Title: Details on search engine manipulation and co-amplification appendix for "Search engine manipulation to spread pro-Kremlin propaganda"
Authors: Evan M. Williams (1), Kathleen M. Carley (1)
Date: February 16th, 2023
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

# Appendix: Details on search engine manipulation and co-amplification

*What is search engine manipulation?*

Search engines can rank websites using a variety of algorithms that are not made public, but by far the most well-known is PageRank, proposed by Google co-founder Larry Page (Page et al., 1999). Google's algorithm has become far more complex since PageRank was initially proposed, but the original PageRank ranking factors, backlink quantity, and quality, remain important. PageRank determines the ranks of websites by both the quantity and quality of their inbound links. Although many other attributes are weighted by search recommendation algorithms, link quality, and quantity remain important attributes. While it may be difficult to obtain a backlink from high-quality domains, link quantity is much easier to game.

Manipulating rankings through the creation of many new backlinks can involve paying third-party services to post a target website across third-party websites, creating websites to amplify a target website, hacking webpages and injecting invisible URLs, and many other maneuvers. Google calls these kinds of algorithmic manipulations "link schemes"ƒ and explicitly forbids them in its webmaster guidelines.[1] When link quantity is manipulated, a website's ranking can be boosted for various keyphrases. This can result in the website appearing more frequently in the first page of search results, thereby increasing site traffic. *Search engine optimization* (SEO) is the broad class of actions one might take to increase the visibility of a website on search engines. We focus on the "link scheme" subset of SEO forbidden by webmaster guidelines (commonly referred to as "black-hat" SEO), which we refer to as *search engine manipulation* (SEM). In this work, our main interest is the prevalence of link schemes created to boost pro-Kremlin domains. For more details, we refer the reader to (Tripodi, 2022).

We also explore *keyphrases*—search terms for which websites rank highly. Conspiracy theories have a distinct advantage in this space, as keyphrases can be highly specific and have very low competition (Golebiewski & Boyd, 2019). For example, if a user in a vaccine hesitancy group or forum sees a post falsely accusing Pfizer of a specific criminal action, a conspiracy website could reiterate the claim and rank highly for its keyphrases, as there's often little keyphrase competition for emergent conspiratorial stories.

*Co-amplification*

To measure the co-amplification of websites, we needed to quantify the shared backlink overlap between domains. We define *backlink co-amplification* as think tank websites that are linked at high volumes by the same domains. For an unweighted adjacency matrix $A \in \mathbb{R}^{m \times n}$ overlap $\mathcal{O}$ can be calculated using $\mathcal{O} = A^T \mathrm{A}$. However, when adjacency matrices are weighted, matrix multiplication can cause the metric to lose interpretability. Intuitively, if a domain was linked to by *i* 10 times and by *j* 100 times, our co-amplification score should not exceed 10. For a vector of weights, it does not make sense for overlap to exceed the total sum of inbound or outbound links present in either of the two domains. Formally, we want to constrain overlap for domain $u \in \mathbb{R}^d$ and domain $v \in \mathbb{R}^d$ so that $\mathcal{O}_{u,v} \leq \sum_{i=1}^{d} u_i \wedge \mathcal{O}_{u,v} \leq \sum_{i=1}^{d} v_i$. To satisfy this constraint, we define a co-amplification score as the overlap of domains $u$ and $v$ as the sum of the minimum pair-wise overlap: $\mathcal{O}_{u,v} = \sum_{i=1}^{d} min(u_i, v_i)$. The diagonal of the matrix was zeroed out to remove self-links. Intuitively, if two domains are heavily linked to by the same websites, this score will be high. If not, the score will be low.

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