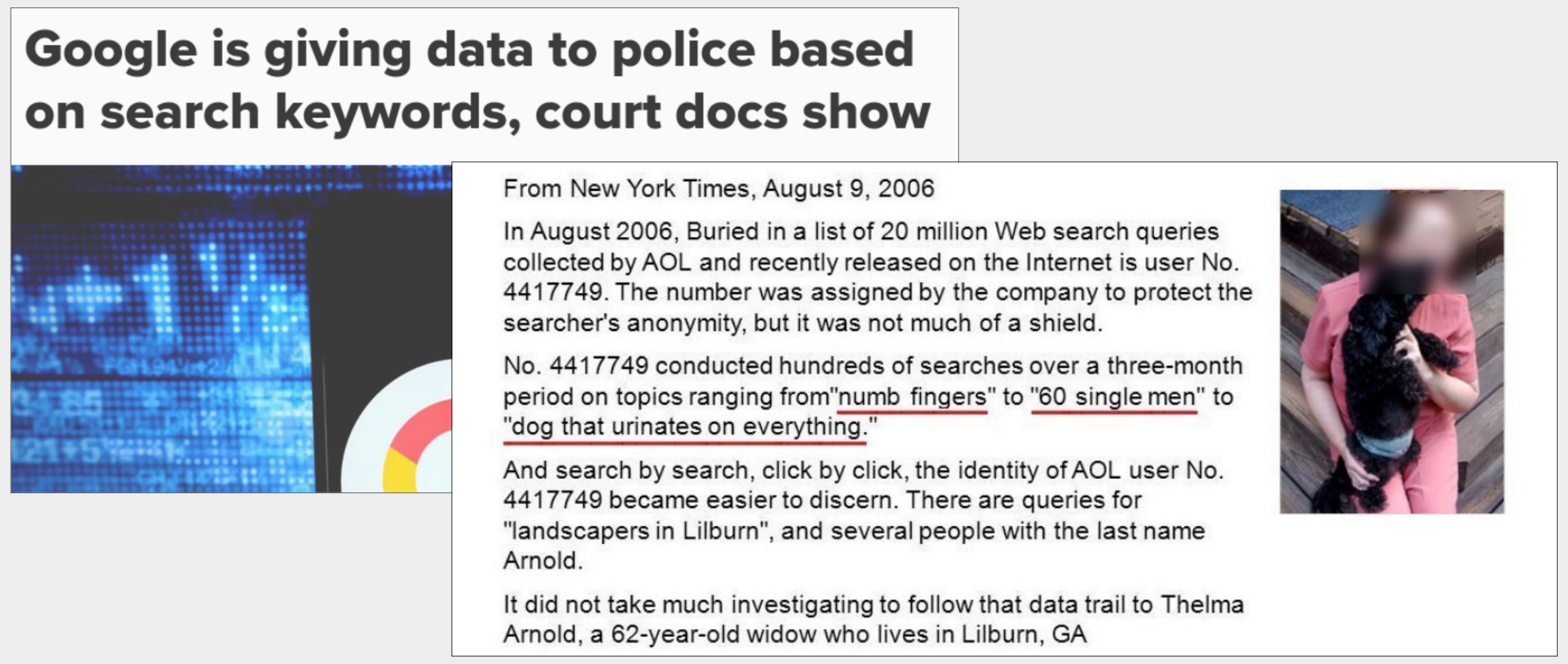
Maik Fröbe   Nicola Lea Libera   Matthias Hagen
maik.froebe@informatik.uni-halle.de

City of Disguise: A Query Obfuscation Game on the ClueWeb

Search Engines—A Thread to Your Privacy?



Google is giving data to police based on search keywords, court docs show

From New York Times, August 9, 2006

In August 2006, Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from"numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn", and several people with the last name Arnold.

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, GA

Automatic Query Obfuscation

There are two approaches for automatic query obfuscation.

Semantic Query Obfuscation
Replace terms in the sensitive query with terms representing more general concepts (E.g., with WordNet).

Statistical Query Obfuscation
Use a private private search engine to identify non-revealing queries that retrieve similar results to the sensitive query.

[Arampatzis et al., IRJ'15]

Can Humans Obfuscate Queries?

Overview of the Game

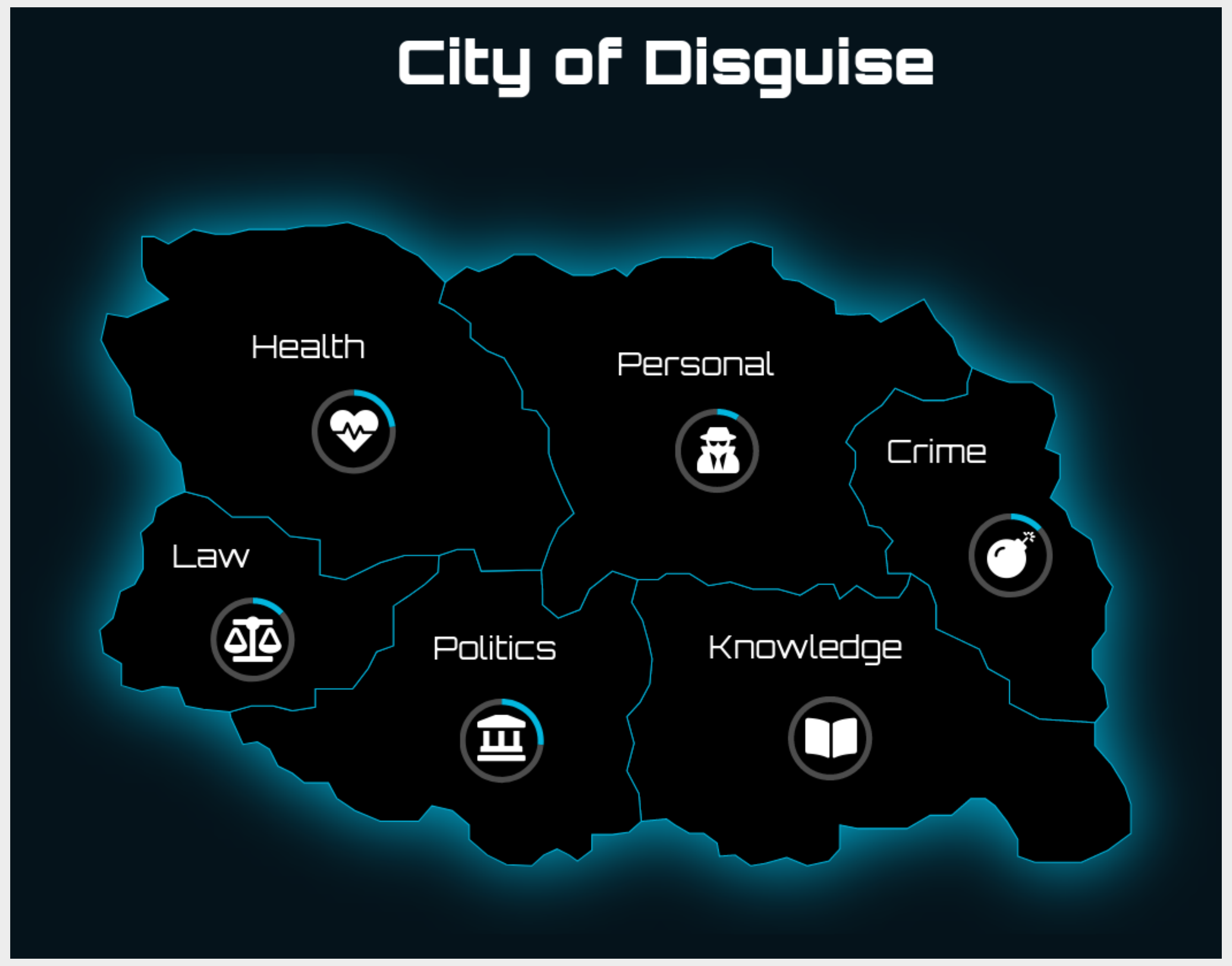
City of Disguise is a retrieval game that tests how well searchers can reformulate some sensitive query in a 'Taboo'-style setup but still retrieve relevant results.

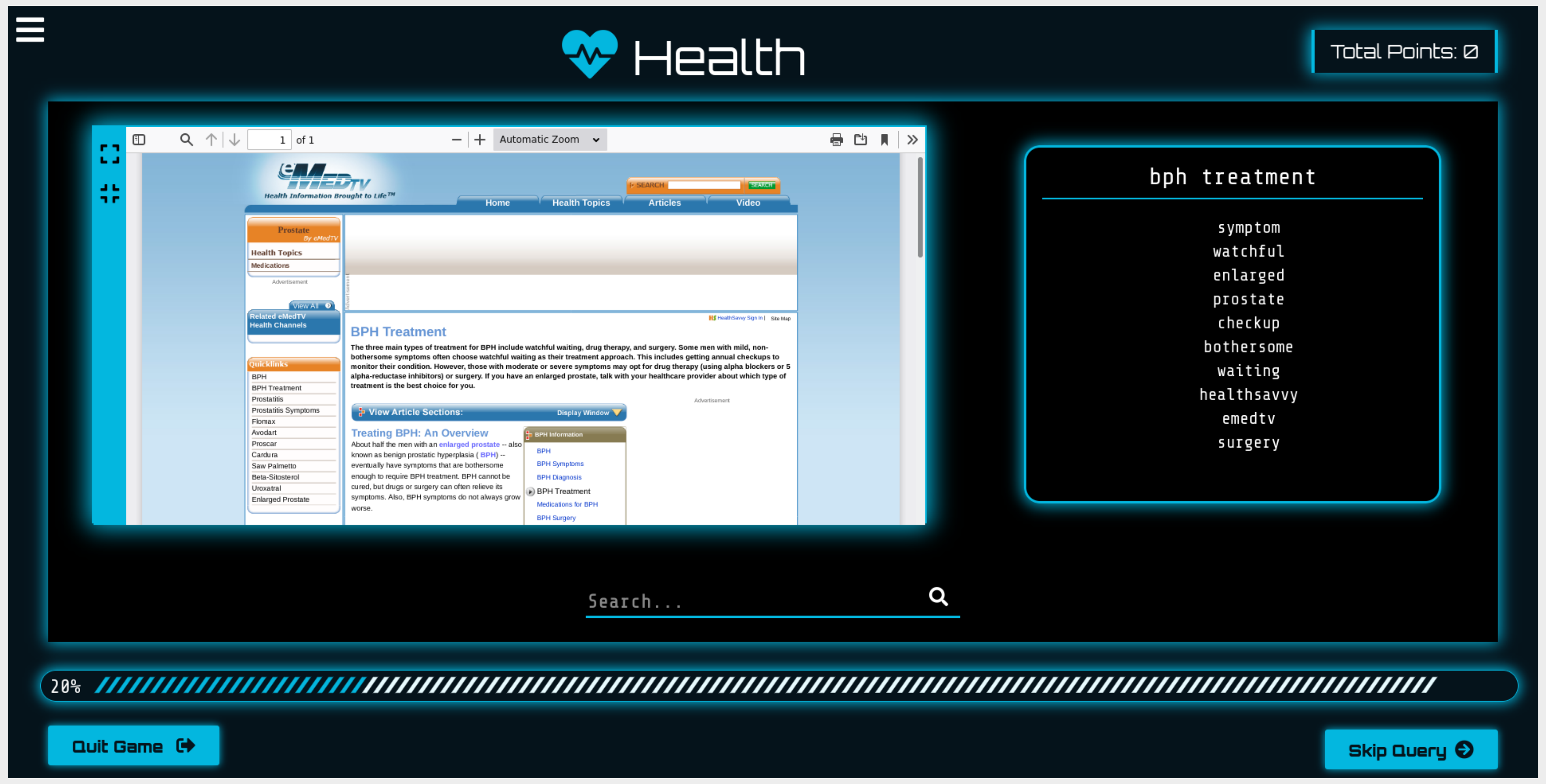
Details:

- Corpus of 0.6 million ClueWeb12 documents
- Documents rendered with the Wayback Machine
- 200 sensitive queries with one relevant document

Scoring Query Obfuscations:

- Position of the relevant document
- Length of the Query
- Recall (top-100 results of sensitive query)
- MAP (top-100 results of sensitive query)





BPH Treatment

The three main types of treatment for BPH include watchful waiting, drug therapy, and surgery. Some men with mild, non-bothersome symptoms often choose watchful waiting as their treatment approach. This includes getting annual checkups to monitor their condition. However, those with moderate or severe symptoms may opt for drug therapy (using alpha blockers or 5 alpha-reductase inhibitors) or surgery. If you have an enlarged prostate, talk with your healthcare provider about which type of treatment is the best choice for you.

Treating BPH: An Overview

About half the men with an enlarged prostate -- also known as benign prostatic hyperplasia (BPH) -- eventually have symptoms that are bothersome enough to require BPH treatment. BPH cannot be cured, but drugs or surgery can often relieve its symptoms. Also, BPH symptoms do not always grow worse.

Evaluation

Pilot Study

- 72 participants from 2 IR courses
- 1,462 obfuscated queries
- 43 seconds on average to formulate an obfuscated query

Main Result

The obfuscation effectiveness decreases the more the players deviate from the game's term suggestions.

Effectiveness of obfuscated queries in ChatNoir and the games' document sample ('Sample') as MRR, the number of documents retrieved for the original query ('Ori.'), and the number of retrieved relevant documents ('Rel.'). We compare automatically obfuscated queries and four types of queries submitted by players.

					Our Sensitive Queries				**Sensitive Web Track Queries**			
		Obfuscated Queries			ChatNoir		Sample		ChatNoir		Sample	
		Count	Length	Time	MRR	Ori.	MRR	Ori.	MRR	Rel.	MRR	Rel.
Players	Only Suggestions	130 / 21	2.42	40.50 s	0.093	5.223	0.325	67.592	0.010	3.094	0.152	3.691
	Some Suggestions	556 / 125	4.53	42.45 s	0.046	4.667	0.258	85.829	0.013	3.632	0.038	3.568
	No Suggestions	576 / 157	2.88	44.39 s	0.029	2.935	0.082	38.932	0.015	1.783	0.024	3.316
	New Word	559 / 158	3.57	46.27 s	0.002	1.517	0.051	49.992	0.002	1.235	0.005	2.790
	Automatic	1025 / 327	2.91	—	0.088	9.229	0.420	84.264	0.014	2.872	0.042	3.743

Can You Beat the Highscore?

demo.webis.de/city-of-disguise/

github.com/webis-de/ECIR-22

Big Data Analytics   www.webis.de   **Martin-Luther-Universität Halle-Wittenberg**   **Bauhaus-Universität Weimar**