What happens if we ask a machine about online reliability a view from the network and a meta-evaluation (PREPRINT VERSION)

Tobias Blanke, Tommaso Venturini

What happens if we ask a machine about online reliability a view from the network and a meta-evaluation

Tobias Blanke, University of Amsterdam, t.blanke@uva.nl
Tommaso Venturini, CNRS, France, Tommaso Venturini, tommaso.venturini@cnrs.fr

Abstract:
This article shows how a machine would reason about the complex social relations of news reliability using a network view. Such a network view promises a topic-agnostic perspective that can be a useful hint on reliability trends and their heterogeneous assumptions. In our analysis, we depart from the ever-growing numbers of papers trying to find machine learning algorithms to predict news sites reliability and focus instead on using machine reasoning to understand the structure of news networks by comparing it with our human judgements. Understanding and representing news networks is not easy, not only because they can be extremely vast but also because they are shaped by several overlapping network dynamics. Here we present a machine learning approach to analyse what constitutes reliable news from the view of a network. This method allows us to test relations from statistical macro perspectives to individual micro relations and to work across a large dimensional space of different features. It can integrate heterogenous features from network relations and can combine several types of network attachments. Our aim is to machine-read a network's understanding of reliability and makes it different from all the other sites. To analyse real-life news sites, we used the Décodex dataset to train a machine learning model so that it can identify reliability drawing on the structure of the underlying network. Finally, we used the results from the machine reading to meta-evaluate the work of human news site assessors.

1. Introduction

'Fake news' seem to be an ever-growing concern in the digital media age. Almost daily there are news stories about how it spreads quickly across the Internet and social media sites. In 2016, nationalreport.net, a site well-known for spreading unreliable information, claimed that customers in Colorado marijuana shops were using food stamps to buy marijuana. This claim had no basis but spread so quickly that Colorado House Representatives proposed legislation to prevent people from using their food stamps to buy marijuana.

With fake news concerns, fact-checking sites have also mushroomed to empower individuals to discern reliable news. The Reporters Lab at Duke University lists almost 300 active sites from all over the world.[1] Fact checkers are, however, confronted by the challenge that on the Internet, the line between journalism and other content has blurred. Usefulness and trustworthiness are at the centre of public concerns regarding fact-checking services (Brandtzaeg and Følstad, 2017). Pennycook and Rand (2019) conducted an online experiment that showed that people believe to make their own choices and think that they prefer traditional news sites. Their crowd-sourcing experiment also demonstrated that 'politically balanced layperson ratings were strongly correlated with ratings provided by professional fact-checkers' (Pennycook and Rand, 2019, p. 2521). Crowdsourcing might thus be an option for establishing the reliability of news sites.

Next to such crowdsourcing work and other digital methods (Bounegru et al., 2017), fake news has been targeted with machine learning to predict 'fake news'. There are many challenges to employing machine learning to detect fake news. Castelo et al. (2019) discuss issues that stem from the dynamic nature of online news. Any consideration of 'correct' facts will quickly become outdated, as new political developments lead to new online discourses. Classifiers will thus age fast. Castelo et al. (2019) show that 'topic-agnostic classification strategies' can offer some remedy. The authors are mainly interested in linguistic features such as 'morphological patterns in texts' or 'readability of texts'. In (Naeem et al., 2020), deep learning is employed to detect 'certain natural language cues' to find patterns of fake news in click bait. Monti et al. (2019) promote another 'topic-agnostic' viewpoint on fake news with 'propagation-based approaches', relying on the different patterns that fake news propagates across social media. We also follow a topic-agnostic view on fake news but, like Kwon et al. (2013), we rely on graph-theoretical features such as centrality, etc. Where Ravandi and Mili (2019) demonstrate how graph analysis can be used to analyse general polarisation issues in simulations of news network, we are interested how graph-based machine learning can be used to analyse a real-life news digital ecosystem.

Lazer et al. (2018) provide a readable, comprehensive overview of scientific challenges to computationally defining the reliability of news that require broad interdisciplinary work. Similarly, Ciampaglia (2018) argues for an increase role of computational social scientists in the fight against fake news. Social and computer sciences must work together to identify generalisable mechanisms that work operate in 'large-scale interactive systems' (Keuschnigg et al., 2018). We follow such demands for a new interdisciplinary research agenda on fake news and add a new viewpoint to the fake news discussion that considers machine learning not so

[1] https://reporterslab.org/fact-checking/

much a tool to filter out fake news but as a tool to understand the relationships between fake and traditional news sites on the web. The above approaches mostly target a prediction of whether a fact, a text or a whole site belong to fake news. This paper takes a different approach and rather employs machine learning to present how a computer would understand *our* understanding of fake news from the point of view of a whole ecosystem of news sites. Based on past work, we selected carefully a number of network features such as neighbourhoods and centrality to establish how a machine would read a news ecosystem. We discover the heterogenous challenges involved in consistently establishing what constitutes 'right' facts vs. fake news, which raises the question who the fact checkers are and who checks on them. This paper therefore targets social questions around the heterogeneity of human decisions on the reliability of news sites rather than the previously cites computer sciences work that uses machine learning to detect and predict fake news.

2. A view from the network on online reliability

Our work on the reliability of news sites focusses on a network view. Using the structure of digital networks to reason about the content of its nodes is not easy, not only because they can be extremely vast but also because its organisation is shaped by two *attachment dynamics* that push online networks in two orthogonal directions. The first dynamics, *communal attachment*, consists of the fact that websites, social media accounts, etc. tend to connect to other websites and accounts that focus on the same topics, issues or matter of interests (Ackland and Shorish, 2014; Centola et al., 2007). Blogs devoted to fly-fishing, for instance, tend to link to fellow fly-fishing blogs more than to blogs dedicated to other types of fishing or other leisure activities. The second dynamics, *preferential attachment*, is related to the fact that websites, for instance, that are already highly cited have a higher probability to attract new hyperlinks (Albert et al., 1999). These two dynamics are equally important but also diametrically opposed (Leskovec et al., 2009; Newman, 2001; Vosoughi et al., 2018). While communal attachment encourages homophily and tends to generate thematic communities where shared interests are discussed by likeminded actors, preferential attachment encourages hierarchy and tends to create a pyramid of attention concentrated around a few hyper-visible actors. Communal attachment goes in circles (within the same community) and creates clustering; preferential attachment goes upward (toward the most visible) and creates ranking.

The difficulty to consider these two dynamics together and to combine their opposed effects has produced a twofold reading of networks. Recently, for example, the mounting alarm about online news misinformation has produced two opposite concerns. On the one hand, commentators have denounced the emergence of increasingly tight news "filter bubbles" (Pariser, 2011; Sunstein, 2001) trapping online users within closed conversations and preventing them from being exposed to different ideas and viewpoints. On the other hand, observers have warned against the amplification of viral stories which capture a disproportionate portion of online attention and reduce the diversity of news consumption (Hindman, 2008; Nelson and Taneja, 2018). However, they are hardly discussed together.

Questions of communal and preferential attachment arise organically in the context of understanding the reliability of news sites as we analyse a series of news sources selected and categorised by *Le Monde* journalists according to their reliability, as well as the network formed by the hyperlinks connecting them. To account for the *combination* of the two types of attachment described above, we use machine learning techniques capable of exploiting in an integrated way network metrics that cover both types of attachments. Here, our objective is not so much to use learning algorithms to predict the reliability of news sources, as has been done in many other studies (Conroy et al., 2015), but rather to exploit them for their capacity to explore data along different and otherwise hardly commensurable dimensions. In this sense, we try to bring together what Wallach (2018) sees as the fundamental methodological difference between interpretation-oriented social scientists and computer scientists, which are mostly concerned with a model that produces great accuracy but do not care as much for the why. We work with a highly curated dataset, we introduce in the next section, rather than concentrate on scale and accuracy.

We want the machine to read the data and learn from this reading about attachment relationships in the news networks. This approach could be called 'distant reading of networks'. The method is inspired by the better-known 'distant reading' methods (Jänicke et al., 2015), used to complement human reading of texts with machine reasoning. The method is called 'distant', because it relies on machines rather than humans to explore relationships in news networks.

The Data: The Décodex Network

The intertwining of communal and preferential attachment, which characterises both social and digital networks, is easily observable in the dataset of news sources that constitutes the case study for this article. We built this network drawing on a list of online news sources catalogued by the 'Décodex' tool.[2] In this fact-checking initiative, the journalist of *Le Monde* reviewed several hundreds of websites active during the 2017 French presidential campaign and evaluated their trustworthiness according to four categories: 'reliable', 'imprecise', 'unreliable' and 'satirical'. News sites have been broadly interpreted by the *Le Monde* team and include gamer community sites or *Yahoo!*. A few months ahead the French elections, we extracted the 667 URLs reviewed by the 'Décodex' and used the Web Crawler Hyphe (Jacomy et al., 2016) to map the hyperlinks connecting them.

The work fits into the large body of current examinations that try to match and even predict whether a news source is reliable (Ahmed et al., 2017; Gilda, 2017; Wang, 2017) but we were interested in understanding how the hyperlink network topology indicates a typology of reliability of sites. Huibers (1996) has demonstrated how this has been successfully done in hypermedia retrieval for over twenty years. The information from hyperlinks and the clusters they form are used to rank documents on the world wide web. The most famous application of this is the original Google PageRank algorithm (Page et al., 1999), which works in two steps. First the content of a site is evaluated, and afterwards the typological neighbourhood of a site is used to reinforce the content-based ranking. The link authority of a page is employed to finally rank it. The higher this authority compared to a random walker the higher the ranking of the page. The authority value is entirely derived from the topology.

PageRank thus combines content and structure in a very successful hypermedia retrieval model. Using the typology alone to evaluate the topological is, however, a further challenge especially in the context of specific decisions such as on reliability. To demonstrate issues with network community detection of the reliability topology in Décodex, let us consider briefly its basic hyperlink relationships based on relatively simple link statistics. We can, for instance, compute a naïve probability of reliability for each source as the simple ratio between the reliable news site directly connected to it and the total number of its neighbours (its degree). This would be a reasonable description of a community approach where reliable sites cluster together and follows on from existing work like (Conroy et al., 2015) where linkages in networks are thought to induce 'trust dimensions'. According to this naïve neighbour classifier, *Le Point* achieves a top score, as it is the least connected to unreliable ones. *Le Point* is a centre-right conservative weekly in France. Discovering it as a reliable news source fits with research such as (Vosoughi et al., 2018), which considers speed of updates to be a distinct feature of unreliable news sites.

[2] www.lemonde.fr/verification

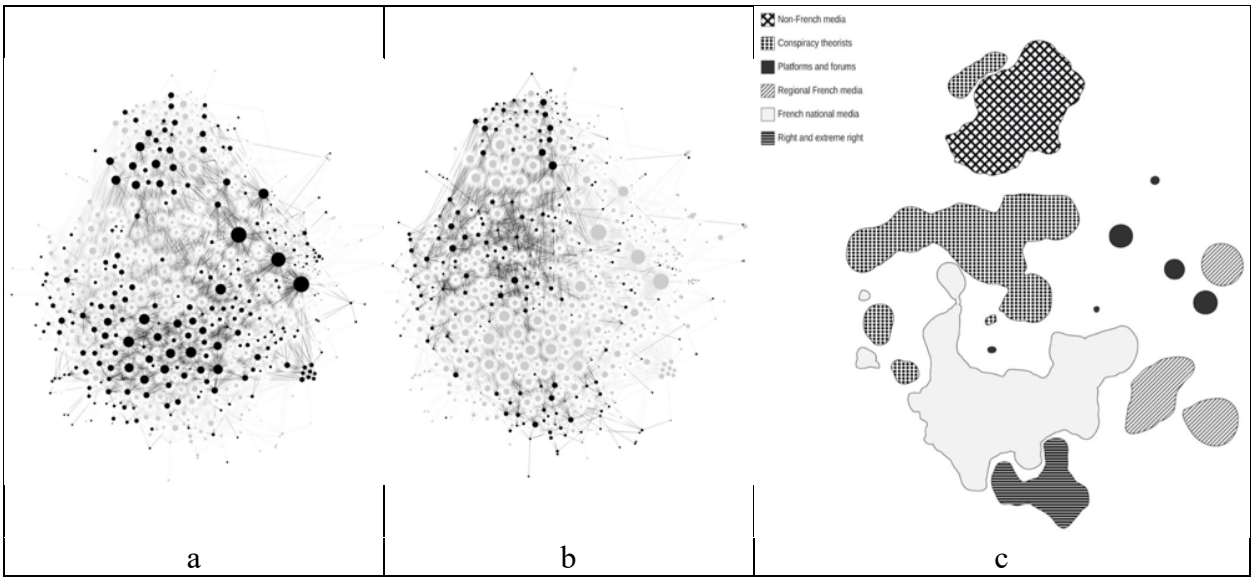
Weekly news sites are less commonly referred to by fake news sources as they lack the novelty and excitement of sites that are more frequently updated.

This statistical method does, however, not yield strong evidence for reliability. The *Area under the Curve* (AUC) value of the Receiver Operating Characteristic (ROC) (i.e. the probability that a randomly chosen reliable site is ranked higher by the model than a randomly chosen less reliable one) is only 0.52 and hardly better than a random selection. We can improve the model by recursively considering the average of the neighbouring nodes' reliability probabilities. The *Le Point* AUC rises to 0.67 but is still not convincing. The simple neighbourhood of a site is a weak indicator of reliability.

Looking at the top 10 most reliable websites in such a statistical neighbourhood ranking demonstrates the scale of the issue. Only 4 out of the top 10 are correctly identified as reliable while the top 10 include some of the clearest cases of unreliable news sites. *Minute* was a journal of the extreme right. *USA News Flash* was a US-based site that distributed conspiracy theories like Pizzagate. Maybe even more interesting are that all the major social media platforms such as Facebook, Twitter but also Wikipedia are at the bottom of the neighbourhood community ranking. This is because they have a lot of unreliable sites linking to them. So, their authority comes as much from unreliable as well as reliable sites.

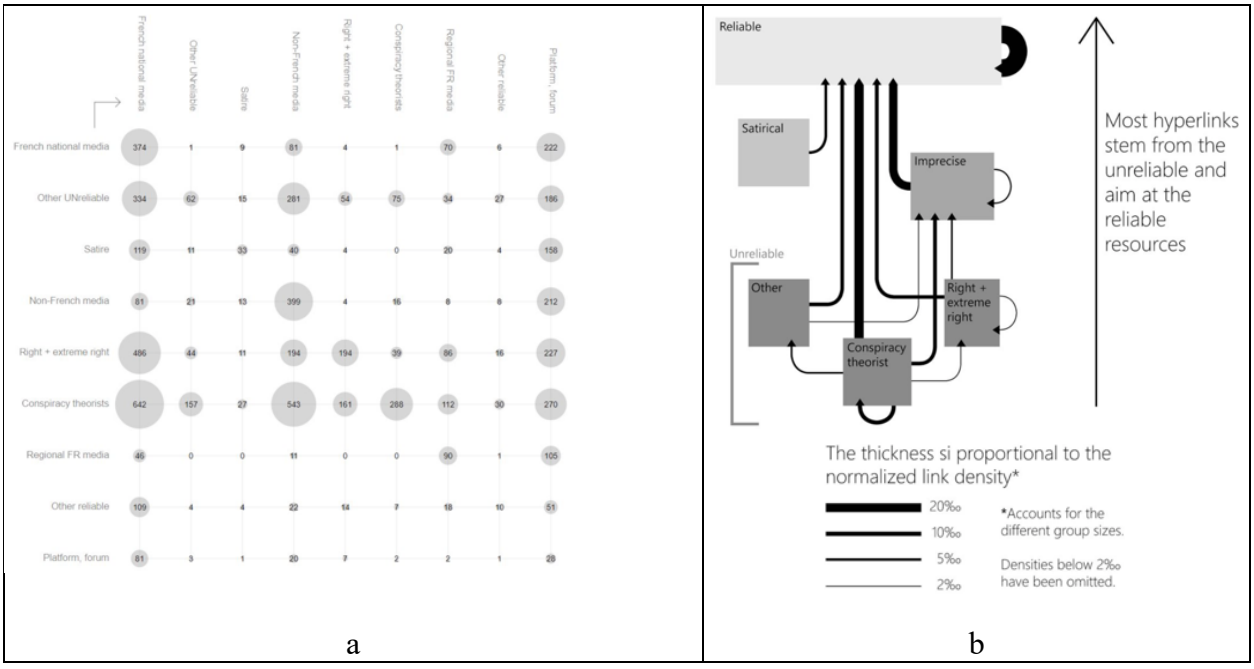
Unreliable websites often link to reliable ones and are therefore in their topological vicinity. We think this is also the reason why we cannot create clearly distinguished zones of reliability, as we visualise the Décodex network. In a previous paper (Venturini, 2018), we analysed the Décodex graph by spatialising it through a force-directed algorithm and submitting the obtained layout to a manual visual analysis (Venturini et al., 2019). The work clearly showed the difficulty of matching the topology of the network with the typology established by the journalist of *Le Monde*. Reliable and unreliable websites appeared as completely mingled in the layout and the disposition of nodes seemed to depend on several other characteristics, such as the geographical span, the language used, the type of website, etc. The best visual partition of the network was thus achieved by including linguistic and political leanings.

Figure 1. The Décodex network with the reliable (a) and unreliable (b) sites highlighted and the diagram of the different regions of the network (c) (extracted from Venturini et al., 2018).



The analysis also revealed that the lack of topological separation between reliable and unreliable websites came from the fact that, while sites in each category tend to cite 'symmetrically' and 'communally' by linking across sites in the same category, sources with lower reliability also link 'asymmetrically' to more reliable sources (but not the other way around). That is why at the bottom of the neighbourhood ranking we find sites which are linked to by reliable and unreliable sites. The visual analysis approach could not easily bring together these two perspectives.

Figure 2. Matrix of citation between the different categories of websites in the Décodex network (a) and diagram of the asymmetrical citation pattern (extracted from Venturini et al., 2018).



Instead of using different techniques for symmetrical relations (force-directed layout) and asymmetry in networks (matrix and diagram), this paper presents an alternative method to consider both in a unified way and cover at the same time communal and preferential attachment in the Décodex network. In this paper, we offer a new technique we developed using predictive analytics to understand how attachments can be used to explain a network view on the reliability of news sites.

3. Methodology

In a previous paper (Blanke, 2018), we introduced a method we called 'Predicting the Past', which allows us to use predictive analytics to analyse heterogenous social spaces. We demonstrated that by using techniques derived from machine learning and originally developed to predict future events we can let machines read complex social relations better than with existing baseline methods. Here we employ a similar approach to analyse relations that make a news site reliable. This method allows us to test relations from statistical macro perspectives to individual micro relations and to work across a large dimensional space of different features. It can integrate heterogenous features from network relations and can combine both hierarchical and community attachments.

The Décodex data contains 305 reliable news sources, 89 imprecise sources, 197 unreliable ones and 76 satirical ones. This means that 305 reliable sites stood against 362 less reliable ones in three categories. Our prediction target will be the reliable sites, which we will compare against all other degrees of unreliability. Our aim is to machine-read a network's understanding of reliability and makes it different from all the other sites. We do not want to develop a model for predicting the websites reliability, but to repurpose the machine learning techniques used for this task in order to derive the hyperlink network view on what constitutes reliability. To analyse news sites, we used the Décodex dataset to train a machine learning model so that it can identify reliability drawing on the structure of the underlying network. Finally, we will use machine learning to meta-evaluate the work of human news site assessors.

Our methodology thus follows the following steps:

1. In order to investigate how a news network's topology influences reliability, we first engineer a number of features that reflect our assumptions on what defines a reliable site according to the hyperlink network. We will present simple features such as the number of in-coming links from reliable sites as well as more advanced features such as modularity and centrality. Recent advancement in graph-based machine learning (Narayanan et al., 2017) engineer their own features, that are, however, not readable for humans. Compared to this work, our dedicated, theory-driven feature engineering will allow us to interpret ourselves the network view on reliability.
2. The second step in machine learning is to choose the right model. Here, we reason that potential models should optimise the machine reading of reliability and consider network linking structures. We first present a baseline advanced ensemble model that optimises performance and prepares the error analysis in a third step. We finally present two simpler but easier to interpret models that reflect the importance of asymmetric relations to determine reliability.
3. The third step is the meta-evaluation of the *Le Monde* evaluators. We employ in particular an error analysis to understand borderline decisions and compare these with human evaluators. Having reduced machine learning errors as far as possible, we manually examine the remaining misclassifications of the algorithm and describe trends in these errors based on a detailed error analysis following Ng (2016).

Features

The first step in our methodology is the featurisation of the dataset, that is the selection and preparation of the network features that will be used as input of the machine learning models and allow us to interpret the machine view on reliability. For our experiments, we introduce several new measures that consider network attachments by including aspects of community formation and preferential relations. To take into account the direction of links, which we have earlier established to be important for reading reliability, we include several edge-based features. This way, we also make use of the best and largest data we have. We have over 14,000 edges compared to 565 nodes.

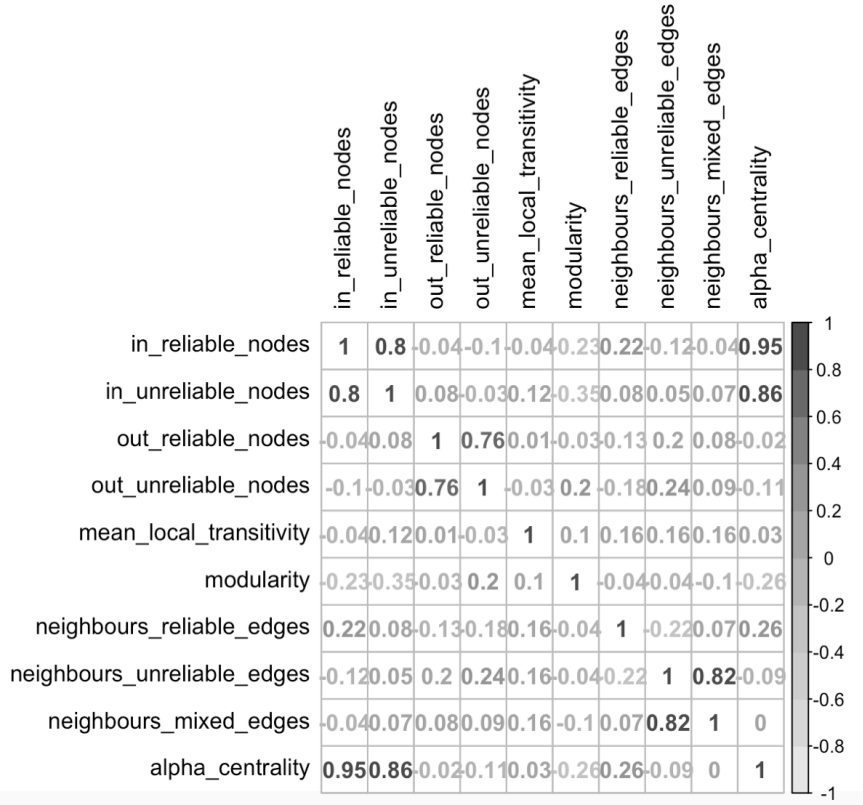
Regarding the features, we concentrate not on the content of the websites (which is most generally used to assess their trustworthiness but also criticised as highly subjective) but exclusively on features derived from their hyperlink profiles.

- We include four direct neighbourhood relations: the number of incoming vs outgoing links from reliable sites vs those from less reliable ones. This feature should detect local symmetries but, as we are following directed hyperlinks, preferential relations are also represented. We expect, for instance, that a reliable site has more links to other reliable sites but is at the same time preferentially linked to by unreliable sites. The features are called: "in_reliable_nodes", "in_unreliable_nodes", "out_reliable_nodes" and "out_unreliable_nodes".
- To move beyond the immediate neighbourhood of a site, we consider the average number of reliable, unreliable and mixed edges of a site's neighbours. Reliable edges connect reliable nodes, unreliable edges connect unreliable nodes and mixed edges are found between reliable and unreliable nodes. This feature follows a similar reasoning as the first one but moves to the neighbourhood. A reliable site's neighbourhood should have a majority of reliable and mixed edges. The features are called "neighbours_reliable_edges", "neighbours_unreliable_edges" and "neighbours_mixed_edges".
- As the only direct centrality measure we consider alpha centrality (called "alpha_centrality"), which is 'an Eigenvector-like measure of centrality for asymmetric relations' (Bonacich and Lloyd, 2001). Alpha centrality is a generalisation of eigenvector centrality for a directed graph with the addition that nodes are imbued with importance from external sources. We are interested in alpha centrality as it was shown to be effective in detecting locally connected and globally interconnected nodes (E. Montijano et al., 2018). Reliable sites should be described by a higher alpha centrality given that they connect both locally and globally. If we rank the data according to alpha centrality, *DailyMail* is at rank 28 the highest ranked site that is considered to be unreliable by the Décodex team. The second highest ranked is *Fox News* at rank 54. We will meet both frequently throughout this article as sites where the machine network view contradicts the human evaluators.
- The modularity is the value of the cut that separates a site and its neighbours from the rest of the network ("modularity"). Modularity should detect a network part that is coined by many local relations (Newman, 2001) and use this to determine reliability. We find the unreliable *Russia Today* at rank 5 and then the *Daily Mail* at rank 14.

- The clustering coefficient of the neighbours is defined by the average/mean local transitivity ("mean_local_transitivity"). It describes the probability that two randomly selected neighbours of a site are neighbours of each other. It is the fraction of pairs of the node's neighbours that are connected to each other. Reliable sites should therefore have a lower clustering coefficient, as they link across neighbourhoods. According to the clustering coefficient, the *Daily Mail* is the highest ranked less reliable site again, this time at rank 7.

The features are designed to test how a machine would establish a hyperlink network view on what constitutes a reliable news site. Part of the machine learning models evaluation will be to find out which of the features provide the best support to describe reliability. As Figure 3 shows, the incoming links from reliable sites correlates strongly (>0.9) with other features. So, we take that feature out.

Figure 3. Correlations between different features.



Modelling

In our experiment, we apply all the steps of traditional predictive analytics to create a stable machine-learning model of the data and avoid overfitting (Blanke, 2018). To avoid overfitting the existing data (that is constructing a model that is excessively influenced by the distribution of the training dataset), we use cross-validation. Here, the data is divided into k subsets or 'folds'. Over a pre-defined number of iterations, one of the folds is held back as test data, while all other

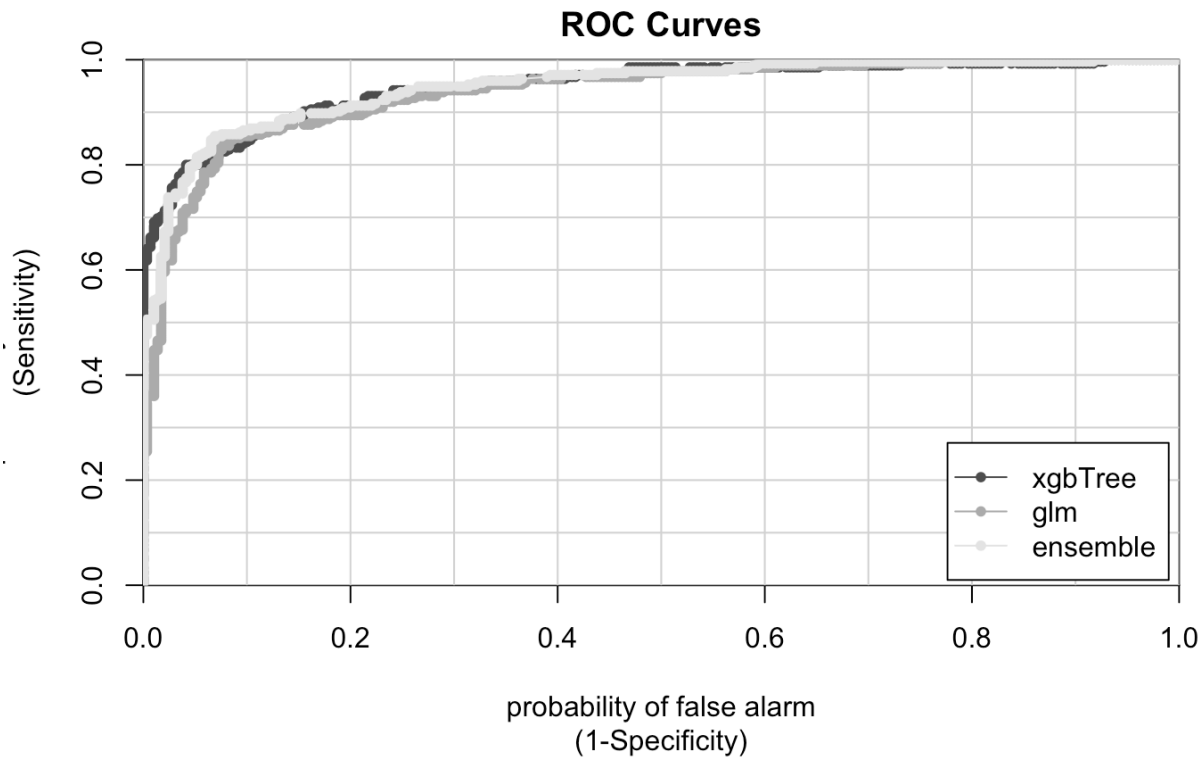
folds are used for training. The test data is then used to perform an independent estimation of the accuracy of the model trained on the rest of the data. Every data item is part of the testing fold exactly once and of the training set in all other the iterations. This reduces underfitting as more data is used for training but also overfitting as all data is at some point used in testing. We hold back 20% of the overall data for an out-of-sample final test to see whether the model is also stable towards completely unknown data.

In this paper, we test several prediction models to identify the ones most suited to describe reliability. We tested generalized linear models (GLM; logistic regression), randomForest (rF) and xgbTree. GLMs are known as more flexible generalizations of ordinary linear regression, Random Forest are ensembles of the well-known decision trees and are often used in prediction experiments, but xgbTree models might be less known. XGB stands for 'eXtreme Gradient Boosting' (Chen et al., 2015) and is also an ensemble version of the decision trees method. It 'boosts' decision tree modelling by running trees in sequence so that each one can learn from the errors of the earlier trees until no further improvement is possible. The two tree-based models are typical rule-based learners that reflect asymmetrical relations while the logistic regression model is a commonly used model to detect symmetric communities based on class descriptions.

Tree models as well as other rule-based learners we introduce in Section 4 represent knowledge as a set of rules or logical if-else statements; described by the antecedent and the consequence (Lantz, 2013, p. 142). So, a simple rule in our case would be "If there are more than 500 links to unreliable sites then this site is also unreliable". For machine learning, this means that the if-statement consists of a logical combination of features (predictors), while the result is a decision on reliability (target). Regression and rule-based classification are among the best performing models for high-dimensional data, which is why we have chosen them for this example (Caruana et al., 2008). Recently, neural networks have become an attractive alternative for high-dimensional data, as a more recent evaluation by Zekić-Sušac et al (2014) shows. In network terms, Narayan et al. (2017) have introduced a novel approach to exploit neural networks based on local walks across the graph. However, these approaches generally learn their own features of the data and thus do not allow us to test our own assumptions about the network view on reliability, for which we rely on our own features.

All these models have a vastly improved ROC/AUC performance compared to the statistical neighbourhood comparison of attachment. However, the random forest model is performing well below the other two models with an average AUC of ~0.84 compared to GLM and xgbTree both performing on average above AUC 0.9. GLM and xgbTree are therefore already showing excellent performance, but we can do even better by stacking the models. The results of the two best performing models XGB and GLM are not correlated (~0.4 on average in our tests). This means we can build a 'meta-model' from them to ensemble a collection of predictive models using the combined strengths of each contributing model. The meta-model will be automatically learned and combine the decisions of the two underlying models by stacking them on top of each other. The idea is that some models are better than others for particular data patterns and that the stacking of them will deliver the best of all worlds. Such a meta-model should in our case integrate a symmetrical view as well as the asymmetrical view.

Figure 4. ROC/AUC performance of the stacked meta-model



The best performing combined model outperforms the two individual models with an AUC of about 0.94-0.95. In the out-of-sample test, we achieved an AUC value of ~0.94, which is close to the cross-validation testing. The model is stable. A pairwise bootstrap test of the ROC curves reveals that the ensemble significantly improves the GLM model but not XGB. Figure 4 shows how the ensemble model smooths out the differences between XGB and GLM but all models perform well with XGB rising more quickly than GLM.

The tests demonstrate how we can generate a highly accurate network view of reliable sites. We apply the model to the whole network with a high accuracy of ~89.4% with the following confusion matrix:

Table 1: Confusion Matrix of Meta-Model

	judged as “reliable”	judged as “unreliable”
predicted as “reliable”	270	20
predicted as “unreliable”	40	235

The accuracy is more impressive if we consider the error analysis, which will later show several systematic issues with the dataset.

XGB contributes to about 43% to the decision, while GLM takes care of the remaining 57% if we linearly combine the two models. Figure 5 demonstrates which of the features are considered most relevant by either GLM or XGB. The most dominant features are how much the neighbours were connected to reliable edges, whether a node link to unreliable nodes, as well as how many mixed edges the neighbours connect with. We can furthermore split the feature importance calculation into the two underlying models in Figure 5 to see how they work in conjunction and which features are preferred by which model. Please, note that this assessment is based on considering the two models as linear combinations.

Figure 5. Feature Importance

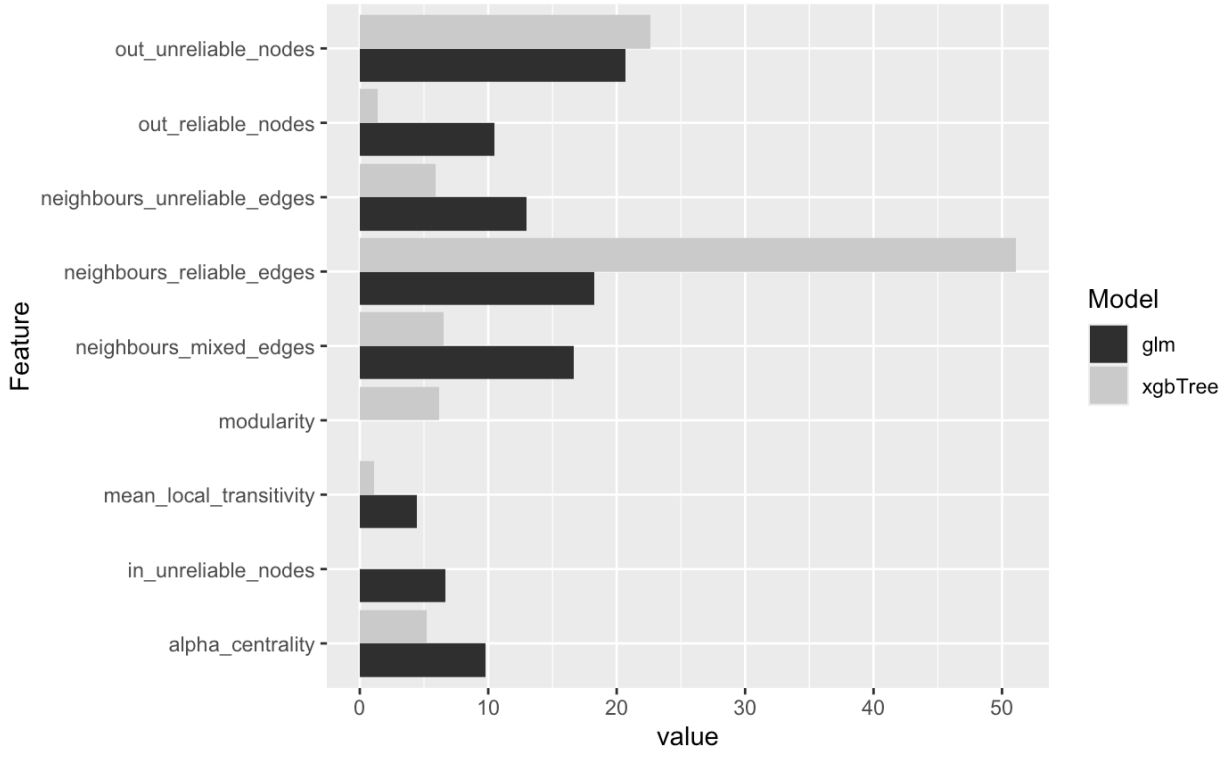


Figure 5 clearly shows that for GLM, there are no particular stand-out features, even though this model seems to work slightly better with the simpler ones. XGB as an advanced model is better at working out which ones are the most important features: “out_unreliable_nodes” and ”neighbours_reliable_edges”. ”out_reliable_nodes” stands out from all the other direct neighbourhood features, while ”neighbours_reliable_edges” dominates the edge-based features. Overall the simpler features such as the reliable edges of the neighbours are preferred, which we explain by the fact that the network is still relatively small. The more complex calculations such as alpha centrality or modularity seem to play no strong role. Together with the connections to unreliable nodes, the neighbours’ reliable edges make up more than 70% of the decision and the two strongest clusters of features. We are especially encouraged by the fact that XGB is able to clearly identify neighbours’ reliable edges as the strongest indicator of reliability. This follows our intuition that incoming hyperlinks to reliable sites are not by themselves a good description of reliability but only if they come from a cluster of reliable websites.

In order to test the importance of the highest ranked features, we remove most features and investigate whether a subset of the top four features is enough to produce an accurate model for predicting the reliability of websites. According to Figure 5, these top four features are: "neighbours_unreliable_edges", "neighbours_mixed_edges", "out_unreliable_nodes" and "neighbours_reliable_edges". With these features only, the number of false negatives slightly increases but the accuracy remains more or less the same at ~88%. To determine how the network reads reliability, the out-going links to reliable sites and the reliability of the neighbourhood are by far the most important features. In Section 4, we will investigate these features further by introducing new machine-reading models that focus on them.

Machine reasoning trends

As described above, a detailed error analysis will help us discover trends in the machine reasoning. Because we are interested in evaluating the human understanding of reliability, errors can be as relevant as correct predictions, as they indicate interesting borderline cases as well as other indicators of challenges in human reasoning about reliability.

Looking at the ensemble model's misclassification, we notice first some interesting assumptions in the judgements of the assessors which stand against a network view of reliability. The model misjudges as reliable several English right-wing news sites such as the *Daily Mail* and *Fox News* as well as French right-wing sites, such as *Valeurs actuelles* and *Contrepoints*. They were all judged to be unreliable by the Décodex journalists. However, according to the confusion matrix in Table 1 the biggest problem for our model seem to be to identify all reliable news sites, as they seem to include several sites that are highly unusual. *Agoravox* and *Mes Opinions*, for instance, were judged as reliable by the Décodex judges but not by the model. *Agoravox* is a citizen journalism website (with some concern about its spread of conspiracy theories), while *Mes Opinions* is an online campaigning site.

At first sight, these errors therefore seem to point to unusual patterns in the dataset based on assumptions of the judges and a favouring of community sites that the model struggles to consolidate with a network view. Overall, we can therefore consider our approach to be better than the accuracy numbers might suggest given such data inconsistencies. The *Daily Mail*, for instance, will be extensively connected to reliable news sites, as it is one of the biggest daily newspapers in the UK but does have a reputation for questionable reporting. The error analysis shows clearly how sensitive the decision on reliability is to the direction of hyperlinks. If the neighbours' reliable hyperlinks are the single most important feature, then the machine-reading must struggle for sites such as *Daily Mail* where these network features contradict the human assessors.

As the ranking of features has shown the reliability of a site's neighbourhood to be the most important feature, we will continue with two rule-based models that allow us to investigate this relationship further, as they are focussed on the most important asymmetric relations in a dataset. This should also support the overall interpretability of our results. Ensembles of machine learning models are generally not easy to follow by humans.

4. Ruling in asymmetries

Let us investigate whether we can reproduce the importance of the edge directedness with two other rule-learner algorithms targeting these specific features. These can be seen as surrogate 'white-box' models that allow us to shed light on the functioning of our ensemble meta-model and baseline of what can be achieved. In this section, we especially target the most important asymmetric reasoning. Otherwise, we follow the same methodology by presenting first the model results and then an error analysis to investigate trends in the reasoning about reliability by the rule-based learners.

The first rule learner algorithm we employ is *OneRule* (Holte, 1993). As the name says, it concentrates on one rule per decision. It simply selects the one rule that most accurately describes the decision based on the fewest prediction errors produced. OneRule has thus shown to be very easy to interpret for humans but to be still very powerful. It identifies the reliable edges of neighbouring nodes as the single most important consideration, as we had anticipated. Using neighbours_reliable_edges to split the network into reliable and less reliable sites, we can find the following rules:

1. (neighbours_reliable_edges < 0.091) $\Rightarrow$ UNRELIABLE
2. (neighbours_reliable_edges $>= 0.091$) AND (neighbours_reliable_edges < 1.442) $\Rightarrow$ RELIABLE
3. (neighbours_reliable_edges $>= 1.442$) AND (neighbours_reliable_edges < 2.178) $\Rightarrow$ UNRELIABLE
4. (neighbours_reliable_edges $>= 2.178$) $\Rightarrow$ RELIABLE
   (471/565 instances correct)

The output clearly indicates that a lower average of reliable edges for neighbouring sites means that the node is likely to be unreliable, while a larger number indicates a higher reliability. Rule 1 states that sites are less reliable if they have neighbours with on average less than 0.091 reliable links. The rule more or less indicates that if the neighbours have no reliable links, the site is also not reliable. The second and third rule point to various more complex cases that confuse the model as it uses only "neighbours_reliable_edges". Rule 4 implies that if a site's neighbours have on average more than 2.178 reliable edges the site is reliable.

Table 2: Confusion Matrix of OneRule

	judged as "reliable"	judged as "unreliable"
predicted as "reliable"	214	33
predicted as "unreliable"	61	257

OneRule identifies 471 of the 565 sites correctly using exclusively the reliable neighbouring edges. That is about 84% of the corpus and demonstrates the power of this simple asymmetric attachment measure considering edges. Overall, the increases in misjudgements compared to the

ensemble model is fairly evenly distributed between false positives and false negatives in Table 2.

The list of misjudgements indicates similar issues as we have discovered before. The problems with sites that have reliable neighbourhoods reappear. *DailyMail*, *Fox News* are now joined by *Russia Today* and *Sputniknews FR*, both of which are accused of pushing Russian state propaganda. They are listed as their asymmetric attachments (neighbours' reliable sources) is high considering that they are part of the broader media ecosystem. The value for the *DailyMail's* neighbours' reliable edges is 6.2 while for *Fox News* it is 3.14, which puts both in rule no 4 of reliable sites with a value larger than ~2.17. As a comparison, the most recognised French reliable news sources from the left and right, *Le Monde* and *Le Figaro*, have both larger values for reliable edges – as we should expect – but are still in the same rule as *DailyMail* and *FoxNews*. *Russia Today* is at the borderline of this classification with 2.44 for neighbours' reliable edges, which also reflects their status as a site whose reliability is generally considered to be in doubt but some of their reporting is also cited by more reliable sites.

At the other extreme are examples of sites with values of less than 0.09 for neighbours' reliable edges. On top this includes the well-known *Breitbard* as well as *ZeroHedge*, a financial Blog accused of distributing right-wing conspiracy theories. For the extremes, the algorithm seems to do a good job using network attachments to distinguish sites.

Between the maximum and minimum values there is more heterogeneity in the decision according to the second and third rule. Here, we also find most of the misjudgements. Rule 2, e.g., has 46% misjudgements. The change of judgements in-between the extreme cases in rules 2 and 3 tells us more about the broader issues of predicting reliability from network attachments but also the biases and issues typical to a human-created dataset such as Décodex. In the group of rule 2, we discover the *New York Daily News*, a US-based newspaper with several Pulitzer prices but also sometimes controversial news stories that the *New York Times* has called 'populist',[3] with a neighbours' reliable nodes value of 1.04. We also find another regional newspaper from France *Le Maine Libre*, which is seen by Décodex as reliable and which has for neighbours' reliable nodes a value of 1.29. Otherwise, we discover the already described citizen news sites *Agoravox* and *Mes Opinions* in this group, as well as gamer sites such as *Jeuxvideo*. For regional sites, we can assume that they are less connected/visible, while we have already spoken about the unusual decision by the Décodex judges to include citizen news sites. These control less how their 'citizens' link out to reliable or unreliable sites. Among the correct classifications in rule 2 is *Génération Identitaire*, a hard-right unreliable political site. Its value is very close to the next group of unreliable sites with a value of 1.43 for neighbours' reliable nodes. The algorithm does a better job here than the human assessors, which have misqualified the site as reliable.

The third rule includes the *New York Post*, a tabloid competitor to the *New York Daily News*. The *Post* is seen as unreliable, with a value of 2.04 for the neighbours' reliable connections. Also, correctly identified is *Novopress*, another site from the same background as *Génération Identitaire*. A wrong classification is the reliable local newspaper *Herault Tribune*, whose value is just inside the wrong classification boundary with 1.5. A soft boundary methodology (rather

[3] https://www.nytimes.com/2016/01/30/business/media/drop-dead-not-the-newly-relevant-daily-news.html

than the hard one employed by OneRule), might be better to consider these kinds of misqualifications.

An expansion to OneRule is JRip, based on the Ripper algorithm (Cohen, 1995). JRip tries all the possible values of all predictors and chooses the ones with the highest gain to generate a multi-condition rule. In the case of the Décodex network, JRip produces four rules with the same top 4 features that we used in our last ensemble experiment:

1. (neighbours_reliable_edges >= 3.558824) and (out_unreliable_nodes <= 0) => RELIABLE (188.0/6.0)
2. (out_unreliable_nodes <= 0) and (out_reliable_nodes >= 4) and (neighbours_mixed_edges >= 32.8) => RELIABLE (23.0/2.0)
3. (out_unreliable_nodes <= 1) and (neighbours_reliable_edges >= 0.181818) => RELIABLE (27.0/7.0)
4. ELSE => UNRELIABLE (327.0/52.0)

The first number in the final brackets is the number of cases covered by the rule and the second number is the number of misclassified cases. The first rule therefore predicts as reliable 188 websites that do not link to less reliable sites and that have some neighbours with reliable relations. This includes 6 incorrect classifications. The second rule uses a combination of 0 links to less reliable sites as well as more than 4 links to reliable sites in combination with the neighbours' mixed edges in order to judge 23 cases as reliable with 2 errors. The third rule continues using the combination of out-links to unreliable nodes and the reliable edges of neighbours to determine 27 reliable nodes with 7 errors. The fourth rule states that in all other cases the nodes are not reliable, but here we also see the largest number of misjudgements. In the following analysis, we would like to concentrate on the misjudgements.

The performance of JRip is very good (with ~88% accuracy) and better than OneRule at identifying unreliable sites through its second condition. Compared to the ensemble model, JRip performs better at avoiding false negatives but significantly worse with false positives. The strength of the baseline ensemble model comes from its dynamic combination of features providing a better balance.

Table 3: Confusion Matrix of JRip

	judged as "reliable"	judged as "unreliable"
predicted as "reliable"	223	15
predicted as "unreliable"	52	275

The false classifications from the first rule include the libertarian *Contrepoints*, a French site, as well as *Investigaction*, a Belgian left-wing site. Both of these are still in the realm of what reliable sites discuss. In particular, they both avoid out-links to any unreliable sites according to the second feature that JRip considers compared to OneRule. So, their classification as unreliable stands out based on the neighbourhoods of their neighbours and points us back to the question

what the judges from Décodex have seen as reliable. If a site links to no unreliable sites, should it really be considered to be unreliable?

The second rule misclassifications are *Demotivateur* and *Brave Patrie. Demotivateur* is an infotainment site, while *Brave Patrie* is a satirical site which presents itself as 'the true journal of the true values of true France'. They can be found in this rule, as they attract a large neighbourhood of mixed edges. They also attract no incoming links from reliable sites while they do have links from less reliable ones. The false classifications from the third rule include the already discussed *Fox News* as well as the *New York Post*. It is also interesting that JRip follows the *Le Monde* evaluators and sees the *Daily Mail* as unreliable. *Fox News* does not cite any unreliable news nodes in this network, while the *New York Post* does but to a very small amount.

The fourth rule's misclassifications include several regional newspapers such as the *Irish Daily Star* or the Austrian *Standard*, which are not well connected to the rest of the network, as well as further community sites such as *Hoaxbuster*, a community platform to limit the circulation of hoaxes. The later strikes us as ironic and also points us to a problem in our original data generation. This site considers a lot of hyperlinks to unreliable sites so that considering them prior to the French elections as a sign of importance of sites seems to run into problems with at least some of sites considered in Décodex. However, just looking at hyperlinking does not actually tell us much what is implied in the relationships. *Hoaxbuster* includes many unreliable sites in order to discuss and most likely criticise them.

Maybe even more strikingly, the fourth rule also misclassifies some of the international news heavy weights such as *International Business Times* or *Politico Europe*. Even the home of Décodex evaluators *Le Monde* belongs to this rule, as it links out to two unreliable sites, although it does have a lot of reliable edges in its neighbourhood.

5. Conclusion

This article has shown how to reason distantly using a network view about a complex social relation, which is the reliability of news. Such a network view promises a topic-agnostic perspective that can be a useful hint on reliability trends and their heterogeneous assumptions. It is based on a more long-term network of hyperlinks that is more difficult to manipulate, as PageRank has already shown, especially as the wider neighbourhood is considered. The article has tried to reconcile the exploration of different attachment structures in online networks through machine learning. We analyse a network of news websites by the journalists of *Le Monde*. In our analysis, we depart from the ever-growing numbers of papers trying to find machine learning algorithms to predict news sites reliability and focus instead on using machine reasoning to understand the structure of news networks by comparing it with our human judgements.

We showed that machine learning can be useful for the reliability-check of the reliability-checkers. Machine learning models can be used to reason about the decisions by human judges. Overall, the machine reading has taught us a lot about the challenges of identifying reliability in a network of news sites. First there is a question of what is included in a corpus of news sites. Our data, for instance, included citizen journalism websites and an online campaigning site. The machine reading struggled to assign these sites to either side of the network, as none of them is in the narrower sense of the word a news site. The problems in the underlying dataset became especially clear when considering *Hoaxbuster*, a community platform to limit the circulation of hoaxes, which was considered to be unreliable by the algorithm as it links to many unreliable sites.

The machine reading tells us how difficult an undertaking the human judgement of reliability of news is. The machine reading struggled to reproduce the evaluators' judgements for sites such as *Fox News* or the *Daily Mail*. Both belong to the official news ecosystems, with reliable sites such as the *BBC*, e.g., often reporting on stories from the *Daily Mail*, but they have a reputation for biased and unreliable reporting. Rather than simply considering these sites as unreliable, the struggle of machine leaning models based exclusively on network information reveals the heterogeneous decision-making of the human assessors, which also took into consideration the contents and the reputation of the sites.

Rather than focussing on fact-checking, we used predictive analytics to analyse existing knowledge about network attachments in a real-life human-created network of news sites. Our approach successfully integrated different network perspectives and is thus a showcase how community and preferential attachment play together. While all models we presented do very well at reading out the network's view on reliability and better so than standard statistical descriptions using a node's neighbours, there are also some limitations of this approach. First and foremost, the network is fairly small and unevenly distributed. It has relatively few nodes, as these were created by human assessors, but lots of edges (14,298) detected by the Hyphe web crawler. This could explain why the edges might play such a big role in the model's judgements. The second limitation is that our analysis shows that we should have extended our modelling using techniques with softer decision boundaries. More work is required here. An interesting follow-on investigation could be to concentrate on only those sites that are at the borderline.

References

Ackland, R., Shorish, J., 2014. Political Homophily on the Web, in: Cantijoch, M., Gibson, R., Ward, S. (Eds.), Analyzing Social Media Data and Web Networks. Palgrave Macmillan UK, London, pp. 25–46. https://doi.org/10.1057/9781137276773_2

Ahmed, H., Traore, I., Saad, S., 2017. Detection of online fake news using N-gram analysis and machine learning techniques. Presented at the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Springer, pp. 127–138.

Albert, R., Jeong, H., Barabási, A.-L., 1999. Diameter of the world-wide web. nature 401, 130–131.

Blanke, Tobias, 2018. Predicting the Past. Digital Humanities Quarterly 12.

Bonacich, P., Lloyd, P., 2001. Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. 23, 191–201. https://doi.org/10.1016/S0378-8733(01)00038-7

Bounegru, L., Gray, J., Venturini, T., Mauri, M., 2017. A Field Guide to Fake News: A Collection of Recipes for Those Who Love to Cook with Digital Methods (Chapters 1-3). Public Data Lab Res. Rep.

Brandtzaeg, P.B., Følstad, A., 2017. Trust and distrust in online fact-checking services. Commun. ACM 60, 65–71.

Caruana, R., Karampatziakis, N., Yessenalina, A., 2008. An empirical evaluation of supervised learning in high dimensions. Presented at the Proceedings of the 25th international conference on Machine learning, ACM, pp. 96–103.

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., Freire, J., 2019. A topic-agnostic approach for identifying fake news pages. Presented at the Companion Proceedings of The 2019 World Wide Web Conference, pp. 975–980.

Centola, D., González-Avella, J.C., Eguíluz, V.M., San Miguel, M., 2007. Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups. J. Confl. Resolut. 51, 905–929. https://doi.org/10.1177/0022002707307632

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., 2015. Xgboost: extreme gradient boosting. R Package Version 04-2 1–4.

Ciampaglia, G.L., 2018. Fighting fake news: a role for computational social science in the fight against digital misinformation. J. Comput. Soc. Sci. 1, 147–153. https://doi.org/10.1007/s42001-017-0005-6

Cohen, W.W., 1995. Fast Effective Rule Induction, in: Prieditis, A., Russell, S. (Eds.), Machine Learning Proceedings 1995. Morgan Kaufmann, San Francisco (CA), pp. 115–123. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Conroy, N.J., Rubin, V.L., Chen, Y., 2015. Automatic deception detection: Methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. 52, 1–4. https://doi.org/10.1002/pra2.2015.145052010082

E. Montijano, G. Oliva, A. Gasparri, 2018. Distributed Estimation of Node Centrality with Application to Agreement Problems in Social Networks, in: 2018 IEEE Conference on Decision and Control (CDC). Presented at the 2018 IEEE Conference on Decision and Control (CDC), pp. 5245–5250. https://doi.org/10.1109/CDC.2018.8619765

Gilda, S., 2017. Evaluating machine learning algorithms for fake news detection. Presented at the 2017 IEEE 15th Student Conference on Research and Development (SCOReD), IEEE, pp. 110–115.
Gray, J., 2018. Tommaso Venturini, Mathieu Jacomy, Liliana Bounegru, and. Routledge Handb. Dev. Digit. Journal. Stud.
Hindman, M., 2008. The myth of digital democracy. Princeton University Press.
Holte, R.C., 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Mach. Learn. 11, 63–90. https://doi.org/10.1023/A:1022631118932
Huibers, T.W.C., 1996. An axiomatic theory for information retrieval.
Jacomy, M., Girard, P., Ooghe-Tabanou, B., Venturini, T., 2016. Hyphe, a curation-oriented approach to web crawling for the social sciences. Presented at the Tenth International AAAI Conference on Web and Social Media.
Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G., 2015. On close and distant reading in digital humanities: A survey and future challenges. Proc EuroVis—STARs 83–103.
Keuschnigg, M., Lovsjö, N., Hedström, P., 2018. Analytical sociology and computational social science. J. Comput. Soc. Sci. 1, 3–14. https://doi.org/10.1007/s42001-017-0006-5
Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y., 2013. Prominent features of rumor propagation in online social media. Presented at the 2013 IEEE 13th International Conference on Data Mining, IEEE, pp. 1103–1108.
Lantz, B., 2013. Machine learning with R. Packt Publishing Ltd.
Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L., 2018. The science of fake news. Science 359, 1094. https://doi.org/10.1126/science.aao2998
Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W., 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Math. 6, 29–123.
Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M., 2019. Fake news detection on social media using geometric deep learning. ArXiv Prepr. ArXiv190206673.
Naeem, B., Khan, A., Beg, M.O., Mujtaba, H., 2020. A deep learning framework for clickbait detection on social area network using natural language cues. J. Comput. Soc. Sci. 1–13.
Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., Jaiswal, S., 2017. graph2vec: Learning distributed representations of graphs. ArXiv Prepr. ArXiv170705005.
Nelson, J.L., Taneja, H., 2018. The small, disloyal fake news audience: The role of audience availability in fake news consumption. New Media Soc. 20, 3720–3737.
Newman, M.E.J., 2001. Clustering and preferential attachment in growing networks. Phys. Rev. E 64, 025102. https://doi.org/10.1103/PhysRevE.64.025102
Ng, A., 2016. Nuts and bolts of building AI applications using Deep Learning. NIPS Keynote Talk.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
Pariser, E., 2011. The filter bubble: What the Internet is hiding from you. Penguin, New York, NY.

Pennycook, G., Rand, D.G., 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc. Natl. Acad. Sci. 116, 2521. https://doi.org/10.1073/pnas.1806781116

Ravandi, B., Mili, F., 2019. Coherence and polarization in complex networks. J. Comput. Soc. Sci. 2, 133–150. https://doi.org/10.1007/s42001-019-00036-w

Sunstein, C.R., 2001. Republic. com. Princeton university press.

Venturini, T., Jacomy, M., Jensen, P., 2019. What do we See when We Look at Networks. An introduction to visual network analysis and force-directed layouts. Introd. Vis. Netw. Anal. Force-Dir. Layouts April 26 2019.

Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359, 1146. https://doi.org/10.1126/science.aap9559

Wallach, H., 2018. Computational social science≠ computer science+ social data. Commun. ACM 61, 42–44.

Wang, W.Y., 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. ArXiv Prepr. ArXiv170500648.

Zekić-Sušac, M., Pfeifer, S., Šarlija, N., 2014. A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem. Bus. Syst. Res. J. 5, 82. https://doi.org/10.2478/bsrj-2014-0021