TENSOR: a solution to dark web investigations

DAY, Tony <http://orcid.org/0000-0002-3214-6667>, DENNIS, Kieran and GIBSON, Helen <http://orcid.org/0000-0002-5242-0950>

Chapter 05. TENSOR: A solution to Dark Web investigations

Tony Day, Sheffield Hallam University, t.day@shu.ac.uk
Kieran Dennis, Sheffield Hallam University, k.dennis@shu.ac.uk
*Helen Gibson, Sheffield Hallam University, h.gibson@shu.ac.uk

Keywords: TENSOR, Social Media, Content, Dark Web, Law Enforcement Agencies

ABSTRACT

A wide range of criminal activities, particularly those related to hate crime, radicalisation, and terrorism, have an ever-increasing online component that needs to be considered, captured and investigated. Activity on the Dark Web is becoming more commonplace and can be a significant factor in investigations. Such information may need to be coupled with other open source information from the surface web and social media. Due to this, law enforcement is undergoing radical changes in their investigatory processes requiring new digital systems; however, aspects of the Dark Web still remain 'off-limits' to investigators as they battle technical, regulatory and organisational challenges. TENSOR is conceived as a system that can address many of these challenges by providing a mechanism for investigators to capture intelligence from online sources including the Dark Web and convert it into structured and coherent data. Within the system, TENSOR implements a unique data model, acquisition and extraction process that enables information to become searchable, analyse-able and relatable. Several analytical capabilities are layered above delivering advanced functionalities including natural language processing and concept extraction, machine translation, image recognition and iconography detection, social network analysis, stylometry, content similarity recommendations and image forensics. Each capability is presented through a unified intelligence dashboard that provides a clear entry point and logical analysis pathway for investigation to ensure TENSOR delivers a system that enables law enforcement to access and analyse complex information from the surface, deep and Dark Web.

1. Introduction

A wide range of criminal activities, particularly those related to hate crime, radicalisation, and terrorism, have an ever-increasing online component that needs to be considered, captured and investigated. Some of this crime is occurring in dark, hidden corners of the internet or is conducted in encrypted private channels, where there is little that can be done to access it in real-time. However, as with many types of crime, there is often a visible surface layer available to investigators that can provide vital intelligence. For example, a private group containing extremists can remain closed sharing their hate speech and radical ideals amongst themselves; but, many radical individuals want to spread their ideals through propaganda or even discuss or inspire others to carry out attacks and illegal activities. This is where such activity may manifest itself in the open, whether on message boards, forums, web sites or social media platforms. Any and all of these methods of communicating and sharing information can be found across the surface, deep and Dark Web, often available to all.

The effect of this sea change in criminal behaviour means law enforcement is having to go through radical changes in their investigatory processes to respond to this challenge and will continue to do so for the foreseeable future. Digital systems to support criminal investigations are becoming commonplace; however, aspects of the Dark Web still remain 'off-limits' to investigators as they battle technical, regulatory and organisational challenges.

TENSOR is conceived as a system (see Akhgar *et al*. (2017) for an initial introduction to the vision for TENSOR) that is able to address many of these challenges by providing a mechanism from which investigators can capture intelligence from online sources including the Dark Web and convert it into structured data with a coherent data model (Section 3). The benefits of this structure are the ease with which the information within TENSOR becomes searchable, analyse-able (both through automated and investigator-led means) and relatable (links between multiple pieces of information are immediately evident). The purpose of this chapter is to walk-through the rationale, key functionalities and potential use cases of TENSOR to distinguish it from existing OSINT or Dark Web investigation platforms while demonstrating the added benefit TENSOR provides to the investigator.

The chapter proceeds as follows, firstly we illustrate the underpinning data model for TENSOR and combine this with a discussion of potential data sources, acquisition and extraction processes, secondly we review the core functionality of the underlying system for analysis, thirdly we cover the user interaction with TENSOR and how analysis is manifested in the intelligence dashboard interface, fourthly we review initial user feedback to motivate future work through a set of lessons learned.

2. The role of TENSOR

TENSOR (reTriEval and aNalysis of heterogeneouS online content for terrOrist activity Recognition) focuses enabling investigators to research and analyse content on publicly accessible virtual spaces where modern technologies are being exploited for nefarious and illicit causes. These virtual spaces occur on social media platforms, the surface web, or on hidden marketplaces or forums on the Dark Web. They represent anywhere that an individual can access relatively easily with simple technologies such as a web browser. Currently, many law enforcement agencies (LEAs) are bound to archaic approaches of searching, extracting and analysing this unstructured online content (Deloitte, 2015), such as through screenshots, which are only easily interpretable by humans and have severe limitations (e.g., see Feldman (2015)).

Recently, LEAs have had access to technologies that better support their capture of data from the web although such methods are still fraught with difficulties. First and foremost, is the fact that no matter how advanced the technologies LEAs have access to the chances are criminals are already two steps ahead making use of new apps, communication channels and the ease with which they can switch aliases. Thus, law enforcement is always playing catch-up. Furthermore, the very tools that are also supposed to help them capture data also drown them in data with extensive configuration required to limit over-collection and a lack of analysis capability. This is often compounded by a misunderstanding of the role of analysts and a lack of training (Belur and Johnson, 2018).Access to social media services for LEAs can also be switched off at a moment's notice which does not help when, after a crime, there are questions from the public and media if a perpetrators intentions were there for all to view. Finally, most tools are not silver bullets and with limited budgets LEAs can only invest in licencing and training for the most broadly used.

TENSOR is a collaborative project between European partners funded by the European Commission's Horizon 2020 Research and Innovation Programme. It aims to extend investigators' capabilities by replacing much of the repetitive and manual work required to understand, acquire, extract and analyse data from the ever-increasing range of open spaces on the web. TENSOR is particularly focused on addressing the proliferation of terrorist content posted online; however, many of the technologies developed within the project are applicable across a wide range of domains. TENSOR brings together a wide range of technical capabilities covering content acquisition and extraction, analysis and visualisation, and intelligence management into a single space known as the TENSOR Intelligence Dashboard.

Using TENSOR, open-source law enforcement investigations can be both speeded and scaled up. Scaling up allows investigators to focus less on structuring and organising content and much more on identifying important patterns and connections between individuals, groups, and the content they produce, share, or interact with. Using a customised data model focused on connectivity, the investigator is able to work with heterogeneous data from the surface web, Dark Web, and social media services whilst only dealing with familiar concepts such as posts, profiles, pages, messages, and hashtags. Taking advantage of a powerful content extraction capability, this content is indexed and categorised efficiently to allow searchability and analysis that simply cannot happen with the status quo of screenshots in Word documents. Using cross-platform fusion, TENSOR simplifies the identification of actors communicating over various systems and platforms with various aliases alongside the content they produce.

Crucially, TENSOR has been informed by law enforcement end users at every stage, taking advantage of the highly collaborative structure of a project between pan-European partners. Not only has this given TENSOR a clear strategic direction, it has also meant ongoing feedback from law enforcement end users through three pilot and evaluation phases over the three years of the project has been directly incorporated during development. With such close participation between legal experts, law enforcement, academia, and industry, TENSOR has utilised privacy-by-design (Langheinrich, 2001) at a time when there has been a major shift in the way data protection is perceived and managed due to the GDPR. TENSOR has had to be particularly mindful of these issues given its capacity to increase the scale and velocity of data collection and analysis of potentially sensitive information. Efforts have been made throughout the development to assure that any potential impacts on personal privacy have been outlined and mitigated, from the design, research and implementation perspectives using both law enforcement focused legal and ethical controls, but also through good security practices within development.

Considering how TENSOR aims to expedite typically menial content acquisition tasks so that highly skilled investigators can focus on their investigations, special emphasis has also been placed on protecting the chain of evidence. As a complex part of policing, it is difficult to replicate procedures for storing of physical evidence with digital evidence, especially that which has been obtained online (given that some standardisation already existing for typical digital forensics tasks relating to content acquisition from hardware). A common solution is digital hashing, employed regularly by law enforcement (Giova, 2011), which helps to maintain the line of authenticity from evidence collection onwards. TENSOR introduces a further security measure of digital signatures integrated into comprehensive auditing system that tracks the introduction and subsequent modifications to each individual piece of content or entity. Protecting the integrity of the content with digital signatures can precisely record when and how the content was acquired, and, crucially, cannot be modified without access to the secret key that created it.

3. Modelling highly connected content

TENSOR has been designed around an abstract data model that emphasises the connectivity between content and the 'things' within it that replicates and amplifies the natural connections made by human investigators. This model is reflected in TENSOR's underlying secure storage repository - the hub of the system. At its core this data model is effectively a triplestore containing relationships structured in subject-predicate-object format. For example, consider the following triple: "Actor A liked Post B". In this triple, Actor A is the subject, 'liked' is the predicate and 'Post B' is an object. The subject and the object both form entities within the system while the predicate represents links. Triples are always directed, and the link can take many forms including "follows", "author of", "shared" and many more.

Triplestores are often represented as a graph (or network) where objects are vertices (nodes), and predicates are edges (links). As predicates denote one-way relationships, a triplestore can be represented as a directed graph. This allows for both a representation of these links in a graph or network visualisation and for processing and exploration of the data, by 'walking' or 'traversing' the graph alongside the application of methods from both graph theory and social network analysis (SNA). For example, to find all friends of friends for a single actor (two steps away from our initial actor) would involve starting at the object representation of that actor, a node in the graph, and then for each of the nodes linked via a predicate of 'friend of', find all the nodes linked from them via the same predicate. These types of queries can be far more complex, and as a result far more powerful, allowing for high-level analysis of the structure of the network.

3.1 Artefacts, entities and links – the TENSOR data model

Building on the idea of triplestores and entity-relationship models, the TENSOR data model is divided into three parts: artefacts, entities, and links (

Figure **1**).

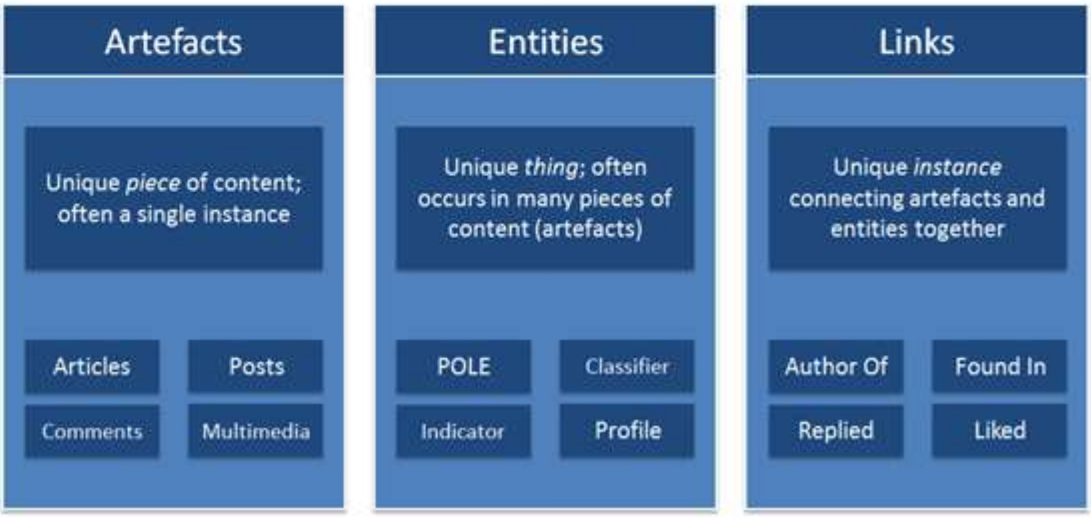
Artefacts
Unique *piece* of content; often a single instance
Articles
Posts
Comments
Multimedia
Entities
Unique *thing*; often occurs in many pieces of content (artefacts)
POLE
Classifier
Indicator
Profile
Links
Unique *instance* connecting artefacts and entities together
Author Of
Found In
Replied
Liked

Figure 1: Abstract types of content in the TENSOR data model

Artefacts represent a unique piece of content ingested by TENSOR's content acquisition functionalities. Each artefact is assigned a unique reference aligned to its source and a content hash is created. For text, artefacts include news articles, social media posts, status updates (e.g., tweets), comments, replies, or messages between individuals. As for multimedia, this includes audio, images and videos but also documents, other files and binary data.

Entities on the other hand are unique '*things'* within an artefact's content or about the content (e.g., a classification or category) as well as things like social media profiles. Within a piece of content, an entity is then an attributable thing which us extractable such as a location, person, group, organisation, date, time, URLS, other social media profiles, domain names or web sites. A common way to think about entities is using the POLE acronym, often referred to by law enforcement (College of Policing, 2019), which standards for "People, Objects, Locations, Events". In short, entities are things that link to content, that are not themselves content.

Finally, between artefacts or entities are the interactions or relationships between them, these are called *'links'*. Links are captured by extracting observed entities mentioned within artefacts or by observing the relationships between artefacts and entities. For example, a social media profile entity is linked to a social media post artefact with the link “author of” if they have written the post. The extraction of links may seem trivial, but it results in a rich data set with extensive opportunities for querying data. For example, links capture huge networks of entities based on a wide variety of relationship types such as profile mentions which can then be exploited by social network analysis to discover communities of social media profiles based around interactions, rather than simple friend/follower relationships.

While the data model may be clear to those with technical background, expecting end users to become experts in understanding and appreciating abstract concepts such as artefacts, entities and links could introduce a steep learning curve, prove a barrier to long term adoption or simply introduce unnecessary confusion during analysis. In the TENSOR Intelligence Dashboard, the main graphical user interface of the system, these abstract notions are hidden and instead are referred to through the natural language of the domain, e.g., posts, messages, pages, URLs, tags, profiles, persons and groups; each with their own specialised interface and representation to help extract the most meaningful knowledge for the user. Such naming conventions are also retained when displaying the graph to the user. The example in Figure 2 shows a simplified version of the complex relationships between only two artefacts and three entities. In reality however, there are vastly more links extracted from even the simplest of artefacts.

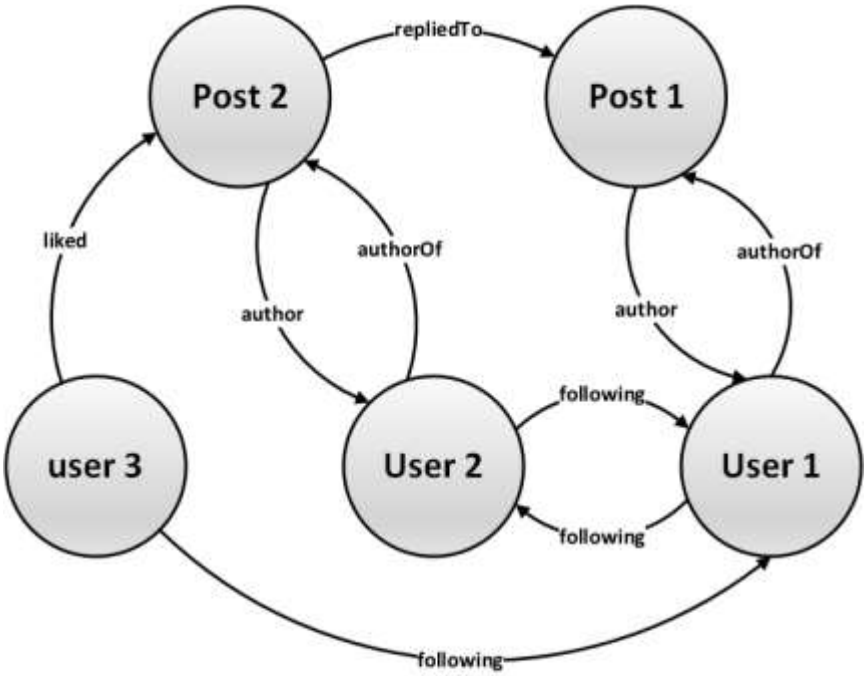

Figure 2: An extract of a set of potential relationships between two posts and their authors

4. Acquiring content

Illegal and harmful content exists everywhere on the Internet, from the open web to social media platforms and forums to the Dark Web. To access such content and assess it for intelligence or evidential value may require an investigator or analyst to manually trawl through site after site, page after page, using their instinct and experience to follow potential leads. Such leads may be profiles representing known individuals or groups, content such as illegal propaganda, or damaging violent images or videos. Not only is this extremely time consuming, but it is also difficult to appreciate the extent of the content the investigator encounters and the vast number of potentially interesting and relevant connections between content, authors, profiles, and other associations. These associations include followers, likes and members, but also a deep network of associations within the content including an image's content, offensive and illegal references and rhetoric, and who is sharing, supporting or being mentioned within it. Therefore, investigators must have access to methods which support easier acquisition of such content.

TENSOR views there as being two overarching goals for the use of the system and, depending on the action an investigator wishes to take, affects how and why content is acquired for the system. Firstly, partially reactive investigations follow a deep and narrow approach to acquiring content in that it will be focused on specific profiles representing individuals (suspects, victims) or groups (organizations). Generally, the acquisition will be interested in the individuals or groups themselves, the content being produced by these sources and any associations shared between them. These shared associates may then become a central part of the ongoing investigation and acquisition. Timeliness and scope of the content is likely to cover a more specific period, the investigation itself may only last weeks or months, whereas the content acquired (i.e., posts) may be part of a narrow window, for example one week in January.

Alternatively, intelligence gathering (or proactive investigations) can be thought of as shallow and wide where larger numbers of profiles may be monitored for extended periods of time, but with less emphasis on their associates in particular. There is expected to be more emphasis on the content itself - profiles may even be generally redacted to avoid collateral collection of those on the fringes. This level of monitoring is likely to observe lots of collateral data that the investigator may never need to look at or analyse. What is more pertinent is general trends and patterns within the content over a much longer time scale, months or years.

5. Understanding content

Imagine investigating a dark marketplace and finding a seller relevant to an ongoing investigation. By capturing a screenshot of their illicit product list alongside their profile name, image and a bitcoin address the 'content' is available for other investigators to see.

This content of the screenshot makes sense to a lone investigator, they can extract the profile image, could quickly recognise if they had seen the image before, assess whether it is a generic image or perhaps specific to that user, read the bitcoin address and recall other sellers of the same products. Even subconsciously categorising the seller based on the spectrum of their products is relatively easy, subtle and instinctive, and happens without effort.

Computers struggle to match humans at this capability. What they are excellent at is processing data and making calculations, much faster than any human can. Even so, there have been many advances in the way computers can automatically interpret natural language in text and objects and scenes in multimedia. These capabilities that allow a computer system to categorise content and extract meaning from it allows investigators to take advantage of the computer's powerful abilities of organizing and sorting data. Where previously an investigator would have to categorise and organise their own screenshots through post-it notes, files, and index cards, with modern natural language processing (NLP) and image recognition much of this can be automated, accelerating the scale at which information can be interpreted and an investigation can operate.

NLP and image recognition are two crucial parts of the TENSOR system as they allow content to be categorised and, therefore, organised conceptually and thematically. The first of these methods allows text to be extracted and interpreted from all types of content. The POLE acronym is an excellent starting point to understand this, it stands for "People, Objects, Locations, and Events" and is often used by LEAs to model interacting elements of an investigation (College of Policing 2019).

Beginning with "people", the investigator is interested in those real-life subjects who may be a suspect, victim, witnesses, by-stander, foot-soldier or a leader of a group. Online, such roles will be taken by authors, sharers or propagators, or even the observers, whether interacting, "liking" or "up-voting" content, or more indirectly through their browser history or router logs. Admittedly, the last points are leaving the realms of open source data, but nonetheless they reveal how open source data can complement closed source data for corroboration. People are also found within content through mentions, direct and indirect references, and response. NLP can support detecting and identifying names or aliases (natural or pseudonyms) their extraction and how they are attributed across content.

"Objects" are then any tangible or intangible *thing* that can be uniquely attributed or attributed as a group of potentially unique *things*. Examples include products (cars and their models, brands, weapons), materials, media (texts, quotes, music, video), concepts (emotions, desires, beliefs, thoughts), and many others extractable *things*. "Locations" categorise named places, meeting places, venues, parks, cities, counties, and countries, whether real or fictional, globally understood (the Statue of Liberty) or only locally by a community (e.g., slang or colloquialisms). Extraction of temporal information helps with extracting "events" and allows the system to learn about things that have happened or may in the future, it is the foundation for plotting the content and *things* within it on a timeline.

Particularly relevant to TENSOR and terrorist content on the Dark Web is the detection of events and their associated timings which could indicate a forthcoming attack.

A key feature of terrorism related content is that it is often talked about and shared in languages other than English. TENSOR currently supports translation from Arabic, French and Turkish to English directly within the system whilst Spanish and German are supported as "core" languages and do not require translating to be processed. The translation components make use of number of core frameworks and services including the statistical machine translation offering Moses (Koehn *et al*., 2007) and a more modern neural machine translation service ModernMT (2019).

Recognising media, particularly images, uses a similar approach to extraction and attribution of textual content, but is instead based on the content within the image or the image as a whole. Building on widely available image recognition models, as with NLP, TENSOR is able to exploit this vast knowledge base to categorise images into many potential groups. These models have been trained using machine learning techniques by attributing existing images, e.g., pictures of hands, landscapes, people, vehicles, and all manner of other *things* in the world, to the concepts they represent. For example, there are thousands of pictures containing a hand that have been labelled with the concept "hand". The machine learning algorithm has used this information to produce a model that, given a new unprocessed image, is able to detect whether or not there *may* be a hand in it. The same model has also been trained with images containing thousands of other concepts, including many that could help the Dark Web investigator.

Using content acquired via TENSOR from the surface web, Dark Web, and some social media platforms, the extraction of concepts and themes from text and images provide the entities and links that can be linked to the body of content. It is these entities and links that provides the data needed by many of TENSOR's analytical capabilities. The richness of these groups and connections between data make the TENSOR Intelligence Dashboard more powerful.

To situate these extraction capabilities, take *Alice* who is operating an illicit store front, focused on selling antique firearms favoured by gangs, on a number of dark markets. She advertises her products for short windows of time and particularly during weekends believing this will keep her off law enforcement's radar. An investigator, concerned with the illegal firearms trade, wants to begin a proactive investigation into firearms sellers on hidden dark markets. This investigator does not have any specific sellers in mind but manually scopes out the URLs of the sites they want to target. Using TENSOR, they enter the relevant URLs and begin crawls to monitor products being posted scheduling re-crawls at an hourly basis. Returning to the system after several days, they can search for images identified as containing guns or ammunition as well as textual mentions of these same concepts, given they have been extracted as entities using TENSOR's capabilities. Navigating the acquired and processed content enables quick identification of pseudonymised profiles on the marketplaces which then may warrant further investigation based on how prolific they are or their connectedness. Furthermore, all mentioned profiles are stored alongside other extracted information such as Bitcoin wallets, specific terminology or keywords, or products which may be vital later in the investigation.

6. Analysing content

Obtaining significant amounts of content has become easier with automated content acquisition capabilities; however, this has led to a major growth in the challenges of handling that data. Historically, systems relied on manually trawling, categorising, and indexing all investigation material but the extent of digital content available through both offline and online sources make analysing such information without automated means near impossible.

Fortunately, automated methods to extract and categorise content are becoming increasingly accurate, but this still leaves the problem of how to find the proverbial needles in the haystack? These needles may be specific pieces of content that are critical to an investigation, perhaps a hint to a potential attack, but they may well be hidden in the complex web of interactions. In TENSOR, the investigator can call on advanced automated analyses which can detect textual and visual patterns, group and organise content, find paths between entities and their content, and uncover organisational structures.

7. Beyond TENSOR and Lessons learned

TENSOR as a system does not standalone and many other systems with similar and complementary capabilities exist. TENSOR is not supposed to be a silver bullet designed to replace all existing systems. Therefore, it is crucial that TENSOR provides ways for investigators to provide ways of seeding and enriching its initial content with other sources they have access to, and in common formats such as CSV (comma-separated value), always allowing the system's potential to be exploited by law enforcement.

TENSOR must also "close the loop" for investigators by supporting and exploiting the enrichment and analysis performed within the system and allowing investigations to continue forwards beyond TENSOR. This is imperative when considering how connections in the discovered content can lead to new relevant content and also new potential leads in the form of profiles, events, and analyses. At the raw-level, content can simply be exported, again in the common CSV format, allowing further exploration in tools the investigator is familiar with such as Microsoft Excel, or using IBM's I2 Analyst's Notebook to further develop elements of the investigation to present in court for example. On top of this, many of TENSOR's visualisations or analytical outputs may be captured or exported to support these purposes including the investigators notes, the commonalities, and visualisations.

Connecting the inputs and outputs of TENSOR with those resources and tools available to investigators at the beginning and the end of an investigation or intelligence gathering exercise demonstrates just how TENSOR can deliver value and save time.

Throughout the development of TENSOR opportunities for new knowledge and insights have been sought not only from a technical perspective, but also directly from the terrorism domain and across the legal, ethics, data protection and other areas. The emphasis here will be more on the technical side, but these are by no means the only important takeaways for this ambitious and challenging project.

First and foremost, has been gaining a deeper understanding of the challenges facing any kind of automated deep web acquisition. The generally accepted model, including the one used in this book, is that of an iceberg with three layers: the surface web, the deep web, and the Dark Web. Whilst this is a reliable model for thinking about the layers of the Web, it is also useful to think about the Dark Web as a more logical layer which can have its own surface and deep web within it. These surface and deep webs may also occur across many darknets as well. Darknets being the additional layers of the Web that are hidden across different protocols, such as Tor's hidden services, I2P, or Freenet. The first lesson here is on the assumption that because a human being may find it relatively straight-forward to access these services in the deep web, this is not the same for a system such as TENSOR. Captchas are a great example; these are the fuzzy texts requesting the user to enter them correctly to produce they are a human. People generally do not struggle with them, computers on the other hand do - the actual point of having a Captcha. These Captchas only exist because they are effective, and many deep web sites and services use them for this reason. Secondly, accessing these services as a human being has a very particular usage profile that is difficult for a computer to replicate. If the acquisition task was to be automated in the same way a human uses the service and via a web browser, it is likely that the pattern of usage would differ in very detectable ways from a human user.

As a result of this challenge, the more common route when it comes deep web sources such as social media platforms is to utilise their existing APIs. This is also based on the assumption that these social media platforms all have APIs, when in fact the majority of them do not. Normally, they are only provided by larger platforms and these are quickly becoming more difficult to use or being ruled out based on their terms of service. For example, many services now declare that law enforcement are not to use the APIs or that they cannot be used for any activities resembling surveillance, i.e., the continued monitoring of an account. Additionally, APIs are also a moving target as platforms continue to develop. APIs also create a development overhead as the more services are integrated the more integrations that need to be maintained.

This leads onto the next emerging challenge with the deep web: the scale of growth in the app-only market and other sources that are challenging to access. Services that operate solely through mobile apps are plentiful and continue to grow in popularity, these include popular services such as WhatsApp and Snapchat, and although this is changing - for example WhatsApp can be accessed through a web browser, access remains limited. Furthermore, with each service operating in a different way it is difficult for investigators to keep track of where illegal actively may be being conducted regardless of whether they can access such content. Another known issue of access to a deep web source is that of in-game communication systems that are increasingly popular in modern advanced gaming but are extremely difficult to extract any information from automatically. Ultimately, better cooperation between service providers and law enforcement, built on a sense of trust from both sides, will allow capabilities such as TENSOR to be utilised safely and fairly by law enforcement for the purposes of protecting the public.

In software development terms, all systems, especially those which have been co-developed by academia and industry, have to find a balance between performance that can be achieved in a laboratory environment and state of the art implementation of algorithms versus the requirement for future operation deployment on systems which may have limited hardware capacity. However, as Goble (2014) notes, developing better research software has a wide range of benefits to both the researchers and science as a whole. TENSOR has also experienced that what works in the lab does not always experience the same high-level performance when tested by real-users or that particular challenges may arise when an investigator needs to re-run an analysis but also maintain the chain of evidence for their existing analyses. Additionally, even only when simulating operational use does it become clear the true extent of the volume of data law enforcement are encountering when they need to investigate openly accessible data sources. Nonetheless, it is also when such volumes of data confront the developers does the requirement to pull from advanced research techniques become even more apparent.

Furthermore, a key lesson from the project has been to realise, often, how seemingly simple functionalities that are neither difficult to imagine nor implement are missing from existing analytical software used by law enforcement. Occasionally, it is these solutions that provide the most value to users, such as the commonalities view in TENSOR, and can actually increase adoption and uptake of the system where the advanced functionalities can come into play.

8. Concluding remarks

This chapter has set down the power of TENSOR as an internet intelligence and investigation tool. It has demonstrated the layered capabilities of TENSOR that is able to collectively deliver a coherent and widely applicable data model that generically supports a wide range of web, social media, forum and marketplace sites across the surface, deep and Dark Web.

On the top layer, TENSOR supports a single central Intelligence Dashboard through which all content can be accessed whilst hiding the complexity of the processing underneath and focuses on delivering value and valuable insights. In the future, some interfaces will support direct access to the specific functionalities of certain components of the system - for example, an investigator may have an image they need to upload to check for evidence of tampering but does not need to acquire any additional data around that image at that point in time.

The development of TENSOR has also demonstrated the ways in which even software still under development can enhance investigatory approaches, especially if the overall system is not intended to be a silver bullet. While some modules offer an experimental approach, taking advantage of the latest research and algorithms, other components, such as the commonalities interface can deliver immediate results. Other aspects are continually improving, specifically; advances in NLP, multimedia extraction and computer vision are almost continuous and can often be easily integrated.

As with any big data system, you cannot avoid capturing data that ultimately turns out to be noise. This is particularly true for the processing steps after content acquisition where false positives of concepts may be extracted from text or objects from images. However, corroboration between data, using techniques such as FCA, can help reduce these false positives by linking content based on similarities and ignoring or diminishing the importance of the outliers.

Ultimately, the goal of the TENSOR system is to ensure that law enforcement can access and analyse complex information from the surface, deep and Dark Web whilst supporting advanced analytical capabilities that deliver real operational value to the user. The expert decision making is left to the human-in-the-loop while the monotonous and data heavy processing are performed by the system. Symbiotically, TENSOR is able to deliver extensive intelligence generated by users without the need for extensive technical knowledge that restricts adoption.

9. References

Akhgar, B., Bertrand, P., Chananouli, C., Day, T., Gibson, H., Kavallieros, D., Kompastsiaris, I., Kyriakou, E., Leventakis, G., Lissaris, E., Mille, S., Tsikrika, T., Vrochidis, S. and Williamson, U. (2017). TENSOR: Retrieval and analysis of heterogeneous online content for terrorist activity recognition. Proceedings Estonian Academy of Security Sciences, 16: From Research to Security Union, 16, 33-82.

Belur, J., and Johnson, S. (2018). Is crime analysis at the heart of policing practice? A case study. Policing and society, 28(7), 768-786.

College of Policing (2019, August) Collection and recording. College of Policing Authorised Professional Practice. Retrieved from https://www.app.college.police.uk/app-content/information-management/management-of-police-information/collection-and-recording/

Deloitte (2015) The Digital Policing Journey: From Concept to Reality - Realising the benefits of transformative technology. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/public-sector/deloitte-uk-ps-digital-police-force.pdf

Feldman, B. (2015, December 11). It Is Incredibly Easy to Fake a Screenshot. Here's How. Intelligencer. Retrieved from http://nymag.com/intelligencer/2015/11/how-to-fake-a-screenshot.html

Giova, G. (2011). Improving chain of custody in forensic investigation of electronic digital systems. International Journal of Computer Science and Network Security, 11(1), 1-9.

Goble, C. (2014). Better software, better research. IEEE Internet Computing, 18(5), 4-8.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Brooke Cowan, Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster sessions (pp. 177-180).

Langheinrich, M. (2001). Privacy by design—principles of privacy-aware ubiquitous systems. In Proceedings of the International Conference on Ubiquitous Computing (pp. 273-291). Springer, Berlin, Heidelberg.

ModernMT (2019) Retrieved from https://www.modernmt.com/