Authoritative Sources in a Hyperlinked Environment
Jon M. Kleinberg

This paper highlights the challenge of finding useful sources of information when there are lots of hyperlinks, calling these authoritative sources, emphasizing the problem that some queries in a search return very few pages, but others return way too many, showcasing the difficulty in identifying the most relevant ones. The author introduces the concept of hubs and authorities as an approach to solve this motivating challenge. The author's Hyperlink-Induced Topic Search (HITS) algorithm defines highly referenced pages that contain authoritative information as authorities and hubs as pages that link to multiple authorities under the premise that a good hub should link to many good authorities and a good authority is linked to by many good hubs. The author provides a math formulation of their algorithm, representing the web as a directed graph and assigns each page an authority score and a hub score, based on in-links and out-links respectively. This formulation assumes the web contains well-defined hubs and authorities, but is limited by the fact that in practice, this distinction may not always be clear. Additionally, the paper assumes that the top authorities retrieved for a query are always useful, but there could be the case that a page with many in-links is still irrelevant. The paper also assumes that iterative mutual reinforcement converges reliably, but there could be a scenario where the algorithm doesn't converge. The author recognizes the impracticality of computing authority scores across the entire web, so they design an approach to a focused subgraph, where there is an initial set of root pages, expanded upon by including pages pointed to by the root pages and pages that point to the root pages. The conclusion is that this is an effective algorithm for returning more relevant information to user queries if applied across the web, or in various subgraphs, however, there is not much commentary passed on its practicality. One of the limitations of the subgraph approach is the scenario that multiple subgraphs are disconnected, making HITS struggle to rank pages relative to each other across subgraphs.

The PageRank Citation Ranking
Brin, Page

This paper introduces another approach to ranking web pages based on their relative importance and serves as the backbone behind Google Search. PageRank models the importance of a web page as a recursive function of the importance of pages linking to it, meaning it's based on the assumption that a link to a page is an endorsement of its importance. The authors offer an alternative interpretation for the reader's intuition, approaching PageRank as a random walk where you start on any page, then with some probability you follow a link off of that page and otherwise you jump to a random page. This addresses the edge case that pages may have links to them but no links from them. One of the key differences between PageRank and HITS is that PageRank is independent of query. The drawback of this is that PageRank is not dynamic, meaning as pages get created or revised, the algorithm must be rerun across a subset of the entire web. PageRank also doesn't classify links as hubs nor authorities, instead, it computes one single importance score. This makes PageRank more effective at handling spam as HITS can lead to some failures in the event that spam hubs are created. One extension to PageRank could be some weight hyperparameter so that not all links are treated the same, mitigating some of the risks involved with spam. This could give PageRank some of the benefits of the flexibility that HITS has without adding too much to the

computational cost. The algorithm is very effective and the theory has it converging quickly which is extremely important in this context of scaling to billions of pages across the entire web. However, the assumptions do not always hold, for example, sometimes a link to another web page is just for navigation, like links in the footer, not necessarily positive endorsements. This could cause spam sites to gain high scores in PageRank. Also, PageRank, unlike HITS, is based on link structure only, making it fail to capture relevance across a specific theme, or subgraph like HITS manages to do.