Data Void Exploits: Tracking & Mitigation Strategies

Abstract

A data void is a gap in online information, providing an opportunity for the spread of disinformation or a *data void exploit*. We introduce lightweight measures to track the progress of data void exploits and mitigation efforts in two contexts: Web search and Knowledge Graph (KG) querying. We use case studies to demonstrate the viability of these measures as data void trackers in the Web search context. To tackle data voids, we introduce an adversarial game model involving two agents: a disinformer and a mitigator. Both agents insert content into the information ecosystem to have their narrative rank higher than their counterpart in search results. At every turn, each agent chooses which content to deploy within their resource constraints, mimicking real-world situations where different entities have varying levels of influence and access to resources. Using simulations of this game, we compare and evaluate different mitigation strategies to recommend ones that maximize mitigation impact while minimizing costs.

1 Introduction

Search begins with keywords. When there is a dearth of information online that is relevant to the keywords, we are in a data void [15]. Data voids are not inherently problematic. A random string such as “xydea8gya8g7” or “battery equator jargon apple” may return no results or a few pages that are irrelevant with respect to the search keywords. We do not care about such information voids. Traveling back in time to the 2016 US Elections, the seemingly random set of keywords “satan pizza hillary children” would have brought users to the carefully constructed content around the pizzagate conspiracy theory, which implicated Hillary Clinton in a child sex ring run from a pizza restaurant [2]. In the short time frame from content creation to its coverage and debunking in main stream media, searchers were directed to the problematic disinformation. We care about these data voids.

Disinformers have capitalized on the presence of data voids and the operation of search engines to drive information seekers to their narratives. Tripodi outlines how political agents have exploited the information consumption habits of Evangelical groups in the US to push right-wing agendas on taxation, liberal corruption, deep-state conspiracies, etc. [40] As information seekers self-discover the content by searching for specific keywords on Google, they deem it authentic as it was actively found rather than passively shared with them [40]. *Thus, an effective data void exploit can have deep and lasting impact on non-suspecting users.*

To understand how a data void exploit occurs and how a typical mitigation response works, consider the keyword search query in Figure 1, circa 2008. Disinformers manipulate search results for a fresh data void: “Obama born Kenya.” They add web pages (red content) with high search relevance for the void’s keywords. They may also deploy these pages in sites with higher PageRanks [30] or use Search Engine Optimization (SEO) [27] tactics to boost the ranking of their narrative. The exploit is not only limited to Web search results. Search engines like Google rely on structured data, stored in Knowledge Graphs (KGs), to extend search beyond merely matching keyword queries to pages, to provide users with faster and richer results [17, 18]. Since KGs suffer from incompleteness [43], they depend on continual data curation and augmentation for accuracy and coverage of new facts. This incompleteness allows attackers to inject fresh facts that “fill up" the data void. In Figure 1, disinformers further manipulate search results by adding KG facts (red edges) with high relevance for the void’s keywords. Mitigators respond to such attacks by also filling up the void with counter-content (green) to rank their narrative higher in search results.

Golebiewski and Boyd argue that *there is no easy “fix” for data voids* and that search platforms and disinformation mitigators need to work together to “identify vulnerabilities and respond to attacks” [15]. Their eye-opening report, however, leaves much to be determined as to how exactly can mitigators monitor and respond to data voids. Current search platforms are centralized corporations that either have limited bandwidth or little incentive to mitigate all forms of disinformation, especially those beyond their regional legal liability.

Starting, however, from the point of mitigators knowing

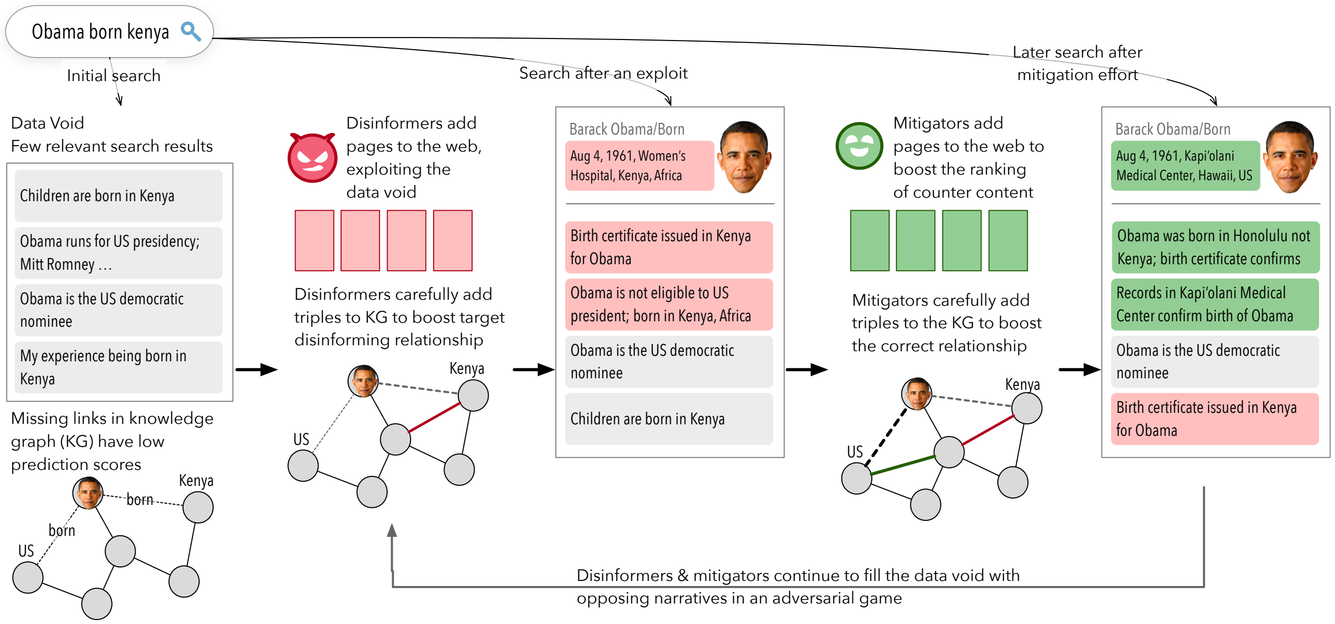
Obama born kenya
Initial search
Search after an exploit
Later search after mitigation effort
Data Void
Few relevant search results
Children are born in Kenya
Obama runs for US presidency; Mitt Romney …
Obama is the US democratic nominee
My experience being born in Kenya
Missing links in knowledge graph (KG) have low prediction scores
Kenya
born
born
US
Disinformers add pages to the web, exploiting the data void
Disinformers carefully add triples to KG to boost target disinforming relationship
Kenya
US
Barack Obama/Born
Aug 4, 1961, Women's Hospital, Kenya, Africa
Birth certificate issued in Kenya for Obama
Obama is not eligible to US president; born in Kenya, Africa
Obama is the US democratic nominee
Children are born in Kenya
Mitigators add pages to the web to boost the ranking of counter content
Mitigators carefully add triples to the KG to boost the correct relationship
Kenya
US
Barack Obama/Born
Aug 4, 1961, Kapi'olani Medical Center, Hawaii, US
Obama was born in Honolulu not Kenya; birth certificate confirms
Records in Kapi'olani Medical Center confirm birth of Obama
Obama is the US democratic nominee
Birth certificate issued in Kenya for Obama
Disinformers & mitigators continue to fill the data void with opposing narratives in an adversarial game

Figure 1: Data voids and how they evolve as disinformers (red) and mitigators (green) act to fill the void with content.

exactly what the problematic data void keywords are[1], we argue that we can use light-weight measures to track a data void and the effectiveness of both disinformation and mitigation efforts on Web and KGs. Given this tracker, we show that we can maximize the effectiveness of mitigation efforts given constraints on resources[2] or actions that one can take on third-party search platforms or KG Q&A systems.

In particular, we first demonstrate that we can use *search result rank* to determine the effectiveness and progress of disinformation or mitigation efforts with respect to a set of data void keywords. We provide historical evidence of Google search rank changes of disinformation and its counter information over time using a canonical data void case study about American politics and a case study around the recent Israel-Gaza War; battlegrounds also play out virtually online with each side promoting different narratives.

On demonstrating that a lightweight measure based on search rank can track data voids, we consequently show how it can also be used to direct how mitigators should respond or *what strategy to employ when promoting counter-content.*

In this paper, we have three main contributions. First, we describe in detail two data void exploit cases and show how search ranking can track the data void progression. In a novel case study, we present search results' ranks on how diverging narratives compete with a neutral one (§2).

Second, we model disinforming and mitigating agents as adversarial agents with limited control over the search or KG platforms and constrained resources. The agents have competing goals of promoting their content over the other. This novel problem formulation allows us to analyze the effects of different strategies in simulated environments and hence inform mitigators how to best tackle data voids (§3). We develop three strategies for each game that make different trade-offs with respect to effort, cost and impact.

Third, we empirically evaluate the effectiveness of different mitigation strategies across web search and KG querying (§4). We validate our simulation with real data from one of the case studies (§4.1.1). Results show that the choice of mitigation strategy is crucial in the initial phases of a data void: an aggressive mitigation strategy outperforms the baseline 95% of the time. We also examine the impact of delaying mitigation actions (§4.2). Finally, we discuss related works and differentiate our problem formulation and contributions (§5).

2 Case Studies

A data void refers to a situation where there is a lack of information on a specific topic in online search results. This gap allows misinformation or biased content to dominate the search results for a particular query.

Since Golebiewski's and Boyd's report on data voids, many researchers have analyzed and presented data void case studies: For example, *"crisis actor"* is a search query co-opted by conspiracy theorists to refer to victims of mass shootings to

[1] Mitigators are often aware of data void keywords: the disinformation narrative is circulating within closed networks or the content may not be easily searched initially [28, 40].

[2] Despite mitigators' awareness of brewing disinformation, resource scarcity plays a factor in deciding whether to create counter-content, since not all disinforming narratives gain sufficient traction to merit mitigation.

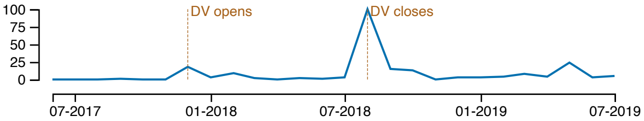

Figure 2: The data void surrounding the search keywords "Nellie Ohr" described in §(2.2) illustrates the typical search trend pattern: a slight uptick in search popularity indicating the opening, followed by a spike in interest and its mitigation.

prove that an event, such as the 2012 US Sandy Hook shootings, was staged [24, 44]. *"Iowa Caucus, rigged"* is another search query co-opted by conservative propagandists to undermine the integrity of the Iowa democratic caucus following a glitch in the app that presented election results [6, 31]. Finally, *"pizzagate"* is a search query curated by conspiracy theorists to connect Hilary Clinton and Jeffrey Epstein to a sex ring operation run from a pizza store [2, 14, 23, 25].

Most trend analyses focus on the popularity of search terms over time, as measured by Google Search Trends [10, 29], with some research examining social media platforms like Twitter or YouTube [21, 45]. In Google Search trends, a unifying picture across several data voids emerges [10]. First, an initial small spike in search interest of the data void terms signals its opening, followed by a long tail of low to moderate search interest, and finally a larger spike in search interest signals the mainstreaming of the keywords and an eventual closure of the void as authoritative sources counter the misinformation with content addressing the void keywords.

2.1 From Trend Analysis to Content Analysis

While search trend analysis has largely helped researchers understand the progression and impact of data voids from a societal interest perspective, it provides little insight into the actual content created by either disinformers or mitigators as the data void evolves. Ideally, we would like to keep track of content as it is created, indexed, searched and accessed [10]. However, this data, even though it can possibly be curated by major search platforms, is not available.

For a *historical analysis*, we consider an approximation of such data through the use of custom data range searches on Google. This is a proxy dataset for what searchers see on searching data void keywords at different points in time. We search for the data void keywords on Google using a custom date range: an unspecified start date and an increasing daily, weekly, or monthly end date. This creates an approximate snapshot of what users would see —- results and their search ranks — if they had queried the data void terms on different past dates. A typical analysis of this data set would show an absence of any relevant information prior to the emergence of the data void, followed by a gradual increase in the search rankings of misinforming content, and possibly an increase in the rankings of counter content. We note that custom range searches suffer from a few issues that might affect the accuracy of this analysis: e.g. (i) some results are missing date meta-tags, (ii) page snippets and historical search indices maintained by the search engine might become invalidated by content changes, and (iii) it is difficult to retrieve the actual contents of a page that were published at a given date and archival sites like the Web Archive have very low coverage.

For a *contemporary analysis* of an ongoing data void exploit, we execute a live search on Google on an hourly or daily basis to capture what and when content is created and ultimately what search results users would see as the data void evolves. No custom ranges are required. Data is pulled until the mitigator terminates the ongoing analysis.

Data Extraction Process. We implemented a pipeline that extracts search rank data for both forms of analysis. Search results are obtained through Selenium web automation [9], which emulates clicks on the Google webpage, and the Google Custom Search API [16]. After collecting each link from the search results within a time range, the pipeline extracts and stores a timestamped copy of the page content using the Trafilatura [3] library to extract text from the raw HTML.

In a contemporary analysis of a data void, the data extraction process runs periodically. This often results in the same URL appearing multiple times across search result snapshots, requiring the handling of various versions for the same webpage. Each newly extracted copy is compared with previously downloaded pages and only pages exhibiting significant differences from earlier versions are marked as new.

Using an LLM (ChatGPT-3.5), the pipeline automatically assesses whether the content of new pages is irrelevant to the data void, or if it falls into one of two or more categories. The categories for the first case study are *disinformation*, *mitigation* and *irrelevant*. For the second case study, we also support three labels: two diverging narratives (*Self-strike* vs *External attack*), plus *Neutral narrative*. Using a template, the analyst provides prompts that describe how to differentiate the categories. The initial zero-shot labeling by the LLM has been manually vetted for accuracy by the authors.

The pipeline generates a visualization with a color-coding of the top-50 search rank results over time and a line chart showing the aggregate inverse rank of pages in each category ($\sum \frac{1}{\text{rank}}$). The higher the sum of inverse ranks of content from one side, the more prominent it is in the search results indicating that it is *"winning"* the other side.

2.2 Case Study 1: Nellie Ohr

Overview: The Trump-Russia collusion investigation began in 2016 to determine if Trump and Russia colluded to manipulate the 2016 US Elections. Steele, a British Intelligence officer working at an intelligence firm, Fusion GPS, produced a 'dossier' with unverified claims on Trump's connections to Russia. This dossier was a point of contentious political discourse. In mid-2017, a name, Nellie Ohr, emerged in a

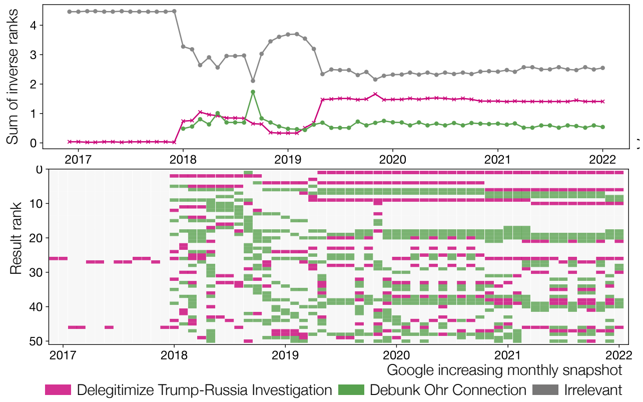

Figure 3: Prior to 2018, Nellie Ohr, was a relatively unknown figure. As conspiracy theorists exploited the "Nellie Ohr" data void, they began pushing disinforming content to discredit Trump-Russia collusion investigations. By August 2018, mitigators managed to subdue this narrative on online searches as mainstream media got involved, only for the disinforming content to get back and dominate web searches for Nellie Ohr.

conspiracy theory connecting FusionGPS, the dossier and the collusion allegations. Nellie Ohr's emergence into the discourse was not haphazard. It was due to a planned and concerted effort of *keyword curation*: Nellie Ohr was a data void — before then she was a relatively unknown figure with scant information about her online — that was ripe for exploitation. Exploiting this data void started by seeding the internet with stories about Nellie Ohr's connection to the Department of Justice through her husband, who was involved in the collusion investigations, across multiple platforms. QAnon, a far-right group, was an early propagator, introducing unverified information about her on platforms like Reddit and Twitter. These deliberate mentions filled the void with a particular narrative before authoritative sources could. Searches for "Nellie Ohr" would land information seekers on content delegitimizing the investigation. In 2018, *The Daily Caller* and other influential conservative platforms consistently echoed her name and the conspiracy theory effectively gaming search algorithms.

Results: The rank analysis in Figure 3 shows how the data void was initially filled by sites supporting the narrative that delegitimizes the investigation. When mitigators start to push content, both narratives climb up in the ranking. Over time, both agents put resources into the game with alternating success until the situation stabilizes after two years. Note that the highest mitigation peak coincides with the search trends pattern of Figure 2, which is attributed to the involvement of main stream media (e.g. NYTimes) in debunking the conspiracy [10]. This provides some evidence of the robustness of our methods in tracking certain historical data voids despite the limitations of custom range searches[3].

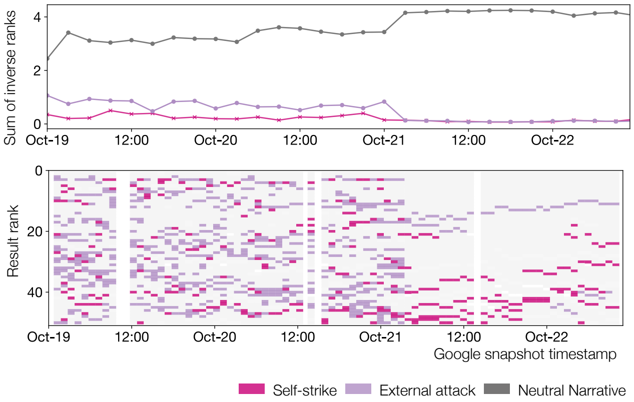

Figure 4: Immediately after the Gaza hospital explosion on Oct 17, 2023, two conflicting narratives started to emerge to fill the void: explosion was due to a *self-strike* or an *external attack* by Israeli forces. By Oct 21, an uptick in neutral narratives that present event timelines without taking sides can be seen dominating the conflicting narratives and filling the void instead. Gaps are due to temporary failures in the extraction pipeline. The aggregated inverse ranks are averaged over 3 hour intervals of the hourly snapshot results.

2.3 Case Study 2: Gaza Hospital Blast

Overview: On October 17, 2023, the Al-Ahli Arab Hospital in Gaza City was the site of a devastating explosion amidst the ongoing Israel-Hamas war [34]. This incident resulted in a high number of casualties among the displaced Palestinians seeking shelter at the hospital. The cause of the explosion is a matter of contention; while some sources attribute it to a failed rocket launch by the Palestinian Islamic Jihad, other sources attribute it to an Israeli airstrike. This event catalyzed widespread protests and had significant political and media repercussions, underscoring its contentious and complex nature.

Results: The rank analysis in Figure 4 shows how the data void is initially filled by three narratives. Both conflicting agents push their own narrative reaching the top ranks in the search results. After about two days, the neutral narrative becomes the dominating one and the conflicting narratives go down in the ranking. Main stream intervention (largely neutral) subdue other narratives, indicating the importance of having "influential pages" (e.g. higher PageRanks).

3 Data Voids as an Adversarial Game

We start by describing the attacker model for our agents. We then discuss the game playing scenarios, including three strategies to win the game. We then dive deep into the simulation models for the Web and the KG settings.

[3] On more sensitive topics such as shootings, search engines typically de-index problematic content, affecting the accuracy of retroactive analysis.

3.1 Attacker Model

We model data void exploits as a game played between two adversarial agents: a disinformer d and a mitigator m who each take turns choosing which content to deploy (e.g., which page in the case of web search, or which triple in the case of the KG). Their goal is to have their own content ranked higher by some user-facing algorithm accessing the information ecosystem, either Web or KG. This game applies to differing narratives, opinions, etc. beyond the narrow lens of factually incorrect as “disinformation" and fact-checks as “mitigation". Our assumption of a fixed set of resources to choose from mimics the resources and access constraints in real-world settings. For example, a disinformation campaign run by a state actor may have access to state-run, media news channels, whereas a political fact-checking team may not.

We study two settings. In the Web setting, agents compete in having their content ranked higher than the counter-part. In the KG setting, a data void is when the two competing, disinformation and mitigation, claims have low *link prediction* scores [43] — the graph may not even have either claim. Since link prediction scores determine the ranking of a claim as an answer to a KG query, we track and measure the effectiveness of disinformation or mitigation actions (e.g. adding new triples to the KG) in terms of this ranking of the competing claims. In the example in Figure 1, the agents add new triples[4] to increase the likelihood of the triple that supports their narrative — (Obama, born, Kenya) vs. (Obama, born, US).

An agent here can represent the actions of multiple decentralized agents with an overlapping agenda. The fact that online content is often produced by multiple different entities who may have little influence on each other is tangential to our analysis. First, prior work does shows that single entities (e.g. a state actors) do independently launch large-scale disinformation campaigns and decentralized campaigns are typically coordinated [28]. Second, the goal of our work is to inform a global strategy, which can direct the efforts of mitigation teams even if they operate independently. Nevertheless, our adversial agents in a game framing is amenable to future extensions where one can study the effects of coordination on disinformation or mitigation strategies.

Next, we define the rules of the game, including the main strategies, then instantiate it in the Web and KG settings.

3.2 The Game-playing Scenario

Five main elements define the game-playing scenario.

Turn. Each agent, d or m, selects a piece of information (x_d or x_m) from their **resource pools** (D, M) to modify the information ecosystem U_{t-1} at each turn t, where $\{t \in \mathbb{Z} | t \geq 1\}$. Let U_0 represent the data void and D_0, M_0 the initial resource pools, then

$$U_t := U_{t-1} \cup \{x_d, x_m\}$$

$$D_t := D_{t-1} - x_d$$

$$M_t := M_{t-1} - x_m$$

The specific choice of what information to use is guided by the agent’s strategy. At each turn both agents act simultaneously. An agent may skip their turn.

Effect. $\mathcal{E} : U \to \mathbb{R}_d \times \mathbb{R}_m$ Each turn alters the state of the information ecosystem, either amplifying the disinformation or its mitigation. The effect of each move is quantitively tracked by reevaluating the rank (a proxy for visibility and therefore influence) of the disinformation and its mitigation after each turn. For notational convenience, $\mathcal{E}_d$, and $\mathcal{E}_m$ respectively refer to the disinformation and mitigation components of effect.

Cost. $\mathcal{C} : X \to \mathbb{R}$ The cost in this game is assumed to be proportional to the *influence* of an information item within the information ecosystem. Often a proxy for influence is used. Metrics such as node centrality, pagerank, and degree can be used to determine how well connected a page is within the Web or a triple is within the KG and thus their influence or capacity to promote a certain narrative. More influential items are more costly, e.g., adding triples associated with a celebrity to a KG may undergo more scrutiny and vetting, hence more costs may be involved to circumvent or pass these checks than adding triples associated with a less-known figure.

Winning. $\mathcal{W} : (\mathbb{R}_D, \mathbb{R}_M) \to \{d, m\}$ Agents measure their success based on the rank of their content in the information ecosystem at each turn. The disinformer is winning when the disinformation has higher ranking, and conversely for the mitigator. Thus, $\mathcal{W}(E(U_t))$ will declare the disinformer agent d as winner if its effect (e.g. ranking) is higher than the mitigator m and the mitigator as winner otherwise at turn t.

While the initial resources an agent has could predetermine the game’s final outcome to a certain extent, we are more concerned about the immediate impact and the maximization of the mitigator limited resources’ effectiveness while minimizing costs. Therefore, it is not just about who wins in the end, but how effectively the actors influence the information ecosystem as the game progresses.

Strategy. $\mathcal{S} : X \times U \to x$ In the context of this game, a strategy is an agent’s set of rules that dictates which content to deploy when it is their turn. It guides the actions of an agent based on the available resources (D_{t-1} for d and M_{t-1} for m) and the current state of the information ecosystem (U_{t-1}). A strategy can take various forms, such as prioritizing deployment of the most crucial information first or deploying content in a random fashion. The choice and effectiveness of a strategy influences the course of the game.

The strategy is the most important element and the one that is controlled by the players to win the game. We study

[4] Most KGs are bootstrapped and updated according to open resources, such as DBpedia and Wikidata. While both agents can add triples, updates must be done carefully and parsimoniously to avoid spam and vandalism detection techniques [1, 19].

the following three noting that our model (an instantiated simulations) can be easily extended to support other strategies:

1. **Random**: In this baseline strategy, an agent chooses a random piece of content to add to their information pool each time. This strategy does not account for the impact or cost of the selected content; therefore, its result can vary greatly from one run to the other.

2. **Greedy**: The ranking of an information item in the ecosystem is often determined by a variety of factors including its relevance to the search query (e.g., keyword match similarity), the item's influence (e.g., pagerank), etc. In this strategy, the resource pool is sorted once, in decreasing order, apriori by a weighted combination of these factors. At each turn, the agents pulls the topmost item from this pool. Note that these factors alone do not determine the exact final ranking of an item: that depends on all the items currently in the information ecosystem U_t. This aggressive strategy often chooses more costly items to add first.

3. **Multiobjective Greedy**: A modification of the Greedy strategy, incorporating cost considerations. It sorts the items in the resource pool, once, in decreasing order, apriori using a weighted combination of search-rank factors and negatively weighted cost. This strategy aims to strike a balance between high impact and low cost at each turn, but may need fine-tuning the weights for different ecosystems.

Using this gamified abstraction of data voids, we build two simulators that model the actions and strategies of disinformers and mitigators in a *realistic* setting. In the simulated environment, the set of pages or claims that an agent has an access to needs to plausibly represent the influence that such a page has in a web setting or a claim has in a KG. We use search over Wikipedia pages as a stand in for searching over the web and link prediction over FB15k-237 as a stand in of KG answering[5]. These datasets are large enough to support rich queries and data void scenarios. They have representative content and graph properties of the Web graph or other KGs. They are also sufficiently small to enable the execution of complex analysis and simulations within a reasonable time frame on accessible computational resources[6].

3.3 Simulating the Web Search Game

We now describe our simulation of the game starting with Web search. A corpus of interlinked webpages, such as Wikipedia, is given as input. In this setting, an agent aims at making their own narrative more supported. Functionally, this means that the webpages supporting their narrative appear higher in search result ranking than the opponent's webpages.

The Resource Pools, D, and M. To compose the set of pages available to a disinformer, D, and to a mitigator M, we pick two seed pages representing divergent or conflicting viewpoints about topics in a domain, e.g., "Declarative Language vs. Procedural Language" or "Rationalism vs. Empiricism". A seed page is labeled arbitrarily as a disinformer, s_d, or mitigator page, s_m. D and M are the disjoint in-neighbors (N^-) of their corresponding seed pages: $D = N^-(s_d) - (N^-(s_d) \cap N^-(s_m))$ and $M = N^-(s_m) - (N^-(s_d) \cap N^-(s_m))$. Let U be the universe of Wikipedia pages, we then identify a set of data void keywords such that each keyword appears in approximately as many pages in D as in M and infrequently in $U - D - M$. Together the keywords cover all pages in $D \cup M$.

We construct the data void by removing all the pages in D and M from U (i.e., $U_0 := U - D - M; D_0 := D; M_0 = M$). At this starting point, a search for the data void keywords will yield results that are by construction irrelevant. As each agent adds a page from their respective set, they change the overall search results and their performance is evaluated by the ranks of their added pages in the search results.

We posit that constructing a data void by subtracting the disjoint in-neighbors of the seed pages realistically approximates an agent's capacity to influence web search. With this construction, we do not assume (i) the distribution of pageranks of the pages available to an agent, (ii) the graph properties of D, M or U, or the (iii) the sizes of D and M relative to U. These are naturally determined by the graph of pages in Wikipedia. This construction might also give one agent more power than the other (e.g, $\mathcal{W}(U) = d$). This is also true of real-world agents who have access to different resources.

Effect. The effect of a mitigator or disinformer's actions at turn t is reflected in the aggregate ranking of disinformer or mitigator pages in U_t. Thus,

$$\mathcal{E}(U_t) = \left(\sum_{x \in D} \frac{\mathbf{1}_{U_t}(x)}{\mathsf{rank}_{U_t}(x)}, \sum_{x \in M} \frac{\mathbf{1}_{U_t}(x)}{\mathsf{rank}_{U_t}(x)} \right)$$

We use the inverse-rank weighted sum of pages as a measure of effect: the higher the search rankings of a disinformer's pages (i.e., the page appears in the search results with lower-valued ranks), the higher the effect of its component and the lower the effect of the mitigator and vice-versa. This is because each rank can only have one page.

In our simulator, given the graph of web pages that contains all pages in U_t, $\mathsf{rank} : x \to \mathbb{N}$ is computed as follows for a collection of data void keywords:

1. Let $\mathsf{relevance} : x \to [0, 1]$ be a measure of how relevant documents are to the data void keywords: higher values mean more text matches are found in the page. We use Postgres's ts_rank, which takes into account lexical, proximity, and structural information [33].

[5] Of course, one can disagree with our choice of Wikipedia as a stand-in for the web and FB15k-237 as a stand-in for KGs in general. Nevertheless, as George Box famously remarks: "all models are wrong; some are useful."

[6] On a 128-node, single GPU, of a shared high-performance computing cluster, a single end-to-end run of all the experiments in this paper needs two weeks from data preparation to simulation and analysis

Narratives		**Size**		**Power**	**Data void keywords**	**Avg keyword frequency in**		
d	m	$\lvert D\rvert$	$\lvert M\rvert$	$\mathcal{E}(U)$		D	M	U_0
Declarative Language	Procedural Language	32	39	$(3.04, 1.36)$	lisp, semantics, javascript, ...	0.35	0.36	0.0006
Optimism	Pessimism	119	133	$(3.42, 2.69)$	nihilism, affective, depressive, ...	0.16	0.20	0.0005
Rationalism	Empiricism	58	117	$(3.04, 2.70)$	descartes, leibniz, gottfried, ...	0.30	0.38	0.0004
Classical Economics	Keynesian Economics	240	90	$(1.41, 4.97)$	maynard, keynes, laissez, faire, ...	0.32	0.33	0.0005
Delegitimize Investigations	Debunk Ohr Connection	13	19	$(1.71, 2.61)$	Nellie, Ohr	0.41	0.34	0.04

Table 1: Properties of the different simulated data void scenarios in the Wikipedia web data set, and the monitored, real Nellie Ohr data void scenario that unfolded between 2017 and 2022.

2. Let $\mathsf{pagerank} : x \rightarrow [0,1]$ be the numerical score assigned to a page by the PageRank algorithm [30]. It represents the likelihood that a random walk on the graph ends at page x. It is a measure of a page's relative influence. More central pages with higher in-degrees typically have higher pageranks. For example, a news media outlet like CNN has higher pagerank than a blog with few followers. Every turn, $U_0, U_1, \ldots$ requires the recomputation of pageranks as adding a page or a node to the web graph also adds the links to and from the page. However, as computing page rank at every turn is computationally expensive, we compute it once for all pages in U, with the assumption that certain pages stay more important than others.

3. The search $\mathsf{score}(x)$ of a page is now
$$\frac{1}{2}(\mathsf{relevance}(x) + \mathsf{pagerank}(x))$$
We sort all pages in descending order of search score breaking ties arbitrarily. Thus $\mathsf{rank}(x_1) < \mathsf{rank}(x_2)$ if $\mathsf{score}(x_1) > \mathsf{score}(x_2)$.

Cost. The cost of a page is determined by its pagerank capturing the intuition that pages with higher pageranks require more effort, access, influence or monetary resources to add.

$$\mathcal{C}(x) = e^{\mathsf{pagerank}(x)}$$

Winning. We determine the winner at every turn as follows:

$$\mathcal{W}(U_t) = \begin{cases} d & \text{if} \quad \mathcal{E}_d(U_t) > \mathcal{E}_m(U_t) \\ m & \text{otherwise} \end{cases}$$

Strategy. We implement the following three strategies in our simulator; at each turn t:

- **Random**. An agent randomly selects a page from its resource pool without replacement and adds its to U_{t-1}.
- **Greedy**. An agent pulls the top page from its pool (ordered in descending order of $\mathsf{pagerank}$) and adds its to U_{t-1}.
- **Multiobjective Greedy**. For each page x, we compute a linear weighted sum of an estimate of its cost and effect:
$$\frac{1}{2}\left(\mathsf{score}(x) - \frac{\mathcal{C}(x) - 1}{e - 1}\right)$$
Pages in the resource pool are ordered in descending order of this weighted sum. At each turn the agent pulls the top page from its pool and adds it to U_{t-1}.

Table 1 provides a summary of the properties of the different resource pools of disinformers and mitigators used in four different data void scenarios. The simulator can easily be extended to support more scenarios, different strategies, etc.

3.4 Simulating the KG Querying Game

We now describe our game simulation in the KG querying setting. A graph of interlinked entitiesis given as input. In this setting, the agents compete on making a triple (e.g., a fact such as the birthplace of a president) more supported and thus higher in the ranking than others by adding new KG triples.

The Resource Pools, D and M. We pick two target claims or triples s_d : (head, rel, tail$_d$) and s_m : (head, rel, tail$_m$) with the same head and relationship, but different tails. A query (head, rel, ?) returns both claims at different ranks according to their likelihood from a link prediction algorithm. For example, the query (Ben Affleck, directed, ?) on the FB15k-237 KG returns target claims (Ben Affleck, directed, Argo) at rank 2 and (Ben Affleck, directed, The Town) at rank 1. The claims are labeled arbitrarily as disinformer or mitigator claims. We choose four such one-to-many or many-to-many claims from FB15k-237 to examine KG data void exploits. For simulation run-time scalability, we subsample U, the universe of claims, such that U is the breadth-first neighborhood of both claims (5% of the entire KG).

We then search for the top-25 triples, $\hat{D}$, that explain the disinformer claim and similarly the top-25 triples that explain the mitigator claim, $\hat{M}$, using Kelpie's *necessary tail explanations* [37]. A triple is a necessary tail explanation if (i) it has the form (head, ?, ?) or (?, ?, head), and (ii) removing it from the KG reduces the tail prediction score of the target claim it explains. For example, given the target claim (Obama, nationality, United States), (United States, hadPresident, Obama) is a necessary tail explanation as it includes Obama and removing it reduces the tail prediction score of the target claim.

For each explanation triple x, Kelpie also produces a

Query		**Target Claim**		**Initial KG State**	**Size**	**Data void State**	**Power**
head	rel	d	m	$\mathcal{E}(\mathsf{KG})$	$\lvert D\rvert, \lvert M\rvert$	$\mathcal{E}(U_0)$	$\mathcal{E}(U)$
Ben Affleck	director	Argo	The Town	(0.5, 1)	21	(0.1, 0.01)	(0.25, 1)
George Clooney	actor	Good Night, and Good Luck.	Ocean's Twelve	(1, 0.24)	21	(0.14, 0.016)	(1, 0.25)
Ben Affleck	producer	Argo	Pearl Harbor	(0.5, 0.2)	23	(0.17, 0.013)	(0.25, 0.2)
Steven Spielberg	director	Saving Private Ryan	Amistad	(0.5, 1)	19	(0.1, 0.077)	(0.25, 0.5)

Table 2: Properties of the different simulated data void scenarios in the FB15k-237 knowledge graph.

relevance score. It is a straightforward score for ranking candidate triples for injection, more intricate methods for determining or ordering candidate triples can be explored in the future [4, 46]. $\mathsf{relevance} : x \to \mathbb{R}^+$ describes how well a triple explains the target claim. It is the expected rank worsening of the target claim associated with removing the triple. A higher value means that removing the triple will cause a higher increase in the rank-value of the target claim. $\hat{D}$ and $\hat{M}$ are the top-25 necessary tail explanations with the highest relevance scores.

We eliminate any triples that exist in both sets such that the disinformer triples are $D = \hat{D} - (\hat{D} \cap \hat{M})$ and mitigator triples are $M = \hat{M} - (\hat{D} \cap \hat{M})$ and we remove $\{s_d, s_m\} \cup (\hat{D} \cap \hat{M})$ from U. We construct the data void by removing the target claims and all the triples in D and M from U ($U_0 = U - D - M; D_0 = D; M_0 = M$). The query $(\mathsf{head},\ \mathsf{rel},\ ?)$ may return the target claims with very low rankings and prediction scores as they no longer exist in the KG and all the supporting (or explanation) triples have been removed. As each agent adds a triple from their set to the KG, the prediction score and the ranking of their target claim increases.

Again, our construction of a void by subtracting sets of triples pre-labeled as disinformer/mitigator based on how well they explain a target claim is a solid approximation of the agents' capacity to influence KGs. With this construction, we do not need to generate triples that an agent can plausibly add to boost a missing claim, we use what already exists. As with Web search, this construction might give one side more power than the other based on the set of triples. This also is true of real-world agents who can access different resources.

Effect. The effect of a mitigator or disinformer's actions at turn t is the inverse-rank of their target claim in U_t. Thus,

$$\mathcal{E}(U_t) = \left(\mathsf{rank}_{U_t}(s_d)^{-1}, \mathsf{rank}_{U_t}(s_m)^{-1}\right)$$

Unlike web search, in KG querying, especially when the KG suffers from incompleteness, *link prediction* is used for query answering. Here, a KG embedding (KGE) facilitates the prediction of a missing tail in a triple [13]. The link prediction algorithms in our simulator predicts answer tails for the given data void query $(\mathsf{head},\ \mathsf{rel},\ ?)$, with a score for each predicted triple [41]. Query answers are sorted in decreasing order of link prediction scores, which allows us to derive the rank of the mitigator or disinformer target claims.

Cost. Let $\mathsf{degree} : x \to \mathbb{Z}^+$ be the degree of the head or tail entity in x that is not the head entity of the data void query. We define cost as follows:

$$\mathcal{C}(x) = \mathsf{degree}(x)$$

This function captures the intuition that higher degree entities are more popular and are often subject to additional scrutiny when their facts are added to the KG. Hence, adding these triples requires more access, influence or resources [8].

Winning. We determine the winner at every turn as follows:

$$\mathcal{W}(U_t) = \begin{cases} d & \mathsf{if} \quad \mathcal{E}_d(U_t) > \mathcal{E}_m(U_t) \\ m & \mathsf{otherwise} \end{cases}$$

Strategy. We implement the following three strategies:

- **Random.** At turn t, an agent randomly selects a triple from their resource pool without replacement and adds it to U_{t-1}.
- **Greedy.** An agent's resource pool is ordered in decreasing order of the triple's relevance. At each turn t, the agent pulls the top triple from the pool and adds it to U_{t-1}.
- **Multiobjective Greedy.** For each triple x, we compute a linear weighted sum as an estimate of its cost and effect:

$$\frac{1}{2}\left(\mathsf{relevance}(x) - \frac{\mathcal{C}(x) - 1}{\max_{y \in U} \mathcal{C}(y) - 1}\right)$$

Triples in the resource pools, D or M, are ordered in decreasing order of this weighted sum. At each turn t the agent pulls the top page from its pool and adds it to U_{t-1}.

Table 2 shows for each of the target disinformer-mitigator claims, their initial effect (inverse ranks), $\mathcal{E}(\mathsf{KG})$, their effect, $\mathcal{E}(U)$, after they are removed, and after their supporting sets D, M are removed, $\mathcal{E}(U_0)$.

4 Experimental Results

Given our framing of data void exploits and responses as an adversarial game between disinformer and mitigator agents, we set out to answer the following research questions:

RQ1. *Which strategies should the agents employ to maximize the effect of their actions?*

RQ2. *What is the impact of a delayed mitigation strategy?*

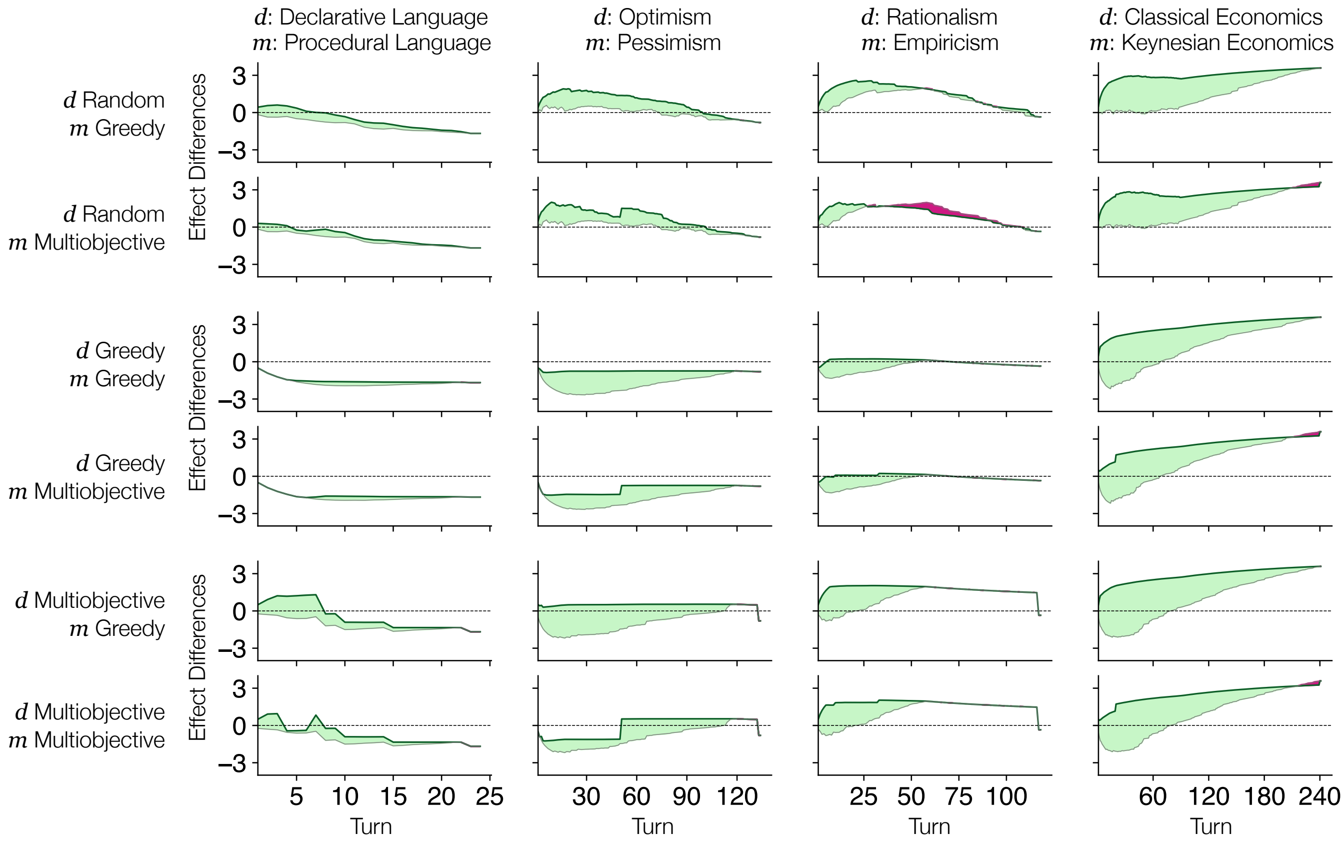

Figure 5: Differences of effects $\mathcal{E}_m(U_t) - \mathcal{E}_d(U_t)$ at every turn t of the Web search simulation across four data void scenarios. A positive difference indicates that the mitigator is *winning*. For every disinformer strategy (*d*: Random, Greedy, or Multiobjective Greedy), we plot the differences of effects for two mitigator strategies (*m*: Greedy and Multiobjective greedy) against the Random strategy baseline. If the evaluated strategy is better than the baseline (i.e. higher effect differences), we shade the area showing how much better it performs in green. If it is performing worse than the baseline, we shade the area showing how much worse in red. Overall, Greedy and Multiobjective Greedy strategies are better than Random ones and they both can help a mitigator win at least initially or if the disinformer is not strategic (e.g. employs a Random strategy) even if the disinformer has a stronger resource pool. For random, we compute the mean of effect differences over 15 simulations.

4.1 RQ1: Winning Strategies

Assuming a fixed set of resources available to each agent, D or M, we know that the final outcome after the agents add all their pages or triples back to the information ecosystem is predetermined: the agent with the relatively more powerful set of resources finally wins. We are therefore not as interested in this final outcome, but are rather more interested in which strategies are more impactful and more cost-effective over the course of the simulation. So, even if an agent never strictly *wins* (§3), they did the best with what they have.

Figures 5 and 7 illustrate the effects of simulating different strategies across the data void scenarios described in Tables 1 (Web setting) and 2 (KG setting). Here, we establish the Random strategy (§3.2) as the baseline strategy to beat. We fix the strategy of the disinformer to one among Random, Greedy and Multiobjective; the choice of the disinformer's strategy does not influence the choice of the mitigator's strategy. When simulating the random strategy, we compute averages and variance over 15 runs in web search and over 10 runs in KG querying. We then evaluate the performance of the mitigator when employing one of two strategies, Greedy or Multiobjective, against its performance when employing the Random strategy. Each plot in Figures 5 and 7 thus shows (i) the baseline performance of $\mathcal{E}_m(U_t) - \mathcal{E}_d(U_t)$ at every turn t when the mitigator is employing the Random strategy — a gray line — and (ii) the performance of the evaluated strategy — a thick green line. The shaded area between the two curves illustrates how well or poorly a strategy performs when compared to the baseline. We shade this region green to indicate that the evaluated strategy is outperforming the baseline and red otherwise. If at turn t, we are above or at the 0 line, the mitigator strategy is also *winning*, in the strict sense of its effect being

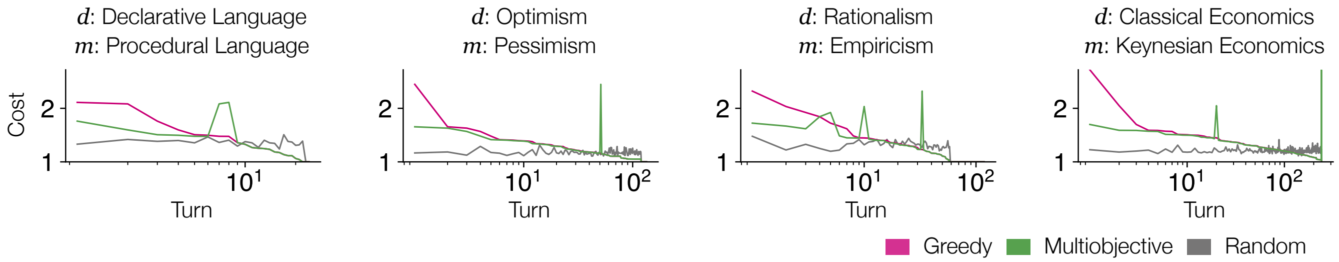

Figure 6: Mitigator costs for different strategies at each turn of the web search simulation across four data void scenarios. Actual disinformer's strategy does not influence the choice of the mitigator's strategy. For random, we plot mean costs over 15 simulations. Greedy selects pages with high Pagerank first and thus spends more initially. Multiobjective strategy balances cost with search relevance; as it runs out of lower-cost pages that are relevant, it adds the higher cost ones thus exhibiting spending peaks at later turns. Turn axis in log-scale to emphasize the initial spending behavior of the different strategies.

greater than or equal to the disinformer strategy at turn t.

In Figures 6 and 8, we also plot the costs of each strategy as more of the mitigator's pages or triples are added at every turn. As the mitigator picks a strategy without information about the disinformer strategy at hand, the cost graphs are independent of the latter.

We find the following insights from the analysis of the results for the Web search game (Figures 5, 6):

▶ Greedy is the most aggressive strategy and allows mitigators to get ahead of an emerging data void scenario even if their resources overall are limited. This is especially true if the disinformer is non-strategic, i.e., using a Random strategy, or cost-cutting with a Multiobjective strategy. In the first three scenarios of the web game (Figure 5), the disinformer ultimately wins. But following a Greedy strategy allows the mitigator to maximize their effect for the longest duration of turns initially. Across all scenarios, Greedy outperforms the other strategies 95% of the time, in the first half of the simulation. This might be important in a situation where the mitigator wishes to get ahead of a trending situation and reach early searchers of the data void, limiting exposure and a possible escalation of the disinforming narrative.

▶ A Random strategy has little chance at outperforming a strategic disinformer with better resource pools — across all scenarios, Random only outperforms other strategies 3% of the time in the first half of the simulation. For example, notice the gray line in the second data void scenario where the disinformer promotes 'Optimism' and uses either a Greedy or a Multiobjective strategy and the mitigator promotes 'Pessimism' and uses a Random strategy.

▶ Multiobjective strategies have less pronounced effects compared to Greedy strategies. While they can outperform Random strategies, they may not yield as many "wins" against the disinformer. Nonetheless, initially, they are less expensive than greedy strategies (Figure 6).

We obtain similar insights from the analysis of the results for the KG query answering game (Figures 7 and 8):

▶ Both Greedy and Multiobjective strategies are more effective than the Random strategy. Across all scenarios, the informed strategies give overall better results.

▶ However, the role of the data is even more important than that played by the strategies. All strategies fall short in the cases where the mitigator's resources are less effective than the disinformer's ones. In the first and last data void scenarios of Figure 7, when the mitigator has overall a more effective set of triples to choose from, a Greedy or Multiobjective strategy performs much better than a Random strategy. However, for the second and third scenarios, the mitigator can barely overtake the disinformer and there is little benefit to Greedy or Multiobjective strategies compared to Random.

▶ As in the Web setting, the Greedy strategy quickly consumes more resources, as depicted in Figure 8. The Multiobjective starts lower while providing comparable performance according to Figure 7.

4.1.1 Validation with Case Study 1

Using the monthly historical search results for the Nellie Ohr data void described in §2.2, we examine in Figure 9 the impact of choosing one of the three strategies on the mitigation efforts in the web search setting.

Unlike the previous evaluation, we set the baseline strategy for disinformer and mitigator to the actual order in which pages started to appear on Google search. We then plot the differences in effect when the mitigator uses one of three strategies (Random, Greedy, Multiobjective). We estimate the Pagerank of every page using Moz URL Metrics [38].

We find the following:

▶ Both Greedy and Multiobjective outperform Random and the observed baseline strategy, but with Greedy the mitigator beats the disinformer for more turns initially.

▶ The observed (baseline) mitigator strategy and random strategy are similar. This is not surprising as we believe different mitigators may add web pages without coordination.

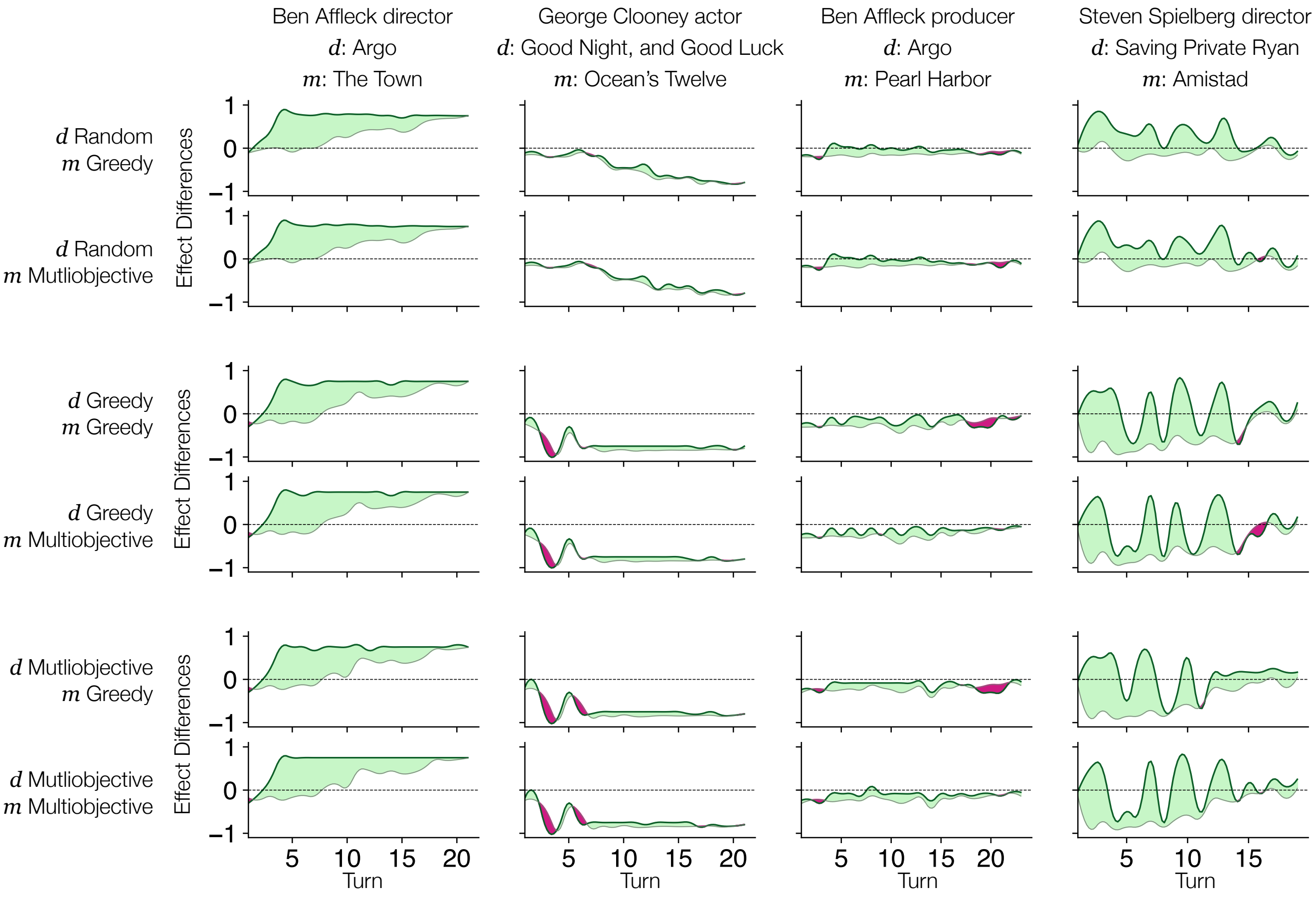

Figure 7: Differences of effects $\mathcal{E}_m(U_t) - \mathcal{E}_d(U_t)$ at every turn t of the KG query answering simulation across four data void scenarios. A positive difference indicates that the mitigator is *winning*. For every disinformer strategy (*d*: Random, Greedy or Multiobjective Greedy), we plot the differences of effects for two mitigator strategies (*m*: Greedy and Multiobjective greedy) against the Random strategy baseline. If the evaluated strategy is better than the baseline (i.e. higher effect differences), we shade the area showing how much better it performs in green. If it is performing worse than the baseline, we shade the area showing how much worse in red. For random, we compute the mean of effect differences over 10 simulations.

Practical Takeaways.

► In both the Web and KG settings, in the case of early detection of a new data void, mitigators have alternative options for *matching* the disinformer depending on the quality of the available resources and budget.

► For better ranking results and faster impact, an informed strategy (Greedy, Multiobjective) should be favored to a Random one, with Multiobjective chosen in cases of a limited budget.

► Determining the relative impact of a triple — a prerequisite for the Greedy strategy in a KG — might be difficult if the mitigators do not have access to the full KG or its link prediction models. Figure 8 shows a correlation between cost, determined by degree, and the order of edges added by Greedy strategy. Hence, triple degree can be an indication of its relevance.

► Greedy strategies rely on access to "influential" pages or entities. For early suppression, mitigators, need to create the infrastructure to support Greedy strategies, for example, by creating consortia, hyperlinking their resources, attracting the sponsorship of high-value entities or nodes, to maximize the ranking of deployed mitigation content by search engines and KG answering systems[7].

[7] In the "Nellie Ohr" case disinformers promoted their narrative on The Daily Caller — a site with high influence, on par with the Washington Post, as determined by Moz's SEO metrics [38]. The Department of Justice, which responded, ranked low on SEO metrics [38]. This disparity can be found globally, e.g. Italy's state-run media Rai ranks as high as the Washington Post, but Pagella Politica, a political fact-checker, is on par with the DOJ.

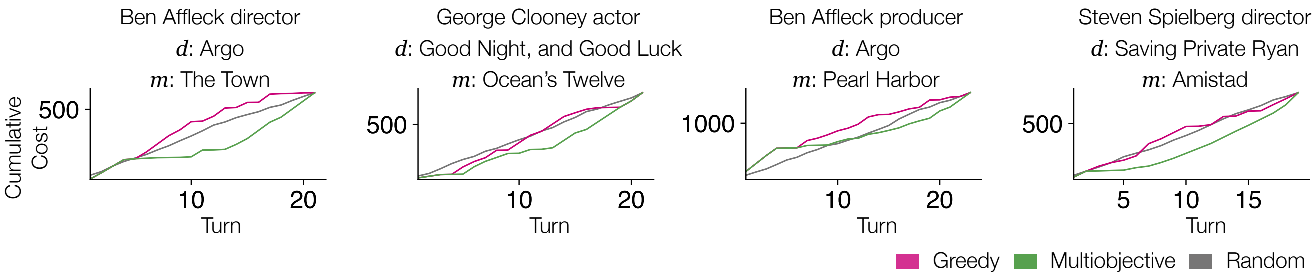

Figure 8: Mitigator's *cumulative* costs for different strategies at each turn of the KG query answering simulation across four data void scenarios. Actual disinformer's strategy does not influence the choice of the mitigator's strategy. For random, we plot mean cumulative costs over 10 simulations. At the end of the simulation, all strategies spend the same amount. Given the degree of a triple, which determines its cost and relevance, Greedy spends more initially. Multiobjective balances cost with relevance, reserving more expensive but less relevant triples to the end and thus under-spending initially w.r.t. Greedy and Random.

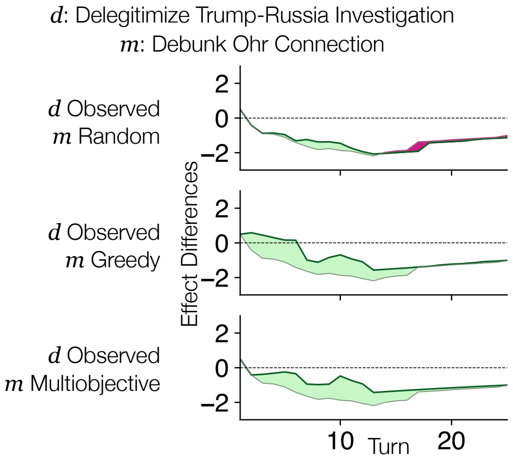

Figure 9: Differences of effects $\mathcal{E}_m(U_t) - \mathcal{E}_d(U_t)$ at every turn t of simulating the web search game with data scraped from the Nellie Ohr case study (§2.2). A positive difference indicates that the mitigator is winning. Disinformer adds pages in the same order that the page appeared during the 2017-2022 timeline. Green denotes the benefit of the mitigator strategy (Random, Greedy, Multiobjective) against the order in which the mitigator's pages appeared during the search timeline. The random strategy is the mean difference over 15 simulations.

4.2 RQ2: The Impact of Delay

Mitigators often wonder whether to engage in an emerging data void scenario or wait in case it fails to yield any traction, hence utilizing their limited resources on other more pressing mitigation efforts [28]. Taking a decision on when to react is difficult, requiring to a certain degree some foresight. However, our simulator can help navigate such complex decision making processes. We illustrate the impact of delay across data void scenarios: mitigators can map an emerging situation to their resources and the disinformer's resources and strategy to best determine their response strategy

In Figure 10, we present the cost that a mitigator must spend to win or overtake the disinformer (i.e., $\mathcal{E}_m(U_t) \geq \mathcal{E}_d(U_t)$) if they delayed their efforts for t turns such that $(U_0 \cup \ldots \cup U_{t-1}) \cap M =$ and at turn t, they add as many pages as needed from their resource pool to U_{t-1} using a Random, Greedy, or Multiobjective strategy). The plots also shows the percentage of mitigator wins: if the entire bar is colored in, the mitigator won 100% of the time; otherwise the proportion of the bar's height that is colored in is the percentage of mitigator wins. For example, in the "Declarative vs. Procedural" data void, after a delay of 12 turns, a Multiobjective mitigator wins in 20% of the 5 random runs by the disinformer.

We find the following:

▶ Given a fixed mitigator resource pool and a strong disinformer, after a certain delay, the mitigator cannot win. This is seen in Figure 10 in the case of the declarative vs. procedural language data void, and in the case of optimism vs. pessimism, when the disinformer is adopting an aggressive Greedy strategy.

▶ A Random response strategy is expensive! In the Classical vs. Keynesian economics data void, a Random mitigator strategy, on average, is 4.4x more expensive than a Greedy one and 4.2x more expensive that a Multiobjective one. It is 2x more expensive than either strategy in the Rationalism vs. Empiricism scenario.

Practical Takeaways. If a mitigator chooses to wait, it is often more cost-effective to respond aggressively after the delay using a Greedy strategy. Here, quality, not quantity is key. A delayed response can thus be beneficial if the mitigator uses the delay to gain access to an influential resource rather than divesting efforts and resources on low-impact counter-content.

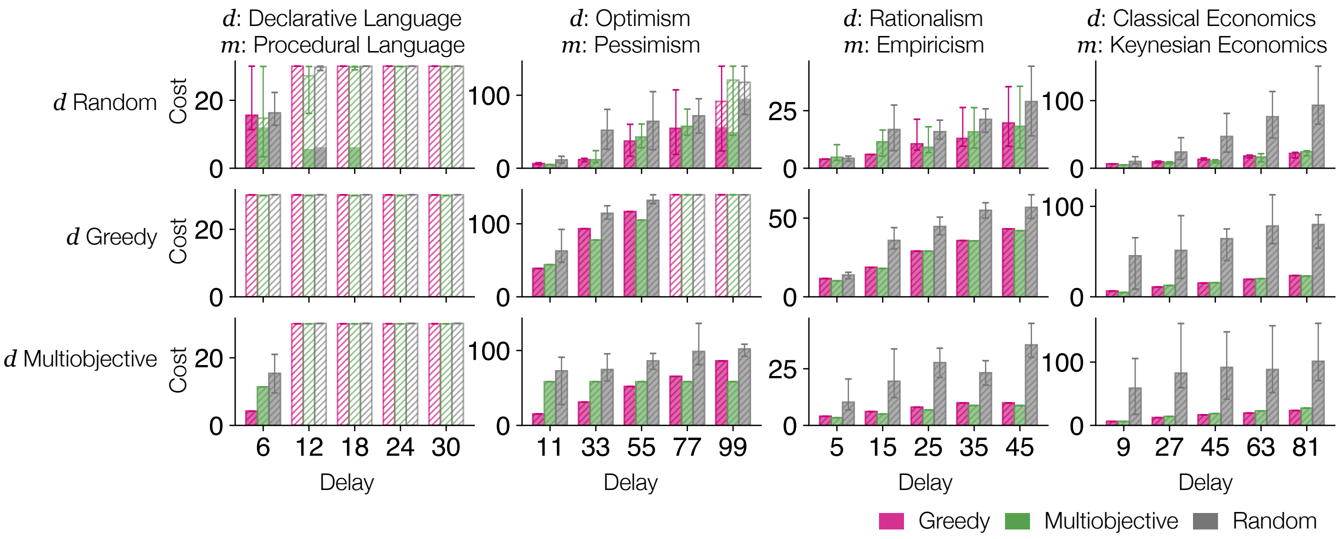

Figure 10: Cost incurred by a mitigator to win after a delay of t turns. Bars are colored in proportion to the percentage of mitigator wins: a fully colored in bar means the mitigator wins 100% of the time. E.g., the Multiobjective mitigator in the top-left plot beat the Random disinformer once across 5 runs (20%) at a delay of 12 turns. Random strategies were run 5 times with error bars showing the range of costs across all runs.

5 Related Work

Our work lies in the intersection of two fields: Web Search and KG querying attacks, and data void studies.

In our Web Search game, the agent's objective is to enhance the ranking of a website. This game aligns with studies exploring the malicious or intentional manipulation of ranking systems, including PageRank vulnerability exploits [7], web spam detection and mitigation [39, 42], and search engine optimizations and poisoning [5, 26, 27]. We opt for straightforward techniques to illustrate our framework. More intricate strategies can be integrated in the future.

In our KG querying game, the agent's actions resemble those of *data poisoning attacks* - a type of adversarial assault on ML models where the attacker manipulates a subset of the training data to tailor the model to fulfill specific goals [46]. More precisely, both agents act as attackers by inserting triplets in the KG. These triplets act as training data for the link prediction model that determines the ranking of the agent's target triple. KG poisoning attacks [4, 11, 32, 46] often study a variety of direct and indirect attacks including deletions and relationship modifications. We simulate agent attacks as selecting which triples to add from a predetermined set of facts derived from a full KG that is reduced for the simulation. We do not investigate how to create "new" triples to manipulate link prediction scores. Future research can integrate these sophisticated tactics into our simulations.

Several studies explore the presence and impact of data void exploits and the challenges of early detection [15, 29, 40]. Beyond web search and KG answering, the impact of data voids on search-adjacent systems such as auto-play, auto-fill has also been studied [20, 22, 35, 36]. A recent system [12], helps mitigators construct the text of the counter-content by using NLP and ML techniques to analyze the disinformation text filling up a data void in terms of racial bias, political leaning, etc. However, we are the only work that proposes a concrete method to track the progress of data voids, and to evaluate mitigation strategies in terms of content deployment, thus providing a tangible resource for mitigators on how to take action beyond the specifics of the actual content creation.

6 Conclusion

We develop rank-based measures to track the progress of disinformers or mitigators in filling up data voids in the Web and KGs. We illustrate the power of such a tracker with real case-studies. We formulate data void exploits and response as an adversarial game between disinformers and mitigators and use a simulator modeled on the game to help mitigators determine effective response strategies given their resource constraints. Future work directions include extending the framework with more sophisticated (and costly) strategies and, given the role of data in our simulations, better estimation of the effectiveness of the content available to agents [4, 11]. Finally, we plan to conduct user studies on emerging real-world data voids with information integrity teams, who list monitoring tools as a missing asset in assessing disinformation threats [28]. With this work, we now have a practical way to forward the conversation on data voids from "no easy fix" [15] to we can find cost-effective ways to tackle them.

References

[1] Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing* (2011), Springer, pp. 277–288.

[2] Aisch, G., Huang, J., and Kang, C. Dissecting the #pizzagate conspiracy theories. *The New York Times* (Dec 2016).

[3] Barbaresi, A. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (2021), Association for Computational Linguistics, pp. 122–131.

[4] Bhardwaj, P., Kelleher, J., Costabello, L., and O'Sullivan, D. Poisoning Knowledge Graph Embeddings via Relation Inference Patterns. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online, Aug. 2021), Association for Computational Linguistics, pp. 1875–1888.

[5] Boucher, N., Pajola, L., Shumailov, I., Anderson, R., and Conti, M. Boosting big brother: Attacking search engines with encodings. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses* (New York, NY, USA, 2023), RAID '23, Association for Computing Machinery, p. 700–713.

[6] Chariton, J. Investigator: Dnc was "directly involved" in iowa caucus app development, countering dnc denial. *The Intercept* (2020). Accessed: 2024-01-15.

[7] Cheng, A., and Friedman, E. J. Manipulability of pagerank under sybil strategies.

[8] Deshpande, O., Lamba, D. S., Tourn, M., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2013), SIGMOD '13, Association for Computing Machinery, p. 1209–1220.

[9] Developers, S. selenium, 2023. Python package version 4.14.0.

[10] EC, Copeland, R., Fan, J., and Jaeel, T. Filling the data void, 2020.

[11] Fang, M., Yang, G., Gong, N. Z., and Liu, J. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th annual computer security applications conference* (2018), pp. 381–392.

[12] Flores-Saviaga, C., Feng, S., and Savage, S. Datavoidant: An ai system for addressing political data voids on social media. *Proc. ACM Hum.-Comput. Interact. 6*, CSCW2 (nov 2022).

[13] Ge, X., Wang, Y.-C., Wang, B., and Kuo, C. C. J. Knowledge graph embedding: An overview, 2023.

[14] Gillin, J. How pizzagate went from fake news to a real problem for a d.c. business. *PolitiFact* (2016). Accessed: 2024-01-15.

[15] Golebiewski, M., and Boyd, D. Data voids: Where missing data can easily be exploited. Tech. rep., Data & Society Research Institute, 2019.

[16] Google. Google custom search json api overview, 2024.

[17] Google Blog. Introducing the knowledge graph: things, not strings, 2012. Accessed on 2024-01-15.

[18] Google Support. The knowledge graph, 2024. Accessed on 2024-01-15.

[19] Green, T., and Spezzano, F. Spam users identification in wikipedia via editing behavior. In *Proceedings of the International AAAI Conference on Web and Social Media* (2017), vol. 11, pp. 532–535.

[20] Hussein, E., Juneja, P., and Mitra, T. Measuring misinformation in video search platforms: An audit study on youtube. *Proc. ACM Hum.-Comput. Interact. 4*, CSCW1 (may 2020).

[21] Juneja, P., Bhuiyan, M. M., and Mitra, T. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on youtube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023), CHI '23, Association for Computing Machinery.

[22] Juneja, P., Bhuiyan, M. M., and Mitra, T. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on youtube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023), CHI '23, Association for Computing Machinery.

[23] Kim, C. Pizzagate, qanon, and the 'epstein list': Why the far right is obsessed with sex trafficking. *Politico* (2024). Accessed: 2024-01-15.

[24] Koebler, J. Where the 'crisis actor' conspiracy theory comes from. *Vice* (2018). Accessed: 2024-01-15.

[25] LaCapria, K. Is comet ping pong pizzeria home to a child abuse ring led by hillary clinton? *Snopes* (2016). Accessed: 2024-01-15.

[26] Leontiadis, N., Moore, T., and Christin, N. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2014), CCS '14, Association for Computing Machinery, p. 930–941.

[27] Lewandowski, D., Sünkler, S., and Yagci, N. The influence of search engine optimization on google's results: A multi-dimensional approach for detecting seo. In *Proceedings of the 13th ACM Web Science Conference 2021* (New York, NY, USA, 2021), WebSci '21, Association for Computing Machinery, p. 12–20.

[28] Mirza, S., Begum, L., Niu, L., Pardo, S., Abouzied, A., Papotti, P., and Popper, C. Tactics, threats and targets: Modeling disinformation and its mitigation. In *NDSS 2023, Network and Distributed System Security Symposium, 27 February-3 March 2023, San Diego, California, USA* (San Diego, 2023), Usenix, Ed.

[29] Norocel, O. C., and Lewandowski, D. Google, data voids, and the dynamics of the politics of exclusion. *Big Data & Society 10*, 1 (2023), 20539517221149099.

[30] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford Digital Library Technologies Project, 1998.

[31] Pager, T. Iowa autopsy report: Dnc meddling led to caucus debacle. *Politico* (2020). Accessed: 2024-01-15.

[32] Pezeshkpour, P., Tian, Y., and Singh, S. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 3336–3347.

[33] PostgreSQL. Text search ranking - postgresql documentation, 2023.

[34] Press, A. Israel-hamas war, 2023. Accessed: January 19, 2024.

[35] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., and Meira, W. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 131–141.

[36] Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., and Wilson, C. Auditing partisan audience bias within google search. *Proc. ACM Hum.-Comput. Interact. 2*, CSCW (nov 2018).

[37] Rossi, A., Firmani, D., Merialdo, P., and Teofili, T. Explaining link prediction systems based on knowledge graph embeddings. In *Proceedings of the 2022 International Conference on Management of Data* (New York, NY, USA, 2022), SIGMOD '22, Association for Computing Machinery, p. 2062–2075.

[38] SEOMoz Inc. Moz url metrics, 2024.

[39] Spirin, N., and Han, J. Survey on web spam detection: principles and algorithms. *SIGKDD Explor. Newsl. 13*, 2 (may 2012), 50–64.

[40] Tripodi, F. B. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. Yale University Press, New Haven and London, 2022.

[41] Trouillon, T., Welbl, J., Riedel, S., Éric Gaussier, and Bouchard, G. Complex embeddings for simple link prediction, 2016.

[42] Vani, M. S., and Babu, O. Y. A survey on link based algorithms for web spam detection.

[43] Wang, Q., Mao, Z., Wang, B., and Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering 29*, 12 (2017), 2724–2743.

[44] Williamson, E. Sandy hook hoax: The power of conspiracy theories. *NPR* (2022). Accessed on 2024-01-15.

[45] Yan, M., Lin, Y.-R., and Chung, W.-T. Are mutated misinformation more contagious? a case study of covid-19 misinformation on twitter. In *Proceedings of the 14th ACM Web Science Conference 2022* (New York, NY, USA, 2022), WebSci '22, Association for Computing Machinery, p. 336–347.

[46] Zhang, H., Zheng, T., Gao, J., Miao, C., Su, L., Li, Y., and Ren, K. Data Poisoning Attack against Knowledge Graph Embedding. 4853–4859.