Universiteit Leiden
Computer Science

Reach and content of Dutch junk news on Facebook

Name: S. Kanhai
Date: 04/07/2018
1st supervisor: Dr. S. Verberne
2nd supervisor: Dr. J.P. Burger

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

[25] contains a recent and extensive review on scientific research on "fake news" on social media. In their review, they list shortcomings in current scientific research on this subject. The most prominent shortcomings they feature are that most work focusses on the effect of "fake news" as opposed to determining its reach and that it is done on the more research-friendly Twitter instead of on Facebook, the world's biggest social network. Additionally, according to [16] "fake news" has become a politicised umbrella term for false content and disinformation.

In this work we address these shortcomings by researching what the reach of junk news on Dutch Facebook is, compared to mainstream news. We define reach as the collective user engagement on all posts of a set Facebook pages in terms of its number of reactions, comments and shares, our *engagement metrics*. Junk news is content that lacks professional journalism traits such as transparency and accountability, has misleading, exaggerating and emotionally driven language and is sourced from untrustworthy media. We research Facebook, the world's biggest social network. More specifically, we analyse Dutch Facebook. Addressing the additional shortcoming of [25] that most research is currently done on the US.

From a seed list of junk and mainstream news domains provided by Nieuwscheckers; two experts on journalism and new media, we collect all posts published by their associated Facebook pages between Jan 2013 and Dec 2017 with their engagement metrics using the Facebook API.

On our 58,986 junk news Facebook posts we got 25,314,481 reactions, 9,725,792 comments and 9,335,548 shares. Comparing these numbers with mainstream news, junk news published 1.36% more posts, while getting 16.6% more reactions, 51.65% more comments and 54.8% more shares; indicating larger user engagement on junk news than on mainstream news. An in-depth comparison between the distributions of junk and mainstream news has shown that the greater user engagement on junk news is a trend.

Thus, junk news on Dutch Facebook is a phenomenon to behold. We find its reach to be even greater than mainstream news. Due to its absolute and relative popularity junk news on Facebook warrants further scientific research.

Contents

1 Introduction

Survey data from the Pew Research Center [19] and the Reuters Digital News Report [15] show that between 40% and 60% of adults in most developed countries, read news on social media, with Facebook being the leading source. Simultaneous to this growth in news consumption on social media we got an increase in false content on social media [1]. [25] contains a recent and extensive review on scientific research on "fake news" on social media. In their review, they list shortcomings in current scientific research on this subject. The most prominent shortcomings they feature are that most work focusses on the effect of "fake news" as opposed to determining its reach and that it is done on the more research-friendly Twitter instead of on Facebook, the world's biggest social network. Additionally, according to [16] "fake news" has become a politicised umbrella term for false content and disinformation.

In this work we address these shortcomings by researching what the reach of junk news on Dutch Facebook is, compared to mainstream news. We define reach as the collective user engagement on all posts of a set Facebook pages in terms of its number of reactions, comments and shares, our *engagement metrics*. Junk news is content that lacks professional journalism traits such as transparency and accountability, has misleading, exaggerating and emotionally driven language and is sourced from untrustworthy media. It does not need to have all described properties as long as it contains most. Our definition of junk news is a variation of its definition in [14].

According to Nieuwscheckers; two experts on journalism and new media, junk news is an important category of news due to its suspected reach and the amount of money revolving around junk news, while it is currently little researched. The greatest distinction between junk news and "fake news" is their most distinguishing property. Junk news' most distinctive property is its low-quality regardless of its truthfulness, while "fake news"' most distinctive property is being false and/or misleading content with no restrictions on its quality. "Fake news" can be low- or high-quality. Both junk and "fake news" can be political. However, the junk news we research is primarily commercially driven with little content that is politically motivated.

Junk news and "fake news" are types of content that both also classify as clickbait. Clickbait is a format of content. It is (internet) content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page. Clickbait can be real, fake, low-quality or high-quality news and is not restricted to specific topics.

We perform our research on Facebook, the leading source for news consumption on social media ([19];[15]). More specifically, we analyse Dutch Facebook. Addressing the additional shortcoming of [25] that most current scientific work focusses on the US omitting similar research in other countries. There are 10.5 Million Facebook users in The Netherlands [21] on a population of 17 Million.

For a frame of reference on the overall popularity of Facebook and our interest in the comparison itself, we compare the reach of junk news with the reach of mainstream news on Facebook. Mainstream news is high-quality, edited, original content, posted by widely recognized news media. All of this results in our central research question:

- What is the reach of junk news on Dutch Facebook compared to mainstream news?

We start our research from a seed list of 63 junk news and 20 mainstream news domains compiled by Nieuwscheckers. For each seed domain we get its accompanying Facebook page(s) and use the Facebook

API[7] to collect all its posts between Jan 2013 and Dec 2017 with their engagement metrics; the collective user engagement.

Over time, the number of users on Facebook in the Netherlands has grown from 9.6 million users in 2014 to 10.4 Million users in 2017 [21]. Such timebased differences could influence the user engagement on junk and mainstream news. As such, we also research the reach of junk and mainstream news over time:

- What is the reach of junk news on Facebook compared to mainstream news over time?

In addition to researching the reach of junk news on Facebook we explore the objective of junk news on Facebook. Currently, no scientific work on the objective of junk news on Facebook exists. The only scientific research on the objective of junk news is on their objective on their own domains. On their own domains, junk news aims for profit-maximizing with their advertising supported and driven sites ([4];[3];[1];[12]). However, with the data we acquire via the Facebook API we can not directly determine the objective of junk news on Facebook. Therefore, we infer the objective of junk news on Facebook using its collective linking behaviour.

The collective linking behaviour is the result of combining the linking behaviour of all individual junk news Facebook pages. We compute the linking behaviour for an individual junk news Facebook page as follows. Each Facebook post can contain text, audiovisual media; photo or video, and outgoing references, its *status links*. For all posts of a Facebook page we categorize its status links into one of four categories and calculate the relative proportions of each category of status link; its linking behaviour. Based on the relative proportions of the four categories we can infer the object of junk news on Facebook:

- What is the collective linking behaviour of our junk news Facebook pages?

To get the collective linking behaviour of junk news on Facebook we create a directed *linking behaviour* network where all Facebook pages and domains we link in the status link are the nodes. In this network we connect all seed domains to their accompanying Facebook page(s) and all Facebook pages to the domain they link in their status link.

During the compilation of the seed list Nieuwscheckers got suspicions of duplicate content among junk news seed domains. We analyse these suspicions by assessing if the content junk news publishes on Facebook is similar between different Facebook pages. With Facebook parsing all linked content, retrieving the title, cover media and part of its body of text for each link we check if junk news Facebook pages share content by checking if they share titles verbatim and answer the research question:

- To what extent is content shared by our junk news Facebook pages?

To answer this question we introduce an undirected *content overlap* network. We use junk news Facebook pages as our nodes and connect those Facebook pages that verbatim share titles. Based on the level of connectivity of the network we know how many Facebook pages share content and how much content is shared. The remainder of this thesis is structured as follows: Chapter 2 discusses background and prior work, Chapter 3 outlines our methods, Chapter 4 discusses our analysis and Chapter 5 concludes our thesis.

2 Background and prior work

To frame our work we elaborate on the three shortcomings of current scientific research on “fake news” on social media: a. most work focusses on the effect of "fake news", b. most work is done on the more research-friendly Twitter and c. "fake news" has become a politicised umbrella term for false content and disinformation .

First, [25] states that most scientific work focusses on the effect of “fake news” as opposed to determining its reach. Currently, most work on this subject is done in the social sciences. Its academic incentives motivate scholarly publications to establish causal relationships on the effect of a phenomenon as opposed to determining its prevalence. However, before establishing causal relationships we need to understand the reach of a phenomenon. Additionally, knowing about the reach of “fake news” enables policymakers to make smart public policy decisions which properly consider the costs and benefits of proposed policy changes [25].

Some work on the reach of “fake news” on social media *is* present. Craig Silverman from Buzzfeed News [20] analysed the top 10 from “fake” and mainstream news posts on the topic of the 2016 presidential elections in terms of its user engagement. He found that the top 10 “fake news” posts got greater user engagement. Furthermore, he also found that the top 10 posts from mainstream news was sensational content. [2], [17] and [23] corroborate the findings of Silverman and elaborate on his research on the reach of “fake news” on the topic of the 2016 presidential elections.

Other work on the reach of “fake news” on social media is from [6]. Their work is most similar to ours. They provide toplevel usage statistics for the most popular sites that independent fact-checkers and other observers have identified as publishers of false news and online disinformation in France and Italy. However, this work has been criticized by [11] who notes their a. lack of verification of the credentials for the list of domains they consider "fake news", b. failure to address the really disturbing sources of "fake news" namely "fake news" produced by mainstream news and politicians, c. exclusion of "fake news" content without its own domain such as YouTube, AlterVista or BlogSpot, d. comparison with mainstream news-websites which themselves publish manipulated news and e. wrong method of measuring the size of "fake news", since they measure the time spend on a website instead of measuring how it is spread.

Second, [25] also states that most research on “fake news” on social media is done on Twitter instead of Facebook. The relative lack of this research on Facebook could be due people’s tendency to interact with just their friends on Facebook as opposed to their predominantly public interactions on Twitter. This difference could be due to Facebook’s insistence for its users to use their real name, location, educational background, and other biographical information, while Twitter encourages tweeting under a nickname and requires only minimal information about its users [25]. As such, Twitter’s API is considered more research-friendly than Facebook’s. Zooming in on computer science research in particular on “fake news” on social media we note that most work is on bot detection algorithms. [5] lists a number of such bot detection algorithms on Twitter. Similar research on Facebook does not exist.

Third, [16] states that most work in social media focusses on “fake news” which has become a politicised umbrella term. It has become a term for false content for advertising revenue [22], government-backed misinformation campaigns, tendentious news coverage, favoring a particular party with false or outrageous statements by politicians, and a weaponized term by critics of established news media to attack and undermine the credibility of professional journalism.

We recognize different levels of intent with regards to the politicalization of news on social media. Content with a purely apolitical intent as done by teenagers in Macedonia who created a range of websites like USConservativeToday.com posting stories favoring either Trump or Clinton earning them tens of thousands of dollars [22]. Content with profit-driven polical intent by publishing more pro-Trump than pro-Clinton content during the 2016 presidential election, since pro-Trump content generated more advertising revenue [12]. Content with purely political intent, such as endingthefed.com which intended to help Donald Trump's campaign [24].

In this work we address these shortcomings by researching what the reach of junk news on Facebook is, compared to mainstream news.

3 Methods

Nieuwscheckers compiled a seed list of 63 known Dutch junk news sites. The experts also compiled a list of 20 Dutch mainstream news websites. These seed domains are listed in Table 6.1 for junk news and Table 6.2 for mainstream news.

For each domain in our seed lists we get the Facebook page(s) for the domain by crawling the site using Selenium [18] to extract all its links to Facebook. We use the Facebook API to download all posts published between Jan 2013 and Dec 2017 by a junk news Facebook page. For each post we get its published date, status message, status link, reaction, comment and share count. For the status links to Facebook we keep the Facebook domain and page name and for status links to an external domain we only keep the domain.

All data collection from the Facebook API was done between December 2017 and January 2018 using Facebook API 2.10 in Python. All data preparation and analysis was done in R.

4 Analysis and discussion

In this section we analyse and discuss our four research questions:

- What is the reach of junk news on Dutch Facebook compared to mainstream news?
- What is the reach of junk news on Facebook compared to mainstream news over time?
- What is the collective linking behaviour of our junk news Facebook pages?
- To what extent is content shared by our junk news Facebook pages?

Each research question gets discussed in its own subsection. Some research questions have additional subquestions.

4.1 Reach comparison between junk and mainstream news

Summarizing all post activity and user engagement on junk and mainstream news for the period between Jan 2013 to Dec 2017 we get the values listed in Table 4.1. This table lists the total number of posts published by our junk and mainstream news Facebook pages with their total number of reactions, comments and shares.

From Table 4.1 we note that junk news published 1.36% more posts than mainstream news with 16.6% more reactions, 51.65% more comments and 54.8% more shares. Recalling that we define reach as the collective user engagement on all posts of a set Facebook pages in terms of its number of reactions, comments and shares, Table 4.1 shows that junk news has greater reach than mainstream news.

However, the results listed in Table 4.1 are summarizations over a period of time which could hide influential datapoints. Outliers for junk and mainstream news such as viral Facebook posts could, for example, be disproportionately responsible for these differences. We are therefore interested in the distributions of our engagement metrics:

Table 4.1: For our junk and mainstream news Facebook pages we list the number of pages, posts, reactions, comments and shares for each set.

	pages	**posts**	**reactions**	**comments**	**shares**
junk news	63	58,986	25,314,481	9,725,792	9,335,548
mainstream news	20	58,186	21,113,471	4,702,733	4,220,128

Question: *What are the distributions for the engagement metrics?*

We can represent the distribution a continuous random variable x either with the *cumulative distribution function (CDF(x))* or *probability distribution function (PDF(x))*. The CDF(x) tells us the odds of measuring any value up to and including x, $P(x \leq X)$. The PDF(x) is the derivative of CDF(x), its magnitude is the relative likelihood of measuring a particular value x. For our use case of comparing the distributions between junk and mainstream news we use CDFs. A comparison using PDFs would introduce the additional complexity of comparing magnitudes per particular value or sets of values.

To research the distributions for the engagement metrics we gather the frequency x for each weight w for all three engagement metrics. $x \in X$ is the set of all possible frequencies and $w \in W$ is the set of all possible weights. To also consider the frequency for each weight, we use a variant of a CDF, the *weighted cumulative distribution function*:

Table 4.2: List the 2.5th, 25th, 50th, 75th and 97.5th percentile for all engagement metrics for both sets of Facebook pages: junk news and mainstream news. Additionally, we also list the mean and standard deviation.

Number of reactions	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	**mean**	**std. dev**
junk news	0	17	70	264	3,369	429.16	1,744.73
mainstream news	0	9	42	174	2,517	362.86	2,093.1
Number of comments							
junk news	0	2	11	53	1,374	164.88	857.62
mainstream news	0	1	8	37	569	80.82	498.24
Number of shares							
junk news	0	2	13	61	895	158.27	1,878.88
mainstream news	0	1	5	24	395	72.53	1,243.79

$$F(x) = \int_{-\infty}^{w} f(y)dy \qquad \text{for all possible } w \text{ values} \tag{1}$$

Due to the greater number of published posts for junk news we measure the differences between junk and mainstream news in proportions.

Figure 4.1d lists the CDFs for all engagement metrics for junk news (*red*) and mainstream news (*blue*). Summary statistics in Table 4.2 for the CDFs from Figure 4.1d list the 50th percentile commonly known as the median, 25th and 75th percentile; the *Inter Quantile Range (IQR)*. Additionally, Table 4.2 lists the weighted mean:

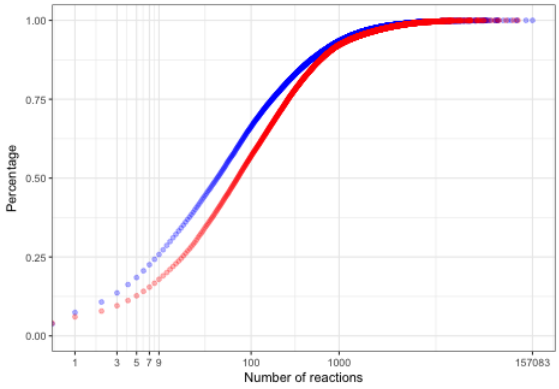
$$\bar{x} = \frac{\sum_{n=1}^{N} w_i x_i}{\sum_{n=1}^{N} w_i} \tag{2}$$

and standard deviation. Large differences between the median and mean and the relatively large standard deviation indicate the presence of large outliers; justifying the logarithmic scales for our CDFs. We observe that all CDFs for junk news are to the right from those for mainstream news; for a similar proportion of junk news we got greater weight for an engagement metric in comparison to mainstream news. However, graphical limitations of the CDF in Figure 4.1d and the limited summary statistics in Table 4.2 make comparing the distributions difficult.

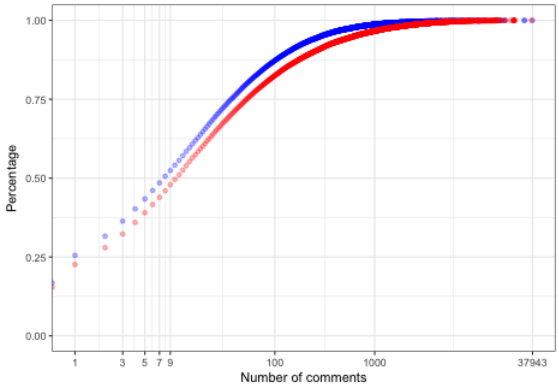
To easier compare our CDFs we use the *relative distribution* ([9]; [8]). With Y_0, Y as the random variables for our CDFs F_0, F we apply F_0 to Y to get the distribution of the random variable R. R is the relative distribution of Y:

$$R = F_0(Y) \tag{3}$$

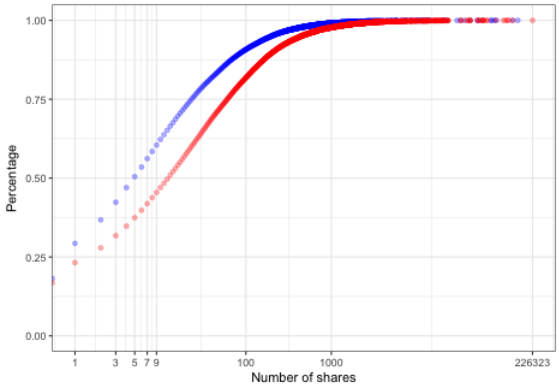
We use mainstream news as our baseline CDF (F_0) and plot the CDF for junk news relative to this baseline. Figure 4.1h shows the relative distributions for all engagement metrics.



(a) reactions

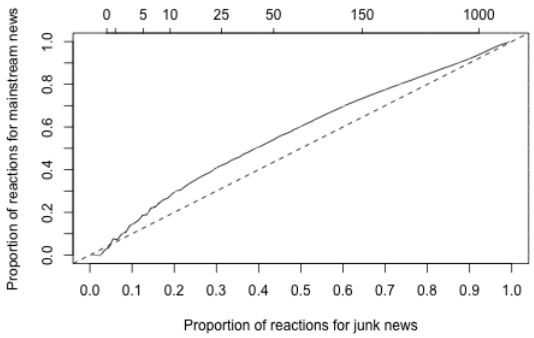


(b) comments

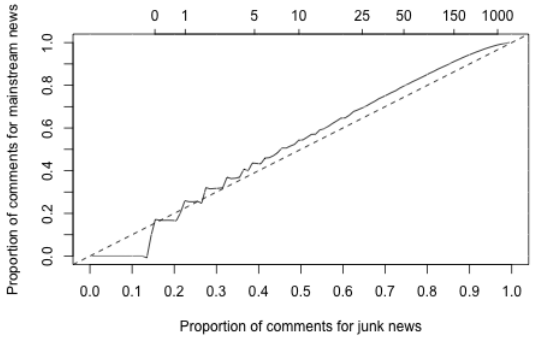


(c) shares

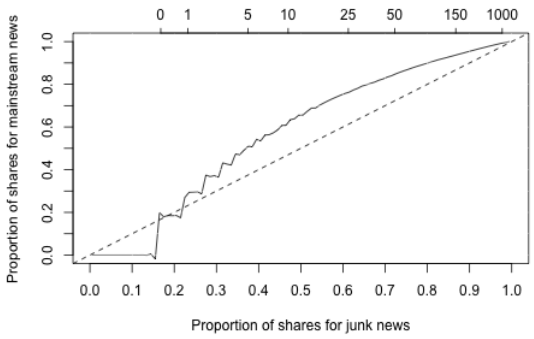
(d) Cumulative densities for all engagement metrics for both sets of Facebook pages: junk news (*red*) and mainstream news (*blue*). The x-axis is the weight of the engagement metric on a logarithmic scale, while the y-axis shows what proportion for a set of Facebook pages have a value up to a particular weight. If a CDF is more located to the right, it means that for a particular proportion of Facebook pages it has greater weight for the engagement metric.



(e) reactions



(f) comments



(g) shares

(h) Relative densities for all engagement metrics. For all engagement metrics we use the CDF for junk news as reference distribution and plot the CDF for mainstream relative to the baseline. The reference distribution is plotted as the 45° line from $(0, 0)$ to $(1, 1)$. The x-axis shows the proportion of Facebook posts for junk news and the y-axis for mainstream news. Since, all engagement metrics have posts with weight 0 for a metric we are not able to represent $(0, 0)$ datapoint. So, we ignore all results before the relative CDF for hits the baseline CDF for the first time. This is for the following proportions on the y-axis; 16.79% for reactions, 16.79% for comments and 18.2% for shares.

The relative distribution zooms in on the ≤ 1000 weights for the engagement metrics. The relative distributions in Figure 4.1h reaffirm that junk news has greater reach than mainstream news; a similar proportion of junk news requires greater weight for an engagement metric in comparison to mainstream news. The jagged plots for the lower weights are due to the relative large differences between junk and mainstream news for the lower weights of the engagement metrics.

Because the differences between the CDFs range over the interval we evluate, we note that the greater user engagement on junk news in comparison to mainstream news is a trend. Each engagement metric does have a distinct growth trend. We compare the growth trends by evaluating them for three weights of the engagement metrics; 25, 50 and 150.

First, we evaluate which of these three weights are closest to the maximum differences for junk and mainstream news. The maximum difference for the reactions is closest to a weight of 50, the maximum difference for the comments is closest to 150 and the maximum difference for the shares is closest to 25. Analysing the growth trends in general and around their maximum differences we note that the reactions grow more steeply towards its maximum difference than it declines, comments grow more gradually towards its maximum difference than it declines and shares are balanced in its growth and decline towards its maximum.

The relatively small difference in reaction engagement (16.6%) for junk news makes reviewing its real-world causes less interesting. More interesting are the relative differences for the comments and shares with their rather similar percentual differences 51.65% respectively 54.8%, but very different relative growth trends as shown in Figure 4.1h. The growth trend for the comments for junk and mainstream news are similarly shaped, while the growth trend for the shares show a clear maximum difference.

If we translate the differences to the real world we think that the difference for the comment engagement is because the greater number of comments per post. Possibly, due to the nature of the content of junk news; having misleading, exaggerating and emotionally driven language, junk news Facebook posts evoke greater comment engagement. Similarly, the active stimulus of junk news to share its posts on Facebook causes these posts to be more likely to get shared. By sampling the status messages for junk and mainstream posts we found that junk news actively stimulates sharing.

All our analysis up until now assumes that engagement metric do not influence each other. However, engagement metrics could have reinforcing or decreasing effects on each other. Posts with more reactions could, for example, be more likely to get shared. Such relationships could skew our observations, we therefore analyse if any of such relationship exists.

Question: *Are there relationships between the engagement metrics?*

Figure 4.2d lists all three engagement metrics in pairs relative to each other. Each dot on the scatterplot represents one post from the sets of Facebook pages; a *red* dot for a junk news Facebook post and a *blue* dot for mainstream news. Posts with higher weights for engagement metrics are located higher and to the right.

Detecting any relationships between the engagement metrics in Figure 4.2d is difficult due to the presence of large outliers. Since it is complex to detect relationships on non-linear axes, we do not change the scale on the x- and y-axis. We rather discard some data. Therefore, we compute the 95% confidence interval around the weighted median. The 95% confidence interval lets us contain most data, while we discard the largest outliers. Therefore, we calculate and list the 2.5th and 97.5th percentile for all engagement metrics for junk and mainstream news in Table 4.2.

Table 4.3: We calculate the $\sim$ 95% confidence interval around the weighted median for each engagement metric. Here, we list what proportion of data remains in the $\sim$ 95% confidence interval.

reactions	**proportion of posts**
junk news	93.5
mainstream news	93.7
comments	
junk news	82.09
mainstream news	80.7
shares	
junk news	80.7
mainstream news	79.3

Table 4.4: We compare all pairs of engagement metrics using the $\sim$ 95% confidence interval around the weighted median for each engagement metric. Here, we list what proportion of data remains in each intersection of these intervals.

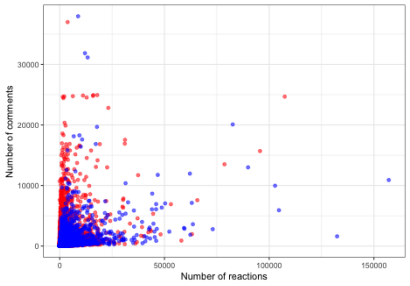
reactions vs comments	**proportion of posts**
junk news	78.75
mainstream news	78.51
reactions vs shares	
junk news	77.92
mainstream news	77.36
comments vs shares	
junk news	70.61
mainstream news	69.95

To get the 95% confidence interval we discard all data with weights in the 2.5th and 97.5th percentile for an engagement metric. However, we can not exactly discard all data in the 2.5th percentile; all engagement metrics for both sets of news have 0 as the maximum value for the 2.5th percentile and posts outside the 2.5th percentile also have 0 as their weight. For this reason we get a custom 95% confidence interval; $\sim 95\%$ *confidence interval* where we keep $< 95\%$ of our data. Table 4.3 lists what proportion of data we keep when we discard all posts with an engagement metric weight of 0. Furthermore, since we intend to compare pairs of engagement metrics we calculate what proportion of data remains if we get the intersection for each $\sim 95\%$ confidence intervals of our pairs of engagement metrics. Table 4.4 lists what proportion of posts remain if we meet this condition.

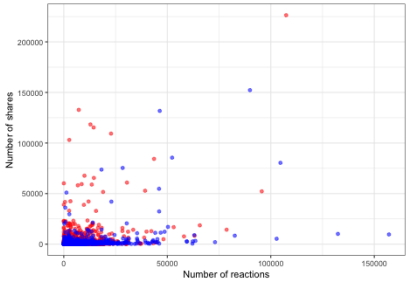
lists the same scatterplot as Figure 4.2d, but just for the custom $\sim 95\%$ confidence interval. Because the values for the engagement metrics for junk and mainstream news differ substantially we plot their scatterplots separately in Figure 4.2l and Figure 4.2p.

For posts with weights in the lower one-thirds proportion for reactions versus comments and comments versus shares, Figure 4.2i respectively Figure 4.2k we can not discern any relationship for the engagement metrics due to the high density of the data in this part of the scatterplots. The maximum weights for these part of the scatterplots are 1000 for the reactions, 500 for the comments and 1000 for the shares. From the CDFs for the engagement metrics in Figure 4.1d we note that these weights cover 90% of the posts for each engagement metric. We therefore conclude that we can not observe any relationship for these pairs of engagement metrics. While the scatterplot for the reactions versus shares in Figure 4.2j is differs more from

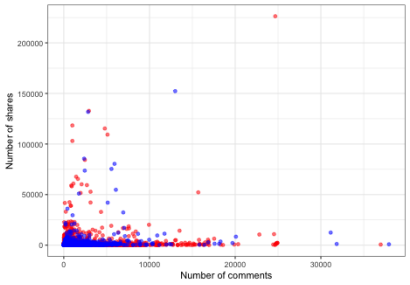
Figure 4.2i and Figure 4.2k it is still to dense to detect a pairwise relationship. Thus, we conclude that no observable reinforcing or decreasing relationships are between any pair of engagement metrics.



(a) reactions vs comments

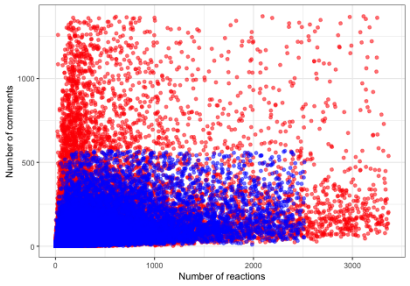


(b) reactions vs share

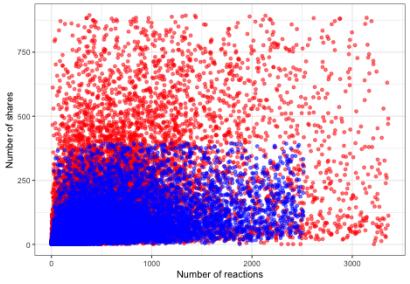


(c) comments vs shares

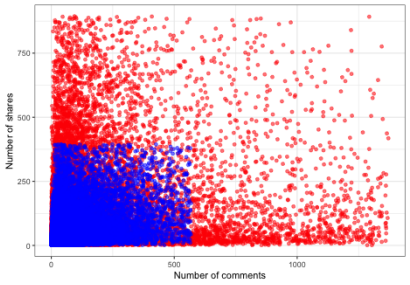
(d) From top to bottom we list scatterplots for pairs of engagement metrics relative to each other for our the $\sim$ 95% confidence interval around the weighted media Table 4.4. Each (*blue*) dot on the scatterplot represents one mainstream news Facebook post. The x- and y-axis represent the weight of an engagement metric where the location of each point is the weight for a Facebook post for an engagement metric. Posts with higher engagement metrics are located further up and to the right.



(e) reactions vs comments

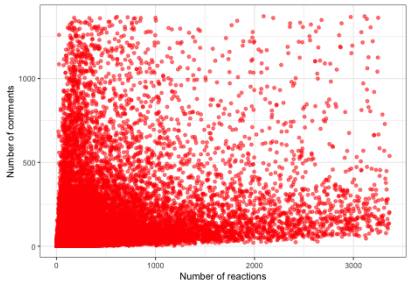


(f) reactions vs shares

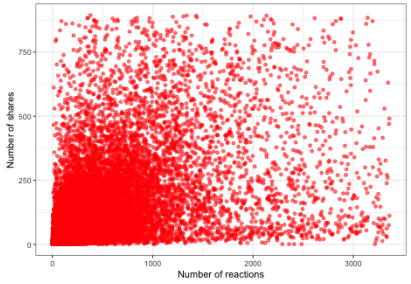


(g) comments vs shares

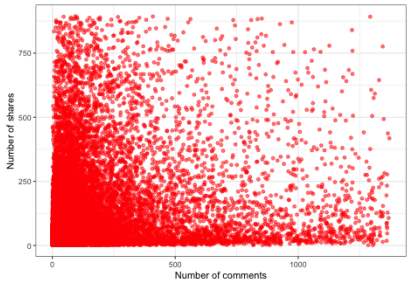
(h) From top to bottom we list scatterplots for pairs of engagement metrics relative to each other for our the $\sim$ 95% confidence interval around the weighted media Table 4.4. Each (*red*) dot on the scatterplot represents one junk news Facebook post. The x- and y-axis represent the weight of an engagement metric where the location of each point is the weight for a Facebook post for an engagement metric. Posts with higher engagement metrics are located further up and to the right.



(i) reactions vs comments

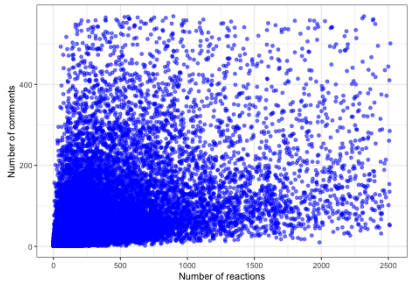


(j) reactions vs shares

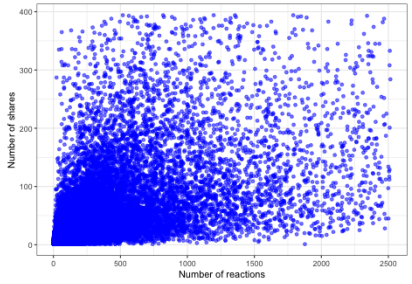


(k) comments vs shares

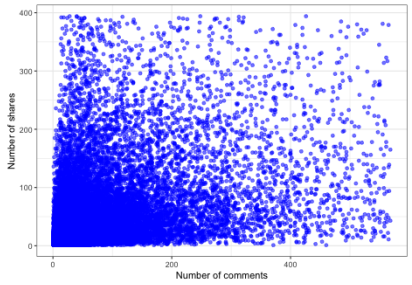
(l) From top to bottom we list scatterplots for pairs of engagement metrics relative to each other for our $\sim$ 95% confidence interval around the weighted media Table 4.4. Each dot on the scatterplot represents a post from our data; a *red* dot for a junk news Facebook post and a *blue* dot for mainstream news. The x- and y-axis represent the weight of an engagement metric where the location of each point is the weight for a Facebook post for an engagement metric. Posts with higher engagement metrics are located further up and to the right.



(m) reactions vs comments



(n) reactions vs comments



(o) reactions vs comments

(p) From top to bottom we *red* dot for a junk news Facebook post and a *blue* dot for mainstream news. The x- and y-axis represent the weight of an engagement metric where the location of each point is the weight for a Facebook post for an engagement metric. Posts with higher engagement metrics are located further up and to the right.

4.2 Reach comparison over time between junk and mainstream news

From Jan 2013 till Dec 2017; our investigative period, Facebooks' popularity has grown from 9.6 million users in 2014 to 10.4 million users in 2017 [21]. Such timebased growth could influence the use of Facebook as a platform for publishing content and consuming said content. As such, we review the post activity; the number of posts published per month, of junk and mainstream news and the user engagement on these posts. We use all post activity by and user engagement on mainstream news as our baseline.

Question: *How does the post activity of junk news on Facebook compare to mainstream news' over time?*

To compare the post activity between junk and mainstream news for our investigative period we plot four figures in Figure 4.3. In the top figure from Figure 4.3 we summarize all posts published in a particular month for junk respectively mainstream news. Below we go in greater detail on the activity of the Facebook pages by showing their post activity distributions using boxplots. We create two boxplot; one where we display the post activity on a linear y-axis and another where we display the post activity on an exponential y-axis, to zoom in on the IQR unhampered by outliers. Finally, we list what number of Facebook pages publish per set of Facebook pages per month.

From Figure 4.3 we note that mainstream news keeps its post activity consistent; ranging around a mean of 970 posts with a standard deviation of 136. Its considerable standard deviation is due to the recent increase in post activity starting at the end of 2017. The post activity boxplots for mainstream news corroborate mainstream news' post activity. Over the different months we got a consistent mean indicating that individual mainstream news Facebook pages keep their post output consistent. Finally, the bottom figure from Figure 4.3 also shows almost all mainstream news Facebook publish each month. Additionally, we note the IQR for mainstream news growing towards a stable range. As time passes, Facebook established itself as a regular publishing platform for mainstream news.

The growth trend for the post activity for junk news on Facebook is more irregular. Growing greater starting from Jan 2013, matching the post activity for mainstream news at the start of 2016 and finally drastically increasing its post activity at the end of 2017. The growing means in the post activity boxplots partially explain the increase; as time passes, individual junk news Facebook pages publish more often on Facebook. However, the real cause for the increased post activity is the increased number of publishing junk news Facebook pages. In Jan 2013 we had just a single junk news Facebook page, we see ≥ 40 publishing junk news Facebook pages for each month in 2017. For junk news we also see that Facebook established itself as a regular publishing platform. Its IQR starts narrow due to the limited number of publishing junk news Facebook pages, widening from the second half of 2014 due to the increasing number of publishing junk news Facebook pages and finally shrinking towards a relatively consistent range from the start of 2017.

Question: *How does user engagement of junk news on Facebook compare to mainstream news' over time?*

Having analysed the use of Facebook as a platform for publishing content we now evaluate how said content is engaged on over time. In the top figures of item 4.4, item 4.5 and item 4.6 we list the total user engagement for junk and mainstream news for the reactions, comments and shares per month. Note that we normalised all engagement metrics by the post activity of each set of news for a month. Below this show the ratio where we divide monthly user engagement for junk news by monthly user engagement. Mainstream news is our baseline. The ratios in these figures mean the following; a. < 1: we got greater user engagement on mainstream news than junk news for that month, b. 1: we got equal user engagement on junk and mainstream

news for that month and c. > 1: we got greater user engagement on junk news than mainstream news for that month. We calculate the ratio, because the total user engagements per month tend to alternate between junk and mainstream news, making comparing them difficult. With the ratios we can also quantify the monthly differences in user engagement. Additionally, in the bottom two figures of item 4.4, item 4.5 and item 4.6 we list the distribution for each engagement metric for junk and mainstream news per month using boxplots. Again, we plot two boxplots; one where we display the engagement metric on a linear y-axis and another where we display the engagement metric on an exponential y-axis, to zoom in on the IQR unhampered by outliers.

Overall, we notice the user engagement per post for each set of news for all engagement metrics grows greater as time passes. Remember that all results we show in item 4.4, item 4.5 and item 4.6 are normalised by the post activity of each set of news for a month. The number of comments and shares seem to grow especially greater. Additionally, the ratios show user engagement on junk news to be greater than mainstream news for each month, starting from the beginning of 2016. With the exception of the numbe of shares which has always been greater for junk news. The differences in engagement as shown by the ratios are at most 2 times as much for some months for junk news and 4 to 5 times as much for the shares for junk news.

The boxplots for the engagement metrics show that junk news has more outliers for its monthly user engagement. This is most likely due to the greater likeliness of junk news Facebook posts to go viral. This possibility of going viral is most apparent for the the number of shares. The IQR of the boxplots for the engagement metrics show user engagement per individual Facebook page to increase. Again, we see the IQR from mainstream news shrinking to a stable range and the IQR of junk news starting from a smaller range, widening as time passes to finally stabilize around a stable IQR.

Contextualizing the timebased changes to real world we see that Facebook established itself as a more popular platform to publish content on by junk and mainstream news. Junk news, in particular, has shown an explosion in how much content it publishes on Facebook. With regard to user engagement we find that over time all individual Facebook posts get more reactions, comments and shares. Starting from the beginning of 2016 we also note that junk news tends to consistently get a greater number of reactions, comments and shares per month in comparison to mainstream news.

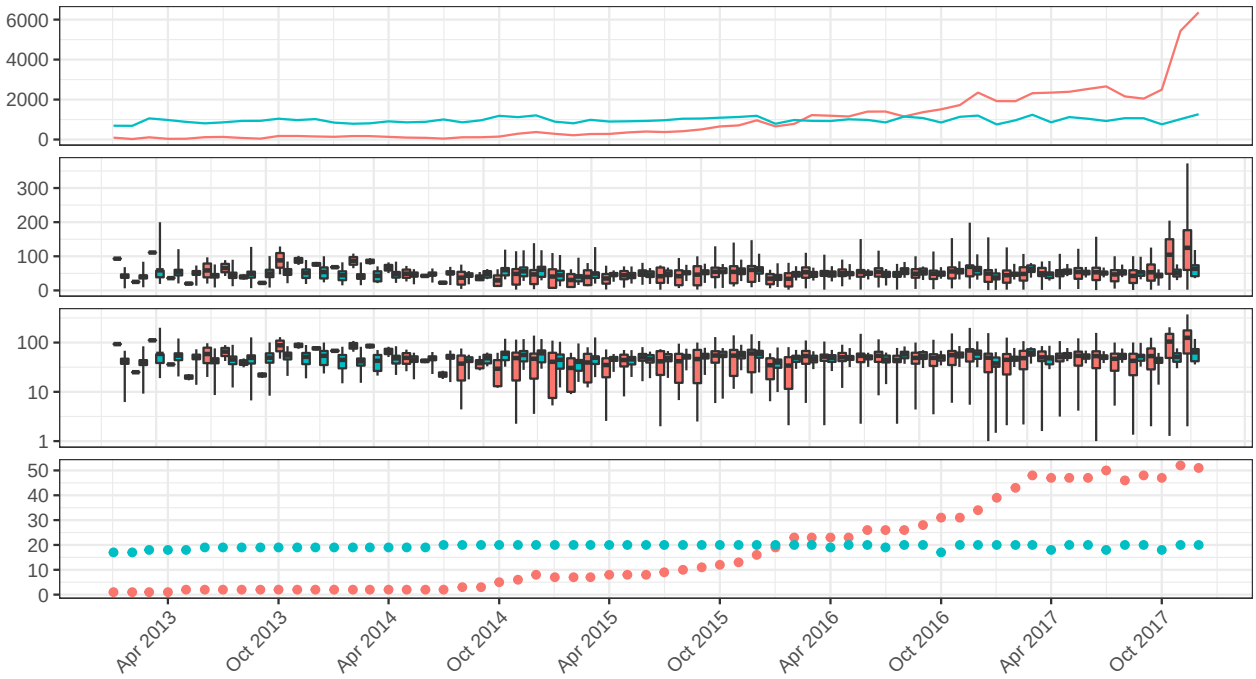


Figure 4.3: From top to bottom we list four timeseries figures. The x-axis represents the period from Jan 2013 till Dec 2017. We list a datapoint for each month. The y-axis in the first three figures represents the post activity in a particular month for each set of Facebook pages. Each figure lists the set of junk (*red*) and mainstream news (*blue*) Facebook pages. In the top figure, we show the summarised post activity for each set of Facebook pages per month. Below we get two monthly boxplots, representing the summary statistics (minimum, maximum, 25th, 50th and 75th percentile) for the distribution of the post activity for each set of Facebook pages. The first boxplot lists the post activity on a linear y-axis and the second on an exponential y-axis. Finally, we list what number of Facebook pages published for every given month. We do so on a linear y-axis and for each set of Facebook pages.

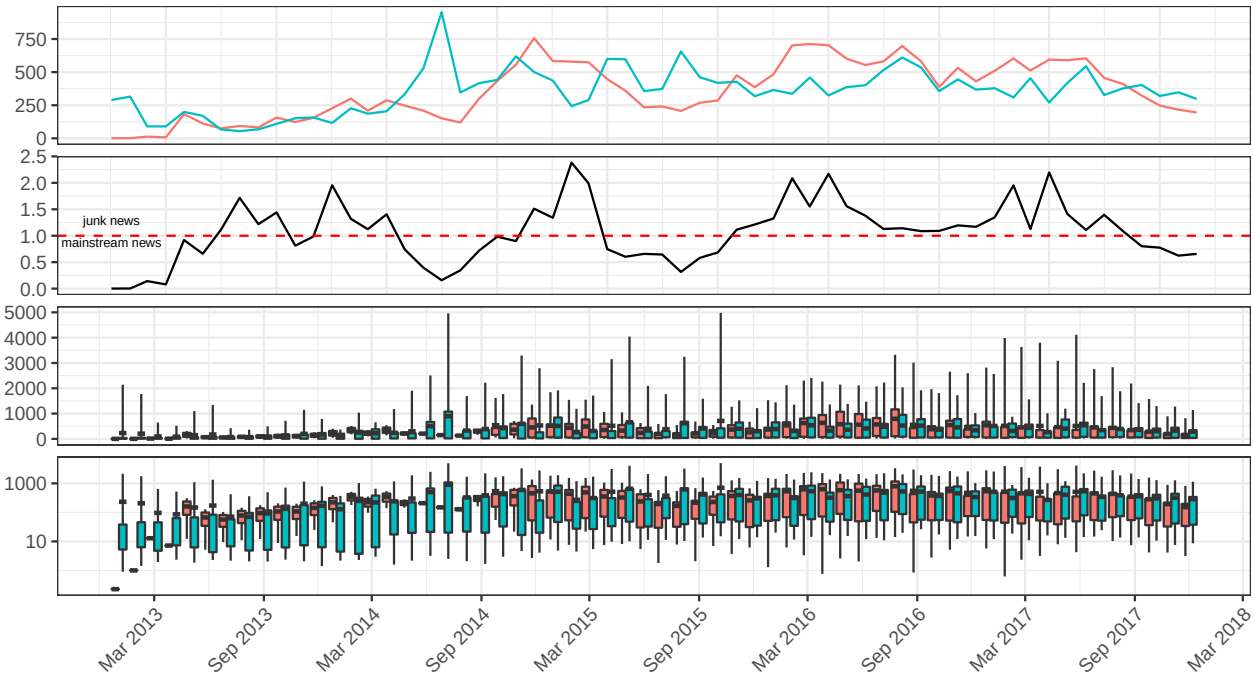


Figure 4.4: From top to bottom we list four timeseries figures. The x-axis represents the period from Jan 2013 till Dec 2017. We list a datapoint for each month. The y-axis in the first, third and fourth figure represents the number of reactions in a particular month for each set of Facebook pages. Each figure lists the set of junk (*red*) and mainstream news (*blue*) Facebook pages. At the top, we show the summarised number of reactions for each set of Facebook pages per month. Below we list the ratio for the number of reactions between junk and mainstream news with mainstream news as our baseline. Per ratio we get: a. < 1: more reactions on mainstream news than junk news for that month, b. 1: an equal number of reactions on junk and mainstream news for that month and c. > 1: more reactions on junk news than mainstream news for that month.Finally, we get two monthly boxplots, representing the summary statistics (minimum, maximum, 25th, 50th and 75th percentile) for the distribution of the number of reactions for each set of Facebook pages. The first boxplot lists the post activity on a linear y-axis and the second on an exponential y-axis. Note that number of reactions have been normalised by the number of posts in each month for each set.

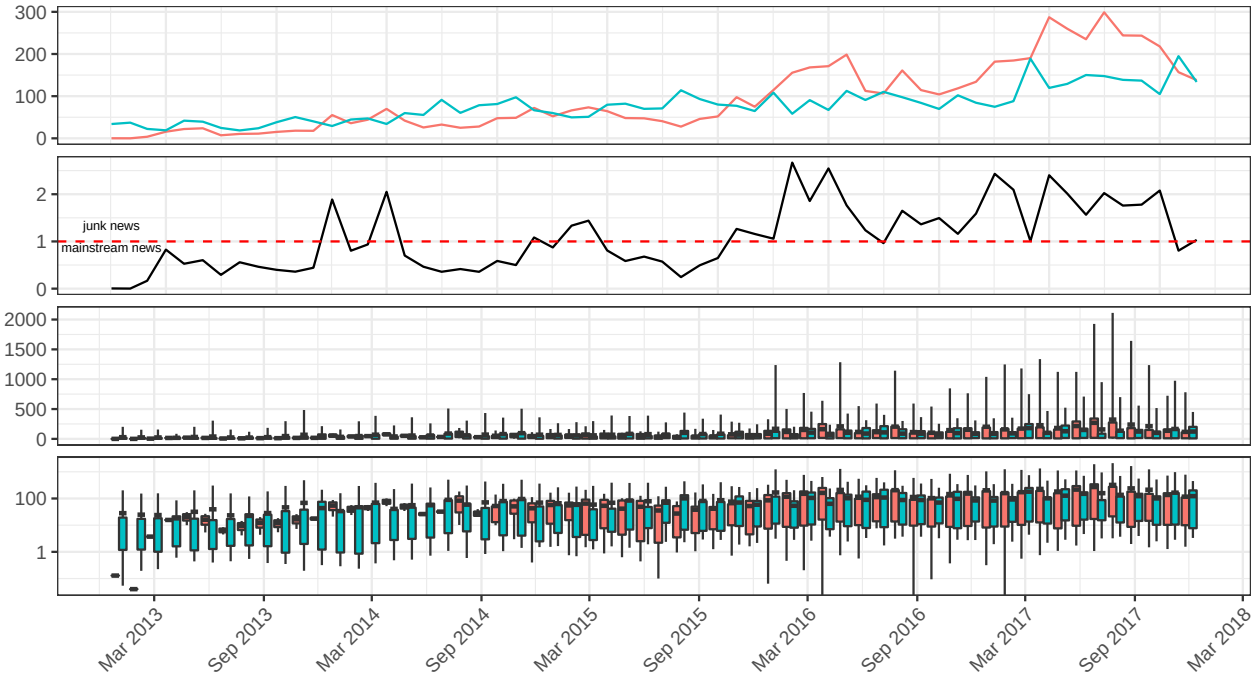


Figure 4.5: From top to bottom we list four timeseries figures. The x-axis represents the period from Jan 2013 till Dec 2017. We list a datapoint for each month. The y-axis in the first, third and fourth figure represents the number of comments in a particular month for each set of Facebook pages. Each figure lists the set of junk (*red*) and mainstream news (*blue*) Facebook pages. At the top, we show the summarised number of comments for each set of Facebook pages per month. Below we list the ratio for the number of comments between junk and mainstream news with mainstream news as our baseline. Per ratio we get: a. < 1: more comments on mainstream news than junk news for that month, b. 1: an equal number of reactions on junk and mainstream news for that month and c. > 1: more comments on junk news than mainstream news for that month.Finally, we get two monthly boxplots, representing the summary statistics (minimum, maximum, 25th, 50th and 75th percentile) for the distribution of the number of comments for each set of Facebook pages. The first boxplot lists the post activity on a linear y-axis and the second on an exponential y-axis. Note that number of comments have been normalised by the number of posts in each month for each set.

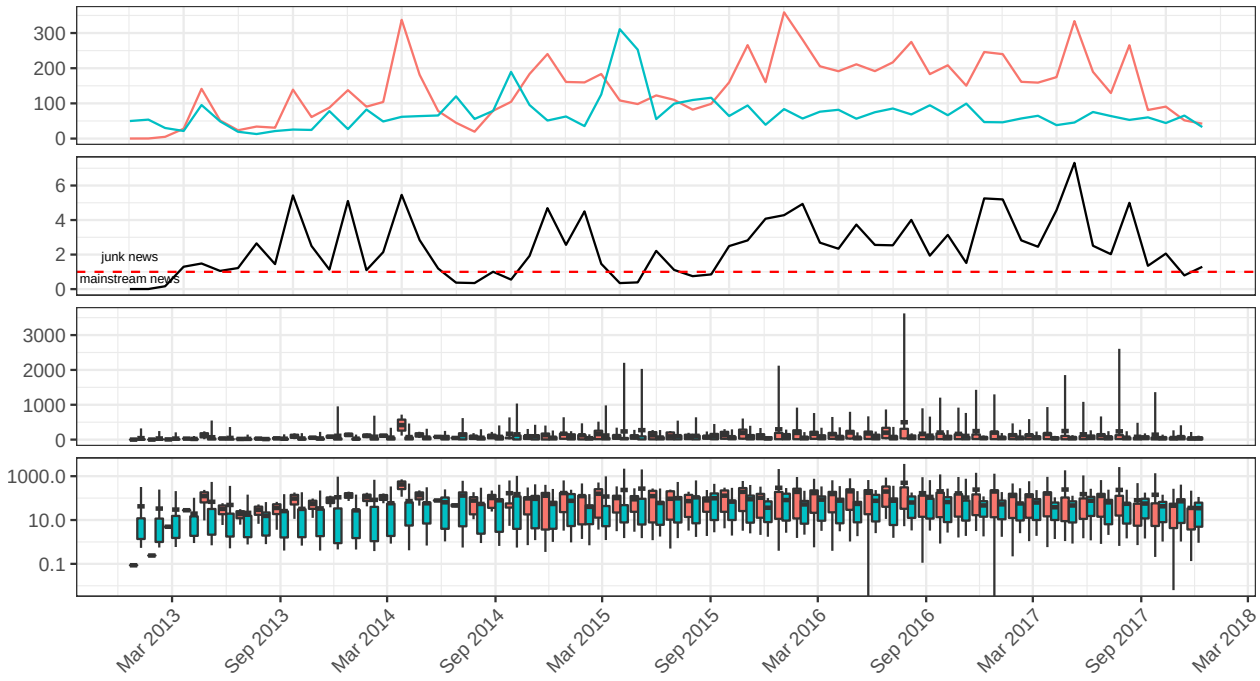


Figure 4.6: From top to bottom we list four timeseries figures. The x-axis represents the period from Jan 2013 till Dec 2017. We list a datapoint for each month. The y-axis in the first, third and fourth figure represents the number of reactions in a particular month for each set of Facebook pages. Each figure lists the set of junk (*red*) and mainstream news (*blue*) Facebook pages. At the top, we show the summarised number of shares for each set of Facebook pages per month. Below we list the ratio for the number of shares between junk and mainstream news with mainstream news as our baseline. Per ratio we get: a. < 1: more shares on mainstream news than junk news for that month, b. 1: an equal number of shares on junk and mainstream news for that month and c. > 1: more shares on junk news than mainstream news for that month.Finally, we get two monthly boxplots, representing the summary statistics (minimum, maximum, 25th, 50th and 75th percentile) for the distribution of the number of shares for each set of Facebook pages. The first boxplot lists the post activity on a linear y-axis and the second on an exponential y-axis. Note that number of shares have been normalised by the number of posts in each month for each set.

4.3 Objective of junk news

With the data we acquire via the Facebook API we can not directly determine the objective of junk news on Facebook. Therefore, we infer the objective of junk news on Facebook using its collective linking behaviour.

We already introduced the collective linking behaviour in the introduction. It is the result of combining the linking behaviour of all individual junk news Facebook pages where we compute the linking behaviour for an individual junk news Facebook page as follows. Each Facebook post can contain text, audiovisual media; photo or video, and outgoing references, its *status links*. For all posts of a Facebook page we categorize its status links into one of four categories and calculate the relative proportions of each category of status link; its linking behaviour.

We recognized four kinds of status links; links to its seed domain, the Facebook page itself, an *external domain*; a non-facebook.com domain that is not its seed domain, and an external Facebook page. An external Facebook page is a facebook.com domain that is not the Facebook page itself. Based on the relative proportions of the four categories we can infer the object of junk news on Facebook.

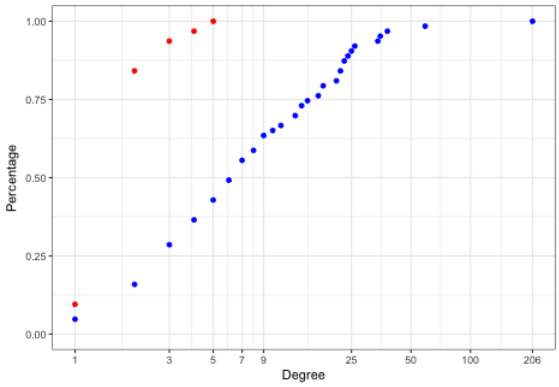
For example, if a high proportion of status links for junk news links to their seed domain we infer that junk news Facebook pages focusses on increasing its profit. Since junk news aims for profit-maximizing on their advertising supported and driven sites ([4];[3];[1];[12]). Similarly, if a high proportions of outgoing links of junk news on Facebook references the pages themselves then it suggests that they focus on increasing their popularity on Facebook. We therefore investigate:

Question: *What is the collective linking behaviour of our junk news Facebook pages?*

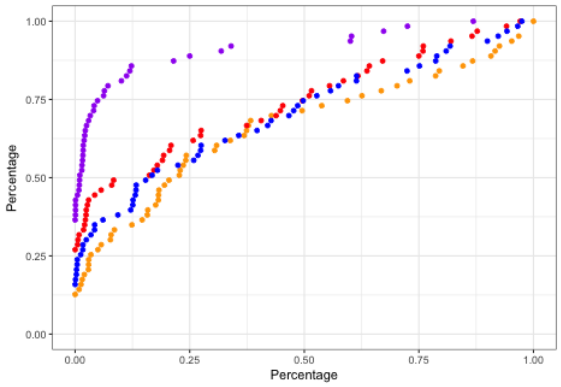
To get all incoming and outgoing connections that form the linking behaviour of a junk news Facebook page we build a directed network. We use a directed network to capture self-references and references from Facebook pages back to their seed domain. We use all distinct Facebook pages and external domains as our nodes. All nodes are connected based on the following connections:

- seed domain - Facebook page; connect each seed domain to its accompanying Facebook page and
- Facebook page - status link; connect each Facebook post to its outgoing connection.

We exempt all posts linking to common content provider (YouTube), non-Facebook social networks (Twitter) and common link shorteners (bit.ly) from our analysis. Because we do not analyse their underlying content, user account respectively domain. All domains discarded are listed in Table 6.3. Due to this decision we discard the 126 posts linking to common content provider, 11 posts linking to other social networks and 85 posts linking to common link shorteners. Additionally, posts with no status link: 63 posts, are also thrown away. As a result, the collective linking behaviour network has 674 nodes N and 940 directed edges, *arcs* A.



(a) Cumulative densities for the indegree (*red*) and outdegree (*blue*) for all junk news Facebook pages in our linking behaviour network. The x-axis is the value of the in- respectively outdegree on a linear scale, while the y-axis shows what proportion for the set of junk news Facebook pages have an in- respectively outdegree up to a particular value.



(b) Zoom in on the outdegree in the linking behaviour network by listing the cumulative densities for the four kind of outgoing connections we recognize; links to the itself (*orange*), its seed domain (*red*), an external Facebook page (*purple*) and an external domain (*blue*). The x-axis is the value of each kind of outgoing connection on a linear scale, while the y-axis shows what proportion of Facebook pages have a particular kind of outgoing connection up to a particular value.

However, before we analyse the collective linking behaviour we review the distinct number of connections for each node using the degree centrality. Separating the incoming and outgoing connections in the indegree respectively outdegree centrality. In Figure 4.7a we list CDFs for the indegree and outdegree for all junk news Facebook pages. Table 4.6 lists the 2.5th, 25th, 50th, 75th and 97.5th percentile for the indegree and outdegree with its weighted mean and standard deviation.

From the indegree centrality CDF in Figure 4.7a we note that only 10% of the Facebook pages have a 1 incoming connection, 75% has 2 incoming connections and the remaining 15% have > 2 incoming connections. Based on how we built the network we infer that 10% of the Facebook pages is linked by just their seed domain, 75% is linked by both their seed domain and the Facebook pages themselves and 15% is linked by additional domains. The additional incoming connections are listed in section 6 in Table 6.4. From Table 6.4 we note that the additional connections have little interaction and therefore do not warrant further analysis.

The CDF for the outdegree centrality in Figure 4.7a show no particular pattern. 5% of the Facebook pages has a 1 outgoing connection, 18% has 1 outgoing connections and its IQR ranges from 3 to 17 connections (Table 4.6). The maximum number of outgoing connection is 206. From the large difference in the mean and median for the outdegree we do note that the maximum number of outgoing connection is an outlier. Without any particular pattern in the distribution of the outdegree we can not ascribe any meaning to the distribution of the outgoing connections as we did for the indegree.

Now, we analyse the collective linking behaviour. We calculate it by counting the number of times a connection occurs; the level of interaction between nodes on a connection. As a result, we get the weighted degree centrality. We categorize all interactions into the four categories and compute the relative proportions of the presence of a category to get the collective linking behaviour. We list the linking behaviour for the individual junk news Facebook pages in section 6 in Figure 6.1 and list the collective linking behaviour for junk news on Facebook in Figure 4.7b.

Most remarkable for the collective linking behaviour is how infrequent pages refer to external Facebook pages. At most 6% of the outgoing links for 75% of all junk news Facebook pages link to an external Facebook page.

The distribution to link to an external Facebook page is logarithmic. All other kinds of outgoing links: links to its seed domain, external domains and self-references are gradual linear where the likeliness to link to any three of these categories of status links differs at most 10%.

Contextualizing the collective linking behaviour to the real-world, we note by combining the proportions of links to the seed and external domains that junk news Facebook pages are primarily interested in guiding users to their own domains and other domains outside the Facebook domain. Thus, junk news Facebook pages focus on profit-maximizing by guiding users to sites outside the Facebook domain ([4];[3];[1];[12]). However, with differences of at most 10% we can not conclusively say that junk news Facebook pages are more likely to refer users to their own seed domain or other non-Facebook domain. We see no signs of junk news Facebook pages explicitly focussing on attracting as many users possible to increase their popularity on Facebook itself. Also, the behaviour for linking to external Facebook pages indicates that cross-referencing Facebook pages is uncommon.

Table 4.5: List the 2.5th, 25th, 50th, 75th and 97.5th percentile for the indegree and outdegree for the linking behaviour network. Additionally, we zoom in on the outdegree and also list the 2.5th, 25th, 50th, 75th and 97.5th percentile for the four kind of outgoing connections we recognize; links to its seed domain, itself, an external Facebook page and an external domain.

Degree	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	**mean**	**std. dev**
indegree	1	2	2	2	5	2.16	1.17
outdegree	1	3	7	17	59	13.92	27.57
Weighted degree (link behaviour)							
seed Facebook page	0	0	16.2	51.06	94.12	26.44	31.69
seed domain	0	3.44	22.73	62.57	100	34.3	34.05
external Facebook page	0	1.22	16.79	52.72	96.56	30.25	32.03
external domain	0	0	1.02	6.21	72.51	9	20.87

item 4.8 visualizes the collective linking behaviour in a network. In item 4.8 we color all nodes belonging to the same community in the same color. We calculate the community of each Facebook page using the *multi-level* community detection algorithm:

> The multi-level algorithm aims to maximize a networks' *modularity* Q. This measure compares the level of connectivity of the communities in our network to a random network with the same weighted degrees where all edges are rewired at random. The random network should be community-free. We optimize the modularity by moving our nodes between communities such that each node makes the local decision to maximize its contribution to Q. The algorithm keeps running until nodes no longer change their membership. If nodes no longer change membership, all communities are collapsed into single nodes and the process continues; its multi-level property. The algorithm itself determines the number of communities it detects.

Notable observations from item 4.8 are: a. a selection of nodes have a large degree, b. nodes with a large degree are not directly connected, c. nodes with a large degree are connected via other nodes and d. several communities of nodes are completely disconnected from the main network.

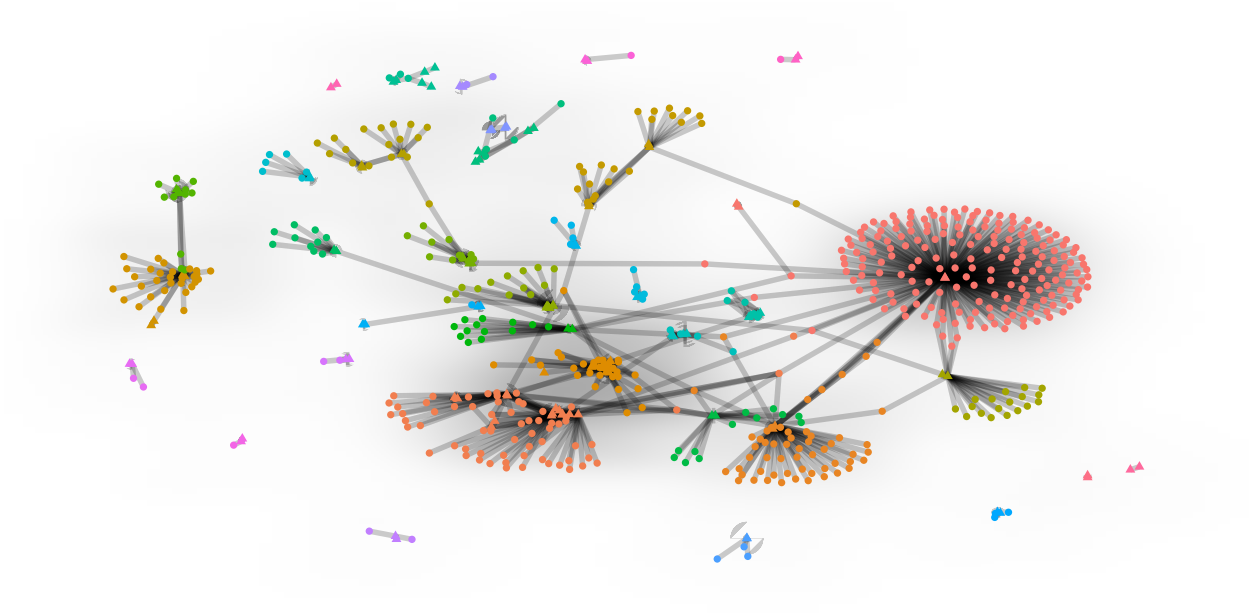


Figure 4.8: Network visualization of our link behaviour network where all distinct Facebook pages and external domains are our nodes which are connected by the two pairwise relations: a. seed domain - Facebook page; this pair connects each seed domain to its corresponding Facebook page, b. Facebook page - status link; this pair represent our linking behaviour.Triangles represent the junk news Facebook pages, circles all other domains and our edges get darker if there is more (undirected) interaction between nodes. Furthermore, its color indicates what community a node belongs to.

4.4 Content overlap between junk news Facebook pages

During the compilation of the seed list Nieuwscheckers got suspicions of duplicate content among junk news seed domains. Here, we research their suspicions by checking if the content junk news Facebook pages publish is similar between different pages. With Facebook parsing all linked content, retrieving the title, cover media and part of its body of text for each link we check if they share content by checking if they share titles verbatim:

Question: *To what extent is content shared by our junk news Facebook pages?*

We split this research question into two subquestions:

- With how many other Facebook pages do the junk news Facebook pages share content?
- How much content is shared between junk news Facebook pages if they share content?

To answer these questions we build an undirected *content overlap* network. Based on the level of connectivity of the network we know how many Facebook pages share content and how much content is shared. We create our content overlap network by first creating an undirected bipartite network with the 60 junk news Facebook pages and 29,273 distinct titles of our Facebook posts as our nodes. In this network all Facebook pages are connected to the titles of the status links they link. We list the titles of the status links verbatim.

For our network we only analyse the Facebook posts of type *link* and not the *photo* posts, *video* posts and *event* posts. On the undirected bipartite network we run bipartite projection to get an unimodal network with just the junk news Facebook pages as our 60 nodes. Here, all junk news Facebook pages which share one or more titles are connected. The edge weight of a connection is the number of distinct titles a pair of Facebook pages share. Using the content overlap network we answer to what extent content is shared by our junk news Facebook pages.

To answer with how many other Facebook pages the junk news Facebook pages share content, we use the degree centrality in Figure 4.10b of the content overlap network. The degree centrality shows that 25% of the Facebook pages has no common titles, 25% shares titles with ≤ 5 Facebook pages, 25% shares titles with $5 <$ Facebook pages ≤ 12 and 25% shares titles with $12 <$ Facebook pages ≤ 24. As for how much content is shared between junk news Facebook pages we use the weighted degree centrality in Figure 4.10d. The weighted degree centrality shows 50% of the Facebook pages share ≤ 2 titles, 25% of the Facebook pages share $2 <$ titles ≤ 10 and some even share > 100 titles. Note that with our approach we are unable to determine if Facebook pages share content by linking to the same domain or if they just copy/steel content among themselves. We are unable to make this distinction due to complexities in how content is posted. Content is, for example, often reposted.

Overall, it is clear that junk news Facebook pages share content. They share content with on average 5 other Facebook pages for on average 2 titles. Figure 4.9 lists our content overlap network as an adjacency matrix. Along its axes we list the names of the junk news' Facebook pages. For each pair of Facebook pages we display the number of shared titles. Having established content overlap among our junk news Facebook pages we also want to check if any particular clustering for Facebook pages exists: do we detect communities among junk news' Facebook pages?

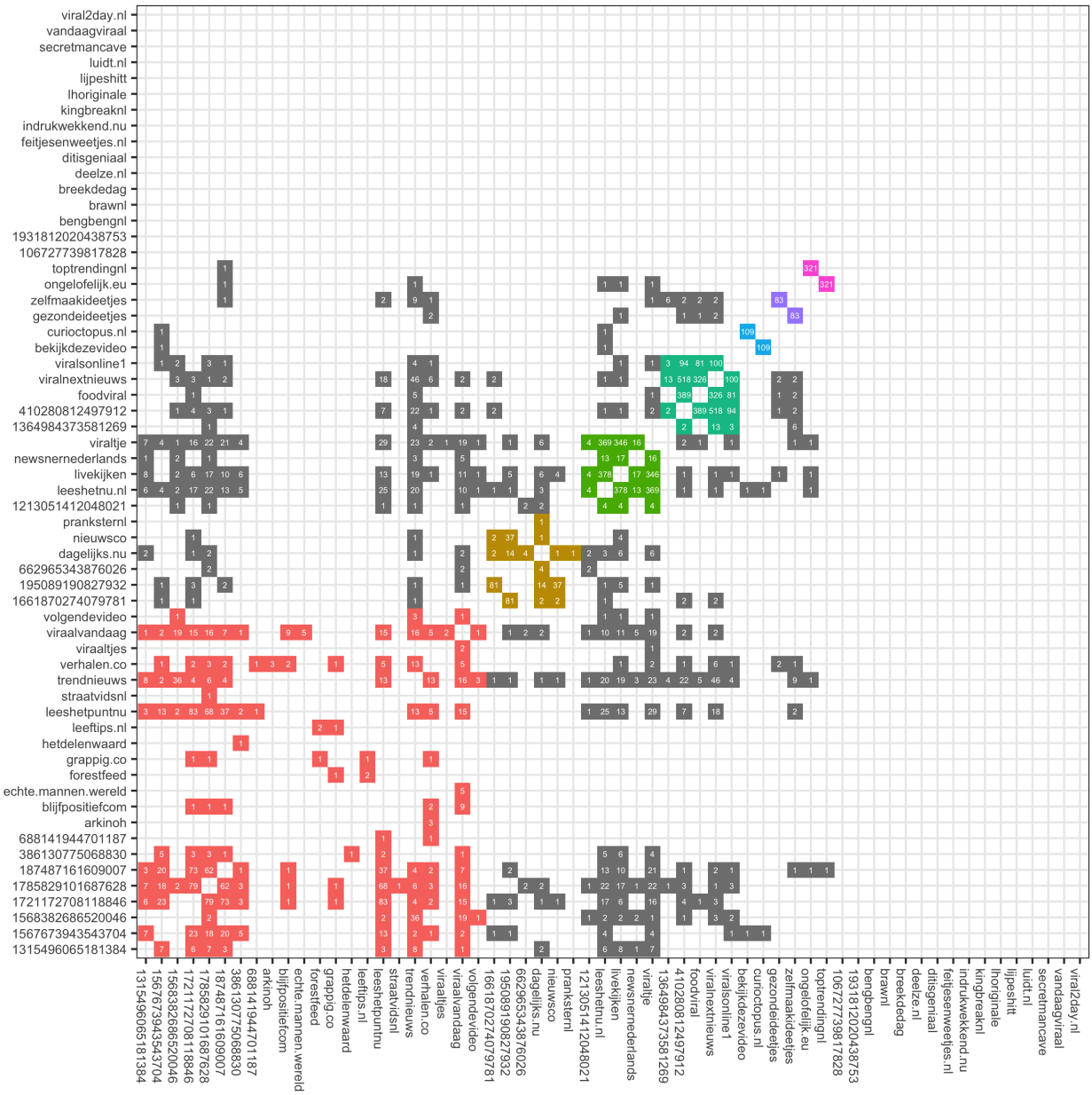
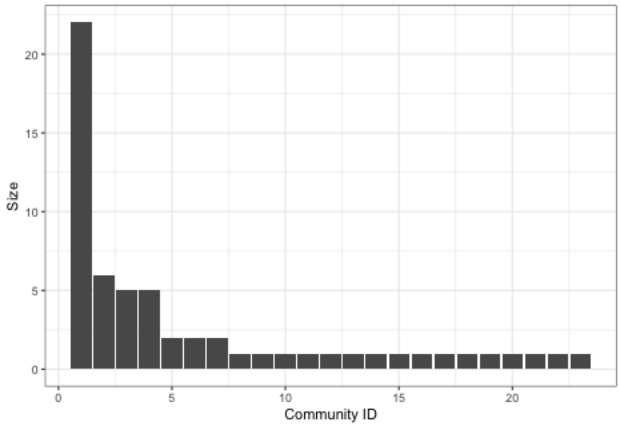


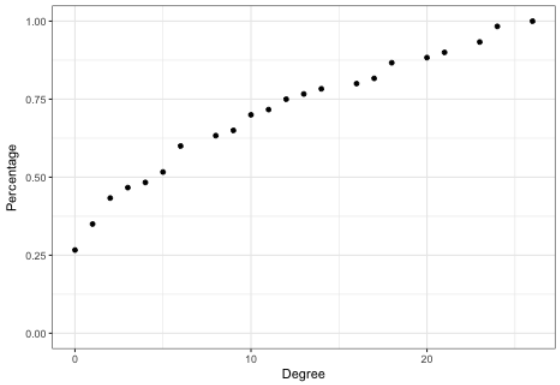
Figure 4.9: Adjacency matrix which visualizes the content overlap network between junk news Facebook pages. Along its axes we list the names of the junk news' Facebook pages. For each pair of Facebook pages we display the number of shared titles and its color indicates the community both Facebook pages belong to. If pages belong to separate communities or no community we color them *gray*. We compute the communities in our network using the $multi-level$ community detection algorithm. Each community gets a numerical value which we use to sort our Facebook pages in descending order. So, Facebook pages belonging to the community with the smallest numerical value (the biggest community) are in the bottom-left corner. Our adjacency matrix is mirrored in its diagonal.

Table 4.6: List the 2.5th, 25th, 50th, 75th and 97.5th percentile for the centrality measures of the content overlap network. Additionally, we list the mean and standard deviation.

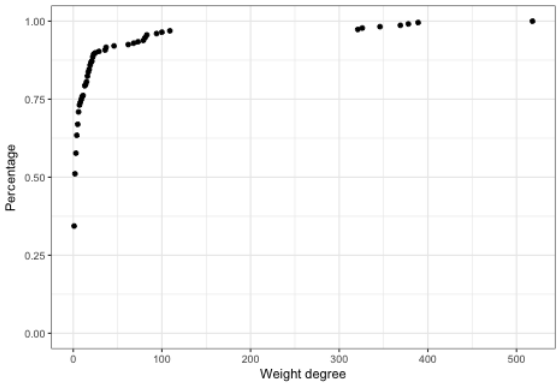
Centrality measure	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	**mean**	**std. dev**
degree	0	0	5	12	24	7.57	8.6
weighted degree	1	1	2	10	326	20.84	73.34
eigenvector	0.0000001	0.000001	0.0001344	0.0095553	0.9645419	0.06	0.22
betweenness	0	0	2.08	40.29	142.95	26.08	42.39



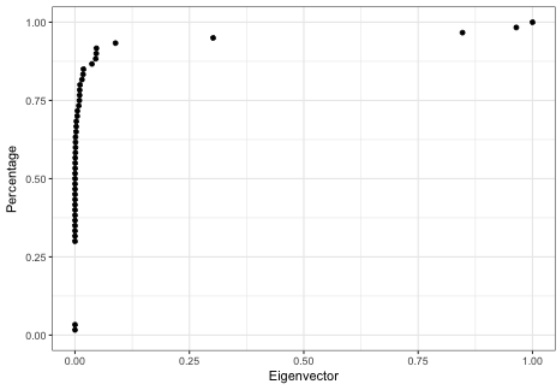
(a) Histogram on the communities in the content overlap network. The x-axis lists all communities in descending order for community size, while the y-axis lists the size of the individual communities.



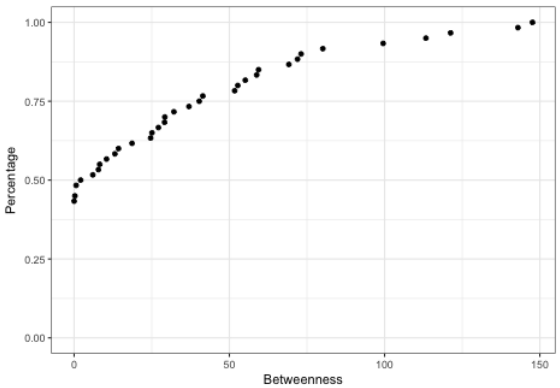
(b) Cumulative densities for the degree for all junk news Facebook pages in our content overlap network. The x-axis lists the value of the degree on a linear scale, while the y-axis lists what proportion for the set of junk news Facebook pages have a degree up to a particular value.



(d) Cumulative densities for the weighted for all junk news Facebook pages in our content overlap network. The x-axis lists the value of the weighted degree, while the y-axis lists what proportion for the set of junk news Facebook pages have a weighted degree up to a particular value.



(c) Cumulative densities for the eigenvector for all junk news Facebook pages in our content overlap network. The x-axis lists the value of the eigenvector, while the y-axis lists what proportion for the set of junk news Facebook pages have a degree up to a particular value.



(e) Cumulative densities for the betweenness for all junk news Facebook pages in our content overlap network. The x-axis lists the value of the betweenness, while the y-axis lists what proportion for the set of junk news Facebook pages have a betweenness up to a particular value.

(f) Centrality measures for the junk news Facebook pages

Question: *Do we detect content sharing communities among junk news' Facebook pages?*

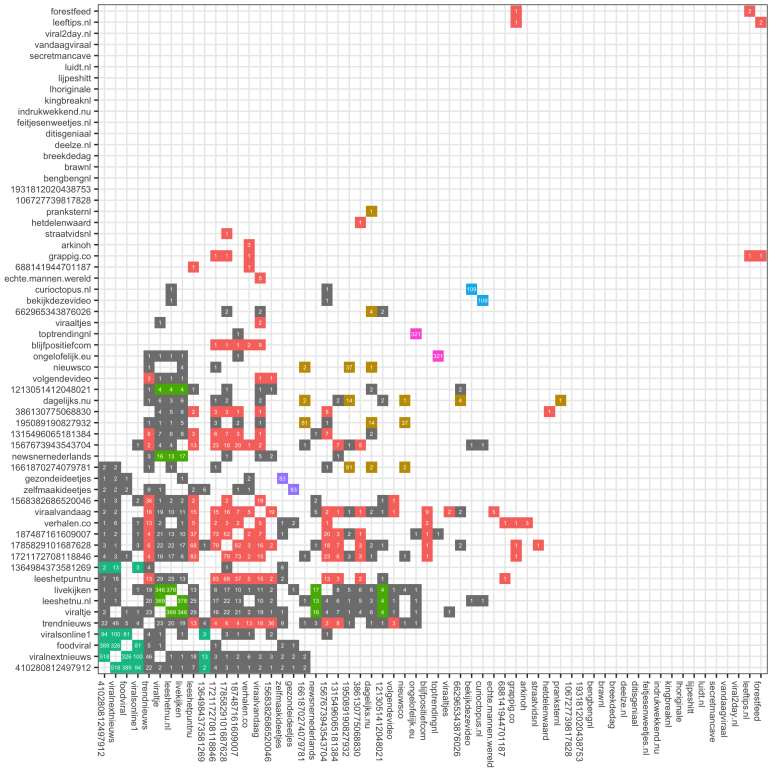
Again, we use the *multi-level modularity* algorithm to look for communities in our network. The color

for each pair of Facebook pages indicates to which community both Facebook pages belong, if any. From our adjacency matrix in Figure 4.9 and barplot for community sizes in Figure 4.10a we find one dominate community containing about one-thirds of all Facebook pages; 24 Facebook pages, three communities with 6 to 5 Facebook pages and three communities of 2 Facebook pages. Communities seem to form around a core of Facebook pages sharing ≥ 100 common titles with the core surrounded by pages with less common titles. As communities shrink in size, Facebook page pairs with less common titles disappear.

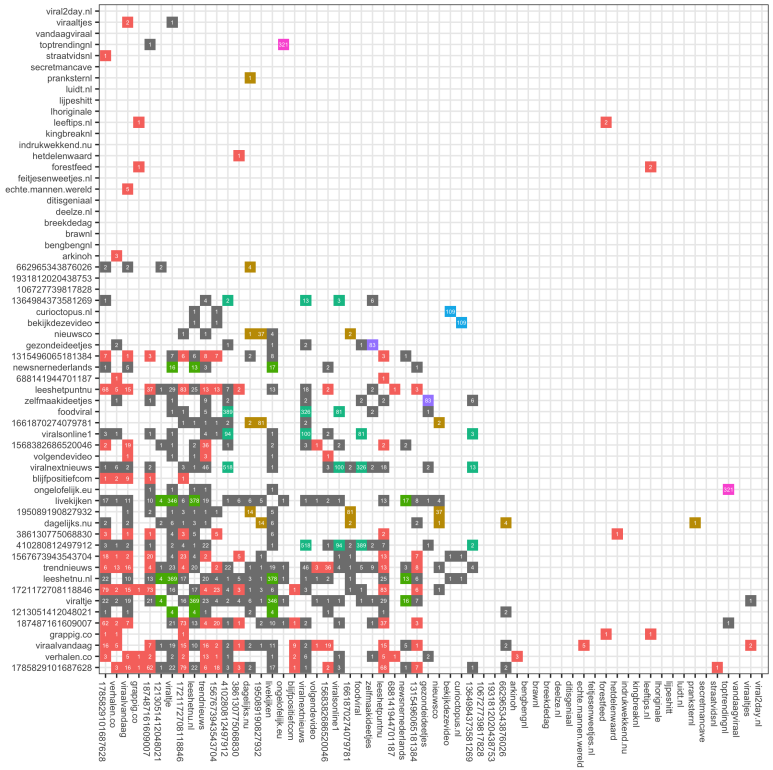
Additionally, we check for the Facebook pages in the content overlap network if there are any which hold a particular role in the network. We are interested in the level of connectivity of a particular Facebook page with (other) important pages; *eigenvector centrality*, and the frequency of a Facebook page to lie on short paths between other pairs of Facebook pages; *betweenness centrality*.

The eigenvector centrality in Figure 4.10c shows that 75% of our Facebook pages hold a value between 0 and 0.01. An insignificant value for the eigenvector, while the other 25% of our Facebook pages have a greater eigenvector. On the adjacency matrix Figure 4.11a we sort the junk news Facebook pages in sort descending order on the eigenvector. Here, we observe that the Facebook pages with greatest eigenvector all belong to the same community. The community of Facebook pages which share the most titles.

The betweenness centrality in Figure 4.10e has greater variance than the eigenvector centrality. 50% of our Facebook pages have a betweenness centrality of ≤ 2, 25% of our Facebook pages have a betweenness centrality of ≤ 40 and 25% of our Facebook pages have a betweenness centrality of ≤ 145. Similar to adjacency matrix in Figure 4.11a we sort the junk news Facebook pages in sort descending order. This time we do so on the betweenness centrality. Interestingly it seems that its distributions aligns nicely with the distribution for the degree in Figure 4.10b. Based on us building the content overlap network using bipartite project and connecting Facebook pages sharing one or more titles, we note that this shape for the CDF of the betweenness centrality is in line with we expected. Finally, we notice that the betweenness centrality adjacency matrix in Figure 4.10e is significant less clustered around its communities than we found for the eigenvector.



(a) Eigenvector adjacency matrix which visualizes the content overlap network between junk news Facebook pages. Along its axes we list the names of the junk news' Facebook pages. For each pair of Facebook pages we display the number of shared titles and its color indicates the community both Facebook pages belong to. If pages belong to separate communities or no community we color them *gray*. We compute the communities in our network using the $multi-level$ community detection algorithm. For each Facebook page we calculate its **eigenvector** which we use to sort our Facebook pages in descending order. So, Facebook pages with the highest eigenvector are in the bottom-left corner. Our adjacency matrix is mirrored in its diagonal.



(b) Betweenness adjacency matrix which visualizes the content overlap network between junk news Facebook pages. Along its axes we list the names of the junk news' Facebook pages. For each pair of Facebook pages we display the number of shared titles and its color indicates the community both Facebook pages belong to. If pages belong to separate communities or no community we color them *gray*. We compute the communities in our network using the $multi-level$ community detection algorithm. For each Facebook page we calculate its **betweenness** which we use to sort our Facebook pages in descending order. So, Facebook pages with the highest betweenness are in the bottom-left corner. Our adjacency matrix is mirrored in its diagonal.

Figure 4.11: Adjacency matrices for eigenvector and betweenness sorted in descending order

5 Conclusions

The central research question for our work was:

- What is the reach of Dutch junk news on Facebook compared to mainstream news?

To answer this question we gathered all posts published by our 63 junk news and 20 mainstream news Facebook pages between Jan 2013 and Dec 2017. In this period junk news published 1.36% more posts than mainstream news with 16.6% more reactions, 51.65% more comments and 54.8% more shares. Recalling that we define reach as the collective user engagement on all posts of a set Facebook pages in terms of its number of reactions, comments and shares, we note that junk news has a greater reach than mainstream news. An in-depth comparison between the distributions of junk and mainstream news has shown that the greater user engagement on junk news is a trend. The differences are not caused by outliers, such as viral posts. All engagement metrics show a trend.

With the growing number of Facebook users in the Netherlands users, going from 9.6 million users in 2014 to 10.4 Million users in 2017 [21] we wondered if timebased changes influence the user engagement:

- What is the reach of junk news on Facebook compared to mainstream news over time?

First, we investigated if the use of junk and mainstream news on Facebook as a platform changed over time by evaluating their post activity; the number of posts they published per month. Thoughout the period we analyse, we saw that the post activity of mainstream news stayed consistent. The post activity of junk news was more irregular; starting with just a few posts per month in Jan 2013, matching the post activity of mainstream news at the start of 2016 and drastically increasing its post activity at the end of 2017. While the post activity of the individual junk news increased, these increases were modest in comparison to the increase in the number of publishing junk news Facebook pages per month. For the monthly junk news publishing Facebook pages we went from a single Facebook page in Jan 2013 to 63 of such pages in Dec 2017.

In addition to its growing post activity we also saw junk news to consistently attract more user engagement on a per post basis from the beginning of 2016. Differences in the user engagement got as great as twice as much for the reactions and comments and as much 4 to 5 times for the shares.

We also explore the objective of junk news on Facebook. However, with the data we acquired from the Facebook API we could not directly determine the objective of junk news on Facebook. Therefore, we infer the objective of junk news on Facebook using their collective linking behaviour:

- What is the collective linking behaviour of our junk news Facebook pages?

For the collective linking behaviour we categorized all status links for all posts of an individual Facebook page into one of four categories and calculated the relative proportions of each category of status link. We recognized four kinds of status links; links to its seed domain, the Facebook page itself, an *external domain*; a non-facebook.com domain that is not its seed domain, and an external Facebook page. An external Facebook page is a facebook.com domain that is not the Facebook page itself.

Most remarkable for the collective linking behaviour is how infrequent pages refer to external Facebook pages. Junk news Facebook pages are unlikely to cross-reference each other. All other kinds of outgoing links: links to its seed domain, external domains and self-references show similar proportions to be linked to for the set of junk news Facebook pages. By combining the proportions of links to the seed and external domains we

note that junk news Facebook pages are primarily interested in guiding users to their own and other sites outside the Facebook domain. Thus, junk news Facebook pages focus on profit-maximizing by guiding users to the advertising supported sites they own outside the Facebook domain ([4];[3];[1];[12]).

Finally, we analysed Nieuwscheckers' suspicions that junk news Facebook pages share content:

- To what extent is content shared by our junk news Facebook pages?

We split this question into two sub-questions: a. With how many other Facebook pages do the junk news Facebook pages share content? b. How much content is shared between junk news Facebook pages if they share content? Junk news Facebook share content if they share titles verbatim for the content they link. We found that 25% of the Facebook pages has no common titles, 25% shares titles with ≤ 5 Facebook pages, 25% shares titles with $5 <$ Facebook pages ≤ 12 and 25% shares titles with $12 <$ Facebook pages ≤ 24. From those which shared content 50% of the Facebook pages share ≤ 2 titles, 25% of the Facebook pages share $2 <$ titles ≤ 10 and remainder sometimes even shares > 100 titles. Thus, junk news Facebook pages share content with on average 5 other Facebook pages for on average 2 titles. Note that with our approach we were unable to determine if Facebook pages share content by linking to each other or if they just copy/steel content among themselves.

5.1 Future work

Future work on the reach of junk news on Facebook could learn a classifier which based on its published content determines if we got a junk or mainstream news Facebook page. The classifier could for example be learnt on the style of the textual content of a Facebook page or it could focus on the content of the text itself using text analysis. With topic modelling we could determine if certain subjects attract greater user engagement.

Apart from analysing just the text, future work could also analyse photos and videos posted by junk news. This is especially interesting when we look at recent advances in modifying and creating photos and video via machine learning ([13]; [10]). With currently most false audiovisual content is created manually, automatic methods to create such content could drastically increase the amount of audiovisual "fake news". Making it an interesting subject to research.

6 Appendix

Table 6.1: List the seed domains for junk news with their accompanying Facebook page.

url	fb_page	page_id
architectdistrict.nl	facebook.com/architectdistrict	architectdistrict
arkinoh.com	facebook.com/arkinoh	arkinoh
bekijkdezevideo.nl	facebook.com/bekijkdezevideo	bekijkdezevideo
bekijkhetnu.com	facebook.com/lijpeshitt	blijfpositiefcom
blijf-positief.com	facebook.com/blijfpositiefcom	lijpeshitt
brakkaboys.nl	facebook.com/106727739817828	106727739817828
braw.nl	facebook.com/brawnl	brawnl
breekdedag.nl	facebook.com/breekdedag	breekdedag
curioctopus.nl	facebook.com/curioctopus.nl	curioctopus.nl
dagelijks.nu	facebook.com/dagelijks.nu	dagelijks.nu
dagelijksfilmpje.nl	facebook.com/386130775068830	386130775068830
deelze.nl	facebook.com/deelze.nl	deelze.nl
ditisgeniaal.nl	facebook.com/ditisgeniaal	ditisgeniaal
doedatzelf.nl	facebook.com/410280812497912	410280812497912
echtemannenwereld.nl	facebook.com/echte.mannen.wereld	echte.mannen.wereld
eetradar.nl	facebook.com/eetradar	eetradar
fantastisch.co	facebook.com/1785829101687628	1785829101687628
feitjes-weetjes.nl	facebook.com/feitjesenweetjes.nl	feitjesenweetjes.nl
forestfeed.nl	facebook.com/forestfeed	forestfeed
gezondeideetjes.nl	facebook.com/gezondeideetjes	gezondeideetjes
grappig.co	facebook.com/grappig.co	grappig.co
hetdelenwaard.net	facebook.com/hetdelenwaard	hetdelenwaard
kijkhet.nl	facebook.com/1567673943543704	1567673943543704
kijkhet.nl	facebook.com/572385656268218	572385656268218
kingbreak.nl	facebook.com/kingbreaknl	kingbreaknl
kookfans.nl	facebook.com/1661870274079781	1661870274079781
leeftips.nl	facebook.com/leeftips.nl	leeftips.nl
leeshet.nu	facebook.com/leeshetpuntnu	leeshetpuntnu
leeshetnu.nl	facebook.com/leeshetnu.nl	leeshetnu.nl
lekkerwonen.org	facebook.com/trendnieuws	trendnieuws
lhviraal.com	facebook.com/lhoriginale	lhoriginale
livekijken.nl	facebook.com/livekijken	livekijken
luidt.nl	facebook.com/luidt.nl	luidt.nl
memesisleven.nl	facebook.com/indrukwekkend.nu	indrukwekkend.nu
niet100.tv	facebook.com/1931812020438753	1931812020438753

Continued on next page

url	fb_page	page_id
nieuws.co	facebook.com/nieuwsco	nieuwsco
nieuwsviraalvandaag.nl	facebook.com/viraalvandaag	viraalvandaag
nl.newsner.com	facebook.com/newsnernederlands	newsnernederlands
ongelofelijk.eu	facebook.com/ongelofelijk.eu	ongelofelijk.eu
ongelooflijk.co	facebook.com/1315496065181384	1315496065181384
prankster.nl	facebook.com/pranksternl	pranksternl
secretmancave.nl	facebook.com/secretmancave	secretmancave
straatmedia.tv	facebook.com/662965343876026	662965343876026
straatvids.nl	facebook.com/straatvidsnl	straatvidsnl
suri.nu	facebook.com/1213051412048021	1213051412048021
tipsenweetjes.nl	facebook.com/195089190827932	195089190827932
toptrending.nl	facebook.com/toptrendingnl	toptrendingnl
vandaagviraal.nl	facebook.com/vandaagviraal	vandaagviraal
verhalen.co	facebook.com/verhalen.co	verhalen.co
videodump.nu	facebook.com/1568382686520046	1568382686520046
viraal.co	facebook.com/187487161609007	187487161609007
viraalnederland.nl	facebook.com/bengbengnl	bengbengnl
viraaltjes.nl	facebook.com/viraaltjes	viraaltjes
viraaltv.nl	facebook.com/688141944701187	688141944701187
viral2day.nl	facebook.com/viral2day.nl	viral2day.nl
viralfood.nl	facebook.com/foodviral	foodviral
viralnext.nl	facebook.com/viralnextnieuws	viralnextnieuws
viralsonline.com	facebook.com/viralsonline1	viralsonline1
viraltje.nl	facebook.com/viraltje	viraltje
volgendevideo.nl	facebook.com/volgendevideo	volgendevideo
vrouwenhumor.com	facebook.com/1364984373581269	1364984373581269
wtfbro.nl	facebook.com/1721172708118846	1721172708118846
zelfmaakideetjes.nl	facebook.com/zelfmaakideetjes	zelfmaakideetjes

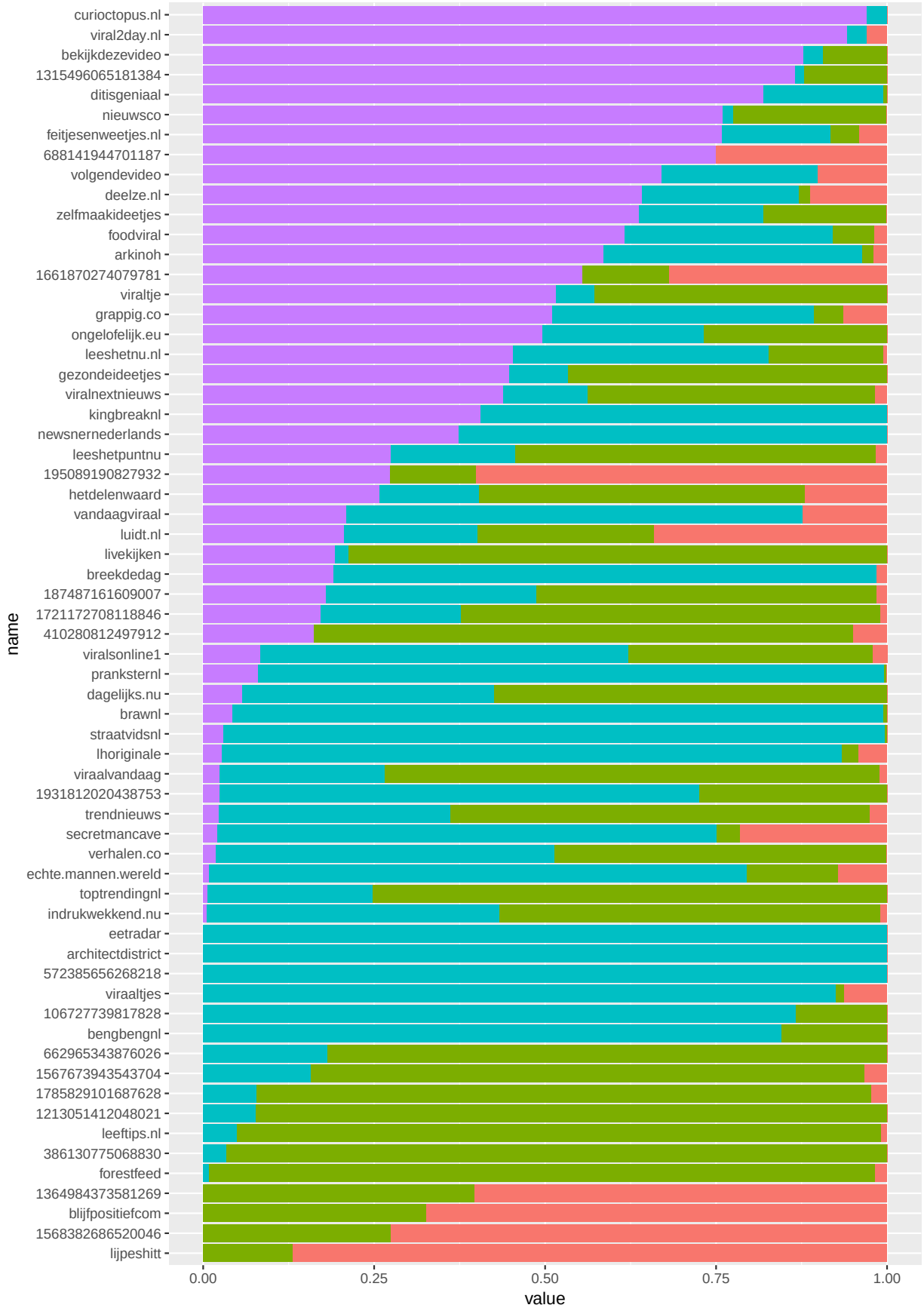


Figure 6.1: List the linking behaviour for the each individual junk news Facebook pages. The x-axis lists the relative proportions for all four kinds of links; links to the itself (*purple*), its seed domain (*blue*), an external Facebook page (*green*) and an external domain (*red*), while the y-axis lists the junk news Facebook pages. Additionally, we sort the relative proportions as follows: Facebook pages with most links to their seed domain are at the top. Since no pages have an exactly similar proportions of links to their seed domain we see no additional sorting among those. Subsequently, we list the Facebook pages with the most self-references, unknown domains and unknown Facebook pages.

Table 6.2: This table lists the seed domains for mainstream news with their accompanying Facebook page-url and name.

url	fb_page	page_id
ad.nl	facebook.com/ad.nl	ad.nl
bnr.nl	facebook.com/bnr.nieuwsradio	bnr.nieuwsradio
decorrespondent.nl	facebook.com/decorrespondent	decorrespondent
elsevierweekblad.nl	facebook.com/elsevierweekblad	elsevierweekblad
fd.nl	facebook.com/hetfd	hetfd
geenstijl.nl	facebook.com/geenstijlnl	geenstijlnl
groene.nl	facebook.com/190231243842	190231243842
hpdetijd.nl	facebook.com/103652819717294	103652819717294
hpdetijd.nl	facebook.com/103652819717294	nporadio1
metronieuws.nl	facebook.com/metro	metro
nos.nl	facebook.com/nos	nos
nrc.nl	facebook.com/nrc	nrc
nu.nl	facebook.com/nu.nl	nu.nl
parool.nl	facebook.com/paroolnl	paroolnl
rd.nl	facebook.com/refdag	refdag
rtlnieuws.nl	facebook.com/rtlnieuws	rtlnieuws
telegraaf.nl	facebook.com/telegraaf	telegraaf
tpo.nl	facebook.com/tponl	tponl
trouw.nl	facebook.com/trouw.nl	trouw.nl
vk.nl	facebook.com/volkskrant	volkskrant

Table 6.3: List the links for the common content provider, non-Facebook social networks and link shorteners we do not use in our content overlap network. Also, we list all content provider and non-Facebook social networks together we their link shortener if applicable.

Content providers
youtube.com (youtu.be)
tumblr.com
giphy.com (gph.is)
gifsnation.com
imgur.com
photobucket.com
Social network
instagram.com
reddit.com (i.redd.it)
pinterest.com (pinimg.com)
twitter.com (t.co)
Link shorteners
bit.ly
buff.ly
goo.gl

Table 6.4: Junk news Facebook pages with an indegree greater than > 2; those linked to by other domains than their seed domain and Facebook page, are scarce. We therefore explicitly list these links in this table with the number of times such a connection occurs.

source	**target**	**count**
facebook.com/1721172708118846	facebook.com/leeshetpuntnu	1
facebook.com/187487161609007	facebook.com/leeshetpuntnu	1
facebook.com/410280812497912	facebook.com/foodviral	1
facebook.com/410280812497912	facebook.com/viralnextnieuws	1
facebook.com/410280812497912	facebook.com/viralsonline1	1
facebook.com/arkinoh	facebook.com/verhalen.co	1
facebook.com/breekdedag	facebook.com/trendnieuws	1
facebook.com/echte.mannen.wereld	facebook.com/vandaagviraal	1
facebook.com/feitjesenweetjes.nl	facebook.com/viraaltjes	1
facebook.com/trendnieuws	facebook.com/viralnextnieuws	1
facebook.com/viral2day.nl	facebook.com/viraalvandaag	1
facebook.com/viralnextnieuws	facebook.com/foodviral	1
facebook.com/viralnextnieuws	facebook.com/viralsonline1	1
facebook.com/viralsonline1	facebook.com/foodviral	1
facebook.com/viralsonline1	facebook.com/viralnextnieuws	1
facebook.com/zelfmaakideetjes	facebook.com/gezondeideetjes	1

References

[1] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–36. DOI: 10.1257/jep.31.2.211. URL: http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[2] Mike McIntire Andrew Higgins and Gabriel J.X. Dance. *Inside a Fake News Sausage Factory: 'This Is All About Income'*. Nov. 2016. URL: https://www.nytimes.com/2016/11/25/world/europe/fake-news-donald-trump-hillary-clinton-georgia.html.

[3] Vian Bakir and Andrew McStay. "Fake News and The Economy of Emotions". In: *Digital Journalism* 6.2 (2018), pp. 154–175. DOI: 10.1080/21670811.2017.1345645. eprint: https://doi.org/10.1080/21670811.2017.1345645. URL: https://doi.org/10.1080/21670811.2017.1345645.

[4] Joanna M. Burkhardt. *Chapter 1. History of Fake News*. URL: https://journals.ala.org/index.php/ltr/article/view/6497/8631.

[5] Emilio Ferrara et al. "The Rise of Social Bots". In: *Commun. ACM* 59.7 (June 2016), pp. 96–104. ISSN: 0001-0782. DOI: 10.1145/2818717. URL: http://doi.acm.org/10.1145/2818717.

[6] Richard Fletcher et al. "Measuring the reach of "fake news" and online disinformation in Europe". In: (2018). URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/Measuring%20the%20reach%20of%20fake%20news%20and%20online%20distribution%20in%20Europe%20CORRECT%20FLAG.pdf.

[7] *Graph API - Documentation*. URL: https://developers.facebook.com/docs/graph-api/.

[8] Mark S. Handcock. *Relative Distribution Methods*. Version 1.6-6. Project home page at http://www.stat.ucla.edu/~handcock/RelDist. Los Angeles,CA, 2016. URL: https://CRAN.R-project.org/package=reldist.

[9] Mark S. Handcock and Martina Morris. *Relative Distribution Methods in the Social Sciences*. ISBN 0-387-98778-9. New York: Springer, 1999. URL: http://www.stat.ucla.edu/~handcock/RelDist.

[10] *Lip-syncing Obama: New tools turn audio clips into realistic video*. URL: https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/.

[11] maicolengel butac maicolengel. *The Reuters Institute for the Study of Journalism vs fake news*. Feb. 2018. URL: http://www.butac.it/the-reuters-institute-for-the-study-of-journalism-vs-the-fake-news/.

[12] Alice Marwick and Rebecca Lewis. *Media Manipulation and Disinformation Online*. URL: https://www.benton.org/headlines/media-manipulation-and-disinformation-online.

[13] Cade Metz and Keith Collins. *How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos*. Jan. 2018. URL: https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html.

[14] V Narayanan et al. "Polarization, Partisanship and Junk News Consumption over Social Media in the US". In: *arxiv.org* (2018). URL: https://arxiv.org/abs/1803.01845.

[15] Nic Newman, David A. L. Levy, and Rasmus Kleis Nielsen. "Reuters Institute Digital News Report 2017". In: *SSRN Electronic Journal* (2017). URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf.

[16] Rasmus Kleis Nielsen et al. *"News you don't believe": Audience perspectives on fake news*. URL: http://www.digitalnewsreport.org/publications/2017/news-dont-believe-audience-perspectives-fake-news/.

[17] Katie Rogers and Jonah Engel Bromwich. *The Hoaxes, Fake News and Misinformation We Saw on Election Day*. Dec. 2017. URL: https://www.nytimes.com/2016/11/09/us/politics/debunk-fake-news-election-day.html.

[18] *Selenium (software)*. May 2018. URL: https://en.wikipedia.org/wiki/Selenium_(software).

[19] Elisa Shearer and Jeffrey Gottfried. *News Use Across Social Media Platforms 2017*. Sept. 2017. URL: http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.

[20] Craig Silverman. *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook*. URL: https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

[21] Statista. *Number of Facebook users in the Netherlands 2014-2018 | Statistic*. 2018. URL: https://www.statista.com/statistics/283635/netherlands-number-of-facebook-users/ (visited on 03/06/2018).

[22] Samanth Subramanian. *The Macedonian Teens Who Mastered Fake News*. May 2017. URL: https://www.wired.com/2017/02/veles-macedonia-fake-news/.

[23] Craig Timberg. *Russian propaganda effort helped spread 'fake news' during election, experts say*. Nov. 2016. URL: https://www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe_story.html?noredirect=on&utm_term=.d2c317fe2fb7.

[24] Tess Townsend. *The Bizarre Truth Behind the Biggest Pro-Trump Facebook Hoaxes*. Nov. 2016. URL: https://www.inc.com/tess-townsend/ending-fed-trump-facebook.html.

[25] JA Tucker et al. *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Mar. 2018. URL: https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.