An LLM-driven Approach to Gain Cybercrime Insights with Evidence Networks

Honghe Zhou [1] Weifeng Xu [2] Josh Dehlinger [1] Suranjan Chakraborty [1] Lin Deng [1]

[1]Towson University, Maryland, USA [2]University of Baltimore, Maryland, USA

Background

- Digital forensics is crucial in the fight against cybercrime.
- Investigators heavily rely on manual processes to identify and analyze pertinent evidence from mobile devices.
- Conventional manual forensic procedures are labor-intensive, error-prone, and time-inefficient.

Motivation

- Developing an automated approach for gaining criminal insights with digital evidence networks harness Large Language Models (LLMs) to learn patterns and relationships within forensic artifacts, automatically constructing Forensic Intelligence Graphs (FIGs).
- FIGs graphically represent evidence entities and their interrelations, providing an intelligence-driven approach to the analysis of forensic data.

Research Questions

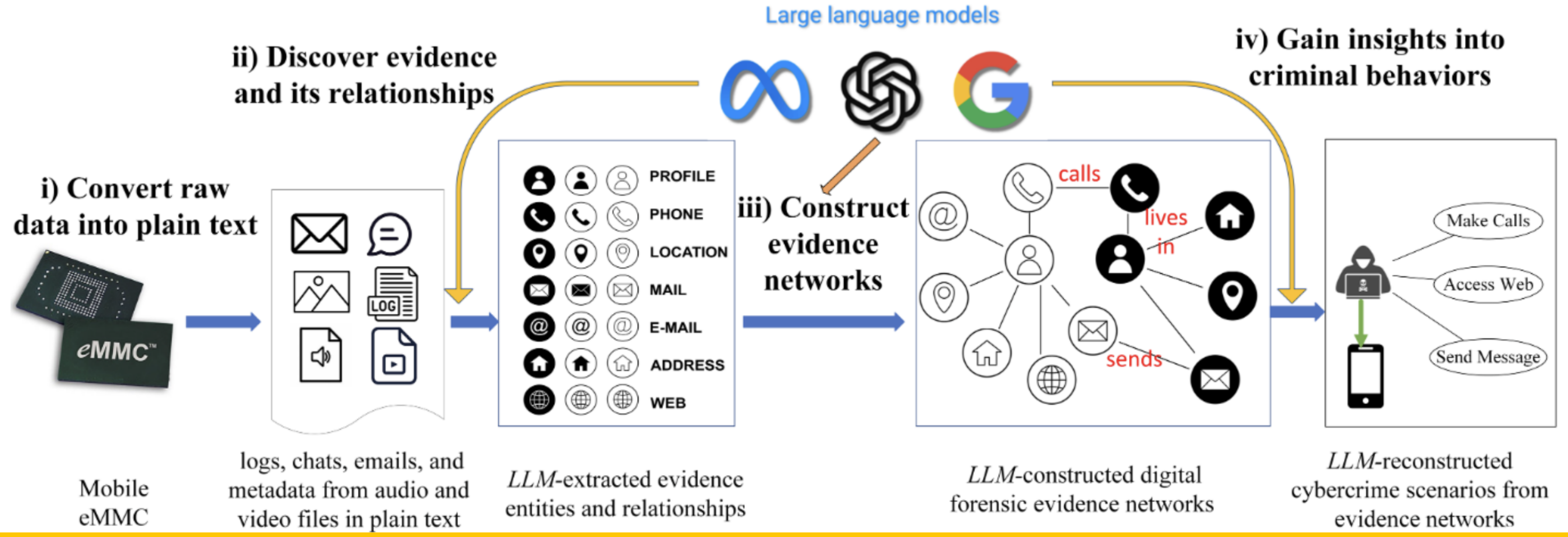
This research aims at revolutionizing digital forensics by harnessing the capabilities of Large Language Models (LLMs) to automate digital evidence discovery by addressing two critical **Research Questions (RQs)**:

- Can LLMs automatically identify various forms of evidence stored in different file types, such as system logs, system configurations, and databases, from mobile devices?
- Can LLMs reconstruct suspects' behavior and reveal valuable insights?

What are Large Language Models?

- A type of AI that can process and produce natural language text.
- It learns from a massive amount of text data such as books, articles, and web pages to discover patterns and rules of language from them.

Flowchart of Proposed Approach


Large language models
ii) Discover evidence and its relationships
iv) Gain insights into criminal behaviors
i) Convert raw data into plain text
eMMC
LOG
PROFILE
PHONE
LOCATION
MAIL
E-MAIL
ADDRESS
WEB
iii) Construct evidence networks
calls
lives in
sends
Make Calls
Access Web
Send Message
Mobile eMMC
logs, chats, emails, and metadata from audio and video files in plain text
LLM-extracted evidence entities and relationships
LLM-constructed digital forensic evidence networks
LLM-reconstructed cybercrime scenarios from evidence networks

Proposed Approach

1. **Convert raw data into plain text**: involves examining the evidence on mobile devices' Embedded MultiMediaCard (eMMC).
2. **Discover evidence and its relationships**: involves creating and testing LLM prompts to extract evidence from text files line-by-line.
3. **Construct evidence networks**: involves the development and testing of prompts aimed at linking isolated evidence to construct evidence networks, representing a unique contribution to the field.
4. **Gain insights into criminal behaviors**: focuses on deriving critical understandings and conclusions regarding criminal activities, behaviors, patterns, and relationships from evidence networks.

LLM Prompt for Discovering Evidence and Its Relationships

Act as an experienced digital forensic investigator. Identify evidence entities, including personal information like names, addresses, and phone numbers, from the given text. Describe any relationships among entities.

Desired output format:
Person's Name: `<person names>`
Address: `<mailing address>`
Phone number: `<phone number>`
Relationship: `<phone number>` ->(relationship description) `<mailing address>`
Text input: *a line of text* from a text file

Forensic Intelligence Graphs (FIGs)

FIGs can effectively represent complex forensic scenarios by mapping entities and their interconnections through labeled edges. A FIG is defined as a graph $G = (V, E)$, where:

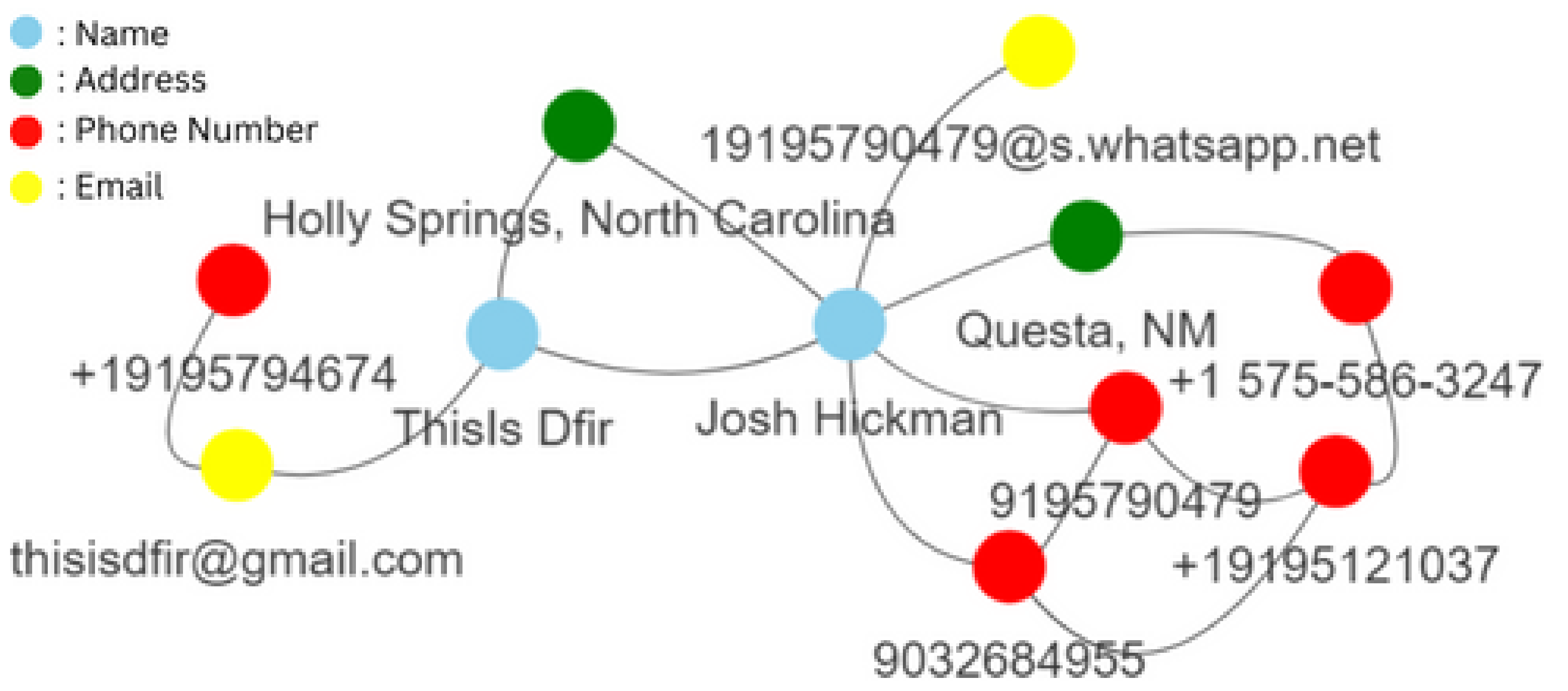
- V is a set of nodes representing evidence entities, such as a person's name, address, and phone number.
- E is a set of edges, where each edge $e \in E$ represents a relationship between two evidence entities.

Each edge e has a label that describes the relationship between the connected evidence entities. Examples: **"owns"**: indicating ownership, e.g., a person owns a phone number. **"lives-in"**: indicating residency, e.g., a person lives in an address.

Empirical Study and Preliminary Results

LLM-reconstructed FIG

Our approach generates an LLM-reconstructed FIG using three folders containing three popular Android apps from an Android 10 mobile phone image, including *Phone*, *Facebook Messenger*, and *Snapchat*. Each edge represents a relationship between these evidence entities.



Empirical Results

The table below shows the number of reconstructed evidence entities and relationships using two different approaches.

- **Baseline** indicates the "truth" (i.e., the initial manual investigation results) provided by the original creator of the Android disk image.
- **LLM-driven** is our automated approach. 'Match,' 'Added,' and 'Missed' indicate how the LLM-driven approach compares to the baseline in terms of matched, newly discovered, and overlooked entities and relationships.

	# of Reconstructed Evidence Entity		**# of Reconstructed Relationship**	
	Baseline	**LLM-driven**	**Baseline**	**LLM-driven**
Match	6	6	4	4
Added	0	5	0	11
Missed	1	0	1	0

The LLM-driven approach can discover additional evidence entities (5) and relationships (11), while only missing 1 evidence entity and relation. We fixed the baseline by adding newly discovered entities and relationships. Overall, our approach achieved:

- **Evidence Entity Coverage**: (6+5) / (6+5+1) = 91.67%
- **Evidence Relationship Coverage**: (4+11) / (4+11+1) = 93.75%

References

[1] André Årnes. *Digital forensics*. John Wiley & Sons, 2017.

[2] Graeme Horsman and Nina Sunde. Unboxing the digital forensic investigation process. *Science & Justice*, 62(2):171–180, 2022.

[3] Jianwei Hou, Yuewei Li, Jingyang Yu, and Wenchang Shi. A survey on digital forensics in internet of things. *IEEE Internet of Things Journal*, 7(1):1–15, 2019.

[4] Jigar Patel. Forensic investigation life cycle (filc) using 6 'r'policy for digital evidence collection and legal prosecution. *Int. J. Emerg. Trends Technol.*, 2(1):129–132, 2013.

[5] Alan Roder, Kim-Kwang Raymon Choo, and Nhien-An Le-Khac. Unmanned aerial vehicle forensic investigation process: Dji phantom 3 drone as a case study. *arXiv preprint arXiv:1804.08649*, 2018.

[6] Sarfraz Shaikh, Lin Deng, and Weifeng Xu. A practical survey of data carving from non-functional android phones using chip-off technique. In *21st International Conference on Information Technology: New Generations*, Las Vegas, Nevada, USA, April 2024.

[7] Norwegian University of Science Svein Y. Willassen and Technology. Cell phones | digital corpora. `https://digitalcorpora.org/corpora/cell-phones/`. Accessed: May 21, 2024.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024, Philadelphia, PA

ldeng@towson.edu