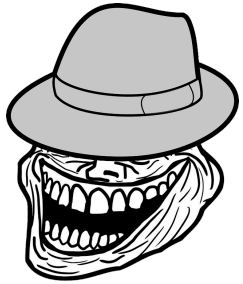
Beyond the Lulz:

Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape

Matt Goerzen & Jeanna Matthews

Media Manipulation Initiative | Data & Society

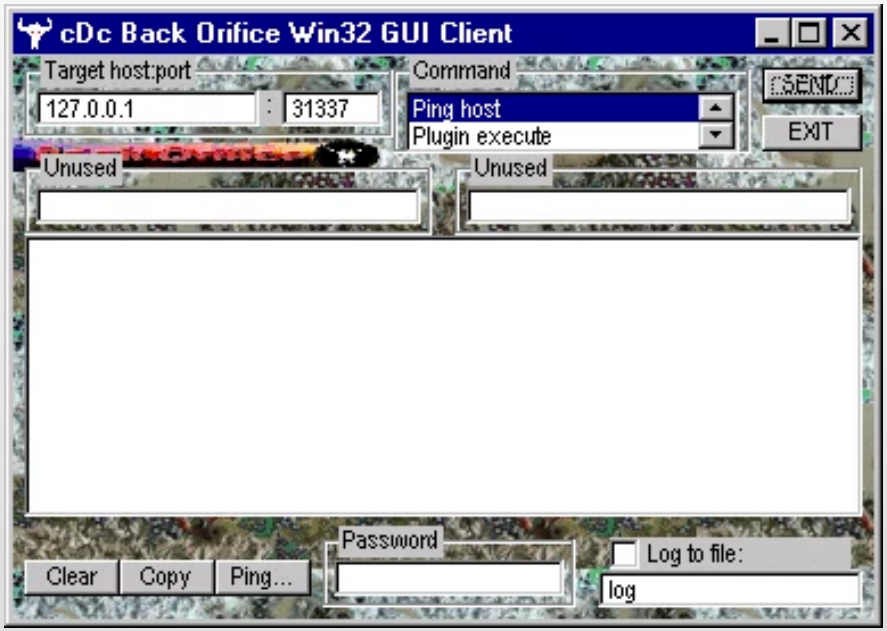
www.datasociety.net



INTRO

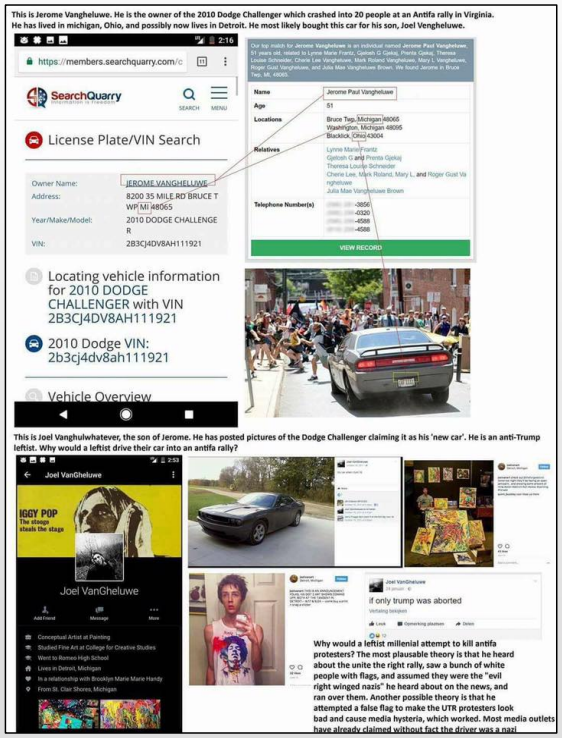
HACKING and TROLLING

BACK ORIFICE



August 1, 1998 @ DEF CON

UNITE THE RIGHT DRIVER



August 12, 2017 @THE INTERNET

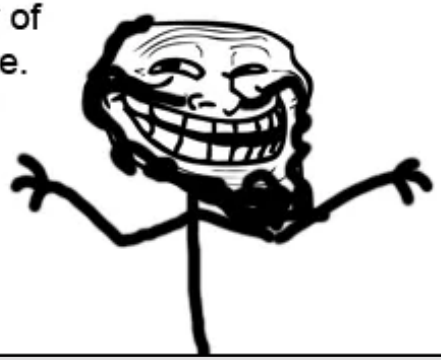
Trolling is Attention Hacking

Hackers have Trolled

Others Also, say Trolls

Socrates

I can't prove that I know anything, therefore the very statement "I know nothing" has the possibility of being false. Problem?



And Others Too, say Others

Need Some Rigor

First: Mode of governing other users on Usenet

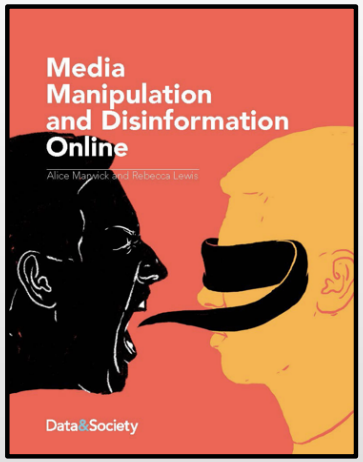
Later:

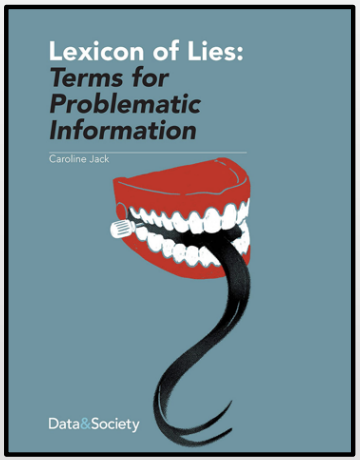
- Imageboard users raiding for “lulz”
- Hacktivists drawing media attention to activist causes
- Networked harassment and abuse (often w/ political intent)
- Using controversy to sow chaos and amplify disinformation
- Baiting journalists into reporting false information
- Manipulating popular opinion

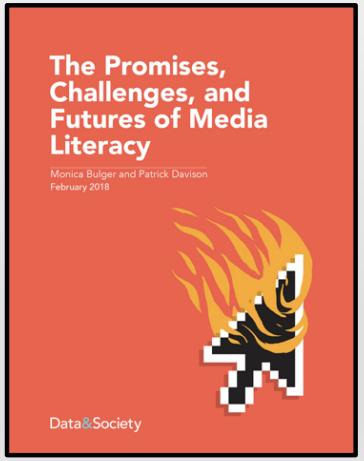
Why We're Here

To Discuss & Differentiate… With You

SOME BACKGROUND







SOME CONTEXT

More about us

What We'll Discuss

1. Media (news, social) as **Security Systems**
2. A range of **Techniques**
3. Possible **Goals** and **Outcomes**
4. Various **Targets**
5. Potential **Vulnerabilities**
6. **Black Hat, Gray Hat,** and **White Hat** dispositions
7. **Ethics** & **Threat Models**
8. Trolling: A **Bug** or a **Feature?**
9. The **Role of the Troll?**

What We'll Draw Upon

40+ CASE STUDIES

1. The Media is a Security System

1.a The Media is a Security System

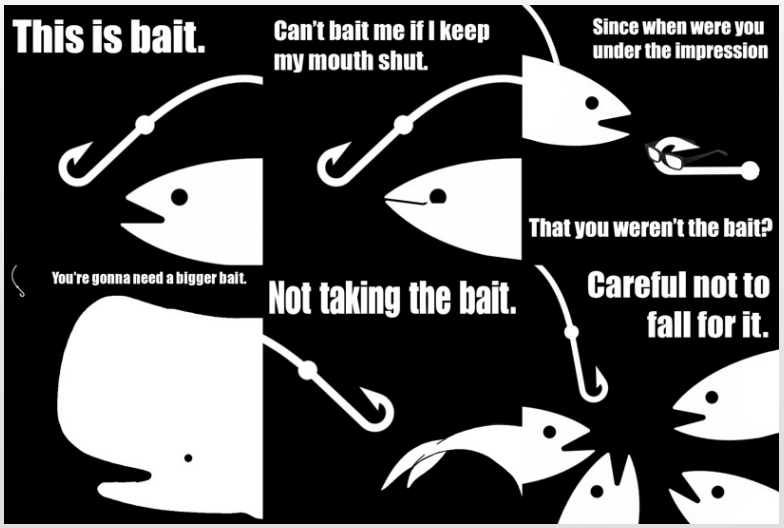
- Trolling is to socio-technical systems what hacking is to technical systems
- "Attention hacking" to set the media agenda, directing journalists, social media users and others to information that serves their interest
- New media = new vulnerabilities
- Advertisement & "If it bleeds it leads" = unpatched vulnerabilities

2.a There are a Range of Techniques

- Source Hacking / Journobaiting
- Attentional Honeypots
- Keyword Squatting
- Capturing the Narrative
- Cognitive Denial of Service
- Brute Force Harassment
- Ironical Bait & Switch
- Gaming Databases / Algorithmic Control
- Memetic Trojan Horses
- Controlled Opposition
- Driving a Wedge / Divide & Conquer
- Concern Trolling
- Brigading / Dogpiling
- Sockpuppetry
- Jiu Jitsu / Self-Victimization
- Doxing
- Context Distortion
- Deep Fakes / Photoshops

2.b There are a Range of Techniques

- Troll
- Bait
- Target
- Third-party Interpreter
- Intended outcome
  - Lulz
  - Politics



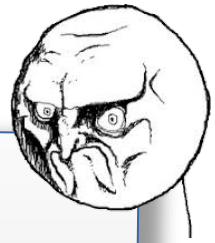
3.a The Motivations May Vary

- Political Maneuvering
- Chaos / Scorched Earth
- Perception Management
- Denial of Service
- Lulz / Drama
- Agenda Setting / Priming
- Incitement
- Response seeking
- Community building
- Community governance
- Demoralization / Outrage
- Profit???

3.b The Outcomes are Often Political

- **Draw** attention to ignored issues (“agenda setting”)
- **Waste** attention and **distract** from issues
- **Direct** attention to **pressure** entities (“security economics”)
- Draw a telling **response** (transparency and accountability)
- **Deny** a response
- **Diminish** an opponent’s resources or morale – perhaps to exhaustion

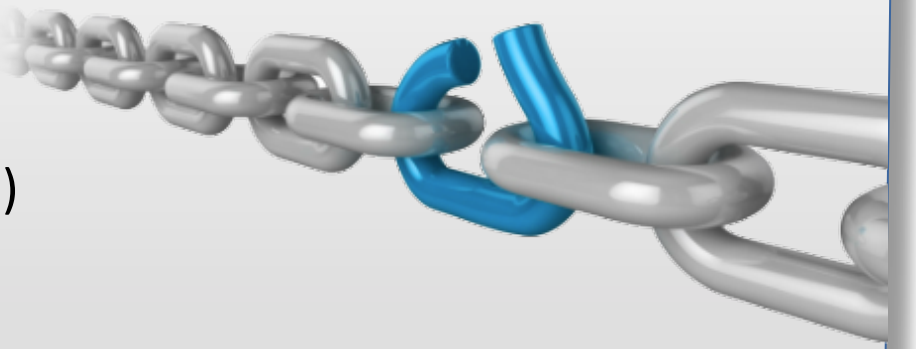
4. There are Various Targets



- **Peers**
- **Journalists** (and their **readers**…)
- Social Media **Users** (potential **allies**)
- Social Media **Users** (**perceived threats**)
- **Politicians** & **Regulators**
- **Companies** and **vendors**

5. There are Vulnerabilities?

- Clickbait / ”Bleeding Leads” / Breaking News
- Network curation
- Asymmetric cost to debunk
- Automated amplification (bots)
- Adtech
- Context collapse / Poe’s Law / Anonymity



6. There are different Ways of Seeing

6.a Black hat

- Push private agenda
- Counter to target's interest
- Non-prescribed engagement

6.b Gray Hat

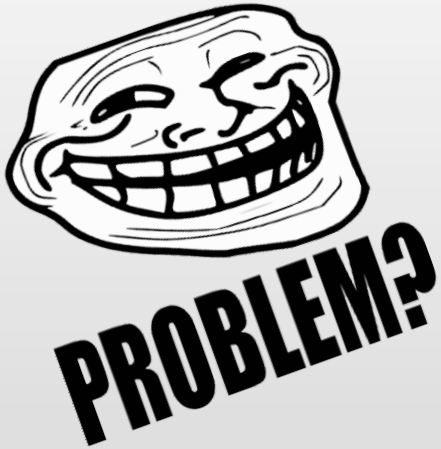
- Non-prescribed engagement
- Sense of public interest
- Demonstrate attack surface (proof of concept)
- Increase pressure on gatekeepers
- Troll to Reveal?

6.c White hat

- Reveal vulnerabilities selectively to platforms
- Disable possibility of non-prescribed engagement through research and education
- Reduce attack surface
- Align system with intent
- Blue team platforms… for profit?

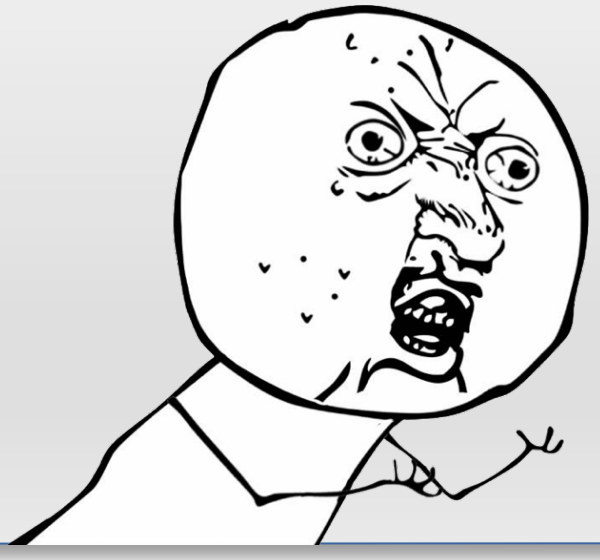
7.a Is there an Ethos?

- Is “I did it for the lulz” sufficient?
- Is “anti-” enough?
- Who judges?
- Does it look different from different perspectives?



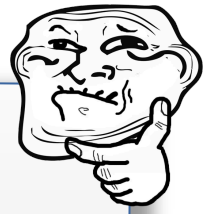
7.b Who Frames It?

- Where do we locate the public good?
- What are the threat models?



8. Trolling: a Bug or a Feature?

8.a Free Speech

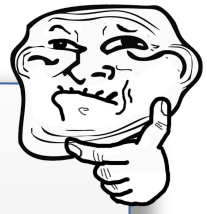


- Right to free speech with your one mouth?
- Right to amplification of that speech?
- Monopolization of attention = DOS of other's free speech?

8.b Everyone an investigative journalist?

- Finite lifetime problem
- Can’t verify/investigate every claim
- Inflammatory lies stickier and more profitable

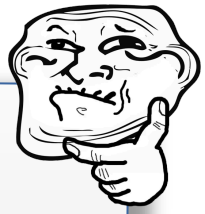
8.c Gatekeeping?

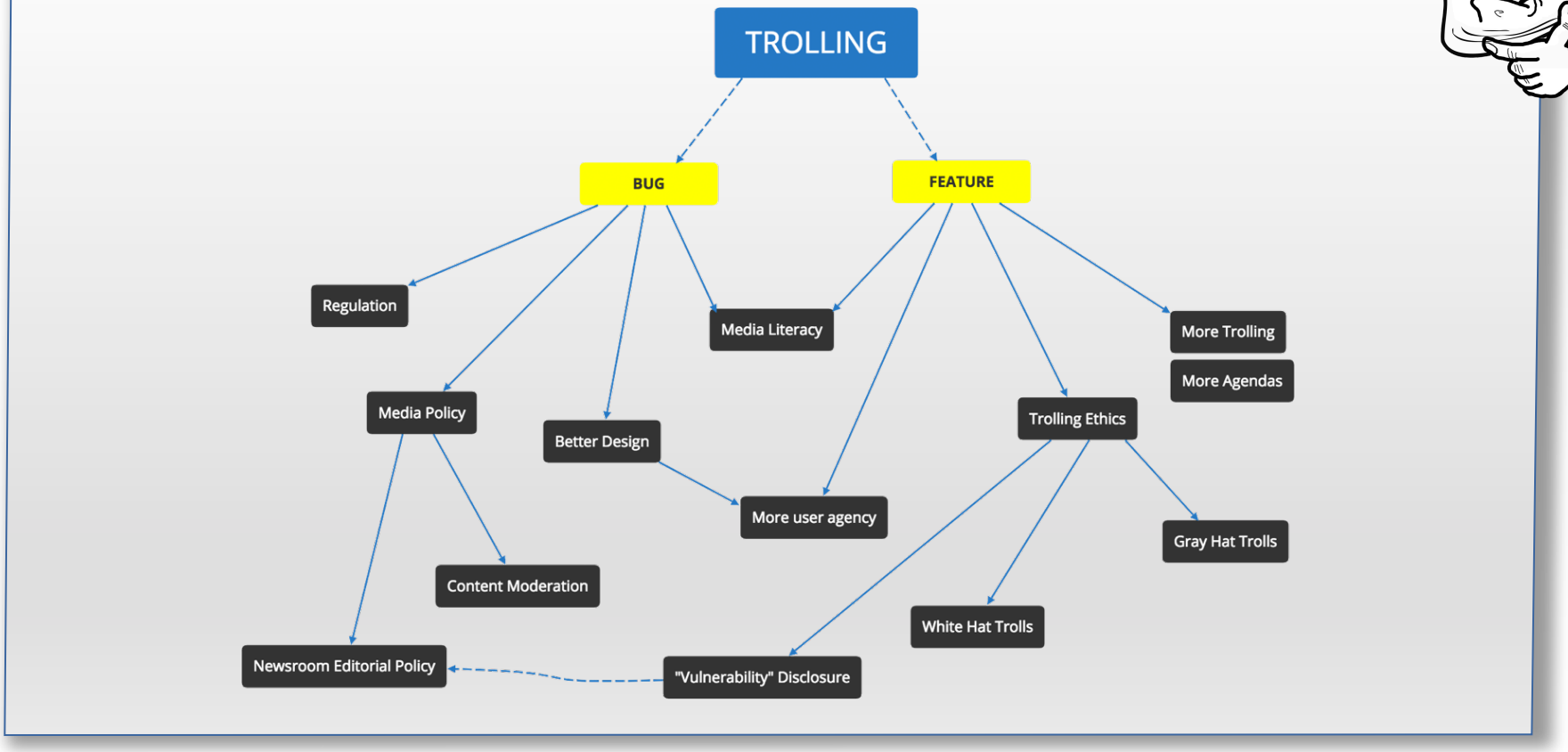


- Gatekeeping under counts
- No gatekeeping over counts
- Do-it-yourself gatekeeping tends to tribalism?
- Old gatekeeping wasn't working. Rules for sharing our attention landscape? Long-term goals? Short-term changes?

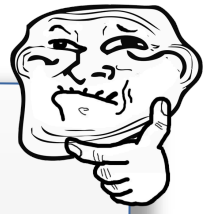
- Gatekeeping but not too much? Cracks are where the light gets through!

9.a And So?



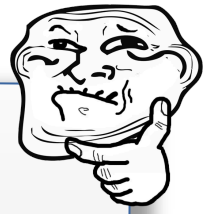

TROLLING
BUG
FEATURE
Regulation
Media Literacy
More Trolling
More Agendas
Media Policy
Better Design
Trolling Ethics
More user agency
Gray Hat Trolls
Content Moderation
White Hat Trolls
Newsroom Editorial Policy
"Vulnerability" Disclosure

9.b Top-down interventions



- Platform re-design
- Media Literacy
- Trust and Verification
- Disrupting Economic Incentives
- Banning Accounts
- Regulatory Approaches
- Don’t Feed the Trolls / Strategic Silence

9.c Bottom-up interventions



- Doxing “bad actor” trolls
- “Vulnerability” disclosure
- Shaming or trolling platforms and newsrooms to change
- P2P moderation
  - White lists
  - Call outs
  - “Countermemes” or “discursive patches” like Godwin’s Law or “Don’t feed…”
  - Redirection / board sliding

9.d Role of the Troll?

- Inform interventions by:
  - “Security research” / “Penetration testing”
  - Informing media companies / platforms of vulnerabilities?
  - Identifying bad actors and their incentives?
- Intervening directly by:
  - Putting pressure on media companies to change?
  - Demonstrating attack vectors as “proofs of concept”?
  - Trolling trolls?