Lowering the Barrier for Web Advertisement Research at Scale

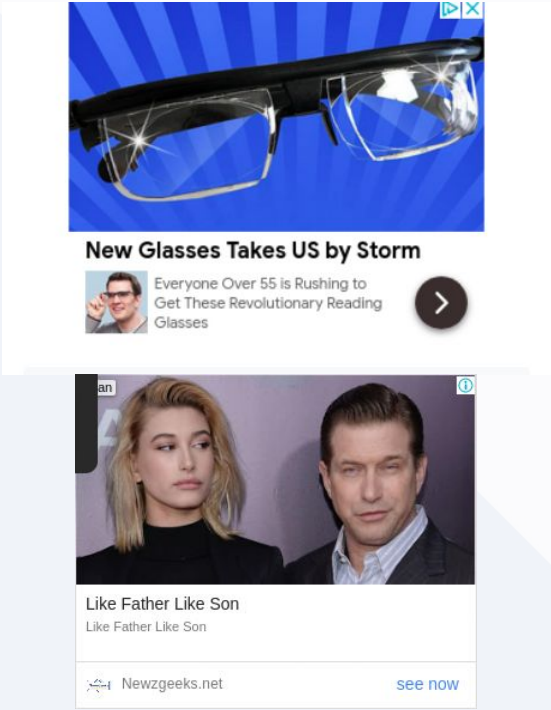
Kevin Feng
kjfeng@princeton.edu
with Arunesh Mathur and Arvind Narayanan
Princeton University

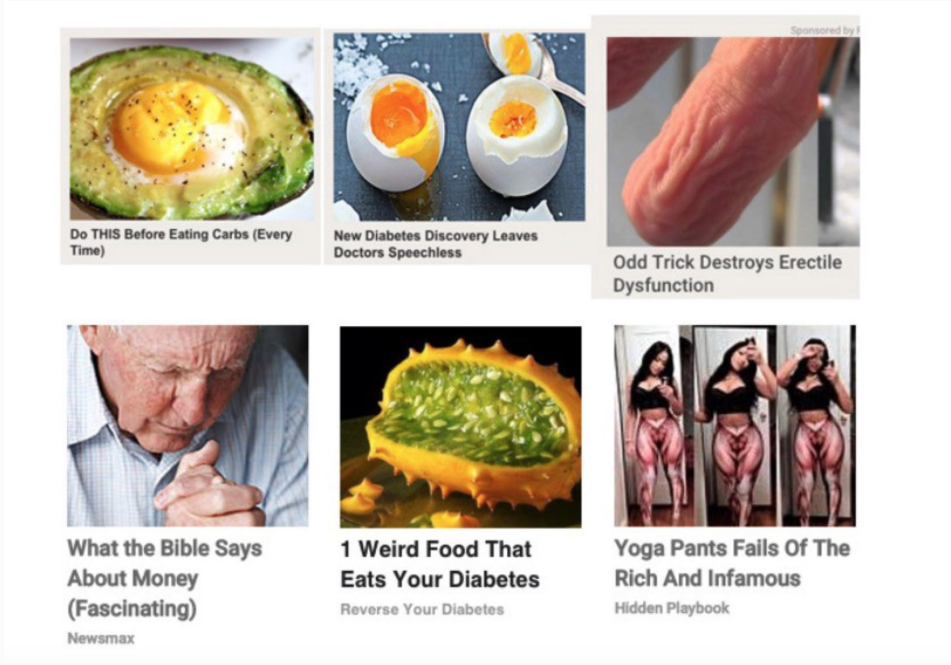
Agenda

- Motivation
- Ad Collection
- Search Interface
- Demo
- Research Questions & Findings
- Policy Recommendations

The Ad Market

- Low quality and deceptive techniques are ubiquitous

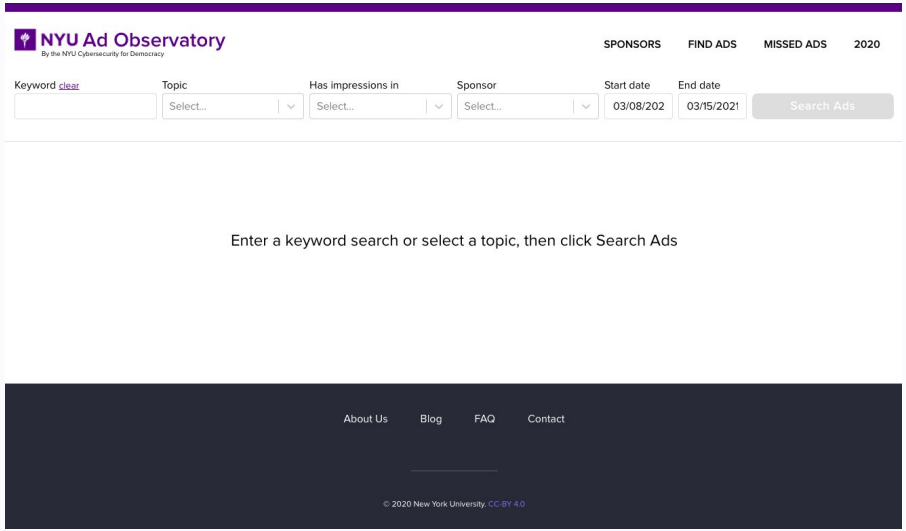




Motivation

- Ads are hard to study because they appear and disappear unpredictably
- Ads are hard to track for policymakers
- Research and policymaking should have more interdisciplinary collaboration

Motivation – Previous Work



NYU Ad Observatory

Motivation – Previous Work

Bad News: Clickbait and Deceptive Ads on News and Misinformation Websites

Eric Zeng, Tadayoshi Kohno, Franziska Roesner
Paul G. Allen School of Computer Science & Engineering
University of Washington

***Abstract*—A key aspect of online ads that has not been systematically studied by the computer security community is their visible, user-facing content. Motivated by anecdotal evidence of problematic content such as clickbait, misinformation, scams, and malware, particularly in native advertising, we conducted a systematic measurement study of ad content on mainstream news sites and known misinformation sites. We provide evidence for significant numbers of problematic ads on popular news and misinformation sites, primarily served through native ad platforms. This work begins a rich, systematic line of inquiry into problematic ad content, ultimately to inform technical and/or regulatory solutions.**

Fig. 1. Portion of a "chumbox" native ad banner, showing four ads that use "clickbait" techniques to entice clicks (such as distasteful imagery, sensationalism, provoking curiosity, and urgency). Such ads often lead to low-quality sites, misinformation, or outright scams.

I. Introduction

Online advertisements are an unavoidable fact of the modern web — they are embedded in and financially support the

have the potential to generate a significant amount of revenue. Prior work has shown, and native ad platforms themselves claim that they generate significantly higher clickthrough rates

What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising

Eric Zeng
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA
ericzeng@cs.washington.edu

Tadayoshi Kohno
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA
yoshi@cs.washington.edu

Franziska Roesner
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA
franzi@cs.washington.edu

ABSTRACT

Online display advertising on websites is widely disliked by users, with many turning to ad blockers to avoid "bad" ads. Recent evidence suggests that today's ads contain potentially problematic content, in addition to well-studied concerns about the privacy and intrusiveness of ads. However, we lack knowledge of which types of ad content users consider problematic and detrimental to their browsing experience. Our work bridges this gap: first, we create a taxonomy of 15 positive and negative user reactions to online advertising from a survey of 60 participants. Second, we characterize classes of online ad content that users dislike or find problematic, using a dataset of 500 ads crawled from popular websites, labeled by 1000 participants using our taxonomy. Among our findings, we report that users consider a substantial amount of ads on the web today to be clickbait, untrustworthy, or distasteful, including ads for software downloads, listicles, and health & supplements.

that 18% of U.S. internet users and 37% of German internet users used an ad blocker [69], a large percentage considering that it takes some initiative and technical knowledge to seek out and install an ad blocker.

There are many drivers of negative attitudes towards online ads. Some users find the mere presence of ads to be problematic, often associated with their (perceived) increasingly disruptive, intrusive, and/or annoying qualities [5] or their impact on the load times of websites [92]. Users are also concerned about the privacy impacts of ads: research in computer security and privacy has revealed extensive ecosystems of tracking and targeted advertising (e.g., [9, 28, 30, 61, 62, 64, 76, 84, 97, 98]), which users often find to be creepy and privacy-invasive (e.g., [29, 96, 100, 101]). The specific *content* of ads can also cause direct or indirect harms to consumers, ranging from material harms in the extreme (e.g., scams [1, 34, 72], malware [65, 74, 104, 105], and discriminatory advertising [3, 57]) to simply annoying techniques that disrupt the user experience

Zeng et al. Conpro 2020 and CHI 2021

Our Contributions

- Large dataset of web ads from popular websites
  - 12K+ ads collected from 638 publishers (crawled from a list of 3330)
  - 8859 processed and available for public browsing
- Automated visual analysis and search interface
- Consequently, research questions and policy implications

Ad Collection

- Crawler built on OpenWPM, runs in FireFox
- Repurposed Cliqz Adblocker JS library to parse EasyList selectors/domains and capture ads instead of blocking them
- Capture ads by taking a screenshot of them and collecting metadata
- Data collected from each ad:
  - Image of ad (screenshot)
  - Publisher
  - Time
  - Location
  - HTML

Visual Analysis

- Assisted by Google Cloud Vision API
- OCR, paragraph segmentation
- Object detection
- Face detection
- Disclosures
  - AdChoices Icon
  - Mute Icon
  - Text disclosures (terms + conditions, “Advertisement”, etc.)
- Brand and industry detection
- Colour palette extraction
- Size classification

AdOculos Search Interface

- Searches on:
  - ad text
- Filters on:
  - brand, industry, size, colour, disclosures, object, face, publisher, date, location

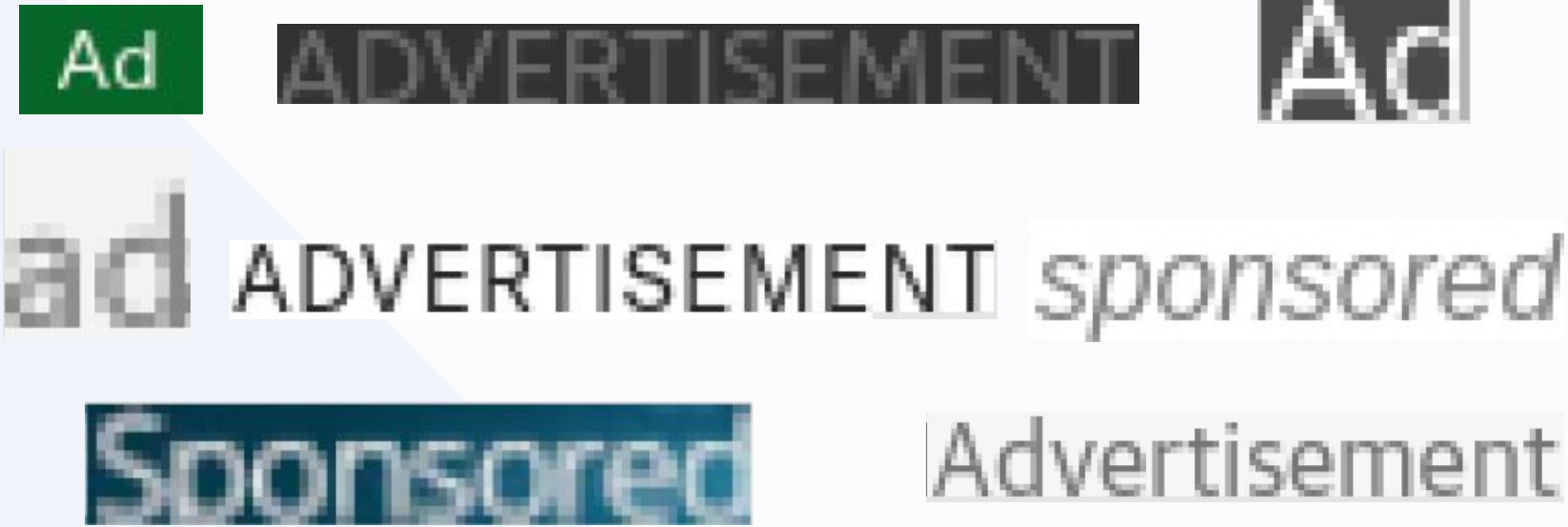
Demo

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

Findings

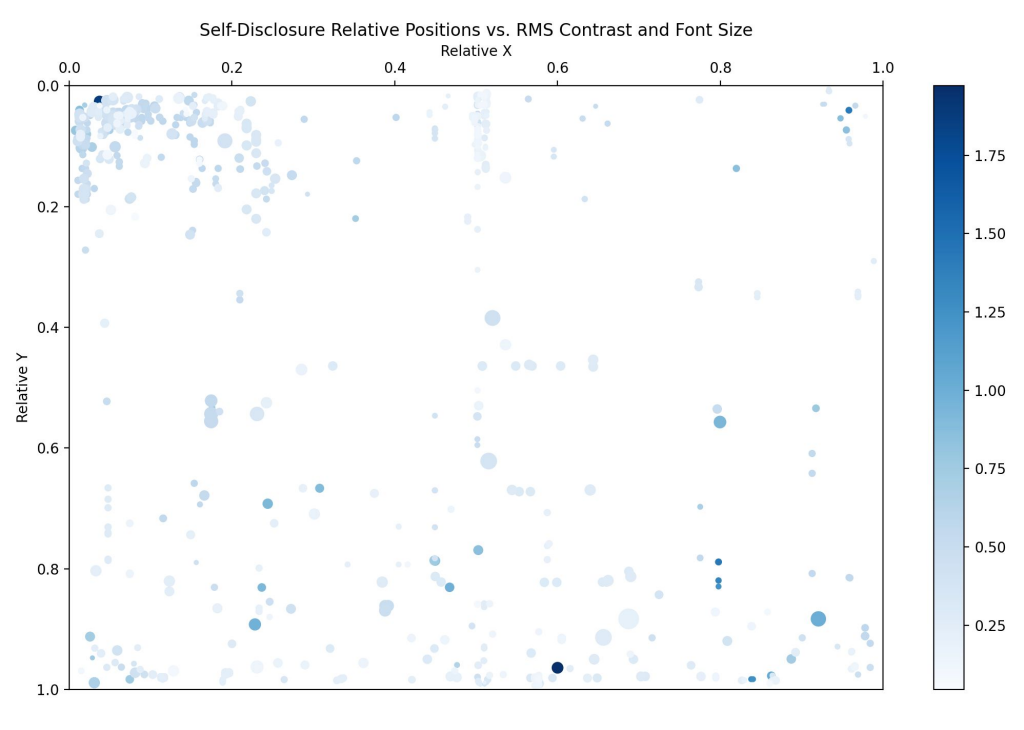
How are self-disclosures represented?

- Ads have various ways of disclosing themselves as ads



Self-disclosures analysis

- Analyzed placement, font size, and contrast



Self-disclosures analysis, cont.

	n	RMS Contrast	Font Size (px)
all	1048	Mean: 0.36 Median: 0.30	Mean: 13.77 Median: 13.0
top quartile	696	Mean: 0.37 Median: 0.37	Mean: 13.30 Median: 13.0
bottom quartile	260	Mean: 0.32 Median: 0.24	Mean: 14.37 Median: 13.0

Ad similarity analysis

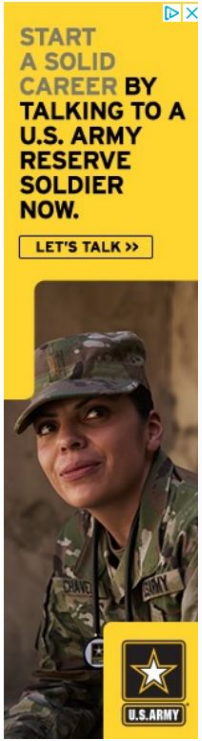
- Within one webpage, how similar are the ads?
- Similar by industry, brand, messaging
- Examined ads from top 30 publishers, each publisher had between 29 and 52 ads, averaging 33.4
- Found that on average, 49.7% of ads on the same page belonged to the same industry
  - 72.8% for publishers with ad-content alignment
  - 36.3% for publishers without

Ad similarity analysis, cont.

- Proportion of ads with the same brand:
  - 37.9% for publishers with ad-content alignment
  - 26.0% for publishers without
- Proportion of ads with same messaging:
  - 23.8% for publishers with ad-content alignment
  - 17.6% without
- Why? Some researchers suggest that revenue maximizing algorithms for clustering ads together and selling slots as "package"

Ad similarity problematic?

- citationmachine.net: collected 15 Army ads (30 total)



Ad similarity problematic?

- nine.com.au: collected 9 blood sugar ads (30 total)

One Simple Method To Keep Your Blood Sugar Below 100

SPONSORED | world-health-wellness.com

Doctors Can't Explain Why This Fruit May Cut Your Blood Sugar By 90%

SPONSORED | Gluco Shield Pro

Start Eating This Everyday And You Will Reduce Blood Sugar

SPONSORED | Gluco Shield Pro

Policy Implications

- Publishers:
  - Obtain concrete evidence of problematic ad practices on their site
  - Create "community standards" for ads on their platform
- Intermediaries:
  - Diversify ad placement on websites by capping the number of ads per cluster
  - Make practices more transparent to other agents in the market
- Consumers:
  - Standardized ad reporting mechanism that communicates directly with regulators
- Regulators:
  - Roll out guidelines for ad self-disclosure style/location, similar to AdChoices

Future Work

- AdOculos UX:
  - Supporting browsing of large quantities of data
  - Group together similar ads
  - Interactive data visualizations
  - API and documentation
- Crawler
  - Dismiss popups
  - Run profiled crawls
  - Run crawls from multiple locations
  - Click into links and pages

Other Research Questions

- Depiction of people
  - Race, gender, emotion, etc.
- Targeting of vulnerable populations
- Imitation of UX (shopping, games)
- Claims – false vs. puffery
- Shady publishers
- Ads that compare their product to others
- Measuring regret in ad interactions
- Understanding ads with computer vision
- New ML challenges (perceptual ad blocking)
- Many more to be explored...


CENTER FOR
INFORMATION
TECHNOLOGY
POLICY
PRINCETON UNIVERSITY

Questions?

Thank you!

Email: kjfeng@princeton.edu

Slack (CITP): @Kevin Feng