Misinformation on Search Engines: Networked Characterization and Identification

Evan M. Williams

CMU-S3D-26-106

April 2026

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley, Chair
Uttara Ananthakrishnan
Francesca Tripodi
Jamie Callan

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing.

This work was supported in part by the Office of Naval Research, MURI: Persuasion, Identity, & Morality in Social-Cyber Environments under grant N000142112749 and by the Knight Foundation. It was also supported by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the US government. Chapter 2 was supported in part by the Stanford Internet Observatory and the Stanford Cyber Policy Center. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, Department of Defense, or the Office of Naval Research.

Keywords: Search Engines, Web Search, SEO, Data Voids, Information Gaps, RAG, GNN, Web Graph, Misinformation

Abstract

I explore how misinformation spreads through websearch, with a focus on how and why unreliable websites can rank highly in search engines. I additionally introduce models that can help search engines improve the reliability of search results. I first explore the limitations of Google's current efforts to combat low-reliability and low-relevance content in its search engine. I then demonstrate both how owners of unreliable websites can manipulate structural authority patterns present in the webgraphs, and introduce content-agnostic Graph Neural Networks to detect unreliable websites using these signals. I then explore the content-based approaches that unreliable websites use to rank on search engines and again propose models for detecting them. I also explore how the presence of unreliable websites on SERPs impacts Google's AI overviews. I then combine the content and webgraph with social media mentions of unreliable websites to create a realistic model of the paths that users can take to reach an unreliable website in hopes of better understanding the signals present in each. Finally, I combine the takeaways from the previous chapters into a conceptual framework that will be integrated into a simulated information environment exercise, where misinformation experts will use the work developed in this thesis to characterize and identify misinformation and narratives.

Acknowledgments

I would like to thank my advisor, Kathleen M. Carley, for her mentorship and support throughout this work and for seeing the value in this line of research. I'd like to thank my committee, Uttara Ananthakrishnan, Francesca Tripodi, and Jamie Callan for their support, encouragement, and guidance. I'd also like to thank each of my collaborators: Peter Carragher, Ronald E. Robertson, Ramon Villa-Cox, David Thiel, and my former interns, Kyle Herdrich, Luke Prakarsa, and David Li.

I would like to thank my labmates who've helped make my time in the program so meaningful. I'd like to thank Peter Carragher for being a close friend throughout this entire process. I'm thankful for Reba Marigliano and John Benson with whom I can always laugh. I'm thankful for Catherine King and her husband Travis and the many board games we've played. I'm grateful for the deep conversations with Samantha Phillips, for the insightful conversations with Lynette Ng, and for the kindness of Daniele Belutta. I'm also thankful for the labmates who paved the way, most notably Charity King, Joshua Uyheng, Janice Blane, Matt Hicks, Tom Magelenski, and Ramon Villa-Cox.

I would like to thank my family who supported me through this process. I would like to thank my parents, Phyllis and Tom. I'm thankful for my dad for always encouraging my curiosity and pushing me to think scientifically. I'm thankful for my mom for teaching me how to care for others. I'd like to thank my sister Carmen for reminding me how to lead a good life, and her husband Brandon, for introducing me to great horror films. I'm also grateful for my aunt Connie's courage and her amazing shirts, and for my cousins Maddy and Ren. Finally, I would like to thank Hunter, who reminds me every day how to find kindness in the world and in myself. I dedicate this to all of you.

Contents

List of Figures

Chapter 1

Introduction

On January 1st, 2024, a 19-year-old college student, Matthew Sachman, who went by Matteo, fell onto the tracks of a New York City subway and was killed by an oncoming train. In the hours after his tragic death, as friends and family scoured the internet for any information on what had happened, they encountered numerous strange and AI-generated obituaries littered with false information; articles claimed that Matteo Sachman was 29 years old, that he was from Nantucket, and that he'd been stabbed on the subway platform [Keh and Thompson, 2024b]. The family was seeing search results populated by "obituary pirates"—individuals who target keywords around recent deaths as an opportunity to generate clicks and ad revenue [Knibbs, 2023]. Obituary pirates exploit what researchers often refer to as *data voids*—search results that are dominated by unreliable, irrelevant, or low-quality websites [Golebiewski and boyd, 2019]. Obituary pirates and data voids exist because search engines assume that every query has a relevant response. However, many queries, such as "pope red shoes human skin", receive coverage on unreliable websites, but not on reliable news sites. In a 2024 editorial, *Nature* called data voids an existential threat to democracy, and called for more research on the topic Nature Editorials [2024].

Search results, both political and non-political, receive a remarkable degree of global trust. Since 2016, global polls have consistently identified search engines as the most trusted source of information, ahead of both social and traditional media [Edelman, 2021, Barometer, 2024, McDuling, 2015]. The order in which search rankings are returned has been shown to impact candidate preference and voting behaviors [Epstein and Robertson, 2015b, Epstein et al., 2017]. High-ranking pages—the pages that appear at the top of a Google search result—are far more likely to be seen. In a 2013 analysis of 300 million search engine clicks, 92% were on the first page of search results, and 51% of those were the first or second result [Insights, 2014].

Additionally, users were found to be 140% more likely to click the last result on the first page than to click the first result on the second page [Insights, 2014]. More recent analyses by Backlinko, an SEO firm, and Ignite Visibility, a digital marketing firm, both found that the click-through rate of the first result was ten times higher than that of the tenth result [Lincoln, 2020, Dean, 2022]. Consequently, misinformation returned via search engines can have substantial impacts on information environments.

Misinformation on search engines has become increasingly concerning in recent years with the rise of what Marwick and Partin call populist expertise, or the rejection of experts and traditional information vectors in favor of alternative, "home-grown" knowledges [Marwick and Partin, 2024]. Phrases like "Do your own research" or "DYOR" have become ways for users to signal skepticism toward medical and scientific expertise, particularly around vaccine safety [Chinn and Hasell, 2023]. In the manifesto published by mass-murderer and neo-Nazi Dylann Roof, Roof attributed his radicalization to the first time he typed the words "Black on White crime" into Google. He recounted that the first website he discovered while searching was The Council of Conservative Citizens—a misinformation news site operated by a white supremacist hate group [Hersher, 2017]. Misinformation on web search can result in real-world harms, but the study of misinformation on search engines has been fairly limited, due in large part to data collection bottlenecks.

Search engine reliability has traditionally been studied through small-scale search engine audits. In these types of studies, researchers select an (often small) set of keyphrases, query the keyphrases in one or more search engines, and compare the top k results for queries using human annotators Makhortykh et al. [2020], Urman et al. [2022a]. While these studies can provide some insights into the relative reliability of search engines given a standardized set of keywords, search engine audits are bottlenecked by the initial selection of queries; many search audit studies contain fewer than a dozen keywords. In this thesis, I rely on search directives and third-party SEO toolkits to conduct the two largest search engine audit studies to date, covering 1.4M and 4.2M unique query-SERP pairs, respectively. By also integrating webgraphs and social media data, I go beyond SERP data, and consider broader paths that users can take to unreliable websites.

There are many ways to get bogged down when thinking about the spread of misinformation on web search: the challenges of defining misinformation, constructing and selecting queries, evaluating Search Engine Results Page (SERP) reliability, accounting for mixed-reliability results, understanding how AI overviews or snippets characterize a concept, or assessing social media content returned in search results, to name a few. Many researchers have proposed

helpful concepts that partition subsets of the broader problem of misinformation on web search. Ideas like *problematic queries* [Golebiewski and boyd, 2019], *data voids* [Golebiewski and boyd, 2019], and *keyword signaling* [Tripodi, 2019a], have been examined in case studies, but can be hard to quantify at scale and do not always lead to unreliable information. The recently proposed concept of *search directives* offers a clear and more observable path by which users are directed to unreliable content. A search directive is content explicitly intended to prompt an online search, such as user a telling user b to "look up Chemtrails on Google" [Robertson et al., 2023a]. However, even in these cases, the resulting search results are not necessarily unreliable. For example, I could instruct a colleague to look up "The set of topological minors for the 1-holed torus" (a search directive). Although this may yield limited results (a data void), the information returned is nonetheless relevant to topology and graph theory.

It is worth noting here that the concept of a "data void" can describe multiple often-overlapping phenomenon. In chapter 2 I'll discuss three types of data voids that Google acknowledges: when search results 1) have low-relevance, 2) have low-reliability, or 3) are rapidly changing. While each of these phenomenon are important considerations, I will focus primarily on data voids that contain unreliable search results, as these have the greatest potential to cause harm through medical misinformation, radicalization, election misinformation, etc. To better align the data with our area of interest, I introduce the concept of *dredge words*—terms or keyphrases for which unreliable domains rank highly in search results. I will return to dredge words frequently throughout this thesis, as they represent, by definition, a primary vector through which misinformation can spread via search. Additional vectors are embedded within HTML elements on Google SERPs, including Google's snippets and AI overviews, as well as video and shopping boxes, which can contain links to dubious content on websites like YouTube and Amazon. Throughout this dissertation, I aim to understand how users access unreliable content on search engines and to introduce methods that increase the reliability of the content returned. Finally, I note that misinformation research is inherently multidisciplinary, and while the methods I propose are aimed at technical audiences, I aim for this work to be of value to social scientists, computational social scientists, policymakers, analysts, and other stakeholders in healthy information environments. My guiding questions are the following:

1. How does misinformation spread on search engines?
2. How do unreliable websites rank highly for content on search engines?
3. How do search engines interact with social media in misinformation spread?

4. How can Graph Neural Networks and LLMs be used to improve the reliability of search engine results?

In this dissertation, I highlight the limitations of Google's current efforts to combat harmful data voids. I then outline the structural authority patterns present in the web graphs of misinformation websites and introduce content-agnostic Graph Neural Networks (GNNs) for detecting unreliable websites, drawing in part on evidence of attempts to game search engine recommendation algorithms. Next, I explore content-based strategies that allow unreliable websites to rank highly on search engines and propose models for identifying such content. I also examine how the presence of unreliable websites in SERPs influences Google's AI overviews. Subsequently, I integrate content signals, web graph structure, and social media mentions of unreliable websites to create a realistic model of the paths users can take to reach an unreliable website, with the goal of better understanding the signals present in each layer. Finally, I synthesize the key findings from the previous chapters into a conceptual framework, which is incorporated into a simulated information environment training exercise. In this exercise, misinformation experts use the tools and models developed in this thesis to characterize and identify misinformation and online narratives.

1.1 Background

I argue that unreliable websites rank highly in search through two interconnected methods: 1) through algorithmic and structural authority and 2) through keyphrase relevance and competition. To increase algorithmic and structural authority, websites can undertake various actions that fall under the umbrella of Search Engine Optimization (SEO). These actions can be "white-hat", meaning in-line with Google's "Search Essentials"[1] like ensuring the website loads fast and is friendly to mobile users. Alternatively, domain owners can take "black-hat" actions that violate Google's "Search Essentials", like paying third-party services to have thousands of websites provide a target website with millions of links. These behaviors affect the authority that search engines assign any given domain. Second, while creating content, websites knowingly or unknowingly use specific keyphrases that can have various degrees of competition. "Best socks" has substantial competition, whereas the competition for "subcortical vascular dementia Hillary Clinton" is more limited. The latter is an example of a "long-tail" keyphrase, as it is likely not often searched and has little competition. A more ambiguous, open-ended question is "how do users wind up searching problematic long-tail

[1] https://developers.google.com/search/docs/essentials

queries"? Certainly, there are many situations online and offline where users could formulate queries that surface unreliable content. In Chapter 2 and Chapter 5, I examine in more depth the usage of queries on social media that also surface unreliable search results.

1.1.1 Graph Neural Networks

Graph neural networks can be expressed as a differentiable variant of belief propagation [Dai et al., 2016]. Consequently, the update function for the hidden state h_u of single node u of a GNN at layer k is often expressed as:

$$h_u^{(k+1)} = \text{UPDATE}(h_u^{(k)}, \text{AGGREGATE}(h_v^{(k)} \forall v \in \mathcal{N}^{(k)}(u)) \tag{1.1}$$

where $\mathcal{N}^{(k)}(u)$ denotes k-order neighbors of u and AGGREGATE and UPDATE are arbitrary differentiable functions. One of the most common instantiations of this formula is Graph Convolutional Neural Networks (GCNs) [Kipf and Welling, 2016]. The update function for the hidden states of a GCN can be written as

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1.2}$$

where $\tilde{D}$ is the diagonal degree matrix, and $\tilde{A}$ is the sum of the adjacency matrix and its identity matrix I. W^l and H^l are the respective weights and hidden state at layer l, where $H^0 = X$.

1.1.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) models combine LLMs with external information retrieval systems in order to ground generated responses in a specified knowledge base [Lewis et al., 2020]. Instead of relying solely on parametric knowledge stored in model weights, RAG systems retrieve relevant documents from a corpus and condition generation on this retrieved context. This has the added benefit of allowing users to see sources used by the LLM. Formally, given a query x, a retriever selects a set of documents $D = \{d_1, \ldots, d_k\}$ from a corpus $\mathcal{C}$ according to a retrieval function $p_\eta(d \mid x)$, typically implemented using dense vector similarity search.

The generator then produces an output sequence y conditioned on both the query and the retrieved documents. The resulting conditional likelihood can be expressed as

$$p(y \mid x) = \sum_{d \in D} p_\eta(d \mid x)\, p_\theta(y \mid x, d) \tag{1.3}$$

where p_η denotes the retriever distribution and p_θ denotes the language model conditioned on retrieved context.

Retrieval is commonly implemented using embedding models that map both queries and documents into a shared vector space. Given embeddings $q = f(x)$ for a query and $e_d = g(d)$ for documents, retrieval selects the top-k documents according to similarity

$$d^* = \arg\max_{d \in \mathcal{C}} \text{sim}(q, e_d) \tag{1.4}$$

where $\text{sim}(\cdot, \cdot)$ is typically cosine similarity or inner product. The retrieved passages are then concatenated with the original prompt and provided as context to the language model. In RAG experiments in this thesis, I will treat retrieval as given, as I am more interested in what RAGs do in cases where retrieved documents are "polluted".

1.2 Data

Each chapter will retain its own data section, but I will briefly cover data sources used here. In this thesis, I circumvent those bottleneck constraints by relying on Ahrefs, a third-party SEO toolkit platform for data access. Ahrefs has the 9th most active webcrawler as of May 2025 [2], and monitors Google search rankings of over 28 billion keywords[3]. From Ahrefs, I am able to extract Keyphrases for which websites rank most highly on Google, as long as SEO features, which I will discuss in more detail in Chapter 3. Along with Ahrefs data, I use data from a diverse set of sources, which I describe below. While there are clear limitations in relying on proprietary data, we demonstrate in work outside this Thesis Carragher et al. [2025] that Ahrefs metrics correlate highly with both other proprietary SEO data providers and with metrics we calculated directly on large-scale Common Crawl Foundation webgraphs.

1.2.1 Search Directives

In collaboration with researchers at Stanford, we collect the first Search Engine Results Page (SERP) on Google for 1.4 million *search directive* [Robertson et al., 2023a] queries using automated requests from a fixed location, extracted features from those SERPs (e.g., web domains), and merged those features with metrics validated in past work (e.g., the domain

[2] https://radar.cloudflare.com/bots#verified-bots

[3] https://ahrefs.com/academy/how-to-use-ahrefs/ahrefs-seo-metrics/keyword-and-search-traffic

quality scores from Lin et al. [2023]). To account for changes in Google's search results and warning banner systems over time [Munger, 2019].

1.2.2 Think Tank Webgraphs

I collect webgraph and SEO data from Ahrefs for four groups of think tanks. These consist of eight of the Kremlin-linked think tanks identified by the Institute of Modern Russia, with primarily domestic Russian audiences [Smagily, 2018a]. These are compared with eight influential Western European think tanks and eight US conservative think tanks. For Europe, I use the top eight Western European think tanks ranked in the University of Pennsylvania's 2020 Global Go To Think Tank Index Report [McGann, 2021a]. To identify US conservative think tanks, I use the eight think tanks that supplied the largest number of staff, cabinet, and political appointees in the first Trump administration [Kravitz et al., 2019a]. I contrast these think tank networks with a network of the seven Kremlin-backed proxy outlets identified in the Global Engagement Center's Pillars of Russia's Disinformation and Propaganda Ecosystem report [U.S. Department of State: Global Engagement Center, 2020b]. I call these pseudo–think tanks, as each of these proxy outlets blurs the lines between news, think tanks, misinformation, and propaganda. The purpose of these pseudo-think tanks is to spread Russian state propaganda to Western audiences.

1.2.3 Website Reliability Webgraphs

I use the SEO toolkit service Ahrefs[4] to extract SEO data for the 11.5k websites in the reliability-labeled dataset proposed by Lin et al. [Lin et al., 2023]. For each of the domains, I also extract the 10 domains that link to each of the 11,520 target domains at the highest volume (the highest-volume back-linking domains) from Ahrefs.

1.2.4 Reliability-labeled Keyphrases and SERPs

Again from Ahrefs, I extract the 1k keyphrases for each of the 11.5k websites with the lowest Google positions (where position 1 would correspond to the top result on the first SERP). Once I have these keyphrases, I scrape Google SERPs and AI overviews for each keyphrase. This dataset contains over 4 million Keyphrases and over 40 million search results. For each

[4] ahrefs.com

of these pages, I also scrape AI Overviews (if present), and all sources cited within each AI overview.

1.2.5 Covid Twitter Data

For social media context, I use a Twitter dataset constructed by querying COVID-related keywords[5] via Twitter's streaming API between January 29, 2020 and June 26, 2022. Due to server issues and API limitations, 121 days over the time period have partial or missing data. However, these gaps are spread relatively evenly over the time period, and so the data still provide strong coverage. Our final Twitter dataset contained 3.6 billion extracted tweets. I extracted all mentions of the 11.5k reliability-labeled websites and I extracted mentions of dredge words from this dataset.

1.2.6 OMEN Response Data

OMEN King et al. [2021] is an information environment exercise where 15-30 expert misinformation analysts attempt to characterize a simulated information environment and provide recommendations based on their findings. I include several findings from this Thesis into the synthetic data generation pipeline Morgan et al. [2025], including dredge words. I provide a survey to analysts about their engagement with dredge words during the exercise and to solicit general feedback.

1.3 Chapter Contributions

Chapter 2 examines Google's efforts to address data voids—queries for which little high-quality information exists, making search results susceptible to manipulation or low-reliability content. In 2020, Google introduced warning banners intended to alert users when search results might be unreliable or rapidly evolving. In collaboration with researchers at Stanford, this chapter presents a large-scale longitudinal audit of these banners across four measurement waves using 1.4 million queries drawn from social media sharing. We demonstrate that while low-quality warning banners were rare, inconsistent, and ultimately removed, low-quality search results themselves persisted over time. While these banners were ultimately ephemeral, few researchers have attempted the identification of data voids at any real-world scale. I construct graph neural network models designed to detect potential data void conditions

[5] coronavirus, Wuhan virus, Wuhanvirus, 2019nCoV, NCoV, NCoV2019, covid-19, covid19, covid 19

using query–domain interaction structure. These models identify data void risk at rates substantially higher than platform-issued warnings, representing the first attempt to measure data voids at meaningful scale.

Chapter 3 examines misinformation in web search from the perspective of strategic manipulation. Rather than focusing on individual misleading articles, the chapter analyzes how adversarial actors attempt to influence search visibility through **structural authority** signals within webgraphs. Using a cross-national dataset of U.S., European, Russian, and Kremlin-linked pseudo–think-tank websites, the analysis documents patterns consistent with coordinated amplification: pseudo-think tanks receive disproportionate backlink volume from low-quality domains, participate in dense internal co-amplification networks, and target highly specific conspiratorial keyphrases that exploit search data voids. Although these strategies substantially increase structural amplification, the affected domains do not consistently achieve high Google rankings, suggesting that search engines partially—but not fully—mitigate such manipulation attempts. Building on these observations, the chapter demonstrates that webgraph structure and SEO-derived features contain measurable reliability signals. Using a partially labeled webgraph of over 43,000 domains, I show a simple two-layer GNN model distinguishes reliable from unreliable sites, enabling language-agnostic domain reliability detection. Together, these results show that while the structural properties of the web can be exploited for manipulation, they also provide scalable signals for identifying unreliable domains.

Chapter 4 examines the other avenue through which adversarial actors can exploit search results: by optimizing language and keyphrases. We introduce a novel methodology for researching and auditing data voids (via "dredge words"—queries for which unreliable domains rank highly) and demonstrate the impact that they have on generative search systems, focusing on Google's AI Overviews. Using a large-scale dataset of AI Overview responses and their cited sources, the chapter demonstrates a relationship between the reliability of traditional search results and the reliability of AI-generated summaries. Queries that surface low-reliability domains in search results are substantially more likely to produce AI Overviews that cite similarly unreliable sources. The chapter also identifies recurring failure modes in AI Overviews, including the presentation of pseudoscience and medical misinformation without adequate contextualization. Through controlled experiments—including leave-one-out domain removal and comparisons between retrieval-based and parametric responses—the analysis shows that retrieval source quality plays a significant role in determining response reliability. The results highlight the sensitivity of generative search outputs to the underlying retrieval

ecosystem and identify specific high-impact domains whose removal would reduce harmful outputs.

Chapter 5 investigates the relationship between social media discourse and unreliable search outcomes. The chapter considers the paths that users could take from social media to unreliable search results through dredge words. To operationalize this connection, the chapter develops heterogeneous graph neural network pipelines that integrate two structural layers: social network relationships among users and hyperlink structures among domains. By jointly modeling these interaction networks, the system identifies patterns linking social amplification to downstream search exposure. Empirical evaluations demonstrate SoTA performance on unreliable website detection and discovery tasks. This chapter provides the first large-scale framework for understanding how users can transition from social media to unreliable search results.

In Chapter 6, I combine insights from the previous chapters and integrate them within real-world software and educational training exercises. I adapt the BEND framework, which has traditionally been used for analyzing social media information operations, to webgraph settings. The chapter introduces interpretable metrics that capture community-level maneuver patterns directly from hyperlink structures, enabling analysts to quantify tactics such as coordinated amplification and low-authority backlink generation. These metrics are validated on two webgraph information operations: the think-tank network from Chapter 3 and the Pravda network Châtelet and Lesplingart [2025]. Next, I combine many findings from my thesis to construct an LLM-based framework for generating synthetic websites at scale for the 2026 OMEN information environment training exercise. I then report survey results from participants.

Chapter 7 synthesizes the findings across the thesis to provide a unified account of the structural conditions that allow unreliable information to persist within modern search ecosystems. It highlights the methodological contributions developed throughout the thesis—including large-scale query collection pipelines, webgraph feature extraction, and graph-based discovery models—and discusses their implications for both research and platform governance. The chapter also outlines key limitations, data access constraints, temporal instability in ranking systems, and challenges associated with multilingual transfer. Together, these considerations frame a research agenda for understanding and mitigating manipulation in increasingly hybrid search environments that combine traditional ranking systems with generative AI.

Chapter 2

Limits of Google's Content Moderation

2.1 Introduction

Between 2020 and 2022, in a series of blog posts, Google rolled out three distinct content advisory banners that appear at the top of Google's search results. These banners, which can appear at the top of Search Engine Result Pages (SERPs), are a form of explicit content moderation, and ostensibly help users identify potential data voids—searches that return low-relevance or low-quality results. These banners help users identify to low-quality [Nayak, 2022], low-relevance [Tucker, 2020], and rapidly-changing data voids [Sullivan, 2021] (See Figure 2.1). Apart from Google's rollout blog posts, Google has released no information on these banners and little is known about when or why these banners appear.

2.2 Chapter Research Questions

RQ2.1 How can we find search queries that produce Google's Content Advisory Banners?

RQ2.2 What query and SERP features are associated with those banners?

RQ2.3 How consistent and stable are Google's banners?

RQ2.4 Can we improve Google's bannering approach?

The content moderation practices of large online platforms, and how they're applied to information on important topics like health and elections, are topics of widespread interest to researchers, the public, and policymakers around the world. One such practice that has been

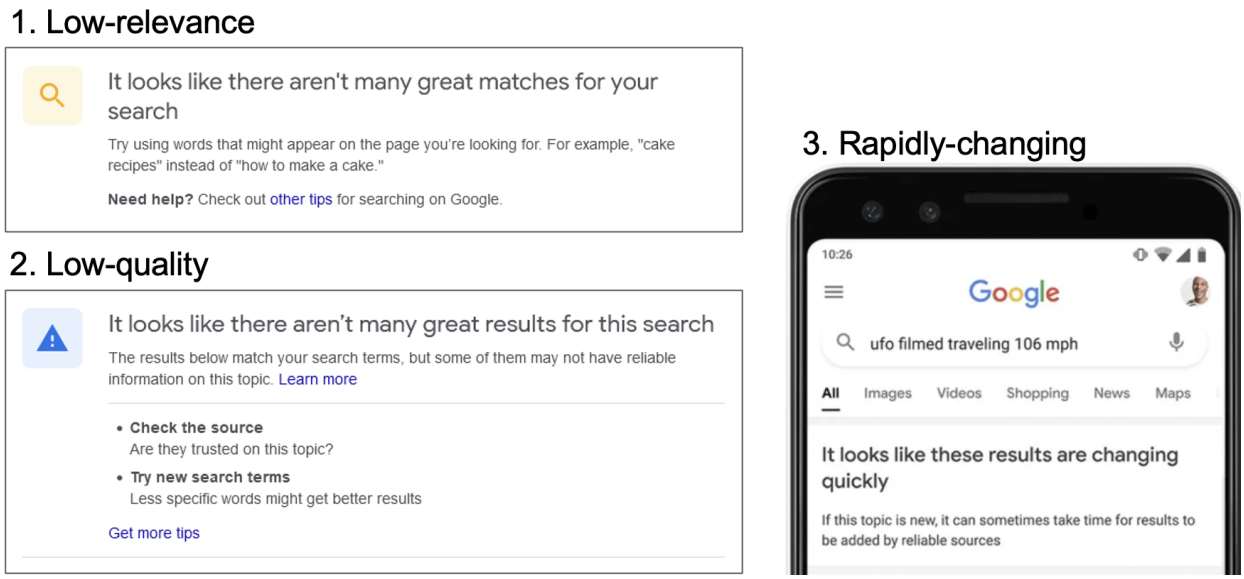


Figure 2.1: Examples of warning banners for low-relevance (1), low-quality (2), and rapidly-changing (3) data voids on Google Search. Google displays the low-quality banner (2) at the top of its results when their "systems don't have high confidence in the overall quality of the results available for the search" [Nayak, 2022], and the low-relevance banner (1) "when Google hasn't been able to find anything that matches your search particularly well" [Tucker, 2020]. Google also has a rapidly-changing banner, but these are rare due to their time-sensitive nature and not a focus of this study

widely debated in particular, is the use of "warning labels" to alert and inform users about the accuracy, context, or quality of a specific piece of content Morrow et al. [2022]. For example, social media sites have placed warning labels on posts classified as inaccurate by professional or crowd-sourced fact-checkers, which recent research finds can be an effective strategy for reducing the belief and spread of misinformation Martel and Rand [2023], Slaughter et al. [2025]. However, the policies determining the use of such warnings, especially fully automated ones, are often opaque or absent Krishnan et al. [2021], and research on how such labels are applied in practice has often been limited due to the challenges of surfacing examples to study Bradshaw et al. [2023].

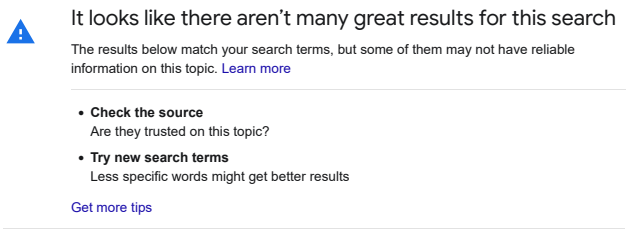
Research in this vein has often focused on social media, but search engines also play a central role in online information seeking and employ similar content moderation practices Urman et al. [2024], Juneja et al. [2024]. Analogous to the warning labels used on social media, Google places warning banners at the top of its search results to address *data voids*: when the search results available for a given query are scarce, unstable, or dominated by unreliable or irrelevant websites Golebiewski and boyd [2019]. While the prevalence of data voids is unclear,

some concerning instances have been covered in case studies, books, and mainstream news articles, including in December 2016, when Google's top-ranked search results for the query "did the holocaust happen" contained known neo-Nazi websites and narratives Mulligan and Griffin [2018]. Such cases highlight the need for further research Orben and Matias [2025], and are especially concerning because web search engines are widely trustedToff and Nielsen [2018], Shearer and Mitchell [2021], Edelman [2021], but can increase belief in misinformation Aslett et al. [2024], influence political preferences Epstein and Robertson [2015b], and be exploited as an indirect intermediary for guiding people into data voids Golebiewski and boyd [2019], Robertson et al. [2023a], Tripodi [2022].

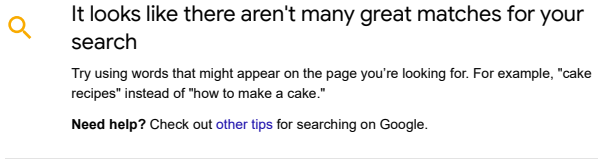
To address concerns about such issues, Google introduced three distinct banner types between 2020 and 2022, tailored for low-quality Nayak [2022], low-relevance Tucker [2020], and rapidly-changing Sullivan [2021] data voids (Figure 2.2). As with warning labels on social media, research has shown that the use of warning banners in web search can generally help users to evaluate low-quality or partisan search results Epstein et al. [2017], Ludolph et al. [2016], though small backfire effects on trust in accurate sources have been found for warnings about unknown reputations Williams-Ceci et al. [2024]. Given Google's reach, their use of these warning banners could potentially help millions of people navigate data voids on their platform. However, aside from Google's brief blog posts announcing the rollout of each banner Nayak [2022], Tucker [2020], Sullivan [2021], little is known about when or why they're deployed in practice.

In this study, we surfaced and evaluated Google's warning banners, built deep learning models to predict their presence, and used those models to examine the prevalence of data voids over time. To surface warning banners, we first collected a dataset of 4M social media posts containing *search directives*—which are defined as attempts to prompt people into conducting an online search (e.g., a post that tells viewers to conduct a search for "vaccines cause autism") Robertson et al. [2023a]. Search directives have been found to regularly occur on social media, contain queries covering a wide range of topics, and occasionally lead to data voids of low-quality results Robertson et al. [2023a], making the search queries they contain the ideal candidates for our purposes. To achieve this scale, we focused on search directives shared as URLs (e.g., `google.com/search?q=vaccines+cause+autism`), which are easier to systematically identify and extract search queries from than text-based search directives.

From the posts we collected, we extracted a diverse set of 1.4M unique search queries that were primarily in English, comparable in length to existing estimates, and overlapped with a large dataset of frequently used queries released by Bing Craswell et al. [2020]. Next,



(a) Low-quality banner, shown when Google's "systems don't have high confidence in the overall quality of the results available" Nayak [2022].



(b) Low-relevance banner, shown "when Google hasn't been able to find anything that matches your search particularly well" Tucker [2020].

Figure 2.2: Examples of Google's (a) low-quality and (b) low-relevance warning banners. Google also has a rapidly-changing banner, but these are rare due to their time-sensitive nature and not a focus of this study (see Appendix A.0.2, Figure A.4).

we collected the first Search Engine Results Page (SERP) on Google for each of our queries following best practices in algorithm auditing Metaxa et al. [2021], including the use of automated requests from a fixed location with no account login. We then extracted features from those SERPs (e.g., web domains), and merged those features with metrics validated in past work (e.g., domain quality Lin et al. [2023]). To account for changes over time, we collected SERPs for the full set of queries in four waves conducted about five months apart, in October 2023, March 2024, September 2024, and February 2025, resulting in a total of 5.8M SERPs and 90.1M search results.

Across waves, 1.1% of our search queries produced a warning banner of any type when searched on Google. Among Google's three distinct warning banners, low-relevance banners were the most common, accounting for 98–99% of the banners we observed in any wave. In contrast, and aligned with their time-sensitive nature, rapidly-changing banners were the least common, and only 2–11 queries produced one in any wave. Low-quality banners accounted for 2.1% of all banners seen in Oct 2023 and 1.5% of all banners seen in Mar 2024, but did not appear in either Sept 2024 or Feb 2025.

Although there was substantial churn in the search results returned each wave, we found only minor changes in average domain quality. Using query-paired comparisons across waves—i.e., comparing the set of search result URLs for the same query in Oct 2023 and in Mar 2023—we found that only 34.4% of the URLs returned at one time step still appeared for the same query about five months later, on average. However, paired comparisons of domain quality over time show that it only improved by an average of 0.13% between waves, raising concerns about the absence of the banners.

When Google's low-quality banners did appear, the average domain quality of the SERPs

they appeared on was 16-20% lower than those without a banner. The queries that received such banners were also more likely to contain conspiracy-related keywords identified in past work Ballatore [2015], Mahl et al. [2021], Urman et al. [2022b] than queries that did not. Despite some queries containing advanced operators that effectively guaranteed a data void when searched—e.g., "covid vaccine detox site:naturalnews.com", which uses the "site:" operator to ensure that all results come from naturalnews.com, a domain with a quality score of zero Lin et al. [2023]—low-quality banners never appeared for search queries containing such operators. The lack of low-quality banners in such cases may provide a loophole for guiding others into data voids, as Google provides little notice that the operator is filtering the results Robertson et al. [2023a], and may follow from the ambiguous attribution provided in the banner's text (Figure 2.2); if a low-quality banner had appeared for a search query containing one or more "site:" operators, then there would be no ambiguity as to which website(s) triggered it.

To identify data voids beyond Google's warning banners, we fine-tuned three deep learning models—using DistilBERT, a homogeneous Graph Neural Network (GNN_{Hom}), and a heterogeneous GNN (GNN_{Het})—to predict the presence of low-quality banners based on search queries and their corresponding search results. The GNN_{Het} model outperformed the other models on a comprehensive set of evaluations, including annotated precision@K, out-of-sample testing, and correlations between prediction confidence and average domain quality. Using this model, we classified 0.57% of SERPs across waves as low-quality data voids, which is 30 times more SERPs than Google placed a low-quality banner on in Oct 2023 and Mar 2024. Despite the absence of the low-quality banners in Sept 2024 and Feb 2025, our model continued to identify data voids at similar rates. These results are consistent with a simpler data voids definition, using an average domain quality cutoff of ≤ 0.5, which identified 0.73% of SERPs across waves as data voids. Together, these results suggest that Google's low-quality banners may have been discontinued despite data voids remaining a problem on the platform.

After initially noticing the absence of Google's low-quality banners, we sought out but were unable to find any relevant announcements. To confirm the status of the banners, while maintaining the independence of our report, we then published the results from our first three waves as a preprint Robertson et al. [2025] and shared it with a journalist who could contact Google for comment. In their response, Google confirmed that they had discontinued the low-quality banners, citing "a ranking quality improvement" that they declined to elaborate on Newton [2025]. However, our model continued to identify data voids after the banners

disappeared, and comparisons of pairwise average domain quality show an increase of only 0.21% between waves when the banners first disappeared, and a subsequent decrease of 0.14% as the banners remained off. In the domain quality scores we used, these differences are minimal, and comparable to the gaps between Infowars (0.046) and Stormfront (0.045), or Reuters (1) and the Associated Press (0.998). Aside from Google's statements regarding our preprint, no other official announcements on the discontinuation of the low-quality banners appear to have been made.

Discontinuing the low-quality banners in the months preceding the 2024 US presidential election may have been an especially impactful time to do so. Events with breaking news updates provide fertile grounds for data voids Golebiewski and boyd [2019], political campaigns have leveraged search engines in the past (e.g., "Google Ron Paul" Baker [2008]), and recent advances in Large Language Models (LLMs) have made it easier than ever to create misleading and persuasive content Feuerriegel et al. [2023], Lin et al. [2025], Hackenburg et al. [2025]. As such, our findings highlight the need for greater transparency around search engines' content moderation policies and their impact on users. Given that we focus on Google Search and use a predominantly English sample of search queries, our results may underestimate the prevalence of data voids for searches in other languages, which can receive fewer resources Borge et al. [2021], or on other search engines, including newer ones that incorporate or feature LLMs. Last, our findings highlight the importance of longitudinal research on online platforms, as their systems for moderating and curating content are moving targets that can frequently undergo rapid, unannounced, and substantive changes Bagchi et al. [2024]. As our study demonstrates, without such research, changes like the one we observed may otherwise go entirely undocumented.

2.3 Methods

To conduct this study, we first needed to gather a large and diverse set of search queries that had the potential to surface Google's warning banners. To that end, we collected a large set of social media posts containing search directives, which are defined as prompts to conduct an online search and have been show to lead to data voids (Section 2.3.1). We then used the 1.4M unique search queries extracted from that dataset as the inputs for an algorithm audit of Google Search to obtain a search results page for each query across four waves (Oct 2023, Mar 2024, Sept 2024, Feb 2025) of data collection (Section 2.3.2). The search data we collected from that process allowed us to evaluate the presence of such warning banners by

using established query (Section 2.3.3) and domain-level metrics (Section 2.3.4) in logistic regression models (Section 2.3.5). Last, we tested the consistency of the low-quality banners over time (Section 2.3.6), and developed deep learning models to identify unlabeled data voids (Section 2.3.7).

2.3.1 Search Directives

We collected search directives from social media posts (Section 2.3.1) to gather a diverse set of search queries (Section 2.3.1) for our study. By specifying a flexible linguistic strategy (prompts to conduct an online search) rather than specific content (search queries), search directives provide a useful tool for surfacing unspecified and unknown content. The search queries used in search directives have been shown to cover a diverse array of topics, ranging from music, sports, and advertising, to medical misinformation about Ivermectin, an emerging conspiracy about the COVID-19 vaccine causing people to "die suddenly," and a cryptocurrency scam Robertson et al. [2023a]. Although the potential harms of people being led into data voids like these have been well documented Golebiewski and boyd [2019], few studies have examined how people can be led into data voids Tripodi [2019b], Tripodi et al. [2023], how to computationally identify data voids Flores-Saviaga et al. [2022], or how to measure the bridge between social media and search engines more broadly Bode and Vraga [2018], Lukito [2020], Mukherjee and Jansen [2017], Riedl et al. [2018], Yarchi et al. [2021], Zuckerman [2021]. Rather than on relying on smaller sets of queries generated by researchers Kawakami et al. [2020], Kravets and Toepfl [2022], Makhortykh et al. [2020], Mejova et al. [2022], Metaxa et al. [2019], Mustafaraj et al. [2020], Perreault et al. [2024], Urman et al. [2022b], Zavadski and Toepfl [2019] or solicited from survey or interview participants Lurie and Mustafaraj [2018], Trielli and Diakopoulos [2020], van Hoof et al. [2022], or medium-sized sets of queries generated via autocomplete Haak and Schaer [2023], Robertson et al. [2019], our use of search directives allowed us to collect 1.4M unique queries without defining the topic space or a starting set of queries to expand upon.

Social Media Posts

We collected a total of 5.25M posts that contained a URL fragment (e.g. `google.com/search`) leading to one of 25 popular search engines. This collection strategy allows for flexibility in subdomains, variability in URL parameters, and enabled us to easily and accurately extract search directive queries. Following past work, we filtered out URLs that did not lead to

a page of search results, including those that did not contain a known query parameter (e.g., "&q={query}" for Google Search) and those that contained a blank query, leaving 4M search directive posts, 4.17M URLs (posts can contain multiple URLs), and 1.44M unique queries that were created by 1.82M unique accounts over a 16.5-year window (2006 to 2023). Advancing on prior work that examined the five most popular modern search engines in the US (Google, Bing, DuckDuckGo, Yahoo, and Brave), we used a list of 25 search engines to collect our dataset (Google, Bing, DuckDuckGo, Yahoo, Brave, AOL, Ask, Baidu, Dogpile, Ecosia, Exalead, Excite, Hotbot, Lycos, Metacrawler, Mojeek, Petalsearch, Qwant, Sogou, Startpage, Swisscows, Webcrawler, Yandex, You, and Youdao), including search engines that are prominent outside of the US (e.g. Yandex), were prominent in the past (e.g. AOL and Ask), or newer search engines that feature large language models (e.g. You.com).

Search Directive Queries

Of the 4M posts that contained a URL fragment and a search query—which excludes links that don't qualify as a search directive (e.g., a search engine homepage)—we obtained a diverse sample of 1.44M unique queries that varied widely in terms of both their content and structure. While not representative of what people search for in general, this sample of search queries shared on social media covers a wide range of topics (including music, sports, and politics), were produced across a 16 year span, and include event-driven bursts (e.g., around the ICC Men's T20 World Cup 2016, a biannual cricket tournament), making them ideal candidates for our discovery-oriented goal of surfacing warning banners. These queries also widely varied in terms of their length, with the average search directive query containing an average of 4.5 words (Figure A.11), which is slightly longer than estimates of query length in the US, which find that 82% of queries are 3 words or less KeywordDiscovery [2020]. The longest query in our dataset was 896 tokens long, and 234 (0.01%) queries were only one character, often an emoji. Notably, Google Search limits queries to 32 words, and that length is counted after processing by an unknown tokenizer. When a query is too long, Google adds a notice at the top of the search results which states: "... (and any subsequent words) was ignored because we limit queries to 32 words." Further details on these queries, including lengths with and without truncation and similarities to a publicly released set of queries from Bing are available in Appendix A.0.3. Details on the features we extracted from our queries and the lexicons we used to evaluate them are available in Section 2.3.3.

2.3.2 Search Engine Results Pages (SERPs)

We used open-source tools to collect our search results (Section 2.3.2), and an iterative approach to discovering and classifying Google's warning banners (Section 2.3.2). To evaluate the rate at which search directive queries produce warning banners and data voids, we used our set of 1.4M unique queries as the inputs for an approach known as the algorithm audit Sandvig et al. [2014], which typically involves collecting and examining the outputs of a black-box system based on some fixed set of inputs Bandy [2021], Metaxa et al. [2021], Mustafaraj et al. [2020], van Hoof et al. [2022]. In this case, the inputs are the search directive queries, the system is Google Search, and the outputs are the Search Engine Results Pages (SERPs) returned by Google.

Collecting and Parsing SERPs

For collecting the search results available for each query, we used WebSearcher Robertson and Wilson [2020]—an open source tool for collecting and parsing SERPs that has been used in prior algorithm audits of Google Search Mejova et al. [2022]—to conduct a search using each query in our set, store the corresponding HTML, and extract details about its corresponding search results (e.g. rank, URL, result type). We also extracted several elements other elements from the SERP, including Google's estimate for the total number of results it found for each query (across its entire index), which could also be indicative of a data void, as the search results for a query with few matches may be easier to manipulate due to the limited competition. As with most algorithm audits of web search, this SERP dataset represents only what someone searching these queries might have seen at the time of our collection. We also searched from a fixed location and do not study localization effects Kliman-Silver et al. [2015]. Details on the number of results we collected are available in Appendix A.0.3, Table A.6.

Identifying Warning Banners

We initially identified banners by checking for the exact phrasing of each warning banner type, and then built phrase-agnostic HTML parsers to extract them across the entire dataset. For low-relevance banners, the phrasing was "It looks like there aren't many/any great matches for your search" Tucker [2020]. Low-quality banners contained similarly phrased language ("It looks like there aren't many great results for this search"), swapping only "matches" with "results." In contrast with the low-relevance banners, we never observed the variation where "many" was replaced with "any" in the low-quality banners, which aligns with the ambiguity of

the banner message ("some of [these results] may not have reliable information", Figure 2.2), and how they were described in Google's blog post announcing their rollout ("This doesn't mean that no helpful information is available, or that a particular result is low-quality" Nayak [2022]). This reluctance to specify which search results are low-quality may also help explain why we never saw a low-quality banner for searches with a "site:" operator that restricted the search results to a specific web domain: doing so would remove the ambiguity of the judgment. In contrast to these warnings about content, the rapidly-changing banner stated: "It looks like the results below are changing quickly" Sullivan [2021]. Additional details and a screenshot of the rapidly-changing banner, as well as details and a screenshot of a low-relevance banner variant that only appeared in our last wave, are available in Appendix A.0.2.

2.3.3 Search Query Text Features

To evaluate query content, we used a dictionary-based approach to identify queries containing partisan and polarizing search terms or conspiracy-related search terms (Section 2.3.3), and extracted other text features, such as advanced query operators (Section 2.3.3).

Political and Conspiracy-related Lexicons

To identify queries around controversial topics that could potentially lead to data voids, we used a dictionary-based approach to tag words and phrases associated with conspiracies and politics in prior work. Specifically, we used: (1) Ballatore's (2015) Ballatore [2015] set of 96 conspiracy-related search queries, (2) Mahl et al.'s (2021) Mahl et al. [2021] set of 44 conspiracy-related hashtags, and (3) Urman et al.'s (2022) Urman et al. [2022b] set of 6 conspiracy-related search queries. We also considered Haak and Schaer's (2023) Haak and Schaer [2023] set of QAnon-related search queries and autocomplete expansions, but the terms were too broad for our purposes. For the terms from Mahl et al. (2021) Mahl et al. [2021], we added non-hashtag versions of each item (e.g. "#vaccineskill" becomes "vaccines kill") and excluded "#dew", which refers to conspiracies around directed energy weapons but produces a high false positive rate due to the popularity of Mountain Dew, a soda brand. Of the 14 conspiracy categories covered by this dictionary—including conspiracies about 9/11, chemtrails, and reptilians—we found at least one search directive query that mentioned each. We also used two existing lexicons of polarized terms—one designed to capture "polarized language" Simchon et al. [2022] and one designed to capture "partisan cues" Hu et al. [2019]—to classify search directive queries as politically related. Combined, we

used these lexicons to classify the full set of queries as related to politics (11.1%), conspiracies (0.11%), both (0.02%), or neither (88.8%).

Advanced Query Operators

Advanced query operators allow searchers to specify additional constraints on their search results. For example, the query "trump site:dailycaller.com" will search for results containing that term ("trump") only within that site ("dailycaller.com"). When considering search directives as an attempt to exert indirect online influence, the use of these operators has strategic value in guiding people to specific content via a trusted search engine: searchers less familiar with these operators may not understand that their results have been filtered, and while some search engines (e.g., DuckDuckGo) display a message to informs users that such a filter is active, Google does not Robertson et al. [2023a]. In total, 1.5% of our queries contained one of 11 advanced operators, and among those, the most common operator was "site:" (92.0%), followed by "inurl:" (2.8%), "filetype:" (1.9%), "intitle:" (1.0%), "ext:" (0.5%), "before:" (0.5%), "source:" (0.4%), "related:" (0.3%), "allintitle:" (0.3%), "after:" (0.2%), and "allinurl:" (0.1%). These queries varied widely in their content and complexity, with some containing multiple operators and others containing only one. For example, one query used the OR operator, parentheses, quotes, and 15 site operators:

```
(mask | vaccine | "death count" | "case count") fraud and evidence election
( site:amac.us | site:townhall.com | site:heritage.org | site:thegatewaypundit.com
| site:oann.com | site:scienceunderattack.com | site:conservativetribune.com
| site:thefederalist.com | site:greatamericandaily.com | site:westernjournal.com
| site:zerohedge.com | site:prageru.com | site:realclearpolitics.com
| site:mercola.com | site:naturalnews.com )
```

Query Language

As many queries are names, fragments, emojis, or are otherwise grammatically incorrect, determining the language of queries can be challenging, and some level of noise is inevitable. To get a general sense of query languages, we used the FastText library (lid.176) Joulin et al. [2016] to predict the most likely language for each query. Across all 1.4M queries, more than 1.2M were predicted to be in English. For the subset of 930K queries where the model returned a confidence of at least 0.5, 875K were predicted to be English. The second most common category in the high and low-confidence query sets was French, with 23K and 9K

queries, respectively. Many of the queries that were classified as French appear to have been classified that way because they use French words or names in an otherwise English-speaking context. For example, of the 55 queries that contained the name "De Blasio"—the former mayor of New York—FastText predicted that 24 were French, including "bill de blasio" and "bill de blasio drops groundhog video". German was the third most popular language, and similar to the French classifications, many of the queries classified as German appeared be English queries associated with American politics like "adolf hitler defund police" and "führermccarthy".

Generalizability of Query Set

Without access to proprietary search engine data, it is impossible to know exactly how many users searched each query in our dataset. However, several large-scale query datasets have been released by search engines in the past. ORCAS, released by Microsoft, contains 10.4 million queries searched by at least "k different users, for a high value of k" on Bing around January 2020 Craswell et al. [2020]. The exact value of k used is not specified. The authors also applied filters to remove potentially offensive queries, like those containing pornography and hate speech. Although the ORCAS list is somewhat sanitized by its k-anonymity and offensive content filters, it allows us to check if any of our search directive queries were widely searched on Bing during that time period.

Among our queries, 154,833 (10.8%) were present in ORCAS, and 1,972 of which were classified by our best-performing model as warranting a low-quality banner. Among those, the 20 most confident GNN predictions included queries like "facebook illuminati," "lizard people conspiracy," and "black groups that hate whites." The presence of these queries in ORCAS show that about 1 in 10 of our queries were widely searched on Bing, and that subset included queries that our models identified as data voids.

2.3.4 Web Domain Features

To evaluate the search results we collected, we extracted the second and top level domain names for each URL (e.g. https://cnn.com/politics → cnn.com) and merged them with several domain-level metrics. This includes news classifications, a domain quality metric (Section 2.3.4), and domain traffic estimates (Section 2.3.4). After merging our search result domains with these metrics, we aggregated those classifications and scores at the SERP-level as counts (e.g. of news domains) and averages (e.g. quality scores) and use those as our

primary unit of analysis. Counts and averages by wave are available in Appendix A.0.3.

News and Quality

To classify domains as news, we used a set of 7,582 news classifications aggregated in past work Robertson et al. [2023b]. Overall, among the 70.5M search results with a domain, 10.4% were news domains. Similarly, to evaluate domain quality, we used a validated set of scores based on a compendium of similar ratings from expert sources Lin et al. [2023]. These scores range from 0 to 1, with higher scores indicating higher quality, and cover 11,519 unique domains (we drop one duplicate that appears with and without a "www." prefix).

When calculating average domain quality at the SERP-level, we excluded the scores of three popular online platforms because they had quality scores that were hard to interpret and frequently appeared, reducing the utility of the averages. Those domains were: `youtube.com`, which has a quality score of 0.375 and accounted for 12.9% of search result domains across all waves, `facebook.com`, which has a quality score of 0.407 and accounted for 2.9% of domains, and `google.com`, which has a quality score of 0.668 and accounted for 0.25% of domains. Across all waves, we were able to match 19.6% of domains to a quality score, and the average domain quality was 0.787 (SD = 0.091), which is comparable to the score for `tabletmag.com` (0.781). As with all domain-level measures, these scores are coarse-grained and do not account for instances, for example, where unreliable domains publish accurate webpages, or vice versa Green et al. [2025].

Search Engine Optimization (SEO) Metrics

Search Engine Optimization (SEO) is a billion dollar industry aimed at improving websites' search rankings by tracking and estimating features such as backlink counts and web traffic. To aid in our investigation, we obtained these SEO features from Ahrefs (`ahrefs.com`) for 9,125 unique domains, covering 72.3% of our search results with a domain. Recent work using data from Ahrefs suggests that its traffic estimates are reliable, and that some of its features are predictive of misinformation Carragher et al. [2024, 2025], **?**. We provide additional details on the SEO features we used, including their validity and use in past work, in Appendix A.0.3.

2.3.5 Logistic Regressions

We used Firth's logistic regression Firth [1993], which uses a penalized likelihood estimator and helps address issues with small sample sizes or separation issues, to examine the factors associated with the presence of Google's low-relevance and low-quality banners across waves. The presence of either banner (low-quality or low-relevance) was our dependent variable, and our independent variables included factors related to both the search query (Section 2.3.3) and the SERP it produced (Section 2.3.4). The features related to the text of the search query were: word count (truncated), presence of political keywords, and presence of conspiracy keywords. The features related to the SERP were: average domain quality, estimated total results (log10), average domain traffic (log10), and news domain count. We used HC1 robust standard errors and trained separate models for each banner type and wave, resulting in a total of six models. Our models for predicting low-relevance banners had pseudo-R^2 values of 0.38, 0.18, 0.32, and 0.37 for waves 1, 2, 3, and 4, respectively. In contrast, and likely due to the smaller sample size, our models for predicting low-quality banners had smaller pseudo-R^2 values of 0.18 and 0.15 for waves 1 and 2, respectively. Complete model results including model fit summaries (Table A.7) and detailed coefficient tables (Tables A.8 and A.9) are provided in Appendix A.0.4.

2.3.6 Measuring Low-quality Banner Consistency

To better understand the consistency of Google's low-quality banners, we collected SERPs for the 301 queries that produced a low-quality banner in wave 1 approximately every four hours from June 7, 2024 to June 24, 2024. For each search, we attempted to collect the first 100 results by modifying a URL parameter in our requests, but many queries consistently had fewer than 100 results. Due to server related errors, we had two gaps in this collection, with the first lasting for approximately 45 hours between June 11 and June 13, and the second lasting about 15 hours between June 17 and 18. Additional details on this data collection are available in Section 2.4.6.

Using the Jaccard similarity index, calculated as the intersection of two sets divided by their union, we found that the minimum similarity in the set of queries receiving a low-quality banner between any two time periods was 0.79, and the average similarity (excluding the diagonal) was 0.88 (Extended Data, Figure A.3a). That is, 4.5 hours after observing a set of queries return low-quality banners, only 79% of those queries still produced a low quality banner, on average. There was only one instance in which two time steps shared the exact

same set of queries.

To compare the search result URLs across timesteps in this dataset, we used rank-biased overlap (RBO) Webber et al. [2010], a measure designed for comparing ranked indeterminate lists. We found that the queries that consistently received banners also returned relatively consistent search results (Extended Data, Figure A.3b). We find similar results in a pilot version of this dataset that we collected in March 2024 over 34 time steps (with about 1.5 hours between each) without any gaps. Additional details on our March 2024 dataset and results, as well as our consistency and similarity measures, are available in Section 2.4.6.

2.3.7 GNN Model Development

To construct models for predicting warning banner presence, we first preprocessed our data (Section 2.3.7) and extracted training and test sets (Section 2.3.7). We then used that data to train a DistilBERT model on the text of the search queries alone (Section 2.3.7), and two Graph Neural Network (GNN) models that incorporated both query text and features related to the search results (Section 2.3.7). Last, we validated the performance of our models using a range of both standard and custom metrics (Section 2.4.4).

Preprocessing

Prior to modeling, we performed two preprocessing operations. First, we calculated embeddings for all queries using a multilingual Sentence-BERT model Reimers and Gurevych [2019]. Second, most results on a SERP include a title—the blue text on Google's SERPs that one clicks to reach a webpage. For each domain that appeared at least twice in the dataset, we create a single string that contains the domain name with a colon followed by titles sampled from the domain with replacement. The intention with this step is to create a feature with some, albeit shallow, notion of the topics covered by the domain.

Train and Test Datasets

After quantitative and qualitative evaluations, we chose a 1:3 positive-to-negative sample to train our classifiers. In predictions on unlabeled data, we observed that the DistilBERT model trained on a 1:1 positive-negative sample seemed to be over-relying on the presence of quotation marks; DistilBERT's most 100 confident banner candidate predictions contained quotation marks and some contained names of movies or books like "the craft" and "scout mindset." We therefore elected to use a 1:3 positive-to-negative sample in order to provide a more

diverse set of negative samples. We use this imbalanced training set for all models; training imbalance has been shown to improve performance in some imbalanced settings Hasanin et al. [2019]. The sample consists of the 301 queries that produced a low-quality banner in wave 1, and an additional 903 queries that were randomly sampled from the queries that did not receive such a banner. We applied a stratified 80/10/10 split to the final set of 1,204 labeled and unlabeled queries.

Query-Only DistilBERT Model

We included a model that uses only the text of our search queries as a baseline that we could use to compare against more complex models. Formally, let $p(B|T)$ be the probability of a low-quality banner B appearing for a given query text T. This assumes that the probability of a banner's presence depends only on semantic cues present in the query text T. Specifically, we fine-tune a DistilBERT model Sanh et al. [2019b] (using `distilbert-base-uncased` from Hugging Face) to predict $p(B|T)$ for each query. The model is trained for two epochs using an Adam optimizer with a learning rate of 2e$-$5 and a linear warm-up scheduler.

Query-SERP GNN Models

Next, we sought a model that could incorporate the assumption that two different queries with highly similar SERPs should likely have the same banner status. We therefore elected to use Graph Neural Networks (GNNs), as this allows us to propagate domain-level context into query representations (e.g., as done with different features in past work Williams et al. [2025]). To do so, we constructed two simple homogeneous (GNN_{Hom}) and heterogeneous (GNN_{Het}) models which incorporate the assumption that the presence of a banner $p(B)$ depends on both the query text (T) and the content associated with returned domains S. These models aim to predict the conditional probability $p(B|S,T)$, integrating information from both sources. We represent the problem as a bipartite query-to-domain graph, with one node set corresponding to queries and the other to domains. Our approach is also similar to the vaccine-related query-click graphs used in prior work Chang et al. [2024], but we elect to leverage DistilBERT for query embedding as our query topics span a broader range of domains.

Given a set of queries $Q = \{q_1, q_2, \ldots, q_n\}$ and a set of returned SERP domains $D = \{d_1, d_2, \ldots, d_n\}$ we construct a homogeneous bipartite graph $\mathcal{G}_{Hom} = (V, E)$ where domains and queries are treated as the same node types. Both node types have text-based node

features, and labels Y are a binary variable indicating the presence of a banner on Q. We additionally construct a heterogeneous graph, $\mathcal{G}_{Het} = (V, E)$ where V and E are associated with a node type mapping function $\Psi : V \rightarrow A$ and an edge type mapping function $\Phi : E \rightarrow \phi$. In our setting, the set of node types are $A = \{Q, D\}$ and the set of edge types are $\Phi = \{domain - to - query, query - to - domain\}$.

To incorporate the assumption that ranking changes in $top - k$ SERP results (with URLs held constant) should not alter banner presence, we do not weight edges in our networks. To allow information to propagate between queries, we exclude "pendulum" domains—those that only appeared once in our SERP data. For each of those domains we sampled at most 10 "titles"—the blue text that appears on Google search results (generally the title of the article or webpage)—embedded the titles with DistilBERT, and took the simple mean. Although this is a relatively simple and coarse-grained approach that excludes many relevant domain-level signals, we demonstrate its effectiveness and leave the incorporation of more nuanced domain-level features to future work.

Both models consist of a GraphSage convolution with a dropout of 0.5 and ReLU activation, followed by a second GraphSage convolution and a final log-softmax activation function Hamilton et al. [2017]. This results in a model with 526k parameters. GNN_{Het} uses a heterogeneous GraphSage convolution Fey and Lenssen [2019b]. In this setting, where there are only two node types with bi-directional ties, this doubles the number of model parameters to 1.05M. We use an Adam optimizer with $\eta = 1\text{e}{-3}$, and a weight decay of $5\text{e}{-4}$, Cross Entropy Loss, and a Cosine Annealing Learning Rate Scheduler with $\eta_{min} = 2\text{e}{-5}$ Loshchilov and Hutter [2016].

2.4 Results

2.4.1 Evaluating Warning Banner Prevalence and Characteristics

Using our dataset of 1.4M unique search queries shared on social media (Methods 2.3.1), we collected the corresponding Search Engine Results Page (SERP) for each query in four primary waves (Oct 2023, Mar 2024, Sept 2024, and Feb 2025). To ensure comparability across waves, we used the exact same set of queries, conducted them from a fixed location in the Northeastern US, and used a fixed user-agent (Methods 2.3.2). To evaluate the factors associated with the presence of Google's low-relevance and low-quality banners, we used a set of features related to each search query (Methods 2.3.3) and the corresponding SERP it

returned (Methods 2.3.4). This includes the presence of conspiracy-related terms in the search query Mahl et al. [2021], Urman et al. [2022b], Ballatore [2015], and the average domain quality score of the information and news URLs present on the SERP Lin et al. [2023].

Google displayed one of its three distinct banner types (Figure 2.2) in the search results for about 1.1% of our search queries, on average across waves (Figure A.8a). Low-relevance banners were the most common type, accounting for 97.9% or more of the banners seen in each wave. Low-quality banners were the next most prevalent, accounting for 2.1% of all banners in Oct 2023 and 1.5% in Mar 2024, but never appeared in Sept 2024 or Feb 2025 (Figure A.8a). Rapidly-changing banners appeared at least once in every wave, but were rarely deployed, appearing only twice in Oct 2023 and Mar 2024, 10 times in Sept 2024, and 11 times in Feb 2025. Given how rare the rapidly-changing banners were, which was expected due to their time-sensitive nature, our main analysis focuses on Google's low-relevance and low-quality banners. Additional details on the banner types and variants we observed are available in Appendix A.0.2.

On average, SERPs with low-quality banners had lower average domain quality scores than SERPs without a banner or SERPs with another banner type (Figure A.8a). In Oct 2023, SERPs with low-quality banners had an average domain quality of 0.582 while SERPs with no banner had an average domain quality score of 0.783. We found similar differences in average domain quality for SERPs with a low-quality banner (0.625) and those without (0.788) in Mar 2024. Examining the cumulative proportion of low-quality banners by their average domain quality, we found that 34.4% of the SERPs with a low-quality banner had an average domain quality score ≤ 0.5, but less than 3% of the SERPs with no banner or a different banner type had scores in the same range (Figure A.8b).

We find further support for a negative association between average domain quality and the presence of low-quality banners using logistic regressions (Methods 2.3.5) that included a range of search query and SERP features as independent variables (Appendix, Figure A.12). While the presence of conspiracy-related keywords in the search query had the strongest positive association with the presence of low-quality banners, the strongest positive association for low-relevance banners was with longer search queries (by truncated word count). For either banner type, Google's estimated total results for the search (their estimate of the total number of search results that exist for a given query) and the number of news domains that appeared on the SERP (using a compendium of existing lists) both had strong negative associations with banner presence across almost all waves. That is, both banner types were more likely to appear when fewer search results were found for the query and when fewer

news domains appeared on the SERP.

Among the 1.5% of search queries in our dataset that contained an advanced operator, none received a warning banner in any wave. This absence occurred despite some of those queries returning SERPs with very low average domain quality scores by design. For example, the "site:" operator, which accounts for 92% of queries containing any operator, can be used to restrict one's search results to a specific website (e.g., "site:cnn.com"), such that the results only consist of matches for that query within that website (Methods 2.3.3). Although queries with advanced operators never returned a warning banner, 1.9% returned SERPs with an average domain quality ≤ 0.5 because of the advanced operator(s) they contained. For example, the queries "ginko site:naturalnews.com", 'site:stormfront.org "sleepy eyes"', and '"deep state" site:infowars.com', each returned search results exclusively from those domains, all of which have quality scores near zero (≤ 0.05), but never produced a low-quality banner. In contrast, we found a query in our dataset that invoked the same domain name as our first example without using a "site:" operator ("naturalnews "the coming delta lockdown is designed to invoke nationwide protests""), and returned SERPs with low average domain quality scores (≤ 0.12) across all waves, but only returned a low-quality banner in Oct 2023 and Mar 2024.

2.4.2 Measuring Warning Banner Consistency

To better understand the placement, consistency, and absence of Google's low-quality warning banners, we examined both changes in the search results that each query returned across waves, and changes in the set of queries that received a low-quality banner. Using paired similarity comparisons of the URLs that each query returned across consecutive waves, only 34.4% (SD=18.7%) of the URLs that were returned for a given query were still returned when we searched the same query in the next wave, on average (Extended Data, Table A.2). This rate of churn varied by banner status, with queries that did not produce a banner generally having higher similarity in their search results between waves (Figure A.7a).

Despite this high rate of churn, the increase in average domain quality—again using paired comparisons across consecutive waves—averaged across all wave transitions was relatively small (0.13%, SD = 5.7%). Paired average domain quality increased by 0.30% ($t = 49.4, P < 0.001$) between Oct 2023 and Mar 2024, increased by 0.21% ($t = 34.1, P < 0.001$) between Mar 2024 and Sept 2024, and decreased by 0.14% ($t = -22.4, P < 0.001$) between Sept 2024 and Feb 2025 (Extended Data, Table A.3). These minimal changes in paired average domain quality suggest that the absence of the low-quality banners in the final two waves was not

due to substantive increases in search result domain quality. However, these averages may obscure larger changes among specific subsets of queries, such as those with a history of producing a low-quality banner.

Across our first two waves, when the low-quality banners still appeared, we found substantial changes in the subset of queries that received such a banner. Among the set of unique queries that received a low-quality banner in Oct 2023 (N=301), over half (N=154) no longer did in Mar 2024, but 74 queries that had not previously received such a banner then did. Across all waves, a total of 375 unique queries produced at least one low-quality banner, the average domain quality for this subset followed a similar pattern to the full set of queries across waves, first increasing, then decreasing in the last wave (Figure A.7a). Despite not receiving a low-quality banner after Mar 2024, this subset of queries continued to return SERPs with a low average domain quality score (Figure A.7b). Among this subset of queries were examples like ""underground war" qanon blessed2teach", which consistently returned SERPs that had an average domain quality score ≤ 0.163 and contained results promoting narratives related to QAnon conspiracies.

Other examples we found were related to health, including the query "former pfizer vp: 'adverse impacts on conception and ability to sustain a pregnancy were forseeable"', which returned SERPs with an average domain quality of 0.49 across waves but only received a low-quality banner in Oct 2023. Although few of the search results for this query had a domain quality score, the titles present in their URLs generally promoted conspiracy theories around the side effects of the COVID vaccine. Similarly, the query "mRNA prions" received a low-quality banner only in Oct 2023, but returned SERPs that had an average domain quality ≤ 0.42 across waves and search results that contained websites promoting the false claim that COVID vaccines contain "manipulative nanoparticles" to induce "magnetism." Additional details and examples are available in Section 2.4.8.

Prior to the disappearance of the low-quality banners, we collected a supplementary dataset to examine their consistency over shorter time intervals. Specifically, using the subset of 301 queries that received a low-quality banner in Oct 2023, we conducted searches for each query on a more rapid data collection schedule—consisting of 73 time steps spaced about 4.5 hours apart—that resulted in 22K SERPs containing 1.05M search results (Methods 2.3.6). Using these data, we found that low-quality banner presence was inconsistent over short time spans and more consistent for queries with stable search results (Extended Data, Figure A.3). We also found that the presence of a low-quality banner could not be explained by the presence or ranking of specific URLs in the search results (Figure 2.4). To further investigate

these results, we tried to collect an additional supplementary wave in August 2024, this time including minor perturbations to the queries, but the sudden absence of the low-quality banners made that impossible. Additional details on this supplementary dataset and analysis are available in Section 2.4.6.

2.4.3 Predicting Warning Banner Presence

To aid in identifying data voids, we built and tested three models to predict the presence of Google's low-quality warning banners. This includes a fine-tuned DistilBERT model trained only on query text (as a baseline), as well as a homogeneous (GNN_{Hom}) and a heterogeneous (GNN_{Het}) Graph Neural Network, both of which were trained on a bipartite query to domain graph. All models were trained to predict low-quality banner presence using data from our first wave (Oct 2023). To reduce overfitting risks due to the limited number of low-quality banners, we repeated training and evaluation ten times with varying negative samples. We provide additional details on model construction in Methods 2.3.7.

Across all evaluation metrics—accuracy, F1, precision, and recall—both GNNs outperformed the DistilBERT baseline, with the GNN_{Het} model achieving the strongest overall performance (Table 2.1). To help establish the utility and validity of our GNN models Inoue and Kilian [2005], we also evaluated their out-of-sample performance using the set of queries that gained a low-quality banner between October 2023 and March 2024 (n=74). Both GNN models identified more of these new low-quality banners among their most confident predictions than DistilBERT, with the GNN_{Hom} model performing best at smaller K thresholds and the GNN_{Het} model at larger ones (Table 2.1). The GNN_{Het} model's top predictions also performed better on annotated precision@K (Table 2.2) Last, we also evaluated the performance of our models by examining their relationship between the confidence of their predictions and the domain quality scores of the corresponding SERPs. We found that the GNN_{Het} model again outperformed the other models in identifying low-quality search results (Extended Data, Figure A.2), with its 500 most confident predictions having lower average domain quality scores than the same top predictions from either the GNN_{Hom} or DistilBERT models, despite our exclusion of that metric while training the models. Additional details on these model evaluations are available in Section 2.4.7.

Table 2.1: GNNs perform best on hold-out data and out-of-sample evaluation. Test set evaluation (left) shows the mean and std. deviation over 10 runs on the hold-out test-set, and the out-of-sample evaluation (right) shows the number of queries (out of the 74 that received a low-quality banner only in Mar 2024) in each model's K most confident predictions. Additional details on the construction of our models are availble in Methods 2.3.7, and additional details on our model evaluations are available in Section 2.4.7. While the GNN_{Hom} model outperformed the GNN_{Het} model in the out-of-sample evaluation on the top 500 and 1k predictions, the GNN_{Het} outperformed it on the top 5k and 10k predictions (Table 2.3).

	Test Set Evaluation				Out-of-Sample Evaluation		
Model	Accuracy	F1	Recall	Precision	Top 100	Top 500	Top 1k
DistilBERT	0.87 ± .01	0.68 ± .03	0.57 ± .04	0.84 ± .07	2	9	12
GNN_{Hom}	0.90 ± .00	0.87 ± .01	0.72 ± .02	0.88 ± .00	5	**13**	**16**
GNN_{Het}	**0.93 ± .01**	**0.90 ± .01**	**0.78 ± .04**	**0.92** ± .00	**6**	11	15

2.4.4 Proactively Identifying Data Voids

After developing and validating our models for predicting low-quality banner presence, we next used them to identify data voids beyond those labeled with a warning banner. To examine these in context, we consider three definitions for a low-quality data void. First, we can rely on Google's classification, and consider a data void to be present when a query returns a low-quality banner. Second, we can consider a data void to be present when a query returns a SERP with a low average domain quality score (≤ 0.5). Third, we can consider a data void to be present when a query returns a SERP that our GNN_{Het} model predicts a low-quality banner for with a high confidence threshold (≥ 0.90).

If we rely on the first definition of a data void—the SERP received a low-quality warning banner—then only 0.021% of our queries in Oct 2023 and 0.015% of our queries in Mar 2024 produced data voids, and data voids ceased to exist in Sept 2024 and Feb 2025 (Figure A.6). If instead, we use our second definition, where the average domain quality of a SERP is ≤ 0.5, then the prevalence of data voids among our queries rises to 0.83% in Oct 2023, 0.72% in Mar 2024, 0.65% in Sept 2024, and 0.73% in Feb 2025. While these numbers remain relatively small, they represent a nearly 40× increase in SERPs classified as data voids relative to the number of SERPs that Google applied a low-quality warning banner to in Oct 2023, and a 48× increase for Mar 2024, suggesting these banners were underused when they were still being applied. More importantly, this definition continues to identify data voids at

comparable rates in Sept 2024 and Feb 2025, despite the disappearance of the low-quality banners.

If we use our third definition—where we define a data void using the predictions from our validated GNN_{Het} model with a high confidence threshold (≥ 0.90)—we find results more comparable to those found with our second definition than our first. Applied across our dataset, this definition suggests that 0.93% of our queries in Oct 2023, 0.39% in Mar 2024, 0.50% in Sept 2024, and 0.46% in Feb 2025 produced data voids. Similar to our second definition, these numbers are relatively small, but represent a 44× increase in SERPs classified as data voids relative to the number of SERPs that Google applied a low-quality warning banner to in Oct 2023, and a 26× increase for Mar 2024. As with our second definition, our GNN_{Het} model continued to identify data voids at comparable rates in Sept 2024 and Feb 2025. Combined with our other findings, these results suggest that the low-quality banners were underutilized prior to their disappearance, and despite being discontinued due to unspecified ranking improvements Newton [2025], the subset of queries that produced them continued to return low-quality search results in Sept 2024 and Feb 2025.

==========

Model Validation

We evaluated our models using standard accuracy, F1, precision, and recall metrics (Table 2.1). To calculate these statistics, we trained and evaluated each model ten times, each time drawing a different negative sample for the non-bannered class to mitigate the possibility of overfitting to a single negative sample. We also evaluated these models by examining the average domain quality of SERPs across their most confident predictions (Extended Data Figure A.2) and evaluating annotated precision@K among their most confident predictions (Table 2.2). When running inference with the GNN models, we only included domains which appeared at least twice across all waves.

While the DistilBERT classifier yielded high accuracy on a small, balanced "banner" vs. "non-banner" query subset (Table 2.2), we also found that it did not learn generalizable patterns from the query text data alone. When we ran inference over the 1.4M unlabeled queries using the fine-tuned DistilBERT model, we found its 100 most confident banner predictions all contained the presence of quotation marks, despite many of those queries not returning unreliable content (e.g., *"phishing attack"*, *"secretary of state"*, and *"data breach"*), highlighting issues with using query text alone. In contrast, our GNN approach allowed us to condition the probability of a banner on both the text of the query and its

returned domains. In Appendix A.0.1, we provide details regarding a subset of related queries that produced a low-quality banner in wave 1, and frequently appeared in the models most confident predictions.

Additional Analyses

2.4.5 SERP Features

Google Search's result size estimates follow a mixed distribution consisting of several heavy-tail distributions (Figure A.9). While this distribution may indicate an underlying binning by Google, these estimates are generated through a non-public process, can vary based on factors like the data center used, and can vary in counter-intuitive ways. For example, longer and more specific queries ("cars -used") can produce larger estimates than more generic shorter queries ("cars") because longer queries may trigger a deeper search that surfaces a larger and more accurate estimate Sullivan [2010]. Google returned an estimate of 0 results for 1% of queries in wave 1 (0.87% in wave 2; 1.3% in waves 3 and 4), but about 50% of those searches still returned at least one result across waves.

2.4.6 Low-Quality Banner Consistency

In this section we provide additional details on our rapid data collection waves (2.4.6), the URL similarity metrics we used and their corresponding results (2.4.6), and the formal URL dependencies we tested along with their corresponding results (2.4.6).

Rapid Data Collection Waves

Prior to the discontinuation of the low-quality banners, we conducted two waves of rapid data collection in March and June 2024. In both waves, we used the 301 search queries that returned a low-quality banner in wave 1 to conduct repeated searches over short intervals and better understand the features associated with low-quality banner placement. In the March 2024 wave, we collected SERPs for this query subset across 34 time steps, spaced about an hour-and-a-half apart, between March 10 and March 12, 2024. In the June 2024 wave (which we also report on in the main text and Methods 2.3.6), we conducted the same query set across 73 time steps that were spaced about four hours apart. Among these searches, three queries in March and five queries in June did not return any search results at any time step, and were therefore dropped from the analysis, leaving us with 298 queries in Match and 296 queries in June. Due to server related errors, we had two gaps in the June 2024 wave, with the first lasting for about 45 hours between June 11 and June 13, and the second lasting about 15 hours between June 17 and 18.

Measuring Warning Banner Stability

We assessed the stability of Google's low-quality banner placement in several ways, including the similarity of queries that received a low-quality banner over time, the similarity of the URLs in the SERPs returned for those queries over time, and the presence of specific URLs or URL pairs in SERPs with and without low-quality banners. Here we provide details on both our stability metrics and results.

Metrics To measure query similarity, we extracted the set of queries that returned a low-quality banner at each time step t, and then calculated the pairwise Jaccard similarity between those query sets for each time step pair within a collection. This measure quantifies the relative stability of the set of queries that received a low-quality banner over time. To measure URL similarity, we used Ranked-Biased Overlap (RBO) Webber et al. [2010] to compare the similarity of the search result URLs that our query subset returned at each time step.

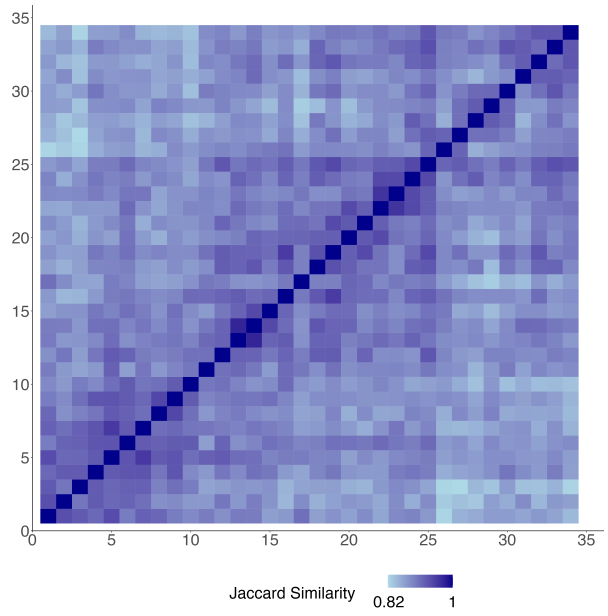
The RBO metric was designed to compare ranked indeterminate lists (i.e. search results) by calculating the agreement between two ranked lists S and U at every level of depth from $1 : D$ and then taking the average. This can be weighted by a user-chosen probability p, and we elected to take the average overlap, which corresponds to RBO with $p = 1$. RBO with $p = 1$ can be written as:

$$RBO(S, T, p = 1) = \frac{1}{D} \sum_{d=1}^{D} \frac{\mid S_{1:d} \cap U_{1:d} \mid}{d} \tag{2.1}$$

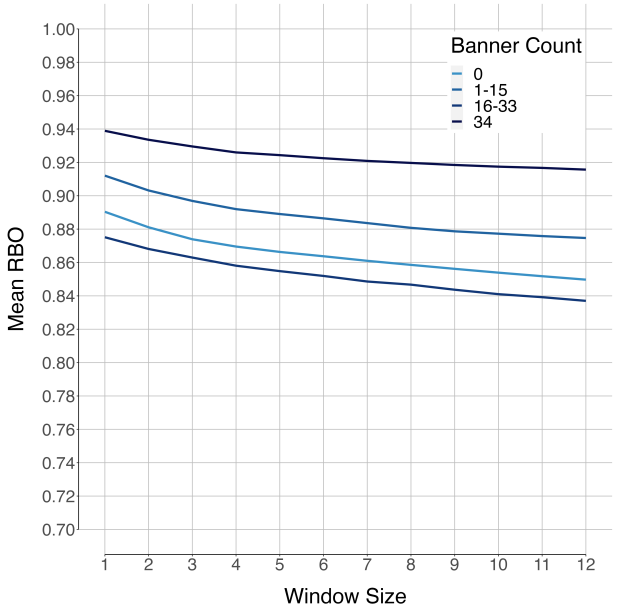
As our interest is in the relative stability of the URLs across all queries, we constructed a windowed metric (RBO_k) to quantify URL similarity across a range of consecutive time steps. Let X_i be the pairwise RBO matrix for URLs returned in the SERPs of query i over T timesteps, and K be window size. To measure the stability of a single query's SERP URLs over K consecutive time steps, we define query-level windowed RBO similarity as:

$$\bar{x}_{i,K} = \frac{1}{2KT} \sum_{t=0}^{T} \sum_{k=1}^{K} (\mathbb{1}_{[t-k \geq 0]} x_{t-k} + \mathbb{1}_{[t+k \leq T]} x_{t+k}) \tag{2.2}$$

where $\mathbb{1}$ is the indicator function. If we set $k = 1$, this would correspond to the average RBO of URLs returned at time t with URLs returned $t - 1$ and $t + 1$ over all time steps. If this number was close to 1, that would indicate a high similarity—both in the set of URLs returned and their ranking—between consecutive SERPs for a given query. Conversely, if



(a) Heatmap cells show the pairwise Jaccard similarity of the set of queries that returned a low-quality banner over 34 time-steps. In no two time steps did the exact same set of queries receive a low-quality banner.



(b) RBO_k (y-axis) over 12 window sizes (x-axis) by the mean RBO over all queries in the 34 time-steps. Queries that always returned a banner had the most stable search results over all window sizes.

Figure 2.3: In March 2024, we repeatedly collected SERPs for queries that produced a low-quality banner in wave 1 (once every 1.5 hours for 34 time steps). The Jaccard similarity plot shows that the identified queries often did not return banners, and banner status changed substantially, even over short intervals. On the right, we see that the stability of SERPs may have some relationship with banner stability.

this number was 0, this would suggest high volatility, and would mean that a query never returns any of the same URLs in consecutive time-steps. Last, to measure the average SERP stability across all queries and K consecutive time steps, we define RBO_k as:

$$RBO_k = \frac{1}{N} \sum_{i=1}^{N} \bar{x}_{i,K} \tag{2.3}$$

Results On average, 4% of queries that had a low-quality banner in one time step did not have that banner type at the next time step within the March wave, which was slightly higher than the 3.2% change we found within the June wave. The minimum Jaccard similarity for the set of queries that returned a low-quality banner in the March wave was 0.82, and the mean was 0.89 (Figure 2.3a); slightly higher than in the June wave (0.79 min and 0.88 mean; Figure A.3a). In the March dataset, there were no instances of two timesteps returning the

same set of queries with low-quality banners, while in the June dataset, this occurred once.

In both datasets, we observed a strongly bimodal distribution for low-quality banner presence: most queries either always received a low-quality banner or never did. For the June wave, 166 queries never received a low-quality banner, 40 always did, and 90 received them intermittently (between 1 and 72 times). In the March wave, 115 queries never received a low-quality banner, 116 always received one, and 67 queries received them intermittently (between 1 and 33 times). This bimodality makes it unlikely that the observed inconsistencies in Google's banner placement result from stochastic variation, A/B testing, or server latency, all of which would be expected to produce a more uniform distribution of banner instability across time steps.

Using our RBO_k metric, we found that the queries that consistently returned a low-quality banner also produced the most stable search results over time. In both the June wave (Extended Data, Figure A.3b) and the March wave (Figure 2.3b), the most stable results were from queries that always returned a banner. However, while the least stable results were from queries that never produced any banners in June, in March the least stable results were from queries that produced 16-33 banners. RBO_k also monotonically increased with window size, underscoring the continual churn in Google's search results. These results suggest that the composition of the search results—rather than the query text itself—plays a larger role in determining whether Google applied a low-quality banner.

Measuring URL Dependencies

In addition to examining the stability of Google's low-quality banner system, we also conducted several tests to identify simple rules that could potentially explain low-quality banner variance across queries. Here we provide additional details on our approach to measuring these rules as URL and ranking dependencies, and the corresponding results.

Metrics For all queries, we attempted to determine whether or not there is 1) a single URL in all SERPs with low-quality banners but never in SERPs without them, 2) a pair of URLs that appears in all SERPs with low-quality banners but never in SERPs without them, and 3) a pair of URLs conditioned on a rank cut-off that appear in all SERPs with low-quality banners but never in SERPs without them, e.g., "If u_i always appears in the top 5 search results and u_j always appears in at a position below 5, is there always a banner?"

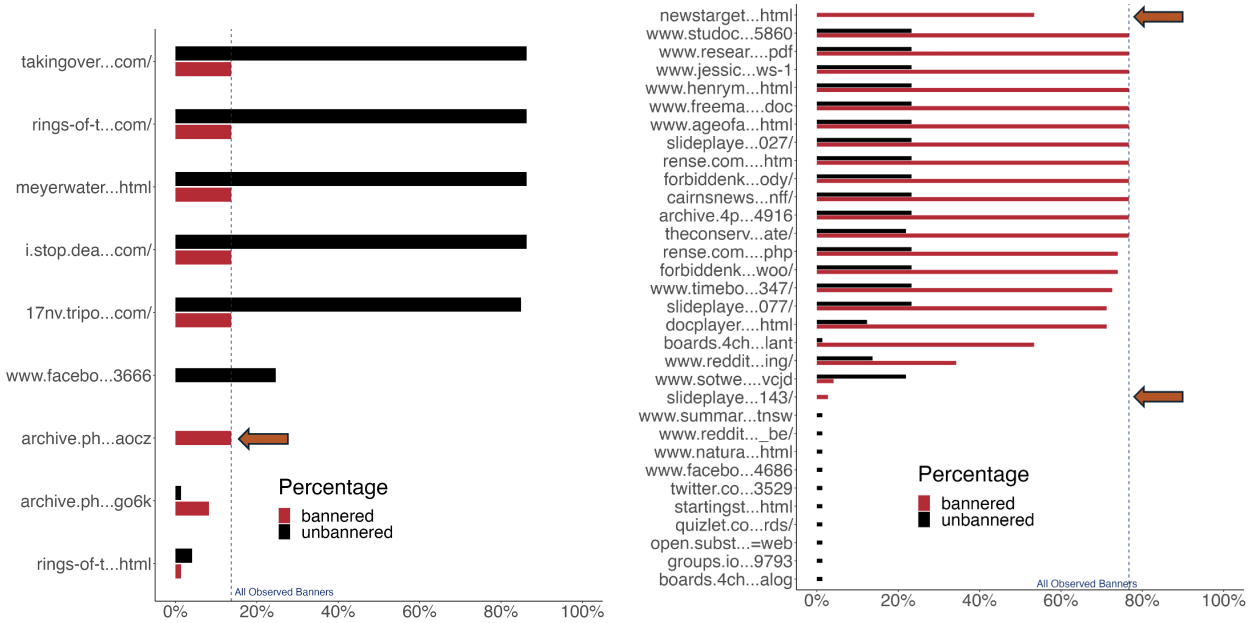
Formally, let $Q = \{q_0, q_1, \ldots, q_D\}$ be the set of 90 queries that returned between 1 and 72 banners, let $S = \{S_1, S_2, \ldots, S_N\}$ be the set of SERPs that returned a low-quality banner

for a query q_j where S_i contains a ranked list of URLs $\{u_1, u_2, \ldots, u_D\}$, and let R be the set of SERPs that did not return a low-quality banner for query q_j. For SERPs that had a low-quality banner, let $S_{:c*} = \{S_{1,:c*}, \ldots, S_{N,:c*}\}$ contain the URLs that appeared below rank c (e.g., $c = 3$ would correspond to the first two URLs that appear on a SERP) concatenated with an arbitrary string 'X', i.e. $S_{i,:c*} = \{\text{CONCAT}(u_1, \text{'X'}) \ldots, \text{CONCAT}(u_{c-1}, \text{'X'})\}$. Conversely, let $S_{c:} = S_{1,c:}, \ldots S_{N,c:}$ where $S_{i,c:} = \{u_c, u_{c+1}, \ldots, u_D\}$ without concatenation. We define $R_{:c*}$ and $R_{c:}$ equivalently over SERPs without a low-quality banner. Finally let $S^*_{i,c} = (u_j \in S_{i,:c*}, u_k \in S_{i,:c*}) \cup (u_j \in S_{i,:c*}, u_k \in S_{i,c:}) \cup (u_j \in S_{i,c:}, u_k \in S_{i,c:})$. Again, we define $R^*_{i,c}$ equivalently using $R_{:c*}$ and $R_{c:}$. Using this formulation, we can formalize our three questions as:

1. : $\forall q \in Q$ does there exist a u such that $\exists u \in \left(\cap_{i=1}^{N} S_i\right) \setminus \cup R$?
2. : $\forall q \in Q$ does there exist a (u_j, u_k) such that $\exists (u_j, u_k) \in \left(\cap_{i=1}^{N} S_i\right) \setminus \cup R$?
3. : $\forall q \in Q$ does there exist a c such that $\exists (u_j, u_k) \in \left(\cap_{i=1}^{N} S^*_{i,c}\right) \setminus \cup R^*_c$?

Results Using the subset of data for the 90 queries that produced SERPs both with and without low-quality banners in the June wave, we found that the presence of specific URLs does not fully explain low-quality banner presence. We use this subset of queries because the variance in their banner presence can be leveraged to compare the conditions under which a banner does or does not appear. For 25 of these 90 queries, we were able to pinpoint a single URL that was always present in its search results when it received a low-quality banner, but never present when it did not (Figure 2.4a). However, the remaining majority of queries did not have a single URL that met the same conditions. For example, the query *"mrna prions"*—which consistently surfaced anti-vaccine articles and social media posts in its top 10 results (Section 2.4.8)—returned a low-quality banner in 56 of the 73 time steps and had a URL that was highly associated with banner presence, but the presence of that URL in the search results for that query did not guarantee a low-quality banner (Figure 2.4b). Extending this measure to account for the presence of two URLs helped to explain low-quality banner presence for 29 of the 90 queries, and extending this measure to three URLs allowed us to explain 67 of the 90 queries.

To help account for search rankings, we also checked for determinative cases where the presence of a pair of URLs within a specific range of rankings always produced a banner. The goal of this was to test questions like "if website i is in the top c search results and website j appears in search results below c, is there always a banner?" For the June 2024 data, we found that at least one rank cutoff could explain the presence of low-quality banners for 65



(a) The presence of the first archive.ph URL ("archive.ph...aocz") can explain all banners displayed for "advanced search result manipulation technology" observed over 73 time steps.

(b) While the newstarget ("newstarget...html") and sideplayer ("sideplaye...vcjd") URLs never appeared without a low-quality banner, neither URL appeared in all SERPs with a low-quality banner.

Figure 2.4: The presence of specific URLs can sometimes fully explain the presence of low-quality banners for certain queries (left), but not for others (right). The vertical line indicates the total number of banners observed for the query. If every time a URL appears there is a low-quality banner (red bar) and never appears when no such banner is present (black bar), the URL can fully explain observed banners for the query.

queries of 90 across all time steps, and a cutoff of $c = 1$ (i.e., the top-ranked URL) was able to explain the largest proportion of those queries (48 of 65). Although some queries were fully explainable with larger cutoff values (i.e. $c > 1$), the proportion of banners explainable by subsequent cutoffs ($c = \{2, \ldots, 50\}$) decreased monotonically. These results show that, while the top-ranked results on a SERP are an important factor in determining low-quality banner placement, they do not fully explain the presence of low-quality banners.

Stability Experiments

In addition to our rapid data collection waves, we also sought to examine the impact that introducing minor changes to a search query could have on the SERP it returned. We were motivated to pursue this extension because deviations in how queries are phrased, or other intentional or unintentional linguistic differences, can substantively impact the SERP

returned when that query is searched. Moreover, past work has found that data voids are often spread through strategic keyword choices Golebiewski and boyd [2019], Tripodi [2022], and that lexical variants are used to evade content moderation more broadly Chancellor et al. [2016].

We designed and launched several experiments in August 2024 to evaluate the stability of a query's SERP under various minor permutations. Similar to the rapid data collection waves, we used the 301 queries that produced a low-quality banner in wave 1 as our starting point. We expanded this into our full query set by pluralizing nouns and noun phrases, introducing keyboard typos (with 10% probability), and adding or removing quotation marks. After collecting the search results for this set, we found that not a single query—original or permutation—generated a low-quality banner. We initially assumed there may have been an error in our data collection, so we tried again using alternative servers and data collection approaches, but again no low-quality banners were returned. As such, we were not able to test the effects of minor query permutations in this study, but these results do allow us narrow the discontinuation date of Google's warning banners to August 2024.

2.4.7 Data Void Models

DistilBERT Model Results

While the DistilBERT classifier yielded high accuracy on a small, balanced "banner" vs. "non-banner" query subset (Table 2.2), we also found that it did not learn generalizable patterns from the query text data alone. When we ran inference over the 1.4M unlabeled queries using the fine-tuned DistilBERT model, we found its 100 most confident banner predictions all contained the presence of quotation marks, despite many of those queries not returning unreliable content (e.g., *"phishing attack"*, *"secretary of state"*, and *"data breach"*). Reliance on such quotations is not ideal, because once a falsehood is subject to widespread fact-checking, such queries can surface reliable or helpful results. In contrast, our GNN approach allowed us to condition the probability of a banner on both the text of the query and its returned domains.

GNN Robustness Checks

Manual Annotation After manually examining the 301 queries that received a low-quality banner in wave 1 (Oct 2023), two annotators labeled the top 20 predictions of each

Table 2.2: Precision of the discovery process for each model at the top 5, 10, and 20 most confident predictions.

	P@5	P@10	P@20
DistilBERT	0	0	0.050
GNN_{Hom}	0.600	0.450	0.575
GNN_{Het}	**0.800**	**0.900**	**0.775**

Table 2.3: The GNN_{Hom} and GNN_{Het} models best identify new out-of-sample bannered queries. This table shows how many of the 74 queries each model identified in its top 10, 50, 100, 500, 1k, 5k, and 10k most confident predictions.

Model	Top 10	Top 50	Top 100	Top 500	Top 1k	Top 5k	Top 10k
DistilBERT	0	1	2	9	12	19	21
GNN_{Hom}	0	**3**	5	**13**	**16**	24	27
GNN_{Het}	0	1	**6**	11	15	**32**	**39**

model based on their corresponding SERPs. The primary label indicated whether or not the SERPs produced by each query should receive a low-quality banner (i.e., returned results that appeared to be low-quality), and there was substantial agreement (Cohen's $\kappa = 0.73$) between the two annotators Landis and Koch [1977]. Both annotators agreed that the GNN_{Het} model yielded the highest precision for the top 5, 10, and 20 predictions (Table 2.2).

Out-of-Sample Banner Identification In wave 2 (Mar 2024) there were 74 queries that received a low-quality banner that did not have a low-quality banner in wave 1 (Oct 2023). To evaluate out-of-sample banner identification, we ran inference over all queries that did not receive a low-quality banner in wave 1 (Oct 2023), and evaluated the number of those queries that received a low-quality banner at time step 2, sorted by confidence. Similar to our other model validation tests, we again find that GNN_{Het} model performed the best (Table 2.3).

Relationship Between Prediction Confidence and Domain Quality Given the focus on reliability in the stated purpose of the low-quality banners (Figure 2.2A), the prevalence of domains with low-quality scores in a query's SERP provides a useful proxy for model evaluation. A model's most confident "low-quality banner" predictions should be associated, on average, with the presence of more low-quality domains. To investigate this

Table 2.4: The heterogeneous model confidently identifies queries associated with lower quality websites. Across all waves the means of the heterogeneous model's most confident 500 query prediction means (Het-Mean) are significantly lower than the means of homogeneous and d-BERT models (Other Means).

Crawl	Het-Mean	Other Means	T	DF	P
1	0.66	0.79	-10.99	632.08	7.64e-26
2	0.71	0.80	-9.11	616.82	1.14e-18
3	0.73	0.80	-7.46	616.11	3.02e-13
4	0.70	0.80	-9.41	634.04	8.62e-20

relationship, we used the rolling mean of average domain quality over the 500 most confident banner predictions obtained from each model, and found the GNN_{Het} model's predictions had the strongest relationship with domain quality (Figure A.2). Average domain quality was calculated for each query as the simple average over all of its returned search results in each wave, and we used a rolling mean with a window size of 10 to account for inherent noise in this metric. Across waves, the GNN_{Het} model displayed the most robustness, and continued to identify queries associated with low-reliability domains on data extracted 5 months after its training data. Not all SERPs had at least one non-social media website with a domain quality score, so those predictions are excluded here. These results suggest that the GNN_{Het} model best learned to identify queries associated with low-quality domains, despite our models not including any explicit domain-level scores, which helps validate our findings and again point to the GNN_{Het} model as the best performer.

Across all four waves, we compared the predicted mean domain quality scores for each of the 500 most confident low-quality query model. For each wave, we performed a t-test comparing the scores from the heterogeneous model against the combined scores of the other two models. In all four waves, the heterogeneous model confidently identified queries with lower mean domain quality scores, and the differences were significant at the 0.001 threshold (Table 2.4). We additionally find that the heterogeneous model prediction ranks are more correlated with domain quality ($r = 0.17, p < 0.001$), than homogeneous ($r = 0.02, p = 0.48$), and DistilBERT ($r = 0.07, p = 0.001$).

Table 2.5: The GNN_{Het} model outperformed other models at identifying Query-SERP pairs that produce low-relevance banners, for wave 1 (Oct 2023).

	Accuracy	F1	Recall	Precision
DistilBERT	0.934	0.902	0.908	0.896
GNN_{Hom}	0.918	0.907	0.870	0.881
GNN_{Het}	0.964	0.959	0.928	0.963

Table 2.6: Few of the queries that our models confidently predicted a low-quality banner for also received a high-confidence prediction for a low-relevance banner.

	Jaccard Similarity of Top N Predictions				
	100	500	1k	5k	10k
DistilBERT	0	0.001	0.005	0.093	0.196
GNN_{Hom}	0	0.000	0.000	0.001	0.004
GNN_{Het}	0	0.005	0.007	0.013	0.027

Low-relevance Banner Models

To better understand how low-relevance banners were distinct from low-quality banners, we also trained a set of models to predict them as well. The training and preprocessing procedures were almost identical, but we elected to a 2-to-1 negative sample ratio. This was for two reasons: first, as there are almost 14k low-relevance banners, data sparsity is not as much of an issue as it was for low-quality banners. Second, we found that a 2-to-1 negative sample outperformed the 3-to-1 negative sample on all evaluation metrics for the low-relevance banner model. As with the low-quality banner models, each model was trained 10 times and we report the average over all runs, and again find that the GNN_{Het} model outperformed the other models on all metrics (Table 2.5).

During our rapid collection, we observed that 93 of the 301 queries returned a low-relevance banner at least once. Of the 21,775 SERPs total collected over the 73 timesteps, 4,170 returned a low-relevance banner. Consequently, we were curious to determine whether models can discriminate between query-SERP cases which warrant a low-relevance banner vs. cases which warrant a low-quality banner. To evaluate the separability of the two classes, we first compare predictions of models trained on quality banners to predictions trained

on low-relevance banners. More specifically, for each model used in our low-quality banner and low-relevance banner experiments, we compared the k most confident predictions made by each model pair (Table 2.6). For k=100 and DistilBERT, this means calculating the Jaccard Similarity of the 100 most confident banner predictions of the DistilBERT trained on low-relevance banners and of the DistilBERT trained on low-quality banners. Despite the low-relevance banners appearing on queries associated with low-quality banners during rapid collection, there was little overlap in the most confident predictions made by each model.

2.4.8 Query Examples

This section provides detailed examples and specific URLs for queries mentioned in the main text that produced low-quality search results across multiple data collection waves.

Among the subset of queries that consistently produced domains with low-quality domains across waves were examples like “”wikileaks is cia””, which produced SERPs with an average domain quality of 0.11 in the first two waves. While the search results for this query in the last two waves did not have any domain quality scores, the top ranked search result in Sept 2024 was a Facebook post by the Wikileaks account containing a cartoon image depicting Barack Obama and Hillary Clinton dressed in trench coats and exchanging dossiers. The majority of the remaining results, many of which led to social media sites or blogs, promoted a false claim that “WikiLeaks is CIA”. The second, fifth, and seventh ranked results were Twitter/X posts that reiterated the claim “WikiLeaks is CIA”.

One search query from our dataset, “mRNA prions,” is likely in reference to debunked claims stemming from a 2021 report on how mRNA vaccines could cause diseases like Alzheimers Funke [2021]. This query consistently returned content promoting debunked claims about COVID-19 vaccines. In waves 1 and 2, when this query received low-quality banners, it surfaced articles alleging that vaccines contain “manipulative nanoparticles” designed to induce “magnetism” in recipients[1]. While the specific magnetism article disappeared in waves 3 and 4, this query continued to surface similar content from prominent conspiracy websites, including articles from naturalnews.com falsely claiming that mRNA vaccines cause “heart attacks, strokes, infertility, turbo cancer, deranged thinking, loss of appetite, and 3-foot-long vascular clots”[2].

[1] https://web.archive.org/web/20210624162453/https://forbiddenknowledgetv.net/magnetism-intentionally-added-to-vaccine-to-force-mrna-through-entire-body/

[2] https://web.archive.org/web/20240919072715/https://www.naturalnews.com/2024-08-23-walzs-minnesota-plandemic-incentives-nationwide-scam-bird-flu.html

As of September 2024, the top 10 search results for "mRNA Prions" still surfaced support for this narrative, including a Research Gate article titled "COVID-19 RNA Based Vaccines and the Risk of Prion Disease," the author of which has also been published by SciVision Publishers, which is included on a list of Predatory Journals and Publishers (`beallslist.net`). The "mRNA prions" query returned a low-quality banner in 56 of the 73 timesteps in our temporally dense dataset from June 2024, and we found that two URLs could explain the majority of instances (43 of 56) in which that query produced a banner using our rank-cutoff approach (Section 2.4.6). In the August 2024 supplementary wave, this query did not display the newstarget or 4chan URLs strongly associated with a banner in Figure 2.4b, but returned three URLS we never observed in the June data: two links to a LinkedIn post and YouTube video posted by an apparent Australian anti-vaccine organization[3] and a link to an 8kun post encouraging users not to get vaccinated[4].

2.4.9 August 2024 Core Update

The disappearance of quality banners coincided with an August 2024 Google core update Schwartz [2024]. In this section, we consider the question "did Google remove quality banners because the August 2024 update removed all unreliable results?" We find that while the update did significantly impact search results for formerly quality-bannered queries SERPs did not necessarily become more reliable.

For each query, for each of the 73 time steps from June 2024, we calculated the Jaccard similarity and RBO between the top 10 SERPs of the June data and the August data (Section 2.4.6). We found that the mean average Jaccard similarity over all queries and time-steps was 0.31 (max 0.75), and the mean RBO was 0.37 (max 0.70). This was only slightly less than the SERPs with low-quality banners from March 2024 (Section 2.4.6), where the average Jaccard similarity and RBO with the June data were 0.27 (max 0.69) and 0.27 (max 0.62), respectively.

In our results section, we stated that 25 queries that displayed between 1 and 72 banners had at least one URL appear in all bannered SERPs that never appeared in unbannered SERPs. We looked for these URLs in the August data, and found that 8 queries displayed those URLs strongly associated with banners in the June data, yet returned no banner. Additionally, we calculated average domain reliability of the March 10th rapid crawl, June 7th rapid crawl, and August rapid crawl. This is a coarse-grained approach, as we do not

[3] https://web.archive.org/web/20240326082456/https://healthallianceaustralia.org/

[4] https://web.archive.org/web/20240703140131/https://8kun.top/freedomzine/res/15841.html

have labels for 81–90% of domains returned in SERPs, but this still provides some frame of comparison. We found that the average domain-level SERP reliability (where 1 is most reliable, and missing values are ignored) were 0.52 (March), 0.44 (June), and 0.47 (August). While the labeled August SERPs results were, on average, slightly more reliable than the June 7th SERPs, they were slightly less reliable than the March SERPs. This provides additional evidence that banners were not turned off in August because the problem was solved.

A more qualitative exploration revealed that while many search results changed between June and August, the August search results often did not correspond with an increase in overall SERP reliability. The query "mrna prions" (Section 2.4.8), did not display the newstarget or 4chan URLs strongly associated with a banner in Figure 2.4b, but returned three URLS we never observed in the June data: two links to a linkedin post and youtube video of posted by an apparent Australian anti-vaccine organization[5] and a link to an 8kun post encouraging users not to get vaccinated[6].

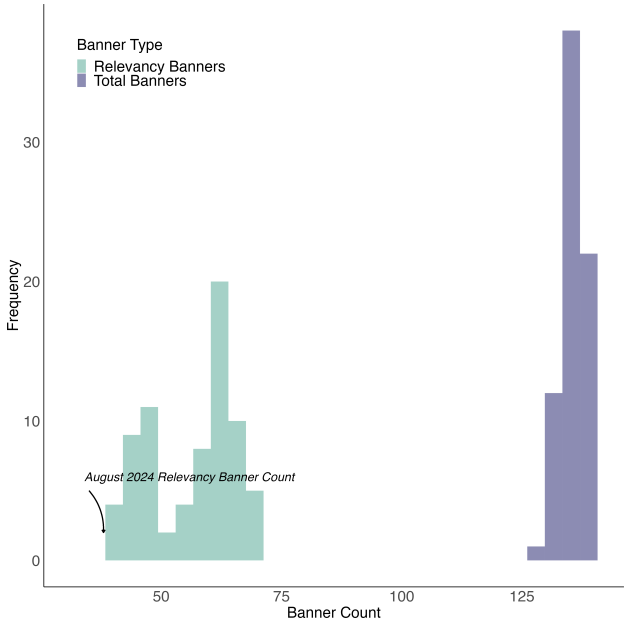


Figure 2.5: Quality banners do not appear to have been subsumed by low-relevance banners. We display histograms of the distributions of the counts of low-relevance banners over 73 time-steps (teal) and the distribution of total banners (i.e., low-relevance + low-quality) (purple). In August 2024, the count of low-relevance banners we observed was lower than we observed across any of the 73 June 2024 time-steps.

[5] https://web.archive.org/web/20240326082456/https://healthallianceaustralia.org/
[6] https://web.archive.org/web/20240703140131/https://8kun.top/freedomzine/res/15841.html

We considered the possibility that quality banners had been subsumed inside low-relevance banners during the August 2024 update. To test this hypothesis, we created one histogram showing the total number of low-relevance banners for the 296 queries we analyzed over 73 time-steps from June 2024 and a second histogram containing the sum of the counts low-quality and low-relevance banners that appeared in each of those 73 time-steps (Figure 2.5). If every quality banner were replaced with a low-relevance banners, we would expect that the total number of low-relevance banners in our August 2024 data should fall within or near the distribution of total banners (i.e., the sum of low-quality and low-relevance banners). This was not what we observed; rather, the number of low-relevance banners returned over the 296 queries in August (38) was lower than the count of low-relevance banners returned in any of the 73 time steps (the minimum was 41).

2.5 Conclusion

In this study, we developed and deployed methods for surfacing and evaluating the warning banners used by web search engines, used that information to design GNN models for identifying unlabeled data voids, and applied those models to examine the prevalence of data voids across four waves between 2023 and 2025. We found that Google's low-quality banners were rare, governed by inconsistent and evadable rules, and disappeared around August 2024 without an official announcement and without a corresponding improvement in the average domain quality of their search results. After sharing our initial findings with a journalist as a preprint Robertson et al. [2025] in early 2025, Google confirmed the banners had indeed been discontinued Newton [2025]. These findings raise broader questions about the reliability of the content moderation systems used by online platforms, the consistency of the policies that govern them, and the timing and transparency of the changes we observed on Google Search in the lead up to the 2024 US Presidential election.

When Google was using its low-quality warning banners, our search queries rarely produced them, and their presence could be inconsistent over short time spans. We also found that Google never displayed a low-quality banner for queries with advanced operators, but we found a number of queries that appear to have intentionally abused this loophole to guide people into data voids (e.g., the query "covid vaccine detox site:naturalnews.com"). The absence of low-quality banners in such scenarios could be framed as respect for user choice, which recent research suggests plays a large role in the content that people consume while on search engines Greene et al. [2024], Robertson et al. [2023b], van Hoof et al. [2025], and Google

has stated that "users may decide to seek and select content that our signals determine to be of low-quality ... we believe it is of fundamental importance to respect their choices" Google [2024]. However, an absolute rule here ignores the social side of search: people regularly share search queries on social media Robertson et al. [2023a]. Not all people who click on a search link, or copy and paste a search query, will understand the impact that advanced operators have on the information they receive.

Although Google noted an unspecified "ranking quality improvement" Newton [2025] as a reason for discontinuing the low-quality banners, we found little evidence of substantive increases in domain quality across waves. Instead, many of the queries that received a low-quality banner in Oct 2023 and Mar 2024 continued returning low-quality results in Sept 2024 and Feb 2025, they just no longer included a warning banner. Similarly, our GNN model identified up to 48× more SERPs as data voids than Google applied its low-quality banners to in the first two waves, and continued to identify data voids at a similar rate afterwards. Another reason Google provided for the discontinuation was that the low-quality banner was "surfacing extremely infrequently and was triggering false positives at a high rate" Newton [2025]. While neither of these provide an obvious rationale for discontinuing the banners rather than improving them, our results align with the infrequency claim (the banners were rare), but do not suggest a high false positive rate (e.g., low-quality banners were associated with low-quality domains).

Our use of search directives to identify a set of search queries is both an advantage and limitation of our study. Compared to prior algorithm audits of web search, where researchers often select their own search queries or solicit them from survey participants Ballatore [2015], Norocel and Lewandowski [2023], Lurie and Mulligan [2021], van Hoof et al. [2022], our use of search directives allowed us to develop perhaps the largest query set ever used in an algorithm audit, which was crucial to our goal of surfacing a sufficient sample of warning banners for training our models. While this approach provided us with a large and diverse set of queries, and avoided researcher-induced selection biases, this design choice also means that our query set only reflects the queries that people were willing to share on a single social media site, and are not representative of the queries that people may have been searching at the time of our study. However, over 10% of our search queries appeared in a dataset of frequently searched queries that Bing released in 2020, suggesting that many of our queries remained relevant and in use (Appendix A.0.3).

While the data void prevalence estimates we find are specific to queries shared on social media—and do not provide a representative sample of real users' queries—they may be viewed

as a lower bound on queries of this type for three reasons. First, we only examine results on Google, which has publicly shared their extensive processes for understanding and improving the quality and reliability of their search rankings Google [2023], and other search engines may lack such resources. Second, only 0.11% of our queries contained a conspiracy-related keyword, suggesting that such queries were not overrepresented in ways that might inflate our estimates. Third, the majority of our queries were in English, which has been shown to produce higher quality results than similar queries in other languages Borge et al. [2021]. Future work should explore independent approaches to collecting exposure and engagement with warning banners among real search engine users Feal et al. [2024], Robertson et al. [2023b], collaborations with industry that could facilitate that research Greene et al. [2024], and examine a broader set of search engines, languages, and locations Kravets and Toepfl [2022], Makhortykh et al. [2020, 2022a], Toepfl et al. [2023], Urman et al. [2022b], Zoubi et al. [2022].

As search evolves and continues to incorporate LLMs and agent-based approaches White [2024], studies like ours may become both increasingly important and difficult to conduct. The methods we developed here are platform agnostic, but LLMs add greater stochasticity to the outputs of search engines, may change how people write their search queries, and introduce a conversational interface that could change fundamental aspects of how we conduct online searchesSpatharioti et al. [2025]. While recent research finds that LLMs can be used to reduce conspiracy beliefs Costello et al. [2024], it is unclear whether popular platforms would incorporate the specific prompts needed to elicit that behavior, especially given recent trends towards less moderation. On the supply side, LLMs may also make data voids easier to create, harder to identify, and more likely to be incorporated into a search results page or a chatbot's response Aggarwal et al. [2024], Keh and Thompson [2024a]. As such, the intersection of data voids and the content moderation practices of both search engines and LLM providers, especially as the two become more intertwined, presents an important and pressing area for future research.

Author Contributions Ronald E Robertson and Evan M Williams designed the research. David Thiel built the infrastructure for collecting social media posts, and Ronald E Robertson built the infrastructure for collecting search results. Evan Williams designed the deep learning models. Ronald and Evan analyzed the data and wrote this paper.

Chapter 3

Structural Authority Search Engine Manipulation

3.1 Introduction

In chapter 2, we explored several concrete ways that Google's algorithm can return unreliable information. In this chapter, I will approach the problem of misinformation spread via websearch through the perspective of the attacker. I begin with a case study exploring manipulation of webgraphs around pro-Kremlin think tanks. I will then generalize the findings to a broader set of news domains.

The Kremlin's use of bots and trolls to manipulate the recommendation algorithms of social media platforms is well-documented by many journalists and researchers, but pro-Kremlin manipulation of search engine algorithms had never been explored. We examine pro-Kremlin attempts to manipulate search engine results by comparing backlink and keyphrase networks of US Think Tanks, European Think Tanks, Russian Think Tanks, and Kremlin-linked Pseudo Think Tanks that target Western audiences. We find evidence suggesting pro-Kremlin pseudo think tanks are being artificially boosted and co-amplified by a network of low-quality websites that generate millions of backlinks to these target websites. We find that Google's search algorithm appears to be penalizing Russian and pseudo-think tank domains.

3.2 Chapter Research Questions

RQ3.1 What SEO strategies do low-reliability websites attempt to rank highly on Google?

RQ3.2 What SEO features are associated with low-reliability news sites?

RQ3.3 Using SEO features and webgraphs, can we build models to identify website attempts to manipulate search rankings?

3.3 Implications

In 2014, Vitaly Bespalov was hired as a writer by a secretive organization, later revealed to be the Kremlin-linked Internet Research Agency (IRA), a troll farm which was indicted by a US federal grand jury for online US election interference in 2018 [United States of America, 2018]. On an average day, Bespalov recounts being tasked with rewriting articles on Ukraine over and over, each time keeping roughly 70 percent of the original text. Bespalov states he was asked to change words like "terrorist" to "militia" or to write "national guard" instead of "Ukrainian army," and he was instructed to never criticize Russia in these articles [Popken and Cobiella, 2017a]. The goal of this operation, according to Bespalov, was to get the articles to the top of search engine results [Popken and Cobiella, 2017a]. These kinds of activities present an important and understudied component of Kremlin-linked attempts to spread online misinformation and propaganda.

In recent years, the Kremlin's influence operations on social media platforms have been the subject of widespread public attention and research. In 2017, an estimated 127 million individuals were exposed to Russian disinformation on Facebook alone [Isaac and Wakabayashi, 2017]. In 2021, Facebook reported in its Threat Report on influence operations that Russia remains the biggest driver of disinformation and detailed the removal of Kremlin-backed networks linked to the IRA and other Russian military intelligence organizations [Facebook, 2021]. To make these information operation attacks successful, IRA trolls rely on a wide range of techniques to manipulate recommendation algorithms [Carley, 2020a, Benigni et al., 2017]. Similarly, using what is known about search engine ranking algorithms, bad actors can attempt to artificially inflate website search result rankings.

Search engine rankings represent an important line of inquiry in misinformation research, as the rankings can substantially impact both what information users consume and what information they believe. High-ranking pages—for example, the pages that appear at the top

of a Google search result—are far more likely to be seen. In an analysis of 300 million search engine clicks, 92% were on the first page of search results, and 51% of those were the first or second result [Chitika Insights, 2013]. Users were also found to be 140% more likely to click the last result on the first page than the first result on the second page [Chitika Insights, 2013]. More recent analyses by Backlinko and Ignite Visibility both found the click-through rate of the first result was ten times higher than that of the tenth result [Dean, 2022, Lincoln, 2020]. It has also been demonstrated that search engine rankings can impact political beliefs and voting patterns: three laboratory experiments with double-blind control group designs found that relatively minor changes in rankings could influence decisions of undecided voters by 20% [Epstein and Robertson, 2015a]. While the magnitude of this effect has been contested [Zweig, 2017], if manipulation of search engines is successful, more individuals can be exposed to artificially amplified content.

This increased exposure can have substantial impacts on information environments and has become increasingly concerning with the rise of what Marwick and Partin [2022] call *populist expertise*, or the rejection of experts and traditional information vectors in favor of alternative "home-grown" knowledges. On QAnon message boards, users implore others to "do the research" to understand conspiracies in Q's messaging. More recently, "do your own research" has become synonymous with vaccine-hesitancy movements that lead users down anti-vax rabbit holes [Ballantyne and Dunning, 2022]. A 2021 search engine audit study found that Google was highly effective at suppressing material that promotes misinformation for conspiracy keyphrases, whereas Bing, Yahoo, Yandex, and DuckDuckGo were far less effective [Urman et al., 2022c]. Consequently, conspiracy communities increasingly urge followers to use search engines that are less effective at suppressing misinformation [Thompson, 2022]. Understanding the environment in which users "do the research" is therefore important.

While Kremlin influence operations on social media platforms have received widespread attention, there is currently very little research on state-backed attempts to manipulate search engine recommendation rankings. Audit studies have compared rankings across engines; for example, Kravets and Toepfl [2021] found that Yandex returned far fewer websites than Google with information on anti-regime Russian protests, and other work examined cross-country differences in pro-Kremlin content suppression [Toepfl et al., 2022, Makhortykh et al., 2022b]. Search engine audits are useful but constrained by the choice of keyphrases. Addressing these limitations, Bradshaw [2017] used proprietary data to examine search optimization strategies of junk news sites, and Hrckova et al. [2021] used Ahrefs to explore linking patterns of partisan news sites. We adopt a similar methodology.

We examine networks of backlinks and keyphrases (search terms for which websites rank highly) for US, European, Russian, and what we refer to as pro-Kremlin "pseudo" think tanks—organizations that blur the lines between news, think tank content, misinformation, and propaganda. Successful search manipulation can result in websites ranking highly on search engines for specific keyphrases. Conspiracy theories have a distinct advantage here, as conspiratorial keyphrases can be highly specific and have very low competition [Golebiewski and Boyd, 2019]. For example, the pseudo-think tank Katehon has over 99% of its links coming from low-quality domains; despite this, as of September 24, 2022, a highly conspiratorial article produced by Katehon appears in the top three results on multiple search engines for the keyphrase "Rothschild criminal."[1]

Our principal recommendation is that further exploration is needed. While Google appears to be penalizing the Russian and pseudo-think tanks in this report, we do not have enough data to determine how widespread these manipulation behaviors are or how Google's penalization algorithm generalizes to other contexts. If Google is penalizing based solely on curated lists—where links between these think tanks and the Kremlin have been publicly documented—then newly created sites or sites not otherwise identified as misinformation may not be penalized. If, in contrast, Google's algorithm is penalizing these domains as a result of their backlink profiles, further research is needed to ensure smaller "innocent" sites are not also being penalized. Finally, we note that even if Google perfectly penalizes these sites, they can still attract new users through other search engines, social media, or direct links from other domains. Search engine companies have monetary interests in stopping search manipulation, aligning the goals of companies, regulators, and stakeholders [Ghosh and Scott, 2018]. This study provides a methodological framework to help researchers, journalists, and other stakeholders further explore and expose state-aligned search manipulation attempts in other country and topic contexts.

3.4 What is search engine manipulation?

Search engines can rank websites using a variety of algorithms that are not made public, but the most well-known is PageRank, proposed by Google co-founder Larry Page [Page et al., 1999]. Google's algorithm has become far more complex since PageRank was initially proposed, but backlink quantity and quality remain important factors. PageRank determines

[1]The article promotes antisemitic conspiracies; the term is defined in ADL's Glossary of Extremism (https://extremismterms.adl.org/glossary/rothschilds).

ranks by both the quantity and quality of inbound links. Although many other attributes are weighted by search recommendation algorithms, link quality and quantity remain important.

Manipulating rankings through the creation of many new backlinks can involve paying third-party services to post a target website across third-party websites, creating websites to amplify a target website, hacking webpages and injecting invisible URLs, and many other maneuvers. Google calls these kinds of manipulations "link schemes" and explicitly forbids them in its webmaster guidelines.[2] When link quantity is manipulated, a website's ranking can be boosted for various keyphrases, increasing the likelihood it appears on the first page of results and thus increasing traffic. Search engine optimization (SEO) is the broad class of actions taken to increase visibility. We focus on the "link scheme" subset of SEO forbidden by guidelines ("black-hat" SEO), which we refer to as search engine manipulation (SEM). For more details, see Tripodi [2022].

We also explore keyphrases—search terms for which websites rank highly. Conspiracy theories have an advantage here, as keyphrases can be highly specific and have very low competition [Golebiewski and Boyd, 2019]. For example, if a user in a vaccine-hesitancy group sees a post falsely accusing a company of a specific criminal action, a conspiracy website could reiterate the claim and rank highly for its keyphrases because there is often little competition for emergent conspiratorial stories.

3.5 Findings

3.5.1 Finding 1: Pro-Kremlin websites are heavily amplified by domains seemingly built for generating backlinks.

We find a highly imbalanced backlink volume across networks. Global Research, a Kremlin-aligned pseudo-think tank, receives 22.1 million backlinks, more than all US, European, and Russian think tanks combined. In total, American think tanks received 14.1 million links, European think tanks received 4.6 million links, Russian think tanks received 1 million links, and pseudo-think tanks received 30.8 million backlinks. Although the pseudo-think tanks received more links than Russian and US think tanks combined, most of these links were from low-quality domains. Ahrefs (see Methods) calculates PageRank-like scores normalized from 0–100, where 100 signifies the most authoritative. Websites that linked only to pseudo-think tanks in our network had a mean score of 17.04, whereas the mean rank for all other

[2] https://developers.google.com/search/docs/advanced/guidelines/link-schemes

backlinking websites was 35.59.

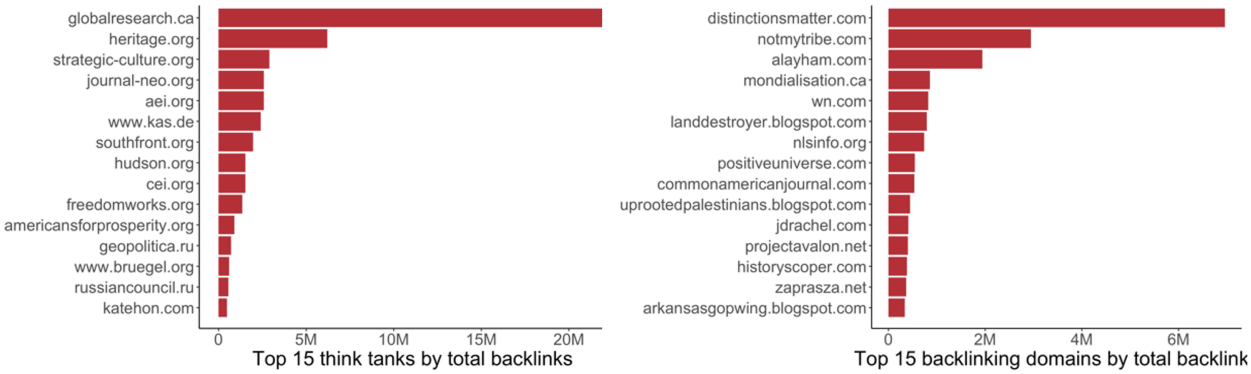


Figure 3.1: Figure 1. Left: Top 15 think tanks by backlink volume. Right: Top 15 backlinking websites.

Distinctionsmatter, the website with the largest volume of outbound links (all to pseudo-think tank domains), bears indicators of a link scheme website. There is no "about" or contact information for the author, there are no ads on the page, and the page generates millions of links to unreliable news sites. Distinctionsmatter posted 6.6 million links to Global Research—more than any other site. Its second most linked-to site is Zero Hedge, a news outlet accused of spreading Russian propaganda by US intelligence agencies [Walsh, 2022].

The site that generates the second most backlinks, `notmytribe.com`, has the same general tone as many IRA bots identified in Twitter operations. The navigation banner on `notmytribe.com`'s landing page has "News," "Culture," "Work," and "Disinfo" dropdown sections, with labels that suggest non-native English or machine translation (e.g., "Info Virus," "On War Machine," "AntiGlobalization," "Nighttime Gardening"). The purpose of the website is almost certainly to boost ranking of other sites, as it generates 2.9 million backlinks to Global Research. We inspected each of the 15 top backlinking domains and found 13 of 15 (with the exceptions of `wn.com` and `nlsinfo.org`) met criteria consistent with pro-Kremlin amplification or disproportionate linking to pro-Kremlin domains.

3.5.2 Finding 2: Keyphrases of pseudo-think tanks exhibit high internal overlap and appear to target conspiracy theorists.

Many top keyphrases shared between think tank groups are names of people, but many are very specific and conspiratorial, particularly within the pseudo-think tank network. These keyphrases appear to be targeting conspiratorial "data voids" where search results are sparse

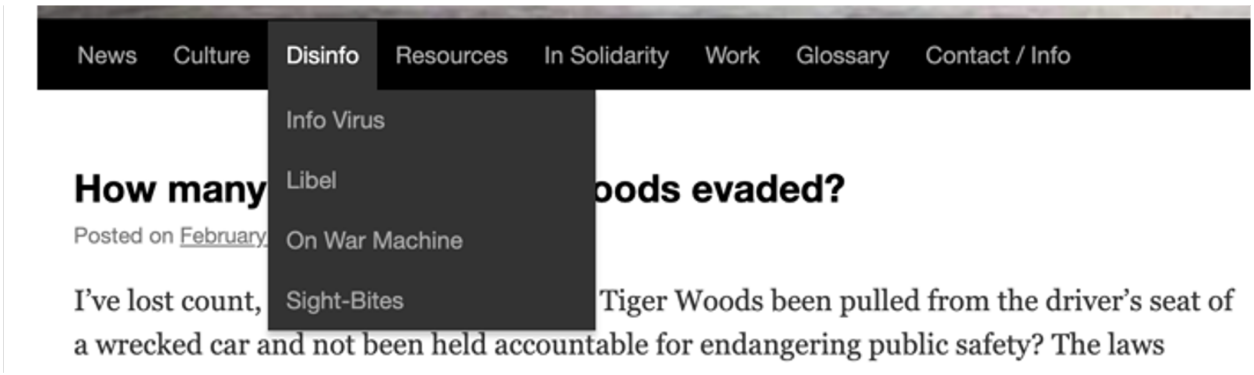


Figure 3.2: Figure 2. Notmytribe.com's site navigation bar contains a "Disinfo" drop-down with conspiratorial subsections.

and low-quality websites can rank highly [Golebiewski and Boyd, 2019]. Among the pseudo-think tank keyphrases are phrases like "Is Zelensky a drug addict," "Zelensky on cocaine," "neutron bomb in Yemen," and "subcortical vascular dementia Hillary Clinton."

The keyphrases network visualizes keyphrases used by two or more think tanks. The Russian think tanks are characterized largely by Russian-language keyphrases and only have one overlapping keyphrase with another think tank group.[3] Several keyphrases cross think tank groups; for example, pseudo-think tanks and `hudson.org` share "climate change money trail," multiple variants of "Mike Pompeo speech," and keyphrases about former CIA director John Brennan voting communist.

We find, on average, Russian and pseudo-think tanks rank much lower for keyphrases than US and EU think tanks. Averaging Google positions over the top 1,000 keyphrases for each network, average positions are: US (7), Europe (11), Russia (35), and pseudo-think tanks (34). This suggests Google is penalizing pseudo-think tank domains despite, or perhaps as a result of, ongoing link scheme manipulation attempts.

3.5.3 Finding 3: Many pseudo-think tanks are strongly amplified by the same websites.

We constructed a co-amplification network for inbound links using pairwise co-amplification weights between each website, and we visualized a filtered version of this co-amplification network. Edges reflect the strength at which think tanks are linked to by the same domains.

[3] `russiacouncil.ru` (Russia) and `ifri.org` (Europe) share "csto."

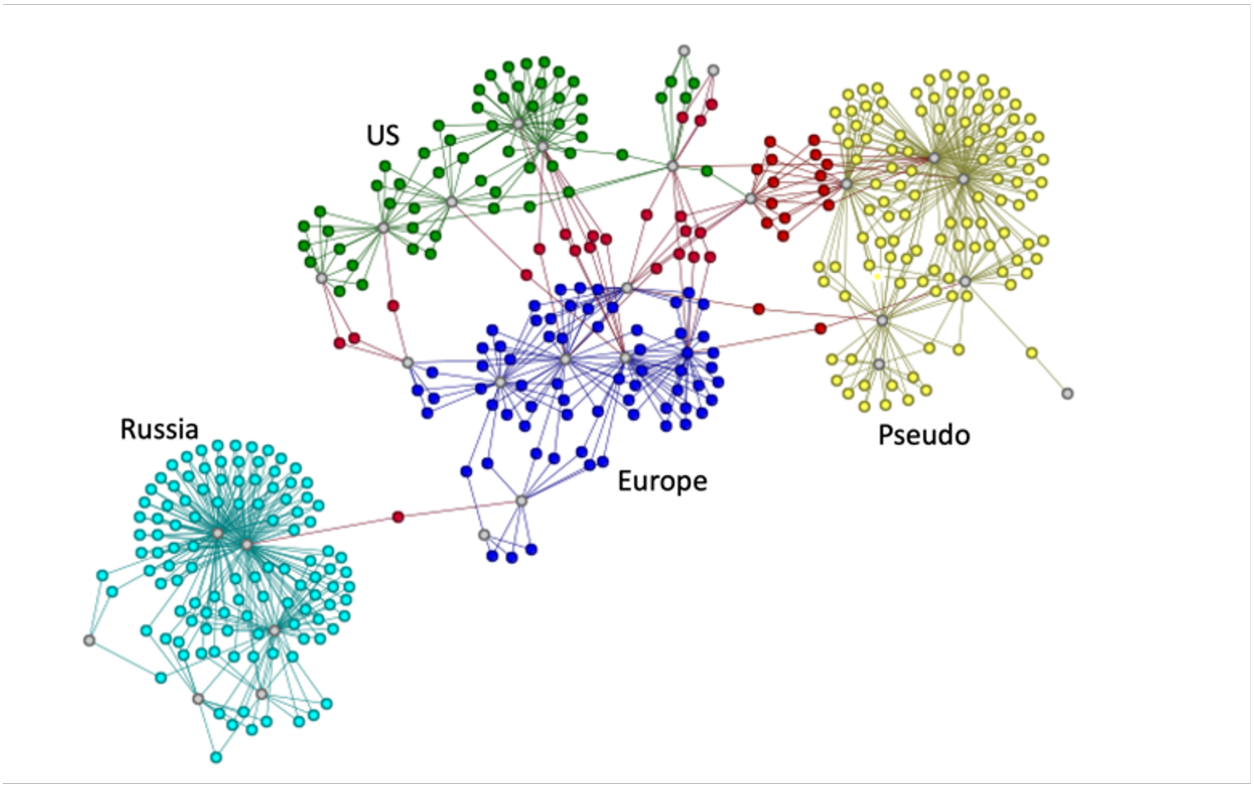


Figure 3.3: Figure 3. Keyphrase network visualization. Grey nodes are think tanks. Blue nodes are EU keyphrases, teal are Russian keyphrases, green are US keyphrases, yellow are pseudo-think tank keyphrases, and red are keyphrases shared across different think tank groups.

We only visualize edges with edge weights larger than 15k.[4]

For an unweighted adjacency matrix $A \in \mathbb{R}^{m \times n}$, overlap can be calculated using $\mathcal{O} = A^\top A$. However, when adjacency matrices are weighted, matrix multiplication can cause the metric to lose interpretability. If a domain is linked to by i 10 times and by j 100 times, a co-amplification score should not exceed 10. Formally, we want to constrain overlap for domains $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ so that

$$\mathcal{O}_{u,v} \leq \sum_{i=1}^{d} u_i \quad \wedge \quad \mathcal{O}_{u,v} \leq \sum_{i=1}^{d} v_i.$$

To satisfy this constraint, we define co-amplification as the sum of minimum pair-wise overlap:

$$\mathcal{O}_{u,v} = \sum_{i=1}^{d} \min(u_i, v_i).$$

[4]This threshold is arbitrarily chosen; it highlights strongest pairwise ties and improves readability.

The diagonal of the matrix is zeroed out to remove self-links. Intuitively, if two domains are heavily linked to by the same websites, this score will be high; if not, it will be low.

The most substantial co-amplification happens within the pseudo-think tank network. Sites with the highest overall co-amplification scores are Global Research (4.4M), Strategic Culture Foundation (4M), Heritage Foundation (3.8M), New Eastern Outlook (3.2M), and American Enterprise Institute (2.7M).

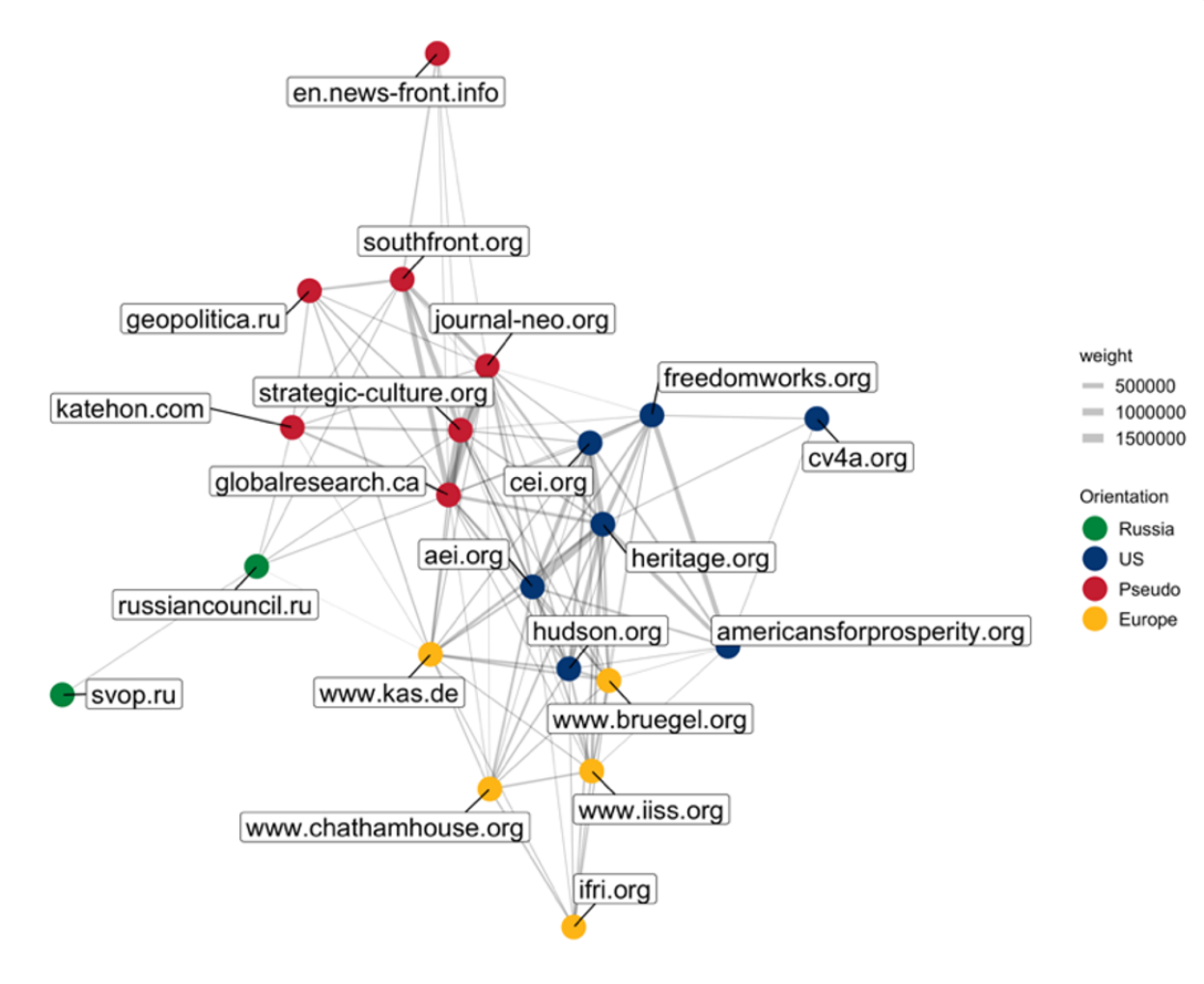


Figure 3.4: Figure 4. Filtered co-amplification network. Each edge indicates at least 15k links from the same set of referring domains. Green nodes are Russian, yellow nodes are European, blue nodes are US, and red nodes are pseudo-think tanks.

While strongest edge weights are between think tanks of the same group, there is notable co-amplification across groups, particularly between pseudo-think tanks and American right-wing think tanks. The Heritage Foundation has an edge weight of 163k with New Eastern

Outlook and 213k with Global Research. The Heritage Foundation, American Enterprise Institute, and New Eastern Outlook all received almost exactly 102.2k backlinks from `historyscoper-islamwatch.blogspot.com`, an anti-Islamic blog with millions of external links to what appear to be conspiracy sites. `oldephartte.blogspot.com` is a more standard SEO blog that links heavily to Global Research, Strategic Culture Foundation, and New Eastern Outlet while also often linking to the Heritage Foundation.

3.6 Methods

3.6.1 Data collection

To choose initial domains of interest, we leveraged the fact that the Kremlin invests in soft-power initiatives, including Kremlin-linked think tanks and pro-Kremlin pseudo-think tanks. We examined eight Kremlin-linked think tanks identified by the Institute of Modern Russia with primarily domestic Russian audiences [Smagily, 2018b]. We contrasted these with eight influential Western European think tanks (top eight Western European think tanks ranked in the University of Pennsylvania's 2020 *Global Go To Think Tank Index Report* [McGann, 2021b]) and eight US conservative think tanks (those supplying the largest number of staff, cabinet, and political appointees in the Trump administration [Kravitz et al., 2019b]). We contrasted these networks with seven Kremlin-backed proxy outlets identified in the Global Engagement Center's *Pillars of Russia's Disinformation and Propaganda Ecosystem* report [U.S. Department of State: Global Engagement Center, 2020a]. We call these pseudo-think tanks, as each blurs the lines between news, think tank content, misinformation, and propaganda. All think tanks and their network assignments are shown in Table 3.1.

All data for this project were collected using the platform Ahrefs.[5] Ahrefs advertises a 12-trillion link database and the second most active commercial web crawler after Google.[6] For each of 31 think tanks, we pulled (1) the top 1,000 backlinking websites by backlink volume and (2) the 1,000 keyphrases with the highest Google positions for each domain. If a website had fewer than 1,000 results for a query, all results were returned. These queries yielded eight networks (two kinds of networks across four groups). Summary statistics are shown in Tables 3.2–3.3. Data were collected on September 22, 2022.

[5] https://ahrefs.com/

[6] https://ahrefs.com/robot

Table 3.1: Table 1. All think tank domains and their corresponding networks.

Russia	**Europe**	**US**	**Pseudo**
cskp.ru	bruegel.org	freedomworks.org	strategic-culture.org
rethinkingrussia.ru	realinstitutoelcano.org	cei.org	globalresearch.ca
russiancouncil.ru	clingendael.org	heritage.org	journal-neo.org
svop.ru	ifri.org	cv4a.org	geopolitica.ru-en
foreignpolicy.ru	chathamhouse.org	hudson.org	southfront.org
iiseps.org	dc.fes.de	americaneconomic freedomalliance.com	en.news-front.info
doc-research.org	kas.de	americansfor prosperity.org	katehon.com-en
eurasian-strategies.ru	iiss.org	aei.org	

3.6.2 News Webgraph GNNs

In the first section of the chapter, we demonstrated that unreliable websites can employ search engine optimization techniques to increase their reach and audience. In this section, I will evaluate how well graph neural networks can differentiate reliable and unreliable websites using only SEO features and webgraph structure. Being able to identify unreliable domains is an important task if the aim is to return reliable information where possible. However, keeping track of unreliable domains is challenging, and lists of unreliable domains quickly go out of date. In 2024, it was found that 50% of domains on unmaintained blocklists published between 2017 and 2019 no longer existed Carragher et al. [2024]. As our approach only uses SEO metrics and webgraph structure, the set-up is fully language-agnostic.

GNNs

For our GNN experiments, our goal is to leverage local homophily present in our partially-labeled SEO network in order to better classify unreliable and biased domains. Formally, for a given task, for website u, our goal is to create node embeddings $z_u \in \mathbb{R}^d$ that map u to its corresponding one-hot-encoded label $y_u \in \mathbb{Z}^c$.

Table 3.2: Table 2. Backlink Network Statistics for Each Think Tank Group

	Russia	**Europe**	**US**	**Pseudo**
Nodes	3,247	5,362	4,729	4,427
Edges	4,170	7,422	7,071	6,995
Total Weighted Degree	1,017,762	4,631,448	14,130,634	30,863,844
Clustering Coef.	0.006	0.002	0.12	0.2
Density	0.0008	0.0005	0.0006	0.0007

Table 3.3: Table 3. Keyphrase network statistics for each think tank group.

	Russia	**Europe**	**US**	**Pseudo**
Nodes	3,184	7,938	6,963	3,834
Edges	3,293	8,000	7,011	3,954
Total Weighted Degree	3,293	8,000	7,011	3,954
Clustering	0	0	0	0
Density	0.0006	0.0003	0.0003	0.0005

We use a simple 2-layer GNN with GraphSAGE convolutions. GraphSage is an inductive graph neural network that learns node embeddings by sampling and aggregating features from local neighborhoods Hamilton et al. [2017]. At each layer, a node updates its representation by combining its previous embedding with an aggregated representation of its neighbors.

Formally, for node v at layer k:

$$\mathbf{h}_v^{(k)} = \sigma\Big(\mathbf{W}^{(k)} \cdot \mathrm{CONCAT}\Big(\mathbf{h}_v^{(k-1)},\ \mathrm{AGGREGATE}^{(k)}\big(\{\mathbf{h}_u^{(k-1)} : u \in \mathcal{N}(v)\}\big)\Big)\Big)$$

where $\mathcal{N}(v)$ denotes the neighbors of node v, $\mathrm{AGGREGATE}^{(k)}$ is mean pooling, $\mathbf{W}^{(k)}$ is a learnable weight matrix, and σ is a nonlinear activation function.

The graph is partially labeled as many backlinks and outlinks do not have reliability or bias labeled. Consequently, unlabeled nodes are masked, but their features are still available to labeled neighbors during the propagation step. To train, we use a transductive random 80:10:10 split and 0.05 for the learning rate. We select models using early stopping with a

patience of 30 and a minimum delta of 1e-4.

Webgraph and Features

We use the domain credibility labels published by Lin et al. [2023]. The authors of the work align the domain reliability ratings of all domains ranked by 6 expert groups and run imputation followed by principal component analysis to generate aggregate ratings for 11,520 domains. While the the sites are largely English-language, there are websites spanning multiple languages and target audiences, including news sites that operate in Italian, Russian, and German. We binarize the news domain rankings as "reliable" and "unreliable" using a threshhold of 0.5162, corresponding to the bottom two quintiles of the data. Following the methodology of Carragher et al. [2024], We use the SEO toolkit service Ahrefs[7] to extract the 10 domains which link to each of the 11,520 target domains at the highest volume (the highest-volume back-linking domains). Ahrefs feature values have been found to correlate strongly with ground-truth pagerank calculations and competing traffic estimate systems Carragher et al. [2025]. In total, we extracted 43,758 domains, each with 23 domain-level attributes. Each edge in this dataset denotes a direct path, as a user on the source site could click a hyperlink to move to the reliability-labeled target site. We were unable to extract features for 193 domains—the domains largely appeared dead or inactive, so we excluded them from the data. This left us with 11,327 labeled domains.

To our knowledge, the list created in Lin et al. [2023] is the most comprehensive public domain reliability rating list as of August 2024. The domains in this list are ranked by the calculated first principal component where low scores correspond to high expert agreement on unreliability and high scores correspond to expert agreement on reliability. While these scores are very sensible across wide principal component score gaps, locally the relative orderings are less clear. Treating website reliability as continuous also muddies the interpretability and the discovery process, as there's not a transparent reason a website should have a score of 0.571 as opposed to 0.570. To address this issue, we elected to binarize the website reliability labels. After manual inspection of the data, we elected to consider the bottom two quintiles as unreliable (corresponding to a principal component threshold of 0.5162).

From Ahrefs, we were unable to pull backlinks for 193 of the domains. A random sample of 10 of 193 domains and found 9 of them were either dead or unreachable. We therefore elected to drop these 193 domains, leaving us with 11,327 unique reliability-labeled domains and 32,431 unique unlabeled backlinking domains for a total of 43,758 domains. For each of

[7] ahrefs.com

these 43,758 domains, we pull 23 attributes, which for an individual domain, contains fields like the total number of backlinks and outlinks, number of backlinks coming from .edu or .gov domains, and number of referring pages [8].

Similar to the backlink network constructed in Carragher et al. [2024], there is clear assortativity across label-reliability groups as can be seen in Figure 3.5. The prevalence of unlabeled nodes in our network overwhelms a node attribute assortativity coefficient calculation $r_A = -0.336$. We re-calculate node attribute assortativity on an induced subgraph that only maintains edges between labeled nodes. The subgraph contains less structure and 176 components, but nevertheless displays strong positive associativity $r_A = 0.376$. In other words, domains of the same (binary) reliability are more likely to link to one another than linking to a domain of a different reliability. This provides a strong justification for the use of network-based models. We perform sensitivity analyses and include thorough discussion of features, feature impact, and the impact of political bias in Carragher et al. [2024].

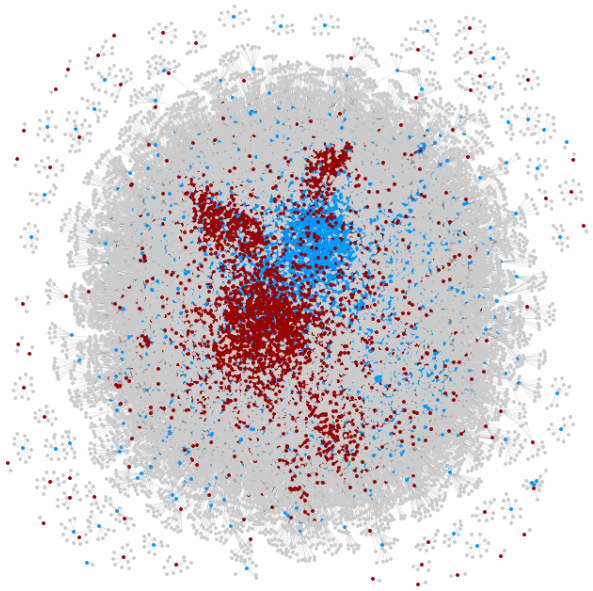


Figure 3.5: Webgraph colored by domain reliability labels. The network contains 6,861 reliable (blue) websites, 4,466 (red) unreliable websites, and 32,431 unlabeled (grey) backlinking websites.

[8] https://ahrefs.com/api/documentation/metrics

Webgraph Results

Over all 100 training runs over 80:10:10 splits of the 11,327 labeled websites, our model yielded an average accuracy of $0.7686 \pm .004$ and an average F1 of $0.7545 \pm .005$. These results are similar work we've done on smaller webgraphs Carragher et al. [2024]. Together, these present strong evidence that there is some website reliability. signal in SEO data and webgraphs. We consider more complex and heterogeneous training set-ups in Chapter 5.

3.7 Limitations

We use website reliability lists curated by third-party sources. For the subset of websites with political bias annotated, unreliable websites heavily skew right-leaning Carragher et al. [2024]. We do not know if this reflects a ground-truth internet distribution, biases of website annotators, or some combination of the two. While there are some low-reliability left-leaning websites and many websites that don't map to any clear political affiliation, right-leaning websites are overrepresented and may bias detection models. In related work, we demonstrate that partisan bias can be predicted using an almost-identical GNN architecture Carragher et al. [2024]. Further, while our GNNs are making decisions based on network structure, we do not attempt to identify structures or patterns (e.g., identifying link farms). We allow the GNNs to determine which structures are relevant.

Additionally, we do not know how Google's proprietary systems work, but the company does have patents that signal that Google uses message passing on webgraphs to label low-quality websites website representations Pennock et al. [2013] and uses embedded website representations Tsykynovskyy [2020]. There is, however, little academic work in this area, and therefore value in testing and evaluating these ideas more openly.

3.8 Conclusion

This chapter examined misinformation in web search from the perspective of strategic manipulation. By analyzing backlink and keyphrase networks across US, European, Russian, and Kremlin-linked pseudo-think tanks, I documented structural patterns consistent with coordinated amplification. Pseudo-think tanks receive disproportionate backlink volume from low-quality domains, exhibit dense internal co-amplification, and target highly specific conspiratorial keyphrases that exploit search "data voids." Although these domains are heavily

amplified, they do not achieve correspondingly high average Google positions, suggesting that search engines partially—but not fully—mitigate such manipulation attempts.

Beyond the descriptive case study, this chapter demonstrated that webgraph structure and SEO-derived features contain measurable reliability signals. Using a partially labeled webgraph of over 43,000 domains, I showed that a simple two-layer GraphSAGE model can differentiate reliable and unreliable websites with consistent performance across randomized splits. Importantly, this approach requires no textual content and is therefore language-agnostic, making it applicable across multilingual and cross-national information environments.

Taken together, these findings highlight two complementary insights. First, adversarial actors can exploit structural properties of the web—particularly link schemes and low-competition keyphrasing—to attempt to influence search visibility. Second, those same structural traces can be leveraged defensively to identify unreliable domains at scale. While SEO manipulation remains an adaptive and evolving threat, webgraph-based detection offers a promising foundation for scalable, content-independent credibility assessment.

Chapter 4

Content Manipulation: Dredge Words and Data Voids

4.1 Introduction

In May 2024, Google rolled out *AI Overviews* in the United States, which generate LLM summaries for users based on the contents of Search Engine Result Pages (SERPs). The roll out was rocky, even by Google's own admission, and only partially because users were seeing AI Overviews instructing them to eat rocks Rogers [2024]. In a Google blog post on 30 May titled "AI Overviews: About Last Week", Google attributed the problems in its AI overviews not to model hallucination, but rather to *data voids*—search results that yield unreliable or irrelevant content Reid [2024], Golebiewski and boyd [2019]. Data voids present a challenging problem for information retrieval systems, and their presence allows the spread of misinformation, propaganda, pseudoscience, bad medical advice, conspiracy, and many other forms of undesirable content. The author of the blog post, aware of the problems data voids pose, linked to a separate Google blog post about the company's efforts to warn users entering data voids by displaying content advisory banners. Independent researchers found that these banners were inconsistent and unstable, and documented their silent discontinuation months before the US 2024 presidential election Robertson et al. [2025].

Google's AI Overviews are becoming ubiquitous, and in March 2025, Semrush and Datos estimated that AI overviews appear on 13% of all U.S. desktop searches. Additionally, AI search is becoming an increasingly popular alternative to search engines, and in February 2025, 27% of Americans reported having used AI tools in the place of a traditional search engine Rowlands [2025]. These AI overviews can be very helpful for users as they save time

by retrieving and synthesizing information from multiple different webpages. However, many audits conducted by SEO companies are finding that AI Overviews are taking web-traffic away from webpages. In other words, after a user enters a query into Google, the user might read the AI overview and terminate the session rather than clicking on the links that the AI overview retrieves. This harms the revenue of websites that serve ads.

In this work we provide the first large-scale, open-domain audit of Gemini's AI Overviews on Google search. We scraped Google search results and AI overviews for over 4.1M unique queries. Over the course of the collection, we captured AI overviews that contained a wide range of conspiracies and pseudoscientific statements: including promoting pseudo-scientific wellness practices and products, promoting the idea that extraterrestrial aliens are influencing human society, and presenting spiritual/ psychic beliefs as factual. We outline common failure modes we identified in AI overviews. We then systematically evaluate the extent to which "information gaps" or "data voids" are responsible for pseudoscientific AI Overiviews and identify the domains most responsible for pseudoscientific RAG responses.

4.2 Chapter Research Questions

RQ4.1 What are properties of dredge words and how are these different from keyphrases of reliable websites?

RQ4.2 How do dredge word SERPs impact the reliability of AI Overviews like Google Gemini?

4.3 Related Works

Retrieval-Augmented Generation (RAG) models are often framed as a solution to hallucination: by grounding responses in an external corpus, they allow generated claims to be traced back to specific documents. However, grounding does not completely eliminate error. Prior work has highlighted the risk of "data pollution," where retrieved documents contain mixtures of true and false statements Pan et al. [2023]. In such cases, generation is constrained by the quality of the corpus.

A large body of work attempts to improve LLM reliability through retrieval or to improve the reliability of RAG outputs directly. Most of this research implicitly assumes that the majority of retrieved documents are themselves reliable. For example, Retrieval-Augmented Correction (RAC) decomposes model outputs into atomic claims and attempts to revise

them using retrieved evidence Li and Flanigan [2024]. This approach, like much of the RAG reliability literature, presumes that the retrieved evidence is trustworthy enough to serve as a corrective signal. However, when retrieved documents are erroneous, outdated, or low-quality, these methods may fail.

Instead of addressing credibility directly, many papers focus on retrieval performance in long-tail or low-coverage settings. These works examine what happens when relevant knowledge is rare or buried in large corpora. For instance, Laban et al. [2024] construct "needle-in-a-haystack" conditions in which relevant context is overwhelmed by unrelated information. Similarly, Li et al. [2024] explicitly model long-tail knowledge and modify retrieval to surface it more effectively. These approaches improve recall under sparse conditions, but they primarily treat the problem as one of coverage rather than one of unreliable evidence.

Other research focuses on improving the quality of retrieved content through better ranking and data quality integration. RankRAG incorporates ranking signals into the generation process and achieves strong benchmark performance Yu et al. [2024]. Related work emphasizes the interaction between data quality frameworks and RAG systems Müller et al. [2025]. While these approaches improve relevance and coherence, they do not directly address scenarios in which relevant documents themselves are misleading or low-credibility.

Accepting that RAG systems are constrained by the corpora they access, some researchers instead evaluate how LLMs behave when retrieved context is flawed. Pan et al. [2024] introduce benchmarks that test model performance under inaccurate or misleading retrieval. A related line of work examines how LLMs handle knowledge conflicts Xu et al. [2024], estimate source reliability Hwang et al. [2025], handle knowledge conflicts Shen et al. [2025], and perform source attribution Wang et al. [2025]. These efforts treat unreliable context as a robustness challenge, but typically do not model environments in which unreliability is pervasive.

Finally, multiple studies demonstrate that deliberate data poisoning—injecting adversarial or false content into retrievable corpora—can significantly degrade RAG performance Zou et al. [2025]. While document poisoning is well studied in controlled settings, many real-world open-domain environments present a more ambiguous problem. Corpora contain not only outright falsehoods, but also outdated material, satire, advertisements, propaganda, user-generated content, and spirituality. In these cases, the boundary between factual and nonfactual is not always clear, and the assumption that most relevant evidence is reliable may not hold.

To our knowledge, there is limited work that directly examines RAG behavior when the

relevant portion of a corpus is systematically low-credibility. Our work addresses this setting explicitly, focusing on "data voids" where retrieval surfaces evidence that is relevant but often unreliable.

4.4 Data

Collecting search engine data has long been a challenge in audit studies, and most audit studies rely on a set of keywords hand-chosen by researchers. While this approach can provide useful insights on the predefined topics selected by researchers, the approach does not scale well and are necessarily limited in scope. In recent years, several researchers have explored ways of scaling search engine data collection. The data collection is largely inspired by the data collection processed proposed in Williams et al. [2025]. In the paper, I propose the concept of *Dredge Words*—queries for which unreliable websites rank highly on search engines—and use the proprietary SEO toolkit Ahrefs[1] to extract 1,000 organic keywords for which a set of unreliable websites rank most highly on Google Search. I'll discuss this concept in more detail in Chapter 5. We follow this process to extract organic Keywords 11,520 reliability-labeled websites in Lin et al. [2023]. Each of these websites has a first principal component (PC1) reliability score between 0 and 1 that was calculated by aggregating website reliability ratings from multiple sources. Reuters, Associated Press, Nature, and Smithsonean Magazine have some of the highest reliability scores in the dataset (closest to 1) whereas websites like "naturalnews.com" and "infowars.com" have some of the lowest reliability scores.

For each website in this list, using Ahrefs, we extract 1,000 organic keywords in the previous 30 days from the collection date sorted by SERP ranking (i.e., prioritizing keywords that appear highest on SERPs) in the United States. For websites with more than 1,000 organic keywords in the first position, we randomly sample 1,000. For websites with fewer than 1,000 organic keywords, we collect all available organic keywords. In line with previous work discussed how quickly website reliability lists become stale or obsolete Carragher et al. [2024], only 6,487 of the labeled domains had organic keywords on Google in the past month. From this set of 6,544 reliability-labeled domains, we extracted 4.88M keywords (4.16M unique). Keywords were collected between from Ahrefs between September 2024 and January 2025.

Ahrefs estimates that the target websites receive a monthly organic traffic of 128M over all target websites, plus an additional 90k in paid traffic. The estimated organic monthly

[1] Ahrefs.com

traffic for "dredge words", keywords extracted from domains with domain reliability scores of less than 0.53 (see Chapter 5 for cutoff rationale), is 24M. These monthly traffic estimates were made at the time of collection, so traffic estimates have likely changed.

Over this collection strategy, we collected 260,714 AI Overviews (≈6% of all scraped SERPs), 258k of which were unique. Duplicated AI Overviews were most often a result of Google modifying queries when it detects a Typo. Our query dataset contained 80 different misspellings of "millennial", all of which were shown the search results for millennial. Given that the AI overview was the same for each of these 80 misspellings, it is likely at least some AI Overviews are cached. For each of these AI Overviews, we scraped all sources that they referenced. We additionally scraped the first 10 search results below each AI overview. We focus our analysis on these 258k AI Overviews.

4.5 Methods

4.5.1 Model-as-Judge

To identify AI Overviews containing pseudoscience and medical misinformation without sufficient context, we employ a LLM-as-judge approach using gpt-5-mini to identify unreliable AI Overviews. A random sample of 20 AI Overviews rated as medical misinformation and an additional 20 rated as pseudo-scientific were annotated by two independent human annotators. On medical misinformation queries, human annotators had substantial agreement (Krippendorf $\alpha = 0.84$). For pseudoscience, annotators were in total agreement that all surfaced documents contained pseudoscience, but only moderately agreed on whether enough context was provided (Krippendorf's $\alpha = 0.44$).

4.5.2 Leave-One-Out Retrieval Experiments

To estimate the marginal contribution of individual domains to response quality, we conducted leave-one-out (LOO) experiments under a controlled RAG setup. For all AI Overviews labeled as pseudoscientific or medical misinformation, we scraped the full text of each cited webpage. For YouTube citations, we retrieved transcripts using the YouTube API. Using these documents, we constructed a retrieval-augmented pipeline with `Llama-3.2-3B-Instruct`[2].

For each query, we generated RAG outputs while excluding all documents from a single target domain. This procedure was applied to the nine most frequently cited domains

[2]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

(excluding one domain for which scraping was incomplete). All outputs were labeled using the same `gpt-5-mini` annotation schema. We report percentage reductions in pseudoscientific and medical-misinformation responses relative to Gemini's original retrieval.

4.6 Results

In this section, we analyze the $n = 225{,}701$ AI Overviews scraped from Google search. The LLM-as-judge approach found only 1.37% contained pseudoscience and 0.14% contained medical misinformation under our LLM-as-judge annotation schema.

A core motivation for this research has been to evaluate how data voids impact RAGs, and whether poor SERP quality is associated with poor Google AI overview quality. In this section, we will comprehensively examine the relationship between SERPs and AI Overviews.

Somewhat surprisingly, overlap between AIO and SERP domain sets is limited. Averaged across queries, 26.78% of AIO domains also appear in SERPs ($n = 260{,}714$), while, on average, 33.33% of SERP domains appear in AI Overviews ($n = 126{,}850$). The mean per-query Jaccard similarity between AIO and SERP domain sets is 12.20% ($n = 260{,}714$).

4.6.1 SERPs tend to be more reliable than AIO sources

Despite only moderate overlap in domains, low-reliability exposure is strongly associated for low-reliability websites. Using bottom-20% floor thresholds for SERPs and AIOs (SERP threshold 0.315; AIO threshold 0.424), the Pearson correlation between SERP and AIO floors (minimum PC1) is $r = 0.240$. The odds ratio for observing a low AIO floor given a low SERP floor is 2.30 (95% CI: $[2.24, 2.37]$; $\chi^2(1) = 3330.65$, $p < 0.01$). 60.29% of queries with low SERP floors also have low AIO floors, compared to 39.75% when SERP floors are not low. This suggests that low-reliability data voids are indeed associated with low-reliability AI Overviews.

At the query level, we compare ($110k$ paired queries), where both at least one SERP domain and at least one AIO domain have a known PC1 score. We obseve that domains cited in AI Overviews exhibit lower average reliability than those appearing in SERPs. Mean reliability (PC1) is 0.685 for AIO domains and 0.711 for SERP domains, yielding a mean difference (AIO − SERP) of -0.026 with a 95% bootstrap confidence interval of $[-0.027, -0.025]$. A paired test on this restricted set yields $t = -59.99$, $p \ll 0.001$. We visualize the distribution difference in Figure 4.1. However, this can largely be explained by

the heavy inclusion of of YouTube in AI Overviews. When YouTube is removed, the mean AI Overview reliability (0.73) is slightly higher than SERPs (0.72).

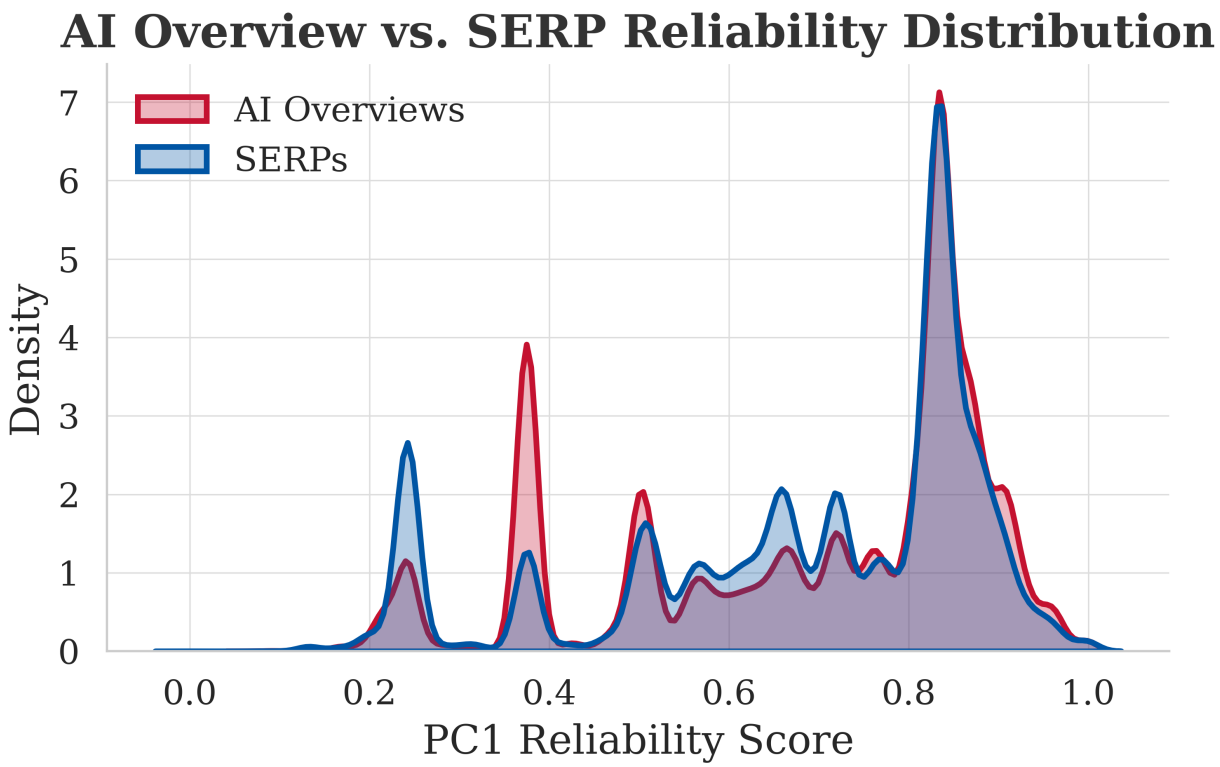


Figure 4.1: Distribution of reliability scores (PC1) for AI Overviews and SERP sources. The red spike at 0.4 is largely driven by AIOs including YouTube more frequently than SERPs.

4.6.2 Observed AI Overview Failure Modes

We employed an LLM-as-judge framework, using GPT-5-nano as the evaluator, to assess whether AI Overviews promoted medical misinformation or pseudoscience without providing adequate contextualization or caveats. Across the corpus, we identified 3,288 AI Overviews that promoted pseudoscientific claims without sufficient context and 337 that promoted medical misinformation. In what follows, we examine the qualitative characteristics of these AI Overviews and their corresponding SERPs, with particular attention to recurring thematic patterns and framing strategies. We note that these categories are overlapping, and single AI Overviews often fall into multiple failure modes.

Medical Misinformation

In the medical misinformation responses we often saw AI Overviews promote fad diets, including promotions of "detoxification" (28 AIOs) and fad diets. The "alkaline diet," which has been promoted in some online communities as a cancer treatment despite lack of scientific support, was referenced in 16 AIOs while promoting pseudoscientific claims. For example: *Barley is alkaline, which means it can help reduce acidity in the body.* Another overview stated: *Some say that a balanced diet includes both acidic and alkaline foods. The body can manage neutral foods and not overly extreme acid or alkaline energy.* In response to the query "what does salt lemon water do," one subsection read: *Detoxification: Some believe that the combination can help detoxify the body by pulling out toxins, although more research is needed to support this claim.* While mild hedging language is occasionally present, these AI Overviews do not provide sufficient scientific context.

In several instances, user queries explicitly sought information aligned with non-evidence-based practices, and the AI Overviews often provided affirmative descriptions without substantive qualification. For the query "yoga for liver detox," the response began: *Yoga poses that may help detoxify the liver include the bridge pose, boat pose, cobra pose, and twists.* Similarly, the AI Overview for "crystals good for migraines" opened with: *Some crystals that are believed to help with migraines include amethyst, selenite, rose quartz, carnelian, lepidolite, moonstone, lapis lazuli, sodalite, and fluorite,* and further advised readers to *Place amethyst on your forehead for 20 minutes to ease tension and anxiety.* In both cases, AI Overview responses promote remedies not supported by science.

AI Overviews around crystals often attributed physiological or psychological effects to specific stones. For the query "what are the properties of red jasper," the response asserted: *Red Jasper is a grounding stone known for its protective and healing properties. It's associated with strength, courage, and passion, and is believed to stimulate the root chakra. Red Jasper also promotes emotional and physical balance, enhances the immune system, and supports digestion.* Likewise, for "what does amethyst crystal do for you," the overview stated: *Amethyst is believed to offer a range of benefits, including calming the mind, reducing stress, and promoting restful sleep. It's also associated with enhancing intuition, clarity of thought, and spiritual awareness. Additionally, amethyst is thought to boost the immune system and improve skin health.* These responses consistently present metaphysical belief systems alongside biomedical-style claims, without clearly delineating the epistemic status of each.

Other examples reflect a similar pattern of presenting alternative health claims in an authoritative tone. One overview stated: *Ginger tincture, made by extracting ginger's*

beneficial compounds into alcohol, offers a range of health benefits, particularly for digestion, inflammation, and immune support. It can help with nausea, reduce inflammation, and potentially aid in blood sugar and cholesterol management. Another AI overview claims that *Tualang honey can help prevent diseases like cancer, coronary disease, and neurological degeneration.* Although some of these claims have varying degrees of empirical support, these responses do not differentiate more speculative claims or provide any weighting for that evidence. In some cases, secondary sources were cited as evidence of scientific support without linking directly to primary research. For the query "tumeric vs nsaids," the overview claimed *Some studies suggest that turmeric may be as effective as NSAIDs for pain and function*, but linked to a commercial sports medicine webpage rather than to the underlying trials. The cited page referenced "clinical studies" without providing any actual citations.

Pseudoscience

The majority of pseudoscientific AI Overviews we observed promote metaphysical or new-age practices, often including dubious claims alongside them. For example, in response to the query "cocoon meditation," the benefits section stated: ***Benefits*** *You may experience a dream-like consciousness You may receive messages from ancestors and the world beyond your ego You may harmonize your emotional body You may integrate your physical, mental, emotional, and spiritual energy bodies You may enhance your healing process.* Similarly, an overview of psychic abilities asserted: *The four clairs are clairvoyance, clairaudience, clairsentience, and claircognizance. They are considered to be the four main psychic gifts, but there are many other psychic abilities [...] These abilities can help people develop a balanced and holistic approach to personal and spiritual growth.* Regurgitating new-age beliefs without context was the most common reason AI Overviews were classified as psuedoscientific. For example, the term "angel number", defined by Google's AI Overviews in March 2026, as *repeating, synchronized, or patterned number sequences (e.g., 111, 444, 1234) believed to be messages from the spiritual universe, guardian angels, or passed loved ones*, appeared in 702 of the 3.3k psuedoscientific AI Overviews. We explore other failure modes of AI Overviews in the rest of this section.

Marketing as Truth

A distinct pattern involves reproducing marketing claims in an informational tone, often drawing on manufacturer websites, commercial blogs, or user reviews. For the query "wah

tah water," the overview stated: *Wat-aah is a brand of water that is marketed as a functional water that can help with energy, hydration, and detoxification* [...] *Wat-aah is said to be polarized, which may help with cellular transport of nutrients and energy* [...] *Wat-aah is vapor distilled and contains oxygen.* Although the overview correctly attributes certain claims to marketing, it does not contextualize or correct the pseudo-scientific implications surrounding "detoxification" or "polarization."

A similar pattern appeared in product reviews. Searching "dom pen review" returned: *Dom Pen, a brand known for disposable cannabis vapes, generally receives positive reviews for its discreet design, ease of use, and range of flavors, with some users noting its effectiveness for pain and anxiety relief.* The next line read: *The Dom Pen's slim profile and lack of buttons make it ideal for stealth vaping.* The remainder of the overview summarized positive product feedback across multiple flavors.

Conspirituality

In some responses, spiritual and conspiratorial themes are intertwined. For example: *Mount Shasta is a place shrouded in legends and mysteries, with claims of hidden underground cities, UFO activity, and connections to ancient civilizations, particularly Lemuria and Telos.* For the query "were pyramids shiny", the AI Overview returns *The pyramids were covered in a layer of white limestone that was polished to a smooth finish. The limestone casing made the pyramids sparkle like diamonds in the sun. The pyramids were originally part of a magical port city that was bathed in sunlight.* Other responses re-articulate theological or supernatural claims without sufficient religious contextual grounding. For example *"888 Lions Gate" refers to a period of heightened spiritual and manifestation energy centered around the annual Lion's Gate Portal, which occurs on August 8th.* These spiritual AI Overviews can veer in conspiratorial directions: *"Spiritual wickedness in high places," from Ephesians 6:12, refers to demonic forces and their influence in the supernatural realm, not physical conflict. It symbolizes powerful evil entities such as rulers, authorities, and spiritual forces of darkness operating in heavenly or spiritual spheres to corrupt and influence individuals and systems on Earth, including governments and institutions.* We would argue such claims made in AI Overviews should be presented alongside, at a minimum, religious or cosmological context (e.g., "Within new age belief systems...".

Information Laundering

Many AI overviews hide attribution of dubious claims by using language like "is believed" or "some say". This can launder unsupported claims without clear attribution. For example, *Sirians are beings from the Sirius star system, and some believe they have influenced human civilization.* The phrase "who some believe" obfuscates the dubious origin of the claim. Other AI Overview responses recount extraordinary allegations without situating them within broader reporting or skepticism. For example: *Claims of "alien attacks" and alleged encounters with "armored aliens" have surfaced in rural Peru, with villagers describing beings with Green Goblin-like features, long heads, and yellowish eyes. One villager even claimed to have shot at one of these beings, which then elevated further and disappeared.* In other cases, AI overviews can blur the line between summary and promotion. For example, when searching a conspiratorial documentary: *Documentary Focus: The documentary, "Bermuda Triangle Revealed: The Devil's Graveyards," investigates 11 "vile vortices" or paranormal regions beyond the well-known Bermuda Triangle. Unexplained Events: These regions are characterized by bizarre animal attacks, unexplained weather patterns, and passenger jets disappearing, among other unusual occurrences.*

Accepting the Premise

It's challenging to delineate when a AI Overviews should aim to correct or push back against user premises. In the query "best crystal for migraine", where there could be medically-harmful consequences of providing a user with bad information, it seems clear that AI overviews should at least in part, refute the user's premise. However, it becomes trickier when spirituality is treated as the premise. For the query, "what sign does libra fall in love with", Google's AI Overview begins with *Libras are typically most compatible with other air signs, like Gemini and Aquarius, as well as fire signs, like Aries and Leo.* The AI Overview does not mention astrology or provide scientific refutation of astrological beliefs, but in this setting, it's also not clear that the AI Overview would need to do so. However, even seemingly innocuous queries can escalate into more extreme pseudoscientific beliefs. The query "sagittarius evil superpower" generated an AI Overview that begins: *A Sagittarius evil superpower could be interpreted as a Sagittarius-driven ability that, while seemingly positive on the surface, could be used for malicious purposes. Given Sagittarius's love for freedom and adventure, one such ability might be Teleportation or Time Travel, allowing them to quickly escape consequences or manipulate situations for personal gain.* The AI Overview takes adds teleportation and

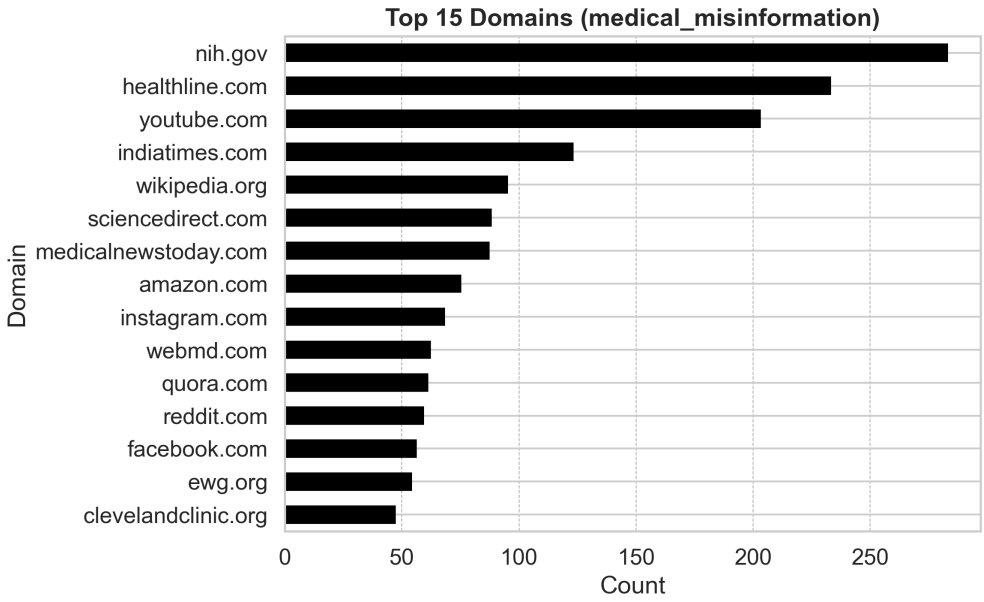
time travel to an astrological query.

Failing to Reconcile Sources

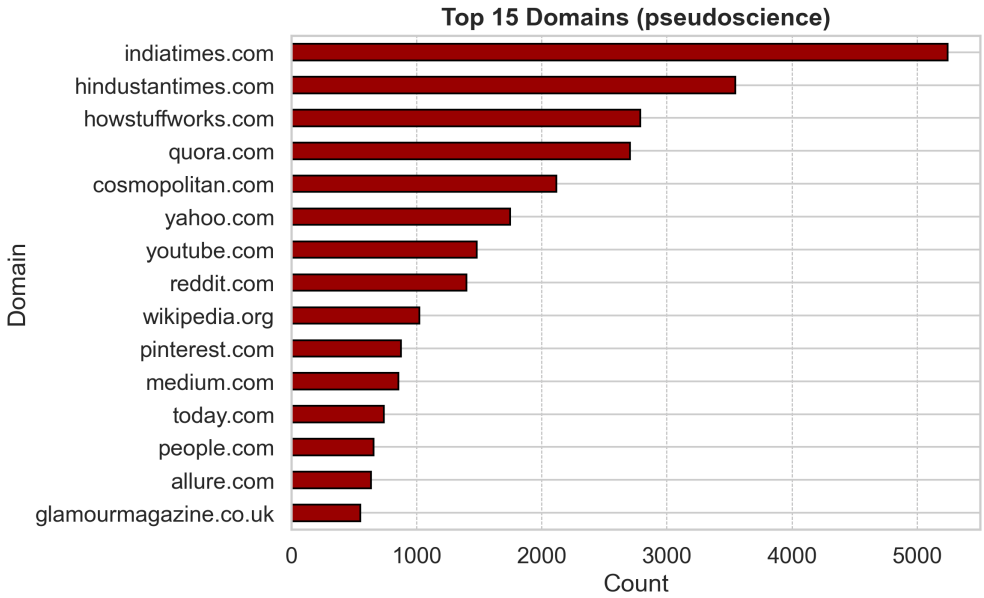
Finally, we observe instances in which disparate retrieved fragments are combined into claims not directly supported by any single source. In one instance, an individual is claimed to be *a prominent figure in both the worlds of policy research and astrology/psychic readings*, when in reality these are two different people with an identical name. An additional AI Overview responding to a query about a Japanese Mukbang claimed that the influencer: *has reportedly been diagnosed with bipolar disorder, which experts suggest can be triggered by excessive eating habits*. No returned source states that "experts" attribute bipolar disorder to excessive eating. However, one source mentions the influencer's bipolar disorder, and another notes that online harassment around eating "triggered" depression. The overview appears to synthesize these elements into a novel, unsupported causal claim.

4.6.3 Properties of Dubious AI Overviews

In Figure 4.2, we display the most frequently cited 15 domains for AIOs containing pseudoscience and medical misinformation. Indiatimes features highly in both, as does youtube, reddit, and quora. Indiatimes and hindustantimes were used as a sources in many AI overviews around astrology and new-age mysticism. We present the domains most strongly associated with pseudoscience in Figure 4.3. The top websites all contain substantial amounts of content devoted to astrology, numerology, wellness, and new-age beliefs. Otterspirit and Wyldemoon both heavily promote crystal healing, the former, in part, to sell crystals.



(a) Most frequently cited domains in AI Overviews promoting medical-misinformation



(b) Most frequently cited domains in AI Overviews promoting pseudoscience

Figure 4.2: Domain distributions by type.

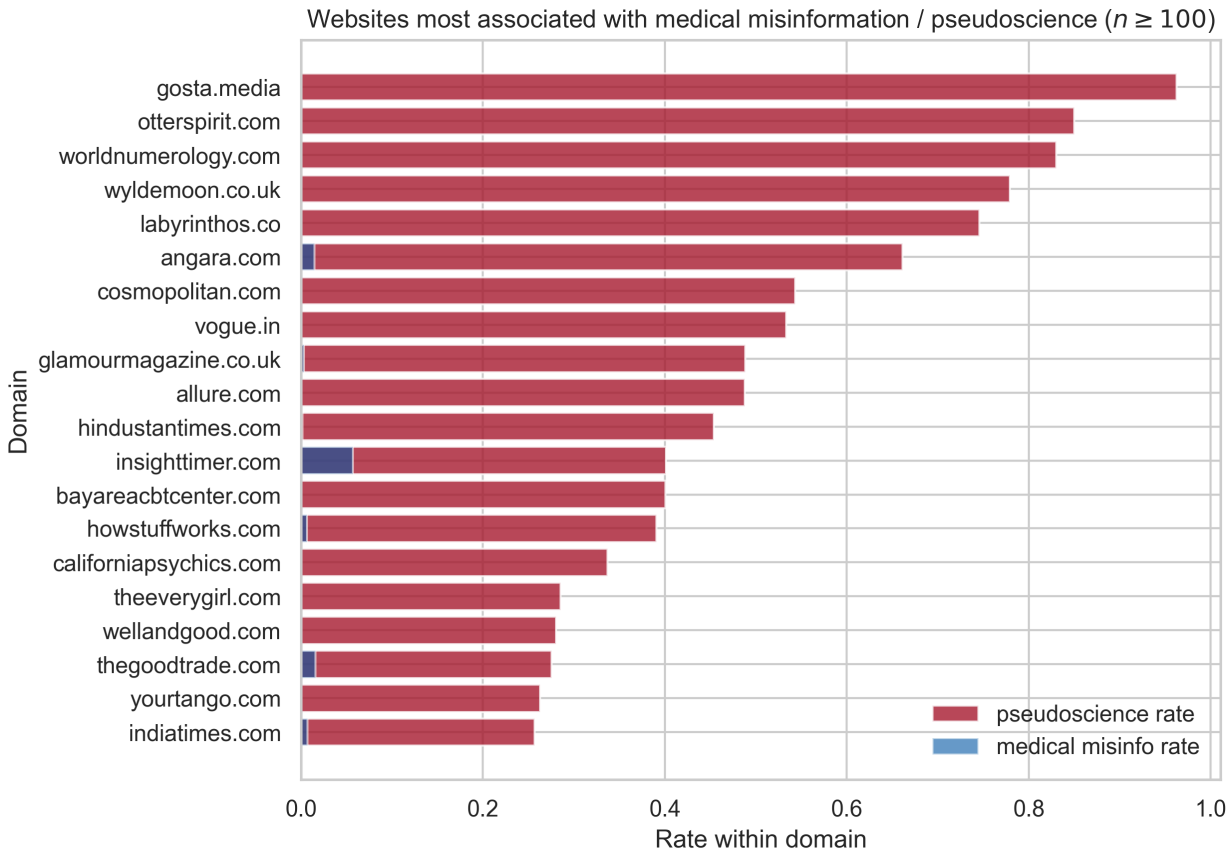


Figure 4.3: Domains most strongly associated with Pseudo-science AI Overviews($n \geq 100$)

If AI Overviews cited at least one low-reliability domain, they were substantially more likely to return pseudoscience or medical misinformation. For each threshold percentile, AI Overview rows are split into low ($\texttt{pc1\_min} \leq t$) and high ($> t$) groups, then a 2×2 table is formed per category and converted to an odds ratio. In Figure 4.4, we show the odds ratio of returning pseudoscience or medical misinformation are highest when at least one extremely low-reliability websites is present in AIO sources.

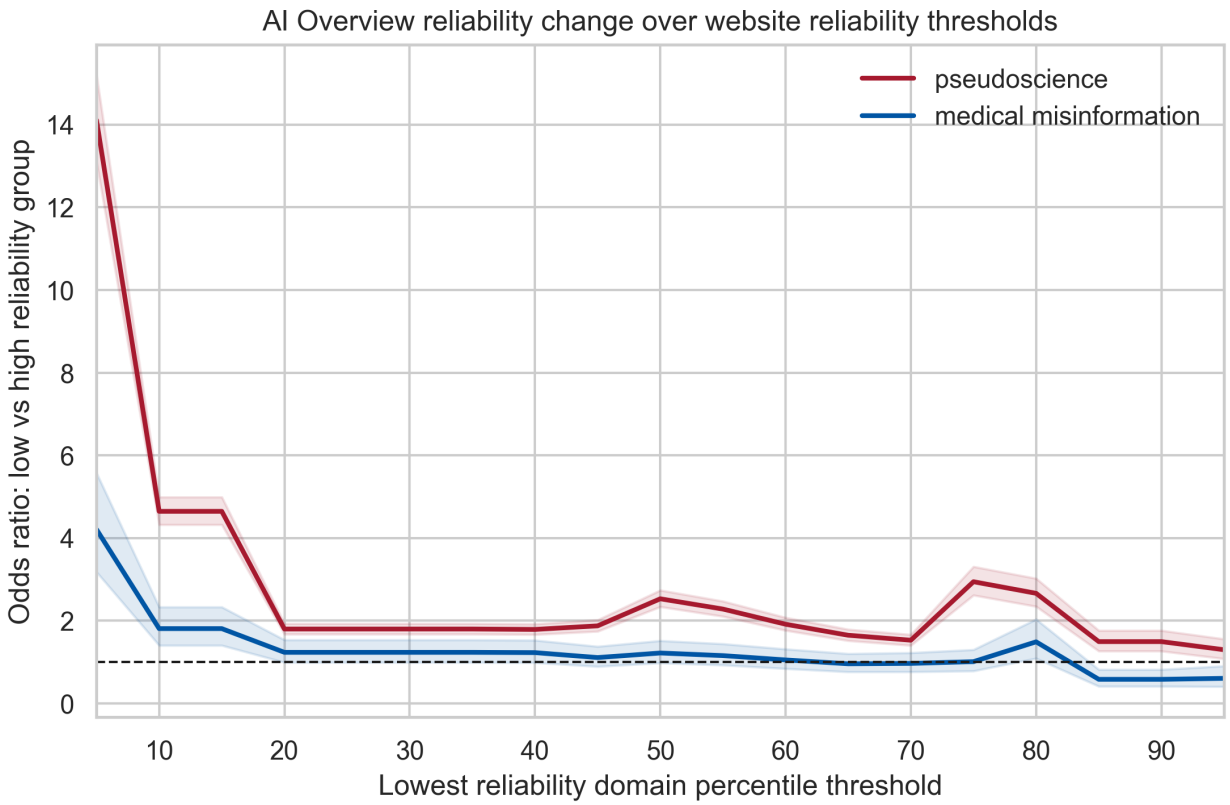


Figure 4.4: AI Overview reliability change over source website reliability thresholds.

There were 793 pseudo-scientific domains for which we had at least one reliability-labeled website in both the AI Overview and in Google's first 10 search results. Of the 793, 261 returned the same or approximately the same low-reliability website (having a PC1 difference of less than 0.02). However, SERPs contained the least reliable website more often (307) than AI Overviews (225). However, pseudo-scientific AI Overviews tended to have lower mean PC1 reliability (448) compared to SERPs (234).

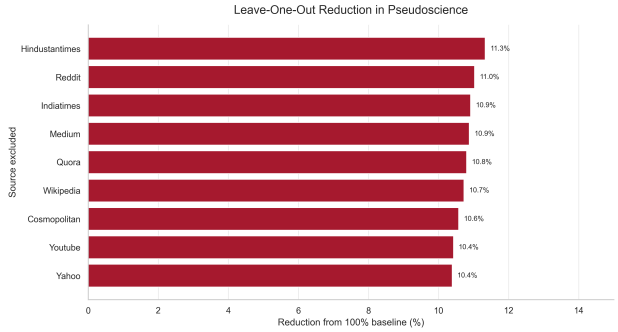
4.7 Results

4.7.1 Leave One Out

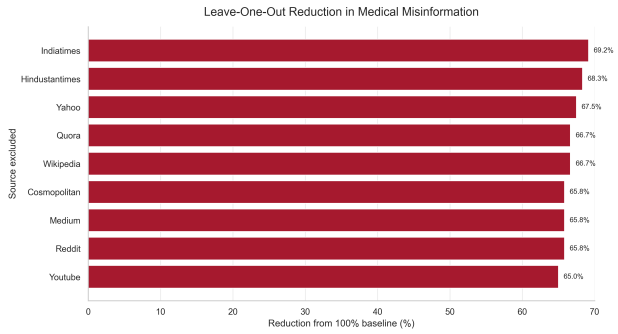
A natural follow-up question, we aim to explore to what extent are retrieved documents responsible for Google returning dubious content and identify the domains most impacting AI Overview reliability. We scraped text from all of the webpages cited in AI overviews promoting pseudoscience and medical misinformation. For YouTube, we collected video transcripts

through the Youtube's API. We run several experiments with `Llama-3.2-3B-Instruct` to better understand how these sources might be impacting AI Overview quality. In the first set of experiments, we attempt to set up a RAG similar to Gemini: we provide the original query and all documents that Gemini retrieved, but we exclude all documents from a single source (leave-one-out). We run this over all dubious queries, and then use gpt5-mini as a judge using the same annotation schema as for the original queries.

For each of the 9 most frequently cited sources, (we exclude howstuffworks.com due to scraping issues). For each of the 9 sources, we exclude the target source and measure the change in dubious responses (Figure 4.5). For pseudoscientific responses, removing Hindustan Times resulted in the largest reduction in pseudoscientific responses (11.3%) whereas remotving YouTube resulted in the most modest drop (10.4%). For medical misinformation, removing Indiatimes resulted in a 69.2% drop whereas removing YouTube resulted in the smallest 65% drop. We note that removing Hindustantimes and Indiatimes are associated with some of the largest reductions in medical misinformation and pseudoscience promotion.



(a) Leave-one out reduction in pseudoscience



(b) Leave-one out reduction in medical misinformation

Figure 4.5: Leave-one out reductions in medical misinformation and pseudoscience. Removing YouTube resulted in smallest reduction in both settings. Removing Indiatimes and Hindustantimes resulted in some of the largest improvements.

4.7.2 Parametric vs. Retrieved Knowledge

As we observed in the Leave One Out experiments, retrieved knowledge has some impact on the factuality of responses. We evaluate Llama-3.2-3B-Instruct on all documents retrieved by Gemini and found approximately 80% of the responses promoted pseudoscience or medical misinformation without sufficient context (Figure 4.6). We repeated this process without providing any documents (i.e. forcing the model to rely on parametric memory), and found over a 10 point drop in unreliable responses. We then repeated these two experiments with an additional line within the model prompt: "you are a scientific debunker". This resulted in a nearly 65 point decrease in the amount of pseudoscientific or medical misinformation responses ('SciPrompt Retrieved Knowledge'). Again performance improved with this prompt when documents were excluded, suggesting RAGs might, in cases of unreliable retrieval, benefit from learning when to ignore retrieved documents.

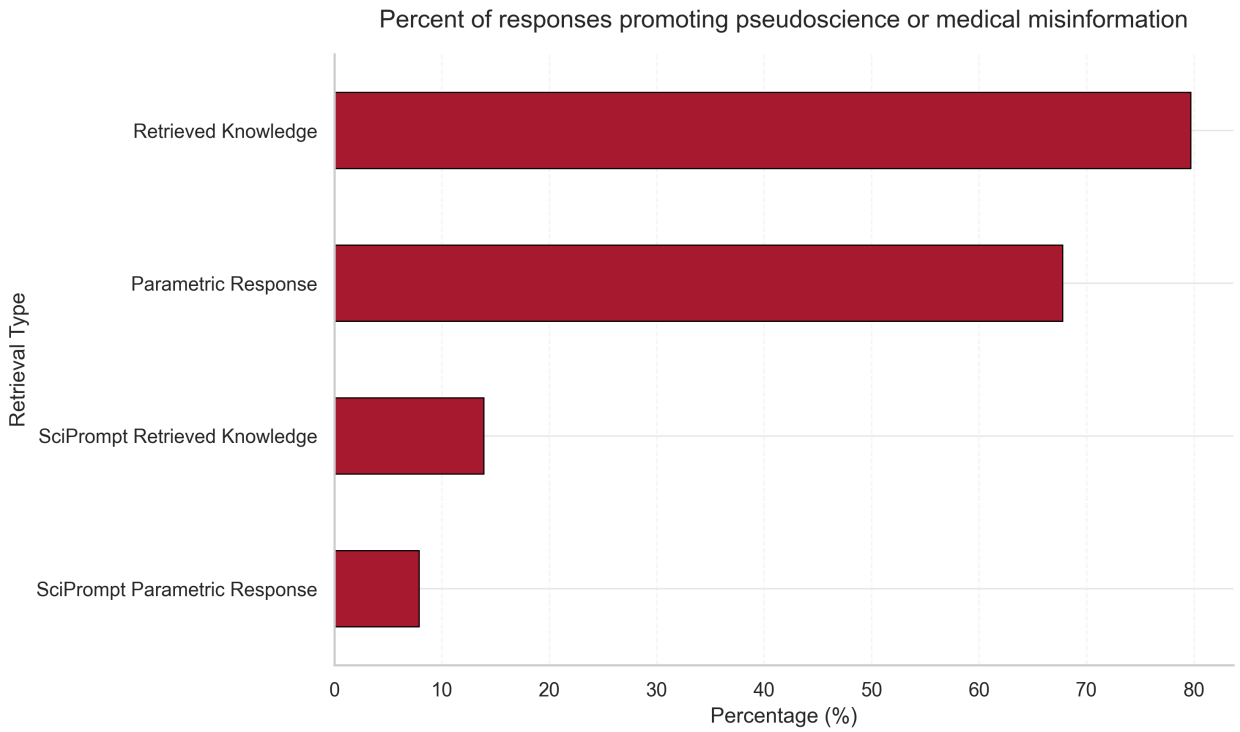


Figure 4.6: Comparisons of retrieved knowledge and parametric knowledge for

4.8 Discussion

There are a number of ways Google or other search engines could intervene to improve its AI Overviews. Many problematic AI Overviews contain forms of hedging, e.g., "some say", but simply banning hedging doesn't overcome the core problem and might make AI Overviews return dubious information with even more confidence. A better approach would be to identify unreliable sources and remove them from retrieval databases. One could also explore LLM-as-judge approaches for evaluating weak or conflicting evidence cited in retrievals.

I further note that categories like "pseudoscience" and "medical misinformation" are not culturally neutral, and some examples may instead reflect knowledge appropriation from non-Western or local practices rather than inherently false claims. Future work should make a clearer distinction between whether a claim is supported by reliable scientific evidence and whether its presentation in AI Overviews could plausibly lead to unsafe actions (e.g., delaying treatment or substituting unproven remedies). However, I would argue that Google should, at a minimum, provide adequate context and disclaimers when claims contradict scientific consensus.

4.9 Conclusion

This paper examined the reliability of Google AI Overviews at scale using a corpus of 225,701 scraped responses, with particular attention to pseudoscience, medical misinformation, and the role of low-quality retrieval. We find dubious outputs were relatively uncommon: 1.37% of AI Overviews promoted pseudoscience and 0.14% promoted medical misinformation without sufficient contextualization. At the same time, these failures were not random. We find that low-reliability SERPs are associated with lower-reliability AI Overview source sets, demonstrating that data voids continue to impact RAG systems. My Leave-one-out experiments and retrieval vs. parametric knowledge experiments demonstrate that in some instances, RAGs are more reliable without retrieval. In future work, I will explore fine-tuning and post-training approaches to help RAGs determine when it is best to use context and when it is best to rely on parametric memory.

Chapter 5

Bridging Social Media and Search Engines

In chapters 2, 3, and 4, we considered ways that attackers can manipulate search results. Conceptually, we can subdivide the approaches attackers take into two broad buckets: attackers can manipulate the structure of webgraphs (e.g. link scheming), or attackers can optimize for low-competition content. I call these two frames "structural authority-based strategies" and "content-based" strategies. In Chapter 3, we primarily focused on structural authority based strategies, and demonstrated the strength of signals of webgraphs alone in predicting website reliability. In chapter 4 and to a lesser extent 3, we focused on content-based strategies employed by low-reliability websites. In this chapter, we attempt to combine these two approaches and present a realistic pathway through which users might access unreliable websites. We combine these signals into a single model to more accurately model user paths to unreliable websites.

5.1 Chapter Research Questions

RQ5.1 What are the paths users take to unreliable websites?

RQ5.2 How can the paths users take to unreliable websites be incorporated into models?

RQ5.3 Can Twitter, SERP, and dredge word signals improve the accuracy of proposed webgraph-based reliability-detection systems?

5.2 Proposed Work

5.3 Introduction

On February 24, 2022, the day Russia began its invasion of Ukraine, Jacob Creech, a fringe QAnon conspiracy theorist, posted a series of unsubstantiated claims on Twitter. These tweets insinuated that the U.S. had created COVID-19 and implied that Russia's invasion was actually an effort to shut down U.S.-funded "biolabs" in Ukraine and prevent another global pandemic League [2022]. Hours after these tweets, the conspiratorial website InfoWars published an article promoting them, crediting Creech for uncovering an "ulterior motive theory" League [2022]. The tweets soon circulated to other conspiratorial sites, and search interest in "U.S. biolabs" and "Ukraine biolabs" spiked in the following days [1]. Conspiratorial sites were likely the predominant search results until Snopes debunked the claim later that day Evon [2022].

In some ways, this represents a success of the current misinformation response paradigm—fact-checkers acted swiftly to address a rapidly spreading falsehood. However, it also highlights the limitations of reactive misinformation interventions. Despite the swift debunking by fact-checkers, the conspiracy theory was still echoed by established news sources like Fox News and later amplified on the floor of the U.S. Senate League [2022]. Researchers later found that a coordinated network of social media accounts helped boost the narrative on Twitter Alieva et al. [2022]. As of 2024, even though fact-checkers have repeatedly debunked the underlying claims Chappell and Yousef [2022], Parachini [2022], a search engine audit study found that Google, Bing, and Yandex still return first-page results that promote the "Ukraine biolabs" conspiracy theory Kuznetsova et al. [2024]. This example highlights an

[1] https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=Ukraine%20biolabs&hl=en

important and understudied aspect of misinformation: the interaction between social media and search engines.

An alternative to reactive fact-checking is proactive algorithmic content moderation. This can involve modifying recommendation and ranking systems to reduce the reach and virality of unreliable information sources. In search engines, this might mean downranking articles from unreliable domains, while on social media, it could mean ranking posts containing unreliable links lower in users' newsfeeds. For proactive approaches to work, platforms need systems that can identify unreliable domains. Both classification and discovery are essential as unreliable websites can and do employ tactics to evade blacklists Carragher et al. [2024]. In this work, we present a novel approach to unreliable domain detection and discovery that leverages signals from both large-scale social media data and webgraph data.

Little is understood about how users transition from social media to search engines to access unreliable websites. However, it is known that directing users to search engines can be an effective tactic for spreading misinformation. Since 2016, global surveys have consistently found that individuals trust search engines more than traditional media Barometer [2024], McDuling [2015]. Thus by directing users to search engines, either explicitly or implicitly, bad actors can foster a false sense of content reliability. Research has only recently begun exploring these pathways, and some helpful concepts like *problematic queries* Golebiewski and boyd [2019], *data voids* Golebiewski and boyd [2019], and *keyword signaling* Tripodi [2019a], have been examined in case studies. The recently proposed concept of *search directives* offers a clearer and more observable path by which users are directed to unreliable content. A search directive is content explicitly intended to prompt an online search, such as user a telling user b to "look up Chemtrails on Google" Robertson et al. [2023a].

However, search directives have two key limitations as content moderation tools. First, while they can lead users to unreliable content, they can also direct them to neutral or beneficial information, such as song lyrics or mathematical theorems Robertson et al. [2025]. Second, as seen in the "Ukraine biolabs" case, users can be driven to unreliable content through search engines even without being explicitly told to search something. To overcome these limitations, we propose the concept of *dredge words*—terms or keyphrases for which unreliable domains rank highly in search results.

By attempting to explicitly incorporate each of the paths that users take to unreliable websites into GNNs, we seek to explore the dynamics that connect the spread of unreliable content on social media and search engines. We demonstrate that our relatively-simple curriculum-based heterogeneous graph model that leverage context from both webgraphs

and social media data achieves SoTA results on the website credibility classification task. Further, our best model model does not incorporate any explicit text or semantic content from the webpages or from social media users, which makes this approach flexible, and easily-extendable to non-English contexts.

Finally, we provide the first exploration of *dredge words* on social media. We incorporate a small set of dredge words into an unreliable domain discovery process in an attempt to mimic how social media users may transition from social platforms to the discovery of misinformation sources via search engines. Surprisingly, we find that dredge words frequently surface social media URLs in top Google Search Engine Result Page (SERP) positions—i.e., the websites returned when a user enters a query on Google Search. This suggests the existence of a bidirectional path that often leads back to social media. We show our heterogeneous model greatly outperforms competing systems in the top-k discovery of unlabeled unreliable websites. We show our heterogeneous model greatly outperforms competing systems in the top-k discovery of unlabeled unreliable websites. We publicly release the code, webgraph data collected for this project, a dataset of 3,939 dredge words for 46 unreliable news domains, and the resulting SERPs for each of the dredge word queries[2]. Our findings demonstrate that both direct and indirect paths to misinformation provide important signals that can be leveraged by researchers and platform designers working to mitigate the spread of misinformation across digital ecosystems.

5.4 Related Works

5.4.1 Social Media and Webgraphs

Classifying individual texts or articles is a core problem addressed by research in misinformation detection. Such detection methods have typically relied on website content and social media data Castelo et al. [2019], Chen and Freire [2020], Silva et al. [2021], Wang et al. [2024]. Castelo et al. proposed a topic-agnostic detection system Castelo et al. [2019] that identifies unreliable articles based on Linguistic Inquiry and Word Count features. Chen and Freire adopted the method for the task of unreliable domain discovery Chen and Freire [2020]. Their discovery system capitalized on user tendencies within the social graph, wherein a user who tweets a URL from a known unreliable source is likely to also tweet URLs from yet unknown sources. Similarly, Silva et al. combine website content and social context resulting in a

[2] https://github.com/CASOS-IDeaS-CMU/DredgeWords

misinformation page detection method that leverages heterogeneous data types Silva et al. [2021], with a focus on early detection across a broad range of topics. The predictive power of webgraphs and social media data have been demonstrated separately in several contexts. Aswani et al. detect SEO manipulation by link-building sites by clustering on Pagerank score and Domain Authority Aswani et al. [2021]. For detecting news site bias, Aires et al. Patricia Aires et al. [2019] scrape cross-links from a list of prominent news sites from Media Bias Fact Check [Zandt, 2022]. Sehgal et al. explore a case where misinformation was spread through coordinated hyperlink and social media networks Sehgal et al. [2021]. Another recent case study demonstrated the manipulated webgraph linkages of unreliable pseudo-thinktanks Williams and Carley [2023]. Additionally, Zhang and Cabage show that social media and SEO promotion have different strengths; they find that social-sharing results in immediate but short-term boosts to traffic, while the benefits of link-building are slower to realize, but last longer Zhang and Cabage [2017]. This hints that a combination of both webgraph and social media data should be used in investigating misinformation sources. However, to our knowledge, this is the first work that attempts to combine the two contexts for unreliable domain classification or discovery.

Despite its importance and relevance, domain-level credibility detection is a relatively uncommon task in the literature—partially as a result of the lack of accepted labels. The most comprehensive label list as of 2024 are those published in Lin et al. [2023], which aggregates the scores of multiple different domain reliability rating lists. The vast majority of work in this space occurs at the article level (e.g., Nakov et al. [2023], Bianchi et al. [2024]). This class of approaches often take a sample of articles from each website and aggregate reliability ratings in some way. These approaches have several important limitations; namely, article-level approaches are restricted to languages with LLM support, model qualities depend heavily on what is sampled, many unreliable news sites often report true information with heavily partisan slants, and articles don't necessarily reveal the goals of the publisher—e.g., knowing a media outlet has strong ties with a adversarial government can change how readers interpret the site's content. Consequently, researchers have recently begun exploring content-agnostic domain classification approaches. In 2023, Yang and Menczer [2023] showed that LLMs could be used to evaluate website credibility. More recently, Carragher et al. [2024] proposed a webgraph-based model for unreliable domain detection and discovery tasks. For discovery, the authors use backlinking domains in the webgraph in a snowball sampling approach. The authors demonstrate that GNN models trained on SEO attributes and webgraph data are effective for detection.

5.5 Data

We construct a heterogeneous graph with different relations based on the direct and indirect paths that users can take to unreliable websites. To capture direct paths to target news sites, we collect the 10 domains which most frequently link to each labeled news site (e.g., website i has linked 1M times to website j). We further extract social media mentions of each target news site (e.g., user k posts a hyperlink to website j). We consider both of these to be *direct paths*, as in both these cases, users would simply click a link to access website j. Finally we consider an *indirect path* through dredge words. Dredge words connect users to websites through Google SERPs. When browsing social media, users can encounter words or concepts with which they're unfamiliar and that exist in data voids—for example, a user might see a Twitter post about "indigo children" and search a variation of that query on Google. The results surface numerous pseudoscientific websites and advertisements (see 5.9.1). We provide a summary visualization of this heterogeneous network in Figure 5.1.

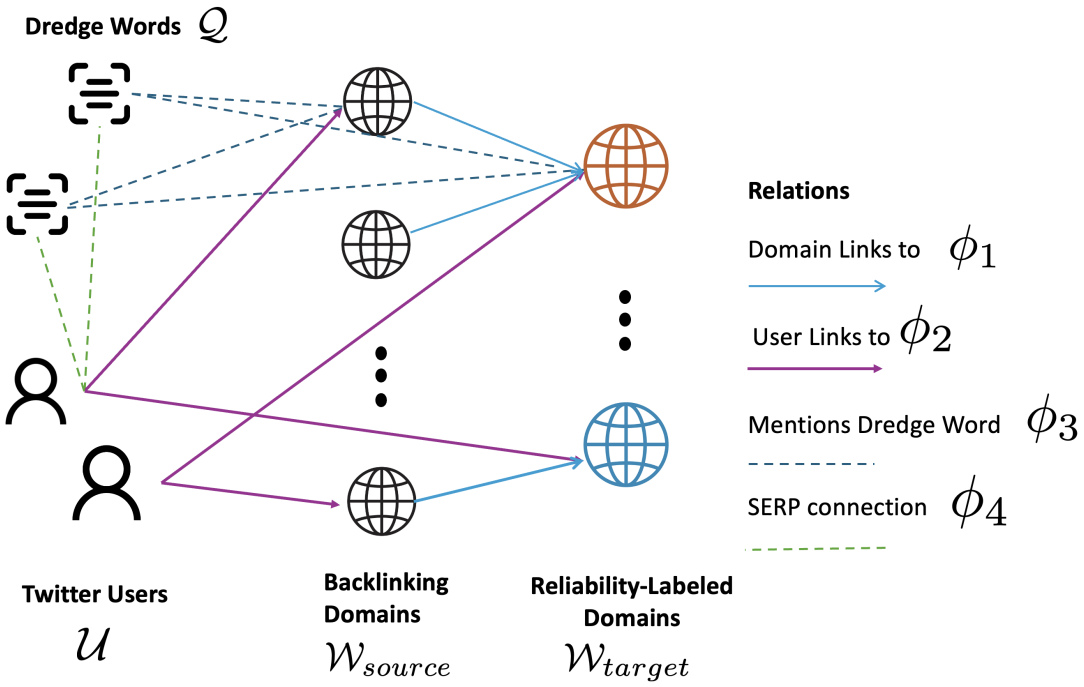


Figure 5.1: A summary of the heterogeneous graph construction process. Solid lines denote direct paths (a user clicks a hyperlink), and dashed lines denote indirect paths (a user sees a post and then queries a subset of that post on a search engine).

5.5.1 Webgraph and Features

We use the dataset discussed in chapter three. In summary, we build on the domain credibility labels introduced by Lin et al. [2023], who aggregate ratings from six expert groups using imputation and principal component analysis to produce reliability scores for 11,520 domains. Because the continuous principal component scores are less interpretable at fine margins, we binarize the labels following prior work, classifying the bottom two quintiles (principal component score ≤ 0.5162) as "unreliable" and the remainder as "reliable." Although the dataset is largely English-language, it includes domains targeting multiple languages and audiences. After excluding 193 domains for which backlink data could not be retrieved (most of which appeared inactive), we retain 11,327 labeled domains.

To construct the network, we use Ahrefs to extract the top ten highest-volume backlinking domains for each labeled site, yielding a graph of 43,758 unique domains (11,327 labeled and 32,431 unlabeled backlinking domains) with 23 domain-level attributes per node. Edges represent direct hyperlink paths from source domains to labeled targets. Consistent with prior findings, we observe assortative structure among labeled domains: when restricting to the labeled subgraph, reliable domains are more likely to link to other reliable domains, and unreliable domains to other unreliable domains (see Figure 3.5). This assortativity motivates the use of network-based modeling approaches in subsequent analyses.

5.5.2 Twitter Data

For social media context, we use a Twitter dataset constructed by querying COVID-related keywords[3] via Twitter's streaming API between January 29, 2020 and June 26, 2022. Due to server issues and API limitations, 121 days over the time period have partial or missing data. However, these gaps are spread relatively evenly over the time period, and so the data still provide strong coverage. Our final Twitter dataset contained 3.6 billion extracted tweets.

From this set of 3.6 billion tweets, we extract all Tweets that either link to one of our 11,327 websites or retweet, reply to, or quote a tweet containing one of our 11,327 websites. This resulted in a dataset of 320M tweets that link to over 840k unique domains. Of the original 11,327 labeled domains, only 5,504 appeared within the Twitter data. To reduce the noise in the dataset we perform several cleaning and filtration operations including dropping all tweets that do not directly link to one of our target sites.

We first extract all tweets that link to or mention one of our 11,327 websites and all

[3]coronavirus, Wuhan virus, Wuhanvirus, 2019nCoV, NCoV, NCoV2019, covid-19, covid19, covid 19

tweets that retweet, reply to, or quote a tweet containing one of those 11,327 websites. Once we have this data, we drop all users that do not explicitly include a link to one of our 11,327 target domains. Second, users with fewer than 10 observed tweets over the time period were dropped. To further condense the dataset, we only include tweets which appeared at least 3 times in the dataset—e.g., if a tweet was reposted or retweeted twice. As reposting and retweeting are important influence metrics on Twitter, this drops the tweets that likely did not receive as much attention. We also dropped domains that were only linked to by a single user and we dropped users that only linked to a single domain, as we hypothesized these pendulum nodes would be of limited use given the size of the graph. As a result of the expense of the webgraph attribute API, we chose to further restrict the Twitter data to only include tweets that mention at least one of the 43,758 domains for which we extracted attributes. This corresponds to keeping tweets that mention at least one of the 11,327 labeled domains for which we have attributes or tweets that mention one of the 11,327 and co-mention any of the 32,431 unlabeled domains. Following these cleaning operations, we are left with 555k users and 4.9M unique tweets which we observed tweeted or retweeted 91.2M times. This reduced dataset contains mentions of 2,475 reliability-labeled domains and 714 unlabeled backlinking domains.

5.5.3 Dredge Words

For the 100 websites with the respective lowest and highest reliability rankings, we extract the top-ranked $k \leq 1{,}000$ Google keyphrases from Ahrefs, for a total of 34,646 keyphrases. We elected to choose a limited set of websites due to an infrastructure bottleneck: querying dredge words over 3.6B tweets is very slow on our hardware. We then used WebSearcher Robertson and Wilson [2020] to query each of these keyphrases on Google and extracted the first 10 URLs returned in each SERP. There was no user account and cookies were not stored across queries; scraping was conducted from an IP address in Pittsburgh, Pennsylvania, which could impact search results. We kept all queries for which the target unreliable domain was returned in the top 10 Google search results. As many of the domains for which we collected dredge words did not have any queries for which they ranked in the top 10 on Google, this resulted in a set of 3,939 *dredge words* spanning 46 unreliable domains. We created a separate query to find mentions of these keyphrases in the 3.6B tweet covid dataset. This yielded 5.7 million tweets containing dredge words. Many of the most common mentions were explicit mentions of domain names or organization names (which can be identical to twitter handles—like “gatewaypundit”, “infowars”, and “nvic” all received over 10,000 mentions.

To filter down the data, we use regular expressions to ensure each drudge word begins with a hashtag, starts a tweet, or is preceded by a white space. This retains 213 dredge words in 421k tweets that qualitatively contain less noise. Of these 421k tweets, only 9,788 (2%) explicitly linked to the unreliable domain associated with the dredge word.

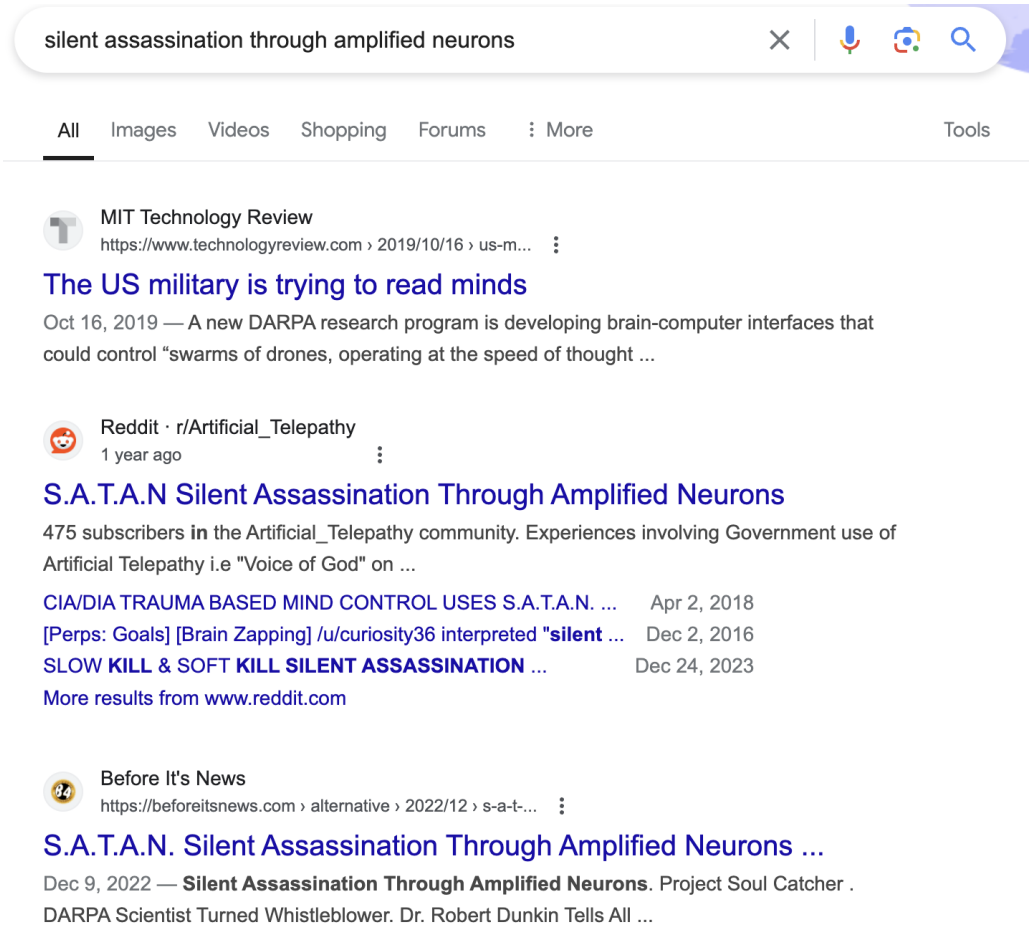


Figure 5.2: The top search results for the dredge word “silent assassination through amplified neurons”. The query surfaces fringe reddit subreddits followed by “beforeitsnews”, an unreliable news source.

This data pipeline attempts to mirror the way a user might access this information “in the wild”. A user might encounter a Twitter post that contains a reference to “silent assassination through amplified neurons”, or “Project S.A.T.A.N.”, search the query on Google, and encounter unreliable results like those in Figure 5.2. Some additional examples of dredge words include “psychic attack” (28 Twitter mentions), “akashic record” (12), “flu shot injury” (6), and “fallcabal” (5). However, we note that our current collection pipeline

is almost certainly underestimating their actual usage as 1) on Twitter, we only capture dredge word usage mentioned alongside COVID keywords, and 2) we extracted dredge words in 2024, 2-4 years after tweets were posted. While the majority of the dredge words are in English, other languages are present in the list, including Chinese, Hindi, and Arabic. We use this condensed twitter dataset and the set of SERP results they yielded in our dredge-word-based unreliable domain discovery process. While we were initially interested in paths from social media to search engines, we find dredge word SERPs surprisingly demonstrate a strong paths to social media. Youtube was by far the most commonly-returned domain (4,304 times). Wikipedia, Reddit, Quora, Twitter, Amazon, and Facebook were also among the most commonly-returned domains, in part due to the widespread commercialization of pseudo-scientific concepts.

5.6 Case Study

To provide a more concrete illustration of the tasks and our motivation, we'll consider some of the ways that users might end up on the unreliable website, Gaia[4]. This website often promotes pseudoscience, vaccine misinformation, and various conspiracy theories. We can imagine two users taking direct paths and a third following an indirect path. The first imaginary user follows paranormal accounts or pages on social media, and sees an account post a link to an article about haunted houses on Gaia. The second individual is interested in UFOs and listens to the "Coast to Coast AM"[5] radio talk show—if that individual visited the website for the talk show, they would find 29.4K hyperlinks that would lead to Gaia. In a third instance, we can imagine a user interested in astrology seeing the post in Figure 5.3.

The user, who is encountering new information, might then be curious and compelled to search "indigo children meaning" or just "indigo children" on Google—both phrases are dredge words that surface Gaia in the top 10 results. Interestingly, after manually searching these "indigo children meaning" in June 2024, we found that Google's definition snippet—a summarization box on the top of some SERPs with the text "From sources across the web"—expresses pseudoscientific concepts. The box states that "Indigo Children are kids with indigo-colored auras". This snippet was followed by YouTube videos promoting Indigo Children pseudoscience and a Reddit post from the /r/Psychic subreddit. Another user who may have searched only "indigo children" would have seen a snippet from Wikipedia

[4]https://mediabiasfactcheck.com/gaia/

[5]https://en.wikipedia.org/wiki/Coast_to_Coast_AM

Figure 5.3: A truncated tweet about "Indigo Children".

which correctly calls indigo children pseudoscientific. However, adjacent to the snippet, the user would have seen four books for sale about indigo children. After that, the second highest-ranking return was an article from an unreliable domain identifying the "13 signs you're an indigo child".

This case study also highlights the importance of conceptualizing misinformation consumption as interconnected; none of our hypothetical users were actively seeking anti-vaccine content, and yet in pursuing their relatively-innocuous interests, all are exposed to medical misinformation. While all of these paths lead to Gaia, the indirect routes may be the most effective at swaying users, as discovering Gaia through a trusted intermediary (in this case, Google) may confer additional trust to the source Robertson et al. [2023a]. The questions we are trying to answer in this work, with respect to this case study, are 1) "How does integrating each of these path contexts into a model impact our ability to identify Gaia as unreliable?" and 2) "What websites that link to Gaia are also promoting misinformation"? Even in this case study, neither tasks are trivial; detecting unreliable websites is challenging, and while many of the websites that link to Gaia are unreliable, others are generic SEO sites. Additionally, there are reliable and .edu sites that link to Gaia in fact-checking articles, which further complicates the task.

5.7 Methods

In order to evaluate the impact of additional levels of context on our models, we construct graphs that capture increasingly granular levels of context present in the data. Let $\mathcal{W}$ be the set of all labeled and unlabeled websites $\{w_1, w_2, \ldots, w_n\}$ contained in the extracted

webgraph data. We define $\mathcal{W}_{target}$ as the subset of $\mathcal{W}$ for which we have reliability labels—i.e., the 11,327 reliability-labeled news domains; we define $\mathcal{W}_{source}$ as the 10 websites that most link to each domain. We note that $\mathcal{W}_{source}$ contains some websites in $\mathcal{W}_{target}$. Next define $\mathcal{U}$ as the set of all Twitter users $\{u_1, u_2, \ldots, u_n\}$ that link to domains $\in \mathcal{W}$. Finally, let dredge words $\mathcal{Q}$ be the set of keyphrases $\{q_1, q_2, \ldots, q_n\}$ connected to SERPs containing URLs $\in \mathcal{W}$.

We define several preliminaries. A Graph Union operation ($G = G_1 \cup G_2$), between graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is defined as $G = (V, E)$ where $V = V_1 \cup V_2$ and $E = E_1 \cup E_2$. A heterogeneous graph is an extension of a homogeneous graph, $\mathcal{G} = (V, E)$, where V and E are associated with a node type mapping function $\Psi : V \to A$ and an edge type mapping function $\Phi : E \to \phi$. In our setting, the set of node types are $A = \{\mathcal{W}, \mathcal{U}, \mathcal{Q}\}$ and the set of edge types are $\Phi = \{\phi_1 : domain - to - domain, \phi_2 : user - to - domain, \phi_3 : user - to - dredgeword, \phi_4 : dredgeword - to - domain\}$. This setup is visualized in Figure 5.1.

All GPU experiments experiments were run on a single NVIDIA GeForce RTX 3080. Given the small number of labels, no model took over 10 minutes to train. GNN experiments were implemented in Python3.10 using Pytorch Geometric Fey and Lenssen [2019a]. We publicly release all code, annotations, and non-Twitter data.

5.7.1 Homogeneous Graphs

We consider three homogeneous baselines—where all nodes are treated as part of the same class—to justify the added complexity of using heterogeneous networks. We define two homogeneous graphs: $\mathcal{H}_{domains}$ contains domain-to-domain relationships, i.e., $\mathcal{W}_{source}$ to $\mathcal{W}_{target}$. $\mathcal{H}_{users}$ contains relationships between Twitter users and the websites they mentioned $\in \mathcal{W}$. Finally, we define $\mathcal{H}_{domains+users} = \mathcal{H}_{domains} \cup \mathcal{H}_{users}$. Clearly, these latter two graphs contain different sets of nodes and relations, but we treat them as equivalent in these experiments as a simple baseline.

As the features of domains and users are drawn from different spaces and have different sizes, we elect to use positional node features calculated individually on each of the three constructed homogeneous graphs. For this graph, we therefore define Z_n as features calculated using using Node2Vec Grover and Leskovec [2016] parameterized with a walk length to 20, context size to 10, and 10 walks per node, and an embedding dimension of 23, chosen because it is identical to the number of features that we extract for each website from Ahrefs.

$\mathcal{H}_{users}$ presents an additional challenge in that not all websites $\in \mathcal{W}_{target}$ are mentioned by users $\in \mathcal{U}$. We include social-media-only baselines, but these are trained and evaluated on

only a network of user interactions with the 2,475 labeled domains mentioned in the social media data, which make these results not directly comparable with our other models.

5.7.2 Additional Baselines

We include comparisons to two past works that explored website credibility classification without including article or website content. First, we re-implemented the GNN for news domain credibility classification and discovery proposed in Carragher et al. [2024], which is a homogeneous model which uses the Ahrefs features of domains. Following the authors, we log-normalize the Ahrefs features of each domain. We train and evaluate their model on all labeled 11,327 reliability-labeled domains.

Second, we compare our results to those from Yang and Menczer [2023], where the authors asked ChatGPT to return the probability that a website is unreliable. Following the authors recommendation, we take a binary cut-off at 0.5 and compare results accordingly. As there is no hold-out set for the ChatGPT data, we consider two comparisons between our best-performing model and this approach. We first compare accuracy and F1 predictions for all 7,318 sites; however, this comparison is problematic as this guarantees leakage—i.e., some of these sites were in our model's training data. To account for leakage, we perform a second evaluation where we only evaluate the predictions of our model and of ChatGPT on the hold-out test set of domains that our model never encountered.

5.7.3 Heterogeneous Graphs

We construct two heterogeneous graphs. $\mathcal{E}_{domains+users}$ is structurally equivalent to its homogeneous counterparts, but explicitly incorporates the two node types ($\mathcal{W}$,$\mathcal{U}$) and their different relations (ϕ_1, ϕ_2). we use a heterogeneous graph-neural network architecture that can treat user nodes and domain nodes as different nodetypes. We additionally construct $\mathcal{E}_{domains+users+dredge}$, which contains all node types ($\mathcal{W}$,$\mathcal{U}$, $\mathcal{Q}$) and all relation types ($\phi_{1:4}$).

In the heterogeneous networks, the features of domain nodes $\mathcal{W}$ are always the logged domain-level features extracted from Ahrefs, as in Carragher et al. [2024]. For user nodes, we ran each model with positional context—Node2Vec features, and separately with text context. To create textual context, for each user, we randomly sample, without replacement, 10 tweets for each observed user and embed them using multilingual distilBERT Sanh et al. [2019a]. This is a naive approach that excludes a large amount of context for some users, particularly as we observed two users tweet over 1 million times in our filtered dataset.

However, considering only 10 tweets is a standard practice in user embedding literature Pan and Ding [2019]. A clear path for future work is to explore the impact of more advanced user embedding strategies within this system. Dredge word embeddings were extracted using multilingual distilBERT, and these embeddings are used as features for dredge word nodes. We note that dredge words were only considered for 46 of the least reliable websites, so their inclusion in models is likely to harm the models, as those features only exist for a small set of the 11,327 labeled domains. We nonetheless report statistics and discovery evaluations.

5.7.4 Graph Neural Network Training

A label-stratified 80/10/10 split on labeled websites is used to create training, validation, and test sets. Nodes that did not have reliability labels—the majority of nodes in all networks—were masked during training. For our one-mode, heterogeneous, and homogeneous experiments we use a simple 2-layer graph neural network using GraphSAGE convolutions proposed in Hamilton et al. [2017]. We ran experiments using graph attention networks Veličković et al. [2017] and heterogeneous graph transformers Hu et al. [2020], but found marginal accuracy gains at the cost of longer training times. As our primary classification interest is evaluating the impact of context—i.e., signals from different sources—on domain credibility classification, rather than finding the best architecture, we elect to use homogeneous and heterogeneous GraphSAGE layers for each respective model. We include a single baseline that uses a heterogeneous graph transformer as well. Each model consists of a SAGEConv layer with dropout, with a hidden dimension of 512, followed by a ReLU activation and a second SAGEConv layer with a log softmax activation function. For models trained with user text, we included an additional linear layer to align input feature sizes. For each model, we train for a maximum of 1,000 epochs with early stopping based on validation loss and a patience of 50. We use an Adam optimizer with a starting learning rate of 1e-3 and a cosine annealing learning rate scheduler[6]. The heterogeneous GNN, the most complex network we consider, contains 3.2M parameters, and no model exceeded 5 minutes of training time to reach our early stopping convergence condition. All models in this paper were run or trained on a single NVIDIA GeForce RTX 3080.

[6]Implementation is available at https://github.com/CASOS-IDeaS-CMU/DredgeWords

5.7.5 Curriculum Learning

We assume that highly reliable and highly unreliable websites are easier to differentiate in webgraphs than those that are mixed, mostly reliable, or mostly unreliable. A manual examination of the labeled URL data leads us to believe that this assumption is very reasonable when website content is considered (see Section 5.5.1). We implement a slightly-modified version of the"Baby Steps" learning curriculum proposed and explored in Spitkovsky et al. [2010], Cirik et al. [2016]. As domain labels from Lin et al. [2023] contain unified principal component scores of expert ratings of domain reliability, we can develop a curriculum that first learns labels of extremely reliable and extremely unreliable domains, and that gradually works from the extremes towards the websites with mixed reliability. Using the original labeled domain dataset $\mathcal{D}$, we calculate quintiles of reliable labels and unreliable labels using principal component scores, and define these as an ordered set of batches $\{d_1, d_2, \ldots, d_{10}\}$, where d_1 is the first quintile of the most reliable domains and d_{10} is the fifth quintile of the unreliable domains. When using curriculum learning, we begin training the model $\mathcal{M}$ using $\{d_1, d_{10}\}$, and following convergence, the model considers $\{d_1, d_{10}\} \cup \{d_2, d_9\}$. The model continues in this fashion until it converges on all available data. In our Baby Steps implementation, the we define model's corresponding convergence as 10 epochs without an improvement in validation loss. Once the curriculum has incorporated all data, the convergence criterion described in the previous section are used. More generally, for a website reliability curriculum $\mathcal{C}$, an even-length dataset $\mathcal{D}$, and a model $\mathcal{M}$, our implementation of the Baby Steps curriculum can be expressed as the procedure in Algorithm 1.

Algorithm 1 Modified Baby Steps Curriculum

input $\mathcal{M}$, $\mathcal{D}$, $\mathcal{C}$
  sort($\mathcal{D}$, $\mathcal{C}$)
  $B \leftarrow \emptyset$
  for $i \in 0, 1, \ldots \frac{k}{2}$ **do**
    while $\mathcal{M}$ not converged **do**
      $B \leftarrow B \cup \mathcal{D}[i] \cup \mathcal{D}[k-i]$
      $\mathcal{M}(B)$
  return $\mathcal{M}$

5.7.6 Unreliable Domain Discovery

While we can evaluate our classifiers on domains in $\mathcal{W}_{target}$, we desire to make a tool that can identify unknown unreliable domains. To evaluate our best-performing classifier, we explore its ability to identify unlabeled unreliable domains over the unlabeled domains in $\mathcal{W}_{source}$. We outline our discovery processes and we benchmark our discovery process against two previous works.

We implement and evaluate two distinct discovery processes. The first is GNN discovery, where we take predictions for unlabeled domains in the graph from the best performing GNN model. We call the second method Dredge Word-Based Discovery (DW-BD$_{lower}$), which mimics the path social media users take to reach unreliable domains by observing dredge words on social media, and querying these terms on a search engine. Using WebSearcher Robertson and Wilson [2020], we create a candidate list of domains from the top 10 SERP results for each of the dredge words we have compiled from our Twitter dataset. We then pull Ahrefs attributes for each of the candidate domains and use the SEO attribute classifiers from Carragher et al. [2024] to filter down the candidate domains.

We add an additional Dredge Word-Based Discovery baseline (DW-BD$_{upper}$), which does not limit the discovery process based on whether or not we observed dredge words mentioned in the Twitter dataset. For this baseline, we extracted Google SERPs for 118k dredge words associted with 1,051 unreliable domains. We then dropped all results where we did not observe the target domain in rank in the first 10 SERP results. This resulted in 73,843 dredge words for which 836 unreliable websites ranked in the top 10 Google results. The resulting dataset contained 884k individual search results. Due to Ahrefs API constraints, we filter randomly drop 30k domains that only appeared once in search results, leaving 854k individual search results linking to 44,503 unique domains. We and run the discovery process described on this dataset above to generate candidate unreliable domains. Whereas DW-BD$_{lower}$ is bounded by Twitter data, DW-BD$_{upper}$ is not.

We compare our discovery processes with the webgraph-based discovery (WG-BD) process proposed in Carragher et al. [2024] and with the social media-based discovery process (SM-BD) proposed in Chen and Freire [2020]. Specifically, we consider Precision@5, @10, @20, and we consider the partial F1 metric proposed in Chen and Freire [2020]. This means we run the discovery process twice, once on the full domain list where we evaluate results manually with top-10 and top-20 accuracies, and again on a restricted domain list to compute the partial F1 metric. We additionally report partial precision and recall.

Due to a lack of ground truth labels for evaluating newly discovered domains, Partial

precision, recall, and F1 measure the ability of the discovery system to find unreliable domains with respect to two known lists of unreliable sources, a seed list and an evaluation list. As defined by Chen and Freire [2020], the seed list is the PoliticalNews dataset and the evaluation list is drawn from unreliable MBFC domains:

$$p = \frac{\#\text{prediction} = \text{fake } \textbf{and } \text{MBFC label} = \text{fake}}{\#\text{prediction} = \text{fake}} \tag{5.1}$$

$$r = \frac{\#\text{prediction} = \text{fake } \textbf{and } \text{MBFC label} = \text{fake}}{\#\text{MBFC label} = \text{fake}} \tag{5.2}$$

$$pf1 = 2 \times \frac{p \times r}{p + r} \tag{5.3}$$

Partial F1, as given by Equation 5.3, measures the ability of a discovery system seeded on PoliticalNews to discover as many MBFC domains as possible (high partial recall), without discovering domains that are not in the MBFC list (high partial precision) [Castelo et al., 2019]. These metrics are further limited in our DW-BD experiments, as many (53 of 74) politifact websites are either inactive or do not have keywords that rank them on Google. For those experiments, we therefore calculate partial metrics only on SERPs corresponding to dredge words from 21 active domains. We discuss the limitations of the partial F1 metric in more detail in Limitations and Appendix C.

5.8 Results

5.8.1 Credibility Classification

We present Accuracy and F1 statistics for website reliability classification for homogeneous models ($\mathcal{H}$) and heterogeneous models ($\mathcal{E}$) in Table 5.1. We find that the heterogeneous model which incorporates both the user network and domain webgraph outperforms all other models with an average accuracy over 10 runs of $0.7865 \pm .002$. However, swapping the Heterogeneous GraphSage Convolutions in this model with a Heterogeneous Graph Transformer ($HetGT$) convolutions yielded a slightly better F1 ($0.789 \pm .003$). While the accuracy of the Social Media user model ($\mathcal{H}_{users}$ is higher, this model is only evaluated the 2,475 labeled domains in $\mathcal{W}_{target}$ to which social media users were connected, and thus are not directly comparable to other models. In the heterogeneous models, we considered embedded user text features (text) as well as user node node2vec features (n2v). Interestingly, including embedded text

of social media users as node performed worse than using features that uniquely consider network position.

Model	Accuracy	F1
$\mathcal{H}_{domains}$	0.7686 ± .004	0.7545 ± .005
Carragher et al.	0.7799 ± .002	0.7714 ± .002
$\mathcal{H}^{*}_{users}$	0.8113* ± .007	0.7582* ± .009
$\mathcal{H}_{domains+users}$	0.7696 ± .006	0.7559 ± .006
$\mathcal{E}_{domains+users(n2v)}$	**0.7865 ± .002**	0.7777 ± .003
$HetGT_{domains+users(n2v)}$	0.7820 ± .003	**0.7895 ± .003**
$\mathcal{E}_{domains+users(text)}$	0.7738 ± .118	0.7657 ± .018
$\mathcal{E}_{domains+users+dredge}$	0.7787 ± .004	0.7675 ± .007

Table 5.1: Mean Accuracy, F1, and standard deviations for each GNN ablation over 10 runs. $\mathcal{H}$ denotes homogeneous and $\mathcal{E}$ denotes heterogeneous. * denotes a social-media-only model run and evaluated on the 2,475 labeled domains mentioned in the Twitter data.

We compare our best performing model with the approach proposed in Yang and Menczer [2023] in Table 5.2. The authors in Yang and Menczer [2023] used ChatGPT to return the probability that 7,318 websites—7,317 of which are in our list—were unreliable labeled domains which we will denote $\mathcal{W}_{LLM}$. Following the authors, we use a cutoff of 0.5 to binarize their ChatGPT predictions. Our model was trained on many of the 7,318 websites that GPT evaluated, so comparing all data results in label leakage. Consequently, we provide a second evaluation, where we only compare labeled websites that our models never encountered in training, i.e., $\mathcal{W}_{LLM} \cap W_{test}$. On both $\mathcal{W}_{LLM}$ and the $\mathcal{W}_{LLM} \cap W_{test}$ evaluations, our model outperformed ChatGPT. Our model yielded an F1 that was 0.03 higher than ChatGPT on $\mathcal{W}_{LLM}$ and 0.04 higher on $\mathcal{W}_{LLM} \cap W_{test}$.

5.8.2 Unreliable Domain GNN Discovery

Two annotators[7] independently annotated the reliability of the top 20 predictions (sorted by prediction confidence) of the top performing model, heterogeneous model domains+users(n2v),

[7]25-30 year-old PhD candidates studying misinformation

	Acc.	F1
GPT ($\mathcal{W}_{LLM}$)	0.81	0.751
$\mathcal{E}_{domains+users(n2v)}$ ($\mathcal{W}_{LLM}$)	**0.837**	**0.780**
GPT ($\mathcal{W}_{LLM} \cap \mathcal{W}_{test}$)	0.782	0.701
$\mathcal{E}_{domains+users(n2v)}$ ($\mathcal{W}_{LLM} \cap \mathcal{W}_{test}$)	**0.819**	**0.745**

Table 5.2: Our best model outperforms the LLM approach proposed in Yang and Menczer [2023]. We evaluate against all 7,318 websites the authors considered ($\mathcal{W}_{LLM}$) and against the subset of $\mathcal{W}_{LLM}$ contained in our test set ($\mathcal{W}_{LLM} \cap \mathcal{W}_{test}$).

over the set of all unlabeled domains. Annotators were asked to assess on a binary scale, whether each identified site was reliable or unreliable. In some of our discovered sites, this distinction is unambiguous (e.g. prominently pro-Q-anon blogs). If the website appeared to express political opinions, the annotators were instructed to search for medical or scientific claims. If authors made conspiratorial (e.g., "covid is part of a great reset") or pseudo-scientific claims (e.g., "global warming is a hoax"), the website was judged to be unreliable. Annotations were done independently, and we report Krippendorf's Alpha on the independently-annotated lists. While we do not control for potential annotator bias, annotators never used political bias as a rationale for rating a site as unreliable. We release all annotations along with notes left by the annotators. Inter-annotator agreement in the first round had a Krippendorff's $\alpha = 0.78$. The annotators then met and resolved the single disagreement. We compare our model with the webgraph-based discovery (WG-BD) approach in Carragher et al. [2024][8]. Again, two independent annotators ranked the results, yielding a Krippendorff's $\alpha = 0.69$.

For comparison with previous unreliable domain discovery approaches, we additionally report partial F1, defined in Chen and Freire [2020] as the set of true positive unreliable domains with credibility labeled "mixed" or worse by Media Bias Fact Check. Thresholding at a prediction confidence level of 0.7 (see Appendix C for sensitivity analysis), the Partial F1 of our top-performing GNN model is 0.25, which is largely previous works; 0.29 for SM-BD Chen and Freire [2020]) and 0.28 for WG-BD Carragher et al. [2024]. As demonstrated in Table 5.3, we observe that our heterogeneous domains+users model outperforms both competing systems at each level of precision we consider.

[8]As that paper's discovery process did not rank discovered domains, we sort results by the misinformation classifier proposed in the work.

Model	P@5	P@10	P@20	PF1	PP	PR
$\mathcal{E}_{d+u(n2v)}$	**1**	**1**	**0.9**	0.25	0.18	0.35
$\mathcal{E}_{d+u(n2v)+dredge}$	*	*	*	0.07	0.04	0.26
SM-BD	-	-	0.7	**0.29**	**0.24**	**0.37**
WG-BD	0.2	0.5	0.65	0.28	0.24	0.31
DW-BD$_{lower}$	0.8	0.7	0.55	0.02	0.17	0.01
DW-BD$_{upper}$	0.6	0.4	0.5	0.12	0.28	0.08

Table 5.3: Precision@K and Partial F1 (PF1), Partial Precision (PP), and Partial Recall (PR) discovery evaluations. Chen and Freire [2020].

5.8.3 Dredge Word Discovery

We consider two dredge word discovery approaches: one approach (DW-BD$_{lower}$ and DW-BD$_{upper}$) adapted from the discovery approach proposed in [Carragher et al., 2024] and the other using the GNN trained with dredge word context ($\mathcal{E}_{domains+users+dredge}$), which we shorten to ($\mathcal{E}_{d+u(n2v)+dredge}$) in Table 5.3. The GNN without dredgewords outperforms all other baselines at every level of precision tested. We note that DW-BD$_{lower}$ yields higher precision than the webgraph based discovery baseline and the DW-BD$_{upper}$ baseline at P@5 and P@10, but the reported social media P@20 outperforms the webgraph and dredge-word discovery processes. Several of the sites returned by DW-BD$_{upper}$ were user-based forums, crypto exchanges, or supplement sellers, which while not necessarily reliable, are out-of-domain and not rated as unreliable.

In the latter approach, as a consequence of 1) only extracting dredge word networks for only 46 of the least reliable domains 2) the dominance of social media and shopping sites in the dredge word SERPs, the dredge word GNN learns to heavily associate social media sites and online shopping platforms with unreliable domains. The 10 most confident predictions the GNN returns contain 7 social media sites, amazon, itunes, and s3.amazonaws. We find at least several of the pseudo-scientific and conspiratorial dredge words are targeted by dubious sellers, podcasters, and social media influencers. As these websites are out-of-domain for our task, but still often contain unreliable content, we chose not to evaluate Precision@k for this set. A deeper exploration of the link between these conspiratorial dredge words and non-news domains would be a fruitful path for future work.

5.9 Analysis and Discussion

We explore the 25 most confident unreliable predictions of our best performing model over the set of all unlabeled domains in $\mathcal{W}_{source}$. We observe that these predictions seem to fall into three broad categories: health misinformation (6 websites), Qanon[9] misinformation (5), and German-language far-right misinformation (9). In all three categories, the majority of the websites express skepticism towards vaccines, the existence of COVID-19, or both. Most interestingly, our investigation of the these categories captures an interplay between websites spreading medical misinformation and the websites profiting off of it.

Six of our model's 25 most confident predictions, are an likely co-owned network of websites selling dubious medical products. The $\mathcal{W}_{target}$ website "holitichealth"[10] offers users advice on how to detox from vaccines, which the site claims "poison your DNA, brain, nervous system and immune system". The six sites that most frequently link to holistichealth all appeared in our model's most confident unreliable predictions—all sites have the same HTML template, end in .one, and resolve to the same IP address, and frequently link to one another, heavily suggesting co-ownership. Our model's second most confident prediction, herbalremedies[11], sells 17 supplements that the site claims kill cancer cells, including colloidal silver and "parasite cleansing herbs". Another .one site, "emfprotection", sells electric and magnetic field protection gear that the site claims will defend customers from psychic mind control and 5G radio waves. While we did not observe any Twitter mentions of the 6 unlabeled sites, holitichealth, was mentioned by 7 Twitter accounts. Interestingly, the 7 Twitter accounts often linked to some reliable news sites, but also linked to Zero Hedge, The Epoch Times, Breitbart, and other low-and-mixed reliability websites.

The second large category of discovered sites promoted conspiratorial—often Qanon Garry et al. [2021]—content. Three of these sites (wakeup.icu, vintel1776.net, and qaggregator.news), are now dead, but these sites were all active during the period that Twitter data were collected. Two of the sites linked heavily to a website containing all of Q's drops[12], and wakeup linked heavily to a Qanon discussion board. The only identified Qanon website that was mentioned on Twitter was Hnewswire[13], which was mentioned by 100 distinct Twitter

[9] A conspiracy theory that claims that a cabal of Satanist, sex-trafficking, child molesters is running the world.

[10] https://web.archive.org/web/20240329171920/https://holistichealth.one/

[11] https://web.archive.org/web/20240225051441/https://herbalremedies.one/herbs-for-cancer/

[12] https://web.archive.org/web/20240911060604/https://qagg.news/

[13] https://web.archive.org/web/20240903154602/https://hnewswire.com/

accounts that also heavily mentioned Zero Hedge, Gateway Pundit, and many other low and mixed reliability domains. Hnewswire's navigation menu contains a trending topics bar which includes categories like "Covid Kill Shot", "Plandemic", and "Demonic Activity End Times". Independent annotations of predictions can be found in the Github repository.

The top 25 discovered websites were mentioned by 173 Twitter users in the COVID data who cumulatively mentioned over 3,165 domains. Interestingly, 42 of these users also used dredge words at least once in COVID-19-related tweets during that time period. Most of these are names of unreliable websites, e.g., "creation", "gateway pundit", "rense", "infowars", and "bitchute", but the dredge words "what really happened", "national vaccine", "being 6", and "missing links" were also used by this set of users.

This analysis highlights the connectivity of COVID-19 misinformation. Both the Qanon websites and the far-right German-language websites (which we do not discuss due to space constraints) actively promoted COVID misinformation, often alongside other conspiracies. This misinformation was then monetized by the .one websites, which sell, for example, "natural Hydroxychloroquine and Ivermectin alternative[s]" to those concerned about COVID[14]. This is a key advantage of the current method over article-based approaches; incentive structures are often built into networks. There is often overlap in those spreading conspiracies and those monetizing them Ballard et al. [2022], and having a system that considers the diverse paths that users can take to unreliable content helps us identify that interplay.

5.9.1 Dredge EDA

We applied the python library langdetect Danilák [2014] to the list of 3,933 dredge words used in the paper, but we observe that automated language extraction of dredge words is a challenge. We observed many dredge words are disembodied fragments, names of people, or simply foreign words adopted by various communities. For example, several Hindi-language-origin yoga poses appear in the dredge word list: parivrtta anjaneyasana and parivrtta utkatasana. These terms surface *gaia*, a conspiracy pseudoscience website[15]. There are additionally arabic and chinese dredge words that surface *creation*, a website that frequently promotes pseudoscience[16]. Qualitatively, hatespeech is fairly uncommon in the dredge words we identified; 11 unique dredge word phrases contain the non-euphemistic writing of the

[14] https://web.archive.org/web/20240420185505/https://holistichealth.one/natural-hydroxychloroquine-and-ivermectin/

[15] https://mediabiasfactcheck.com/gaia/

[16] https://mediabiasfactcheck.com/christian-ministries-international/

n-word and 3 others contain a homophobic slur.

5.9.2 Curriculum EDA

Among the websites rated most reliable, the style, content, and widespread name recognition would very likely result in most annotators to correctly identifying the sites as reliable (e.g., reuters.com, nasa.gov, smithsonianmag.com, nature.com). The landing pages of some of the least reliable labeled domains are also typically easy to categorize, with websites that contain toxic green backgrounds accompanied by blurry photos of George Soros [17], proclamations of white nationalism [18], or full links to alien siting websites and paid psychic services [19]. Though our webgraph model is content-agnostic, we hypothesize that the webgraph and site attributes will exhibit similarly "easy" patterns across the extremes. To test this hypothesis, we implement a curriculum learning batching procedure Bengio et al. [2009].

5.9.3 Partial F1 threshhold sensitivity

To investigate the partial F1 metric we experiment with various classifier confidence thresholds. Naturally, as confidence increases, we see higher precision and lower recall (Figure 5.4). The disparity between partial recall and partial precision implies that our discovery process is not very precise. However, in the manual evaluation of Precision@k, we find precision is actually very high. This indicates a fundamental issue with the partial precision metric; it underestimates model performance, particularly as the MBFC reference list ages. This is in line with previous findings that unreliable domain lists quickly become outdated [Carragher et al., 2024], and it complicates the evaluation of discovery processes. Further work on domain reliability discovery metrics is needed.

For Precision@k assessments, two annotators were asked "is this website unreliable" and asked to provide a rationale for why or why not. Annotations involved exploration of the domain and lateral reading—looking at what other reliable sources have written about the domain. We publicly release these annotations alongside annotator rationales [20].

[17] https://web.archive.org/web/20240112053008/http://www.endgamethemovie.com/
[18] https://web.archive.org/web/20240108044820/https://www.stormfront.org/forum/
[19] https://web.archive.org/web/20240116140224/https://rense.com/
[20] https://github.com/CASOS-IDeaS-CMU/DredgeWords

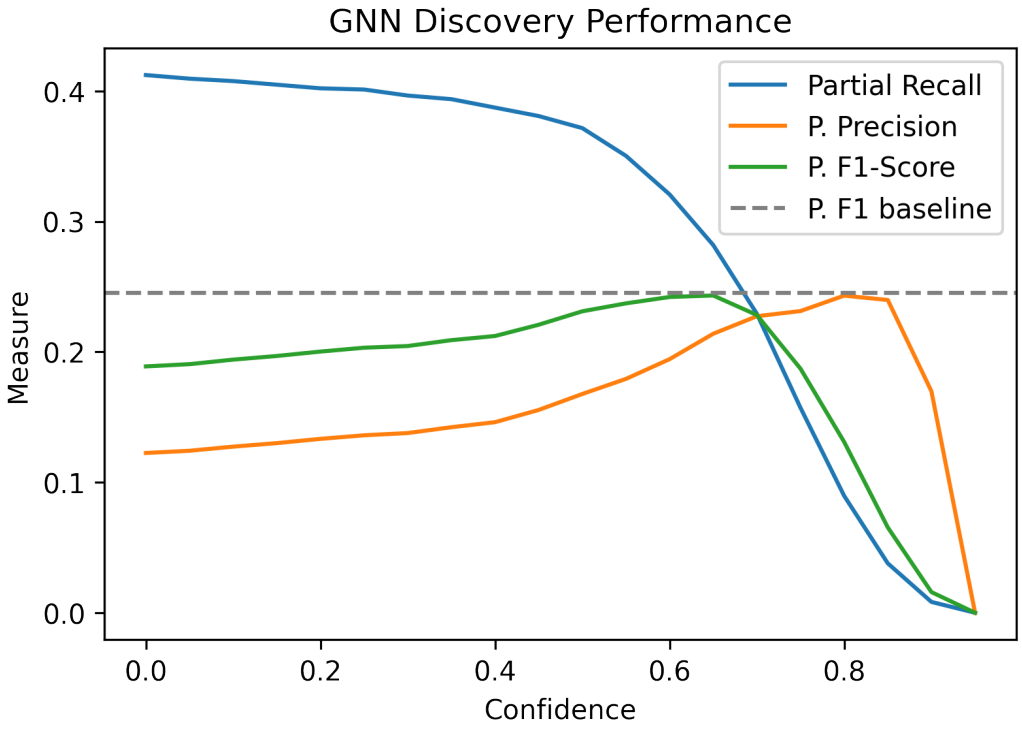


Figure 5.4: GNN discovery performance vs. classifier confidence reveals that the Partial F1 metric is precision bounded.

5.10 Limitations

While we attempted to mitigate limitations where possible, several substantial ones exist. First, the Twitter data were collected from 2020-2022, but the dredge words used in this paper were based on Google SERPs in 2024. We cannot say that the dredge words we extracted surfaced the same content in 2024 as when the phrases were initially used on Twitter. Further analysis on the temporal alignment of keyword rankings and is warranted and would be a promising avenue for future research.

As with the Ukraine Biolabs example, there are also likely many cases where unreliable websites ranked highly, but were overtaken by fact checkers; our approach would not capture this. Future work could explore the impacts of better temporal alignment, which would have the added benefit of covering data voids formed around breaking news stories. Future work could also attempt to expand this work beyond Google. We note that Google has a 2013 patent to use message passing to improve website quality signals Pennock et al. [2013], so this work could simply augment that approach. However, our model only incorporates information from two-hop neighborhoods of labeled websites (given that we use 2-layer GNNs) which makes these less effective on websites that appear on the periphery of webgraphs. Being less connected to unreliable websites might therefore be an effective camouflage strategy.

A key limitation is that the current process for extracting dredge words relies on paid third-party services. While we validated each of the extracted dredge words to ensure they are ranking in Google's top-10, approaches to extract keyphrases from, for example, common-crawl webgraphs would provide more democratization of this line of research. We publish all dredge word data collected for this project in line with that goal. We additionally note that due to resource and domain expert availability constraints, the discovery process was only evaluated by two domain experts; we release all annotations and annotator notes alongside the code. Additionally, the Partial F1 metric we use for discovery evaluation is limited, as previous work has shown many PoliticalNews and MBFC sites on which the metric relies are dead or no longer active Carragher et al. [2024]. However, to our knowledge, no better metric exists for evaluating unreliable domain discovery systems. This is a needed avenue for future work.

Finally, Additionally, the inclusion of Twitter data resulted in a relatively modest increase in performance over the domains-only model. This is surprising given how much more context Twitter provides, and I will explore why improvement is marginal in future work. Additionally, dredge words were only extracted for 46 of the least reliable 11,327 labeled

domains. Their inclusion in our models is therefore not particularly helpful. We elected to include the results from these models because they point to an interesting and understudied phenomenon: bidirectional paths between social media and search engines, which appear to be often targeted by monetary interests. In future work, we plan to repeat this process with dredge words extracted for a larger set of unreliable domains, along with keywords extracted from reliable domains.

5.11 Conclusion

Proactive content moderation requires systems that rapidly and continuously identify unreliable information sources. By considering additional direct paths that users can take to unreliable websites, we can improve the ability of models to identify these sources. We propose the idea of *Dredge Words* and highlight their bidirectional connections with social media. Finally, we demonstrate that our best-performing model outperforms competing systems on the tasks of domain reliability classification and unreliable domain discovery. This work demonstrates the signals present in the direct and indirect pathways that users follow to unreliable websites. Better understanding these paths could be a promising area of collaboration between platforms, organizations, and researchers intent on mitigating the spread of misinformation across digital ecosystems.

Chapter 6

BEND for Webgraphs and OMEN Evaluation

6.1 Introduction

Russia's Internet Research Agency (IRA)—which was indicted by Robert Mueller for its role in influencing the 2016 US election—gained widespread attention for its manipulation of social media platforms. However, the IRA's influence efforts extended beyond social media: it also created and maintained numerous websites and blogs aimed at shaping the views of foreign audiences. In addition to promoting this content on social media, the IRA actively sought to manipulate search engine rankings to boost the visibility of its sites on search engine results pages (SERPs) Adler et al. [2018]. A former employee who worked on Search Engine Optimization (SEO) at the IRA recounts rewriting articles for a series of blogs spoofing Ukrainian locations, to promote Kremlin narratives in the country Popken and Cobiella [2017b], Adler et al. [2018]. More recently, the pro-Kremlin, multilingual "Pravda Network" of news domains aimed to influence global views in line with Kremlin geopolitical interests Klen [2024]. More broadly, given the rise of generative AI, these campaigns have broader implications than websearch; researchers found information from Pravda-network websites was used as sources on Wikipedia and that content from the websites was repeated uncritically by LLMs like ChatGPT, Copilot, Perplexity, and Gemini Châtelet and Lesplingart [2025]. Attempts to manipulate webgraphs can have many downstream impacts, but analysts lack shared quantitative metrics to characterize actions taken to manipulate information environments at this level.

Experts and analysts often characterize social media Information Operations using estab-

lished conceptual frameworks like DISARM Terp and Breuer [2022], SCOTCH Blazek [2021], and BEND Carley [2020b], Blane [2023]. While each has advantages and disadvantages, BEND is often favored by network scientists because it quantifies how actors on social media attempt to influence both narrative and community (network) structures. Although social media is a key component of modern information environments, it represents only one modality through which attackers seek to manipulate the broader information landscape. While BEND is useful conceptually, its metrics are currently proprietary, and only accessible through the ORA-PRO software toolkit Carley et al. [2018]. I include a light discussion of the process to extract these metrics in Appendix A.1. There is therefore both a need for BEND metrics that extend beyond social media settings as well as open-source BEND metric definitions that can democratize its use.

In this chapter, I propose Web-BEND, quantitative definitions of community-level BEND metrics in webgraph settings, and demonstrate how the framework can be used to analyze attempts to manipulate web-based information environments. BEND comprises two components: narrative maneuvers and community maneuvers. I adopt the simplifying assumption that the original BEND narrative metrics can be applied to website content without substantial modification, using article headlines or other extracted textual fields as analogues to social media posts. However, I argue that websites require new community-level metrics as existing BEND community maneuvers were developed for social interaction networks, which are different in kind and interpretation from links within webgraphs. The development of these new webgraph community-level metrics is therefore a the central focus of this chapter. More broadly, I argue that SEO-driven attacks can be understood as efforts to influence either narratives or website-level structural authority.

To evaluate the proposed metrics, I apply both Web-BEND and the original BEND metrics across two case studies. First, I analyze the think tank webgraph networks introduced in Chapter 3, which are well-suited for evaluation because prior work identified clear signs of coordinated SEO activity within both networks, including websites created primarily to generate inbound links and artificially amplify authority signals Williams and Carley [2023]. Second, I apply the proposed Web-BEND metrics to communities within the Pravda Network, identified using the Louvain method, and show that they exhibit strong face validity in this setting as well. Together, these case studies demonstrate the utility of Web-BEND in characterizing distinct forms of webgraph manipulation.

This chapter makes three primary contributions. First, it provides the first publicly available formulation of BEND metric definitions, as existing implementations are proprietary.

Second, it shows that the proposed webgraph metrics achieve improved face validity across multiple measures in both the think tank and Pravda networks. Together, these contributions offer a standardized framework for detecting and characterizing webgraph manipulation using common SEO tactics. Finally, I take the insights from this chapter and the previous five chapters and integrate them into a larger information environment training exercise, which is a simulation designed to mimic a real-world online information ecosystem (e.g., websites, social media, and competing narratives). In these exercises, participants are given access to a synthetic but realistic stream of content and are tasked with analyzing the information, identifying narratives and potential misinformation, and proposing responses or actions. I outline the website simulation procedure I developed and report survey results from participants of the training.

6.2 Chapter Research Questions

RQ6.1 How can the BEND framework be extended to websites and webgraphs?

RQ6.2 Can the BEND framework be applied to a webgraph of think tanks?

RQ6.3 Can insights from this thesis help analysts in information environment excercises

6.3 Background

6.3.1 BEND

The BEND Framework Carley [2020b], Blane [2023] defines 16 maneuvers: 8 impacting community structure and 8 impacting narratives. Community maneuvers either build structure {Build, Bridge, Boost, Back} or fragment it {Negate, Neutralize, Narrow, Neglect}. Narrative maneuvers aim to evoke either positive emotions {Excite, Engage, Explain, Enhance} or negative ones {Dismay, Distort, Dismiss, Distract}. I provide definitions for all narrative maneuvers in Table 6.2 and all community maneuvers in Table 6.1. BEND describes information environments in terms of "maneuvers", where maneuvers are actions with effects. The BEND maneuvers are designed to be descriptive of patterns one sees in social networks and do not assume intent on the part of users. In other words, one could view them as descriptive labels that map to observable effects. For example, an actor can serve as a bridge between two groups without intending or realizing it, and the existence of that bridge

can have an impact on how information is transmitted through a network. Following this assumption, Web-BEND will treat maneuvers as descriptive labels for observed structural effects, independent of their source. For example, a webpage may function as a "Bridge" by connecting predefined communities. Quantitative metrics for each BEND network on social media data are contained within the proprietary network analysis software ORA-PRO Carley et al. [2018].

Table 6.1: BEND community maneuvers.

Valence	**Maneuver**	**Description**
Affirmative	Back	Discussion or actions that increase the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic.
Affirmative	Build	Discussion or actions that create a group, or the appearance of a group, where there was none before.
Affirmative	Bridge	Discussion or actions that build a connection between two or more groups or create the appearance of such a connection.
Affirmative	Boost	Discussion or actions that increase the size of a group and/or the connections among group members, or the appearance of such.
Adverse	Negate	Discussion or actions that decrease the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic.
Adverse	Neutralize	Discussion or actions that cause a group to be, or appear to be, no longer of relevance, e.g., because it was dismantled.
Adverse	Narrow	Discussion or actions that lead a group to be, or appear to be, more specialized, and possibly to fission, or appear to fission, into two or more distinct groups.

Valence	Maneuver	Description
Adverse	Neglect	Discussion or actions that decrease the size of a group and/or the connections among group members, or the appearance of such.

Table 6.2: BEND narrative maneuvers.

Valence	Maneuver	Description
Affirmative	Excite	Discussion or actions related to a community or topic that cause the reader to experience a positive emotion such as joy, happiness, liking, or excitement.
Affirmative	Explain	Discussion or actions that clarify a topic to the targeted community or actor, often by providing details on, or elaborations on, the topic.
Affirmative	Engage	Discussion or actions that increase the relevance of the topic to the reader, often by providing anecdotes or enabling direct participation and so suggesting that the reader can impact the topic or will be impacted by it.
Affirmative	Enhance	Discussion or actions that provide material that expands the scope of the topic for the targeted community or actor, often by making the topic the master topic to which other topics are linked.
Adverse	Dismay	Discussion or actions related to a community or topic that cause the reader to experience a negative emotion such as worry, sadness, disliking, anger, despair, or fear.

Valence	Maneuver	Description
Adverse	Distort	Discussion or actions that obscure a topic to the targeted community or actor, often by supporting a particular point of view or calling details into question.
Adverse	Dismiss	Discussion or actions that decrease the relevance of the topic to the reader, often by providing stories or information that suggest that the reader cannot impact a topic or be impacted by it.
Adverse	Distract	Discussion or actions that redirect the targeted community or actor to a different topic, often by bringing up unrelated topics and making the original topic just one of many.

The BEND framework has been widely used to characterize social media information environments across diverse contexts. Prior work has applied BEND to Twitter discussions of the U.S. COVID-19 vaccine rollout Blane et al. [2022], Chinese state media activity Phillips et al. [2023], and bot-like behavior surrounding Russia's 2022 invasion of Ukraine Marigliano et al. [2024]. Beyond Twitter, researchers have also used BEND to study Pink Slime news activity on Facebook Lepird and Carley [2024] and pro- and anti-Russian communities on Telegram Kloo and Carley [2023]. However, to date, BEND has not been extended beyond social media platforms.

In this work, I adapt the original BEND metrics to web-based information environments. I make the strong assumption that narrative maneuvers can be mapped to website headlines and articles, while community maneuvers require a different treatment. Specifically, I argue that community maneuvers on the web should be defined in terms of the actions groups of websites take to build or undermine structural authority within webgraphs. Because the original BEND metric formulations are proprietary, direct comparison is not possible. Instead, I use the ORA-PRO and NetMapper toolkits to compute BEND metrics for selected think tank groups using both headline text and webgraph structure Carley et al. [2018]. The original community-level BEND metrics also rely on a combination of textual and network

data, therefore headlines (or some other form of text) are needed to compare our proposed Web-BEND metrics with the original BEND metrics.

6.3.2 Louvain Community Detection method

In this chapter, I propose metrics that rely on some user-defined community constructed over a network. In some cases, like for the think tank networks discussed in Chapter 3, communities are known in advance. However, this is often not the case, and researchers often therefore rely on unsupervised "community detection" approaches. In this chapter, I use the Louvain method on the Pravda network case study Blondel et al. [2008]. The Louvain method is a greedy, modularity-based algorithm that partitions a network into communities by iteratively optimizing the modularity objective, which measures the density of edges within communities relative to a null model of random connections. It remains one of the most popular and well-studied community detection algorithms and is widely used in largescale network analysis (e.g., Sánchez et al. [2016], Sliwa et al. [2024].

6.4 Data

I use an updated version of the think tank webgraph introduced in Chapter 3. Following the same methodology, I re-extracted webgraphs for each think tank in May 2025, along with SEO attributes for both the think tanks and their backlinking domains using the SEO toolkit Ahrefs[1]. In Chapter 2, I examined European, American, Russian, and Kremlin-linked "pseudo" think tanks aimed at influencing Western audiences. In this work, I will compare our proposed Web-BEND community metrics with the original proprietary BEND metrics. However, computing the original metrics with the ORA and Netmapper toolkits requires both textual and network inputs. To ensure comparable evaluations, I exclude the Russian think tank group, since those websites are predominantly in Russian and target domestic audiences Williams and Carley [2023]. I further exclude three U.S. think tanks that are no longer active or no longer operate as think tanks: FreedomWorks, American Economic Freedom Alliance, and Concerned Veterans for America. For the remaining websites, I scraped homepage headlines on June 2, 2025 and use these as the textual input for the original BEND metrics (which require text). As the original BEND metrics were originally developed for social media data, this requires the simplifying assumption that headlines serve as a reasonable

[1] ahrefs.com

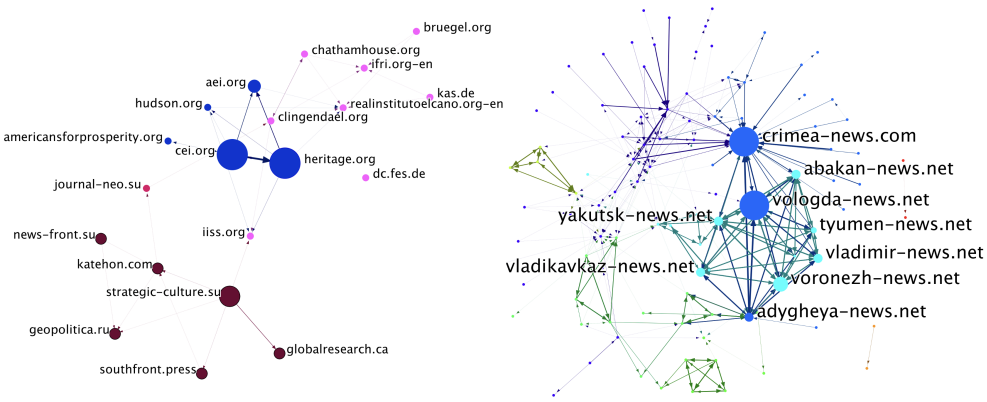


Figure 6.1: Left: Links between think tanks only with websites sized by degree, colored by grouping, and edge-width scaled by logged link volume. Right: Links between 99 Pravda sites only, colored by Louvain groupings. Nodes are sized by in-degree and nodes with in-degree of over 1,000 have visible labels.

proxy, given their brevity and similarity in form. Alternative approaches, such as using full articles or reformulating headlines into social media-style posts with a large language model, are possible; however, since the focus of this analysis is on community-level metrics, such extensions are considered beyond the scope of this work. A visualization of the resulting network is provided in Figure 6.1.

Second, I extract webgraphs and SEO metrics for the Pravda Network, a network of 193 domains which have published millions of articles in line with Kremlin geopolitical interests Klen [2024]. While there are websites in the network that target many different regions and languages, I chose to only consider the 99 Pravda websites that have at least one link to or from at least one other Pravda website. Unlike the think tank network, I do not have starting categories into which each of the target websites can be grouped. I therefore use the Louvain Method Blondel et al. [2008] on this network of 99 Pravda websites (excluding non-Pravda backlinking websites) and treat the resulting communities as website groups.

6.5 Methods

In this section, I outline equations for identifying the eight Web-BEND community maneuvers 6.1 in a webgraph setting. Web-BEND assumes user-defined groupings of target websites; these groupings could be based on region, shared characteristics, or generated through unsupervised clustering algorithms. While our definitions are framed around think tanks, they generalize to other contexts. Web-BEND metrics consider both global network structure

and group-level connections between websites of interest (Figure 6.1). Let the set of think tank websites be denoted by $W = \{w_1, w_2, \ldots, w_{20}\}$, partitioned into three groups $C = \{c_1, c_2, c_3\}$, where each $c_j \subset W$. For each target website $w_i \in W$, define the set B_i as the collection of websites that most frequently link to w_i. I also make the strong assumption that all websites within a group have the same incentives and are likely to co-amplify one another. For example, I assume that any Pseudo think tank would favorably promote any other Pseudo think tank. However, it is important to note Web-BEND metrics are descriptions that map to observable effects. Neither the original nor Web-BEND maneuvers do not necessarily imply intent on the part of social media users or website owners. The aim of the maneuvers is help analysts quantitatively characterize patterns in online information environments.

6.5.1 Positive Community Maneuvers

Back

Back are actions that increase the importance of a website relative to a community or topic. To measure *Back* in a webgraph setting, I consider the volume at which a target website w_i links to other target websites within its group c_i relative to how often it links to all other think tanks. Specifically, for a website $w_i \in c_l$, I define:

$$\text{Back}(w_i) = \frac{\sum_{w_j \in c_l,\, j \neq i} \text{links}(w_i \to w_j)}{1 + \sum_{w_k, k \neq i} \text{links}(w_i \to w_k)}$$

where $\text{links}(w_i \to w_j)$ denotes the number of links from w_i to w_j. The numerator represents the number of outlinks from w_i to other members of its own group, while the denominator captures outlinks from w_i to all other think tanks, with 1 added for smoothing.

Build

Build are actions that create the appearance of a group or community. Building websites to amplify a set of target websites or paying a third party to engage in link-scheming both accomplish this end. I define *Build* as the average Jaccard similarity between B_i, the set of domains that link to w_i where $w_i \in c_l$, and B_j, the set of domains that link to each $w_j \in c_l$. I define Build as:

$$\text{Build}(w_i) = \frac{1}{|w \in c_l| - 1} \sum_{\substack{w_j \in c_l \\ j \neq i}} \frac{|B_i \cap B_j|}{|B_i \cup B_j|}$$

Bridge

Bridge are actions that build connection between groups or create the appearance of such a connection. Bridge can therefore be defined as the proportion of links from think tank $i \in c_l$ to think tanks outside its group $j \notin c_l$. For a target website $w_i \in c_l$, define:

$$\text{Bridge}(w_i) = \frac{\sum_{w_j \notin c_l} \text{links}(w_i \to w_j)}{1 + \sum_{w_k, k \neq i} \text{links}(w_i \to w_k)}$$

where $\text{links}(w_i \to w_j)$ denotes the number of links from w_i to w_j. The function captures outlinks from w_i to members think tanks of other groups normalized by its outlinks to all think tanks.

Boost

Boost are discussion or actions that increase the size of a group and/or the connections among group members. A *Boost* activity in a webgraph might be paying a third-party SEO service to co-amplify a set of websites to increase their relative authority. To measure *Boost*, I consider the percentage of in-group co-amplification. In other words, I consider what percentage of backlinks to a target think tank come from websites that link to other think tanks in the same group. This definition of Boost aims to uncover signals that would be indicative of within-group co-amplification. However, if all of these backlinks were "toxic", this could actually be a coordinated attempt to harm website rankings. For this reason, Boost should be read alongside Negate, for a more complete characterization of within-group backlinking behavior. For a website $w_i \in c_l$, let B_i be the set of websites that link to w_i, and let

$$b' = \{b \in B_i : \exists w_k \in c_l \setminus \{w_i\},\ \text{links}(b \to w_k) > 0\}$$

be the subset of backlink sources that link to w_i *and* at least one other website in the same community. I define *Boost* as:

$$\text{Boost}(w_i) = \frac{\sum\limits_{b_c \in b'} \text{links}(b_c \to w_i)}{1 + \sum\limits_{b \in B_i} \text{links}(b \to w_i)}$$

6.5.2 Negative Community Maneuvers

Negate

Negate are actions that decrease the importance or effectiveness of a website relative to a community or topic. Common negative SEO strategies to do this include linking to broken

pages of competitors or paying third-party services to provide low-quality backlinks from "toxic" websites to harm a competitor's structural authority rankings, but can also result from coordinated link-building practices that rely on networks of low-quality domains. If a target website were being acted upon by a *Negate* maneuver, one would expect to see a preponderance of backlinks coming from low-quality domains. For this metric, I use Ahrefs' Domain Rank, which is a measure of website authority that ranges from 0-100 where 100 is the maximum possible domain authority. I note that this metric is similar to that of Boost, but is more concerned with the quality of backlinks. If a website were to have high Boost and high Negate, that would suggest that websites in a group were being co-linked by low-quality backlinks. For website w_i, let D be the unique domain authorities of all backlinking websites in B_i and let $d(b)$ be the domain rating of backlinking domain $b \in B_i$. To ensure that high values of *Negate* correspond to more of the maneuver, I define *Negate* as the expected value of the Domain Rank of B_i normalized between 0 and 1 subtracted from 1:

$$\text{Negate}(w_i) = 1 - \frac{\sum_{k \in D} k \cdot \sum_{\substack{b \in B_i \\ d(b)=k}} \text{links}(b \to w_i)}{100 \sum_{b \in B_i} \text{links}(b \to w_i)}$$

Neutralize

Neutralize are actions that cause a group to be, or appear to be, no longer relevant. In a webgraph, this could be losing links from websites in your group between two time periods. For a target website w_i, I define *Neutralize* as the share of links lost from other domains in its group as a share of total links lost from all of its back-linking domains B_i between two arbitrary time periods t_1, t_2:

$$\text{Neutralize}(w_i) = \frac{\sum_{w_j \in c_l} \text{links}_{t_1}(w_j \to w_i) - \text{links}_{t_2}(w_j \to w_i)}{1 + \sum_{b \in B_i} \text{links}_{t1}(b \to w_i) - \text{links}_{t2}(b \to w_i)}$$

Narrow

Narrow are actions that cause a group to be, or appear to be, no longer relevant. For this metric, I consider the entropy of the backlink distribution. To gain authority in webgraphs, it helps to receive high volumes of links from a wide set of domains. If links are evenly distributed across all backlinking websites, I speculate that the website is more likely to cover diverse topics or be an authority on a single topic. Conversely, if nearly all backlinks come from a single website, one could reasonably assume that the website covers fewer topics or is

not as authoritative. For a website w_i with backlinking websites B_i, I define a distribution p_j over all backlinks and then calculate *Narrow* by subtracting the normalized entropy of p_j from 1:

$$\text{Narrow}(w_i) = 1 - \frac{\sum_{b \in B_i} p_b \log p_b}{\log |B_i|}, \quad \text{with } p_b = \frac{\text{links}(b \to w_i)}{\sum_{b \in B_i} \text{links}(b \to w_i)}$$

Neglect

Neglect are actions that decrease the size of the group and connections among group members, or the appearance of such. For a target website w_i, I define *Neglect* as the number of backlinks lost from backlinking websites B_i between time periods t_1, t_2 over the total number of backlinks of w_i in t_2. While this metric is not theoretically bounded between 0 and 1, in practice, I did not observe any instances of websites losing more than their total links in the second time period. I define *Neglect*:

$$\text{Neglect}(w_i) = \frac{\sum_{b \in B_i} \text{links}_{t_1}(b \to w_i) - \text{links}_{t_2}(b \to w_i)}{\sum_{b \in B_i} \text{links}_{t_2}(b \to w_i)}$$

6.6 Results

6.6.1 Think Tank Network

In this section, I examine the differences between our proposed Web-BEND metrics and the original BEND community metrics, and I demonstrate the face validity of our proposed metrics. As the original BEND metrics are proprietary, I can only compare the outputs of the two approaches. In the ORA-PRO software, the original BEND metrics are reported as counts. To facilitate comparison, I min-max normalize each of the original BEND metric calculations. I provide a visualization of the distribution of differences between the proposed webgraph and the original BEND metrics in Figure 6.2. I find that the three websites with the most substantial mean absolute changes in score magnitude are all Pseudo think tanks: globalresearch.ca ($\mu = 0.54$), southfront.press (0.49), and news-front.su (0.45).

I observe that our proposed Web-BEND metrics better capture the phenomenon that they aim to measure in webgraphs. In the original BEND metrics, *Back* was 0 for all domains, but as Figure 6.1, (left panel) clearly demonstrates, there are think tanks within each group that link to one another. Our *Back* metric is only 0 for the 9 think tanks that do not link to

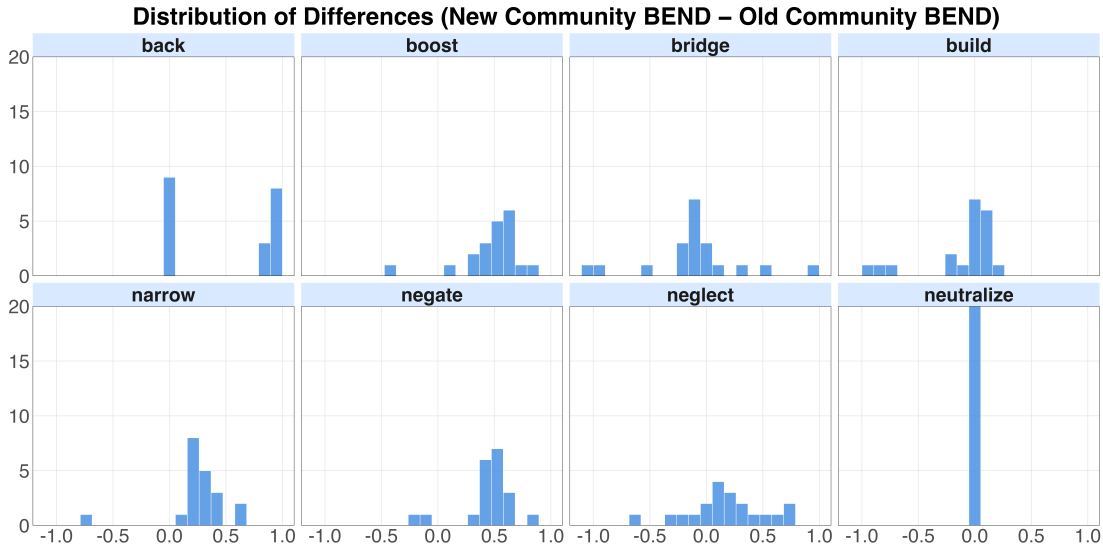


Figure 6.2: Distribution of Differences (Web-BEND - Original BEND) across all think tanks on each of the Community-level BEND Metrics.

any other think tanks within their group and is highest for websites like cei.org, which only link to think tanks inside their group (1.6k links to other American think tanks). Similarly, in the original BEND calculation, globalresearch.ca and southfront.press had the two highest *Bridge* scores despite not linking to any think tanks outside their community. As our proposed *Bridge* calculation considers links external to communities, both globalresearch.ca and southfront.press receive *Bridge* scores of 0.

Further, I find that our metrics identify several signals of information environment manipulation around the Pseudo think tanks, in line with findings from Williams and Carley [2023]. Our *Boost* calculation, which incorporates within-community backlink co-amplification, i.e., the backlinking websites that link to multiple think tanks in a given group, was highest on average for the Pseudo think tanks, with 3.4k backlinking websites that linked to two or more Pseudo think tanks. This was slightly more than 3.3k websites that linked to two or more EU think tanks and 2.8k for US think tanks. Within this network, *Boost* is strongly negatively correlated[2] with the number of backlinks received from .gov ($\rho_s = -0.61$, $p = 0.0039$) and .edu ($\rho_s = -0.74$, $p = 0.0002$) domains, indicating Pseudo think tanks as a group receive fewer co-amplifying links from authoritative official websites.

Negate explicitly considers the domain-level PageRank authority of backlinking websites (Domain Rating), which Ahrefs presents on a scale of 0 to 100 where 100 is the maximum

[2]I report Spearman's rank correlation.

Domain Rating. Websites linking to Pseudo think tanks had the lowest average Ahrefs Domain Rating (31.33)—and consequently the highest *Negate* scores. On average, websites linking to Pseudo think tanks had Domain Ratings 10 points below both average US backlinking Domain Ratings (43.39) and average EU backlinking Domain Ratings (45.78). These patterns are captured in our Web-BEND *Negate* metric, which expresses the normalized expected value of backlinking Domain Ratings subtracted from 1; Pseudo think tanks have the highest average score (0.66), followed by US (0.61) and EU (0.57). These Negate scores indicate that pseudo think tanks operate within backlink environments dominated by low-authority domains, consistent with prior findings of coordinated SEO activity involving networks of sites created primarily to generate inbound links Williams and Carley [2023]. However, this metric alone does not help determine if those low-quality domains were intended to promote or harm the overall structural authority of the website.

6.6.2 Pravda Network

I next apply our metrics to the Pravda network to demonstrate the face validity and usefulness of our proposed metrics in a setting where website communities are unknown. As a result of constructing groups using the Louvain method, Bridge is low for all groups on average (less than 0.22). Within the Pravda network visualization (Figure 6.1, right panel), I see what looks like both link farming (sets of websites all boosting one another, i.e. cliques) and link scheming (attempting to use smaller sites to boost a target website). Our metrics capture these phenomena and provide deeper insights; *Back* quickly identifies groups with clique-like structures with the top 3 groups having average *Back* scores of 0.98, 0.78, and 0.75 respectively. The 10 nodes with the highest internal degree fall into 2 of these groups. The group containing sites with the largest in-degree (Figure 6.1, right panel) (crimea-news.com and vologda-news.net), has the highest average *Boost* score (0.81), surfacing signs of co-amplification in backlinking domains across the entire network. This group also has the highest *Negate* score (0.83), reflecting that backlinking websites in this group are lower quality on average than backlinking sites in other groups. Again, Web-BEND metrics exhibit face validity and surface insights useful for information environment analysts.

6.7 Real-world Training Simulation

The Web-BEND framework and metrics have been integrated into the ORA-PRO software Carley et al. [2018] and deployed in trainings to academics, industry professionals, and members of NATO. In this chapter, I focus on OMEN, an information environment simulation [King et al., 2021] in which several dozen domain experts interact with synthetic data constructed around complex geopolitical scenarios and must derive analytic judgments and intervention strategies Morgan et al. [2025]. This section describes one of several updates to make the simulation more realistic Marigliano and Carley [2025], specifically the use of LLMs for large-scale synthetic website generation.

I developed a semi-automated pipeline for generating synthetic news articles designed to emulate a heterogeneous online media ecosystem with varying levels of credibility, editorial stance, and narrative framing. The generation process is conditioned on scenario metadata, narrative descriptions, and event information. The pipeline consists of four primary components: (1) scenario and narrative ingestion, (2) synthetic outlet construction, (3) narrative–event conditioning, and (4) article generation using a large language model.

6.7.1 Simulation Specification

Scenario and Narrative Ingestion

The generation process begins by loading structured scenario data. Event metadata within the scenario are ingested, as is each narrative around the event. For example, an event might be an unintentional military ship collision, and narratives might be things like: "ship A was responsible", "ship B was responsible", "the collision was because sailors are always intoxicated", etc. Different groups of users in the scenario spread different narratives. I create buckets of these narrative patterns and the system retrieves the narratives associated with that group and constructs narrative–event pairs. When a narrative is associated with multiple events, the pipeline generates a pair for each combination. Each pair therefore represents a specific contextual configuration consisting of (i) the narrative title, (ii) the narrative description, (iii) the event title, and (iv) the event description. These pairs serve as the semantic inputs for downstream article generation.

Synthetic News Outlet Construction

To simulate an information ecosystem containing outlets of varying reliability, the system programmatically generates a collection of synthetic news websites. Each outlet is characterized by four attributes: domain name, reliability classification, editorial stance, and target audience. Websites are assigned reliability labels according to a user-specified distribution across three categories: *reliable*, *unreliable*, and *conspiracy*. The distribution determines the number of outlets instantiated within each category. Domain names may either be generated automatically or specified explicitly in advance. For each outlet, the system constructs a short textual description that reflects its editorial style and credibility level. For example, outlets labeled as reliable emphasize factual reporting and conventional journalistic standards, whereas conspiracy-oriented outlets are described as promoting speculative or conspiratorial interpretations of events. These outlet descriptions are subsequently incorporated into article-generation prompts to induce stylistic variation across sources.

Narrative–Event Conditioning

Prior to article generation, narrative and event metadata are transformed into structured textual representations. Each narrative block contains the narrative title, optional description, stance information, and references to related events. These elements are concatenated into a formatted prompt segment that contextualizes the narrative framing expected in the generated article. For each stance group, the pipeline iterates over all narrative–event pairs and combines them with the corresponding outlet metadata. This process produces a set of candidate article-generation contexts, each defined by a unique outlet, narrative, and event combination.

Article Generation

Articles are generated using a large language model; I designed the system using GPT4o-mini, but any LLM could be used in principle. For each generation context, the model receives a prompt containing: (i) a high-level scenario description, (ii) the narrative and event information to be incorporated into the article, (iii) the synthetic outlet description, including reliability classification and editorial stance, and (iv) explicit instructions specifying the desired output structure.

The model generates a headline, an article URL, and a short article body consisting of approximately three to four paragraphs. Prompt instructions further direct the model

to produce text consistent with the editorial style and credibility level of the outlet. For example, conspiracy-oriented outlets may incorporate speculative language or narrative keywords associated with conspiratorial framing. The model returns a structured JSON response containing the generated headline, URL, and article text. Each article is assigned a timestamp, typically spaced at fixed intervals to simulate a chronological stream of news content.

We use the following LLM template, where each field in brackets is a predefined selection (reliability), something the user generates at another step (stance, audience, website description, dredge words), or information retrieved from the broader scenario construction (scenario, narratives):

```
Generate a news article for a {reliability} news website with {stance} stance
targeting {audience} audience.

Scenario: {scenario}

Narratives and Events to incorporate: {narratives}

Website description: {website_description}

Generate: 1. A compelling headline 2. A realistic article URL 3. Article
text (3--4 paragraphs) that incorporates the narrative descriptions and event
details

Return as JSON: {"headline": "Your headline here",
"url": "https://example.com/article-url",
"text": "Your article text here..."
}

Additional condition (only for conspiracy sites):

Incorporate these keywords naturally and without special formatting:
{dredge_words}
```

We found that generating website descriptions appears to improve the realism of articles, but including so much context can cause the LLMs to spend too much time summarizing the entire scenario rather than reacting dynamically to an event, as would be desired. Adjusting the prompt and context provided would likely be fruitful avenues for future research.

Output Storage

Generated articles are serialized into JSON files and stored within a scenario-specific directory structure and then are mapped to the articles themselves are written to HTML format for

the excercise. Each output file contains metadata identifying the stance group, generation timestamp, and the full set of generated articles.

6.7.2 OMEN 2026 Examples

The system just described was used as part of the exercise 2026 OMEN information environment training exercise about international conflict after naval ship collisions in the arctic. In creating stance groups, the user can enter or have an LLM generate dredge words, based on the description of the stance group, that will be referenced by both conspiratorial articles and by a class of social media bot called "dredgers". Below are some of the dredge words used by a conspiracy theory stancegroup to promote conspiracy theories throughout the simulation:

> ice lies; arctic secrets; buried under ice; northern cover-up; frozen truth; silent arctic; five eyes many lies; cold war hot secrets; nazis beneath the north; hollow earth hidden kings; blood on the ice; trafficked in the tundra; war beneath the world; fish kill cover-up; silent spill slaughter; poisoned seas plot; dead fish don't lie; murder in the water; spill and krill; false flag flood; nazi wake-up; extinction by design; diplomat death diversion; assassins in plain sight; maritime death zone

Articles were displayed on simple HTML backgrounds with a single photo appearing on a fraction of the articles. A set of photos were synthetically generated or collected manually, alongside text descriptions, and mapped to articles post-hoc by semantic similarity. An example article header is shown in Figure 6.3 and an example article is shown in Figure 6.4. Bots with aligned stances that are active on the same narratives would draw from the same list of dredge words to produce tweets conditioned on the scenario like the following:

> Why are Russian expeditions suddenly swarming the Arctic? Something's being hidden. They're digging for what's buried—the *hollow earth hidden kings* and that old Nazi base they don't want anyone to find. Wake up.

6.7.3 OMEN Survey Results

During the February 2026 OMEN exercise, held for information environment experts and analysts, participants were administered 5 yes/no survey questions. Each of these questions aimed to assess the helpfulness of websites within the exercise and of dredge words. On all 5 survey questions, the majority of respondents ($n = 19$) expressed that websites were a beneficial addition to the exercise (Full results and questions in table 6.3). The two questions

Figure 6.3: Title and cover image of an example article generated from the scenario

with the highest positive responses were "is the concept of dredge words valuable for your analysis" (17 Yes, 2 No) and that the presence of synthetic websites in the training data enhanced understanding of the overall information environment (15 Yes, 4 No). It appears that the work in this thesis is of value to analysts.

Table 6.3: OMEN survey results about the usefulness of websites and dredge words within the simulation

Question	Yes	No	Percent Yes
Do you mostly agree or disagree with the following statement: "The presence of websites in the data enhanced my understanding of the overall information environment?"	15	4	78.95%
Do you mostly agree or disagree with the following statement: "The websites were helpful for interpreting claims or narratives within the dataset?"	13	6	68.42%

Question	Yes	No	Percent Yes
Did the inclusion of websites improve your ability to evaluate the credibility of information in the dataset?	14	4	77.78%
"Dredge words" are keyphrases for which unreliable websites rank highly (and have no reliable competition). Did you notice hyper-SEO-sounding phrases used by unreliable websites in your analyses?	10	9	52.63%
Is the concept of "Dredge words" valuable for your analyses?	17	2	89.47%

While these survey results are positive, the small sample size and complexity of the system limit their helpfulness. In the future, it would be ideal to track and log more quantitative metrics during the training exercise. The goal of the exercise is to teach participants methods for information environment analytics, and while articles are a part of information environment, the quality of the articles does not necessarily correspond to any specific learning objective. This creates challenges in evaluation with respect to the exercise. If we assume that higher article quality relates to more attention to articles, we could track things like the number of times websites and articles are discussed in the analytical reports teams produce at the end of each day, how often article content is referenced, or even log the amount of time analysts spend on article pages. However, even this would require a more rigorous pipeline for evaluating article quality. Without time and monetary constraints, it would be helpful to track these and other metrics in two versions of the scenario: one with websites and one without to compare differences, but this comes with its own constraints.

6.8 Limitations

Data acquisition is the primary challenge in extending these metrics to dynamic settings. Neglect and Neutralize are already defined as differences between two time periods, but any metric I proposed could be presented in terms of its change. Links are constantly changing: link distributions can can shift wildly between time periods as websites shut down, move, re-organize, or emerge. However, crawling the entire internet is inherently slow and requires substantial compute. While SEO toolkit providers sell historical SEO data, it is far more expensive than the data I collected. A more economical approach would be to use Common Crawl Foundation snapshots, which are collected monthly. However, the uncompressed December 2025 snapshot was just over 400.22TB and weights must be derived manually

By ironarchive.org | 2032-06-01 08:13

As global tensions heighten in early June 2032, the Arctic has emerged as a focal point of geopolitical intrigue, intertwining with the longstanding conspiracy theories surrounding the Hollow Earth. With Russia's assertive maneuvers in the Arctic region—bolstered by a clandestine fleet and extensive operations in newly navigable waters—questions arise not only about the resource wealth hidden beneath the seabed but also about the motivations behind Moscow's aggressive posture. Observers note that the melting ice is unveiling not only potential riches but also fueling speculation about what lies beneath the surface, both literally and figuratively.

Meanwhile, China's ongoing military activities in the Western Pacific contribute to a climate of uncertainty and suspicion. Frequent incursions near Taiwan and escalating tensions with U.S. allies have created a perception of a calculated strategy aimed at normalizing aggression in the region. Critics argue that these actions are not merely routine maneuvers but part of a broader narrative to assert dominance over critical maritime corridors. This geopolitical backdrop, coupled with the Arctic's emerging significance, has ignited discussions about the implications of resource competition and environmental sustainability.

As the Five Eyes nations prepare for Operation Sea Dragon, a large-scale naval exercise designed to bolster regional security, analysts are divided over the implications of such a show of force. While some representatives insist that the operation is merely defensive, aimed at ensuring freedom of navigation and deterring potential aggression, others caution that the timing could be perceived as provocative amid the escalating Sino-Russian military cooperation. The exercise underscores a complex interplay of deterrence and provocation, raising questions about the stability of alliances in a rapidly changing global landscape.

Amid these tensions, the narrative surrounding the Arctic and the so-called Hollow Earth conspiracy continues to capture the public imagination. Theories suggesting hidden civilizations and advanced technologies beneath the Earth's crust have gained traction in fringe circles, further complicating the geopolitical discourse. While the scientific community remains skeptical, the intersection of conspiracy theories with real-world events highlights a growing trend where myth and reality converge, reflecting the anxieties of a world grappling with uncertainty and the implications of uncharted territories.

Figure 6.4: Example text from an LLM-generated article.

from historical data and there are fairly substantial compute restraints to doing analyses over time. However, this could be a promising avenue for future research. Using Common Crawl would allow for more accurate calculations of the two temporal metrics (Neutralize and Negelct), and instead of using Ahrefs' Domain Rating, a user could use normalized pagerank scores calculated over Common Crawl webgraphs. It has been shown in previous work that Common Crawl Pagerank scores correlate strongly with Ahrefs Domain Ratings Carragher et al. [2025].

A big advantage of Web-BEND over the original BEND metrics is that they are not proprietary and can therefore be used and evaluated by anyone. Further, I show that, at least in our think tank case study, Web-BEND seem to offer improvements over the original BEND metrics in a webgraph setting. However, users should be aware of several important considerations. First, the method used to define groups can significantly influence the resulting metrics. For example, using Louvain Method communities often yields low bridge scores within the Pravda network, due to the algorithm's tendency to form tightly connected clusters. This matters because our approach assumes that websites within a group have incentives to co-amplify one another—a strong assumption that may not hold in all contexts. While this assumption is reasonable in the case of Kremlin proxy sites, where incentives are likely aligned, the assumption does not hold in many settings Carragher and Carley [2024]. Our measures are defined relative to group membership and therefore cannot be used directly to assess how a site interacts with those outside its group. Additionally, while these metrics describe observable effects, there is ambiguity that results from the analyst not knowing the intent of any given website. As discussed in the chapter, Negate provides a description of backlink quality, but it does not determine whether the low-quality backlinks are a result of positive or negative SEO. Future work should examine how these metrics evolve over time and how they perform in settings where the shared incentive assumption does not hold. Additionally, I only look at community-level BEND metrics in this work; future work could propose open narrative-level BEND metrics and explore them over different website-level text sampling approaches.

Finally, there are several limitations in the synthetic article generation pipeline. The current simulation lacks realistic website interfaces and largely omits images, both of which are central to how users interpret online content. The text generation itself is also limited: many articles exhibit repetitive phrasing (e.g., "In a shocking turn of events"), while increasing diversity via higher temperature often introduces entities or events not grounded in the scenario. In addition, about half of analysts failed to identify dredge words, suggesting they

are not sufficiently salient in the current design. These issues reflect a broader limitation: there are no well-defined evaluation metrics to jointly optimize realism, diversity, and fidelity to the scenario, which constrains systematic improvement of the generated content.

6.9 Conclusion

I propose Web-BEND quantitative metrics that extend the BEND Framework to webgraph settings. These metrics are interpretable in webgraph settings and demonstrate face validity in instances where the original BEND scores fall short in webgraph settings. These metrics provide a shared quantitative language to characterize actions taken to manipulate information environments in webgraphs. By making webgraph manipulation more measurable and comparable across cases, Web-BEND enables analysts to better monitor information environments and better develop more targeted interventions. Further, I integrate the proposed Web-BEND metrics into the OMEN training exercise, along with a realistic website generation pipeline that draws from work of previous chapters.

Beyond methodological contributions, this chapter also demonstrates how these concepts can be operationalized within a realistic training environment. The proposed Web-BEND metrics were integrated into the ORA software and are being used in network analytics trainings. In addition, I created a semi-automated synthetic website pipeline for generating synthetic news articles for the OMEN information environment exercise. Results from a survey of OMEN participants suggest that the inclusion of synthetic websites meaningfully improved the training exercise. Most respondents reported that the presence of websites enhanced their understanding of the overall information environment and aided in interpreting narrative claims within the dataset. Participants also indicated that the concept of dredge words, a key contribution of this thesis, was useful. Together, these results suggest that integrating structural webgraph analysis with realistic synthetic data generation can improve both the analytical tools available to researchers and the practical training environments used to prepare analysts for real-world information operations.

Chapter 7

Conclusion

7.1 Discussion

Throughout this Thesis, I have aimed to provide a deeper understanding of how unreliable content and misinformation can spread via search engines and how we can build robust systems to mitigate this spread. I demonstrate the limitations of Google's bannering and ranking algorithms, and demonstrate the need for better algorithmic interventions to identify data voids on search engines. To this end, I demonstrate the reliability signals present in webgraph and SEO data. I further introduce the concept of *dredge words* and explore how unreliable search results in SERPs impact AI overviews. Next I put forward a unified model that leverages signals from SERPs, keyphrases, and social media data and build SoTA website reliability classification models. Finally, I demonstrate how these concepts can be integrated into the BEND framework and evaluate how expert analysts engage with the framework and concepts I introduce in a simulated information environment exercise. In this conclusion, I will attempt to answer the questions put forth in the introduction, discuss takeaways, policy recommendations, limitations, and future work.

7.1.1 How does misinformation spread on search engines?

There are several mechanisms through which unreliable content can appear in search engine results pages (SERPs). At the most extreme, one could construct a search engine that indexes only unreliable sources, as was the case with the search engine `GoodGopher.com`. In such environments, intervention is largely meaningless because the system itself is designed to surface unreliable material.

Navigational search is another pathway for which the intervention space is limited. As briefly discussed in Chapter 2, users can employ advanced search operators like `site: infowars.com` to intentionally retrieve results from a specific domain. In such cases, the user's intent is unambiguous, which makes suppression may be perceived as censorious or paternalistic. It is, in the author's view, not obvious that platforms *should* intervene in navigational search results, except in cases there is ambiguity in a query (for example, a query containing the term "blaze" may not refer to th unreliable news website *The Blaze*), or when the destination site distributes illegal content, malware, or other harmful material.

Data voids constitute a more concerning pathway through which misinformation spreads. Data voids occur when a query yields results that are irrelevant or unreliable. Data voids reflect a structural limitation of information retrieval systems: retrieval systems assume every query has a relevant response. When authoritative sources have not yet produced content on a particular topic or when the query itself is obscure, newly emerging, or was deliberately targeted by adversarial actors, the retrieval system may return unreliable content.

7.1.2 How do unreliable websites rank highly for content on search engines?

There are two key mechanisms through which unreliable websites can rank: keyphrase targeting and manipulation of structural authority signals within the webgraph. Keyphrase targeting exploits the linguistic structure of search queries. Adversarial can target highly specific or newly emerging keyphrases—particularly those associated with conspiracy, rumor, or pseudoscientific claims. I outline many examples of this in chapters 2-5. This strategy allows adversarial publishers to capture search traffic from users investigating unfamiliar or emerging topics, as was the case with obituary pirates.

A second pathway involves the manipulation of structural authority signals that search engines derive from hyperlink networks. As search ranking algorithms use links to calculate to calculate structural authority scores for websites, actors seeking to influence search visibility can therefore attempt to artificially inflate these signals by manipulating links into and out of target websites. We discuss this form of manipulation extensively in Chapters 3 and 6.

7.1.3 How do search engines interact with social media in misinformation spread?

In chapter 5, I look at the pathways users can take to unreliable websites. Users can take a "direct" path by clicking links they encounter on social media or some other website, but they can also encounter new terms or concepts on social media that they then enter into search engines. By considering these pathways, we can construct more realistic models of information diffusion and more realistic models for predicting website reliability. Misinformation research has overwhelmingly focused on social media, often without considering the ways information ecosystems extend beyond it.

7.1.4 How can Graph Neural Networks and LLMs be used to improve the reliability of search engine results?

In chapter 2, I introduced GNNs for identifying data void queries and demonstrated that they outperformed Google's identification system at that point in time. This was the first published attempt to identify data voids at scale. In chapters 3 I introduce a language agnostic graph neural network over webgraphs for predicting website reliability, an essential task when identifying data voids. I extend this model to further include incorporate query context and social media data. These simple lightweight architectures can allow search engines to better identify, downrank, and flag unreliable domains in search results. Finally, I use fine-tune RAGs in chapter 4 in an attempt to identify unreliable AI Overviews and to make AI Overviews more robust to retrieval in data void settings.

7.2 Policy Recommendations

What is most needed is transparency. Data voids are poorly understood, and exist in proprietary data that are not shared with most researchers, think tanks, or other institutions. Data sharing agreements between platforms and independent researchers would be extremely beneficial in helping to identify and mitigate data void problems. Search engines provide very limited public access to data about ranking behavior, query ecosystems, or domain-level signals. Structured research partnerships, privacy-preserving query datasets, and expanded transparency reports would enable external evaluation of search reliability and facilitate the development of independent detection tools As a result of the lack of transparency, it is hard

to quantify the extent to which data voids pose a problem or how ubiquitous they are in English as well as in other languages.

For platforms, we have three primary suggestions. First, while Google's initial bannering contained flaws, it's laudable that the company attempted to warn users about data voids. There is substantial research showing the benefits of warning labels on social media platforms, and there's no reason they shouldn't help users approach results more critically on search engines. Second, we recommend that AI search engines and companies deploying AI search also take data voids seriously. There is ongoing legal debate over whether or not AI web search and Google's "AI overviews" constitute editorials on the part of the search engine company. If that's the case, AI overviews could be liable under section 230 Ryan [2024]. Given the problems we demonstrated with AI Overviews, we argue there should be more safeguards in place for AI overviews, particularly around medical information. Finally, we would argue more attention should be paid to when and why low reliability websites appear in search results, and better models should be deployed to identify and down-rank them. I have proposed and open-sourced several approaches in this thesis that could be of use on both of those fronts.

7.3 Limitations

There are a number of limitations in this work, several of which we outline here. We don't have access to Google's Algorithm and cannot evaluate the effectiveness of our interventions. Consequently, our keyphrase explorations rely on a third-party data (Ahrefs). As a validity check, we verify that each website appears on the top SERPs for each extracted query and perform additional validation checks on the data [Carragher et al., 2025]. Future work in this area would benefit from large-scale user data, like the data being collected at the National Internet Observatory Feal et al. [2024]. Additionally, while Keyphrases are predominantly (though not exclusively) in English, our structural authority models are designed to be language-agnostic and our keyword based models can be extended to other languages. Another limitation is that we rely on domain-level reliability metrics for our systems, which are not fixed and not temporally-robust. While I touch on discovery systems in chapter 5 and in other work, the dynamic nature of online environments makes auditing challenging. As discussed in Chapter 3, websites can shut down, move, or change domain names, meaning reliability lists can quickly become outdated or obsolete.

While the dredge word sampling procedures proposed are novel and would, in principle,

generalize to any traditional search engine, the nature of search is changing. With AI search, queries are often getting far more specific and being provided much more context. Information retrieval remains an important component of grounding LLM responses, but there's no guarantee the short queries we focused on will be representative of queries in the LLM age. Additionally, I do not explicitly consider multi-modal search. When users are searching for images or videos using text queries, data voids can clearly still be present, but it's less obvious how systems could respond with multimodal queries that many VLMs now support. These directions would both provide compelling directions for future research.

One of the largest limitations is the lack of temporal coverage for dredge word analyses. Rankings were extracted for a moment in time, but as rankings are always changing, many of these data voids may have been filled, and new ones arisen. This also means we did not study rapidly-changing result data voids, as was the case with obituary pirates. Nobody has yet studies this at scale and it I would argue it is an important area for future research. Additionally, SERPs change constantly and can change substantially after large changes after Core updates, which may impact replication. To mitigate this limitation we release all data collected for all experiments.

We do not devote sufficient attention to Google's advertising on webpages. Some ads observed over the course of these projects have been deeply concerning, including a fake charity advertising on hurricane disaster keywords that we reported to relevant government agencies. Further, I lacked the space to sufficiently engage with valid criticism from other disciplines. For example, the ways that search engines encode and amplify racial and gender hierarchies Noble [2018], Sweeney [2013] or the ways in which search engines and AI search engines launder biased and partisan content as factual Beers [2025], Tripodi [2022].

As the reader may have noticed at multiple points in this dissertation, misinformation and disinformation can be profitable. Workers can be paid to produce it by ideological or for-profit organizations, or workers can produce it to enrich themselves (e.g., obituary pirating). I focus on the roles that algorithms play in this ecosystem, but influencers, crowds, incentives, and poverty are all an important and under-discussed dimension of the broader overall phenomenon. More interdisciplinary work is needed to flesh out how different incentives shape search information environments.

7.4 Future Work

7.4.1 Extending the work

A better understanding the paths that users take across search engines, social media, and now LLMs. This work focuses heavily on Google, with one chapter exploring dredge word usage on Twitter. To better understand broader information environments, it would be worthwhile to explore other social media platforms (e.g., Reddit, Facebook, etc.) and other search engines (e.g., Bing, Perplexity, etc.). Researchers have demonstrated that users can prioritize search engines for different things, and that users seeking misinformation often promote search engines other than Google Williams and Carley [2025]. Finally, little is known about how LLMs and AI search engines fit into this broader information environment.

7.4.2 Temporality

All of my research so far has largely considered data voids that are static, but many data voids, including obituary pirates, are known to be dynamic. In the early hours after an event, before reliable coverage, unreliable websites can fill data voids. During the horrific school shooting in Uvalde, officials released the name of the perpetrator, and almost immediately, far-right Twitter personalities immediately began wrongly accusing the shooter of being a transgender person in Mexico. Google was ill-equipped to handle situations like this, and for hours, as I searched the shooter's name, the top image result was an innocent woman holding a transgender flag. I'm really interested in how harmful data voids can be discovered and combatted in real time; such a system could substantially improve the reliability of both websearch and social media.

7.4.3 RAGs

As AI search engines and LLM chatbots are increasingly used for information retrieval, more work is needed to understand how the information LLMs receive shapes the worldviews they espouse. It's well understood that RAGs prioritize "retrieved knowledge" (retrieved documents) over "parametric knowledge" (knowledge from pre-training), so when, for example, alt-tech social media platforms build LLMs that seem to only retrieve articles from dubious sources, it can treat those documents as fact and in turn shape an alternate reality on the platform. This has important downstream implications for polarization, democratic norms, and general information health.

7.4.4 Combatting Data Voids

In my research I've focused largely characterizing and discovering data voids, but the natural follow-up question, and one which requires institutional buy-in, is "how can we fill them?" I've been in discussions with medical researchers at Stanford about how data voids might lead people to medical misinformation, and the ways in which AI could help us 1) map out medical data voids and 2) proactively fill data voids in partnerships with reliable institutions. Reliable institutions, which are favored in search results, could play an important role in creating a healthier information ecosystem.

Appendix A

Appendix A

A.0.1 Children's Immortality Project

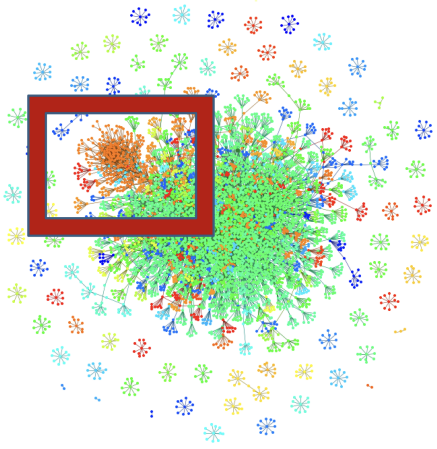
Among the set of 301 queries that produced a low-quality banner in wave 1, 141 appeared to be aimed at directing users to a network of strange websites aimed at boosting awareness of the "Children's Immortality Project" (CIP), which argues that children only die because adults teach them about death (see Appendix A.0.1). These queries often contained quoted keywords present in websites and were often surrounded by quotes (e.g., "creating all children to die", "the virtue bios and mortality resolution" and clarity genius of sedona "interneted"). Given that these CIP queries generated a large portion of the observed low-quality banners in wave 1, we also annotated whether each of those queries produced a SERP containing websites related to the CIP. We found substantial agreement between annotators on this task ($\kappa = 0.80$), and the GNN_{Hom} model identified the most CIP queries in its top 20 predictions (16), followed by the GNN_{Het} (15) and DistilBERT (5) models.

The websites associated with Children's Immortality Project (CIP) often claimed that Google was censoring them, and sought to create SEO websites so that more people would learn about CIP. We found hundreds of keyword-heavy SEO sites related to CIP that heavily linked to one another and often had questionable domain names (e.g, howtokillchildrenlegally-breed.blogspot.com[1] and keywordsarefortakingovertheinternet.blogspot.com[2]). CIP has also been discussed at length on forums like 4chan [3]

[1] https://web.archive.org/web/20240411055816/http://howtokillchildrenlegally-breed.blogspot.com/

[2] https://web.archive.org/web/20240605114527/http://keywordsarefortakingovertheinternet.blogspot.com/

[3] Explicit content: https://web.archive.org/web/20100728210054/http://4chanarchive.org/



(a) Graph of the domains returned in low-quality SERPs and the 10 websites that 10 domains that most frequently link to each of them.

(b) The old landing page of the google-bomb.com, one of the Children's Immortality Project Websites, available at: https://web.archive.org/web/20141217044958/http://google-bomb.com/

Figure A.1: Children's Immortality Project SERP network and landing page.

For each of the domains returned alongside a low-quality banner, we extracted the 10 domains which link most frequently to the target domain using Ahrefs. We used this data to create a bipartite network of second and top level domains to queries and ran Louvain community detection on the resulting graph. One cluster stood out as highly interconnected and contained many of the domains amplifying the CIP (Figure A.1). While we demonstrate that the backlink network displays some signals of coordination, better understanding how coordination manifests in search directives would be a promising avenue for future research. We note that not all the queries associated with the CIP required a banner at the time of this writing, as some were related to keywords whose search rankings have become highly competitive over time, such as "weaponized artificial intelligence."

Extended Data

brchive/dspl_thread.php5?thread_id=3595190&x=Childrens+Immortality+Project

Table A.1: Warning banner types observed by wave for our 1.4M search queries.

	Any Banner		Low-relevance		Rapidly-changing		Low-quality	
Wave	N	%	N	%	N	%	N	%
Oct 2023	14,424	1.00	14,121	0.9821	2	0.0001	301	0.0209
Mar 2024	13,629	0.95	13,406	0.9323	2	0.0001	221	0.0154
Sept 2024	18,603	1.29	18,593	1.2931	10	0.0007	-	-
Feb 2025	14,342	1.00	14,331	0.9967	11	0.0008	-	-
Overall	60,998	1.06	60,451	1.0510	25	0.0004	522	0.0091

Table A.2: Paired similarity across data collection waves. We measure paired similarity by obtaining the set of URLs that appeared on each SERP, then compare those sets across consecutive waves for each query using the Jaccard index, and summarize those results across all queries. SERPs without URLs were excluded.

Wave Comparison	Mean	95% CI	Std.
Oct 2023 to Mar 2024	0.3536	(0.3533, 0.3539)	0.1830
Mar 2024 to Sept 2024	0.3373	(0.3370, 0.3375)	0.1787
Sept 2024 to Feb 2025	0.3406	(0.3403, 0.3409)	0.1868
Average	0.3438	(0.3436, 0.3440)	0.1830

Table A.3: Paired comparisons of average domain quality. Similar to our paired similarity measure, we compared the average domain quality score of the SERPs each query produced across consecutive waves and used paired t-tests to evaluate those differences (all $P < 0.001$).

Wave Comparison	Mean	95% CI	Std.	t	df	Cohen's d
Oct 2023 to Mar 2024	0.0030	(0.0029, 0.0032)	0.0571	49.4	863,711	0.053
Mar 2024 to Sept 2024	0.0021	(0.0020, 0.0022)	0.0567	34.1	832,958	0.037
Sept 2024 to Feb 2025	-0.0014	(-0.0015, -0.0013)	0.0571	-22.4	810,095	-0.025
Average	0.0013	(0.0012, 0.0014)	0.0570	35.8	2,506,766	0.023

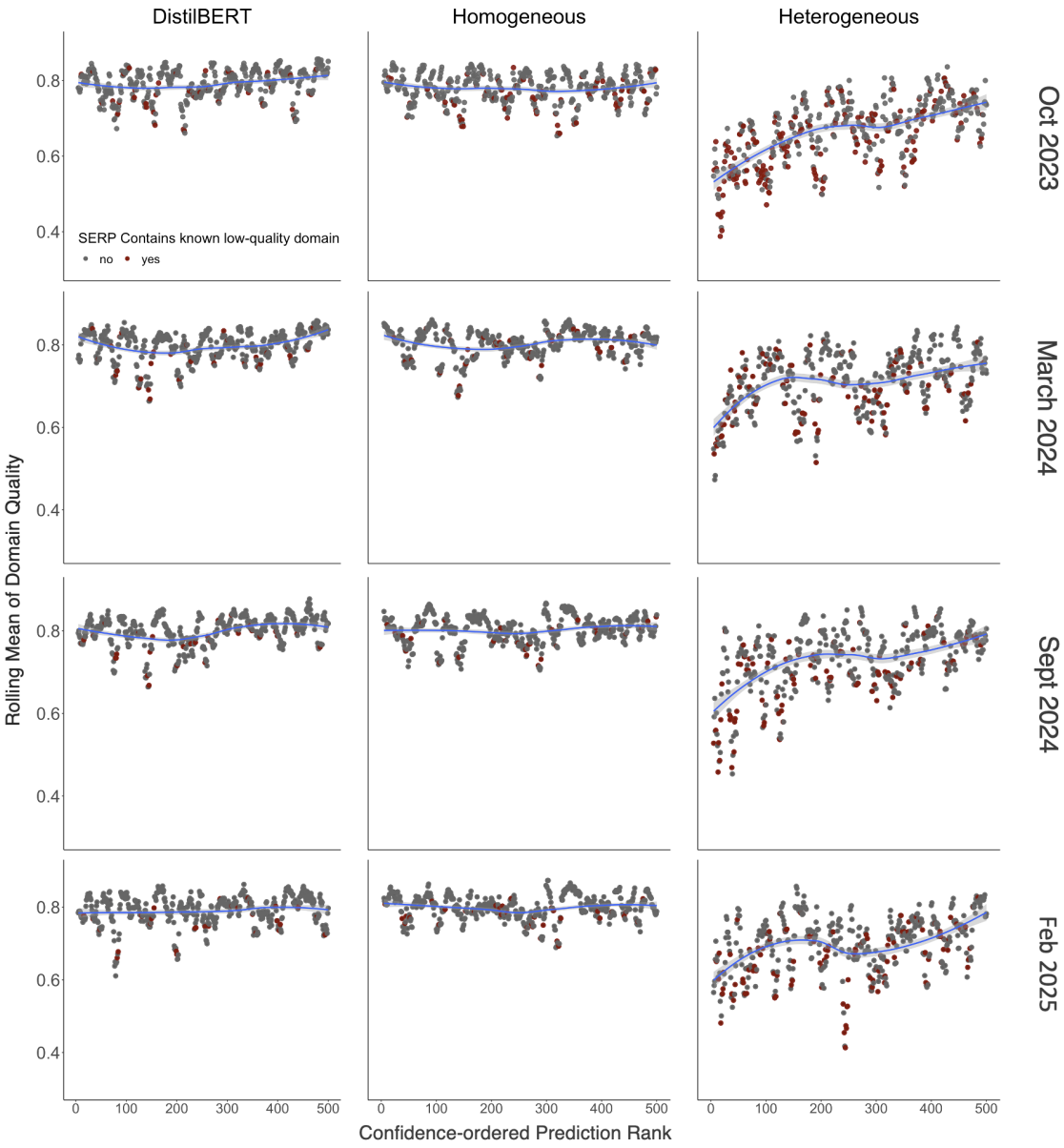
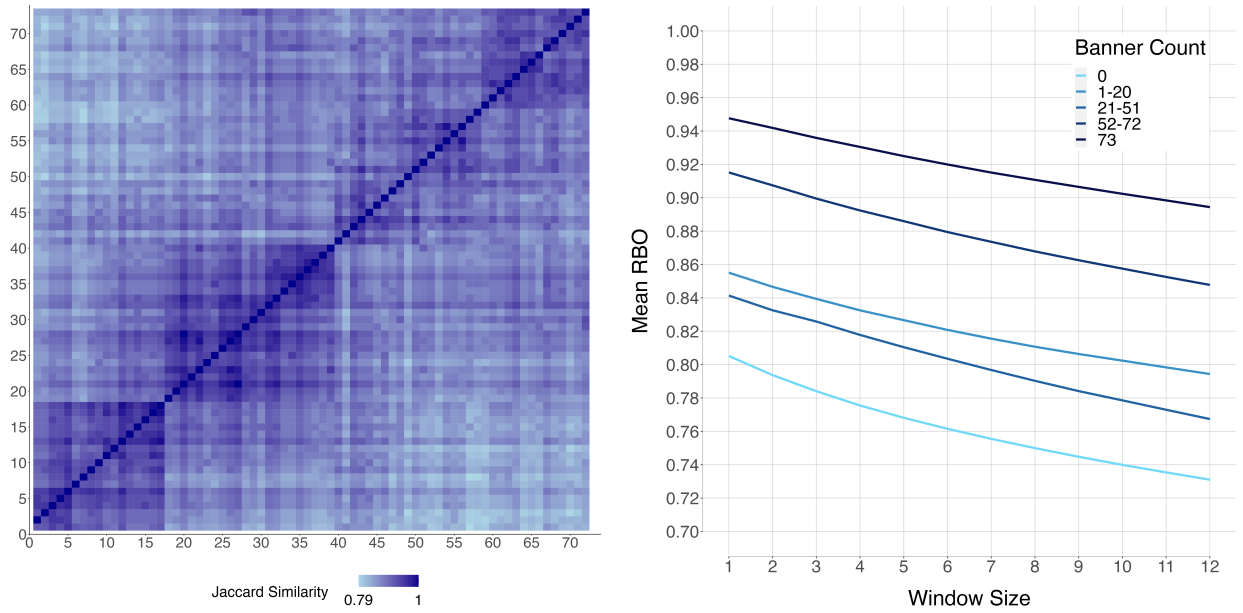


Figure A.2: The GNN_{Het} model best identified SERPs associated with low-quality domains. Each subplot displays the rolling mean (y-axis; window size of 10) of average domain quality scores over the 500 highest-confidence model predictions (lower scores indicate less reliable websites). The left-most point in each subplot is the query the model was most confident should receive a low-quality banner. Individual points show the average domain quality of the SERPs produced by each query. Points colored red returned at least one domain with a domain quality score lower than 0.5. W e evaluated low-quality banner predictions for all waves with models trained on wave 1 (Oct 2023).

Table A.4: Data void prevalence estimates based on high confidence (≥ 0.90) predictions from our validated GNN_{Het} model, and based on a low average domain quality threshold (≤ 0.5 on a scale from 0–1).

Data Void Definition	Wave	Count	Prevalence	SEM	Std.
GNN_{Het} Confidence ≥ 0.9	Oct 2023	13,404	0.932%	0.008%	9.610%
	Mar 2024	5,672	0.394%	0.005%	6.268%
	Sept 2024	7,218	0.502%	0.006%	7.067%
	Feb 2025	6,545	0.455%	0.006%	6.731%
	Overall	32,839	0.571%	0.003%	7.535%
Average Domain Quality ≤ 0.5	Oct 2023	11,977	0.833%	0.008%	9.089%
	Mar 2024	10,400	0.723%	0.007%	8.474%
	Sept 2024	9,286	0.646%	0.007%	8.010%
	Feb 2025	10,546	0.733%	0.007%	8.533%
	Overall	42,209	0.734%	0.004%	8.535%



(a) Heatmap cells show the pairwise Jaccard similarity of the set of queries that returned a quality banner over all 73 time-steps. Data collection issues resulted in gaps at steps 18 and 41, which account for the steepest plot gradient changes (see Section 2.4.6).

(b) Similarity of SERPs measured by using mean RBO_k (y-axis) over 12 window sizes (x-axis). Queries were divided into groups based on whether the SERPs they produced received a banner in 0, 1-20, 21-51, 52-72, or all 73 time steps (legend).

Figure A.3: The set of queries that received a low-quality banner frequently changed over short time spans (A.3a), and queries that consistently received banners also returned relatively consistent search results (A.3b). These data come from a supplementary dataset that we collected in June 2024 by repeatedly conducting the subset of queries that produced a low-quality banner in wave 1 on a more rapid data collection schedule (once every 4.5 hours for 73 times steps). We provide additional details on this dataset and analysis in Methods 2.3.6 and Section 2.4.6

A.0.2 Warning Banners

Detailed counts and percentages (of all SERPs) for the warning banners we collected in each wave are provided in Extended Data, Table A.1. In this section, we provide details and screenshots for Google's rapidly-changing banners (A.0.2), as well as the low-relevance banner variants we observed across waves (A.0.2). The prevalence of the low-relevance banner variants we observed across waves is available in Table A.5.

Rapidly-changing Banners

We provide a limited examination of Google's rapidly-changing banners (Figure A.4) in the main text because we only observed two instances in the first two waves and 10 instances

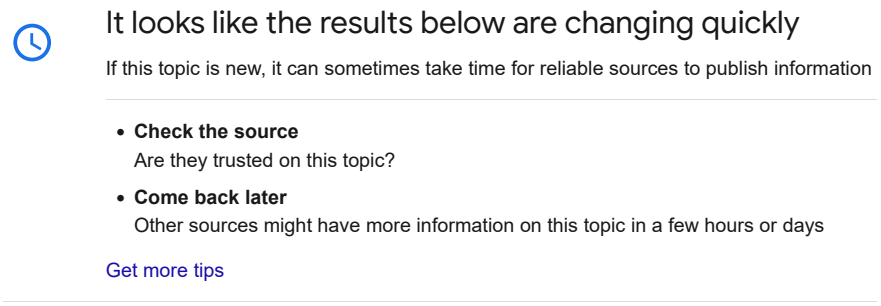


Figure A.4: Example of a rapidly-changing warning banner on Google Search. This banner is displayed when Google's systems detect "a topic is rapidly evolving and a range of sources hasn't yet weighed in" Sullivan [2021].

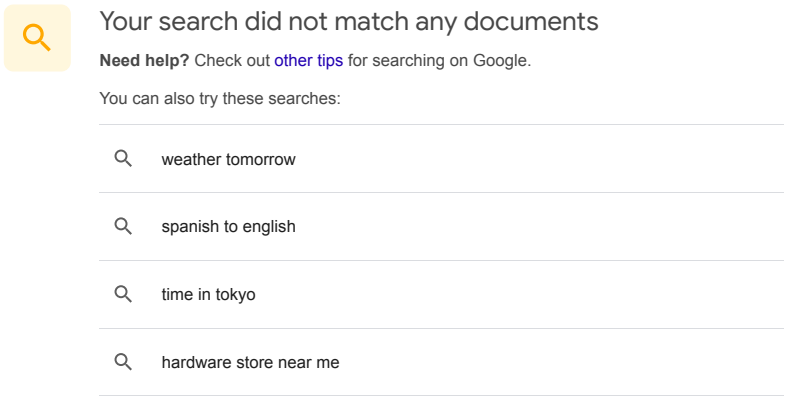
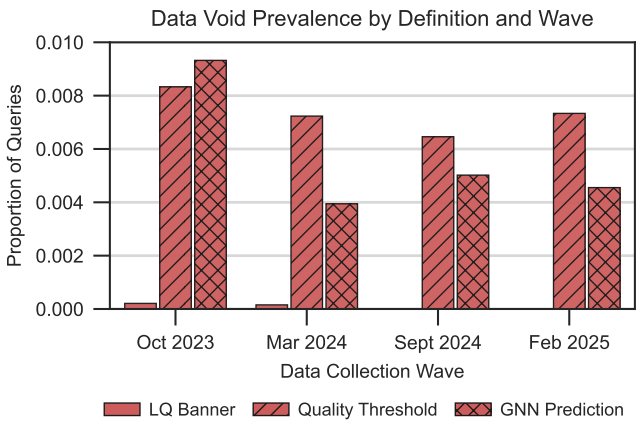
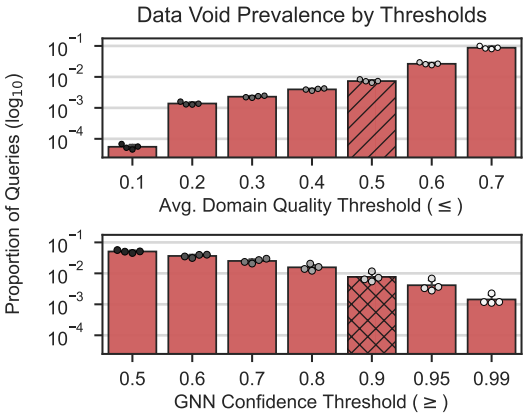


Figure A.5: Example low-relevance banner variant that we observed only in Sept 2024.

in the third wave. These 14 queries were unique with no repeats across waves. The queries triggering these banners were often political, with 11 of the 14 contained either "trump" or "biden" in the query. For wave 1 (October 2023), the queries were "biden zionist" and "dr. reiner fuellmich". For wave 2 (March 2024), the queries were "republicans are russian asset" and "trump word salad". In wave 3 (Sep 2024), the queries were: "trump people magazine 1998", "'lifelong republican"', "trump people 1998 quote", "trump republicans dumbest voters", "trump republican voters dumbest", ""trump will win election"", "donald trump people magazine quote 1998", "trump wanders away—trump wanders away", "biden with young girls", and "trump tell republicans dumbest voters". Last, in wave 4 (Feb 2025), the queries were: "#impeachtrumpnow", "37 year old arrested", "plane "blame trump"", "91 year old died", "59 year old arrested", "rodneyudell", "trump melt down", "woman charged after allegedly sexually assaulting", "trump ruano faxas", "trump ruanofaxas", and "arrest



(a) Data void prevalence (y-axis) was highest when defined by the average domain quality (x-axis) and lowest when defined by Google's low-quality banners. The legend distinguishes the three different definitions: (1) LQ Banner, searches that received a low-quality warning banner, (2) Quality Threshold, searches that produced results with average domain quality scores ≤ 0.5, and (3) GNN Prediction - searches predicted as data voids by our GNN_{Het} model with confidence ≥ 0.90.



(b) Sensitivity analysis of data void prevalence by threshold parameter for two definitions. The top panel shows prevalence by average domain quality threshold and the bottom panel shows the prevalence rates by GNN_{Het} confidence threshold. The bars with hatch patterns highlight the threshold values we used. Error bars show 95% CI and the individual points on each bar show values by wave.

Figure A.6: Data void prevalence by definition and wave (A.6a), and threshold (A.6b). Prevalence statistics for the threshold definitions are available in Extended Data, Table A.4.

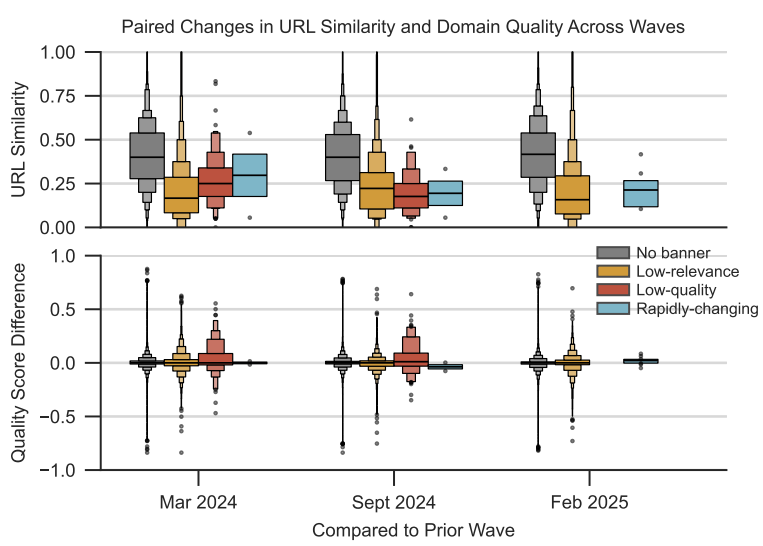
traffickin". The time-sensitive nature of these warnings is exemplified by the query "biden zionist", which received a rapidly-changing banner when conducted on October 17, 2023, shortly after the October 7 attack on Israel, but not in any subsequent waves. Although our approach provides a limited sampling of these banners, the few we surfaced largely involve political actors and should be investigated in future work.

Low-relevance Banner Variant

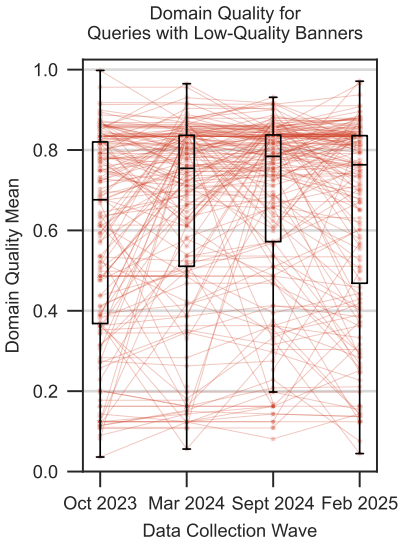
In the third wave, we observed a variant of the low-relevance banners that had not previously appeared (Figure A.5). This banner contained the same icon as the other low-relevance banners (a magnifying glass with a yellow background), but instead of stating that there were not many or any great matches, it stated that "Your search did not match any documents" and provided a consistent list of generic suggestions for alternative searches (e.g., "weather tomorrow"). This variant appeared for 6,081 queries in wave-3, accounting for 32.7% of the 18,593 low-relevance banners we observed in total for that wave. The presence of this variant was largely responsible for the overall increase in warning banner presence in that wave (Table A.5), despite low-quality banners disappearing. Although the exact language used in the banner is "Your search did not match any documents" (Figure A.5), we observed five cases in which the SERP contained an ad (e.g., Dell, ADT).

Table A.5: Low-relevance banner subtypes observed by wave across all queries. Partial view - some suggestions cropped for space.

	"Not many great matches"		"Not any great matches"		"Did not match any documents"	
Wave	N	%	N	%	N	%
Oct 2023	14,062	0.9780	59	0.0041	-	-
Mar 2024	13,348	0.9283	58	0.0040	-	-
Sept 2024	12,468	0.8671	44	0.0031	6,081	0.4229
Feb 2025	14,270	0.9924	61	0.0042	-	-
Overall	54,148	0.9414	222	0.0039	6,081	0.1057

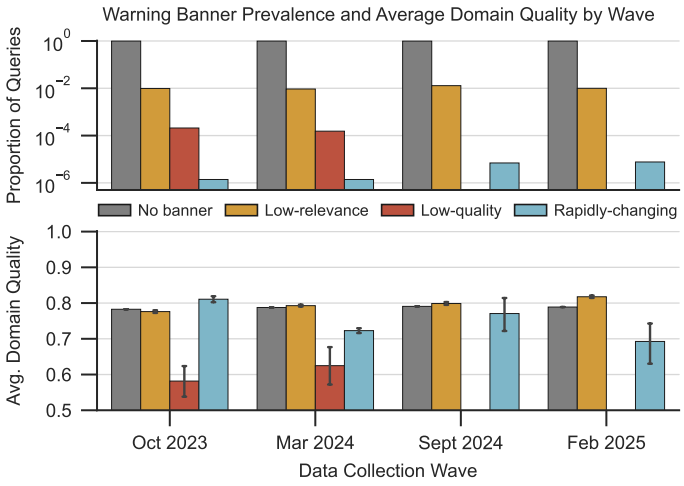


(a) Paired comparisons of SERPs between consecutive data collection waves show substantial churn in search result URLs but minimal changes in average domain quality. The top panel shows paired URL similarity (Jaccard index), indicating the proportion of shared URLs between consecutive waves. The bottom panel shows paired changes in average domain quality between consecutive waves. Legend colors correspond to the banner type observed in the first wave of each comparison (e.g., the red plots for Mar 2024 show URL similarity and quality differences for queries that received a banner in Oct 2023). In letter-value plots, boxes represent letter-value quantiles, outliers are shown as individual points, and the black line in each box shows the median Hofmann et al. [2017].

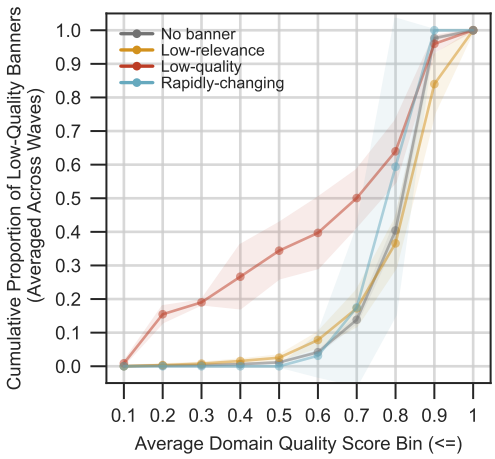


(b) Average domain quality improved among the subset of queries that received a low-quality banner in at least one wave, but many queries continued to produce low-quality results after the banners disappeared. Each red line shows the average domain quality of the search results returned for a single query over time. Boxplots summarize the distribution and median of average domain quality for this subset.

Figure A.7: Comparisons of URL similarity and average domain quality across waves.



(a) Warning banner prevalence varied by type (top), and SERPs with low-quality warning banners had the lowest average domain quality (bottom). The top panel shows the proportion of our 1.4M queries (y-axis) that received each banner type (legend) during each data collection wave (shared x-axis). The bottom panel shows the average domain quality (y-axis) of the SERPs that received each banner type in each wave. Error bars in the bottom panel show 95% confidence intervals.



(b) SERPs with low average domain quality scores account for a substantial proportion of SERPs with low-quality banners but not other banner types. The x-axis shows the upper bound of each domain quality bin, and the y-axis shows the cumulative proportion of SERPs across those bins for each banner type. Lines show averages across waves, shaded areas are 95% confidence intervals.

Figure A.8: Warning banner prevalence and domain quality. Counts and percentages for warning banner prevalence can be found in Appendix A.0.2, Table A.1.

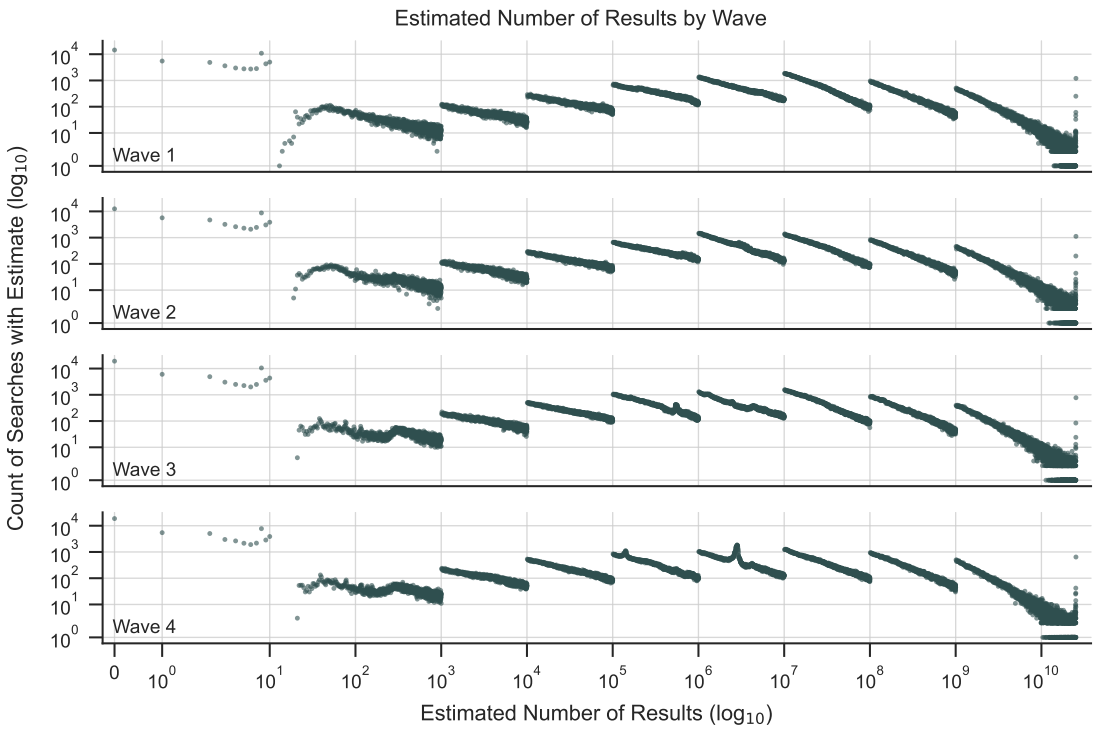


Figure A.9: The estimated number of search results available for a given query—provided by Google in each SERP—follows a mixture distribution consisting of nine distinct distributions (e.g., 0 to 10 and 11 to 1000), suggesting different mechanisms for estimating the number of results within each band. The median number of estimated results was 4.6M in Wave 1 (Oct 2023), 3.3M in Wave 2 (Mar 2024), 1.9M in Wave 3 (Sept 2024), and 2.6M in Wave 4 (Feb 2025). Means ranged from 281M in Wave 1 to 204M in Wave 3, and standard deviations ranged from 1.4B in Wave 1 to 1B in Wave 4. Our data suggests a ceiling on Google's estimates, and the max value we observed in any wave was 25.27B.

A.0.3 Descriptive Statistics

In this section we provide additional details about our dataset, characteristics of the search queries we used (A.0.3), and the SEO metrics we used (A.0.3). Details on our datasets, including counts and averages for key measures by wave, are provided in Table A.6 and Figure A.10.

Table A.6: Dataset counts and averages by wave for our 1.4M search queries show substantial decreases in the number of results returned over time alongside minimal changes to average domain quality. Details on Avg. Domain Quality are available in Methods 2.3.4, and details on Avg. Estimated Total Results are available in Section 2.4.5.

Wave	Search Results	Total Domains	News Domains	Avg. Domain Quality	Avg. Estimated Total Results
Oct 2023	26,119,263	19,945,139	2,392,470	0.783	280.8M
Mar 2024	22,231,957	16,905,242	1,810,744	0.788	289.9M
Sept 2024	20,621,796	16,820,589	1,595,996	0.791	204.4M
Feb 2025	21,143,317	16,813,318	1,498,731	0.789	214.3M

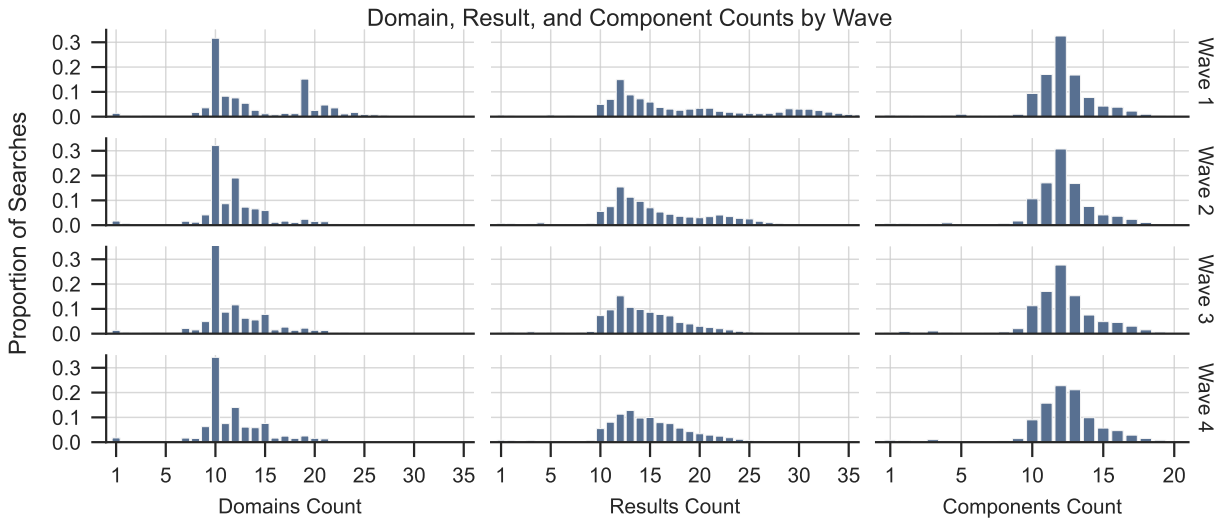
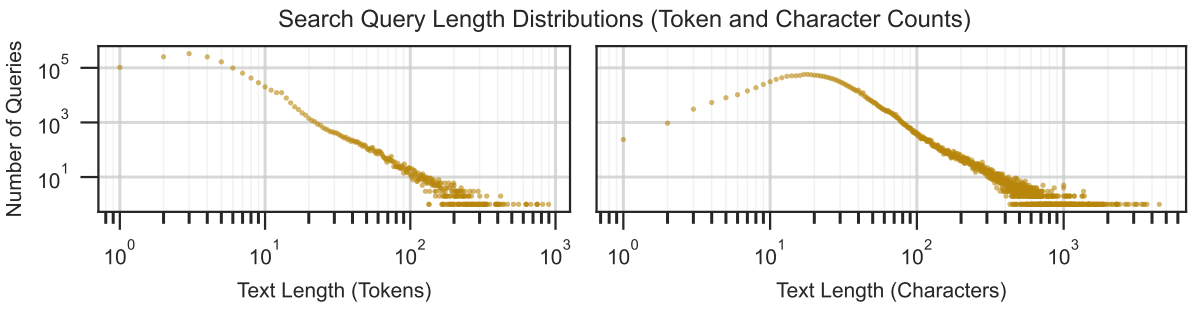
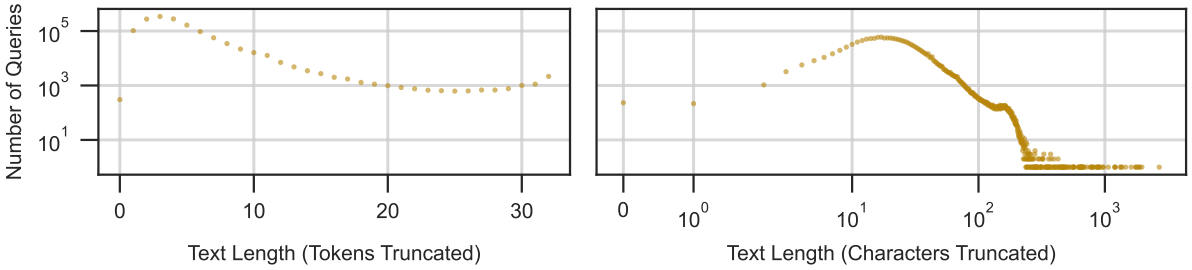


Figure A.10: Distribution of domain, result, and component counts across Search Engine Result Pages (SERPs) by data collection wave. Components represent a section of a SERP that can contain several results and domains (e.g., a Top Stories carousel can contain several results but only counts as one component).



(a) Full search query length distribution.



(b) Truncated search query length distribution.

Figure A.11: Search query length distributions for our set of 1.4M unique queries. The full distribution (a) shows the raw query lengths (measured via token and character counts), while the truncated distribution (b) shows the same distributions truncated at Google's 32 token query limit (see Methods 2.3.1). Some queries had a truncated token and character count of 0 because they only contained punctuation or emojis (which were removed during tokenization).

Search Queries

The length of the queries in our dataset (by token and character count) followed a heavy-tail distribution in its raw form (Figure A.11a) and remained skewed after truncating at Google's 32 token limit (Figure A.11b). The earliest search directive with a conspiracy-related query we found in our dataset was posted on July 19, 2009, and involved the longstanding conspiracy that 9/11 was an "inside job" Ballatore [2015], Mahl et al. [2021]. The post asked a question ("Was 9/11 an inside job?") and provided a Google Search link with "9/11 inside job proof" as the query. We provide additional query examples in Appendix 2.4.8.

Search Engine Optimization (SEO) Metrics

SEO is an understudied dimension of information environments and an important consideration in the broader field of social cybersecurity Carley [2020b]. Many SEO indicators require external data to calculate, which makes these indicators infeasible for researchers to collect given limited compute and storage resources, e.g., the number of backlinks a target domain receives is the total number of links that point towards the target domain from all other websites on the internet. Consequently, organizations and researchers often must rely on third-party SEO toolkits to obtain such indicators.

Here we used SEO metrics from Ahrefs[4], which had the 5th most active commercial webcrawler in October 2023 by number of requests according to Cloudflare Radar[5]. Specifically, we used Ahrefs' API to obtain the total number of backlinks and traffic estimates for a set of domains. Due to time and budget constraints, we collected this data only for the 9,125 unique domains that appeared in the results at least 10 times for queries containing political or conspiracy-related keywords.

While the only way to obtain true website traffic is through website owners, a non-peer-reviewed case study by AuthorityHacker, which used traffic data provided by 47 website owners, found Ahrefs' traffic estimates to be the most accurate of any SEO toolkit Webster [2022]. These estimates are calculated using position-weighted click-through rate estimates for Google search volumes of all keyphrases for which a website appears in the top 100 Google search results. In a report Ahrefs published on the accuracy of its own traffic estimates, which was conducted across a larger set of 1,635 domains, they found results that were similar to the AuthorityHacker case study[6]. These traffic estimates may contain noise, but are the among the best estimates currently available, and have been used in past work on SERP reliability Carragher et al. [2024, 2025], Williams and Carley [2023].

[4] https://ahrefs.com
[5] https://radar.cloudflare.com/traffic/verified-bots
[6] https://ahrefs.com/blog/traffic-estimations-accuracy/

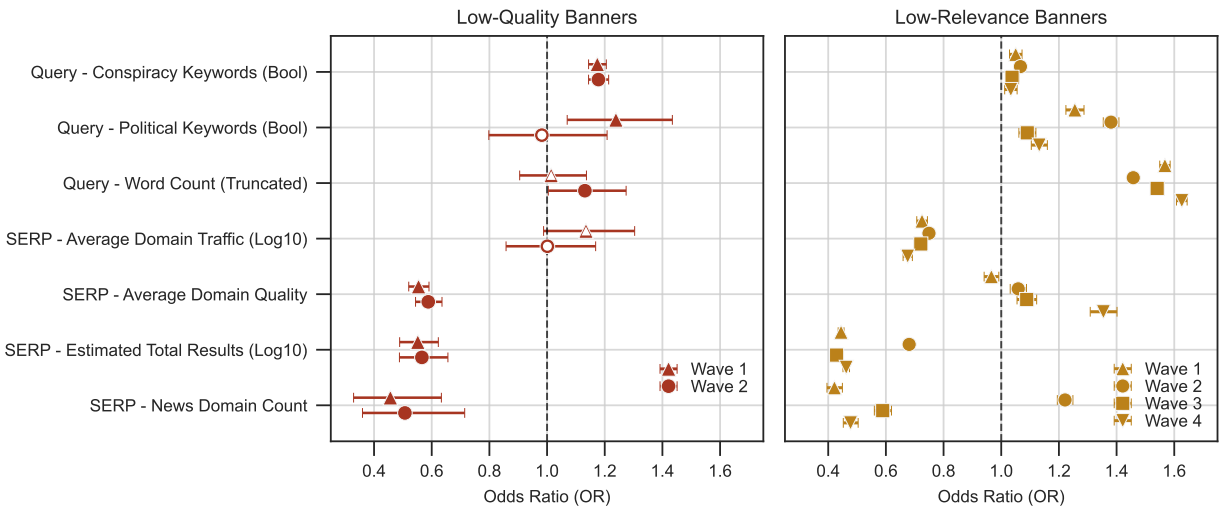


Figure A.12: The average domain quality of a SERP was strongly associated with the presence of low-quality banners (left), and longer search queries were strongly associated with the presence of low-relevance banners (right). Points represent Odds Ratios (OR) with 95% confidence intervals from logistic regression models (Methods 2.3.5). The vertical dashed line at OR=1 indicates no relationship, white-filled markers indicate coefficients that did not reach statistical significance ($p \geq 0.05$), and features are ordered by their average OR in the low-quality banner models. Additional model details and regression tables are available in Appendix A.0.4.

A.0.4 Logistic Regressions

Summary statistics for all models are presented in Table A.7, including sample sizes, degrees of freedom, pseudo-R^2 values, log-likelihood (LL), and information criteria (AIC). Our models for predicting low-relevance banners demonstrated better fit than the low-quality banner models in terms of pseudo-R^2 values. Odds ratios (OR), confidence intervals (95% CI), and P-values for our low-quality banner models are shown by wave in Table A.8. The same results for our low-relevance banner models are available in Table A.9.

Table A.7: Logistic regression model summary statistics.

Outcome	Wave	N Obs	N Events	AIC	Deviance	Pseudo R^2
Low-quality Banner	Oct 2023	956,661	140	2,281.6	2,265.6	0.177
	Mar 2024	911,885	96	1,683.3	1,667.3	0.145
Low-relevance Banner	Oct 2023	956,661	4,136	33,019.9	33,003.9	0.381
	Mar 2024	911,885	5,577	56,118.0	56,102.0	0.175
	Sept 2024	883,272	3,779	33,000.4	32,984.4	0.324
	Feb 2025	868,734	4,873	37,893.2	37,877.2	0.371

Table A.8: Logistic regression results for low-quality banners across waves.

(a) Wave 1

Feature (SERP / Query)	OR	95% CI	P
Conspiracy Keywords†	1.174	(1.144, 1.205)	0.000
Political Keywords†	1.239	(1.070, 1.435)	0.004
Word Count (Truncated)	1.014	(0.905, 1.137)	0.808
Avg. Domain Traffic‡	1.135	(0.988, 1.304)	0.074
Avg. Domain Quality	0.554	(0.520, 0.591)	0.000
Estimated Total Results‡	0.552	(0.489, 0.623)	0.000
News Domain Count	0.457	(0.329, 0.634)	0.000

†Boolean indicator; ‡Log10 transformed

(b) Wave 2

Feature (SERP / Query)	OR	95% CI	P
Conspiracy Keywords†	1.178	(1.144, 1.214)	0.000
Political Keywords†	0.982	(0.798, 1.208)	0.863
Word Count (Truncated)	1.131	(1.004, 1.274)	0.042
Avg. Domain Traffic‡	1.001	(0.858, 1.168)	0.988
Avg. Domain Quality	0.588	(0.544, 0.636)	0.000
Estimated Total Results‡	0.566	(0.488, 0.656)	0.000
News Domain Count	0.507	(0.360, 0.714)	0.000

†Boolean indicator; ‡Log10 transformed

Table A.9: Logistic regression results for low-relevance banners across waves.

(a) Wave 1

Feature (SERP / Query)	OR	95% CI	*P*
Word Count (Truncated)	1.568	(1.550, 1.586)	0.000
Political Keywords†	1.255	(1.224, 1.287)	0.000
Avg. Domain Quality	0.966	(0.941, 0.992)	0.010
Conspiracy Keywords†	1.050	(1.029, 1.072)	0.000
Avg. Domain Traffic‡	0.725	(0.707, 0.744)	0.000
News Domain Count	0.422	(0.396, 0.449)	0.000
Estimated Total Results‡	0.445	(0.435, 0.455)	0.000

†Boolean indicator; ‡Log10 transformed

(b) Wave 2

Feature (SERP / Query)	OR	95% CI	*P*
Word Count (Truncated)	1.458	(1.442, 1.474)	0.000
Political Keywords†	1.381	(1.354, 1.408)	0.000
Avg. Domain Quality	1.059	(1.031, 1.088)	0.000
Conspiracy Keywords†	1.067	(1.053, 1.080)	0.000
Avg. Domain Traffic‡	0.749	(0.732, 0.767)	0.000
News Domain Count	1.221	(1.195, 1.249)	0.000
Estimated Total Results‡	0.681	(0.667, 0.695)	0.000

†Boolean indicator; ‡Log10 transformed

(c) Wave 3

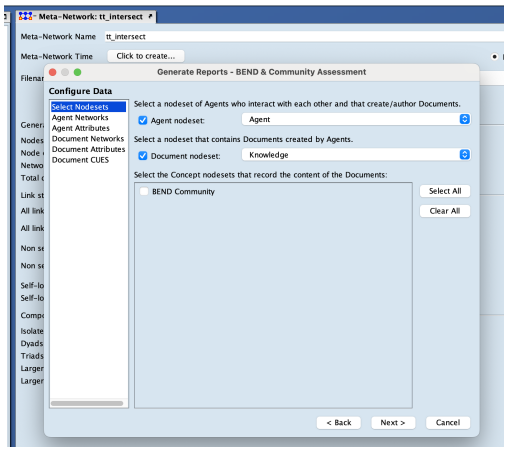
Feature (SERP / Query)	OR	95% CI	*P*
Word Count (Truncated)	1.541	(1.522, 1.560)	0.000
Political Keywords†	1.090	(1.061, 1.120)	0.000
Avg. Domain Quality	1.088	(1.055, 1.123)	0.000
Conspiracy Keywords†	1.038	(1.015, 1.061)	0.001
Avg. Domain Traffic‡	0.721	(0.703, 0.740)	0.000
News Domain Count	0.589	(0.560, 0.619)	0.000
Estimated Total Results‡	0.429	(0.418, 0.441)	0.000

†Boolean indicator; ‡Log10 transformed

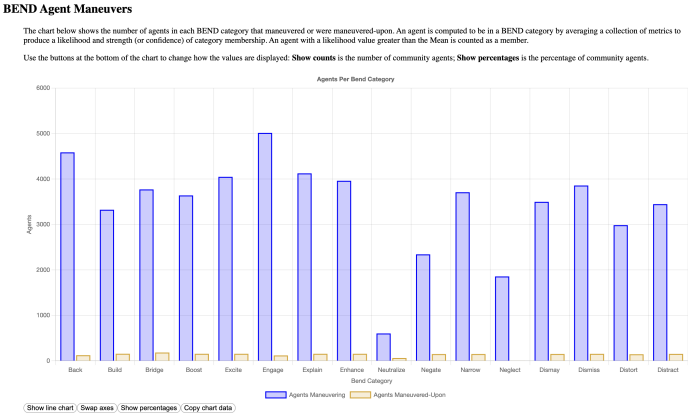
(d) Wave 4

Feature (SERP / Query)	OR	95% CI	*P*
Word Count (Truncated)	1.626	(1.608, 1.645)	0.000
Political Keywords†	1.132	(1.104, 1.160)	0.000
Avg. Domain Quality	1.354	(1.309, 1.402)	0.000
Conspiracy Keywords†	1.033	(1.012, 1.055)	0.002
Avg. Domain Traffic‡	0.676	(0.660, 0.692)	0.000
News Domain Count	0.477	(0.452, 0.504)	0.000
Estimated Total Results‡	0.462	(0.451, 0.474)	0.000

†Boolean indicator; ‡Log10 transformed



(a) ORA-PRO's "BEND & Community Assessment report" GUI.



(b) Barplot of BEND metrics returned in the "BEND & Community Assessment report" over a synthetic dataset

Figure A.13: ORA GUI and output of BEND report

A.1 BEND Report

We discuss running the original proprietary BEND metrics in chapter 6. I will provide more details here on what those look like in practice. After a user extracts CUES from ORA-PRO's sister tool Netmapper Carley et al. [2018], the user reads networks, text, and CUES into the the ORA-PRO software and runs the "BEND & Community Assessment" analysis. ORA-PRO then calculates BEND metrics on the observed data. I provide example screenshots of the process in Figure A.13. On the left panel, I present a screenshot of ORA-PRO's GUI as a user navigates through running the "BEND & Community Assessment" report. On the right, I show an example BEND output on a synthetic network. The two bars demonstrate the raw counts between agents "maneuvering" and agents being "maneuvered-upon". Again, these do not imply intent; only that the stated BEND maneuver has been observed. The reports allow analysts to show additional information for each metric. In Figure A.14, I display the "Build" normalized metrics ORA-PRO calculated for the Think Tanks discussed in chapter 6.

Build Agent Analysis

Discussion or actions that create a group, or the appearance of a group, where there was none before.

Number of maneuvering agents:	4	22.22% (of total agents)
Number of maneuvered-upon agents:	4	22.22% (of total agents)

Most Maneuvering Agents

This shows the agents maneuvering the most in the Build category.

If the node of interest has a higher than normal value (greater than 1 standard deviation(s) above the mean) the row is colored red. The row is green if the node is within 1 standard deviation of the mean. Finally, the row is colored blue if the node has a lower than normal value (less than one standard deviation(s) below the mean).

Node name markers: [+] Verified, [*] News Agency, [gov] Government Actor, [bot] BOT

Copy Print Search:

Rank	Agent ID	Agent Label	Value
1	strategic-culture.su	strategic-culture.su	0.776
2	hudson.org	hudson.org	0.561
3	cei.org	cei.org	0.500
4	clingendael.org	clingendael.org	0.200
5	chathamhouse.org	chathamhouse.org	0.065
6	journal-neo.su	journal-neo.su	0.062
7	americansforprosperity.org	americansforprosperity.org	0.038
8	news-front.su	news-front.su	0.032
9	katehon.com	katehon.com	0.024
10	southfront.press	southfront.press	0.020

Showing 1 to 10 of 13 entries Previous 1 2 Next

Figure A.14: Barplot of BEND metrics returned in the “BEND & Community Assessment report” over a synthetic dataset

Bibliography

Simon Adler, Annie McEwen, Becca Bressler, and Charles Maynes. The curious case of the russian flash mob at the west palm beach cheesecake factory, February 2018. URL https://radiolab.org/podcast/curious-case-russian-flash-mob-west-palm-beach-cheesecake-factory. Podcast episode produced by Radiolab.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5–16, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3637528.3671900.

Iuliia Alieva, Lynnette Hui Xian Ng, and Kathleen M Carley. Investigating the spread of russian disinformation about biolabs in ukraine on twitter using social network analysis. In *2022 IEEE international conference on big data (big data)*, pages 1770–1775. IEEE, 2022.

Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995):548–556, 2024. doi: 10.1038/s41586-023-06883-y.

R. Aswani, S.P. Ghrera, S. Chandra, and A.K. Kar. A hybrid evolutionary approach for identifying spam websites for search engine marketing. *Evolutionary Intelligence*, 14(4): 1803–1815, 2021. ISSN 1864-5909. doi: 10.1007/s12065-020-00461-1.

Chhandak Bagchi, Filippo Menczer, Jennifer Lundquist, Monideepa Tarafdar, Anthony Paik, and Przemyslaw A. Grabowicz. Social media algorithms can curb misinformation, but do they?, 2024.

Loren Baker. Google Ron Paul : Connecting with Searchers, 2008. URL https://www.searchenginejournal.com/google-ron-paul-connecting-with-searchers/6177/.

Nathan Ballantyne and David Dunning. Skeptics say, 'do your own research.' it's not that simple. The New York Times, 2022. URL https://www.nytimes.com/2022/01/03/opinion/dyor-do-your-own-research.html.

Cameron Ballard, Ian Goldstein, Pulak Mehta, Genesis Smothers, Kejsi Take, Victoria Zhong, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. Conspiracy brokers: Understanding the monetization of youtube conspiracy theories. In *Proceedings of the ACM Web Conference 2022*, pages 2707–2718, 2022.

Andrea Ballatore. Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday*, 20(7), 2015. doi: 10.5210/fm.v20i7.5597.

Jack Bandy. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):74:1–74:34, 2021. doi: 10.1145/3449148.

Edelman Trust Barometer. 2024 edelman trust barometer global report, 2024. URL https://www.edelman.com/trust/2024/trust-barometer.

Anna Beers. Anti-transgender disinformation in the age of algorithmic search summaries. *Social Science Research Council*, 2025. doi: 10.35650/JT.3083.d.2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Matteo Benigni, Jennifer Kenny, and Kathleen M. Carley. Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PLOS ONE*, 12(12):e0181405, 2017. doi: 10.1371/journal.pone.0181405.

John Bianchi, Manuel Pratelli, Marinella Petrocchi, and Fabio Pinelli. Evaluating trustworthiness of online news publishers via article classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 671–678, 2024.

Janice T Blane. Social-cyber maneuvers for analyzing online influence operations, 2023.

Janice T Blane, Daniele Bellutta, and Kathleen M Carley. Social-cyber maneuvers during the covid-19 vaccine initial rollout: content analysis of tweets. *Journal of Medical Internet Research*, 24(3):e34040, 2022.

Sam Blazek. Scotch: A framework for rapidly assessing influence operations. *Atlantic Council*, 2021.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Leticia Bode and Emily K. Vraga. Studying Politics Across Media. *Political Communication*, 35(1):1–7, 2018. doi: 10.1080/10584609.2017.1334730.

Olivia Borge, Victoria Cosgrove, Elena Cryst, Shelby Grossman, Shelby Perkins, and Anna Van Meter. How Search Engines Handle Suicide Queries. *Journal of Online Trust and Safety*, 1(1), 2021. doi: 10.54501/jots.v1i1.16.

Samantha Bradshaw. Government responses to malicious use of social media, 2017. Referenced in paper as proprietary-data study of junk news SEO strategies.

Samantha Bradshaw, Shelby Grossman, and Miles McCain. An investigation of social media labeling decisions preceding the 2020 U.S. election. *PLOS ONE*, 18(11), 2023. doi: 10.1371/journal.pone.0289683.

Kathleen M. Carley. Social cybersecurity: An emerging science. *Computational and Mathematical Organizational Theory*, 26(4):365–381, 2020a. doi: 10.1007/s10588-020-09322-9.

Kathleen M. Carley. Social cybersecurity: An emerging science. *Computational and Mathematical Organization Theory*, 26(4):365–381, 2020b. doi: 10.1007/s10588-020-09322-9.

L Richard Carley, Jeff Reminga, and Kathleen M Carley. Ora & netmapper, 2018.

Peter Carragher and Kathleen M Carley. Accountability in search engine manipulation: A case study of the iranian news ecosystem. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 154–163. Springer, 2024.

Peter Carragher, Evan M. Williams, and Kathleen M. Carley. Detection and Discovery of Misinformation Sources Using Attributed Webgraphs. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:214–226, 2024. doi: 10.1609/icwsm.v18i1.31309.

Peter Carragher, Evan M. Williams, and Kathleen M. Carley. Misinformation resilient search rankings with webgraph-based interventions. *ACM Trans. Intell. Syst. Technol.*, 16(1): 10:1–10:27, 2025. doi: 10.1145/3670410.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980, 2019.

Stevie Chancellor, Jessica Annette Pater, Trustin A Clear, Eric Gilbert, and Munmun De Choudhury. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 1199–1211, San Francisco, California, USA, 2016. ACM Press. doi: 10.1145/2818048.2819963.

Serina Chang, Adam Fourney, and Eric Horvitz. Measuring vaccination coverage and concerns of vaccine holdouts from web search logs. *Nature Communications*, 15(1):6496, 2024. doi: 10.1038/s41467-024-50614-4.

Bill Chappell and Odette Yousef. How the false russian biolab story came to circulate among the u.s. far right, 2022. URL https://www.npr.org/2022/03/25/1087910880/biologicalweapons-far-right-russia-ukraine.

Zhouhan Chen and Juliana Freire. Proactive discovery of fake news domains from real-time social media feeds. In *Companion Proceedings of the Web Conference 2020*, pages 584–592, 2020.

Sedona Chinn and Ariel Hasell. Support for "doing your own research" is associated with COVID-19 misperceptions and scientific mistrust. *Harvard Kennedy School Misinformation Review*, 2023. doi: 10.37016/mr-2020-117.

Chitika Insights. The value of google result positioning, 2013. URL https://research.chitika.com/wp-content/uploads/2022/02/chitikainsights-valueofgoogleresultspositioning.pdf.

Valentin Châtelet and Amaury Lesplingart. Russia-linked pravda network cited on wikipedia, llms, and x, Mach 2025. URL https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/. Accessed: 2025-06-04.

Volkan Cirik, Eduard Hovy, and Louis-Philippe Morency. Visualizing and understanding curriculum learning for long short-term memory networks. *arXiv preprint arXiv:1611.06204*, 2016.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), 2024. doi: 10.1126/science.adq1814.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. ORCAS: 18 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2983–2989, New York, NY, USA, 2020. Association for Computing Machinery. doi:

10.1145/3340531.3412779.

Hanjun Dai, Bo Dai, and Le Song. Discriminative embeddings of latent variable models for structured data. In *International conference on machine learning*, pages 2702–2711. PMLR, 2016.

Michal Danilák. langdetect. https://github.com/Mimino666/langdetect, 2014.

Brian Dean. We analyzed 4 million google search results: Here's what we learned about organic click through rate, 2022. URL https://backlinko.com/google-ctr-stats. Accessed: 2025-05-12.

Edelman. Edelman trust barometer 2021. Technical report, Edelman, 2021. URL https://www.edelman.com/trust/2021-trust-barometer.

Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015a. doi: 10.1073/pnas.1419828112.

Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), 2015b. doi: 10.1073/pnas.1419828112.

Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (SEME). In *Proceedings of the ACM on Human-Computer Interaction*, volume 1, pages 1–22, 2017. doi: 10.1145/3134677.

Dan Evon. Ukraine, us biolabs, and an ongoing russian disinformation campaign, 2022. URL https://www.snopes.com/news/2022/02/24/us-biolabs-ukraine-russia/.

Facebook. Threat report: The state of influence operations 2017–2020, 2021. URL https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf.

Alvaro Feal, Jeffrey Gleason, Pranav Goel, Jason Radford, Kai-Cheng Yang, John Basl, Michelle Meyer, David Choffnes, Christo Wilson, and David Lazer. Introduction to National Internet Observatory. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*, 2024.

Stefan Feuerriegel, Renée DiResta, Josh A. Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Pröllochs. Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, 7(11):1818–1821, 2023. doi: 10.1038/s41562-023-01726-2.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric.

In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019a.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019b.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. doi: 10.1093/biomet/80.1.27.

Claudia Flores-Saviaga, Shangbin Feng, and Saiph Savage. Datavoidant: An AI System for Addressing Political Data Voids on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29, 2022. doi: 10.1145/3555616.

Daniel Funke. The COVID-19 vaccine could lead to prion diseases, Alzheimer's, ALS and other neurodegenerative diseases., 2021. URL https://www.politifact.com/factchecks/2021/feb/26/j-bart-classen/coronavirus-vaccine-doesnt-cause-alzheimers-als/.

Amanda Garry, Samantha Walther, Rukaya Rukaya, and Ayan Mohammed. Qanon conspiracy theory: examining its evolution and mechanisms of radicalization. *Journal for Deradicalization*, pages 152–216, 2021.

Dipayan Ghosh and Ben Scott. Digital deceit: The technologies behind precision propaganda on the internet. New America, 2018. URL https://d1y8sb8igg2f8e.cloudfront.net/documents/digital-deceit-final-v3.pdf.

Michael Golebiewski and Danah Boyd. Data voids: Where missing data can easily be exploited. Data & Society Research Institute, 2019. URL https://datasociety.net/library/data-voids/.

Michael Golebiewski and danah boyd. Data voids: Where missing data can easily be exploited. Technical report, Data & Society, 2019. URL https://datasociety.net/library/data-voids.

Google. Search Quality Rater Guidelines: An Overview. Technical report, Google Search, 2023.

Google. Information Quality & Content Moderation, 2024. URL https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/Info_quality_content_moderation-whie_paper-UPDATED.pdf.

Jon Green, Stefan Mccabe, Sarah Shugars, Hanyu Chwe, Luke Horgan, Shuyang Cao, and David Lazer. Curation Bubbles. *American Political Science Review*, pages 1–19, 2025. doi: 10.1017/S0003055424000984.

Kevin T. Greene, Nilima Pisharody, Lucas Augusto Meyer, Mayana Pereira, Rahul Dodhia, Juan Lavista Ferres, and Jacob N. Shapiro. Current engagement with unreliable sites from web search driven by navigational search. *Science Advances*, 10(44):eadn3750, 2024. doi: 10.1126/sciadv.adn3750.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

Fabian Haak and Philipp Schaer. Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 239–244, Austin TX USA, 2023. ACM. doi: 10.1145/3578503.3583628.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Tawfiq Hasanin, Taghi M Khoshgoftaar, Joffrey L Leevy, and Richard A Bauder. Severely imbalanced big data challenges: investigating data sampling approaches. *Journal of Big Data*, 6(1):1–25, 2019.

Rebecca Hersher. What happened when dylann roof asked google for information about race?, 2017. URL https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylann-roof-asked-google-for-information-about-race. Accessed: 2025-05-12.

Heike Hofmann, Hadley Wickham, and Karen Kafadar. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3):469–477, 2017. doi: 10.1080/10618600.2017.1305277.

Andrea Hrckova, Robert Moro, Ivan Srba, and Maria Bielikova. Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in central europe. *Online Information Review*, 46(5):954–973, 2021. doi: 10.1108/OIR-10-2020-0441.

Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. Auditing the partisanship of Google Search snippets. In *The World Wide Web Conference*, pages 693–704, San Francisco, CA, USA, 2019. ACM Press. doi: 10.1145/3308558.3313654.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34267–34291, 2025.

Atsushi Inoue and Lutz Kilian. In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402, 2005. doi: 10.1081/ETC-200040785.

Chitika Insights. The value of google result positioning, 2014. URL https://research.chitika.com/wp-content/uploads/2022/02/chitikainsights-valueofgoogleresultspositioning.pdf.

Mike Isaac and Daisuke Wakabayashi. Russian influence reached 126 million through facebook alone. The New York Times, 2017. URL https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Prerna Juneja, Wenjuan Zhang, Alison Marie Smith-Renner, Hemank Lamba, Joel Tetreault, and Alex Jaimes. Dissecting users' needs for search result explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, Honolulu HI USA, 2024. ACM. doi: 10.1145/3613904.3642059.

Anna Kawakami, Khonzodakhon Umarova, and Eni Mustafaraj. The media coverage of the 2020 US presidential election candidates through the lens of Google's top stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 868–877, 2020. URL https://aaai.org/ojs/index.php/ICWSM/article/view/7352.

Andrew Keh and Stuart A. Thompson. He Died in a Tragic Accident. Why Did the Internet Say He Was Murdered? *The New York Times*, 2024a. URL https://www.nytimes.com/2024/01/25/nyregion/obituary-pirates-matteo-sachman.html.

Andrew Keh and Stuart A. Thompson. He died in a tragic accident. why did the internet say he was murdered?, 2024b. URL https://www.nytimes.com/2024/01/25/nyregion/obituary-pirates-matteo-sachman.html?searchResultPosition=1.

KeywordDiscovery. Keyword usage statistics on the average number of keywords per search

phrase by Country, 2020. URL https://www.keyworddiscovery.com/keyword-stats.html?date=2020-01-01.

Catherine King, Christine S. Lepird, and Kathleen M. Carley. Project omen: Designing a training game to fight misinformation on social media, 2021. URL http://reports-archive.adm.cs.cmu.edu/anon/anon/usr0/ftp/home/ftp/isr2021/CMU-ISR-21-110.pdf.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Michel Klen. Portal kombat: la nouvelle offensive de désinformation menée par la russie. *Revue Défense Nationale*, 869(4):108–113, 2024.

Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference - IMC '15*, pages 121–127, Tokyo, Japan, 2015. ACM Press. doi: 10.1145/2815675.2815714.

Ian Kloo and Kathleen M Carley. Social cybersecurity analysis of the telegram information environment during the 2022 invasion of ukraine. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 23–32. Springer, 2023.

Kate Knibbs. The bizarre cottage industry of youtube obituary pirates, 2023. URL https://www.wired.com/story/youtube-obituary-pirates/.

Daria Kravets and F. Toepfl. Gauging reference and source bias over time: How Russia's partially state-controlled search engine Yandex mediated an anti-regime protest event. *Information, Communication & Society*, 25(15):2207–2223, 2022. doi: 10.1080/1369118X.2021.1933563.

Daria Kravets and Florian Toepfl. Gauging reference and source bias over time: How russia's partially state-controlled search engine yandex mediated an anti-regime protest event. *Information, Communication & Society*, 25(15):2207–2223, 2021. doi: 10.1080/1369118X.2021.1933563.

D. Kravitz, A. Shaw, C. Perlman, and A. Mierjeski. Trump town: Track white house staff, cabinet members and political appointees across the government, 2019a. URL https://projects.propublica.org/trump-town/.

Daniel Kravitz, Al Shaw, C. Perlman, and A. Mierjeski. Trump town: Track white house staff, cabinet members and political appointees across the government. ProPublica, 2019b.

URL https://projects.propublica.org/trump-town/.

Nandita Krishnan, Jiayan Gu, Rebekah Tromble, and Lorien C. Abroms. Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, 2021. doi: 10.37016/mr-2020-85.

Elizaveta Kuznetsova, Mykola Makhortykh, Maryna Sydorova, Aleksandra Urman, Ilaria Vitulano, and Martha Stolze. Algorithmically curated lies: How search engines handle misinformation about us biolabs in ukraine. *arXiv preprint arXiv:2401.13832*, 2024.

Philippe Laban, Alexander Richard Fabbri, Caiming Xiong, and Chien-Sheng Wu. Summary of a haystack: A challenge to long-context llms and rag systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, 2024.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Anti-Defamation League. Unmasking "clandestine," the figure behind the viral "ukrainian biolab" conspiracy theory, 2022. URL https://www.adl.org/resources/blog/unmasking-clandestine-figure-behind-viral-ukrainian-biolab-conspiracy-theory.

Christine Sowa Lepird and Kathleen M Carley. Comparison of online maneuvers by authentic and inauthentic local news organizations. *Computational and Mathematical Organization Theory*, pages 1–15, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Changmao Li and Jeffrey Flanigan. Rac: Efficient llm factuality correction with retrieval augmentation. *arXiv preprint arXiv:2410.15667*, 2024.

Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, et al. On the role of long-tail knowledge in retrieval augmented large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–126, 2024.

Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. High level of correspondence across different news domain quality rating sets. *PNAS nexus*, 2(9):pgad286, 2023.

Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas

Costello, Gordon Pennycook, and David G. Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, 648(8093):394–401, 2025. doi: 10.1038/s41586-025-09771-9.

John E. Lincoln. Google click-through rates (ctr) by ranking position [2020], 2020. URL https://ignitevisibility.com/google-ctr-by-ranking-position/. Accessed: 2025-05-12.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ramona Ludolph, Ahmed Allam, and Peter J Schulz. Manipulating Google's Knowledge Graph Box to Counter Biased Information Processing During an Online Search on Vaccination: Application of a Technological Debiasing Strategy. *Journal of Medical Internet Research*, 18(6):e137, 2016. doi: 10.2196/jmir.5430.

Josephine Lukito. Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017. *Political Communication*, 37(2):238–255, 2020. doi: 10.1080/10584609.2019.1661889.

Emma Lurie and Deirdre K. Mulligan. Searching for representation: A sociotechnical audit of Googling for members of U.S. Congress, 2021.

Emma Lurie and Eni Mustafaraj. Investigating the effects of Google's search engine result page in evaluating the credibility of online news sources. In *Proceedings of the 10th ACM Conference on Web Science*, pages 107–116, Amsterdam, Netherlands, 2018. ACM Press. doi: 10.1145/3201064.3201095.

Daniela Mahl, Jing Zeng, and Mike S. Schäfer. From "Nasa Lies" to "Reptilian Eyes": Mapping Communication About 10 Conspiracy Theories, Their Communities, and Main Propagators on Twitter. *Social Media + Society*, 7(2):20563051211017482, 2021. doi: 10.1177/20563051211017482.

Mykola Makhortykh, Aleksandra Urman, and Roberto Ulloa. How search engines disseminate information about COVID-19 and why they should do better. *Harvard Kennedy School (HKS) Misinformation Review*, 2020. doi: 10.37016/mr-2020-017.

Mykola Makhortykh, Aleksandra Urman, and Mariëlle Wijermars. A story of (non) compliance, bias, and conspiracies: How google and yandex represented smart voting during the 2021 parliamentary elections in russia. *Harvard Kennedy School Misinformation Review*, 3(2): 1–16, 2022a.

Mykola Makhortykh, Aleksei Urman, and Mariëlle Wijermars. A story of (non) compliance,

bias, and conspiracies: How google and yandex represented smart voting during the 2021 parliamentary elections in russia. *Harvard Kennedy School (HKS) Misinformation Review*, 3(2), 2022b. doi: 10.37016/mr-2020-94.

Rebecca Marigliano and Kathleen M. Carley. Aurora: Enhancing synthetic population realism through rag and salience-aware opinion modeling. *2025 Winter Simulation Conference (WSC)*, pages 2563–2574, 2025. URL https://api.semanticscholar.org/CorpusID:284968238.

Rebecca Marigliano, Lynnette Hui Xian Ng, and Kathleen M Carley. Analyzing digital propaganda and conflict rhetoric: a study on russia's bot-driven campaigns and counter-narratives during the ukraine crisis. *Social Network Analysis and Mining*, 14(1):170, 2024.

Cameron Martel and David G. Rand. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 54:101710, 2023. doi: 10.1016/j.copsyc.2023.101710.

Alice E. Marwick and William C. Partin. Constructing alternative facts: Populist expertise and the qanon conspiracy. *New Media & Society*, 2022. doi: 10.1177/14614448221090201.

Alice E Marwick and William Clyde Partin. Constructing alternative facts: Populist expertise and the qanon conspiracy. *New Media & Society*, 26(5):2535–2555, 2024.

John McDuling. Google is now a more trusted source of news than the websites it aggregates, 2015. URL https://qz.com/329211/google-is-now-a-more-trusted-source-of-news-than-the-websites-it-aggregates.

J. G. McGann. 2020 global go to think tank index report, 2021a. URL https://repository.upenn.edu/think_tanks/18/.

James G. McGann. 2020 global go to think tank index report. Technical report, Think Tanks and Global Societies Program, University of Pennsylvania, 2021b. URL https://repository.upenn.edu/think_tanks/18/.

Yelena Mejova, Tatiana Gracyk, and Ronald E. Robertson. Googling for Abortion: Search Engine Mediation of Abortion Accessibility in the United States. *Journal of Quantitative Description: Digital Media*, 2, 2022. doi: 10.51685/jqd.2022.007.

Danaë Metaxa, Joon Sung Park, James A. Landay, and Jeff Hancock. Search media and elections: A longitudinal investigation of political search results. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–17, 2019. doi: 10.1145/3359231.

Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021. doi: 10.1561/1100000083.

Jonathan H. Morgan, Jake Shaha, Rebecca Marigliano, Matthew S. Hicks, and Kathleen M. Carley. Ghostfield: Simulating an integrated information training environment at scale. In *Proceedings of the 18th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2025), Working Papers*, Pittsburgh, PA, USA, 2025. URL https://sbp-brims.org/2025/papers/working-papers/2025_SBP-BRiMS_paper_8.pdf.

Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386, 2022. doi: 10.1002/asi.24637.

Partha Mukherjee and Bernard J. Jansen. Conversing and searching: The causal relationship between social media and web search. *Internet Research*, 27(5):1209–1226, 2017. doi: 10.1108/IntR-07-2016-0228.

Leopold Müller, Joshua Holstein, Sarah Bause, Gerhard Satzger, and Niklas Kühl. Data quality challenges in retrieval-augmented generation. *arXiv preprint arXiv:2510.00552*, 2025.

Deirdre K Mulligan and Daniel S Griffin. Rescripting search to respect the right to truth. *Geo. L. Tech. Rev.*, 2:28, 2018. URL https://ssrn.com/abstract=3228671.

Kevin Munger. The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society*, 5(3), 2019. doi: 10.1177/2056305119859294.

Eni Mustafaraj, Emma Lurie, and Claire Devine. The case for voter-centered audits of search engines during political elections. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 559–569, 2020.

Preslav Nakov, Firoj Alam, Giovanni Da San Martino, Maram Hasanain, Dilshod Azizov, Rabindra Nath Nandi, and Panayotov Panayot. Overview of the clef-2023 checkthat! lab task 4 on factuality of reporting of news media, 2023.

Nature Editorials. How online misinformation exploits 'information voids' — and what to do about it. *Nature*, 625:215–216, 2024. doi: 10.1038/d41586-024-00030-x. URL

https://www.nature.com/articles/d41586-024-00030-x. Editorial.

Pandu Nayak. New ways we're helping you find high-quality information, 2022. URL https://blog.google/products/search/information-literacy/.

Casey Newton. Google gives up on data voids, 2025. URL https://www.platformer.news/google-data-voids-warning-banners-2024-election/.

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, 2018.

Ov Cristian Norocel and Dirk Lewandowski. Google, data voids, and the dynamics of the politics of exclusion. *Big Data & Society*, 10(1):20539517221149099, 2023.

Amy Orben and J. Nathan Matias. Fixing the science of digital technology harms. *Science*, 388(6743):152–155, 2025. doi: 10.1126/science.adt6807.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching llms credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863, 2024.

Shimei Pan and Tao Ding. Social media-based user embedding: A literature review. *arXiv preprint arXiv:1907.00725*, 2019.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*, 2023.

John V. Parachini. Debunking russian lies about biolabs at upcoming u.n. meetings, 2022. URL https://www.rand.org/pubs/commentary/2022/09/debunking-russian-lies-about-biolabs-at-upcoming-un.html.

Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. A link-based approach to detect media bias in news websites. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 742–745, 2019.

Christopher C. Pennock, Jeremy Hylton, and Corinna Cortes. Website quality signal generation, May 2013.

Brooke Perreault, Johanna Hoonsun Lee, Ropafadzo Shava, and Eni Mustafaraj. Algorithmic Misjudgement in Google Search Results: Evidence from Auditing the US Online Electoral Information Environment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 433–443, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3630106.3658916.

Samantha C Phillips, Joshua Uyheng, Charity S Jacobs, and Kathleen M Carley. Chirping diplomacy: Analyzing chinese state social-cyber maneuvers on twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 95–104. Springer, 2023.

Ben Popken and Kelly Cobiella. Russian troll describes work in the infamous misinformation factory. NBC News, 2017a. URL https://www.nbcnews.com/news/all/russian-troll-describes-work-infamous-misinformation-factory-n821486.

Ben Popken and Kelly Cobiella. Russian troll describes work in the infamous misinformation factory, 2017b.

Elizabeth Reid. Ai overviews: About last week, 2024. URL https://blog.google/products/search/ai-overviews-update-may-2024/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Christoph Riedl, Johannes Bjelland, Geoffrey Canright, Asif Iqbal, Kenth Engø-Monsen, Taimur Qureshi, Pål Roe Sundsøy, and David Lazer. Product diffusion through on-demand information-seeking behaviour. *Journal of The Royal Society Interface*, 15(139), 2018. doi: 10.1098/rsif.2017.0751.

Ronald E Robertson and Christo Wilson. WebSearcher: Tools for Auditing Web Search. In *Proceedings of Computation + Journalism Symposium (C+J '20)*, page 4, 2020.

Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. In *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, pages 235–244, Boston, Massachusetts, USA, 2019. ACM Press. doi: 10.1145/3292522.3326047.

Ronald E. Robertson, Amy Dunphy, Shelby Grossman, Renée DiResta, and David Thiel. Identifying Search Directives on Social Media. *Journal of Online Trust and Safety*, 2(1),

2023a. doi: 10.54501/jots.v2i1.133.

Ronald E. Robertson, Jon Green, Damian J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 618:342–348, 2023b. doi: 10.1038/s41586-023-06078-5.

Ronald E Robertson, Evan M Williams, Kathleen M Carley, and David Thiel. Data voids and warning banners on google search. *arXiv preprint arXiv:2502.17542*, 2025.

Reece Rogers. Google admits its ai overviews search feature screwed up, 2024. URL https://www.wired.com/story/google-ai-overview-search-issues/.

Chris Rowlands. Goodbye google? people are increasingly switching to the likes of chatgpt, according to major survey – here's why, 2025. URL https://www.techradar.com/tech/people-are-increasingly-swapping-google-for-the-likes-of-chatgpt-according-to-a-major

Graham Ryan. Generative ai will break the internet: Beyond section 230. 2024.

Daniel López Sánchez, Jorge Revuelta, Fernando De la Prieta, Ana B Gil-González, and Cach Dang. Twitter user clustering based on their preferences and the louvain algorithm. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 349–356. Springer, 2016.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *64th Annual Meeting of the International Communication Association*, 2014.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019a.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019b.

Barry Schwartz. Google august 2024 core update rolling out now, 2024. URL https://searchengineland.com/google-august-2024-core-update-rolling-out-now-445221.

Vibhor Sehgal, Ankit Peshin, Sadia Afroz, and Hany Farid. Mutual Hyperlinking Among Misinformation Peddlers, April 2021. URL http://arxiv.org/abs/2104.11694. arXiv:2104.11694 [cs].

Elisa Shearer and Amy Mitchell. News Use Acrosss Social Media Platforms in 2020. Technical

report, Pew Research Center, 2021. URL https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2021/01/PJ_2021.01.12_News-and-Social-Media_FINAL.pdf.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. Reliabilityrag: Effective and provably robust defense for rag-based web-search. *arXiv preprint arXiv:2509.23519*, 2025.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

Almog Simchon, William J Brady, and Jay J Van Bavel. Troll and divide: The language of online polarization. *PNAS Nexus*, 1(1), 2022. doi: 10.1093/pnasnexus/pgac019.

Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. Community notes reduce engagement with and diffusion of false information online. *Proceedings of the National Academy of Sciences*, 122(38):e2503413122, 2025. doi: 10.1073/pnas.2503413122.

Karolina Sliwa, Ema Kusen, and Mark Strembeck. A case study comparing twitter communities detected by the louvain and leiden algorithms during the 2022 war in ukraine. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1376–1381, 2024.

K. Smagily. Hybrid analytica: Pro-kremlin expert propaganda in moscow, europe and the us: A case study on think tanks and universities, 2018a. URL https://www.underminers.info/publications/hybridanalytica.

Kristina Smagily. Hybrid analytica: Pro-kremlin expert propaganda in moscow, europe and the us: A case study on think tanks and universities. Institute of Modern Russia, 2018b. URL https://www.underminers.info/publications/hybridanalytica.

Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Yokohama Japan, 2025. ACM. doi: 10.1145/3706598.3714082.

Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, 2010.

Danny Sullivan. Why Google Can't Count Results Properly, 2010. URL https://searchengineland.com/why-google-cant-count-results-properly-53559.

Danny Sullivan. A new notice in Search for rapidly evolving results, 2021. URL https://blog.google/products/search/new-notice-search-rapidly-evolving-results/.

Latanya Sweeney. Discrimination in Online Ad Delivery. *Queue*, 11(3):10, 2013. doi: 10.1145/2460276.2460278.

Sara-Jayne Terp and Pablo Breuer. Disarm: a framework for analysis of disinformation campaigns. In *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 1–8. IEEE, 2022.

Stuart A. Thompson. Fed up with google, conspiracy theorists turn to duckduckgo, 2022.

Florian Toepfl, Daria Kravets, Anastasia Ryzhova, and Anja Beseler. Who are the plotters behind the pandemic? comparing covid-19 conspiracy theories in google search results across five key target countries of russia's foreign communication. *Information, Communication & Society*, 2022. doi: 10.1080/1369118X.2022.2065213.

Florian Toepfl, Daria Kravets, Anna Ryzhova, and Arista Beseler. Who are the plotters behind the pandemic? Comparing COVID-19 conspiracy theories in Google search results across five key target countries of Russia's foreign communication. *Information, Communication & Society*, 26(10):2033–2051, 2023. doi: 10.1080/1369118X.2022.2065213.

Benjamin Toff and Rasmus Kleis Nielsen. "I Just Google It": Folk Theories of Distributed Discovery. *Journal of Communication*, 68(3):636–657, 2018. doi: 10.1093/joc/jqy009.

Daniel Trielli and Nicholas Diakopoulos. Partisan search behavior and Google results in the 2018 U.S. midterm elections. *Information, Communication & Society*, 25:145–161, 2020. doi: 10.1080/1369118X.2020.1764605.

Francesca Tripodi. Devin nunes and the power of keyword signaling, 2019a. URL https://www.wired.com/story/devin-nunes-and-the-dark-power-of-keyword-signaling/.

Francesca Tripodi. Devin Nunes and the Power of Keyword Signaling | WIRED, 2019b. URL https://www.wired.com/story/devin-nunes-and-the-dark-power-of-keyword-signaling/.

Francesca Tripodi. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. Yale University Press, New Haven, 2022. URL https://yalebooks.yale.edu/book/9780300248944/the-propagandists-playbook.

Francesca Tripodi, Lauren C. Garcia, and Alice E. Marwick. 'Do your own research': Affordance activation and disinformation spread. *Information, Communication & Society*, 0(0):1–17, 2023. doi: 10.1080/1369118X.2023.2245869.

Yevgen Tsykynovskyy. Website representation vector, February 2020.

Elizabeth Tucker. Getting to great matches in Google Search, 2020. URL https://blog.google/products/search/getting-great-matches-google-search/.

United States of America. United states of america v. internet research agency llc et al. (case 1:18-cr-00032-dlf). United States District Court for the District of Columbia, 2018. URL https://www.justice.gov/file/1035477/download.

Aleksandra Urman, Mykola Makhortykh, Roberto Ulloa, and Jisun Kulshrestha. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, 72:101860, 2022a. doi: 10.1016/j.tele.2022.101860. URL https://doi.org/10.1016/j.tele.2022.101860.

Aleksandra Urman, Mykola Makhortykh, Roberto Ulloa, and Juhi Kulshrestha. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, 72:101860, 2022b. doi: 10.1016/j.tele.2022.101860.

Aleksandra Urman, Aniko Hannak, and Mykola Makhortykh. User Attitudes to Content Moderation in Web Search. *Proceedings of the ACM on Human-Computer Interaction*, 8 (CSCW1):146:1–146:27, 2024. doi: 10.1145/3637423.

Aleksei Urman, Mykola Makhortykh, Roberto Ulloa, and Juhi Kulshrestha. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, 72:101860, 2022c. doi: 10.1016/j.tele.2022.101860.

U.S. Department of State: Global Engagement Center. Pillars of russia's disinformation and propaganda ecosystem, 2020a. URL https://www.state.gov/wp-content/uploads/2020/08/Pillars-of-Russia%E2%80%99s-Disinformation-and-Propaganda-Ecosystem_08-04-20.pdf.

U.S. Department of State: Global Engagement Center. Pillars of russia's disinformation and propaganda ecosystem, 2020b. URL https://www.state.gov/wp-content/uploads/2020/08/Pillars-of-Russias-Disinformation-and-Propaganda-Ecosystem_08-04-20.pdf.

Marieke van Hoof, Corine S Meppelink, Judith Moeller, and Damian Trilling. Searching differently? How political attitudes impact search queries about political issues. *New Media & Society*, 2022. doi: 10.1177/14614448221104405.

Marieke van Hoof, Damian Trilling, Corine Meppelink, Judith Möller, and Felicia Loecherbach. Googling politics? Comparing five computational methods to identify political and news-related searches from web browser histories. *Communication Methods and Measures*, 19(1): 63–89, 2025. doi: 10.1080/19312458.2024.2363776.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Joe Walsh. U.s. reportedly accuses financial blog zero hedge of publishing russian propaganda. Forbes, 2022. URL https://www.forbes.com/sites/joewalsh/2022/02/15/us-reportedly-accuses-financial-blog-zero-hedge-of-publishing-russian-propaganda/.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, 2025.

Qi Wang, Chenxin Li, Chichen Lin, Weijian Fan, Shuang Feng, and Yuanzhong Wang. A news media bias and factuality profiling framework assisted by modeling correlation. *Computers, Materials & Continua*, 81(2), 2024.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

Mark Webster. Are keyword tools traffic estimates accurate? (case study), 2022. URL https://www.authorityhacker.com/traffic-estimates-accuracy/.

Ryen W. White. Advancing the Search Frontier with AI Agents. *Commun. ACM*, 67(9): 54–65, 2024. doi: 10.1145/3655615.

Evan M Williams and Kathleen M Carley. Search engine manipulation to spread pro-kremlin propaganda. *Harvard Kennedy School Misinformation Review*, 2023.

Evan M Williams and Kathleen M Carley. How alt-tech users evaluate search engines: Cause-advancing audits. *HARVARD KENNEDY SCHOOL MISINFORMATION REVIEW: Shorenstein Center for Media, Politics, and Public Policy*, 2025.

Evan M Williams, Peter Carragher, and Kathleen M Carley. Bridging social media and search engines: Dredge words and the detection of unreliable domains. In *Proceedings of*

the International AAAI Conference on Web and Social Media, volume 19, pages 2030–2043, 2025.

Sterling Williams-Ceci, Michael W. Macy, and Mor Naaman. Misinformation does not reduce trust in accurate search results, but warning banners may backfire. *Scientific Reports*, 14 (1):10977, 2024. doi: 10.1038/s41598-024-61645-8.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, 2024.

Kai-Cheng Yang and Filippo Menczer. Large language models can rate news outlet credibility. *arXiv preprint arXiv:2304.00228*, 2023.

Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, 38(1-2):98–139, 2021. doi: 10.1080/10584609.2020.1785067.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

Dave Van Zandt. Media bias fact check: A comprehensive media bias resource., 2022. URL https://mediabiasfactcheck.com/methodology.

Andrei Zavadski and Florian Toepfl. Querying the Internet as a mnemonic practice: How search engines mediate four types of past events in Russia. *Media, Culture & Society*, 41 (1):21–37, 2019. doi: 10.1177/0163443718764565.

Sonya Zhang and Neal Cabage. Search Engine Optimization: Comparison of Link Building and Social Sharing. *Journal of Computer Information Systems*, 57(2):148–159, April 2017. ISSN 0887-4417, 2380-2057. doi: 10.1080/08874417.2016.1183447. URL https://www.tandfonline.com/doi/full/10.1080/08874417.2016.1183447.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844, 2025.

Kawther Zoubi, Aviv J. Sharon, Eyal Nitzany, and Ayelet Baram-Tsabari. Science, Maddá, and 'Ilm: The language divide in scientific information available to Internet users. *Public*

Understanding of Science, 31(1):2–18, 2022. doi: 10.1177/09636625211022975.

Ethan Zuckerman. Why study media ecosystems? *Information, Communication & Society*, 24(10):1–19, 2021. doi: 10.1080/1369118X.2021.1942513.

Katharina A. Zweig. Watching the watchers: Epstein and robertson's 'search engine manipulation effect'. AlgorithmWatch, 2017. URL https://algorithmwatch.org/en/watching-the-watchers-epstein-and-robertsons-search-engine-manipulation-effect/.