

Aiding Informed Critical Thinking: Mining and Visualizing the Evolution of Online Content

Master Thesis

Krystof Spiller



Aiding Informed Critical Thinking

Mining and Visualizing the Evolution of Online Content

Master Thesis

January, 2022

By:

Krystof Spiller

Supervised by:

Andrea Burattin & Ekkart Kindler

Copyright: Reproduction of this publication in whole or in part must include the customary bibliographic citation, including author attribution, report title, etc.

Cover photo: Freepik / lemonsoup14 <https://bit.ly/3zqeEGA>

Published by: DTU, Department of Applied Mathematics and Computer Science,
Richard Petersens Plads, Building 324, 2800 Kgs. Lyngby Denmark
www.compute.dtu.dk

Abstract

False information is a pervasive problem which only becomes more serious with passing years. Being able to navigate the modern information landscape is imperative for maintaining stable democratic society. There is currently no effective and widespread technological solution addressing the issue. The thesis discusses the flaws of existing non-technical solution and the immaturity of most technical solutions and argues for a as-of-yet untried highly collaborative solution where the main goal is not prescribing truth but informing and supporting individuals in critical thinking by considering credibility of the entities involved in a narrative.

The thesis proposes a conceptual solution by considering what is needed for building a technological solution to encourage critical thinking. It discusses considerations for collection of relevant data and proposes inclusion of some often disregarded metadata helpful for credibility assessment. It proposes to communicate information about a narrative using visualization in form of a graph for a single narrative. This graph offers interactive elements through which the user can approximately set their beliefs about importance of different aspects affecting information credibility. The proposed solution is also designed from the perspective of being highly collaborative which is considered and argued to be superior because of increased transparency as well as potential for fast growth and improvement because of the community involved.

At the end, the proposed solution is largely positively evaluated by a set of interviews which shows potential and encourages further work in this direction.

Contents

Abstract	ii
1 Introduction	1
1.1 Narrative example	5
2 Background	7
2.1 Approaches for addressing false information	7
2.2 Glossary	14
2.3 Problem statement	15
2.4 Research questions	15
2.5 Methodology	16
3 Related work	17
3.1 Exploratory work	17
3.2 Aiding fact-checking	18
3.3 Automated fact-checking	18
3.4 Related journalistic work	20
3.5 Active intervention via tool	20
4 Conceptual solution	21
4.1 Domain model	21
4.2 Data collection	25
4.3 Credibility graph	31
4.4 Addressing cognitive bias	41
4.5 Collaborative project organization	42
4.6 Proposed usage	46
5 Implementation	56
5.1 Credibility graph visualization	56
5.2 Configuration file	56
5.3 User experience simplifications	58
6 Evaluation	59
6.1 Verification	59
6.2 Validation	59
6.3 Limitation	66
6.4 Future work	67
7 Conclusion	69
Bibliography	70

1 Introduction

The problem of false information is not a new one. It has been with us since language first evolved but the scale of the problem has only become greater thanks to the internet and related changes to, for example, digital journalism, as well as new technologies, prime example of which is social media. Some false information is recurrent and so persistent that it has been around even before the advent of internet and has only changed the media through which it spreads¹.

The interest around false information has risen especially in the last five years due to massive disinformation campaigns for the political cause of Brexit [4] and 2016 US presidential elections [5]. Nevertheless, political causes are not the only topic touched by false information. There are also instances affecting areas of financial markets [6]–[8], natural disasters [9], [10], specific individuals [11] or just common misconceptions [12].

The impact of false information can range from inconsequential to severe, both on the personal and societal level. It can negatively affect the path through which one's life or the whole society proceeds. Let's explore the *impact spectrum* with a few increasingly more impactful examples:

- Being misconceived about how a microwave works, where the commonly-held misconception is that microwave heats up the food by operating at a special resonance frequency of water, does not affect anyone's life in any meaningful way and is an innocent belief².
- Staying in the realm of microwaves, believing that looking into a microwave can damage one's eyes is also relatively innocent but it can be a source of anxiety, especially in the situation of a caring parent trying to protect their children. In some cases, such a worry can outright prevent purchase and usage of microwave even though it would otherwise be a net positive for the family or individual involved.
- Being confused about how tax brackets work and falsely believing that an increase in gross income might reduce one's post-tax earnings due to moving to a higher tax brackets can re-enforce a belief that a different taxation policy, such as a flat tax rate, is more favorable and can therefore affect for whom such an individual votes in the elections, possibly choosing a candidate they would not choose otherwise had they been well-informed about the workings of progressive taxation. Such a belief might not be consequential in isolation but if a critical mass of people follow the same misconception, it can result in affecting the policy and having an outcome that the same group of people would consider as undesirable, i.e. higher taxation rate with flat tax resulting in a reduced post-tax earnings compared to a progressive tax.
- Having read controversial information about vaccine safety, side effects and effectiveness can prevent people taking a vaccine. This increases a risk of severe illness and mortality directly for the individual as well as indirectly for the community due to decreased chance of achieving herd immunity.
- Believing in a conspiracy theory that a certain establishment is involved in human

¹So-called zombie claims [1, p. 47], example of which is the cabbage law length myth [2], [3].

²Possible exception is if you are directly involved in designing microwaves.

trafficking and is a part of a child sex ring can result in an individual trying to rescue the allegedly captive children by using excessive force, such as was the case in the infamous Pizzagate incident [13], [14] where a man fired a rifle inside a pizzeria allegedly involved in such illicit activities. The pizzeria staff also received death threats. This is obviously an extreme example of the impact false information can have on a behavior of an *individual*.

- The 2021 United State Capitol attack [15] is an instance of when such conspiracy theories affect large group of people and an angry mob determined to address the situation forms. In this instance, most members of the mob believed the false claims of 2020 presidential election fraud and wanted to overturn the result of the election. The attack resulted in multiple deaths and many more injuries, damages worth \$30 million and is recognized as a historical event that garnered the attention of international press. It also has a significant political impact where the whole democratic system is under large scale attack. The scale of this event showcases the impact false information can have on a *crowd*.³

To relate the issue of false information to a topical matter that affects everyone's life in the last two years, consider the infodemic around COVID-19 [17], especially concerning vaccinations:

Recent misinformation induced a decline in intent of 6.2 percentage points (95th percentile interval 3.9 to 8.5) in the UK and 6.4 percentage points (95th percentile interval 4.0 to 8.8) in the USA among those who stated that they would definitely accept a vaccine. ([18])

Another general effect of false information is lowered trust in democratic institutions [19] disrupting the health of society.

If an individual is well-informed and thinks critically, the chance they will succumb to such false information decreases [20]–[22]. The assumption here is that both informedness and critical thinking are imperative to resist false information. Being only well-informed but not thinking critically about an issue is not helpful since the information is not evaluated with reason and any outcome of such reasoning is possible. Critically thinking without having all the information means operating with limited information, where crucially important information that might change the result of such reasoning is missing, which might again lead to an unreasonable conclusion. The goal should then be to both increase informedness and promote critical thinking.

There are various strategies to address the false information in the contemporary digital information landscape among which are supporting investigative and fact-checking journalism, reducing financial incentives for spreading false information, improving media literacy of the general public, social media platforms attempting to reduce the spread of false information or governments passing anti-disinformation laws and establishing task forces to address such information.

All such strategies come with their pros and cons. Governmental action, especially one that involves blocking and removing content, is problematic since it can often be viewed as censorship limiting free speech [23]. Social media platforms face the same problem where it becomes very difficult for them to be fair across the thematic spectrum and without

³Last example is (thankfully) fictional and has been brought up by the recent movie Don't Look Up [16] where a comet is on a collision course with Earth and no one seems to care. The movie brings about issues of society distracted by social media, problems with the constant 24-hour news cycle but to a certain degree the problem of false information as well.

appearing as an arbitrary moderator of truth. Addressing financial incentives is a tall order in a market where just two companies (Google and Facebook) take one third of the global advertising spend [24] and which are also almost exclusively reliant on the revenue from selling ads [25] forcing them to optimize for selling as much ads as possible creating an environment ripe to be abused by creators of viral false information. In addition, both governments and social media companies are seen as top-down entities which can often be associated with some ulterior motives.

In journalism, the comparatively recent discipline of modern fact-checking⁴ attempts to gather the facts, preferably from public disclosable sources, connect the dots and present a neutral conclusion to a claim being checked. Such a presentation of a narrative can then have a positively informative effect on an audience that is, briefly said, in constant search and understanding of reality. Such a *reality-based audience* is generally open-minded, does not have many dogmatic beliefs, has respect for the truth and mostly believes in science and intuitively understands the scientific method. This is a requirement since if a reader only selectively considers certain information it is easy to omit other important information that can change the perspective. This therefore means it is problematic to reach everyone, let alone equally. It is also more likely that the hard to reach audience is affected more by information found further on the severe side of the impact spectrum.

However, when the audience meets such a requirement they are offered a holistic view of the narrative involving the nuances of the often messy reality as well as a conclusion for the narrative given by the fact-checker. While this is certainly helpful, two caveats can be identified.

First, it might miss people with beliefs on the fringe of the impact spectrum whose beliefs are often stubbornly held and hard to challenge, and thus remain largely unaffected by such fact-checking venture. Only other interventions of a psychological kind can affect these individuals and make them open and think critically from a different point of view [27], [28]. The only hope fact-checking brings in this regard is that the way of thinking espoused by the reality-based audience spreads also in the communities around the fringe of the impact spectrum.

Second, not only do the fact-checking publications gather information that provides a more comprehensive view of the narrative but they also largely assess the collected information and evaluate it in form of some rating — ordinary true to false spectrum, PolitiFact's Truth-O-Meter spectrum with its notorious "pants on fire" rating [29] or Washington Post's zero to four Pinocchios [30]. While it is understandable that these fact-checking publications do this mainly to help their readers digest the information, it can also be argued that in this way they stipulate what readers should think about the discussed narrative and are therefore not much better than other more questionable sources that lay out what is truth as they see fit. Even though it can be argued that such an approach is justifiable due to the fact that fact checkers provide the sources they used to evaluate the narrative and reach their claim, it is easy to see why some can see fact-checking organizations as yet another entity imposing the truth based on some hidden agenda.

Another challenge in dealing with false information is presented by the myriad of cognitive biases [31]. It is not desirable that all individuals in the society think homogeneously as discussion of diverse ideas is vital to a healthy democracy and society, and difference of opinion should therefore be encouraged. All human beings, however, are subject to

⁴That is organizations and projects either exclusively or largely focused on publishing fact checks. For example, FactCheck.org which is credited with starting the modern fact-checking movement was founded in 2003 [26].

various cognitive biases that affect how we approach different narratives. Even if an individual approaches a certain topic with rational and clear thinking it is not guaranteed such mode of thinking will be applied to other topics. There can therefore be a topic where a conclusion drawn to many of its narratives is a non sequitur. Meaning, that if the identical mode of thinking would be applied to such narratives as is applied to others, where one's thinking is clear and rational, it would lead to a different conclusion.

Since cognitive bias is often deeply ingrained and difficult to deal with as an individual, such non sequiturs can be identified often only while discussing ideas with others who know them and their way of thinking well and can identify holes and jumps in their reasoning. Not everyone might be fortunate to find themselves in such company or at least not very often. Their way of thinking then does not get challenged often enough preventing them to see such issues clearly.

In order to attempt to address this issue, let's briefly introduce one way how to look at reasoning, specifically forming conclusions to narratives. There are multiple entities in the information landscape that all play their role in the reasoning process. Let's first have a look at these entities.

Major role plays the information itself which is most often in the form of text but could be accompanied or outright replaced by other media like images, audio or video. This information relates to a one or more stories, also often called narratives. At the center of a narrative is a single unique statement identifying it. Although in reality a document relating to a narrative might often have more than one statements, the assumption here is that these can be separated into multiple narratives. Individually, the information appears in documents that can be uniquely identified by a persistent identifier like a Digital Object Identifier (DOI). These documents can support or oppose the central statement of a narrative. They can also provide mix of the two or give a neutral account for the narrative. The documents are authored and published by some entity. This can both be a physical person or a group of them as well as an organization or a company. These entities and the relation among them play a significant role to assess the credibility of a statement.

The way how the documents relate to the central statement of a narrative can also differ. Consider a narrative started by one document such as a tweet. Other subsequent documents relating to the narrative might be a response to the first document. In such a way, a chronological thread of documents can form. These documents can also build upon the first document and use it for additional arguments. These are just two examples of a more complex structure emerging.

These entities relating to a single narrative therefore form a graph where the entities represent the set of vertices and their relationships represent the set of edges. Each entity, or a vertex, has a weight associated with it, representing how much a person trusts and believes in it. The quality of trusting and believing is defined as credibility so this weight will be from now on referred to as such. If we were to ask different people about their level of trust and belief in different entities, each person would likely assign different credibility to each entity. Once we have these initial credibilities assigned it is easy to see how they can propagate and affect one another through the connections, i.e. edges, in the graph as will be introduced next on two examples. Since the graph in question has a specific use in regards to propagation of credibility among the interconnected entities, it can be denoted as *credibility graph*.

The credibilities assigned by different entities might vary in scope. It can be fairly small and specific, such as a high credibility for a certain favorite author and their blog, or rather

large and generic, such as generally higher credibility towards one side of the political spectrum. One can then imagine an example of a narrative where a document is written by this favorite author and published on their blog and such document will therefore have a higher credibility compared to others. In another example, for a person who might have a preference for conservatively leaning sources, such authors and publishers would be assigned higher probabilities and these probabilities would then propagate and affect the resulting credibility of the documents they authored and published. The propagation of credibilities does not stop there, after what can be called one iteration of credibility propagation, and these documents might then affect credibilities of other connected documents and crucially, the statement at the center of a narrative.

Once these credibilities converge, i.e. they remain approximately identical after another iteration of credibility propagation, we are left with some credibility for the statement at the center of a narrative. The higher the value, the more credible the statement is. If this value is high enough, the statement can be considered credible and vice versa. In this model, such a result represents the reached conclusion to a narrative. Importantly, such a model can be used to address cognitive bias, a concept explored further in the thesis.

1.1 Narrative example

An example of a narrative that is used throughout the thesis to explain various concepts is described next. It takes place in November 2016 in the midst of US presidential election and it is a false claim first debunked by a fact-checking website Snopes [32] and later covered by The New York Times as a case study for how fake news (their term) goes viral [33].

It starts with a photograph being shared on Twitter by Eric Tucker, at the time an average user with only few dozens of followers, on November 9 after 8 pm showing a group of buses that were allegedly used to ship in paid anti-Trump protesters to Austin, Texas. The tweet was then posted to Reddit on November 10 shortly after midnight and later in the morning a link to the Reddit thread was posted on a conservative discussion forum Free Republic. Various Facebook pages linked to the Free Republic thread and over 300,000 people have shared this link.

Later that day, some opposition to the prevailing narrative started to appear. Sean Hughes, the director of corporate affairs for the bus company Coach USA North America whose buses were on the shared photo, responded with a statement that "at no point were Coach USA buses involved in the Austin protests". Eric Tucker, the original poster of the photograph, was replying to queries on Twitter and admitted further lack of evidence by stating that he "did not see loading or unloading". Both of these statements, however, did little in the way of stemming the online frenzy.

Still on November 10 at around 6 pm, the conservative blog Gateway Pundit posted a story using the original shared image under the title "Figures. Anti-Trump Protesters Were Bussed in to Austin #FakeProtests". The post added another conspiratorial claim mentioning involvement of George Soros' money, a frequent trope of conservatively leaning conspiracy theories. Other conservative blogs shared the same narrative, incorporating Eric Tucker's original tweet to posts about paid protesters.

Shortly after 9 pm, then-president Donald Trump sent a tweet referring to "professional protesters". This emboldened Mr. Eric Tucker who considered removing the tweet at that time.

On November 11, Doreen Jarman, a spokeswoman for a software company Tableau

which was at the time organizing a conference in Austin, issued a statement to the local television station KVUE and a major daily Austin newspaper The Austin American-Statesman, saying that the buses were used for the company's conference. The American-Statesman posted an article shortly after noon. Around 2 pm, Mr. Tucker tweeted a link to his blog where he acknowledged he could have been "flat wrong". This is also the time when Snopes posted their rebuttal regarding the claim.

Snopes pointed out tweets by one Twitter user who reported helping load and unload attendees for the Tableau conference. In an update to the original rebuttal, Snopes also linked to a similar rumor posted on November 13 showing a large number of buses parked on a street and explains that the street where the video was taken is one of the closest spots to downtown that allows buses to park for free, which is supported by Google Street screenshots from the same road in different years.

After midnight on November 12, Mr. Tucker deleted his original tweet and instead posted an image with the word "false" written over it. Compared to the original tweet which was shared and liked over 10,000 times, it didn't garner much attention as it wasn't shared and liked even 100 times.

The reason for choosing this narrative is its reasonable complexity (i.e. neither too simple nor too comprehensive while offering many of the elements treated further in the thesis) and availability of the case study analysis performed by The New York Times as well as Snopes. The political topic of the narrative is not of any particular interest.

If we were to judge where on the impact spectrum this example falls, it would probably be between middle and high impact since at best, people affected by such narrative might lean towards lending their support to and voting for Donald Trump, and at worst, such narrative might compel people to organize a counter-protest that could result in clash and possible injuries or even casualties.

Chapter 2 provides a more comprehensive and thorough background for the thesis, specifying the problem statement, research questions and the methodology used. Chapter 3 explores related work within the area of false information and relates the different tasks to the various approaches for addressing the problem of false information. Chapter 4 explores the conceptual solution for the identified problem. It introduces a domain model together with the assumptions applicable for the task at hand. Chapter 5 focuses on implementation of a solution for the credibility assessment task relying on the conceptual solution identified in Chapter 4. It focuses on modeling the credibility propagation between the different entities in the credibility graph. Chapter 6 evaluates the results of the thesis by discussing verification, validation and limitations of the solution. Based on this, it also considers the possible future work. Chapter 7 concludes the thesis.

2 Background

A more detailed treatment of the introduced topic is given here together with context relating the selected approach in this thesis to other approaches. The terms developed in this chapter are listed and explained in section 2.2. The problem statement is succinctly stated in section 2.3. Research questions are given and their relevance explained in section 2.4. Lastly, the chosen methodology is explained in section 2.5 together with reasoning for why it was chosen.

2.1 Approaches for addressing false information

As illustrated with the examples for the impact spectrum in the introduction, it is well worth addressing the problem of false information. However, instead of approaching the topic from the perspective of decidedly determining what is true or false, this thesis makes the assumption that approaching it by aiding people being well-informed and think critically is a better goal to strive towards.

Let's discuss the space of addressing false information more broadly at first. The following is an attempt at a high-level categorization of approaches available for dealing with and understanding false information or its specific sub-domains. As already suggested, one dimension for this categorization is whether the goal is to understand, **explore** and study false information, or whether the goal is e.g. to **assess** veracity of a given claim or categorize claims and narratives. Another dimension in this categorization is whether the approach is manual or automatic. Table 2.1 provides a single example for each combination but there can be other approach falling into the same combination. This categorization also simplifies the fact that the two dimensions exist on a continuous spectrum instead of being discrete as shown. For example, majority of manual approaches involves some level of automation but not enough to be generally considered as automatic. In the following list, these examples are further explained and instances of research are given for each.

Approach	Explore	Assess
Manual	<i>Field experiments</i>	<i>Fact-checking</i>
Automatic	<i>Online content research</i>	<i>Automated fact-checking</i>

Table 2.1: High-level categorization of approaches for addressing false information.

1. *Field experiments* - [34] is an example of a field experiment exploring the ability of young people to judge credibility of online information. Generally, such a field experiment involves gathering participants in the real world, testing them, observing their behavior and analyzing their responses.
2. *Fact-checking* - Relatively recent journalistic discipline that is done either externally often by dedicated non-profit organizations (e.g. [FactCheck.org](#) or [PolitiFact](#) in the US, [Full Fact](#) in the UK or [TjekDet](#) in Denmark) or internally by the publisher¹. This process entails a person systematically exploring the claim and, preferably using only publicly available information, neutrally describe the argument and draw a conclusion, establishing veracity of the claim.

3. *Online content research* - [35] is an example of a research analyzing spread of information on Twitter, here specifically with the conclusion that false information spreads more pervasively than the truth and that human behavior contributes more to the spread than automated (bot) activity. Generally, such a research often involves performing statistical analysis on data from an online platform and the amount of data is usually of a size which would make it infeasible to gather and go through manually.
4. *Automated fact-checking* - Often abbreviated as AFC. This is most likely an area with the biggest research interest and variety of solutions, some of which are explored in chapter 3. The research community strives for an AFC solution that would mimic the manual approach of fact-checking while avoiding the bias introduced by a human being as well as being much more scalable than a team of fact-checkers. Creating such a system is a tall order due to the number of complex sub-systems that need to be involved and we are nowhere close to having such a system available. Much of the research now focuses specifically on some of the necessary sub-systems or problems closely related.

Each category contributes in a different way to address the problem of false information. Exploratory approach studying for example how false information spreads and why do people believe it often gives us insights for how to fight it but it does not directly move the proverbial needle in the struggle against false information.

On the other hand, fact-checking addresses particular instances of false information and its direct effect on reality-based audience is non-negligible although even more utility from such fact-checking might be brought by holding other accountable and promoting good journalistic practices. Since the modern fact-checking movement started with founding of FactCheck.org in 2003, fact checkers have focused primarily on education of their readers. While educating and informing readers remains an important part of work for fact checkers, as argued in [36], the so-called second generation fact-checking generation now also focuses on agitating their regular readers and organizing them to challenge those promoting false information. But it also acknowledges the need for what it calls the third generation of fact-checking that needs to work on an internet scale, be massively collaborative and work across international borders. This could be understood as only the fact-checking organizations being more open and collaborative with each other but could also include collaboration with their readers and anyone involved in consuming information. Additionally, the fact checkers alone with their limited bandwidth and their manual, i.e. non-automated, work are not enough to meet these challenges and they require additional help, part of which falls within automation.

What can then be said about the stated example falling into the assess and automatic category from table 2.1, i.e. AFC? Since no clear and agreed-upon definition of AFC exists, most likely because a true AFC system does not currently exist, and different authors think about it in different terms, it is first necessary to define the term. For our purposes it will suffice that AFC denotes the computer science discipline imitating the journalistic discipline of fact-checking. Nevertheless, as can be seen from the previous paragraph, the fact-checking discipline itself is evolving so even the understanding of what fact-checking is, and consequently what AFC is by our definition, can vary. However, drawing on how most people perceive fact-checking from the fact checks they have seen, it is usually a

¹Many established publishers followed best practices and code of honor making them operate with the same integrity that modern fact checkers do. However, the modern notion of fact-checking where a fact check is published all by itself and especially organizations that are exclusively focused on publishing such content is an innovation of the early 20th century.

rather prescriptive practice in a sense where a conclusion about the truth value is given. Such understanding can also be seen from the wealth of research that analyzes and categorizes claims as true or false, and that incidentally often associates itself with the discipline of automated fact-checking, subject that is more closely explored in chapter 3.

The following section 2.1.1 focuses directly on this issue of truth prescriptiveness and emphasizing the different approach taken in this thesis. Section 2.1.2 discusses the issue of a mode of organization or management of major fact checking organizations related to how overbearing they can be and offering an alternative way. These two sections together with the overarching goal to aid informed critical thinking provide a basis for the proposal explored in this thesis. A great challenge that makes critical thinking difficult is presented by cognitive bias, an issue explored in section 2.1.3. Since the ultimate goal is to *actively* aid people in informed critical thinking, how the proposed solution can be used in people's daily life needs to be considered, an issue discussed in section 2.1.4. Lastly, a summary of requirements for the proposed solution is given in section 2.1.5.

2.1.1 The issue of truth prescriptiveness

This prescriptive quality of fact-checking, where it collects relevant facts, analyzes the soundness of reasoning and optionally does some of its own and draws a conclusion about the truth value of a narrative, is, as already mentioned in the introduction, a double-edged sword. On the one hand, it helps with the fact check being easy to digest as well as lending itself to being easily shareable as a response or statement by others, assuming a headline summarizing the reached conclusion is provided. On the other hand, the exact same practice garners number of opponents precisely because of its prescriptiveness.

This thesis approaches the problem from the perspective that the goal is *not* to prescribe a certain truth and with it ideally make everyone have the same beliefs and think alike. Rather, the goal is to inform and encourage critical thinking. So while the goal of AFC is to establish veracity of a certain claim, i.e. say whether something is true or false, this thesis deals with a comparatively weaker relation, only saying whether a claim could be trusted or not when critically thought through while establishing credibility for sources and documents involved in a narrative. Considering the main goal of the proposed solution, it will be from now on referred to as *Critical Thinking Support System* (CTSS for short).

AFC is, however, worth pursuing, especially since number of sub-tasks, described in more detail in section 3.3.2, is relevant for CTSS as well. AFC tries to build upon and exploit the gains the AI field has made in the last decade using deep learning and big data. Specifically, the sub-field of Natural Language Processing (NLP) is of special interest for processing textual information and forms a backbone of many systems solving a sub-problem in the topic of AFC. Other machine learning techniques can be involved when considering other media like audio, images and video. NLP is then used to analyze content of the document but it might also involve additional information such as title, number of authors, date and time of posting, URL, alleged country of origin or users' comments. This analysis is then used for different purposes, such as evidence extraction, stance detection, bias estimation or claim detection to name a few. It will become evident how some of these sub-tasks are also relevant for CTSS in the subsequent chapter.

2.1.2 The issue of overbearing organizations

Another issue closely relating to the truth prescriptiveness relates to its source. If we were to rate the general perception of how overbearing governments, social media platforms and fact checkers are, arguably one of the most prominent bodies in the false information discussion, they would come in the specified order with the governments on top and fact checkers at the bottom. Both governments and social media are massive entities with

a lot of power behind and they are generally perceived as not being afraid to (ab)use it whereas fact checkers are relatively much less powerful which leads to the perception of them not being able to act as overbearing as governments and social media platforms. The mission and stemming practices of fact checkers also play its role when compared to other traditional media and are trusted more as a result [37]. Nevertheless, even fact checkers can be seen as overbearing, a feeling that everyone who has ever read a fact check they at first did not agree with surely experienced, and a reality that fact checkers themselves are well aware of [36].

The approach of CTSS, where the goal is not to prescribe the one and only truth helps in this regard, but another element that can aid with this perception is the way it is organized. As previously mentioned and supported by a group of fact-checking organizations [36], there is a promise in the next chapter of addressing false information and fact-checking (the so-called third generation) for the new solutions to be more collaborative. Imagine then a crowd-sourced fact-checking where interested information consumers get involved. Such an approach would therefore be more wiki-like. The collaborators could help with identification of new narratives, collection and organization of relevant data. More technically savvy collaborators could help with development of the system. Lastly, everyone would have access to see and analyze how the whole machinery works, improving its transparency.

There are two main benefits to such an organization approach. First, it helps to curb the issue of overbearingness where it is less valid to point at a community-run platform and state that it operates with ulterior motives and a hidden agenda when the whole system is open for everyone to examine and when that individual can become part of the community and attempt to right the alleged wrongs they see. Second, community-run platforms have an immense potential to grow and surprise in unexpected ways with its impact. Such a community can identify and address current narratives and issues with great speed since it is a much higher chance that some members of the community are close to the center of events. It can also identify and explore various ways of platform-wide abuse and report it such that it can be promptly addressed. Not to mention the technically-savvy members might get directly involved in addressing the issue. The main disadvantage is that once (and if) the community truly gets engaged, the platform suddenly has a life of its own, the initial tight control is lost and the unexpected and abusive uses of the platform become evident. However, Wikipedia has demonstrated that it is possible for such a community-driven collaborative project to thrive and address issues caused by such organization as they arise and although the challenge remains, it should not be seen as a roadblock for such an approach.

2.1.3 The issue of cognitive bias

The problem of various cognitive biases as it relates to critical thinking can be summarized by saying that a different mode of reasoning is applied to selected narratives. As this effect is undesirable, it is relevant to come up with a solution addressing this issue. Intuitively, if the problem is a different mode of reasoning for different narratives, the straightforward solution is to identify when this different mode of reasoning is applied and instead apply the standard mode not affected by cognitive biases. This is obviously more easily said than done but certain strides towards such a solution can be made.

The proposed solution can be gradually built starting with the concept of credibility ascribed to the various entities in the information landscape. These entities have relations to one another through which the credibilities can propagate. This propagation of credibility is governed by various rules which strength can be adjusted and which are applied

universally (in every relevant situation of a narrative) and uniformly (with the same strength in every case). These entities together with the relations can be mapped to a graph where nodes represent the entities and edges represent the relations between them. Since the graph mainly relates to credibility and its propagation, it is accordingly called credibility graph. The credibility graph can be understood as a simple model of the reasoning process that leads to forming of conclusion to a narrative.

The proposed solution then builds on the concept of credibility graph presented in the introduction which should be understood as a simple model of the reasoning process and of how are conclusions to narratives formed. This credibility graph has adjustable rules governing the propagation of credibilities. Before getting to the solution, let's dig further into credibility and credibility propagation rules.

Credibility itself can be split into an objective and a subjective part. For example, if we have the same content in two documents, one with and the other without supporting and meaningful sources, the former would generally be considered to have a higher credibility². Another example is relating to publisher entities, where certain initiatives to address false information can be assessed as objectively positive, at least when compared to not having any initiatives at all. The attempts by social media platforms to stymie spread of false information by various measures can be generally seen as good and as inherent to these publishers which should therefore to a small degree contribute to the overall credibility. On the other hand, if a source, both an author and a publisher, historically supported a significantly larger number of claims that were later debunked, this should lead to a general decrease in the overall credibility.

As for the credibility propagation rules, the two examples given in the introduction show the intuition behind how a rule in plain language ("I believe in my favorite author and their blog" and "I generally trust conservative sources more") affects the credibility of documents these sources produce. Other rules affecting the propagation of credibilities can be thought of and categorized by how they affect the propagation. This categorization is shown in table 2.2.

In the implementation of such rules, various parameters affect their behavior — whether they are enabled in the first place and how pronounced they are. The collection of these parameters for all rules will be shortly denoted just as *set of parameters*.

Now onto the proposed solution. The idea is to use the same set of parameters across various credibility graphs representing different narratives while ensuring the conclusion to each of the narratives agrees with that of the individual. A parallel could be drawn from this to software testing. The narratives represent unit tests where each is asserted on the desired conclusion set by the user and only the set of parameters affect the outcome of these unit tests. The more narratives we have in such a unit test, the more constrained the space of acceptable model parameters becomes. The idea then is that when a non sequitur conclusion comes into picture, the space of acceptable model parameters attained from previous narratives does not allow for this narrative to get to the intended conclusion

²A caveat to this has been demonstrated by a research discussed in section 3.1.2.

³Also, the credibility of the author and the publisher could be affected, although this is harder to judge and depends on the context. Such a behavior could be a sign of undesirable haphazard attitude flipping and could forecast occurrence of the same behavior in the future. In such a case, the author's and/or publisher's credibility should decrease. On the other hand, it could be a sign of an intellectual honesty when an honest mistake has been made and an attempt to right the past mistakes has been made in which case, the author's and/or publisher's credibility should increase.

⁴This does not have to be necessarily direct reference and can go through other sources — the size of the self-referential loop increases.

Credibility propagation rule category	Explanation & example
<i>Grouping</i>	Affects credibility for entities fulfilling a specific condition, e.g. "I believe in my favorite author and their blog" (small group) or "I generally trust conservative sources more" (large group).
<i>Aggregation</i>	Aggregates, i.e. merges, entities based on some condition, e.g. "Similar tweets (a form of document) from non-verified accounts do not add to the total credibility" or a more general version "Similar documents from low-quality sources do not add to the total credibility".
<i>Exclusion</i>	Excludes or ignores certain entities from contributing to the credibility propagation, e.g. "A document from a source supporting a claim can be excluded if there is another later document from the same source opposing the claim" ³ .
<i>Miscellaneous</i>	Any other rule not befitting other rule categories, e.g. "Documents referencing each other have generally decreased credibility" ⁴ .

Table 2.2: Categorization of credibility propagation rules with an explanation and an example for each rule.

and thus a non sequitur is found.

The identification of cognitive bias represents one use case for having the credibility graph and all the credibility propagation rules. In general though, it might be interesting to see the end result, i.e. credibility of the claim center to the narrative. Such information is informative by itself, more so if the set of parameters has been constrained on other narratives already. Outside of the use for an individual, it is also possible to imagine using it on experimental base, seeing how different convictions (and with it the parameters setting those convictions) affect the perceived credibility on different narratives.

2.1.4 The issue of usage

The last relevant issue has to do with *how* should the solution discussed so far be used. The assumption here is that addressing the false information right at the time when an individual is exposed to it is the best way to prevent such information taking root and affecting the individual. When considering this, we have to keep in mind how information being consumed and build from there.

Major change in consumption of information has been caused by social media as well as emergence and subsequent growth of podcasts and even more recently the success of Substack offering independent individuals the ability to publish written content while making a living through subscriptions. The underlying driver of these changes is without a question rise of the now ubiquitous smartphone and fast wireless networking. This development drives all kinds of factors such as the possibility to consume audio-visual media everywhere and at any time. On the one hand, TikTok, Instagram Reels or YouTube Shorts push short and easily digestible content which in the information landscape (i.e. ignoring entertainment) leads to additional speedup of the 24/7 news cycle shorter attention spans making it harder for the same audience to consume expansive content. On the other hand, podcasts have conclusively demonstrated the thirst for long-form content where the most popular podcast, The Joe Rogan Experience [38], has an average episode length of more

than 2½ hours [39]. This shows that people are interested in extensive conversations beyond what traditional cable TV or even radio offers. Such long conversations offer plenty of space for nuance.

Smartphones and the related technologies like podcasts and social media have caused a massive shift in the way how information is being consumed. Especially the young generation relies on smartphones, and by extension on social media, much more than the older generations who have a preference for television. In order to address false information holistically, it is important to consider these different channels and what challenges do they bring.

The most relevant challenge for the context of this thesis is the difference in applications being used for the consumption of information on mobile and computer devices⁵.

Whereas computer operating systems and their applications are much more likely to provide a way to modify and extend their many functionalities, the situation on mobile devices is more limited. Consider the situation for arguably the most used application on any device, the internet browser. Apple only recently⁶ added web extensions for the Safari browser that can be downloaded from the App Store [40]. On Android, more browsers have some support of extensions, but it is generally not as rich as on computer browsers and the possibilities are also more limited. This situation will no doubt improve and especially the recent changes make it possible to address the false information with the help of a browser extension. Nevertheless, this represents only a small win in the context of information consumption through mobile devices since most applications are not used through the mobile browser but a dedicated application which is sandboxed from other applications and therefore does not allow any modification or functionality extension. As of now, there does not seem to be a way for external parties to address this issue and it is only up to the application developer and not the user to choose how will they attempt to address false information. It is the belief of the author that in the ideal case the end user has more freedom to choose how they desire to deal with false information and it should be possible for them to choose a third party solution to help them deal with it.

However, since browser extensions offer a meaningful way to deal with false information and browsers are used for consumption of a lot of information, they offer the most relevant solution. It is imagined the proposed solution that would address the problem of false information could use these extensions to show the end users additional information to aid them in the process of informed critical thinking across the web. The imagined usage flow is presented in section 4.6.

As already mentioned, the web and social media consumed through internet are not the only popular channels through which information, and with it false information, is consumed. Other highly popular information channel is television. It is possible to imagine here that in the case of a smart TV an application showing an overlay can be shown, again for the purpose of encouraging critical thinking. Additionally or alternatively, there could be a mobile application showing this information which could also show more information relating to the content seen once it becomes available.

The last point from the previous paragraph is worth closer attention. A narrative evolves over time. While at one point there might be information that is uncertain or that leads to

⁵Mobile meaning smartphones and tablets, generally devices running operating systems like Android and iOS, and computer meaning desktops and laptops, generally devices running operating systems like Linux, Mac OS and Windows.

⁶With the release of iOS 15 on September 20, 2021.

one conclusion, later on a new piece of information can appear that makes the previous information more solid or that changes the perspective. By keeping track of narratives an individual has seen it becomes possible to notify that individual when a narrative develops and the conclusion changes or shifts significantly.

2.1.5 Summary

To summarize the observations and discussion above, CTSS needs to:

- be informative - i.e. be able to provide the relevant context to a narrative
- not be prescriptive
- be collaborative and open-source for transparency
- address cognitive bias
- present relevant information in an apt and timely manner

2.2 Glossary

Definitions for certain terms and abbreviations follows organized by category and ordered alphabetically.

- *Abbreviations*
 - *AFC* = Automated Fact-Checking. Denotes the computer science discipline imitating the manual journalistic discipline of regular fact-checking. Introduced in section 2.1 and further discussed in section 2.1.1.
 - *CTSS* = Critical Thinking Support System. Represents the main contribution of the thesis and denotes the conceptual solution developed throughout the following pages. Reasoning for the naming is given in section 2.1.1.
 - *NLP* = Natural Language Processing. Mostly relevant relevant for data collection and therefore described in section 4.2.
 - *PID* = Persistent Identifier. Uniquely identifies a source for its reliable and efficient retrieval.
- *Attribute* - associated to a node, describing some characteristic of it. Intentionally flexible part of the domain model allowing for an ad hoc instantiation.
- *Credibility* - indicates the extent to which an individual believes a given entity. It can be assigned a numerical value indicating the strength of the belief.
- *Credibility graph* - the nodes (see below) relating to a single narrative also relate to one another in various ways and form a credibility graph. The nodes have credibility assigned to them which can propagate through the graph according to credibility propagation rules (see below).
- *Credibility propagation rule* - rules describing how can credibility between entities propagate in a credibility graph. See table 2.2 for categorization of them including an example for each.
- *Edges* - the nodes in the credibility graph relate to each other in various ways and these relations are captured by edges in the credibility graph summarized in table 4.3. There are edges representing relations that have a stable meaning and one catch-all edge representing all other ad hoc relations.

- *Entity* - the author, publisher and other nodes since they are instances of legal entities, i.e. natural entities (also called natural or physical person) or juridical entities (also called juridical or fictitious person).
- *False information* - relies on an intuitive understanding of what such a term means and it encompasses misinformation, disinformation and other forms of spurious information like hoaxes, myths, fallacies and the nowadays popular term fake news. Even though the intuition of different readers can diverge, it is the belief of the author that individuals from the targeted audience would in majority of cases agree on which instances constitute false information which suffices for the purposes of this thesis. There isn't other uniformly agreed upon alternative term among researchers or journalists. The term is chosen for its intuitiveness, brevity and lack of baggage associated with terms such as fake news [41], [42].
- *Impact spectrum* - defines a continuous spectrum that a piece of information or narrative has on one's life and goes from inconsequential to severe affecting crowd of people. For simplicity, only the single dimension is considered although an argument can be made for two-dimensional spectrum reconciling impact and reach.
- *Narrative* - mostly self-explanatory, synonym to story or account. It was chosen because of its frequent use in the academic sphere. For the context of this thesis, a narrative is uniquely identified by a single claim it relates to.
- *Nodes* - generally anything involved in a narrative. These nodes usually have a distinctive role based on which it can be categorized. They also relate to other nodes where some recognizable patterns can be observed. They also have various attributes, often uniquely relevant only to that nodes. The nodes are elaborated in section 4.1 where they are gradually developed in table 4.2.
- *Reality-based audience* - the audience which is uniquely predisposed to be positively affected by the proposed solution thanks to their open-mindedness and search for the truth. This is the targeted audience for the proposed solution.
- *Rule* - captures the belief affecting credibility of nodes in a deterministic way and has a parameter affecting the strength of its application on the credibility

2.3 Problem statement

Contemporary digital information landscape offers an exorbitant wealth of information which is difficult to navigate and understand. This understanding is further complicated by cognitive bias. The most controversial claims are dealt with by fact checkers who provide more context and information based on which they draw their own conclusion. However, although definitely useful, these fact-checks do not necessarily encourage critical thinking and they also do not consider the problem of cognitive bias. It is important for people not to have misconceptions and to think critically about issues in order to maintain stable and healthy democratic society. This thesis therefore proposes a conceptual solution with the main goal of promoting informed critical thinking while mitigating the effect of cognitive bias on reasoning.

2.4 Research questions

RQ1 What components are necessary for a tool supporting critical thinking?

RQ2 What is relevant data for such a tool and which channels can be used to obtain it?

RQ3 What is an effective way to aid people in informed critical thinking?

2.5 Methodology

Design science is the selected methodology for this thesis. Design science focuses on development and subsequent evaluation of an effective artifact to solve a specific problem within a certain context and provide utility. Design science can be used for addressing unsolved problems [43]. The artifact of this thesis is a conceptual solution named Critical Thinking Support System (CTSS) to aid in critical thinking in order to address the problem of false information. The context is that of a digital information consumer using a personal computer. The unsolved problem this thesis addresses is the lack of widely used system to address the problem of false information.

The approach chosen for exploring the topic at hand is one not adopted before and itself represents a contribution. The treatment of the topic cannot directly rely on existing work and instead develops a new conceptual framework. Afterwards, a primitive prototype of the proposed conceptual framework is implemented.

The evaluation of the result is twofold. First, the conceptual framework is verified, i.e. whether fulfillment of requirements is met. This is done by ensuring that structurally different narratives from the real world can be captured by the framework and different modes of credibility propagation can capture various behaviors. Second, a small set of interviews focused on qualitative evaluation of the utility of the proposed solution are conducted. Given the time window available, conducting a more systematic evaluation with bigger randomized sample size over longer period of time was not feasible.

3 Related work

Considering the given background, there are a few areas of work related to the thesis which relate to the categories identified in table 2.1.

3.1 Exploratory work

3.1.1 Civic online reasoning

[34] is a field experiment research exploring civic online reasoning, the ability to judge the credibility of information on young people's devices. Various tasks have been administered to middle school, high school and university students and 7804 responses were gathered. The report goes into detail for three of its tasks, each for a different school group.

The discussed task for middle schoolers is to identify an advertisement on a news website. While some students show mastery of such recognition, others are less sure or wrong, e.g. identifying a "sponsored content" as not being an advertisement. It should be obvious how this is a relatively straightforward problem to help address if given the ability to annotate content on a website — sponsored content could be annotated or even replaced by the word "advertisement" together with a link to an in-depth explanation.

The next discussed task aimed at high schoolers shows an image on Imgur of an ill-formed flower with the a title of "Fukushima Nuclear Flowers" without any further context. Students are asked whether the image provides a strong evidence about the conditions near the Fukushima Power Plant and explain their reasoning. Given the context of a random internet user (with a curious nickname in addition) posting an image without providing any further context and evidence about its authenticity and location, it is to be judged as non-credible. Students again demonstrate different levels of reasoning. This is much more difficult task, especially considering its potential automation. There are many individual subtasks to be carried out and make sense of which our current technological capabilities are not able to do. It might, however, be possible to identify a claim is being made (as opposed to an image of a cat intended for entertainment) and annotating a general caution in believing it.

The last closely discussed task aimed at college students concerns an reading a tweet and evaluating whether it might or might not be a useful source of information. The presented tweet is from a liberal advocacy organization MoveOn.org, reads "New polling shows the @NRA is out of touch with gun owners and their own members"¹, includes a graphic with a text "Two out of three gun owners say they would be more likely to vote for a candidate who supported background checks" and also contains a link to a press release by the poll's sponsor, the Center for American Progress, another liberal advocacy organization. The graphic as well as the linked press release indicate the poll, which main finding is presented in the graphic, was conducted on 816 gun owners by Public Policy Polling, a U.S. private polling firm affiliated with the Democratic party. Mastery is shown by both arguing for its usefulness since it is based on polling data, as well as on acknowledging the bias introduced by affiliation of all three entities to liberal ideology and by extension to the Democratic party and stronger gun control laws. While an end-to-end general evaluation of such situations is difficult, it is at least comparatively straightforward to identify the entities involved in such a claim and collect basic background information about them, such as their political views and affiliations. It might also be possible to identify the topic

of discussion and generalize that a certain political bias predicts a certain opinion on the matter, i.e. liberals are generally against guns and for stricter gun control measures, which is indeed the case in this scenario and makes the reached conclusion less important.

3.1.2 Importance of lateral reading

[45] studied how professional fact checkers, Ph.D. historians and Stanford University undergraduates fare in three tasks evaluating credibility of information online. The main finding was the difference in behavior of the fact checkers compared to both historians and Stanford undergraduates. Official looking sites with well-crafted logos and official looking domain names often tricked the latter group. Fact checkers, in contrast, employed lateral reading, i.e. leaving a site after a quick scan and searching to judge credibility of the original site. This was observed to be the crucial difference that allowed fact checkers to arrive at warranted conclusions in a fraction of a time.

One of the tasks was to assess the reliability of information from two groups: the American Academy of Pediatrics (AAP), the largest professional organization of 67000 pediatricians in the world, and the American College of Pediatricians (ACPeds), a socially conservative advocacy group of around 500 pediatricians splintered from AAP that is ideologically opposed to LGBT. Both historians and students were deceived by, among others, the sources used on ACPeds article and they cited ACPeds as more credible and reliable compared to AAP. The fact checkers employed lateral reading and spent the least amount of time on both the ACPeds and AAP websites and searched other information about the sources which helped them figure out that AAP is the more reliable of the two.

This work points to an importance of lateral reading that is then to be encouraged, a fact relied upon in the thesis.

3.2 Aiding fact-checking

There are a few approaches going in the direction of automated fact-checking while already being helpful to human fact checkers. Since fact checkers have limited bandwidth and tackle the false information at a slower pace than at which it comes, anything that makes fact checkers' work more effective is a success. Let's consider the fact-checking process to be broadly composed of stages for (1.) monitoring news and other information, (2.) detecting claims, (3.) checking these claims and (4.) finally creating and publishing fact checks regarding these claims [46]. [47] focuses on aiding the second stage, i.e. claim detection. Another fact checker, Africa Check, uses an automated video creation platform [Lumen5](#) to create short clips for their fact checks and publish them on their [YouTube channel](#) [48]. This approach and the chosen platform therefore aids in the fourth stage of the fact-checking process.

This work has an implication for data collection as the task of claim detection is an important part of an automated solution. An automated video creation points out the reality of today where video, especially short-form video, is a dominant media type, fact that is important to consider for successful dissemination of correct information to counter the false information.

3.3 Automated fact-checking

3.3.1 Surveys

Surveys on dealing with "fake news" provide a good overview of the tasks that the research community is interested in. They examine the area and the existing research

¹NRA being the National Rifle Association, a major gun rights advocacy group of 5.5 million members based in the United States [44].

using different perspectives, most of which relates in one way or another to AFC. [49] surveys the existing research using four perspectives to detect false information: by the false knowledge it carries, by its writing style, by its propagation patterns, and by the credibility of the source. [50] considers the core fact-checking tasks to be document retrieval, evidence extraction, stance detection and claim validation. [51] approaches the area with an interest in proactive intervention strategies and dynamic knowledge bases. As can be seen from these examples, the research approach can be quite varied.

It is important to keep in mind that AFC has a different goal than the one pursued within this thesis, the distinction that most closely related to the issue of truth prescriptiveness described in section 2.1.1. However, the tasks studied in the broader research relate to this thesis since some of them are useful for automating the data collection.

Datasets

These surveys also list the available datasets for the tasks studied. Different datasets are applicable for different tasks, examples of which are fake news detection (mostly binary and in a few cases multiclass classification), rumor classification, fact extraction or stance detection. [50] discusses the problems with available datasets regarding their size, the relevant domain (often too narrow such as only sourced from Twitter or Wikipedia or concerning only politics) and unavailability of relevant metadata such as the related documents, evidence and its source, stance of the documents and evidence, and lastly the inter-annotator agreement.

3.3.2 Relevant tasks

The tasks, many of which rely on NLP, encountered in the existing research relevant to this thesis are listed below. Some general grouping as to what the task is largely relevant to is applied:

- Author (i.e. legal entity) verification
 - Automation detection - distinguishing between a human, bot or cyborg (hybrid)
 - Bias estimation - can be of different kinds, such as political or occupational, found by association to other entities and relevant for identifying characteristics of an entity
 - Diffusion role categorization - distinguishes between intentional spreaders such as spammers and trolls, unintentional spreaders, i.e. unaware victims, clarifiers pointing out falsehoods and persuaders trying to change beliefs.
 - Dubiousness detection - marking dubious entities by association with spreading or creating false information
 - Echo chamber and filter bubble effect estimation - measured by association with other entities and representing an important characteristic of an entity
 - Identity identification - simple such as verified vs. unverified social media account, or more complicated, multiple linked identities for a single person such as mother on Facebook, entrepreneur on Twitter as well as a policewoman on Twitter.
 - Reliability estimation - e.g. a metric measuring an association with verifiably true information
- Claim or statement
 - Claim detection - important for monitoring and subsequent tasks of grouping

- Claim similarity grouping - relating similar claims into one group
- Document
 - Clickbait detection - especially those with misleading intent
 - Document retrieval - provides implications for data collection
 - Metadata retrieval - such as that of an author, publisher, time of creation, etc.
 - Satire detection - a very distinct category of content important to be ignored from serious consideration
 - Stance detection - relevant for automatic detection of support or opposition towards a statement as well as identification of bias
 - Topic identification - useful for grouping, could be a subset of the more important claim similarity grouping task
 - Virality detection - important for correct application of "innoculation" measures slowing down its virality and mitigating its negative consequences

3.4 Related journalistic work

Interesting and relevant work is also done in journalism outside of the fact-checking realm. Organizations like AllSides and Ad Fontes Media rate media sources based on their political bias and in the latter case based on reliability as well. Such a work is useful, just consider the task described in section 3.1.1 given to the college students which involved consideration of bias of the involved entities. AllSides is also interesting from the perspective that it presents stories covering the same narrative from different sources with various political bias. These stories give a different point of view on the matter, present different arguments and as such can promote a more careful consideration, or in short, informed critical thinking.

These ventures represent an interesting source of information affecting credibility of entities as well as give an encouragement with their selected approach emphasizing being informed from multiple differently biased sources.

3.5 Active intervention via tool

One of the most apt and popular means for an active intervention to aid the problem of false information in one's daily life is through a usage of browser extensions. Searching Chrome Web Store for extensions relating to bias checking, "fake news" detection or news evaluation reveals a large number of such solutions. Probably the most popular one is called NewsGuard created by NewsGuard Technologies founded recently in 2018 which rates the credibility of news and information sources and provides this information in form of a "nutrition label" explaining the given rating [52]. It focuses on media sources in US, Germany, France, Italy and the UK, allegedly covering 95% of the online engagement with news. This solution also partially relies on their users flagging potentially false stories and with it improving the service [53]. Their browser extensions are also available for Microsoft Edge, Safari and Firefox. It is also included in the mobile version of Edge allowing users to use it outside of desktops. There are also namesake mobile apps for iOS and Android.

This points out the possible approach to take for effective and active intervention in users' everyday lives.

4 Conceptual solution

Recall the summary in section 2.1.5 that serves as a useful guide for the development of a conceptual solution in this chapter and it influences how it is structured.

The first point is for CTSS, the proposed conceptual solution, to be informative. Recall the goal of aiding in *informed* critical thinking and the argument made earlier for why critical thinking by itself is not enough without all of the relevant information. In order to represent this information, a domain model needs to be developed (section 4.1). Afterwards, it is important to consider how to get the real-world information into such a domain model to use it (section 4.2).

Once the domain model is in place, attention can be given to credibility graph and consideration of the different scenarios that can arise in the domain model and how do they map onto and propagate in the credibility graph (section 4.3).

At this moment, with both the domain model and credibility graph, the scene has been set up to look into addressing the problem of cognitive bias (section 4.4).

Although the preceding section reference some of the ideas from the collaborative nature of CTSS, more ideas are further developed and summarized in section 4.5.

Lastly, the question of how could CTSS be used in the wild, considering the various real-world constraints, is discussed in section 4.6.

4.1 Domain model

Some parts of the domain were outlined already in the previous chapters. The following is a more systematic treatment of each part.

The only way, except for in-person conversation, how people get information is that they consume it in some media form. This could be in a written form but also audio-visual form¹. These forms could be combined in various ways. There is a huge variety in terms of specific media types these forms could take:

- **written** - e.g. a printed book, newspaper or a journal as well as electronic versions of the former together with a blog article, social media post or a message from instant messaging app
- **audio** - e.g. a radio, podcast, voicemail or an audio message²
- **visual** - e.g. a printed illustration or a photograph as well as digital image like computer graphics, screenshot or a meme.
- **audio-visual** - e.g. a recorded or an animated video, a movie.

These forms when presented to a consumer come in some unified package that will be denoted as a document. This is therefore not a document in the commonly used sense of

¹Using the commonly cited (and misconceived) Aristotle's categorization of senses, sight and hearing are the most important senses for information consumption. However, blind people can learn Braille and use their sense of touch to read which would therefore allow them to consume information in a written form as well, apart from an audio. For simplicity's sake though, such complications are omitted from consideration here.

²Voicemail is specifically related to cell phone services while audio message is more diverse and nowadays mostly used through social media services like WhatsApp or Facebook.

the word and for example none of the definitions for "document" given on Merriam Webster truly fit [54]. The definition given on Lexico is more general and works well:

A piece of written, printed, or electronic matter that provides information or evidence or that serves as an official record. ([55])

These documents make statements, which is:

A definite or clear expression of something in speech or writing. ([56])

Such a definition works well for explicitly articulated statement. Except not all statements are that clearly and directly expressed, often making only an implicit statement. Phrases such as "we all know who's behind it" are ambiguous but in a certain context can be understood as an implicit statement about specific entity and leave no doubt as to what was meant. However, such a statement is enigmatic and difficult to parse for anyone not within that context. This makes the relationship between documents and certain statements often unclear and while it is possible to say that a particular document exists in the objective reality, which statements it makes can be more subjective.

The documents, and by extension statements as well, relate to certain topics as well. This is an intuitive everyday abstraction that helps in communication and here it is helpful for generalizations. For example, while one can be interested in the topic of celebrity gossip, and even though an epitome of false information³ still putting a lot of trust in it, the same person might not be interested in political topics, especially from the political party disliked by them. Another person might be a polar opposite. In such a case, the generalization through a topic can be useful since other generalization of the same broad extent through an author or publisher might not be possible.

All these documents are also created by someone or something⁴ and published somewhere. For example, person writes an article, with it becomes its author, and publishes it on Medium, an online publishing platform for social journalism [59]. There could be more authors contributing to a document. A document can be published by multiple publishers. The authors could be part of an organization or a company, in general a so-called juridical person as opposed to a natural person. In some cases, the author and publisher can be the same, such as when an organization publishes a document on their own site and no other author is mentioned. [60] is an example of such a document.

The document serves as a record, as given in the definition above, and can therefore be retrieved from one of its sources. This therefore limits the set of documents to be those for which a source is available so an unrecorded phone call or private message that cannot be retrieved and was not leaked does not count as documents. This source needs to be referenced in order to be able to get the underlying document.

The best form of a reference is a persistent identifier (PID) since it allows for a reliable and efficient retrieval of the referenced document [61]. Examples of such PIDs are Digital Object Identifiers (DOI) often used in academic sphere, International Standard Book Number (ISBN) used for books or BitTorrent magnet links. An ordinary URL also identifies a document but there are two problems associated with URLs that make them unreliable:

- **Link rot** - denotes the problem when a previously accessed web address goes dead — the content has been moved and its address has changed, it has been removed

³Wikipedia concurs since the gossip magazine article contains a reference to an article on disinformation [57].

⁴Automatic generation of certain articles is being done for at least the last 6 years. [58].

or the whole website is no longer live.

- **Content drift** designates a problem when the document changes over time. So while the URL still retrieves a document, it might be different and it is impossible to state just how different it can be. The changes can be visual, such as in the case of website redesign, supplemental, for example when an article gets updated with further information, corrective, when a piece of information in the original article is corrected (good practice is to leave note at the beginning or an end of such article about the correction), but it can also be manipulative and deceptive and the document can be changed to such a degree that it is unrecognizable to its original form.

Some websites maintain themselves a permalink system intended for long-lasting references, examples of which are Wikipedia or StackExchange. There are also web archiving services that create a Persistent URL, or PURL, such as Internet Archive's Wayback Machine, WebCite, archive.today or perma.cc.

These sources, which inherently relate to publishers, can be partially assessed objectively. The practices and initiatives they undertake as well as historical accuracy all play a role. For example, while Twitter and Facebook are two platforms and publishers (even though they do not call themselves as such [62]) that are involved in a lot of false information dissemination, they at least have initiatives to curb this problem [60], [63]. The same cannot be said about other sources, such as Gab, the alternative social media platform for the far-right [64]. Such initiatives make the former sources objectively better from an editorial quality perspective. Another question is to what degree is this editorial quality important which gives us another instance for generalization rule, i.e. "I do (not) put weight on editorial quality and do (not) consider sources with more practices and initiatives for improving it as better".

Intuitively speaking, authors, publishers and communities are at some point represented by some people. These people could actual humans, or natural entities, but they could also be part of an organization or a company, or juridical entities. Communities could be a collection of both natural and juridical entities. In this domain, all of these are a form of entity. This is useful, since these different entities can have various relations between each other. For example, an author could be affiliated with an organization. This organization is owned or controlled by someone. When this affiliated author writes about the owner, this relationship gives us additional context to work with, namely there exists some suspicion of bias and that the author did not give a fully independent account of the owner.

Now we have in place the most important parts contributing to the dissemination of information that could be objectively assessed, given some caveats regarding ambiguity. Everyone in the reality-based audience that this thesis works with can agree on these parts: "Yes, there is this document with this content, concerning these topics, from this author, by this publisher, appearing on this source. It definitely makes these statements and it is possibly making these other statements as well". Let's move onto a part of most importance, the subjective credibility. Credibility means:

The quality of being trusted and believed in. ([65])

The credibility of a document and its statement is affected by credibility of the provided evidence. This evidence can be both in support or in opposition to the statement. The evidence also comes in a form of a document with all of its associations to sources, statements, authors and publishers which have their own credibility. What represents suitable

evidence is subjective, both (1) from the perspective of whether it is even relevant to the statement in the first place and (2) whether it comes from a credible source. While the second relation concerns credibility which is explicitly modelled and can therefore be directly expressed in the domain model, the first relation deals with an uncertainty as in the case between documents and ambiguous statements.

To demonstrate on a simple example, let's say we have a claim that COVID-19 vaccines are deadly. While a numerical data point of significantly high value based on high enough sample size stating how many people died in the following week after taking a vaccine would be considered a suitable evidence, an anecdote about a senior dying after taking a vaccine would not (1st relation). No matter what source the anecdote comes from, it is statistically insignificant and no overarching conclusion can be drawn from it. The credibility of evidence given by the numerical data point depends on what source does it come from (2nd relation). The source might be relatively unknown or it could be an impostor of an otherwise established and credible source and once assessed as so be deemed not credible.

See the domain model in a simplified UML diagram in fig. 4.1 and the entity detail in fig. 4.2.

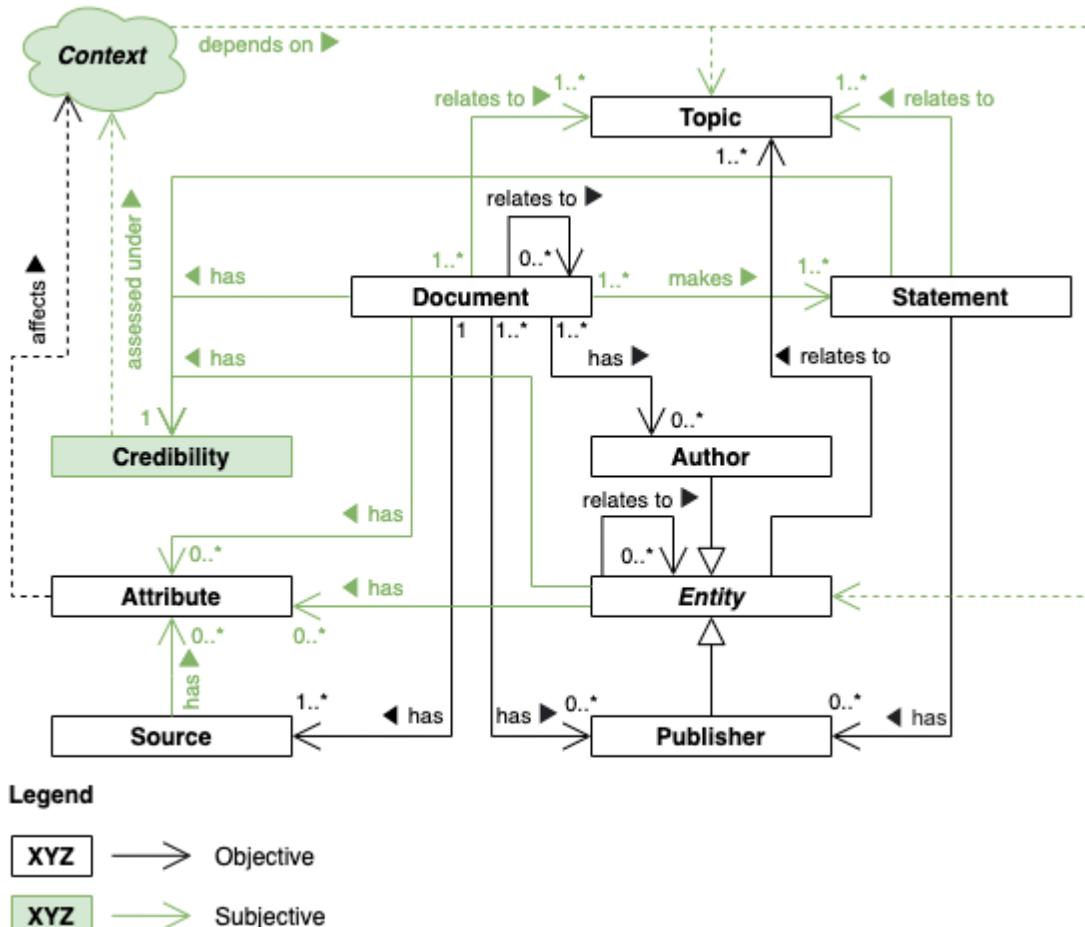


Figure 4.1: Domain model using a UML class diagram. Note the legend in the bottom left corner. For simplicity's sake, complexity around the abstract *Entity* has been omitted from this diagram and can be seen in fig. 4.2.

Notice that *Entity*, *Publisher* and *Author* are *abstract*, whereas *NaturalEntity* and *Juridi-*

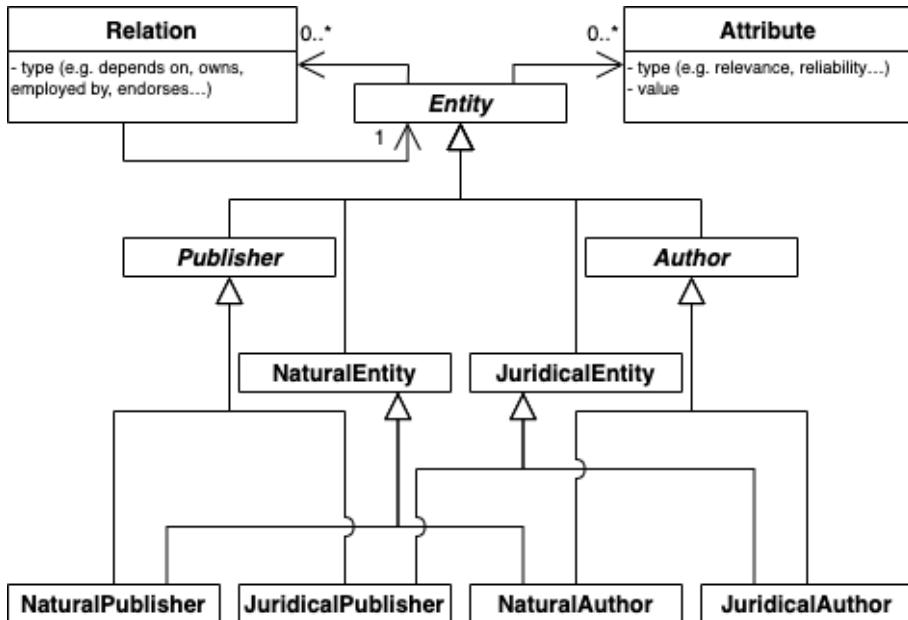


Figure 4.2: Domain model - *Entity* detail.

calEntity are not. This allows for existence of an entity which is not necessarily an author or a publisher, such as the aforementioned owner of an organization.

The individual classes in the domain model also have various attributes that are not shown on the figures. In general, any attributes relevant to that class could be associated. From the obvious, like author's name, to the nonobvious, like author's number of Twitter followers.

4.1.1 Domain model used on narrative

Let's see how does the narrative example from section 1.1 fit in the domain model.

4.2 Data collection

Now that we have an idea about the domain model it is time to consider how could it be filled with real-world data.

Since the number of narratives already out there and also of narratives started every day is absurdly high and even teams of cumulatively hundreds of people, namely fact checkers, struggle with making sense of it all, a different approach must be taken. Various parts of the data collection pipeline can be automated to a different degree and the missing parts can be aided by community contribution.

In the long run, the community contribution should be replaced by further automation except in places where it is technically not feasible due to technical reasons. For example, private chain messages on WhatsApp, Telegram or Facebook Groups are impossible to monitor comprehensively — even the platforms themselves essentially have no remedy that would not result in privacy violation or disabling a feature beneficial for other use cases. Any attempt to tackle such an issue will by definition lead to an incomplete resolution.

It is also important to keep in mind that collection of data for addressing false information is and forever will be constantly evolving process for at least these three reasons:

1. False information itself will evolve and new approaches and information will be needed to address this evolution.
2. Flaws and opportunities for improvement in the established data collection process will be found and will need to be addressed.
3. New technologies and approaches for new and more accurate data will be implemented.

With these caveats out of the way, let's plunge into automated mining in section 4.2.1 and afterwards consider the contributions that could be made by the community in section 4.2.2.

4.2.1 Automated mining

Let's first consider what fact checkers do as part of the whole fact-checking process. This serves as a useful guide for plenty of ideas for automation.

At first, a relevant narrative to focus on needs to be identified. The problem therefore becomes to get a good, ideally complete, picture of current narratives by monitoring news and other sources of information as well as choose the most relevant narrative to work on, one that will possibly have the highest impact in terms of how many people it will affect and by how much. This is the case for a human fact checker due to time constraints. An automated system with a processing capacity much higher than that of a person is not limited in such a way and does not need to select narratives based on relevancy. However, the relevancy might still be useful to know about for the end user and the community.

Second, relevant information for the identified narrative needs to be found and retrieved. One part of this is to find what is already out there as the narrative might have started in one way and continued to evolve with more (false) information gradually accumulated. The second part is to identify other sources of information to either support or oppose the claims made in the narrative. The first part is only complicated by not being able to access closed communities such as private Facebook or WhatsApp groups and therefore not being able to see whether such a narrative is discussed there and in what way. However, just a simple keyword search, where keywords could be unique terms relating to the narrative but also e.g. a URL of a tweet that started a particular narrative, can retrieve plenty of context. The second part is arguably more difficult as different narratives require information to be retrieved from varying sources. The fact checker with their highly flexible skills understands the narrative and is able to identify the claims that need to be checked as well as the possible approaches for how to do so. This can take many forms as demonstrated further.

Using the example from section 1.1, the most relevant information to be checked there is about the buses. There are many questions from which the query could be started and the following is just one example which closely mimics how the story actually developed:

- **Q:** Whose buses are they?
A: The bus company Coach USA North America.
- **Q:** What does the owner (Coach USA North America) of the buses state?
A: Sean Hughes, director of corporate affairs for Coach USA North America, states that at no point were their buses involved in the Austin protests.
- **Q:** What were they then involved in?
A: A big software conference with attendance above 10000 people.
- **Q:** Can this be confirmed?

A: Yes, the spokeswoman for the company organizing the conference issued a statement confirming they hired the buses for transporting people to the conference.

- **Q:** Can the original claim be confirmed?

A: No concrete proof of evidence is apparent.

Whereas answering these questions required getting in touch with relevant entities or at least monitoring whether these entities state anything related, a completely different approach might be necessary for a different narrative. Suppose a narrative revolves around a numerical claim though, e.g. nowadays relevant COVID-19 related claims. Such claims might be possible to be verified by checking against data from national and international statistical services.

The third part has to do with compiling all the information and making sense of it in such a way so that it is easily presentable and digestable for the general public. The role of presentation is taken by visualization using credibility graph that is discussed in section 4.3.

Let's have a look at possibilities for data collection automation for the individual parts.

Get currently relevant narratives

Consider some of the most relevant sources of narratives⁵:

- Daily newspapers and online news - New York Times, USA Today, Breitbart
- Blogs - Crooks and Liars, The Gateway Pundit
- Social media itself - Twitter, Snapchat, Reddit
- TV - CNN, NBC, Fox News
- Radio, podcasts - New York Times' The Daily, NPR's Weekend Edition, The Ben Shapiro Show

Since for start, we want to only recognize the themes and keywords of narratives, getting the information from aforementioned sources in ordinary textual form is acceptable and higher dimensional methods can be ignored. This self-imposed limitation also allows for using more established methods working with text rather than relying on more experimental solutions. Let's consider how can we get the mentioned sources into a textual form using NLP techniques.

A lot of previously mentioned sources offer RSS feed (e.g. [New York Times](#), [Crooks and Liars](#) or [The Gateway Pundit](#)) or an API (e.g. [Twitter](#) or [Reddit](#)) for consumption. For other websites, web scraping is a possibility. It might more efficient to use an aggregator which might offer a more unified experience for data mining. These methods, especially web scraping, generally cover textual publicly accessible sources. These sources as well as purely image driven communities might have documents accompanied by images from which a text can be extracted by optical character recognition, where the state of art is represented by [Google Tesseract](#) [66], or the image can be summarized by automatic image annotation [67], [68]. As for audio media like radio and podcasts, some transcripts of varying quality exist and it is possible to apply general speech recognition itself for getting more of such content [69]. As for the mixed media like TV, a simplification is afforded by treating it as audio or image content only and subsequently applying the methods from

⁵An attempt has been made to order the examples for the different categories (except for social media which vary by content creator) by their political bias by checking against [AllSides](#) and [Ad Fontes Media](#) bias ratings pages. Political sources have been chosen as they are among the most popular but other non-political sources could have been chosen to serve as an example as well.

above — such an approach, in the case of TV news, would be enough to identify the spoken words of news presenters and news headlines.

Once we have these various documents identified, they need to be aggregated by a narrative they relate to. There are many possible approaches but the general process would include text vectorization with a subsequent grouping by nearest neighbor search. This could be interlaced by text classification in order separate the content to a few broad categories. An example of this can be seen in [70]. Other approach more in line with our definition of a narrative is to identify claim(s) being made as in [47] and aggregate based on those. The aggregated sources could then be summarized into a single headline [71], preferring a format of a claim.

As for measuring the aforementioned relevancy to a general populace, an engagement on social media and the general interest through e.g. Google Trends could be measured for the identified keywords and topics and serve as a proxy for the relevance.

A simple approach that is rather reactive and dependent on other credible sources (mainly fact-checkers) is to collect data based on narratives covered by these sources.

Get relevant documents to an identified narrative

Although we might already have multiple documents identified from the previous step, this step used some limited amount of sources to identify a narrative. Once we know a narrative, we can use it to find other relevant sources. Specifically, we can use the summarized name for the narrative, headlines from the already identified sources and their URLs and query other sources through a search engine. A high-level overview of this process can be seen in fig. 4.3.

This most likely results in an explosion of content found and some reasonable filtering would need to be applied. On the other hand, it might be helpful to do such a query recursively, that is to repeat the process once we have found new sources, since there might be a thread of information that can be found only with this recursive search.

Get supporting and opposing evidence

Based on these relevant documents, a retrieval of evidence in both support and opposition to the claim of a narrative can begin. As already mentioned, this is a difficult task since different narratives require evidence from different sources. One issue is to know when to use which source and the other is to actually use it. It might only be possible to identify potentially relevant sources of evidence based on the already identified documents relating to the narrative but this is a difficult problem to generalize and human assistance, such as one from the community contribution, is needed. While some sources are amenable for automation, e.g. the aforementioned data from statistical offices, others, like getting in touch with a relevant person and asking for a statement relating to a narrative, are not.

However, automated systems can help with metadata retrieval that is comparatively easy to do for computers and often not done by human fact checkers, possibly because of lack of skills, knowledge or foresight.

Obvious metadata to retrieve are name of authors, publishers and timestamps for a document. There might be a profile or an about page for authors and publishers within the same page or social media page where especially LinkedIn should not be left out. Such pages can provide various context including biography/lifetime details or self-proclaimed opinions.

However, more interesting and often disregarded metadata can be identified by learning from the evolving false information landscape and especially the past successes of false information. The following lists first point out some past successes of false information

INFORMATION TYPE PIE CHART

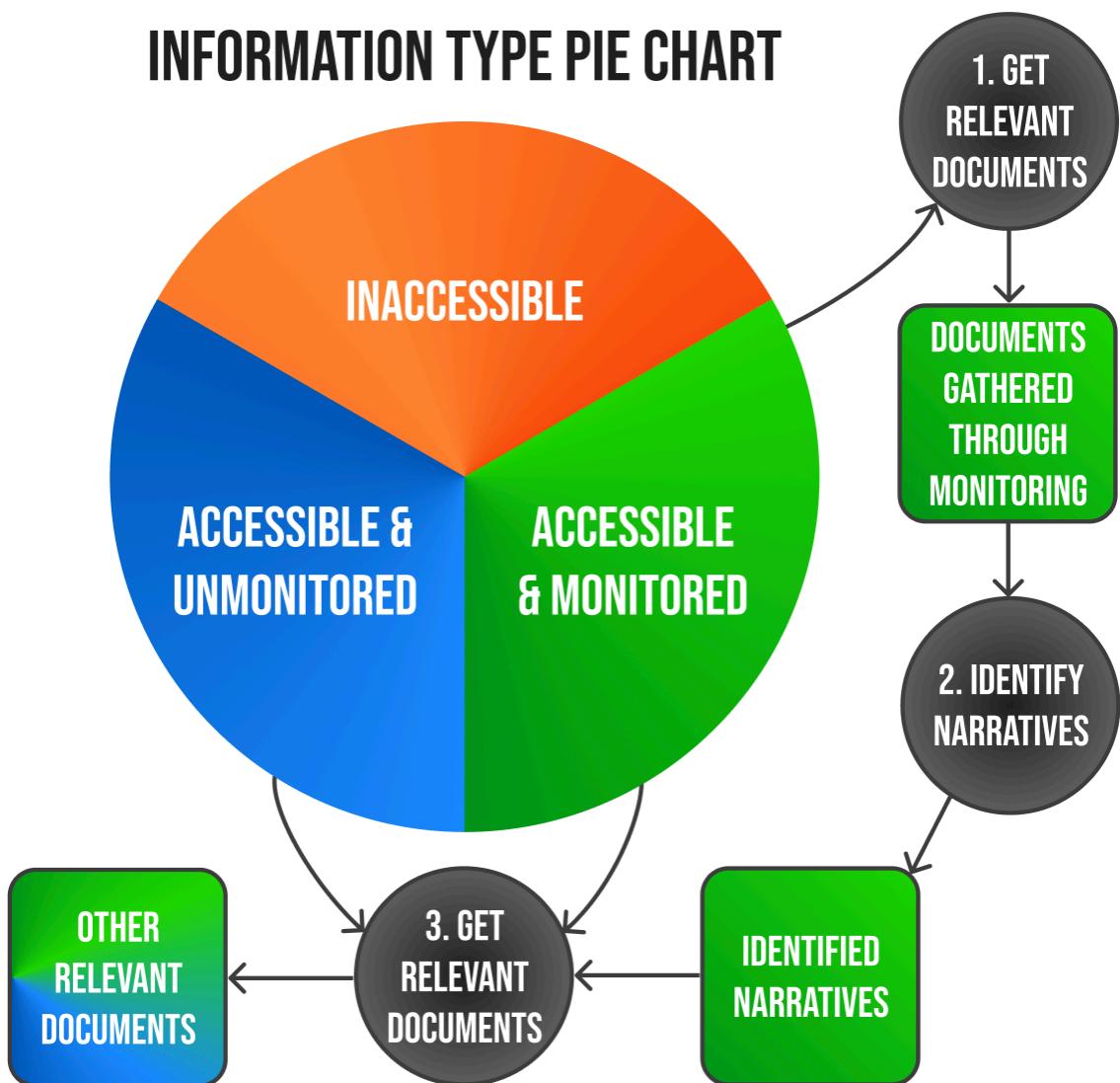


Figure 4.3: High-level overview for getting relevant docs. In the set of all the information out there, part is inaccessible and from the accessible part, only certain subpart is easily monitored using sensible amount of effort and the rest is searchable. First a set of documents is gathered through the monitoring from the "accessible & monitored" part, narratives are identified from these and other relevant documents to these documents are searched across all of the accessible documents. The individual parts are not necessarily to scale.

spread and the other discusses various ideas to detect such abuse using automation that would then decrease this site's credibility:

1. Fraudulent website impersonating an established one:
 - (a) abcnews.com.co instead of abcnews.com [72]
 - (b) fraudulent polling website impersonating IPSOS [73]
2. Website taking on names that sound credible - National Report, World News Daily [72].
3. Seemingly legitimate website judged by its design which however contained only a single fraudulent article and everything else was a carefully designed facade leading to nowhere.
 - Domain name check - can be easily automated and one of the most used browser extensions previously mentioned in chapter 3, NewsGuard, relies on this as it has a database of the most popular sources representing majority of the traffic in the information dissemination landscape. For the particular use case of abusing similar domain names (item 1 above), an appropriate string similarity metric can be calculated against such a database and reported as abuse if under certain threshold.
 - WHOIS query - helpful for understanding who is behind a site as well as the temporal characteristics of the site. This particular strategy was used in investigation of the IPSOS impersonation (item 1b above). The query might be blocked by WhoIsGuard in which case the most relevant piece of information still available is the date of registration for such domain. If a domain has been registered recently, the chance it is used for fraudulent purposes rises⁶. The maintained database of news sources and other websites could be expanded with data about their date of founding, place of origin and associated URLs and social media accounts that are all potentially available from a WHOIS query and provide additional context when estimating the objective measure of credibility.
 - Checking Wikipedia - many major sources of information, as well as various fraudulent and misinformation sources have their own page on Wikipedia. Checking for an existence of a page on the site and potentially analyzing what it says and fetching it in a structured format could provide some initial context. This could only be relied upon as an additional check since such a strategy would otherwise be too vulnerable for an abuse. However, it is a good cross-check and prevents too high of a reliance on the ad hoc maintained database.
 - Search engine query - could be performed with the site name and checked whether it appears on other verified sites from the maintained database of sources and in what context.
 - Crawl site to estimate amount of content - A site could be crawled when encountered for the first time to answer a question "How many articles are on this site?". The answer would provide the necessary context to prevent single article site frauds (item 3 above).
 - Check the About page - also a very naive strategy that could not be exclusively relied upon as the site creator is in control of it and can write whatever they want. However, in some cases, About pages are surprisingly elucidating and reveal helpful context. This should obviously still be verified against other sources.

- Social media statistics - this encompasses statistics for a particular article (number of engagements, e.g. on Facebook [74]) or entity. This provides information about popularity which is relevant for assessing the potential for virality.

4.2.2 Community contribution

While automation can help with data ingress, making sense of all of this data is comparatively much more difficult and a task by and large still only suitable for human reasoning. This is where a community, of both professional fact checkers as well an ordinary consumer, steps in.

The automation with the currently available technology can monitor information sources, aggregate information broadly by topic or less accurately by a narrow narrative and provide context not easily accessible for humans but fairly straightforward to obtain using automation. The community can first of all discuss and suggest improvements to this process, even directly as source code. They could also provide data that the current data collection pipeline did not recognize. Such manual entries would be flagged as opportunities for improvements and should be gradually substituted by automated solutions. However, applying their highly flexible human capability for reasoning to connect relevant entities and make sense of the data is the more important part. It also provides the opportunity to audit this process and use the collected data as a basis for an automated system that could gradually take over even from these community contributions.

4.3 Credibility graph

With the domain model in place, the credibility graph idea can be conceptually developed to be more concrete. This section develops various ideas for credibility graph in two steps and uses the narrative example from section 1.1. Section 4.3.1 introduces basic ideas for structure of credibility graph, setting initial values for leaf nodes, propagation of credibilities within and discusses some issues with it. This leads to section 4.3.2 which addresses the identified issues in two subsections.

4.3.1 Credibility graph introduced

Figure 4.4 shows the first iteration of a credibility graph for our narrative. This iteration considers publishers, authors, documents and statement. The only supported connections are from authors to documents, from publishers to documents and from documents to a statement. Only the latter can take a form of either support or opposition. The former, connecting both authors and publishers to documents, are neutral. This is summarized in table 4.2 and table 4.3 which however already include more considerations considered only later in section 4.3.2.

The following formalizes the calculation of credibilities for nodes in a credibility graph. Refer to table 4.1 for an overview of used variables and their meaning.

Figure 4.4 shows a credibility value from 0 to 1 for every node. The value of most interest is that of the statement. The way to obtain it is from its related documents. These documents in turn obtain their credibility from their related entities, i.e. authors, publishers and other legal entities. These entities have by default $Cred_{\text{default}} = 0.5$, i.e. the middle value from the interval $[0, 1]$ which is the codomain of all credibilities. This default credibility is further affected by its associated attributes and rules. These rules can be set by the user. By clicking a node, the attributes and rules appear as additional nodes connected to the node that was clicked as shown in fig. 4.5. The rule node represented by a diamond shape has a slider next to it that allows user to change its strength.

⁶The same logic can be applied in other domains, such as for automated Twitter account detection.

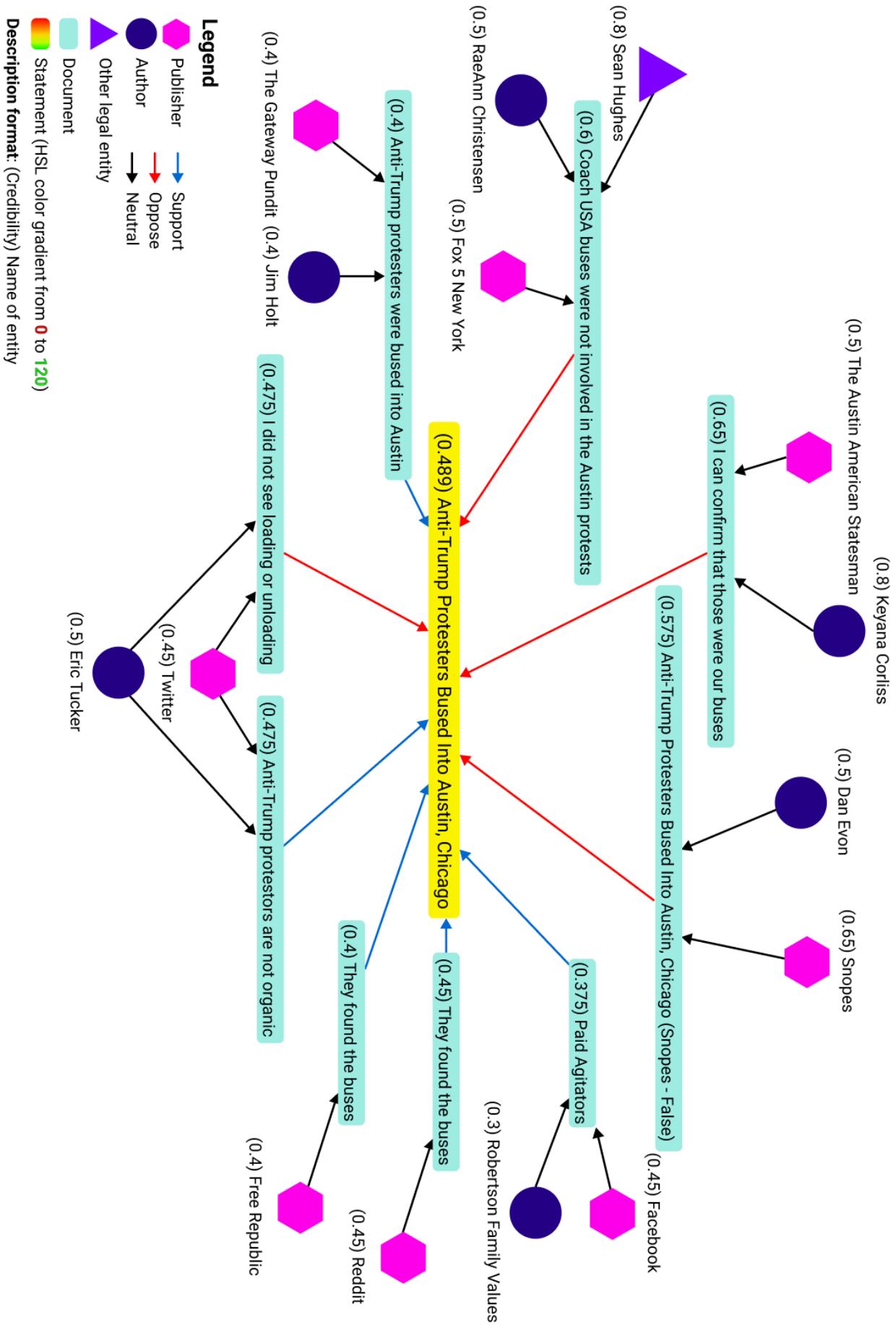


Figure 4.4: Simple credibility graph for a narrative described in section 1.1. Note the legend in the bottom left corner.

d	A document	e	An author, publisher or other legal entity
s	A statement	n	A node (strictly speaking, excluding statement node)
r	A rule	a_r	An attribute associated to rule r within an interval $[0, a_{rmax}]$
a_{rmax}	Maximum value an attribute can have. It is chosen based on the seemed importance of an attribute.		
The codomain of all Credibilities is $[0, 1]$			
Cred_d	Document credibility	Cred_e	Entity credibility
Cred_s	Statement credibility	$\text{Cred}_{\text{default}}$	Default credibility 0.5
D_s	Set of documents relating to statement s	E_d	Set of entities related to document d
R	Set of rules	A	Set of attributes
T_n	Set of tuples (r, a_r, a_{rmax}) for node n		
In the following, we can ignore rules where $r = 0$ since they do not have an effect anyway.			
$T_{n\text{positive}} \subseteq T_n = \{(r, a_r, a_{rmax}) \mid r > 0\}$	Only those tuples that have positive rules		
$T_{n\text{negative}} \subseteq T_n = \{(r, a_r, a_{rmax}) \mid r < 0\}$	Only those tuples that have negative rules		
$D_{\text{opposing}_s} \subseteq D_s$	Set of opposing documents related to statement s		
$D_{\text{supporting}_s} \subseteq D_s$	Set of supporting documents related to statement s		

Table 4.1: Variables and their meaning needed for section 4.3.1

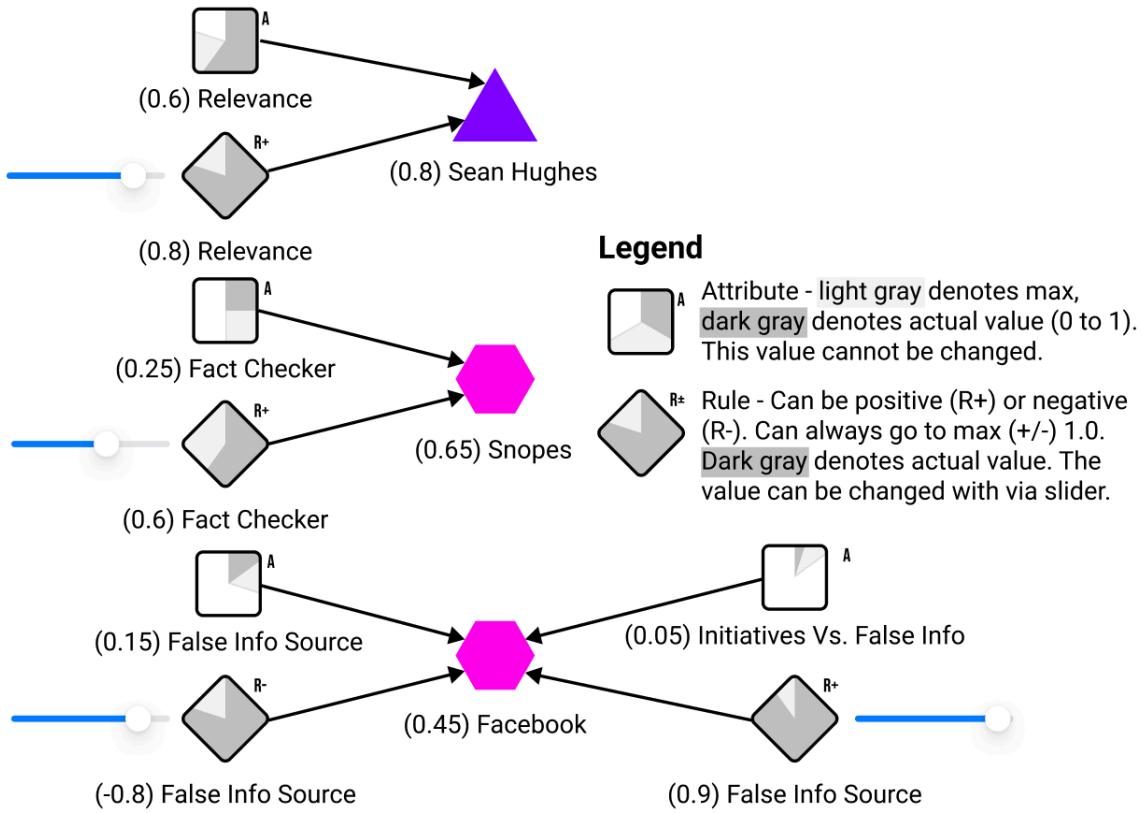


Figure 4.5: Node attributes and rules appearing when a node is clicked.

A rule r can be either of positive nature (R^+), such as being a fact checker, or negative nature (R^-), such as being a false information source, as seen in fig. 4.5. A positive rule is in the interval $[0, 1]$ and negative rule in the interval $[-1, 0]$ where 0 means it is disabled in both cases. The basic interaction rule r and its associated attribute a_r have is $r * a_r$. This product is added to $\text{Cred}_{\text{default}}$. This can, however, result in values > 0 or < 0 and need to be normalized. In order to simplify the resulting formula, it is split in in three parts, the default credibility $\text{Cred}_{\text{default}}$, credibility contributed by the positive rules Cred_{R^+} and credibility contributed by the negative rules Cred_{R^-} .

The credibility contributed from positive rules Cred_{R^+} is calculated by iterating through the set $T_{n\text{positive}}$ where a sum of the products $a_r * r$ is calculated as well as a sum of the maximum values the attributes can take. Assuming it is always the case that $\text{Cred}_{\text{default}} = 0.5$, then it is enough to divide the two sums and additionally divide by 2 so that the range of possible values is halved to $[0, 0.5]$ so that $\text{Cred}_{\text{default}} + \text{Cred}_{R^+} \leq 1$. Therefore,

$$\text{Cred}_{R^+} = \frac{\sum_{(r, a_r) \in T_{n\text{positive}}} a_r * r}{2 * \sum_{a_{r\max} \in T_{n\text{positive}}} a_{r\max}}. \quad (4.1)$$

And similarly, for the credibility contributed from negative rules:

$$\text{Cred}_{R^-} = \frac{\sum_{(r, a_r) \in T_{n\text{negative}}} a_r * r}{2 * \sum_{a_{r\max} \in T_{n\text{negative}}} a_{r\max}}. \quad (4.2)$$

The resulting normalized credibility for entity is then calculated like

$$\text{Cred}_e = \text{Cred}_{\text{default}} + \text{Cred}_{R+} + \text{Cred}_{R-} \quad (4.3)$$

Now that we have credibilities for entities, credibility of a document is calculated using a mean average of related entities like

$$\text{Cred}_d = \frac{\sum_{e \in E_d} \text{Cred}_e}{|E_d|}. \quad (4.4)$$

Finally, based on the credibility of documents the credibility of a statement can be calculated. The mean average is also used, but every supporting document adds 1 to the sum and each document counts twice, i.e.

$$\text{Cred}_s = \frac{\sum_{d_{\text{opposing}} \in D_{\text{opposing}_s}} (1 - \text{Cred}_{d_{\text{opposing}}}) + \sum_{d_{\text{supporting}} \in D_{\text{supporting}_s}} (1 + \text{Cred}_{d_{\text{supporting}}})}{2 * (|D_{\text{opposing}_s}| + |D_{\text{supporting}_s}|)}. \quad (4.5)$$

To understand the idea behind making each document count twice and supporting document adding 1 to the sum, consider what happens if a only a mean average is taken as in eq. (4.4). Even if we would have only supporting documents and these documents would in addition be of varied credibilities, let's also assume normally distributed, the credibility of the statement would tend to 0.5 as predicted by central limit theorem. This, however, does not make a lot of sense intuitively speaking as we only have supporting documents for the statement of interest and not a single opposing document. This should make the statement more credible. With the modification made in eq. (4.5), each supporting document makes the credibility tend to 1 and in contrast, each opposing document makes the credibility tend to 0. This does not solve the general problem of the mean average tending to 0.5 when a lot of documents split approximately equally between supporting and opposing but rather only the specific problem when there is a predominant majority of supporting or opposing documents.

A detail for how each credibility is set, including additional context of the node, is given in an overlay that appears on hover over that node, see fig. 4.6.

Issues

Let's discuss some of the issues with this iteration.

Using the mean average propagation rule has the issue that with rising number of nodes involved in the calculation, the value tends to the middle value, i.e. 0.5. Such a result is not of a great interest and does not provide much insight. Few approaches to mitigate this problem which also expand the possibilities in terms what narratives can be modelled are described. The first two are then more concretely conceptualized in section 4.3.2.

First, the simple and rigid propagation rules that do not take into consideration some of the complexities occurring in the real world can be improved. For instance, a situation in the narrative example occurs where two documents from the same author and publisher differ in their stance towards the statement, see detail in fig. 4.7. This situation is exactly the example given in table 2.2 for *Exclusion* propagation rule. By applying this rule it would mean that in this narrative, the document starting the whole narrative would be removed. This should, in turn, affect the credibility calculation directly as well as indirectly affect credibility of other sources relying on it. This directly leads to another way of alleviating the aforementioned problem.

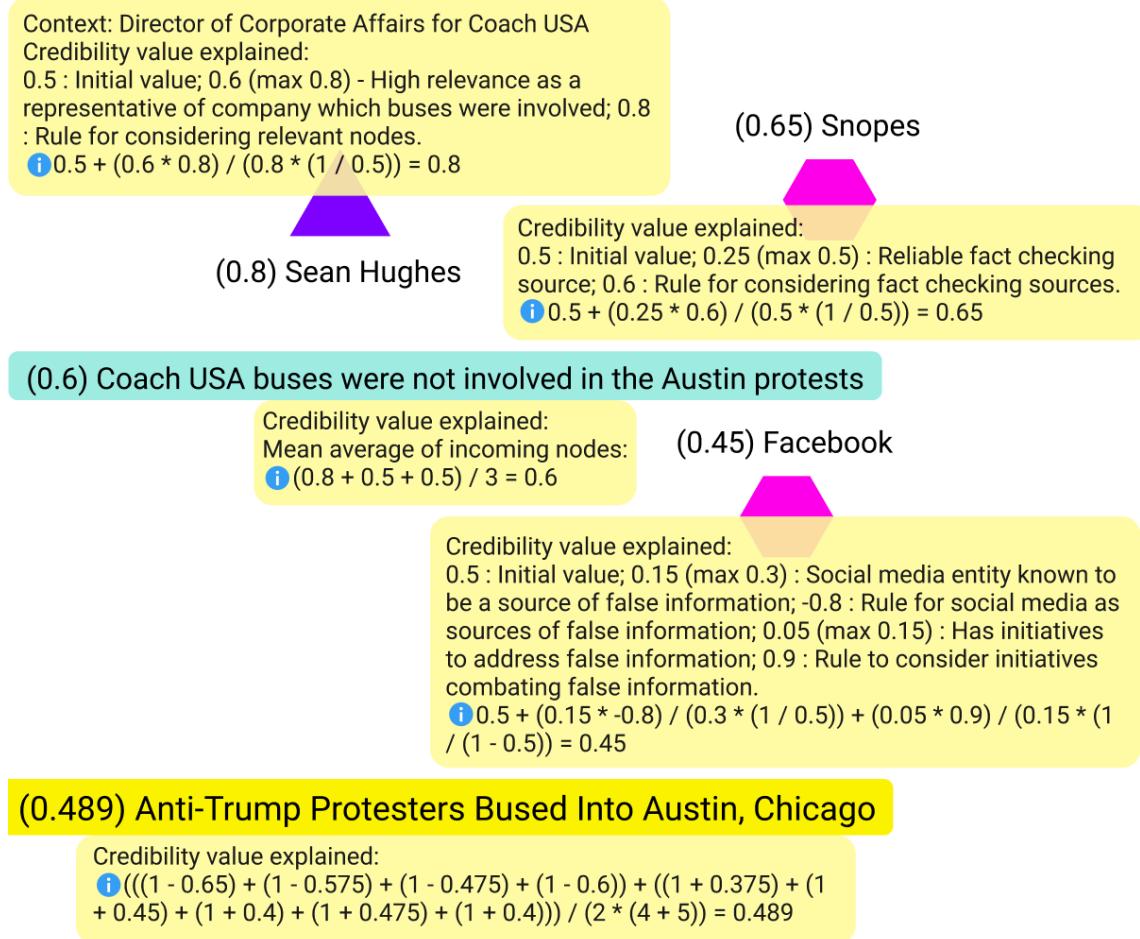


Figure 4.6: Node overlay appearing on hover giving context relevant for that node including explanation of how that node's credibility is calculated. The figure shows five examples, all taken from fig. 4.4.

Second, it should be possible to include other edges in the credibility graph capturing other relations and dependencies from the domain model. For example, every other source supporting the statement in the example narrative directly depends on the initial Eric Tucker's tweet. If it is affected in some way, for example by an activation of the *Exclusion* propagation rule or by an actual removal of the original document in the real world (as indeed happened [75]), this should affect credibility of other sources to such a degree as to completely discount the statement.

Third, the *Aggregation* rule from table 2.2 also helps with the issue since it reduces the number of nodes involved in the calculation and therefore, by definition, mitigates the problem. Looking at the type of sources and documents in the narrative example, there are opportunities to (1) aggregate over documents essentially just repeating other documents (all supporting documents largely just repeat the original tweet from Eric Tucker) and (2) aggregate over similar sources, such as social media (Facebook, Reddit and Twitter), daily newspapers (The Austin American Statesman and Fox 5 New York), but also by their political leaning (Free Republic is a forum for self-described US conservatives [76] and The Gateway Pundit is a far-right news website recently known for spreading false information about the 2020 U.S. presidential election [77]).

And lastly, it is not necessary to rely solely on mean average. Another simple function is to take a maximum credibility of the incoming nodes, for example

$$\text{Cred}_d = \max_{e \in E_d}(e). \quad (4.6)$$

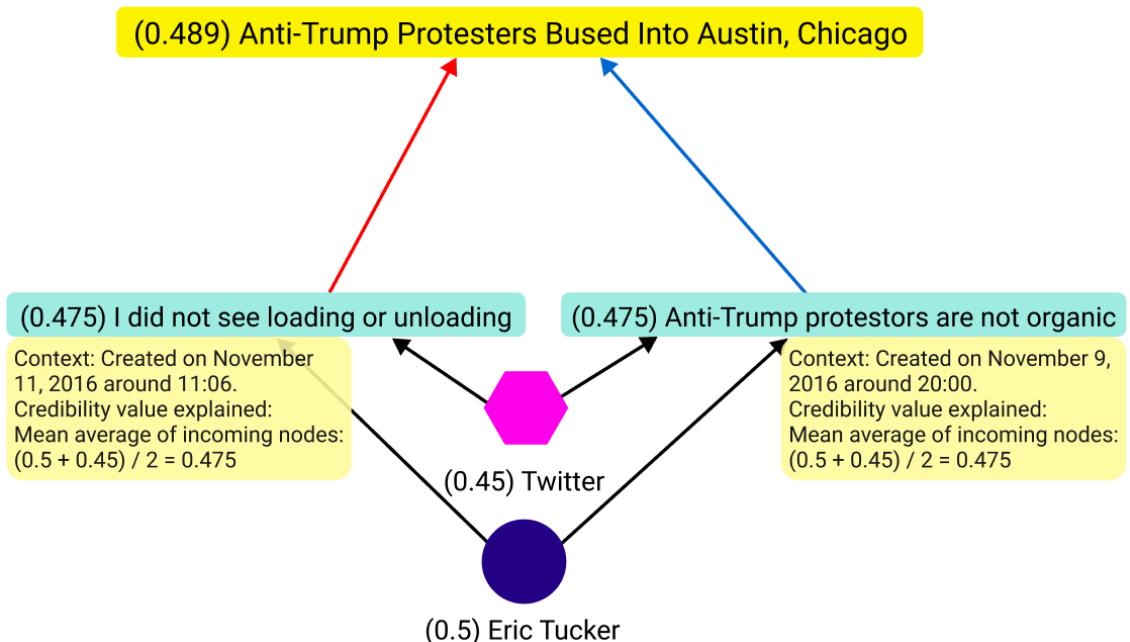


Figure 4.7: Detail of problematic situation in the narrative example suitable for application of the *Exclusion* rule discussed in table 2.2. Notice that while the first document (tweet) published on November 9 (can be seen in the overlay) is in support of the statement, the second document published two days later is in opposition of it.

4.3.2 Credibility graph improved

Let's see the improvements for the issues brought up in section 4.3.1 applied individually, one by one. For reference, the nodes and edges are summarized in table 4.2 and table 4.3,

Node	Description
Statement	At the center of a narrative which it uniquely identifies and which credibility is the most important.
Document	Uniquely identifiable object containing information in written but also audio-visual form.
Author	A legal entity or collection of them that have composed a document
Publisher	Most likely a juridical person who published a document
Other legal entity	Legal entity somehow otherwise related to a narrative. Sean Hughes from our narrative example is an instance of such entity since he is not an author of an article but provided a highly relevant information related to the buses.

Table 4.2: Credibility graph nodes and their descriptions

respectively.

Applying Exclusion rule

The *Exclusion* rule in table 2.2 suggests to "exclude or ignore certain entities from contributing to the credibility propagation" with the example of "a document from a source supporting a claim can be excluded if there is another later document from the same source opposing the claim" that is directly applicable to the situation in fig. 4.7. See fig. 4.8 for how this could look like.

Compared to rules discussed previously which were enabled on a continuous spectrum, this particular *Exclusion* rule is discrete-binary since the situation either occurs or it does not and the rule only decides whether to exclude or not exclude the document. It is difficult to imagine how could this rule work otherwise on a continuous spectrum.

Capturing additional relations

Although the *Exclusion* rule helps to capture a very intuitive idea and indeed decreases credibility of the statement, it is by itself not enough to significantly discredit it, since it only works in isolation and does not affect other nodes. If additional relations between the nodes are introduced, such as those of content dependency, and these in turn affect the connected documents, effects such as those of the *Exclusion* rule can have a much higher impact, see fig. 4.9.

This content dependency relation is shown with a gray dashed line starting with a semi-circle and ending with a triangle arrow. A rule that affects how much are documents, that depend on another document that is excluded, discredited (and their credibility lowered) is enabled at exactly half its strength and the credibilities of the depending documents are halved. This affects the credibility value for the statement and lowers it again from the previous 0.458 in fig. 4.8 to the current 0.376. These calculations are formalized in eq. (4.8) for documents and eq. (4.10) for statement. Only parts of interest are shown clearly and not blurred.

Let r_{DED} be a rule activation for the rule described in fig. 4.9 where DED is an abbreviation for "discredit excluded documents". A document d has a set $d_{D_{dependent}}$ of all documents that d depends on via the content dependency relation indicated by the gray arrows on fig. 4.9. $d_{D_{excluded}}$ denotes a set of excluded documents where $d_{D_{excluded}} \subseteq d_{D_{dependent}}$. Let d_{DED} be the discredit excluded documents factor for a document d defined as

Edge kind	Source	Target	Meaning
Solid red ●	Document	Statement	Document opposes a statement so its higher credibility decreases statement's credibility
Solid blue ●	Document	Statement	Document supports a statement so its higher credibility increases statement's credibility
Solid black ●	Entity (i.e. author, publisher or other legal entity)	Document	Marks the obvious relation of authorship, publication or direct mention of other legal entity
Solid gray ●	Either document or entity	Either statement or document	Relationship is excluded and has no effect
Dashed gray ●	All	All (except itself)	Captures other relationship types such as ownership or content dependency (used in the example)

Table 4.3: Credibility graph edges, their kind, source, target and meaning

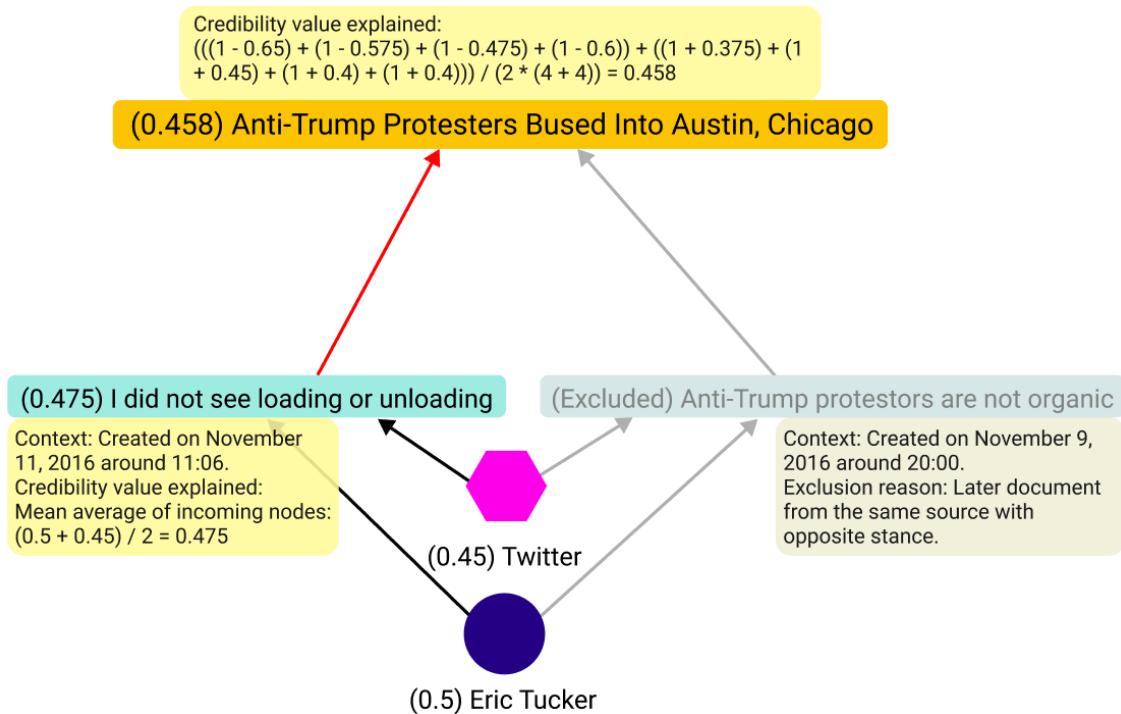


Figure 4.8: Applying *Exclusion* rule to the problematic situation. The excluded document is grayed out together with the ingoing and outgoing edges. Its overlay, that would only appear when hovering over it, has only grayed out background color to keep the legibility high. Instead of explanation for how the credibility value is calculated, it now contains a reason for excluding it. Notice also a change in the credibility value of the statement which decreased from 0.489 to 0.458.

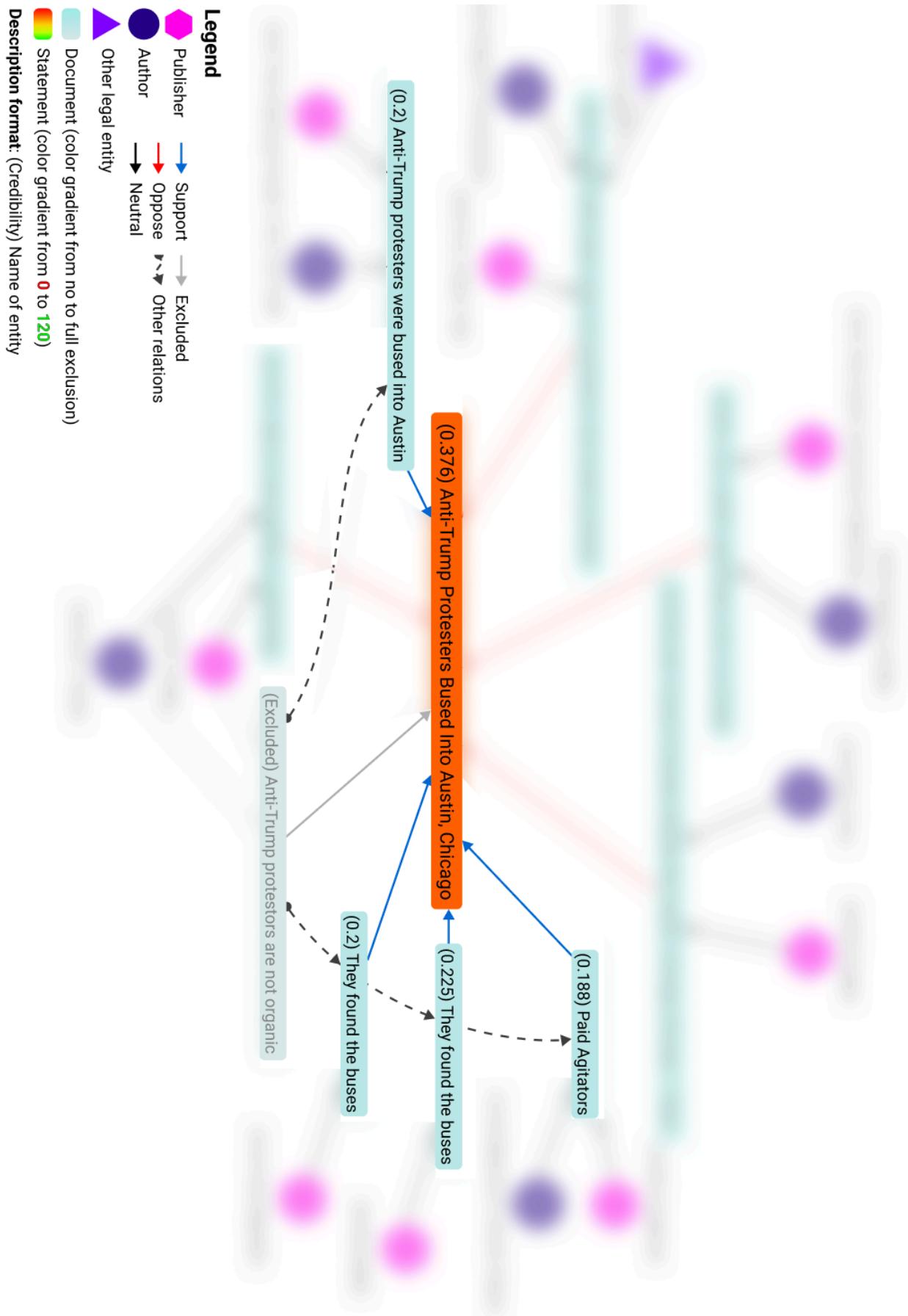


Figure 4.9: Additional relations added for Exclusion

$$d_{DED} = \begin{cases} \frac{|d_{D_{excluded}}| * r_{DED}}{|d_{D_{dependent}}|} & \text{if } |d_{D_{dependent}}| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.7)$$

The credibility for a document d is then calculated as

$$\text{Cred}_d = \frac{\sum_{e \in E_d} \text{Cred}_e}{|E_d|} * (1 - d_{DED}). \quad (4.8)$$

For example, see the credibility calculation for document "Paid Agitators":

$$\begin{aligned} \text{Cred}(\text{Paid Agitators}) &= \frac{0.45 + 0.3}{1 + 1} * \left(1 - \frac{1 * 0.5}{1}\right) \\ &= 0.375 * 0.5 = 0.1875 \approx 0.188. \end{aligned} \quad (4.9)$$

The credibility for a statement s considering excluded documents is calculated as

$$\text{Cred}_s = \frac{\sum_{d_{opposing} \in D_{opposing_s}} (1 - \text{Cred}_{d_{opposing}}) + \sum_{d_{supporting} \in D_{supporting_s}} (1 - d_{DED} + \text{Cred}_{d_{supporting}})}{2 * (|D_{opposing_s}| + |D_{supporting_s}| - \sum_{d \in D_s} d_{DED})}, \quad (4.10)$$

The credibility for the statement in the example is then calculated as

$$\begin{aligned} \text{Cred}(\text{Anti-Trump Protesters Bused Into Austin, Chicago}) &= \frac{((1 - 0.65) + (1 - 0.575) + (1 - 0.475) + (1 - 0.6)) + ((0.5 + 0.188) + (0.5 + 0.225) + (0.5 + 0.2) + (0.5 + 0.2))}{2 * \left(4 + 4 - 4 * 0 - 4 * \left(\frac{1 * 0.5}{1}\right)\right)} \\ &= \frac{1.7 + 2.813}{2 * 6} \approx 0.376. \end{aligned} \quad (4.11)$$

4.4 Addressing cognitive bias

Having the credibility graph in place, attention can be given to the idea of addressing cognitive bias. This can be done using credibility graphs and rules affecting credibility propagation within them.

Consider the associated credibility graphs for a set of narratives NAR the end user is interested in. The user considers each narrative using the information the credibility graph G_{cred} shows and through critical thinking reaches a conclusion to the statement s representing a narrative, either opposing or supporting it. Afterwards, they set the available rule parameters for that particular credibility graph G_{cred} such that credibility of the statement Cred_s leans towards their conclusion, i.e. $\text{Cred}_s < 0.5$ if they oppose the statement s and $\text{Cred}_s > 0.5$ if they support the statement s . Not every rule has an effect on credibility graph G_{cred} since it depends on what is included in G_{cred} . For example, a rule regarding perceived trustworthiness of tabloid sources does not affect in any way every narrative

that does not involve any tabloid source.

There is limited number of rules $|R|$ with their associated parameters, creating a r -dimensional space S_R . Only a proper subset $S_{Rs} \subset S_R$ produces a desired output Cred_s in a credibility graph G_{cred} . Since these subsets are different from one another, the more credibility graphs there are, the smaller the intersection of all these subsets. Figure 4.10 shows this on a simple example with just two rules and three narratives.

The assumption is that under coherent reasoning (and therefore setting of rule parameters), there

$$\exists S_{\text{applicable}} \mid \left((S_{\text{applicable}} = \bigcup_{s \in \text{Set}_s} S_{\text{Applicable}(s)}) \wedge S_{\text{applicable}} \neq \emptyset \right) \quad (4.12)$$

where Set_s is a set of all narratives with their respective statements and

$$\text{Applicable}(s) = \begin{cases} \text{Cred}_s > 0.5 & \text{if user supports narrative with statement } s \\ \text{Cred}_s < 0.5 & \text{otherwise} \end{cases} \quad (4.13)$$

and $S_{\text{Cred}_s > 0.5}$ is then to be interpreted as a subset of S_R where $\text{Cred}_s > 0.5$.

The condition of being under coherent reasoning is crucial here, since this is exactly what can then be used to discover incoherent reasoning that, given the narrative has been thoroughly thought through, indicates presence of cognitive bias, see fig. 4.11.

4.5 Collaborative project organization

This section serves as a single place for all the collaborative aspects of CTSS, summarizing mentions from other sections as well as broaching some not mentioned before. The reasoning for why collaborative approach seems as a promising approach is discussed in section 2.1.1 and section 2.1.2.

- Data collection
 - Manual contribution of data
 - Cleaning automatically collected data
 - Filtering collected data relevant to a narrative
 - Discussing and suggesting improvements to a data collection pipeline
 - Implementing said suggestions ↑
- Credibility graph
 - Creating new credibility graphs
 - Improving existing credibility graphs, e.g. by finding new connections in data
 - Creating and tweaking behavior of credibility propagation rules
 - Voting on arguments affecting objective credibility
 - Discussing and suggesting improvements to credibility graphs, credibility prop-

⁸3D is hard™

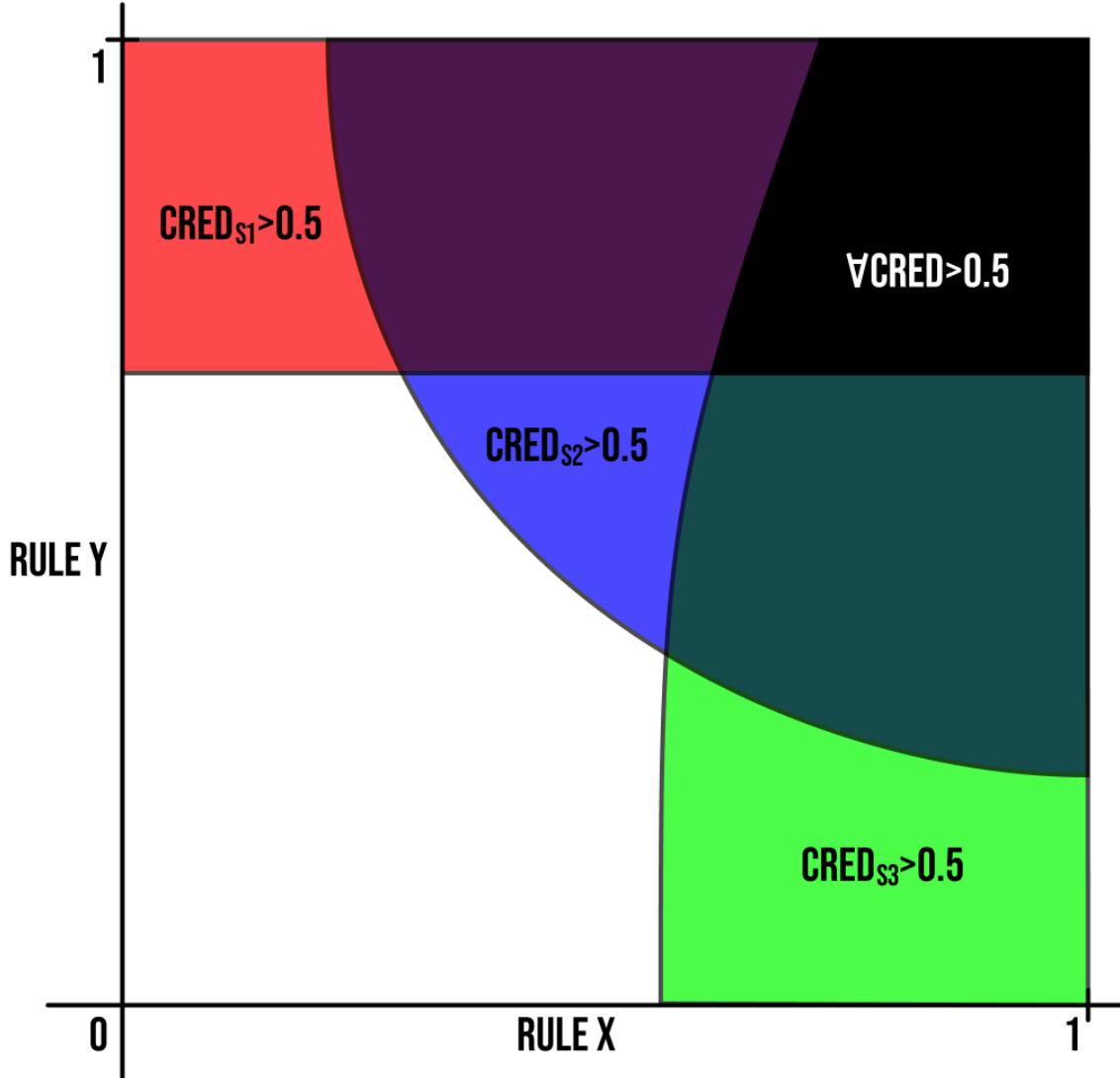


Figure 4.10: Narratives constraining space of applicable rule parameters. Only involves two rules $R = \{\text{rule Y}, \text{rule X}\}$ for legibility⁸ but the concept is generalizable to r -dimensional space. There are three narratives with statements s_1, s_2 and s_3 . Let's assume that the desired outcome for each of these is $\forall_{s \in \{s_1, s_2, s_3\}} \text{Cred}_s > 0.5$. Where they attain $\text{Cred}_s > 0.5$ is shown by red S_{red} , blue S_{blue} and green S_{green} blobs, respectively, each subset of S_r . S_{red} shows the example where a rule has no effect since at any value of rule Y, changing the rule X does not change Cred_{s_1} — rule X could therefore be omitted when setting up credibility graph for narrative relating to statement s_1 . The black area represents the space of applicable rule parameters $S_{\text{applicable}} = S_{\text{red}} \cap S_{\text{blue}} \cap S_{\text{green}}$.

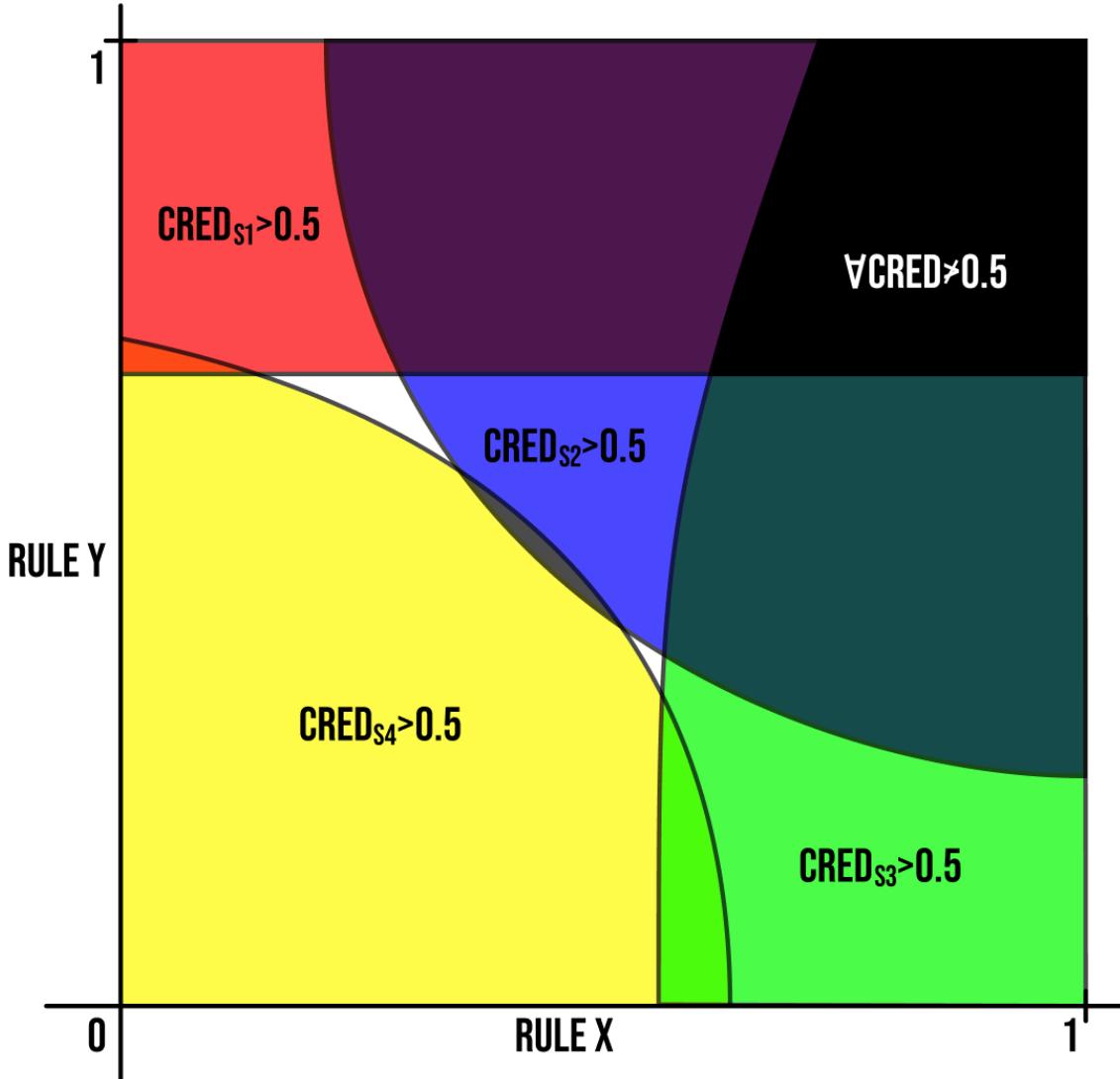


Figure 4.11: Narratives not intersecting in any applicable rule parameters space. Compared to fig. 4.10, this figure contains another narrative with statement s_4 and the desired outcome of $\text{Cred}_{s_4} > 0.5$ being represented by yellow blob $S_{\text{yellow}} = S_{\text{Cred}_{s_4} > 0.5}$. However, $S_{\text{red}} \cap S_{\text{blue}} \cap S_{\text{green}} \cap S_{\text{yellow}} = \emptyset$ which violates the assumption in eq. (4.12). This therefore means that the addition of the narrative represented by statement s_4 is incoherent with the rest of the narratives. User's desired outcome is for $\text{Cred}_{s_4} > 0.5$ which is not possible with constraints set by the other narratives and given the narrative has been carefully thought through, it hints at cognitive bias being introduced.

agation rules and objective credibility arguments

- Implementing said suggestions ↑

- Using CTSS - although principally helpful to the end user, the data on subjectively set rule parameters could become a signal of its own and (1) affect setting of objective attribute and (2) function as a social consensus signal.
- Open source - anyone can contribute, anyone can audit and verify functionality. Makes it fully transparent. Also allows for the system to be installed and used by everyone locally or at least separately in a cloud sandbox environment. This is helpful for development, small-scale experiments and academic research.

Proposing a solution through such a collaborative platform comes with all kinds of considerations complicating matters. There is a problem of abuse discussed more closely in section 4.5.1. There are also relevant considerations about how should versioning work. For example, one user opens a credibility graph about some narrative and sets up the rules such that they get the desired outcome. Other user modifies the credibility graph for this narrative in some way, or credibility propagation rule used there is modified. One solution is that the version of everything used is fixed to a version most up-to-date at the time of credibility graph usage by the user. There could then be an option to update to most up-to-date version when it becomes available, e.g. from within the narrative overview view in fig. 4.21.

4.5.1 Abuse potential

As with any collaborative project where the original creator does not remain in full control and where *anyone* can contribute, there is a potential for abuse. The area of false information is a sensitive topic and it is important to be careful and thoughtfully design in checks and balances that would prevent such abuse. The following is an incomplete list of potential abuse vectors and ideas on how to prevent them:

- Organized false information campaign - An organized group of professional false information creators could as a part of their campaign create supporting material on CTSS and if not only to circumvent the protection that CTSS should offer, they could even directly use it as an additional support and proof that the information is not false.

Multiple ideas arise for protection against such abuse:

- Warn about fresh and developing narratives - When a narrative is new, a general warning about such fact could be presented to the users of CTSS with an encouragement to wait for more information to come in and for the narrative to be reviewed and confirmed. This would not need to be necessarily applied to every narrative and some measure of its virality could be used for deciding whether to show it or not.
- Institute a reputation or privilege system - Inspired by Stack Exchange or Wikipedia, respectively. Certain actions would only be allowed with certain level of reputation or privileges which would be relatively hard to obtain. It would then be possible to identify actions prone for abuse, rank them by their product of likelihood and severity of abuse, and use that for deciding the amount of reputation needed to perform them.
- Establish a review process - Narratives and its nodes could be created by anyone but would have to meet a certain level of peer review in order to be ap-

proved for use. However, it is important to strike a balance between being overly cautious having long, thorough review process and approving information in a timely manner. Such a process could be sped along by users with high reputation (see above).

- Denigration - An established node could be denigrated by modification of its attributes, i.e. adding a negative one, removing a positive one or changing strength of an existing one. This could be prevented by:
 - Attribute elections - When an attribute is to be modified, especially one that would have a meaningful impact (something that could be measured since it would be possible to calculate the total change in outcome of narratives), short elections in the community could be held to decide whether to modify it or not. To prevent the simplest of election frauds when massive amount of accounts is created for such purpose, certain reputation (see above) could be required.

4.6 Proposed usage

This section demonstrates an imagined usage of the proposed solution. It is also used for evaluation of and serves as a basis for conducting qualitative interview as described in section 6.2.

Let's say an individual would encounter the example narrative from section 1.1 on Twitter as seen in fig. 4.12 without CTSS browser extension installed. Compare that to fig. 4.13 that shows an overlay with links to related documents categorized by either their support or opposition to the narrative.



Figure 4.12: Example narrative on Twitter *without* CTSS¹⁰

¹⁰The original tweet is no longer available as the tweet was later removed by Eric Tucker. This figure has been obtained using Internet Archive.

 **erictucker**
`@erictucker`

[Follow](#) 



Anti-Trump protestors in Austin today are not as organic as they seem. Here are the busses they came in. #fakeprotests #trump2016 #austin



RETWEETS **2,587** LIKES **2,405**



6:43 pm - 9 Nov 2016



Related: [See the full picture](#)

Support

- BREAKING: They found the buses! ... (Reddit)
- Anti-Trump Protesters Were Bussed in to Austin (Gateway Pundit)
- Paid agitators (Facebook)
- They found the busses (Free Republic)

Oppose

- Follow-up tweet
- "I can confirm that those were our buses" (Austin American-Statesman)
- Fact-check (Snopes)
- Coach USA buses were not involved in the Austin protests (Fox 5 NY)

Figure 4.13: Example narrative on Twitter *with* CTSS. As compared to fig. 4.12, there is an icon next to the tweet. The little number in it indicates how many related documents are available. After clicking the icon, an overlay appears containing links to these related documents categorized by either their support or opposition to the narrative. It also contains a link to the CTSS website, particularly a credibility graph for this narrative.

Listing these related documents already sets a stage for encouraging critical thinking. By following the "See the full picture" link a new page would appear with a credibility graph for this particular narrative as shown in fig. 4.14.

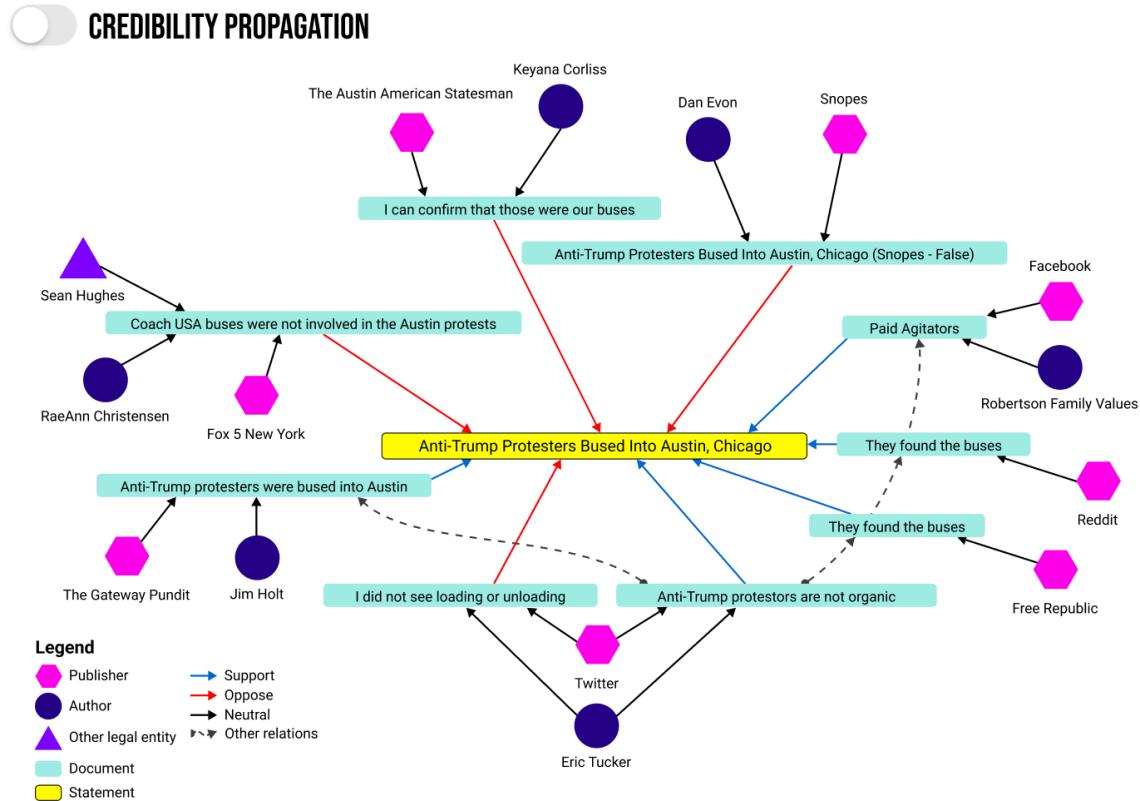


Figure 4.14: Credibility graph for example narrative (section 1.1) with credibility propagation turned off (note the switch in the top left corner).

At first a simplified view of the credibility graph is shown which does not involve the credibilities. This view is useful for providing an overview of the narrative enhancing the mere list of related documents from fig. 4.13 but without other unnecessary information. The credibility propagation can be turned on and the view transforms to that of fig. 4.15.

This view causes the following changes:

- The credibility value is shown next to a node description.
- Statement credibility can change. Since statement is a node of the greatest interest, it is suitable to highlight the changing credibility value, e.g. by changing its background color. A red to green color gradient is chosen since (1) it is big enough to notice relatively small change in the value and (2) the statement denotes either the opposition or the support for which the most suitable colors seem to be red and green, respectively. Note that by default (i.e. using RGB color model) such a color gradient would create a brown color as the middle value. On the other hand, HSL color model creates a gradient going through a yellow color and is preferable choice. See fig. 4.16 for comparison.
- The rules get activated and start to affect the calculation of the credibilities for all of the nodes in the credibility graph as explained in section 4.3. This means that the Exclusion rule comes into effect. A cyan-to-gray color gradient is applied to a document background color based on whether it is excluded or not, or the amount

CREDIBILITY PROPAGATION

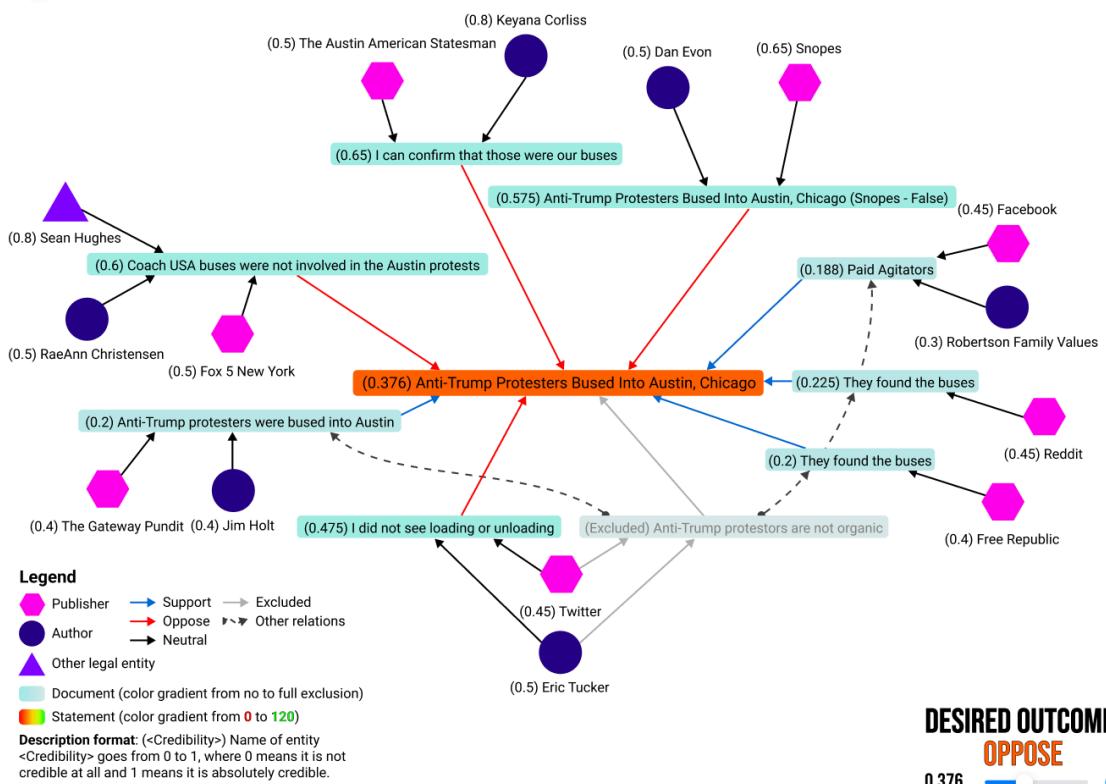


Figure 4.15: Credibility graph for example narrative with credibility propagation turned on (note the change on the switch in the top left corner).

by which it is affected by an excluded document. For example, fig. 4.15 shows a fully excluded gray ● document "(Excluded) Anti-Trump protestors are not organic", four half-excluded half-gray-half-cyan ● documents, one of which is "(0.188) Paid Agitators" and four cyan ● unaffected documents, one of which is "(0.475) I did not see loading or unloading".

- The excluded documents have the ingoing and outgoing edges also changed to a gray color.
- All of the above is added for explanation to the legend.
- Finally, a desired outcome element with an associated interactive slider appears in the bottom right corner. This can be used by the user to indicate their desired outcome for this narrative which is then used in the consideration of cognitive bias as explained in section 4.4. Until a user explicitly sets a desired outcome on the slider for the narrative, the value follows that of the statement's credibility.



(a) RGB red to green gradient

(b) HSL red to green gradient

Figure 4.16: RGB vs. HSL color gradients

In this view it also becomes possible to click on the individual nodes which makes them active. Figure 4.17 shows an example where the publisher node "Snopes" is clicked and activated.

An active node becomes highlighted by a blue ● outline and the associated attribute and rule nodes affecting its credibility are shown. A legend is expanded to explain these attribute and rule nodes. A strength of a rule can be changed by moving the knob on the associated slider. An example of this is shown on fig. 4.18.

The value of the Reliable Fact Checker rule is changed from +0.6 to +0.2 which causes the following:

- The proportion of background that is dark gray in the diamond shape used for the Reliable Fact Checker rule is changed to the new proportion, i.e. in this case from 60% to 20%.
- Snopes's credibility decreases according to eq. (4.3) from 0.65 to 0.55. All changes to the credibility are temporarily highlighted by a bold font.
- Change to Snopes's credibility affects the credibility of the connected document which according to eq. (4.8) decreases from 0.575 to 0.525.
- This decrease of credibility of an opposing document positively affects the credibility of the statement which according to eq. (4.10) increases from 0.376 to 0.38.
- The value of the desired outcome follows that of the statement's credibility and therefore also changes to 0.38.

Let's say that in this hypothetical usage scenario, the user really does not believe the statement and wants to decrease its credibility a lot. This user is of the opinion that the situation captured in fig. 4.8 should affect the credibility of the related documents much more. They click on one of the documents having a relation to the excluded documented

CREDIBILITY PROPAGATION

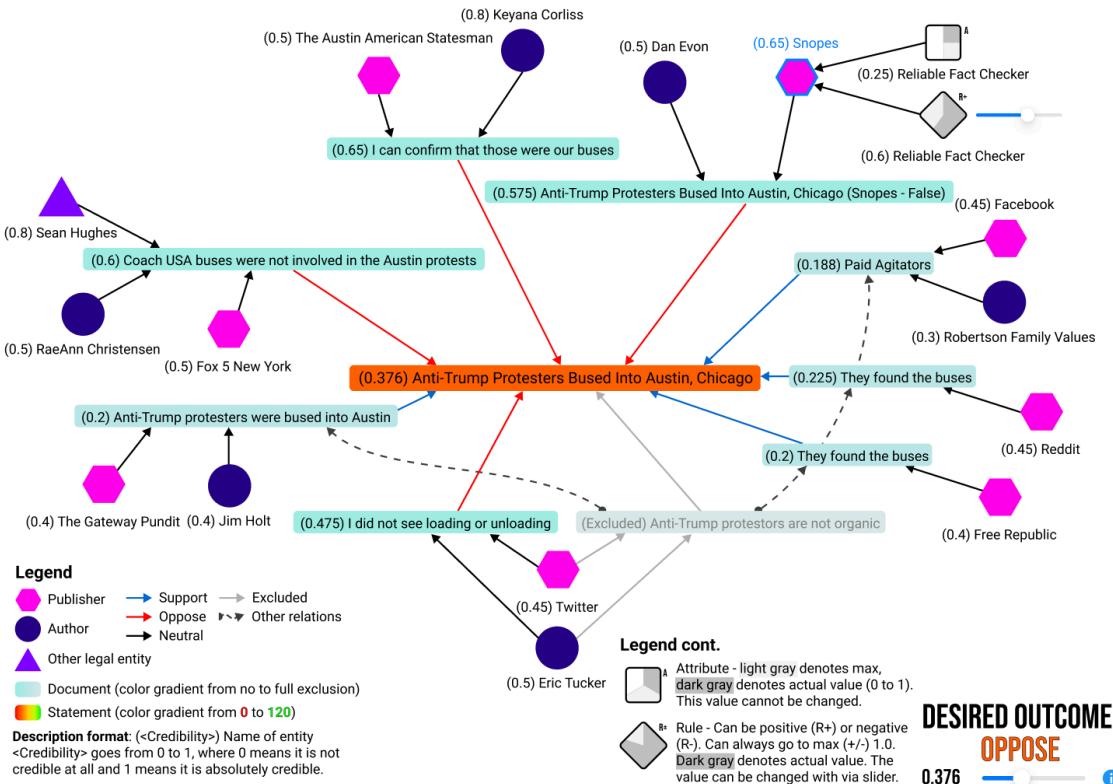


Figure 4.17: Snopes publisher node is activated by a mouse click in the credibility graph

which activates the documents and results in fig. 4.19.

The document is highlighted as previously by a blue ● outline and the associated rule "Credit Excluded Documents" is shown. This document does not have any associated attribute so none is shown. A strength of this rule is again changed on the associated slider which is shown on fig. 4.20.

The value of the Credit Excluded Documents rule is changed from +0.5 to +0.2 which causes the following:

- The proportion of background that is dark gray in the diamond shape used for the Credit Excluded Documents rule is changed to the new proportion, i.e. in this case from 50% to 20%.
- Credibility of all four documents decreases according to eq. (4.8) by a relative decrease of 60% in every case which corresponds to the 60% decrease in the rule strength. All changes to the credibility are temporarily highlighted by a bold font.
- This decrease of credibility of the supporting documents negatively affects the credibility of the statement which according to eq. (4.10) decreases from 0.38 to 0.299. This also affects the background color which changes from orange ● to red ●.
- The value of the desired outcome follows that of the statement's credibility and therefore also changes to 0.299. The text also changes from just "Oppose" to "Strongly oppose" and its color changes in accordance to the statement's background color.

In such a way, the credibility of the statement was changed from the initial 0.376 to the

CREDIBILITY PROPAGATION

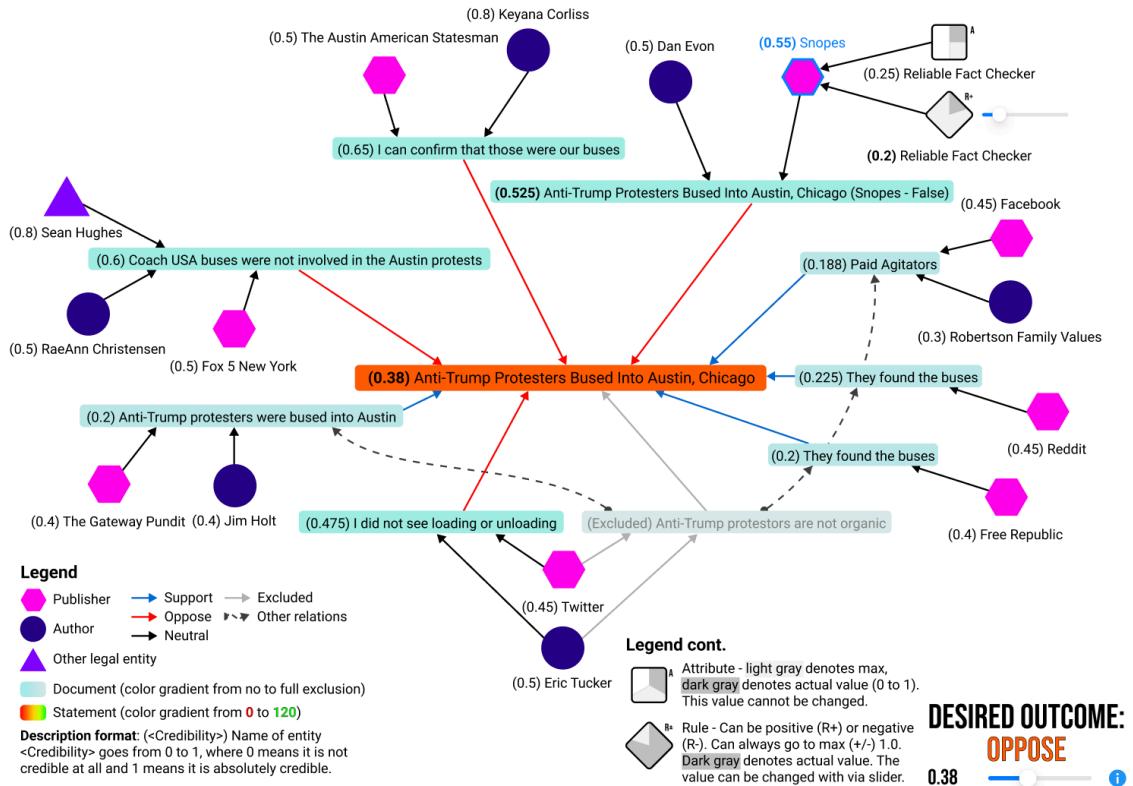


Figure 4.18: Reliable Fact Checker rule associated with Snopes is changed

current 0.299. It is assumed that these actions a user does are done with a serious intent and represent how this user believes the rules should be set. As such, the credibility of a statement is taken as the desired outcome for this narrative and is stored¹¹ for the user.

Let's now suppose that the same user would use CTSS for other narratives. It is then meaningful to show an overview of narratives the user showed an interest in comparing current and desired outcomes. An example of such an overview with six narratives is shown in fig. 4.21. Each row has a narrative summarized by its statement, its desired outcome which is the credibility value that has been stored after the user first encountered the narrative in CTSS, its current outcome, i.e. the credibility of the statement calculated based on the current settings of the rules, and finally a status based on comparing these two values. Three of the narratives are reported with an "OK" status having a green background which means that the desired and current outcome do not differ by much — these narratives in the example differ by a maximum of 0.01 point. Two narratives have an orange status, signifying a warning. The status message for these two narratives is different, the first one reading "Weaker Than Desired" and the second being "Stronger Than Desired". Looking at the desired and current outcome columns, the first narratives has a desired outcome that is smaller than the current outcome by $0.38 - 0.3 = 0.08$. So in this case the current outcome is weaker than the desired outcome, therefore the status message. The case is reversed for the second narrative where the desired outcome is bigger than the current outcome by $0.4 - 0.33 = 0.07$ so the current outcome is stronger than desired and thus the status message. Lastly, the third narrative from the top has a red

¹¹A user account would be needed for a long-term storage which is omitted as an uninteresting aspect of the system.

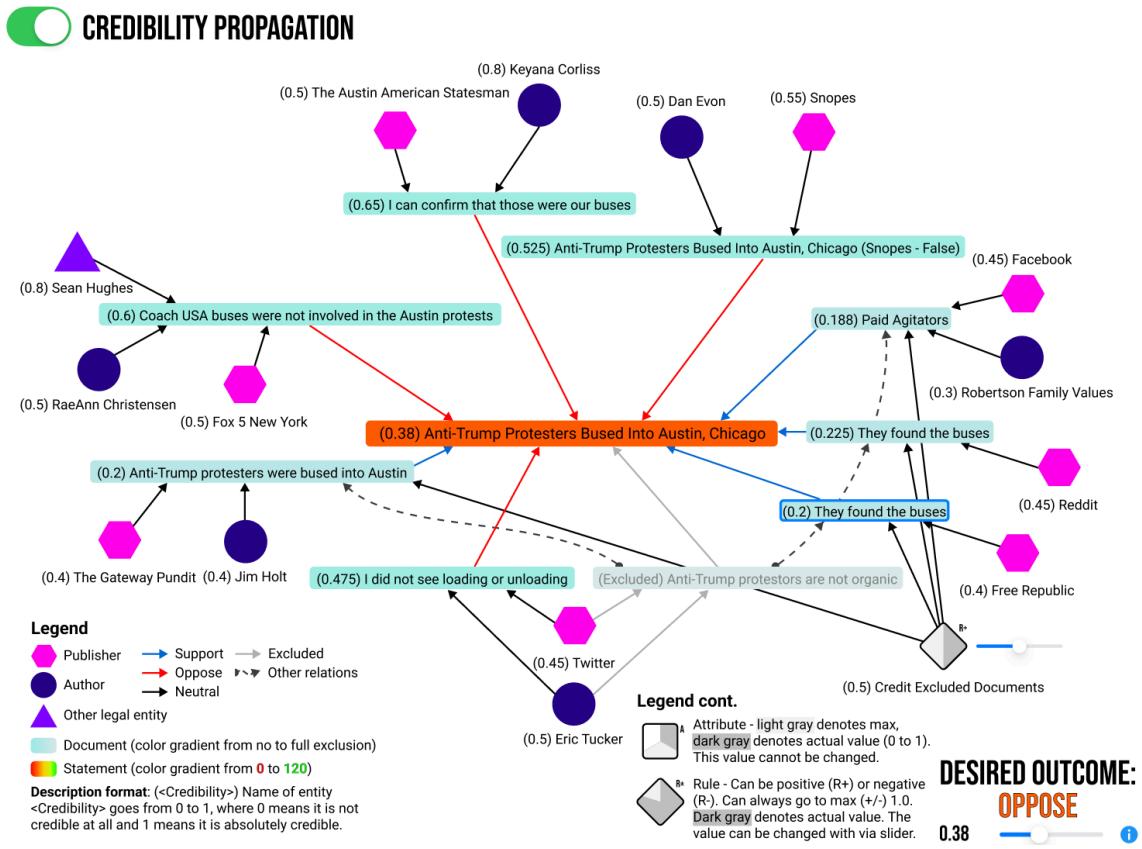


Figure 4.19: The document "(0.2) They found the buses" related to the original tweet document "(Excluded) Anti-Trump protestors are not organic" is activated

CREDIBILITY PROPAGATION

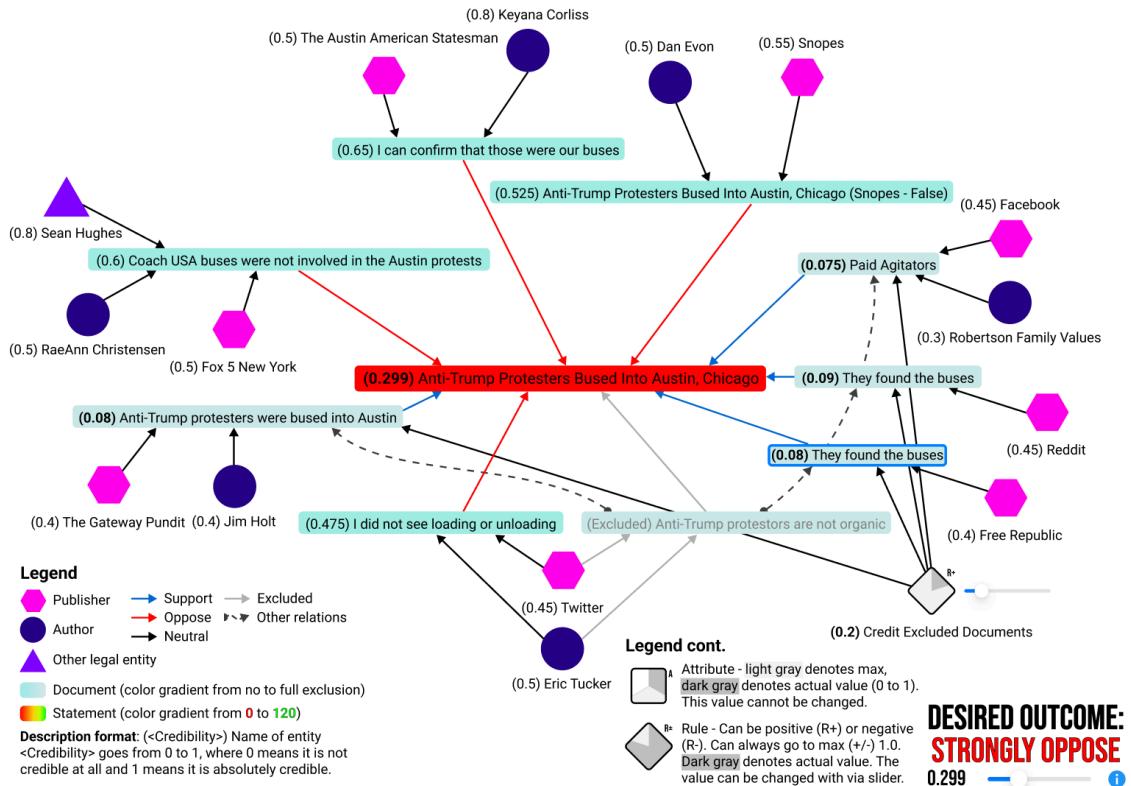


Figure 4.20: Credit Excluded Documents rule is changed

status, indicating a problem. Even though the difference between the desired outcome and current outcome in this case is just $0.51 - 0.45 = 0.06$ which is a lower value than for the just discussed orange narratives, the crucial difference here is that while the desired outcome is smaller than 0.5 and therefore means an opposition to the statement, the current outcome is bigger than 0.5 which means a support towards the statement. Because of this difference, the status message is "Different Than Desired".

YOUR NARRATIVES

NARRATIVE NAME	DESIRED OUTCOME	CURRENT OUTCOME	STATUS
Anti-Trump protesters bused into Austin, Chicago	Strongly Oppose (0.3)	Oppose (0.38)	Weaker Than Desired
Taliban seized \$85 billion of U.S. weapons	Strongly Oppose (0.3)	Oppose (0.31)	OK
COVID-19 originates from Wuhan laboratory leak	Lightly Oppose (0.45)	Leaning Support (0.51)	Different Than Desired
Uncertainty makes you seem weak to others	Oppose (0.4)	Oppose (0.33)	Stronger Than Desired
Afternoon alertness drop is not connected with lunch	Support (0.6)	Support (0.59)	OK
Trickle-down economics works	Strongly Oppose (0.3)	Strongly Oppose (0.3)	OK
Homeopathy works	Strongly Oppose (0.3)	Lightly Oppose (0.45)	Weaker Than Desired

Figure 4.21: User's narratives overview comparing current vs. desired outcome

Such an overview then shows to the user where the initially set desired outcome differs from the currently calculated calculated credibility. The user then has an option to revisit the narrative and do either:

1. Change the rules such that the current outcome is closer to the desired outcome. Note than when revisiting, the desired outcome value does not change anymore in accordance to the current outcome. This only happens when encountering the narrative for the first time when it is assumed that the reached value of a statement also represents the desired outcome for the narrative.
2. Change their opinion about the narrative and change the value of the desired outcome accordingly such that it is closer to the current outcome which can be done through the slider in the bottom right corner as shown in fig. 4.22.

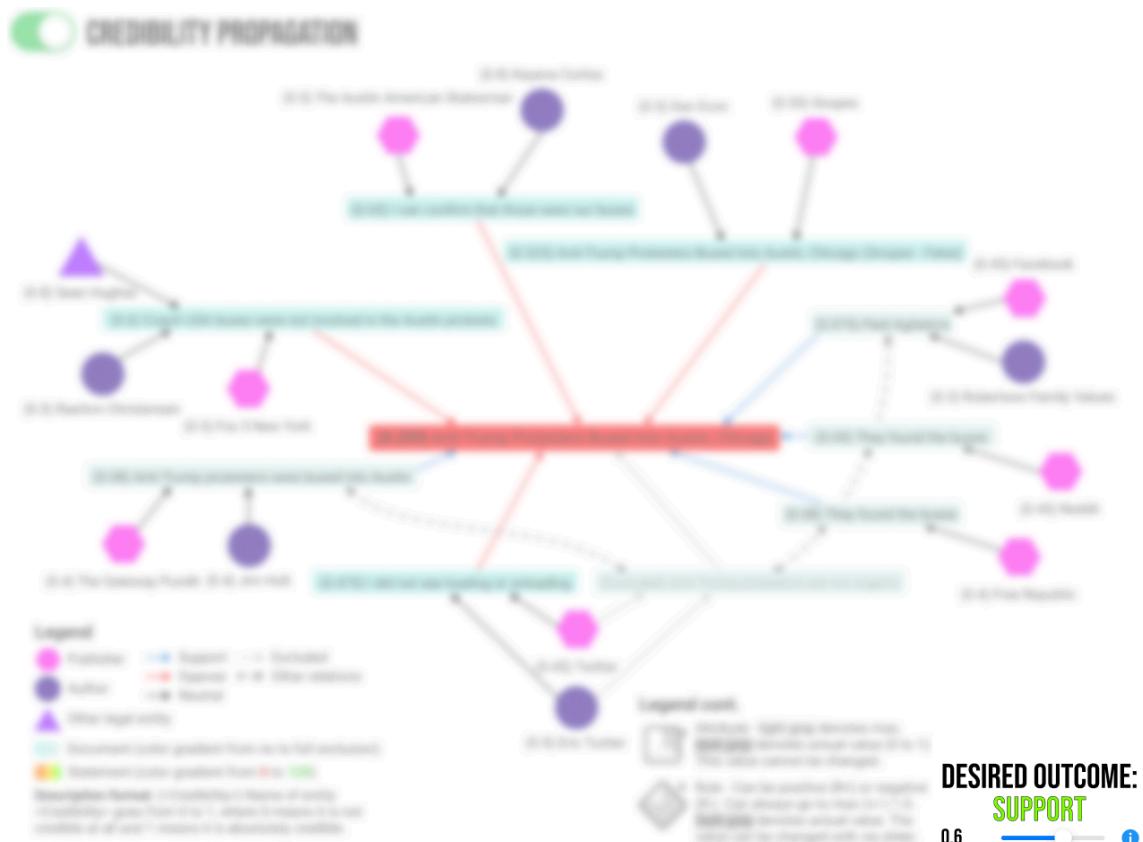


Figure 4.22: Change of desired outcome

5 Implementation

The implementation focuses on a visualization prototype of credibility graph from section 4.3. It is implemented as a web application using JavaScript framework React and a visualization library for the credibility graph visualization. It uses the example narrative from section 1.1 and encodes it in a configuration file.

5.1 Credibility graph visualization

A [Network component from vis.js library](#) is used for the credibility graph visualization. It offers a simple interface for network (graph) configuration with a support for the necessary entity shapes and colors.

By default, it positions the nodes randomly when initialized and uses [physics simulation](#) to position the nodes and edges and afterwards allows for a manual movement of the nodes. This has the following undesired effects:

- when loaded, there is a few seconds of wild and disruptive movement until the nodes and edges stabilize
- when refreshed, since the initialization is random, a different result is obtained
- the stabilized results are of questionable legibility oftentimes requiring a lot of manual adjustments.

For this reason, the positions of the individual nodes are hardcoded in the configuration file.

Using the vis.js library provides a simple and efficient way to approximate the desired solution presented in chapter 4. However, some things are clumsy and to satisfy certain requirements it would require similarly clumsy and hacky solutions, if it would be possible to satisfy them in some way in the first place. For this reason, an alternative and probably highly customized solution for credibility graph visualization would be superior. Such a solution should solve the general problem of positioning nodes to provide a compact yet legible results.

5.2 Configuration file

The configuration file stores both the credibility graph of a narrative as well as the rules affecting credibility propagation. A JSON5 format is used which is a more flexible and forgiving format, as evidenced in [78], superior to JSON when it comes to the configuration use case JSON is often used for.

The format of the file is shown in listing 5.1 with example values and clarifying comments. Ellipsis (...) is used when the same key was already defined and its value is the same as before or for next values in a list.

```
{  
  nodes: {  
    statement: { // there is only a single statement  
      summary: 'Anti-Trump Protesters Bused Into Austin',  
      id: 'statement node', // unique value across nodes  
      position: { // hardcodes the position on the vis.js canvas  
        x: 0,  
        y: 0,  
      }  
    }  
  }  
}
```

```

    },
},
documents: [ // there can be multiple documents, authors and publishers
{
  title: 'Protesters not as organic as they seem',
  id: ...,
  timestamp: '2022-01-21T12:47:02Z', // ISO 8601 date and time in UTC
  position: ...,
  deleted: true, // bool or timestamp of deletion if known
  other: { // for other data, used only to show it in an overlay
    sources: [ // for multiple links, e.g. the original and archived
      'https://example.com/tweet',
      'https://example.com/archived-tweet'
    ],
    times-shared: {
      facebook: 350000,
      twitter: 16000,
    },
    followers-at-time: {
      twitter: 40,
    },
  },
},
...
],
authors: [
{
  name: 'Eric Tucker',
  id: ...,
  position: ...,
  // names correspond to rules
  // attributes seen in the rest of thesis used as example
  // value in range [0,1], 0 value attributes can be omitted
  attributes: {
    relevance: 0,
    reliable-fact-checker: 0,
    false-information-source: 0,
    initiatives-against-false-information: 0,
  },
  other: {
    age: 35,
    job: 'Co-founder of marketing company',
    context: [ // any other relevant context
      'I am just a private citizen with tiny Twitter following',
    ],
  },
},
...
],
publishers: [
{
  name: ...,
  id: ...,
  attributes: ...,
  position: ...,
  other: {
    homepage: 'https://example.com/publisher-homepage',
    context: ...,
  },
},
...
],

```

```

},
edges: [
{
  from: 'node ID',
  to: 'node ID',
  type: 'neutral', // or 'oppose', 'support' or 'other'
},
...
],
rules: { // collection of all attributes used in nodes
  relevance: {
    type: 'positive' // or 'negative'
    strength: 0.5 // value in range [0,1]
  },
  reliable-fact-checker: ...,
  false-information-source: ...,
  initiatives-against-false-information: ...
}
}
}

```

Listing 5.1: Configuration file format

5.3 User experience simplifications

Certain simplifications affecting the user experience were made. Due to limitations and unfamiliarity of the vis.js library, the interactivity and resulting user-friendliness envisioned in chapter 4 is replaced by a workaround.

First, an editable text area containing the full configuration file is relied upon. Since the configuration file contains the rules it is possible to edit the value there instead of interacting with more user-friendly sliders.

Second, the interactive element where it is possible to click on a node and attributes and rules affecting it appear is also not included. It is, however, quite easy to see the attributes (and with it the namesake rules) in the aforementioned text area.

6 Evaluation

6.1 Verification

The main components necessary to support critical thinking (RQ1) is a relevant domain model having consequences for the data needed (RQ2) and the associated perspective used for looking at the problem that has consequences for the way of communicating and encouraging people in informed critical thinking (RQ3). These guide development of components needed further. The domain model used for development of credibility graph is derived from general observations of the information landscape from the perspective of credibility which is just one of the possible perspectives to consider for the problem at hand [79]. Still, the information landscape being as complex as it is where all kinds of aspects can affect the perceived credibility, certain simplifications in the domain model are in order. As such, the domain model attempts to capture the most relevant elements affecting perception of credibility. At the same time, the domain model is intentionally flexible by allowing addition of ad hoc relations and attributes to those elements.

The domain model provides basic guidance for development of the data collection pipeline component and the credibility perspective leads to credibility graph component that can show the domain model entities together with their credibility in a meaningful way.

The ideas proposed for the data collection pipeline are not verified neither for its utility nor feasibility and only provide a rough outline for solving this major task which would itself be deserving of a separate thesis. Through the encountered research a few unique ideas for often disregarded metadata helpful in establishing credibility are given.

The amorphous context from the domain model affecting the credibility is realized through attributes and rules which deterministically govern the credibility propagation in a credibility graph. The outlined mathematical modelling for this credibility propagation is limited by considering only linear functions for the rule propagation and using average or maximum aggregation functions. This could be expanded to other functions but this would at the same time affect the user friendliness, both for the end users as well as the collaborators contributing and maintaining such a system.

From the general observations made which derived the domain model it is not possible to conclusively state that it would be possible to model any narrative, apply the relevant beliefs in form of rules and see a reasonable and expected outcome while ensuring the tool would not contradict itself at a certain point. A specific case proving such contradiction and showing the possible workaround is overly elaborate would need to be shown which would require further work. The functionality and the ad hoc rules and attributes derived throughout the thesis were mainly thought of in the context of the narrative example from section 1.1 that guided development of most of the ideas presented in the thesis. As such, the proposed solution was shown to apply well to this situation, which is however not generalizable to other narratives.

6.2 Validation

Fulfillment of the stated objective of the proposed solution was validated through qualitative interviews.

The qualitative interview guided the interviewees and largely followed section 4.6, i.e. the imagined usage flow, which itself relies yet again on the narrative example from sec-

tion 1.1. The interview process is described in the following list, where each step explains what was introduced and shown to the interviewee, what questions were asked (optionally a reason is given for why the question is asked) and lastly, in the case the interviewee missed something and it is important for the rest of the interview to proceed from a common point, further explanation filling in the missing information is provided.

1. Introduction

- **Introduced:** General introduction of the topic, the approach and intended audience with the following snippet: "The topic of the thesis has to do with what you might have heard being talked about as fake news, mis- and dis-information, propaganda, myths, urban legends and definitely many other similar terms. I will refer to these as false information. It is a topic we have heard much more about in the last 6 years, that is since the 2016 Brexit referendum and US elections. It is also a topic that affects all of us personally to a different degree. Nowadays, everyone gets subjected to false information and it is easy to not realize this and believe it. It is neither a problem where smart people necessarily excel at avoiding believing false information. Nevertheless, there are many people, for example on the political fringe, that maybe because of their identity, maybe because of the people and community they associate with, hold some extreme beliefs. These are people that are not considered and not the focus of the proposed solution. The proposed solution for addressing the problem of false information is intended to help reasonable open-minded people that can be argued with, that respect the truth and that do not construct their own alternative reality with informed critical thinking. I would like you to keep this in mind while answering the questions and evaluating the proposed solution."

2. Twitter post

- **Introduced:** Screenshot of the original tweet from Eric Tucker is shown (fig. 4.12). It is explained that this tweet is from November 9, 2016, a day after the US elections between Trump and Clinton, when massive protests erupted across the United States against the election of Trump.
- **Questions:** Can you give a basic summary of what you see? (Asked to establish a common understanding)
- **Explanation:** The tweet allegedly shows a photo of buses that transported paid protesters to protest Trump being elected a president. That would mean the protests are mainly artificially organized (paid for) and the general population is not really upset as compared to the protests naturally (organically) occurring and people being upset about the election.

3. Twitter post with CTSS browser extension

- **Introduced:** An edited version of the previous screenshot with the CTSS overlay is shown (fig. 4.13).
- **Questions:** What does it show? Does this help you to get a more accurate picture of the situation? Why so or why not?
- **Explanation:** It shows related articles in two categories, one supporting the statement and the other opposing it, and a "See the full picture" link.

4. Credibility graph introduced

- **Introduced:** "See the full picture" link from the CTSS overlay is followed and a credibility graph for that narrative, first without the credibility propagation turned on, is shown (fig. 4.14).
- **Questions:** Can you have a look and explain what you see? Does this give you any more information than the previous view? Is there anything interactive? What do you think happens when the switch is flicked and credibility propagation turned on?
- **Explanation:** The interviewee is made aware of a legend in the bottom left. There is a statement at the center, documents, publishers and authors. The statement is meant to summarize the narrative. These are connected with arrows of different colors with different meanings: support, oppose, neutral and other relations. The other relation dashed line in this case indicates a dependency where the linked documents rely on the original tweet. There is also a switch in the top left corner with a description credibility propagation.

5. Credibility propagation mode turned on

- **Introduced:** The credibility propagation mode is turned on through the switch and the view changes (fig. 4.15).
- **Questions:** What has changed compared to the previous view? What do you think the numbers mean? Which number is of the most interest? Is there anything interactive (apart from the switch)?
- **Explanation:** Numbers to the left of all the descriptions are shown. These represent the credibility. These credibilities propagate from the author and publishers to the documents and to the statement. The most important number is that of the statement. The statement changes color to orange. One of the documents and arrows going into it and out of it are gray. The reason for why the document is grayed out is explained (exclusion rule). Finally, the desired outcome in the bottom right appears that mimics the credibility of the statement. The interviewee should wonder about how are these values calculated. An example of this is shown next.

6. Snopes' attributes and rules

- **Introduced:** The Snopes node is clicked on and activated (fig. 4.17). What is Snopes is explained.
- **Questions:** What has changed compared to the previous view? Can you explain what you see? What do you think is the meaning of these attributes and rules? Is there anything interactive? What do you think happens when we change the value on the slider?
- **Explanation:** Two additional nodes connected to Snopes appear. The legend is extended with explanations for these nodes. An attribute is something that characterizes that particular node so the attribute here says that Snopes is a fact checker and since the maximum value that can be possibly set for this attribute is 0.5, indicated by the light gray color, the value 0.25 indicates we consider Snopes to be a sort of reliable fact checker. This attribute is also unchangeable and where does it come from will be discussed at the end. The rule is interactive and you as a user can change it. It indicates how important for you personally is the fact that a publisher in this case, is a fact checker. This

allows you to modify the credibility of the different nodes that then spreads through the rest of the structure as indicated by the arrows.

7. Changing a rule affecting single publisher

- **Introduced:** The slider value is changed from 0.6 to 0.2 (fig. 4.18).
- **Questions:** What has changed compared to the previous view? Do the changes in the credibility values make sense to you? What do you think of the change in the statement credibility (+0.004) based on the change of the rule (-0.4)? Is it interesting and useful to you that you can change this value and see the result?
- **Explanation:** The credibility of Snopes decreased from 0.65 to 0.55, the credibility of document "Anti-Trump Protesters Bused Into Austin, Chicago (Snopes - False)" decreased from 0.575 to 0.525 and the credibility of the statement increased from 0.376 to 0.38. With the last change, the value for the desired outcome also changed. The basic idea about the calculations and why the values changed in a way that they did is explained.

8. Document's rules

- **Introduced:** The document "They found the buses" related to "Free Republic" is clicked on fig. 4.19.
- **Questions:** What has changed compared to the previous view? What do you think is the meaning of the rule? What is the rule connected to? Can you identify why is it connected exactly to these four documents? Does it make sense to you? What do you think happens when we change the value on the slider?
- **Explanation:** A "Credit Excluded Documents" rule appears. Since it is a rule, it can be changed via the slider appearing next to it. It has four arrows pointing to documents that are all connected with the dashed arrow to the "(Excluded) Anti-Trump protesters are not organic" original document. At different points in the interview, but at latest in this step, an explanation was given as for the reason why one of the documents (the original tweet) is excluded and grayed out.

9. Changing a rule affecting multiple documents

- **Introduced:** The slider value is changed from 0.5 to 0.2 (fig. 4.20).
- **Questions:** What has changed compared to the previous view? Do the changes in the credibility values make sense to you? What do you think of the change in the statement credibility (-0.081) based on the changed of the rule (-0.3)? Is it interesting and useful to you that you can change this value and see the result?
- **Explanation:** Credibility of the four documents connected to the rule decreases by 60%, which correspond to our change from 0.5 to 0.2. This also decreases the credibility of the statement from 0.38 to 0.299 and changes the background of the statement from orange to red. With the last change, the value for the desired outcome also changed and the text changed from "Oppose" to "Strongly oppose" and it also changed its color from orange to red.

10. Overview of narratives

- **Introduced:** Describe the usage scenario where an interviewee uses this on other narratives as well and can get an overview of the narratives they showed an interest in (fig. 4.21).
- **Questions:** What do you think of such an overview? Do you understand the difference between the current and desired outcome? Does the reported status make sense to you? Is such an overview useful? Does it make sense that the rules apply across other narratives with the same strength?
- **Explanation:** First of all, when setting rules, a rule applies with the same strength in all other narratives where it fits the situation that it was made for. For example, the reliable fact checker rule we have seen would apply with the same strength in other narratives when a fact checker source, which could be a different one than Snopes, is involved. So when a user encounters another narrative and sets the rules in accordance to their own conclusion, some of these rules might be present and apply in other narratives the user has already gone through and therefore changing the outcome there as well.

The overview is an example where seven narratives are shown, each in its own row. Each row then has the statement of the narrative, currently calculated credibility of the statement, the desired outcome which is the credibility that the statement reached when encountered for the first time and which can also be explicitly changed (fig. 4.22), and finally a status for the narrative comparing the current and the desired outcome. When the difference between the two values is not too big, it is reported as "OK" with green background. When both of the values still lean towards the same conclusion, either support or oppose, and the difference between the two values is lower than 0.1, it is reported as a warning with orange background and a relevant text. Finally, when the difference between the two values is bigger than 0.1 or when the two values differ in the stance they lean towards, e.g. one leaning support and the other leaning oppose, this is reported as a problem with a red background.

11. Collaborative aspect

- **Introduced:** It was previously mentioned that the attributes are set to some value unchangeable by the user. The idea behind how all this works is that it is a collaborative Wikipedia-like project where it is the collaborators that not only create and assign the attributes to authors and publisher and who discuss and decide what specific value it should take, but also contribute the narratives, with their documents, authors and publishers, and with the different relationships between them, as well as create the rules and program their behavior. It is directly likened to how collaboration works on Wikipedia where Wikipedians discuss articles and other contributions on [Talk pages](#).
- **Questions:** What do you think of such a collaborative approach to solve this problem? What are the pros and cons? Would you be comfortable using such a system knowing it is maintained by a community of ordinary people or this a problem requiring a different solution than that of Wikipedia?

12. General evaluation and grading

- **Questions:** In general, what do you think of the system? How useful would it be in your everyday life? What do you like and dislike? What would be your suggestions for improvement? Finally, how would you grade the proposed

solution on a scale from 0 to 10, 10 being the best?

In total, five interviews were conducted and evaluated. The interviewees were author's friends and colleagues from work. Given the time window available, having a bigger and randomized sample size was not feasible.

The interviews are discussed in the following list which references the interview steps from above. The interviewee's answers are summarized and commented.

2. Some interviewees had a problem understanding the tweet and the underlying claim it makes. After clarification from the interviewer, everyone was able to summarize the claim correctly. The degree of misunderstanding could have been decreased by using a more recent topic which is also more close to European audience.
3. All except one interviewees were able to recognize what the overlay shows. All agreed that having this additional contextual information is better than not having it. Two specifically mentioned they would be curious to click and read the opposing information.
4. Two interviewees had no problem recognizing what is shown and were able to describe everything of interest. Others described the credibility graph and rest of the interface only partially. When asked about a possible improvement for better understanding, some stated they are "just not used looking at such things" and that it would be better if they "could investigate it interactively" expecting additional explanations provided when elements are clicked or hovered. Everyone recognized more information is given than in the previous view but uncertainty was expressed about its utility. Everyone also recognized the credibility propagation switch is an interactive element (some needed to be nudged to pay attention to top left corner).
5. By having the two figures next to each other, everyone was able to tell the differences. As for the meaning of the numbers, one did not know at all and the rest said "it has something to do with credibility". After encouraging everyone to have a look at the legend, some intuitive explanations for what the number means were given. One interviewee guessed the number "measures the proportion of previous articles that turned out to be correct". Two interviewees did not have an idea about how the number was reached but they expressed an interest to find out more, one specifically mentioned "by clicking or hovering it". All except one recognized the credibility value of statement is the most important. One stated that "it is helpful to see the conclusion by color" and that they "immediately knew what to think of the issue". On the one hand, it is encouraging to see the interest to form an informed conclusion, on the other hand, this is a sign of overconfidence in the tool itself which goes against the main purpose of aiding informed critical thinking. As for interactive elements, few guessed that interacting with the elements might give them additional information about how was its credibility reached and everyone noticed the desired outcome slider in the bottom right corner.
6. Almost everyone complained about the attribute and rule element to be overly complicated with how it shows the value with the proportionate color fills. No one was able to accurately explain the meaning of attributes and rules, although everyone understood that attribute is immutable and rule is mutable through the slider. Only two understood that rule could be positive or negative. Another interviewee thought that one rule can be both positive and negative, depending on the value chosen on the slider. After explaining the attributes and rules more thoroughly, three interviewees got a good understanding of its meaning but said they would never be able

to understand it just from the legend, two were still unsure, especially about the meaning of a rule.

7. Everyone was able to find the differences by comparing the two figures, understood why it changed all the values that it did and also why it increased the credibility of the claim as a response to decreasing importance of fact checkers. No one mentioned anything about the magnitude of the change. Some interviewees stated it is interesting to be able to change the value and see the response. Some started to express a concern that this would lead to certain people setting the parameters such that it shows what they want and disregarding any connection to how they actually think about it.
8. All interviewees recognized the active document that was clicked and that a rule appeared next to it. No one but one was able to express what it means except that it would affect credibility of the documents when changed. One interviewee saw a connection to the excluded document right away and was also able to connect that the rule points to those documents which are related to the excluded document. Others found the connection when more context was explained and when explicitly asked about it. In the end everyone understood the meaning of the rule, correctly predicted what would happen when the rule is changed and were also able to explain why it would happen. However, interviewees reiterated that only the additional explanation made it clear to them and not anything from the figure itself.
9. As mentioned in the previous point, everyone already accurately predicted the behavior and the change made sense to them. Two noted a difference in what the desired outcome said. Other two explicitly mentioned the difference in magnitude of the change that it caused as compared to changing the first rule and were more satisfied with a comparatively bigger change.
10. Two did not understand what the overview showed and what the outcome and status columns meant. The remaining three made the connection and were able to correctly describe what current and desired outcome referred to. Two of them also guessed correctly that the status refers to the difference between the current and desired outcome. Four interviewees said they understood the overview after the explanation as well as its utility and agreed it is a good idea to have such an overview. Although the last one understood what the overview showed, they did not see any value in it. One interviewee also expressed confusion with the rules applying to other narratives with the same strength stating that "in different situations, different context applies so I would expect to be able to change the rules individually".
11. No one expressed any deep concern or that they would prefer a different approach. Everyone agreed that "two heads are better than one" as the proverb goes and that such an approach is more in line with such thinking. Two expressed that "if it works for Wikipedia, why couldn't it work for something else". The fact that anyone can join in was positively mentioned by one interviewee. The concerns that were expressed had to do with potential abuse which go in line with the expectations outlined in section 4.5.1.
12. Everyone expressed an interest in such a solution, notable quote is "I'm surprised such a thing doesn't already exist and isn't in wide use". Two interviewees expressed their frustration of not knowing what and who to trust and said they would welcome such a solution that would help them in this. There was a split between the interviewees in terms of appreciating the utility of being able to change the rules themselves.

One side said they would not use it and do not see any value in it and would rely solely on the community setting the default outcome in such a way that represents an average opinion on the narrative. The other side expressed an interest since it would allow them to set their own preferences and also understand the thought process of others assuming it would be possible to share a narrative with their settings. At the same time, this side appreciated it would be of value to be able to have access and link to the "default" or "average" evaluation of a narrative in the case where one would try to disseminate neutral information. Everyone disliked the view of credibility graph with credibility propagation mode disabled and said that "it doesn't show anything interesting".

As for the final question about grading, as previously mentioned, all of the interviewees have a direct relationship with the interviewer and author of the thesis which introduces bias. When the final question for grading the proposed solutions was asked, interviewees were encouraged to judge it objectively and disregard the existing relationship. Even so, this intervention does not eliminate the bias. With that said, based on a sample size of five interviewees, the proposed solution attained an average rating of 7.8.

On balance, the interviewees liked the idea when they understood it. However, this understanding presents a challenge since the individual ideas and building blocks had to be thoroughly explained before a satisfactory understanding was attained and the tool itself was not intuitive and self-evident. The interviewees only had the patience to understand the solution since they were in the artificial setting of an interview instead of in their everyday life. It remains an open question as to what would improve an understanding in a natural setting.

Another major concern is that some interviewees thought of the proposed solution as a shortcut to getting a quick conclusion instead of something that should encourage their critical thinking. This goes against the fundamental objective of the solution. Thus, it seems that while users of such system would end up being more accurately informed about narratives, it would not necessarily increase the time and effort spent on critical thinking for everyone.

As such, usage of the proposed solution as shown in section 4.6 involving a browser extension and visualization of credibility graph offers only partial answer to RQ3 ("What is an effective way to aid people in informed critical thinking?"). While the tool makes it easier to start the process of critical thinking for those already inclined towards such activity, it does not seem to encourage it for others.

6.3 Limitation

The proposed solution looks at the problem of false information from a standpoint of credibility only considering legal person entities and not considering the content. A document could be anonymous, i.e. both author and publisher unknown, but contain good substantiated arguments. Additionally, since the solution involves subjectivity, these arguments would not necessarily have to be good and substantiated, and just play into biases of a user who might like them and therefore consider such document highly impactful to their perception of a narrative. Looking at the problem of subjectivity and false information from a standpoint of logically sound argumentation is not considered and leaves many narratives unresolved.

Even if the proposed solution is implemented as conceptualized it does not change the fact that the task of aiding critical thinking remains difficult. Although the tool can certainly

Criteria	Considered?	Explanation
Social consensus: Do others believe it?	No	Not considered at any point directly.
Support: Is there much supporting evidence?	Partially	Considered through the relations possible in a credibility graph. However, this relies on the credibility graph being accurate and the proposed solution does not in any way propose how to ensure this condition.
Consistency: Is it compatible with what I believe?	Partially	Considered through the constraints put on the rule parameter space which can then inform about incompatible beliefs. However, it was not proved that these rules can model compatibility flawlessly.
Coherence: Does it tell a good story?	No	Not considered since, as mentioned previously, the proposed solution does consider and analyze the content itself for logical coherence.
Credibility: Does it come from a credible source?	Yes	The proposed solution is built mainly around this consideration and is its main focus.

Table 6.1: Consideration of five criteria people use for judging truth. Criteria names and their descriptions are taken directly from [79, p. 87]. See section 6.4 for discussion of future work to improve the coverage of these five criteria.

be helpful in explicitly showing refutation of certain evidence crucial to certain incoherent beliefs, it is an open question whether this would help with belief revision:

“Once formed,” the researchers observed dryly, “impressions are remarkably perseverant.” ...Even after the evidence “for their beliefs has been totally refuted, people fail to make appropriate revisions in those beliefs,” the researchers noted. ([27])

From a psychological standpoint, considering the five criteria people use for judging truth identified in [79] the proposed solution only considers some as can be seen in table 6.1.

6.4 Future work

During the qualitative interviews, given the state of the implemented prototype, it was not feasible to test the tool itself. It would be interesting to see whether the largely positive evaluation of the solution persist while using the actual tool instead of partially imagining its functionality from wireframes. A possible idea for future work is to implement a mature prototype which could be used exactly as proposed in section 4.6 and subject it to a further validation. It also turned out that the selection of a six year old American narrative for European audience was not a great fit as European audience is not necessarily interested in and aware of American issues. The solution should then be validated on a more relevant narrative.

A few interesting ideas were brought up as part of the qualitative interviews. One interviewee expressed an interest in being able to share their views (using their setting of rules) on the narrative with others as well as being able to see the views of others. This social aspect is no doubt interesting to consider as it directly relates to encouraging discussion which by extension could lead to more critical thinking.

As summarized in table 6.1, since the proposed solution looked at the problem of false information only from the perspective of credibility, other considerations were left intact and in the best only accidentally partially addressed. It would be interesting to consider further possibilities for improvements through the other considerations.

Let's briefly consider what would a look from the coherence perspective mean. This would involve analyzing the content through a formal logic which would bring with it many complications. Nevertheless, even a partial chain of logically proved reasoning could be an interesting addition. Considering such an addition from the perspective of the established framework, such addition could contribute to its credibility through an attribute for its logical soundness.

Considering the consistency criteria which is considered in this thesis through the development of an argument about rules constraining the space of acceptable parameters, it would be interesting to go in more depth and see what improvements could be brought by considering a more flexible non-linear functions for the rules.

7 Conclusion

The complexity and the sheer magnitude of the researched problem offer various approaches for addressing it. After initial considerations of other approaches not discussed in the thesis, the selected approach was chosen for its virtues of transparency, unassertiveness in stipulating truth and mass collaboration.

Based on the largely positive evaluation from the qualitative interviews, the approach seems promising. The most worthwhile next step regarding future work seems to be further validation of the solution on a mature prototype that would be in line with the proposed design and user flow, which would model a more recent narrative more relevant for the test audience (European audience is not as aware of American issues).

The thesis also develops an argument for addressing the interesting problem of cognitive biases through technological solution whereby the more egregious of cognitive biases that would often be associated with incoherent thinking could be uncovered.

Although imperfect in many aspects, this work represents a solid attempt to address the problem of false information in a unique way. The most encouraging finding from the interviewees worth wrapping up with is based on paraphrasing the unanimously asked rhetorical question "how does this not exist already?".

.

Bibliography

- [1] Full Fact. (Apr. 2020). "The Full Fact Report 2020 - Fighting the causes and consequences of bad information." From: <https://web.archive.org/web/20210616040905/> <https://fullfact.org/blog/2020/apr/full-fact-report-2020/>.
- [2] ——, (Aug. 2019). "The cabbage myth returns: EU cabbage regulations don't have 26,911 words."
- [3] Snopes. (Mar. 2001). "Are Government Regulations for Cabbage 27,000 Words Long?"
- [4] M. T. Bastos and D. Mercea, "The Brexit Botnet and User-Generated Hyperpartisan News," *Social Science Computer Review*, vol. 37, no. 1, pp. 38–54, 2019. DOI: 10.1177/0894439317734157.
- [5] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, May 2017. DOI: 10.1257/jep.31.2.211.
- [6] S. Kogan, T. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," *SSRN Electronic Journal*, Jan. 2018. DOI: 10.2139/ssrn.3237763.
- [7] C. Carvalho, N. Klagge, and E. Moench, "The persistent effects of a false news shock," *Journal of Empirical Finance*, vol. 18, no. 4, pp. 597–615, 2011, ISSN: 0927-5398. DOI: 10.1016/j.jempfin.2011.03.003.
- [8] S. Langlois, *This day in history: Hacked AP tweet about White House explosions triggers panic*, <https://www.marketwatch.com/story/this-day-in-history-hacked-ap-tweet-about-white-house-explosions-triggers-panic-2018-04-23>, Apr. 23, 2018.
- [9] M. Takayasu, K. Sato, Y. Sano, K. Yamada, W. Miura, and H. Takayasu, "Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study," *PLOS ONE*, vol. 10, pp. 1–18, Apr. 2015. DOI: 10.1371/journal.pone.0121443.
- [10] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13 Companion, Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 729–736, ISBN: 9781450320382. DOI: 10.1145/2487788.2488033.
- [11] A. Y. Chua, S.-M. Cheah, D. H.-L. Goh, and E.-P. Lim, *Collective rumor correction on the death hoax of a political figure in social media*, <https://aisel.aisnet.org/pacis2016/178/>, Pacific Asia Conference on Information Systems, 2016.
- [12] Wikipedia contributors, *List of common misconceptions* — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=List_of_common_misconceptions&oldid=1054322276, [Online; accessed 9-November-2021], 2021.
- [13] M. Fisher, J. W. Cox, and P. Hermann, *Pizzagate: From rumor, to hashtag, to gunfire in D.C.* https://web.archive.org/web/20210405194518/https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunned-down-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html, Dec. 6, 2016.
- [14] Wikipedia contributors, *Pizzagate conspiracy theory* — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Pizzagate_conspiracy_theory&oldid=1058537214, [Online; accessed 5-December-2021], 2021.
- [15] ——, *2021 United States Capitol attack* — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=2021_United_States_Capitol_attack&oldid=1058876651, [Online; accessed 6-December-2021], 2021.

- [16] ——, *Don't Look Up* (2021 film) — Wikipedia, The Free Encyclopedia, [https://en.wikipedia.org/w/index.php?title=Don%27t_Look_Up_\(2021_film\)&oldid=1064842030](https://en.wikipedia.org/w/index.php?title=Don%27t_Look_Up_(2021_film)&oldid=1064842030), [Online; accessed 10-January-2022], 2022.
- [17] World Health Organization, *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation*, <https://www.who.int/news-room/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation>, Sep. 23, 2020.
- [18] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nature Human Behaviour*, vol. 5, no. 3, pp. 337–348, Feb. 2021. DOI: 10.1038/s41562-021-01056-1. [Online]. Available: <https://doi.org/10.1038/s41562-021-01056-1>.
- [19] K. Ognyanova, D. Lazer, R. E. Robertson, and C. Wilson, *Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power*, <https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/>, Jun. 2, 2020.
- [20] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, 2019, The Cognitive Science of Political Thought, ISSN: 0010-0277. DOI: 10.1016/j.cognition.2018.06.011.
- [21] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science*, vol. 31, no. 7, pp. 770–780, 2020. DOI: 10.1177/0956797620939054.
- [22] J. Baldasty, *Fake news and misinformation: Why teaching critical thinking is crucial for democracy*, <https://www.washington.edu/provost/2018/04/23/fake-news-and-misinformation-why-teaching-critical-thinking-is-crucial-for-democracy/>, Apr. 23, 2018.
- [23] M. Ingram, *Europe tries to fight hate, harassment, and fake news without killing free speech*, <https://www.cjr.org/innovations/europe-youtube-facebook-free-speech.php>, 2018.
- [24] W. Turvill, *Quintopoly? Five tech companies now earn 46% of global ad revenues as news media left behind*, <https://pressgazette.co.uk/global-advertising-spend-2020-quintopoly/>, 2021.
- [25] J. Dunn, *The tech industry is dominated by 5 big companies — here's how each makes its money*, <https://www.businessinsider.com/how-google-apple-facebook-amazon-microsoft-make-money-chart-2017-5>, 2017.
- [26] B. Jackson, *Is this a great job, or what?* <https://www.factcheck.org/2003/12/is-this-a-great-job-or-what/>, Dec. 5, 2003.
- [27] E. Kolbert, *Why Facts Don't Change Our Minds*, <https://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds>, Feb. 19, 2017.
- [28] E. Barker, *This Is How To Change Someone's Mind: 6 Secrets From Research*, <https://www.bakadesuyo.com/2019/12/change-someones-mind/>, Dec. 2019.
- [29] A. D. Holan, *The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking*, <https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/>, 2018.

- [30] G. Kessler, *About The Fact Checker*, <https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker/#:~:text=to%20the%20editors.-,The%20Pinocchio%20Test,-Where%20possible%2C%20we>, 2017.
- [31] C. Hallman, *50 Cognitive Biases to be Aware of so You Can be the Very Best Version of You*, <https://www.titlemax.com/discovery-center/lifestyle/50-cognitive-biases-to-be-aware-of-so-you-can-be-the-very-best-version-of-you/>, 2020.
- [32] D. Evon, *Anti-Trump Protesters Bused Into Austin, Chicago*, <https://www.snopes.com/fact-check/anti-trump-protesters-bused-into-austin/>, Nov. 11, 2016.
- [33] S. Maheshwari, *How Fake News Goes Viral: A Case Study*, <https://www.nytimes.com/2016/11/20/business/media/how-fake-news-spreads.html>, Nov. 20, 2016.
- [34] S. Wineburg and S. McGrew, *Evaluating information: The cornerstone of civic online reasoning*, <https://purl.stanford.edu/fv751yt5934>, 2016.
- [35] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. DOI: 10.1126/science.aap9559.
- [36] Full Fact, Africa Check, and Chequedo, *Fact checking doesn't work (the way you think it does)*, <https://fullfact.org/blog/2019/jun/how-fact-checking-works/>, 2019.
- [37] D. Flamini, *Most Republicans don't trust fact-checkers, and most Americans don't trust the media*, <https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-dont-trust-the-media/>, Jul. 3, 2019.
- [38] A. Tapia, *The 20 Most Popular Podcasts in America Right Now*, <https://www.newsweek.com/most-popular-podcasts-america-right-now-joe-rogan-daily-crime-1650687>, Nov. 28, 2021.
- [39] JRELibrary, *Joe Rogan Experience Podcast Stats*, <https://jrelibrary.com/articles/stats/>, Dec. 31, 2021.
- [40] Wikipedia contributors, *iOS version history — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=IOS_version_history&oldid=1062984387#iOS_15_-_iPadOS_15, [Online; accessed 1-January-2022], 2021.
- [41] D. Funke, *Should we stop saying 'fake news'*? <https://www.poynter.org/fact-checking/2017/should-we-stop-saying-fake-news/>, Dec. 14, 2017.
- [42] European Commission and DG Connect, *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation*. Publications Office, 2018. DOI: doi/10.2759/0156.
- [43] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, ISSN: 02767783. [Online]. Available: <http://www.jstor.org/stable/25148625>.
- [44] Wikipedia contributors, *National Rifle Association — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=National_Rifle_Association&oldid=1064542728, [Online; accessed 10-January-2022], 2022.
- [45] S. Wineburg and S. McGrew, "Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information," *Teachers College Record*, vol. 121, no. 11, pp. 1–40, 2019. DOI: 10.1177/016146811912101102.
- [46] M. Babakar and W. Moy, *The State of Automated Factchecking*, https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf, From: <https://fullfact.org/blog/2016/aug/automated-factchecking/>, Aug. 17, 2016.
- [47] L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga, *Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection*, 2020. arXiv: 1809.08193 [cs.CL].
- [48] D. Flamini, *9 fact-checking lessons from Global Fact 6*, <https://www.poynter.org/ifcn/2019/9-fact-checking-lessons-from-global-fact-6/>, Jun. 21, 2019.

- [49] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," 2018. DOI: 10.1145/3395046. eprint: arXiv:1812.00315.
- [50] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, *A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking*, <https://arxiv.org/abs/1911.01214>, 2019.
- [51] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating Fake News: A Survey on Identification and Mitigation Techniques," *CoRR*, 2019. [Online]. Available: %5Curl%7Bhttps://arxiv.org/abs/1901.06437%7D.
- [52] Wikipedia contributors, *NewsGuard — Wikipedia, The Free Encyclopedia*, <https://en.wikipedia.org/w/index.php?title=NewsGuard&oldid=1064045102>, [Online; accessed 10-January-2022], 2022.
- [53] P. Mathur, *Startups battle the spread of fake news surrounding the pandemic*, <https://pitchbook.com/news/articles/how-startups-are-battling-the-spread-of-fake-news-during-the-pandemic>, Apr. 6, 2020.
- [54] Merriam-Webster, *Document*, in *Merriam-Webster.com dictionary*. [Online]. Available: <https://www.merriam-webster.com/dictionary/document> (visited on 01/04/2022).
- [55] Lexico, *Document*, in *Lexico.com dictionary*, Oxford University Press. [Online]. Available: <https://www.lexico.com/en/definition/document> (visited on 01/04/2022).
- [56] ——, *Statement*, in *Lexico.com dictionary*, Oxford University Press. [Online]. Available: <https://www.lexico.com/en/definition/statement> (visited on 01/04/2022).
- [57] Wikipedia contributors, *Gossip magazine — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Gossip_magazine&oldid=1059439233, [Online; accessed 13-January-2022], 2021.
- [58] K. Finley, *This News-Writing Bot Is Now Free for Everyone*, <https://www.wired.com/2015/10/this-news-writing-bot-is-now-free-for-everyone/>, 2015.
- [59] Wikipedia contributors, *Medium (website) — Wikipedia, The Free Encyclopedia*, [https://en.wikipedia.org/w/index.php?title=Medium_\(website\)&oldid=1062113822](https://en.wikipedia.org/w/index.php?title=Medium_(website)&oldid=1062113822), [Online; accessed 4-January-2022], 2021.
- [60] Twitter, *COVID-19 misleading information policy*, <https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy>, Dec. 2021.
- [61] Wikipedia contributors, *Persistent identifier — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Persistent_identifier&oldid=1049039303, [Online; accessed 4-January-2022], 2021.
- [62] S. Levin, *Is Facebook a publisher? In public it says no, but in court it says yes*, <https://www.theguardian.com/technology/2018/jul/02/facebook-mark-zuckerberg-platform-publisher-lawsuit>, Jul. 3, 2018.
- [63] N. Clegg, *Combating COVID-19 Misinformation Across Our Apps*, <https://about.fb.com/news/2020/03/combating-covid-19-misinformation/>, Mar. 25, 2020.
- [64] A. Selyukh, *Feeling Sidelined By Mainstream Social Media, Far-Right Users Jump To Gab*, <https://www.npr.org/sections/alltechconsidered/2017/05/21/529005840/feeling-sidelined-by-mainstream-social-media-far-right-users-jump-to-gab>, May 21, 2017.
- [65] Lexico, *Credibility*, in *Lexico.com dictionary*, Oxford University Press. [Online]. Available: <https://www.lexico.com/en/definition/credibility> (visited on 01/04/2022).
- [66] E. Ferri, *How to extract text from memes with Python, OpenCV and Tesseract OCR*, <https://towardsdatascience.com/extract-text-from-memes-with-python-opencv-tesseract-ocr-63c2ccd72b69>, Dec. 9, 2020.
- [67] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," *Pattern Recognition*, vol. 79, pp. 242–259, 2018, ISSN: 0031-

3203. DOI: <https://doi.org/10.1016/j.patcog.2018.02.017>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320318300670>.
- [68] Wikipedia contributors, *Automatic image annotation — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Automatic_image_annotation&oldid=1022624316, [Online; accessed 8-January-2022], 2021.
- [69] D. Subramanian, *Easy Speech-to-Text with Python*, <https://towardsdatascience.com/easy-speech-to-text-with-python-3df0d973b426>, Apr. 7, 2020.
- [70] I. Iljin, *Building a news aggregator from scratch: news filtering, classification, grouping in threads and ranking*, <https://towardsdatascience.com/building-a-news-aggregator-from-scratch-news-filtering-classification-grouping-in-threads-and-ranking-7b0bbf619b68>, Jan. 27, 2020.
- [71] J. Tan, X. Wan, and J. Xiao, “From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 4109–4115. DOI: 10.24963/ijcai.2017/574. [Online]. Available: <https://doi.org/10.24963/ijcai.2017/574>.
- [72] J. Murtha, *How fake news sites frequently trick big-time journalists*, https://www.cjr.org/analysis/how_fake_news_sites_frequently_trick_big-time_journalists.php, May 26, 2016.
- [73] E. Mugendi, *A Fake Poll Website Shows Just How Crafty Kenyan Fake News Is Getting*, <https://pesacheck.org/a-fake-poll-website-shows-just-how-crafty-kenyan-fake-news-is-getting-35cf90aeb64>, Oct. 5, 2017.
- [74] Meta, *URL - Graph API*, <https://developers.facebook.com/docs/graph-api/reference/url/#fields>. (visited on 01/14/2022).
- [75] E. Tucker, *Why I'm removing the "Fake Protests" Twitter post*, <https://erictucker.wordpress.com/2016/11/11/why-im-considering-to-remove-the-fake-protests-twitter-post/>, Nov. 11, 2016.
- [76] Wikipedia contributors, *Free Republic — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Free_Republic&oldid=1061471704, [Online; accessed 15-January-2022], 2021.
- [77] ——, *The Gateway Pundit — Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=The_Gateway_Pundit&oldid=1064948391, [Online; accessed 15-January-2022], 2022.
- [78] A. Kishore, M. Bolin, D. Crockford, M. Nanassy, A. Eisenberg, and J. Tucker, *Summary of features - JSON5 | JSON for Humans*, <https://json5.org/#summary-of-features>. (visited on 01/21/2021).
- [79] N. Schwarz, E. Newman, and W. Leach, *Making the truth stick & the myths fade: Lessons from cognitive psychology*, <https://behavioralpolicy.org/articles/making-the-truth-stick-the-myths-fade-lessons-from-cognitive-psychology/>, Feb. 17, 2017.

Technical
University of
Denmark

Richard Petersens Plads, Building 324
2800 Kgs. Lyngby
Tlf. 4525 1700

www.compute.dtu.dk