

Вебсаетес оф тхе Формер Совиет Унион & Еастерн Еуропе

Team компромат

Websites of the Former Soviet Union & Eastern Europe

Team компромат

Research Question:

What countries are linking to each other in the collection?

Methods

1. Reverse engineered the seed list with Seed ID in WARC filename + Archive-It API
2. Learned about WARCs and the AUT derivatives
3. Made derivatives with AUT, extracted all URLs from HTML
 - a. 93% parseable, 7 million URLs
4. Manual coding of Seed List, Top Level Domains and Crawled URLs (Google Sheets)
5. [Notebooks](#) to parse URLs with coded TLDs to edge list that we can't run
6. Visualised Data via Gephi, Google Maps, Charts
7. Visualisations helped inform further research questions
8. Took some Paracetamol ; ate Chocolate
9. Made our presentation



```
In [60]: fullurls = "/mnt/data/ivy/former-soviet-union/derivatives/all-domains/11360-fullurls.txt"

output_csv(get_counts('\.(\w+)$', fullurls), 'tlds.csv')
```

```
org,108119
hu,76011
ru,61108
pl,47130
com,26025
ua,25600
al,19623
ba,11268
rs,8297
mk,7930
me,4918
info,1800
net,723
ro,282
tv,135
us,13
fm,12
io,7
ca,6
de,6
uk,6
ru,6
```



URLS



File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive



Share



100% \$ % .0 .00 123 Arial 10 B I S A

fx http://prozhitto.org/

	A	B	C	D	E	F	G	H	I	J
1	URL	Country	Former Bloc		Eastern Bloc	Yugoslavia	USSR			
2	http://dostajebilo.rs/	Serbia	Yugoslavia			1				
3	http://www.aitrus.info/	Russia	USSR				1			
4	http://www.polk.ru/	Russia	USSR				1			
5	https://www.gay-serbia.com/	Serbia	Yugoslavia			1				
6	http://www.fronta.ba/	Bosnia	Yugoslavia			1				
7	https://www.poslednyadres.ru/	Russia	USSR				1			
8	http://prozhitto.org/	Russia	USSR				1			
9	https://kph.org.pl/	Poland	Eastern Bloc		1					
10	http://www.labrizs.hu/	Hungary	Eastern Bloc		1					
11	http://www.demokrate.me/	Montenegro	Yugoslavia			1				
12	https://www.memo.ru/	Russia	USSR				1			
13	https://www.dissernet.org/	Russia	USSR				1			
14	http://kompost.ru/	Russia	USSR				1			
15	http://www.lsv.org.rs/	Serbia	Yugoslavia				1			
16	http://www.sds-org.rs/	Serbia	Yugoslavia			1				
17	http://www.sdp.ba/	Bosnia	Yugoslavia			1				
18	http://www.sdsm.org.mk/	North Macedonia	Yugoslavia			1				
19	http://urokiistorii.ru/	Russia	USSR				1			
20	http://www.lmbtszovetseg.hu/	Hungary	Eastern Bloc		1					
21	http://ksmrus.ru/	Russia	USSR				1			
22	http://zekovnet.ru/	Russia	USSR				1			
23	http://podem.org.mk/	North Macedonia	Yugoslavia			1				
24	http://www.pd.al/	Albania	Eastern Bloc		1					
25	http://www.aleancaiqbt.org/	Albania	Eastern Bloc		1					



Seed List

Seed List TLDs

Crawled TLDs

.com

.com Link Outs

.org



Explore

<https://archive-it.org/collections/11360>

[HOME](#) [EXPLORE](#) [LEARN MORE](#) [CONTACT US](#)

[Login](#)

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive



[Explore](#) >> [Ivy Plus Libraries Confederation](#) >> [Websites of the Former Soviet Union & Eastern Europe](#)



ARCHIVE-IT

Websites of the Former Soviet Union & Eastern Europe

Collected by: [Ivy Plus Libraries Confederation](#)

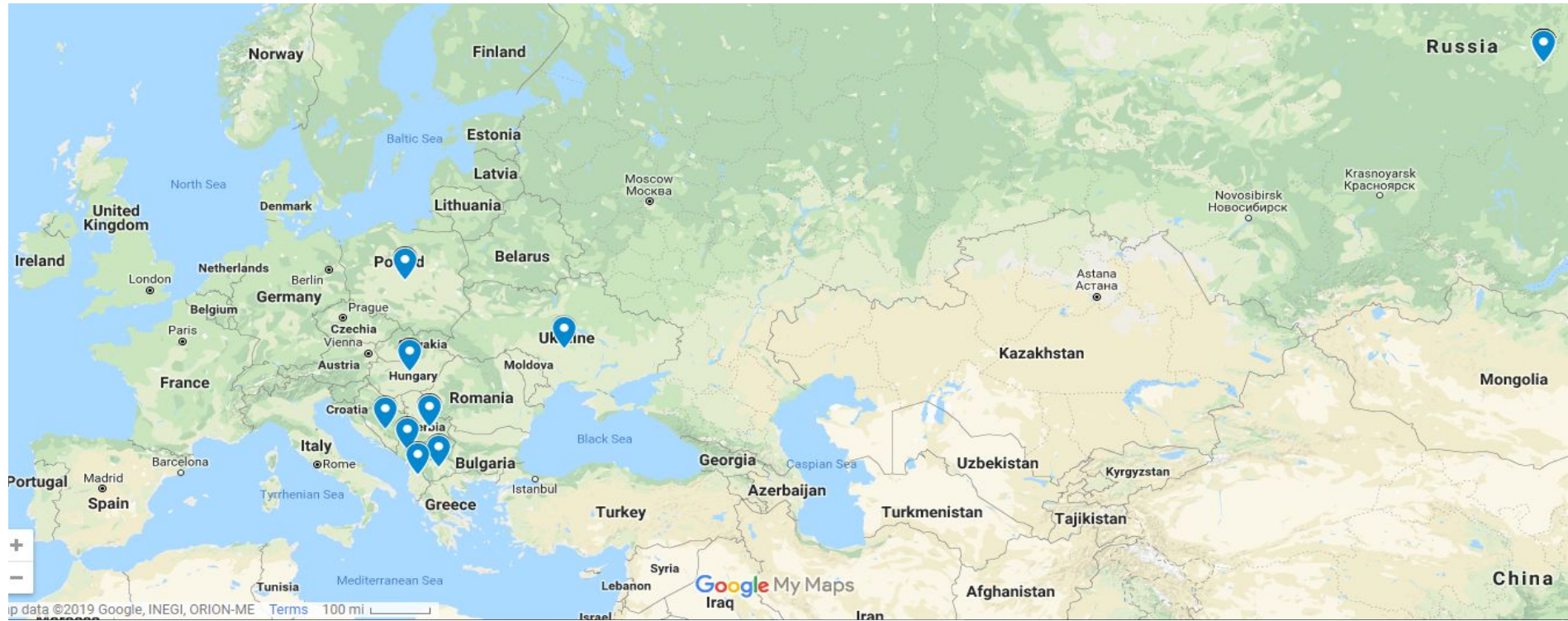
Archived since: Nov, 2018

Description: The Eastern Europe and Former Soviet Union Web Archive is an initiative developed by librarians at Columbia, Princeton, Yale, and New York Universities, and the New York Public Library, in partnership (as the Ivy Plus Libraries Confederation), with Brown University, the University of Chicago, Cornell University, Dartmouth University, Duke University, Harvard University, Johns Hopkins University, and the University of Pennsylvania. The collection is curated by Thomas Keenan (Princeton), Robert Davis (Columbia), Anna Arays (Yale), Alla Royslance (New York University), and Bogdan Horbal (New York Public Library). The Archive represents an effort to preserve research-valuable web content from Eastern Europe and the territories of the Former Soviet Union by a group of research librarians responsible for that part of the world. The countries of the region in recent years have been publishing a wide variety of websites likely to be of value to contemporary and future humanities, social science, and history projects, and this Archive has been established as an attempt to identify, capture, and preserve this material. The thematic and generic scope of the archive is deliberately broad, and includes websites published by political parties, non-governmental organizations and activist groups, artists and cultural collectives, and historians, philosophers, and other intellectuals.

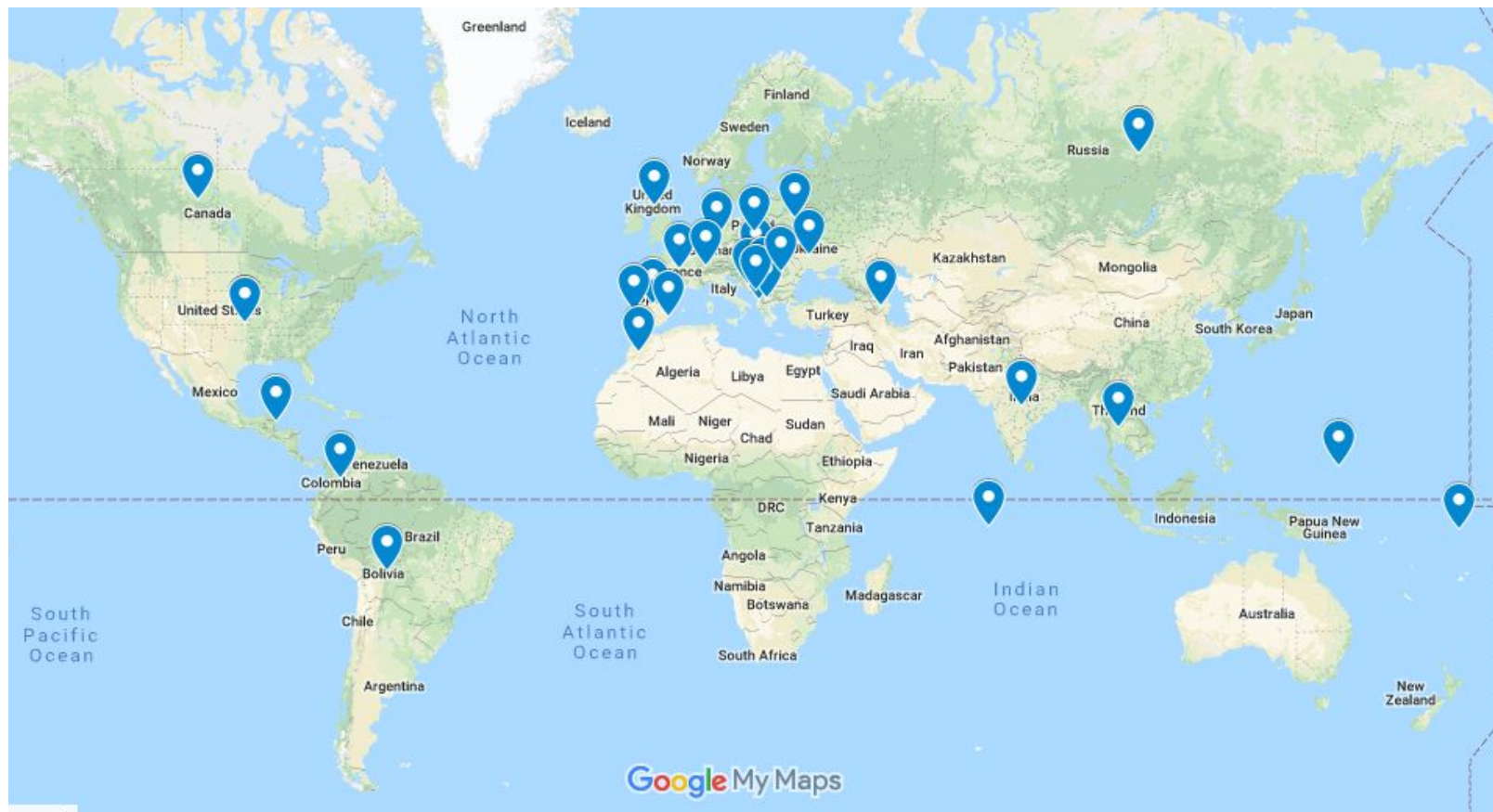
Subject: [Arts & Humanities](#), [Society & Culture](#), [Government](#)

Creator: [Ivy Plus Libraries Confederation](#)

Seed List Map

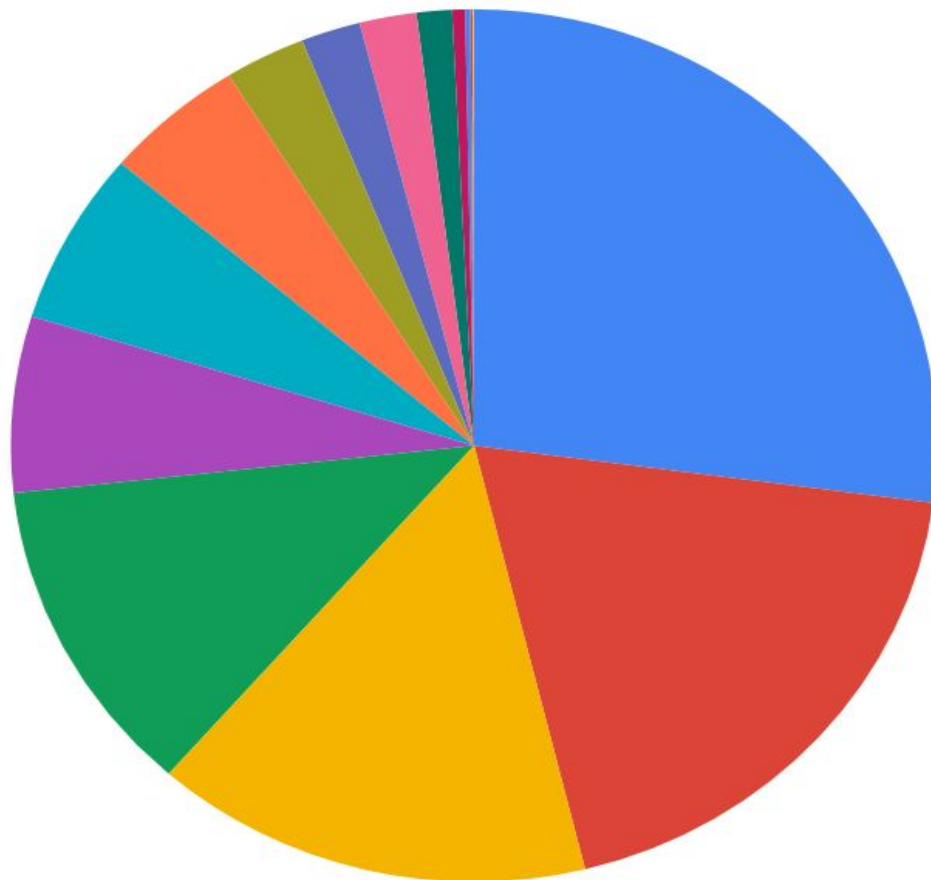


Crawled URLs



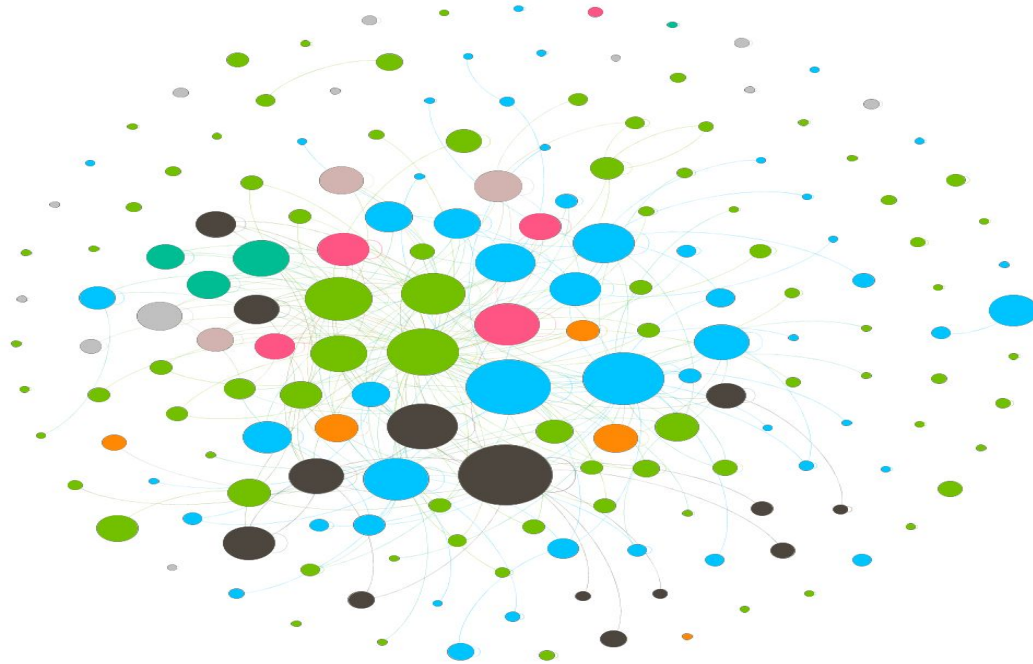
Crawled URLs

TLD	Number
ru	15
org	8
hu	8
rs	4
al	3
mk	3
pl	3
ba	2
me	2
com	1
info	1
ua	1

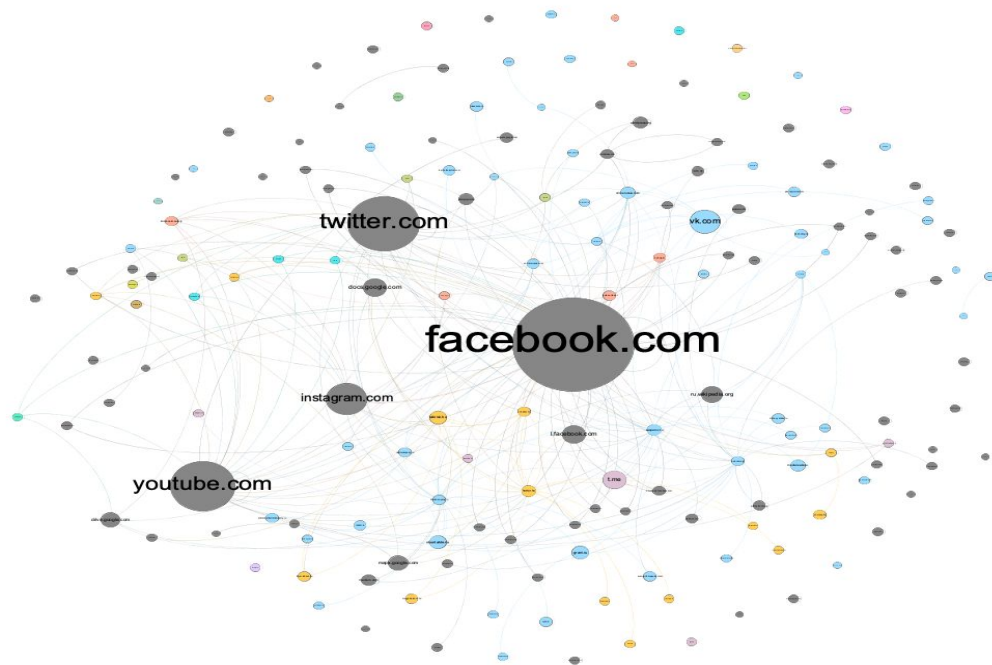


- org
- hu
- ru
- pl
- com
- ua
- al
- ba
- rs
- mk
- me
- 28 more

Network Analyses of What is in the collection



Links out from the archived content

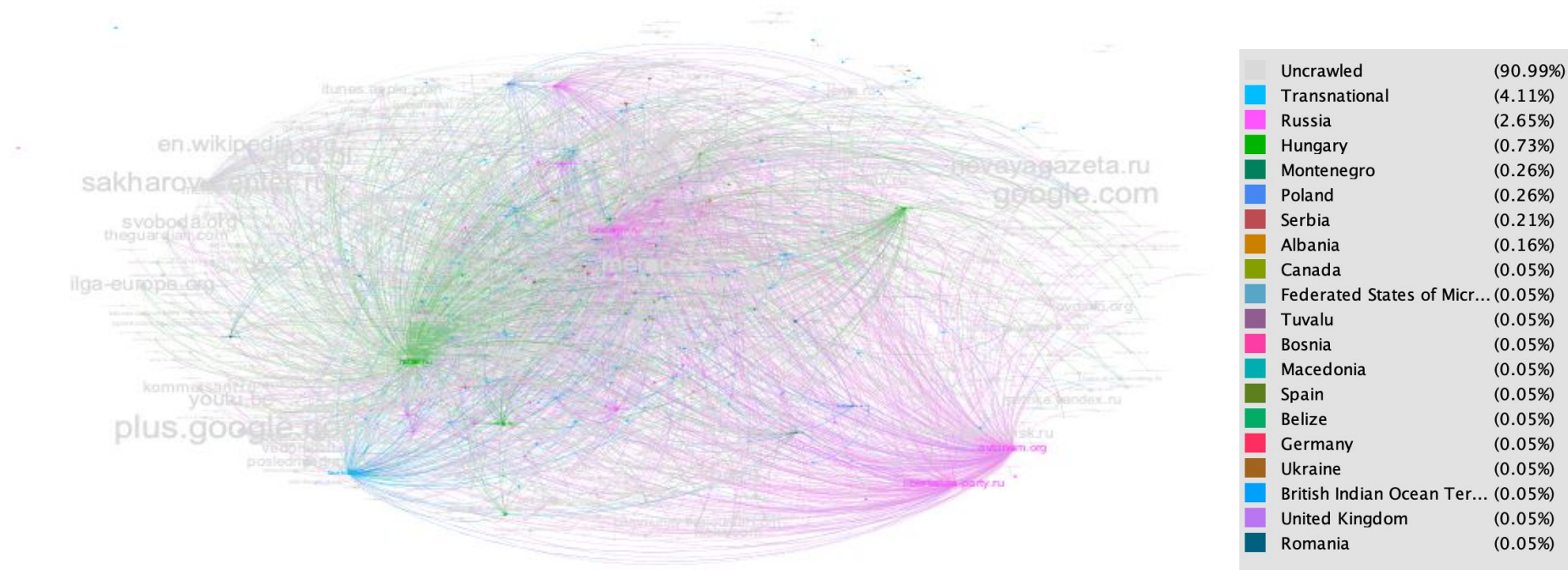


Uncrawled	(90.99%)
Transnational	(4.11%)
Russia	(2.65%)
Hungary	(0.73%)
Montenegro	(0.26%)
Poland	(0.26%)
Serbia	(0.21%)
Albania	(0.16%)
Canada	(0.05%)
Federated States of Micronesia	(0.05%)
Tuvalu	(0.05%)
Bosnia	(0.05%)
Macedonia	(0.05%)
Spain	(0.05%)
Belize	(0.05%)
Germany	(0.05%)
Ukraine	(0.05%)
British Indian Ocean Territory	(0.05%)
United Kingdom	(0.05%)
Romania	(0.05%)

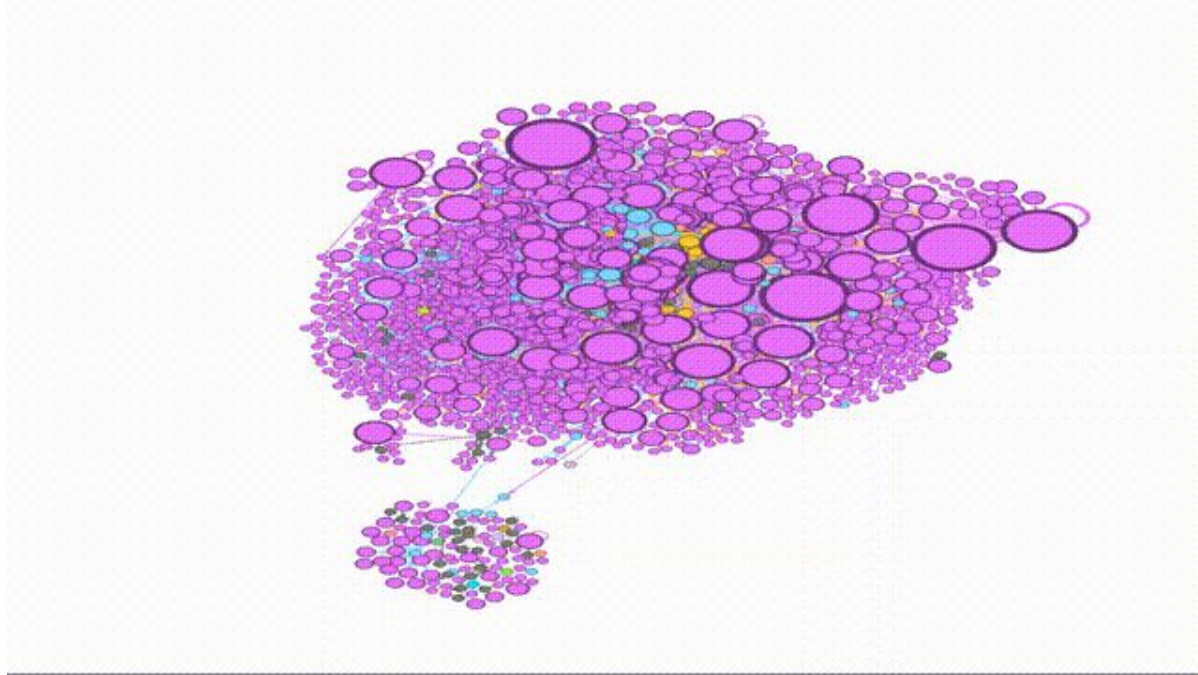
.com links out distribution



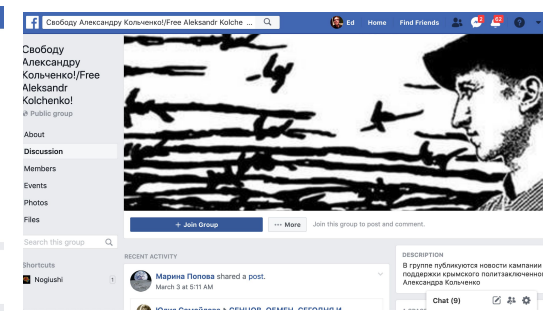
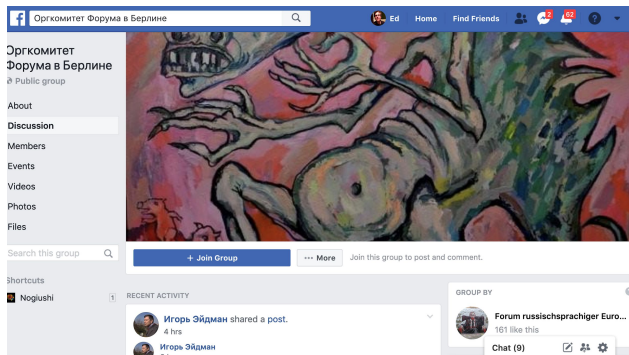
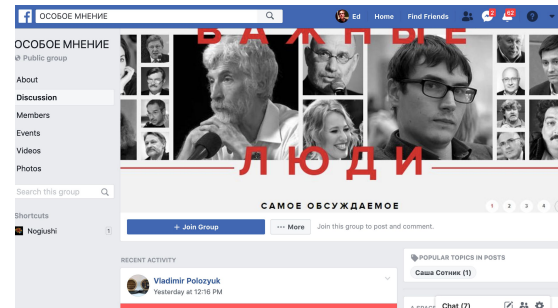
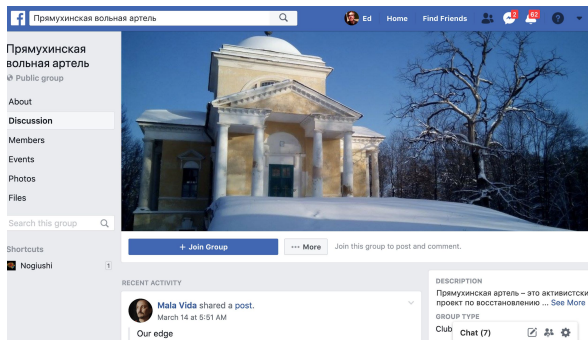
Uncrawled Domains



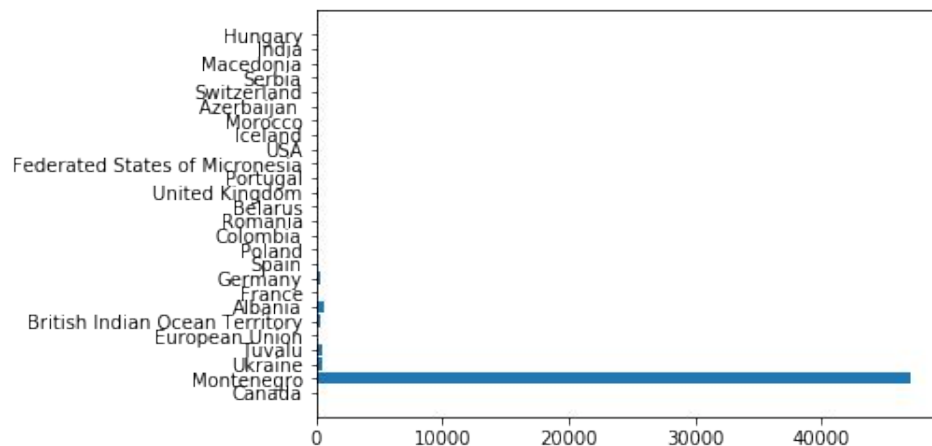
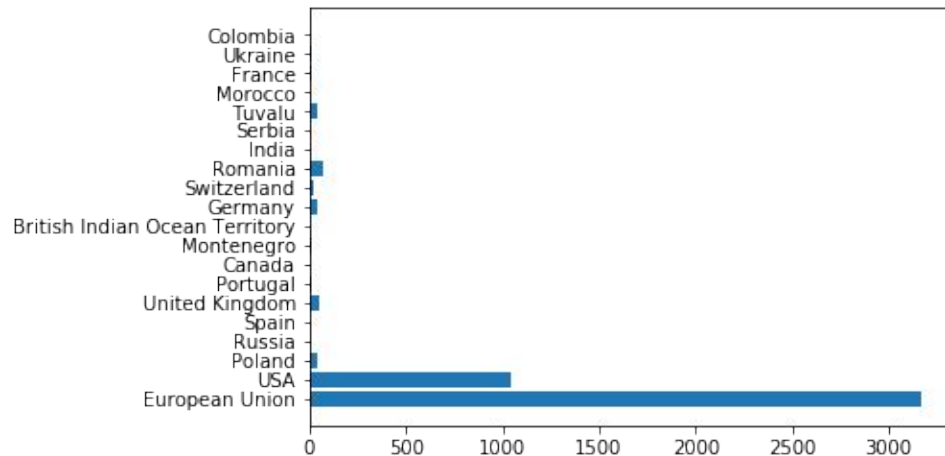
Uncrawled Hot Mess



Facebook Groups Not Crawled (35)



Outlinks: Hungary vs. Russia TLDs



Lessons Learned

Provenance of collections is important to inform your research questions

Defining geographic regions is challenging as many interpretations and politically charged

The web is not easy to define to one geographic region as there are lots of transnational web platforms

Still need to do a visual analysis of a sample of the content and manually code content

Supplied AUT domain visualization included uncrawled content that was not part of the collection.



Q&A

Thank you for listening!