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

Nicholas Rivard, Hoda Hamouda, Victoria Lemieux, Tracey P. Lauriault, and Michel Barbeau

Citizen Journalism Videos (CJVs)

https://clp.mk/citizen-journalism-is-a-powerful-tool-for-revealing-abuses-and-election-irregularities/?lang=en

Video Verifiers: professionals specialized in verifying online CJ videos

Video Verifiers:

- Investigative journalists, Fact-checkers, …

They rely on methods from the fields of:

- Journalism, Digital forensics, …

Archival diplomatics can contribute to video verification practices

The Research Theoretical Framework: Archival Diplomatics

_ It is the study of the origins, the forms, and the transmission of documents and the relationship between the document and the facts represented in them and their relation to the persons who created them to examines the authenticity and trustworthiness of analogue and digital records.

To infer the authenticity of online videos

Video verifiers examine the <u>content</u>.

- Date and location
- Source
- Locate the original video

Archival diplomatics examines the <u>content</u> <u>(to an extent)</u> **<u>AND:</u>**

- Contexts
- Fixity of the video after publishing
- Persons
- Relationships to other records
- The purpose

(Hamouda, 2023)

To infer the authenticity of online videos

Archival diplomatics examines the content AND:

- Contexts:
  - **The administrative and juridical context**
  - The Provenancial context
  - The Technological context
  - The Procedural context
  - The Documentary context

Prior work: the Authenticity Test Framework

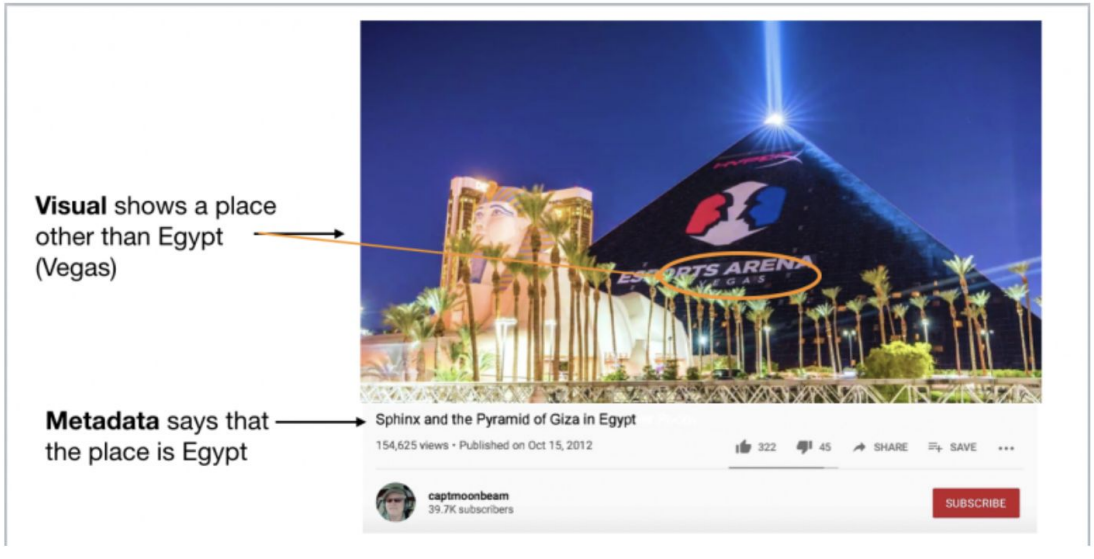
A video is fake when there is inconsistencies between The audio, visual, or metadata components.

Video components of the examined video	**Video components of the same video**		
	Visual	***Audio***	***Metadata***
Visual	**VV Test** (1) check visual (V) against visual (V) inconsistencies	((duplicate of 2)	(duplicate of 3)
Audio	**AV Test** (2) check audio (A) against visual (V) inconsistencies	**AA Test** (4) Tests to check audio (A)against audio (A)	(duplicate of 6)
Metadata	**MV Test** (3) Check metadata (M) against visual (V) inconsistencies	**MA Test** (5) Check metadata (M) against audio (A)	**MM Test** (6) Check metadata (M) against metadata (M) inconsistencies

TABLE I. Test Matrix for Detecting Fake Videos

The Authenticity Test Framework

This **inconsistency between the Visual and Metadata** makes it a **fake video**.



Testing the framework with human subjects

The goal was to:

_ Inform the automation of the classification of fake videos.

	Name of the test
Video 1	VV Test Visual-to-visual (in)consistency check
Video 2	VM Test Visual-to-metadata (in)consistency check
Video 3	AM Test Audio-to-Metadata (in)consistency check
Video 4	AA Test Audio-to-Audio (in)consistency check
Video 5	MM Test Metadata-to-Metada ta (in)consistency check

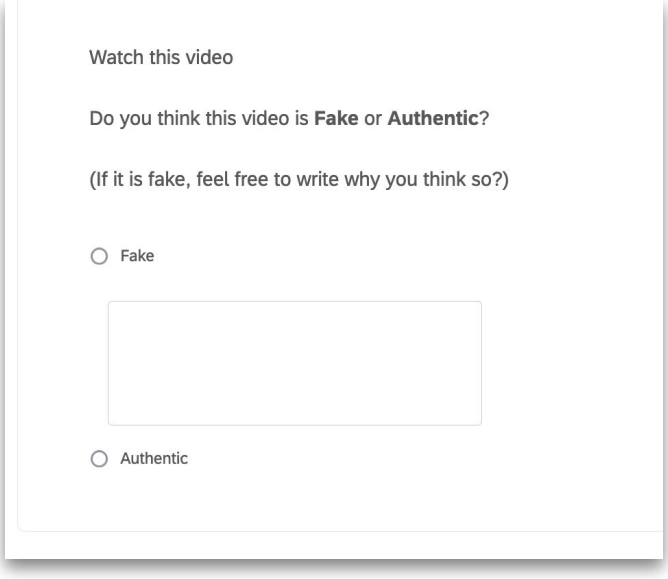
Pilot Study with human participants

_ Online survey.

_ Classify videos as either authentic or fake.

_ Most of the videos sent contained one of the inconsistencies listed in the framework.

_ Text input for every video to write why they think the video is fake (i.e., justification).



justification

Pilot Study Output

The output of this was the data that was then processed

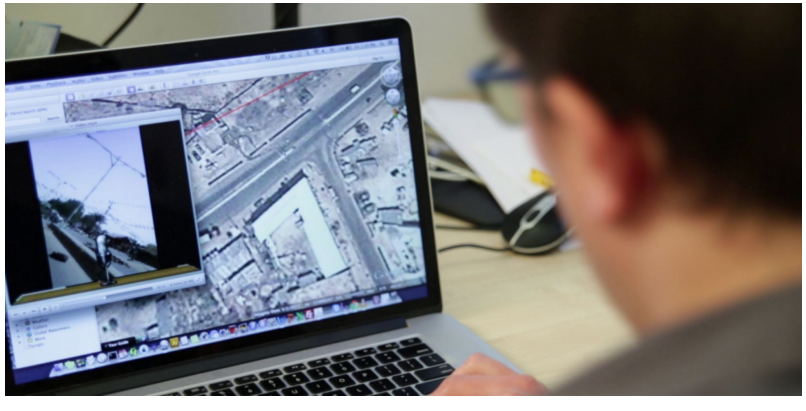
by the algorithm that will be shortly discussed by Nick.

Use case

A group of video verifiers are tasked to:

_ Classify a group of videos as authentic or fake (i.e., inauthentic) each one on their own

_ Provide their justification on why the video is authentic or fake



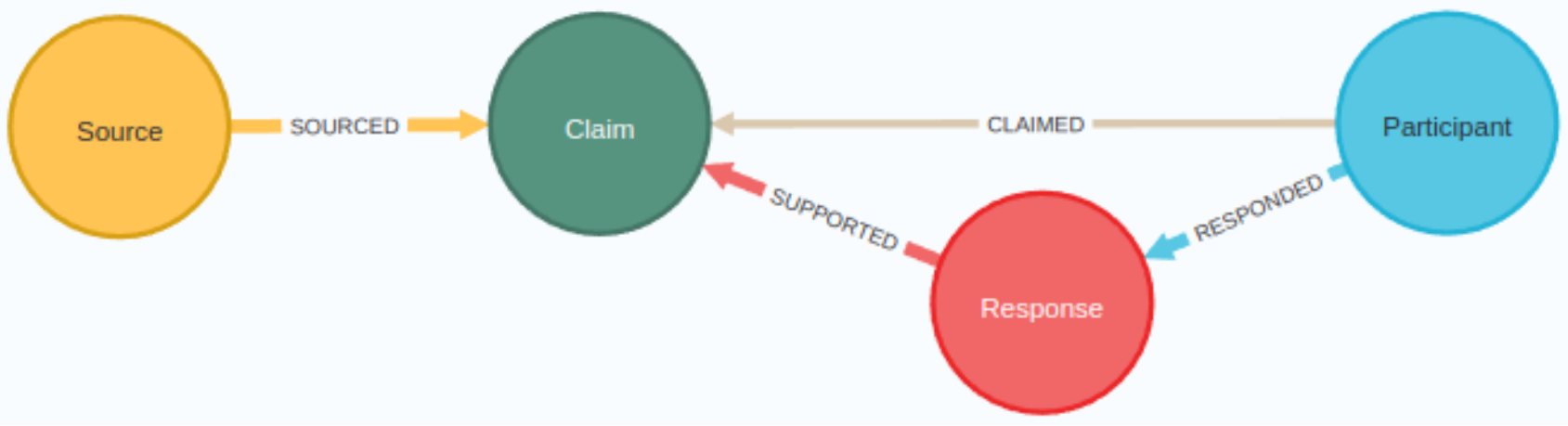
Scoring Trustworthiness of an Entity

1. Initialization
   1.1 Using Oracles (human [expert], decision system, NLP)
   1.2 Building on trustworthy historical data
2. Inference using Graph analytics (Li *et al.*, 2016; Needham and Hodler, 2019; Malewicz *et al.*, 2010; Yu *et al.*, 2014)

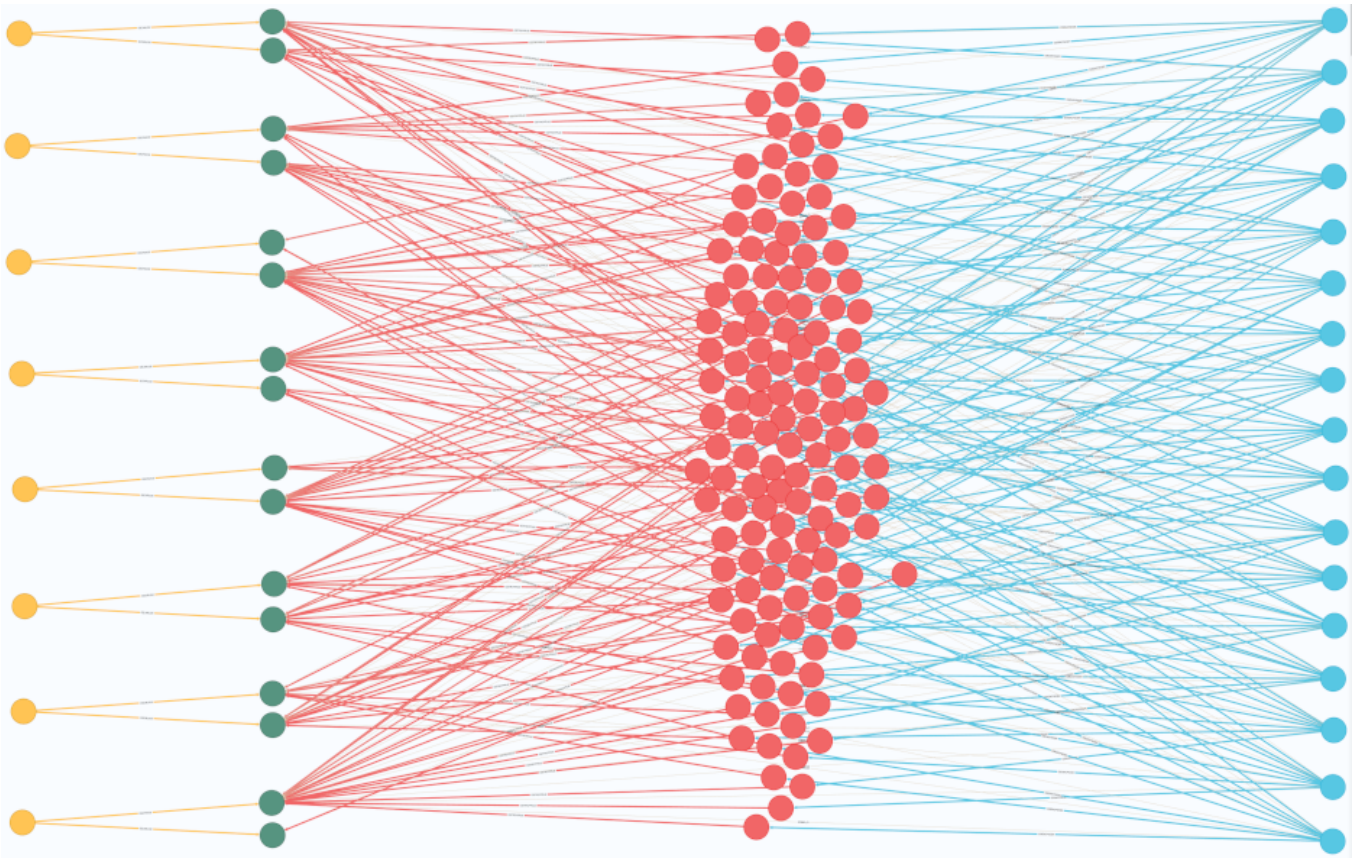
Framework (Rivard *et al.*, 2024)

1. Knowledge graph representation
2. Research on truth finding (Yu *et al.*, 2014; Li *et al.*, 2016)
3. Two stage-inference score process:
   3.1 Initialization with scores derived from user studies (H. Hamouda *et al.*, 2019; H. A. Hamouda, 2023)
   3.2 Iterative propagation: graph analytics (Similarity, PageRank, Propagation)

Schema


Source
SOURCED
Claim
CLAIMED
Participant
SUPPORTED
RESPONDED
Response

Knowledge Graph



Initialization

1. Source: starts with score $1/m$ (m is number of sources)
2. Response: 1.0 (plausible justification); 0.1 (plausible justification)
3. Participant:
   3.1 Participant-to-participant similarity measure (Similarity algorithm)
   3.2 Participant authority measure (PageRank algorithm)

Propagation algorithm

1. Response r:

$$c(r) = (1 - \lambda_1)c_0(r) + \sum_{p \in N(r)} c_0(p)$$

   using initial score c_0; λ_1 in $[0, 1]$; $p \in N(r)$ is a Participant node p related to r

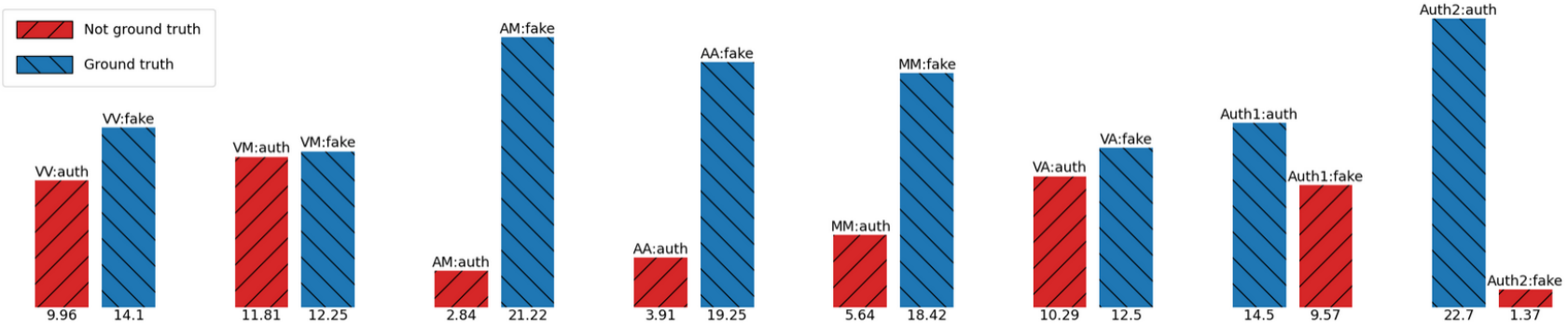
2. Claim ℓ:

$$c(\ell) = \sum_{r \in N(\ell)} c(r)$$

   where $r \in N(\ell)$ is a Response node related to ℓ

3. Claim score to video: "fake score" and "authentic score"

Results


Not ground truth
Ground truth
VV:auth
VV:fake
9.96
14.1
VM:auth
VM:fake
11.81
12.25
AM:auth
AM:fake
2.84
21.22
AA:auth
AA:fake
3.91
19.25
MM:auth
MM:fake
5.64
18.42
VA:auth
VA:fake
10.29
12.5
Auth1:auth
Auth1:fake
14.5
9.57
Auth2:auth
Auth2:fake
22.7
1.37

Conclusion

Claim scores generally reflected ground truth

Results indicative but not definitive

Limitations and Future Work:

1. Small size sample
2. Potential vulnerabilities (coalition attack)
3. Scalability of visualizations

References I

[1] Y. Li *et al.*, "A survey on truth discovery," *SIGKDD Explor. Newsl.*, vol. 17, no. 2, pp. 1–16, Feb. 2016, ISSN: 1931-0145. DOI: 10.1145/2897350.2897352. [Online]. Available: https://doi.org/10.1145/2897350.2897352.

[2] M. Needham and A. Hodler, *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media, 2019, ISBN: 9781492047636. [Online]. Available: https://books.google.ca/books?id=z4WZDwAAQBAJ.

[3] G. Malewicz *et al.*, "Pregel: A system for large-scale graph processing," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 135–146.

References II

[4] D. Yu *et al.*, "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*, 2014, pp. 1567–1578.

[5] H. Hamouda *et al.*, "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)*, IEEE, 2019, pp. 3087–3097.

[6] 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)*, IEEE, 2023, pp. 2036–2046.

References III

[7] N. Rivard, H. Hamouda, V. Lemieux, T. P. Lauriault, and M. Barbeau, *Towards a Prototype to Leverage Archival Diplomatics to Develop a Framework to Detect and Prevent Fake Video*, Submitted to: IEEE Conference on Digital Platforms and Societal Harms, 2024.