The Disinformation Laundromat

An OSINT tool to expose mirror and proxy websites

SUMMARY AND FINDINGS

A consistent challenge in tackling the spread of disinformation, state-sponsored propaganda, and extremist content is our inability to rapidly detect and expose the ways in which bad actors launder harmful content into the online information ecosystem through proxies, cut-outs, aggregators, and mirror sites. This is particularly salient in the context of the war in Ukraine, where Kremlin-affiliated actors have used mirror sites to circumvent EU and tech company bans intended to limit the spread of Russian propaganda. This has allowed disinformation about the war to continue to proliferate across the internet, reaching European audiences through links and websites that obfuscate their affiliation with the Russian government.

Part of the problem in detecting this activity is that the process of identifying networked websites is largely manual, resource intensive, and limited in scale. Although, there are existing OSINT and media tracking tools that can provide useful pieces of the intelligence puzzle, there is not a single analytic tool that pulls together the various threads that analysts need to investigate linkages between suspicious websites.

The purpose of the Disinformation Laundromat is therefore to provide the OSINT community with a tool that can more effectively identify connections—both at the narrative and technical level—between seemingly unrelated websites. In the past ;^) week, collaborators reviewed tools that already exist to examine what data can be procured about a webpage. We further examined the webpages of news websites identified as mirrors in comparison with legitimate news sources, to see what data can be used to point to common news production. We found that similarities in meta data can, in some cases, be immediately useful to discover common ownership. Such as shared IP addresses, registration locations, AdSense and verification IDs. In other cases, it may be productive to do a content-level analysis.

MAIN RESEARCH QUESTION

What commonalities do mirrored sites have that can be leveraged to identify them as mirrors?

RESEARCH SUB-QUESTIONS

Are there shared attributes on a content level that point towards the same production tactics?

Are there technical indicators like metadata and verification IDs that connect disparate sites to the same owner?

Data collection

The first step was a tool review. We searched for existing tools that get various types of metadata about a webpage. These were found through browser queries. We tested out the following tools: WIG, Photon, CTFR, TheHarvester, BuiltWith, DNSlytics, URLscan, WPScan, IntelOwl, IntelX, Sn0int, and SimilarWeb. We found a host of metadata that could be useful. However, we ran into some roadblocks in testing these tools out. Some are oriented towards brand integrity protection and have subscription models oriented towards a company budget rather than individual users. These were completely blocked without a subscription. Others do have subscription models and allow for running a few requests without payment. Two of these - BuiltWith and URLScan were further incorporated into the tool. BuiltWith scans a webpage and reports back all the external technology it is using. This includes ad delivery and analytics plugins from Google and payment gateways embedded on the page. URLScan provides information about variable names used in the webpage and the certificates of the webpage and the technologies used on the webpage. These can be matched against other websites to see if they are set up with the same structure or use the same certificate for a technology.

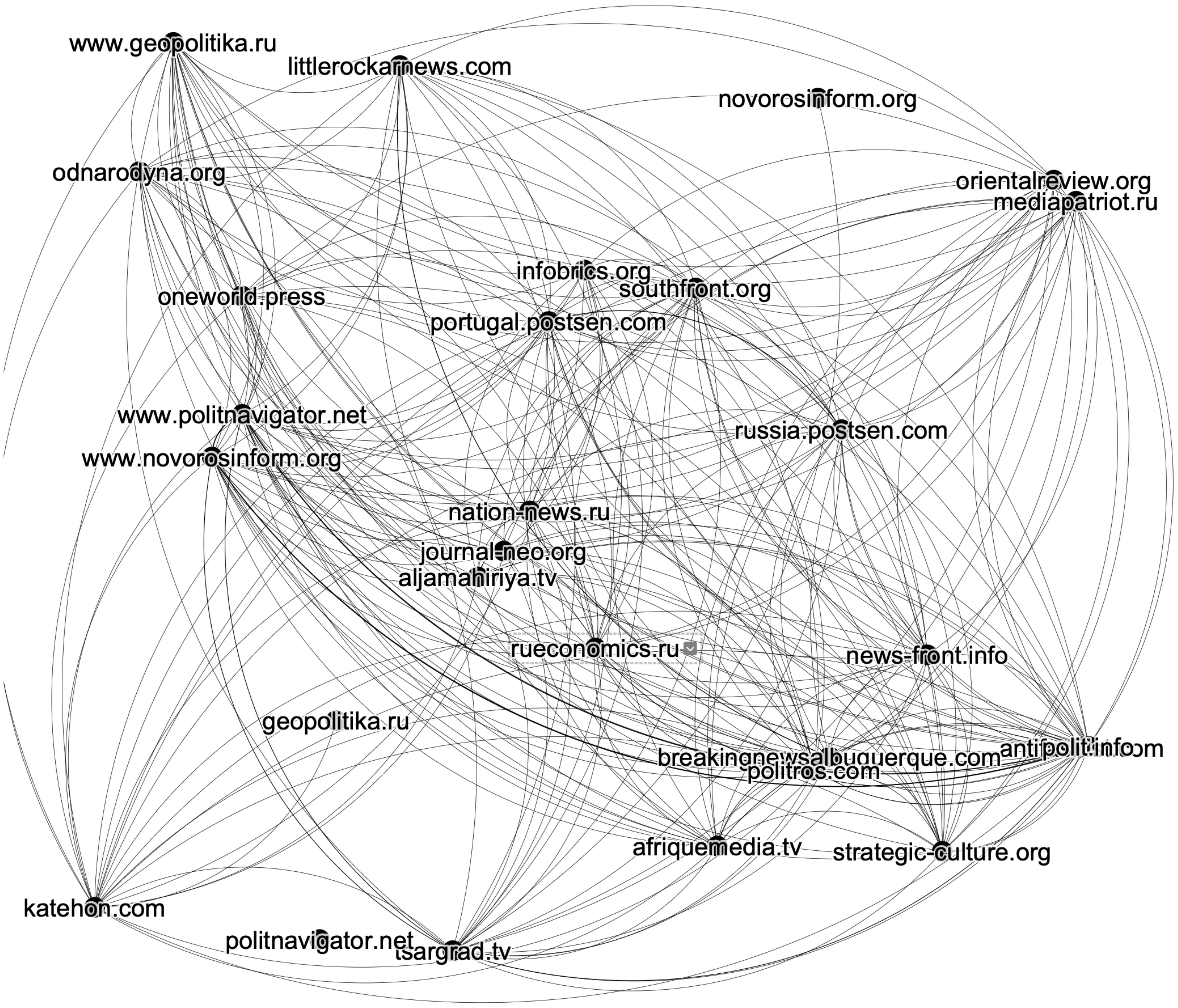
Data curation

We classified the attributes we found into a tier listed based on how conclusively they can say that the owners of a collection of sites is the same. The first tier, entitled “conclusive”, is metadata that, if matched, demonstrates a high level of probability that a collection of sites is owned by the same entity. This includes information like shared analytics and search engine IDs. They are conclusive because each ID is associated with only one account. The second tier is “Associative” data. These indicators point towards a reasonable likelihood that a collection of sites is owned by the same entity. This information can be useful if they have highly similar patterns of sourcing and structuring their data. Using the same source for images for instance is not suspicious in one instance, however, it can be if it exhibits a highly similar pattern of content production processes. These tend to be indicators linked to shared content delivery networks and meta tags in the HTML. Tier 3 are “Tertiary” indicators that could be circumstantial and should be substantiated with indicators of higher certainty. These include shared architecture such as plugins and CSS classes. PHashing is also used to determine whether the images they use are similar because often images might be slightly altered to avoid content detection.

Visualisation and Analysis

To test out the tool we built, we ran it on ~50 domains to see what similarities cropped up. We added a minimum requirement of at least an 90% match between tier one indicators. Using this information, we assigned each comparison a weight and visualised it in Gephi as shown below. The graph demonstrates that some of these interconnected websites are quite connected. This is the case for websites that present aesthetically to be different but are in fact owned and run by the same entity. Some of these website are not connected with other mirrors. These pose interesting questions about future work and what other indicators can be incorporated to track proxy activity. The tool currently does not account for shared authors and users across websites, privacy polices, external endpoint calls, sitemaps, and content. We believe that in doing so we will be able to account for the disconnected nodes. Additionally, we are yet to test out the tool against legitimate sites. This may prove some of our indicators as redundant or requiring more fine tuning.

VISUALISATIONS AND ANALYSIS



PARTICIPANTS

Alicia Bargar, Peter Benzoni, Gaurika K. Chaturvedi, Helena Schwertheim, Bret Schafer