# Notes on Decentralized Search

Sarah Jamie Lewis *†

March 6, 2024

## 1 The High Level Problem

To boil it down to its core, there exists a massive collection of documents, distributed over millions of nodes, through which a separate collection of billions of people wish to conduct a guided search.

A significant portion of these nodes cannot be trusted to deliver suitable content, as such the rest of this note will assume we are working within a malicious risk model.

The current front runner solution for this problem is a collection of centralized services that actively crawl some subset of nodes, index the content, and permit searches over that index - usually in exchange for both search history data and the opportunity to substitute paid advertisements for search results.

This state of affairs has resulted in a large scale privacy and incentives problem for search, and an interest in alternatives that bring about both better privacy and better search results.

## 2 A (Very) Brief History

Research in decentralizing search and search engines has a long history, with considerations for peer to peer solutions[2] being proposed as early as 1997.

Much of this early work also intersects the development of distributed lookup systems like Chord[20] and the general rise of peer to peer systems such as Gnutella[6]

Purely cooperative search algorithms that do not consider malicious participants such as those proposed by Unger[22] dating back to at least 2003.

As an evolution, systems like Maay[13], and YaCy[7] rely on trust peer-to-peer nodes with various heuristics to improve the reliability and safety of the search, but without a technical solution to the problem of sybil attacks in general, and malicious indexes specifically.

More recently, proposals such as Desearch[11] use trusted execution environments and epochs to generate verifiable indexes. Espresso [17] suggests an approach where services generate indexes themselves which are then collected and distributed via a federated network.

Finally, (large) language models have been put forward as an alternative to search engines entirely[19] with promising, if limited, efficacy.

## 3 Is Google Getting Worse?

Since its introduction, PageRank[14] has been heralded as the pinnacle of search rankings. Pagerank is based on a theory that sites with more inbound links have greater authority than those that have fewer links. The integration of pagerank into search engines was quickly met with a rise in "search engine optimization" techniques designed to increase the number of back-links to a site, and was in turn met with various proposals to combat the resulting tide of spam[25][1].

Many now consider that pagerank is unsuitable as a ranking technology for the modern web[15], with most attributing the rise of advertisement-centric search and the factor of competitiveness between sites subverting the underlying theory.

This belief is finding support in newer longitudinal studies into the quality of search results[3] finding the search engines studies have significant problems with highly optimized (affiliate) content, and highlighting that "web search is a dynamic game with many players, some with bad intentions"



*Open Privacy Research Society (sarah@openprivacy.ca)
†Blodeuwedd Labs (sarah@blodeuweddlabs.com)

# 4 New Directions

In this section I would like to discuss certain interesting technological developments that have more recently come into focus and how they may be applied to the search problem.

## 4.1 Efficient Tokenization

Sub-word tokenization techniques like BytePair Encoding[18] and more recently SentencePiece[10] have received increased attention in recent years due to the investment in transformer architectures[23] and related fields that gain significant performance advantages when the input context size can be losslessly compressed and allows documents in various language and localizations to be subject to identical processing.

Importantly, unlike other compression techniques that distribute and/or aggregate information in lookup tables, sub-word tokenization does not change the semantic properties of the underlying context (and so models can be built over the compressed representation saving both memory and processing time).

In addition, part of what makes search at scale difficult is the long-term storage (e.g. for generating summaries, infoboxes, page caches etc.) and analysis (e.g. finding similar entries) of source documents. Representing such documents in their tokenized form, while not as space efficient, does allow more immediate access to the data for processing.

## 4.2 Semantic Embeddings

Semantic analysis also has a long history, though it is worth bringing specific attention to Word2Vec[12] and the introduction of word embeddings and the equivalent rise of language models[1] such as BERT[5]. GPT3[4] and later Llama[21] and derivatives e.g. Mistral [9].

It is clear that these models are capable of some amount of sophisticated semantic analysis of text, and as an offshoot of the generalized chat and instruction interfaces that are most prevalent, we have also seen research into adapting these semantic embeddings for document search[8].

It is also clear that these models are as, if not more, prone to malicious and crafted input as both human curators and previous categorization schemes with both prompted and unprompted hallucinations being a particular cause of concern[24], and there is currently no reason to believe that those issues do not also apply to embeddings derived from such models.

Further, depending on the exact method used to construct or extract embeddings from the underlying model, the resulting embeddings are typically large relative to the input document, and are scale dependent i.e. the embedding for a sentence will be much more concentrated that the embedding for a paragraph, and as such multiple embeddings are required per-document to represent the true semantics of the underlying content.

Nevertheless, we have never had such ability to at-scale categorize and semantically analyze documents before, and the applications of such a capability to search are obviously compelling.

Solutions for capturing and representing documents as vector embeddings are already gaining traction in communities enthusiastic about local language models and it is likely they will continue to improve.

## 4.3 Zero Knowledge Proofs

Directly applicable to the idea of federated or distributed search is the challenge of verifying that such a search really took place and that the indexing was performed honestly. Generating proofs for such activities still mostly resides in the realm of fiction[2] unless great liberties are taken when defining the system - or the actions within the system are highly restricted to a given domain (e.g. honestly modifying the state of a global ledger)

However, that does not mean that all is lost when considering the application of ZKPs to decentralized search. Voting schemes are of particular interest as they can be utilized to anonymously rank the quality of indexes, and the underlying vote can capture that such a ranking was performed to a given specification (with proofs being efficiently checkable).

[1] not to mention more popular proprietary models

[2] See Also: Attacks on Trusted Execution Environments

### 4.4 Federated Networks

While I am still inclined to hold the position that federation is the worst of all worlds[3] the rise of Mastodon and other ActivityPub[?] based systems has demonstrated a significant interest in such software.

Additionally this rise has also triggered a new generation of software that speaks ActivityPub, going beyond RSS feeds to establish a richer semantic context for actions as diverse as publishing a video, or opening a pull request.

It is worth seriously noting that not everyone in these spaces is open to the idea of targeted search of these activities and there is a significant `#NOBOT` movement from the joined perspectives of both privacy and consent[4] - in turn many frameworks have aligned themselves with an opt-in structure for content distribution.

For those who do opt-in however (and for systems where individuals are less focused e.g. Lemmy), there is clearly an opportunity for far stronger search capabilities than currently exist- with the openness of these networks leading to far stronger curation of the semantic and authoritative nature of shared documents than was traditionally accessible via closed platforms.

## 5 Challenges

Regardless of how a decentralized search engine is constructed there are a set of universal risks to be overcome. This section briefly outlines the major areas of concern that any new solution should address.

### 5.1 Index Integrity / Censorship

A major problem with existing centralized providers is censorship. Different sites will be displayed to different users in different countries. Certain websites will not be made accessible at all. While some search engines have taken steps to document instances where they have been forced to remove sites from their index due to jurisdictional pressure or local legislation, search engines have also been suspected to downrank or hide content that they dislike for any reason[5].

These concerns also directly impact decentralized engines. Indexing nodes can choose fully to not index a given set of sites or to not serve those indexes to individuals in favour of sites they otherwise wish to promote.

In that light it is clear that indexing and ranking must be isolated activities with different trust boundaries and verification steps.

### 5.2 Privacy

Highly correlated to index integrity and censorship is the problem of privacy. In centralized search engines searchers have no realistic expectation of privacy. Search queries are stored, mined, and distributed to advertisers, local authorities and other parties without the searcher having any input.

Because privacy in these systems is non-existent then the ability to perform censorship or otherwise compromise the integrity of the index is much easier.

Even in decentralized systems the expectation of privacy remains low, only in (mostly academic, proposed, hypothetical systems) that attempt some variant of privacy-preserving search are resistant to well documented attacks.

As such any decentralized system likely requires that the actual searching happens locally (or at least federally[6]), requiring that indexes be small enough that they can be distributed efficiently and perhaps adapting techniques from PIR to minimize privacy loss even in a federated solution.

### 5.3 Scale

The internet is large and the number of documents to index is even larger. Not counting of course the ever growing number of audiovisual content[7].

It is likely that some of the techniques above can help us here, but it is also perhaps worth turning to a more bottom-up approach for content indexing. How useful a document is, is subjective. Community curation is an oft-brought up solution to

[3] `https://pseudorandom.resistant.tech/federation-is-the-worst-of-all-worlds.html`

[4] Privacy is Consent...`https://leanpub.com/queerprivacy/overview`

[5] Because the internals of search engines are opaque the motivations for such activities remain a mystery

[6] referring, of course to trusted federated servers not jurisdictional boundaries

[7] Although it is now possible to obtain transcripts from such content efficiently thanks to models like Whisper[16]

the deluge of content and it is clearly an attractive choice.

It is also at this point perhaps worth mentioning that any solution doesn't have to be universal. A search engine for papers relevant in a specific domain has utility by itself, as does a search for book that are similar to another.

Any decentralized solution likely requires that it can be built up from disparate initiatives, perhaps only providing a common protocol to capture and disseminate results.

# References


[1] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, 2008.

[2] Jean Marc Andreoli, U Borghoff, Remo Pareschi, Stefano Bistarelli, Ugo Montanari, and Francesca Rossi. Constraints and agents for a decentralized network infrastructure. In *Constraints and Agents: Papers from the 1997 AAAI Workshop*, pages 39–44. Citeseer, 1997.

[3] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. Is google getting worse? a longitudinal investigation of seo spam in search engines. In *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024)(Lecture Notes in Computer Science). Springer*, 2024.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Justin Frankel and Tom Pepper. Gnutella. *Web Site-www. gnutella. com*, 1999.

[7] Michael Herrmann, Kai-Ching Ning, Claudia Diaz, and Bart Preneel. Description of the yacy distributed web search engine. *Technicla report. KU Leuven ESAT/COSIC, iMinds*, 2014.

[8] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561, 2020.

[9] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[10] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[11] Mingyu Li, Jinhao Zhu, Tianxu Zhang, Cheng Tan, Yubin Xia, Sebastian Angel, and Haibo Chen. Bringing decentralized search to decentralized services. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pages 331–347, 2021.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[13] Frédéric Dang Ngoc, Joaquín Keller, and Gwendal Simon. Maay: a decentralized personalized search system. In *International Symposium on Applications and the Internet (SAINT'06)*, pages 8–pp. IEEE, 2006.

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bring order to the web. Technical

report, Technical report, stanford University, 1998.

[15] Vladimir Prelovac. The age of pagerank is over [manifesto].

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[17] Mohamed Ragab, Yury Savateev, Reza Moosaei, Thanassis Tiropanis, Alexandra Poulovassilis, Adriane Chapman, and George Roussos. Espresso: A framework for empowering search on decentralized web. In *International Conference on Web Information Systems Engineering*, pages 360–375. Springer, 2023.

[18] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[19] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

[20] Ion Stoica, Robert Morris, David Karger, M Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. *ACM SIGCOMM computer communication review*, 31(4):149–160, 2001.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] Herwig Unger and Markus Wulff. Towards a decentralized search engine for p2p-network communities. In *Eleventh Euromicro Conference on Parallel, Distributed and Network-Based Processing, 2003. Proceedings.*, pages 492–499. IEEE, 2003.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] Simon Willison. Prompt injection: What's the worst that can happen?

[25] Baoning Wu and Brian D Davison. Identifying link farm spam pages. In *Special interest tracks and posters of the 14th International Conference on World Wide Web*, pages 820–829, 2005.