

SocialFiVis: A Visual Analytics Sandbox for LLM-Grounded Multi-Agent Simulation in Social Finance

Yi-Fan Cao , Qing Shi , Liangwei Wang , Leo Yu-Ho Lo , Lin Chen , Yuzi Han ,
Yang Wang , and Kani Chen 

Abstract—The emergence of social finance (SocialFi) transforms online communities into complex socio-economic systems. Within these spaces, collective decisions shape a “digital commons” characterized by social capital (e.g., community trust) and financial health (e.g., market liquidity). Governing such hybrid ecosystems is challenging because real-world interventions are costly and irreversible. While counterfactual simulation is essential for exploring alternative governance strategies, existing approaches fail to capture the complex, non-linear interplay between governance rules, individual behaviors, and emergent economic outcomes. To systematically unpack this complexity, we operationalize the Institutional Analysis and Development (IAD) framework as our theoretical foundation, synthesizing prior literature with insights from formative expert interviews. Built on this framework, we present *SocialFiVis*, an IAD-embedded visual analytics sandbox. It introduces a robust model to quantify the dual-track digital commons, coupled with a two-phase simulation engine. This engine combines LLM-derived personas with a mechanism-guided *Perception–Reasoning–Action* (PRA) runtime to simulate heterogeneous, context-aware agents empirically grounded in the retained messaging cohort. A hierarchical multi-view interface with interpretable reasoning pathways enables stakeholders to explore counterfactual policies and trace system-level outcomes back to individual behavioral rationales. We validate *SocialFiVis* through case studies, expert interviews, and a user study. Results demonstrate that *SocialFiVis* supports fine-grained behavioral attribution and helps explain emergent phenomena such as the structural decoupling of social capital and the resilience of messaging members under localized governance shocks.

Index Terms—LLM-grounded agent simulation, SocialFi, counterfactual reasoning, digital commons governance

1 INTRODUCTION

Traditional online communities are gradually evolving from simple social networks into socio-economic systems that incorporate decentralized finance (DeFi) attributes [36]. Known as SocialFi, this emerging paradigm assigns economic value to digital identities and social relationships, transforming participants into co-owners and governance decision-makers of community assets [30]. This dynamic is particularly visible in Non-Fungible Token (NFT) communities [9], the empirical setting examined in this work. In these ecosystems, collective decisions shape what the community jointly owns and manages as a “digital commons” [44]. We characterize this commons through two tightly coupled dimensions: *social capital*, reflected in participation, consensus, and trust [14]; and *financial health*, reflected in floor price, liquidity, and the number of holders [12, 30]. These dimensions co-evolve; strong social capital can mitigate market panic, while healthy financial conditions sustain participation incentives [11].

Governing this intertwined digital commons is exceptionally difficult. Communities face rapid macro-market swings and sentiment shifts [39], leaving *community operators* (e.g., project leads and moderators) to rely on limited experience or intuition when steering governance [22]. Furthermore, real-world interventions (e.g., modifying royalty fees or token-gating access) are costly and risky. Testing these governance changes typically requires real budgets, and poorly designed policies can trigger irreversible collapse of both social trust and financial liquidity [44]. Consequently, a visual sandbox is essential for operators to explore alternative governance policies through retrospective counter-

factual simulations without consuming actual resources.

However, existing approaches struggle to capture the complex, non-linear interplay between governance rules, individual behaviors, and emergent economic outcomes [74]. Traditional *Agent-Based Modeling* (ABM) provides a principled way to model emergence, but its hand-specified behavioral rules often fall short in reflecting the nuanced, text-driven social consensus and speculative behaviors in online communities. While Large Language Model (LLM)-based agents can generate plausible reasoning traces, current simulations mostly focus on purely conversational networks [45] or isolated strategic tasks [16]. Crucially, they lack an explicit institutional structure to translate individual motives into collective outcomes. To bridge this gap, we draw on the Institutional Analysis and Development (IAD) framework [55], a classic theory explaining how governance rules shape collective behavior and outcomes. We operationalize this framework as the theoretical backbone of our system, synthesizing prior literature with insights from a formative expert study. Yet, realizing this construct within a visual analytics system presents three challenges:

First, defining and quantifying the digital commons is inherently complex. Measuring abstract properties like social capital and financial health in SocialFi without standard metrics is non-trivial. Simple data aggregations can easily mask underlying liquidity crises or be skewed by localized anomalies, necessitating robust mathematical modeling [59]. Second, achieving ecological validity in multi-agent simulations poses severe algorithmic hurdles [15, 16]. It requires systematically extracting data-driven personas and endowing them with authentic social and financial reasoning. Crucially, the simulation architecture must explicitly model how heterogeneous individual behaviors adapt to governance rules and aggregate into emergent economic outcomes, thereby capturing their complex, non-linear interplay. Third, designing intuitive visual interactions for uncertainty-aware governance analysis is demanding. The system must facilitate flexible hierarchical exploration, seamlessly connecting collective policy outcomes with the understandable reasoning pathways of individual agents. Bridging these macro consequences and micro behaviors, without instilling a false sense of precision, is vital for effective decision-making [71].

To address these challenges, we conducted in-depth interviews with domain experts to distill design requirements and define the evaluation metrics for the digital commons. We propose a soft-penalty geometric

- Yi-Fan Cao, Leo Yu-Ho Lo, Yuzi Han, and Kani Chen are with The Hong Kong University of Science and Technology (HKUST). E-mail: caoyifan@ust.hk, yhlload@cse.ust.hk, yhanam@connect.ust.hk, and makchen@ust.hk.
- Qing Shi and Liangwei Wang are with The Hong Kong University of Science and Technology (Guangzhou), HKUST(GZ). E-mail: {qshi118, lwang344}@connect.hkust-gz.edu.cn.
- Lin Chen is with Northeastern University. E-mail: l.chen2@northeastern.edu.
- Yang Wang is with The University of Hong Kong (HKU). E-mail: yang.wang@hku.hk.

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxx

blend model to robustly quantify social capital and financial health. We then developed a two-phase multi-agent simulation engine. Grounded in a seven-dimensional persona codebook, it combines expert-validated, LLM-derived personas with a mechanism-guided runtime to simulate agent behaviors and communication networks. Finally, we integrate these models into *SocialFiVis*, a multi-view visual analytics system that connects the IAD-based governance parameters with the simulated counterfactual outcomes. The system comprises an *Event Timeline View* for specifying governance interventions, a *Persona View* for examining heterogeneous agent profiles, a *Behavior View* for revealing behavioral divergence across agent groups, and a *Communication Network* capturing agent interactions alongside interpretable reasoning pathways to facilitate fine-grained behavior attribution.

Our primary contributions are summarized as follows:

- **IAD-Driven Socio-Economic Modeling:** Based on the IAD framework, we formalize SocialFi communities as a dual-track digital commons and operationalize its social and financial dimensions through robust quantitative metrics.
- **LLM-Grounded Visual Analytics Sandbox:** We develop *SocialFiVis*, a multi-view sandbox integrating a two-phase simulation engine that combines LLM-derived, expert-validated personas with a mechanism-guided runtime to model how governance rules shape individual behaviors and collective outcomes.
- **Empirical Insights and Validation:** Through case studies and domain expert interviews, we demonstrate the system’s efficacy in revealing critical governance trade-offs among messaging members.

2 RELATED WORK

We situate this research at the intersection of three areas: the governance of SocialFi digital commons, agent-based social simulation, and visualization for counterfactual reasoning.

2.1 SocialFi and Digital Commons Governance

SocialFi represents the convergence of social networks and tokenized economies, where digital assets (e.g., NFTs) carry both economic value and governance rights, transforming users into co-owners and decision-makers of community resources [30, 58]. In these ecosystems, the shared resources managed by participants constitute a “digital commons” featuring two tightly coupled dimensions: *social capital*, commonly characterized by participation, consensus, and trust [14, 37]; and *financial health*, typically comprising asset floor price, liquidity, and holder count [12] (both quantified in Sec. 5.2). Empirical evidence shows that sustained engagement in such communities relies heavily on social identity formation rather than purely financial speculation, necessitating governance structures that balance both objectives [9]. To analyze this dual-track commons, we draw upon Elinor Ostrom’s Institutional Analysis and Development (IAD) framework, which explains how governance rules, actors, and shared resources interact to shape collective outcomes [48, 55]. While the IAD framework has been widely applied to online creation communities [51, 60], existing quantification efforts in SocialFi predominantly rely on ad-hoc, isolated indicators (e.g., tracking price or sentiment independently). These approaches fail to capture the systemic vulnerability where the collapse of one dimension catastrophically impacts the overall ecosystem. Building on the IAD framework, we propose a soft-penalty geometric blend model that robustly quantifies the dual-track digital commons.

2.2 Agent-Based Modeling and LLM-Based Simulation

ABM provides a principled bottom-up framework for studying how individual behavioral rules, under institutional constraints, aggregate into macro-level social dynamics [7, 20, 47]. LLM-based agents have recently extended classical ABM by replacing hand-coded heuristics with language-grounded reasoning, enabling emergent sociolinguistic behaviors in simulated communities [13, 26, 31, 57]. While they have shown promise as behavioral proxies in socio-economic settings [24, 78], validating simulation fidelity against real-world trajectories remains challenging due to the stochastic nature of LLM outputs [4, 23, 29, 42, 74]. One promising direction is to improve ecological validity by initializing agents with empirically grounded personas derived from com-

munity interaction data [15]. While these approaches better reflect the diversity of real community members, they continue to model agent interactions without explicit institutional structures, focusing solely on conversational exchanges [66, 75] or isolated market behaviors [16]. Consequently, how institutional governance shapes the transition from individual decisions to collective outcomes remains largely unexplored [65]. We address this gap with a two-phase architecture: LLM-derived personas establish heterogeneous agent motives, while a mechanism-guided runtime operationalizes IAD governance rules to systematically bridge this micro-macro divide.

2.3 Visualization for Counterfactual Reasoning

Visualization research for financial technology has matured across distinct analytical tasks. Existing systems predominantly address transaction network analysis [10, 67, 76], market metric monitoring [53], and fraud detection [72, 77]. However, these tools typically lack a governance context, treating transactions as isolated events rather than outcomes of collective decision-making processes. Social visualization research has explored the relationship between online engagement and market dynamics, including sentiment contagion mapping [43], but often overlooks the institutional rules shaping participant behavior. Recent advances in visual causal inference rely heavily on counterfactual reasoning to isolate the effects of historical interventions [8, 25, 40, 63, 70, 71]. Despite these theoretical advances, few systems support policy backtesting with traceability from macro outcomes to micro behavioral rationales [27]. Furthermore, visualizing aggregated simulation outputs presents a documented risk of false precision, as deterministic visual encodings can project unwarranted certainty over stochastic outcomes [3]. Current tools for agent-based simulations typically focus on aggregate outcome presentation rather than enabling drill-down exploration of individual agent decision pathways. We address these limitations through *SocialFiVis*, which integrates IAD-based policy configuration, multi-agent simulation, and hierarchical visual analytics with interpretable reasoning traces for uncertainty-aware counterfactual reasoning.

3 FORMATIVE STUDY

To ground the system in practice, we adopted a participatory design approach centered on *community operators* (project leads and moderators) who diagnose community conditions and compare governance interventions. Analysts, traders, and community members are secondary users who interpret or audit these decisions. We collaborated with three practitioners with hands-on SocialFi community governance experience: E_1 (a former project lead), E_2 (COO of a crypto AI infrastructure project), and E_3 (a core community contributor and institutional liaison). A computational social science and ABM researcher (E_4) served as a methodological co-design advisor. Interviews with E_1 – E_3 surfaced the governance workflows and operational bottlenecks.

3.1 Design Goals

Synthesizing the needs of our practitioners (E_1 – E_3) with methodological input from our simulation advisor (E_4), we distilled three high-level design goals (DGs) to guide our system development.

DG1: Quantify the Coupling Between Social and Financial Dynamics. All practitioners emphasized that social behaviors and financial metrics are deeply intertwined yet are tracked in isolation. E_1 noted that these data streams are “*completely siloed*,” lacking a unified tool to examine their underlying coupling. E_3 echoed this, describing community dynamics and financial outcomes as “*mutually influencing each other*.” The system must integrate and visualize the co-evolution of social capital and financial health within a unified analytical framework.

DG2: Enable Counterfactual Policy Evaluation with Retrospective Grounding. The practitioners expressed a critical need to move from post-hoc analysis toward simulation-based strategy assessment. E_2 described her current workflow as “*entirely reactive*,” while E_1 sought to “*back-test strategies*” rather than rely on intuition. From a modeling perspective, E_4 emphasized that assessing alternative interventions requires a controlled, historically grounded counterfactual

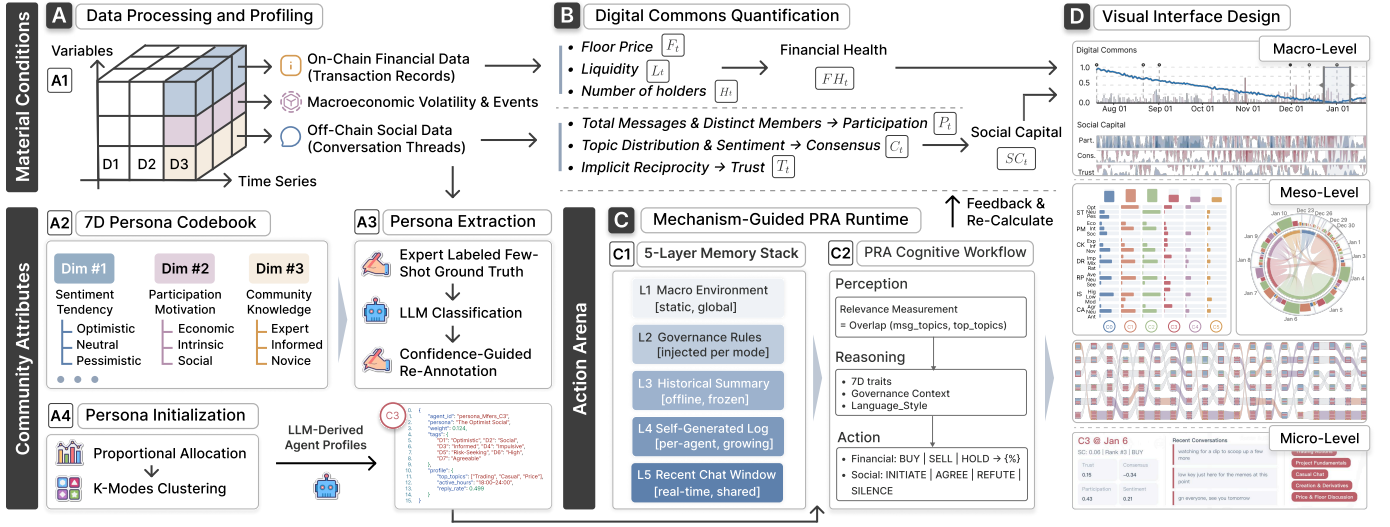


Figure 1: IAD-embedded system overview. (1) Data Processing and Quantification (A, B) establish baseline *Material Conditions* and extract *Community Attributes*. (2) An *Action Arena* (C) simulates agent behaviors under counterfactual policies. (3) The Visual Interface (D) presents *Evaluative Outcomes*. A closed feedback loop (C → B) aggregates simulated micro-actions to iteratively update the *Material Conditions* tick-by-tick.

simulation. The system must therefore enable operators to compare governance policies in a risk-free sandbox.

DG3: Support Heterogeneous User Profiling and Behavioral Attribution. The practitioners stressed that modeling the diverse behavioral patterns within SocialFi networks is critical for understanding market shifts. Rather than treating the community homogeneously, E_2 and E_3 highlighted the need to examine how specific cohorts react differently to market events and governance changes. To operationalize this for simulation, E_4 identified the central technical challenge as “capturing the heterogeneity among community members... to faithfully reproduce macro-level dynamics.” Thus, the system must derive distinct persona types from empirical data and trace how macro-level outcomes emerge from their differentiated behaviors.

3.2 The IAD Framework as an Analytical Lens

To achieve the goal of a unified framework that disentangles social and financial dynamics, we adopted Elinor Ostrom’s IAD framework [55]. We selected this theoretical lens in close consultation with our practitioners because it directly addresses their analytical needs. E_1 noted that Ostrom’s focus on the decentralized governance of common-pool resources naturally aligns with the organizational logic of SocialFi communities. Furthermore, E_2 and E_3 valued the framework’s analytical cycle [65], which supports quantitative attribution by mapping exogenous variables to action situations and evaluating resulting outcomes.

E_4 further emphasized that the strength of the IAD framework lies in its systematic translation of institutional concepts into computational components. This principled mapping allows governance rules, community attributes, and environmental conditions to be operationalized consistently across the two-phase multi-agent simulation.

3.3 Analytical Tasks

Building on the design goals, we refined six analytical tasks (T1–T6) through iterative discussions with E_1 – E_3 . Following the IAD framework, we organize these tasks across three nested analytical scales (macro, meso, and micro). These scales mirror how operators naturally reason about their communities: from diagnosing aggregate outcomes, to identifying cohort-specific reactions, and finally tracing individual motives to overcome the anonymity of on-chain data.

Macro-Level Community Dynamics. Operators seek to understand how governance configurations dictate aggregate outcomes.

- **T1. Temporal Metric Exploration (DG1).** Quantify and visually track the dual-track digital commons (social capital and financial health). The system must support interactive exploration of temporal patterns to ground reasoning about commons evolution.

- **T2. Cross-Metric Policy Impacts (DG2).** Support comparative sandbox analysis to examine the interplay between metrics under varying rules-in-use. This enables users to test alternative governance configurations and observe their systemic effects.

Meso-Level Persona Patterns. Operators investigate how large-scale, heterogeneous participation drives these macro-level shifts.

- **T3. Data-Driven Persona Derivation (DG3).** Derive representative persona types from empirical behavioral and investment patterns, surfacing influential cohorts and their characteristic behaviors to support heterogeneity-aware analysis.
- **T4. Behavior-Outcome Backtesting (DG2, DG3).** Enable retrospective analysis to map specific persona actions to macro-level shifts, uncovering behavioral drivers underlying fluctuations in the digital commons.

Micro-Level Cognitive Pathways. Operators aim to interpret the latent decision-making rationales that shape observed behaviors.

- **T5. Policy-Behavior Differentiation (DG2, DG3).** Reveal how shifts in rules-in-use trigger heterogeneous responses across distinct persona types, exposing specific behavioral nuances that are often obscured in aggregate metrics.
- **T6. Decision Pathway Interpretation (DG1, DG3).** Unpack the persona-conditioned reasoning pathways of agent types to offer an explanatory context for why specific behaviors emerge, bridging actions with underlying rationales.

4 SYSTEM OVERVIEW

Guided by the analytical tasks, we instantiate the IAD framework as an interactive visual analytics sandbox consisting of three tightly coupled modules (Fig. 1). (1) The data processing and quantification module establishes the system’s empirical baseline. It parses heterogeneous on-chain and off-chain data to construct the *Material Conditions*, which quantify the community’s social capital (SC_t) and financial health (FH_t). Meanwhile, it distills the *Attributes of the Community* by clustering users into representative archetypes using a seven-dimensional persona codebook. (2) The two-phase multi-agent simulation engine serves as the *Action Arena*. It pairs LLM-derived personas with a mechanism-guided runtime executing a *Perception–Reasoning–Action* (PRA) pipeline. Conditioned on user-injected counterfactual governance policies (*Rules-in-use*), agents produce heterogeneous social and financial interactions (*Action Situations*). (3) The visual analytics dashboard presents the *Evaluative Outcomes* through coordinated views that support multi-level (macro-meso-micro) policy exploration. Crucially, within any selected time window, these three modules form a

closed-loop co-evolutionary system. Simulated micro-actions are mathematically aggregated to perturb the baseline SC_t and FH_t tick-by-tick, continuously updating the *Material Conditions* for the internal ongoing simulation. This architecture empowers target users to dynamically evaluate the cascading impacts of governance interventions.

5 DATA ANALYSIS

This section details the collection of heterogeneous data and the mathematical formulation of community metrics.

5.1 Data Collection and Preprocessing

We constructed a heterogeneous dataset capturing the digital footprints of SocialFi communities across a ten-month period (Jul. 2022 – Apr. 2023), integrating three core contexts (Fig. 1A1):

Off-Chain Social Data: We collected chat logs from the official communities of two representative SocialFi projects during their respective active phases. *Mfers* [49] was selected as a Western blue-chip project demonstrating sustained market presence, while *Mimic Shhans* [50] represents a prominent Asian project captured during its accelerated growth phase.

On-Chain Transaction Data: Utilizing the NFTGo and OpenSea APIs, we extracted daily financial metrics for both projects, including *floor price* (F_t), *liquidity* (L_t), and the *number of holders* (H_t), where t denotes the daily time step. All values were normalized to a $[0, 1]$ scale.

Macroeconomic Environment: We compiled a chronological dataset of major events impacting the broader SocialFi market (e.g., Federal Reserve rate hikes, the FTX collapse) sourced from authoritative platforms (CoinDesk, Bloomberg). This serves as the global context layer for our simulation.

5.2 Mathematical Modeling of Community Metrics

To translate unstructured behaviors into quantifiable time-series, we formalize social capital (SC_t) and financial health (FH_t) through a 7-dimensional vector space (Fig. 1B). This balanced design prevents index skewing by localized anomalies, providing a robust, multidimensional foundation for subsequent visual encoding.

5.2.1 Social Capital (SC_t) Modeling

Serving as a robust proxy for intangible community cohesion, SC_t quantifies the structural, cognitive, and relational resilience that sustains a project’s long-term viability beyond mere financial speculation. We rigidly map daily SC_t to a $[-1, 1]$ interval. SC_t comprises three metrics:

Participation (P_t) captures structural engagement [38]. Relying solely on message frequency invites spamming by a few hyperactive users [56]. Thus, we incorporate N_t^{unique} (the number of distinct daily senders) alongside total messages (N_t^{msg}). Here μ_{msg} is the community’s historical mean daily message volume, a static baseline for activity fluctuations. A tanh caps activity spikes; the two ratios carry equal weight (high scores need both intensity *and* breadth), and the 0.5 offset centers a typical day near zero:

$$P_t = \tanh \left(\ln \left(1 + \frac{N_t^{\text{msg}}}{\mu_{\text{msg}}} \cdot \frac{N_t^{\text{unique}}}{N_{\text{members}}} \right) - 0.5 \right) \quad (1)$$

Consensus (C_t) deconstructs cognitive consensus into topical and emotional dimensions [6, 28]. For the topical dimension, we apply LDA to obtain a daily distribution over n topics, $p_1, \dots, p_n$, and compute its Shannon entropy $\mathcal{H}_t = -\sum_{i=1}^n p_i \ln(p_i)$. A lower H_t denotes discourse concentrated on few topics (high consensus), whereas entropy approaching the maximum $\ln(n)$ reflects fragmented attention. For the emotional dimension, we compute the standard deviation (σ_t) of daily sentiment scores from a domain-adapted, SocialFi-tuned classifier; a low σ_t implies strong emotional alignment even amid widespread panic. We weight the two dimensions equally, so that high consensus requires both focused discourse and emotional alignment:

$$C_t = 0.5 \cdot \left(1 - 2 \frac{\mathcal{H}_t}{\ln(n)} \right) + 0.5 \cdot (1 - 2\sigma_t) \quad (2)$$

Trust (T_t) adapts to platforms lacking explicit reply structures [2]. We define *implicit reciprocity* (IR_t) as the density of reciprocated directed edges among members, established by sequential, topically related posts within a five-minute sliding window and normalized to $[-1, 1]$. T_t is formalized as an equal-weighted blend of IR_t and the daily mean sentiment polarity (μ_t). This balancing mechanism serves as a safeguard: high trust scores require both constructive engagement and non-negative sentiment, filtering out “flame wars” where turn-taking may be frequent but sentiment polarity is negative:

$$T_t = 0.5 \cdot IR_t + 0.5 \cdot \mu_t \quad (3)$$

To synthesize the overall social capital (SC_t), we first normalize P_t , C_t , and T_t to $[0, 1]$. We then apply a *soft-penalty geometric blend* to formulate the intermediate index SC'_t . By the *arithmetic mean–geometric mean* (AM–GM) inequality, a purely geometric aggregate can over-penalize routine sub-metric variation, shifting SC_t downward relative to arithmetic aggregation and, in our data, toward pessimistic values. We therefore set $\alpha = 0.5$ as a symmetric blend, tempering this bias while retaining geometric weak-link sensitivity to sudden crises:

$$SC'_t = \alpha \cdot \left(\frac{P'_t + C'_t + T'_t}{3} \right) + (1 - \alpha) \cdot \sqrt[3]{P'_t \cdot C'_t \cdot T'_t} \quad (4)$$

where $X'_t = (X_t + 1)/2$ for $X \in \{P, C, T\}$. The result is subsequently mapped back to the diverging $[-1, 1]$ visual space:

$$SC_t = 2 \cdot SC'_t - 1 \quad (5)$$

5.2.2 Financial Health (FH_t) Modeling

Derived from on-chain transaction data (Sec. 5.1), FH_t evaluates market vitality on a strict $[0, 1]$ scale. Informed by domain experts, we synthesize three orthogonal metrics, *floor price* (F_t), *liquidity* (L_t), and *holder count* (H_t). As these metrics are individually normalized and no domain rationale privileges one, we combine them with an *unweighted* geometric mean. The geometric form enforces a “shortboard penalty”: a collapse in any component triggers a non-linear plunge, mitigating the “paper-wealth trap” in which an artificially high floor price masks a liquidity crisis, and keeping isolated anomalies visually salient:

$$FH_t = \sqrt[3]{F_t \cdot L_t \cdot H_t} \quad (6)$$

5.2.3 Sensitivity and Robustness Analysis

We assess the effect of these parameter choices through sensitivity analyses. Sweeping $\alpha \in [0, 1]$ preserves the temporal ordering of SC_t ($\rho \geq 0.991$), showing that the observed SC_t trends are not sensitive to this midpoint setting. Reweighting the three financial inputs preserves the FH_t ranking ($\rho \geq 0.988$). For P_t , C_t and T_t , moderate alternatives around the equal split ($w \in [0.25, 0.75]$) retain substantial downstream SC_t rank agreement ($\rho \geq 0.913$); the P_t centering offset ($c \in [0.3, 0.7]$) is likewise immaterial ($\rho \geq 0.999$). The *Supplement* verifies that these perturbations preserve the key case-level SC_t – FH_t interpretations.

6 TWO-PHASE LLM-GROUNDED MULTI-AGENT SIMULATION

Our sandbox enables counterfactual simulation through a two-phase architecture: an LLM derives expert-validated personas, and a mechanism-guided PRA runtime executes agent interactions without per-tick LLM calls. This separation grounds agent behavior in real community semantics while keeping the runtime controllable, auditable, and fast enough for interactive analysis.

6.1 Phase I: Contextual Persona Extraction

To simulate what-if scenarios, we must instantiate agents with empirically grounded behaviors [15]. To prevent cross-contamination, we extracted personas from the two communities independently. Informed by domain experts and sociological literature, we defined a closed-set codebook across seven 3-class dimensions: *sentiment tendency*, *participation motivation*, *community knowledge*, *decision rationality*, *risk preference*, *influence seeking*, and *communication agreeableness* (validated by ten domain experts; see Fig. 7 in Appendix).

Table 1: Simulation validation results: Mode₀ vs. ground truth.

Community	SC			P	C	T	FH _t		Overall Fidelity
	ρ	DTW	JSD	ρ	ρ	ρ	ρ	JSD	
<i>Mfers</i>	0.877	0.109	0.389	0.910	0.758	0.421	1.000	0.000	0.746
<i>Mimic Shhans</i>	0.865	0.095	0.530	0.917	0.902	0.739	1.000	0.000	0.759

* All $p < 0.01$ (except T in *Mfers*: $p = 0.073$). FH_t matches ground truth by design in Mode₀.

To ensure coding reliability, we deployed a confidence-guided human-in-the-loop (HITL) [52] pipeline. After experts annotated a few-shot ground truth, a long-context LLM (Moonshot API) labeled the remaining users. We excluded silent and near-silent accounts (< 15 messages), whose sparse text provides insufficient evidence for reliable persona inference. The retained cohort comprises about half of all accounts (see Table 2 in Appendix) while contributing over 95% of all messages in both communities. The LLM outputs were triaged by confidence: high-confidence samples were auto-retained, while uncertain cases (confidence < 0.60) underwent expert re-annotation. This data-centric iterative refinement maintained high annotation consistency without requiring model retraining.

Finally, we applied *K*-Modes [33, 34] clustering to the labeled features. Based on multi-criteria evaluations (inertia, silhouette score, and ARI), we identified $K = 6$ for *Mfers* and $K = 7$ for *Mimic Shhans* as optimal balances of compactness and interpretability. Robustness checks confirmed that the dominant archetypes remain stable across random seeds. These structural archetypes were then enriched with dynamic traits extracted from chat logs, such as *top topics*, *active routine*, *reply rate*, and *language style*. Together, they form the foundational templates for agent instantiation. This synthesis of structural archetypes and nuanced linguistic styles helps reduce generic “AI tone,” yielding authentic, domain-specific language for the *Phase II* runtime.

6.2 Phase II: Simulation Engine and Cognitive Workflow

Building upon interactive social media simulation frameworks [45], we design a structured **Perception–Reasoning–Action** (PRA) pipeline to govern agent behavior over time (Fig. 1C). The simulation progresses in discrete time steps (ticks), with each step representing a daily window to capture fine-grained social dynamics.

6.2.1 Simulation Engine Pipeline

The composition of the simulated population is determined by two constraints. To preserve ecological validity, agents are allocated in proportion to the empirical distribution of *K*-Modes archetype clusters. This choice follows a core principle in agent-based modeling: emergent phenomena, such as influence propagation and consensus formation, are shaped by the relative prevalence of agent types; deviating from this distribution (e.g., via uniform allocation) would systematically bias collective outcomes [7, 21, 46]. At the same time, to ensure statistically stable estimation of per-archetype behavioral patterns, we enforce a minimum of $N_p \geq 5$ agents per archetype [1, 17]. Together, these constraints establish a principled lower bound on the total population.

Within this temporal structure, interactions unfold asynchronously rather than through global synchronization. To reflect this property, we introduce a *coordinator* that selects the next active agent from the evolving conversational context. The coordinator does not generate content; it only regulates turn-taking, thereby preserving realistic communication dynamics without interfering with agent behavior.

Each agent operates with a structured memory comprising five layers: (1) macro environment, (2) user-injected governance rules, (3) historical summaries, (4) self-generated past statements, and (5) the recent chat window. These layers jointly support consistent behavior under long-context interactions. Conditioned on this memory, agents execute the PRA pipeline in a tightly coupled manner: during *Perception*, incoming information is filtered according to archetype-specific topical relevance; during *Reasoning*, agents weigh their seven-dimensional traits against persona-specific behavioral thresholds; and during *Action*, they select financial decisions (*buy*, *sell*, *hold*) and social interactions (*initiate*, *agree*, *refute*, *silence*), with posts drawn from context-conditioned,

LLM-distilled message templates. Relational social actions (e.g., *agree*, *refute*) are subsequently mapped to directed interactions, yielding a daily-resolved network that directly drives the frontend visualization and aligns with downstream financial metrics.

These simulated actions are then fed back into our mathematical models (Sec. 5.2), continuously recalculating the dynamic shifts in SC_t and FH_t to complete the visual analytics loop. Specifically, aggregated financial actions (*buy/sell*) perturb the baseline on-chain floor price (F_t) and liquidity (L_t) metrics. Simultaneously, the newly generated semantic edges alter the structural implicit reciprocity (IR_t), while the simulated agent statements shift the collective sentiment variance (σ_t). This closed-loop mechanism enables users to visually trace how micro-level interventions cascade through individual cognition to drive macro-level shifts in community vitality.

6.2.2 Simulation Evaluation

We iteratively refined the simulation across multiple versions to ensure ecological validity, correcting systemic issues including zero sell-rate bias, absence of bear-market baselines, and missing counterfactual reference modes. The final version introduces GT-anchored environmental calibration, a standard ABM practice [23, 74] wherein empirically observed activity levels, sentiment, and topic concentration serve as exogenous environmental inputs. To validate fidelity, we compare Mode₀ (no governance) against empirical ground truth using three complementary metrics: Spearman’s ρ for trend fidelity [74], normalized Dynamic Time Warping (DTW) for shape similarity [5], and Jensen–Shannon Divergence (JSD) for distributional fidelity [19]. Results confirm strong trend fidelity for core metrics across both communities (see details in Tab. 1), with overall composite fidelity scores of 0.746 and 0.759 respectively. Trust (T_t) is the only metric not reaching significance in *Mfers* ($\rho = 0.421$, $p = 0.073$), likely because LLM-distilled message templates smooth the fine-grained sentiment dynamics that drive trust [26, 57]. This limits trust-specific interpretation for *Mfers*, although the aggregate SC_t trajectory remains well aligned with ground truth ($\rho = 0.877$), supported by stronger recovery of P_t and C_t . FH_t achieves perfect calibration in Mode₀ by design, as it uses unmodified real-world financial data, establishing a principled baseline against which governance interventions are evaluated.

7 VISUAL DESIGN

This section presents the visual design of *SocialFiVis*, which consists of four coordinated views designed to support the hierarchical workflow.

7.1 Control Panel and Persona View

The *Control Panel* serves as the entry point for scenario selection and data instantiation (Fig. 2A). Selecting a preprocessed case determines the temporal scope and the community, triggering the persona construction whose results are visualized in the *Persona View* (Fig. 2B).

The *Persona View* is a matrix-based visualization that characterizes the extracted persona clusters and their multi-dimensional trait compositions (**T3**). We employ a structured grid layout. Each column represents a distinct persona, assigned a consistent categorical color to preserve identity across the system. The rows correspond to the seven personality dimensions (as defined in Sec. 6.1). Within each intersecting cell, a group of horizontal bars encodes the probability distribution of the three categorical states for that specific dimension. Additionally, the solid-colored block at the top of each column indicates the persona’s relative population size, which directly determines the agent-proportion initialization in the downstream simulation. The matrix extends horizontally with the number of personas; once it exceeds the palette’s distinguishable hues, personas are grouped into higher-level archetypes.

Interaction enables progressive disclosure of agent details. Hovering over a horizontal bar triggers an overlay showing the absolute member count and its relative percentage within the persona (Fig. 2B1). Selecting a persona column opens a *Persona Card*, which provides a textual summary of the overarching archetype. The card further decomposes the abstract configuration into a detailed *Personality Profile*, listing the dominant state and its exact percentage per dimension, alongside grounded *Behavioral Metrics*, such as primary discussion topics, peak

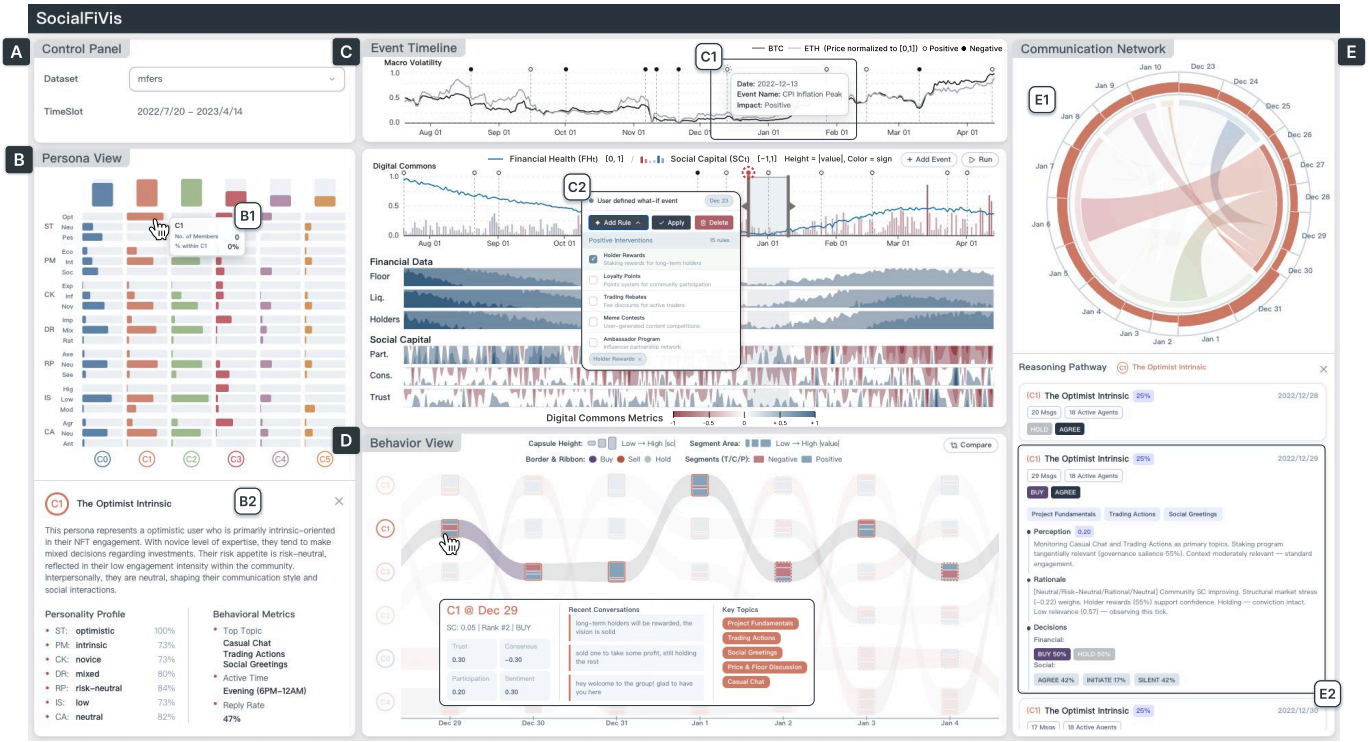


Figure 2: The *SocialFivis* interface. (A) *Control Panel* for case selection; (B) *Persona View* of persona-trait compositions, with (B1) population details on hover and (B2) a *Persona Card* of profile and behavioral metrics; (C) *Event Timeline* of (C1) macro crypto trends and milestones over community metrics, with (C2) user-injected interventions; (D) *Behavior View* tracking persona trajectories; and (E) *Network and Reasoning Views* pairing (E1) the communication topology with (E2) per-agent reasoning.

active hours, and reply rates (Fig. 2B2). This design bridges multi-dimensional trait distributions with interpretable behavioral descriptors.

7.2 Event Timeline View

The *Event Timeline View* supports the temporal exploration of the digital commons within a macroeconomic context (T1) and serves as the primary sandbox interface for comparative policy evaluation (T2). It helps stakeholders pinpoint critical historical milestones and transition fluidly from observational analysis to counterfactual backtesting.

The visual design adopts a multi-tier layout that structurally decouples macroeconomic context from localized community dynamics (Fig. 2C). The top-tier tracks global cryptocurrency trends using line charts, with historical milestones embedded as circular glyphs: hollow circles for positive events and solid black dots for negative ones (Fig. 2C1). The middle tier depicts the community’s overarching trajectory via a primary line chart paired with diverging bars. The bottom tier decomposes the digital commons into its six foundational metrics: *financial health* (F_t , L_t , and H_t), and *social capital* (P_t , C_t , and T_t). To maximize data density and facilitate cross-metric comparison within a constrained vertical space, these metrics are visualized using horizon graphs. A diverging red-to-blue palette encodes each metric on a shared $[-1, 1]$ scale, enabling rapid identification of synchronized shocks or systemic decoupling across the dual-track commons.

Interactions within this view drive the analytical workflow through two complementary modes. In the exploratory phase, brushing a time window updates the downstream *Behavior View* and *Communication Network*, allowing users to inspect historical agent interactions under empirical ground truth. To initiate a counterfactual backtest, users insert a customized intervention at a chosen timestamp, visually marked as a distinct red dot (Fig. 2C2). The simulation then produces alternative trajectories: the community-level timeline and horizon graphs update to overlay these counterfactual macro-metrics against historical baselines, while the *Behavior View* and *Communication Network* simultaneously transition to display the newly simulated micro-level data.

7.3 Behavior View

The *Behavior View* supports meso-level behavioral analysis by tracking the evolving trajectories of persona groups and their heterogeneous responses to policy interventions (T4, T5). While existing approaches capture temporal shifts via asset-centric transfer flows [73] or aggregated Sankey diagrams [61], they inherently obscure the identity-preserving trajectories and multi-dimensional evolutions of specific cohorts by aggregating transitional volumes. To address this, we introduce a *capsule-and-ribbon* design (Fig. 2D). By adapting rank-flow paradigms [68], this layout preserves cohort identity over time while revealing how persona trajectories bifurcate, enabling comparison of the three social-capital metrics and counterfactual branching.

A customized ribbon chart plots persona groups along a chronological x-axis, vertically ranking them by overall social capital (SC). At each tick, a spatially efficient multi-attribute capsule glyph encodes the persona’s state. Its total height represents absolute SC_t , while its interior is vertically partitioned into *participation* (P_t), *consensus* (C_t), and *trust* (T_t) segments. This partitioning deliberately mirrors the visual hierarchy of the macro-level horizon graphs, ensuring cognitive consistency across views. Segment fill colors indicate metric polarity. To preserve identity and encode overall SC_t polarity, the capsule’s border applies the persona’s palette using a solid stroke for positive SC_t and a dashed stroke for negative SC_t .

Ribbons connect these capsules over time, with cross-overs illustrating rank transitions. In the empirical ground-truth mode, the ribbon’s fill defaults to the persona’s palette to emphasize structural continuity (Fig. 3B1). Upon toggling to the counterfactual simulation mode, the ribbon dynamically shifts to encode the agent’s simulated financial action at the originating node (Fig. 3B2). This design distinguishes data provenance while highlighting causal behavioral shifts.

Interaction techniques facilitate in-depth exploration. Selecting a persona icon on the left y-axis highlights its continuous temporal flow by attenuating others for focused tracking (Fig. 3B3). A dedicated toggle seamlessly switches between ground-truth and counterfactual modes, isolating the precise impact of injected governance rules.



Figure 3: Behavior View workflow. (A1) Time window selection loads ground truth community dynamics (B1), while (A2) injecting a negative event (e.g., “Key Member Exit”) generates counterfactual socio-financial behaviors (B2). Persona highlighting (B3) enables micro-analysis: reacting to the shock, pessimistic C0 agents display heightened, negative social engagement coupled with reactive “SELL” strategies (orange ribbons).

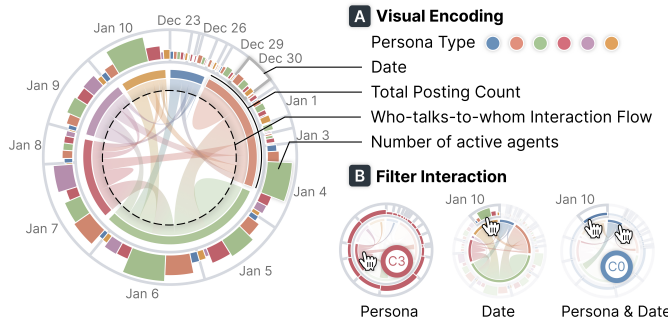


Figure 4: Visual design of the *Communication Network*. (A) Visual encodings show persona-driven social activation via posting volumes, interaction flows, and active-agent counts. (B) Cross-view persona/date filters support focused exploration of reasoning pathways.

Justification. To encode the three-dimensional metrics (P_t , C_t , T_t), we considered alternatives like heatmaps and treemaps. However, heatmaps compromise quantitative precision by over-relying on color channels, while treemaps disrupt cross-temporal tracking by dynamically altering spatial layouts. Our vertically partitioned capsule elegantly resolves these issues by maintaining a consistent spatial mapping and prioritizing geometric length over color intensity for accurate longitudinal comparison.

7.4 Communication Network

The *Communication Network View* unpacks the micro-level communication dynamics and semantic rationales underlying meso-level behavioral shifts (T5, T6). This module juxtaposes aggregate interaction topologies with persona-grounded reasoning pathways, enabling users to trace how policy interventions propagate through modeled agent behaviors to system-level metric shifts (Fig. 2E).

Inspired by [11], we designed a simplified multi-ring chord diagram to visualize the interaction topology (Fig. 4A). Functioning primarily as a hierarchical information filter, its inner ring segments encode the total message volume per persona within the selected time window, with connecting ribbons mapping the communication flows; the outer ring, segmented by day as a circular bar chart, encodes each persona’s daily active-agent count. The radial layout co-locates per-day activity with the persona interaction topology in one compact view, linking the outer and inner rings through shared persona colors rather than radial alignment, which favors space efficiency and cross-cohort comparison.

The diagram controls the adjacent *Reasoning Pathway* panel, which displays the agents’ behavioral rationale and action distributions.

Interactions facilitate a fluid drill-down from macro-topology to individual reasoning traces (Fig. 4B). Intra-view, selecting an outer date arc or an inner persona segment filters the details panel to isolate rationales by day or cohort; combining both pinpoints a specific persona’s cognition on an exact date. Cross-view coordination further drives the analytical workflow. In counterfactual mode, injecting a timeline strategy synchronizes updates across both the network and the reasoning pathways, allowing users to seamlessly trace how policy interventions catalyze micro-level cognitive shifts and dialogues (Fig. 2E2).

8 EVALUATION

We evaluated the effectiveness and usability of *SocialFiVis* through two case studies, followed by a user study and semi-structured interviews.

8.1 Case Studies

E_1 and E_5 conducted exploratory case studies to evaluate *SocialFiVis* in real-world analytical workflows. As the simulated populations are instantiated from messaging members (Sec. 6.1), the case studies characterize how social and financial dynamics co-evolve within this messaging population under different governance interventions.

8.1.1 Case I: Policy Efficacy and Diminishing Returns

Focusing on policy strategy comparison during a stagnant market phase in the *Mimic Shhans* community, E_1 discovered that overlapping incentive strategies yield diminishing marginal returns and that positive macro-policies can be covertly exploited by deceptive individual behaviors (DG2, DG3).

Actionable Insight 1: Stagger incentive releases rather than stacking them. Examining the Aug. 10–28, 2022 phase, E_1 compared three counterfactual conditions: “Holder Rewards” alone, “Loyalty Points” alone, and the two combined. In the *Event Timeline* and *Horizon Graph* (Fig. 5A), each incentive on its own lifted the social-capital trajectory, yet the combined run tracked the stronger single policy rather than exceeding it and left consensus slightly lower rather than stronger. E_1 noted, “Deploying two favorable policies simultaneously violates the economic principle of marginal utility; it dilutes the intended signal and scatters community consensus.” The simulation accordingly guided E_1 toward staggering incentive releases rather than stacking them.

Actionable Insight 2: Monitor sentiment–action divergence during incentive campaigns. Transitioning to the meso-level *Behavior Ribbon* (Fig. 5B), E_1 compared how different personas reacted to the injected liquidity of the “Holder Rewards.” While personas C0 and

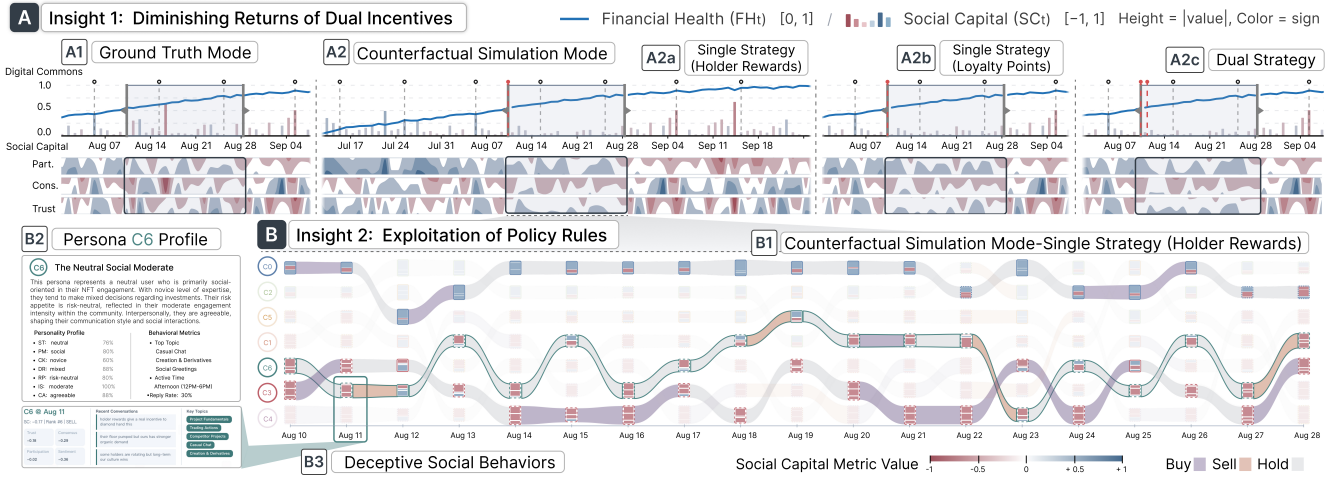


Figure 5: Illustration of Case I. (A) Counterfactual simulations in the *Event Timeline* reveal the diminishing marginal returns of the dual incentive strategy (A2c) compared to single positive interventions (A2a, A2b). (B) Under a single incentive policy (B1), the *Behavior View* exposes policy exploitation, illustrating how speculative agents (B2) leverage deceptive social behaviors (B3) for financial gain.

C3 stabilized the community as “core promoters,” E_1 observed anomalous fluctuations in C6, characterized by discordant color encodings between sentiment and financial action. By interactively filtering C6 in the *Communication Network* and inspecting its micro-level *Reasoning Pathway*, E_1 uncovered a highly deceptive pattern: C6 actively preached “diamond hands” (holding) in the chat to inflate sentiment while simultaneously executing aggressive sell orders. E_1 concluded that operators should monitor sentiment-action divergence to flag such manipulative archetypes before they exploit macro-level policies.

8.1.2 Case II: Aggregate Resilience and Latent Fragility

E_5 examined how the *Mfers* messaging cohort weathered volatility from Dec. 23, 2022 – Jan. 10, 2023, and how macro-level momentum shaped social stability under adverse conditions (DG1, DG3).

Actionable Insight 1: Watch trust-participation decoupling despite stable engagement. Against a baseline of natural macro-level recovery (Founder’s Return), E_5 evaluated resilience by comparing two counterfactual scenarios: a negative “Key Member Exit” shock and a positive “Holder Rewards” incentive. In the negative scenario, E_5 found that the *participation* and *consensus* metrics remained stable in the *Event Timeline View*, maintaining levels comparable to the positive intervention. In contrast, the *trust* index exhibited a marked and isolated decline. This structural decoupling shows E_5 that strong macro-positive signals can sustain engagement and consensus even when interpersonal trust erodes under a localized negative shock. E_5 accordingly flagged the isolated trust decline as a signal to monitor, since participation and consensus alone would mask it.

Actionable Insight 2: Prioritize vulnerable archetypes after a negative shock. By cross-referencing the *Behavior View* with the *Reasoning Pathway*, E_5 found the trust erosion concentrated in specific archetypes rather than uniform. “Optimist” believers (C3) held steady through the shock, whereas “pessimist” members (C0) reacted sharply, selling off and refuting peers as their *trust* plunged and recovered only after the macro-milestone signal propagated (Fig. 3B2). “Band traders” (C4) likewise turned net sellers, thinning support rather than cushioning the downturn. E_5 noted that this archetype-level attribution identifies C0 and C4 as the cohorts to monitor first when a negative shock lands.

8.2 User Study and Expert Interviews

To evaluate the system across a broader audience, we conducted a user study supplemented by semi-structured interviews.

8.2.1 Participants and Procedure

In addition to E_1 – E_4 , we recruited two independent domain experts (E_5 , E_6) and seven users (U_1 – U_7). By role and experience, E_5 , U_1 , U_4 , U_6 and U_7 were operations-experienced participants, whereas E_6 ,

U_2 , U_3 , and U_5 were secondary users (analysts, traders, and ordinary members) who interpret or audit governance analyses. This anchored the evaluation in our primary operator audience while probing broader applicability. Following established ethical guidelines, our study posed minimal risk to participants: participation was voluntary, all participants provided informed consent, and all collected data were anonymized.

The evaluation employed a two-track procedure to accommodate participants’ varying levels of prior engagement. For the 11 participants who did not conduct the exploratory case studies (i.e., E_2 – E_4 , E_6 , and U_1 – U_7), the session began with a 15-minute tutorial. They then completed 25 minutes of predefined analytical tasks using a think-aloud protocol. These tasks were directly mapped to our design goals: (1) correlating macro financial metrics with social capital (DG1); (2) using the sandbox to backtest historical interventions (DG2); and (3) tracing the heterogeneous decision pathways of specific personas (DG3).

E_1 and E_5 bypassed these predefined tasks, as their prior in-depth case studies (Sec. 8.1) far exceeded the complexity of the baseline tasks. Finally, all 13 participants completed a 5-point Likert scale questionnaire spanning five evaluation dimensions [32, 41], followed by a 10-minute post-study interview.

8.2.2 Results and Feedback

Quantitative results (see Fig. 6) highlight the system’s successes in *visualization informativeness* ($M = 4.69, SD = 0.48$) and *adoption willingness* ($M = 4.69, SD = 0.48$; $M = 4.77, SD = 0.44$ for *recommendation*). Qualitative feedback supports these results. Participants consistently highlighted the system’s ability to link macro market events with micro behavioral trajectories as its most valuable contribution. They found the *Event Timeline* and *Behavior View* particularly effective for this purpose. The two-phase agent simulation was praised for capturing behavioral heterogeneity and decision-making nuance that purely statistical models cannot represent. We synthesized the feedback below.

System Usefulness and Effectiveness. Participants highly rated the system’s ability to model the interplay between governance rules and macro outcomes ($M = 4.31, SD = 0.85$). Notably, experts found the multi-level workflow closely aligned with their analytical reasoning processes ($M = 4.46, SD = 0.52$). E_3 and E_4 noted that it “provides a quantitative baseline [for] backtesting,” while E_2 appreciated the persona classification, emphasizing that tracking who does what is “genuinely helpful and intuitive for reporting to management.”

Visual Design and Interaction. Participants praised the visualizations for their rich information density ($M = 4.69, SD = 0.48$). For instance, U_1 noted that the *Behavior View* “reveals social relationships... which is something you cannot see intuitively,” while E_5 found the *Persona View*’s fine-grained classification highly aligned with real-world community experience. However, intuitiveness scored lower

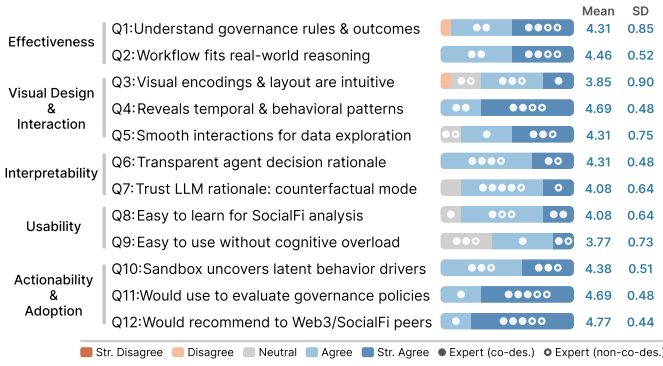


Figure 6: User study results (N=13). Q1–Q12 are rated on a 5-point Likert scale, assessing *SocialFiVis* across five dimensions.

($M = 3.85, SD = 0.90$), reflecting a trade-off with informativeness: the dense, multi-view layouts mobilize cognitive resources but carry a learning curve, particularly for community builders and other less data-oriented users who focus on engagement rather than analysis. This adaptation is expected, as their prior practice often leaned on trial-and-error, intuition, and one-on-one interviews rather than systematic visual analytics. U_4 captured the balance: the *Event Timeline* “gives you the big picture,” though its density demands familiarization.

Interpretability and Trust. The system successfully established transparency in agent decision-making ($M = 4.31, SD = 0.48$) and trust in the agent simulation ($M = 4.08, SD = 0.64$). Observing a simulated agent’s actions, E_5 remarked, “The AI is remarkably true-to-life... its stated sentiment, reasoning, and final action all line up.” To further enhance this utility, E_1 suggested incorporating a proactive policy recommendation feature. Since formulating governance policies is a high-frequency daily task, such automated guidance would significantly streamline operators’ routine analysis workflows.

Usability. Learnability scored positively ($M = 4.08, SD = 0.64$), though cognitive load ratings ($M = 3.77, SD = 0.73$) reflected the aforementioned learning curve. Crucially, usability perceptions varied by professional role. For data-savvy operators and analysts, the density is a strategic asset; E_3 found it “genuinely useful for my day-to-day work,” and E_2 noted that “for someone in a marketing role, this is very usable,” even expressing a willingness to pay for a commercial version. Conversely, less data-oriented users would benefit from a simplified dashboard to ease their initial cognitive load.

Actionability and Adoption. The sandbox’s ability to uncover latent catalysts scored highly ($M = 4.38, SD = 0.51$). E_6 expressed strong intent to recommend the system to executive leadership, emphasizing that its rich insights make it highly “suited for decision-makers”. However, participants also identified data dependency as a critical prerequisite. E_5 and U_7 cautioned that during data collection, operators must carefully filter out chatbots and airdrop hunters (speculators) to prevent them from polluting the user personas and skewing the simulation outcomes.

9 DISCUSSION

We reflect on the broader implications of *SocialFiVis*, detailing its methodological generalizability and limitations guiding future research.

9.1 Significance and Generalizability

Beyond addressing specific governance challenges in tokenized communities, the core contribution of *SocialFiVis* lies in its highly transferable methodology for analyzing the co-evolution of social and financial dynamics in broader decentralized environments.

Significance of the Work. *SocialFiVis* operationalizes the IAD framework into a visual analytics pipeline, bridging the micro-macro gap by coupling *social capital* with *financial health*. Crucially, the counterfactual sandbox provides a risk-free environment for policy testing. This directly addresses a longstanding practitioner pain point: the “attribution problem” of tracing mechanistic pathways between specific

community interventions and subsequent financial outcomes. Furthermore, by grounding agents in LLM-derived, expert-validated personas, the system represents behavioral heterogeneity within the retained messaging cohort and supports inspection of modeled decision rationales.

Generalizability of the Methodology. We designed the system’s core algorithms to generalize far beyond financialized ecosystems. Because the IAD framework inherently governs common-pool resources [55], our approach is naturally applicable to various decentralized organizations, creator economies, and open-source communities. Specifically, the persona extraction and PRA pipeline can serve as a template to simulate heterogeneous populations in corporate organizational behavior or consumer research. Additionally, our soft-penalty geometric blending model can be adapted to quantify other forms of intangible capital [35], such as user engagement in online learning platforms or the health metrics of corporate cultures.

9.2 Limitations and Future Work

While our evaluation confirms the system’s utility, it also highlights limitations that pave the way for future enhancements.

Simulation Boundaries and Predictive Validity. Instantiated from messaging members, the simulation excludes silent and near-silent accounts below our persona-inference threshold (Sec. 6.1); it therefore reflects expressed dynamics, not silent disengagement [54, 64]. This exclusion likewise bounds the aggregate resilience reported in our case studies (Sec. 8.1). Within this cohort, the PRA pipeline produces agents’ posts and trades from a reasoning step but maintains no *persistent belief state* separable from expression. Such a state would model agents who hold views without posting and make discrepancies between words and actions a systematic, auditable signal for users rather than an incidental finding. Validity is likewise bounded: our validation targets aggregate trends, not persona-level fidelity to real users. Specifically, we calibrate the engine against ground truth (Table 1) since these counterfactuals never occurred, and present them as exploratory reasoning aids instead of forecasts [23, 74]. Even so, communicating aggregate uncertainty to calibrate user trust remains an open challenge [69]. We also acknowledge that the LLM may have seen these communities’ public messages in training, a general concern for pretrained models; we temper but cannot fully remove this exposure by confining the LLM to a study-specific coding scheme, while the mechanism-guided runtime generates the simulated behaviors and trajectories.

System Complexity and Learning Curve. The system’s high information density entails a learning curve. Although its coordinated views are organized around progressive disclosure [18, 62], the lower intuitiveness score ($M = 3.85$) shows that the entry state remains demanding for some practitioners. A deployment-ready design therefore requires a compact summary layer (e.g., directional bias indicators, aggregate sentiment–action divergence, and flagged milestones) before users opt into the denser persona, behavior, and reasoning views. The same practical lens applies to evaluation: although we recruited independent operators and secondary users, the co-designers’ involvement may bias feedback favorably, and a larger participant pool studied in operators’ day-to-day workflows would further corroborate the findings.

10 CONCLUSION

In this paper, we presented *SocialFiVis*, a visual analytics sandbox that operationalizes the IAD framework to quantify the dual-track digital commons and model the non-linear interplay between governance rules, individual behaviors, and economic outcomes. Its two-phase simulation pairs LLM-derived, expert-validated personas with a mechanism-guided PRA runtime, letting users trace how interventions propagate through agent rationales into shifts within the messaging cohort. Across two case studies and expert interviews, *SocialFiVis* surfaced actionable insights, turning theoretical governance into concrete behavioral attribution. Ultimately, it establishes a novel foundation for simulation-based decision support at the intersection of visual analytics, agent-based modeling, and digital commons governance.

SUPPLEMENTAL MATERIALS

The supplemental materials provided with this submission contain additional methodological and study details, including task traceability, metric formulation, and sensitivity analyses, persona and simulation validation, and the complete user-study instruments.

REFERENCES

- [1] A. Agresti. *Categorical data analysis*. John Wiley & Sons, 2013. doi: 10.1002/0471249688 5
- [2] S. Al-Oufi, H.-N. Kim, and A. El Saddik. A group trust metric for identifying people of trust in online social networks. *Expert Syst. Appl.*, 39(18):13173–13181, 2012. doi: 10.1016/j.eswa.2012.05.084 4
- [3] A. Atrey, K. Clary, and D. Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning, 2019. arXiv preprint. doi: 10.48550/arXiv.1912.05743 2
- [4] S. Barde and S. van der Hoog. An empirical validation protocol for large-scale agent-based models. Technical Report 1712, School of Economics Discussion Papers, 2017. 2
- [5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proc. KDD Workshop*, pp. 359–370. AAAI Press, 1994. doi: 10.5555/3000850.3000887 5
- [6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. doi: 10.5555/944919.944937 4
- [7] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.*, 99(suppl_3):7280–7287, 2002. doi: 10.1073/pnas.082080899 2, 5
- [8] D. Borland, A. Z. Wang, and D. Gotz. Using counterfactuals to improve causal inferences from visualizations. *IEEE Comput. Graph. Appl.*, 44(1):95–104, 2024. doi: 10.1109/MCG.2023.3338788 2
- [9] K. Brahmstaedt. Community and consumer dynamics in NFTs: Understanding digital asset value through social engagement. *J. Consum. Behav.*, 24(4):1630–1655, 2025. doi: 10.1002/cb.2482 1, 2
- [10] Y. Cao, Q. Shi, L. Shen, K. Chen, Y. Wang, W. Zeng et al. NFTracer: Tracing NFT impact dynamics in transaction-flow substitutive systems with visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 31(8):4369–4386, 2025. doi: 10.1109/TVCG.2024.3402834 2
- [11] Y. Cao, M. Xia, K. Shigyo, F. Cheng, Q. Yu, X. Yang et al. NFTeller: Dual-centric visual analytics for assessing market performance of NFT collectibles. In *Proc. VINCI*, art. no. 20, 8 pp. ACM, New York, 2023. doi: 10.1145/3615522.3615578 1, 7
- [12] H. Chen, C. Zhou, A. El Saddik, and W. Cai. Decentralized Web3 non-fungible token community for societal prosperity? a social capital perspective. *Proc. ACM Hum.-Comput. Interact.*, 9(2), art. no. CSCW058, 36 pp., 2025. doi: 10.1145/3710956 1, 2
- [13] L. Chen, Y. Zhang, J. Feng, H. Chai, H. Zhang, B. Fan et al. AI agent behavioral science, 2025. arXiv preprint. doi: 10.48550/arXiv.2506.06366 2
- [14] C.-M. Chiu, M.-H. Hsu, and E. T. Wang. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decis. Support Syst.*, 42(3):1872–1888, 2006. doi: 10.1016/j.dss.2006.04.001 1, 2
- [15] Y. Choi, E. J. Kang, S. Choi, M. K. Lee, and J. Kim. Proxona: supporting creators’ sensemaking and ideation with LLM-powered audience personas. In *Proc. CHI*, art. no. 149, 32 pp. ACM, New York, 2025. doi: 10.1145/3706598.3714034 1, 2, 4
- [16] M.-L. Chu, L. Terhorst, K. Reed, T. Ni, W. Chen, and R. Lin. LLM-based multi-agent system for simulating and analyzing marketing and consumer behavior. In *Proc. IEEE ICEBE*, pp. 72–79. IEEE, Piscataway, 2025. doi: 10.1109/ICEBE68123.2025.00018 1, 2
- [17] W. G. Cochran. The χ^2 test of goodness of fit. *Ann. Math. Stat.*, 23(3):315–345, 1952. doi: 10.1214/aoms/1177729380 5
- [18] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Trans. Vis. Comput. Graph.*, 16(3):439–454, 2010. doi: 10.1109/TVCG.2009.84 9
- [19] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860, 2003. doi: 10.1109/TIT.2003.813506 5
- [20] J. M. Epstein. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60, 1999. doi: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F 2
- [21] J. M. Epstein and R. Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996. doi: 10.7551/mitpress/3374.001.0001 5
- [22] M. Esposito, T. Tse, and D. Goh. Decentralizing governance: exploring the dynamics and challenges of digital commons and DAOs. *Front. Blockchain*, 8, art. no. 1538227, 13 pp., 2025. doi: 10.3389/fbloc.2025.1538227 1
- [23] G. Fagiolo, P. Windrum, and A. Moneta. A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Comput. Econ.*, 30(3):195–226, 2007. doi: 10.1007/s10614-007-9104-4 2, 5, 9
- [24] A. Filippas, J. J. Horton, and B. S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? In *Proc. ACM EC*, pp. 614–615. ACM, New York, 2024. doi: 10.1145/3670865.3673513 2
- [25] J. Gajcin and I. Dusparic. Redefining counterfactual explanations for reinforcement learning: Overview, challenges and opportunities. *ACM Comput. Surv.*, 56(9), art. no. 219, 33 pp., 2024. doi: 10.1145/3648472 2
- [26] C. Gao, X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanit. Soc. Sci. Commun.*, 11(1), art. no. 1259, 24 pp., 2024. doi: 10.1057/s41599-024-03611-3 2, 5
- [27] L. W. Ge, M. Easterday, M. Kay, E. Dimara, P. Cheng, and S. L. Franconeri. V-FRAMER: Visualization framework for mitigating reasoning errors in public policy. In *Proc. CHI*, art. no. 390, 15 pp. ACM, New York, 2024. doi: 10.1145/3613904.3642750 2
- [28] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. Natl. Acad. Sci.*, 101(suppl_1):5228–5235, 2004. doi: 10.1073/pnas.0307752101 4
- [29] M. Guerini and A. Moneta. A method for agent-based models validation. *J. Econ. Dyn. Control*, 82:125–141, 2017. doi: 10.1016/j.jedc.2017.06.001 2
- [30] B. Guidi and A. Michienzi. SocialFi: towards the new shape of social media. *ACM SIGWEB Newsl.*, 2022(Summer):1–8, 2022. doi: 10.1145/3545196.3545201 1, 2
- [31] Ö. Gürkan. LLM-augmented agent-based modelling for social simulations: challenges and opportunities. In *Proc. FAIR Conf. on AI Systems for the Social Good*, pp. 155–168. IOS Press, 2024. doi: 10.3233/FAIA240190 2
- [32] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: Challenges and prospects, 2018. arXiv preprint. doi: 10.48550/arXiv.1812.04608 8
- [33] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283–304, 1998. doi: 10.1023/A:1009769707641 5
- [34] Z. Huang and M. K. Ng. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans. Fuzzy Syst.*, 7(4):446–452, 1999. doi: 10.1109/91.784206 5
- [35] L. Hunter, E. Webster, and A. Wyatt. Measuring intangible capital: a review of current practice. *Aust. Account. Rev.*, 15(36):4–21, 2005. doi: 10.1111/j.1835-2561.2005.tb00288.x 9
- [36] A. Imani Rad and S. Banaeian Far. SocialFi transforms social media: an overview of key technologies, challenges, and opportunities of the future generation of social media. *Soc. Netw. Anal. Min.*, 13(1), art. no. 42, 2023. doi: 10.1007/s13278-023-01050-7 1
- [37] S. W. Jeong, S. Ha, and K.-H. Lee. How to measure social capital in an online brand community? a comparison of three social capital scales. *J. Bus. Res.*, 131:652–663, 2021. doi: 10.1016/j.jbusres.2020.07.051 2
- [38] R. Jones, S. Sharkey, J. Smithson, T. Ford, T. Emmens, E. Hewis et al. Using metrics to describe the participative stances of members within discussion forums. *J. Med. Internet Res.*, 13(1), art. no. e3, 18 pp., 2011. doi: 10.2196/jmir.1591 4
- [39] A. Kapoor, D. Guhathakurta, M. Mathur, R. Yadav, M. Gupta, and P. Kumaraguru. TweetBoost: Influence of social media on NFT valuation. In *Proc. WWW Companion*, pp. 621–629. ACM, New York, 2022. doi: 10.1145/3487553.3524642 1
- [40] S. Kaul, D. Borland, N. Cao, and D. Gotz. Improving visualization interpretation using counterfactuals. *IEEE Trans. Vis. Comput. Graph.*, 28(1):998–1008, 2022. doi: 10.1109/TVCG.2021.3114779 2
- [41] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Trans. Vis. Comput. Graph.*, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279 8
- [42] M. Larooij and P. Törnberg. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artif. Intell. Rev.*, 59(1), art. no. 15, 31 pp., 2026. doi: 10.1007/s10462

- [43] R. Li, S. Ye, Y. Lin, B. Zhou, Z. Kang, T.-Q. Peng et al. Causality-based visual analytics of sentiment contagion in social media topics. *IEEE Trans. Vis. Comput. Graph.*, 32(1):35–45, 2026. doi: [10.1109/TVCG.2025.3633839](https://doi.org/10.1109/TVCG.2025.3633839) 2
- [44] S. Li and Y. Chen. Governing decentralized autonomous organizations as digital commons. *J. Bus. Ventur. Insights*, 21, art. no. e00450, 2024. doi: [10.1016/j.jbvi.2024.e00450](https://doi.org/10.1016/j.jbvi.2024.e00450) 1
- [45] Z. Lin, Y. Shan, L. Gao, X. Jia, and S. Chen. SimSpark: Interactive simulation of social media behaviors. *Proc. ACM Hum.-Comput. Interact.*, 9(2), art. no. CSCW168, 32 pp., 2025. doi: [10.1145/3711066](https://doi.org/10.1145/3711066) 1, 5
- [46] C. Macal and M. North. Introductory tutorial: Agent-based modeling and simulation. In *Proc. WSC*, pp. 6–20. IEEE, Piscataway, 2014. doi: [10.1109/WSC.2014.7019874](https://doi.org/10.1109/WSC.2014.7019874) 5
- [47] C. M. Macal and M. J. North. Tutorial on agent-based modelling and simulation. *J. Simul.*, 4(3):151–162, 2010. doi: [10.1057/jos.2010.3](https://doi.org/10.1057/jos.2010.3) 2
- [48] M. D. McGinnis. Connecting commons and the IAD framework. In *Routledge Handbook of the Study of the Commons*, pp. 69–86. Routledge, 2019. doi: [10.4324/9781315162782-5](https://doi.org/10.4324/9781315162782-5) 2
- [49] MFERS. mfers. <https://mfers.art/>. Accessed: 2026-03-15. 4
- [50] Mimic Shhans. Mimic shhans. <https://mimicshhans.com/>. Accessed: 2026-03-15. 4
- [51] M. F. Morell. Governance of online creation communities for the building of digital commons: Viewed through the framework of institutional analysis and development. In *Governing Knowledge Commons*, pp. 281–306. Oxford Univ. Press, 2014. doi: [10.1093/acprof:oso/9780199972036.003.0009](https://doi.org/10.1093/acprof:oso/9780199972036.003.0009) 2
- [52] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4):3005–3054, 2023. doi: [10.1007/s10462-022-10246-w](https://doi.org/10.1007/s10462-022-10246-w) 5
- [53] Y. Ni, P. Chiang, M.-Y. Day, and Y. Chen. Using big data analytics and heatmap matrix visualization to enhance cryptocurrency trading decisions. *Appl. Sci.*, 14(1), art. no. 154, 16 pp., 2024. doi: [10.3390/app14010154](https://doi.org/10.3390/app14010154) 2
- [54] J. Nielsen. The 90-9-1 rule for participation inequality in social media and online communities. Nielsen Norman Group, 2006. <https://www.nngroup.com/articles/participation-inequality/>. 9
- [55] E. Ostrom. The institutional analysis and development framework and the commons. *Cornell Law Rev.*, 95(4):807–856, 2010. 1, 2, 3, 9
- [56] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Political communication on social media: A tale of hyperactive users and bias in recommender systems. *Online Soc. Netw. Media*, 15, art. no. 100058, 2020. doi: [10.1016/j.osnem.2019.100058](https://doi.org/10.1016/j.osnem.2019.100058) 4
- [57] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proc. UIST*, art. no. 2, 22 pp. ACM, New York, 2023. doi: [10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763) 2, 5
- [58] L. Ricci, B. Guidi, A. Michienzi, A. Tagarelli, and S. Gaito. AWESOME: Analysis framework for Web3 social media. In *Proc. ACM WebSci*, pp. 355–364. ACM, New York, 2024. doi: [10.1145/3677117.3685010](https://doi.org/10.1145/3677117.3685010) 2
- [59] N. Sánchez-Arrieta, R. A. González, A. Cañabate, and F. Sabate. Social capital on social networking sites: A social network perspective. *Sustainability*, 13(9), art. no. 5147, 2021. doi: [10.3390/su13095147](https://doi.org/10.3390/su13095147) 1
- [60] E. Schlager and M. Cox. The IAD framework and the SES framework: an introduction and assessment of the ostrom workshop frameworks. In *Theories of the Policy Process*, pp. 215–254. Routledge, 2018. doi: [10.4324/9780429494284-7](https://doi.org/10.4324/9780429494284-7) 2
- [61] R. Sheng, Y. Wang, X. Wang, S. Dai, Q. Guo, T.-Q. Peng et al. EMINDS: Understanding user behavior progression for mental health exploration on social media. *IEEE Trans. Vis. Comput. Graph.*, 32(2):2284–2299, 2026. doi: [10.1109/TVCG.2025.3630646](https://doi.org/10.1109/TVCG.2025.3630646) 6
- [62] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. Visual Languages*, pp. 336–343. IEEE, Piscataway, 1996. doi: [10.1109/VL.1996.545307](https://doi.org/10.1109/VL.1996.545307) 9
- [63] J.-T. Sohns, C. Garth, and H. Leitte. Decision boundary visualization for counterfactual reasoning. *Comput. Graph. Forum*, 42(1):7–20, 2023. doi: [10.1111/cgf.14650](https://doi.org/10.1111/cgf.14650) 2
- [64] V. Soroka and S. Rafaeli. Invisible participants: how cultural capital relates to lurking behavior. In *Proc. WWW*, pp. 163–172. ACM, New York, 2006. doi: [10.1145/1135777.1135806](https://doi.org/10.1145/1135777.1135806) 9
- [65] P. Taillandier, J. D. Zucker, A. Grignard, B. Gaudou, N. Q. Huynh, and A. Drogoul. Integrating LLM in agent-based social simulation: Opportunities and challenges, 2025. arXiv preprint. doi: [10.48550/arXiv.2507.19364](https://doi.org/10.48550/arXiv.2507.19364) 2, 3
- [66] J. Tang, H. Gao, X. Pan, L. Wang, H. Tan, D. Gao et al. GenSim: A general social simulation platform with large language model based agents. In *Proc. NAACL (System Demonstrations)*, pp. 143–150. ACL, 2025. doi: [10.18653/v1/2025.naacl-demo.15](https://doi.org/10.18653/v1/2025.naacl-demo.15) 2
- [67] N. Tovanich, N. Heulot, J.-D. Fekete, and P. Isenberg. Visualization of blockchain data: A systematic review. *IEEE Trans. Vis. Comput. Graph.*, 27(7):3135–3152, 2021. doi: [10.1109/TVCG.2019.2963018](https://doi.org/10.1109/TVCG.2019.2963018) 2
- [68] N. Tovanich, N. Soulié, N. Heulot, and P. Isenberg. MiningVis: visual analytics of the Bitcoin mining economy. *IEEE Trans. Vis. Comput. Graph.*, 28(1):868–878, 2022. doi: [10.1109/TVCG.2021.3114821](https://doi.org/10.1109/TVCG.2021.3114821) 6
- [69] E. Wall, L. Matzen, M. El-Assady, P. Masters, H. Hosseinpour, A. Ender et al. Trust junk and evil knobs: Calibrating trust in AI visualization. In *Proc. IEEE PacificVis*, pp. 22–31. IEEE, Piscataway, 2024. doi: [10.1109/PacificVis60374.2024.00012](https://doi.org/10.1109/PacificVis60374.2024.00012) 9
- [70] A. Z. Wang, D. Borland, and D. Gotz. An empirical study of counterfactual visualization to support visual causal inference. *Inf. Vis.*, 23(2):147–163, 2024. doi: [10.1177/14738716241229437](https://doi.org/10.1177/14738716241229437) 2
- [71] A. Z. Wang, D. Borland, and D. Gotz. A framework to improve causal inferences from visualizations using counterfactual operators. *Inf. Vis.*, 24(1):24–41, 2025. doi: [10.1177/14738716241265120](https://doi.org/10.1177/14738716241265120) 1, 2
- [72] X. Wen, T. D. Nguyen, S. Ruan, Q. Shen, J. Sun, F. Zhu et al. Ponzilens+: Visualizing bytecode actions for smart ponzi scheme identification. *IEEE Trans. Vis. Comput. Graph.*, 31(9):6451–6465, 2025. doi: [10.1109/TVCG.2024.3516379](https://doi.org/10.1109/TVCG.2024.3516379) 2
- [73] X. Wen, Y. Wang, X. Yue, F. Zhu, and M. Zhu. NFTDisk: Visual detection of wash trading in NFT markets. In *Proc. CHI*, art. no. 215, 15 pp. ACM, New York, 2023. doi: [10.1145/3544548.3581466](https://doi.org/10.1145/3544548.3581466) 6
- [74] P. Windrum, G. Fagiolo, and A. Moneta. Empirical validation of agent-based models: Alternatives and prospects. *J. Artif. Soc. Soc. Simul.*, 10(2):1–8, 2007. 1, 2, 5, 9
- [75] X. Zhang, J. Lin, X. Mou, S. Yang, X. Liu, L. Sun et al. SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users, 2025. arXiv preprint. doi: [10.48550/arXiv.2504.10157](https://doi.org/10.48550/arXiv.2504.10157) 2
- [76] Z. Zhong, S. Wei, Y. Xu, Y. Zhao, F. Zhou, F. Luo et al. SilkViser: A visual explorer of blockchain-based cryptocurrency transaction data. In *Proc. IEEE VAST*, pp. 95–106. IEEE, Piscataway, 2020. doi: [10.1109/VAST50239.2020.00014](https://doi.org/10.1109/VAST50239.2020.00014) 2
- [77] F. Zhou, Y. Chen, C. Zhu, L. Jiang, X. Liao, Z. Zhong et al. Visual analysis of money laundering in cryptocurrency exchange. *IEEE Trans. Comput. Soc. Syst.*, 11(1):731–745, 2023. doi: [10.1109/TCSS.2022.3231687](https://doi.org/10.1109/TCSS.2022.3231687) 2
- [78] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? *Comput. Linguist.*, 50(1):237–291, 2024. doi: [10.1162/coli_a_00502](https://doi.org/10.1162/coli_a_00502) 2

A APPENDIX



Figure 7: Expert validation of the 7D persona codebook (N=10).

Table 2: Statistics of the data filtering and routing pipeline.

Community	Raw Dataset		Messaging Filtering		Routing (Active Nodes)		
	Text Msgs	Nodes	Dropped (< 15)	Retained (≥ 15)	High-Conf. (≥ 0.60)	Low-Conf. (< 0.60)	Manual (Subset)
Mfers	47,646	395	194 (49.1%)	201 (50.9%)	123 (61.2%)	78 (38.8%)	24 (11.9%)
Mimic Shhans	41,740	435	218 (50.1%)	217 (49.9%)	114 (52.5%)	103 (47.5%)	33 (15.2%)