Towards a Prototype to Leverage Archival Diplomatics to Develop a Framework to Detect and Prevent Fake Videos

Nicholas Rivard
School of Computer Science
Carleton University
Ottawa, Ontario, Canada
0009-0002-3068-5496

Hoda Hamouda
School of Information
The University of British Columbia
Vancouver, British Columbia, Canada
0000-0002-0927-9744

Victoria Lemieux
School of Information
The University of British Columbia
Vancouver, British Columbia, Canada
0000-0003-1339-6289

Tracey P. Lauriault
School of Journalism and Communication
Carleton University
Ottawa, Ontario, Canada
0000-0003-1847-2738

Michel Barbeau
School of Computer Science
Carleton University
Ottawa, Ontario, Canada
0000-0003-3531-4926

***Abstract*—This paper is about the trustworthiness, particularly authenticity, of videos produced and posted by citizen journalists. An archival diplomatics perspective is adopted. It focuses, among other things, on authentication by means of assessment of the form of records rather than simply their content. First, we focus on a study examining whether citizen journalists can ascertain if a news video is fake. The results of that study informed an approach to automate the assessment of a news video's authenticity, building on archival diplomatics' principles and methods of formal analysis. The study data and a set of authenticity criteria are then modeled into a knowledge graph. We suggest that graph analytics techniques may complement archival diplomatics in helping to assess the authenticity of videos. They can produce relative authenticity scores that may help computational archivists and others appraise the veracity and authenticity, of multi-media records with precision.**

***Index Terms*—Trustworthiness, Truth finding, Veracity, Authenticity, Fake video, Citizen journalism video, Graph analytics, Knowledge graph, Archival diplomatics**

I. Introduction

Identifying fake *online citizen journalism videos* is challenging in an era where Artificial Intelligence (AI)-generated content is relatively easy to create. For example, despite the authenticity of citizen-captured footage, some videos may falsely connect to entirely different events [1]. Also, new AI-based tools make it increasingly easy to alter the visual, audio, and metadata components of an online video [2], highlighting the need for a thorough assessment of the veracity of multimedia artifacts beyond a content analysis.

This contemporary digital and AI context makes it imperative to develop new approaches to assess the veracity of online videos, most notably their authenticity.

We acknowledge financial support from Carleton University's REALISE Seed Grant program and InterPARES Trust AI project.

This paper builds upon earlier efforts to apply archival science approaches [3], specifically archival diplomatics, to assess the authenticity of Citizen Journalism Videos (CJVs) uploaded into online platforms.

Archival diplomatics is the science and practice concerned with the creation and use of documents. It is defined as the discipline that studies the genesis, form, and transmission of records. It also encompasses their relationship with the facts represented in them, including information about their creator, to identify, evaluate, and communicate their true nature [3]. In archival science, records are recorded information (documents) regardless of form or medium created, received and maintained by an agency, institution, organization or individual in pursuance of their legal obligations or business transaction [4]. Archival diplomatics approaches, therefore, consider the context of a video's creation (i.e., its origins or provenance) and not only the video's content. Importantly, archival diplomatics also assesses the elements of the video's documentary form, such as its audio-visual layers, frame rates, audio sampling rates, and descriptive and technical metadata [5]. We suggest that archival diplomatics augments approaches that focus solely on the assessment of a record's content or provenance, especially for situations where a video's creator or precise origins cannot be ascertained, as is often the case with CJVs [6], -authentic, or AI-generated fake videos uploaded to online platforms, such as *X* (formerly Twitter) or Youtube.

Building on prior work [1], the authors here feature a prototype tool applying archival diplomatics principles to determine the authenticity of videos, thus applying a previously proposed framework [7] for authenticity testing which presents a method for intelligently aggregating information from multiple channels. This approach involves analyzing the

context surrounding a video's creation and subsequent use, to the extent possible, including elements of its documentary form, rather than relying solely upon its content, to provide a confidence rating of its authenticity.

The paper is divided into six sections. In Section II, we provide a literature review. Section III provides a brief overview of our prior work. In Section IV, we outline the results of authenticity tests conducted with human participants. Section V, we present a prototype of our proposed mixed methods approach to ascertain the authenticity of online videos created by citizen journalists. In Section VI, we evaluate the prototype. Finally, in Section VII, we present our conclusions along with the next steps of our ongoing collaborative and transdisciplinary research.

II. Literature Review

There are different ways to create fake and manipulated videos. In Table I, we provide a list of techniques and typical real-world examples for each. Typologies to classify fake and manipulated video are evolving [8]. Hamouda created the list as part of a literature review [5].

Visual forgery and manipulation are common disinformation tactics grouped into two broad categories: shallowfakes and deepfakes [14]. Table I outlines different types for each based on their manner of generation. In addition to fake or manipulated visual elements, inauthentic CJVs may also have fake or manipulated audio elements. This may consist of anything from completely fake audio elements to slowed playback speeds to distort the audio channel of a video. Finally, metadata associated with CJVs may also be fake or manipulated.

The literature on detecting malicious deepfakes and countermeasures focused on detection, including those that focus on identifying artifacts (e.g., unintended features; for examples, see below) in deepfakes. Deepfakes often generate artifacts that may be undetectable by humans but can be detected using AI and forensic analysis. Mirsky and Lee [2] identify seven types of artifacts in two broad groups: (1) spatial artifacts in i. blending, ii. environments, and iii. forensics; (2) temporal artifacts in iv. behavior, v. physiology, vi. synchronization, and vii. coherence. Another approach involves training deep neural networks as generic classifiers to let them identify features in an unsupervised way instead of focusing on a specific artifact [2]. Using an unsupervised approach, researchers generally have taken one of two paths: classification or anomaly detection [2]. Another approach focuses on data provenance, that is, the relationships between records and the organizations or individuals that created, accumulated and/or maintained and used them in the conduct of personal or corporate activity [4]. To prevent deepfakes, some have suggested that data provenance should be tracked using distributed ledger technology [15]–[17].

The approach we propose adds to this literature and builds upon our prior work by utilizing archival diplomatics formalisms to analyze elements of provenance and documentary

TABLE I
Types of video forgery and manipulation

Type	**Examples**
Shallowfakes	
Slowing down the playback speed of video visual frames	Video of US Democratic House Speaker Nancy Pelosi was slowed down to give the impression that she was intoxicated. [9]
Cutting visual frames	Video instance tweeted by former United States Press Secretary Sarah Sanders from the Infowars site in November 2018. It shows Jim Acosta, CNN's Chief White House Correspondent holding a microphone that an intern is trying to take away from him during a press conference with former US President Donald Trump. Frames of this video were cut to make it appear that Acosta had aggressively placed his hands on the intern. [10]
Altering audio	In the original Acosta video, Acosta says 'Pardon ma'am' to the intern, but the instance of the video edited by Infowars mutes Acosta's voice.
Misrepresentation through alteration of descriptive text	Video posted by Facebook user Hendry Moya Duran purporting to show the devastation caused by 'Hurricane Irma', which took place in September 2017. The original video was actually captured in Uruguay after a tornado hit Dolores in April 2016.
Deepfakes	
Reenactment	Video of former US President Barack Obama produced by Jordan Peel - an American actor, comedian, and filmmaker - in which Obama appears to call his political successor, Donald Trump, a 'dipshit'. [11]
Replacement	Video of former US President Donald Trump's State of the Union address that replaces his face with actor Nicolas Cage's [12].
Editing and synthesis	Altered video tweeted by former US President Donald Trump showing a Nickelback video in which the photo in the video has been doctored to feature a photoshopped image designed to promote the claim that former US Vice-President and now President Joe Biden was involved in corruption in Ukraine. [13]

form, such as the types of artifacts and anomalies that could be detected using the above-mentioned approaches.

III. Prior work to Leverage Archival Diplomatics to Develop a Framework to Detect and Prevent Fake Video

Prior work to develop a framework to assess the authenticity of CJVs was conducted as phase I of a research project entitled *Extending the Scope of Computational Archival Science: A Case Study on Leveraging Archival and Engineering Approaches to Develop a Framework to Detect and Prevent "Fake Video"* (principal investigators were Victoria Lemieux and Chen Feng). The project was funded by the Government of Canada's Defence Excellence and Security (IDEaS) program. A framework [1] arising from this work leveraged archival diplomatics, and proposed an approach to assess the authenticity of CJVs in two rounds as follows:

> **Round 1** is an internal consistency check consisting of a pairwise comparison of the characteristics of each component (visual, audio, metadata) within the same video, see Table II.
>
> **Round 2** is an external consistency check which is a pairwise comparison of the characteristics of each component between one instance of a video and another instance of a near-duplicate video if one is available.

In more detail, Round 1 is an internal consistency check that consists of two steps:

> In **Step 1**, information about the context of the video, in the form of metadata, is extracted, examined, and compared (for example, the title, date, location, and author of the video). The metadata is checked to see if there are any inconsistencies within the video itself. An example of an inconsistency might be if the title of a video states that it was captured in 2019, but the video's publication date on YouTube states that it was published in 2018.
>
> In **Step 2** the video is checked to see if there are any inconsistencies between the metadata (extracted from Step 1) and its visual components. An example of inconsistency might be a video title stating that it was captured in Cairo while the visuals show landmarks in Tunisia. The test also checks for inconsistencies between the metadata and audio, for example, if the video was captured in India. In contrast, the anchor in the video (i.e., audio component) states that they are in Pakistan.

In this paper, we focus only on Round 1 (i.e., internal consistency) and will build on Round 2 in future work.

IV. Testing Video Authenticity with Human Participants to Create a Benchmark

To provide a guide for the development and benchmark for the performance of the solution presented in this paper, we rely upon the results of a 2019 study involving testing of the framework presented in [1] with human subjects. The study was to test the impact of the framework developed in [1] on the evaluation of the authenticity of CJVs and inform the automation of the identification of inauthentic videos. The study was conducted as a phase II of the aforementioned research project *Extending the Scope of Computational Archival Science: A Case Study on Leveraging Archival and Engineering Approaches to Develop a Framework to Detect and Prevent "Fake Video"*.

The following describes the qualitative methodological approach adopted for the pilot study. The methodological approach and online survey were designed by Hoda Hamouda in consultation with Victoria Lemieux, the project's principal investigator, and Heather O'Brien, a faculty member at UBC's School of Information with expertise in user experience and user interface design and testing. The approach and survey are based on typical user testing methods in the field of human-computer interaction [18], [19].

Surveys were used in previous experiments that involve human identification of fake multimedia content such as fake news or videos carrying false information [20]. Participants in these studies were presented with information and asked to classify it (for example either as fake or authentic) by responding to the survey. Our survey was developed based on prior work such as Sütterlin et al. [21], on fake multimedia recognition, and Khodabakhsh et al. [22], on subjective evaluation of fake multimedia.

Using the approach outlined in Section III and discussed in Ref. [1], we recruited 20 participants using a post on the Graduate Student Community forum of The University of British Columbia (UBC) (community.grad.ubc.ca). Participants started the survey by answering some basic demographics questions about their age, level of education, sex, and vision deficiencies. The majority of the participants were between 31 and 40 (45%) and 18 and 30 (30%) years old, with a smaller percentage in the 41-60 range (20%). Regarding the highest completed education degree of participants: 37% of participants hold a Bachelor's degree, half of participants (52%) hold a Master's degree and a few (10%) hold a Doctorate. Regarding the participants' sex, the number was about the same, 53% of participants were male, and 47% of participants were female. Participants were asked in the survey if they have any vision deficiencies (such as color blindness, or blurred vision), the majority (90%) reported that they do not have any vision deficiencies while about (10%) reported that they have vision deficiencies.

To conduct the testing, we divided the participants into two groups, a control group and an intervention group. We asked them to respond to an online survey to classify eight videos as either authentic or fake. We defined to participants in the survey that

1) an authentic video is what it claims to be and is free from manipulations, and
2) a fake video is not what it claims to be or has been manipulated.

The videos used for testing purposes included two authentic and six inauthentic videos, i.e., fake, that typified the manipulation techniques discussed in Table I. Table II lists the types of inconsistencies between the elements of the videos, used in the study, that render the videos fake. The videos were presented to participants in a YouTube interface with room to display the video's metadata. The user interface of YouTube allowed participants to see the video and metadata such as the title, published date, description, and channel name, in addition to other information as shown in Fig. 1. The duration of each video was from 30 to 60 seconds. Participants were provided a text box for every video to optionally input why they think the video is fake. Participants were asked to complete the survey on a computer screen and to turn the computer audio on. Finally, to eliminate order bias, the play sequence of the videos was randomized in each test. Order bias is the effect of question ordering on the response of participants, in either interviews or self-administered surveys [23].

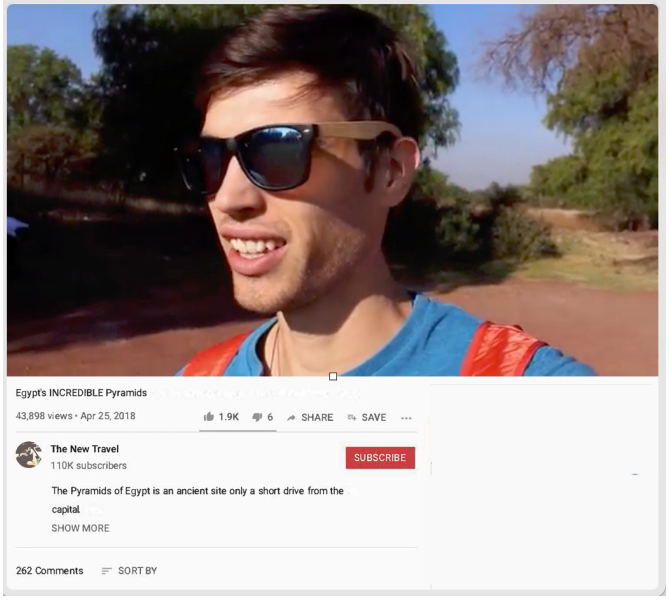

Fig. 1. Each of the eight videos was presented in the YouTube interface which showed the video metadata. This is a rendering of a video in Table II, that was used for the Audio-Metadata test (AM). It showed inconsistencies between what the person was saying, i.e., audio, and the title and description of the same video, i.e., metadata.

The test was unsupervised and was conducted using Qualtrics (www.qualtrics.com). A link was sent to the participants to respond online. Gift cards were given to participants as an honorarium. The control group was asked to respond to the survey immediately after they watched each video. The intervention group, on the other hand, was provided with a tutorial that included textual and visual information, see Fig. 2, that explained each of the six categories of inauthentic video types listed in Table II, based on the framework presented in [1].

We use the results of the pilot study described in this section as a guide to the development of our solution and as a benchmark against which to compare the output of the automated solution we present in this paper.

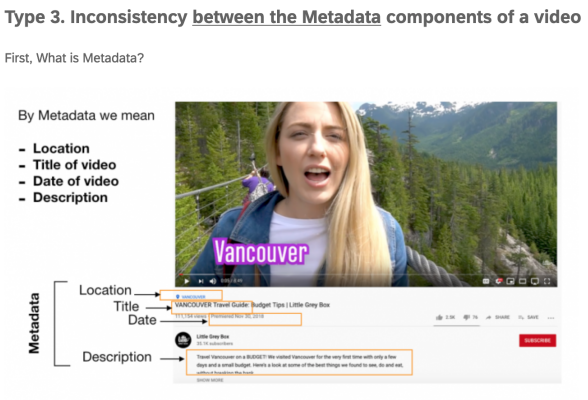

Fig. 2. The intervention group was presented with a tutorial that explained the six types of inauthentic videos. The figure is a screen capture from the explanation of the Metadata-Metadata test (MM) to explain possible inconsistencies between the metadata components of a video.

TABLE II
FAKE VIDEO DETECTION TESTS

Test	Description
VV	Visual against visual
VM	Visual against metadata
AM	Audio against metadata
AA	Audio against audio
MM	Metadata against metadata
VA	Audio against visual

The tests that were submitted to the participants are listed in Table II.

V. A PROTOTYPE TO LEVERAGE ARCHIVAL DIPLOMATICS TO DEVELOP A FRAMEWORK TO DETECT AND PREVENT FAKE VIDEO

In this section, we outline the approach employed in our solution.

A. Approach

Our approach to scoring authenticity draws from research on truth finding [7], [24]. It involves two stages: bootstrapping and iterative propagation. In the bootstrapping stage, initial authenticity scores are assigned to entities. An oracle can assign this initial score, derived from a trusted source of knowledge such as a human expert, a decision system, or the result of natural language processing. Alternatively, it can be determined by leveraging trustworthy historical data, which relies on the past to predict the future. Without oracles or historical data, an ambivalent score can be assigned, serving as a neutral starting point. This initial scoring process ensures that the individual authenticity of each entity is accurately reflected in the model.

The iterative propagation stage produces conclusions based on the initial state created by bootstrapping. This inference

process can be accomplished by considering three main non-exclusive approaches: graph analytics, optimization, and probabilistic [24]. Graph analytics is rich in algorithms that score entities according to their context, which is determined by their relationships with other entities [25], [26]. The algorithms include Similarity, PageRank, and Propagation. In this paper, we follow the graph analytics approach. In particular, we build upon the truth-finding model of Yu et al. [7].

B. Data Model

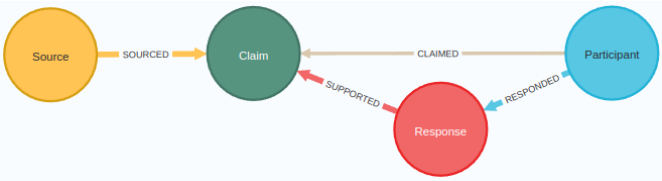

Fig. 3. Knowledge graph schema.

The schema of the knowledge graph is shown in Fig. 3. There are four types of nodes, namely, Source, Claim, Response, and Participant. The Source nodes are the test types (VV, VM, AM, AA, MM, and VA). The Claim nodes are the outcomes for every test type, e.g., VV is authentic or VV is fake. The Response nodes are the textual justifications provided by the Participant for their Claim. The Participant nodes represent individuals who provided claims and responses ($p1, p2, \ldots, p8, p11, \ldots, p19$). There are four types of relationships represented by the edges SOURCED, CLAIMED, SUPPORTED and RESPONDED. A SOURCED relationship connects a Source node, i.e., a test type, to a corresponding Claim node for the test type, i.e., authentic or fake. A CLAIMED relationship associates a Participant, e.g., $p1$, to a specific Claim node, e.g., VV is authentic or VV is fake, but not both. A SUPPORTED relationship indicates that a Response was given to justify a Claim. A RESPONDED relationship associates a Participant node, e.g., $p1$, to a specific Response node.

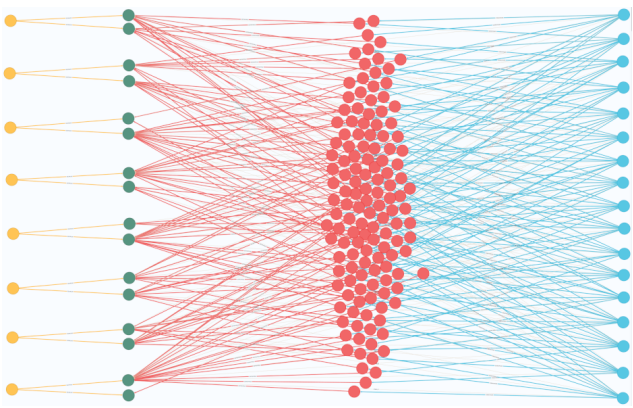

Fig. 4. Knowledge graph.

Fig. 4 Depicts an instance of the schema of Fig. 3 instantiated with the data set introduced in Section IV. Color of node instances in Fig. 4, are matching the colors of the node types, in Fig. 3.

C. Truth Finding Model

Every node n receives an authenticity score $t(n)$. The initial authenticity score $t_0(n)$ is determined according to the type of the node n. In the bootstrapping stage, the nodes are ranked according to the following logic.

1) Source: Every Source node starts with the score value $1/m$, where m is the number of Source nodes.
2) Claim: Each Claim node starts with score $1/n$, n is number of Claim nodes.
3) Response: The initialization of the score of every Response node is 1.0 when the participant has provided no or a plausible justification. It is 0.1 when the participant provides a nonsensical or non-logical justification.
4) Participant: To determine the initial scores of Participant nodes, we measure participant-to-participant similarity concerning their responses, using the Similarity graph analytics algorithm. We create a participant-to-participant relationship weighted by their degree of similarity. Finally, we apply the weighted PageRank graph-analytics algorithm [25]. The rank determined by the PageRank algorithm determines the initial authenticity score of every Participant node. This acknowledges a consensus of opinions among the participants. The higher the ranking, the more Participant nodes agree with others on responses.

The next stage is the iterative propagation of the authenticity score and the propagation logic applies two heuristics [14].

1) A response is more likely to be true if derived from many trustworthy sources. A source is more likely to be trustworthy if many of the responses it provides are true.
2) A response is more likely to be true if it is obtained by many trustworthy participants. A participant is more likely to be trustworthy if many of the responses it generates are true.

Authenticity score propagation follows a logic based on the propagation algorithm of Yul et al. [7]. Related nodes mutually reinforce their authenticity. The use of the Similarity and PageRank graph analytics algorithms to accomplish this stage is detailed below.

D. Use of Similarity and Weighted PageRank Algorithms

A similarity score is assigned to every pair of Participant nodes using the Similarity graph analytics algorithm. The similarity score quantifies the degree of agreement among the responses provided by the two participants. Two participants are considered similar when they share several responses.

Definition 1 (Similarity score)**.** *Let p_1 and p_2 be two Participant nodes. Let their Claims be the sets C_1 and C_2,*

respectively. The similarity of p_1 and p_2 is defined as the following ratio

$$Sim(p_1, p_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}. \quad (1)$$

This definition of similarity is based on the Jaccard metric.

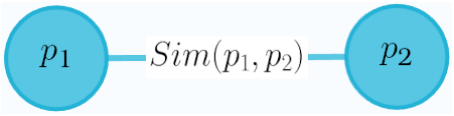

Fig. 5. Similarity score.

A graph $G = (V, E)$ is created where the set of nodes V contains all participants and E is the set of relationships. A relationship is created between every pair of nodes representing participants p_1 and p_2.

The relationship is weighted by the score $Sim(p_1, p_2)$, see Fig. 5. The resulting graph is fully connected with no self-loop.

In this case, Fig. 6 visually illustrates the Similarity of two Participants, p_1 and p_{12}. The two Claims (green) to the right of the Participants (blue) are common to p_1 and p_{12} as indicated by the Responses (red) corresponding to a given test being connected to the same Claim, e.g., *MM is fake*. Additionally, each of the two corresponding Sources (yellow) is connected to only one Claim (its contrary Claim, e.g. *MM is auth*, and is not shown in this case to avoid clutter but is included in the overall graph).

The left side of Fig. 6 contains the Claims which differ between the two Participants. For each of these, only one Participant has a Response connected to it. Additionally, each of the corresponding six Sources needs to be connected to two Claims to be connected to both Participants.

Thus, in this example, the Similarity is

$$Sim(p_1, p_{12}) = \frac{2}{14} = 0.142857143 \quad (2)$$

The PageRank algorithm assigns a rank to every graph node according to its importance, considering the global information represented in the graph. The PageRank algorithm is applied to the graph G. It assigns to every participant p a rank $r(p)$. Let $N(p)$ be the set of neighbors of p with edges incoming to p, i.e.,

$$N(p) = \{v \in V : (v, p) \in E\}. \quad (3)$$

The rank of participant p is determined as the sum

$$r(p) = (1 - d) + d \cdot \sum_{v \in N(p)} \frac{Sim(p, v) \cdot r(v)}{\sum_{q \in N(v)} Sim(q, p)} \quad (4)$$

The factor d has a damping role (e.g., 0.85). The equation is applied iteratively until convergence or a maximum number of iterations is reached. In the main summation, the rank of a neighbor v is multiplied by the weight of the relationship connecting v to p, i.e., $Sim(p, v)$. This product is divided by the sum of the weights of outgoing vertices of v. For a given Participant, more similar neighbors carry a higher weight.

E. Authenticity Propagation Algorithm

The authenticity score is obtained using an adaptation of the Yul et al. propagation algorithm [7]. The initial score of every participant p has been determined by the PageRank algorithm, i.e., $r(p)$. The initial score of every source is $1/m$, where m is the number of sources. According to the plausibility of justifications, every response has an initial score of 0.1 or 1. The propagation algorithm proceeds in three steps. At Step 1, when r is a Response node (with p is a Participant node), its score is determined as

$$c(r) = c_0(r) + \sum_{p \in N(r)} c_0(p). \quad (5)$$

The score c, is determined using the initial scores c_0. The expression $p \in N(r)$ denotes a Participant node p related to r.

Step 2 is completed by propagating scores of Response nodes to their associated Claim nodes, where the corresponding scores are aggregated for each Claim. Let ℓ denote a Claim node (with r a Response node), its score is determined as

$$c(\ell) = c_0(\ell) + \sum_{r \in N(\ell)} c(r). \quad (6)$$

The expression $r \in N(\ell)$ denotes a Response node r related to ℓ.

Step 3 is completed by propagating the corresponding scores of the Claim that the video is authentic and the Claim that the video is fake. A Source, e.g. VV, receives two scores: one for *VV is authentic* and one for *VV is fake*. A higher score reflects greater confidence. A Source with an *authentic* score larger than its *fake* score is deemed authentic. A Source with a higher *fake score* than an *authentic score* is deemed fake.

F. Limitations

Our approach is subject to several limitations. First, the sample size of our benchmark human test is very small. Furthermore, it is comprised of graduate students. A larger and more heterogeneous human test would provide a better benchmark of our approach and remains as future work. Additionally, our approach has some vulnerabilities, including potential coalition attacks. Indeed, it is possible that a group of participants might artificially increase their initial score, maximizing the agreement of a collective response, or that malevolent participants might try to deny the vote of honest participants. Thus, a detailed threat analysis must be carried out to develop a threat model and identify possible risk mitigation strategies. Additionally, the evaluation of archival diplomatics elements of form in our solution still needs further refinement. This will be a focus of future work. Finally, the use of visualizations to represent the output of graph analytics in our solution may not be scalable to analyzing a higher number of videos. We acknowledge that future work will need to focus on human-tractable visual rendering over a large volume of videos.

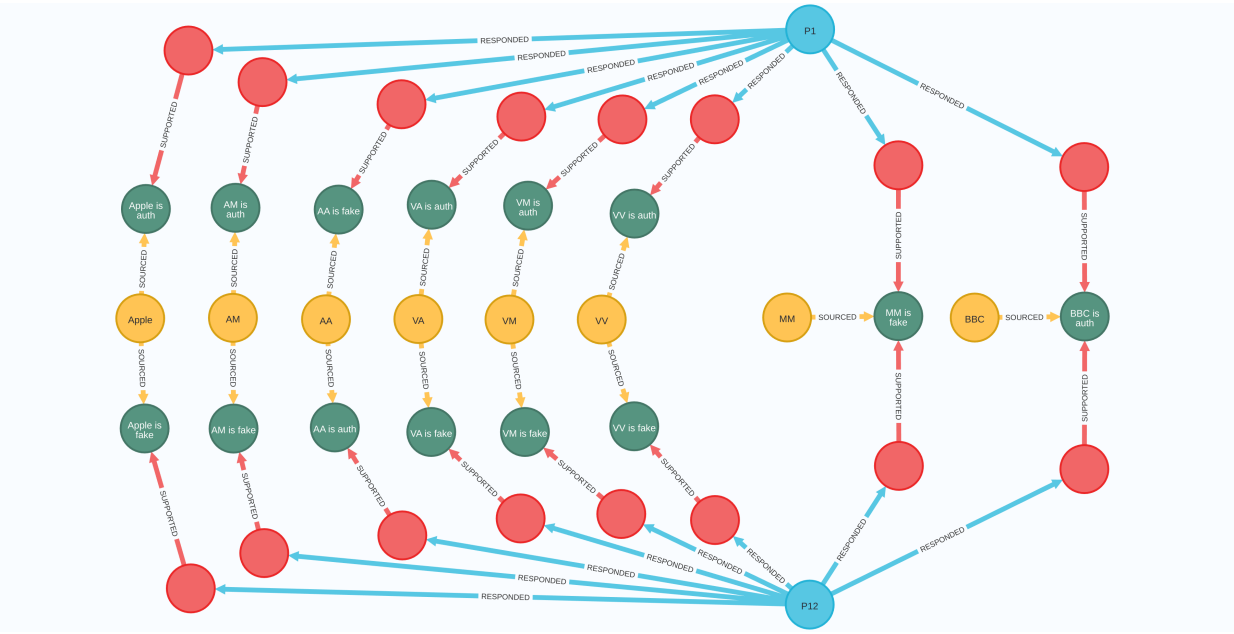

Fig. 6. Similarity of two Participants (p_1 and p_{12}).

VI. Evaluation

The algorithm converged towards ratios of *fake* and *authentic* Claim scores, which generally reflected the ground truth, as seen in Fig. 7, with genuinely authentic Sources, i.e., BBC (Auth2 in Fig. 7) and Apple (Auth1 in Fig. 7), receiving larger *authentic* scores (first bar in each pair in Fig. 7) and lower *fake* scores (second bar in each pair in Fig. 7).

In contrast, the genuinely fake Sources received lower *authentic* scores and higher *fake* scores. The Claims that correspond to the ground truth are blue with backslash hatching in Fig. 7 while those that are incorrect are red with forward-slash hatching.

Apple's relatively low score compared to BBC may be due to any of several factors. In the future, a larger data set would better allow for a more nuanced discrimination between conflicting interpretations.

VII. Conclusion

We addressed the problem of automating the evaluation of the authenticity of CJVs of uncertain origins. Our approach builds on archival diplomatics theory and past research on truth-finding methods. We represent the information using the knowledge graph model. The knowledge graph model was selected because it can summarise and capture dataset semantics with a strong emphasis on relations between entities. The initial scores are derived from human responses to evaluation tests. Graph analytics methods bootstrap and perform the iterative propagation of authenticity scores. The final results, however, are only indicative and not definitive. Nonetheless, the prototyped approach can be seen as a decision-support tool, supporting humans who ought to remain in the loop and carefully read and interpret the results before making assumptions and any final decisions based on them. Also, the choice of initial scoring methods and exact iterative propagation algorithms are critical, and interventions by human experts are essential to design a sequence of logical steps. The software for this study is available [27].

References

[1] H. Hamouda, J. Bushey, V. Lemieux, J. Stewart, C. Rogers, J. Cameron, K. Thibodeau, and C. Feng, "Extending the scope of computational archival science: A case study on leveraging archival and engineering approaches to develop a framework to detect and prevent "fake video"," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3087–3097, IEEE, 2019.

[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[3] L. Duranti, "Concepts and principles for the management of electronic records, or records management theory is archival diplomatics," *Records Management Journal*, vol. 20, no. 1, pp. 78–95, 2010.

[4] InterPARES, "Interpares project: Terminology database," 2024.

[5] H. A. Hamouda, "Authenticating citizen journalism videos by incorporating the view of archival diplomatics into the verification processes of open-source investigations (OSINT)," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 2036–2046, IEEE, 2023.

[6] S. Ahmed and E. Sonuç, "Trustworthiness of citizen journalists videos from the perspective of archival science," *IEEE International Conference on Big Data*, pp. 4403—-4407, 2018.

[7] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail, "The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1567–1578, 2014.

[8] J. Bushey, "Ai-generated images as an emergent record format," *IEEE Xplore*, p. 2020–31, 2023.

[9] W. Post, "Pelosi videos manipulated to make her appear drunk are being shared on social media." https://www.youtube.com/watch?v=sDOo5nDJwgA, 2019.

[10] W. Post, "Watch two versions of acosta video side-by-side." https://www.youtube.com/watch?v=aXZ2jRZMLrg, 2018.

[11] BuzzFeedVideo, "You won't believe what obama says in this video!." https://www.youtube.com/watch?v=cQ54GDm1eL0, 2018.

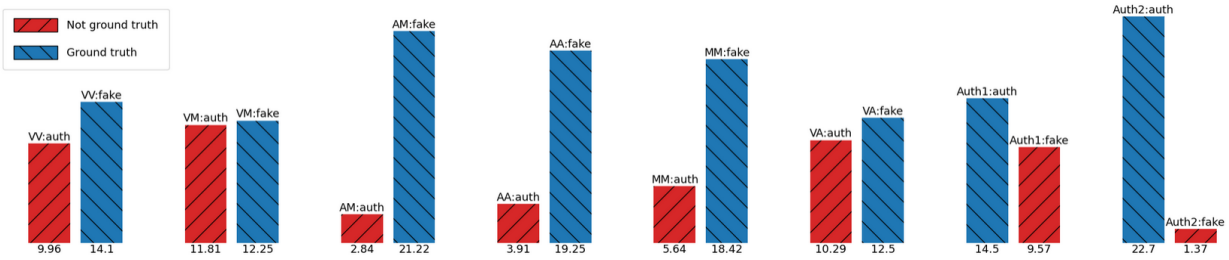

Fig. 7. Authenticity scores, where each test receives two columns: red with forward hatch indicates not ground truth and blue with back hatch indicates ground truth.

[12] S. Ahmed and E. Sonuç, "Evaluating the effectiveness of rationale-augmented convolutional neural networks for deepfake detection," *Soft Computing*, 2023.

[13] Today7, "Video: Trump hilariously tweets Biden photo in Nickelback music video." https://www.youtube.com/watch?v=HUHZNSJGFzQ, 2019.

[14] V. L. Lemieux, *Searching for trust: blockchain technology in an age of disinformation*. Cambridge University Press, 2022.

[15] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-fidelity face manipulation with extreme poses and expressions," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2218–2231, 2021.

[16] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, vol. 7, pp. 41596–41606, 2019.

[17] C. C. K. Chan, V. Kumar, S. Delaney, and M. Gochoo, "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media," in *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*, pp. 55–62, IEEE, 2020.

[18] J. Muirhead, "Assessing online content quality through user surveys and web analytics," *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, p. 433–436, 2019.

[19] V. V. M. Callegaro, K. Lozar Manfreda, "SAGE research methods complete A-Z list, and SAGE research methods core, web survey methodology," *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 2015.

[20] K. A. C. Persen and S. C. Woolley, "Computational propaganda and the news: Journalists' perceptions of the effects of digital manipulation on reporting," in *Affective politics of digital media*, pp. 245–260, Routledge, 2020.

[21] S. Sütterlin, T. F. Ask, S. Mägerle, S. Glöckler, L. Wolf, J. Schray, A. Chandi, T. Bursac, A. Khodabakhsh, B. J. Knox, *et al.*, "Individual deep fake recognition skills are affected by viewer's political orientation, agreement with content and device used," in *International Conference on Human-Computer Interaction*, pp. 269–284, Springer, 2023.

[22] A. Khodabakhsh, C. Busch, and R. Ramachandra, "A taxonomy of audiovisual fake multimedia content creation technology," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 372–377, IEEE, 2018.

[23] E. L. Landon Jr, "Order bias, the ideal rating, and the semantic differential," *Journal of Marketing Research*, vol. 8, no. 3, pp. 375–378, 1971.

[24] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explor. Newsl.*, vol. 17, p. 1–16, feb 2016.

[25] M. Needham and A. Hodler, *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media, 2019.

[26] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for large-scale graph processing," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 135–146, 2010.

[27] N. Rivard, "Prototype-to-Leverage-Archival-Diplomatics-to-Develop-a-Framework-to-Detect-and-Prevent-Fake-Video." https://github.com/nicholasrivard/Prototype-to-Leverage-Archival-Diplomatics-to-Develop-a-Framework-to-Detect-and-Prevent-Fake-Video, 2024.