Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Sister Conferences Best Papers Track

Data Void Exploits: Tracking & Mitigation Strategies (Extended Abstract)*

Miro Mannino[1], **Junior Garcia**[1], **Reem Hazim**[1], **Azza Abouzied**[1] and **Paolo Papotti**[2]
[1]NYU Abu Dhabi
[2]EURECOM
{miro.mannino, juniorgarcia, rh3015, azza}@nyu.edu, papotti@eurecom.fr

Abstract

In the evolving landscape of online information, disinformation is a growing concern. A concept central to this challenge is the "data void", a situation where there is a lack of relevant information online regarding certain search terms. This creates an opportunity for misleading or false narratives to fill the gap, often influencing public perception. In this work, we present methods to track and mitigate data voids in Web search settings.

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 [Golebiewski and Boyd, 2019; Norocel and Lewandowski, 2023]. Not all data voids are problematic. For instance, random strings or uncommon phrases might not yield any meaningful search results without consequence. However, when the gap involves significant or emerging topics, it becomes a target for exploitation. In the historical example of "Pizzagate," disinformation filled the void around conspiracy keywords, thus directing users towards false narratives about a political figure [Aisch *et al.*, 2016].

Disinformers have capitalized on the presence of data voids and the operation of search engines to drive information seekers to their narratives [Williamson, 2022; Koebler, 2018; Chariton, 2020; Pager, 2020]. Tripodi outlines how political agents have exploited the information consumption habits of Evangelical groups to push right-wing agendas [Tripodi, 2022]. As information seekers self-discover the content by searching for specific keywords, they deem it authentic as it was actively found rather than passively shared with them.

To understand how a data void exploit occurs and how a 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 high PageRanks [Page *et al.*, 1998] to boost the ranking of their narrative. The exploit is not limited to Web search results. Search engines rely on structured data, stored in Knowledge Graphs (KGs), to extend search beyond matching keyword queries to pages, to provide users with faster and richer results [Google Blog, 2012; Google Support, 2024]. Since KGs suffer from incompleteness [Wang *et al.*, 2017], they depend on continual data curation and augmentation for accuracy and coverage of new facts [Ortona *et al.*, 2018]. 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.

There is no easy "fix" for data voids and search platforms and mitigators need to work together to "identify vulnerabilities and respond to attacks" [Golebiewski and Boyd, 2019]. 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 either have limited bandwidth or awareness to mitigate all forms of disinformation, especially those beyond their regional legal liability.

Starting, however, from the point of mitigators knowing exactly what the problematic data void keywords are [Tripodi, 2022; Mirza *et al.*, 2023], we argue that we can use lightweight 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 or actions that one can take on third-party search platforms.

In particular, we demonstrate that *search result rank* can 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 data void case study about American politics.

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 abstract, originally presented as a full paper at CIKM 2024 [Mannino *et al.*, 2024], we report three main contributions. First, we describe a data void exploit case and show how search ranking can track the data void progression (§2). Second, we model disinformers and mitiga-

*Original paper presented at CIKM 2024 [Mannino *et al.*, 2024].

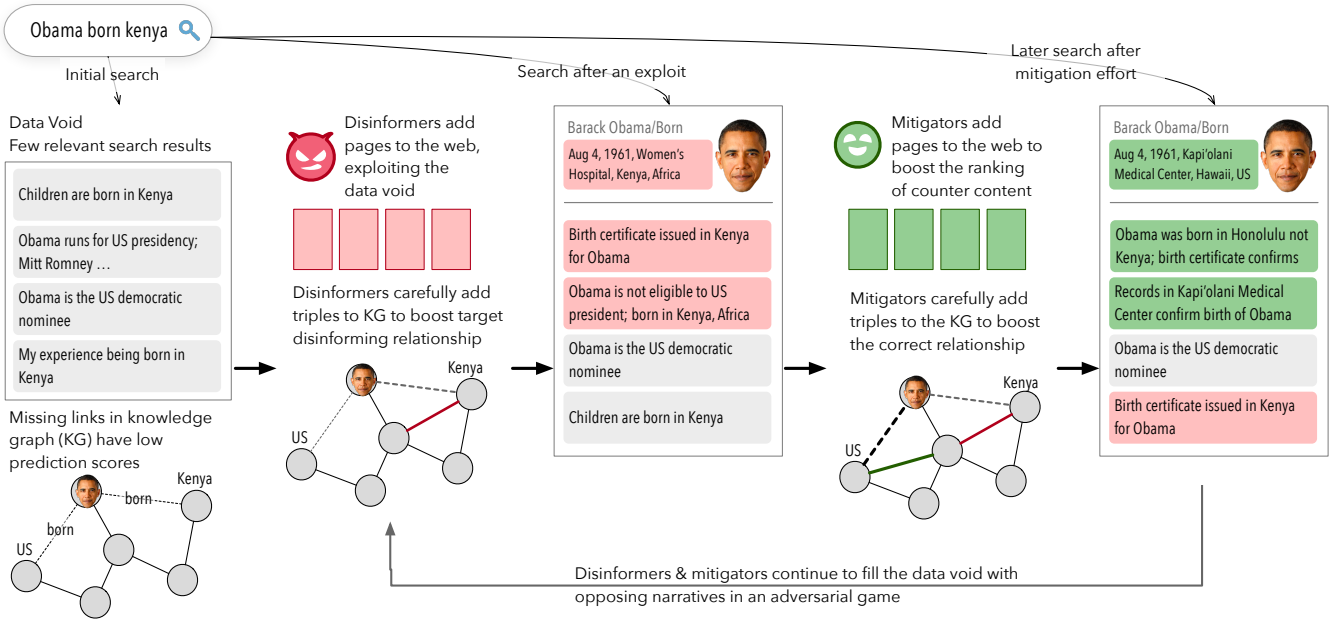

Figure 1: Data voids and how they evolve as disinformers (red) and mitigators (green) act to fill the void with content.

tors as adversarial agents with limited control over the strategies in simulated environments and hence inform mitigators how to best tackle data voids (§3). Third, we empirically evaluate the effectiveness of different mitigation strategies across web search (§4). 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. Code and data are available at https://github.com/huda-lab/datavoids.

2 Case Study Analysis

For historical analysis of data voids, we consider an approximation of such data through the use of custom data range searches on a major search engine. 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 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 implemented a pipeline that extracts search rank data. 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 can be labeled as *disinformation*, *mitigation*, or *irrelevant*. 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

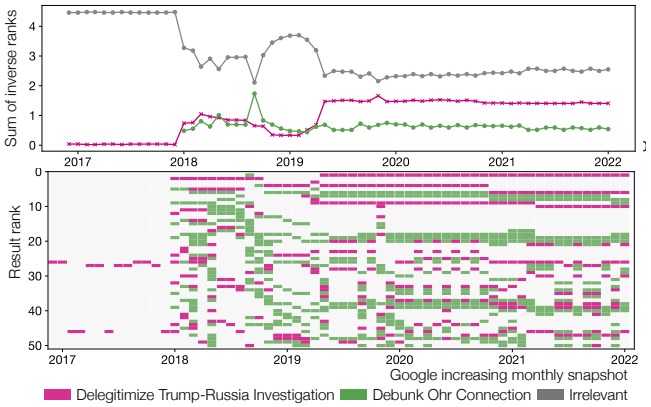

Figure 2: Tracking the effectiveness of disinformers and mitigators in filling the data void and influencing search ranking.

content from one side, the more prominent it is in the search results indicating that it is *"winning"*.

The Nellie Ohr Data Void. The Trump-Russia collusion investigation began in 2016 to determine if they colluded to manipulate the 2016 US Elections. Steele, a British Intelligence officer, produced a 'dossier' with claims on Trump's connections to Russia. In mid-2017, a name, Nellie Ohr, emerged in a conspiracy theory connecting the dossier and the collusion allegations. Exploiting this data void started by seeding the internet with stories about Ohr's husband, who was involved in the collusion investigations. QAnon, a far-right group, was an early propagator, introducing unverified information about her on Reddit and Twitter. These 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, influential conservative platforms consistently echoed her name and the conspiracy theory, effectively gaming search algorithms.

The rank analysis in Figure 2 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 a peak in search trends, which is attributed to the involvement of main stream media (e.g. NYTimes) in debunking the conspiracy [EC *et al.*, 2020]. This provides some evidence of the robustness of our methods in tracking certain historical data voids despite the limitations of custom range searches.

3 Data Voids as an Adversarial Game

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. 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.

In the Web setting, agents compete in having their content ranked higher than the counter-part. An agent here can represent the actions of multiple decentralized agents with an overlapping agenda.

The Game-playing Scenario. Five 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 $U_t := U_{t-1} \cup \{x_d, x_m\}$, $D_t := D_{t-1} - x_d$, and $M_t := M_{t-1} - x_m$. 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.* $\mathcal{E} : U \rightarrow \mathbb{R}_d \times \mathbb{R}_m$ Each turn alters the state of the information ecosystem, either amplifying the disinformation or its mitigation. The effect of each move is 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.* $\mathcal{C} : X \rightarrow \mathbb{R}$ The cost in this game is assumed to be proportional to the *influence* of an information item within the information ecosystem. Metrics such as node centrality, pagerank, and degree can be used as a proxy to determine how well connected a page is within the Web and thus its influence to promote a certain narrative. *Winning.* $\mathcal{W} : (\mathbb{R}_D, \mathbb{R}_M) \rightarrow \{d, m\}$ 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, $\mathcal{W}(\mathcal{E}(U_t))$ will declare an agent winner if its effect (e.g. ranking) is higher than the other agent at turn t.

The Game Strategies. In this game, a strategy ($\mathcal{S} : X \times U \rightarrow x$) 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}). The strategy is the most important element and the one that is controlled by the players to win the game. We study the following three. *Random*: an agent chooses a random piece of content to add to their information pool each time. *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. This aggressive strategy often chooses more costly items to add first. *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.

Using this abstraction of data voids, we build simulators that model the actions of disinformers and mitigators in a *realistic* setting. In the simulated environment, the set of pages or claims available to an agent plausibly represent the influence that such a page has in a web setting. We use search over Wikipedia pages as a stand in for searching over the web.

4 Experimental Results

We are interested in which strategies are more impactful and 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. Figure 3 illustrates the effects of simulating different strategies across the data void scenarios in the Web setting. The Random strategy is the baseline to beat. We fix the strategy of the disinformer to one among Random, Greedy, and Multiobjective; this choice 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. 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 Figure 3 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*, i.e., its effect being greater than or equal to the disinformer strategy at turn t.

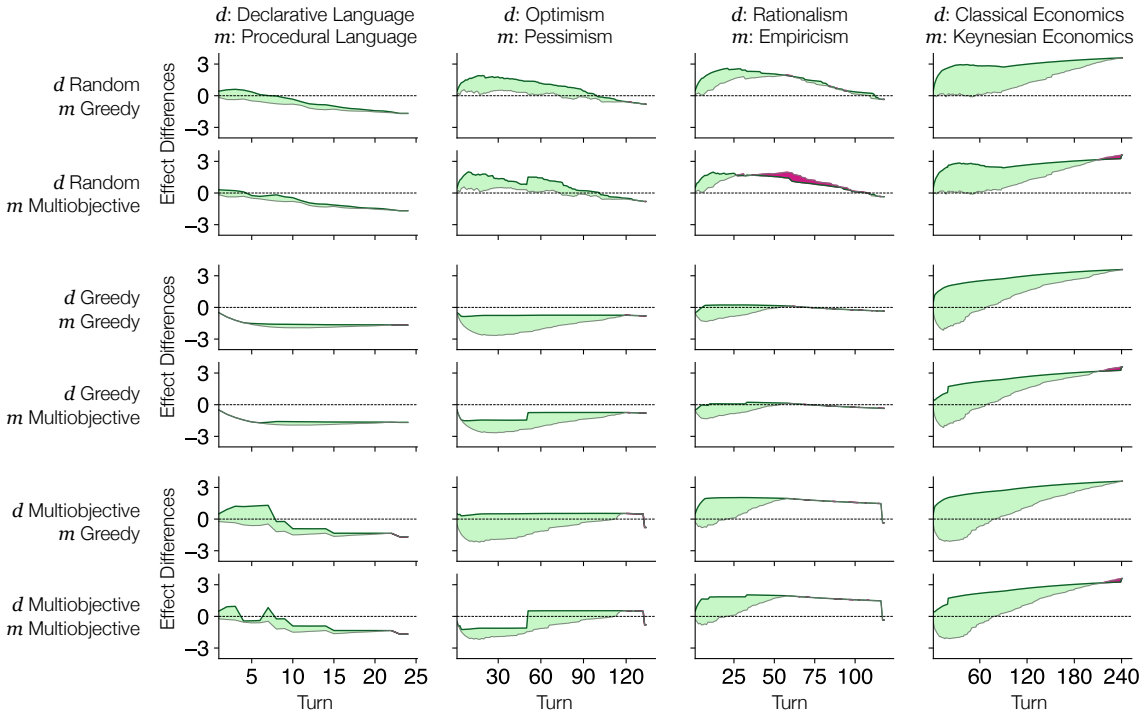

Figure 3: Differences of effects $\mathcal{E}_m(U_t) - \mathcal{E}_d(U_t)$ at every turn t of the Web search simulation across four data void scenarios. Greedy and Multiobjective allow mitigators to get ahead of an emerging data void scenario.

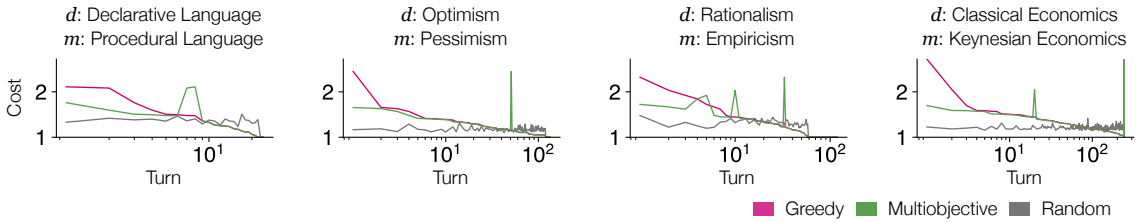

Figure 4: Mitigator costs for different strategies at each turn of the web search simulation across four data void scenarios. Multiobjective is less expensive than Greedy in the initial turns.

Figure 4 reports the cost of each strategy as more of the mitigator's pages 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:

- ▶ 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.
- ▶ Greedy strategies need access to "influential" pages or entities. In a greedy suppression, mitigators need to create consortia, hyperlink their resources, attract the sponsorship of high-value entities or nodes, to maximize the ranking of deployed mitigation content by search engines.

5 Conclusion

We develop rank-based measures to track the progress of disinformers or mitigators in filling up data voids in the Web. 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 includes extending the framework with more sophisticated strategies and better estimation of the effectiveness of the content available to agents [Bhardwaj *et al.*, 2021; Fang *et al.*, 2018; Cheng and Friedman, 2006; Boucher *et al.*, 2023; Leontiadis *et al.*, 2014; Flores-Saviaga *et al.*, 2022].

Acknowledgments

This work was supported by the ASPIRE Award for Research Excellence (AARE-2020) grant AARE20-307, NYUAD CITIES through Tamkeen - Research Institute Award CG001, and in part by the ANR project ATTENTION (ANR-21-CE23-0037).

References

[Aisch *et al.*, 2016] Gregor Aisch, Jon Huang, and Cecilia Kang. Dissecting the #pizzagate conspiracy theories. *The New York Times*, Dec 2016.

[Bhardwaj *et al.*, 2021] Peru Bhardwaj, John Kelleher, Luca Costabello, and Declan O'Sullivan. 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)*, pages 1875–1888, Online, August 2021. Association for Computational Linguistics.

[Boucher *et al.*, 2023] Nicholas Boucher, Luca Pajola, Ilia Shumailov, Ross Anderson, and Mauro Conti. Boosting big brother: Attacking search engines with encodings. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID '23, page 700–713, New York, NY, USA, 2023. Association for Computing Machinery.

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

[Cheng and Friedman, 2006] Alice Cheng and Eric J. Friedman. Manipulability of pagerank under sybil strategies. 2006.

[EC *et al.*, 2020] EC, Rafiq Copeland, Jenny Fan, and Tanay Jaeel. Filling the data void, 2020.

[Fang *et al.*, 2018] Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th annual computer security applications conference*, pages 381–392, 2018.

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

[Golebiewski and Boyd, 2019] Michael Golebiewski and Danah Boyd. Data voids: Where missing data can easily be exploited. Technical report, Data & Society Research Institute, 2019.

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

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

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

[Leontiadis *et al.*, 2014] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 930–941, New York, NY, USA, 2014. Association for Computing Machinery.

[Mannino *et al.*, 2024] Miro Mannino, Junior Garcia, Reem Hazim, Azza Abouzied, and Paolo Papotti. Data void exploits: Tracking & mitigation strategies. In Edoardo Serra and Francesca Spezzano, editors, *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1627–1637. ACM, 2024.

[Mirza *et al.*, 2023] Shujaat Mirza, Labeeba Begum, Liang Niu, Sarah Pardo, Azza Abouzied, Paolo Papotti, and Christina Popper. Tactics, threats and targets: Modeling disinformation and its mitigation. In Usenix, editor, *NDSS 2023, Network and Distributed System Security Symposium, 27 February-3 March 2023, San Diego, California, USA*, San Diego, 2023.

[Norocel and Lewandowski, 2023] Ov Cristian Norocel and Dirk Lewandowski. Google, data voids, and the dynamics of the politics of exclusion. *Big Data & Society*, 10(1):20539517221149099, 2023.

[Ortona *et al.*, 2018] Stefano Ortona, Venkata Vamsikrishna Meduri, and Paolo Papotti. Robust discovery of positive and negative rules in knowledge bases. In *34th IEEE International Conference on Data Engineering, ICDE*, pages 1168–1179. IEEE Computer Society, 2018.

[Page *et al.*, 1998] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

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

[Wang *et al.*, 2017] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

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