# Mapping and visualizing the news coverage of events

## Project BigData 2019

**Marit Beerepoot**
**Jeffrey Bosman**
**Jessy Bosman**
**Delano de Ruiter**
**Group 16**

**Abstract**

Where newspapers used to send reporters to every event, technology and the internet makes it easier than ever to be informed about the latest events that are happening. This development led to more and more news sources covering the same events, but with this development, coverage biases were introduced. So far there is no tool developed to analyze the news coverage and possible biases. This research uses the Global Database of Events Language and Tone (GDELT) to answer the question: "Is it possible to create a model that finds similarity in reported events between different news sources in real time, based on high-level event data?". The framework introduced by this study uses the cosine similarity of news outlets and events to visualize the news coverage by creating a network based on it. The network can adapt to new incoming data by using a lambda structure, and therefore includes the big data dimensions volume and velocity. The study found that the network that is created is a clear network that can be used as a starting point or base to do further research on.

# 1. Introduction

The past few years it has become easier for journalists to read and learn about the work of competing journalists. Back when only newspapers existed, new sources did not know about what other news outlets were writing until the newspaper came out the next day. When a news source did not send out a reporter, it did not know the details about recent events. News sources have started to communicate more and more over the years to share stories and information, so they were able to write about more recent events. Until about a decade ago most journalists were dependent on face-to-face conversations, yet now technology and the internet make it easier to be informed about the most recent events. Journalists are monitoring each other more than ever and a consequence of this is that more and more news sources write about the same events (Boczkowski, 2009).

Even though more news sources write about the same events, it is still not the case that all events are covered by the same amount. It was, for example, found that, especially with political events, television news sources intentionally only cover a small set of events, and thereby create some kind of bubble of what a person gets to know and not know. This could be quite problematic since people have the right to access unbiased political information (Hofstetter & Buss, 1978). Next to that, it was also found that there is a structural bias in the news: the government dominates the media attention. This domination can cause other events, for example, events happening in the opposition, might not be covered by the news (Van Dalen, 2012).

Interesting is that most of the research into bias and similar news coverage are based on qualitative analysis. This research, therefore, focuses on a more quantitative approach by mapping news sources that write about

the same events based on big data analysis. This will be done using the Global Database of Events, Language and Tone (GDELT). The research question that will be used during this study is:

*"Is it possible to create a model that finds similarity in reported events between different news sources in real time, based on high-level event data?"*

Creating such a model could be interesting for different reasons. First of all, it could be interesting to detect news bubbles or biases and if there are certain news outlets that only cover a small set of events about certain topics. Users can thereby gain new insights into biased news outlets. Next to that can this model be used as a starting point for plagiarism detection since this is a big problem in journalism (Girard, 2004). The network can map out news outlets that only cover the same events. This can then be used to extract the URL to the articles, which then can be used to detect textual similarities. Newspapers can hereby gain a starting point to detect plagiarism, which is faster than searching for similar articles manually. Another way to use the model can be to see how events progress over time. It can, for example, be interesting to look at which events are covered by all news outlets immediately and which events never make it to more than a local news outlet. Newspapers and users can both gain insights into what kind of events get covered more.

This paper starts off by describing some related work. First, literature about news biases and framing will be discussed. After that, some research that is already done based on the GDELT dataset will be highlighted. This section ends with an introduction into the lambda architecture that can be used to process new data. After the related work section, the data will be discussed. Thereafter the design of the framework will be described, by looking into the plan, the components that are needed and the architecture that is needed. Next the implementation details will be discussed, followed up by the results and examples of the created network. Finally, this paper is finished up with a conclusion and some future work directions.

# 2. Related work

In the past, all (findable) research about news coverage differences has been done based on qualitative analysis. This research focuses more on quantitative analysis, by looking at the differences in the coverage of certain events. First, there will be looked into the effects of bias, framing, and bubbles, to emphasize on why creating a news coverage model is important. After that, the lambda architecture, an architecture that can be used to process real-time data, will be explored.

### 2.1 News biases and framing

A lot of research has been done into the news coverage of events, especially the news coverage of political effects. Framing and biases seem to be a recurring problem in the news coverage. Entman (1993) described framing as "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (p.52) Framing can have different causes. McLead and Detenber (1999) for example found that different degrees of news coverages can partly be caused by preexisting cognitive orientations. An example of this is when there is a pervasive social bias against a social group, the news coverage will be minimized. Another factor that they found is preexisting knowledge about the event. More knowledge will lead to more news coverage and also better quality articles.

Next to framing bias can be a problem. Ball-Rokeach and DeFleur (1976) found that the media can easily influence the news coverage of events, for events where the target audience of the media has no direct contact with or personal history with. This means that as long as there is a certain distance between the

target audience and the event, the audience would not blame the news outlet by not covering it and that the outlets can switch up the facts without a lot of backlashes. Next to that, it was found that, especially with political events, television news sources intentionally only cover a small set of events, and thereby create some kind of bubble of what a person gets to know and not know. They decide which events to cover based on their target audience. This could be quite problematic since people have the right to access unbiased political information (Hofstetter & Buss, 1978).

A big danger of uneven news coverage of events is that the audience forms an opinion based on what they see and hear. When they only get positive information about events, they will automatically form a positive opinion, while there could be a negative side to the story. Instead of searching for negative information, they will try to confirm their hypothesis: that a positive event happened. This can also be seen as a confirmation bias (Tversky & Kahneman, 1977). For these reasons, it is important that news outlets cover as many events as possible, without tampering with the details. The framework and model proposed in this research can help with visualizing which news outlets could be biased.

### 2.2 The GDELT dataset

In this research the Global Database of Events, Language and Tone dataset has been used. The dataset contains data about all the news in the world reported by broadcast, print, and websites (more information about the dataset can be found in section 3). Since the dataset contains data from 1972 till the present day, there is a lot of historical data available which makes it an interesting dataset to use for predictions. Yonamine (2013) for example used the dataset to predict future levels of violence in Afghanistan, based on historical data of violence. Next to that Qiao (2017) and Galla and Burke (2018) used the dataset to find patterns in news reports to predict social unrest.

Next to the research into predicting events, there is also a lot of research done based on analysis of the events and news outlets. Kwak and An (2016) for instance analyzed the photos of the news articles and were able to find what kind of pictures are used the most, what sentiment these pictures have. Next to that, they analyzed whether this sentiment aligned with the tone text. This made it possible to do research into how political candidates are portrayed by the news outlet. Finally, there is also a lot of research done with GDELT data to analyze the news coverage of a specific event. For example, Sagi and Labeaga (2016) who analyzed the news coverage and opinions in those articles of the energy policies from the Spanish Government.

The research described above covers just a small glimpse into all the research that has been done with GDELT. This study tries to contribute to these and new researches by creating a framework that can visualize the data from GDELT in real time. This framework can then be used to support further research and help other researchers in obtaining insight into how to interpret their findings. For specific researches, like the research into the energy policies, the network that is visualized can be filtered, so it only shows information relevant to the research to justify their findings. The network and framework are thus forming a base to work on and can be adapted for specific needs.

### 2.3 The lambda architecture

Creating a scalable real-time big data architecture can however be challenging. A popular way is by using the Lambda Architecture. The Lambda architecture is mainly focussed on computing functions on a dataset in real time, where all the data is available at all times. The architecture consists of three layers: the batch layer, the serving layer, and the speed layer. The batch layer should be able to store a master copy of the dataset, that is constantly growing. Next to that, it should be able to compute some functions on the dataset (for example to preprocess the data). This layer is thus mainly focussing on storing and updating the dataset and creating views. The serving layer can be seen as storage, where the output of the batch layer is stored

to be queried later. The views that are created here are called 'batch views'. The speed layer does computations on real-time data that is coming in, to ensure that it is possible to query over all the data including the recently updated data. The speed layer has high fault tolerance and focuses more on that the data should be available at all times than that it is the correct data. Mistakes that are made in the speed layer are corrected as soon as the batch layer is done with generating new views (Marz & Warren, 2015).

# 3. The data

The dataset that will be used for this research is the Global Database of Events, Language and Tone (GDELT). The data is gathered by the GDELT Project, which is supported by Google Jigsaw. The dataset contains data about all the news in the world reported by broadcast, print, and websites. Next to that, it contains data about the people, organizations, themes, sources, locations, quotes, emotions, counts and images about events. The dataset contains data from January 1st, 1979 till the present day and is updated every quarter of an hour. It contains more than a quarter billion records and currently contains more than 52TB of data (GDELT, n.d.). The GDELT dataset is one of the largest datasets on the planet (Skvorc, 2014).

For the analysis in this research, the "eventmentions" table is used. This is a reference table of sources (websites) and a unique event id specifying the subject of the article. This data ranges from the 19th of February, 2015, and is still updated today. The dataset is 278GB, with 1,331,664,109 articles[1]. The table contains, in addition to the source and the specific event mentioned, a small overview of the content identifiers of the articles, such as the length of the article, the tone of the article, date published, and the positions (number of characters) before the event and/or actors are named.

### 3.1 Big data analysis

This study can be seen as a big data problem, since it consists of two of the four big data dimensions, namely velocity and volume. The dimensions will both be discussed separately.

**Velocity**
The GDELT dataset is updated in real time, which means that new data keeps arriving continuously. Since it could be interesting for users to directly analyze which news outlets cover an event after it just happened, it was decided to analyze the data in real time. This enables the visualization of events to be observed as time is progressing, giving new insights into the way event coverage is developing in news.

**Volume**
The dataset is really big since it covers every mention of every news outlet of every event. The dataset consists of 52 TB. Since that is way too much to analyze in the scope and timeframe of this research, a subset was taken. The subset consists of all the events from February 17th, 2015 until now. Since that dataset it still too big, it was decided to split the dataset up so that it only consists of events that were covered by news sources in the Netherlands. This subset contains 2 million mentions of events, reported by 1393 unique news sources. The total dataset contains 32 billion mentions of events, reported by 150 thousand unique news sources.

[1] https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.eventmentions?tab=details

# 4. Design

### 4.1 The plan

In this research, the focus lies on creating a model from the GDELT data, that visualizes the similarities in coverage of different events of different news outlets. The visualization will be in the form of a network, where the nodes visualize the news outlets, and the edges visualize how similar the events, the news outlets write about, are.

The nodes will have a size, determined by the number of events they post about. The biggest node thus is the news outlet that writes about the most events and thus has the highest news coverage rate in general.

To determine the edges between nodes, a customizable threshold is set. The similarity of the event coverage of the nodes is calculated (details described in 5. Implementation). If this similarity score reaches the threshold, an edge is drawn. Optionally, the network allows the edgeless nodes to be dropped (thus no similarity > threshold), to reduce processing time and to create a network that is more clear with less clutter.

The edge size is based on a similarity calculation: how more similar the events two news outlets write about, how broader the edge. There are two factors that come into play with the edge size: how much the news outlets write about the same event and how similar the content of those articles is. The narrowest edge will thus be an edge between two nodes that do write about similar events but have a different look on the event determined by the content of the article. This similarity calculation also takes into account how similar the content of the news reports about events are, not only how much events they both wrote about.

### 4.2 Real-time aspect

The system can handle real-time data since it repeats retrieving data every 15 minutes (the same rate as the GDELT dataset is updated) so that the data is always up to date. It can, however, take a while to do the node size and event size calculations. This is why it was decided to implement the lambda architecture, which has 3 layers:

- The batch layer calculates the actual node size and edge size.
- The serving layer contains the stored data in the form of dataframes and can be used to query and create the network on.
- The speed layer does an estimate of the node size and edge size, by using a min-max scaler on the data: instead of doing a similarity calculation it sorts the data values from 0 to 1. Nodes still need to have a certain threshold to be visible in the network and are closer to each other in the network when the scaled values are close to each other. The edge with the min-max scaled value 1 is the broadest and with 0 the narrowest.

### 4.3 Components of the architecture

To be able to realize the plan while taking the real-time data processing into account, the following components are needed:

Software:

- A way to retrieve and preprocess the data.
- A way to handle the real-time data that comes in.
- An algorithm that creates the network, which has 3 subcomponents:
    - An algorithm that counts the events a news outlet wrote about and determines the size of the nodes.

- An algorithm that calculates the similarity of the events different news outlets wrote about, to determine the size of the edges and to determine if the threshold is met.
- An algorithm that creates the network itself and changes the color of new incoming nodes while working with real-time data.

Hardware:

- Storage
- Computation hardware

### 4.4 The architecture

In this section, the architecture is introduced in the form of a pipeline. The pipeline is shown in figure 1 and the steps are further clarified below.

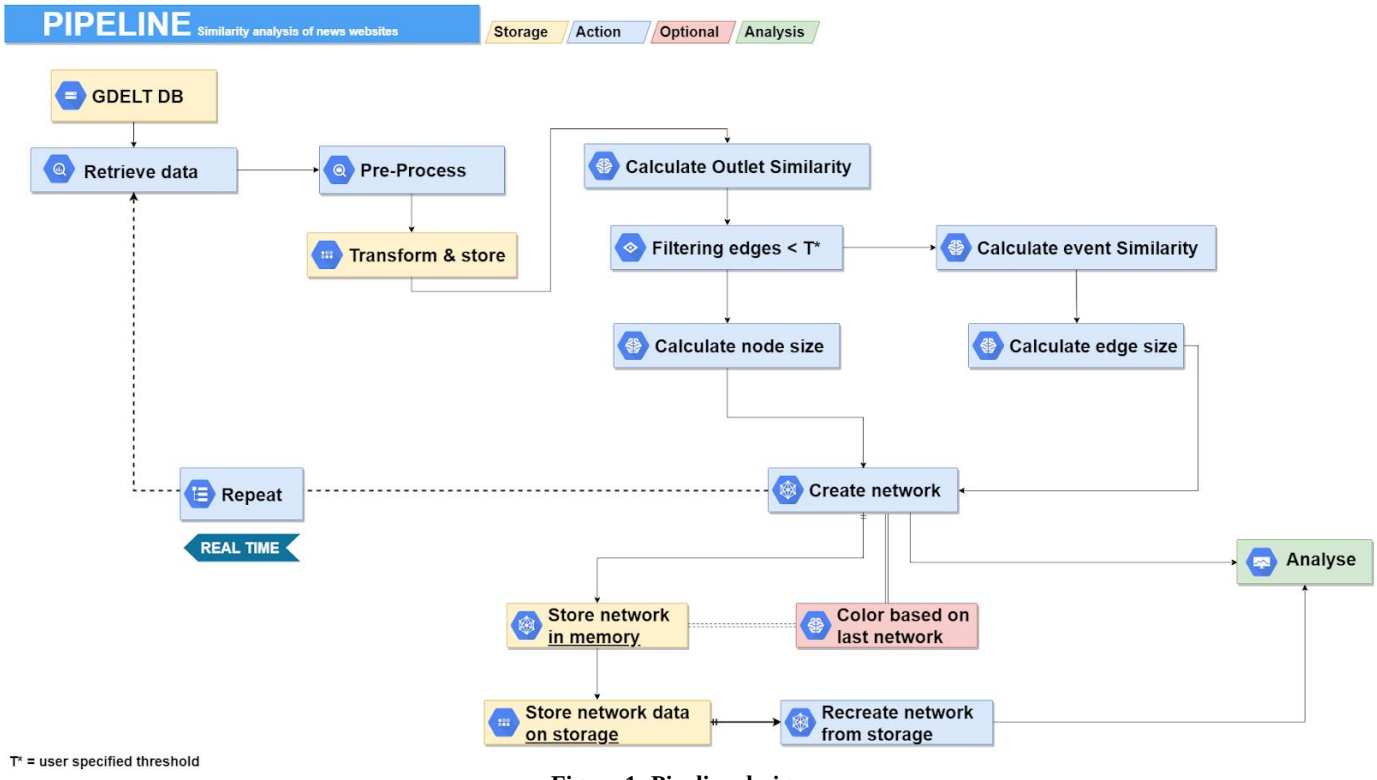



**Figure 1: Pipeline design**

The pipeline starts in the upper left corner with the GDELT DB. The GDELT DB contains the data as described in section 3. This is then retrieved with BigQuery API. Pre-processing consists of creating a dataframe with the columns that are used during the calculation. This dataframe is then stored.

After the data frame is stored, the algorithm to create the network can come into action. It first calculates the similarity between the outlets, and filters out edges that are below a set threshold. Nodes without edges are also deleted from the graph. After the filtering, the node size and edge size are calculated, which makes it possible to create the network. This is further explained in section 5. Since the model supports real-time analysis, the creation of the network is repeated to include new data.

After a model is created, the network is stored in memory. This stored network can detect which nodes and edges are new relative to the old network. These can then be colored differently. The network data is then stored, so networks created in the past can be easily recreated for further analysis.

# 5. Implementation

This section described how the designed network and real-time implementation (section 4) are realized, together with a description of upscaling, a performance analysis and an estimated cost analysis.

### 5.1 BigQuery

The GDELT dataset is hosted in Google BigQuery, which is an enterprise data warehouse that stores massive datasets. In BigQuery the dataset is continuously updated every 15 minutes. Each user receives 1 TB free BigQuery processing every month. This allows us to gather the data into python, which is done using the BigQuery REST API. This REST API allows us to take subsets of the data using the SQL language. For our analysis, a selection and count of all the occurrences per news source per event is made.

### 5.2 Cosine similarity

After cleansing and reshaping the data, the similarity between news outlets is calculated in python using the sklearn cosine similarity function. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The full equation is as follows:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

### 5.3 Network

In this section, the inner workings of the network visualization are explained. First of all, the actual visualization is created using the python package pyvis, a framework build on top of the python library NetworkX. Pyvis allows a more interactive, customizable network visualization, which is used to create a more clear picture of the network, with less overlapping nodes and text.

#### 5.3.1 Nodes

To visualize and map the news' coverage of events, the website-website matrix of cosine similarities is used to initiate a network visualization. First of all, each website is presented by a node (circle). The node size is calculated by querying a count on each source at the given time interval (the time between the last supplied date and the current supplied date). After all the node occurrences are counted, the counts are normalized using the min-max normalization, which is a technique that scales the data between zero and one whereby the minimum value is mapped to 0 and the maximum value to 1. This ensures that the nodes are scaled in a way that makes sure that the nodes are all visible and not too large or small.

#### 5.3.2 Edges

Thereafter, to ensure that the network is not clouded by a large number of edges (lines), a customizable threshold is set. If the cosine similarity between two websites is higher than the threshold, an edge is drawn. Additionally, to determine the weight (size) of the edge, the more informative data (such as people mentioned in the event) is extracted. The events are grouped by event id and the cosine similarity is calculated over all event data for all events for each website pair. These values are min-max normalized and used for the edge thickness.

#### 5.3.3 Update

After the first network is generated and saved, the implementation process described above can be repeated with new data (for example for a real-time implementation). Optionally, if supplying an old network, for example, the past day, the new network is generated using the old network as a reference. This enables coloring of the new nodes and edges, to create an overview of what has changed since the last iteration. This can ideally be used to map the propagation of a certain network. To reduce visualization time, it is also possible to hide the nodes not connected with edges.

### 5.4 Real-time aspect

To handle the real-time aspect of this model, a timer was placed on all code to retrieve the data again after one hour. Every hour the calculation is thus done again. As described in section 4.2, a speed layer, batch layer, and serving layer are used. The speedlayer makes calculations based on a min-max scaler, instead of the cosine similarity and at the same time, the batch layer calculates the actual cosine similarity. The user sees both networks as soon as they are generated. To indicate new nodes that enter the network over time, it was decided to color these new nodes orange instead of blue. This way chances over time can be easily seen.

### 5.5 Upscaling

Scaling to a bigger region and broader timespan increases the size of the network and thereby increase computation time for both the network itself and the preprocessing. To counteract this, a certain type of scalability is needed, which can be achieved using cloud computing in clusters. This section is divided into two parts. First, a global overview is given of the architecture of the implementation in the cloud. Secondly, upscaling services are discussed.

#### 5.5.1 Architecture on the cloud

Clusters can be implemented in two different sections of the pipeline process: cosine similarity and the network loop.

The cosine similarity section 5.2 used to calculate the website event similarity can be split into groups of website pairs since the score is based on the differences between two websites. This can be optimally divided over as much as one pair per cluster computer. This creates a speedup of (***n*n - n)/2*** [optimized matrix iteration, see 5.6 performance] (all pairs on a single machine) to ***n*** (one pair on each pc). Additionally, the data for the creation of each network can be generated individually of one another (for example a network for each hour of a day). The network data can then be saved and later recreated/compared for analysis.

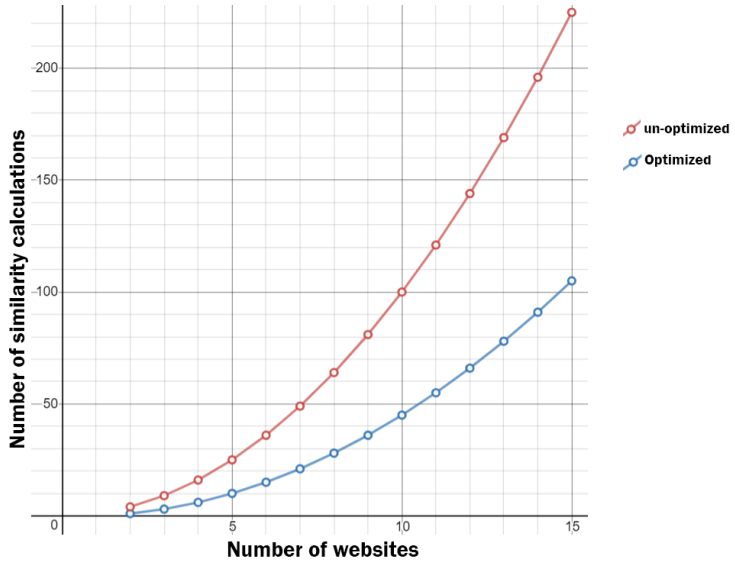
#### 5.5.2 Cloud services

Clustering can be achieved using Amazon Web Services, which is a cloud service platform that offers compute power, database storage, and other functionality to scale data. The storage is done using Amazon S3 and offers scalability, data availability, security, and performance. Amazon has already stored the GDELT data in S3, which can be queried on using a cloud dataware house services (Redshift, Hive, etc).

For pre-processing, calculating similarity and building a network functional programming is required. This can be achieved on an EC2 cluster using the spark framework, which automatically optimizes functional programs over a cluster infrastructure and can be integrated with python and its packages. The advantage of using EC2 and AWS bucket storage is that there are both relatively cheap compared to the other AWS services.

### 5.6 Performance

One of the major issues is the performance as the scope (number of events, timeframe and area/countries) increases. For example, the number of calculations needed to calculate the cosine similarity between all websites increases quadratic (**N*N**) for each added website. Optimization is used to reduce the number of calculations needed to an exponential increase as (***n*n - n)/2***, with n as the number of websites, self-calculations (A→A) and skipping doubles (A→B & B→A). This reduces the number of calculations (as illustrated in graph 1) but does not solve the issue of decreasing performance per website.




**Graph 1: Relationship number of calculation * the number of nodes**

There is also a difference in the number of events per country, with larger amounts increasing computation time of each cosine similarity calculation. Besides the increased preprocessing time, network visualization also takes longer as the data becomes larger. Some of the features implemented to increase the performance of the network are the optional exclusion of nodes without edges in combination with a (user-defined) threshold, which drops edges based on the threshold being lower than the calculated cosine similarity.

### 5.7 Cost analysis

The cost of the process is highly dependant on the scope of the visualization (as described in 5.6 Performance) and the timeframe between real-time simulations. Processing time is not an issue because the process steps are mostly scalable (see 5.5 Upscaling). The quality of the hardware is dependent on the number of calculations required and the time limit. However, since the network creation itself cannot be distributed, it is better to use more powerful computers in less quantity, to ensure that no computers are idle during the time it takes to create a network. Since the costs are process dependent, exact cost estimation is difficult. Therefore, an estimated example, with minimization suggestions, is shown below.

Example (200 nodes with a timeframe of 1 hour, real-time for 12 hours): Suppose the calculation of the preprocessing (network nodes and edges) of one website pair takes 10 seconds with the selected hardware and the network itself is created in 15 minutes. data transfer between machines also takes 1 minute. The time available for the preprocessing is thus 44 minutes. For 200 nodes, (n*n-n/2) = 19900 pair calculations are required. If, as is stated, each calculation takes 10 seconds, with 44 minutes available, this would require (19900 / (44*60/10 = ) 76 computers. This is optimized to use exactly

the time available, to ensure that computers are idle during the 12 hours real time period. However, since the creation of the network can only be done on one machine, 75 are idle, which is a waste of resources.

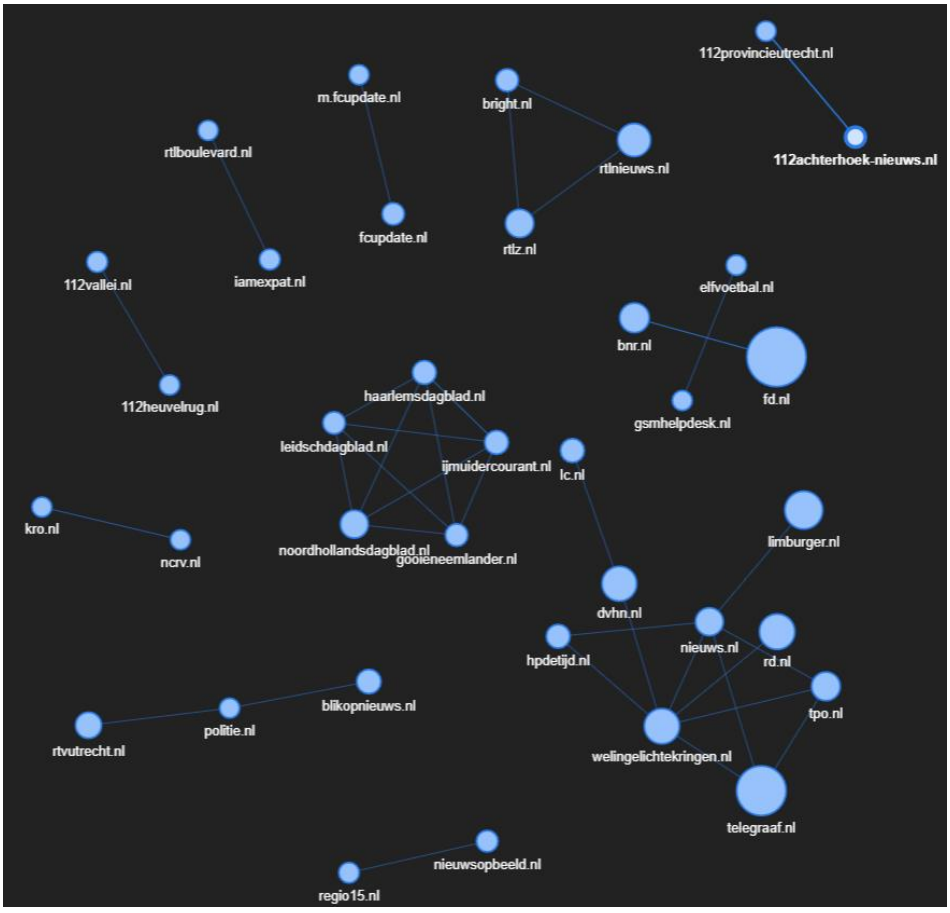
A method to ensure no idle machines is to for example create an overflow of timeframes, where the network is calculated at the start of the new hour, on one machine, while the rest of the machines are calculating the similarity for the new hour. However, this is a tradeoff since it creates a delay in the real-time simulation.

# 6. Results

This section discusses the results and interesting findings in the network visualization and is divided into two subsections: network results and (real-time) simulation results.

### 6.1 Network results

The similarity between news outlets is visualized using a network graph. To illustrate an example of a possible network, a simple, clear network is extracted. Nodes without edges are dropped. This results in 35 Dutch news sites, that have the most events in common in the first quarter of 2019. Note that this is a small-scale visualization, which can be run locally, to ensure it is clear and organized. The threshold (to exclude edges) is set high to remove less significant edges. Also, the optionality to drop edgeless nodes is enabled.




**Network 1: High ranking Dutch news websites 2019**

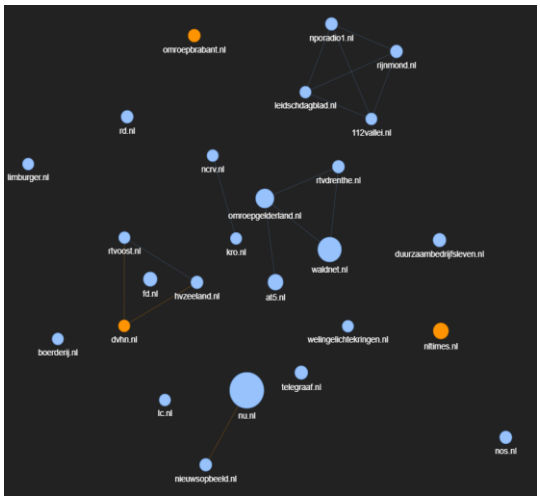
A total of 12 clusters can be identified in the resulting network (1). Using domain knowledge, a possible explanation for some of the clusters can be given. For instance, the central star-shaped figure (with i.a. 'haarlemsdagblad.nl' and 'noordhollandsdagblad.nl') is a collection of regional newspapers in the west of the Netherlands. A possible explanation for the clustering is that more local than national news is covered. This would result in a high clustering between nodes, with no out-links to other clusters, which is exactly what can be seen in the network, supporting the claim. An example of a more national news cluster is the cluster in the bottom right corner, including for example telegraaf.nl and news.nl. Since the Telegraaf node has a more than average node size it shows that it covers a large number of events, which could be a possible explanation for the clustering with other national news websites. Another interesting finding is that 112 websites (Dutch police) are connected, which may support the correctness of the similarity scores.

Remarkably, some of the biggest Dutch news outlets (nu.nl, nos.nl, and ad.nl) are not shown in the graph. This is likely caused by the news websites not reaching the similarity threshold required to be included. This might suggest that these news websites have a low fraction of overlapping news events.

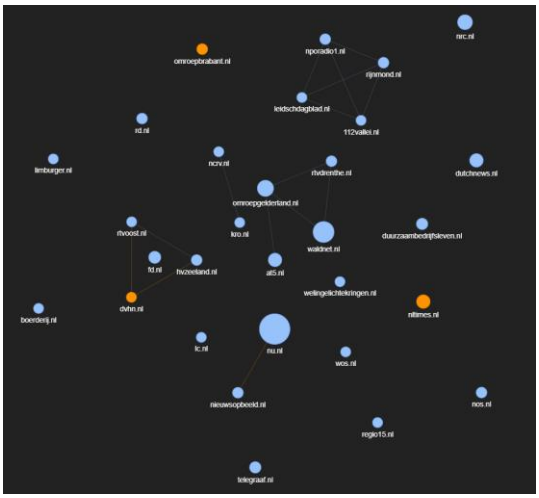
### 6.2 Simulation results

It was decided to do a real-time simulation for a time period of 24 hours. The simulation was done on data from Dutch newspapers on the 1st of January, 2019. As described in section 5.4, the data is retrieved every hour. The resulting networks of the batch layer are visible in appendix A. The newly added nodes and edges are displayed in orange.

It was found that the speed layer min max scaler indication of the similarity works pretty well. Below two images are shown: the left image is the network resulting from the speedlayer, on the right from the batch layer. As visible, a lot of the same edges are formed. The main difference is the spacing, a lot of nodes have a different place in the network. There are also a few nodes missing in the speedlayer, these nodes did not pass the threshold. It was found that on both a system that used 8 GB of ram and 16 GB of ram, using the min-max scaler was 1.5 times as fast then the computation (tried on 5 different sizes of datasets).




**Network 2: speed layer result of 9.00**




**Network 3: Batch layer result of 9.00**

The network propagation can be described as follows (visualizations in appendix A): At 01:00 the network starts off with three nodes. When hours pass new nodes are added. At 05:00 the first clear similarity is spotted: an edge between rtvoost.nl and hvzeeland.nl. At 9:00 the first clusters start to form between nodes, with more than two edges. The hours after that more nodes are added and more edges are forming. At 19:00 a big cluster is visible, indicating that a lot of Dutch news sites did write about similar events in similar

manners that day. The hours after 19:00 the amount of new nodes and new edges added to the network gets smaller than de rest of the day (only 1 or 2 per hour). Finally at 00:00, the 2nd of January 2019, a big cluster is visible in the middle, with some small clusters around it. It is visible that not all news outlets write about the same events since not all nodes are connected.

It is interesting that the news coverage of events can take some time. The cluster in the middle slowly grew, meaning that more and more news outlets write about the same event as time passes. Another interesting observation is that some news outlet do not have a significant similarity with other news sources after 24 hours. This is for example interesting for the bigger Dutch news outlets, like the nrc.nl and metronieuws.nl. This could indicate two things: they don't cover the same events, or they write so differently about the events that the cosine similarity was not high enough between those outlets and other nodes.

# 7. Conclusion and future work

This research showed that it is possible to create a model that finds similarity in reported events between different news sources in real time, based on high-level event data. The model was created using the GDELT dataset. This study thus tries to contribute to existing GDELT research by creating a framework that can visualize the data from GDELT in real time. It thereby contains the big data dimensions volume and velocity. The similarity between different data points/news outlets in GDELT was calculated by using the cosine similarity and the real-time aspect was implemented using a simple lambda structure. To make new nodes and edges emerging over time visible, new nodes and edges are indicated by using a different color. Even though the current implementation was done on a local scale, it is possible to upscale this model to use more of the data. The data can be stored in an AWS bucket, and the Jupyter notebook with the local implementation can be coupled to an EC2 cluster, by using Spark.

This framework can then be used to support further research and help other researchers in obtaining insight into how to interpret their findings. Next to that it can be used as a starting point for a lot of different researches. As discussed in the related work, news coverage can be biased and news frames often exist. Since this model already visualizes the news coverage, researchers that want to do research into specific events only have to limit the data to data from a specific date or event, and can immediately see a visualization of the news coverage. This model can thus be used as a starting point for a lot of different kinds of researches, for example for detecting plagiarism, biases and to see events progress over time. The network and framework form a base to work on and can be adapted for specific needs.

# References


Ball-Rokeach, S. J., & DeFleur, M. L. (1976). A dependency model of mass-media effects. Communication research, 3(1), 3-21.

Barlow, M. (2013). Real-time big data analytics: Emerging architecture. " O'Reilly Media, Inc.".

Boczkowski, P. J. (2009). Technology, monitoring, and imitation in contemporary news work. Communication, Culture & Critique, 2(1), 39-59.

Galla, D., & Burke, J. (2018, July). Predicting Social Unrest Using GDELT. In International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 103-116). Springer, Cham.

GDELT. (n.d.). The GDELT Story: About the GDELT Project. Retrieved March 10, 2019, from https://www.gdeltproject.org/about.html

Girard, N. J. (2004). Plagiarism: an ethical problem in the writing world. AORN journal, 80(1), 13-15.

Hofstetter, C. R., & Buss, T. F. (1978). Bias in television news coverage of political events: A methodological analysis. Journal of Broadcasting & Electronic Media, 22(4), 517-530.

Ken Hes. (2016) Hadoop vs Spark: The New Age of Big Data. https://www.datamation.com/data-center/hadoop-vs.-spark-the-new-age-of-big-data.html

Kwak, H., & An, J. (2016). Revealing the Hidden Patterns of News Photos: Analysis of Millions of News Photos Using GDELT and Deep Learning-based Vision APIs. arXiv preprint arXiv:1603.04531.

Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co..

McLeod, D. M., & Detenber, B. H. (1999). Framing effects of television news coverage of social protest. Journal of communication, 49(3), 3-23.

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting social unrest events with hidden Markov models using GDELT. Discrete Dynamics in Nature and Society, 2017.

Skvorc, B. (2014, 30 mei). Google’s BigQuery Provides Free Access to GDELT. Geraadpleegd op 26 maart 2019, van https://www.sitepoint.com/googles-bigquery-provides-free-access-gdelt/

Van Dalen, A. (2012). Structural bias in cross-national perspective: How political systems and journalism cultures influence government dominance in the news. The International Journal of Press/Politics, 17(1), 32-55.

Yonamine, J. E. (2013). Predicting future levels of violence in afghanistan districts using gdelt. Unpublished manuscript.

# APPENDIX

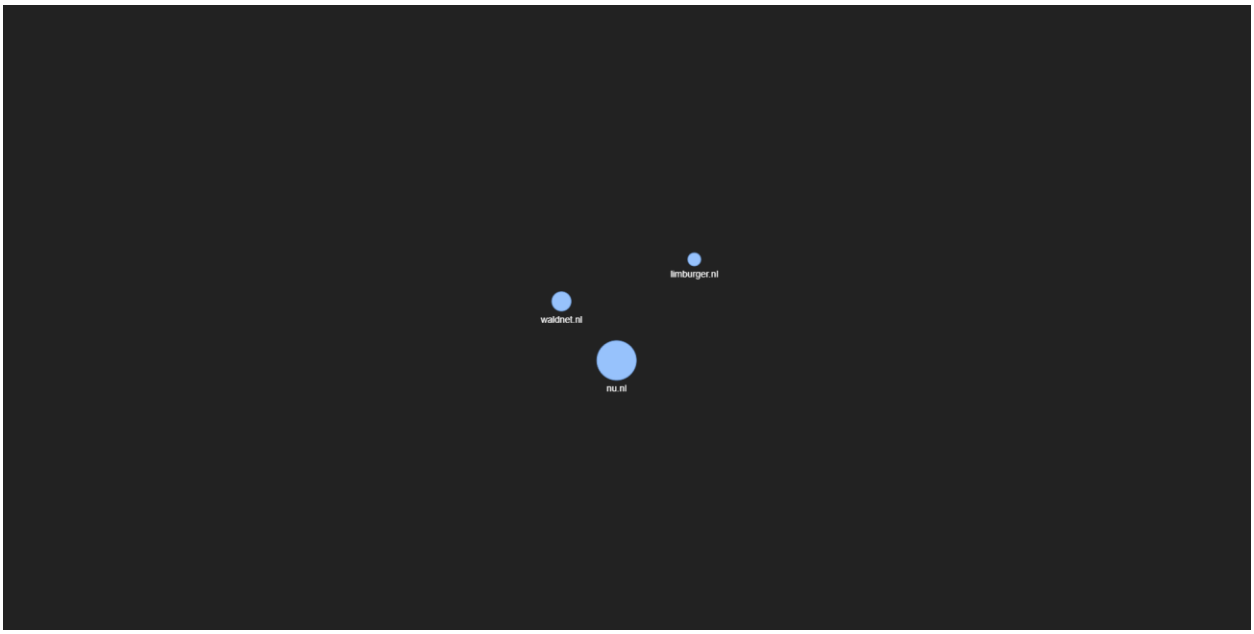
### A. Team member contributions

Sadly enough the workload was not evenly spread between all for team members. Delano did not really contribute to the project. Below the programming contributions per person are shown, with percentages how we interpreted the workload was spread. Jessy, Jeffrey and Marit did write the paper together.

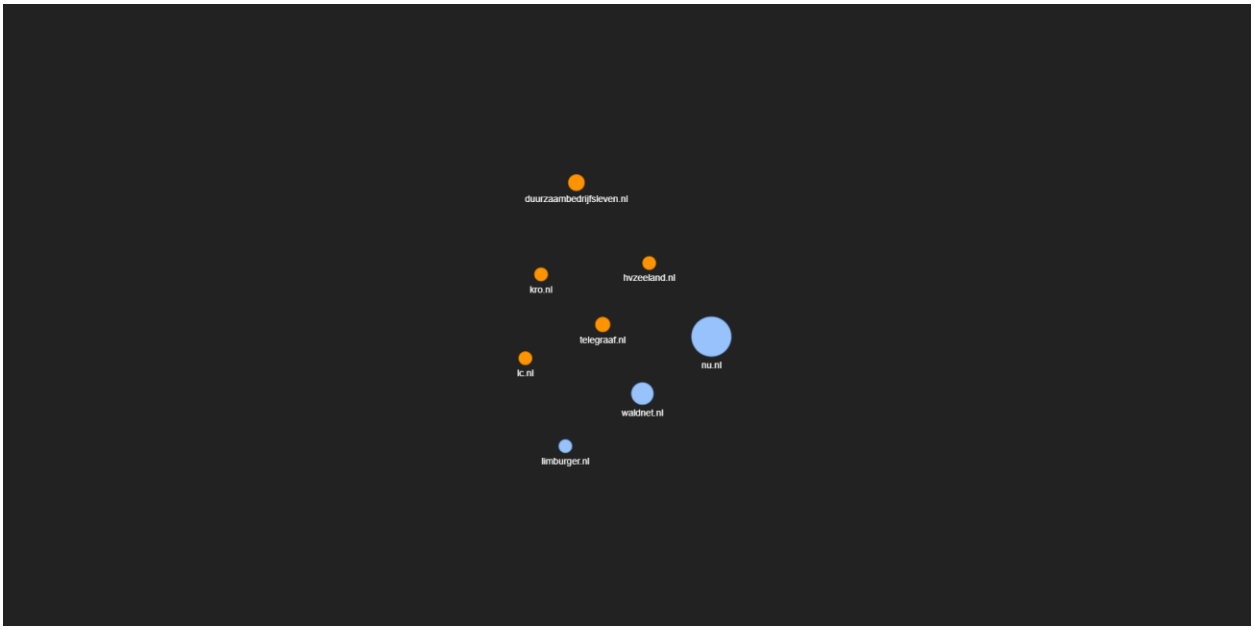
- **Jessy (~30%):**
  - Graph implementation: node size, edge size, overall looks
  - Pipeline design
- **Jeffrey (~30%):**
  - Graph implementation: similarity calculations
  - Upscaling research
- **Marit (~30%):**
  - Real-time implementation: lambda architecture
  - Pipeline design
- **Delano (~10%):**
  - Finding out how the API worked
  - Helped a bit with brainstorming in the beginning, but that's all.

## B. Real time simulation

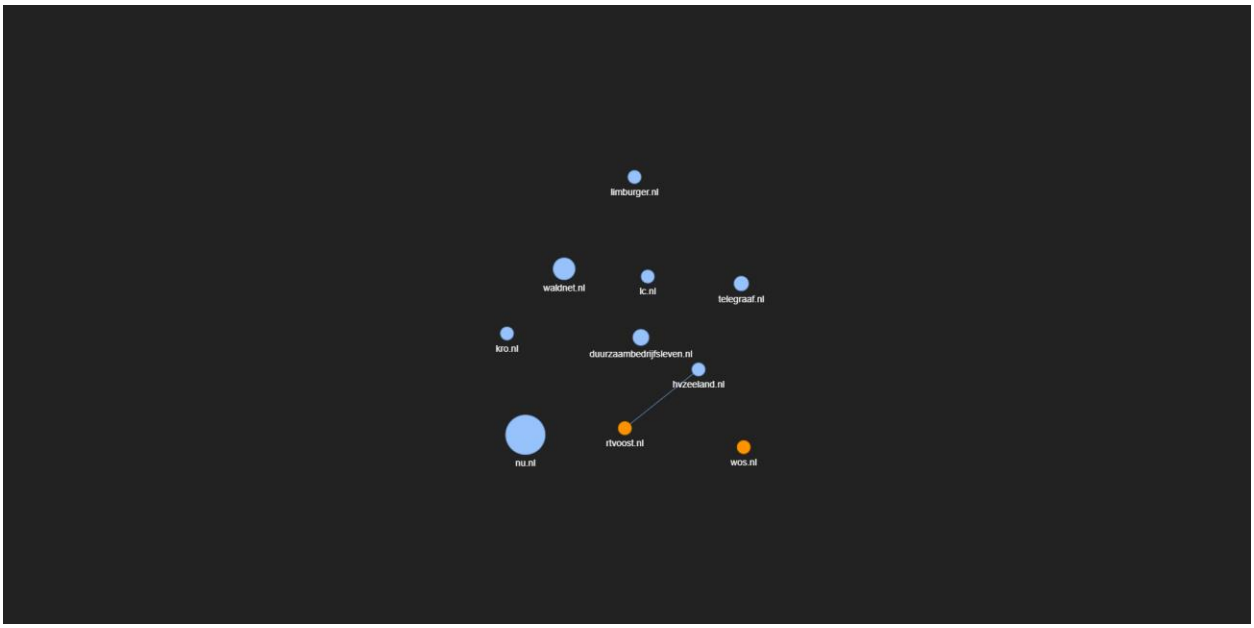
The time of the network is displayed below every image.




**01-01-2019 - 01:00**



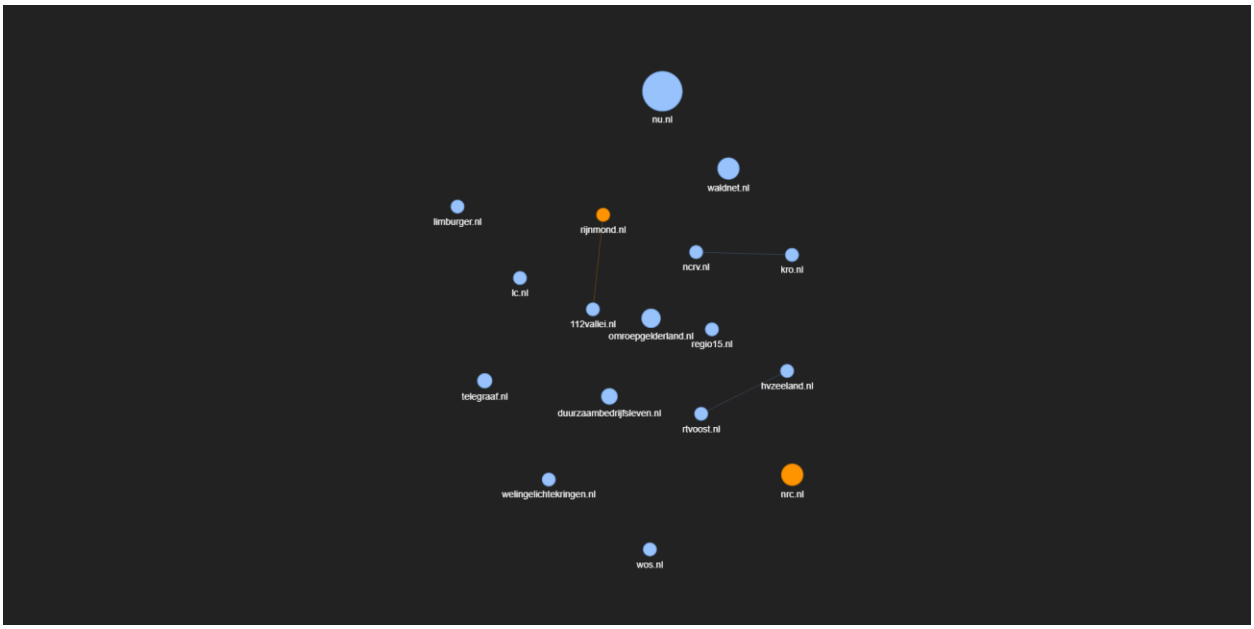

**02:00**


limburger.nl
waldnet.nl
lc.nl
telegraaf.nl
kro.nl
duurzaambedrijfsleven.nl
hvzeeland.nl
rtvoost.nl
nu.nl
wos.nl

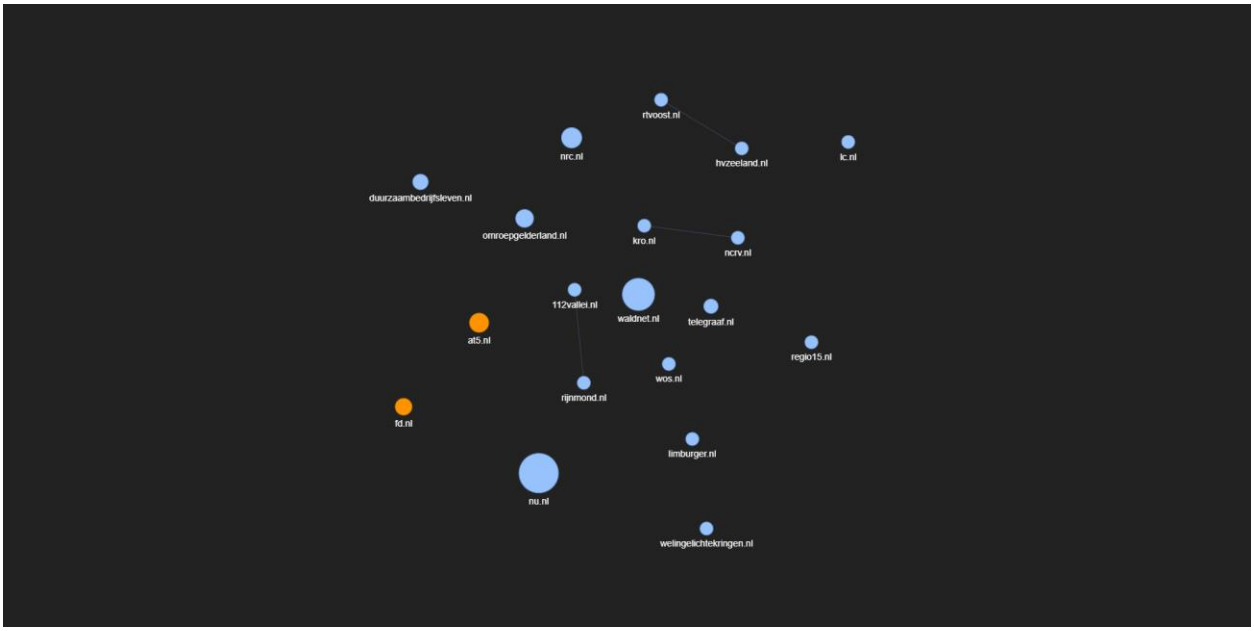
**03:00**


telegraaf.nl
regio15.nl
duurzaambedrijfsleven.nl
kro.nl
welingelichtekringen.nl
ncrv.nl
lc.nl
waldnet.nl
wos.nl
omroepgelderland.nl
hvzeeland.nl
rtvoost.nl
limburger.nl
nu.nl
112vallei.nl

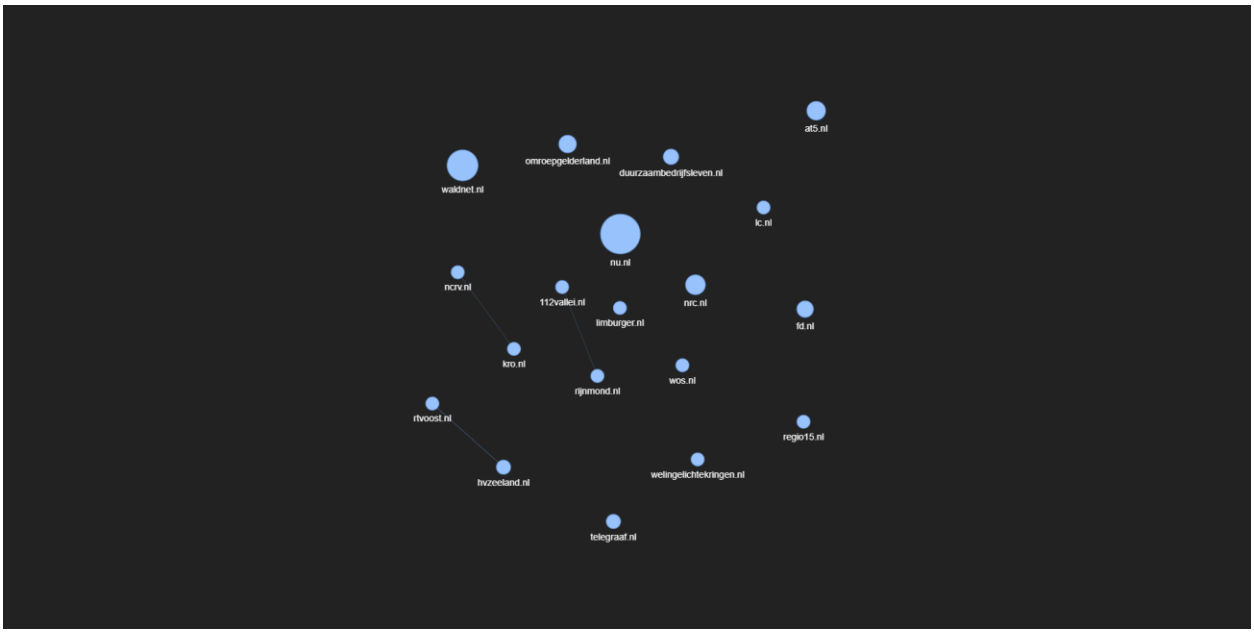
**04:00**


nu.nl
waldnet.nl
limburger.nl
rijnmond.nl
ncrv.nl
kro.nl
lc.nl
112vallei.nl
omroepgelderland.nl
regio15.nl
telegraaf.nl
hvzeeland.nl
duurzaambedrijfsleven.nl
rtvoost.nl
nrc.nl
welingelichtekringen.nl
wos.nl

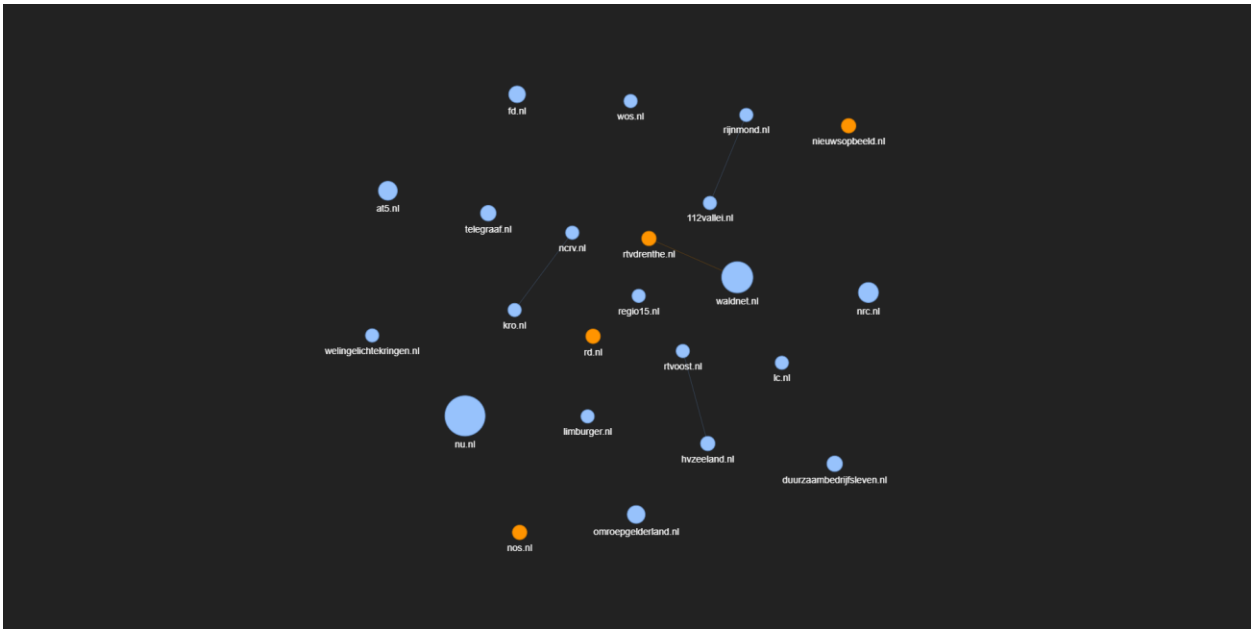
**05:00**


rtvoost.nl
nrc.nl
hvzeeland.nl
lc.nl
duurzaambedrijfsleven.nl
omroepgelderland.nl
kro.nl
ncrv.nl
112vallei.nl
waldnet.nl
telegraaf.nl
at5.nl
regio15.nl
wos.nl
rijnmond.nl
fd.nl
limburger.nl
nu.nl
welingelichtekringen.nl

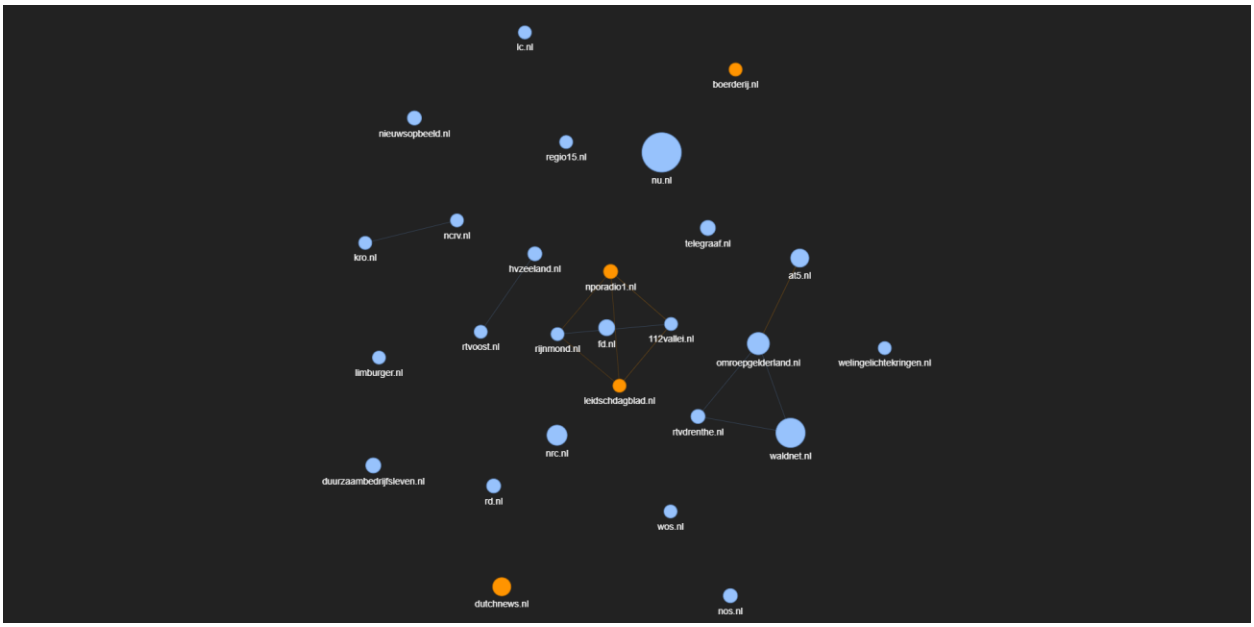
**06:00**


at5.nl
omroepgelderland.nl
duurzaambedrijfsleven.nl
waldnet.nl
lc.nl
nu.nl
ncrv.nl
112vallei.nl
nrc.nl
limburger.nl
fd.nl
kro.nl
wos.nl
rijnmond.nl
rtvoost.nl
regio15.nl
hvzeeland.nl
welingelichtekringen.nl
telegraaf.nl

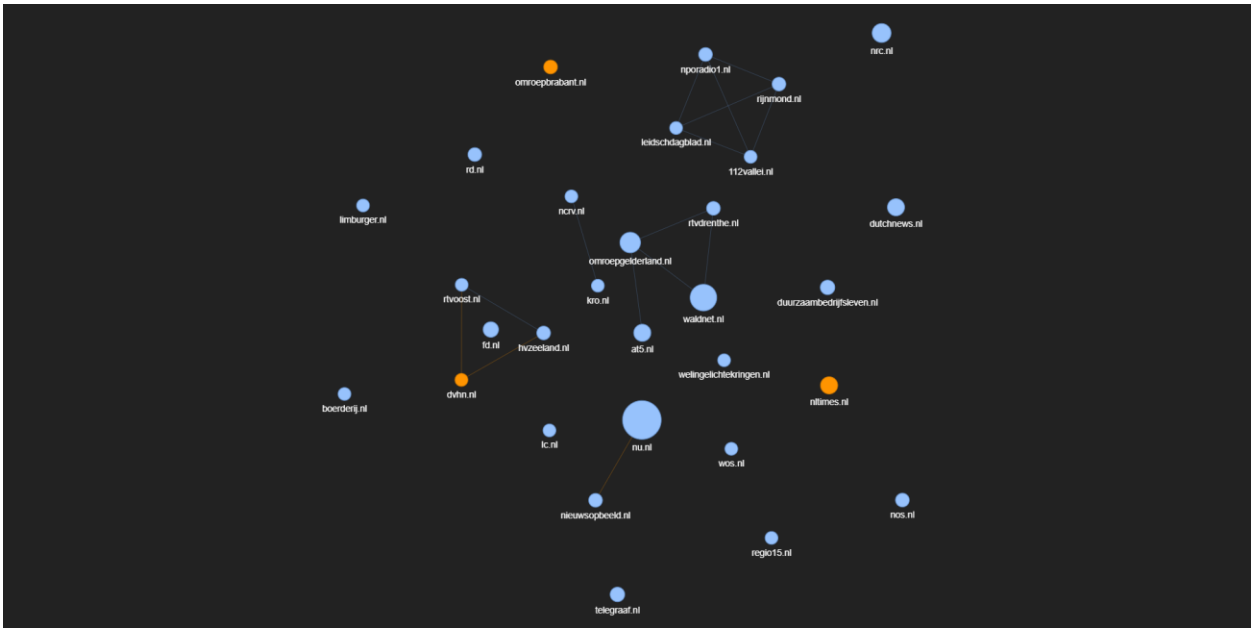
**07:00**


fd.nl
wos.nl
rijnmond.nl
nieuwsopbeeld.nl
at5.nl
telegraaf.nl
112vallei.nl
ncrv.nl
rtvdrenthe.nl
waldnet.nl
nrc.nl
regio15.nl
kro.nl
welingelichtekringen.nl
rd.nl
rtvoost.nl
lc.nl
nu.nl
limburger.nl
hvzeeland.nl
duurzaambedrijfsleven.nl
omroepgelderland.nl
nos.nl

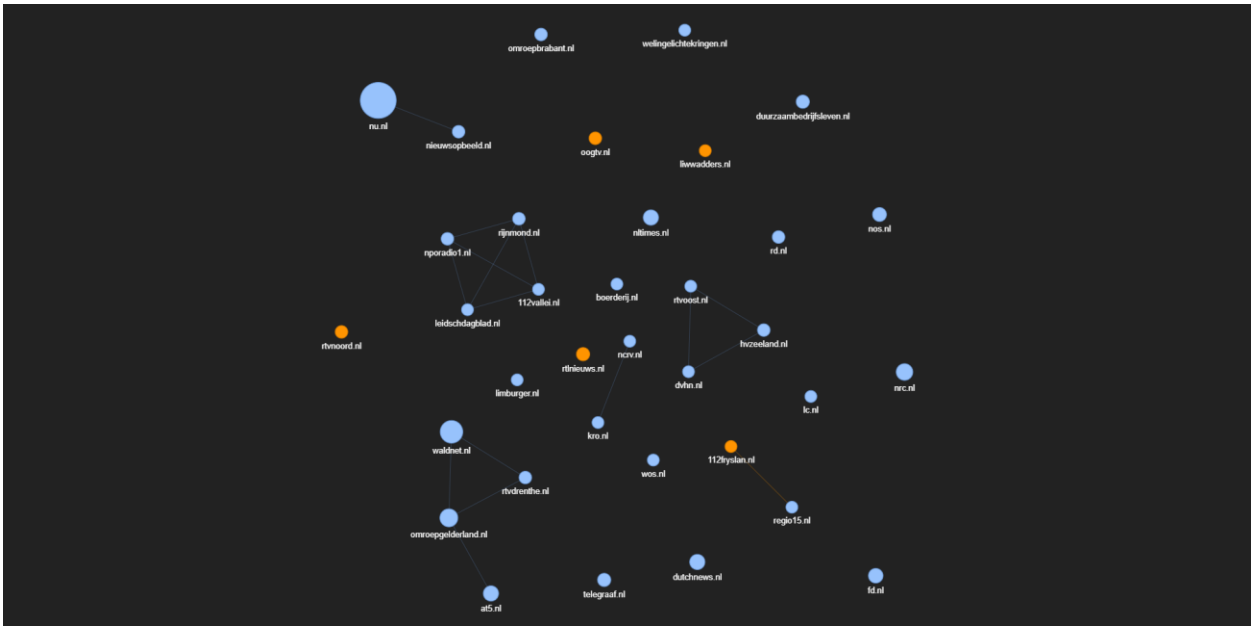
**08:00**


lc.nl
boerderij.nl
nieuwsopbeeld.nl
regio15.nl
nu.nl
ncrv.nl
kro.nl
telegraaf.nl
hvzeeland.nl
at5.nl
nporadio1.nl
rtvoost.nl
rijnmond.nl
fd.nl
112vallei.nl
omroepgelderland.nl
wellingelichtekringen.nl
limburger.nl
leidschdagblad.nl
rtvdrenthe.nl
waldnet.nl
nrc.nl
duurzaambedrijfsleven.nl
rd.nl
wos.nl
dutchnews.nl
nos.nl

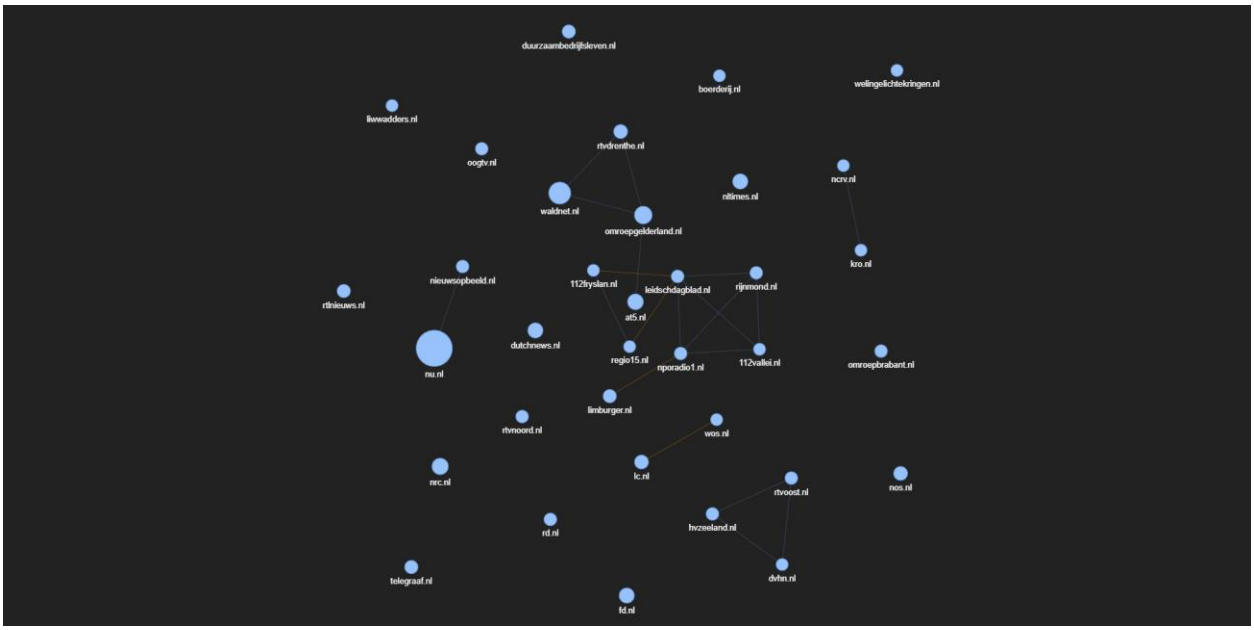
**09:00**


nrc.nl
omroepbrabant.nl
nporadio1.nl
rijnmond.nl
leidschdagblad.nl
rd.nl
112vallei.nl
limburger.nl
ncrv.nl
rtvdrenthe.nl
dutchnews.nl
omroepgelderland.nl
rtvoost.nl
kro.nl
waldnet.nl
duurzaambedrijfsleven.nl
fd.nl
hvzeeland.nl
at5.nl
wellingelichtekringen.nl
dvhn.nl
nltimes.nl
boerderij.nl
lc.nl
nu.nl
wos.nl
nieuwsopbeeld.nl
nos.nl
regio15.nl
telegraaf.nl

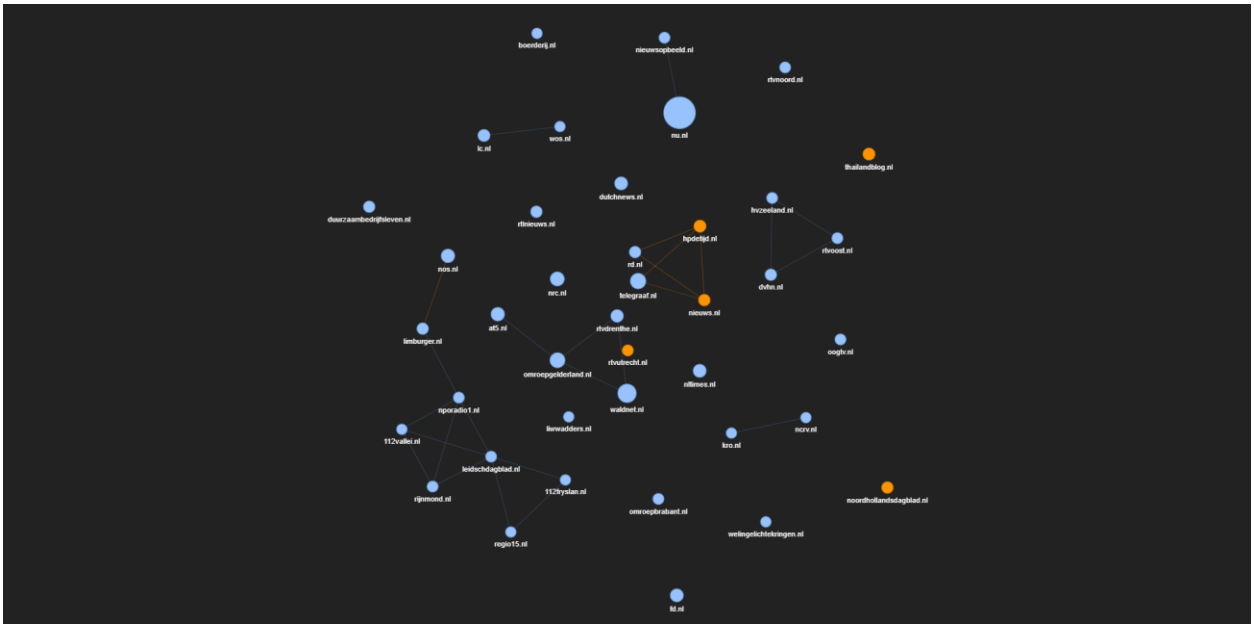
**10:00**


omroepbrabant.nl
welingelichtekringen.nl
nu.nl
duurzaambedrijfsleven.nl
nieuwsopbeeld.nl
oogtv.nl
liwwadders.nl
rijnmond.nl
nltimes.nl
nos.nl
nporadio1.nl
rd.nl
112vallei.nl
boerderij.nl
rtvoost.nl
leidschdagblad.nl
rtvnoord.nl
hvzeeland.nl
ncrv.nl
rtlnieuws.nl
dvhn.nl
nrc.nl
limburger.nl
lc.nl
kro.nl
waldnet.nl
112tryslan.nl
wos.nl
rtvdrenthe.nl
regio15.nl
omroepgelderland.nl
dutchnews.nl
telegraaf.nl
fd.nl
at5.nl

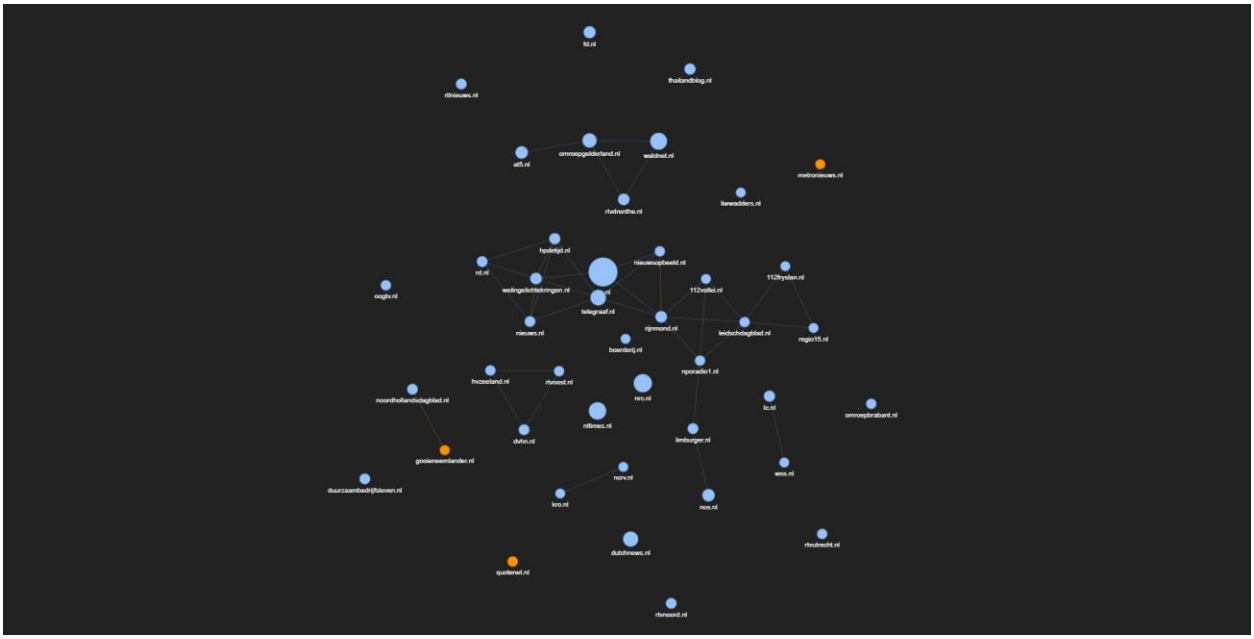
**11:00**


duurzaambedrijfsleven.nl
boerderij.nl
welingelichtekringen.nl
liwwadders.nl
rtvdrenthe.nl
oogtv.nl
ncrv.nl
nltimes.nl
waldnet.nl
omroepgelderland.nl
kro.nl
nieuwsopbeeld.nl
112tryslan.nl
leidschdagblad.nl
rijnmond.nl
rtlnieuws.nl
at5.nl
dutchnews.nl
regio15.nl
112vallei.nl
omroepbrabant.nl
nu.nl
nporadio1.nl
limburger.nl
rtvnoord.nl
wos.nl
nrc.nl
lc.nl
rtvoost.nl
nos.nl
rd.nl
hvzeeland.nl
telegraaf.nl
dvhn.nl
fd.nl

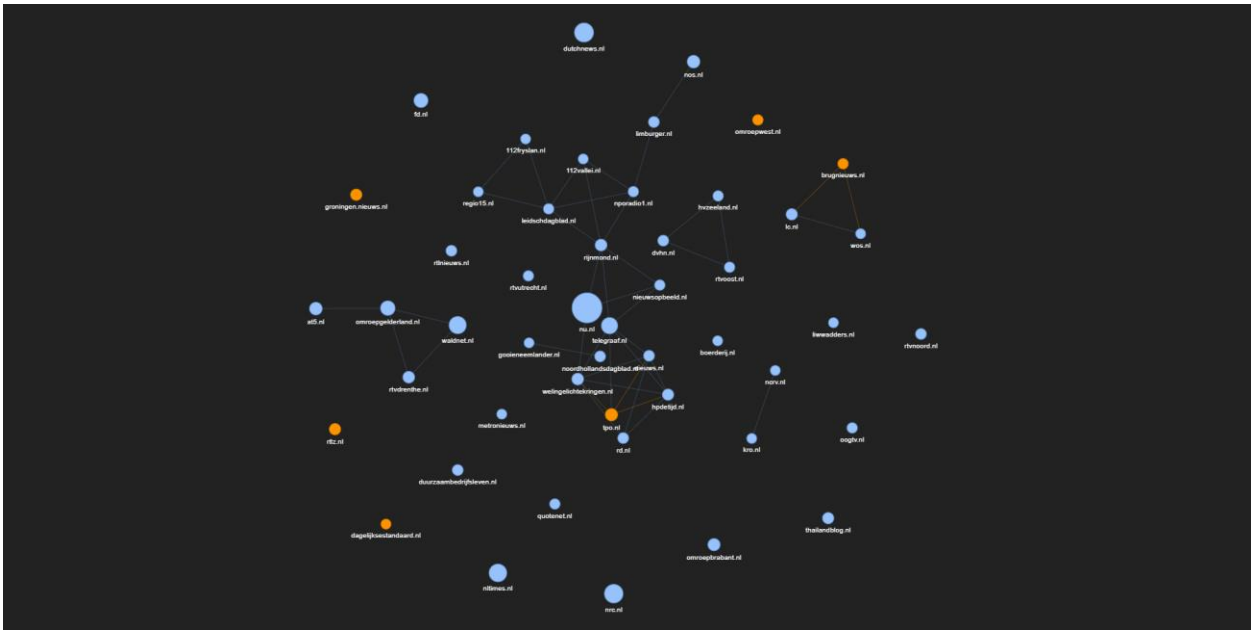
**12:00**


boerderij.nl
nieuwsopbeeld.nl
rtvmoord.nl
nu.nl
wos.nl
lc.nl
thailandblog.nl
dutchnews.nl
duurzaambedrijfsleven.nl
rtlnieuws.nl
hvzeeland.nl
hpdeltijd.nl
rtvoost.nl
nos.nl
nrc.nl
telegraaf.nl
dvhn.nl
nieuws.nl
at5.nl
rtvdrenthe.nl
limburger.nl
rtvutrecht.nl
omroepgelderland.nl
ooghr.nl
nltimes.nl
waldnet.nl
nporadio1.nl
liewadders.nl
ncrv.nl
112vallei.nl
kro.nl
leidschdagblad.nl
rijnmond.nl
112tyslan.nl
noordhollandsdagblad.nl
omroepbrabant.nl
weringelichtekringen.nl
regio15.nl
fd.nl

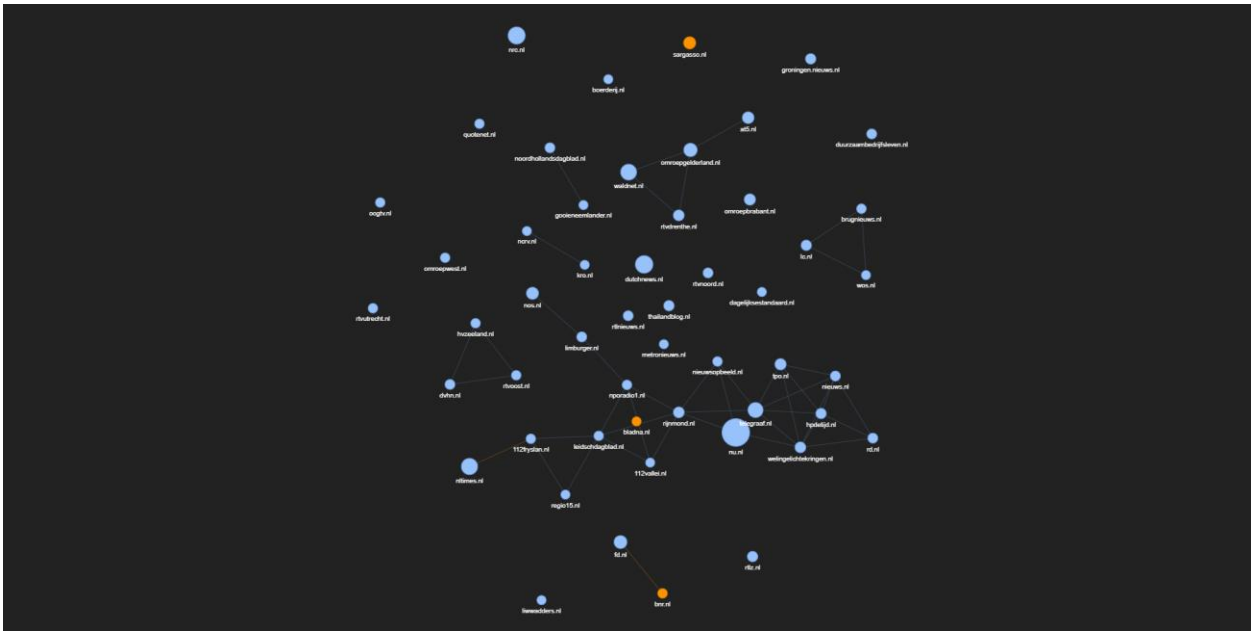
**13:00**


fd.nl
thailandblog.nl
rtlnieuws.nl
at5.nl
omroepgelderland.nl
waldnet.nl
metronieuws.nl
rtvdrenthe.nl
liewadders.nl
hpdeltijd.nl
nl.nl
nieuwsopbeeld.nl
weringelichtekringen.nl
112tyslan.nl
ooghr.nl
112vallei.nl
telegraaf.nl
nieuws.nl
rijnmond.nl
leidschdagblad.nl
regio15.nl
boerderij.nl
nporadio1.nl
hvzeeland.nl
rtvoost.nl
nrc.nl
noordhollandsdagblad.nl
lc.nl
omroepbrabant.nl
nltimes.nl
dvhn.nl
limburger.nl
goeieneenlander.nl
wos.nl
ncrv.nl
duurzaambedrijfsleven.nl
kro.nl
nos.nl
dutchnews.nl
rtvutrecht.nl
quitenet.nl
rtvmoord.nl

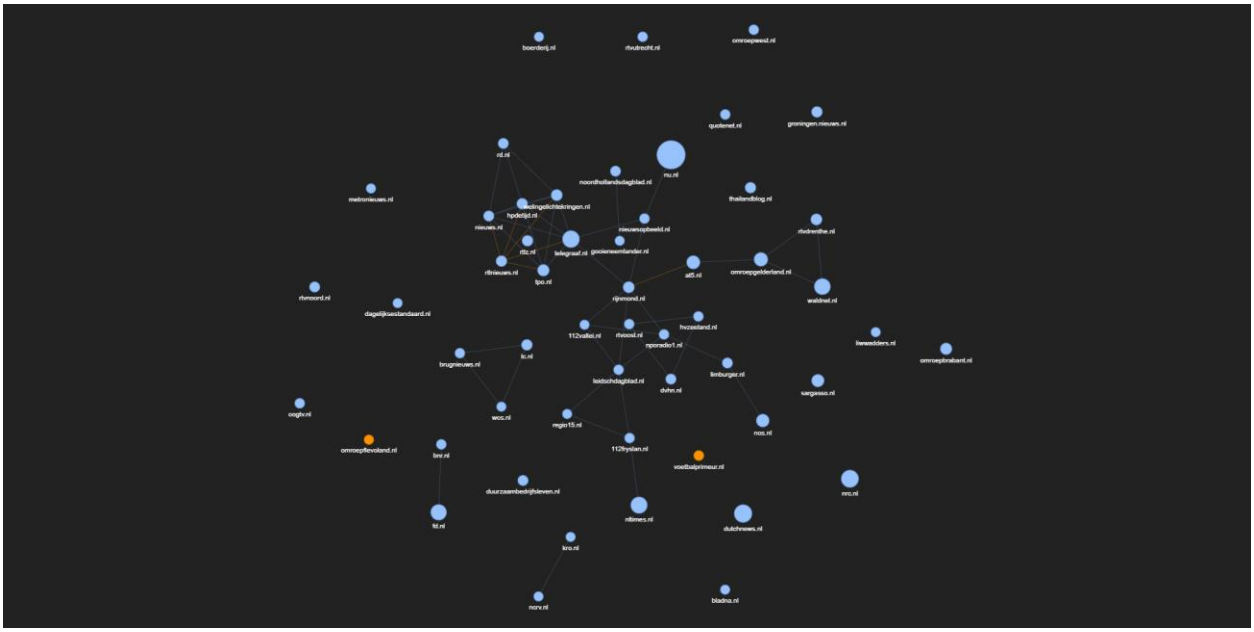
**14:00**


dutchnews.nl
nos.nl
ftd.nl
limburger.nl
omroepwest.nl
112hyslan.nl
112vallei.nl
brugnieuws.nl
groningen.nieuws.nl
regio15.nl
nporadio1.nl
hvzeeland.nl
leidschdagblad.nl
lc.nl
wos.nl
rijnmond.nl
dvhn.nl
rtlnieuws.nl
rtvoost.nl
rtvutrecht.nl
nieuwsopbeeld.nl
at5.nl
omroepgelderland.nl
waldnet.nl
nu.nl
telegraaf.nl
liewadders.nl
rtvnoord.nl
gooieneemlander.nl
boerderij.nl
noordhollandsdagblad.nieuws.nl
rtvdrenthe.nl
weringelichtekringen.nl
norv.nl
hpdelijst.nl
rtlz.nl
metronieuws.nl
tpo.nl
rd.nl
kro.nl
oogtv.nl
duurzaambedrijfsleven.nl
quotenet.nl
dagelijksestandaard.nl
thailandblog.nl
omroepbrabant.nl
rtlltmes.nl
nrc.nl

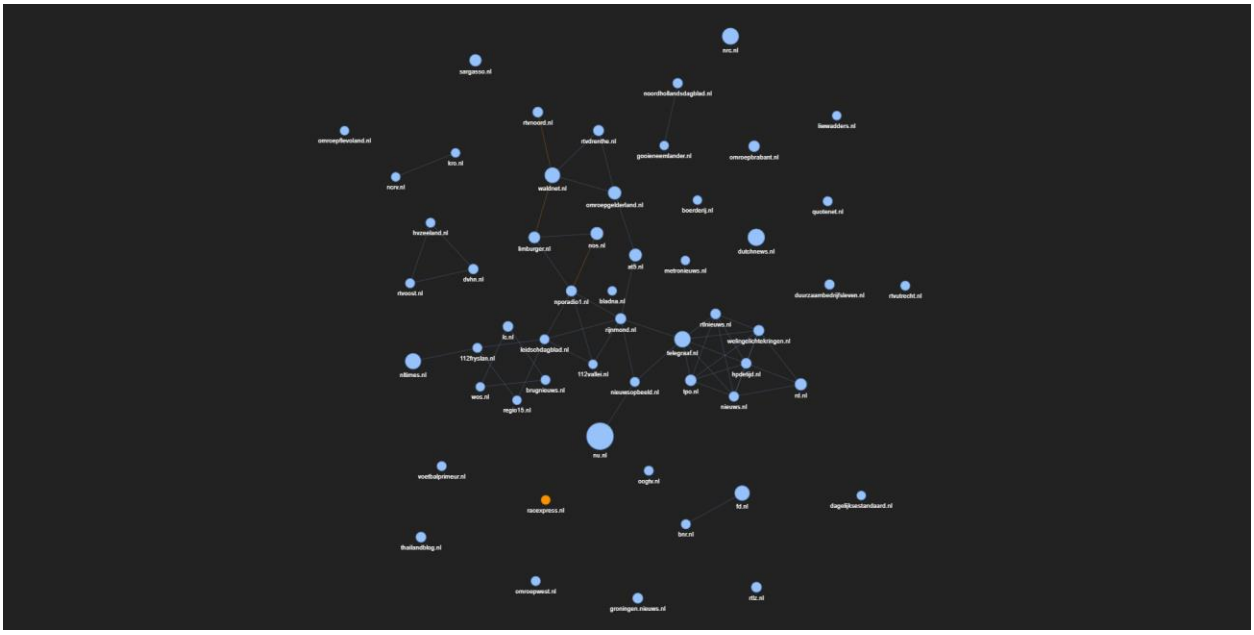
**15:00**


nrc.nl
sargasso.nl
groningen.nieuws.nl
boerderij.nl
quotenet.nl
at5.nl
duurzaambedrijfsleven.nl
noordhollandsdagblad.nl
omroepgelderland.nl
waldnet.nl
oogtv.nl
gooieneemlander.nl
omroepbrabant.nl
rtvdrenthe.nl
brugnieuws.nl
norv.nl
lc.nl
omroepwest.nl
kro.nl
dutchnews.nl
rtvnoord.nl
wos.nl
dagelijksestandaard.nl
nos.nl
rtvutrecht.nl
thailandblog.nl
rtlnieuws.nl
hvzeeland.nl
limburger.nl
metronieuws.nl
nieuwsopbeeld.nl
tpo.nl
dvhn.nl
rtvoost.nl
nieuws.nl
nporadio1.nl
rijnmond.nl
telegraaf.nl
hpdelijst.nl
bndnu.nl
leidschdagblad.nl
nu.nl
rd.nl
112hyslan.nl
weringelichtekringen.nl
rtlltmes.nl
112vallei.nl
regio15.nl
ftd.nl
rtlz.nl
liewadders.nl
bnr.nl

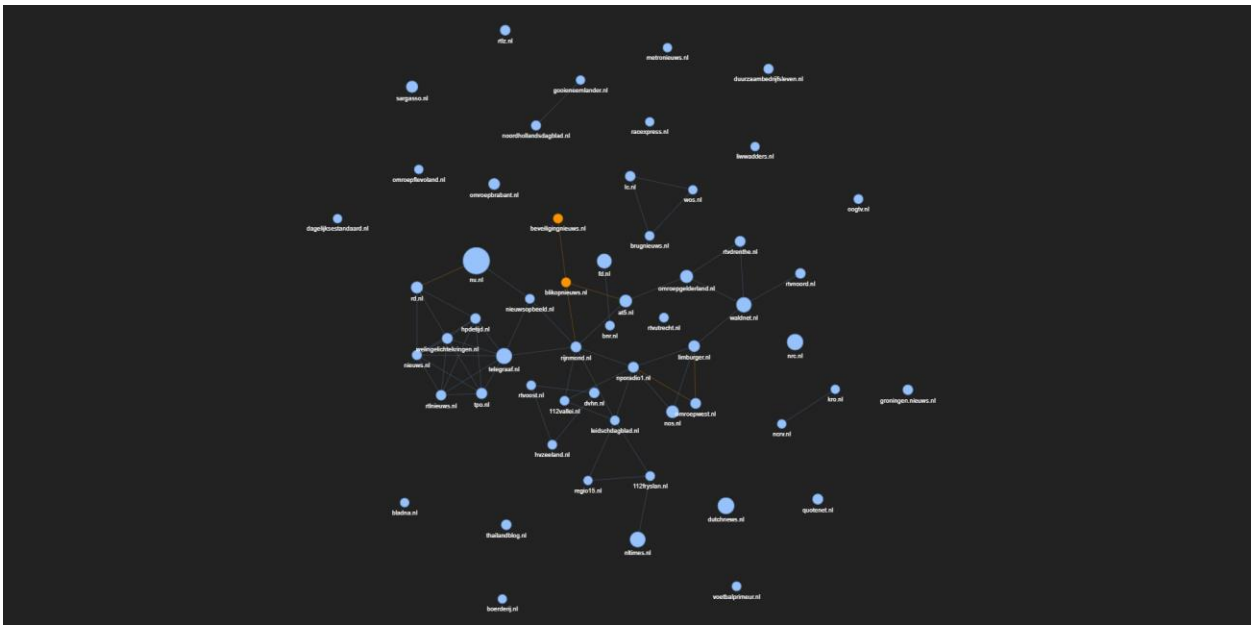
**16:00**


boerderij.nl
rtvutrecht.nl
omroepwest.nl
quotenet.nl
groningen.nieuws.nl
nl.nl
nu.nl
noordhollandsdagblad.nl
metronieuws.nl
thailandblog.nl
hpdeltijd.nl
nieuws.nl
rtvdrenthe.nl
rtlz.nl
telegraaf.nl
gooieneemlander.nl
rtvnieuws.nl
tpo.nl
at5.nl
omroepgelderland.nl
rijnmond.nl
waldnet.nl
rtvmoord.nl
dagelijksestandaard.nl
112valleii.nl
rtvoost.nl
hvzseeland.nl
nporadio1.nl
liewadders.nl
omroepbrabant.nl
brugnieuws.nl
ic.nl
limburger.nl
leidschdagblad.nl
dvhn.nl
sargasso.nl
oogle.nl
wcs.nl
regio15.nl
nos.nl
112hyslan.nl
omroepflevoland.nl
bnr.nl
voetbalprimeur.nl
duurzaambedrijfsleven.nl
nrc.nl
ntlmes.nl
dutchnews.nl
fd.nl
kro.nl
norv.nl
bladnu.nl

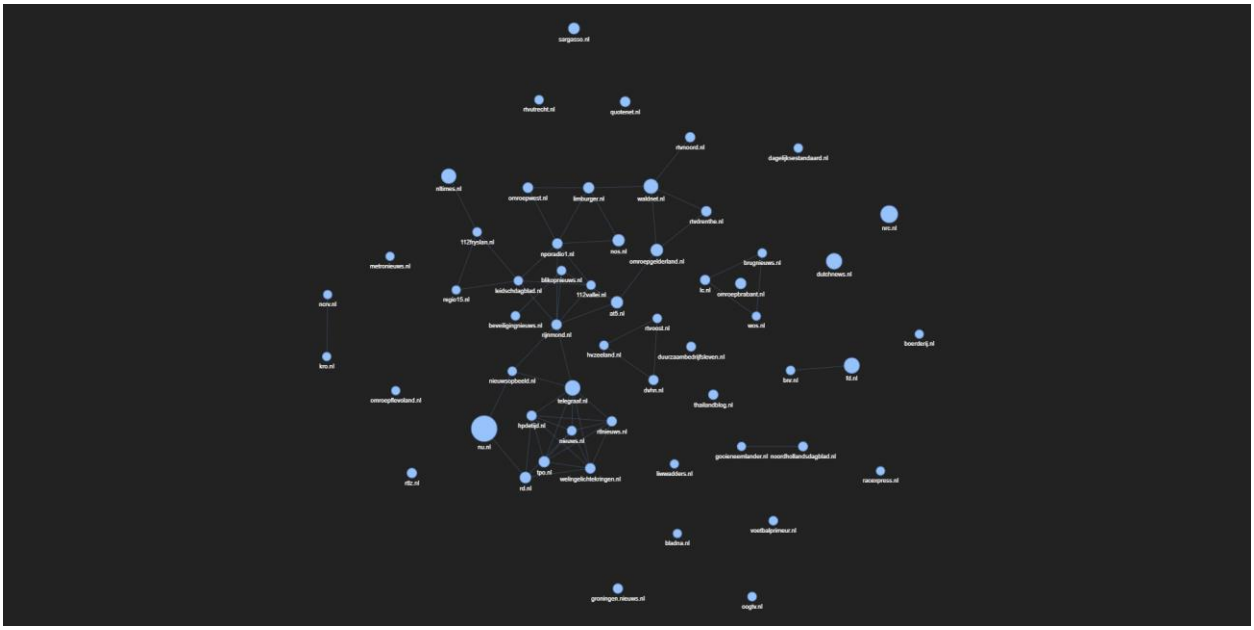
**17:00**


sxs.nl
sargasso.nl
noordhollandsdagblad.nl
rtvmoord.nl
liewadders.nl
omroepflevoland.nl
rtvdrenthe.nl
kro.nl
gooieneemlander.nl
omroepbrabant.nl
norv.nl
waldnet.nl
omroepgelderland.nl
boerderij.nl
quotenet.nl
hvzseeland.nl
limburger.nl
nos.nl
at5.nl
dutchnews.nl
dvhn.nl
metronieuws.nl
rtvoost.nl
nporadio1.nl
bladnu.nl
duurzaambedrijfsleven.nl
rtvutrecht.nl
rijnmond.nl
rtvnieuws.nl
ic.nl
wolingelichtekringen.nl
leidschdagblad.nl
telegraaf.nl
112hyslan.nl
ntlmes.nl
112valleii.nl
hpdeltijd.nl
brugnieuws.nl
nieuwsopbeeld.nl
tpo.nl
nl.nl
wcs.nl
nieuws.nl
regio15.nl
nu.nl
voetbalprimeur.nl
oogle.nl
fd.nl
dagelijksestandaard.nl
racexpress.nl
bnr.nl
thailandblog.nl
omroepwest.nl
rtlz.nl
groningen.nieuws.nl

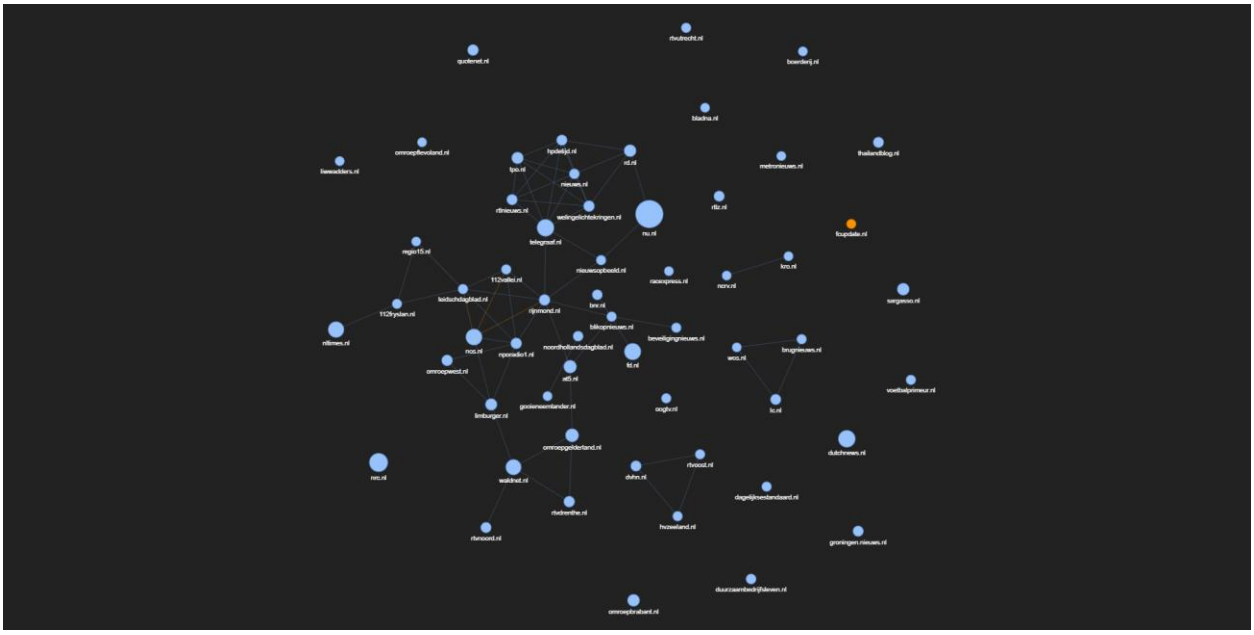
**18:00**


rtlz.nl
metronieuws.nl
duurzaambedrijfsleven.nl
sargasso.nl
gooienoordenlander.nl
noordhollandsdagblad.nl
racenpress.nl
liewoddens.nl
omroepflevoland.nl
omroepbrabant.nl
tc.nl
wos.nl
oogtv.nl
dagelijkseostendaard.nl
bevelligingsnieuws.nl
brugnieuws.nl
rtvdrenthe.nl
nu.nl
fd.nl
rtvnoord.nl
nl.nl
bilkopnieuws.nl
omroepgelderland.nl
nieuwsopbeeld.nl
at5.nl
weldnet.nl
hpdetijd.nl
rtvutrecht.nl
bnr.nl
weringelichtekringen.nl
rijnmond.nl
limburger.nl
rnc.nl
nieuws.nl
telegraaf.nl
rtvoost.nl
rtvnieuwsnl
tpo.nl
dvhn.nl
112vallei.nl
omroepwest.nl
nos.nl
kro.nl
groningen.nieuws.nl
leidschdagblad.nl
norn.nl
hvzoeland.nl
regio15.nl
112rijnlan.nl
quotenet.nl
bladna.nl
dutchnews.nl
thailandblog.nl
nltimes.nl
voetbalprimeur.nl
boerderij.nl

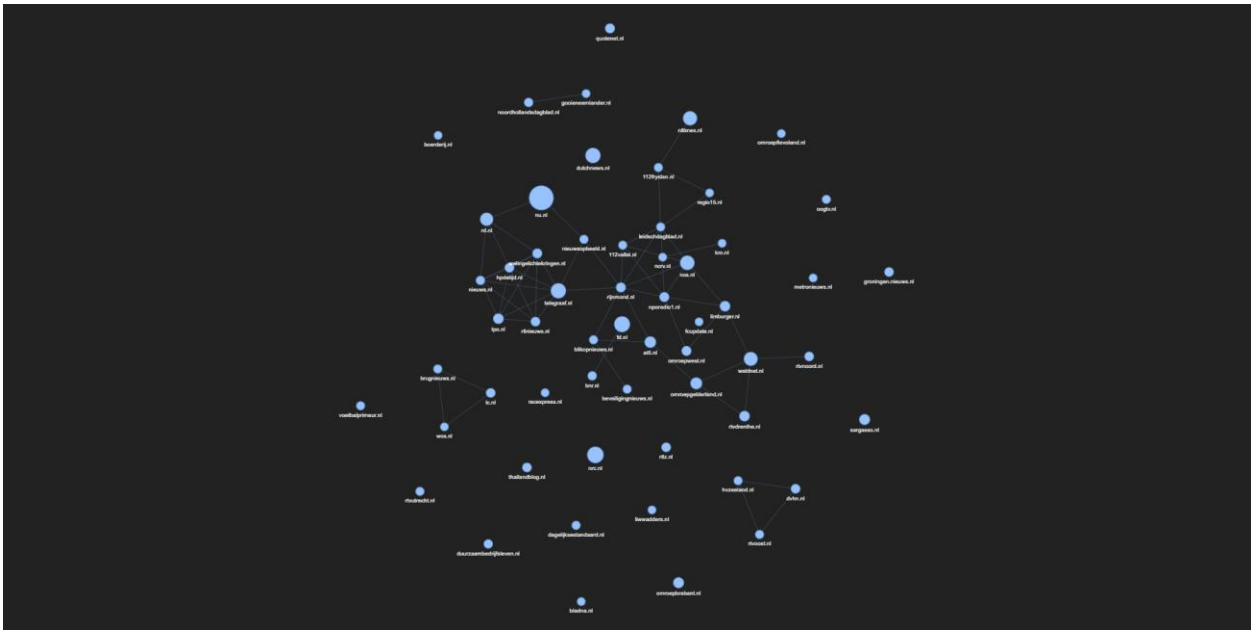
**19:00**


sargasso.nl
rtvutrecht.nl
quotenet.nl
rtvnoord.nl
dagelijkseostendaard.nl
nltimes.nl
omroepwest.nl
limburger.nl
weldnet.nl
rtvdrenthe.nl
rnc.nl
112rijnlan.nl
nos.nl
omroepgelderland.nl
brugnieuws.nl
dutchnews.nl
metronieuws.nl
bilkopnieuws.nl
tc.nl
omroepbrabant.nl
regio15.nl
leidschdagblad.nl
112vallei.nl
at5.nl
wos.nl
norn.nl
beveiligingsnieuws.nl
rijnmond.nl
rtvoost.nl
boerderij.nl
kro.nl
hvzoeland.nl
duurzaambedrijfsleven.nl
bnr.nl
fd.nl
nieuwsopbeeld.nl
dvhn.nl
omroepflevoland.nl
telegraaf.nl
thailandblog.nl
hpdetijd.nl
rtvnieuwsnl
nu.nl
nieuws.nl
gooienoordenlander.nl
noordhollandsdagblad.nl
rtlz.nl
tpo.nl
weringelichtekringen.nl
liewoddens.nl
racenpress.nl
nl.nl
voetbalprimeur.nl
bladna.nl
groningen.nieuws.nl
oogtv.nl

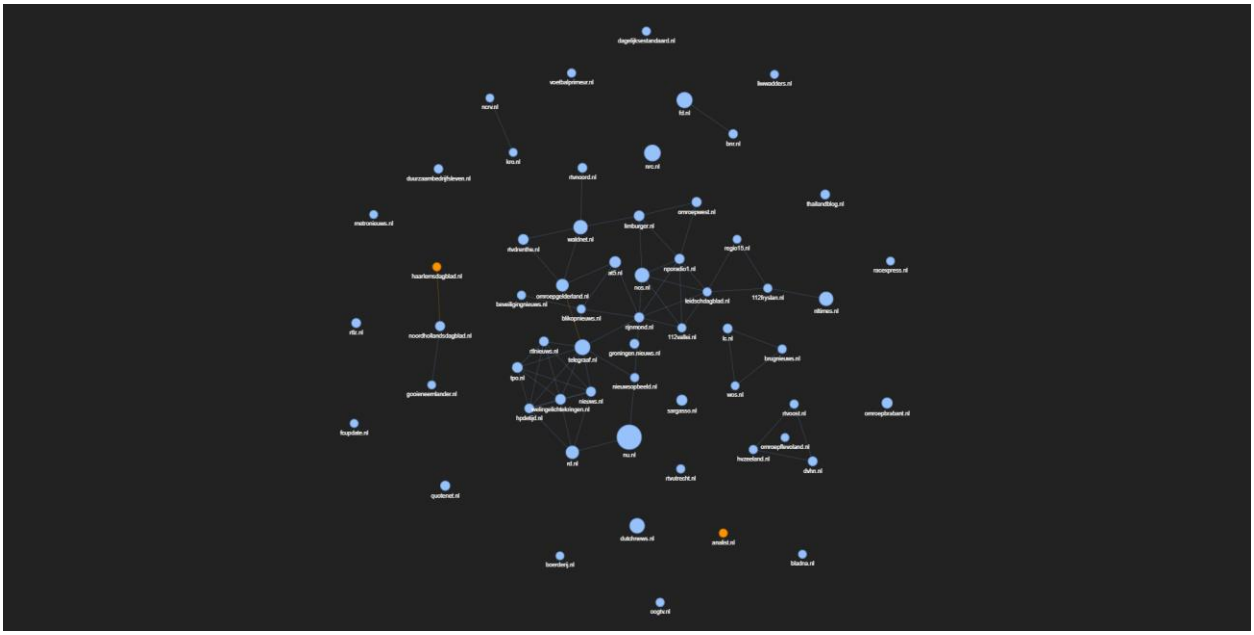
**20:00**


rtvutrecht.nl
quotenet.nl
boerderij.nl
bladna.nl
omroepflevoland.nl
hpdelijd.nl
nl.nl
thailandblog.nl
lawaaiders.nl
tpo.nl
metronieuws.nl
nieuws.nl
rtlz.nl
rtlnieuws.nl
weringelichtekringen.nl
nu.nl
telegraaf.nl
regio15.nl
nieuwsopheald.nl
kro.nl
112vallei.nl
racexpress.nl
nonv.nl
leidschdagblad.nl
sargasso.nl
1129ysian.nl
rijnmond.nl
bnr.nl
nltimes.nl
bilkopnieuws.nl
beveiligingnieuws.nl
nos.nl
noordhollandsdagblad.nl
wos.nl
brugnieuws.nl
nporadio1.nl
omroepwest.nl
ftd.nl
at5.nl
voetbalprimeur.nl
limburger.nl
groeneeemlander.nl
oogtv.nl
tc.nl
omroepgelderland.nl
dutchnews.nl
rtvoost.nl
nrc.nl
waldnet.nl
dvhn.nl
degelijkoostendaard.nl
rtvdrenthe.nl
hvzoeland.nl
rtvnoord.nl
groningen.nieuws.nl
duurzaambedrijfsleven.nl
omroepbrabant.nl

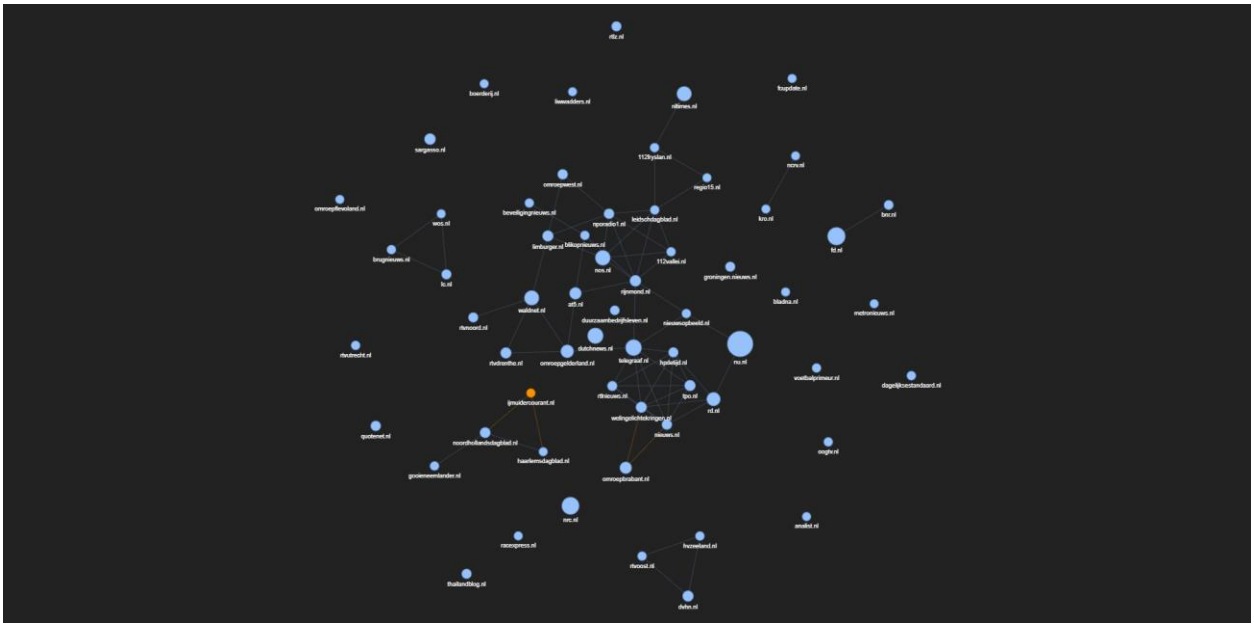
**21:00**


quotenet.nl
groeneeemlander.nl
noordhollandsdagblad.nl
rtlnieuws.nl
boerderij.nl
omroepflevoland.nl
dutchnews.nl
112vallei.nl
regio15.nl
nu.nl
oogtv.nl
nl.nl
leidschdagblad.nl
nieuwsopheald.nl
bnr.nl
nonv.nl
nos.nl
weringelichtekringen.nl
hpdelijd.nl
metronieuws.nl
groningen.nieuws.nl
nieuws.nl
telegraaf.nl
rijnmond.nl
nporadio1.nl
limburger.nl
tpo.nl
rtlnieuws.nl
ftd.nl
liveupdate.nl
bilkopnieuws.nl
at5.nl
omroepwest.nl
waldnet.nl
rtvnoord.nl
brugnieuws.nl
bnr.nl
tc.nl
racexpress.nl
beveiligingnieuws.nl
omroepgelderland.nl
voetbalprimeur.nl
wos.nl
rtvdrenthe.nl
sargasso.nl
rtlz.nl
nrc.nl
thailandblog.nl
rtvutrecht.nl
hvzoeland.nl
dvhn.nl
lawaaiders.nl
degelijkoostendaard.nl
rtvoost.nl
duurzaambedrijfsleven.nl
omroepbrabant.nl
bladna.nl

**22:00**



**23:00**



**02-01-2019 - 00:00**