BACHELOR INFORMATICA

UNIVERSITY OF AMSTERDAM

Virtual silos on the web

Jelmer Neeven

June 6, 2017

Supervisor(s): Martin Lopatka (Mozilla), Maarten Marx (UvA), Hosein Azarbonyad (UvA)

Signed:

Abstract

This thesis project explores possible ways of gaining insight into the relationship between the structural organisation of websites and the tracking networks they employ. To do so, 355,000 URLs across roughly 87,000 domains are crawled, capturing their third-party cookies and scripts as well as their external (i.e. to another domain) hyperlinks. The crawler is started with a list of 40,000 seed URLs, gathered from several selected communities of the popular website `reddit.com`. The collected data is used to construct different graphs, on each of which the Louvain Modularity community detection algorithm is applied, the results of which are then compared.
It turns out that many of the resulting communities (*virtual silos*) are not interesting to look at in the context of tracking code, but that advertisement networks can be identified in sub-communities, hidden within bigger silos. However, manually inspecting these bigger silos ranges from difficult to impossible as they become too big to visualise. Although none of them are applied to this project, possible solutions are suggested, establishing this thesis as a basis for further research.

Contents

Chapter 1

Introduction

As described by [11]:
"*Third-party web tracking is the practice by which third parties like advertisers, social media widgets, and website analytics engines — embedded in the first party sites that users visit directly — re-identify users across domains as they browse the web.*"
Not only have tracking techniques become significantly more advanced since the dawn of the internet [4], their prevalence has since become widespread, making it near impossible for users to avoid being tracked [11, 4, 13][1].
As user behaviour data significantly improves advertisement relevance [9] and therefore increases revenue, gathering more data about website visitors is beneficial for both the advertiser and the owner of the website on which the advertisements are placed, encouraging the latter to employ one or several tracking networks in order to increase the revenue earned by website visits. The entities owning these third-party tracking networks may then use the gathered data to offer advertisement services to companies, creating a profit for all three parties involved[2].

1.1 Hypothesis: virtual silos and tracking code

We hypothesise that a relationship exists between the structural organisation of the Web and a revenue model based on advertisement and tracking of users. We predict the existence of heavily interlinked clusters of web pages with very few outgoing hyperlinks, further referred to as (virtual) *silos*, and a connection between the structure of these silos and the tracking networks used by the websites within.
To gain insight into the plausibility of this hypothesis, the following research will be conducted:

- A set of seed URLs is gathered from multiple diverse sources such that they are likely to form a representative sample of the Web.
- These seed URLs are crawled, collecting both their hyperlinks to other domains as well as the third-party cookies and embedded page code (further referred to as *scripts*) they use. The collected hyperlinks are fed back into the crawler, which will continue to crawl all URLs until it surpasses the time limit set for the collection process.
- The collected data is used to build a communication graph of domains and tracking code hosts.
- Through community detection algorithms, *virtual silos* are identified. One set of silos will be detected using just the hyperlinks, and others sets will also incorporate the cookie and script data.
- The resulting silos are analysed and compared with the aim of providing insights into the way that web content is organised in the context of driving market forces that directly or indirectly draw revenue from web content. As a minimum, this analysis should provide directions for further research.

[1] `https://trackography.org/`, accessed on 4th June 2017, shows which entities might potentially track user behaviour on several media websites.
[2] The fourth involved party being the user, who does not necessarily gain any benefits from the process.

Ideally, this thesis project would look at solely those scripts and cookies that are tracking user behaviour, as these are the keystones of the online advertisement revenue model [13, 14, 9]. However, because identifying tracking code is not a trivial task [13], making a distinction between cookies and scripts that are indeed tracking the user and those that are not exceeds the scope of this thesis. Whenever a web page uses a cookie or script that is hosted on another domain (i.e. third-party cookies or scripts), the cookie or script is suspected of tracking the user. Consequently, any occurrence of the term *tracking code* in this thesis will henceforth refer to third-party cookies and scripts in general.

Chapter 2

Collecting data

As the Web consists of at least several billion pages[1], it is not feasible to analyse them all. Apart from the technical limitations such as the size of the resulting dataset and the time required to thoroughly analyse it, crawling a set of websites has several practical implications (most of which caused by the loading time of pages and the time in between HTTP requests and responses) that limit the amount of pages that can be crawled in a given amount of time.
Therefore, it is essential that a subset of pages is selected that is as general as possible (i.e. concerning different content, owned by different parties, etc.) to prevent the results from being heavily biased. For this project in particular, this subset should contain pages from several different silos in order for us to be able to identify them as such. As it is unknown which websites these silos might contain before they are identified, a sensible approach is to pick websites that are likely to form a *community*.

In network theory, many different definitions of a community exist, which are not agreed upon unanimously [7]. Generalising the different definitions, many agree that a community is a group of nodes that is more likely to connect to each other than to nodes from another group, which is also roughly the definition of a silo as described previously.[2] Research has not only shown that these communities are present in real-life social networks (i.e. networks representing interpersonal interactions), but also that they can be accurately identified by community detection algorithms [8, 3, 1].[3]
While there is no guarantee that websites referenced in communication within a specific community of people also form a community of websites, crawling URLs found across several different social communities should presumably result in observable clusters of websites that may or may not directly correspond to the social communities in which they were found. Consequently, a set of (assumed) online communities will be selected from which the set of seed URLs will be collected. Recursively crawling these seed URLs and their external (i.e. to a different domain) hyperlinks should hopefully lead to several identifiable communities (i.e. silos) of websites.

2.1 Reddit

As described on Wikipedia, *"Reddit is an American social news aggregation, web content rating, and discussion website."*[4] It allows its users to post virtually any kind of content, which may then be upvoted or downvoted by other users (respectively representing enjoyment or dissatisfaction of the viewer) to reflect its popularity. This score then determines the position of the content (referred to as a *post*) on overview pages, which by default prioritise popular content (although users can choose different sorting mechanisms).

[1] `http://www.worldwidewebsize.com`, accessed 24th May 2017
[2] The silo definition is deliberately vague at this point, as results will show what exactly are appropriate definitions of *very few outgoing links* and *heavily interlinked*.
[3] [1] does not present any new research, but refers to several existing sources (chapter 9, p.3).
[4] `https://en.wikipedia.org/wiki/Reddit`, accessed 26th May 2017

What makes Reddit interesting in the context of this thesis is that it is divided into different *subreddits*, each of which corresponds to a specific area of interest. By subscribing to specific subreddits, the user controls the content they see, and as every user has the ability to create them, there is no limit to the amount of subreddits with more than a million of them existing today.[5]. Because these subreddits all revolve around a specific theme, be it broad (i.e. science, politics) or specific (i.e. Hillary Clinton, the *A song of ice and fire* book series), they are 'populated' by users that are interested in that specific type of content. It is because of this shared interest that these users can be considered to form a community together, although the 'strength' of the community may vary depending on the topic. Furthermore, because of Reddit's upvote system, the content is moderated directly by the users, providing a simple way to measure the relevance of a post to the given context.

These properties allow us to select a set of subreddits to represent different communities and collect the URLs shared by their users. Using the *score* of each post and its comments, which corresponds to the amount of upvotes subtracted by the amount of downvotes, a relevance threshold can be determined to filter out posts and comments that do not necessarily belong to this community.

2.1.1 Selecting subreddits

To prevent the data from being biased, it is important to include data from both popular and unpopular sources. To this end, URLs are gathered from both the *top* section of Reddit, which sorts posts in descending order of upvotes, as well as the *new* section of Reddit, which sorts posts by the time they were posted, both sections displaying posts irrespective of their subreddits. By selecting posts from *new* that are at least a week old but received a below average amount of upvotes, their content has shown to be unpopular, with the posts from *top* covering the other part of the spectrum.
Given the limited timespan of this project and the amount it takes to crawl the subreddits for their URLs, both Reddit sections have been crawled for an hour on the 14th of April 2017, yielding approximately 3000 and 6500 seed URLs respectively.

Selecting appropriate subreddits to represent different communities is a more difficult task, because they have to be as diverse as possible. Ideally, the data would include both active popular communities as well as active fringe communities. To determine whether a community is popular or fringe, a method has been explored to analyse the URLs shared in a specific subreddit, counting the URLs to other subreddits. Using this feature, subreddits could be chosen that contain many posts (i.e. are *active*), but rarely link to any other subreddits, making them isolated communities (i.e. *fringe*). However, this process was too lengthy because of the time required to analyse the subreddits.
Instead, to maximise diversity, subreddits have been chosen from the different categories under *sfw subreddits* at `redditlist.com`. It contains twenty different categories, each of which displays a list of the subreddits under that category. By matching the subreddit against each other, only those subreddits that are specific to one of the categories (i.e. do not appear under another category) have been chosen, as those are most likely to form isolated communities. Picking two or three of them per category, supplementing them with a few more general, popular subreddits such as `/r/science` and `/r/gaming`, and manually intervening when needed (for example, *alternativeart* was under pictures rather than art), this results in a list of 51 subreddits, divided in the following categories:[6]

```
'Art': ['alternativeart', 'graphic_design'],
'Culture': ['cyberpunk', 'opieandanthony'],
'Discussion': ['rant', 'socialskills', 'mensrights'],
'Gaming': ['leagueoflegends', 'casualnintendo', 'gaming'],
'Humor': ['scenesfromahat', 'blackpeopletwitter', 'bikinibottomtwitter'],
'Info': ['abrathatfits', 'explainlikeimfive'],
'Lifestyle': ['fitness', 'makeupaddiction', 'relationship_advice'],
'Location': ['losangeles', 'croatia', 'turkey'],
```

[5] `http://redditmetrics.com/history`, accessed 26th May 2016
[6] Note that there are only 19 categories, because the *meta* category did not contain any category-specific subreddits.

```
'Movies': ['movies', 'netflixbestof'],
'Music': ['music', 'kpop', 'popheads'],
'News And Politics': ['socialism', 'hillaryclinton', 'the_donald'],
'Pictures': ['perfectfit', 'highqualitygifs', 'abandonedporn'],
'Q And A': ['iama', 'samplesize'],
'Read And Write': ['writing', 'fountainpens'],
'Science': ['science', 'space', 'chemistry'],
'SFW Porn': ['historyporn', 'gentlemanboners', 'militaryporn'],
'Sports': ['mma', 'eagles', 'reddevils'],
'Technology': ['android', 'jailbreak', 'windowsphone'],
'TV': ['strangerthings', 'community', 'rickandmorty']
```

Each of them is crawled for up to an hour, or until 5000 URLs have been gathered. Combined with the *top* and *new* sections, they yield roughly 250,000 URLs.

2.1.2 Filtering URLs

To decide which content is relevant to a community, an upvote threshold needs to be set. Because this is difficult to do manually, the same principles are followed as in a popular analysis of posts related to Donald Trump[7]: only posts with a score of at least 25 are selected, ignoring links to Reddit or popular image hosting websites. Because the amount upvotes per post varies per subreddit, the distribution of upvotes has been examined to determine what the equivalent of this threshold is for every subreddit:

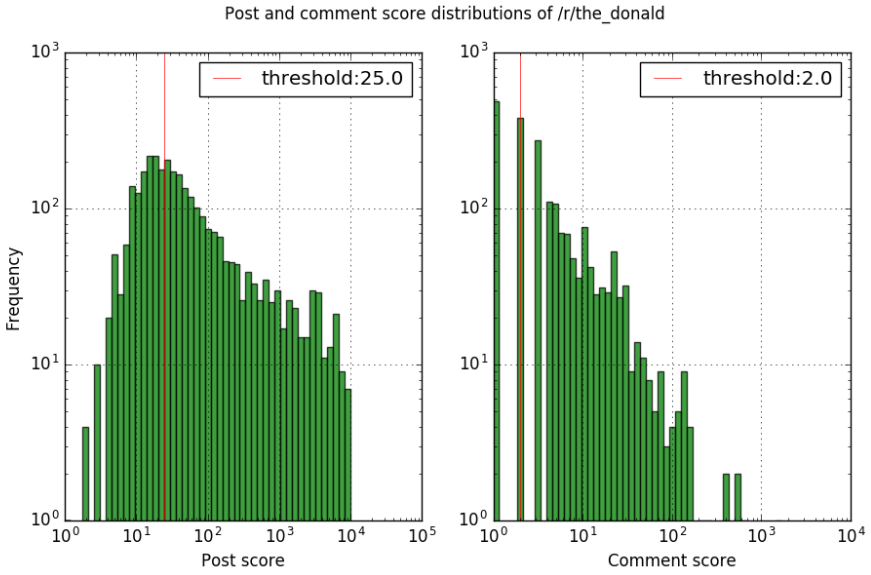

Figure 2.1: Post and comment score distributions of the */r/the_donald* subreddit. Thresholds correspond to 42.08% of the posts or comments having a score less than the threshold.
Both axes use logarithmic scale.

A threshold of 25 upvotes for posts in `/r/the_donald` corresponds to 42.08% of the posts having a score below this threshold. Using this metric, the threshold can be calculated separately for every individual subreddit. While post and comment score do not follow the exact same distribution, this threshold will be generalised to both posts and comments for lack of a better approach.
As figure 2.2 shows, these distributions are not unique to `/r/the_donald`, but apply to other subreddits as

[7] `https://www.reddit.com/r/dataisbeautiful/comments/5x2sie/the_most_linked_sites_this_month_by_the_donald_vs/`, accessed 12th of April 2017

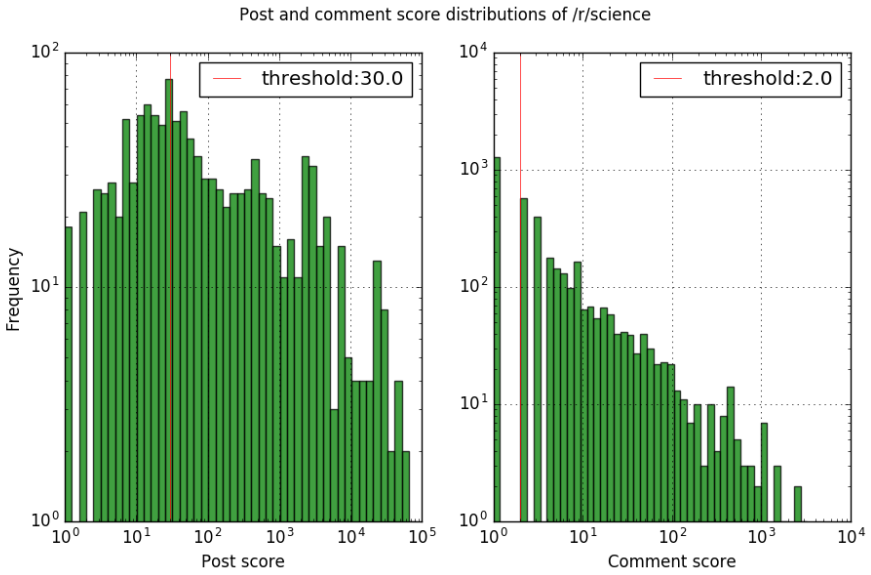

Figure 2.2: Post and comment score distributions of the */r/science* subreddit. Thresholds correspond to 42.08% of the posts or comments having a score less than the threshold. Both axes use logarithmic scale.

well. While some subreddits vary more from this example than others, their distributions are generally similar, meaning the specified threshold approach will suffice for this thesis project. Of course, there is much room for improvement, but analysing the exact distributions and properties of subreddits exceeds the scope of this thesis.

Having filtered all contextually irrelevant URLs out of the collected data, it still contains many links to Reddit itself as well as media hosting websites and social networks, both of which will not be interesting in the context of this project. Media hosting websites simply host media, with different pages representing different files. They are unlikely to contain many hyperlinks, and usually each page contains the same content (except for the hosted file). However, by filtering them out from the seed URLs, we do not blacklist them completely; they may later be found in the crawling process and still be crawled. The same is true for social media websites, although their contents may differ per page. However, because their content is dynamic and may change depending on authentication (i.e. if the visitor has the rights to view a certain profile), the results will be relatively meaningless in the context of this thesis.
Filtering out Reddit links is straightforward, but deciding which websites to filter out is not. As displayed in figure 2.3, the uninteresting websites are in fact the most linked to. To filter them out, all links that end with a video, audio or image format are removed (i.e. *.jpg*, *.mp4*). Then, websites from the top 50 are blacklisted when they are manually judged to meet one of the following criteria:

- It is a popular social media website (i.e. Facebook, Twitter, Instagram)
- It exists solely to host media files (i.e. Imgur, Gfycat)

This results in the following blacklist:

```
['youtube.com', 'youtu.be', 'reddit.com', 'instagram.com', 'imgur.com', 'streamable.com',
'redd.it', 'twitter.com', 'facebook.com', 'reddituploads.com', 'gfycat.com',
'giphy.com', 'staticflickr.com', 'twimg.com', 'pinimg.com', 'vimeo.com']
```

In hindsight, there are many more websites in the top 50 that could have been blacklisted, such as pastebin and soundcloud. This is due to the fact that only the top 15 was originally inspected rather than the top 50. However, because the frequency of these links roughly follows a power law distribution and the top 15 has been correctly

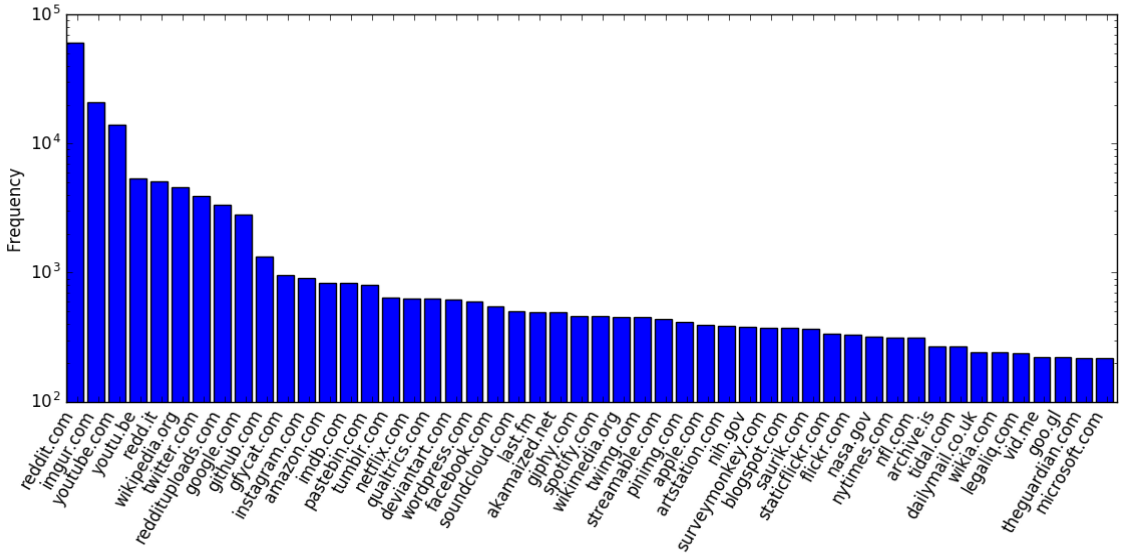

Figure 2.3: Top 50 most linked websites across all selected subreddits and their frequencies, before applying the blacklist. Frequency is on a logarithmic scale.

filtered, the remaining websites occur relatively infrequently and will not cause problems. Furthermore, the domain limit that is discussed later prevent the crawler from visiting these websites more than a thousand times anyway.

Finally, after filtering the duplicate URLs (i.e. exact duplicates, but also HTTP and HTTPS links to the same URL), roughly 40,000 seed URLs remain to use in the crawling process, distributed as displayed in figure 2.4.

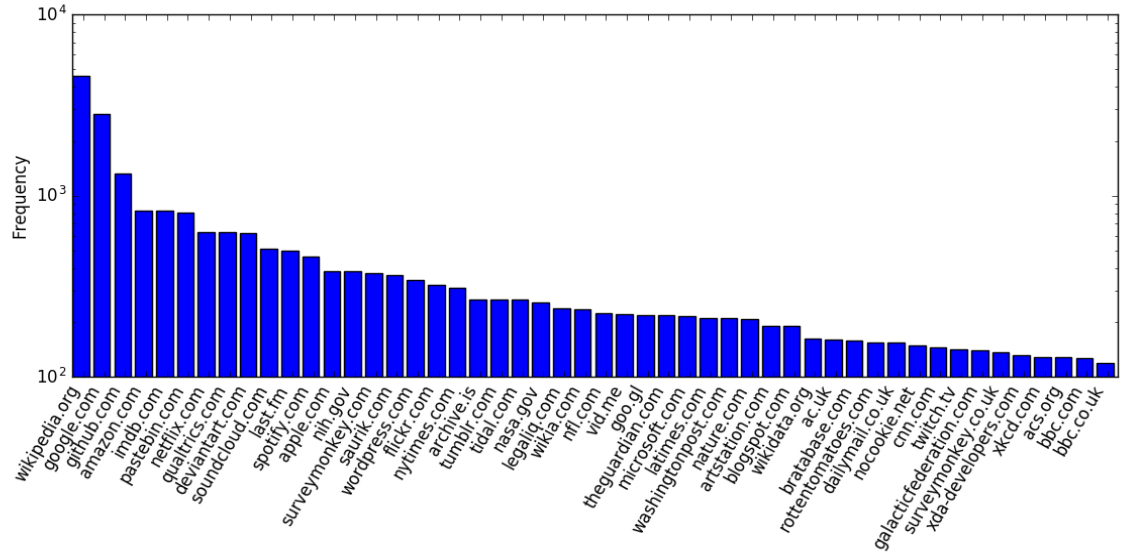

Figure 2.4: Top 50 most linked websites across all selected subreddits and their frequencies, after applying the blacklist. Frequency is on a logarithmic scale.

2.2 Writing a crawler

Now that a set of seed URLs has been selected, they can be used to gather the required data. For this thesis project, several crawler implementations have been explored, but due to platform-specific restrictions, the final choice is by far not the optimal crawling strategy.[8]

To clarify what is meant by *third-party* in the context of this project, it is important to specify what is considered the domain of a given URL, and therefore what is considered to be a different domain. For the purpose of this thesis, a domain has been defined as the second-level domain followed by the top-level domain, i.e. `example.com`. In cases where domain-name registries use the second-level domain to indicate the type of entity associated with the domain (i.e. .co.uk, .ac.jp), the third-level domain is also included. This means that any subdomain is removed, reducing `nl.facebook.com` to `facebook.com`, for example.
While this is favourable in most cases (the aim is to detect third-party tracking code, and a tracker found on `facebook.com` that originates from `nl.facebook.com` is not exactly third-party), it may occur that every subdomain of a given is a completely different website. For example, the websites `jelmer.blogspot.com` (A) and `hosein.blogspot.com` (B) are effectively separate websites hosted under the same domain, and are not necessarily related. By removing the subdomain from the domain definition, a link from A to B will not be considered a link to a different website. However, determining the difference between different subdomains exceeds the scope of this thesis, and this difference is not expected to significantly affect the results of this exploration.

Using the PhantomJS framework, a parallel crawler has been implemented that uses a master process and several worker processes, where the master process handles administrative tasks (i.e. preventing the same URL from being crawled twice) and each worker process crawls a single URL at a time. From each URL, the data is collected in the following manner:

- The page is loaded (excluding images) and its JavaScript code is executed.
- A list of third-party cookie sources is extracted. To capture this information, all incoming HTTP response headers are parsed for the *set-cookie* property. If this header is encountered in a response from a different domain than the crawled URL, the source URL of the response is captured.[9]
- The sources of third-party scripts embedded on the web page are collected. After fully loading the page and executing its JavaScript code, the resulting HTML is returned. Using XPath syntax, the `<script>` tags are extracted from the HTML code and if they have a third-party origin, their source is captured.[10]
- Similar to the scripts, the hyperlinks are extracted by filtering for `<a>` tags. Only the external hyperlinks (i.e. to other domains) are captured, but both internal and external hyperlinks are counted for potential later use.

All collected data is then returned to the master process, which writes it to file and assigns the worker a new URL:

- For every found hyperlink, its stripped version[11] is compared to the stripped version of the current total list of websites (both the uncrawled and crawled part of the queue). If the URL is not yet in the list, does not end with a media file extension and less than a thousand URLs of its domain have been crawled, the original URL (i.e. before stripping it as described) is added to the end of the queue.
- The master process selects the first URL in the queue and checks whether its domain has been crawled in the last fifteen seconds to prevent overloading the server.

[8] While this thesis project uses PhantomJS, the original chosen framework was a combination of Scrapy and Splash. This allowed for a much higher degree of parallelisation, increasing the expected amount of crawled URLs tenfold. PhantomJS was chosen in the end because of compatibility issues with the cluster on which the software was run.

[9] The URL is stripped of its protocol, *www* prefix and trailing slash. For the cookie URLs, the query parameters are also removed: whenever a #, :, ? or ; is encountered, the rest of the URL is ignored.

[10] See footnote 9.

[11] See footnote 9

- If it hasn't, its `robots.txt` file is consulted to determine whether the crawler is allowed to crawl it. For the sake of this project, the URL is crawled either way, but the information is written to file.
- The URL is passed to the worker, returning to the start of the cycle.

This setup was then run on the SURFsara Lisa cluster (a parallel computing facility), on which it ran for the maximum allowed time of 120 hours with 15 worker processes. It crawled 490,821 URLs, but as approximately 135,000 of them returned a *404 not found* error, data has been recorded for roughly 355,000 URLs.

2.3 Processing the data

The resulting output data is imported into a database on a domain basis, with a total of 83,657 unique crawled domains. The amount of external hyperlinks, cookies and scripts on each domain is distributed as displayed in figure 2.5.

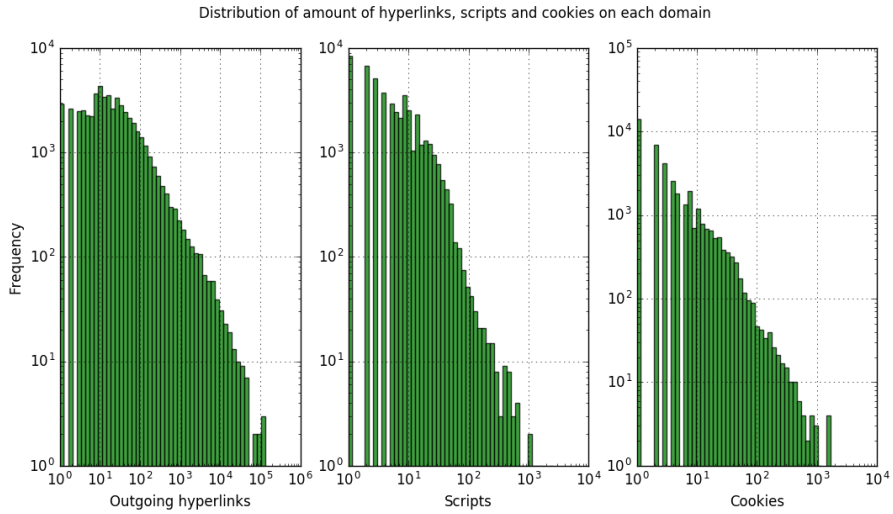

Figure 2.5: The distribution of the amount of hyperlinks, third-party cookies and third-party scripts employed on each of the crawled domains. Both axes use logarithmic scale.

As shown in figure 2.7 and 2.6, the top 50 most used cookies and scripts contains many (in)famous sources such as Doubleclick and Google Analytics. Surprisingly, jQuery, the most popular JavaScript library[12], is not present in the top 50. One might have expected such popular pieces of non-tracking third-party scripts to be detected as tracking code many times, thereby negatively affecting the quality of the resulting findings. Fortunately, this does not seem to be the case, or at least not at a large scale, as most listed cookie and script sources do indeed seem to be tracking code.

[12] `https://www.singsys.com/blog/jquery-most-popular-javascript-library-now-10-years-old/`, accessed 27th May 2017

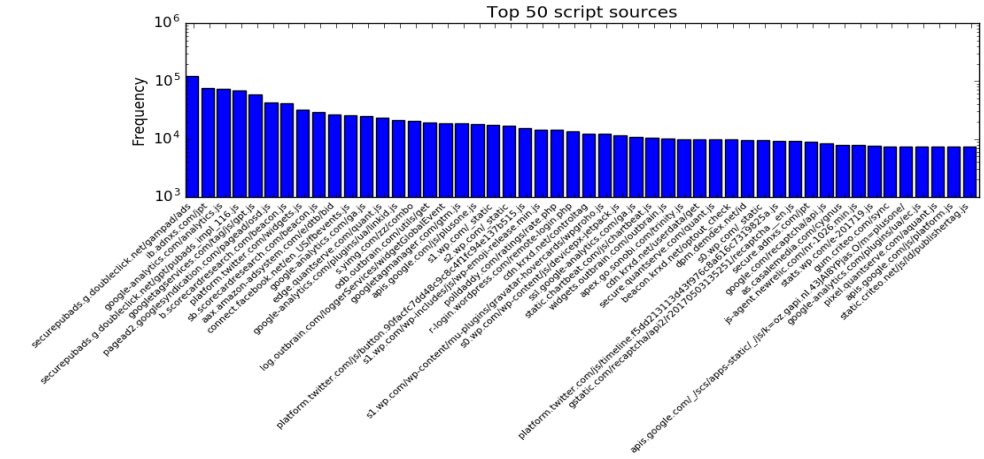

Figure 2.6: The distribution of the top 50 most used script sources across all crawled domains. Frequency uses logarithmic scale.

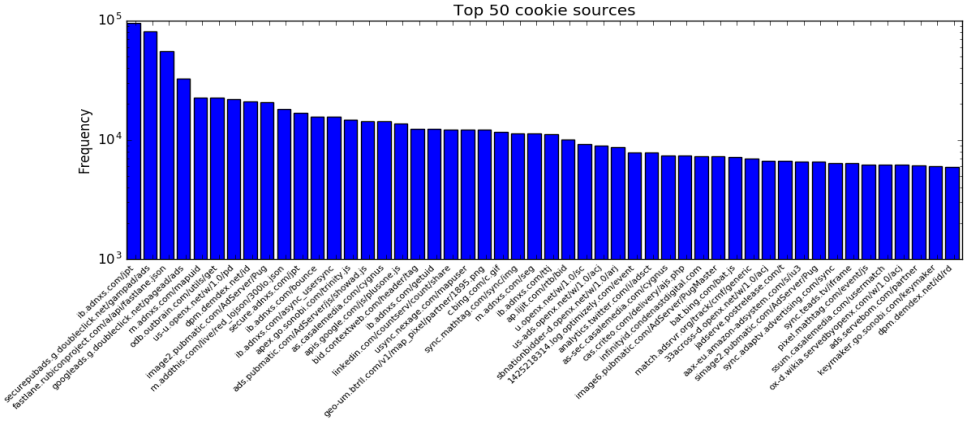

Figure 2.7: The distribution of the top 50 most used cookie sources across all crawled domains. Frequency uses logarithmic scale.

Chapter 3

Analysis of results

3.1 Creating a communication graph

Now that all the data is accessible, steps can be made towards a meaningful analysis. As the aim is to gain insight into the structure of the Web, and the Web is most definitely a network, the first model to represent the data that comes to mind is a graph. There are multiple ways in which the collected data can be turned into a graph, each of which is elaborated below.

3.1.1 Hyperlinking

It will be useful to first look at the most obvious way in which websites are connected. As hyperlinks are usually meant to be clicked on by (and are therefore visible to) the visitor, any connection showing up in a graph based on hyperlinks should be manually verifiable.
Constructing a graph from the hyperlinks is fairly straightforward. Every web page or domain is a node, and when a hyperlink exists between two web pages, their nodes are connected by an edge. As hyperlinks often only go in one direction (i.e. a certain web page links to another, without that page linking back to the first), a connection between two nodes is best represented by a directed edge.
The most direct way to represent the data would be having a separate node for every unique URL. However, because this exploration is aimed at the tracking code deployed across multiple domains rather than intra-domain differences between web pages, it makes more sense to create a node for each unique domain rather than URL.
Because a (set of) web pages can potentially contain multiple hyperlinks to the same domain or even URL, certain links can be stronger than others. This is represented in the graph by a weighted edge: the weight of an edge from one domain to another is equal to the amount of hyperlinks found on the first domain that link to the second domain.

This approach leads to the weighted directional graph shown in figure 3.1 with 259987 nodes and 833028 edges, of which 82026 are symmetrical (i.e. 41013 domains are linked in both directions). While only 83657 domains have been crawled, a total of 259987 domains is represented in the graph, because uncrawled domains that have been linked to are also included.
Because there are so many nodes in the graph, figure 3.1 is not too useful. In figure 3.2, the same graph is displayed, but nodes with a degree less than 2500 are filtered out. As the highest degree is 34117, this clears up the image enough to add the domain labels, showing us which websites are most prominent in the data.

There are 19 nodes after filtering, but because Blogspot and Wordpress are in there multiple times with different domains, there are 16 unique websites. It looks accurate, but two things should be noted:

- Only Facebook and Twitter are blue, while all nodes have a high in-degree. This is simply because of the way the Gephi ranking works. The colour is not affected by the ratio of in-degree to out-degree, so all we can derive from it is that out of all nodes in the displayed sub-graph, Facebook and Twitter have the

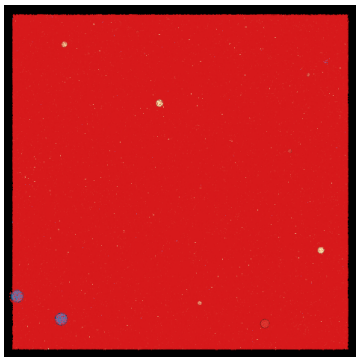

Figure 3.1: Hyperlink graph, in which every node is a domain and an edge represents a hyperlink.
Size is determined by degree and colour is determined by in-degree. Red is a low in-degree, white is neutral, and blue is high.

highest in-degrees. This seems reasonable, because they are both extremely popular websites, and likely more so than the other nodes.

- `bit.ly` shows up in this graph, and is therefore one of the more popular (i.e. more often linked to) websites. It has an in-degree of roughly 1200 and an out-degree of roughly 1500. However, when visiting a typical `bit.ly` URL, it will instantly redirect the visitor to the URL it has shortened (i.e. one belonging to another domain). All non `bit.ly` hyperlinks found there will be regarded as being external, because they are from a different domain than the one we originally visited. Therefore, the outgoing links from `bit.ly` are inaccurate. However, this problem also works the other way around: all links to `bit.ly` are actually links to the associated URL. The best way to solve this problem would be to have the crawler detect when the current URL is different from the originally visited URL (i.e. it has been redirected). If this is the case, an edge can be created between the original URL and the URL it redirects to.

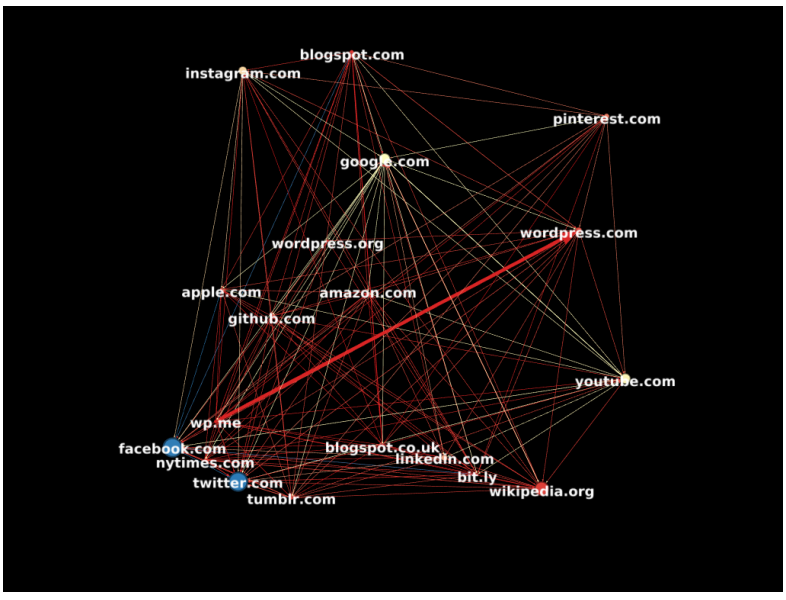

Figure 3.2: Hyperlink graph.
Nodes with a degree less than 2500 have been filtered out. Edge colour corresponds to the node it leaves from, and label colour is not dependent on any attributes.

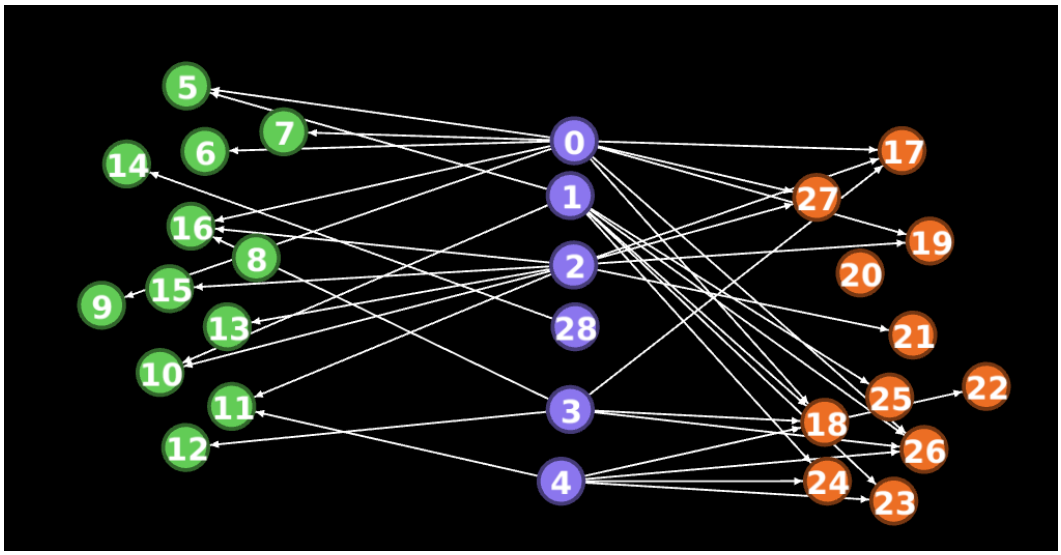

Figure 3.3: Example tripartite graph.
Green nodes are cookies, blue nodes are domains and orange nodes are scripts.

3.1.2 Tracking code

To now incorporate the tracking code data as well, several options exist. It may be interesting to create a graph that links websites together solely based on their tracking code, leaving out hyperlinks altogether. This graph would be an *affiliation network*: a tripartite graph that represents the affiliation of a domain with different tracking code sources, whereas the hyperlink graph can be considered a *social graph* [6](p.93-97). Multipartite graphs are graphs with different types of nodes, in this case domains, cookie sources and script sources, that link to nodes of other types, but never to nodes of the same type. For example, the graph could look like figure 3.3, in which none of the domains are connected. However, node 1 and 4 both (partially) link to the same set of scripts, which indirectly connects them. Using community detection on this graph, if two websites are determined to belong to the same community, this indicates that there somehow is a similarity in the tracking code.

Another possible approach is to create a graph with edges for both hyperlinks and tracking code. Essentially combining the previously mentioned social and affiliation networks, this would result in an (aptly named) *social-affiliation network* [6](p.93-97), providing insight in both the direct relations between websites through their hyperlinks and their indirect relations through their tracking code.
The benefit of that approach is that the community results can be compared one on one with the hyperlink communities; if a community has suddenly gained or lost domains, that indicates that something interesting is going on with the tracking code used on those domains.
The benefit of the first (tripartite) approach is that it solely looks at the tracking code, therefore indicating only the 'hidden' connections. If websites that seem completely unrelated use a highly similar set of tracking code, this indicates that something suspicious is happening. By comparing the resulting graph with the hyperlink graph, we can identify websites that are in different communities regarding the hyperlinks, but in the same community when using the tracking code.

As both approaches seem equally appropriate, both will be explored. The first approach will be referred to as the *tracking* graph, and the second will be referred to as the *total* graph.

Tracking graph

Linking domains together solely based on their tracking code results in a graph with 310735 nodes and 696197 edges as seen in figure 3.4. Many of the displayed nodes are script sources. Interestingly, `sportklub.hr` has

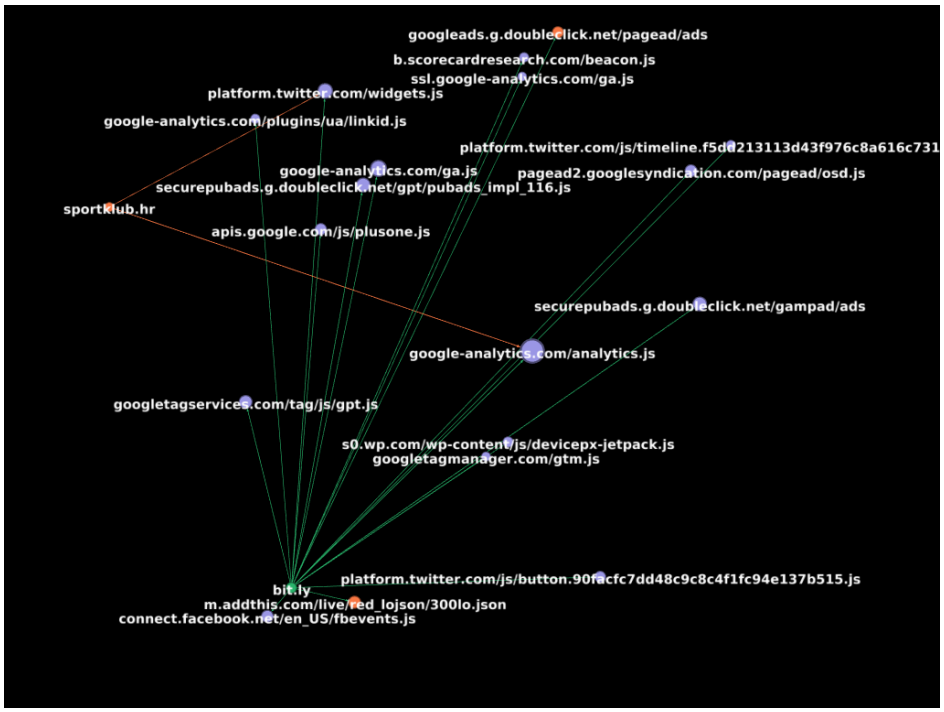

Figure 3.4: Tracking graph, in which a node is either a domain, cookie source or script source. Edges represent the code being deployed on the linked domain. Nodes with a degree less than 2500 are filtered out. Size is determined by degree and colour is determined by type. Green is a domain, blue is a script and red is a cookie.

outbound edges, but is a cookie. This means that it is both a crawled website and a third-party cookie host. Looking at the website, it is likely that they host some kind of sports widget people can include on their site. Out of all displayed nodes, only `bit.ly` is a domain, and it connects to almost all the other nodes. This makes sense, because different sites linked on `bit.ly` can be completely unrelated and are therefore likely to use many different types of tracking code.

Total graph

Using both hyperlinks and tracking code usage as links between nodes (and therefore using both domains and tracking code sources as nodes), a graph is constructed with 512934 nodes and 1528226 edges. Because this graph is essentially a combination of the other two graphs, it does not provide any useful insights. As can be seen in figure 3.5, no new features have shown up. The potential benefit of using this graph will only become clear once the community detection has been carried out.

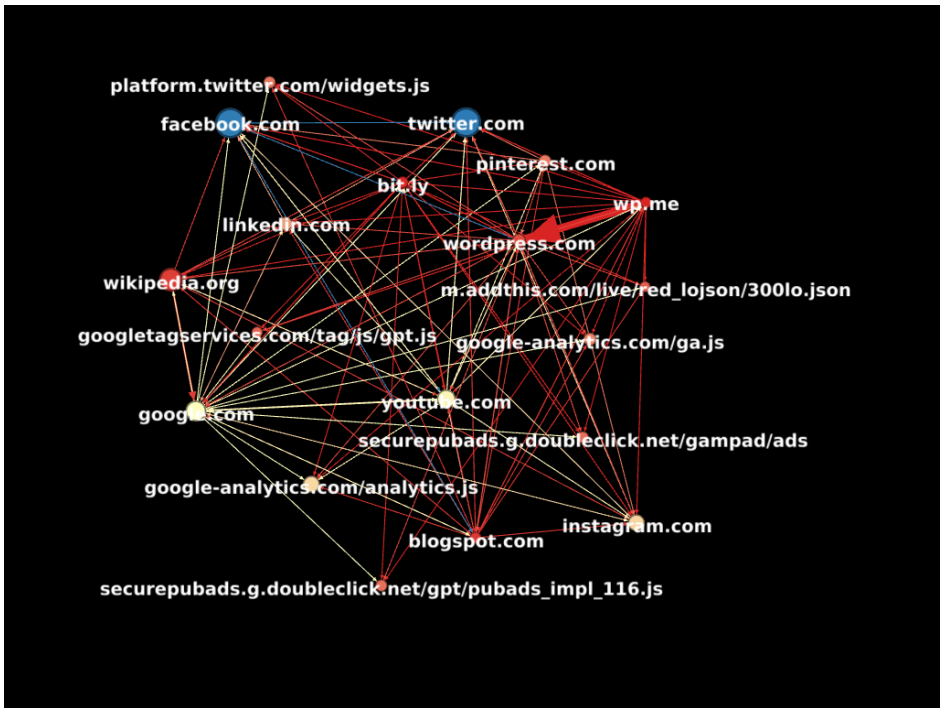

Figure 3.5: Total graph, in which a node is either a domain, cookie source or script source.
Edges represent hyperlinks or usage of the code source on the given domain.
Minimum degree has been set to 5000 as there are twice as many 'popular' nodes as in the previous graphs.

3.2 Community detection

As the aim is to find interesting clusters of domains in the context of tracking code, it is important to specify when a certain domain belongs to a certain cluster. This problem is known as *community detection*, and several approaches exist to divide a graph into these communities [12, 10, 7]. The following three categories will be considered for use in this project:

- *Hierarchical clustering* is a class of algorithms that iteratively split or merge clusters, depending on whether the chosen algorithm is divisive or agglomerative. They work using a similarity (also called distance) metric to determine which clusters should be merged or split. A popular graph algorithm is *Girvan-Newman*, which starts off with the biggest possible clusters, and iteratively removes edges[1] based on *betweenness*, which measures the amount of node pairs in the graph of which the shortest path passes through this specific edge. The algorithm usually stops when the clusters have reached maximum *modularity*[2], which measures the quality of a cluster, and has a complexity of $O(N^3)$ [10].

- *Partitional clustering* is a class of clustering methods that optimise a predefined amount of clusters. Nodes are assigned to the clusters such that the distance between the clusters is as large as possible, meaning that nodes inside one cluster have a low similarity with nodes from another cluster. An interesting algorithm in the context of this thesis is *k-clique-percolation*, which finds adjacent *k-cliques*. K-cliques are clusters of k nodes that are fully connected, and adjacent k-cliques then form a community together. Using this method, it is possible for a node to be part of multiple overlapping communities, and different levels of hierarchy can be detected by varying the value of k [12].
  As mentioned in [10]: "*The complexity of this procedure can be high, as the computational time needed to*

[1] Therefore, it is a divisive algorithm.
[2] Making this a modularity optimisation as well as hierarchical clustering method.

find all k-cliques of a graph is an exponentially growing function of the graph size [31], but in practical applications the method is rather fast, enabling one to analyze systems with up to 10^5 *nodes."*

- *Modularity optimisation* is a problem that tries to reach the optimal *modularity* for each cluster which, as mentioned previously, assesses its quality.
  "In the case of weighted networks, it is defined as [12]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

  where A_{ij} *represents the weight of the edge between* i *and* j, $k_i = \sum_j A_{ij}$ *is the sum of the weights of the edges attached to vertex* i, c_i *is the community to which vertex* i *is assigned, the* δ*-function* $\delta(u, v)$ *is* 1 *if* $u = v$ *and* 0 *otherwise and* $m = \frac{1}{2} \sum_{ij} A_{ij}$*."[3]*
  A popular algorithm from this class is the *Louvain Modularity* algorithm, which begins with each node as a separate cluster[3], assigning them to the neighbouring cluster that results in the highest gain in modularity. When no more moves are possible that increase the modularity, a new graph is constructed with a single node for every cluster, and the process is repeated. The key feature of this algorithm is that it uses a very efficient modularity computation method, resulting in a significantly faster clustering process [3].

Perhaps the most appropriate community detection algorithm would be the clique percolation method. After all, it should be possible for a website to be part of multiple different clusters, i.e. a science news website could well be part of both some cluster of news websites and a cluster of science websites. Unfortunately, it is computationally expensive when run on a graph of several hundred thousands of nodes and edges. Furthermore, the k parameter needs to be experimented with, further increasing the required time. Due to the limited time available for this thesis project, it has shown not to be feasible to use this algorithm for the community detection. From the other two classes, the most popular algorithms happen to both be hierarchical and optimise modularity. This is a useful combination, because it removes the need to speficy appropriate stopping criteria. While *Girvan-Newman* has a complexity of $O(N^3)$ and therefore takes a long time to compute on the described graphs (as each contains several hundred thousand nodes), the Louvain Modularity is incredibly quick, and running it graphs of this size is only a matter of seconds [3]. At the same time, it outperforms other algorithms in terms of reached modularity as well. While it limits nodes to be part of only one community, it is the most appropriate algorithm to use given the scope of this project. For future research, it would be interesting to explore other community detection algorithms as well.

To use Louvain Modularity, the graphs have been converted to undirected graphs: every pair of edges is merged into one edge with their summed weights. While Louvain Modularity can theoretically be modified to support directed graphs [5], the python-louvain package used in this project does not support directed networks, and writing our own implementation of the algorithm would exceed the timescale of this project. However, this undirectionality should not negatively effect the quality of the resulting clusters; in fact, many of the domains present have not been crawled and will therefore be registered as having no external hyperlinks. This affects many properties of the graph like connectivity, for example because no path will be possible between two of such fringe nodes. Removing the directionality solves this problem, and as the weights are updated correspondingly, only a minimal amount of data is lost.

So far, three graphs have been described, each representing the collected data in a separate way. While the amounts and types of nodes and edges differ per graph as they have different structures, this is not problematic for the community detection. However, to calculate meaningful metrics that will give insight into the differences between these graphs, it is desirable to keep as many factors constant as possible. To this end, a fourth graph will be constructed, henceforth referred to as the *induced graph*. This induced graph will contain exactly one node for every domain, and none for the cookie and script sources. An undirected edge will exist between two domains if a hyperlink is present between the two or if both domains use the same tracking code source, resulting in three different types of edges.
To create this graph, a very small portion of the data has been ignored. If a link is created between every of the n

[3] Therefore, it is also an agglomerative hierarchical clustering algorithm.

neighbours of a cookie or script node, this creates a fully connected sub-graph with n nodes. As a fully connected graph contains $\frac{n(n-1)}{2}$ edges and the biggest script node has a degree of 14,117, this node alone would create 99 million edges in the induced graph. To ensure that the graph will not become too big to use in the computations, all cookie and script nodes with a degree of 500 or higher have to be filtered out, removing a total of 162 cookies and scripts. This is problematic, because especially the cookies and scripts with a high degree are likely to play a big part in establishing a community. In the case of a cluster existing solely around a high-degree cookie or script, this community will be largely disconnected in the induced graph. This is really not ideal, but will have to do for this thesis. Luckily, it only affects the metrics, not the structure of the detected communities.

3.2.1 Hyperlink community graph

Using Louvain Modularity on the previously created hyperlink graph, 671 communities are detected with a modularity of 0.697. Constructing a new communication graph from these communities results in the graph displayed in figure 3.6.

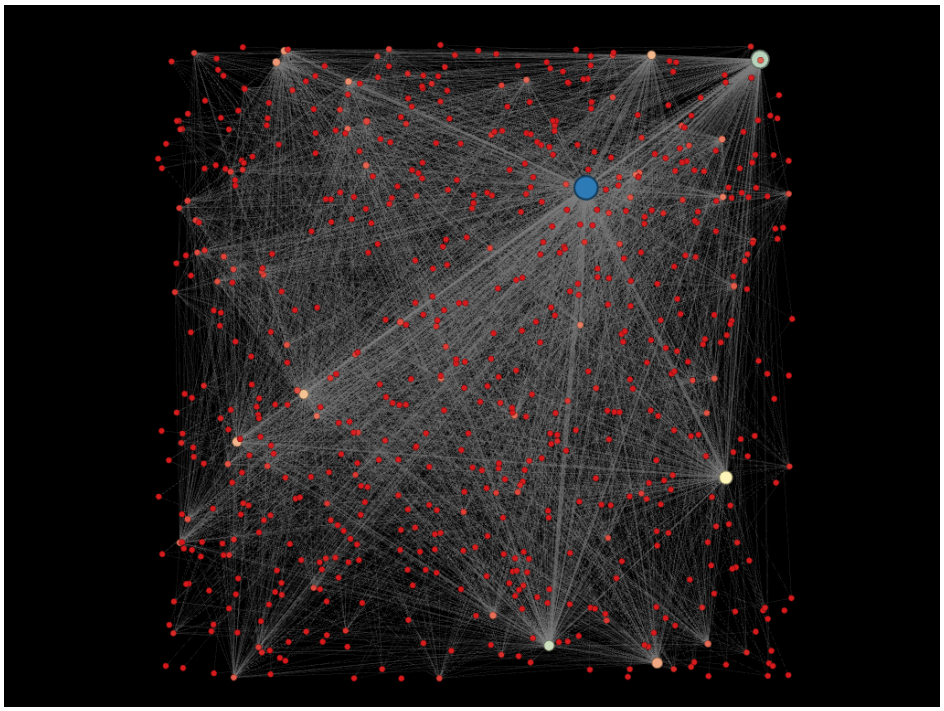

Figure 3.6: Communities detected in the hyperlink graph.
Size corresponds to the amount of domains in the community, and colour is determined by degree. Edge colour is not dependent on any attributes.

Although there is no overlap in the domains within different communities, the figure shows that there are many edges between them, with the average degree being 13.24. As figure 3.7 depicts, the degree of a node and its size (the amount of domains belonging to the community it represents) seem to be heavily correlated. Although they are not far off, they are not quite power law distributions, meaning the communities do not form a scale-free network, contrary to what one might expect. As scale-free networks often consist of several *hub* nodes with high degree that are connected to many nodes with low degree, this description would fit figure 3.6, but it does make sense. If the network of communities would consist of many hubs, we might as well replace every hub and its surrounding nodes with a single node, representing a new community. Because the Louvain Modularity algorithm has optimised the graph modularity, it follows automatically that it is likely not beneficial to further group any of these nodes into a new community.

Because a silo has been defined *a cluster of heavily interlinked websites with very few outward links*, perhaps the

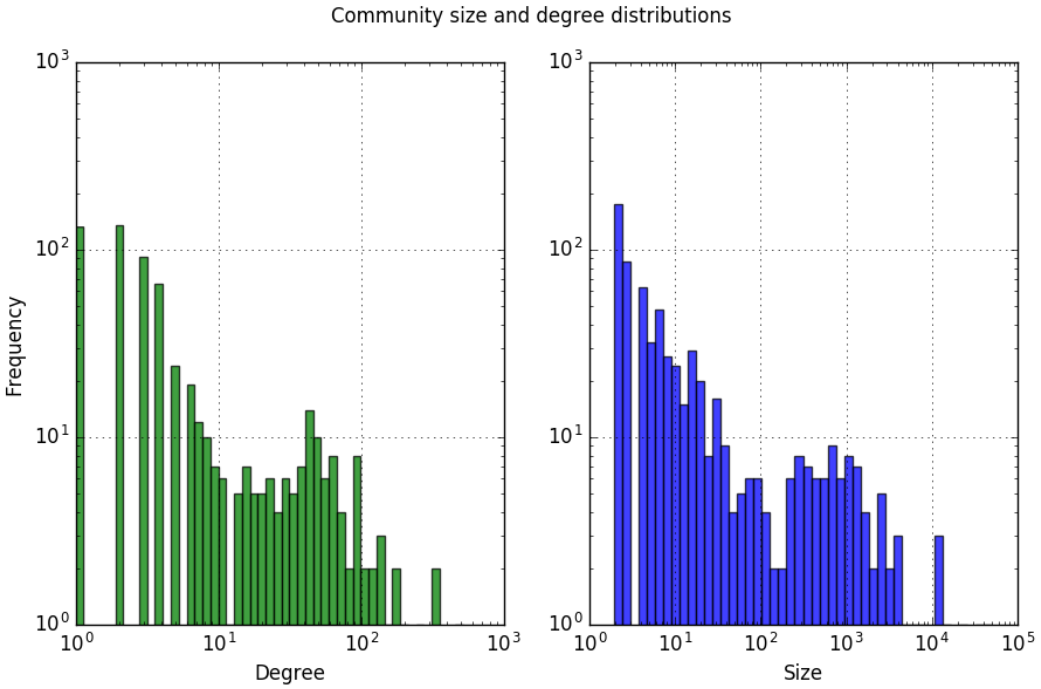

Figure 3.7: Hyperlink community size and degree distribution.
Average degree: 13.26
Average size: 387.46

Density:	0.59
Size:	7.98
Average degree:	2.35
Average connectivity:	1.38

Table 3.1: Average sub-graph metrics of hyperlink communities

most interesting nodes are those with a low degree, thus relatively isolated from the rest of the network[4]. If all nodes with an above average degree are filtered out, the network becomes fully disconnected (i.e. there are no edges), leaving just the isolated nodes. However, as isolation is only one of the silo criteria, let's have a look at some other properties of these communities, displayed in figure 3.8 and table 3.1.

For all of the communities, the following properties have been measured:

- Density, which is the ratio between the actual amount of edges and the amount of possible edges
- Size, which is the amount of domains in the community.
- Average degree, which is the average amount of incoming and outgoing edges per node.
- Average connectivity, which is calculated as follows: for every pair of nodes A and B in the graph, calculate the minimum amount of nodes that need to removed from the graph in order to ensure that there is no path between A and B [2].

[4]More importantly, communities with a high degree are often also very large, making it impossible to carry out a manual analysis. For completeness, communities with a higher degree are also briefly explored later.

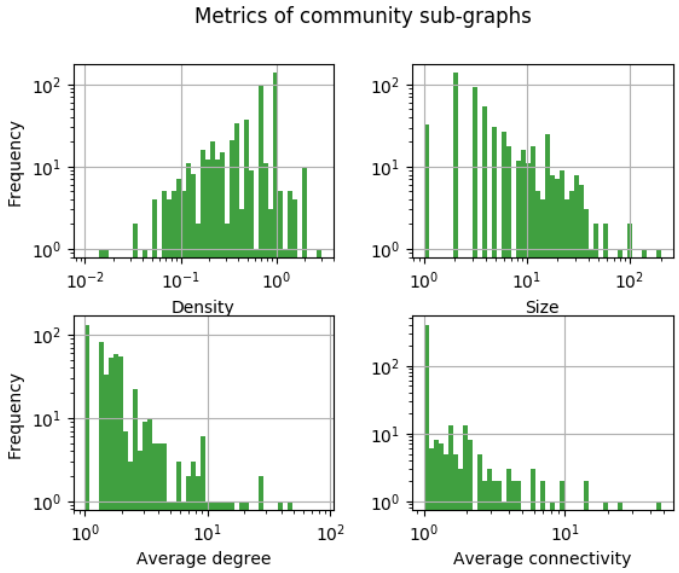

Figure 3.8: Distributions of sub-graph metrics for hyperlink communities with below average degree, calculated on the induced graph.

For a community to qualify as a silo, its websites need to be *heavily interlinked*. Therefore, it makes sense to consider the density of the communities to define heavy linkage; after all, density directly measures the amount of links in a graph.
If all communities with less than average density are filtered out, 300 nodes remain. However, many of them consist of just two websites, and will not be interesting communities. Therefore, let's redefine the definition of a silo: *a cluster of websites that, compared to other clusters, has a less than average degree, consists of at least four websites, and has an above average density.* This leaves us with the nodes displayed in figure 3.9.

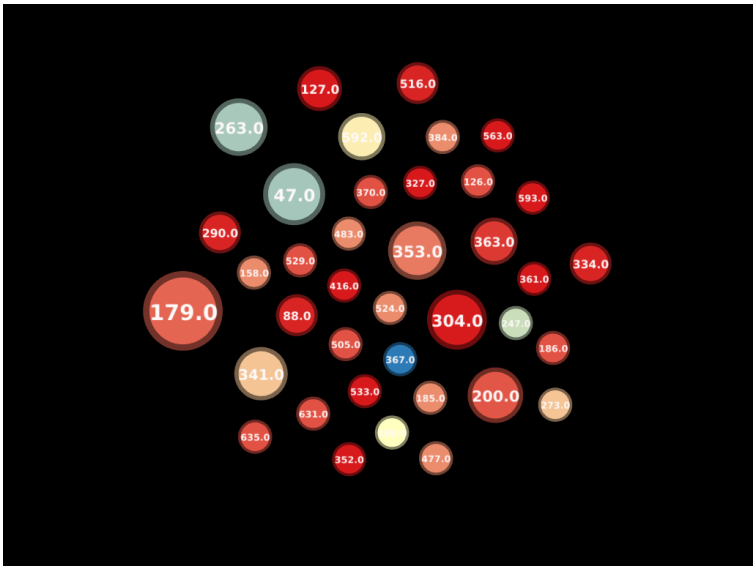

Figure 3.9: Detected silos from the hyperlink graph with below average degree, above average density and at least four websites. Size corresponds to the amount of websites, colour corresponds to density. The node label is the silo ID.

Silo inspection

The most interesting silos are those high in both size and density, of which the silos 47 and 263 are prime examples. Upon further inspection, 47 is not that interesting at all; it simply contains 21 domains of `blablacar` with a different top-level domain. Silo 263, however, does show interesting results. It contains the following domains:

```
"fresh.co.nz",       "bravotv.co.nz",
"thebreeze.co.nz",   "threenow.co.nz",
"radiolive.co.nz",   "thesound.co.nz",
"magic.co.nz",       "morefm.co.nz",
"therock.net.nz",    "theedge.co.nz",
"newshub.co.nz",     "mediaworks.co.nz",
"maifm.co.nz",       "3news.co.nz",
"georgefm.co.nz"
```

While all websites share an obvious property, they are not solely connected because of their top-level domain. The crawler has gathered data of 574 `.co.nz` domains, and a total of 837 `.nz` domains, and this silo only contains 15 of them. Let's have a look at the sub-graph, displayed in figure 3.10. The network is not strongly connected, but does contain a large strongly connected component, that is in fact fully connected. Interestingly, `3news.co.nz` links to all other websites, but is never linked to, and `bravotv.co.nz` and `fresh.co.nz` are linked to by all other websites, but do not link to any others.
On manual inspection of the data, the crawler did not detect any hyperlinks on `fresh.co.nz`, and simply did not crawl `bravotv.co.nz`, which explains why they do not have any outbound links. `3news.co.nz` is a news website, so it makes sense that it has many outbound links without necessarily having any incoming edges from the other websites. However, `newshub.co.nz` is also a news website, but has incoming edges from all of the purple nodes. It seems that for some reason, the websites prefer `newshub.co.nz` over `3news.co.nz`.

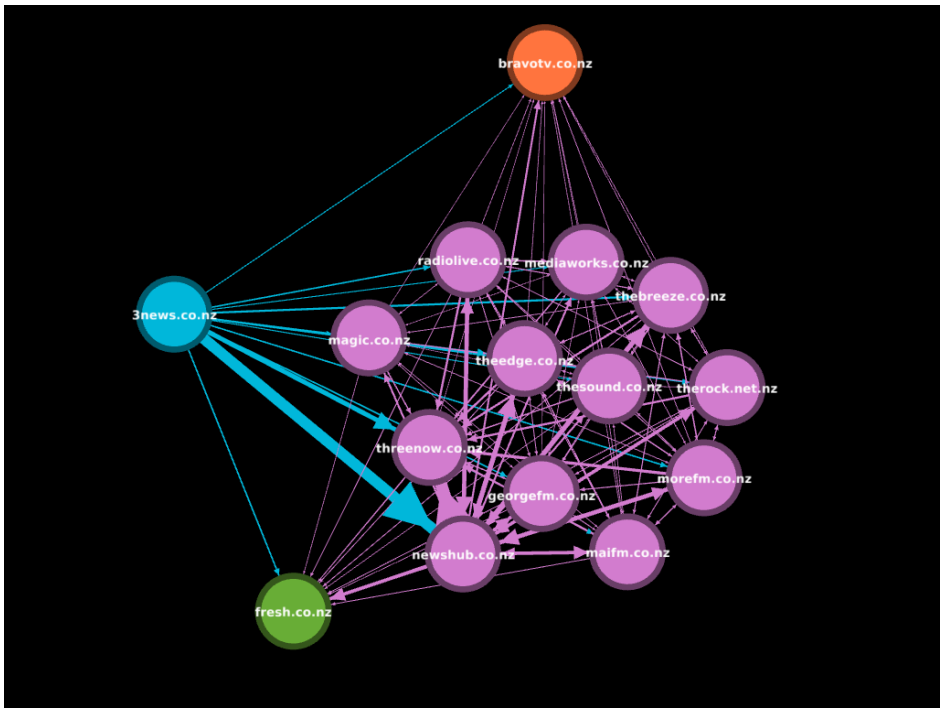

Figure 3.10: Hyperlink silo 263
Node colour corresponds to the strongly connected sub-network it is part of.

No in-depth analysis of the other silos in figure 3.9 will be provided, but some of their graphs are displayed in the *Hyperlink silos* section of the appendix. Note that all communities and sub-communities have been detected by Louvain Modularity, but their layout is processed by the Fruchterman-Reingold algorithm for improved clarity. All of them definitely are or contain silos, but most are either very small (less than 10 websites) or mainly consist of different websites from the same company or organisation. Furthermore, collecting seed URLs from the selected subreddits has not resulted in any obvious corresponding communities. Let's see what happens when the density filter is discarded, and we simply look at communities with at least 20 websites. This threshold should be high enough to prevent a few domains from the same organisation or company from showing up as a community, but is low enough to still yield a significant amount of communities. This results in a graph similar to 3.9, but with many edges between the nodes (visualised in appendix 6.3).

There are 114 nodes that meet these criteria and are therefore potentially interesting silos. Because there are too many of them to analyse them all, a few will be manually selected for inspection. Looking at cluster 79 as displayed in figure 3.11, it is a weakly connected network. With a total of 309 nodes, there are 213 different strongly connected components, most of which consist of just one node without any outbound connections, likely because it hasn't been crawled. As the prominent central blue network is the connecting factor between these nodes, it would not be considered a silo according to any of the previous definitions; after all, it has many outbound links. However, if (more pages of) the connected domains had been crawled, they might well have links back into the network, therefore no longer qualifying as outbound links, and the network would indeed be considered a silo.
This is a difficult issue, because one can never crawl the full internet, therefore always leaving fringe nodes that might not actually be fringe. A possible solution would be filtering out all hyperlinks to domains that have not been crawled, but that would significantly reduce the amount of useful data. After all, when many websites in a community link to a specific domain, that is a strong indication that that domain might also belong to the same community.

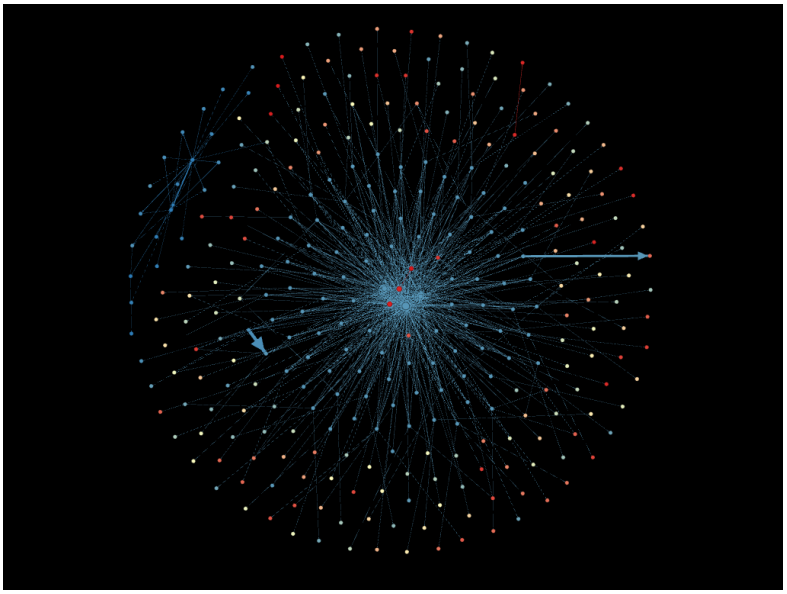

Figure 3.11: Hyperlink cluster 79
Node colour corresponds to the strongly connected sub-network it is part of.
Communities are detected by Louvain Modularity, layout is defined by the Fruchterman-Reingold algorithm.

For the purpose of this research, let us just ignore this issue and look at the core of this network. As figure 3.12 shows, it mainly consists of English news websites, which definitely form a community, albeit not a silo by itself. More complex structures exist, for example in cluster 69, displayed in figure 3.13. There are clearly multiple

Figure 3.12: Central network of hyperlink cluster 79

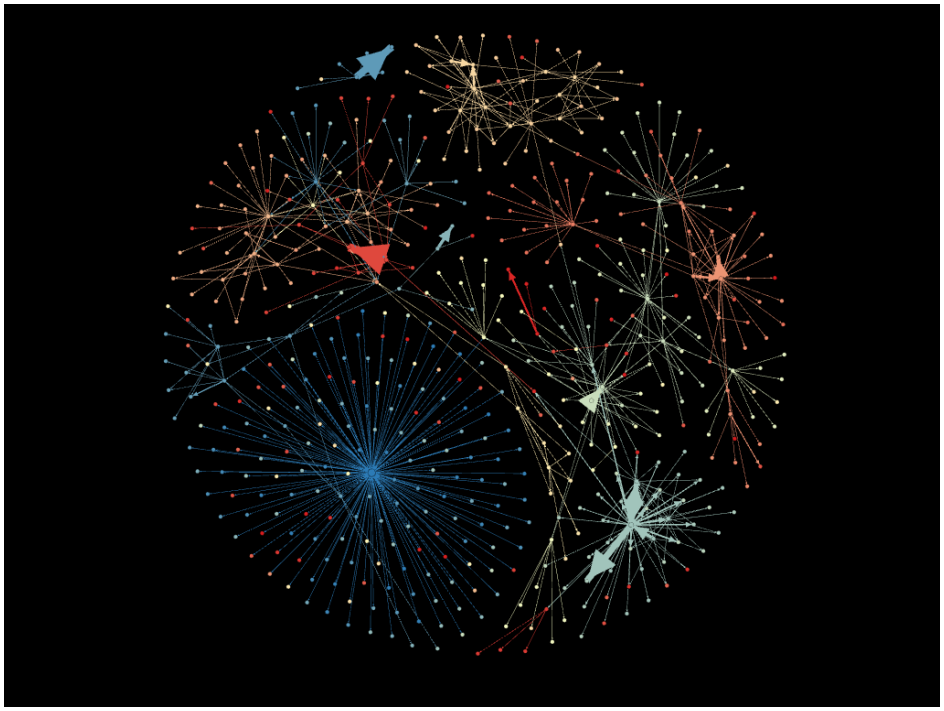

Figure 3.13: Hyperlink cluster 69

separate clusters within this graph, and although they are not strongly connected, they definitely represent different communities. Upon manual inspection, it seems to be divided in the following communities:

- German martial arts
- Korean lifestyle
- Korean news
- Vietnamese news
- Vietnamese tourism
- International martial arts
- Turkish government and news websites
- Websites specifically about Istanbul

It is quite difficult to manually separate the graph into the respective communities, but running Louvain Modularity once again finds 8 communities with a modularity of 0.814, which is quite accurate. Some communities that had been manually identified separately, such as Vietnamese news and Vietnamese tourism, have been detected as one big community. This is not necessarily wrong, as there are edges in between the two, so it is difficult to determine whether they should be considered separate communities or not.
These results show that although the quality of the community detection seems poor at first, it would be greatly improved by adding a level of recursion. As Louvain is in fact a hierarchical clustering method, these nested clusters have actually already been identified separately by the algorithm. However, because it is difficult to visualise nested communities in a graph, the hyperlink community graph only shows the highest level of the hierarchy. The exact results of the community detection can be found in the appendix section *Nested communities*.

3.2.2 Tracking community graph

Using Louvain Modularity on the previously created tracking graph, 4169 communities are detected with a modularity of 0.616. Constructing a new communication graph from these communities results in a graph similar to figure 3.6. However, where the hyperlink communities have an average degree of 13.26, the tracking communities have an average degree of only 0.93, with 86.78% of the nodes being completely isolated.
As figure 3.14 depicts, size and degree follow roughly the same distribution as the hyperlink communities, but there are many nodes with a size of zero or one. Because the average degree is already so extremely low, it does not make sense to only look at communities with a below average degree, as that would only leave the completely isolated communities. Instead, let's use the average degree from the hyperlink communities, 13.26, as the threshold. This filters out only 1.15% of the communities, and all the nodes that remain can be considered fairly isolated (i.e. *very few outbound links*).

Graph:	**Hyperlinks**	**Tracking**
Density:	0.59	0.08
Size:	7.98	1.08
Average degree:	2.35	0.27
Average connectivity:	1.38	0.18

Table 3.2: Average sub-graph metrics of tracking communities.

The average properties of these isolated nodes are displayed in table 3.2, with their distributions displayed in figure 3.15. It is immediately clear that all of the metrics are significantly lower, which makes sense considering that many communities consist of just one domain.

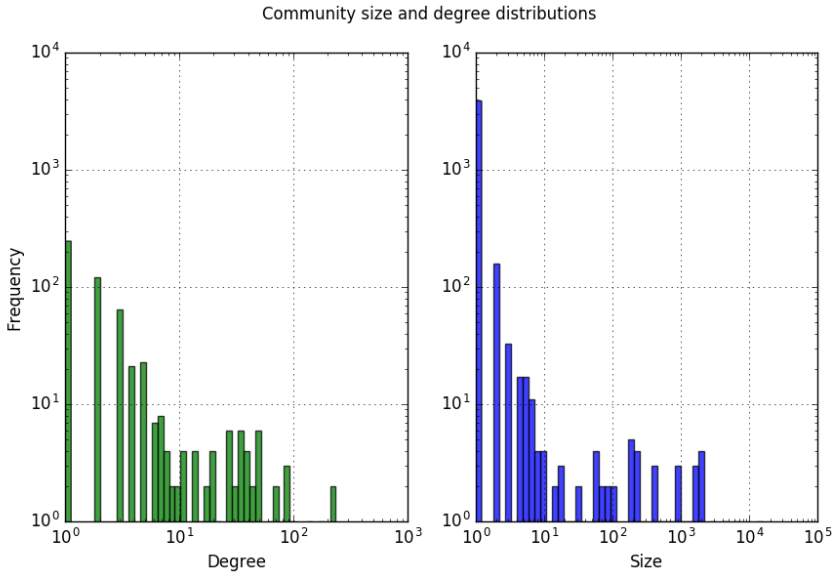

Figure 3.14: Tracking community size and degree distributions.
Size refers only to the amount of domains in a node, not the amount of cookies and scripts.
Both axes use logarithmic scale.
Average degree: 0.93
Average size: 13.74

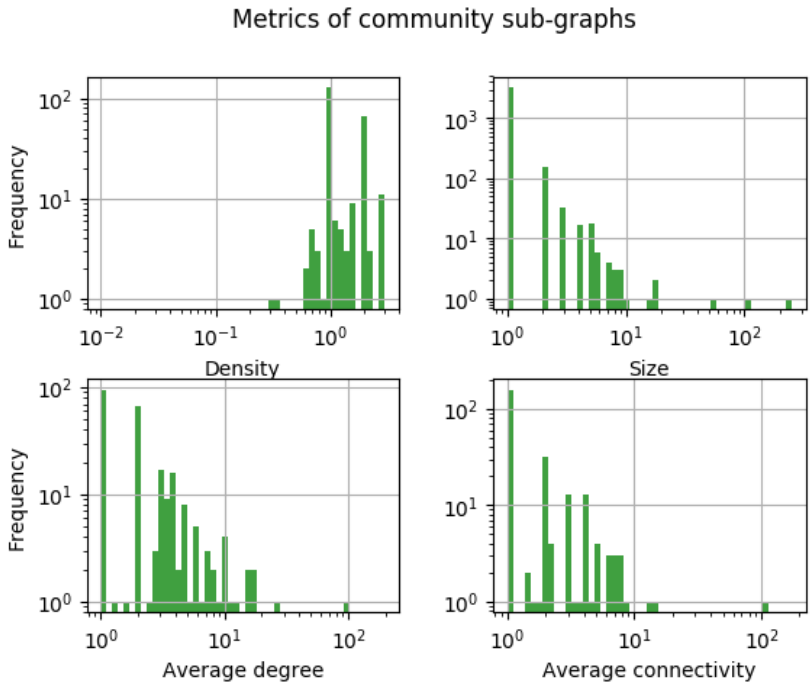

Figure 3.15: Distribution of sub-graph metrics for tracking communities with degree below 13.26, calculated on the induced graph.

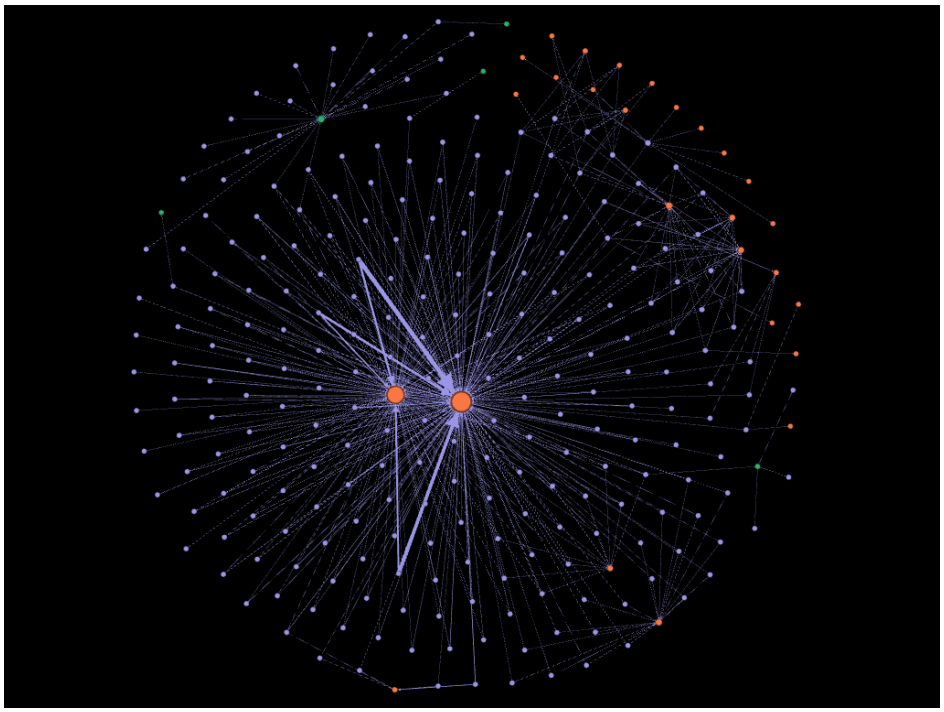

Figure 3.16: Tracking silo 212
Different colours represent node types. Blue is a domain, orange is a script, green is a cookie. Node size corresponds to degree.

Filtering out communities with less than 20 websites and a degree higher than 13.26, only three nodes remain. The biggest cluster, visualised in figure 3.16, contains 248 domains, 5 cookies and 27 scripts.
The graph is weakly connected; every node is a separate strongly connected component. This makes sense, because it is a directional graph and the cookies and scripts only have incoming edges, so no strongly connected components are possible.
While there are a few clusters, most of the websites in this graph are only linked because they use both of the centermost script nodes, being `img.sedoparking.com/js/jquery-1.4.2.min.js` and `google.com/adsense/domains/caf.js`.

Running Louvain Modularity detects 11 clusters with a modularity of 0.29, which is very low. Five of those clusters contain more than three domains, and three of those clusters are linked because of a jQuery version. The remaining two clusters are linked by respectively `parkingcrew.com` and `sarah.tnctrx.com/tr`, which describe themselves as a domainparking platform and a traffic marketplace respectively. Upon inspection, both seem to serve the same purpose: visiting the website either redirects the visitor to an advert, or simply notifies them that the domain is not yet in use.

The second biggest node of the filtered community graph, visualised in figure 3.17, contains 103 domains and 12 cookies, all of which are from `static.hugedomains.com`. While Louvain detects four communities, the modularity is only 0.015, and the graph cannot actually be separated into different clusters, because most domains are linked to all the cookie sources.
The cookie sources seem harmless utilities, but it is suspicious that they set cookies. No conclusions will be drawn, but it is an interesting structure.

The third and last remaining silo contains 55 domains, 20 cookies and 205 scripts, and can be divided into ten separate communities using Louvain, resulting in a modularity of 0.761, which is reasonable. Only four of these communities contain more than three websites, and all of the communities are domains linked by cookies and

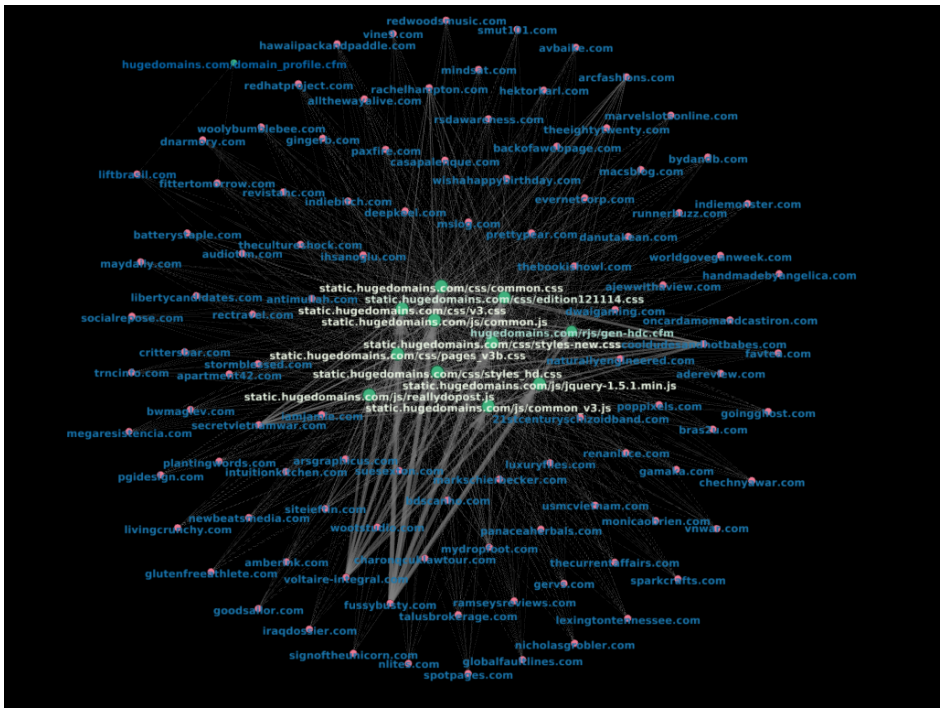

Figure 3.17: Tracking silo 114
Different colours represent node types. Green is a cookie, blue is a domain. Size corresponds to degree, label colour is just for readability.

scripts from `wsimg.com`, which is a domain used by hosting website `godaddy.com`. The situation is very similar to the previous example; various websites make use of utilities hosted on those servers. This time, however, all the utilities are scripts rather than cookies, which is not suspicious. Therefore, no visualisation of this graph will be provided.

In conclusion, linking domains together solely by the cookies and scripts they use proves to be ineffective in detecting interesting tracking code networks. Crawling roughly 87,000 domains has resulted in only three proper (i.e. 20 or more domains) silos, and only one of those has an interesting structure. Even in that case, it is likely that there is a simple explanation for the fact that these utilities show up as cookies.

3.2.3 Total community graph

Running Louvain Modularity on the total graph, which incorporates both hyperlinks and potential tracking code, results in 453 detected communities and a modularity of 0.638, which is slightly higher than the tracking communities and slightly lower than the hyperlink communities. Again, the resulting graph is similar to that in figure 3.6, and is not interesting to visualise.

Although the size and degree (figure 3.18) distributions still look similar, the average degree has increased by 35% while the average size has increased by 48% compared to the hyperlink communities. However, as shown by table 3.3, the average size of the isolated communities has decreased, meaning nodes have shifted towards communities with a higher degree.

Interestingly, as shown by figure 3.19, there are communities with an average connectivity of less than one, meaning they contain isolated nodes. This is unexpected, because if there are no links between a node and any other node in a community, they have no reason to be part of the same community.

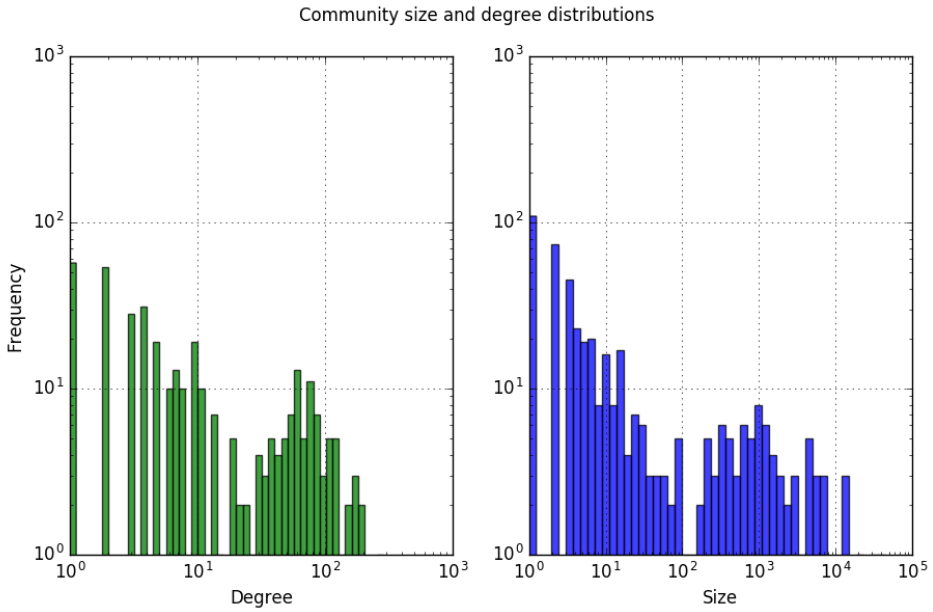

Figure 3.18: Total community size and degree distribution.
Size refers only to the amount of domains in a community.
Average degree: 17.95
Average size: 572.77

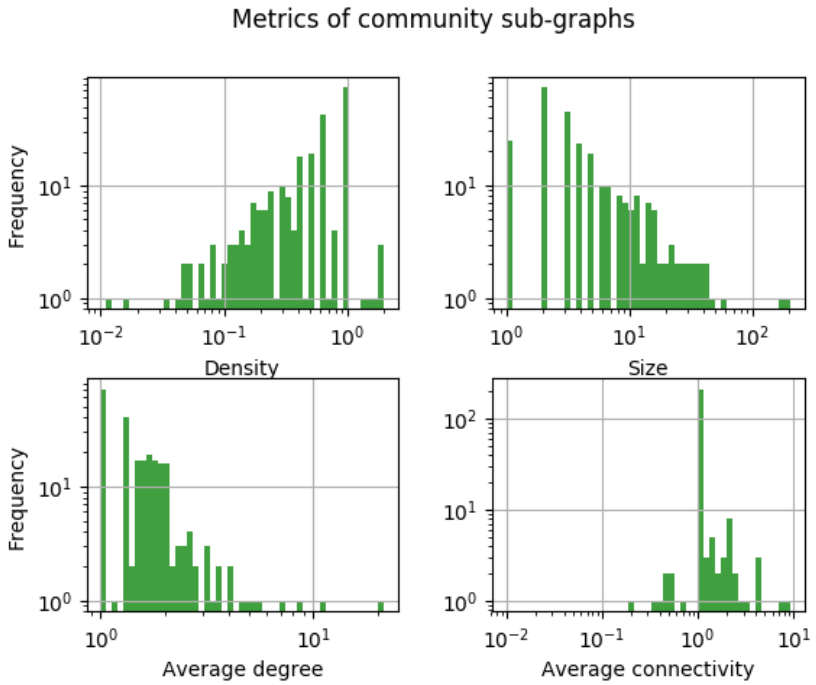

Figure 3.19: Distribution of sub-graph metrics for total communities with below average degree, calculated on the induced graph.

Graph:	Hyperlinks	Tracking	Total
Density:	0.59	0.08	0.42
Size:	7.98	1.08	5.98
Average degree:	2.35	0.27	1.61
Average connectivity:	1.38	0.18	0.81

Table 3.3: Average sub-graph metrics of total communities.

Figure 3.20: Silo 40 as extracted from the induced graph (left) and the total graph (right).

As seen in figure 3.20, this is not a mistake; the community does indeed contain three different sub-networks that are not connected in any way. A plausible explanation might be that grouping these nodes together resulted in an increase in modularity of another cluster that is very densely connected. By removing the separate cluster from the dense cluster, its modularity might increase more than the modularity of the disconnected community decreases, because it was already low to begin with. However, this is just a guess, and the exact reason remains unclear.

Filtering out all the communities with an above average degree leaves 357 nodes with no edges between them, 20 of which contain at least 20 domains. The biggest of those nodes is a silo that revolves around the company 3M (`3m.com`) and its range of products. As it is a big company with products in many different categories, many of which have their own website, this results in quite a large silo. While it does contain a few scripts and cookies, this silo is also present in the hyperlinking communities, so nothing new has been detected here.
The second biggest one is a simple structure: the website `harisingh.com` and all the other domains it links to. None of the sites it links to has any outbound links, and the silo does not contain any cookies or scripts. This silo is also present in the hyperlink communities, but is in a sub-cluster there rather than a separate silo. It is perfectly detectable by the recursive Louvain approach (i.e. it has been detected separately, but was merged later), but apparently adding cookies and scripts has 'untangled' that network into separate silos, removing it from its parent community into the highest level of hierarchy.

Somewhat similarly, the cluster in figure 3.21 was also part of a bigger silo in the hyperlink graph, but has been detected as a separate community because of its cookies and scripts. Most likely, many of the scripts and cookies displayed here (mostly from `burstnet.com` and `bravenet.com`, both hosting websites) are specific to this silo, strengthening the cohesion of the silo enough for the community detection algorithm to detect it as a separate community.
Likewise, another detected cluster that revolves around `odebrecht.com.br` has been detected separately from its surrounding silo because the sites all use cookies and scripts hosted on `odebrecht.com`, which is obviously owned by the same entity.

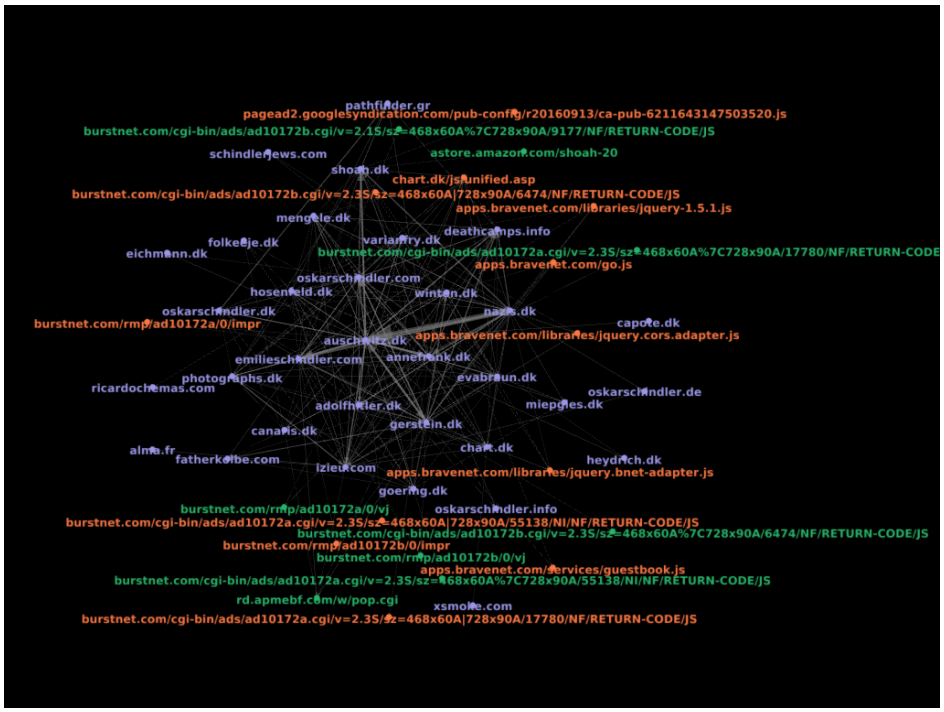

Figure 3.21: Holocaust-themed cluster of domains and tracking code sources. Domains are blue, cookies are green, scripts are orange.

Interestingly, one of the detected silos is only a fraction of its bigger hyperlink counterpart, which has been divided into multiple silos. The original silo, shown in figure 3.22, contains multiple sub-communities as identified by Louvain. For readability, not all nodes have been labeled, but the network consists solely of websites from the UK, a big part of which are English news websites. Using the total graph has split up the silo into different parts, two of which shown in figure 3.23 and 3.24. The first consists mainly of news websites from Wales and Scotland (which are nested communities and have been detected separately by the algorithm, but were merged later), and the second contains news websites from around West-England and London (to which the same applies). Again, this points out that using only the top-level of the community hierarchy as detected by Louvain does not necessarily yield the most interesting results. Many more different communities can be identified by also inspecting lower layers of the hierarchy, as will be confirmed later.

Interestingly, one of the silos was originally part of a bigger silo (although it had been detected separately before being merged) but the new silo, displayed in figure 3.25, has gained two new domains that were not present in the same parent silo. This is unexpected, because the only shared connection between the two clusters is a domain, and therefore also present in the hyperlink graph. It is unclear why this happens; several re-runs of the community detection result in the exact same cluster every time.

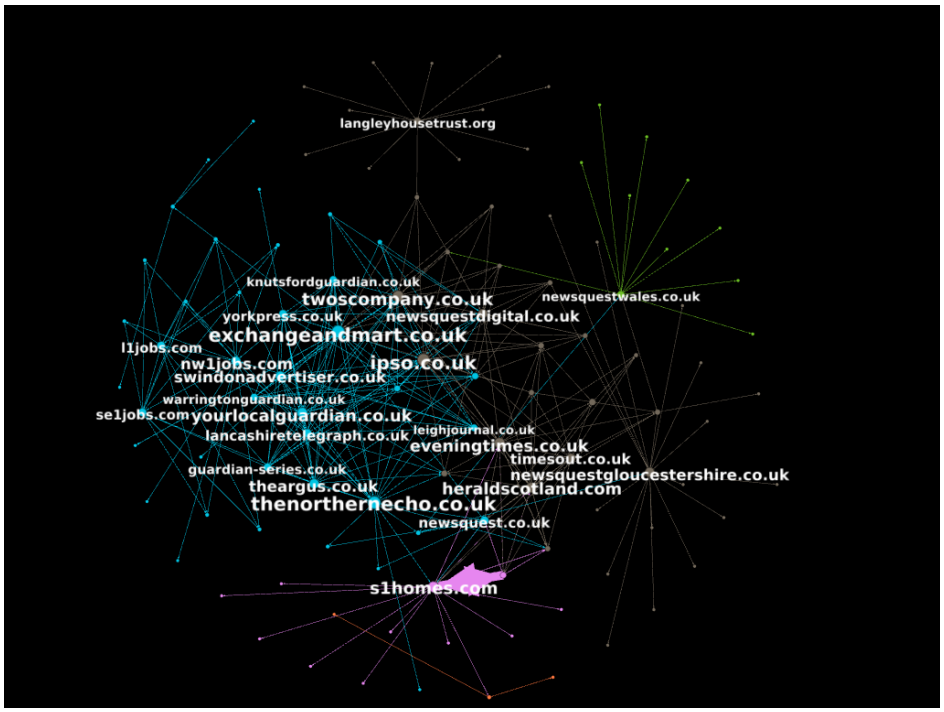

Figure 3.22: Silo of UK websites.
Different colours represent different detected sub-communities.

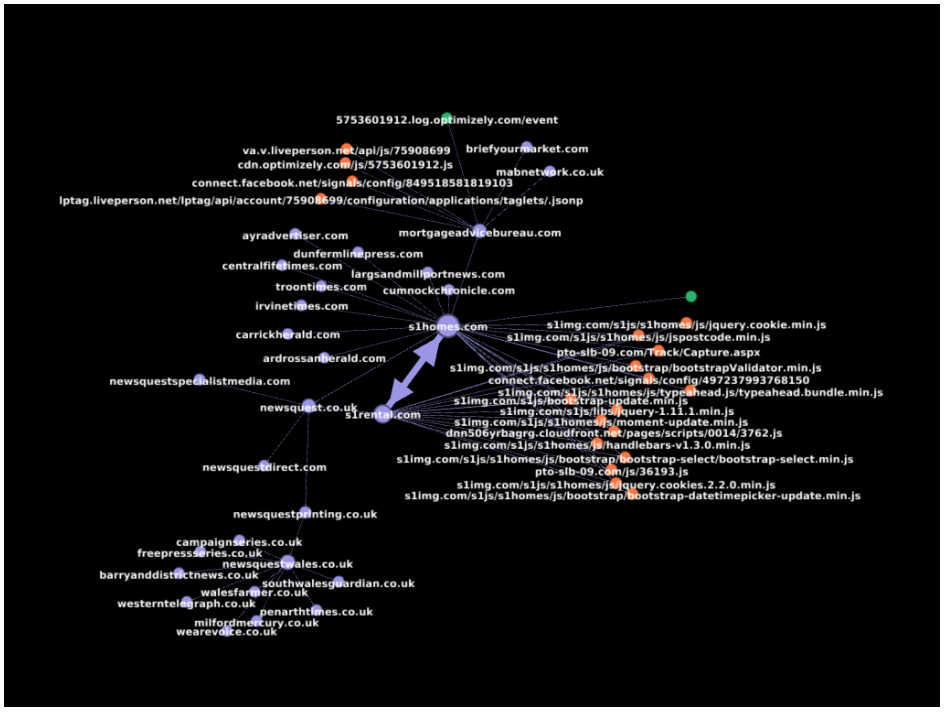

Figure 3.23: Part of the UK websites silo, detected as an individual silo using the total graph.
Domains are blue, scripts are orange, cookies are green.

Figure 3.24: Another part of the UK websites silo, also detected as an individual silo when using the total graph. Domains are blue, scripts are green, cookies are orange.

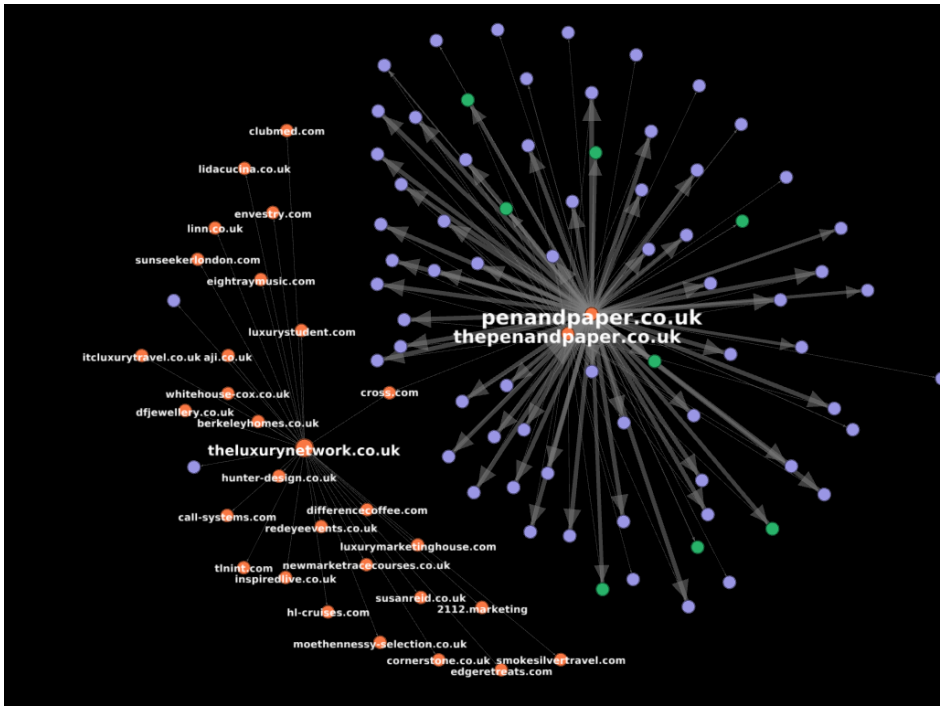

Figure 3.25: The two different clusters are linked solely by hyperlinks, yet they are only clustered together when using both hyperlinks and tracking code data. Domains are orange, scripts are blue, cookies are green.

The remaining fifteen silos are distributed as follows:

- Six of the silos are exactly the same in the total graph as in the hyperlink graph (i.e. they do not contain any scripts or cookies).
- Five of the silos contain exactly the same domains, but now also include cookies and scripts.
- Three silos were previously part of a bigger silo, from which they can successfully be extracted with another iteration of community detection (although one required a resolution of six instead of one). In the total graph, these are already separated into individual silos.
- From one of the silos, a domain has disappeared because although it is not connected to any domains outside of this silo, it uses many popular cookies and scripts.

3.2.4 Communities with higher degree

So far, it seems that including the tracking data in the community detection does not improve the quality of the communities in ways that are not possible using solely the hyperlinks. In fact, it even caused some strange behaviour: websites that are not connected in any way (neither domains nor scripts or cookies) sometimes end up in the same cluster. However, only the communities with a below-average degree have been inspected so far, so let's look at some of the higher degree communities. Unfortunately, any community with more than 5000 domains will be too big to visualise and manually inspect, so those will still have to be filtered out. Excluding the twenty communities with a low degree that were already inspected, this leaves 73 communities in the total graph.

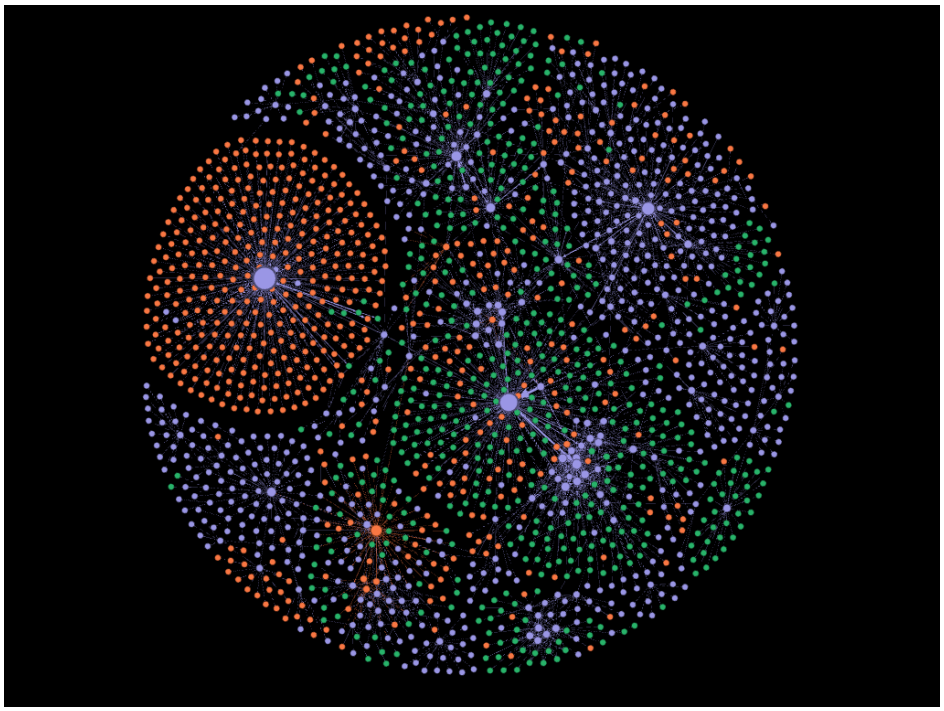

Figure 3.26: Silo 63 of the total graph, contains roughly 2000 domains. Orange nodes are cookies, domains are blue and scripts are green. Node size corresponds to degree.

As figure 3.26 shows, silo 63 is a big silo of 2000 domains, containing multiple sub-communities. It is hard to display such details because there are so many nodes, but all of the communities are solely connected by hyperlinks.

Each of the communities has its own *hub*, the core that connects the surrounding nodes. One might expect that if the same tracking network is deployed across different sites, many websites link to the same (group of) tracking code. Therefore, there should be two possible community structures that correspond to such a situation:

- A group of nodes surrounds a cookie or script node, which is therefore the hub of the community.
- A group of nodes surrounds a group of cookie or script nodes, which therefore form the hub of the community together.

As figure 3.26 shows, one of the cookie and script nodes is indeed a hub; the big orange node in the bottom left of the picture. However, this is partially a false alarm; it is actually both a website and a cookie source. Its degree is 107, but 92 of those edges are actually outbound hyperlink edges. Still, it is interesting to see the network surrounding it, displayed in figure 3.27.

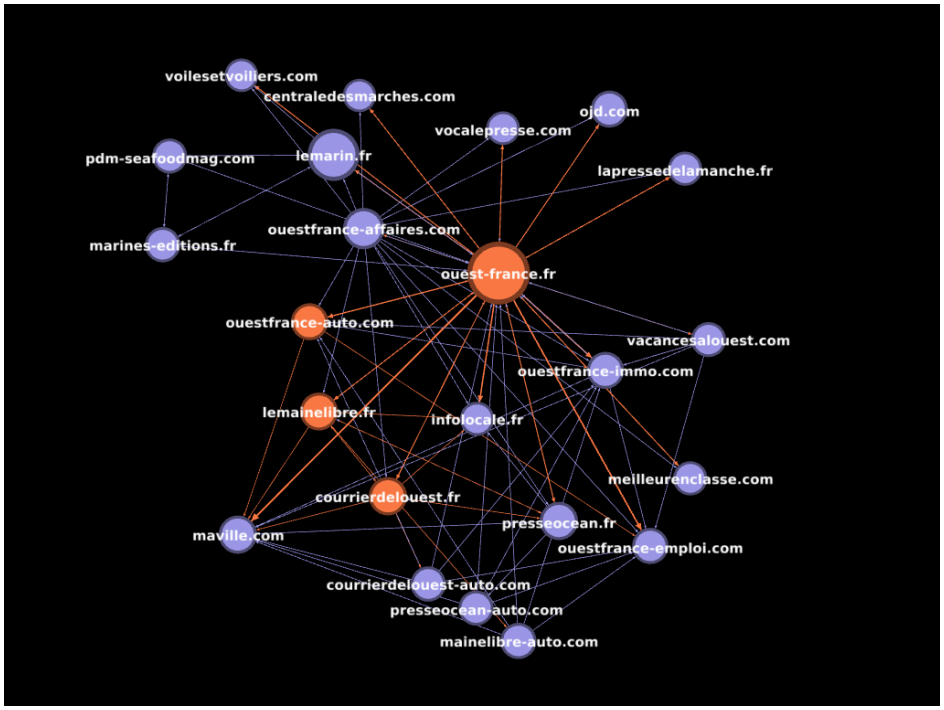

Figure 3.27: All nodes surrounding the cookie/domain hybrid `ouest-france.fr` with an outdegree of at least one.

While all websites are French and many seem to belong to the same entity, it is unclear whether this is true for all of them. As some of these nodes, such as `lemarin.fr`, are hubs of different small sub-communities themselves, this would theoretically allow a structure of the following kind:
A user browses to `lemarin.fr`, which uses a cookie from `ouest-france.fr`. Then, because `lemarin.fr` is the hub of another small community, the user browses to one of its surrounding websites, which is itself not connected to `ouest-france.fr`, for example `webasto.fr`[5]. Imagine `webasto.fr` links to another website that is part of a similar cookie-hub network. If this cookie is owned by the same entity as `ouest-france.fr`, the website could potentially notice that the user came from `webasto.fr`, inspect a database of `ouest-france.fr` to see if anyone has recently clicked on a link to `webasto.fr` from one of the sites using the *ouest* cookie, and trace these visits back to the same user.
While it may seem far-fetched, this is actually a scenario that does occur online[6]. However, it likely does not

[5] This domain is not actually displayed in this sub-graph.

[6] `https://www.slideshare.net/ODBA/measuring-theimpactgoogleanalytics`, accessed on 2nd June 2017, shows an example of this referrer-based tracking using Google Analytics.

apply to this specific case, as only one cookie-hub seems to be present in the graph. Still, it could be possible that one of the websites as mentioned in the scheme links to a website outside of this community, which would be much harder to detect.

Silo 24 contains roughly 4800 domains of different kinds, separable into different sub-communities. One of its sub-communities (figure 3.28) contains the perfect example of the described cookie-hub structure: different sub-communities[7] in the figure surround a small cluster of script nodes, in this case belonging to the Google Maps API. It is no surprise that Google would be able to track users on a large amount of websites, but this does confirm that such structures do indeed exist.

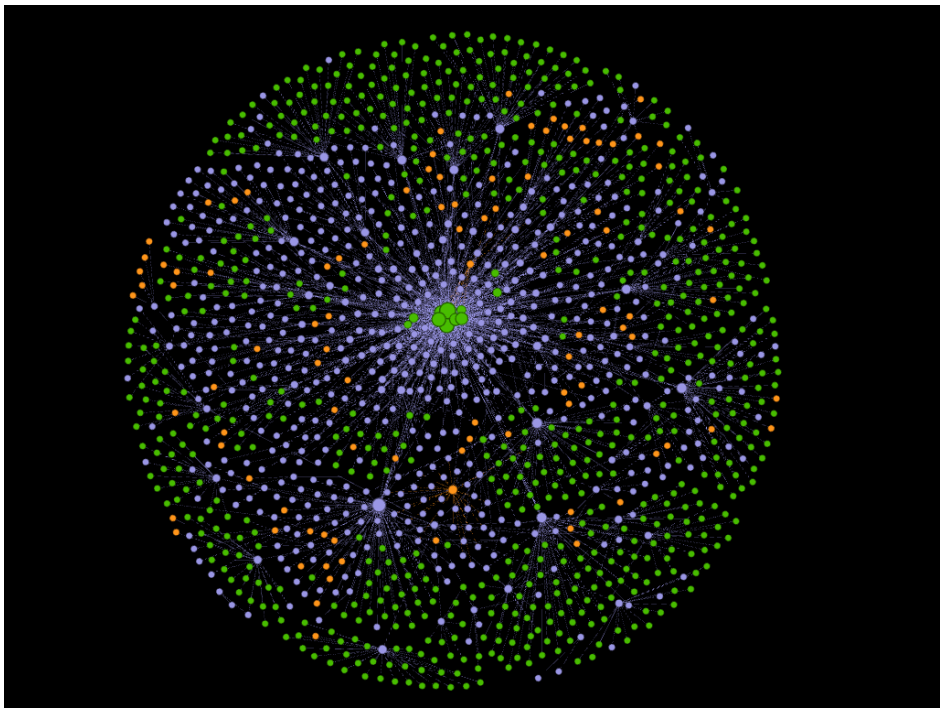

Figure 3.28: A sub-graph of silo 24. Orange nodes are cookies, domains are blue and scripts are green. Node size corresponds to degree. The dense cluster of script nodes in the centre indicates a tracking network.

In another silo, a similar situation has been encountered with Google Adsense and two jQuery versions. While they do not form a single hub together, each is the hub of a different sub-community.
A different silo contains a community of UK websites, linked together by a `voucherism.co.uk` script. None of the website nodes has any other edges, which is interesting.

Unfortunately, manually inspecting all the communities is very time-consuming, and is not feasible given the limited time available for this project. So far, however, the results are promising, indicating that this approach might be useful in future research.

Tracking graph

Applying the same approach to the tracking graph, this leaves 41 silos. Again, there are communities around hubs of several popular scripts and cookies, as shown in figure 3.29.
As the core consists of mainly Wordpress site utilities, this is not an interesting cluster in tracking context. However, figure 3.30 has a core that is actually an advertisement network, consisting of scripts and cookies from

[7] Therefore, they are are actually sub-sub-communities.

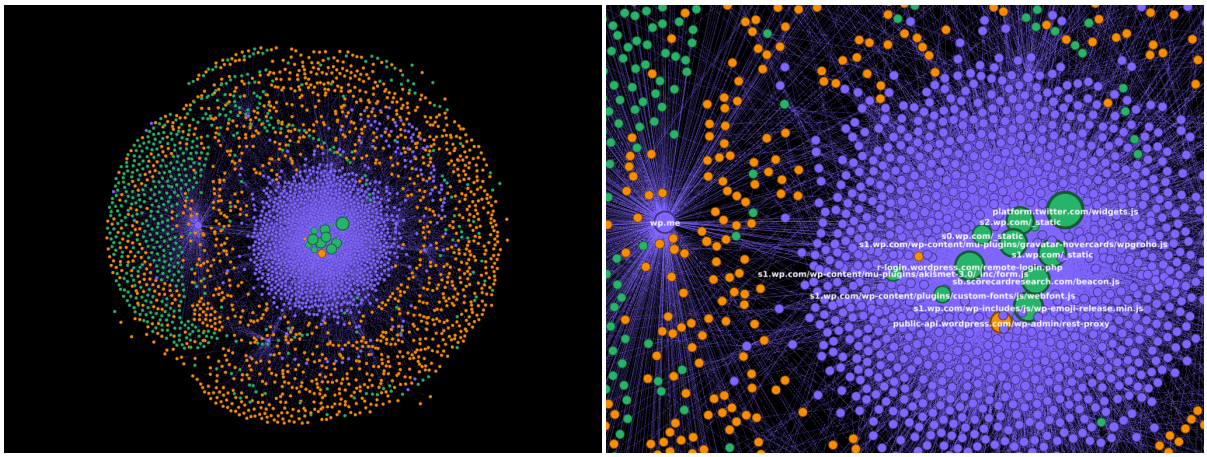

Figure 3.29: A sub-graph of tracking silo 34. It is centered around a hub of scripts and cookies from Wordpress and Twitter. Domains are blue, scripts are green, cookies are orange.

different entities. As there are over 300 websites connected to this hub of tracking code nodes, this is a very interesting finding. Unfortunately, the timeframe of this thesis project is not large enough to further analyse these domains and the hyperlinks between them[8], but this definitely shows that this approach can be used to identify tracking networks.

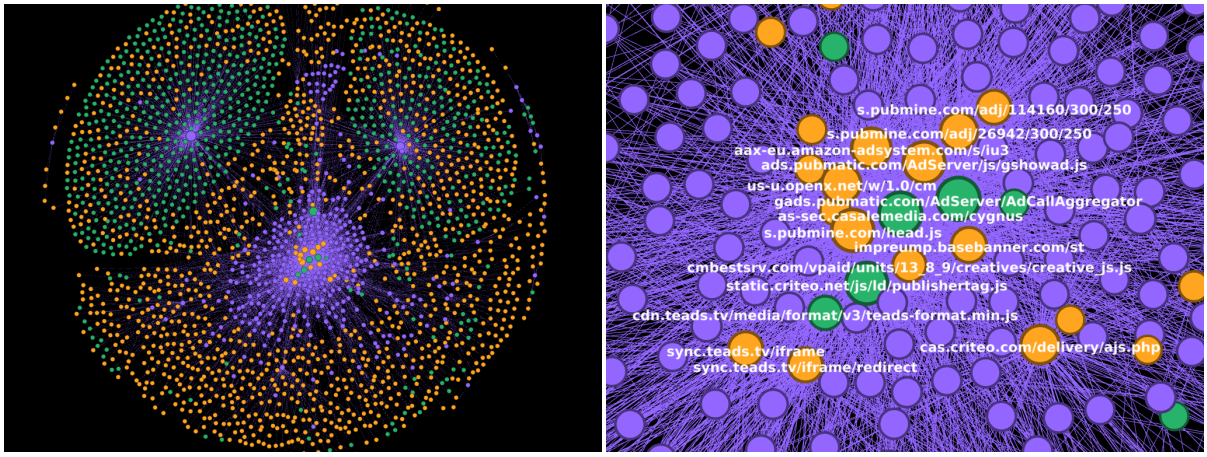

Figure 3.30: Another sub-graph of tracking silo 34 with a tracking core. This core, however, consists of advertisement scripts and cookies rather than general utilities. Domains are blue, cookies are orange, scripts are green.

Five other silos have been explored, but did not display any interesting results. Often, the silos are centered around a single popular piece of tracking code such as a twitter button.

[8] In the total graph, the community around these cookies and scripts is too big to visualise.

3.2.5 Reddit communities

The seed URLs were generated from the following set of subreddits:

```
'Art': ['alternativeart', 'graphic_design'],
'Culture': ['cyberpunk', 'opieandanthony'],
'Discussion': ['rant', 'socialskills', 'mensrights'],
'Gaming': ['leagueoflegends', 'casualnintendo', 'gaming'],
'Humor': ['scenesfromahat', 'blackpeopletwitter', 'bikinibottomtwitter'],
'Info': ['abrathatfits', 'explainlikeimfive'],
'Lifestyle': ['fitness', 'makeupaddiction', 'relationship_advice'],
'Location': ['losangeles', 'croatia', 'turkey'],
'Movies': ['movies', 'netflixbestof'],
'Music': ['music', 'kpop', 'popheads'],
'News And Politics': ['socialism', 'hillaryclinton', 'the_donald'],
'Pictures': ['perfectfit', 'highqualitygifs', 'abandonedporn'],
'Q And A': ['iama', 'samplesize'],
'Read And Write': ['writing', 'fountainpens'],
'Science': ['science', 'space', 'chemistry'],
'SFW Porn': ['historyporn', 'gentlemanboners', 'militaryporn'],
'Sports': ['mma', 'eagles', 'reddevils'],
'Technology': ['android', 'jailbreak', 'windowsphone'],
'TV': ['strangerthings', 'community', 'rickandmorty']
```

So far, the subreddits for which corresponding communities have been encountered are *turkey* (appendix 6.6, 6.12) and *mma* (appendix 6.8), but it would be interesting to see to what extent the other subreddits can be identified in the data. Because both mentioned examples were found as part of a bigger community using the hyperlink graph, the same approach will be used to search for the other subreddits: a list of all lowest-level hierarchy communities will be created from the hyperlink graph, which will be manually inspected with the aim of finding communities related to the original subreddits.

For every subreddit, a selection of three websites will be made from the top 15 most linked domains in that subreddit. Only websites that are manually determined to be specific enough to this subreddit or category are chosen, including only those websites that do not return a 404 response when visited.
If multiple sites are present that obviously belong to the same entity, i.e. `blogspot.com` and `blogspot.net`, or `steampowered.com` and `steamcommunity.com`, only one will be inspected. The exact URLs chosen and their corresponding findings are listed in the appendix section *Reddit community results*.
Note that these results are not objective, as there is no accurate way to programmatically determine whether a detected community corresponds to a given subreddit. The described approach is likely far from optimal, but it is all that can be done given the scope of this thesis project.

Figure 3.31 shows an example of these top 15 linked domains for two subreddits. Following the described approach, `artstation.com`, `deviantart.com` and `360artgallery.com` are chosen to be inspected for *alternativeart*. They are chosen in decreasing order of frequency, leaving out `tumblr.com`, `github.com` etc. because they are not specific to this community.
For *casualnintendo*, there is only one community-specific domain: `mariomayhem.com`, which is therefore the only domain that will be inspected.

Let's look at the results per category:

Art

alternativeart was not found to have a corresponding community. *graphic_design*, however, was a specific sub-community somewhere, meaning that community detection had to be performed once more on an existing community to extract it.

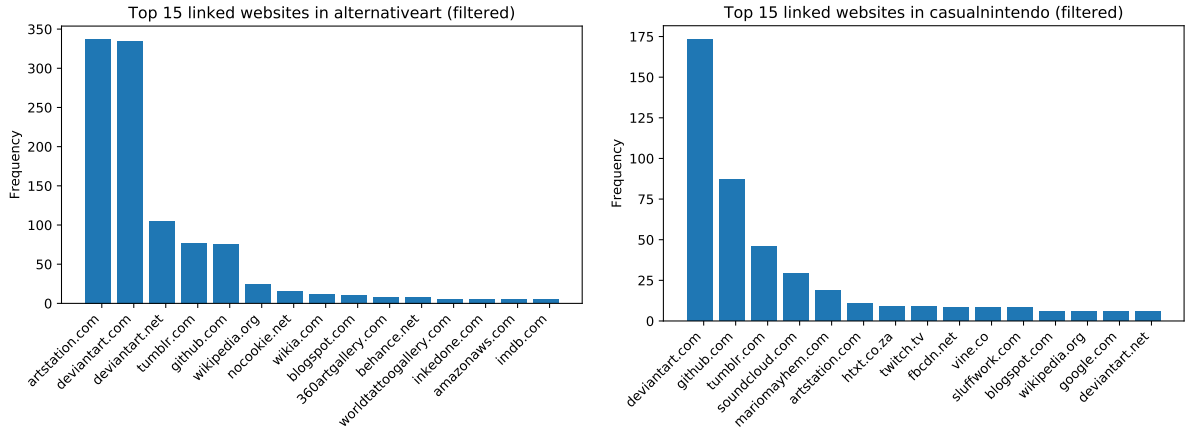

Figure 3.31

Culture

Neither *cyberpunk* nor *opieandanthony* have any community-specific websites, and have no corresponding communities.

Discussion

Both *rant* and *mensrights* do not have any community-specific websites. *socialskills* does have some, but the only found community consists of only seven websites. As only communities with at least twenty websites have been considered before, let's keep consistent and not count this as a corresponding community.

Gaming

All three subreddits have their own corresponding communities, although the *leagueoflegends* community contains only slightly more than twenty websites.

Humor

scenesfromahat and *bikinibottomtwitter* do not have any community-specific sites. While *blackpeopletwitter* does have some, none of these is part of a specific humour-themed community.

Info

Neither subreddit has a corresponding community.

Lifestyle

fitness and *makeupaddiction* do have corresponding communities[9], *relationship_advice* does not.

Location

All three subreddits have corresponding communities.

[9] *makeupaddiction* is only a community to some extent, as described in the appendix.

Movies

The *movies*-subreddit has a corresponding community, *netflixbestof* does not. As one of the most linked Netflix websites is in the movie community, these two subreddits likely form a bigger community together.

Music

music as a subreddit has a corresponding community, but *kpop* and *popheads* do not.

News and politics

There is a community of socialist websites, but none specifically about Hillary Clinton or Donald Trump. However, as many news communities have been found, it is fair to say that this entire category exists in the form of multiple communities.

Pictures

None of the subreddits has a corresponding community, likely (partially) due to the fact that media hosting websites have been filtered out of the seed URLs[10].

Q and A

Neither of the subreddits has a corresponding community.

Read and write

Both subreddits have corresponding communities.

Science

Science communities have been detected in the data, therefore corresponding to the *science* subreddit. However, the other subreddits have shown to be part of these communities as well, indicating that the three subreddits are not individual communities, but rather one large one.

SFW porn

Only the *militaryporn* subreddit has a corresponding community. This is likely due to the fact that the other subreddits are not specific enough: history is often related to military (i.e. the World Wars), and *gentlemanboners* concerns pictures of celebrities, which may also fall under lifestyle or news categories.

Sports

While communities have been found for all three sports, only *MMA* had a community specific to the subreddit. The other two subreddits are specifically about certain clubs, which are part of larger communities regarding their respective sports.

Technology

No specific communities could be found for any of the subreddits, as they form one large community rather than three separate communities.

[10] Furthermore, any URLs that end with a media file extension have also been ignored in the crawling process.

TV

None of the subreddits has a corresponding community, but *community* and *rickandmorty* did not have any community-specific websites. Instead, the subreddits have a tendency to link to the same websites as the *movies* and *netflixbestof* subreddits (i.e. `imdb.com`), indicating that TV might simply be part of the same community as Movies.

Other communities

Apart from those already identified before, the following communities have been encountered in the analysis:

- Virtual Reality
- Coding
- Flying
- Website development
- Baseball
- Healthcare (heart diseases, surgery, physiotherapy, etc.)
- Pornographic material
- Motorsports
- Psychedelics

Chapter 4

Findings

4.1 Community silos

A total of twelve out of the nineteen categories have corresponding communities as defined by the Louvain Modularity algorithm using the specified parameters. Only seventeen of the fifty-one subreddits have their own specific community, and in some cases it appeared that the community consists of the whole category rather than one of its specific subreddits.

This shows a difference between the definition of a community as used by the Louvain algorithm (i.e. a structure-based inter-linkage threshold) and organic user-defined communities. A feasible explanation would be that the seed URLs contain too much *noise*. It is easy to imagine that two users of a specific subreddit engage in a conversation that is not related to the subreddit (this is known online as *off-topic* and is analogous to people at a sports club discussing, for example, the weather), yet enjoyed by other users (e.g. because it contains jokes), therefore exceeding the upvote threshold.
Furthermore, it might simply be the case that it is impossible to capture the essence of certain subreddits by just looking at the hyperlinks that are shared within. The differences between certain subreddits may be so subtle that they require contextual clues to identify.

Therefore, the approach taken in this thesis project is not an accurate way of identifying existing communities. However, although the detected communities do not necessarily correspond to those used to gather seed URLs, this approach still fulfilled its purpose of generating a list of URLs from different communities, as many different communities have been encountered in total. This shows that the websites crawled in this project are very diverse in theme, and the findings should therefore not be biased towards any particular type of Reddit community.[1]

Furthermore, it appears that even though Louvain Modularity is a hierarchical community detection algorithm, many communities were part of a bigger silo and needed a second iteration of community detection to be detected. We posit two feasible explanations:

- The Louvain Modularity algorithm as used in this project is an insufficient method in detecting online communities through hyperlinks.
- While the different communities in a silo are not related content-wise, they are apparently still sufficiently connected through hyperlinks to form one large community.

The first case is plausible, given the fact that the algorithm does not support overlapping communities and that the graphs have been converted to undirected graphs before the community detection could be performed, which may have affected the results. Especially the overlap is expected to have a big influence, as we know

[1] It is important to keep in mind that Reddit communities are inherently biased by being part of the site itself, which may be biased towards certain demographics (i.e. over-representation of North-American users).

from real-life examples that it is very much possible for communities to overlap. For example, if a person is part of a football-related community, this does not exclude them from being part of other communities such as computer science fanatics, and similar rules should also apply to websites.
Furthermore, the algorithm uses a *resolution* parameter, which affects the size of the detected communities. It has consistently been kept at the default value of one for this thesis project, but it may be that this is not the optimal value in this scenario[2].
The second case is less plausible, but more interesting in the context of this thesis. After all, the aim was to gain insight in the structure of the Web in relation to driving market forces. It could theoretically be the case that the different communities existing within the detected silos are connected by something less obvious than the nature of their content; for example a mutual owner, stakeholder, or tracking network (although none have been identified as such). In that case, the entity owning these trackers or websites (or both) would benefit from users visiting one website in the network having an increased chance of also visiting another website in the network, and might thus provide motivation for websites in the same silo having hyperlinks to each other, even though they do not appear to be related.

4.2 Tracking networks

At first, only the communities with a below average degree were observed, as many of those with a higher degree are too big to visualise and inspect manually. Then, using solely the tracking code to link nodes together resulted in communities that are significantly lower in size, average degree and average connectivity compared to the communities based on hyperlinks (figures 3.8, 3.15)[3]. Upon manual inspection, the few communities of significant size (i.e. more than twenty domains) were linked on popular utilities rather than tracking code, which therefore did not provide any useful insights.
However, it turns out that within the bigger communities, different sub-communities can often be detected by running the community detection another time, and that these communities may indeed be centered around tracking networks (figure 3.30). Although it is not feasible to inspect all of these manually[4], devising some automated way of detecting the described *tracking code hub* structure would potentially provide a great, scalable way to gain insight in the tracking networks deployed across the Web, which may then be compared to the structural organisation of the involved websites through the hyperlink or total graph.

The total graph, involving both tracking and hyperlink data, proved not to provide any insights for the lower-degree communities either. While the resulting communities do differ from the purely hyperlink-based communities, this is only due to the fact that the community detection did not 'dissect' these communities to their smallest forms. If the community detection would be run iteratively (i.e. detecting communities within the resulting communities, possibly adding multiple levels of recursion), most communities would remain almost exactly the same, rendering this approach at most as useful as the hyperlink based graph. In fact, the total graph even performs worse in some cases, as communities exist where sub-clusters[5] are completely isolated from the other sub-clusters in the community.
However, while no such networks have been identified during the conducted analysis of the higher-degree communities, figures 3.27 and 3.28 show that this approach could potentially identify tracking networks similarly to the tracking graph. Based on these findings, it is not possible to conclude which approach may perform better using an automated method.

The results have also shown that in both approaches, page templatisation plays a significant role in the community results (figures 3.17, 3.23). Certain web services (e.g. Wordpress) allow their users to create their own website by filling in a template. The resulting websites then only differ in the content that has been filled in by the user, which likely does not include the cookies and scripts. Therefore, many different (otherwise unrelated)

[2] A very brief exploration of different values has shown that this results in the highest modularity, so one would expect that this is indeed the optimal value.

[3] Although it is important to keep in mind that these metrics are affected by the mentioned problem with the induced graph, and may therefore no longer be a good indication of the community quality.

[4] In fact, it is impossible as the graphs become too big to visualise.

[5] Which take the form of single, but also groups of nodes.

websites may contain the same set of scripts and cookies, served by the template host.

In conclusion, while this thesis project has not managed to provide any insight into the possible relationship between the structural organisation of the Web and an advertisement-based revenue model, the results have shown that the methods as described and conducted in this project can form a solid basis for potential future research to do so.

Chapter 5

Future research

The experiments conducted in this thesis project have been affected by several practical limitations.

An important limitation is the assumption that any third-party cookie or script is tracking user behaviour. Although the fifty most used scripts and cookies do indeed seem to be tracking code rather than common utilities, the community analysis has shown that in many cases, these utilities still play an important role. Therefore, it may be worth exploring ways to filter out popular utilities from the captured cookies and scripts. It may even be desirable to ignore popular tracking scripts and cookies such as Google Analytics, because although they are known to track user behaviour, they are used by such a large amount of websites[1] that they are unlikely to form a tracking network of any interest[2].

Furthermore, different community detection algorithms may significantly alter the resulting communities. As mentioned before, Louvain Modularity does not support overlapping communities, while those are not at all an unrealistic scenario. Exactly how much the results differ when using overlapping communities cannot be predicted, but it is definitely worth exploring. K-clique-percolation does not scale well, so especially if a bigger set of domains is to be crawled, an highly efficient algorithm is desired.
Even within the frame of using Louvain Modularity, different approaches could be explored. As it appears that many communities are encapsulated in a bigger community that does not necessarily seem related, it would be useful to construct a recursive Louvain method. This approach would run further iterations of Louvain on previously detected communities for a given amount of times, depending on the resulting sub-communities. As these results differ per community (it may be that the detected community is already optimal and does not contain any sub-communities), the resulting modularity should be examined for different cases, consequently defining a modularity threshold value to decide how many iterations should be run on the given community. This should result in more and smaller communities while also increasing their coherence (i.e. two websites in the community are more likely to share properties), making it easier to compare individual communities from separate graphs.

These improved communities should then be analysed for the described tracking code hubs. Using measurements like *betweenness centrality*[3], it should be possible to programmatically detect these hubs, significantly reducing the amount of communities that needs to be manually inspected. When such structures have been identified, the hyperlink structure of the associated websites may be inspected, providing insights into the way that users might be motivated to browse more websites from this specific tracking network.

Lastly, a very interesting metric would be to calculate the probability of certain websites sharing a piece of

[1] `https://www.slideshare.net/ODBA/measuring-theimpactgoogleanalytics` (accessed on 2nd June 2017) shows that Google Analytics was used on roughly 30% of 48 million inspected websites. For popular websites, the prevalence was as high as 65%, and the research was done several years ago.

[2] They do form a tracking network consisting of all websites that use the specific tracking code, but as this network spans half of the internet, it will not provide any new insights.

[3] As described on `https://en.wikipedia.org/wiki/Betweenness_centrality`, accessed 2nd June 2017.

tracking code. This could be calculated for any two websites in general, two websites that are directly connected by hyperlinks, and websites from the same (hyperlink) community, and should give insight into the relationship between hyperlink-based and tracking-based connections between websites.

Bibliography

[1] Albert-László Barabási. *Network Science (PDF)*. Cambridge University Press, 2015.

[2] Lowell W Beineke, Ortrud R Oellermann, and Raymond E Pippert. The average connectivity of a graph. *Discrete mathematics*, 252(1-3):31–45, 2002.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), 2008.

[4] Károly Boda, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. User tracking on the web via cross-browser fingerprinting. In *Nordic Conference on Secure IT Systems*, pages 31–46. Springer, 2011.

[5] Nicolas Dugué and Anthony Perez. *Directed Louvain: maximizing modularity in directed networks*. PhD thesis, Université d'Orléans, 2015.

[6] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[7] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):pages 9–13 and 19–20, 2010.

[8] Michelle Girvan and Mark Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[9] Avi Goldfarb and Catherine E Tucker. Privacy regulation and online advertising. *Management science*, 57(1):57–71, 2011.

[10] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[11] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, 2016.

[12] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[13] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12. USENIX Association, 2012.

[14] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pages 261–270. ACM, 2009.

Chapter 6

Appendix

6.1 Hyperlink silos

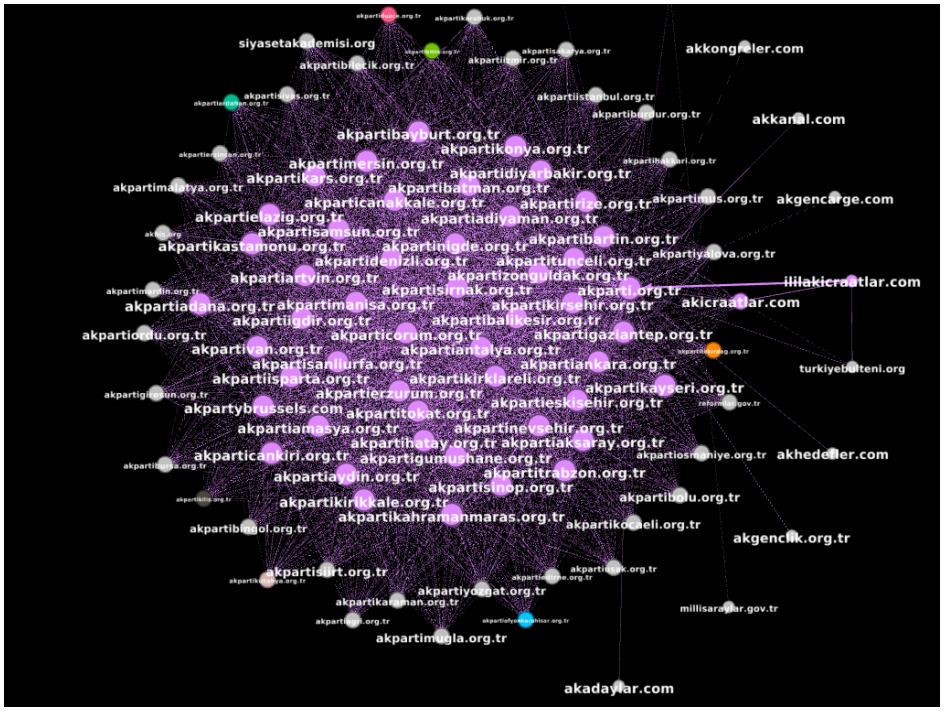

Figure 6.1: Hyperlink silo 179
Different colours represent different strongly connected networks. Many of the websites seem to belong to the same organisation.

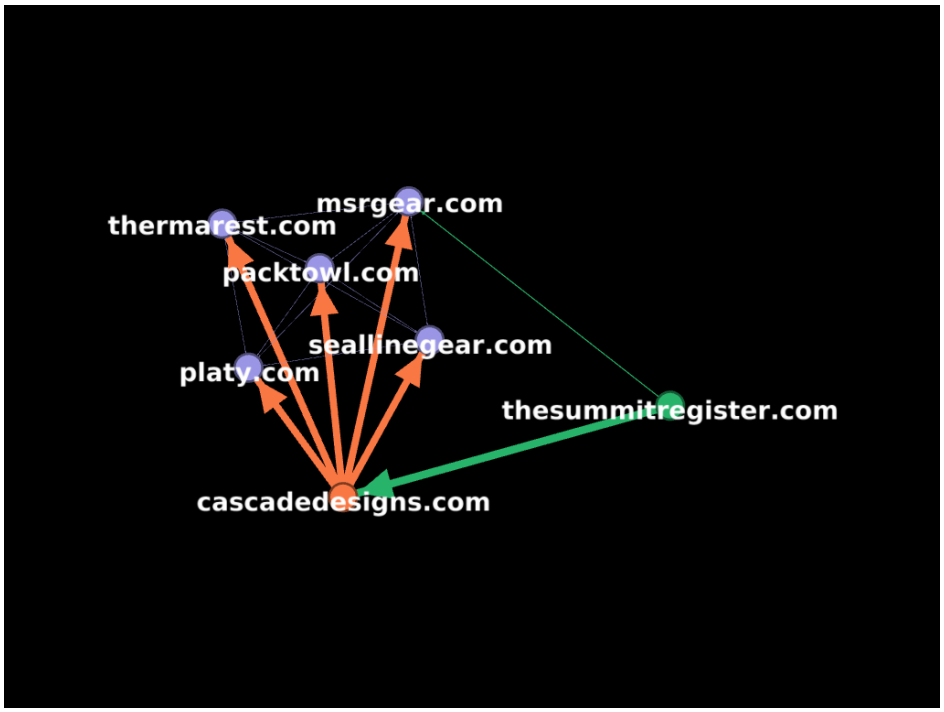

Figure 6.2: Hyperlink silo 592
Different colours represent different strongly connected networks. The network is small, and consists mainly of websites belonging to the same organisation (the purple sub-graph).

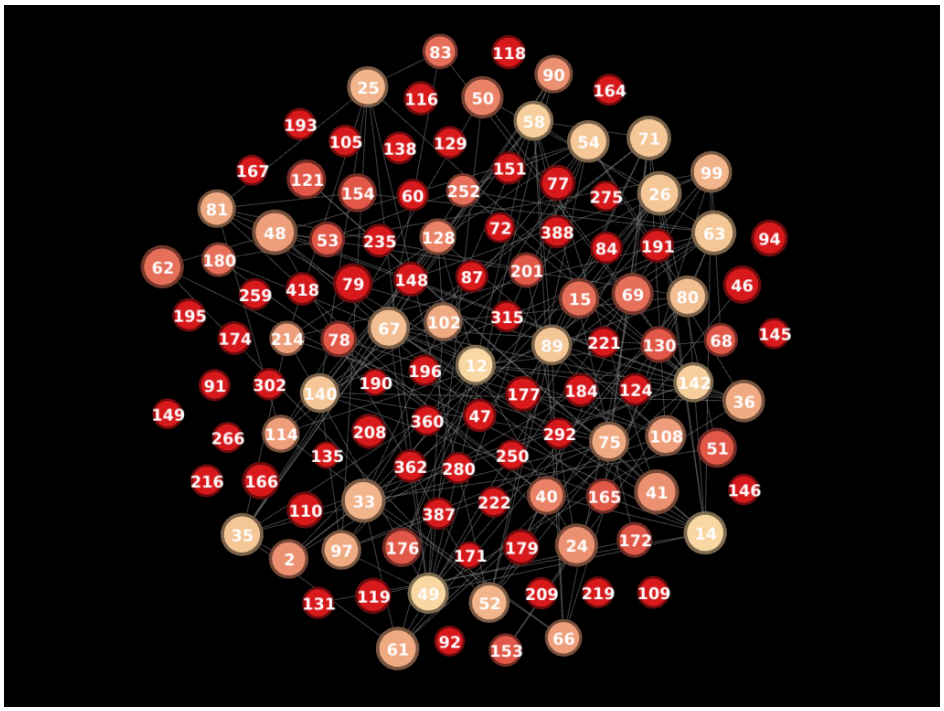

Figure 6.3: Hyperlink communities with at least 20 websites and below average degree.
Colour corresponds with degree, size with amount of websites.

6.2 Nested communities

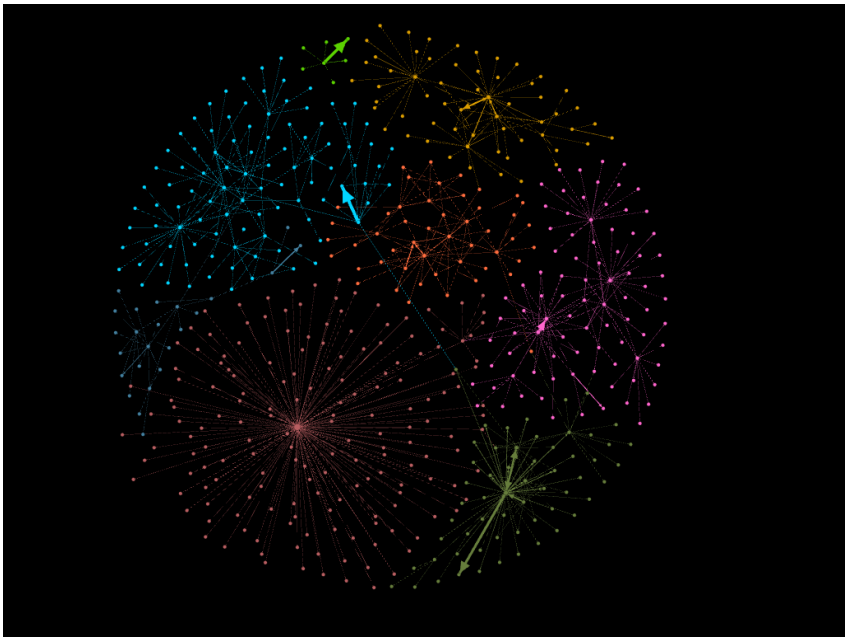

Figure 6.4: Detected communities in hyperlink silo 69.
Different colours represent different communities as detected by Louvain Modularity, calculated in Gephi using edge weights, randomisation and a resolution of 6. Layout is determined by Fruchterman-Reingold.

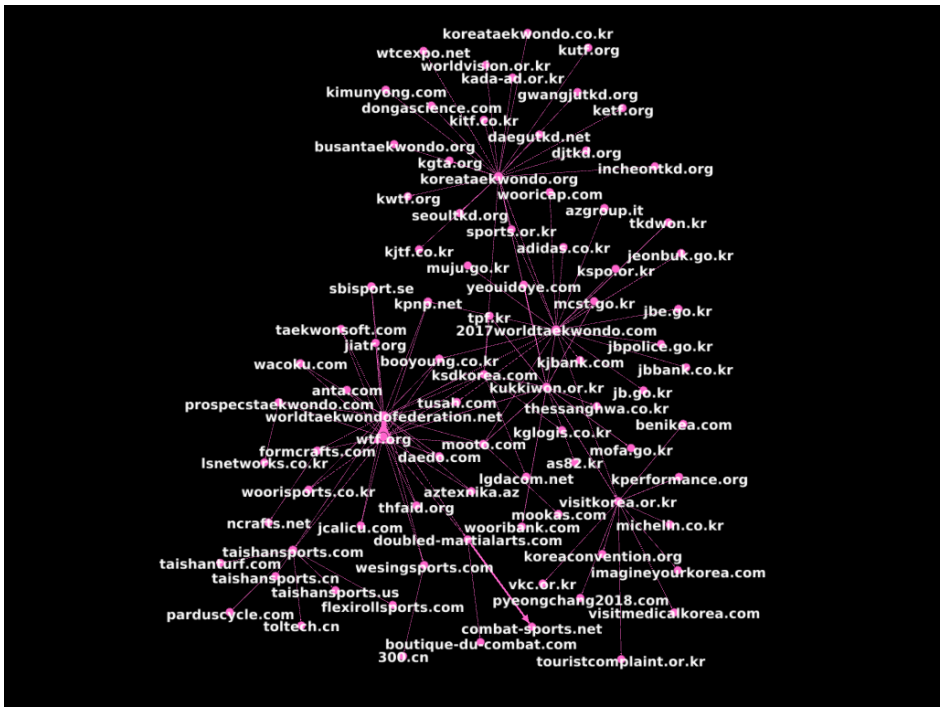

Figure 6.5: Cluster 69.0

Figure 6.6: Cluster 69.1

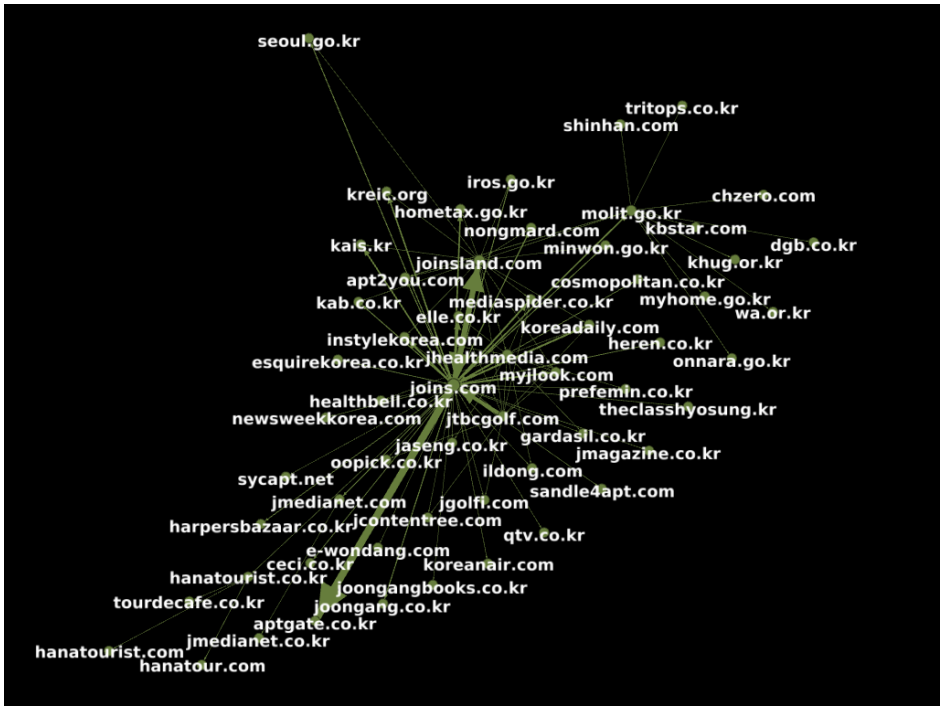

Figure 6.7: Cluster 69.2

Figure 6.8: Cluster 69.3

Figure 6.9: Cluster 69.4

Figure 6.10: Cluster 69.5

Figure 6.11: Cluster 69.6

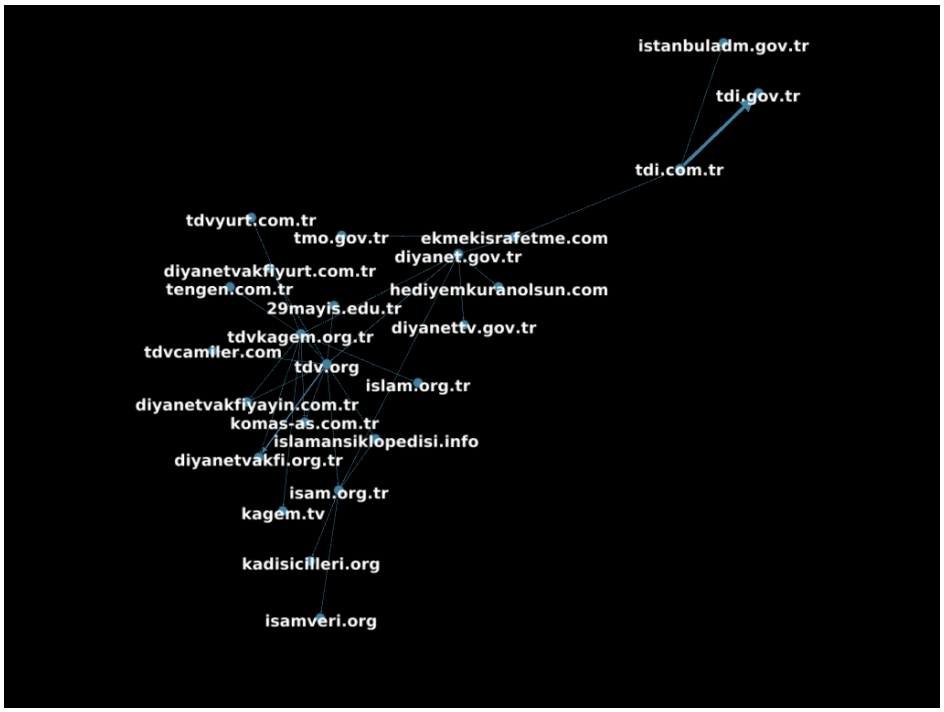

Figure 6.12: Cluster 69.7

6.3 Reddit community results

Art

Alternativeart:

- `artstation.com` is in a huge silo (several thousand domains) that do not share a general theme. Such cases will further be referred to as 'huge random silos'.
- `deviantart.com` is in a cluster that mostly contains sites about fantasy, stories, etc., but also some about completely different topics like pregnancy. A fantasy-themed sub-community is detectable, but no art-themed one.
- `360artgallery.com` is in a cluster that does seem to have many art websites, and is centered around `tumblr.com`. However, no specific art-related community is detectable within this cluster.

Graphic design:

- `behance.net` is in a huge cluster that does seem to contain some design websites. It is centered around `instagram.com`, and a design-themed sub-community is detectable, displayed in figure 6.13.
- `underconsideration.com` is in the same cluster.
- `dribble.com` is also in the same cluster.

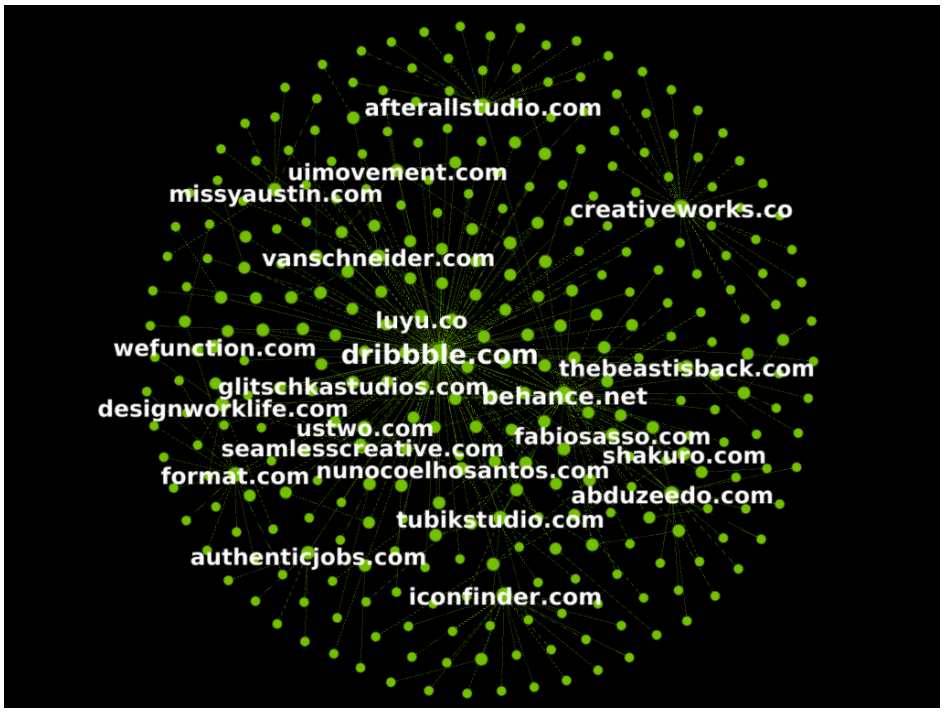

Figure 6.13: Detected design sub-community. Only hub labels are displayed.

Culture

Cyberpunk:

- Does not contain any community-specific sites. Art-wise, it has Artstation and Deviantart, and technology-wise, it has `arstechnica.com` and `wired.com` All four are in different communities, none of which corresponds to a cyberpunk community.

Opieandanthony:

- Does not have any community-specific websites. The only non-general (i.e. not Google, Wikipedia, etc.) are `redbarradio.net` and `tmz.com`, a radio channel and a celebrity news website. Both are in different communities.

Discussion

Rant:

- Does not have any community-specific websites, but `cnn.com`, `legaliq.com` and `bk.com` are all in different communities.

Socialskills:

- `aminl.com`, `becomemorecompelling.com` and `socialcirclemaker.com` are all in different communities.
  `becomemorecompelling.com` is part of a cluster of only 7 websites that do seem to share the life advice theme, but the others are not in any related clusters. Because the only found cluster is so small, it will not be counted as being a corresponding community.

Mensrights:

- Does not have any community-specific websites, and `dailymail.co.uk`, `nih.gov` and `telegraph.co.uk` are all in different communities.

Gaming

Leagueoflegends:

- `lolesports.com` and `leagueoflegends.com` are in the same community of roughly 20 *League of Legends*-related sites.
- `stage.gg` is separate in a random cluster.

Casualnintendo:

- Contains only one community-specific website, `mariomayhem.com`, which is in a cluster of roughly 20 nintendo/retro-games related sites.

Gaming:

- `bulbagarden.net` is in a cluster of roughly 25 Pokémon related websites. That is specifically Nintendo, but does of course qualify as gaming.
- `steampowered.com` is in a cluster of 430 domains, some of which do seem to be gaming. Indeed, a gaming sub-community is detectable, consisting of many gaming websites centered around this website.
- `polygon.com` is in a big cluster of unrelated websites.

Humor

Scenesfromahat:

- Contains no community-specific websites, and `xkcd.com`, `jshell.net` and `explainxkcd.com` are all in separate clusters.

Blackpeopletwitter:

- `snopes.com` is in a huge random cluster.
- `knowyourmeme.com` is in a small cluster of 5 meme sharing websites.
- `memesuper.com` is in a small cluster of seemingly unrelated media websites, which is interesting. Why is it part of this cluster?
  Upon inspection, the cluster is centered around `ziffdavis.com` and `ign.com`. Apparently, `memesuper.com` has been linked to from `ign.com` four times, while not connected to any other nodes in the network.

Bikinibottomtwitter:

- No community-specific websites.

Info

Abrathatfits:

- `bratabase.com` is in a big random cluster.
- `brabandproject.com` different cluster, same situation.
- `brasandbodyimage.com` different cluster, same situation..

Explainlikeimfive:

- No real community-specific websites. `stackexchange.com` is part of a cluster of 237 seemingly unrelated websites. A big part is about coding, but others are about history, science etc. as well. These might fit this subreddit, but the cluster contains completely irrelevant websites as well, and no related sub-community can be detected.
- `scientificamerican.com` is in a cluster of mainly science websites, and `nasa.gov` is in a big cluster of unrelated websites.

Lifestyle

Fitness:

- `strongerbyscience.com` is in a big cluster of fitness/bodybuilding websites.
- `t-nation.com` is in a big cluster of unrelated websites.
- `bodybuilding.com` is in another big cluster of unrelated sites.

Makeupaddiction:

- `allure.com` is in a cluster of fashion, cosmetics and travel websites. However, this cluster also contains other websites, so it may be interesting to inspect further. There are several sub-communities that are shopping/fashion/cosmetics related. Personally, I would say that the community does partially exist, albeit not in the form of any individual community.
- `teenvogue.com` is also in this cluster.
- `sephora.com` is in a cluster of 130 seemingly unrelated shopping websites. Upon inspection, only very few of them (< 5) are actually shopping sites, making this cluster irrelevant to the subreddit.
- Interestingly, a beauty-themed community was actually found during inspection of another subreddit. However, this is centered around a single website, `nerdylibrariangirl.ca`, and is therefore not a dense community.

Relationship_advice:

- `captainawkward.com` is in a huge cluster of unrelated websites.
- `artofmanliness.com` is in the same cluster.
- `loveisrespect.org` is also in this cluster. Despite all three websites being part of the same silo, no corresponding community could be identified as such.

Location

Losangeles:

- `latimes.com` is in a cluster of 220 websites, only a few of which seem to be related to LA. However, it is difficult to judge just from the domain names, because some are companies in california. Upon visualising, many of the domains seem to be LA (or at least California) based.
- `laweekly.com` is in a similar situation of roughly 400 websites, some of which may be LA related. Indeed, upon visualising, many of the domains are LA (or at least California) based. While not a dense community, this does count as a community.
- `laist.com` is in a huge cluster of unrelated sites.

Croatia:

- `jutarnji.hr` is in a cluster of roughly 200 websites, most of which are also Croatian.
- `vecernji.hr` is in a similar cluster, containing roughly 40 websites.
- `index.hr` is in a huge random community.

Turkey:

- `t24.com.tr` is in a huge cluster of unrelated sites.
- `cumhuriyet.com.tr` is in a cluster of five Turkish sites, three of which have *cumhuriyet* in the name[1], so are likely owned by the same organisation, and the other two are `yaynet.com.tr` and `yesilbilgi.org`.
- `hurriyetdailynews.com` is in a cluster of about 50 Turkish websites.

[1] According to `https://en.wikipedia.org/wiki/Cumhuriyet` (accessed 29th May 2017), it is a Turkish newspaper.

Movies

Movies:

- `imdb.com` is in a huge (i.e. several thousands of websites) community of unrelated sites.
- `variety.com` is in a cluster of 400 movie-related sites.
- `hollywoodreporter.com` is in a huge random cluster.

Netflixbestof:

- `netflix.com` is in a small cluster of 12 websites, only 6 of which are actually netflix-related.
- `rottentomatoes.com` is in the same cluster as `variety.com`, containing many movie-related sites.
- `metacritic.com` is in a huge cluster of unrelated sites.

Music

Music:

- `last.fm` is in a huge cluster of unrelated websites.
- `soundcloud.com` is in a cluster of 109 websites, only a handful of which (i.e. 15 or less) seem to regard music.
- `spotify.com` is in a cluster of roughly 50 sites, most of which are indeed about music.

Kpop:

- `soompi.com` is in huge cluster of random sites.
- `vlive.tv` is in a small cluster of Korean websites, not necessarily about kpop.
- `allkpop.com` is in huge random cluster.

Popheads:

- `8tracks.com` is in a cluster of roughly 800 unrelated sites.
- `pitchfork.com` is in the cluster of travel, cosmetics and fashion (as found previously when checking `allure.com`).
- `genius.com` is in a huge random cluster.

News and politics

Socialism:

- `marxists.org` is in a cluster of 166 socialism sites.
- `jacobinmag.com` is in a smaller cluster of roughly 56 websites. It is hard to determine how many share the theme, but there are definitely multiple socialism themed websites.
- `telesurtv.net` is in a huge random cluster (isn't actually socialism themed itself either).

Hillaryclinton:

- `washingtonpost.com` is in a cluster of blogs and news websites, most of which do seem to have a political undertone. However, none seem to regard Hillary Clinton specifically.

- politico.com is in a cluster of many *.gov* websites, many of which seem to be health-themed.
- nytimes.com is in a cluster of 400 domains, but it is difficult to determine what they are about. Upon inspection, it seems that it is just a collection of domains that were linked to by nytimes.com, containing few other links.

The_donald:

- breitbart.com is in a huge random cluster.
- foxnews.com is in a cluster of 750 websites that do not seem related.
- dailymail.co.uk is in the same random cluster as breitbart, which contains roughly 3000 websites. No specific Trump-themed communities can be identified, but in general, the dataset contains many news communities.

Pictures

Perfectfit and Highqualitygifs:

- Both contain only non-community-specific websites with very low frequency (i.e. ikea.com with a frequency of 2). This makes sense, because the crawler and seeds were filtered for image hosting sites, and these subreddits are meant for sharing images.

Abandonedporn:

- freaktography.com is in a huge random cluster.
- abandonedkansai.com is somehow not present in the data at all.
- Does not have any other community-specific domains.

Q and A

Iama:

- Does not have any specific community websites.

Samplesize:

- qualtrics.com is in the same community as the movie websites identified before.
- surveymonkey.com is in a cluster of 528 science and healthcare websites. While this partially fits the theme, there is no specific community about asking and answering questions.
- usabilityhub.com is in a cluster of 50 human-computer interaction themed sites.

Read and write

Writing:

- davidfshultz.com is in a huge cluster (10000+ sites) of unrelated sites.
- wattpad.com is in a cluster of 1000 unrelated sites.
- goodreads.com is in a cluster of 500 sites. Not all of them are about reading and writing, but there is definitely a significant amount. Upon inspection, the cluster is actually centered around goodreads.com, but many websites are also individually interlinked. Such many sites are book/writing-themed that this community definitely corresponds to the subreddit.

Fountainpens:

- gouletpens.com is in a cluster of 1000 sites. Some are about writing/blogging, but relatively few. Upon inspection, the cluster is centered around pinterest.com and contains many book-themed websites, but no specific sub-community can be extracted. None of them seem to be specifically about pens.
- fountainpennetwork.com is in a cluster of 100 websites most of which are definitely about pens/writing.
- jetpens.com is in a cluster of 800 websites. Most of them are unrelated, but there are definitely some writing-themed sites in there. However, no specific sub-community can be detected.

Science

Science:

- nature.com is in a cluster of 1600 websites, most of which are about science.
- sciencemag.org is in a cluster of 150 science sites.
- wiley.com is in the same cluster as nature.

Space:

- nasa.gov is in a huge random silo.
- space.com is in small silo of 100 domains, some are about space/science, but not many.
- phys.org is in a cluster of roughly 50 science sites.

Chemistry:

- acs.org is in the cluster of 1600 science sites.
- sigmaaldrich.com is in a cluster of 700 English and American websites, a few of them seem to be about science.
- rsc.org is in a small cluster of chemistry, pharmacy and general science websites. However, of the 27 domains, only five or so are specifically about chemistry.

SFW porn

Historyporn:

- rarehistoricalphotos.com is in a huge random cluster.
- loc.gov is in a cluster of 100 websites, only a handful of which (< 10) are about history.
- It does not contain any other community-specific websites, as most of the links from this subreddit are to Wikipedia.

Gentlemanboners:

- celebmafia.com is in a cluster with many italian websites that do not seem related to this subreddit.
- purepeople.com is in a cluster that does contain some men's lifestyle websites, but by far not all of them are.
- gossipcop.com is in a huge random cluster.

Militaryporn:

- defense.gov is in a cluster of 200 military sites.
- navy.mil is in the same cluster.
- dvidshub.net is in the same cluster.

Sports

MMA:

- mmadecisions.com is part of a huge random cluster.
- mmafighting.com is part of a another huge random cluster.
- bloodyelbow.com is part of this cluster as well. There might be an MMA sub-community in this cluster. Indeed, although it is centered mainly around these two websites and is not a very strong community, an MMA sub-community can be detected.

Eagles:

- nfl.com is part of the same cluster as mmadecisions.com. It might not be completely random after all, as it does contains different sports websites. Upon inspection, there is an American Football themed sub-community, but it is not specifically about the Eagles.
- bleedinggreennation.com is also in this cluster.
- philadelphiaeagles.com is also in this cluster.

Reddevils:

- manutd.com is in a cluster of English websites, not specifically about sports.
- skysports.com is in a cluster of 50 sports sites.
- premierleague.com is in a cluster of 30 football websites, but they are not specifically about Manchester United.

Technology

Android:

- xda-developers.com is in a large random silo.
- androidpolice.com is in the same silo.
- androidcentral.com is in a silo of 117 sites about technology, gadgets and phones.

Jailbreak:

- saurik.com is in a cluster about webdesign and webprogramming, but it seems fairly random (not all seem to fit the theme).
- apple.com is in a seemingly random cluster of 300 sites.
- appsto.re is in a tiny small cluster with two domains of *urbandecay* as well as upsellit.com.

Windowsphone:

- microsoft.com is in a seemingly random cluster of 1600 sites.
- windowscentral.com is in a tiny cluster with

```
domains "recruiterbox.com"
domains "windowscentral.com"
domains "teslacentral.com"
domains "connectedly.com"
domains "mobilenations.com"
domains "crackberry.com"
domains "vrheads.com"
```

- aka.ms is in a cluster of 600 domains, of which a significant amount seems to be about (mobile) technology.

TV

Strangerthings:

- `ew.com` is in a seemingly random cluster of 1100 sites.
- `dazeddigital.com` is in a huge random cluster.
- `yenniper.com` is in a huge random cluster.

Community:

- No specific websites.

Rickandmorty:

- No specific websites.