



The Right to Be Forgotten: A Data-Driven Study

Jason Xue, ECNU and NYUSH


Gabriel Magno, UFMG

Evandro Cunha, UFMG

Virgilio Almeida, UFMG

Keith W. Ross, NYU and NYUSH

The RTBF Law: Motivation

- Suppose when people google your name, the first link points to an article:
 - about you going bankrupt 20 years ago.
 - or a crime for which you were acquitted.
 - or a minor crime committed as a child.
 - or personal private information, such as your home address or sexual orientation.
- What can you do besides cry? 



**Right to
be
Forgotten**

LAW

RTBF: The Process

- Began in May 2014
- Submit a web-based form:
 - Your name, citizenship, URL you want removed, justification
- Committee at Google approves if:
 - Private or sensitive information.
 - Content that relates to minors.
 - Minor crimes occurring when requester was a minor.
 - Acquittals, exonerations, and spent convictions.
- Google may decline to delist if it determines the page contains information which is in the public interest.

RTBF: The Implementation

- URLs are only delisted in response to queries relating to an individual's name.
 - If Google grants a request from Joe Doe to delist an article about a trip to Shanghai, Google would not show the URL for queries including [john doe].
 - But would show URL for query like [trip to shanghai].
- RTBF does not affect the original published content.
- Links remain listed on google.com

RTBF: Notification

- Google notifies webmasters when pages from the webmasters' sites are delisted.
- “In order to respect the privacy of the individuals who have made removal requests, Google only sends the affected URLs, not the requester's name.”

RTBF: Requests and Removals

- Over 1.5 million link-removal requests from individuals in the European Union
- Approximately 43% of the requested URLs have been removed.



RTBF Debate: Freedom of Speech

- RTBF allows individuals to become directly involved in the privacy process.
- Can minimize the damage associated with inaccurate or outdated public information.
- But some argue that the RTBF restricts the right to freedom of speech.

RTBF Debate: International Law

- RTBF demonstrates limits of national privacy laws in a world of transnational data flows.
- French data protection authority CNIL has ordered Google to delist links from all of its geographic properties including google.com.
- Google has so far refused, and the dispute is likely to end up in European courts.

Data-Driven Study

- This paper is the first data-driven study of the RTBF.
- We hope the results in this paper can inform the debate and future evolutions of RTBF laws.

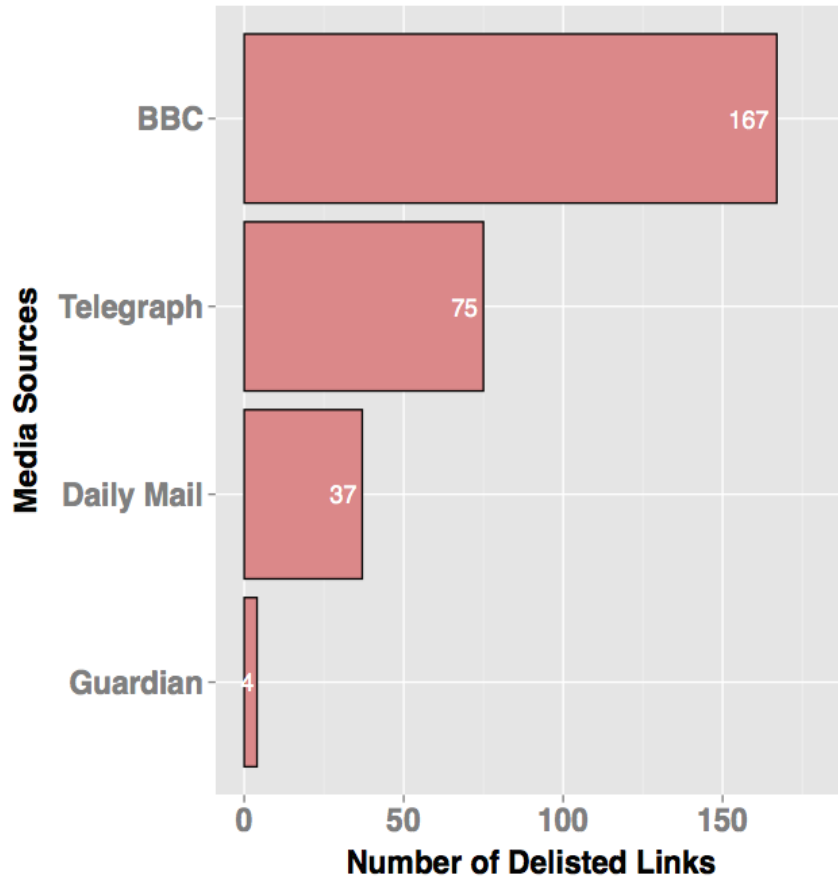
Our Main Result: Data-Driven Attack

- With moderate resources and hacking skills, a transparency activist can determine delisted URLs and the names of the people who requested the de-listings.
- The activist can then republish on URLs and names on his own well-known website.
- Can lead to the Streisand effect, putting the efficacy of the RTBF law into question.

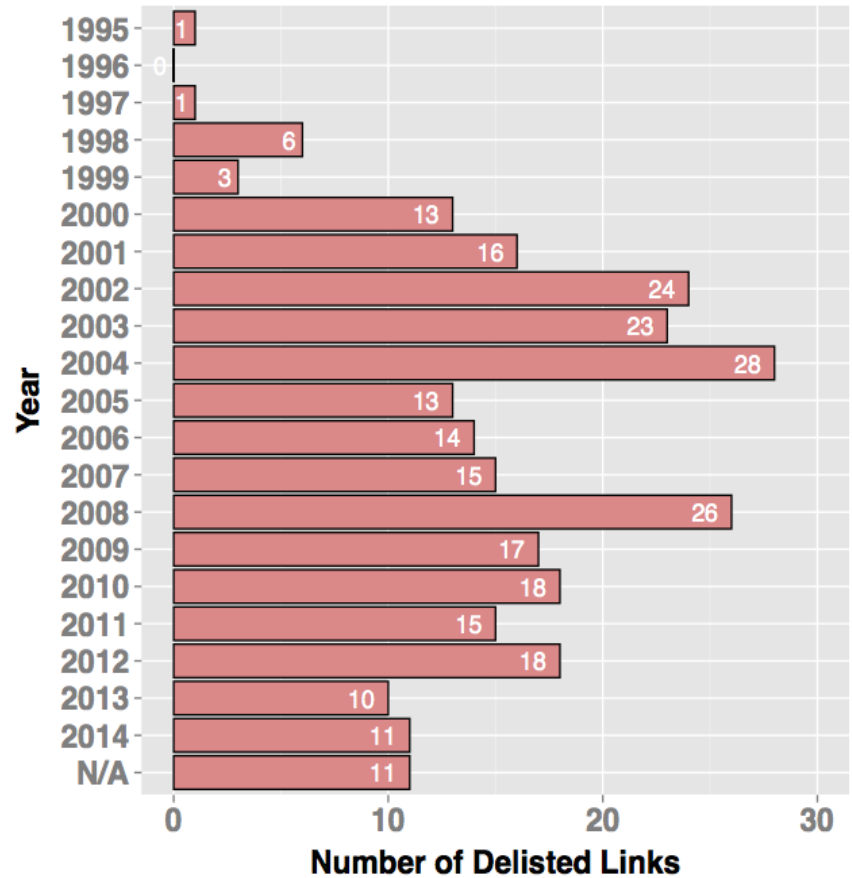
Additional Results

- Content analysis on 283 known delisted UK articles.
- Determine topic themes:
 - Manual investigation
 - Latent Dirichlet Allocation (LDA)
- Determine 80 requesters for 103 articles.
 - Demographic analysis
- Develop methodology to quantify the Streisand effect

The Dataset (1)

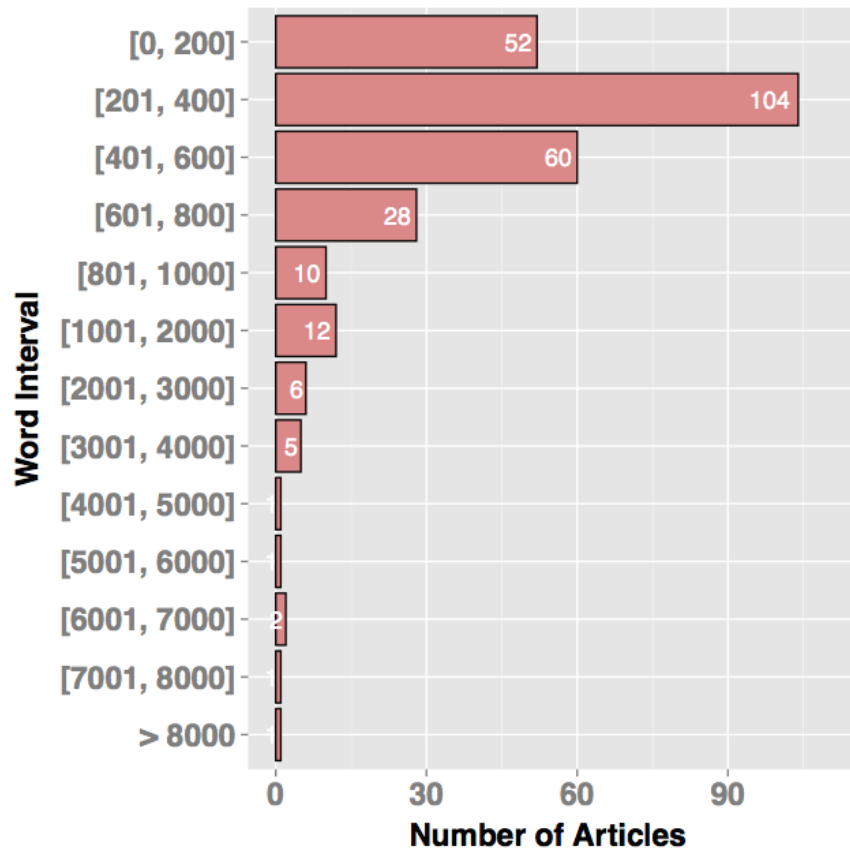


(a) Distribution of delisted links across media sources

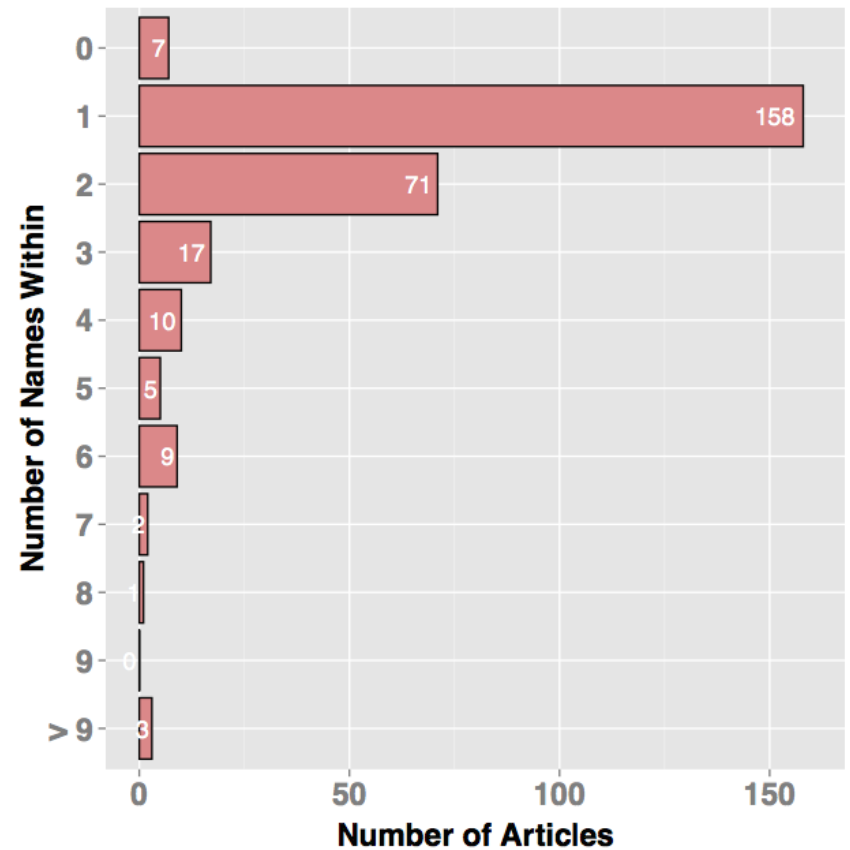


(b) Distribution of delisted links by year

The Data Set (2)

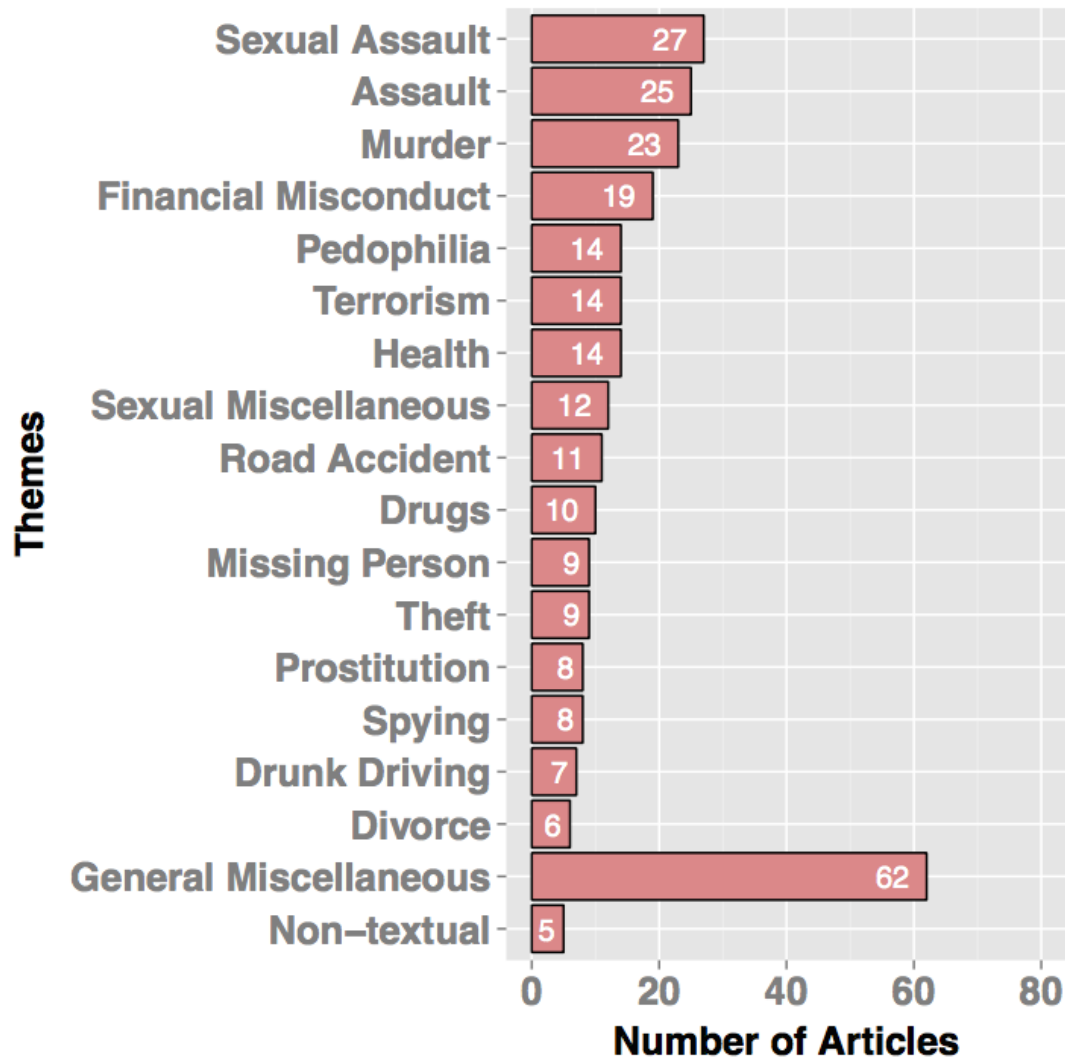


(a) Distribution of articles by number of words



(b) Distribution of articles by number of names within

Manual Topic Analysis



LDA Topic Analysis

Topic ID (Rank)	Cumulative Weight	Top Terms	Most Relevant Articles (Document ID)	LDA Theme
1	12.62	indecent, image, suspect, murder, inspector	4, 6, 89, 92, 93, 142 195, 226, 255, 268, 271	Violent Crime
2	11.95	drive, smash, car, cash, drug	25, 45, 49, 52, 133 149, 184, 220, 221	Drunk/Drugged Driving
3	11.78	debate, referee, heroin, wife, inject	1, 98, 103, 128, 213 233, 234, 245, 248, 265	Domestic Drug Use
4	11.00	jail, murder, corruption, convict, prosecution	96, 101, 113, 114, 123 136, 183, 196, 238, 250	Murder
5	11.00	passenger, escort, seat, benefit, solicitor	24, 63, 71, 164, 172 180, 190, 209, 239, 261	Prostitution
6	10.88	wife, bank, transfer, money, reassure	42, 70, 94, 100 151, 167, 207, 247	Financial Misconduct
7	10.58	woman, bed, charge, explosive, rape	13, 41, 46, 65, 83 87, 91, 223, 256	Sexual Assault
8	10.55	sex, rape, girl, deny, flat	57, 120, 159 162, 212, 270	Sexual Assault

Table 1. LDA Topics for Delisted Content

Attack: Identify the Requester for Known Delisted Link

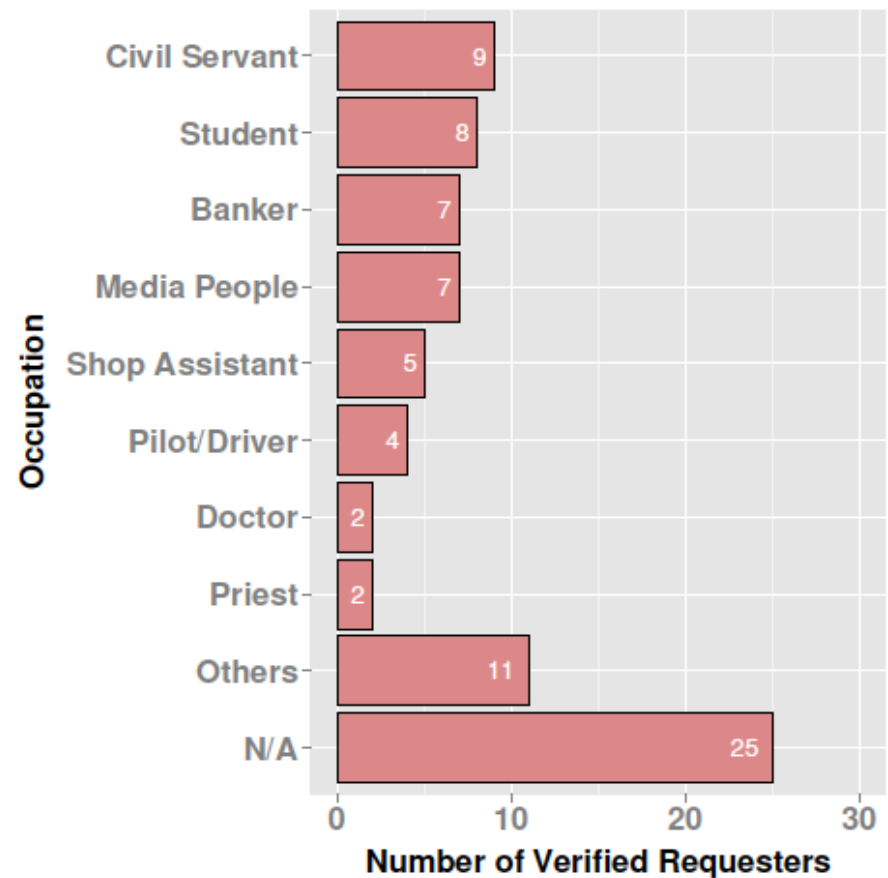
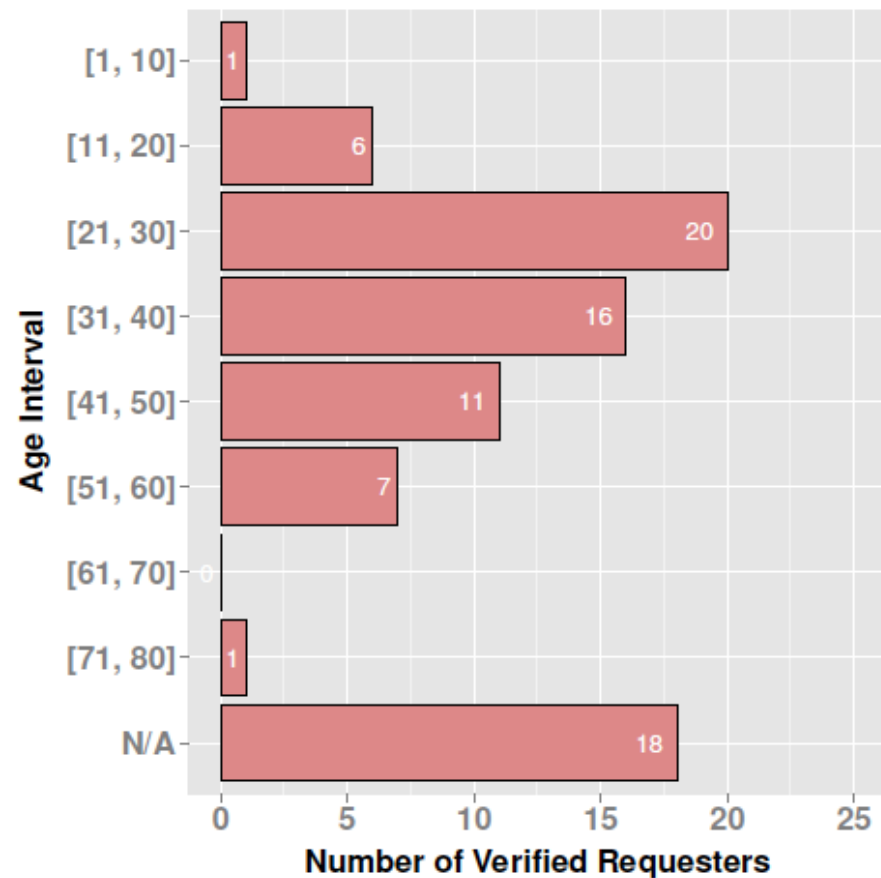
1. Select delisted article
2. Determine all names in article.
3. For each name, query google.uk with {"name" "article title"}
4. If article URL is not listed, then the name is the requester.

Result:

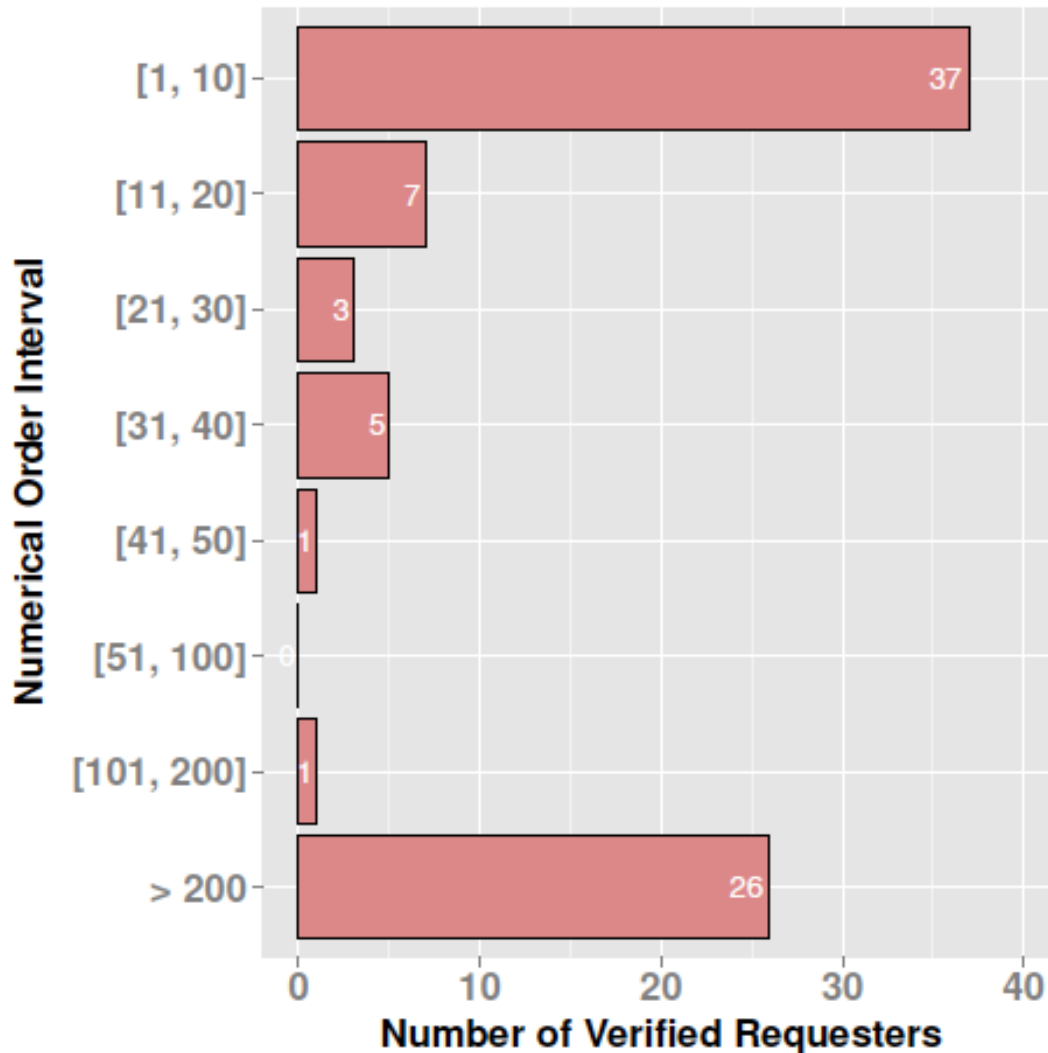
- For 283 articles, determined 80 names for 103 articles.
- Remaining articles have maybe been re-instated.

Demographic Analysis of Requesters

- 87.5% (of 80) of the requesters are male. Only 73% of names are male in news articles.



Extend RTBF to google.com?



But extending to google.com would have no bearing on attacks. RTBF could still be broken.

Attack: Determine Previously Unknown Delisted Articles

1. Crawl and download all the digital articles of a media source such as El Mundo.
2. Put all the articles in a database and search on terms related to topics typically requested for removal.
3. From those articles, automatically collect all the names mentioned.
4. Write a script that puts each name and article title into google.es and checks the links returned.
5. If for any of these articles, if the article URL is not listed, then the URL is delisted and name is the requester

Computational Issues

- The querying effort depends on the number of candidate articles C and names in articles.
- Search engines employ rate limiters, preventing bots from querying from the same location too frequently.
- Attacker could lease a black-market botnet to send queries from many different random locations.
- We believe large-scale attack covering most of the major European newspapers is feasible.

Proof of Concept: El Mundo

- Query El Mundo with 37 Spanish terms related to crime and download 85K articles.
- Extract names from articles.
- To reduce query effort, use heuristic filters to reduce title+name combinations to 6,410 queries. (4,164 articles)
- Sequentially query google.es from a single machine.
- Discover 2 previously unknown RTBF delisted links.
- Double check by querying google.es with title only and google.com with name only.

Discussion

- RTBF can be broken for links to news media.
 - No apparent defense.
 - Attack is still valid if RTBF is extended to google.com.
- But appears to be much more difficult for links to social media and profiling sites.
 - 95% of the URLs are for personal private information.
 - RTBF serves an importance purpose here and is working.
- Privacy laws can backfire: RTBF, COPPA
 - R. Dey, Y. Ding, K.W. Ross, *The High-School Profiling Attack: How Online Privacy Laws Can Actually Increase Minors Risk*, IMC , 2013



Thank You !!!