Measuring and Mitigating Structural Effects of Credibility, Bias, and Ownership on the Media and its Audience

Peter Carragher

Carnegie Mellon University

Roadmap: media dimension-behavior mapping

Dimension	Measure	Detection	Intervention	User Perception
Source Credibility	Sourcing, linkage, factual claims	Webgraph, Knowledge conflict	Resilient search rankings (TIST)	Informational trust assessments
Political Bias	Sharing across ideological divides	Convergence, Framing detection	Crosscutting recommendations	Stance recognition
Corporate Structure	Consolidation (IC2S2)	Ownership in citation networks	Ownership transparency	Media literacy tools

Part 1: Dimensions of the media landscape (Top-down, content creation)

- Source **credibility**
  - Detection & Discovery (ICWSM)
  - Dissemination (TIST)
  - Dredge words (ICWSM)
  - LLM credibility (memorization / hallucination / injection + poisoning)
- Political **bias**
  - Editorial decisions (journal tbd):
    - Event coverage in GDELT
    - Citations in CommonCrawl
    - Social engagement in FB URLs
  - Credibility X Bias (TIST)
- Corporate **ownership**
  - *Creation*: acquisitions + revenue (IC2S2, journal tbd)
    - Credibility X Bias X Corporate
  - *Consumption*: ads + monetization (Google)?
- **Media type**
  - **Modulates** the expression of core dimensions
  - *Online* media (interactive) vs *traditional* media (non-interactive)

1.1: Source Credibility (the epistemic dimension)

- Content Classification
  - Problems: proliferation of LLM hallucinations, fake ads
  - Solutions: counterfactual reasoning in LLMs, fake ad detection
- Network-based Context Classification
  - Problem: exit & evasion tactics make blocklists redundant, bulk ad accounts creation
  - Solution: network-based detection & discovery of low credibility sites, collusion detection
- Combining Content and Context
  - Problem: coded messaging targets data voids (social media → search engine)
  - Solution: cross-media dredge word discovery pathway
- Intervention
  - Problem: blackhat SEO boosts low credibility sites
  - Solution: blackhat SEO detection → PageRank intervention
- Complication: low credibility sites skew (extreme) right wing

1.2: Political Bias (the ideological dimension)

- Credibility X Bias
  - Problem: low credibility sites skew right wing
  - Solution: debiasing detection systems using ML-based fairness methods
- Bias Label Convergence
  - Problem: news sites may change their political leaning over time
  - Solution: propagation of bias labels based on co-event coverage temporal network (GDELT)
- Credibility-pluralism tradeoff (journal tbd)
  - Editorial decision data: event coverage in GDELT + Citations in CommonCrawl
  - Problem: credibility-pluralism appear to be competing ideals, requiring tradeoff
  - Solution: partially explained by the relationship between editorial decisions and social media engagement data (FB URLs)
- Complications:
  - Social vs Editorial: 'out-group' interaction with news media increasing on social media but decreasing in editorial decisions
  - Ownership: sites making similar bias-based editorial decisions often have the same owner

1.3: Corporate Ownership (economic dimension)

- Credibility X Bias X Ownership
  - Problem: there is no centralized media ownership dataset
  - Solution: scrape wikipedia data and validate with human annotation (93% accuracy)
- Regulation on Media Concentration (IC2S2, journal tbd)
  - Problem: FCC regulations on audience limits are hard to impose (measurement issues)
  - Solution: incorporating network affiliation data into measure of ownership concentration demonstrates how the FCC can detect when regulations are circumvented
- Disincentivizing Concentration
  - Problem: financial incentives encourage circumvention of regulations
    - For instance, some media companies make majority of revenue from affiliate agreements
  - Solution: profile acquisitions + revenue data (EDGAR, S&P datasets) and ad monetization opportunities (Google)
- Complication: it turns out to be more efficient for companies to specialize on a specific medium. Large media conglomerates often have specialized subsidiaries that exclusively operate in print media, or radio, or television etc.

CASOS IDeaS
Carnegie Mellon University

1.4: Media Types (technological dimension)

*"The medium is the message… the personal and social **consequences** of of any medium result from the **new scale**."*

~ Marshall McLuhan

- Media type modulates other dimensions
- Different media lead to distinct operating environments for each dimension
  - Online media (interactive): news sites, social media, search engines, LLMs
  - Traditional media (non-interactive): news papers, radio, television
- Detecting / constructing dimensions in these environments requires bespoke datasets and methodologies

1.4: Media type modulates other dimensions

Dimension	Traditional Media	Interactive Media	LLM-Mediated
Credibility	Editorial gatekeeping metrics; Fact-checking processes; Source reputation	Link-based trust signals; User ratings; Citation patterns; Social validation	Knowledge conflict detection; Hallucination rates; Source attribution capability; Retrieval accuracy
Political Bias	Editorial position statements; Story selection patterns; Framing choices	Network homophily patterns; Algorithmic amplification; Echo chamber formation	Prompt sensitivity; Training data biases; Response variance across political topics; Framing consistency
Ownership	Media conglomerate disclosures; concentration; advertiser influence	Platform governance structures; Monetization models; Content moderation policies	Training data provenance; API access patterns; Deployment business models; Parameter licensing

1.4 Agent Control pressures depend on media type

Agent	Search	Social	News	LLM
Adversarial agent	SEO, ads	bots	blog networks	RAG poisoning
Content Creators	clickbait	audience modeling	event coverage	Jailbreak
Platforms	pagerank	social feed	ownership	RLHF, data

Part 2.1: Multi-agent dynamic scenarios

- Building on the core dimensions of news media
- Simulate how various actors use adversarial attacks in information environments
  - blackhat SEO, ads, dredge words, pink slime, etc
- Present a system for learning effective policies for combating adversarial manipulation, trained on the simulation environment
- Scenarios will be validated, with individual trajectories / runs through the simulation being developed enough to work as individual scenarios in OMEN setting

Part 2.2: Behavioral determinants and interventions for media users (Bottom-up)

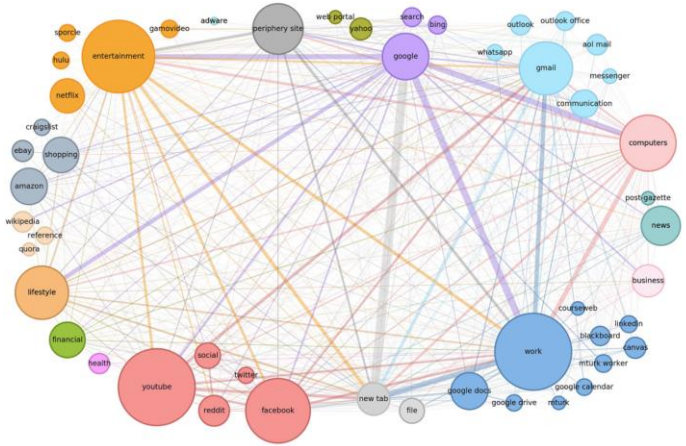
- How do these dimensions affect user traffic? (TIST, clickstream data)
  - What types of user engagement are most affected?
    - Clicking, viewing, commenting, sharing (FB URL), purchasing (Ads)
  - Is the uses and gratification theory of media consumption being undermined?
- Can users perceive the effect of these dimensions on their consumption?
  - Accountability in adversarial information retrieval: can experts tell? (OMEN)
  - Design a media literacy tool to help users *discover* influences on their consumption
- Can AI detect influences (given network context)?
  - Input credibility: Yes, but only if trained to reason over knowledge context (ICWSM)
  - Blackhat SEO: evaluate / train LLMs on reasoning over network data
    - Complication - Lack of accountability in SEO
    - Iranian news network analysis (SBP-BRiMs)
  - Design a tool where LLMs help users *mitigate exposure* to influences

OMEN Scenario Study

- If users were aware of manipulation, would they change their behavior?
  - SEO -> search strategies
  - Clickbait -> clicks
  - Source credibility / ownership -> news source selection
- Setup information environment with credibility, bias, and ownership influence
- Scenario: should country X cancel their election?
  - Modeled on the cancelled 2024 Romanian presidential elections
- Control group: browser without any plugins
- Discovery group: browser with *media literacy plugin* that highlights influences
- Mitigation group: browser with *LLM agent* that detects and mitigates exposures
- Pre/Post Measurements:
  - Instrument for measuring media influence perceptions
  - Test users on their ability to identify instances of media influence

Simulation and Model validation with clickstream data

- Alessandro Acquisti's dataset (CMU)
  - Focus on Ads, requires IRB
- National Internet Observatory (Northeastern)
  - General online interactions, requires RDA
- Uttara Ananthakrishnan's dataset (CMU)
  - Lowest friction, also potential for followup on ads click-through datasets
- Example →



How do home computer users browse the web?
K Crichton, N Christin, LF Cranor
ACM Transactions on the Web (TWEB), 2021

CASOS IDeaS
Carnegie Mellon University

Theme 1: dimensions and behaviors are connected via platform control pressures

- Low credibility sites have less traffic but can use blackhat SEO to boost traffic without having to fear reputational damage
- High credibility sites link to centrist sources and must fear reputational damage, so cannot use blackhat SEO
  - If they did use blackhat SEO, they may face boycotting and/or deplatforming
- Conspiracy sites can sell snakeoil to skeptical audience members who refuse advice from health institutions
  - Meanwhile such products are prohibited via Google Ads terms of service
- Extremely biased news sites link to other news sites from across the political spectrum, but have low reliability (credibility pluralism tradeoff)

Theme 2: control pressures and user needs

- Uses and gratification theory: “users seek out media that meets their needs” vs.
- Control pressures: “incentive structures and algorithmic recommendation systems determine what users see”
- Do control pressures reflect user needs? Do they exist because of user demand?
- Or, are control pressures interfere with a users needs, presenting an actual harm?
  - Do control pressures exist because of financial incentives (manufactured consent) and/or foreign interference and information operations? Are data voids created or exploited?
- Do trends in social media engagement data explain news media editorial decisions?
  - Do ownership changes lag or lead changes in user demand for certain topics?