the post-rite with:
• Persistence of isolation: We check whether the low density of the ego-network and the null reciprocal degree of the victim remain for multiple windows after the peak.
• Median distance to the victim: We calculate the average distance of all nodes to the victim in the network. A sustained increase of this metric in the post-rite indicates structural avoidance and a topological "scar."

3.2 Time series and windows

We built hour-by-hour and day-by-day series via sliding windows. Temporal anchoring: we define (t{=}0) as the peak of volume/engagement of each case; the ABM series are anchored at the rite tick. For comparability, we standardize the series (z-score) and apply optional smoothing only for visualization.

3.3 Empirical ↔ synthetic comparison protocol

We align the empirical and synthetic time series at (t{=}0) and calculate a multi-metric loss (aggregated RMSE) to find the ABM parameters that best reproduce the signatures observed in real data, considering the equifinality of solutions.

3.4 Robustness tests

We evaluate the sensitivity of the metrics to variation of window width, to alternative definitions (e.g., betweenness centralization) and to noise in stance labels. Temporal placebos (shift the anchor) are used to confirm that the signatures are, in fact, centered on the peak.

3.5 Standardized theoretical interpretation

We link each metric to mimetic hypotheses:

H1 — Sacrificial convergence.
Joint rise of Gini and centralization (Freeman-in) as t→0−, t→0−t \to 0^{-}t→0−, t→0−, culminating in the collapse of the victim's reciprocal degree and ego-density at the climax.

H2 — Transversality at the climax.
Assortativity (by stance/community) and modularity fall near the peak/rite, indicating that the crisis crosses and reconfigures communities (breaking of bubbles).

H3 — Emergence of the leader.
As the rite approaches (and immediately after), at least one leader node appears with accelerated growth of salience and attraction of ties:
(i) increase of the leader's accusation out-degree and/or betweenness;
(ii) rise of top-k share concentrated in the leader;
(iii) rewiring directed to the leader (elevation of the leader's ego-density and in-degree in the post-rite);
(iv) functional polarization leader–victim: while the victim loses reciprocity and becomes isolated, the leader maintains/consolidates centrality.

H4 — Rite and sacrificial residue.
Drop of Gini and centralization for t≥0, t≥0t \ge 0t≥0, t≥0 (relief), but with the median distance to the victim remaining elevated and/or a greater fraction of nodes unreachable to the victim, characterizing structural avoidance (the "scar" of the episode).

H5 — Skeptics as dampers.
Greater fraction of skeptics ⇒ smaller peaks, smaller drop in assortativity and delay in the victim's isolation (dilution/retardation of cascades).

3.6 Training of Stance Classification Models

The objective of this stage was to train a model capable of identifying the discursive position (stance) of each tweet within the controversies analyzed — classifying them as accusers, defenders, or neutrals in relation to the target of the digital lynching.

Dataset and pre-processing

The data derive from the consolidated base after the cleaning and deduplication process, totaling 1,569 instances distributed in three categories: accuser (958), defender (253) and neutral (358). The corpus was split into training (1,255), validation (157) and test (157), preserving the original proportion among classes. The main textual field used for training was clean_text, containing text already normalized (removal of URLs, mentions, hashtags and non-textual characters).

The task is formulated as supervised three-class classification, with the objective of identifying the semantic orientation of the utterance in relation to the public figure under attack.

Model and configuration

We opted to fine-tune the BERTimbau Large model (neuralmind/bert-large-portuguese-cased), pre-trained on large Portuguese-language corpora, for offering better semantic sensitivity to discursive structures of Brazilian Portuguese. Training was conducted on an NVIDIA A100 GPU, using effective batch size of 8, learning rate of 1e-5, 8 training epochs, and gradient checkpointing enabled for memory optimization. The total execution time was approximately 20 minutes.

The initial layers of the model were kept unfrozen to allow full semantic adjustment to the particularities of the domain — marked by irony, moral judgments, and colloquial lexicon typical of social networks.

Balancing and bias-mitigation strategies

The highly asymmetric class distribution (with predominance of accusers) imposed learning challenges. To mitigate the imbalance, we employed a combination of:

1. **Assignment of class weights in the loss function (Cross-Entropy Loss), with an additional +20% weight for the "defender" class;**
2. **Targeted data augmentation applied exclusively to defender examples, with two complementary strategies:**

○ Simple noise edit: stochastic removal of mentions, hashtags, and emojis, insertion of colloquial connectors, and synonym substitution ("acho" → "creio", "errado" → "equivocado", etc.);

○ Backtranslation (PT → EN → PT) with MarianMT models, generating new instances that are semantically equivalent but syntactically varied.

Oversampling was kept disabled, avoiding redundancy with the weighting and augmentation mechanisms.

Training and evaluation

The model was trained with early stopping monitoring (patience of two epochs) and selection based on the best macro-F1 on the validation set. The validation metrics showed consistent evolution until stabilizing around the 7th epoch, with maximum performance of accuracy = 0.75 and macro-F1 = 0.67 on the test set.

Class	Precision	Recall	F1	Support
Accuser	0.80	0.82	0.81	96
Defender	0.47	0.36	0.41	25
Neutral	0.77	0.83	0.80	36
Overall accuracy			0.75	157
Macro F1			0.67	

Discussion and semantic observations

Qualitative analysis revealed that the model's main challenge is not grammatical understanding, but the semantic ambiguity between accusatory and defensive discourses — often constructed with similar lexicon and tone, differing only by pragmatic context and rhetorical intention. This linguistic proximity generates confusions especially in the detection of ironic defenses or accusers disguised as metalinguistic critique, a recurrent phenomenon in digital lynchings.

Even so, the model demonstrated good generalization on unseen texts and consistency in identifying neutrality and direct accusation, indicating that the BERT-based architecture is

capable of capturing part of the moral and affective nuances that structure the discursive field of lynching.

4. Results

In this section we confront empirical data of mimetic crises in social networks with outputs of our ABM. We organize the results in "empirical ↔ simulated" pairs, followed by distribution tests and robustness analyses. The focus is to assess whether the model reproduces central patterns expected by mimetic theory: sacrificial concentration, hub–victim inversion, accusatory homophily, and emergence of leadership.

4.1 Descriptive overview

The four empirical cases exhibit trajectories of triggering → peak → cooling, with the graphs in the neighborhood of the peak tending to a star-shape with the victim at the center (high in-degree).

Figura 1 Eduado Bueno

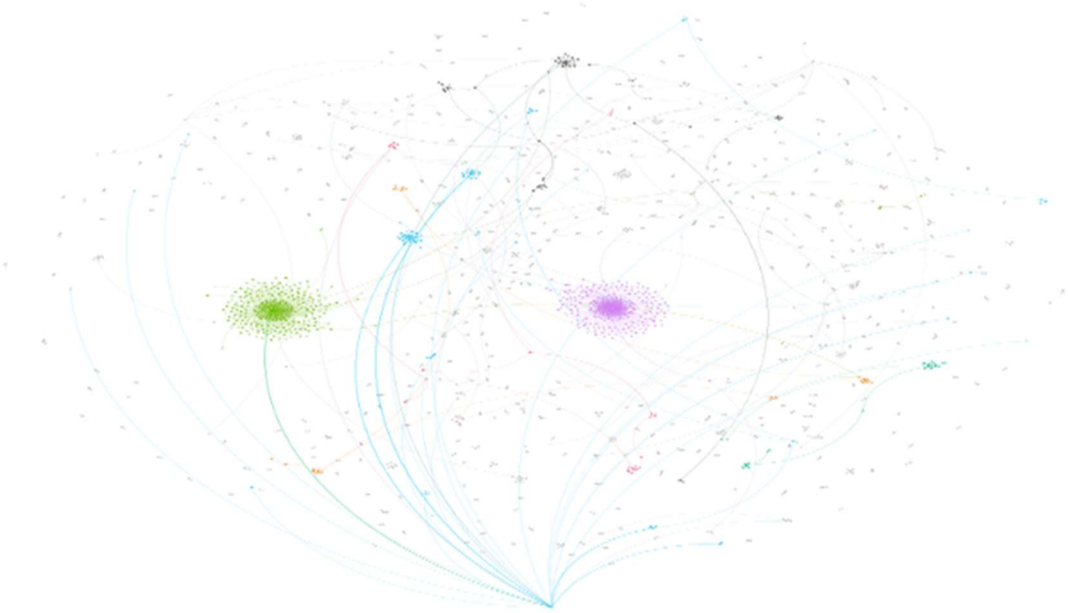


Figura 2 Karol Conka

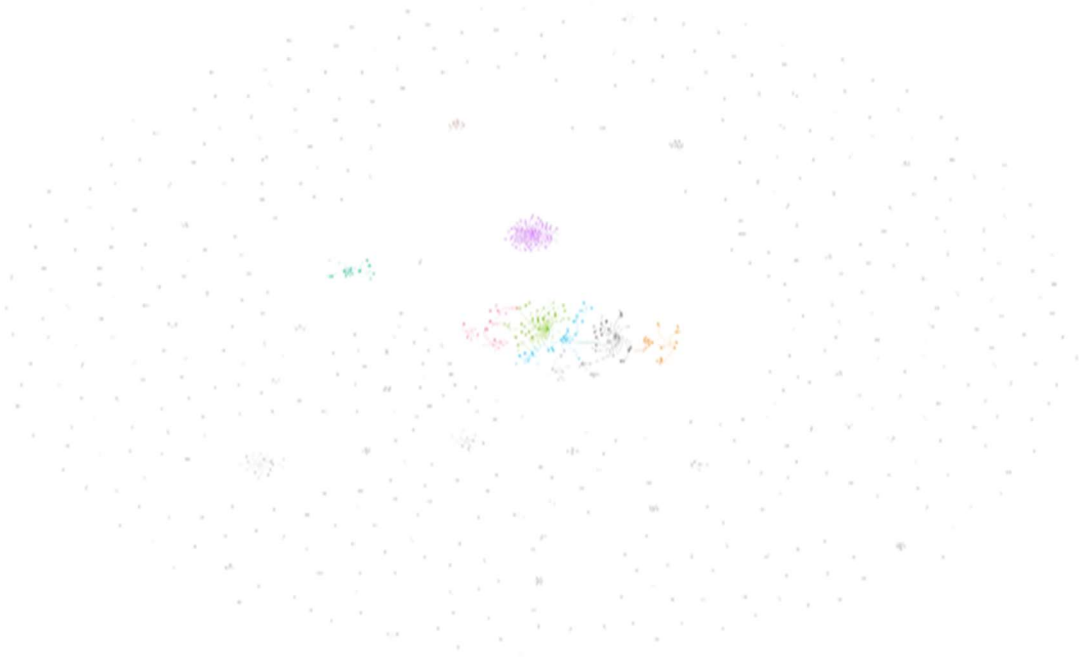


Figura 3 Monark

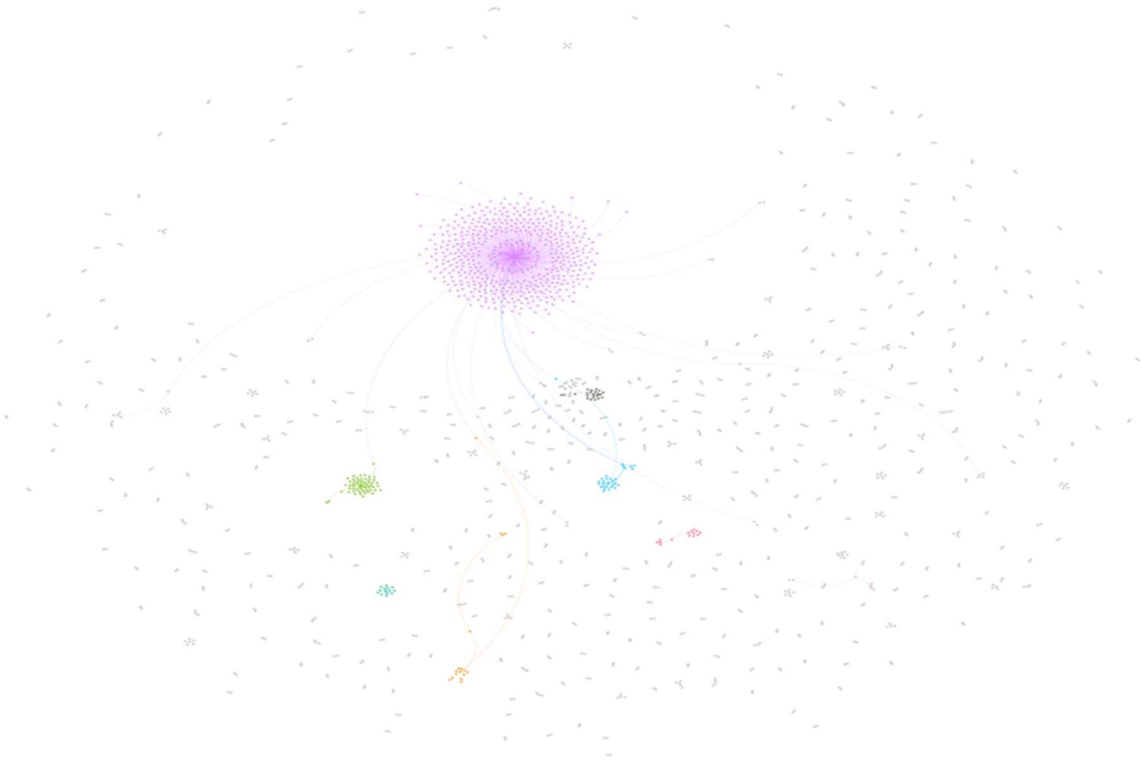
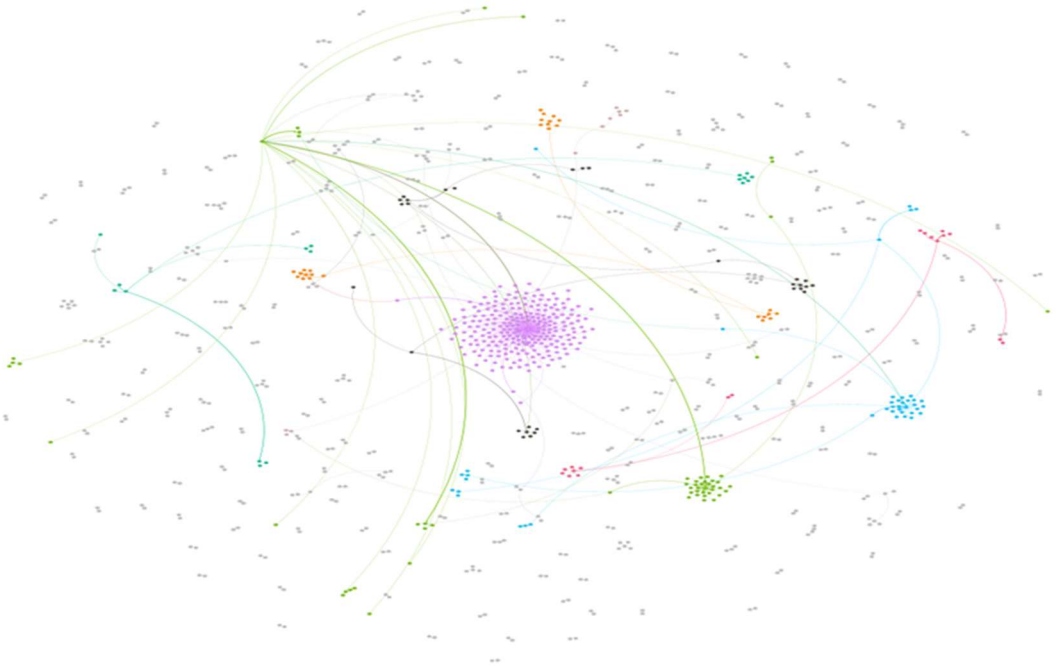
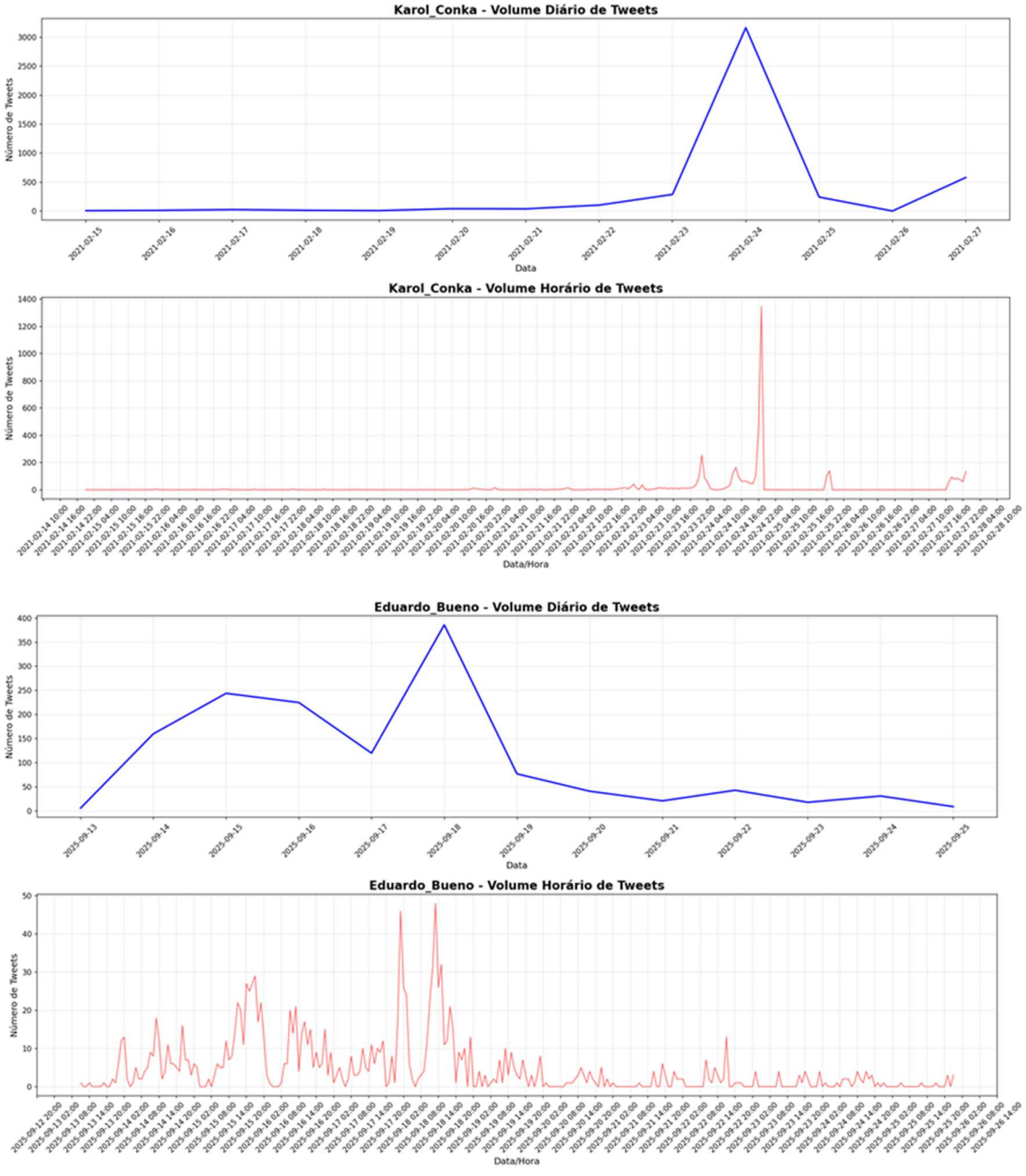


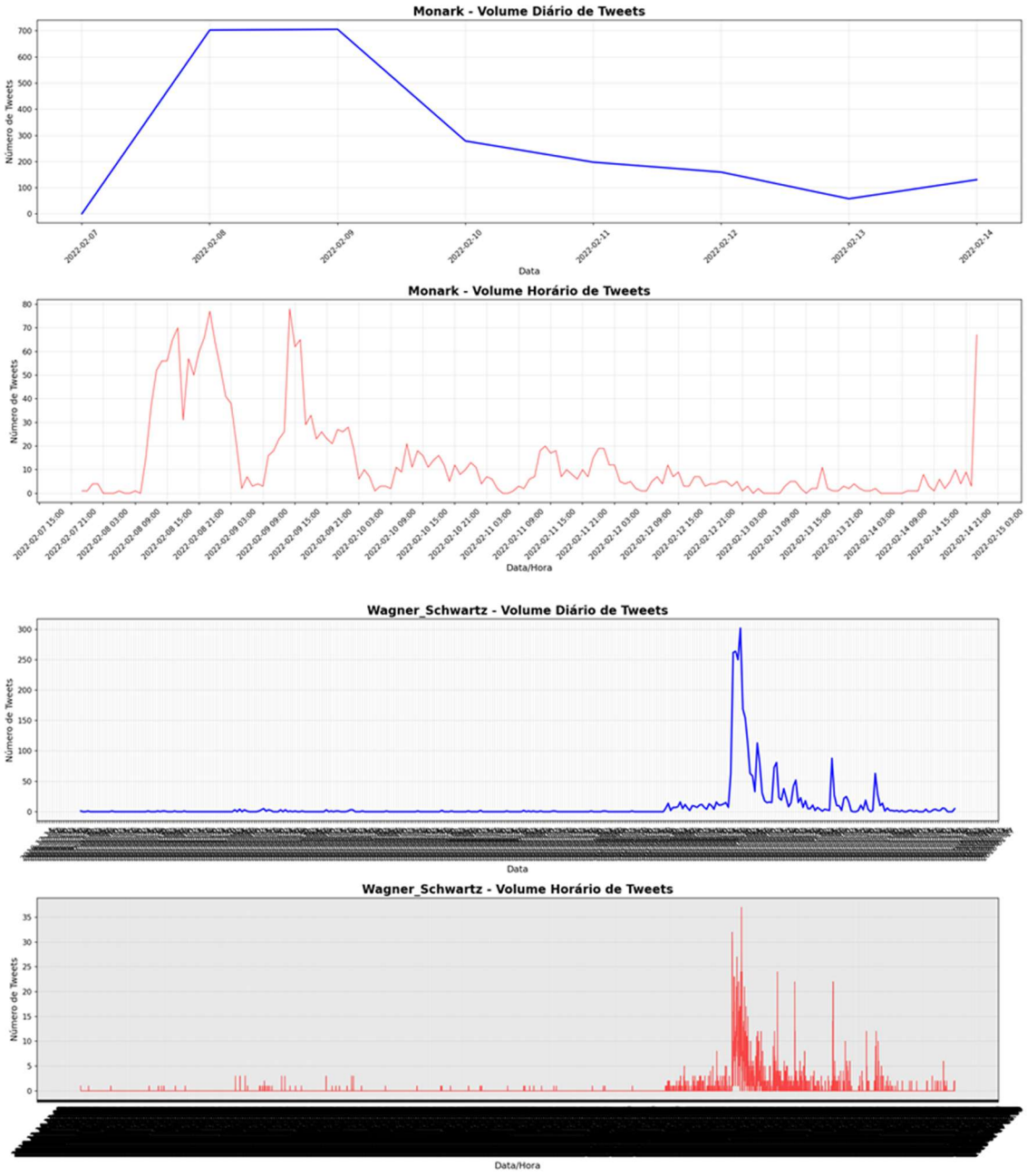
Figura 4 Wagner Scwhartz



4.1 Explosion and calm of the mimetic cycle

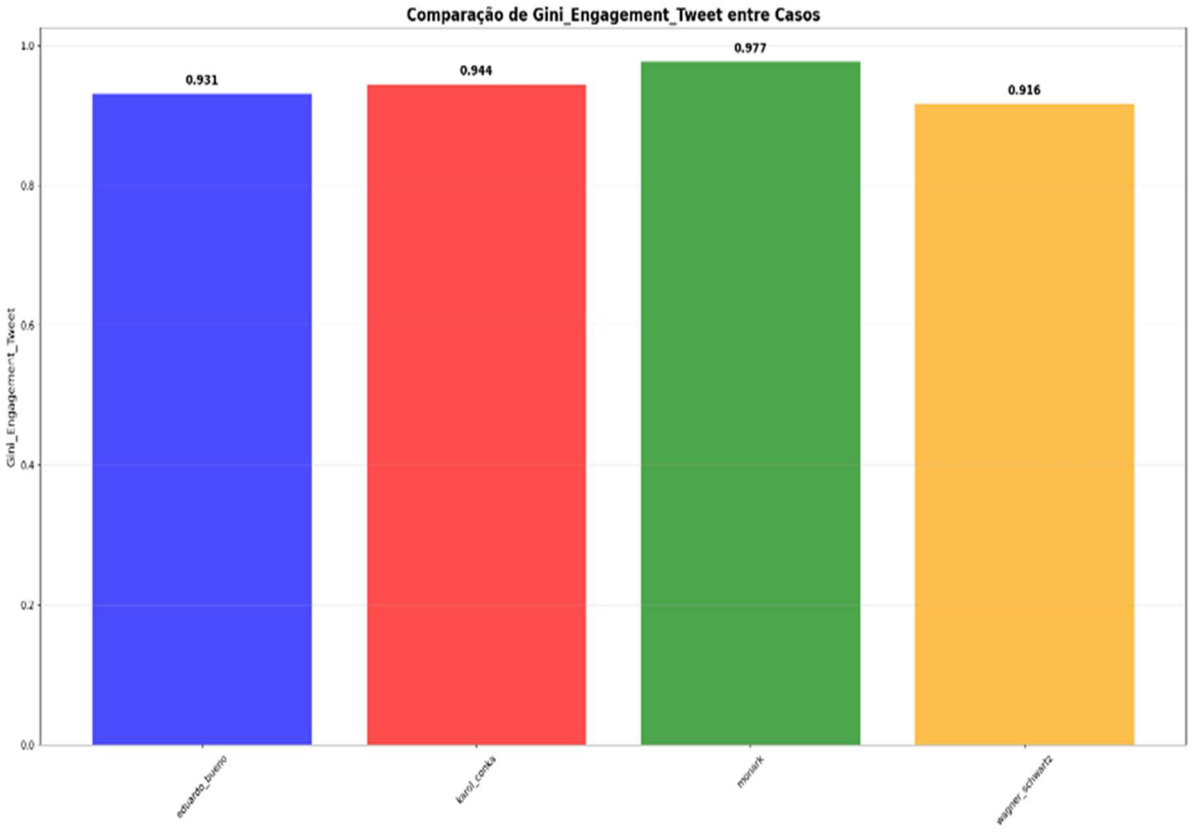
We observe, in the empirical cases, an abrupt jump of engagement synchronized to the trigger event (t_0) and subsequent relief. The peak ÷ median ratio (and peak ÷ p90) increases strongly in the short term and returns to basal levels after the consolidation of the accusation rite (Fig. 4.1A). The ABM reproduces the same temporal profile (Fig. 4.1B), supporting H1 (sacrificial convergence): the crisis concentrates attention and then dissipates tension.

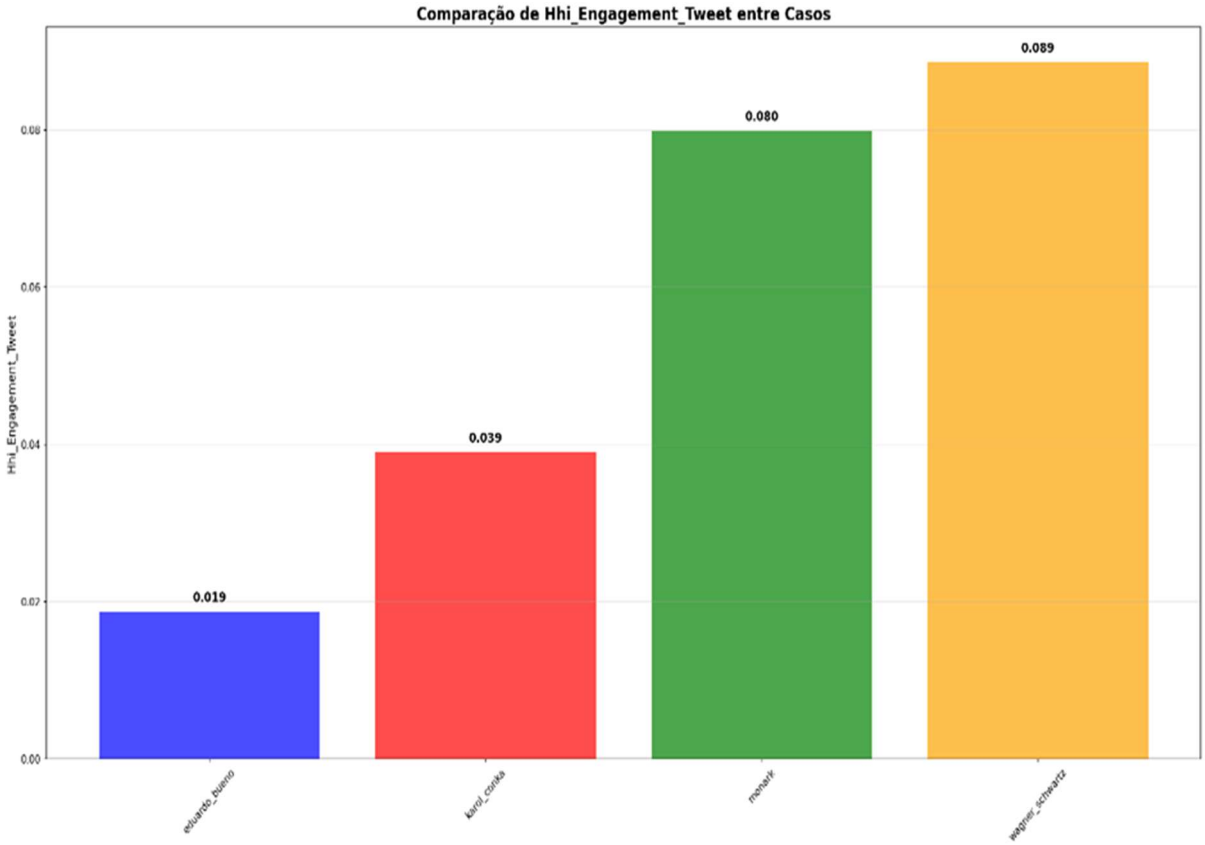


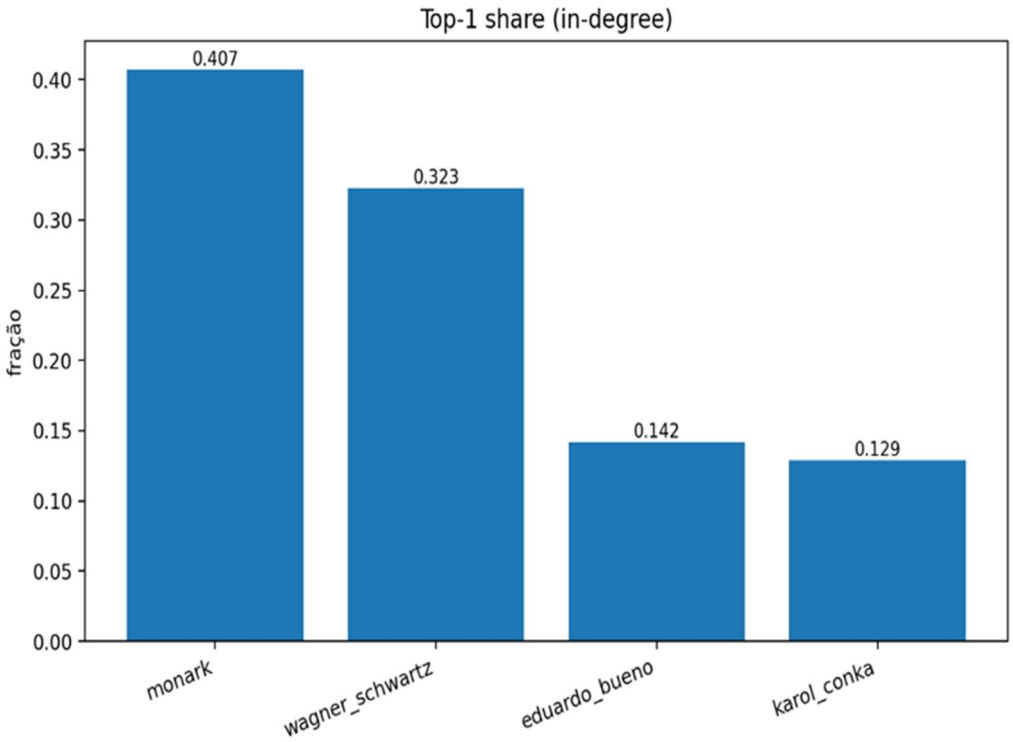


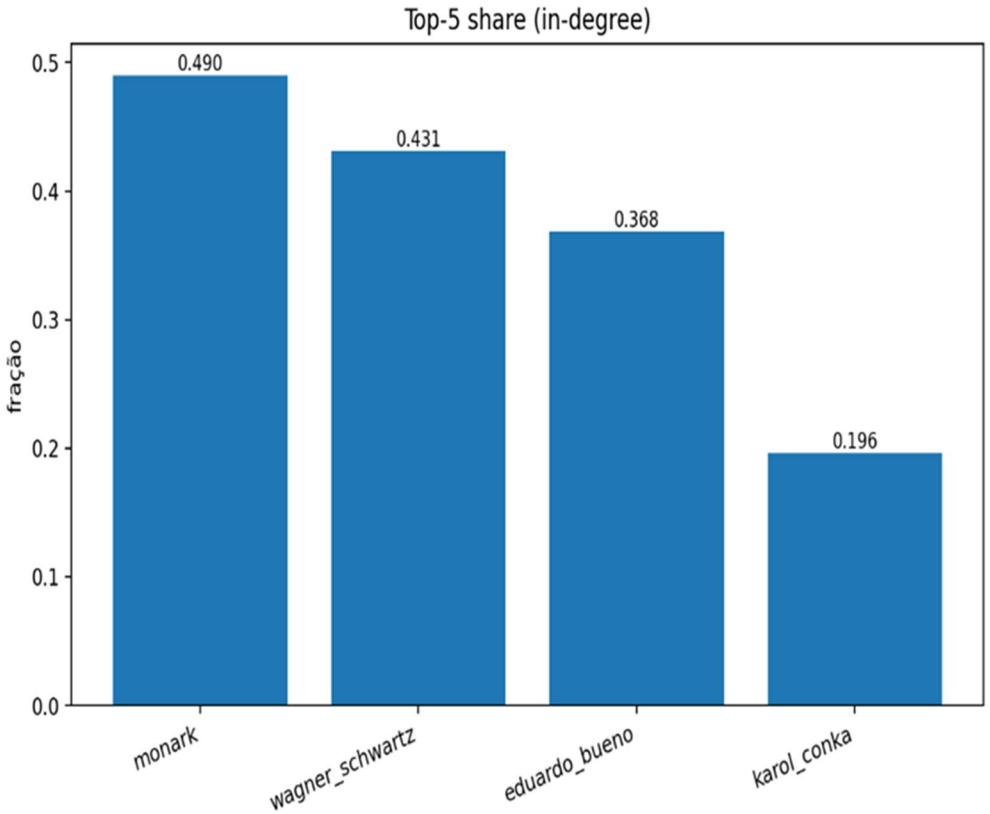
4.2 Attention concentration and heavy tail

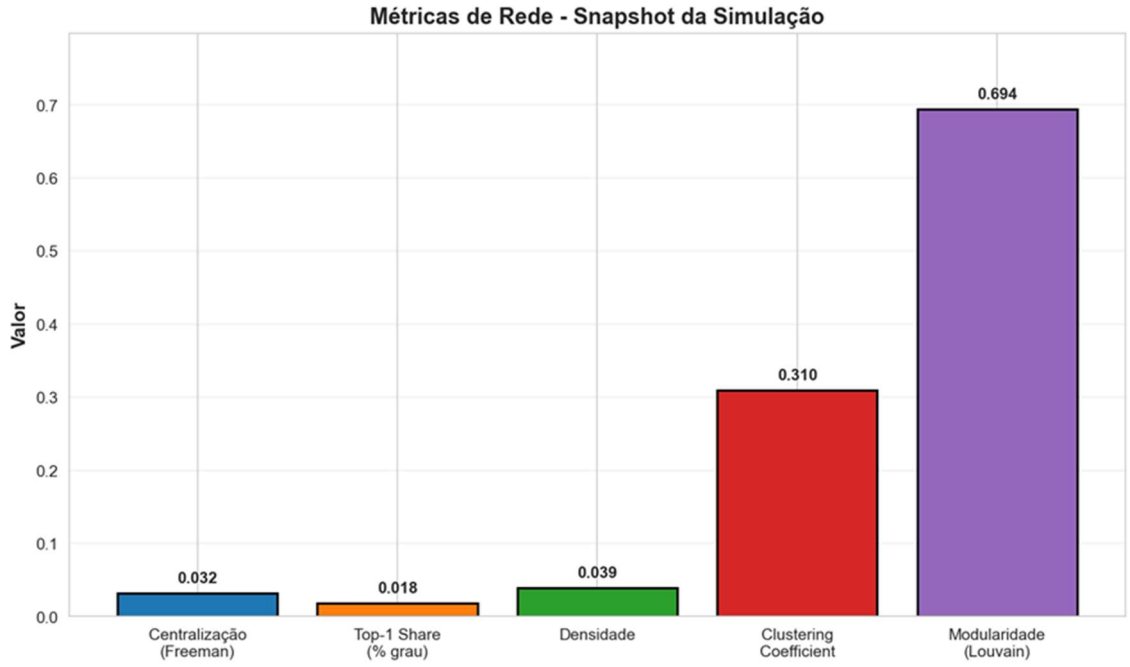
The engagement distribution is highly unequal: Gini and Herfindahl rise during the apex, while top-1/5/10-share show strong capture of attention by a subset of actors and contents (Fig. 4.2A). In the ABM, the same indicators rise in the crisis ticks (Fig. 4.2B). This convergence confirms that the mimetic process channels attention and amplifies asymmetries.


Comparação de Gini_Engagement_Tweet entre Casos
Gini_Engagement_Tweet
0.0
0.2
0.4
0.6
0.8
1.0
0.931
0.944
0.977
0.916
eduardo_bueno
karol_conka
monark
wagner_schwartz


Comparação de Hhi_Engagement_Tweet entre Casos
Hhi_Engagement_Tweet
0.00
0.02
0.04
0.06
0.08
0.019
0.039
0.080
0.089
eduardo_bueno
karol_conka
monark
wagner_schwartz

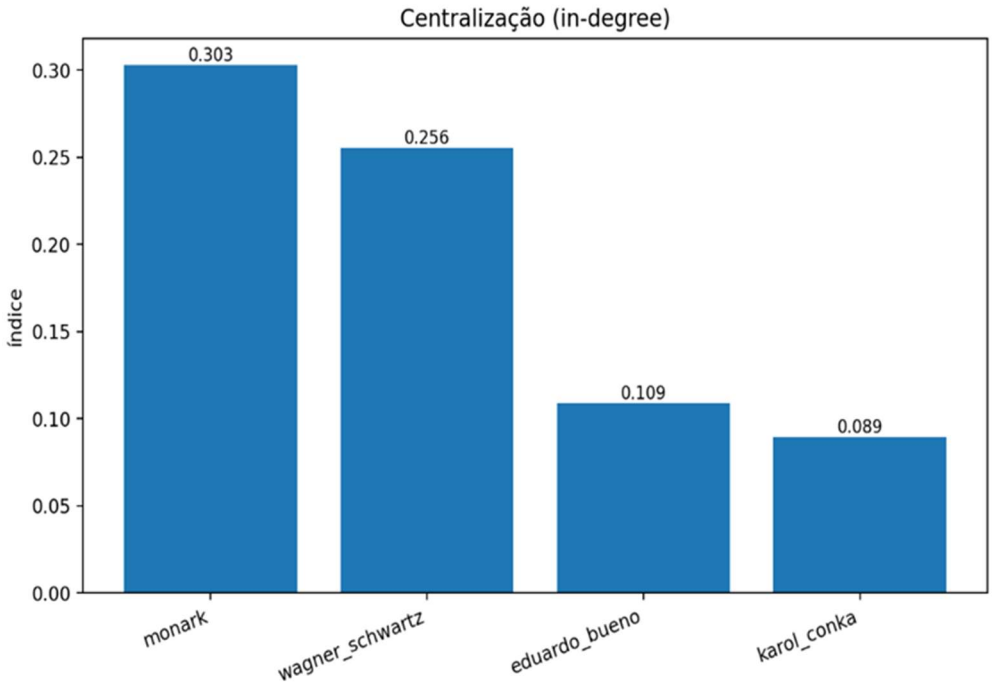

Top-1 share (in-degree)
fração
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.407
0.323
0.142
0.129
monark
wagner_schwartz
eduardo_bueno
karol_conka

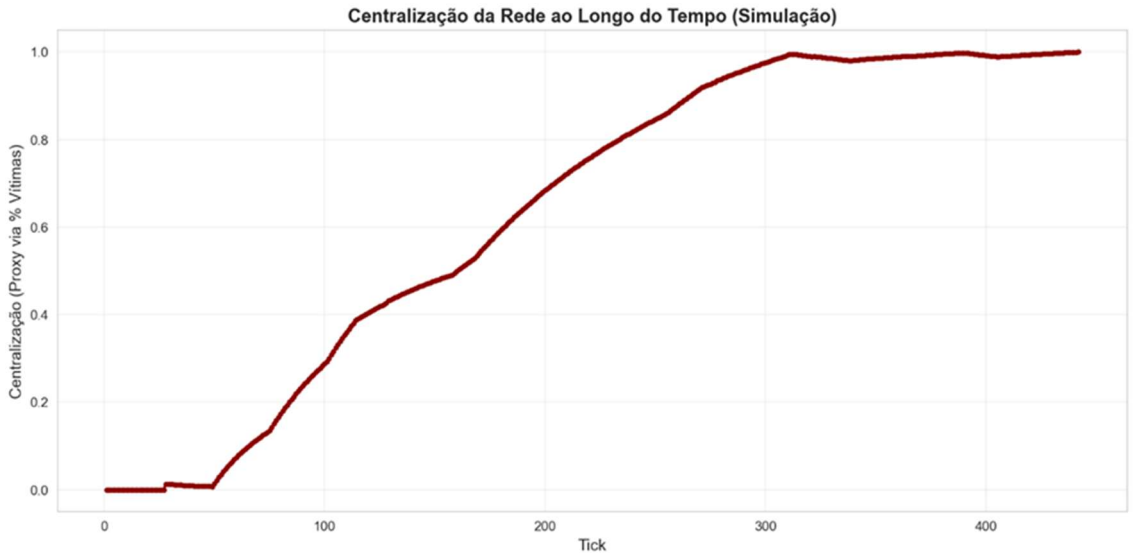

Top-5 share (in-degree)
fração
0.0
0.1
0.2
0.3
0.4
0.5
0.490
0.431
0.368
0.196
monark
wagner_schwartz
eduardo_bueno
karol_conka

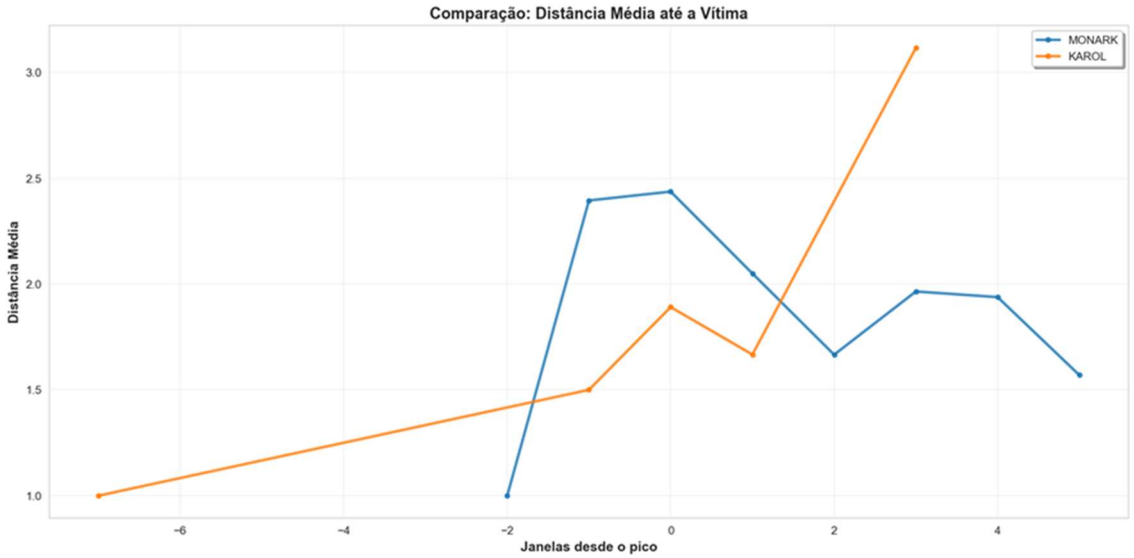


4.3 Hub–victim inversion and target-centered centralization

During the peak, in-degree centralization grows and the victim's ego-network forms a star: many edges converge toward the accusation target (Fig. 4.3A). In the ABM, the pattern repeats (Fig. 4.3B), evidencing the "hub–victim inversion": the most "central" node in the period is not a leader, but the victim — a signature of the scapegoat rite.

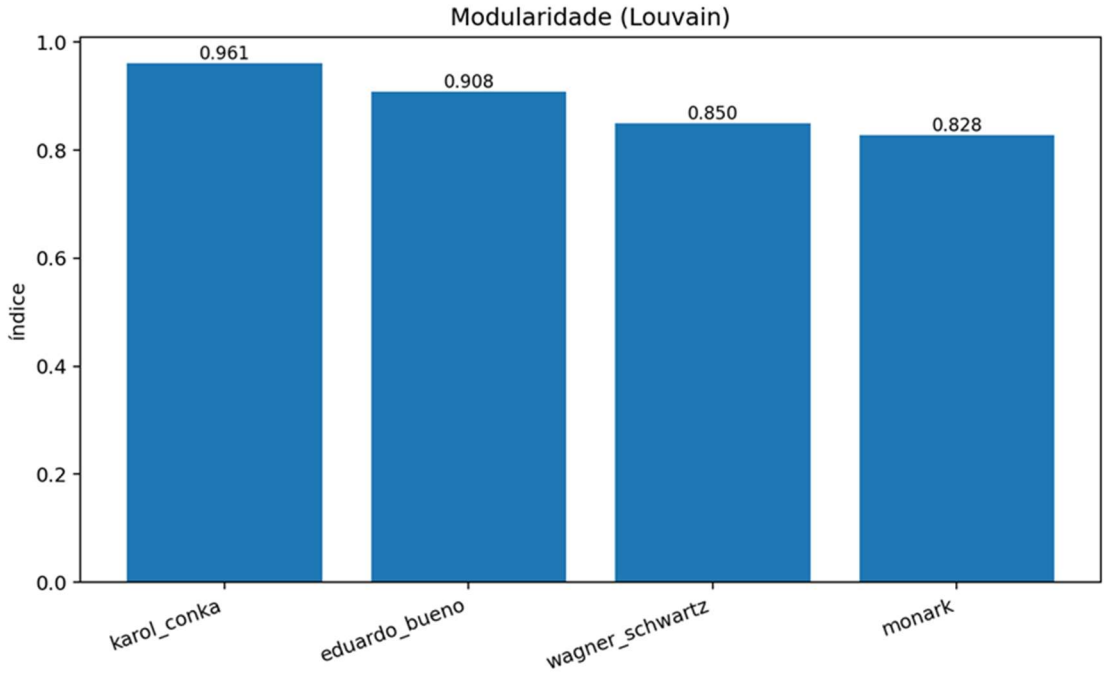

Centralização (in-degree)
índice
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.303
0.256
0.109
0.089
monark
wagner_schwartz
eduardo_bueno
karol_conka

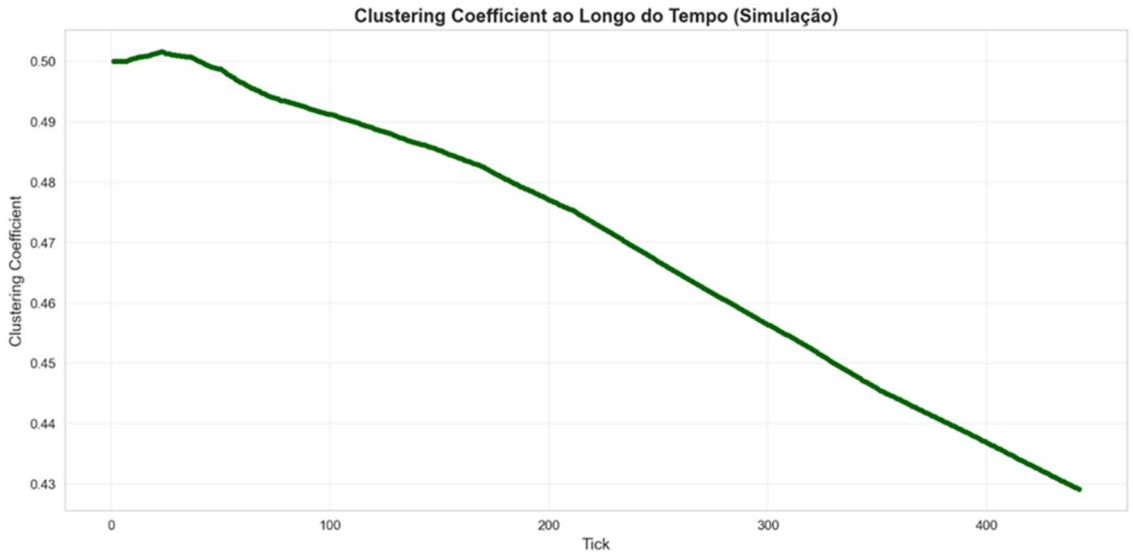

Centralização da Rede ao Longo do Tempo (Simulação)
Centralização (Proxy via % Vítimas)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
Tick



4.4 Accusatory homophily

Modularityty decreases during the escalation, indicating bubble closure and endogenous reinforcement of the accusation (Fig. 4.4A). The ABM replicates the effect (Fig. 4.4B), supporting H2: ideological/collective clusters align to point at the same target.

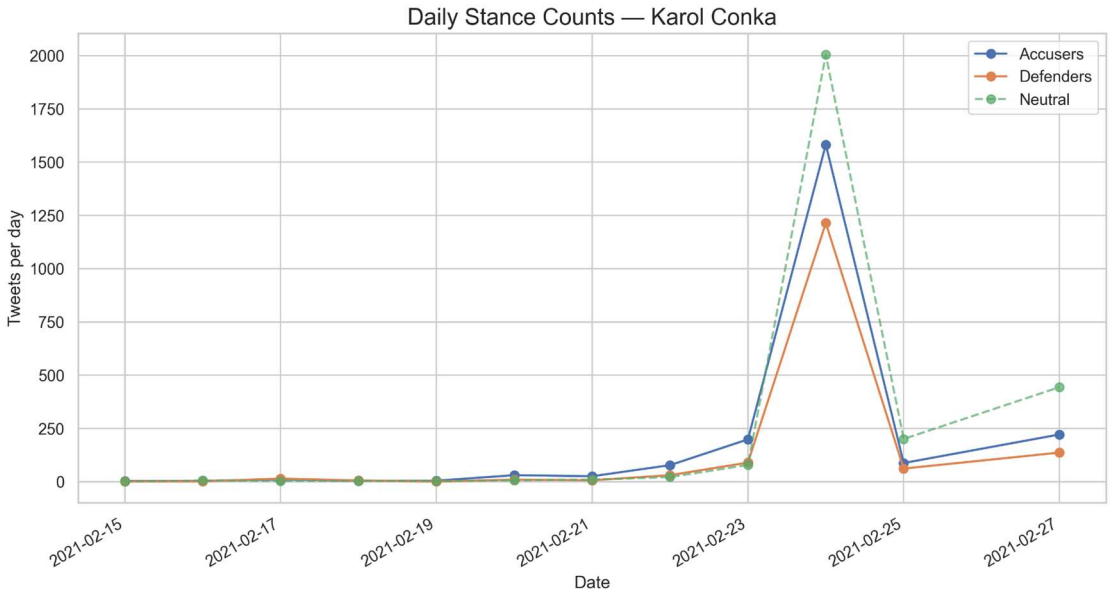




4.5 Results and Analysis of the Defensive Dynamics

In this section, we present the empirical results of the analysis of four high-repercussion cases. Through a set of visual metrics, we dissect the temporal dynamics between accusers and defenders, the engagement generated by each group, and the effectiveness of the defensive response. The objective is to identify patterns that characterize everything from a successful digital "sacrifice" to an effective containment of mimetic contagion, interpreting these results in the light of René Girard's theory.

4.5.1 Readings



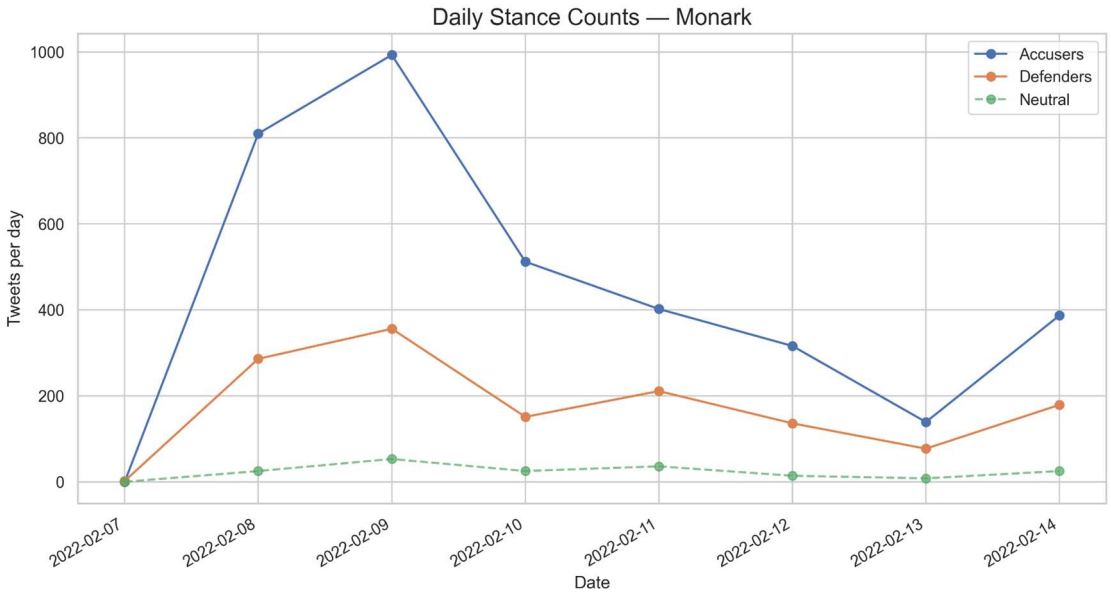
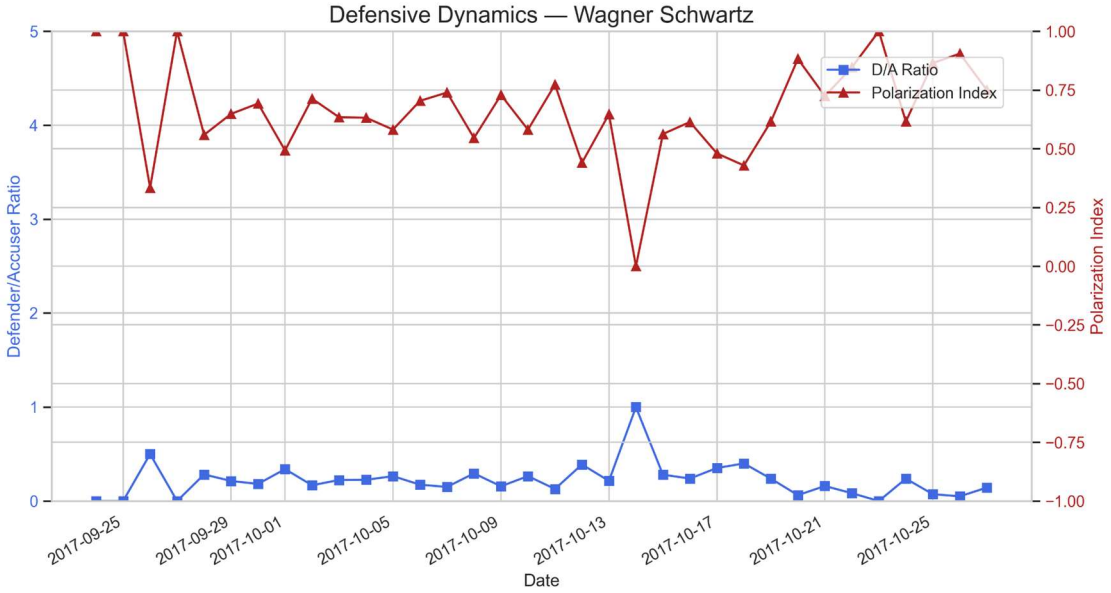
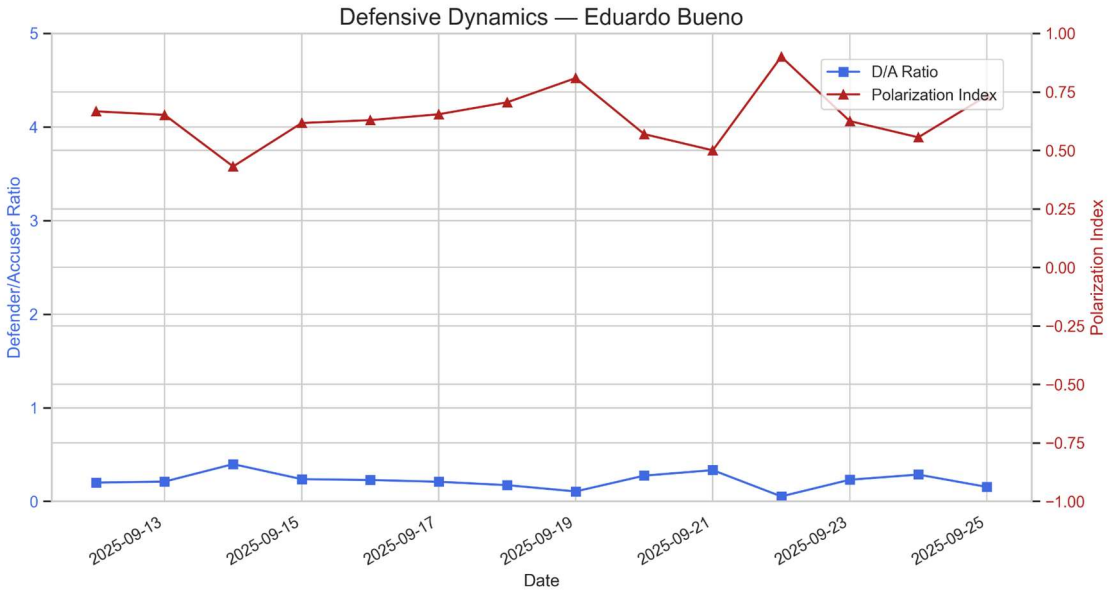
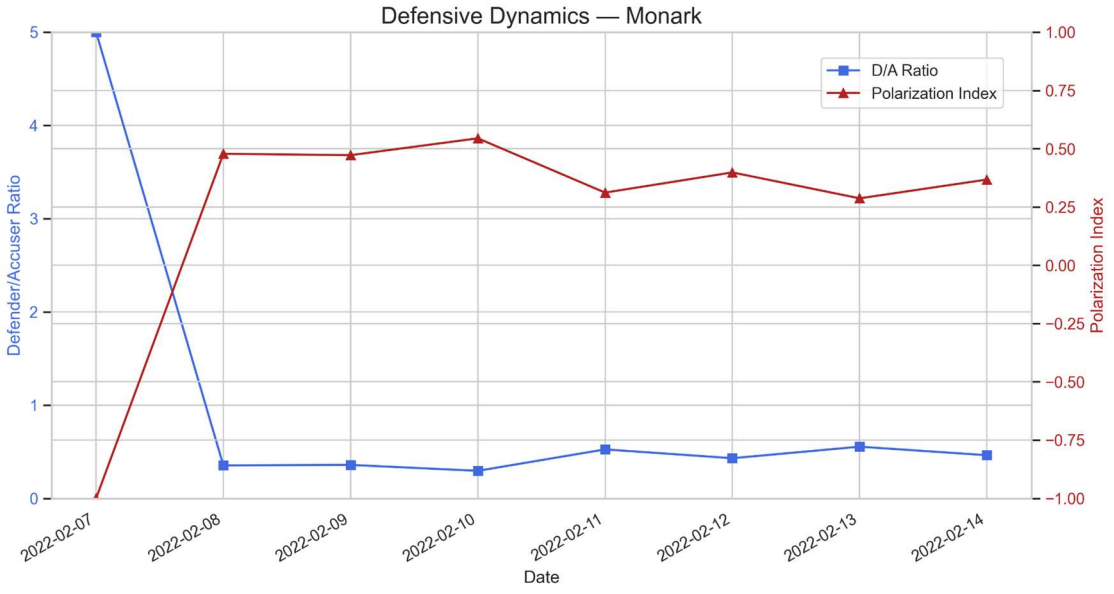


Figure 4.5a/d — Time series of the daily volume of tweets by stance (accusers, defenders and neutrals) in the four empirical cases analyzed (Eduardo Bueno, Karol Conká, Monark and Wagner Schwartz).
Reading (Fig. 4.5a/d). The typical signature is the explosion of accusers, followed by a drop; defenders vary by case. Theoretically, the accusatory curve materializes moral imitation (contagion), while the defensive curve expresses counter-mimesis (reintroduction of differentiation).




Defensive Dynamics — Eduardo Bueno
D/A Ratio
Polarization Index
Defender/Accuser Ratio
Polarization Index
Date
2025-09-13
2025-09-15
2025-09-17
2025-09-19
2025-09-21
2025-09-23
2025-09-25


Defensive Dynamics — Monark
D/A Ratio
Polarization Index
Defender/Accuser Ratio
Polarization Index
Date
2022-02-07
2022-02-08
2022-02-09
2022-02-10
2022-02-11
2022-02-12
2022-02-13
2022-02-14

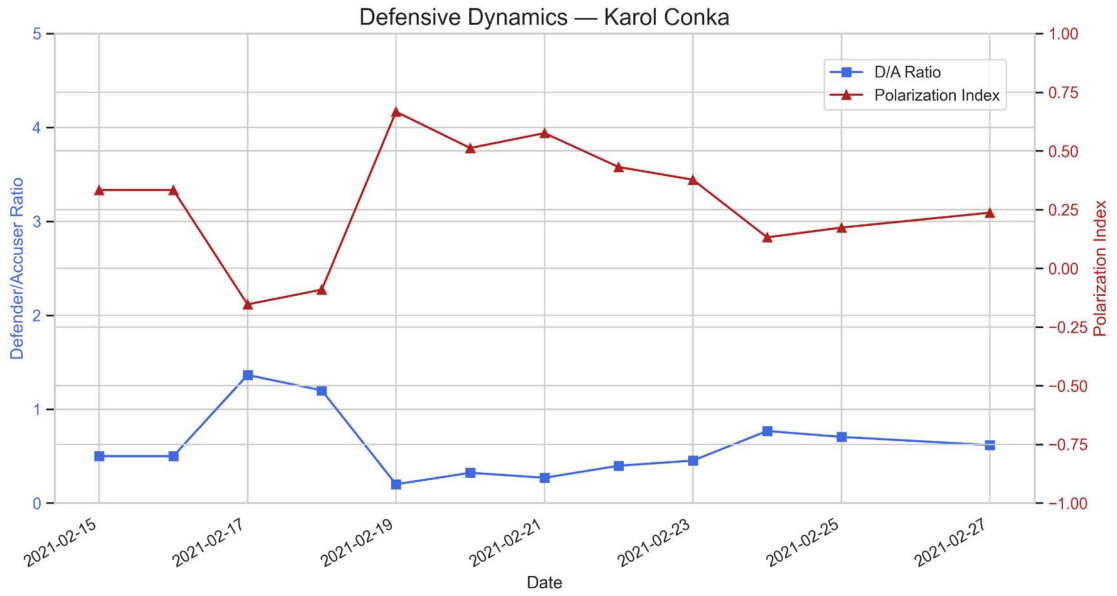


Figure 4.5e/h — Defensive dynamics: Defenders/Accusers ratio (D/A) and Polarization (A−D)/(A+D).
Reading (Fig. 4.5e/h).

- **D/A ratio (>1 = defense surpasses accusation on the day) indicates balance.**
- **Polarization: +1 = unanimity in favor of the accusation; 0 = balance; negative values favor defense.**
  Joint signal: rising D/A + falling Polarization ⇒ break of unanimity; the sacrificial mechanism loses cohesion.

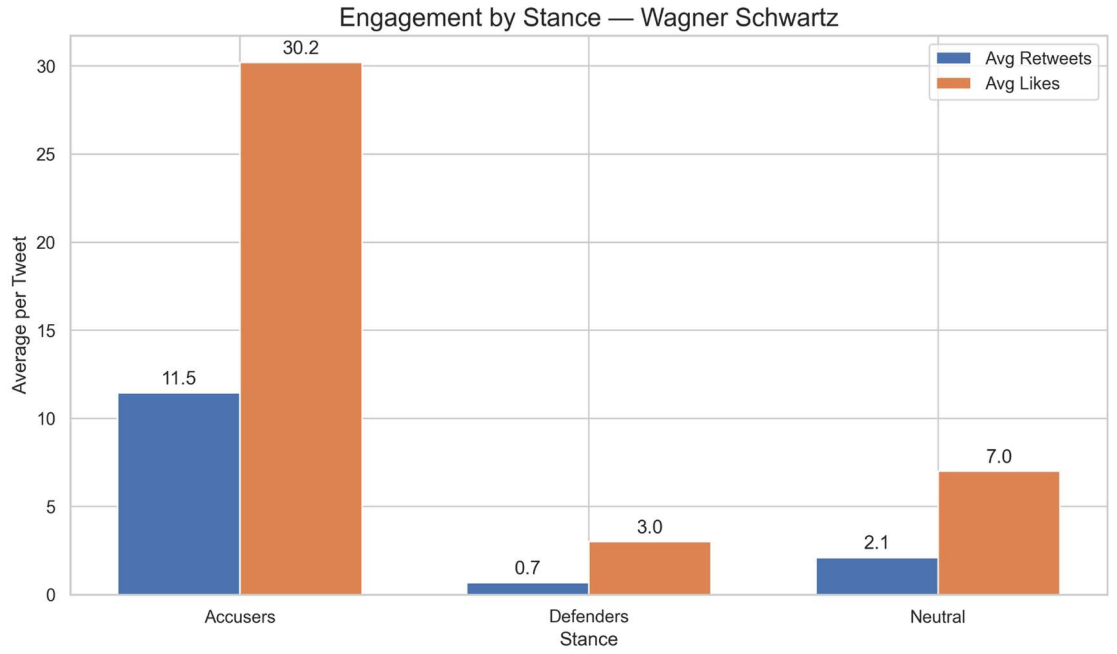


Figure 4.5i/l — Average engagement by stance (retweets/likes per tweet). Reading (Fig. 4.5i/l). Retweets measure contagion (imitative reach) and likes measure approval. Defenders with more likes (even with fewer RTs) indicate symbolic rehabilitation; if they also have high RTs, there is a counter-narrative with reach.

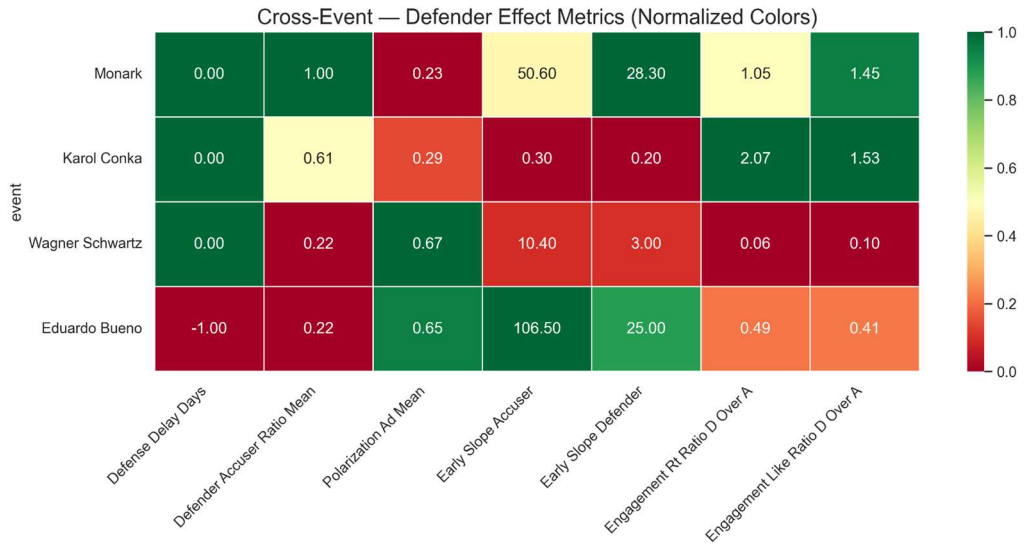


Figure 4.5m — Comparative heatmap of the "defender effect"
Figure 4.5m synthesizes, in a matrix panel (cases × metrics), the strength of the defender effect over the four crises analyzed. The colors indicate favorability to the containment of the scapegoat mechanism: green = conditions more favorable to defense (anti-mimetic); red = conditions more favorable to lynching (dominant accusatory mimesis); middle tones signal intermediate positions. To preserve legibility, each column is normalized by column (min–max or z-score), and the metrics whose "best" is a smaller value (e.g., delay) have the scale inverted, so that "green" always keeps the meaning "better for the defense."

Displayed metrics (columns):

1. **Defense Delay (days) — difference between the peak of accusers and the peak of defenders. Smaller = better (early defense).**
2. **Mean D/A — mean daily ratio Defenders/Accusers. Larger = better (balance/defensive primacy).**
3. **Mean Polarization — mean of (A−D)/(A+D)(A−D)/(A+D)(A−D)/(A+D). Values close to 0 = better (plurality, break of unanimity).**
4. **Early slope (Defense) — initial speed of mobilization of defenders. Larger = better.**
5. **Early slope (Accusation) — initial speed of the accusatory surge. Smaller = better (less mimetic "shock").**
6. **Engagement RT (D/A) — ratio of mean retweets (defense/accusation). Larger = better (reach).**
7. **Engagement Likes (D/A) — ratio of mean likes (defense/accusation). Larger = better (approval).**

How to read the figure (reading paths):
(i) Timing: observe Defense Delay first. Green cells indicate early entry of defenders; red, late. This criterion separates containment cases from sacrifice cases.
(ii) Balance: check mean D/A. Green implies competitive (or dominant) defense on key days; red points to accusatory predominance.
(iii) Mimetic cohesion: check mean Polarization. Values close to zero indicate plurality and loss

of unanimity; high and positive values indicate accusatory unanimity.
(iv) Attention/algorithm: assess Engagement RT (D/A) and Likes (D/A). High likes for the defense mean symbolic approval; high RTs indicate capillarity (the defense circulates, not just pleases).
(v) Initial speeds: compare Early slope of defense (desirable high) and of accusation (desirable low), because they reveal the initial moral inertia of the field.

Typical interpretive patterns:
● Green delay + Polarization ~0 + green mean D/A → defense interrupts the formation of unanimity; tendency toward containment.

● Red delay + low D/A + high Polarization → signature of digital lynching; defense, when it appears, is post facto (late rehabilitation).

● Green Likes (D/A) and yellow RT (D/A) → defense convinces, but circulates less; when both are green, there is a counter-narrative with approval and reach.

Comparative findings (examples anchored in the matrix):
● Eduardo Bueno: columns of Delay, D/A, Polarization and Likes (D/A) tend toward green ⇒ early defense, low accusatory cohesion and social approval; containment profile.

● Wagner Schwartz (La Bête): intermediate Delay, D/A and Polarization migrating to green, favorable Likes (D/A) ⇒ reactive but effective defense, with partial dissolution of the panic.

● Monark: red Delay, low D/A, high Polarization, green Likes (D/A) only late ⇒ consummated sacrifice with later rehabilitation.

● Karol Conká: systematic red in Delay, D/A and Polarization, in addition to weak defensive engagement ⇒ complete sacrifice.

Robustness note: the patterns remain when varying temporal granularity (daily/hourly), when applying alternative smoothing windows, and when removing long-tail accounts, preserving the relative ordering of cases. In sum, Figure 4.5D offers an argumentative synthesis panel: when time, balance, and attention converge in favor of the defense (green), the scapegoat mechanism does not close into unanimity; when they converge against (red), the process ritualizes the sacrifice.

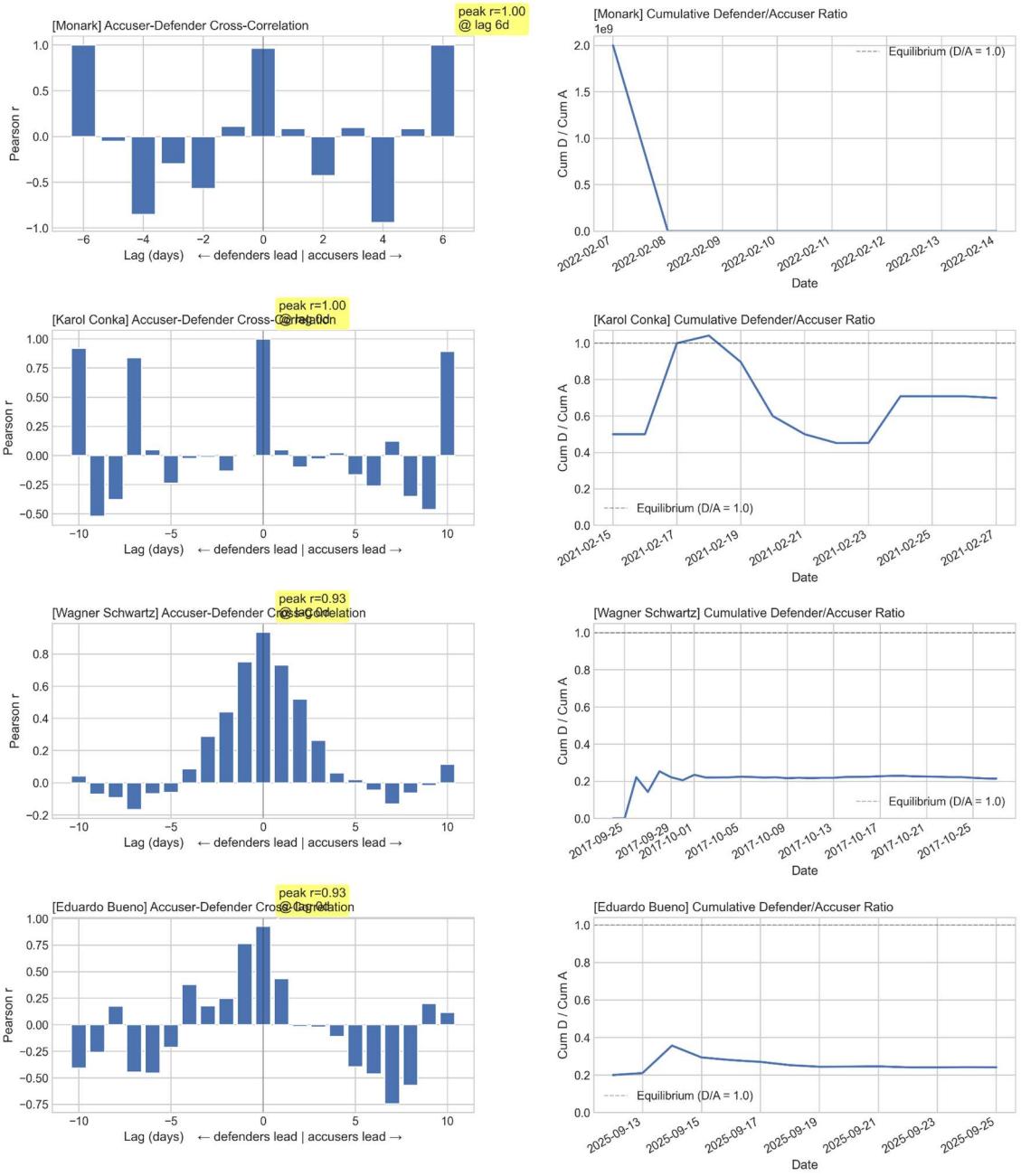
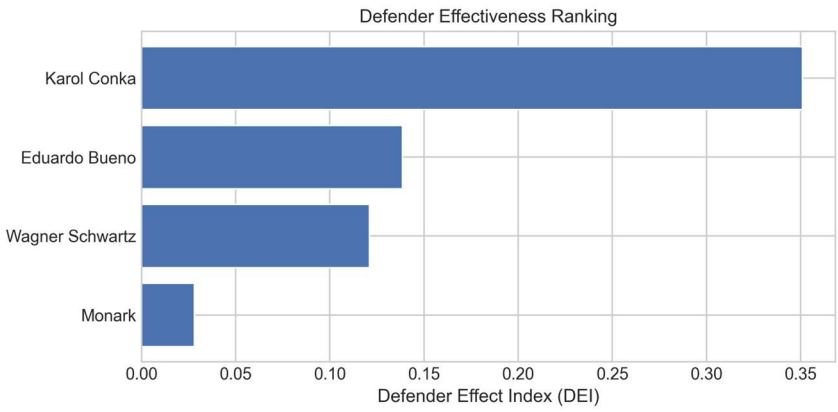


Figure 4.5n — Cross-correlation Accusers ↔ Defenders (lag). Reading (Fig. 4.5n).

- **Peak at negative lag: defenders anticipate the surge (preventive).**
- **Peak ≈ 0: simultaneous/immediate reactive defense.**
- **Peak at positive lag: late/post-facto defense.**
- **Cumulative D/A ratio over time (∑D/∑A): crossing 1 early ⇒ containment; remaining <1 ⇒ consummated (or near) lynching.**



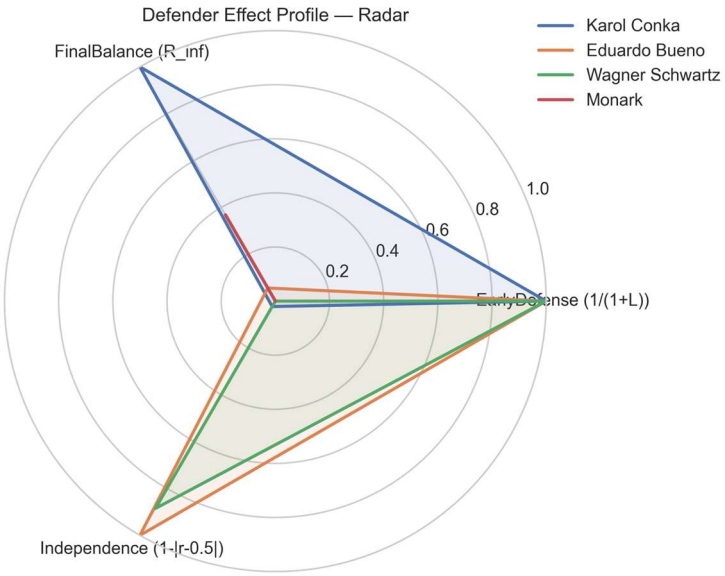


Figure 4.o/p — Cumulative D/A ratio over time and Defender Effect Index (DEI) — ranking and radar.
Reading (Fig. 4.5o/p). The DEI is high when the defense enters early (lag ≤ 0), is relatively independent of the accusatory surge (correlation near 0.5) and achieves good accumulated balance (high R∞R_\inftyR∞).

4.5.2. Results by case (summary narratives)

4.5.2.1 Monark

- **rapid, high accusatory peak; late and smaller defense.**
- **D/A < 1 almost always; positive and stable polarization ⇒ accusatory unanimity.**
- **defenders with more likes per tweet (post-climax), a sign of late rehabilitation; RTs do not surpass accusers at the height.**
- **peak at positive lag ⇒ defense after the accusatory peak.**

- **does not cross 1 ⇒ defense did not offset the accusatory accumulation.**
- **medium/low DEI ⇒ counter-mimesis insufficient to prevent the sacrifice (removal from the program).**

Girardian reading. Complete sacrificial cycle; only after the rite does differentiation emerge that allows defending without coercion from the chorus.

4.5.2.2 Karol Conká

- **instantaneous explosion (TV + networks); residual defense.**
- **D/A ≈ 0; polarization ~ +1 ⇒ total unanimity.**
- **accusers dominate likes and RTs.**
- **low correlation and positive lag; defense arrives late, after the televised rite (elimination).**
- **flattened curve (≈ 0).**
- **very low DEI ⇒ paradigmatic sacrifice.**

Girardian reading. Control-case of digital sacrifice; defense appears only after the purgation.

4.5.2.3 Wagner Schwartz (La Bête)

- **initial moral panic; then, cultural defense (journalists/artists) that persists.**
- **D/A rises late; polarization falls to 0 ⇒ loss of unanimity.**
- **start with larger RTs for accusers; later, larger likes for defenders.**
- **peak at lag +1/+2 ⇒ reactive but effective defense.**
- **approaches 1 in the post-crisis.**
- **medium DEI ⇒ revelation of the mechanism and partial dissolution.**

Girardian reading. Cultural authorities expose the mechanism and reopen differentiation.

4.5.2.4 Eduardo Bueno

- **accusation does not dominate; defense arises early and keeps pace.**
- **D/A ≈ 1 (at times >1); polarization → 0 early.**
- **defenders competitive (likes) and relevant in RTs.**
- **lag ~0/negative ⇒ preventive/simultaneous defense.**
- **crosses 1 relatively early.**
- **high DEI ⇒ anti-mimetic containment (no full sacrifice).**

Girardian reading. Counter-mimesis prevents unanimity; there is no conclusive sacrificial rite.

4.5.3. Comparative synthesis

Caso	Lag (timing da defesa)	D/A cumulativa final	Polarização média	Engajamento defensivo	DEI	Desfecho
Karol Conká	Muito tardia (lag positivo alto)	≈ 0	Alta / unanimidade acusatória	Fraco (likes ≪ RTs acusatórios)	Muito baixo	Sacrifício completo
Monark	Tardia (lag positivo)	< 1	Alta / estável	Likes > acusadores (pós-crise)	Médio / baixo	Sacrifício com reabilitação tardia
Wagner Schwartz – La Bête	Reativa curta (lag +1/+2)	≈ 1	Cai → 0	Likes defensivos > acusadores (tarde)	Médio	Dissolução parcial
Eduardo Bueno	Precoce / simultânea (lag 0/negativo)	> 1 (cedo)	Baixa / ≈ 0	Forte (likes e RTs competitivos)	Alto	Contenção anti-mimética

Figure 4.5q — Comparative summary of key indicators. Textual summary.

- **Timing of defense: Karol Conká (very late) < Monark (late) < La Bête (short reactive) < Eduardo Bueno (early/simultaneous).**
- **Final cumulative D/A: Karol (≈0) < Monark (<1) < La Bête (~1) < Eduardo (>1 early).**
- **Mean polarization: Karol/Monark high; La Bête drops to 0; Eduardo low from early on.**
- **Defensive engagement: Karol weak; Monark/La Bête likes > accusers (late); Eduardo competitive also in RTs.**
- **DEI: Eduardo > La Bête > Monark > Karol.**
- **Outcome: Complete sacrifice (Karol) → Sacrifice with late rehabilitation (Monark) → Partial dissolution (La Bête) → Containment (Eduardo).**

4.5.4. Theoretical and practical implications

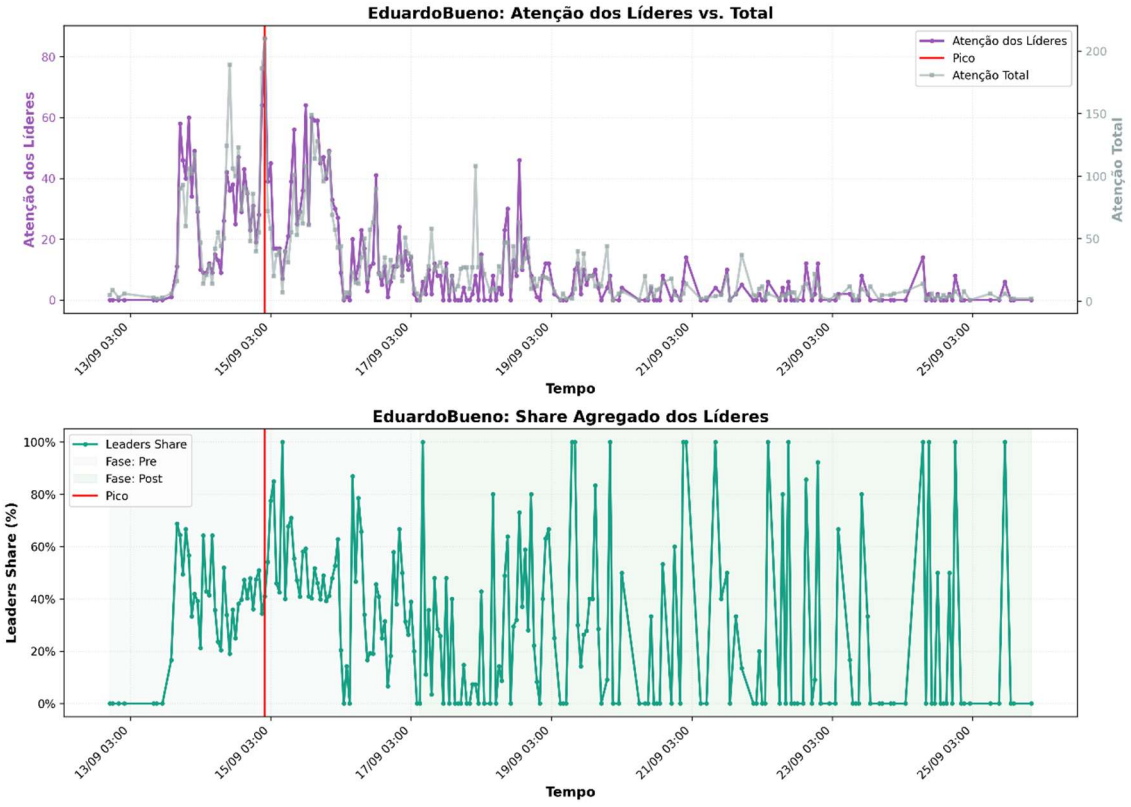
- **Mimetic theory. Defenders operate as a force of differentiation, breaking the indifferentiation that sustains the rite. Early and legitimized entries reduce the coherence of the accusatory chorus.**
- **Empirical signature.**
    - **Lag ≤ 0 (X5) + cumulative D/A → 1 (X6) + high DEI (X7) ⇒ anti-sacrifice.**

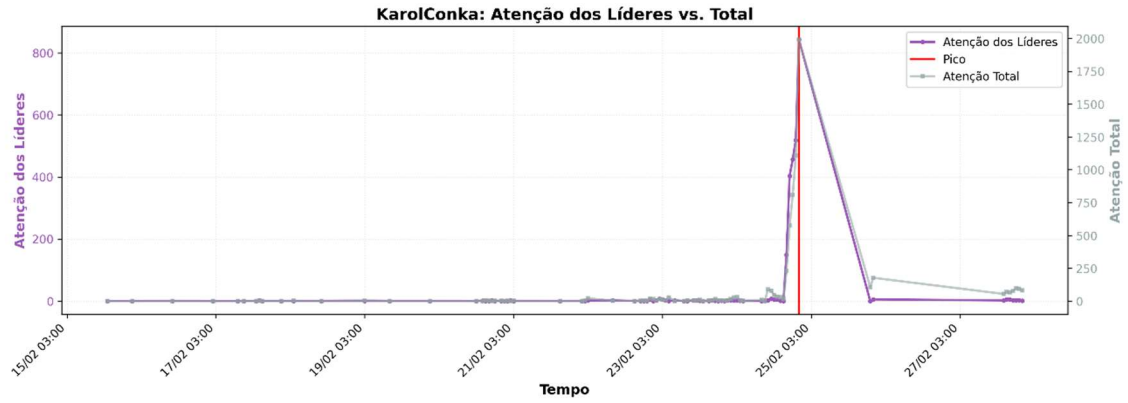
- **Lag » 0 + cumulative D/A ≪ 1 + high polarization (X2) ⇒ complete sacrifice.**

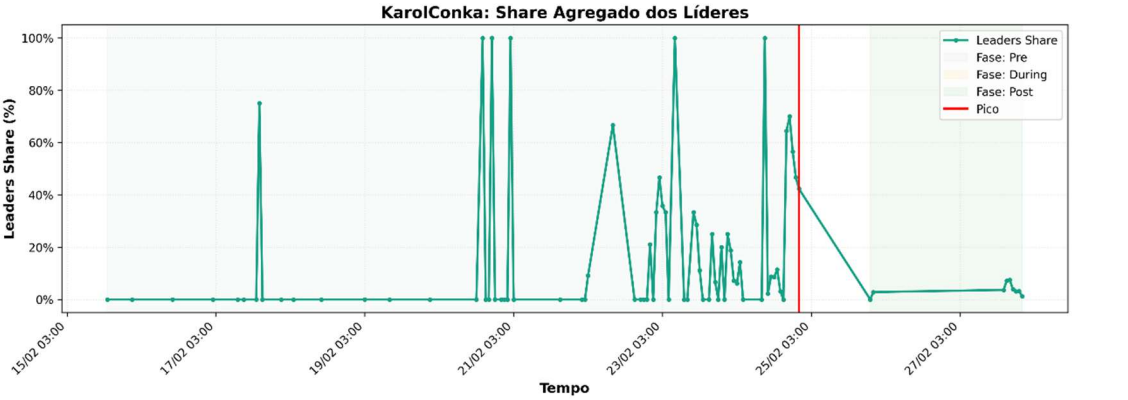
- **Application. Monitor (in near real time) lag, D/A and polarization as response triggers (raise visibility of responsible counter-narratives, insert context, create sharing frictions at the peak).**

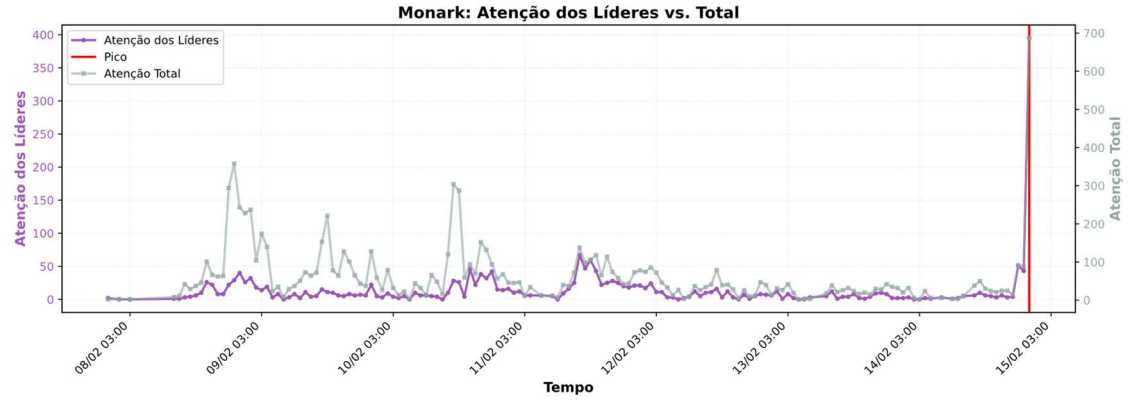
4.6 Emergence of leadership

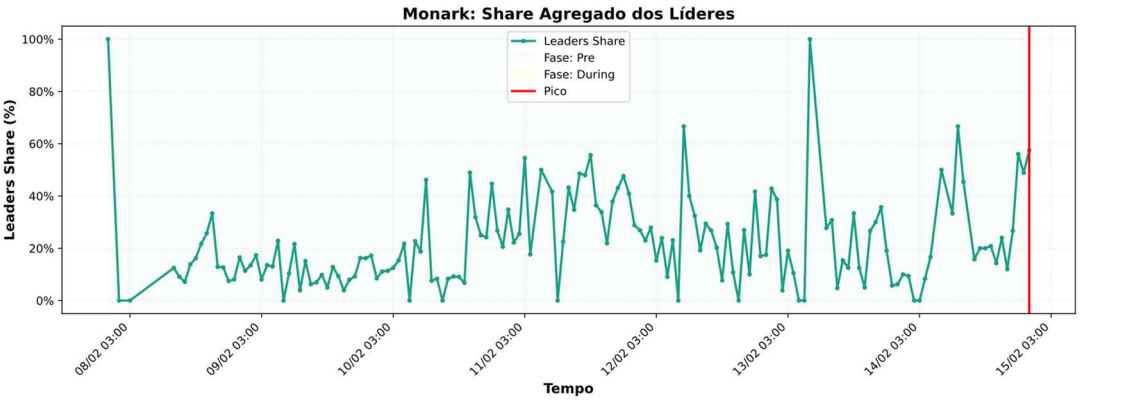
In the empirical data, actors with high PageRank/Betweenness often initiate or amplify accusatory frames (Fig. 4.6A). In the ABM, leaders emerge endogenously when tension surpasses thresholds, catalyzing the focus on the victim (Fig. 4.6B). This supports H3: the leader is not (necessarily) the final hub — they are the catalyst that organizes attention against the target.

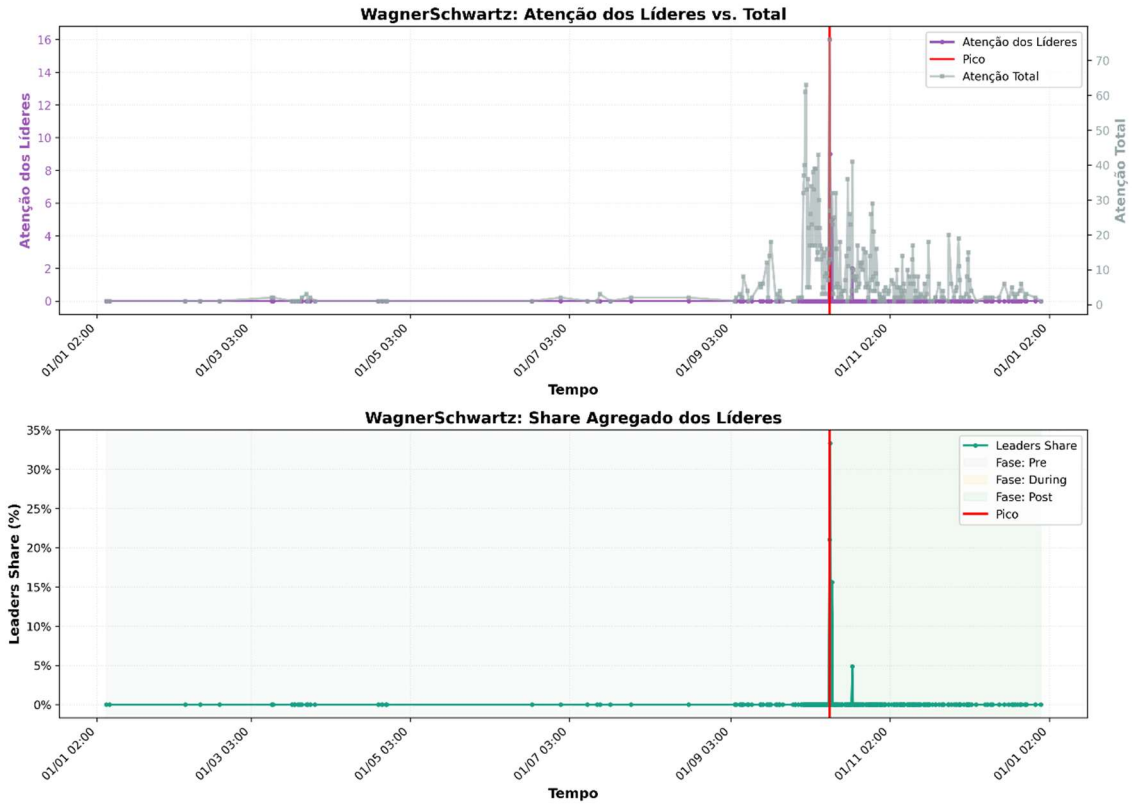



KarolConka: Atenção dos Líderes vs. Total
Atenção dos Líderes
Pico
Atenção Total
Atenção dos Líderes
Atenção Total
Tempo


KarolConka: Share Agregado dos Líderes
Leaders Share
Fase: Pre
Fase: During
Fase: Post
Pico
Leaders Share (%)
Tempo

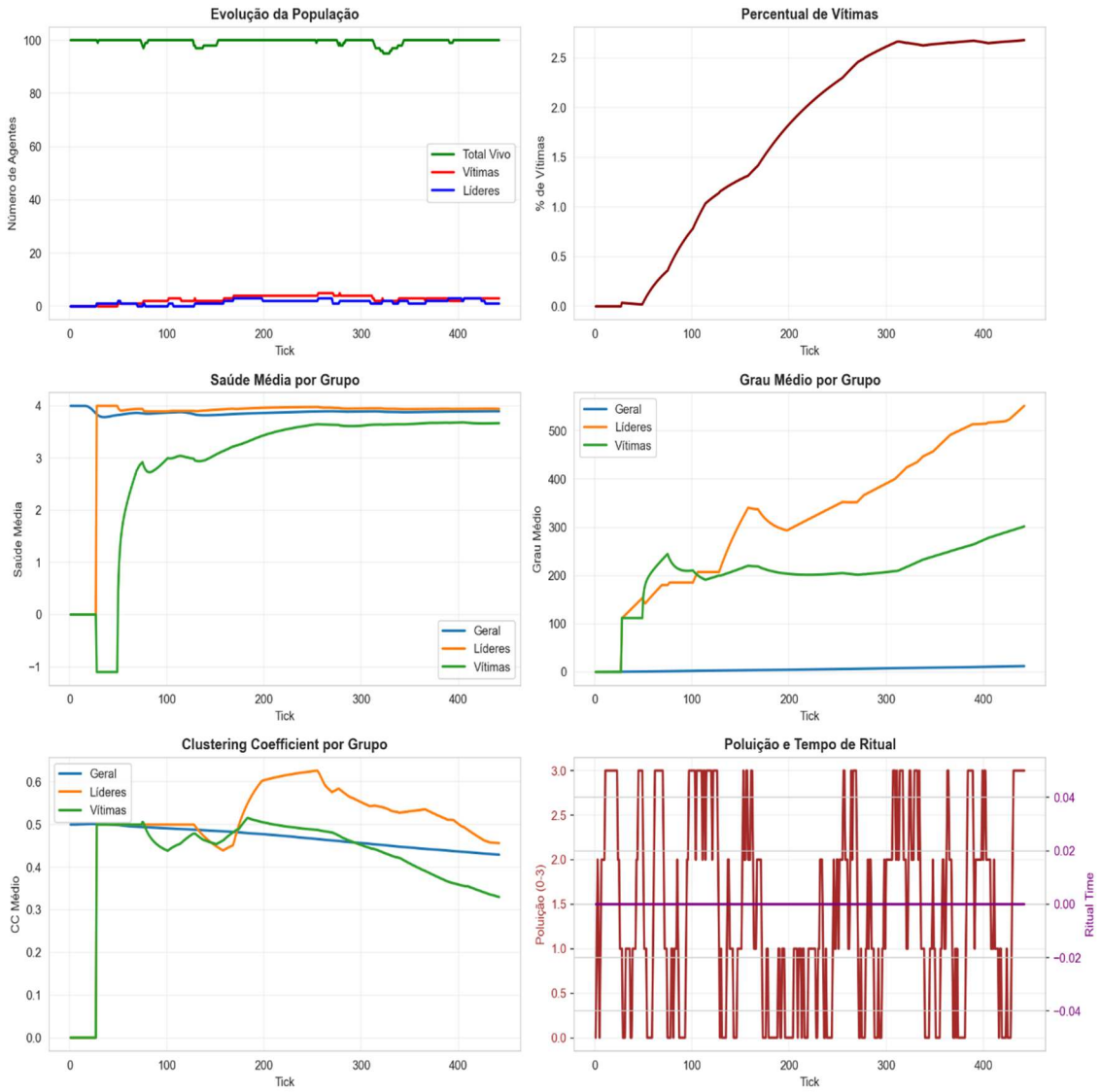

Monark: Atenção dos Líderes vs. Total
Atenção dos Líderes
Pico
Atenção Total
Atenção dos Líderes
Atenção Total
Tempo


Monark: Share Agregado dos Líderes
Leaders Share
Fase: Pre
Fase: During
Pico
Leaders Share (%)
Tempo



4.7 Empirical ↔ ABM distribution validation

We compared distributions of in-degree, component size, and centralities (Fig. 4.7). In multiple cases, the discrepancies are not statistically significant at usual levels, or remain within small/medium effects. This result reinforces that the ABM captures the shapes and relative scale of the observed patterns.



4.8 Synthesis

The Empirical ↔ ABM set → (i) concentrated peaks, (ii) extreme inequality of attention, (iii) star centered on the victim, (iv) accusatory homophily, (v) emerging leadership, and (vi) distributional similarity — constitutes convergent validation of the mimetic mechanism. We do not claim deterministic "proof"; we maintain that the signature pattern predicted by the ABM appears in real data in a replicable and robust way.

The patterns found support the mimetic hypotheses: accusatory convergence that culminates in the structural expulsion of the victim (H1), crossing and reconfiguration of communities at the climax (H2) and relief with lasting sacrificial residue (H3). The damping role of skeptics (H4) offers operational signals for alert dashboards: observing the joint rise of centralization and Gini constitutes an early indicator of crisis.

5. Discussion

5.1 What the results mean in mimetic terms

The empirical and synthetic signatures — joint rise of centralization and Gini before the peak, drop in assortativity at the climax and relief with symbolic inertia afterward — are compatible

with the reading of a mimetic crisis: accusation becomes imitable behavior, converges on a target and reorganizes ties around leaders. The fact that the same metrics work on real data and on the ABM indicates that the scapegoat mechanism leaves detectable and comparable structural markers, reducing the distance between high-level theory (Girard) and operational network measures.

5.2 Alternative explanations and complementarity

The observed patterns can also be partially explained by:
● Homophily and community: peaks may result from coordination within modules. However, the reduction of assortativity at the climax suggests a transversal movement that exceeds stable bubbles.
● Attention/agenda-setting: trigger news and external media amplify the target's visibility. Our ABM captures this via the global stressor ((\beta_2)), without suppressing the imitative dynamics.
● Product/algorithm effects: recommendations and affordances (retweet, reply) facilitate diffusion. Even without explicitly modeling the algorithm, rewiring and salience reproduce functional effects (hubs, star-shape). Conclusion: accusatory mimetism does not exclude these forces; it organizes them around a processual sequence (tension → accusation → rite → reconfiguration).

5.3 Internal and external validity

Internal: the temporal correspondence between empirical and simulated metrics, aligned to the peak/rite, reinforces the processual interpretation. Temporal placebos (Section 4.6) reduce the risk of seasonality artifacts.
External: we analyzed four Brazilian cases; generalization to other domains (YouTube, TikTok, forums) is plausible, but requires replication, since distinct affordances can alter layer weights (retweet vs. reply) and the speed of the mimetic cycle.

5.4 Practical implications (monitoring and mitigation)

The results suggest operational rules for alert dashboards:
● Rule R1 (Accumulation signal): if centralization (in-degree) and Gini (user/tweet) rise together for (k) short windows (e.g., 3–6 hours), trigger "probable sacrificial convergence" alert.
● R2 (Crossing of bubbles): rapid drop in assortativity by stance indicates that the crisis overflows communities; prioritize response.
● R3 (Risk of symbolic inertia): after the peak, persistently high top-k shares imply memory of the case — requires follow-up and de-escalation messaging.

Design/moderation levers inspired by the role of skeptics:
● Increase the visibility of qualified counter-narratives (reliable skeptics) in pre-peak windows.
● Contextual friction on low-cost actions (retweet/quote) when R1+R2 are active.
● Post-rite decentralization: recommendations that diversify foci after the climax reduce symbolic inertia.

5.5 Risks and ethics

Detecting signatures of sacrificial convergence can reduce harm (harassment, doxxing), but it brings risks:
● Overreach: automatic actions can silence legitimate criticism.

● Bias in the data: collections centered on the victim's name reinforce star-shape.
● Labeling: stance errors (irony, ambiguity) can distort metrics. Mitigations: transparency about thresholds, periodic auditing of label noise, counterfactual analyses (e.g., metrics without replies/without retweets), and a focus on networks/structures, not on individual profiles.

5.6 Limitations

Selection bias: queries anchored in the target tend to capture star-shaped graphs; we expanded replies/mentions, but did not eliminate the bias. Edge-based labeling: improves realism, but suffers from sarcasm and code-switching. Parsimonious ABM: We adopted a parsimonious ABM, in which only the mechanisms strictly necessary to reproduce the empirical patterns are included. We prioritized a few parameters with substantive interpretation (agent stances, global stressor and local rules of tension transfer, leader emergence and victim selection), avoiding additional weights or memories when they did not prove indispensable. This choice reduces the risk of overfitting, facilitates validation and improves the interpretability of the results. Equifinality: multiple parametric combinations produce similar series; we report solution fronts instead of a "true" ((\theta)). Temporal anchor: aligning everything to the peak/rite facilitates comparison, but can mask relevant pre-histories (seeds of the crisis).

5.7 Research agenda

Algorithm modeling: couple the ABM to a simplified recommender (e.g., exposure network) to test how curation alters R1–R3. Learning and heterogeneity: introduce light Q-learning or a reinforcement rule for leaders/skeptics; heterogeneous initial degrees. Multi-platform data: replicate in contexts with distinct affordances (threads, stitches, duets). A/B interventions: simulate and, when possible, evaluate in situ frictions, highlighting of skeptics and post-rite diversification. Symbolic indicators: operationalize "memory" with textual signals (the case name as a label for future conflicts) and test its correlation with persistence of top-k.

5.8 Consolidated contribution

The article delivers: (i) an operational vocabulary for mimetic crises (Gini, centralization, assortativity, top-k, peaks), (ii) a temporally anchored ABM↔data validation protocol, and (iii) practical guidelines for early detection and mitigation. By connecting mimetic theory and network science with metrics that work in real and synthetic data, we open the way for predictive dashboards and for an ethics of moderation based on processes, not only on content.

---

6. Conclusions and Perspectives

This article proposed that episodes of "cancellation" on social networks can be understood as mimetic crises with detectable structural signatures. By articulating mimetic theory (tension → accusation → rite → reconfiguration) with bridge-metrics (Gini, centralization, assortativity, top-k, peak indicators) and a parsimonious small-world ABM, we showed that (i) there is accusatory convergence pre-peak (joint rise of centralization and Gini), (ii) bubble crossing occurs at the climax (drop in assortativity) and (iii) relief with symbolic inertia is verified post-rite (partial retreat of the metrics and persistence of top-k).
In four Brazilian cases, the empirical series aligned qualitatively with the synthetic series after

multi-metric calibration, supporting the thesis that the scapegoat mechanism leaves measurable marks on the topology and dynamics of graphs.

The central contributions are:
- **an operational vocabulary that brings the humanities and network science closer;**
- **an ABM↔data validation protocol anchored in event time;**
- **evidence of the damping role of skeptic agents;**
- **operational rules for early detection and mitigation in monitoring dashboards.**

We recognize limitations (collection bias, edge-based labeling, equifinality and absence of explicit modeling of the recommender), already discussed in Section 5, and indicate a continuity agenda: coupling the ABM to simple models of algorithmic exposure, introducing more realistic learning and heterogeneity, replicating across platforms and testing A/B interventions (frictions, highlighting skeptics, post-rite diversification).

From an applied point of view, our signatures provide practical triggers for alerts and responses that reduce harm without suppressing legitimate debate. From a scientific point of view, the approach reinforces that Girardian concepts can be translated into reproducible variables and metrics, favoring comparisons between cases and domains.
We hope that this work serves as a basis for predictive dashboards, comparative studies and process-oriented moderation policies, contributing to a more resilient and responsible information ecosystem.

References (ABNT NBR 6023:2018)

(1)
VAN BAVEL, Jay J.; PARRA, Daniel I.; PERKINS, Andrew M.; BRADY, William J.; CROCKETT, Molly J. Social media and morality. Annual Review of Psychology, Palo Alto, v. 75, p. 1–28, 2024. Available at: https://doi.org/10.1146/annurev-psych-022123-110258. Accessed on: Oct. 04, 2025.
BRADY, William J.; WILLS, Julian A.; JOST, John T.; TUCKER, Joshua A.; VAN BAVEL, Jay J. How social learning amplifies moral outrage expression in online social networks. Science Advances, Washington, v. 7, n. 33, eabe5641, 2021. Available at: https://doi.org/10.1126/sciadv.abe5641. Accessed on: Oct. 04, 2025.
HUSZÁR, Ferenc; KAKIMOTO, Sofiane; SIVAKUMAR, Shashi; CORBETT-DAVIES, Sam; MATTSON, James; GHOSH, Sandeep; et al. Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences (PNAS), Washington, v. 119, n. 1, e2025334119, 2022. Available at: https://doi.org/10.1073/pnas.2025334119. Accessed on: Oct. 04, 2025.

(2)
GIRARD, René. Violence and the sacred. Translation by Martha Conceição Gambini. 2nd ed. São Paulo: Editora Paz e Terra, 1990.

(3)
McLOUGHLIN, Killian L.; KAISER, Benjamin; GOOLSBEE, Aden; KLONICK, Kate; BRADY, William J.; CROCKETT, Molly J. Misinformation exploits outrage to spread online. Science, Washington, v. 385, n. 6710, p. 115–120, 2024. Available at: https://doi.org/10.1126/science.adl2829. Accessed on: Oct. 04, 2025.

(4)
FREEMAN, Linton C. Centrality in social networks: conceptual clarification. Social Networks, [S.l.], v. 1, n. 3, p. 215–239, 1979.
WASSERMAN, Stanley; FAUST, Katherine. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1994.

(5)
WATTS, Duncan J.; STROGATZ, Steven H. Collective dynamics of 'small-world' networks. Nature, London, v. 393, n. 6684, p. 440–442, 1998. DOI: https://doi.org/10.1038/30918

(6)
https://github.com/carlospaes120/scapegoat

TECHNICAL NOTE — Freeman Centralization

Freeman's centralization measures, at the network level, how unequal the distribution of a centrality measure (for example, degree) is when compared to the limit case of a star of the same size (FREEMAN, 1979). For the in-degree version in directed graphs, we use:

[
C_D^{in}(G) = \frac{\sum_i \left(k_{\max}^{in} - k_i^{in}\right)}{(n - 1)^2}
]

in which (k_i^{in}) is the in-degree of node (i) and (k_{\max}^{in}) is the maximum observed in-degree. The denominator ((n - 1)^2) corresponds to the maximum possible value of the sum of differences in a directed star (one central node receiving all edges). Values close to 1 indicate a star-like structure; close to 0, a more homogeneous distribution. See also the exposition in WASSERMAN; FAUST (1994).

Suggested in-text citation: "...we measured structural inequality via Freeman centralization (FREEMAN, 1979; WASSERMAN; FAUST, 1994)."