CRAWLING THE INTERNET FOR EXIF DATA AND CONTEXTUAL MISMATCHES

Mohamed Amine Aissati
Msc. Computer Security
University Of Kent
Canterbury - UK

Supervisor: **Dr. Julio Hernandez-Castro**

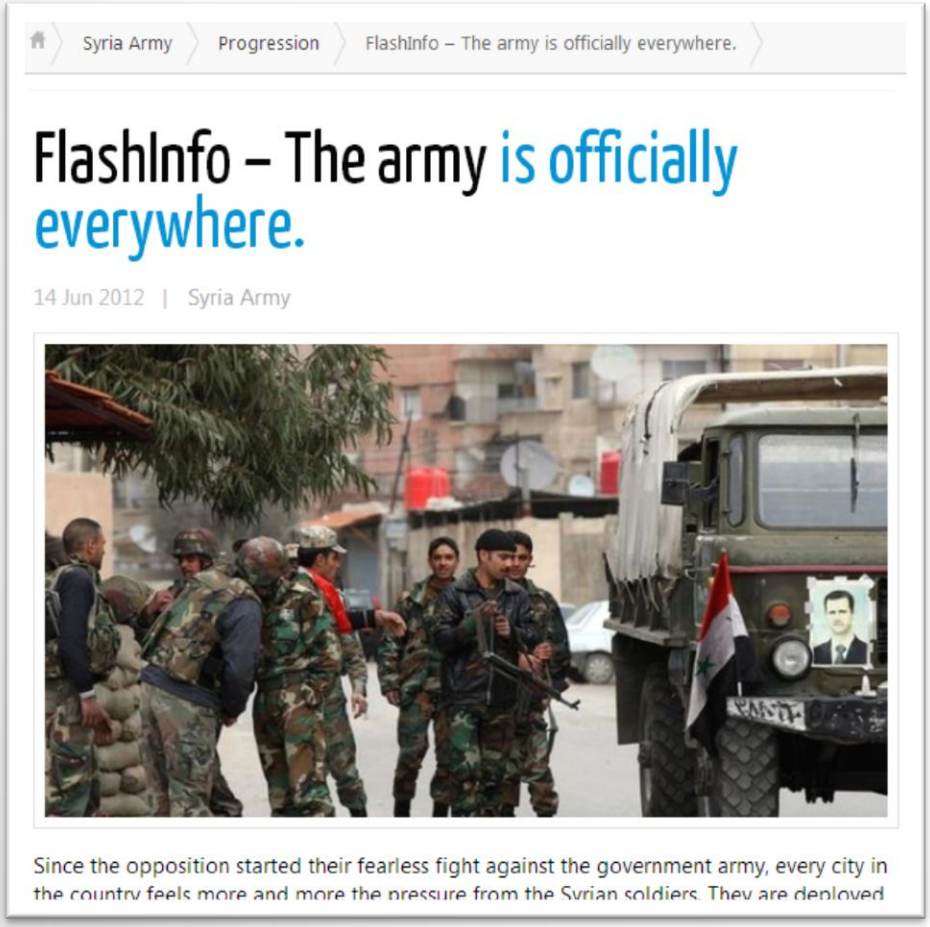
Content

1. Introduction
2. What are EXIF data ?
3. Overview of the problem
4. The general approach
   - Analyzing the EXIF data
   - What about the contextual mismatches ?
5. Results
6. Back to our story
7. Conclusion

M.A. Aissati
Contact: maa48@kent.ac.uk

Introduction

- Project inspired from recent events : Civil war in Syria



Fictional article, the screenshot above is not the original post

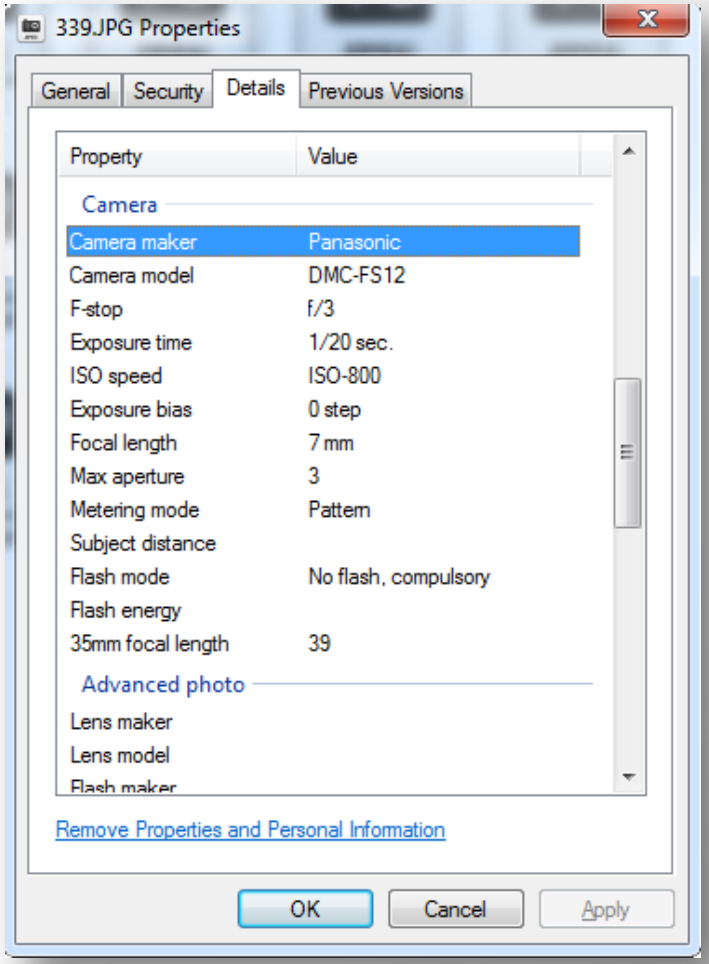
- Story of a Syrian **blogger** who wrote:
  "(…) and here is a **picture** of the opposition army I took with my **iPhone** directly **from my window**. The whole neighbourhood is officially dead. […]"

What is potentially wrong with that ?

M.A. Aissati
Contact: maa48@kent.ac.uk

What are EXIF data?

- **Exchange Image File format**
- Specification for file format used by digital cameras
- Add metadata **tags** to the file to have additional information:
  - Constructor
  - Model
  - Date and time
  - Flash
  - …
- JPEG, TIFF, and RIFF (wave)



M.A. Aissati
Contact: maa48@kent.ac.uk

Overview of the problem

- People are more and more concerned about their privacy
- How EXIF data can reveal private information ?
- Are these data in a correct context ?
- A **proof of concept** and a **statistical study**
- One of the main and private information in EXIF tags are the geolocation data (**GPS**)

M.A. Aissati
Contact: maa48@kent.ac.uk

The general approach

The Spider

The spider will act as a Web Crawler.

It will only aim blogs and personal websites.

It will run for 1 – 2 weeks to collect at least 1,000 domains.

Process Images

Two stages:

1- Extract GPS data

2- Analyse context for any incoherence (with support of different languages)

To Database

Build an output database with the results for each domain

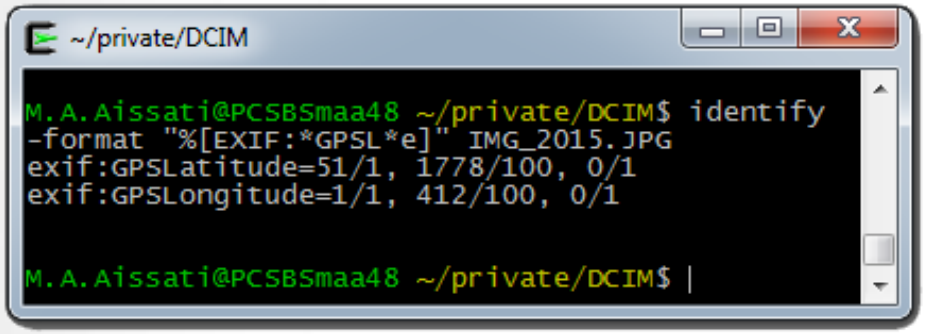
Technology:

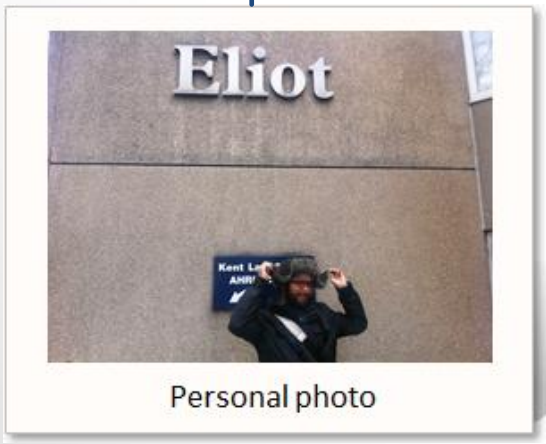
- Spider – Main code in Python using Scrapy Framework
- Output databases: MangoDB (noSql), Mysql, CSV, XML
- Use of Google Search API

M.A. Aissati
Contact: maa48@kent.ac.uk

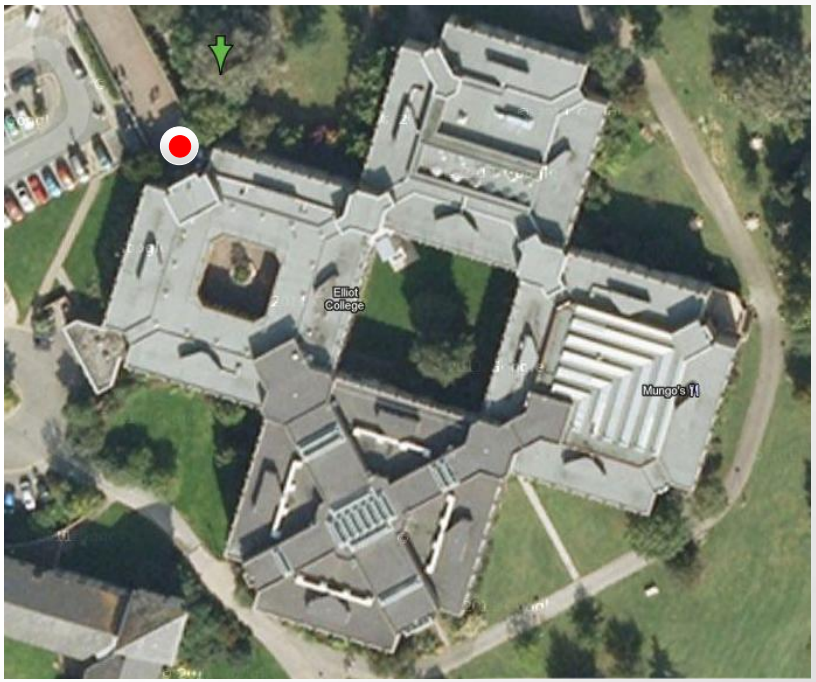
Analysing EXIF data

- GPS tags in EXIF: **GPSLatitude**, **GPSLongitude**





	Float Value
GPSLatitude	51.29633333333333
GPSLongitude	1.068666666666666



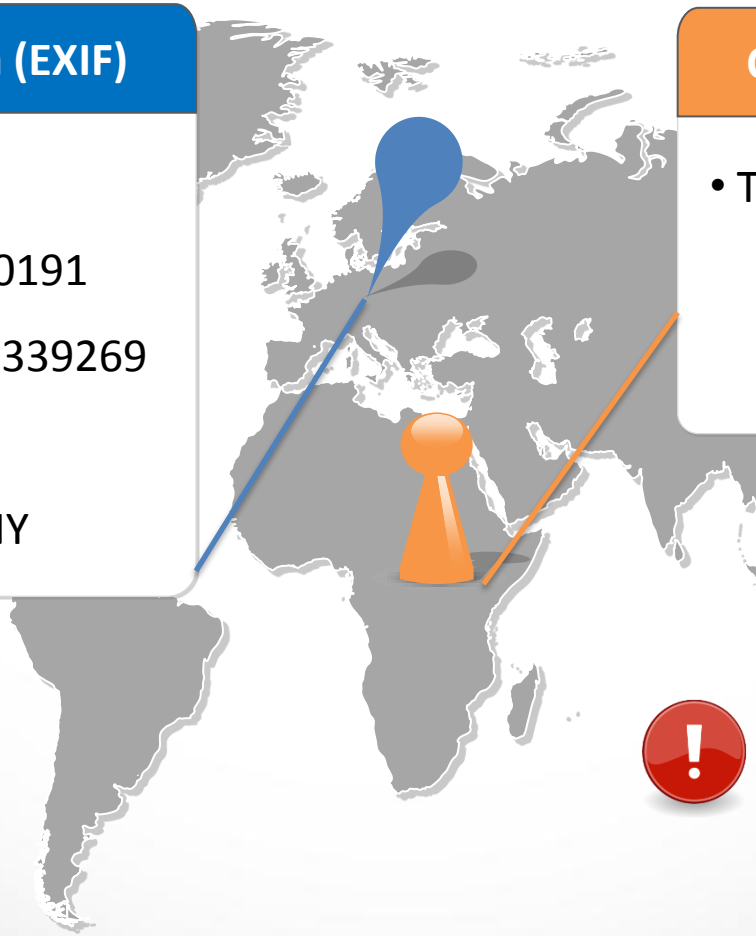
M.A. Aissati
Contact: maa48@kent.ac.uk

What about the contextual mismatches ?

- Imagine someone claims on his blog: “This is a picture of an elephant during my trip to **Kenya** !”

PICTURE GPS data (EXIF)

- GPS coordinates:
  - Latitude: 52.50191
  - Longitude: 13.339269
- City: Berlin
- Country: GERMANY



Context (Name analysis)

- Token “Kenya”
  - Latitude: 0.505365
  - Longitude: 37.968750

Possible mismatch !

- Different countries
- Distance: >10 000 km

M.A. Aissati
Contact: maa48@kent.ac.uk

Results

- Automatic Post-process Databases Analysis
  - Percentage of web domains with EXIF content embedded
  - Percentage of web domains that have GPS information stored in their pictures
  - Percentage of contextual mismatches (alerts, false alarms, ambiguous alerts)

- **Human analysis** for sensitive disclosures and privacy issues

M.A. Aissati
Contact: maa48@kent.ac.uk

Back to our story

- The syrian blogger case
  - **iPhone** : Geo location is probably enabled
  - **From my window** : We assumed the picture was taken from its home

- Analysis of the picture

  POSITIVE ! The picture has GPS data in it.

- This might be compromising for the author.

  He could be arrested by the government for being part of the Opposition Movement .

M.A. Aissati
Contact: maa48@kent.ac.uk

Conclusion

- The project itself:
  - It tempts to be as **abstract** as possible for future studies (modular implementation)
  - Explosion of smartphone users with cameras and GPS chips
  - Hope to make people more aware of the picture they put online
  - Delete EXIF data from files is trivial

- Related work:
  - Steven J. Murdoch & Maximillian Dornseif (2004), "**Hidden Data in Internet Published Documents**", Chaos Communication Congress 2004
  - Arai, I., Fujikawa, K., & Sunahara, H. (2008). "**Proposal of time-crawler which collects an event time by reading exif data in blogs**". *2008 Annual IEEE Student Paper Conference*.

Special thanks to **Dr Julio Hernandez-Castro** for supervising me in this project !

M.A. Aissati
Contact: maa48@kent.ac.uk

THANK YOU!

M.A. Aissati
Contact: maa48@kent.ac.uk