Journal of Computational Social Science
https://doi.org/10.1007/s42001-021-00116-w

RESEARCH ARTICLE

A network view on reliability: using machine learning to understand how we assess news websites

Tobias Blanke[1] · Tommaso Venturini[2]

Received: 25 August 2020 / Accepted: 17 March 2021

Abstract
This article shows how a machine can employ a network view to reason about complex social relations of news reliability. 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 the reliability of news 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. We present a machine learning approach to analyse what constitutes reliable news from the view of a network. Our aim is to machine-read a network's understanding of news reliability. To analyse real-life news sites, we used the Décodex dataset to train machine learning models from the structure of the underlying network. We then employ the models to draw conclusions how the Décodex evaluators came to assess the reliability of news.

Keyword Network Analysis · Distant Reading · News ecosystem · Fake News

Introduction

'Fake news' [5] has become the latest form of what McLuhan has called media moral panic [25]. Almost daily, a new alert is raised about its unstoppable spread across the Internet and social media. In 2016, nationalreport.net, a site well-known

✉ Tobias Blanke
t.blanke@uva.nl

Tommaso Venturini
tommaso.venturini@cnrs.fr

[1] University of Amsterdam, Amsterdam, The Netherlands

[2] Center for Internet and Society, CNRS, Paris, France

Published online: 06 May 2021

for spreading unreliable information, claimed that customers in Colorado were using food stamps to buy marijuana. While this claim had no basis, it spread so widely that Colorado House Representatives proposed legislation to stop these non-existent purchases. With fake news concerns, fact-checking sites have also mushroomed to distinguish between reliable and less reliable news sites. The Reporters Lab at Duke University lists almost 300 active sites from all over the world.[1] Fact checkers are, however, confronted by the fact that on the Internet the line between journalism and other content is blurred and that the efficacy and usefulness of fact-checking is frequently questioned [6]. Pennycook and Rand [34], for instance, conducted an online experiment using crowdsourcing to find out whether lay people could also provide fact-checking. It suggested that 'politically balanced layperson ratings were strongly correlated with ratings provided by professional fact-checkers' [34]: 2521).

Next to such human evaluation approaches, fake news has also been targeted with machine learning. There are many challenges to employing machine learning to detect fake news. Castelo et al. [8], for instance, discuss issues that stem from the dynamic nature of online news where 'correct' facts quickly become outdated, as new political developments lead to new online discourses. Classifiers will thus age fast. Castelo et al. [8] 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 [27], deep learning is employed to detect 'certain natural language cues' to find patterns of fake news in clickbait. Monti et al. [26] 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. [22], we rely on features derived from graph theory such as centrality. Where Ravandi and Mili [35] demonstrate how graph analysis can be used to analyse polarisation in simulated news networks, we are interested in how graph-based machine learning can support an analysis of an actual news ecosystem for the reliability of its sites.

Lazer et al. [23] provide an overview of the scientific challenges related to computationally defining the reliability of news and claim this operation requires broad interdisciplinary collaboration. Similarly, Ciampaglia [10] argues for an increased role of computational social scientists in the fight against fake news. Social and computer sciences must work together to identify generalizable mechanisms capable of operating in 'large-scale interactive systems' [21]. In this paper, we follow such demands for interdisciplinary research on fake news and claim that machine learning can be used not only to filter fake news but also and more productively to understand the relationships between fake and traditional news sites. Rather than trying to detect whether a fact, a text or a whole site belong to fake news, this paper employs machine learning to explore how a computer would understand *our* understanding of fake news. Based on previous work, we carefully selected a number of network features from interlinked news sites such as neighbourhoods and centrality to establish how a machine would read a news ecosystem. We discover the heterogeneous

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

challenges involved in consistently establishing what constitutes valid vs. fake news, which raises the question who the fact checkers are and who checks on them. This paper, therefore, addresses social questions about the heterogeneity of human decisions on the reliability of news sites rather than the previously cited computing sciences work that uses machine learning to detect fake news.

A view from the network on online reliability

Using the structure of digital networks to categorise websites is far from easy because this structure is shaped by two opposite *attachment dynamics*. The first dynamics, *communal attachment*, consists of the fact that websites, social media accounts, etc. tend to connect to other sites that focus on the same topics, issues or matter of interests [1, 9]. 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*, describes the fact that websites that are already highly cited have a higher probability to attract new hyperlinks [3]. These two dynamics are equally important but also diametrically opposed [24, 30, 39]. While communal attachment encourages homophily and tends to generate thematic communities where shared interests are discussed by like-minded actors, preferential attachment encourages hierarchy and tends to create a pyramid of attention concentrated around a few hyper-visible nodes. Communal attachment goes in circles (within the same community) and creates clustering, pr eferential attachment goes upward (toward the most visible) and creates ranking.

The difficulty to consider these two dynamics together and combine their opposed effects has produced a twofold reading of networks and two opposite concerns about online news misinformation. On the one hand, commentators have denounced the emergence of increasingly tight news ’filter bubbles’ [33, 36] 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 [16, 29].

Questions of communal and preferential attachment also arose in an analysis we carried out exploring a series of news sources selected and categorised by *Le Monde* according to their reliability, as well as their network of hyperlinks [37]. This analysis provides the data for our machine-learning approach to detect reliable news sites according to *Le Monde*. To account for the combination of the two types of attachment described above, we use several 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 detect the reliability of news sources, as has been done in many other studies [13], 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 [40] sees as the fundamental methodological difference between interpretation-oriented social scientists and computer scientists, who are mostly concerned with obtaining models that produce great accuracy but disregard the reasons for doing so. We work with a highly curated dataset rather than concentrate on scale and accuracy.

Working with a highly curated dataset, we employ machine learning as a tool to gain new insights into the attachment relationships of our news networks and how they determine reliability. This approach could be called ‘distant reading of networks’. The method is inspired by the better-known ‘distant reading’ methods [20], used to complement the human reading of texts with machine reasoning. The method is called ‘distant’ because it relies on machines rather than humans to explore relationships.

The data: the Décodex network

The intertwining of communal and preferential attachment, which characterises digital networks, is easily observable in the dataset of news sources that we focus on in this article. We built a network drawing on a list of online news sources catalogued by the ‘Décodex’ initiative.[2] In this fact-checking project, 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’. A few months ahead of the French elections, we extracted the URLs reviewed by ‘Décodex’ and used the Web Crawler Hyphe [19] to map the hyperlinks connecting them. All the websites from the Décodex corpus have been crawled at a depth of one-click starting from their homepage. This means that we have considered all the hyperlinks present on the homepage and in the pages directly accessible from the homepage.

Our work falls into the large body of examinations that try to detect the predictability of news sources [2, 15, 41] through the topological analysis of the hyperlink network connecting them. Huibers [18] has demonstrated how this approach has been successfully employed for over twenty years in the retrieval of hypermedia contents. Using the information about hyperlinks and their clusters is a well-known solution to rank search results from the world wide web, at least since Google introduced its PageRank algorithm [32]. PageRank uses the typological neighbourhood of a site to reinforce content-based rankings. The link authority of a page is employed to 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 network structure in a very successful hypermedia retrieval model. Relying on the topology to match typology, is, however, far from straightforward especially in decisions on reliability. To demonstrate this difficulty in the Décodex network, let us consider its hyperlink relationships based on simple link statistics. We can, for instance, compute a naïve probability of reliability for each site as the simple ratio between the reliable news sites directly connected to it and the total number of its neighbours (its

[2] www.lemonde.fr/verification

degree). This would be a reasonable description of a community approach where reliable sites cluster together and follows existing work like [13] on news sites, where a 'trust dimension' is inferred from the links in the network. According to this naïve neighbour classifier, *Le Point* (a centre-right conservative weekly) achieves top scores, as it is the least connected to unreliable ones. Discovering it as a reliable news source agrees with research such as [39], which considers the speed of updates as a distinct feature of unreliable news sites. 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. In this paper, to assess the capacity of different techniques to identify reliable news sites we will generally observe the *Area under the Curve* (AUC) measure 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. AUC is a better measure than simple accuracy in our case, as accuracy only looks at fractions of correct classifications. The AUC value for the statistical methods described above 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 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 according to the statistical neighbourhood ranking demonstrates the problem. Only 4 out of the top 10 are also categorized as reliable by *Le Monde*, while the list includes 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. Even more tellingly, all major social media platforms (such as Facebook, Twitter but also Wikipedia) are at the bottom of this simple neighbourhood ranking, receiving links from a large number of unreliable sites.

In fact, unreliable websites often link to reliable ones and are therefore in their topological vicinity [38]. This is the reason why we cannot create clearly distinguished zones of reliability in the Décodex network. In a previous paper [37], we analysed the Décodex graph through a force-directed layout and later a visual network analysis [38]. The work clearly showed the difficulty of matching the visual topology of the network with the typology established by *Le Monde*. Reliable and unreliable websites appeared as completely mingled in the layout, and the disposition of nodes seemed to depend instead on other characteristics, such as the geographical span, the language, the type of website, etc. The best visual partition of the network was achieved by a combination of media types, language used and political leanings according to Fig. 1.

The analysis 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 sites in the same

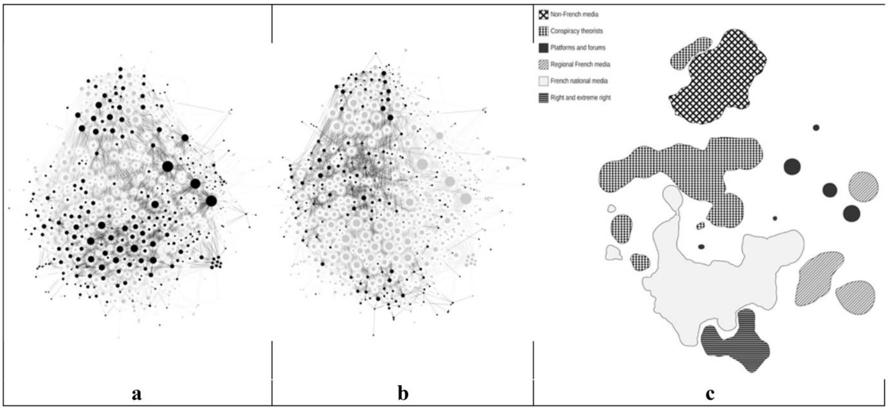


Fig. 1 The Décodex network with the reliable (**a**) and unreliable (**b**) sites highlighted and the diagram of the different regions of the network (**c**) (taken from Venturini et al. [37])

category, sources with lower reliability also link ’hierarchically’ and ‘asymmetrically’ pointing to more reliable source without being linked to in return.

Methodology

In a previous paper [4], we introduced a method we called ‘Predicting the Past’, which allows us to use predictive analytics to analyse heterogeneous social spaces. We demonstrated that techniques derived from machine learning and originally developed to predict future events can be used for the distant reading of social relations. Here, we employ a similar approach to analyse relations among news websites. This method allows us to test network features ranging from macro-statistical perspectives to individual micro relations and to work across a large dimensional space of different features. It can integrate heterogeneous features from network relations and combine hierarchical and community attachments.

The Décodex data for our experiment contains 305 reliable news sources, 89 imprecise sources, 197 unreliable ones and 76 satirical ones. This means that 305 reliable sites stand vs. 362 less reliable ones in three categories.[3] Our prediction targets reliable sites, which we compare against all other degrees of unreliability. Our aim is to investigate the way in which our machine learning model comes to define reliability and to compare this distant-reading with the close reading of human fact-checkers. We do not want to develop a model for predicting websites’ reliability but to repurpose machine learning techniques to devise a view on trustworthiness drawing on the structure of the underlying news ecosystem. We can then use machine learning to understand the work of human news site assessors better.

[3] Note that for the later analysis we drop 102 websites that are not linked to the main body of the news ecosystem.

Our methodology thus follows the following steps:

1. To investigate how a news network's topology can help decide on the reliability of different sites, we first engineer a number of features that reflect our assumptions on what defines a reliable site according to the hyperlink network. We work with 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 [28] are based on 'deep features' engineered by the machine itself, which can be highly effective in classification tasks but are unreadable for humans. Our theory-driven feature engineering allows us to interpret the network view on reliability.
2. The second step in machine learning is to choose the right model. Here, we consider models capable of optimising the machine reading of reliability by considering network linking structures. We first present an advanced ensemble model that optimises detection performances and then two simpler but easier to interpret models that reflect the importance of asymmetric relations to determine reliability.
3. The third step is a comparison between the results of our model and the evaluation of *Le Monde*, paying particular attention to borderline decisions. 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 [31].

Features

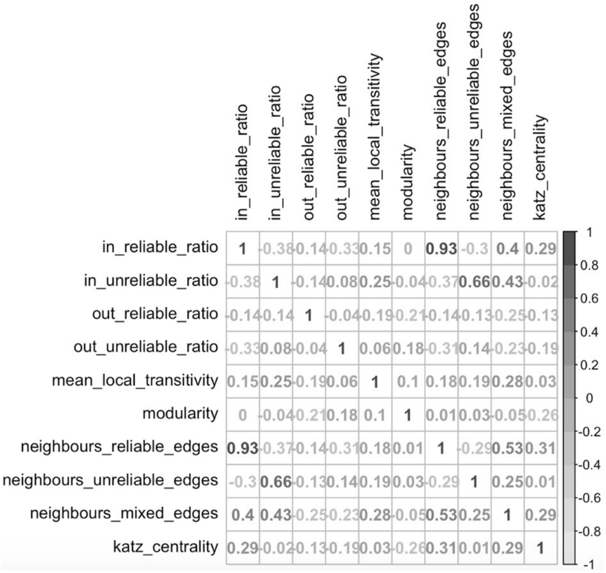
The first step in our methodology is the featurisation of the dataset, that is the selection and preparation of the network features that will be used as input for the machine learning models. For our experiments, we introduce several new measures that consider network attachments by including aspects of community formation and preferential relations. We furthermore compute several features that consider the direction of edges, which we have earlier established to be important for readability.

- Our four direct neighbourhood relations are the number of incoming and outgoing links from reliable and unreliable sites. Because the websites in our network can have a very different number of neighbours (varying from one to more than two hundred)-depending on their editorial practices and how they have been captured by our crawl -, we computed for each website the ratio between each type of neighbour and the total number of incoming or outgoing links of the website. These features are called: 'in_reliable_ratio', 'in_unreliable_ratio', 'out_reliable_ratio' and 'out_unreliable_ratio'.
- To move beyond the immediate neighbourhood of websites, we consider all the edges to and from the nodes in the 'in-neighbourhood' of each website, defined as the ensemble of sites citing the node under examination. We decided to focus on the website citing a source (rather than those cited by it) because our previous work Venturini et al. [37] suggested that their nature is

a better indicator of the source trustworthiness. As websites can choose who to link to, unreliable websites often try to pass themselves as reliable by citing authoritative sources. Websites, however, cannot choose who cite them and thus unreliable websites are generally incapable to attract citations from reliable ones.

- We consider, in particular, the average number of reliable, unreliable and mixed edges in relation to all edges in a node's in-neighbourhood. Reliable edges connect reliable nodes, unreliable edges connect unreliable nodes and mixed edges are found between reliable and unreliable nodes. The in-neighbourhood of a reliable site, we expect, should contain 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 Katz centrality. Katz centrality is a generalisation of eigenvector centrality for a directed graph. It measures the number of direct neighbours of a node and all other nodes that connect to that node through its neighbours. Setting the parameter alpha = 0.09, we penalise the connections with distant neighbours. We are interested in Katz centrality as it was shown to be effective in detecting both locally and globally interconnected nodes [14]. Reliable sites should be described by a higher Katz centrality given that they connect both locally and globally. If we rank the data according to Katz centrality, *Daily Mail* 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 choice of our model differs from those of the human evaluators.
- Modularity is computed for the cut that separates a site and its neighbours from the rest of the network ('modularity'). It is the fraction of the edges that fall within the cut minus the expected fraction if edges were distributed at random. Modularity should detect a network part that is highly integrated [30] and use this to determine reliability. On average the modularity of reliable sites is lower than the modularity of unreliable ones. We find the unreliable *Russia Today* at rank 5 and the *Daily Mail* at rank 14.
- The clustering coefficient of the neighbours is defined by the average local transitivity ('mean_local_transitivity'). It describes the probability that two randomly selected neighbours of a site are also direct neighbours. It is the fraction of pairs of the node's neighbours that are connected to each other. The most reliable sites should therefore have a lower clustering coefficient, as they are connected across the network and receive links from all types of sources. On average reliable sites have a 'mean_local_transitivity' value of 0.21 and unreliable a value of 0.23. 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 decide on the reliability of websites based on their hyperlink connections. Part of the machine learning models evaluation will be to find out which of the features provide the best support to describe reliability. As Fig. 2 shows, the ratio of incoming links from reliable sites correlates strongly (0.93) with other features. So, we remove that feature.



	in_reliable_ratio	in_unreliable_ratio	out_reliable_ratio	out_unreliable_ratio	mean_local_transitivity	modularity	neighbours_reliable_edges	neighbours_unreliable_edges	neighbours_mixed_edges	katz_centrality
in_reliable_ratio	1	-0.38	-0.14	-0.33	0.15	0	0.93	-0.3	0.4	0.29
in_unreliable_ratio	-0.38	1	-0.14	0.08	0.25	-0.04	-0.37	0.66	0.43	-0.02
out_reliable_ratio	-0.14	-0.14	1	-0.04	-0.19	-0.21	-0.14	-0.13	-0.25	-0.13
out_unreliable_ratio	-0.33	0.08	-0.04	1	0.06	0.18	-0.31	0.14	-0.23	-0.19
mean_local_transitivity	0.15	0.25	-0.19	0.06	1	0.1	0.18	0.19	0.28	0.03
modularity	0	-0.04	-0.21	0.18	0.1	1	0.01	0.03	-0.05	-0.26
neighbours_reliable_edges	0.93	-0.37	-0.14	-0.31	0.18	0.01	1	-0.29	0.53	0.31
neighbours_unreliable_edges	-0.3	0.66	-0.13	0.14	0.19	0.03	-0.29	1	0.25	0.01
neighbours_mixed_edges	0.4	0.43	-0.25	-0.23	0.28	-0.05	0.53	0.25	1	0.29
katz_centrality	0.29	-0.02	-0.13	-0.19	0.03	-0.26	0.31	0.01	0.29	1

Fig. 2 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. 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 estimation of the performance of the model. Every data item is part of the testing fold exactly once and of the training set in all other 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 finally 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 that best describe reliability. We tested generalized linear models (GLMs), randomForest (rF) and xgbTree. GLMs are more flexible generalizations of ordinary linear regressions, rF are the ensemble version of the well-known decision trees technique and are often used in prediction experiments, but xgbTree might be less known. XGB stands for 'eXtreme Gradient Boosting' (Chen et al. [11] 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 in a network while the logistic regression model is a commonly used model to detect symmetric communities based on class descriptions.

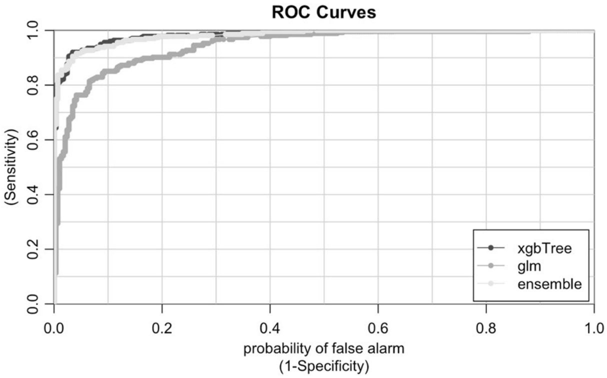


Fig. 3 ROC/AUC performance of the stacked meta-model

Tree models as well as other rule-based learners we introduce in "Ruling through the asymmetry of hyperlinking practices" represent knowledge as a set of rules or logical if-else statements, described by an antecedent and a consequence. So, a simple rule in our case would be: 'If a website links to more than 500 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 [7]. Recently, neural networks have become an attractive alternative for high-dimensional data, as a more recent evaluation by Zekić-Sušac et al [11] shows. In network terms, Narayan et al. [28] have introduced a novel approach to exploit neural networks based on local walks across the graph. However, these approaches generally learn their own representations of data rather than relying on a set of manually defined features and thus do not allow us to test our own assumptions about the network view on reliability.

All three models show an improved ROC/AUC performance compared to the statistical neighbourhood comparison of reliability. However, the random forest model is performing well below the other two models with an average AUC of ~ 0.88 compared to GLM and xgbTree both performing on average above 0.91 AUC. GLM and xgbTree are therefore already showing excellent performance, but we can do even better by stacking the models. We can build a 'meta-model' using the combined strengths of each individual model. 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. Because the results of the two best performing models XGB and GLM are slightly correlated (~ 0.65), we do not expect a major gain compared to the best XGB model.

The best performing stacked ensemble model outperforms its two contributing individual models. In the out-of-sample test of the ensemble, we achieve an AUC value of ~ 0.96 compared to XGB's 0.94. Figure 3 shows how the ensemble model smoothes the differences between XGB and GLM but all models perform well with XGB rising more quickly than GLM. Overall, the ensemble does not significantly

Table 1 Confusion matrix of meta-model

	Judged as 'reliable' by Décodex	Judged as 'unreliable' by Décodex
Categorized as 'reliable' the model	278	12
Categorized as 'unreliable' the model	25	250

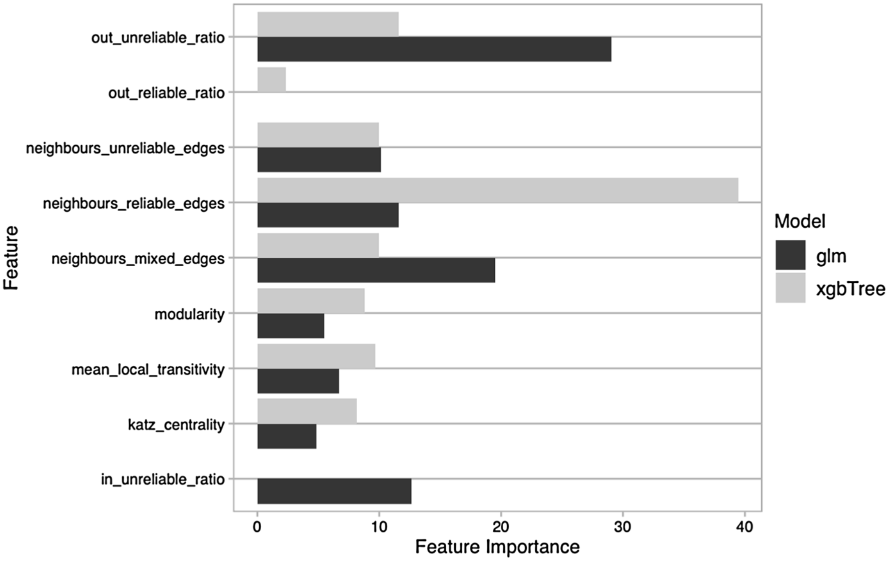


Fig. 4 Feature importance for GLM and XGB models

improve on XGB. XGB contributes about 76% to the decision of a linear combination ensemble, while GLM takes care of the remaining 24%.

The tests demonstrate how we can generate a highly accurate network view of reliable sites. We apply the ensemble model to the whole network with a very high accuracy of ~ 93% according to the confusion matrix in (Table 1).

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

Figure 4 shows which of the features are considered most relevant by GLM and XGB. For both models, the dominant features are the fractions of reliable and mixed edges among the neighbours of a website as well as the ratio of out-links to unreliable nodes.

For GLM, 'out_unreliable_ratio' is the most important feature, but this model seems to rely on almost all features to a similar degree. XGB as a more advanced model is better at differentiating between features with 'neighbours_reliable_edges' the most important one. Overall, the ensemble model prefers simpler features such as the ratio of reliable edges in the neighborhood of a website. The more complex calculations such as Katz centrality or modularity seem to play a lesser role. Together with the ratio of connections to unreliable nodes, the neighbours' reliable edges make up more than 70% of the decision. The fact that XGB clearly identifies neighbours' reliable edges as the strongest indicator of reliability confirms our intuition that receiving hyperlinks from reliable sites is a worse proxy of reliability than the fact of being linked to a cluster of reliable websites. 'out_reliable_ratio' is then also by far the least important feature for both XGB and GLM. Both unreliable and reliable nodes link out to reliable nodes.

To test the importance of the highest-ranked features as a proxy of reliability, we create a new ensemble model but this time using only the four top features: 'neighbours_unreliable_edges', 'neighbours_mixed_edges', 'out_unreliable_ratio' and 'Neighbours_reliable_edges'. For the ensemble model based only on these four features, the number of false negatives and false positives increases but the overall accuracy is only slightly reduced to ~ 89%.

Patterns of algorithmic reasoning machine reasoning

A thorough error analysis helps us discover patterns of algorithmic reasoning and in particular borderline decisions that show how human and machine evaluations differ. Because we are interested in evaluating the human definitions of news reliability, errors can be as revealing as correct predictions.

Looking at the misclassification of the ensemble model, we notice first some interesting assumptions in the judgements of the assessors which stand against a network view of reliability. The model miscategorize as reliable right-wing news sites in both English (e.g., *Fox News*) and French (e.g., *Contrepoints*). However, according to the confusion matrix in Table 1 the biggest problem for the ensemble model seems to be the fact that several reliable news sites are missed. *Agoravox*, for instance, was categorized as unreliable by the model but judged as reliable by Décodex, although it also has concerns about its possible spread of conspiracy theories. As a citizen journalism site hosting the contributions of tens of thousands of volunteer journalists, the question is really whether *Agoravox* can be judged as a whole. At first sight, such false negatives seem to be related to the tendency of the judges of Décodex to favour community sites that the model struggles to consolidate with its network view.

Considering underlying structural issues with the data, our approach seems to be better than suggested by the accuracy numbers. *Fox News*, for instance, will be extensively connected to reliable news sites, as it is one of the biggest daily TV stations in the US 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 to consolidate the network connections of *Fox News* with the Décodex judgement.

An interesting case that shows that the model can correct human judgements is *Génération Identitaire*, a hard-right unreliable political site. It is counted as unreliable by the ensemble model, while *Le Monde* incorrectly records it as reliable. This suggests that the method can discover clear errors in the human-created data.

As the ranking of features has shown the reliability of a site's neighbourhood to be the most important feature, we will move on to consider 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 directly interpretable by humans and, in our error analysis, we could only present the overall decision but not the steps involved.

Ruling through the asymmetry of hyperlinking practices

Let us investigate whether we can reproduce the importance of edge directedness with two other rule-learner algorithms specifically targeting these features. These can be seen as surrogate 'white-box' models that allow us to shed light on the functioning of our ensemble meta-model. We follow the same methodology by first presenting the results of the model results and then discussing its errors to reveal trends in algorithmic reasoning.

The first rule learner algorithm we employ is *OneRule* [17]. It simply selects the one rule that most accurately describes the decision based on the fewest prediction errors. OneRule is very easy to interpret for humans but can still be very powerful. As anticipated, it identifies the reliable edges between neighbouring nodes as the single most important feature.

Using neighbours_reliable_edges to split the network into reliable and less reliable sites, we retrieve the following four rules.[4]:

1. (neighbours_reliable_edges < 7.5%) -> UNRELIABLE (323/61)
2. (neighbours_reliable_edges >= 7.5%) AND (neighbours_reliable_edges < 11.5%) -> RELIABLE (21/8)
3. (neighbours_reliable_edges >= 11.5%) AND (neighbours_reliable_edges < 14.8%) -> UNRELIABLE (11/5)
4. (neighbours_reliable_edges >= 1.48%) -> RELIABLE (210/14)

The first number in the final brackets per rule is the number of cases covered, while the second number is the number of misclassified cases. The output clearly indicates that the ratio of reliable edges in a neighborhood is a good proxy of reliability. According to rule 1, a site is not reliable if the percentage the neighbours' reliable edges is lower than 7.5%. Rule 4 states that sites should be categorized as reliable if more than ~ 14.8% of the edges in their neighbourhood are reliable. The second and third rule reveal more complex distinctions for the model. We can link these to idiosyncrasies in the data (Table 2).

[4] Please, note that we transformed the output into percentages rather than values between 0 and 1 to improve readability.

Table 2 Confusion matrix of OneRule

	Judged as ’reliable’ by Décodex	Judged as ’unreliable’ by Décodex
Categorized as ’reliable’ by the model	209	22
Categorized as ’unreliable’ by the model	66	268

OneRule identifies 477 of the 565 sites correctly using exclusively the reliable neighbouring edges. This is about 84% of the corpus and demonstrates the power of this simple measure exploiting the asymmetry of hyperlinking practices. Overall, the additional misjudgements (compared to the baseline ensemble model) are fairly evenly distributed between false positives and false negatives as well as across the rules, though rule 2 and 3 clearly stand out has having the relatively highest numbers of misclassifications.

The list of misclassifications indicates first similar issues as we have discovered before. Sites that are judged to be unreliable but have reliable neighbourhoods are difficult to classify correctly. The *Daily Mail* is a UK newspaper that has a reputation for bad reporting but is also part of the commonly cited UK news ecosystems. The machine categorises it as reliable against the judgment of *Le Monde* because its asymmetric attachments (neighbours’ reliable edges) reflect that it is part of the broader media ecosystem. The value for the *DailyMail’s* neighbours’ reliable edges is 0.18. As a comparison, the most authoritative French news sources from the left and right, *Le Monde* and *Le Figaro* have both much larger values for neighbouring reliable edges—as expected—but still fall in the same category as *Daily Mail* in rule 4.

With regard to rule 1, we find many examples of sites whose neighbourhood contains less than 7.5% of reliable edges. These include the well-known *Breitbard* as well as *ZeroHedge*, a financial Blog accused of distributing right-wing conspiracy theories. OneRule identifies them correctly as unreliable. In terms of misjudgements in rule 1, the *New York Daily News* is a US-based newspaper with several Pulitzer prices but sometimes controversial news stories that the *New York Times* has called ‘populist’.[5] It has a neighbours_reliable_edges value of 0.05 and is wrongly categorized as unreliable. How the *New York Daily News* should be classified, can therefore be seen as a friction reflected in the disagreement between OneRule and Décodex. In rule 1, we can also find the tabloid newspaper *Irish Daily Star*, which is judged by Décodex as reliable but has a neighbours’ reliable edges value of less than 0.07. Regional news is less connected at the international level and thus less visible in the French media system. Regional and local news sources are therefore easily misjudged by this network-based methodology.

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

The change of judgements in rules 2 and 3 tells us more about the broader issues of predicting reliability from network attachments but also about the biases and issues characteristic of a human-created dataset such as Décodex. This is expressed as much in the correct classifications as in the wrong ones. A correct classification in rule 2 is, for instance, *Russia Today*, which is often seen as a Russian propaganda instrument but can also be cited by mainstream news sites. *Russia Today* is at the borderline of this classification with 0.07 for neighbours' reliable edges, which reflects their status as a site whose reliability is generally considered to be doubtful rather than clearly unreliable. Rule 2 also tries to correct the misclassification of *Génération Identitaire*, which OneRule assigns to the reliable sites following the error in the Décodex original classification. Its value is very close to the next group of unreliable sites with a value of 0.09 for neighbours' reliable edges. Rule 2 also has 8 misjudgements. These include, e.g., the *New York Post*, a tabloid competitor to the *New York Daily News*. The *Post* is judged as unreliable by *Le Monde* against OneRule, with a value of 0.09 for the neighbours' reliable connections. A soft boundary methodology (rather than the hard one employed by OneRule), might be better to consider such borderline cases.

Compared to the ensemble method, *Fox News* is now judged correctly as unreliable. Its neighbours have only about 13% reliable edges, thus activating rule 3. *Sputnik News* is also covered by rule 3, though as a borderline case with 0.117 for neighbours_reliable_edges. Next to these agreements between OneRule and the human judges, we also find several errors in rule 3. *The Daily Beast*, e.g., is judged as reliable by *Le Monde* but is categorised as unreliable according to rule 3 of OneRule. Rule 3 also highlights *Change.org*, which *Le Monde* has included for some reason in its list of news websites. As a crowdsourcing site, it will contain petitions of varying reputations. It has therefore a neighbours_reliable_edges value of 0.13. We have already talked about the unusual decision by the Décodex judges to include citizen sites, which have less control over the editorial and hyperlinking choices of their multiple authors.

An expansion to OneRule is JRip, based on the Ripper algorithm [12]. JRip tries all 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 three rules based on neighbours_reliable_edges and out_unreliable_ratio, which are also the two most important features according to the ensemble model. The other two most important features from the ensemble 'neighbours_mixed_edges' and 'neighbours_unreliable_edges' can be found in rule 2 next to 'modularity'.

1. (neighbours_reliable_edges >= 6.4) AND (out_unreliable_ratio <= 7.4%) -> RELIABLE (216/9)
2. (neighbours_mixed_edges >= 68.6%) AND (neighbours_unreliable_edges <= 39.9%) AND (modularity <= 0.00) -> RELIABLE (15/2)
3. ELSE -> UNRELIABLE (334/55)

The first rule predicts as reliable 216 websites that link to less than 7.4% unreliable sites and that have neighbours with strong reliable relations. This includes

Table 3 Confusion matrix of JRip

	Judged as ’reliable’ by Décodex	Judged as ’unreliable’ by Décodex
Categorized as ’reliable’ by the model	220	11
Categorized as ’unreliable’ by the model	55	279

9 incorrect classifications. The second rule is a more complex combination of ’neighbours_mixed_edges’, ’neighbours_unreliable_edges’ and ’modularity’. It tries to capture reliable sites that are in the neighbourhood of a significant number of unreliable edges but not too many. 13% of the cases in rule 2 are misclassified. It is at first sight not clear why modularity was chosen for rule 2. However, without the constraint on the modularity, the rule would identify 20 sites and 50% of them would be misclassified. The third rule states that in all other cases the nodes are not reliable, but it also contains the largest number of errors.

The performance of JRip is very good (with ~ 88% accuracy) and better than OneRule at both identifying reliable and unreliable sites. Compared to the ensemble model, JRip performs slightly better at avoiding false positives but significantly worse at identifying reliable sites. The strength of the baseline ensemble model comes from its dynamic combination of features providing a better balance, which also leads to a much better overall performance. It is able to identify other features than ‘modularity’ to make overall better decisions (Table 3).

Within the wrong classifications of the first rule, we find *Fox News* again because its fraction of reliable neighbouring edges is much larger than 7% and because it does not link out to unreliable sites. The misclassifications furthermore include the French libertarian *Contrepoints* and the Belgian left-wing site *Investigaction*. Both of these are still within the boundaries of reliable news discussions. In particular, they both avoid out-linking to unreliable sites according to the second feature considered by JRip. So, their classification as unreliable is based on their neighbourhoods raising again the question of what the judges from Décodex consider as reliable.

The second rule is small with only nine members and two misqualifications. It is again a reaction of the algorithm to challenges in the data. The two misqualifications are *Antipresse*, a site for conspiracy theories, and *Infos Bordeaux*, a hard-right site. It is not immediately clear why they are classified as reliable by JRip but they are for all three values in rule 2 borderline cases with both having a modularity of just below 0 and values of neighbours_mixed_edges of around 0.7 and of neighbours_unreliable_edges of around 0.36. More interesting is why the rule exists in the first place. The rule covers sites such as the personal website of Tariq Ramadan, which *Le Monde* thought to be relevant, and regional newspapers such as the *Houston* Chronicle which are not well connected to the rest of the network.

Misclassifications of the third rule include community sites such as *Hoaxbuster*, a community platform to limit the circulation of hoaxes. The classification of a fact-checking site as unreliable is actually very telling. While *Hoaxbuster* cites unreliable sites with the purpose of debunking their stories, it may, for the very practice of linking them, giving them more attention and thus ending up favouring the spread of misinformation [5].

Maybe even more strikingly, JRip also misclassifies some of the international news heavyweights such as *CNN* or *Politico Europe*. *Politico* had at the time of harvesting the hyperlinks a too small neighbourhood of reliable links, which indicates that the harvesting of links should be repeated at different time intervals. *CNN* has an out_unreliable_ratio of 0.167 though its neighbourhood of reliable edges is very strong with 0.34. Even the Blog page of the home of Décodex evaluators *Les blogs du Monde Diplomatique* belongs to this rule, as it links out to unreliable sites, even though it does have a lot of reliable edges in its neighbourhood.

JRip, as said, also improves the ensemble model. It judges correctly the website of conspiracy theorist Alex Jones, which the ensemble model categorized as reliable. Its neighbours do not have enough reliable edges and too many unreliable links.

Conclusion

This article has explored a network view on the complex politics of association involved in the reliability of news sites. Such a network view promises a topic-agnostic perspective that can provide useful insights on reliability and its 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 understanding the structure of news networks by comparing it with human judgements. Machine learning models can be used to reason about the decisions by human judges and identify underlying contradictions and challenges.

Overall, our machine reading has taught us a lot about the challenges of identifying reliability in a network of news sites. There is, for instance, the question of what should be included in a corpus of news sites. Our data included citizen journalism websites and an online campaigning site. The machine reading struggled to assign these sites to either side of the classification, as none of them is, strictly speaking, a news site. This problem is especially clear when considering a site like *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 highlights how difficult it is to judge the reliability of news. The machine reading struggled to reproduce the evaluators' judgements for sites such as *Fox News* or the *Daily Mail*. Both belong to official news ecosystems, with reliable sites such as the *BBC*, e.g., often reporting on stories from the *Daily Mail*. Both also have a reputation for biased and unreliable reporting. Rather than simply considering these sites as unreliable, the struggle of machine learning models based exclusively on network information reveals the heterogeneous decision-making behind news reliability.

Rather than focussing on fact-checking, we used predictive analytics to analyse existing knowledge about networks in a real-life human-created network of news sites. Our approach successfully integrated different network perspectives as well as community and preferential attachments. While all models we presented do very well in providing a network's view on reliability (and much better so than simple statistical descriptions), there are also limitations to this approach. First and foremost, the network is fairly small and unevenly distributed. It has relatively few nodes, as these were created by human judges, but many more edges detected by the Hyphe web crawler. This could explain why the edges might play such a big role in the judgements of the different models. 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 borderline sites.

Funding This work was partially funded by the projects SobigData++ (EC, H2020 n. 871042) and AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911).

References

1. Ackland, R., & Shorish, J. (2014). Political homophily on the web. In M. Cantijoch, R. Gibson, & S. Ward (Eds.), *Analyzing social media data and web Networks.* (pp. 25–46). London: Palgrave Macmillan UK.
2. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using N-gram analysis and machine learning techniques. In: *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 2017, pp. 127–138. Springer.
3. Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature, 401*(6749), 130–131 Nature Publishing Group.
4. Blanke, T. (2018). Predicting the past. *Digital Humanities Quarterly* 12(2).
5. Bounegru, L., Gray, J., Venturini, T., et al. (2017). *A field guide to fake news: a collection of recipes for those who love to cook with digital methods (Chapters 1–3).* . Research Report: Public Data Lab.
6. Brandtzaeg, P. B., & Følstad, A. (2017). Trust and distrust in online fact-checking services. *Communications of the ACM, 60*(9), 65–71 ACM New York, NY, USA.
7. Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine learning*, 2008, pp. 96–103. ACM.
8. Castelo, S., Almeida, T., & Elghafari, A., et al. (2019). A topic-agnostic approach for identifying fake news pages. In: *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 975–980.
9. Centola, D., González-Avella, J. C., Eguíluz, V. M., et al. (2007). Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution, 51*(6), 905–929 SAGE Publications Inc.
10. Ciampaglia, G. L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science, 1*(1), 147–153.
11. Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd SIGKDD international conference on knowledge discovery and Data Mining, ACM, pp. 785–794.

12. Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Machine learning proceedings 1995.* (pp. 115–123). San Francisco (CA): Morgan Kaufmann.
13. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: methods for finding fake news. In *Proceedings of the Association for Information Science and Technology*, John Wiley & Sons, Ltd *52*(1), 1–4.
14. Montijano, E., Oliva, G., & Gasparri, A. (2018). Distributed estimation of node centrality with application to agreement problems in social networks. In: *2018 IEEE Conference on Decision and Control (CDC)*, 17 December 2018, pp. 5245–5250.
15. Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In: *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, 2017, IEEE, pp. 110–115.
16. Hindman, M. (2008). *The myth of digital democracy*. Princeton University Press.
17. Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90.
18. Huibers, T. W. C. (1996). An axiomatic theory for information retrieval. Universiteit Utrecht Press, Utrecht
19. Jacomy, M., Girard, P., Ooghe-Tabanou, B., et al. (2016). Hyphe, a curation-oriented approach to web crawling for the social sciences. In: *Tenth International AAAI Conference on Web and Social Media*, 2016.
20. Jänicke, S., Franzini, G., Cheema, M. F., et al. (2015). On close and distant reading in digital humanities: a survey and future challenges. In *Proceedings of of EuroVis—STARs*: 83–103.
21. Keuschnigg, M., Lovsjö, N., & Hedström, P. (2018). Analytical sociology and computational social science. *Journal of Computational Social Science, 1*(1), 3–14. https://doi.org/10.1007/s42001-017-0006-5.
22. Kwon, S., Cha, M., Jung, K., et al. (2013). Prominent features of rumor propagation in online social media. In: *2013 IEEE 13th International Conference on Data Mining*, 2013, IEEE, pp. 1103–1108.
23. Lazer, D. M. J., Baum, M. A., Benkler, Y., et al. (2018). The science of fake news. *Science, 359*(6380), 1094.
24. Leskovec, J., Lang, K. J., Dasgupta, A., et al. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics, 6*(1), 29–123 Taylor & Francis.
25. McLuhan, M. (1994). *Understanding media: the extensions of man*. MIT Press.
26. Monti, F., Frasca, F., Eynard, D., et al. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint*.
27. Naeem, B., Khan, A., Beg, M. O., et al. (2020). A deep learning framework for clickbait detection on social area network using natural language cues. *Journal of Computational Social Science, 3*(1), 231–243 Springer.
28. Narayanan, A., Chandramohan, M., Venkatesan, R., et al. (2017). graph2vec: Learning distributed representations of graphs. *arXiv preprint*.
29. Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society, 20*(10), 3720–3737 SAGE Publications Sage UK: London, England.
30. Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102 American Physical Society.
31. Ng, A. (2016). Nuts and bolts of building AI applications using Deep Learning. *NIPS Keynote Talk*.
32. Page, L., Brin, S., Motwani, R., et al. (1999). *The PageRank citation ranking: bringing order to the web.* Stanford InfoLab, available at https://www.citeseer.nj.nec.com/page98pagerank.html.
33. Pariser, E. (2011). *The filter bubble: what the internet is hiding from you*. New York, NY: Penguin.
34. Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, 116*(7), 2521. https://doi.org/10.1073/pnas.1806781116.
35. Ravandi, B., & Mili, F. (2019). Coherence and polarization in complex networks. *Journal of Computational Social Science, 2*(2), 133–150. https://doi.org/10.1007/s42001-019-00036-w.
36. Sunstein, C. R. (2001). *Republic com*. Princeton University Press.
37. Venturini, T., Mathieu, J., Liliana, B., & Gray, J. (2018). *The Routledge handbook of developments in digital journalism studies*. Routledge.
38. 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. *An introduction to visual network analysis and force-directed layouts (April 26, 2019)*. Available at SSRN: https://doi.org/10.2139/ssrn.3378438.

39. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146.

40. Wallach, H. (2018). Computational social science≠ computer science+ social data. *Communications of the ACM, 61*(3), 42–44 ACM New York, NY, USA.

41. Wang, W. Y. (2017). 'Liar, liar pants on fire': a new benchmark dataset for fake news detection. *arXiv preprint*.

42. Zekić-Sušac, M., Pfeifer, S., & Šarlija, N. (2014). A comparison of machine learning methods in a high-dimensional classification problem. *Business Systems Research Journal, 5*(3), 82. https://doi.org/10.2478/bsrj-2014-0021.